

Information Processing

VOLUME II

The Maximum Entropy Principle

DAVID J. BLOWER

© David J. Blower, *Third Millennium Inferencing*,
Pensacola, Florida, May 2013.

DEDICATED TO MY SON

PATRICK

Preface

My goal is to speculate about the nature of our evolutionary successors. In the Preface to Volume I, they were not given very romantic names. I called them *Advanced Information Processors* (AIPs).

Engaging in such a highly speculative enterprise is fraught with perception through a very dim glass indeed. However shrouded in vagueness our musings must be, it seems in the end, that these advanced IPs must be epistemological creatures.

They will demand of themselves that they reason in a consistent and logical manner. If they can solve a problem through deduction, they will choose that path. Just like us, their primitive forebearers, if the paths to deduction have been closed off, they will have to resort to generalizing Classical Logic in order to make inferences. Probability theory, buttressed by recourse to information theory, will be the foundation for their generalizations, as it was for us.

Volume I was concerned mainly with a rather generic and abstract rationale for probability. I thought it helpful to motivate an examination of the formal rules by borrowing ideas from Boolean Algebra and Classical Logic. Wolfram's elementary cellular automata were introduced next. They were briefly analyzed as about the easiest examples of deductive ontological systems. Despite their outward simplicity, these elementary CA led to non-trivial and non-predictable behavior.

Another of my objectives was to demonstrate that these cellular automata, along with Classical Logic, could be generalized by probability theory. Working in this vein, it seemed appropriate to stress formal manipulation rules. The complementary problem of actually assigning legitimate numerical values to probabilities was given short shrift.

Volume II rectifies that oversight with a detailed discussion of the *Maximum Entropy Principle*. The first few Chapters mark an initial, rather discursive, and essentially non-technical introduction to the Maximum Entropy Principle, henceforth abbreviated to MEP. The more technical details are developed slowly after gaining some familiarity with the important conceptual hallmarks of the MEP.

It seems to me that our evolutionary successors will need to make *inferences* as opposed to *deductions* in order to reconstruct us and the world we lived in. Today, for example, we are taking the very first steps in this journey when we make better and better inferences about, say, what the dinosaurs really looked like, and the kind of world they inhabited.

Implicit in this belief that the AIPs will make inferences as opposed to deductions is that *information* must play a role. If information is to play a role, then *missing information* in the form of *entropy* will play a complementary role.

I hinted at the vital importance of missing information in the last Chapter to Volume I, when we were at the brink of despair over our inability to predict the far future behavior of even a simple elementary cellular automata. It is a curious, but fascinating, question to contemplate: What will the AIPs leave out as missing information when they reconstruct us and our world?

My opinion here differs from Tipler's viewpoint that the computing power in the far future at the Omega Point will be so all-comprehensive that the Omega Point will not only be able to *emulate* us, but every logically conceivable being, together with every logically conceivable universe they lived in! The analogy is very roughly that of the distinction between *inferences* made by statistical thermodynamics (my opinion) versus the *deductions* made through a dynamical simulation of every atom (Tipler's opinion).

With great regret, I now must descend from such lofty heights of speculation, however much fun it might be, and deal with the mundane practicalities of how to implement the principle of maximum entropy. This Volume strives to present a clear, and by that I mean a non-mysterious, explication of the MEP. If this task had already been accomplished by someone else, then there would be no point to my effort here.

*David John Blower
Pensacola, Florida, USA
May 2013*

An Author's Apologia

Why do the words *Information* and *Processing* appear in the titles to these books? Here, in Volume II, we will address the definition of *information*. Volume I was all about how to *process* that information. In a sense, we put the cart before the horse, and started discussing the formal manipulation rules for processing information before we even knew what information was!

Loosely speaking, information is something epistemologically useful created by a conscious entity called an *Information Processor* (IP). The IP inserts this consciously created information into a probability distribution for the sole purpose of forming a tentative state of knowledge. This tentative state of knowledge, or degree of belief, concerns those statements about which the IP wishes to make inferences.

It is very important to emphasize the conceptual distinction between information and data. The information inserted into a probability distribution by the IP is not data! Information is epistemological; data is ontological.

If information is the *positive* something injected by the IP into a probability distribution, then the *negative* something not there because it was left out by whatever information *was* proposed is labeled as *missing information*. This missing information is measured quantitatively by information *entropy*.

The principle underlying a good algorithm for making numerical assignments to probabilities is to *maximize* this quantitative measure of the missing information. At the same time, all of the desired information must be retained. The IP captures the desired information, and excludes all of the unwanted missing information, about the statements in the state space within some probability distribution. The processing of that information is always regulated by appeal to the formal manipulation rules.

The Apologia for Volume I emphasized that it was incumbent upon an author to clearly state his purpose and goals in what he seeks to present to his reader. In that initial Volume, the focal point was on what I termed the first grand concept in probability theory, the formal manipulation rules. Here, in Volume II, we treat the second grand conceptual notion: **How can legitimate numbers be assigned to abstract probabilities from the information provided under some model?**

My purpose in Volume II is to present what I believe is the best way to accomplish a satisfactory resolution to this second grand conceptual notion. I am going to try to clearly explain what the Maximum Entropy Principle (MEP) is all about.

In doing this, I must show you as well an algorithm that relies on the core rationale of the MEP. The finest conceptual principles in the world will not do us any good if not backed up by practical methods. Thus, a programmable algorithm for assigning legitimate numerical values for the probabilities to all the statements in the state space must also be within our grasp.

The core notion is that of *information* inserted into a probability distribution by some model that is itself characterized by a dual set of parameters. If this notion is considered the foreground, then the background is equally essential.

The background notion is that of *missing information*. As already mentioned above, *information entropy* is the way that an IP quantitatively measures how much *missing information* is in some probability distribution. When an IP maximizes the amount of information entropy, it is maximizing the amount of missing information reflected in the numerical assignments to the probabilities for all the statements in the state space. In other words, this procedure reflects something important about the IP's degree of belief in all of the statements in the state space.

It comes as no surprise then that this concept called information takes center stage. Furthermore, its absence is equally important if missing information is such a vital concept in the MEP algorithm. The picture crystallizes only if both the background of missing information, and the foreground of active information are correctly portrayed.

My curious position regarding Jaynes

Practically everything presented here is directly due in some way or another to Edwin T. Jaynes's seminal contributions. Nearly everything of any merit I know about the MEP, I learned from reading Jaynes. I will make very clear the few places where I diverge from Jaynes's position.

At the end of every Chapter, I will include a section called **Connections to the Literature**. More often than not, the floor is given over to Jaynes for the simple reason that he must be credited, first and foremost, with explaining to all of us just why the MEP is so important, and how it works in all of its details.

Now, anyone who has read my works knows that I stand in awe of Professor Jaynes. He cut through the dense fog obscuring most of the core concepts which we will discuss in information processing. He understood better than anyone, I believe, what his predecessors got right, but more crucially, where they mangled things badly. His writings spanning over forty years are a gold mine which I have barely begun to tap. And yet ...

I am forced into the curious, and rather uncomfortable, position of both praising Jaynes and criticizing him at the same time. For example, one of the major influences in both Volume I and this work is Jaynes's *magnum opus*,

Probability Theory: The Logic of Science

Shockingly, this is one of the *worst* books ever written on topics that should have been given the quintessential treatment by the master himself!

It remains to this day a major mystery to me why the man who understood so much of this, and who had it in his power to explain it better than anyone else, managed time and time again to fall short. There will be many occasions in the upcoming discussions where I will point out where I think Jaynes was wrong, misleading, or confusing. And it is necessary to repeat, he seriously misled us, most perplexingly, in those very areas where he practically "owned the territory" in his role as the leading proponent of the MEP.

The most glaring example, discussed extensively here, was his repeated assertion that DATA IS INFORMATION. When information is defined in the MEP algorithm, it becomes clear that DATA IS NOT INFORMATION.

In short, as we saw in Volume I, data are known observations on statements in the state space. Their role in the formal manipulation rules of probability theory is to modify the relative standing of all the models in model space.

As Jaynes himself emphasized over and over again, data and frequency counts are fundamentally ontological in nature. On the other hand, information is fundamentally epistemological in nature. Technically, information is defined as an average of a mapping from each statement in the state space to real numbers. THIS IS NOT THE SAME AS DATA.

That is why information is allowed to make an appearance in our definition of how numerical assignments arise within probability distributions. Probability distributions capture an IP's particular state of knowledge, independent of any data not yet observed, and certainly not yet relevant, about what might eventually be observed. So, at the very heart, there is a clash between core notions!

The data *are* allowed to affect our degree of belief in how true a model may be, but the data *are* never allowed to participate in the original definition of a state of knowledge about statements in the state space. By definition, this kind of establishment of a state of knowledge must occur prior to, and before, any observations. Information must be intimately involved in generating any possible state of knowledge; however, most of the models containing their respective foreground active information and background missing information will never survive the onslaught of the data.

Probability vs. frequency

One of our main objectives here in Volume II is to disambiguate, to whatever extent possible, the massive and pervasive conceptual confusion over frequencies and probabilities. Following Jeffreys, our motto always is: **Frequencies are NOT probabilities and probabilities are NOT frequencies.**

Frequencies are ontological in character; they are actual counts of things that have actually happened. Excluding recording mistakes, frequencies do not change. There is absolutely no need to invoke any concepts like models, entropy, information, or states of knowledge, when counting how many times something happened. To whatever extent the human mind (or generally an information processor) is capable, it *knows* that something has definitely happened. Rather prosaically, we make a mark to record that happening.

Probabilities, on the other hand, are epistemological in character; they reflect a state of knowledge of an information processor conditioned on whatever information has been inserted into a probability distribution under some model. The information processor's main concern is not whether something has happened. Its main interest is always in predicting whether something *will* happen, not in knowing that something *has* happened.

Frequencies based on a correct count *do not change*. On the other hand, the *sine qua non* of probabilities is that they *must change* when different information is incorporated within a model.

These two concepts of frequency and probability are about as orthogonal as one could imagine. And yet, over 400 years of debate has still not settled the issue. I don't expect that my feeble attempts will make much of a dent either. Nonetheless, like Sisyphus rolling the rock up the hill, one must make the effort.

Where do probability distributions come from?

From the moment I first started to become interested in probability, I always wondered about the origin of those myriad distributions with exotic names like the Gaussian, the Chi Square, the Poisson, the Weibull, the Student-*t*, the *F*, the Gumbel, Pearson's system of distributions, and on, and on, and on, seemingly without end. To be a master of probability, it seemed that one would have to be a master of this incredible panoply of probability distributions.

To understand the Gaussian, one must master the Central Limit Theorem, to master the Poisson, one must master queuing theory, to master the *F* distribution one must master the analysis of variance, and on, and on, and on. And from this position it certainly seemed unarguable that mastery of the mathematics of each of these probability distributions would distinguish the *cognoscenti* from the idle dilettantes. No textbook you read tried to dissuade you from this conviction. The situation has hardly changed to the present day.

Well, it turns out that we were all treated to a huge helping of bunkum. Detailed knowledge of all of these named distributions was probability's version of the classic red herring. It led you down a glittering path of no return where at the end of the journey lay the sickening realization that true understanding was hidden away somewhere else.

The Maximum Entropy Principle and how, in applying it, one can derive any probability distribution, was, at least for me, a truly eye-opening experience. Now the mysterious origin of all these distributions and how to generate them at will became a liberating piece of knowledge. And the core idea that any probability distribution was a repository of active information against a background of missing information made the mysteries go poof in the night!

One of my most favorite phrases to describe this situation is that, prior to the MEP, probability distributions were literally the *deus ex machina* of the Greek playwrights. They appeared on the scene, mysteriously descending from the clouds, with neither rhyme nor reason, to solve our inferential problems. “Oh, you need a Pearson Type VII distribution to answer that question! Since you are not an acknowledged master of Pearson Type VII distributions, I’m afraid that you won’t be able to solve that problem and you will be forced to rely on my expertise.”

Since it would take several books on their own to fully exploit the development of deriving prominent probability distributions from the MEP, I merely broach the issue at the end of this Volume with the now classic derivation (at least for MEP proponents) of the Gaussian distribution. This procedure shocked the statistical Establishment as an early example of how to insert active information into a probability distribution against the backdrop of maximum missing information.

To provide my own minor poke in the eye of the Establishment, I also show in some detail how the Cauchy distribution is derived via the MEP. This kind of MEP derivation of a somewhat obscure probability distribution is not new with me, but it is not very well known. It should be broadcast to a wider audience.

The reverberations from these revelations of how numerical assignments to probability distributions actually came about, as opposed to somebody’s mathematical *deus ex machina*, continue to ripple down to the present. Rather amusingly, these ripples continue to provoke some unpleasant reactions from the entrenched interests.

Obviously, I am going to present the rationale for the MEP in a very positive light. Then, surprisingly, I am going to argue in a sense for its dethronement as the primary way of setting up numerical assignments prior to the data. The MEP actually becomes superfluous whenever an integration is going to range over all the possible assignments made by the MEP!

Nonetheless, the MEP approach remains extremely valuable for two reasons. First, it does show us how all of those mysterious probability distributions, which have been the false hallmark of the central essence of probability all these years, can be generated at will from an informational perspective. Secondly, after the data have been processed, there still remains the practical requirement for generating one, or a few, probability distribution(s) best supported by the data.

Interestingly, the MEP provides the IP with some clues as to how this important, post-data, distribution could be interpreted from a causal modeling philosophy. That is, the MEP's technique of model construction lends itself quite readily to including associations among variables thought to be relevant in the inference.

Why include some statistical mechanics?

I have been of two minds with regard to including some material on statistical mechanics. In the end, I decided to go ahead with it after all. I hope I don't regret that decision.

First of all, when one barges in where angels fear to tread, we all know how the world will end up labeling you. Secondly, this is a very well-developed area of physics where subject matter expertise is a highly prized commodity. And I will be the very first to admit that I am definitely not that subject matter expert.

But I couldn't resist the lure of probing Erwin Schrödinger's thought processes as he explained probability, and what we now recognize as a precursor to the MEP in his classic text, **Statistical Thermodynamics**. Here is a notorious example of a renowned thinker who could not grasp the idea that probability was essentially an epistemological concept. Rather, he adhered firmly to grounding probability as essentially a frequency concept. You will see his type of argument repeated unquestioned in (almost?) every statistical mechanics textbook.

I am willing to live with the consequences of following the consistent story that the MEP tells to us about information, entropy, and its implications for physics. What I am certain of is that this attempt at consistency will not please very many people.

Mathematica

In Volume II, we continue our introductory tutorial to *Mathematica*. As in Volume I, most of this is relegated to the various appendices. Although, here in Volume II, there is more *Mathematica* material making an appearance outside of the appendices, and, perforce, within the Chapter text and the Exercises at the end of each Chapter.

The material is obviously slanted towards an exposition on how one actually computes numerical assignments to probabilities with the MEP. This material also advances the complexity of what can be accomplished with *Mathematica* code over and above the elementary introduction in Volume I.

As I emphasized in Volume I, if some concept can't be implemented and then computed within *Mathematica*, then we really don't understand what we are talking about. We see that we are able to implement a version of an MEP algorithm in *Mathematica* that provides us with numerical assignments to probabilities under the information resident in some model.

The power of *Mathematica* to assist in our arguments really starts to show here in Volume II. For example, there are integrations related to Schrödinger's derivations which are easily carried out with *Mathematica*, but would be a considerable burden if we had to do them ourselves by hand.

Style in mathematical derivations

My style in presenting this material has not really changed all that much from Volume I. The essential ideas are developed slowly in a rather discursive manner. I will try my best to show much greater detail than you will find elsewhere. Fully solved numerical examples abound.

Many mathematical texts lose the reader early on because no rationale is given for why such an intense effort should be invested in understanding long and obscure derivations. My hope is that someone reading this text will persevere in untangling *my* long and obscure derivations of the MEP.

The motivation should be an intense desire to understand how we can process information in an optimal manner. As I try to demonstrate in the opening Chapters, the numerical assignments made by the MEP algorithm seem quite reasonable from any perspective.

As a cognitive psychologist, I realize that it is quite a rare event when we frail creatures happen to stumble upon the right way of doing things. We should strive to reward ourselves with some satisfactory level of understanding before we endow our intelligent machines with information processing improvements; improvements that evolution seems to have left up to us to pass on.

The style of my mathematical derivations are long and somewhat demanding of your attention span. However, if pursued with diligence, they *are* easier than going directly to the literature. The hallmark of my derivations, if not evident to you by now, is the rather long, slow, and discursive treatment. It will also become evident that re-visiting old problems is another trademark of mine. For this, I make no apology.

There is a saying I once read where someone was commenting about the difficulty of a particular mathematical treatise. It went something like this. “It is a *little* book that is *little* read because it is a *little* hard.” A beginner should always be wary of a thin book on mathematics. A thick book, on the other hand, holds out some promise for the reader that things may be spelled out in excruciating detail. And excruciating detail is always to be preferred over obtuseness.

In the literature, it is far too frequent that important concepts are presented as bald mathematical products devoid of any preparatory remarks. My job is to cushion the surprised “Huh? Where did that come from?” that such isolated pronouncements inevitably arouse.

It may seem that the discussions are too long winded or involved to warrant your attention, but that is the tradeoff one must make. You can either puzzle over a mathematical statement and accept it on faith, or spend some time and concentration to follow the threads of the argument. I believe the latter to be more enjoyable for author and reader alike.

Please indulge me when I propose this analogy to the uncertainty principle in quantum mechanics where, for example, the more precise your measurement of position, the less precise your measurement of momentum, or vice versa. A similar uncertainty principle seems to operate in learning new ideas.

You can accept strange things on faith and be highly puzzled by them, but in the end you are spared tiresome minutiae. Or, you can reduce your puzzlement by increasing your skepticism, but at the cost of more time engrossed in piddling details.

Now, it is time to begin discussing that second grand conceptual underpinning of inference: How can an IP assign legitimate numerical values to probabilities?

Contents

Preface	i
An Author's Apologia	iii
List of Figures	xvii
List of Tables	xix
17 How To Assign Numerical Values to Probabilities	1
17.1 Introduction	1
17.2 The State Space	2
17.3 Introducing the MEP formula	5
17.4 The Important Lessons	9
17.5 The MEP and Data	10
17.6 Connections to the Literature	12
17.7 Solved Exercises for Chapter Seventeen	18
18 The MEP and Models for Coin Tossing	35
18.1 Introduction	35
18.2 The MEP and the Coin Toss	36
18.3 Definition of Information Entropy	38
18.4 The Number of Constraints	41
18.5 Probability of Future Events	41
18.6 Connections to the Literature	47
18.7 Solved Exercises for Chapter Eighteen	50

19 The MEP and Models for Rolling Dice	59
19.1 Introduction	59
19.2 The Fair Model and Two Competing Models	60
19.3 The MEP Assignments in Rolling the Die	61
19.4 Updating the Three Models	66
19.5 Information Entropy Given the Models	67
19.6 The Missing Information in an Assignment	68
19.7 Constraint Functions and Physical Reasoning	70
19.8 Connections to the Literature	72
19.9 Solved Exercises for Chapter Nineteen	75
20 Information and Data	89
20.1 Introduction	89
20.2 The MEP, Data, and Information	91
20.3 Validating Jaynes's Numerical Results	91
20.4 Where Jaynes Was Wrong	95
20.5 The Probability of the Data	98
20.6 Large N Examples	98
20.7 The Ensuing Confusion	100
20.8 Connections to the Literature	102
20.9 Solved Exercises for Chapter Twenty	108
21 The MEP and the Kangaroos	125
21.1 Introduction	125
21.2 Contingency Tables	126
21.3 The MEP Algorithm and Different Models	127
21.4 Probability for Future Kangaroos	135
21.5 Connections to the Literature	137
21.6 Solved Exercises for Chapter Twenty One	141

22 MEP Models and Correlation	149
22.1 Introduction	149
22.2 Setting up the Problem	150
22.3 Correlations Via MEP Models	154
22.4 The Probability Based on all Models	161
22.5 Connections to the Literature	162
22.6 Solved Exercises for Chapter Twenty Two	165
23 Logistic Regression	177
23.1 Introduction	177
23.2 Setting Up the Problem Conventionally	178
23.3 Setting Up the Problem via Bayes and MEP	181
23.4 Numerical Example	184
23.5 Details of the MEP Models	186
23.6 The Data	191
23.7 Connections to the Literature	194
23.8 Solved Exercises for Chapter Twenty Three	196
24 The Legendre Transformation	203
24.1 Introduction	203
24.2 Jaynes's Proof	204
24.3 A Mathematical Definition	208
24.4 Connections to the Literature	211
24.5 Solved Exercises for Chapter Twenty Four	213
25 Deriving the Maximum Entropy Principle	223
25.1 Introduction	223
25.2 Overview of Constrained Optimization	224
25.3 The MEP and its Origin in Calculus	225
25.4 The Derivation	226
25.5 State Space of $n = 3$ with Two Constraints	232
25.6 Final Formula	242

25.7 Connections to the Literature	243
25.8 Solved Exercises for Chapter Twenty Five	244
26 Statistical Mechanics and the MEP Formula	265
26.1 Introduction	265
26.2 The Transition to Statistical Mechanics	267
26.3 Information and Data “Reconciled”	277
26.4 Interesting MEP Relationships	280
26.5 Connections to the Literature	286
26.6 Solved Exercises for Chapter Twenty Six	290
27 Schrödinger’s Statistical Thermodynamics	311
27.1 Introduction	311
27.2 Schrödinger’s Chapter II	312
27.3 Understanding Schrödinger’s Equations	314
27.4 Schrödinger’s Examples	317
27.5 Thermodynamic Example Revisited	323
27.6 Connections to the Literature	328
27.7 Solved Exercises for Chapter Twenty Seven	339
28 Fisher Information and Relative Entropy	367
28.1 Introduction	367
28.2 Fisher Information	369
28.3 Kullback’s Divergence Measure	374
28.4 Data as Information	382
28.5 Connections to the Literature	383
28.6 Solved Exercises for Chapter Twenty Eight	390
29 Relative Entropy and Correlational Models	403
29.1 Introduction	403
29.2 The Background Concepts	404
29.3 Probability Distributions as Points	405

29.4 Canonical Divergence	409
29.5 The Pythagorean Relationship	411
29.6 Projection of Complicated Models	412
29.7 What About the Data?	416
29.8 Connections to the Literature	419
29.9 Solved Exercises for Chapter Twenty Nine	420
30 The Gaussian Distribution	455
30.1 Introduction	455
30.2 From the Discrete to the Continuous	456
30.3 A Standard MEP Formula	457
30.4 MEP Characterization of a Gaussian	458
30.5 Probability under a Gaussian Model	461
30.6 Relative Entropy Expressions	463
30.7 Connections to the Literature	466
30.8 Solved Exercises for Chapter Thirty	467
31 The Cauchy Distribution	473
31.1 Introduction	473
31.2 MEP Approach to a Cauchy Distribution	475
31.3 The Derivation	476
31.4 A Geometric Motivation	477
31.5 More General Cauchy Distributions	478
31.6 A Rationale from the Geometric Layout	480
31.7 The Lighthouse Example	482
31.8 Conceptual Issues	486
31.9 Connections to the Literature	488
31.10 Solved Exercises for Chapter Thirty One	493
32 Discovering Causal Models with the MEP	509
32.1 Introduction	509
32.2 The Data and the Contingency Table	511

32.3 An Inference about the Next Kangaroo	514
32.4 MEP Models	518
32.5 Search for Good Models	520
32.6 Signal Plus Noise Models	521
32.7 Relative Status of Models	522
32.8 Connections to the Literature	523
32.9 Solved Exercises for Chapter Thirty Two	533
A The Initial MEP Formula and <i>Mathematica</i>	557
B Manipulating the Lagrange Multiplier Parameters	563
C <i>Mathematica</i> and the Legendre Transformation	569
D The Nested Structure of <i>Mathematica</i> Programs	575
REFERENCES	579

List of Figures

20.1 <i>Feller's abstract representation for an elementary point in the sample space as it applies to data for the dice scenario</i>	115
21.1 <i>An example of a contingency table where $M = 16$ kangaroos will be categorized according to beer and hand preference</i>	127
21.2 <i>A joint probability table for the kangaroo scenario assigning numerical values to the probability of all four cells</i>	132
22.1 <i>An eight cell joint probability table for one personality trait and two physical traits in the kangaroo scenario</i>	152
22.2 <i>An eight cell contingency table containing the data from past observations of the physical and personality traits of $N = 100$ kangaroos .</i>	153
22.3 <i>The joint probability table under a model implementing a correlation between beer preference and fur color</i>	168
22.4 <i>The joint probability table under a model implementing correlations between beer preference and hand preference as well as between beer preference and fur color</i>	169
23.1 <i>A joint probability table for the logistic regression example</i>	184
23.2 <i>Observed frequency counts from an hypothetical experiment involving 1,000 patients in the logistic regression example</i>	191
23.3 <i>Tree structure breakdown for the generic statement notation in the logistic regression example</i>	197
24.1 <i>The eight cells of the joint probability table for the enhanced kangaroo scenario under a model incorporating all three double interactions . .</i>	221

29.1 <i>Information geometry characterization of the Pythagorean relationship among three points representing probability distributions</i>	412
29.2 <i>“Triangular” relationship among the four points p, q, r and s representing probability distributions. The distribution p in \mathcal{S}^8 is projected onto point q in \mathcal{S}^4.</i>	415
31.1 <i>The geometry of the right triangle used to set up the physical motivation for the Cauchy distribution</i>	478
31.2 <i>A sketch of the geometric layout for the lighthouse scenario</i>	482
31.3 <i>Contour plot of the unnormalized posterior distribution of y, location of lighthouse out to sea, and z, location of lighthouse along shore, given the data from three detected flashes and a Cauchy model</i>	485
31.4 <i>Cauchy probability density function with $y = 1$ and $z = 0$</i>	495
31.5 <i>General Cauchy density function with $y = 4$ and $z = 2$</i>	498
31.6 <i>Regions of equal, but non-normalized, likelihood for the location of the lighthouse based on $N = 50$ flashes</i>	501
31.7 <i>A plot of the highly non-linear relationship between Kapur’s parameters b and c in the MEP derivation of the Cauchy distribution</i>	506
31.8 <i>A Cauchy distribution that looks like the Student-t distribution</i>	507
31.9 <i>Another general Cauchy distribution</i>	508
32.1 <i>The contingency table for the numerical example of assessing a kangaroo’s beer preference when conditioned on three known physical traits as well as the data</i>	512
32.2 <i>The contingency table containing the data for a new numerical example of assessing a kangaroo’s beer preference when conditioned on known physical traits</i>	546
32.3 <i>A joint probability table showing the numerical assignments under a noise model and a signal plus noise model</i>	551
B.1 <i>Mathematica output using the Manipulate[] command to implement the MEP formula for the die scenario</i>	568

List of Tables

17.1 <i>The effect that varying λ has on both the assigned numerical value Q_1 and the average of the constraint function</i>	8
18.1 <i>The effect that varying λ has on assigned numerical values and the average of the constraint function when the mapping has changed</i>	37
19.1 <i>The details of the MEP computation for the die problem with only one constraint. This is model \mathcal{M}_B</i>	64
19.2 <i>The details of the MEP computation for the die problem with two constraints. This is model \mathcal{M}_C</i>	65
19.3 <i>Another legitimate assignment of probabilities to the six faces of the die similar to model \mathcal{M}_C, but which contains extra information above and beyond what is in model \mathcal{M}_C. Its entropy therefore must be lower than model \mathcal{M}_C's</i>	68
20.1 <i>The eight patterns of data that produce a data average of 4.5 in four rolls of a die. The number of ways that each pattern could occur is also shown</i>	100
21.1 <i>The MEP calculations for a specific model inserting information about two marginal probabilities. The Lagrange multipliers for this model are found to have the values $\lambda_1 = \lambda_2 = 1.098612$</i>	131
21.2 <i>How the probability for a future frequency count of $\boxed{9} \boxed{3} \boxed{3} \boxed{1}$ changes as the α_i parameters increase in the Dirichlet distribution capturing the probability of the models</i>	146
23.1 <i>The “dummy” coding variables used for the two predictor variables in a conventional logistic regression program</i>	179

23.2 <i>The first nine of the seventeen constraint functions for the $2 \times 3 \times 3$ joint probability table. These are the marginal probability constraints for A, B, and C, together with the constraints for the AB and AC interactions</i>	188
23.3 <i>The last eight of the seventeen constraints for the $2 \times 3 \times 3$ joint probability table. These are the BC and ABC interactions</i>	189
23.4 <i>The pattern of Lagrange multipliers in the numerators of the assigned numerical values to the two relevant probabilities in Bayes's Theorem under a main effects model</i>	200
24.1 <i>A numerical illustration of the fact that $\ln x \geq (1 - \frac{1}{x})$</i>	205
25.1 <i>The match-ups between the calculus problem of maximizing a function subject to side conditions and finding a distribution with maximum entropy subject to constraints</i>	227
25.2 <i>A grid search calculation for the value of λ_1 under some model so that a numerical assignment can be made for the probabilities of three statements in the state space</i>	239
25.3 <i>The MEP algorithm assigns numerical values for probabilities when given information reflected in the Lagrange multiplier</i>	241
26.1 <i>Using the MEP formula to assign probabilities to energy levels in a statistical mechanics problem where $kT = 2$</i>	273
26.2 <i>The first set of a few of the close to three trillion possible frequency counts of 200 molecules at eight energy levels. These counts all have a relatively high probability of occurring</i>	275
26.3 <i>The second set of a few of the close to three trillion possible frequency counts of 200 molecules at eight energy levels. These counts all have a relatively low probability of occurring</i>	276
26.4 <i>A numerical experiment looking at the probabilities for the specific frequency count found as the expectation based on the MEP's Q_i . . .</i>	304
26.5 <i>The beginning of an induction comparing the log maximum multiplicity factor versus the log of the sum of all the multiplicity factors . . .</i>	308
26.6 <i>A numerical induction comparing the sum of all multiplicity factors versus the maximum multiplicity factor</i>	309
27.1 <i>Using the MEP algorithm to assign probabilities to the eight energy states in the state space for the Schrödinger scenario</i>	326

27.2 Schrödinger's table that sets up the canonical problem in statistical mechanics	330
27.3 All nine elementary points in Baierlein's sample space illustrating an example of a system with two bosons	337
28.1 The computational details for g_{21} , the covariance between the second and first constraint functions under the independence model, for the simplified kangaroo scenario of Chapter Twenty One	372
28.2 The computational details for g_{31} , the covariance between the third and first constraint functions under the correlation model in the simplified kangaroo scenario of Chapter Twenty One	394
28.3 How the probabilities for two models are re-ordered after a single observation in Jaynes's die scenario	397
29.1 Checking that the probabilities for all six possible frequency counts on the next two coin flips add up to 1	448
29.2 A few of the models that played a role in the probability for the next two tosses of the coin	449
31.1 The kind of x values to expect from a Cauchy distribution as the angle θ varies uniformly over -90° through $+90^\circ$	493
31.2 The kind of x values to be expected from a Cauchy distribution as the angle θ varies uniformly over -90° through $+90^\circ$. The y and z locations are changed from $y = 1$ and $z = 0$ to the lighthouse scenario where $y = 4$ and $z = 2$	497
32.1 Some significant marginal sums from the data in the contingency table of Figure 32.1	513
32.2 The probability for any next, that is, the $(N + 1)^{st}$ kangaroo, to prefer Foster's given observation of the three physical traits of hand preference, fur color, and intelligence. The data were collected from $N = 1000$ kangaroos	517
32.3 The decomposition of $m = 55$ constraint functions into main effects together with all double, triple, and quadruple interactions following a typical ANOVA schema	533
32.4 The details of the computation to find the log likelihood of the noise model versus the fair model	549
32.5 Successive signal models showing the improvement over the previous baseline model. The improvement stops at $m = 12$	550

Chapter 17

How To Assign Numerical Values to Probabilities

17.1 Introduction

In Volume I, numerical assignments to probabilities were made at point blank range with no offer of a justification or a rationale. In fact, neither justification nor rationale were required within the rules of the game as laid down at that juncture. The numerical assignments just had to be *legitimate*, that is, they had to be real numbers between 0 and 1 such that the sum of the assignments to the statements in the state space equaled 1.

Numerical assignments to probabilities may come from anywhere. They *may* come from an algorithm like the Maximum Entropy Principle (MEP). Or they may arrive from intuition, from some underlying theory, from guesswork, from your Uncle Ed, or from divine revelation if you are so lucky. If the assignments are legitimate probability assignments, these various sources are all valid. The MEP is just *one method* for assigning legitimate numerical values to probabilities.

However, the MEP has many features that make it quite attractive, and therefore distinguish it from its more haphazard, *ad hoc* bedfellows. It is an algorithm, easily programmed, which can automatically generate the necessary legitimate numerical assignments to probabilities for a state space of any dimension. The key feature is that the numerical assignments are based on *information* inserted by some model.

The MEP is a disciplined way of allowing the IP to insert some desired information into a probability distribution while, at the same time, excluding all unwanted information from that probability distribution. I like to think of it as taking out an insurance policy. The IP has an iron-clad guarantee that the MEP assignment has *the most missing information* compared to any other assignment.

This is just another way of saying that if the IP didn't consciously place the desired information into the probability distribution over the state space, then the MEP prevents that unwanted information from finding its way into the probability distribution. To boot, all conceivable unwanted information is excluded from the probability distribution.

It must always be kept in mind that the MEP, in its role of assigning legitimate numerical values to abstract probabilities, is conceptually *orthogonal* to the formal manipulation rules like the **Sum Rule**, the **Product Rule**, and **Bayes's Theorem** as introduced in Volume I. These manipulation rules are valid across the board for any and all legitimate probabilities. These rules could care less about the intricacies of how the numbers are assigned. These rules don't care about numbers, and they never get around to even addressing how numbers might be assigned.

A wonderful analogy, not original with me, is to think of the formal manipulation rules as the grammatical rules of a language. Such rules dictate the manner in which something can be expressed. Numerical assignment routines, like the MEP, are analogous to a language's vocabulary. It allows us to say something specific within the overall confines imposed by the grammar.

17.2 The State Space

An absolutely fundamental concept that underlies everything that is done within probability theory is the notion of the *state space*. The state space consists of n statements, or, more generally, n joint statements, where each statement is either TRUE or FALSE. This is familiar to us from Volume I. The statements *define* what could happen, what could be observed, or what could be measured on any one observation or trial.

Take the case of rolling a die. The state space is defined as a listing of all the possible events that could be observed in the roll of the die. Obviously, a state space is allowed to be defined differently depending on how the problem is understood, as well as the level of detail to be included.

For example, the conventional state space for the die rolling scenario lists six possibilities. One of the faces lands squarely face up, and the number of spots on that face is accurately recorded.

However, the state space could be expanded to include the possibility that the die might be lost after being rolled. Or, to include a finer level of detail, the possibility that the die could land tilted resting against a wall could be included in a revised state space. We could keep extending the state space to cover every conceivable physical possibility up to recording every quantum state of the die if so desired.

The canonical example of how different state spaces come into play resides in statistical physics. Historically, physicists first defined state spaces in thermodynamics in terms of small elements of location and momentum. This is the classical phase space of Boltzmann and Gibbs.

Later, as quantum mechanics developed, the state space was redefined to take into account these new theoretical concepts. As far as inferencing is concerned, the state space is a malleable entity adjusting to the requirements demanded of it.

In Volume I, we used the notation $(A = a_i)$ for some statement A which could take on the observed value or measurement of a_i . In order to emphasize the grammatical fact that A was a statement, punctuation marks were used. Such statements always ended with a period, and were always enclosed by quotation marks.

For example, $(A = a_1)$ might be the statement, “The coin shows HEADS after being tossed.” with $(A = a_2)$ the statement, “The coin shows TAILS after being tossed.” Each of these two statements must, in principle, be capable of being judged, or determined, as TRUE or FALSE.

There are no other alternatives for the categorization of a result from the coin toss since the state space was defined as having dimension of only $n = 2$. We are certainly allowed to expand the state space if we wish. If the state space is expanded to $n = 3$ in order to include the possible observation, “The coin landed on its EDGE after being tossed.”, then the two statements in the original state space could both be FALSE.

For joint statements, the notation $(A = a_i, B = b_j, C = c_k, \dots)$ was employed. For example, $(A = a_2, B = b_1)$ might be the joint statement, “The kangaroo uses its right hand to drink Corona beer.” Once again, in this state space defined by $n = 4$, each of the four possible joint statements must be judged as TRUE or FALSE. In addition, every observation of a kangaroo, by definition, must belong to one of these true joint statements. These are the *mutually exclusive* and *exhaustive* properties.

To avoid an unwieldy growth in the number of symbols, we prefer to use just one generic variable to represent a joint statement. In all of our future examples, the following notation will be employed. $(X = x_i)$ will stand for the i^{th} statement in the n -dimensional state space.

Suppose that the state space for the die rolling scenario has been defined to consist of $n = 6$ statements. Thus, $(X = x_1)$ stands for the statement, “The ONE spot showed face up after the die was rolled.” and $(X = x_6)$ stands for the statement, “The SIX spot showed face up after the die was rolled.”

Suppose further that a state space has been defined to cover joint statements about the roll of a die and a toss of the coin. This state space has been defined to have a dimension of $n = 12$. Thus, $(X = x_{10})$ stands for the joint statement, “The die showed a FOUR spot and the coin showed TAILS.”

The probability operator is wrapped around some statement in the state space, $P(X = x_i)$, to indicate the IP’s degree of belief that the statement in question is TRUE. Nevertheless, such a notation can only refer to an abstract probability. If we intend that this probability should be understood as having a legitimate numerical value, then it must be conditioned on the truth of some other statement.

17.2.1 Probabilities conditioned on a model

Let me repeat what I just said. If a probability is understood as having a legitimate numerical value, then it must be conditioned on the truth of some other statement. One hears the echo from Jaynes's exhortations that one should never write out something like $P(X = x_i)$, translated verbally as "*the probability of* $X = x_i$," but rather one should always write $P(X = x_i | \mathcal{I})$, translated verbally as, "*one assigned probability for* $(X = x_i)$ *conditioned on the information in* \mathcal{I} ." Thus, Jaynes taught us an important moral: There cannot be one absolutely true probability.

As a corollary, Jaynes constantly strove to drive home the distinction between ontology and epistemology. He repeatedly emphasized that writing $P(X = x_i)$, without specifying any conditioning information, revealed an IP's predilection for viewing probabilities with an ontological mind set. Under this delusion, probability was inherently no different than any definite physical property of an object.

From a physics and ontological viewpoint, one might imagine a function $C(X)$ that returns *the color of* an object, or a function $W(X)$ that returns *the weight of* an object. Then, it is quite legitimate to say that this die weighs two ounces and $W(X) = 2$ and, furthermore, the die is red and $C(X) = 653$ nm. Again, we feel completely at ease claiming that these assertions are about *the physical property of* the object.

On the other hand, probabilities written as $P(X = x_i | \mathcal{I})$, are epistemological in nature. They represent a *state of knowledge* held by an information processor conditioned on whatever information \mathcal{I} the IP is using. There is no notion that a probability assigned for X is something that can be measured, observed, discovered, or estimated by examining the physical properties of X . Thus, assigned numerical values to probabilities are neither right nor wrong; rather, one can only say that they are compatible with some assumed information.

The epistemological notion is based firmly on the view that probability simply represents the IP's belief that $(X = x_i)$ is true based on some partial and inadequate information. Probability is neither logical nor physical, but rather informational. When fully adequate information becomes available, then $P(X = x_i | \mathcal{I})$ will become 0 or 1. Until then, one cannot claim that anybody's probability assignment, however arrived at, is necessarily right or wrong.

In scientific inference, as a pragmatic instantiation of the information \mathcal{I} , models are introduced that assign legitimate numerical values to the abstract probabilities. What was formerly written in the generic representation as $P(X = x_i | \mathcal{I})$ now becomes $P(X = x_i | \mathcal{M}_k)$. Paraphrasing Dennis Lindley, it becomes convenient to "introduce models into the scientific conversation."

The bulk of the discussion in this Volume then centers on how these models \mathcal{M}_k assign legitimate numerical values for all of the probabilities of the joint statements in the state space. The core concept is that models simultaneously *supply* information to, and *exclude* missing information from, probability distributions.

17.3 Introducing the MEP formula

Before we can progress any further in our discussion, we need to introduce the notation for the simplest possible version of the MEP formula. Other more complicated versions with possibly differing notational elements will eventually find their way into the formula as it develops in scope.

Here is an exponential formula for assigning a numerical value to a probability for observing that X is actually measured or observed as $X = x_i$. An explanation of the components involved in the formula follows.

$$Q_i \equiv P(X = x_i | \mathcal{M}_k) = \frac{e^{\lambda F(X=x_i)}}{Z(\lambda)} \quad (17.1)$$

An abbreviated version for the complete expression $P(X = x_i | \mathcal{M}_k)$ is required for future usage, and the notation Q_i fills that role. When Q_i appears, one should immediately surmise that a definite numerical value has been made, under the auspices of the information contained in some model, to what was formerly an abstract probability.

The notation on the right hand side of Equation (17.1) is as follows: λ is called the Lagrange multiplier, and referred to as a *parameter* of the model. $F(X = x_i)$ is called the *constraint function*. This constraint function contains a statement as its argument. Finally, $Z(\lambda)$ is called the *partition function*. It is a normalization factor forcing legitimate numerical assignments to probabilities between 0 and 1.

The partition function is defined to be the sum over all n elements in the state space.¹ The sum $Z(\lambda)$ consists of all the expressions that appeared in the numerator of Equation (17.1).

$$Z(\lambda) = \sum_{i=1}^n e^{\lambda F(X=x_i)} \quad (17.2)$$

There is an additional critical component in the MEP formula that is not visible. It is also labeled as a parameter, and it is directly linked to λ . Technically, it is called a *dual parameter*. Nevertheless, do not take from this that there are two fundamentally different sets of parameters motivating the MEP formula. In reality, each model can be expressed via parameters in two different guises.

This dual parameter is the probabilistic expectation of the constraint function,

$$\langle F \rangle = \sum_{i=1}^n F(X = x_i) P(X = x_i | \mathcal{M}_k) \equiv \sum_{i=1}^n F(X = x_i) Q_i \quad (17.3)$$

where the angle bracket notation from physics is used to indicate the mathematical expectation, or average, of the constraint function.

¹The letter Z for the partition function derives from the German word *Zustandssumme*, translated literally as “sum over states.” The term originates from early work in statistical mechanics.

17.3.1 Assigning numerical values to a small state space

As always, these notions are best understood through an easy numerical example. Consider the situation where the state space consists of $n = 2$ statements. Obviously, we are thinking about a conventional coin toss here.

Any numerical assignment for this coin tossing scenario that we might have made in Volume I, say, for example, $Q_1 = 3/4$ and $Q_2 = 1/4$, was simply picked out of the blue by some model \mathcal{M}_k . Such an assignment was made with no apparent rhyme nor reason other than qualifying as a legitimate numerical value.

With this proverbial toss of the coin, we are once again engaging in the canonical pedagogical probability exercise. Let the state space be discrete, and consist of the aforementioned $n = 2$ propositions, or statements:

1. $(X = x_1) \equiv$ “HEADS showed face up after the coin was tossed.”
2. $(X = x_2) \equiv$ “TAILS showed face up after the coin was tossed.”

Each of these statements is either TRUE or FALSE upon observation. There is no intermediate alternative in our Aristotelian logic. Boolean Algebra, if you remember, did permit us to choose an alternative to T or F as an assigned value to a variable.

Before the coin is tossed, the information processor is uncertain about which face will show, and the IP captures that uncertainty with a probability distribution over the statements in the state space. After the coin has been tossed, all uncertainty vanishes; it is taken for granted that the IP is fully capable of determining which face appeared.

It is important to emphasize that the state space consists of *statements*, not *numbers*. The total probability of 1 is distributed over all n statements in the state space. Every coin toss must be placed into one, and only one, of the two available categories.

In general, every observation or measurement must be placed into one, and only one, of the previously defined n categories. If the IP can imagine an outcome of the coin toss that could not be categorized as either HEADS or TAILS, but rather, say, something like “The coin landed on its EDGE.”, then the state space must be enlarged to cover these other possibilities.

How does the MEP formula make the assignment $Q_1 \equiv P(X = x_1 | \mathcal{M}_k)$? From this point forward, *every* single Q_i will be found by resorting to the MEP algorithm. Consider the constraint function first. The constraint function makes some functional assignment to the two *statements* $(X = x_1)$ and $(X = x_2)$.

The constraint function implements a mapping from statements to the real numbers. This mapping operation is quite abstract and general. For example, we might simply let $F(X = x_1) = 1$ and $F(X = x_2) = 2$.

The IP will insert the desired information into a probability distribution through this constraint function and its average. Suppose that the average of the constraint function is specified as $\langle F \rangle = 1.5$. Thus, from Equation (17.3),

$$\langle F \rangle = (Q_1 \times 1) + (Q_2 \times 2) = 1.5$$

We can see immediately that the only legitimate values of Q_1 and Q_2 satisfying this constraint are $Q_1 = 1/2$ and $Q_2 = 1/2$. But, of course, we want to illustrate how Equations (17.1) and (17.2) work when things are not so obvious.

Solve the numerical assignment problem with Equations (17.1) and (17.2) by substituting the values for the constraint functions under model \mathcal{M}_k ,

$$\begin{aligned} Q_1 \equiv P(X = x_1 | \mathcal{M}_k) &= \frac{e^{\lambda F(X=x_1)}}{Z(\lambda)} \\ &= \frac{e^{\lambda \times 1}}{Z(\lambda)} \\ Q_2 \equiv P(X = x_2 | \mathcal{M}_k) &= \frac{e^{\lambda F(X=x_2)}}{Z(\lambda)} \\ &= \frac{e^{\lambda \times 2}}{Z(\lambda)} \\ Z(\lambda) &= \sum_{i=1}^2 e^{\lambda F(X=x_i)} \\ &= e^{\lambda \times 1} + e^{\lambda \times 2} \end{aligned}$$

Now set the Lagrange multiplier λ to 0 to see that when $\lambda = 0$, the partition function is equal to,

$$Z(\lambda) = e^\lambda + e^{2\lambda} = 1 + 1 = 2$$

Under this model, the numerical assignments to HEADS and TAILS becomes,

$$Q_1 = \frac{e^{\lambda F(X=x_1)}}{Z(\lambda)} = 1/2$$

and,

$$Q_2 = \frac{e^{\lambda F(X=x_2)}}{Z(\lambda)} = 1/2$$

The Lagrange multiplier and the constraint average are dual parameters. Thus, they are linked in some fundamental sense. We could set the Lagrange multiplier at some value, and then find the resulting constraint average. Or, we could do it the other way around by first setting the constraint average, and then find the Lagrange multiplier. If we set the Lagrange multiplier at $\lambda = 0$, then $\langle F \rangle$ must equal 1.5, and vice versa.

17.3.2 Varying the parameters

How would Q_1 be assigned the numerical value of $3/4$ through the MEP formula? Suppose that $\langle F \rangle = 1.25$ is given as the information. In general, a numerical optimization program would find the corresponding λ .

Otherwise, for our elementary purposes here, it is instructive to simply watch the assigned numerical value Q_1 , as well as the constraint function average $\langle F \rangle$, change in response when λ is varied. When we reach the desired $Q_1 = 3/4$, we stop to see what the accompanying λ parameter is, and then verify that the dual parameter has the correct value of 1.25.

Table 17.1 shows what happens to Q_1 and $\langle F \rangle$ when λ is varied between $-\infty$ and $+\infty$. We see that when $\lambda = -1$, Q_1 has the assigned numerical value of 0.7311, therefore λ must be somewhere in the vicinity of $\lambda = -1$ to achieve the desired value of $Q_1 = 3/4$. In fact, as the table shows, when $\lambda = -1.09861$, $Q_1 = 3/4$, $Q_2 = 1/4$, and the expectation of the constraint function is $\langle F \rangle = 1.25$.

Table 17.1: *The effect that varying λ has on both the assigned numerical value Q_1 and the average of the constraint function.*

λ	Q_1	$\langle F \rangle$
$-\infty$	1.0000	1.0000
-3	0.9526	1.0474
-2	0.8808	1.1192
-1.09861	0.7500	1.2500
-1	0.7311	1.2689
0	0.5000	1.5000
+1	0.2689	1.7311
+2	0.1192	1.8808
+3	0.0474	1.9526
$+\infty$	0.0000	2.0000

It is also clear that just as Q_1 and Q_2 can only vary between 0 and 1, so the dual parameter $\langle F \rangle$ can only vary between 1 and 2. The Lagrange parameter λ can vary between $-\infty$ and $+\infty$, although, as seen in this example, the absolute value of λ doesn't have to reach a very high value before probabilities attain numerical assignments close to 1 or 0.

17.4 The Important Lessons

This initial example using the MEP to assign numerical values to probabilities serves to point out some extremely important concepts. These are the concepts that will be constantly highlighted in the subsequent development of the MEP algorithm.

Bring to mind that over-arching concept of marginalization as discussed extensively in Volume I. Models and parameters will eventually be swept under the rug through marginalization, as in the following canonical expression,

$$P(X = x_i) = \sum_{k=1}^{\mathcal{M}} P(X = x_i | \mathcal{M}_k) P(\mathcal{M}_k) \quad (17.4)$$

In Volume I, we took the liberty of calling such a formula, relying as it does upon the formal rules for probability manipulation, a *prediction* equation. Before any data were observed, we called it a *prior prediction* equation.

First, we just want to reiterate that the MEP is an algorithm for operationally implementing a numerical assignment via some given model. So far, it has been used solely for the first term in the prior prediction equation,

$$Q_i \equiv P(X = x_i | \mathcal{M}_k)$$

Such a notation obviously and unequivocally dictates that the numerical values assigned to the probabilities for the statements in the state space are conditioned on the assumed truth of the information contained in some k^{th} model.

The second term in the prediction equation, $P(\mathcal{M}_k)$, has yet to be discussed. When relying upon the MEP, we are *not* assigning numerical values to some *joint distribution* consisting of both models and the state space, written in our usual notation as $P(X = x_i, \mathcal{M}_k)$, but to events in the state space given that model \mathcal{M}_k is assumed true.

Thus, it is very important to maintain the conceptual separation between,

$$P(X = x_i | \mathcal{M}_k) \text{ and } P(\mathcal{M}_k)$$

Thus far, the MEP has been involved only in implementing a numerical assignment given that some model has been picked out. It says nothing at all yet about how to judge the relative merits of all the competing models.

Secondly, the *parameters* of a model are defined strictly as either of the dual parameters λ or $\langle F \rangle$ appearing in the MEP algorithm. Models and parameters have a one-to-one relationship. Every different parameter setting results in a different model, and vice versa.

Models are not observables; they are expedient from a mathematical standpoint. All of them eventually get swept under the rug during the fundamental averaging procedure. Nevertheless, models and parameters are not just *convenient* fictions, they are *necessary* fictions.

Thirdly, and most important of all, observed data do not play any kind of role whatsoever in how the MEP assigns numerical values! Information is inserted through the constraint function(s) and the associated constraint function average(s). Neither of these are data!

There is, no doubt, a relationship between the observed data and the numerical assignments effected through the auspices of the MEP. This relationship will be explored in much detail later on. For a start, look at Exercises 17.7.8 through 17.7.13. But it remains the most fundamental criterion of a *prior* assignment to the state space, if the word *prior* is to mean anything at all, that any possible future observations or measurements not be involved in that initial numerical assignment.

17.5 The MEP and Data

Let us begin to disambiguate the confusion surrounding the MEP and data. As a refresher on notation, consider the probability of obtaining TAILS on the first trial, and then HEADS on the final two trials in three future tosses of a coin. Thus, we would write,

$$P(X_3 = x_1, X_2 = x_1, X_1 = x_2)$$

as the expression for this joint probability where the subscript on X_t indicates the trial number, and as usual, $(X_t = x_1)$ is the statement, “HEADS shows after the coin is tossed.” and $(X_t = x_2)$ is the statement, “TAILS shows after the coin is tossed.”

The formal manipulation rules of probability stipulate that if we want to include statements about a model assigning numerical values, then,

$$\begin{aligned} P(X_3 = x_1, X_2 = x_1, X_1 = x_2) &= \sum_{k=1}^{\mathcal{M}} P(X_3 = x_1, X_2 = x_1, X_1 = x_2, \mathcal{M}_k) \\ P(X_3 = x_1, X_2 = x_1, X_1 = x_2, \mathcal{M}_k) &= P(X_3 | X_2, X_1, \mathcal{M}_k) \times P(X_2 | X_1, \mathcal{M}_k) \times \\ &\quad P(X_1 | \mathcal{M}_k) \times P(\mathcal{M}_k) \\ P(X_3 = x_1, X_2 = x_1, X_1 = x_2) &= \sum_{k=1}^{\mathcal{M}} P(X_3 | \mathcal{M}_k) \times P(X_2 | \mathcal{M}_k) \times \\ &\quad P(X_1 | \mathcal{M}_k) \times P(\mathcal{M}_k) \\ &= \sum_{k=1}^{\mathcal{M}} (Q_1 \times Q_1 \times Q_2) \times P(\mathcal{M}_k) \end{aligned}$$

If, in fact, there is only one model making the assignments, then the integration replacing the sum over \mathcal{M} models, becomes,

$$P(X_3 = x_1, X_2 = x_1, X_1 = x_2) = \int_0^1 q_1^2 q_2 \delta(q_i - Q_i) dq_i = Q_1^2 Q_2$$

As we discovered in Volume I, this type of argument is easily generalized to,

$$P(N_1, N_2, \dots, N_n) = W(N) Q_1^{N_1} Q_2^{N_2} \cdots Q_n^{N_n} \quad (17.5)$$

This is the probability for the *data*, or the frequency counts N_1 through N_n , when the averaging with respect to $P(\mathcal{M}_k)$ is done not over all models, but only over one model.

With the introduction of the MEP, the Q_i no longer have to remain symbolic. For the simple case of $n = 2$ we are considering here, they are assigned numerical values according to the MEP formula in Equation (17.1) as,

$$Q_i = \frac{e^{\lambda F(X=x_i)}}{Z(\lambda)}$$

Substituting into Equation (17.5), we have,

$$P(N_1, N_2) = W(N) \times \left[\frac{e^{\lambda F(X=x_1)}}{Z(\lambda)} \right]^{N_1} \times \left[\frac{e^{\lambda F(X=x_2)}}{Z(\lambda)} \right]^{N_2} \quad (17.6)$$

This expression is considerably more transparent by taking the log transform for the probability of the future data,

$$\begin{aligned} \ln [P(N_1, N_2)] &= \ln [W(N)] + N_1 [\lambda F(X = x_1) - \ln Z] + N_2 [\lambda F(X = x_2) - \ln Z] \\ &= \ln [W(N)] + N_1 [\lambda F(X = x_1)] + N_2 [\lambda F(X = x_2)] - N \ln Z \end{aligned} \quad (17.7)$$

Notice the explicit appearance of the multiplicity factor $W(N)$, the frequency counts N_1 and N_2 , and the total number of observations N . In other words, the explicit appearance of *data*. So data do make an appearance in calculating probabilities, but only *after* the Q_i were first given their assignments via the MEP.

This above expression for the probability of frequency counts is not what the MEP is concerned about. The MEP operates at the level of the n -dimensional state space, not at the level of the N -dimensional data space. Unceasing attempts to apply the MEP to the latter space have resulted in massive and pervasive confusion as to what the MEP is trying to accomplish.

17.6 Connections to the Literature

As already mentioned, the bulk of the references in Volume II will be to Jaynes's seminal work on the Maximum Entropy Principle. Fourteen of Jaynes's papers focusing on probability and maximum entropy, written from 1957 to about 1980, were edited and then collected together in the book [13],

E.T. Jaynes: Papers on Probability, Statistics and Statistical Physics

Many of the references appear as part of that book. So, it makes sense to refer to a particular paper of Jaynes by its original publication data, which is then followed by where it appears in the above book.

Jaynes mentions in personal reminiscences that he first had the idea of the MEP in 1951, but, because of the opposition of Establishment figures within the Physics community, and his own inability to clearly articulate what were the first glimmerings of the MEP, delayed publishing until 1957. When the first of these two articles [14] finally saw the light of day, it laid out the essential rationale for what we are calling the MEP algorithm.

I wish that scientists were more forthcoming, as Jaynes was below, in describing the real world difficulties in getting ideas out to a wider audience [18, pg. 237].

All this was clear to me by 1951; nevertheless, no attempt at publication was made for another five years. There were technical problems in extending the formalism to continuous distributions and the density matrix that were not solved for many years; but the reason for the initial delay was quite different.

In the Summer of 1951, Professor G. Uhlenbeck gave his famous course on Statistical Mechanics at Stanford, and following the lectures I had many conversations with him, over lunch, about the foundations of the theory and current progress in it. I had expected, naively, that he would be enthusiastic about Shannon's work, and eager as I to exploit these ideas for Statistical Mechanics. Instead, he seemed to think that the basic problems were, in principle, solved . . . and adamantly rejected all suggestions that there is any connection between entropy and information. . . .

It was my total inability to communicate this argument to Professor Uhlenbeck that caused me to spend another five years thinking over these matters, trying to write down my thoughts more clearly and explicitly, and making sure in my own mind that I could answer all the objections that Uhlenbeck and others had raised. Finally, in the Summer of 1956 I collected this into two papers, sending the first off to the Physical Review on August 29.

Jaynes chose to title his first two papers written in 1957 introducing the MEP as ***Information Theory and Statistical Mechanics***. He clearly is indicating his motivation, as a physicist, for his views on probability. The fundamentals of the MEP are presented in section 2 of the first, and easier, of these two papers.

Jaynes's second paper [15] delves into the MEP as he applied it to quantum mechanics. Jaynes himself admits that he tried to cover far too much material in this second paper. I will not refer to this second paper because I do not understand all aspects of it thoroughly. However, the parts that I do partially comprehend seem to contain an untold wealth of fascinating material.

For example, it seems to me that Jaynes, writing in 1957, adumbrated the following notions: the quantum to classical transition, the loss of quantum entanglement, the extremely fast relaxation times of quantum systems in contact with the environment, and last, but not least, the idea of *decoherence*. All of these notions were tied up with the central explanatory concept of *information loss* in irreversible systems.

Nonetheless, the introduction to the MEP in Jaynes's first paper is certainly adequate to our needs. In his Equation (2.1), he clearly states that *information* is going to be defined as a constraint function average. Unfortunately, over the succeeding years he wavered from this correct definition by co-mingling it with the idea that information could also be *data*. This mistake has turned out to be an endless source of confusion.

Later on, Jaynes would sometimes change the language about information as averages of the constraint functions to simply “numbers given in the statement of the problem.” This phraseology is less clear than his initial explanation that the constraint function average was the “information” to be inserted by some model into a numerical assignment for the probabilities. In any case, in 1957 Jaynes was telling us that information was definitely NOT synonymous with data.

Jaynes is at his best in highlighting fundamental conceptual distinctions that all of us should keep in mind. He constantly emphasized the vital distinction between probability as an *ontological* concept (false) versus thinking about probability as an *epistemological* concept (true). Once again, we cannot improve on Jaynes's own explanation [14, pg. 8].

The “objective” school of thought regards the probability of an event as an objective property of that event, always capable in principle of empirical measurement by observation of frequency ratios in a random experiment. . . .

On the other hand, the “subjective” school of thought regards probabilities as expressions of human ignorance; the probability of an event is merely a formal expression of our expectation that the event will or did occur, based on whatever information is available. . . . the purpose of probability theory is to help us in forming plausible conclusions in cases where there is not enough information available to lead us to certain conclusions...

The next fundamental point that Jaynes makes is that the quantitative measure of the “amount of uncertainty” is best captured within Shannon's formulation of *information entropy*. In words, Jaynes foreshadows the derivation of the MEP formula by telling us that [14, pg. 9],

It is now evident how to solve our problem; in making inferences on the basis of partial information we must use that probability distribution which has maximum entropy subject to whatever is known. This is the only unbiased assignment we can make; to use any other would amount to arbitrary assumption of information which by hypothesis we do not have.

My notation and Jaynes's notation in his first 1957 paper coincide for the most part. The dimension of the state space is labeled as n , while the number of constraint functions is labeled as m , where $m < n$. Jaynes uses the notation $f(x_i)$ for constraint functions, whereas I employ $F(X = x_i)$. Jaynes uses p_i for the numerical values assigned to probabilities, whereas I use Q_i . I also follow Jaynes's lead by using λ for the Lagrange multiplier, and $Z(\lambda)$ for the partition function.

In Equation (17.1), I present the simplest version of the formula. Only one constraint function and one parameter are present, so $m = 1$. Jaynes writes out a more general MEP formula with m constraint functions and m Lagrange multipliers in his Equation (2–10) [14, pg. 9] as,

$$p_i = \exp \{ -[\lambda_0 + \lambda_1 f_1(x_i) + \cdots + \lambda_m f_m(x_i)] \}$$

Jaynes chose to show the Lagrange multipliers with a minus sign, whereas I prefer to preface them with a plus sign. The probabilities come out the same; with my plus sign the Lagrange multipliers I calculate are the negative of those that Jaynes calculates, and *vice versa*. Sometimes, Jaynes does show positive λ_j in the MEP formula, for example, his Equation (D7) in [18, pg. 292], and his Equation (17) in [17, pg. 120].

In the above way of writing the formula, λ_0 represents the logarithmic transform of the partition function,

$$\begin{aligned} e^{-\lambda_0} &= \frac{1}{e^{\lambda_0}} \\ e^{\lambda_0} &= Z \\ \lambda_0 &= \ln Z \\ p_i &= \frac{1}{Z} e^{-\lambda_1 f_1(x_i) - \lambda_2 f_2(x_i) - \cdots - \lambda_m f_m(x_i)} \end{aligned}$$

As Chapter Twenty Five will demonstrate, my numerical value for λ_0 is different than Jaynes's.

One important notational subtlety to keep in mind is that the argument to each constraint function is x_i where the subscript i is the same as the subscript i appearing in p_i . This is the same i that indexes the statements in the state space as they run from $i = 1$ through $i = n$.

Thus, there is a summation in the numerator over the m constraint functions and m parameters in the more general formula, but x_i is fixed at the i^{th} statement

in the state space. In the denominator, there is a double summation because we are not only summing over all m constraint functions, but also summing over the dimension n of the state space as well.

Jaynes is not quite so clear about the x_i , merely commenting that they are “quantities capable of assuming discrete values.” As the reader by now is well aware, I always emphasize the fact that $(X = x_i)$ refers to a *statement*. The “discrete values” only enter when the constraint functions $F_j(X = x_i)$ are defined.

In his commentary, Jaynes uses the phrase “maximum entropy predictions.” [14, pg. 5]. This is confusing. The MEP confidently assigns numerical values to probabilities; no predictions are involved. Any “predictions” should refer only to probabilities for future frequency counts.

Jaynes also talks about the “failure of the maximum entropy prediction.” There can be no “failure” of any kind when the MEP algorithm is correctly applied. The resulting assignment represents a state of knowledge about the joint statements in the state space given that the information in some model is assumed true. It can never be wrong in this sense. The state space as set up by the IP may certainly be ill-conceived, or Nature may prefer a different model, but the numerical assignments made by the MEP are not to blame for that.

It is quite clear that Jaynes meant rather that some particular model as captured by an MEP assignment may fail, that is, may not be supported by observed facts. But this does not imply that the original numerical assignment conforming to the information given was in any way wrong.

Unfortunately, Jaynes also uses the phrase “maximum entropy inference” on occasion. The MEP is NOT directly involved in an inference.

Inferencing should be considered to take place within that first conceptual realm of the formal manipulation rules as we were at pains to emphasize in Volume I. For example, $P(A | B, A \rightarrow B)$ is an inference, as opposed to a deduction, involving the statement A when conditioned on the assumed truth of another statement B and a model of logical implication. Information entropy plays no role whatsoever at this level; it’s all up to the formal manipulation rules.

Entropy may play a role when different models are considered for implementing variations on logical implication. Perhaps the marginal probabilities for $A = T$ and $B = T$ are not set at 1/2 within the joint probability table for A and B . The MEP algorithm may be employed to find different numerical assignments, but the form of the inference has already been dictated within that first conceptual realm.

Far more serious is Jaynes’s [14, pg. 8] incorrect characterization of Laplace’s *Principle of Insufficient Reason* in the context of explaining the MEP formula.

The problem of the specification of probabilities in cases where little or no information is available, is as old as the theory of probability. Laplace’s “Principle of Insufficient Reason” was an attempt to supply a criterion of choice, in

which one said that two events are to be assigned equal probabilities if there is no reason to think otherwise. However, except in cases where there is an evident element of symmetry that clearly renders the events “equally possible,” this assumption may appear just as arbitrary as any other that might be made.

This issue was thoroughly discussed in Volume I. Laplace was referring to the model space, and most certainly NOT to the state space when he advocated equal probabilities to all conceivable models through the *Principle of Insufficient Reason*.

The phrase “insufficient reason”² refers to a lack of support for belief in any one *cause* for the event to occur over belief in any other *cause*. Thus, Laplace assigned a probability of $1/2$ to events in the state space of dimension $n = 2$ not because there were only two possibilities; not because there was no reason to think otherwise; not because of any notion involving symmetry at the level of statements in the state space, but rather because $1/2$ is the answer that probability theory supplies when all possible models assigning probabilities to statements in the state space are given equal weight.

All of these verbal gyrations are made clear, as demonstrated in Volume I, when we calculate the “prediction,” or, more correctly, the state of knowledge, for $(X = x_i)$,

$$P(X = x_i) = \sum_{k=1}^{\mathcal{M}} P(X = x_i | \mathcal{M}_k) P(\mathcal{M}_k)$$

The *Principle of Insufficient Reason* applies to $P(\mathcal{M}_k)$ when we use a uniform distribution over model space to capture Laplace’s notion that the IP does not know the cause of $(X = x_i)$. It does NOT apply to $P(X = x_i | \mathcal{M}_k)$, as Jaynes said, which is where we are using the MEP formula. The “little or no information” in the case of $P(X = x_i | \mathcal{M}_k)$ might mean, using our current coin tossing example, that $\lambda = 0$ and $Q_1 = Q_2 = 1/2$.

Obviously, the reason why $P(X = \text{HEADS}) = 1/2$ is not the same reason why $P(X = \text{HEADS} | \mathcal{M}_k) = 1/2$ in this case. The only case where they would be the same is where, contrary to Laplace’s supposition of an insufficient reason to believe in any one cause of the phenomenon, there *is* a sufficient reason to believe in *one*, and *only one*, particular cause (model) to the exclusion of all others.

Then, the mathematics of the prediction equation using a Dirac δ -function for the distribution of models in model space does not result in an averaging over all

²Philosophers attribute the *Principle of Sufficient Reason* to Spinoza and Leibniz. It represents a philosophical axiom (which I agree with) that every event must have some reason for its occurrence. Jacob Bernoulli apparently played on this principle in adapting it to probability by coyly naming it the *Principle of Insufficient Reason* to indicate an indifferent preference over all the events that might take place. Various people, just like Jaynes in the above quote, then attributed this *Principle of Insufficient Reason* to Laplace for the wrong reason. Laplace correctly stated that when there was no reason to believe in any one *cause* for some event happening, then all causes (models) for the event ought to be given equal weight.

assignments as made by all the models, but in an “averaging” over one assignment made by one model. This specific model is obviously one where the information being inserted into the probability distribution is $\lambda = 0$.

Along these same lines, it now becomes clear that Jaynes is misleading us when he says, [14, pg. 9]

The principle of maximum entropy may be regarded as an extension of the principle of insufficient reason (to which it reduces in case no information is given except enumeration of the possibilities x_i), with the following essential difference. The maximum-entropy distribution may be asserted for the positive reason that it is uniquely determined as the one which is maximally non-committal with regard to missing information, instead of the negative one that there was no reason to think otherwise.

As just discussed, the principle of maximum entropy cannot be an extension of the principle of insufficient reason because the principle of insufficient reason just doesn't apply at the same level where the MEP is used. It is terribly confusing to mix up the state space and the model space. It is even more confusing to mix up information about what a model says about the statements in the state space with different information about the model itself in relation to all other conceivable models.

If you are totally uninformed about the models, then you cannot be totally uninformed about the state space. If you are completely knowledgeable about the models, then you may be completely uninformed about the statements. See how confusing this all is if you don't keep the distinction uppermost in your mind at all times about what space the *Principle of Insufficient Reason* really refers to.

Disambiguating the confusion surrounding these notions was the center piece of the discussions in Chapter Fifteen of Volume I. What is most perplexing of all is that it was Jaynes himself who disentangled the web of confusion!

My final quibble with Jaynes occurs when he says, [14, pg. 8]

... henceforth we will consider the terms “entropy” and “uncertainty” as synonymous.

I think it would have been better if Jaynes had said, “henceforth we will consider the terms “entropy” and “missing information” as synonymous.

Let me reiterate my quandary as first set out in the *Apologia*. In these criticisms of Jaynes, my intent is not to denigrate the far greater importance of the overall substance of what he has passed down to us concerning the MEP. Nevertheless, it is incumbent on me to point out where I think Jaynes tended to confuse us, or was just plain wrong.

17.7 Solved Exercises for Chapter Seventeen

Exercise 17.7.1: Change the constraint function for the standard coin toss. Demonstrate how the information must then change to retain the same probability assignments.

Solution to Exercise 17.7.1

Instead of the constraint function vector $(1, 2)$ that was initially used, suppose that a new constraint function vector $(2, 4)$ is suggested. Thus, $F(X = x_1) = 2$ and $F(X = x_2) = 4$. The mapping from the statement, “The coin shows HEADS.” to a real number is 2, and the mapping from the statement, “The coin shows TAILS.” to a real number is 4.

If the information represented by the constraint function average is changed to $\langle F \rangle = 3$, then $Q_1 = Q_2 = 1/2$. The universal constraint function that the assignments must sum to 1 is satisfied. The given constraint function average is also satisfied,

$$\langle F \rangle = [F(X = x_1) \times Q_1] + [F(X = x_2) \times Q_2] = 3$$

The Lagrange multiplier remains at $\lambda = 0$.

Exercise 17.7.2: What makes sense as the mapping for the universal constraint function?

Solution to Exercise 17.7.2

The universal constraint that the assignments must sum to 1 means that,

$$Q_1 + Q_2 = 1 \rightarrow F_0(X = x_1) Q_1 + F_0(X = x_2) Q_2 = \langle F_0 \rangle = 1$$

The mappings take the form of $F_0(X = x_1) = 1$ and $F_0(X = x_2) = 1$. The universal constraint function is represented by the vector $(1, 1)$.

Exercise 17.7.3: Show the details of the calculation that leads to an assignment of $3/4$ for a probability of seeing HEADS in Table 17.1.

Solution to Exercise 17.7.3

The probability for HEADS under some model \mathcal{M}_k is $Q_1 \equiv P(X = x_1 | \mathcal{M}_k)$. The information inserted under model \mathcal{M}_k is specified by some constraint function average. Here, it is $\langle F \rangle = 1.25$. The dual parameter that implements the same information under this model is the Lagrange multiplier λ where $\lambda = -1.09861$.

The numerator for Q_1 is,

$$\begin{aligned} e^{\lambda F(X=x_1)} &= e^{-1.09861 \times 1} \\ &= e^{-1.09861} \\ &= 0.333334 \end{aligned}$$

and the numerator for Q_2 is,

$$\begin{aligned} e^{\lambda F(X=x_2)} &= e^{-1.09861 \times 2} \\ &= e^{-2.19722} \\ &= 0.111112 \end{aligned}$$

The denominator is the calculated value for the partition function $Z(\lambda)$,

$$\begin{aligned} Z(\lambda) &= \sum_{i=1}^n e^{\lambda F(X=x_i)} \\ &= e^{(-1.09861 \times 1)} + e^{(-1.09861 \times 2)} \\ &= 0.333334 + 0.111112 \\ &= 0.444446 \end{aligned}$$

Since $Q_1 = e^{\lambda F(X=x_1)} / Z(\lambda)$, we arrive at the answer of,

$$Q_1 = \frac{0.333334}{0.444446} = 0.75$$

Of course, $Q_2 = 1 - Q_1 = 1/4$, but Q_2 could be calculated in exactly the same manner as Q_1 through the formula. The constraint function average is,

$$\begin{aligned} \langle F \rangle &= \sum_{i=1}^n F(X = x_i) Q_i \\ &= (1 \times 3/4) + (2 \times 1/4) \\ &= 1.25 \end{aligned}$$

Exercise 17.7.4: Construct a state space of dimension $n = 4$ for the coin tossing scenario.

Solution to Exercise 17.7.4

Suppose we envision a coin tossing situation that is slightly more involved than the standard generic $n = 2$ situation of merely recording HEADS or TAILS. The coin is

thicker than usual, so it is not outside the realm of possibility that it might land on its edge. Furthermore, we are tossing the coin onto a table, and if it should happen that the coin rolls off the table, we would like to record that as well.

This is obviously just an initial foray into more complicated state spaces where we are paying closer attention to the physical details that could be measured or observed. Remember that the (current) ultimate extension of this idea is to list all of the quantum states that can be measured for the phenomenon under investigation.

Statement 1. ($X = x_1$) “The coin lands showing HEADS after being tossed.”

Statement 2. ($X = x_2$) “The coin lands showing TAILS after being tossed.”

Statement 3. ($X = x_3$) “The coin lands on its EDGE after being tossed.”

Statement 4. ($X = x_4$) “The coin ROLLS OFF the table after being tossed.”

Exercise 17.7.5: Consider the expanded state space for the coin toss discussed in the last exercise. What is the value of the dual parameter if the Lagrange multiplier is set at $\lambda = -1$?

Solution to Exercise 17.7.5

There are now four possibilities that might be observed on any one trial of the coin toss. The coin might show HEADS, TAILS, land on its EDGE, or ROLL OFF the table. The dimension of the state space is $n = 4$. One, and only one, of these possibilities must be recorded as the data point at each trial.

Suppose that the constraint function vector is $(1, 1, 4, 5)$. This means that the mapping from the four statements in the state space to real numbers is,

$$F(X = x_1) = 1$$

$$F(X = x_2) = 1$$

$$F(X = x_3) = 4$$

$$F(X = x_4) = 5$$

The probabilities can be assigned some legitimate numerical values, (not, in any sense, thought of as the *true* numerical values), under some model by the MEP formula,

$$\begin{aligned} Q_i &= P(X = x_i | \mathcal{M}_k) \\ &= \frac{\exp [\lambda F(X = x_i)]}{\sum_{i=1}^n \exp [\lambda F(X = x_i)]} \end{aligned}$$

Since the information specified under the given model \mathcal{M}_k is that $\lambda = -1$, the numerators for Q_1 through Q_4 are respectively,

$$e^{-1}, \quad e^{-1}, \quad e^{-4}, \quad \text{and } e^{-5}$$

The partition function is,

$$Z(\lambda) = e^{-1} + e^{-1} + e^{-4} + e^{-5}$$

The numerical assignments under this model work out to,

$$Q_1 = 0.483535$$

$$Q_2 = 0.483535$$

$$Q_3 = 0.024074$$

$$Q_4 = 0.008856$$

The parameter dual to λ is $\langle F \rangle$. The constraint function average can now be calculated as,

$$\begin{aligned} \langle F \rangle &= \sum_{i=1}^n F(X = x_i) Q_i \\ &= (1 \times 0.483535) + (1 \times 0.483535) + (4 \times 0.024074) + (5 \times 0.008856) \\ &= 1.10765 \end{aligned}$$

The information inserted into a probability distribution under model \mathcal{M}_k is represented by either $\lambda = -1$ or $\langle F \rangle = 1.10765$. The numerical assignments are the same in either case.

Exercise 17.7.6: Discuss what a new constraint function $(1, 1)$, as opposed to the previously used constraint function $(1, 2)$, implies for the standard coin toss.

Solution to Exercise 17.7.6

This problem was broached in Exercise 17.7.2 with a discussion of the universal constraint function. This exercise asks us to treat $(1, 1)$ as just another constraint function, not really any different conceptually than our original constraint function of $(1, 2)$.

With this new constraint function, the numerical values for the probabilities, according to the MEP formula, are,

$$Q_1 = Q_2 = \frac{e^\lambda}{e^\lambda + e^\lambda} = 1/2$$

The constraint function average must be 1,

$$\begin{aligned}\langle F \rangle &= \sum_{i=1}^n F(X = x_i) Q_i \\ &= (1 \times Q_1) + (1 \times Q_2) \\ &= Q_1 + Q_2 \\ &= 1\end{aligned}$$

However, Q_1 and Q_2 separately could be anything as long as they satisfy the constraint of adding to 1. The assignments Q_1 and Q_2 might be 1/2, or they might be $Q_1 = 3/4$ and $Q_2 = 1/4$. They might be $Q_1 = 0$ and $Q_2 = 1$ for that matter. So why does the MEP formula come up with an answer of $Q_1 = 1/2$ and $Q_2 = 1/2$?

As its name implies, the MEP formula finds the assignment with the maximum value of the information entropy satisfying all of the constraints. In this case, the assignment with the largest possible value of the information entropy satisfying the constraint is the assignment found by the MEP formula. Every other assignment satisfying the constraint, such as $Q_1 = 3/4$ and $Q_2 = 1/4$, possesses a lower value for the entropy of its distribution.

As discussed elsewhere throughout this Volume, the constraint function (1,1) is implementing the so-called universal constraint function, that is, the constraint function enforcing the requirement that all of the numerical assignments sum to 1. Thus, if no further information other than the universal constraint function is specified by a model, the numerical assignment of a probability to all n statements will always turn out to be $1/n$.

In a sense, λ doesn't even exist, so set it to $\lambda = 0$. The resulting MEP assignment with $\lambda = 0$ is, of course, the correct one. This is the same answer we obtained with the original constraint function (1, 2) when it was not counted because $\lambda = 0$. The only constraint operating in that case was the universal constraint function, and we are back where we started.

Exercise 17.7.7: Discuss the Fermi oscillator in terms of the familiar coin tossing scenario.

Solution to Exercise 17.7.7

The relevance of something called a “Fermi oscillator” is not apparent now, but will become so when we treat the topics in Chapter Twenty Seven. In any case, the MEP formula as it was introduced here in this opening Chapter can be applied to this mysterious problem without any cause for concern. Use the following function to map the two statements in the coin tossing scenario to new values,

$$F(X = x_1) = 0$$

$$F(X = x_2) = \epsilon$$

Calculate the partition function Z first,

$$\begin{aligned} Z &= \sum_{i=1}^2 e^{\lambda F(X=x_i)} \\ &= e^{\lambda F(X=x_1)} + e^{\lambda F(X=x_2)} \\ &= e^{\lambda \times 0} + e^{\lambda \times \epsilon} \\ &= 1 + e^{\lambda \epsilon} \end{aligned}$$

Thus, relying on the MEP formula, the numerical assignment to a probability for HEADS is Q_1 ,

$$Q_1 \equiv P(X = x_1 | \mathcal{M}_k)$$

and the corresponding numerical assignment for TAILS is Q_2 ,

$$Q_2 \equiv P(X = x_2 | \mathcal{M}_k)$$

with the expression on the right hand side showing the strict conditioning under model \mathcal{M}_k . The MEP formula provides us with the assignment under such a model,

$$\begin{aligned} Q_1 &= \frac{1}{1 + e^{\lambda \epsilon}} \\ Q_2 &= \frac{e^{\lambda \epsilon}}{1 + e^{\lambda \epsilon}} \\ Q_1 + Q_2 &= \frac{1}{1 + e^{\lambda \epsilon}} + \frac{e^{\lambda \epsilon}}{1 + e^{\lambda \epsilon}} \\ &= 1 \end{aligned}$$

The constraint function average $\langle F \rangle$ must be specified as between 0 and ϵ . The actual numbers from the constraint function don't make any difference as long as the average is adjusted accordingly. For the sake of a numerical example, let $\epsilon = 50$ with $\langle F \rangle = 40$. Then, $\langle F \rangle = (0 \times 0.20) + (50 \times 0.80)$ if $Q_1 = 0.20$ and $Q_2 = 0.80$.

What is the value of the Lagrange multiplier that results in such an assignment?

$$\begin{aligned}
 Q_1 &= \frac{1}{1 + e^{\lambda\epsilon}} \\
 &= \frac{1}{1 + e^{50\lambda}} \\
 \frac{1}{1 + e^{50\lambda}} &= 0.20 \\
 0.20(1 + e^{50\lambda}) &= 1 \\
 0.20 + 0.20e^{50\lambda} &= 1 \\
 0.20e^{50\lambda} &= 0.80 \\
 e^{50\lambda} &= 0.80/0.20 \\
 \ln(e^{50\lambda}) &= \ln 4 \\
 50\lambda &= 1.38629 \\
 \lambda &= 0.027726
 \end{aligned}$$

As emphasized in the final section, none of this development relied upon any past observations. In other words, the data did NOT play any role whatsoever in the MEP assignment of numerical values to a probability for HEADS and a probability for TAILS. The probability for future observations can be cast into an expression that takes account of our new MEP formula as shown in the next few exercises.

Exercise 17.7.8: What is the probability of observing 30 HEADS and 70 TAILS in the next 100 coin flips under this last model?

Solution to Exercise 17.7.8

If the above model with the value of the parameter $\langle F \rangle = 40$ or, equivalently, $\lambda = 0.027726$ is taken to be the only model under consideration (that is, the IP is certain that this statement is TRUE, “This model assigning $Q_1 = 0.2$ and $Q_2 = 0.8$ as the probability for HEADS and TAILS is the correct assignment.”), then the binomial distribution will provide the probabilities asked for.

Instead of the correct notation of M and M_i as introduced in Volume I for *future* frequency counts, we will accede to convention and use N and N_i . The total number of future coin flips is $N = 100$ with $N_1 = 30$ HEADS and $N_2 = 70$ TAILS. The binomial probability for these future frequency counts of 30 HEADS and 70 TAILS under the one model specified above is then,

$$\begin{aligned}
 P(N_1 = 30, N_2 = 70 | \mathcal{M}_k) &= W(N) \times Q_1^{N_1} Q_2^{N_2} \\
 &= \frac{100!}{30! 70!} \times (0.2)^{30} (0.8)^{70} \\
 &= 0.00519
 \end{aligned}$$

Exercise 17.7.9: Rework the above calculation using the MEP.

Solution to Exercise 17.7.9

With the MEP formula for the Q_i at our disposal, this calculation of the probability for future observations of HEADS and TAILS can be reworked into an interesting alternative expression. Proceed small step by small step in the spirit of Wolfram's advice on theorem proving: applying an allowed transformation at each step and substituting. Technically, we should have already established each transformation as an axiom, lemma, or theorem (see Chapter One, Volume I, where this technique was illustrated for Boolean Algebra.)

It is quite helpful for later discussions to distinguish information as constraint function averages (correct) *vis-à-vis* information as data (incorrect). Please take care to note in the following derivation that no data enter the picture at any stage. It is confusing when frequency counts like N , N_1 and N_2 , and sample averages like \bar{F} enter into expressions because it does seem like data is playing a role. That is why I emphasized the distinction in the notation between future frequency counts, the M_i , and past data, the N_i .

In the last exercise, we left off with,

$$P(N_1, N_2 | \mathcal{M}_k) = W(N) \times Q_1^{N_1} Q_2^{N_2}$$

Apply a log transform, (to unclutter leave out the conditioning on the model),

$$\ln P(N_1, N_2) = \ln W(N) + N_1 \ln Q_1 + N_2 \ln Q_2$$

and then substitute the MEP formula for the Q_i ,

$$\ln P(N_1, N_2) = \ln W(N) + N_1 [\lambda F(X = x_1) - \ln Z] + N_2 [\lambda F(X = x_2) - \ln Z]$$

Begin a series of algebraic transformations to rework the probability for the future frequency counts into the alternative expression,

$$\begin{aligned}
\ln P(N_1, N_2) &= \ln W(N) + N_1 \lambda F(X = x_1) - N_1 \ln Z + N_2 \lambda F(X = x_2) - N_2 \ln Z \\
&= \ln W(N) + N_1 \lambda F(X = x_1) + N_2 \lambda F(X = x_2) - [N_1 \ln Z + N_2 \ln Z] \\
&= \ln W(N) + N_1 \lambda F(X = x_1) + N_2 \lambda F(X = x_2) - N \ln Z \\
&= \ln W(N) + \lambda [N_1 F(X = x_1) + N_2 F(X = x_2)] - N \ln Z \\
\frac{\ln P(N_1, N_2)}{N} &= \frac{\ln W(N)}{N} + \lambda \left[\frac{N_1}{N} F(X = x_1) + \frac{N_2}{N} F(X = x_2) \right] - \ln Z \\
\bar{F} &= \frac{N_1}{N} F(X = x_1) + \frac{N_2}{N} F(X = x_2) \\
\frac{\ln P(N_1, N_2)}{N} &= \frac{\ln W(N)}{N} + \lambda \bar{F} - \ln Z \\
P(N_1, N_2 | \mathcal{M}_k) &= \exp \left[N \left(\frac{\ln W(N)}{N} + \lambda \bar{F} - \ln Z \right) \right]
\end{aligned}$$

At this juncture, it is imperative to check that this new (but still unfinished) expression for the probability of the future frequency counts coincides with the previous calculation. The problem specified these particulars for the future frequency counts,

$$N = 100$$

$$N_1 = 30$$

$$N_2 = 70$$

and these values for the constraint function and the model parameter for the one model that inserted its information into the resulting probability distribution of $Q_1 = 0.20$ and $Q_2 = 0.80$,

$$\epsilon = 50$$

$$\lambda = 0.027726$$

Now it is just a matter of substituting these values into the various terms of the reworked expression,

$$\begin{aligned}
\ln Z &= \ln \left(1 + e^{50\lambda} \right) \\
\bar{F} &= \left(\frac{30}{100} \times 0 \right) + \left(\frac{70}{100} \times 50 \right) \\
&= 35 \\
W(N) &= \frac{100!}{30! 70!} \\
\frac{\ln W(N)}{N} &= 0.5864 \\
P(N_1, N_2 | \mathcal{M}_k) &= \exp \left[N \left(\frac{\ln W(N)}{N} + \lambda \bar{F} - \ln Z \right) \right] \\
&= \exp [100 (0.5864 + (0.027726 \times 35) - \ln (1 + \exp [50 \times 0.027726]))] \\
&= 0.00519
\end{aligned}$$

Apparently, the long series of symbolic manipulations was not in error since we arrive at the same probability for 30 HEADS and 70 TAILS in 100 future coin tosses. The important expressions to take note of in the subsequent development are the terms inside the parentheses, $\ln W(N)/N$, $\lambda \bar{F}$, and $\ln Z$. It is not too hard to see that these terms involved in the probability for $N = 100$ future frequency counts can be confused with similar terms arising in the MEP formula for determining the numerical assignments to the probability for the $n = 2$ statements in the state space.

Exercise 17.7.10: Consider an “assembly” of N coins where each coin is considered as one “system.” Show the interrelationship of the various combinatorial formulas for the sample space.

Solution to Exercise 17.7.10

As we will see in Chapter Twenty Seven, this strange language employing words like “assembly” and “system” appears here because it was used by Schrödinger in his explication of statistical mechanics. For an easy numerical example, examine an assembly of $N = 4$ systems. We are simply talking about four coins tossed at the same time. The dimension of the state space remains at $n = 2$. The total number of elementary points in the *sample space* is $n^N = 2^4 = 16$.

This number $n^N = 16$ can be broken down into a sum of multiplicity factors,

$$n^N = \sum_{j=1}^u W_j(N)$$

where the upper limit u to the sum is the total number of possible frequency counts,

$$u = \frac{(N+n-1)!}{N!(n-1)!} = 5$$

Thus, the sum consists of five multiplicity factors,

$$n^N = \sum_{j=1}^5 W_j(N) = 16$$

Explicitly, these five multiplicity factors for the five possible frequency counts in the assembly of $N = 4$ systems are attached to the events of all four HEADS and no TAILS, three HEADS and one TAILS, and so on.

$$\begin{aligned} W_1(N_1 = 4, N_2 = 0) &= 1 \\ W_2(N_1 = 3, N_2 = 1) &= 4 \\ W_3(N_1 = 2, N_2 = 2) &= 6 \\ W_4(N_1 = 1, N_2 = 3) &= 4 \\ W_5(N_1 = 0, N_2 = 4) &= 1 \end{aligned}$$

Failure to understand the relationships involved in n , n^N , $\frac{(N+n-1)!}{N!(n-1)!}$, and $W(N)$ leads to much confusion surrounding the MEP.

Exercise 17.7.11: Point out what seem to be the most salient aspects in disambiguating information as understood by the practitioners of the MEP and those who are wedded to a frequency interpretation.

Solution to Exercise 17.7.11

The state space is well-defined with dimension n . The IP relies on the MEP to find a numerical assignment of a legitimate probability for all n statements in the state space. This assignment is found by conditioning on a model \mathcal{M}_k which inserts information into the distribution of probabilities Q_i .

The information is well-defined in the sense of a mathematical expectation of arbitrary constraint functions mapping statements to numbers. The information is NOT the data. Conceptually and fundamentally, the information exists completely independently of the data. The data may as well not even exist as far as the MEP is concerned.

The MEP is employed *just once* to make assignments at the level of the state space. After this task has been accomplished, the MEP is no longer needed at any further stage of the inferential process.

Frequency counts, both future and past, are handled through the formal manipulation rules of probability. The most basic relationship is,

$$P(N_1, N_2, \dots, N_n) = \sum_{k=1}^{\mathcal{M}} P(N_1, N_2, \dots, N_n | \mathcal{M}_k) P(\mathcal{M}_k)$$

In the example just discussed, this took the form of,

$$P(N_1, N_2, \dots, N_n) = \int \cdot \int W(N) \prod_{i=1}^n q_i^{N_i} \delta(q_i - Q_i) dq_i = W(N) \prod_{i=1}^n Q_i^{N_i}$$

To the frequentist, the *sample* space, the data, the multiplicity factor, and the sample averages are all seen as important factors. The data *are* the information. The total number of observations N is important. Finding the average frequency counts is paramount.

To the MEP practitioner, on the other hand, it is not the *sample* space, but rather the *state* space and its dimension n that is important. The total number of observations N does not even exist. The Shannon entropy, not the multiplicity factor, is the objective function to be maximized.

The mathematical expectation of the constraint functions, that is, the average $\langle F \rangle$, with respect to the Q_i , and its associated dual parameter, the Lagrange multiplier λ is critical, not the sample average, \bar{F} , taken with respect to N_i/N . Finding the Q_i representing the degree of belief that the statements in the state space are true is the overriding concern. Once the Q_i are found, all else follows from the formal rules. Most especially, any inference regarding frequency counts, both past and future, follow from these rules.

By contemplating these last few exercises, we obtain clues about the places where things tend to get confused in the two camps. What is the state space? What is the sample space? What is the objective function that gets maximized? What kind of averages are the constraints? What is the role of $\ln Z$?

Exercise 17.7.12: Clarify exactly what probabilities we are talking about.

Solution to Exercise 17.7.12

Are we talking about the probabilities assigned to the statements in the state space under the information from some given model? Are we talking about the probabilities of future frequency counts bereft of any previous data, together with the assumption of being totally ignorant about the causes? Are we talking about the probabilities of future frequency counts without any data, but not uninformed about the causes? Are we talking about the probabilities of future frequency counts taken in the context of having observed some data? Are we talking about just the very *next* trial taken in the context of having observed some data?

The MEP finds the Q_i that are the probabilities assigned to the $n = 2$ statements in the state space under the restriction of the information from some given model.

If we want the probabilities for $N = 100$ future counts of HEADS and TAILS having observed no data, and furthermore consider ourselves completely uninformed about the relative status of what is causing HEADS or TAILS, then the probability of $P(N_1, N_2) = 1/101$ is the state of knowledge for every possible frequency count.

If we want the probabilities for $N = 100$ future counts of HEADS and TAILS, but we do not consider ourselves totally ignorant about what is causing HEADS or TAILS, and, in fact, we consider ourselves so well-informed that one model is sufficient for the causal effects, then the probability is,

$$P(N_1, N_2) = W(N) Q_1^{N_1} Q_2^{N_2}$$

If we want the probabilities for $N = 100$ future frequency counts, and once again do not consider ourselves totally ignorant about what is causing HEADS or TAILS, but not so well-informed that one model is sufficient for the causal effects, then the probability is,

$$P(N_1, N_2) = \sum_{k=1}^{\mathcal{M}} P(N_1, N_2 | \mathcal{M}_k) P(\mathcal{M}_k)$$

If we want the probabilities for $M = 100$ future frequency counts, and possess data from $N = 100$ previous coin flips having started out before the data completely uninformed about the causes, then the probability is,

$$P(M_1, M_2 | N_1, N_2)$$

If we want the probability for the very next occurrence of HEADS, still having observed $N = 100$ previous coin flips, as well as having started out before the data completely uninformed about the causes, then the probability is,

$$P(M_1 = 1, M_2 = 0 | N_1, N_2) = \frac{N_1 + 1}{N + 2}$$

Exercise 17.7.13: How is this scenario explained in statistical mechanics?

Solution to Exercise 17.7.13

This final exercise was inspired by an explanation of the “canonical ensemble” from a text on statistical mechanics by Chandler [5, pp. 66–68]. It is also serves as another example of the “Fermi oscillator.”

At this early stage, we are merely going to give the solution the same “flavor” as it would be explained in statistical mechanics. However, you will lose nothing by simply viewing the analysis as yet another example of the humble coin tossing

experiment where the MEP assigns numerical values to the probabilities for HEADS and TAILS.

Within statistical mechanics, the N flips of the coin are now to be thought of as N distinguishable particles, each of which can exist in two states separated by an energy ϵ . As belabored in Volume I, N may refer to N trials of one coin over time, or N coins flipped at the same time. So, in a closer analogy to statistical mechanics, suppose that $N = 100$ coins have all been tossed at the same time.

Both cases, that is, either the same coin flipped repeatedly 100 times, or 100 coins all flipped at once, are considered to be the same “assembly.” This is allowed within our framework where an “assembly” can be composed of separate “systems” where the “systems” can be coins, dice, kangaroos, students, or atoms. The “assembly” is made up of separate “systems,” courtesy of the definition of the sample space. The system at each trial, or as its distinguishable self among N such systems, may exist in only one of the n different states.

Let us try to be clear about the appropriate state space. In our application of the MEP formula to the coin toss scenario, the state space had dimension of $n = 2$. The statement ($X = x_1$) was, “HEADS showed after the toss.”, while ($X = x_2$) was, “TAILS showed after the toss.”

As part of the MEP formalism, we had to define a mapping for each of these two statements in the state space to numbers, $F(X = x_1)$ and $F(X = x_2)$. So, in analogy to the way the problem is set up in statistical mechanics, we may define a mapping from HEADS to 0 and a mapping from TAILS to ϵ ,

$$F(X = x_1) = 0 \text{ and } F(X = x_2) = \epsilon$$

The partition function is then defined by,

$$Z = \sum_{i=1}^2 e^{\lambda F(X=x_i)} = e^{(\lambda \times 0)} + e^{(\lambda \times \epsilon)} = 1 + e^{\lambda \epsilon}$$

The probability for HEADS under this model is then,

$$P(X = x_1 | \mathcal{M}_k) = \frac{1}{1 + e^{\lambda \epsilon}}$$

and the probability for TAILS,

$$P(X = x_2 | \mathcal{M}_k) = \frac{e^{\lambda \epsilon}}{1 + e^{\lambda \epsilon}}$$

The denominator appearing in the probability for either a HEADS or a TAILS on each trial is $1 + e^{\lambda \epsilon}$. The probability for any number of HEADS and TAILS in N tosses of the coin must then involve the expressions $(1 + e^{\lambda \epsilon})^N$ or $N \ln(1 + e^{\lambda \epsilon})$.

However, in [5] the state of the assembly is specified by listing the outcome of the $N = 100$ tosses of the coin in sequential trial order, such as by the list,

$$(\text{HEADS}_1, \text{TAILS}_2, \dots, \text{HEADS}_t, \dots, \text{TAILS}_N)$$

or as in,

$$(n_1, n_2, \dots, n_s, \dots, n_N) \equiv (0, 1, \dots, 1, 0)$$

where each n_s , the “quantum number,” is either 0 or 1 corresponding to HEADS or TAILS. Each n_s must also be the same dimension as the state space, that is, it can only take on two values.

Again, in analogy with statistical mechanics, the total energy E of this assembly could then range from 0ϵ when all HEADS occurred up to a maximum of 100ϵ when all TAILS occurred.

In the previous exercises, we asked about the assembly of $N = 100$ coin tosses consisting of 30 HEADS and 70 TAILS, with ϵ arbitrarily set at 50. Thus, the total energy of such a system is,

$$E = (30 \times 0) + (70 \times 50) = 3500$$

Over $N = 100$ trials, the joint probability for any particular sequence of HEADS and TAILS will then have this expression in the denominator,

$$[1 + \exp(\lambda\epsilon)]^N = [1 + \exp(50\lambda)]^{100}$$

in the denominator. The log transform of this denominator in the joint probability for N trials is now,

$$N \ln Z = N \ln [1 + \exp(\lambda\epsilon)] = 100 \ln [1 + \exp(50\lambda)]$$

We saw in Exercise 17.7.9 that such a term involving $N \ln Z$ arising from the MEP formula does appear in the probability for $P(N_1, N_2)$.

We can begin to feel ourselves starting to slide down a slippery slope of increasing confusion. Nonetheless, let us continue on to see where the explanation given by statistical mechanics eventually leads. One of the primary features of the MEP is that taking the derivative of $\ln Z$ with respect to the Lagrange multiplier returns the constraint function average. Despite the fact that everything is falling into disarray, take the derivative of $N \ln Z$,

$$E = \frac{\partial (N \ln Z)}{\partial \lambda}$$

to see that it does return the total energy of our assembly of 100 systems,

$$\begin{aligned} E &= \frac{\partial (N \ln Z)}{\partial \lambda} \\ &= \frac{\partial (100 \ln (\exp [1 + 50\lambda])))}{\partial \lambda} \\ &= \frac{5000 e^{50\lambda}}{1 + e^{50\lambda}} \\ E &= 3500 \end{aligned}$$

In order for the total energy to have the required value of,

$$E = (N_1 \times 0) + (N_2 \times \epsilon) = (30 \times 0) + (70 \times 50) = 3500$$

the Lagrange multiplier must equal $\lambda = 0.016946$.

You can discern that the overriding objective here was to obtain a “canonical ensemble” formula that “looked just like” our original MEP formula for the probability of HEADS and TAILS. Its goal was to return the frequentist definition for the probability of HEADS and TAILS based solely on observed data of 30 HEADS and 70 TAILS.

And indeed, this new Lagrange multiplier of $\lambda = 0.016946$ does result in a probability for HEADS of 0.3 and 0.7 for TAILS which obviously was the intent of the whole exercise from the outset. The “most probable frequencies” are then said to be,

$$\begin{aligned} \text{HEADS} &= N \times \frac{\exp[0.016946 \times 0]}{\exp[0.016946 \times 0] + \exp[0.016946 \times 50]} \\ &= 30 \\ \text{TAILS} &= N \times \frac{\exp[0.016946 \times 50]}{\exp[0.016946 \times 0] + \exp[0.016946 \times 50]} \\ &= 70 \end{aligned}$$

But this entire approach is different from the way the original problem was set up which returned $\lambda = 0.027726$ with $Q_1 = 0.20$, $Q_2 = 0.80$. A correct application of the rules of probability already gives one the answer that this approach mistakenly tries to derive, namely,

$$P(M_1 = 1, M_2 = 0 | N_1 = 30, N_2 = 70) = \frac{N_1 + 1}{N + 2} = \frac{31}{102}$$

We experience here a harbinger of the attendant confusion when information is defined as *data* represented in N known outcomes versus information defined by a constraint function average. The latter does not depend on data in any sense.

The MEP principles, which should be applied at the level of the $n = 2$ state space, are instead applied to the much larger total energy space. Everything is couched in terms of the *data*, N , N_1 , and N_2 , and the total energy of the assembly,

$$E = [N_1 \times F(X = x_1)] + [N_2 \times F(X = x_2)] = 3500$$

as opposed to the *information* in the constraint function expectation for *one* model,

$$\langle F(X = x_i) \rangle \equiv [F(X = x_1) \times Q_1] + [F(X = x_2) \times Q_2]$$

where, for example, our previous model in Exercise 17.7.8 inserted the information that,

$$\langle F \rangle = 40$$

The intent of the frequentist mind-set is to select the so-called “most probable distribution of counts” satisfying both $E = 3500$, and maximum multiplicity. That is, the intent is to select *one* set of frequency counts $N_1 = 30$ and $N_2 = 70$.

The MEP practitioner, on the other hand, calculates the *probability* for *any* possible future frequency count as conditioned on some model. This IP finds, for example, that $P(N_1 = 30, N_2 = 70 | \mathcal{M}_k) = 0.00519$.

If, instead, a different model had inserted the information $\langle F \rangle = 35 \equiv \overline{E} = 35$, then the *probability* for the particular frequency count of $N_1 = 30$ and $N_2 = 70$ would have had the largest probability of any of the 101 possible frequency counts.

Chapter 18

The MEP and Models for Coin Tossing

18.1 Introduction

In the coin tossing example of the last Chapter, an IP could have easily assigned numerical values to the probabilities $P(X = x_1)$ and $P(X = x_2)$ directly. But for real world problems in scientific inference, directly assigning numerical values to probabilities becomes prohibitive.

What is needed is a defensible technique allowing an IP to assign legitimate numerical values to an extremely large number of abstract probabilities. We have chosen to label these numerical values as Q_i when they are assigned under the information resident in some model. Moreover, it would be an added bonus if such a technique were grounded in Information Theory, and relatively easy to program.

The Maximum Entropy Principle (MEP) is just such a technique. And it *has* been used to assign probabilities within a rather wide range of scientific disciplines. We continue our in-depth, but rather slow and discursive discussion of the MEP, by continuing to concentrate on coin tossing.

The more technical aspects involved in the derivation of the MEP will be deferred to a later Chapter. At this point, we have faith that the MEP will take over the job of assigning legitimate numerical probabilities, no matter how complicated the real world problem.

An IP will always face the practical question of how to assign numerical values to probabilities when making inferences. The rules of probability theory handle the *symbol manipulation* aspects of the problem. We have encountered many examples of these formal manipulation rules already in our use of the **Sum Rule**, **Product Rule**, and **Bayes's Theorem** in Volume I.

But probability theory never advances beyond the symbol manipulation stage. It is silent on the issue of assigning numerical values. Its domain is to take what is given, and then operate on the abstract symbols in order to promote the idea of generalizing logic. It could care less whether the numerical value assigned to a particular probability is 0.0001 or 0.9999.

The MEP, however, *does* address the issue of assigning numerical values. It accomplishes this objective by incorporating information the IP might wish to insert into a probability distribution. It satisfies all of the information requirements laid down by the IP. Interestingly, at the same time it is doing this, the MEP is also maximizing the entropy of the resulting probability distribution.

Roughly speaking, by maximizing the entropy of the probability distribution, the total amount of missing information reflected in this state of knowledge is maximized as well. The net result is that this particular probability distribution *includes* only the information the IP wanted to include. The MEP *excludes* all the information the IP didn't want to include in the probability distribution. These are very desirable features of the MEP.

All of this sounds very mysterious, and we shall not attempt to convince you of its efficacy in one large dose. The MEP will be introduced slowly and gradually in order to develop an intuitive feel for whether it is doing a good job of assigning probabilities. At this beginning stage, we make use of the MEP formula as if it came to us engraved on stone tablets from on High. Later, after some familiarity with the practical aspects of the MEP, we shall return to examine its theoretical underpinnings.

18.2 The MEP and the Coin Toss

To begin, let's delve a little bit more into how the MEP behaves in the coin tossing scenario. Everything is the same as it was initially presented in the beginning Chapter. However, many of the same things are gone over again so that concepts can sink in through repetition.

The official technical name for λ is the *Lagrange multiplier*. We shall also call λ a *parameter* of the model. Specifying the parameter is synonymous with specifying the model that is assumed true. All we need to know about the parameter right now is that it can be adjusted in order to make the probability distribution conform to the information the model wants to insert.

There are only two probability values we are trying to assign: Q_1 , the numerical value assigned as a probability for HEADS, and Q_2 , the numerical value assigned as a probability for TAILS. Because the state space has been defined to be of dimension $n = 2$, there are only two observables for $(X = x_i)$, $x_1 = \text{HEADS}$ or $x_2 = \text{TAILS}$.

$F(X = x_1)$ is the function mapping the *statement*, “The coin was tossed and it landed with HEADS up.” to some number. Likewise, $F(X = x_2)$ is the function mapping the *statement*, “The coin was tossed and it landed with TAILS up.” to some other number. To illustrate the abstractness of the mapping, change it from the examples used in the opening Chapter. For this example, let $F(X = x_1) = 3$ and $F(X = x_2) = -5$, instead of the previous $F(X = x_1) = 1$ and $F(X = x_2) = 2$, or $F(X = x_1) = 0$ and $F(X = x_2) = \epsilon$.

With this specification of the constraint function, the MEP formula tells us that the numerical assignments will be,

$$Q_1 = \frac{\exp(3\lambda)}{Z(\lambda)}$$

$$Q_2 = \frac{\exp(-5\lambda)}{Z(\lambda)}$$

$$Z(\lambda) = \exp(3\lambda) + \exp(-5\lambda)$$

Since a model \mathcal{M}_k will set the value of λ , and there are an infinite number of values for λ , there are an infinite number of models as well. If a model sets the value of λ at 0, then $Q_1 = Q_2 = 1/2$ just as before. Label such a model as \mathcal{M}_A .

Table 18.1 illustrates the heuristic method of adjusting λ , and then watching what happens to both Q_1 and Q_2 . It is also possible to monitor the behavior of the constraint function average $\langle F \rangle$ as well.

Table 18.1: *The effect that varying λ has on assigned numerical values and the average of the constraint function when the mapping has changed.*

λ	Q_1	Q_2	$\langle F \rangle$
$-\infty$	0	1	-5.00
-0.500	0.018	0.982	-4.86
-0.250	0.119	0.881	-4.04
-0.137	0.250	0.750	-3.00
-0.100	0.310	0.690	-2.52
0	0.500	0.500	-1.00
+0.100	0.690	0.310	+0.52
+0.137	0.750	0.250	+1.00
+0.250	0.881	0.119	+2.05
+0.500	0.982	0.018	+2.86
$+\infty$	1	0	+3.00

18.2.1 The constraint function average changes

First of all, we see that changing the constraint function did not affect the MEP algorithm's ability to assign any legitimate numerical value to the probabilities for the statements in the state space. What it did affect was the specification of the constraint function average to achieve the same assignment. In the previous example, specifying that $\langle F \rangle = 1.5$ was equivalent to $\lambda = 0$ and $Q_1 = Q_2 = 1/2$. Now, $\langle F \rangle$ must be specified as -1.00 to achieve the same goal.

Secondly, look at the respective variations in the dual parameters. λ extends across the real line from $-\infty$ to $+\infty$ in order to reach the two anchor point assignments of $Q_1 = 0$, $Q_2 = 1$ and $Q_1 = 1$, $Q_2 = 0$. However, this is more of a technical mathematical requirement than a practical one because values of λ from -1 to $+1$ cover most of the range in the numerical assignments from 0 to 1 .

The parameter dual to λ , the constraint function average $\langle F \rangle$, varied over the range from -5 to $+3$. This makes perfect sense because when $Q_1 = 0$, the constraint function average consists of just the one term,

$$F(X = x_2) \times Q_2 = -5$$

Likewise, when $Q_1 = 1$, the constraint function average consists of just the one term,

$$F(X = x_1) \times Q_1 = +3$$

Table 18.1 shows a discrete breakdown of the parameter λ into eleven categories. Each one of these particular λ settings defines a different model \mathcal{M}_k . But λ could have taken on an infinite number of values besides those shown in the discrete sampling. For every setting of λ , there would have been a model corresponding to that setting.

The sixth row where $\lambda = 0$ and $\langle F \rangle = -1.00$, we labeled as model \mathcal{M}_A . This is the model of a “fair” coin where $Q_1 = Q_2 = 1/2$. The eighth row where $\lambda = +0.137$ and $\langle F \rangle = +1.00$ is labeled as model \mathcal{M}_B . This is the model of a “biased” coin where $Q_1 = 3/4$ and $Q_2 = 1/4$. The coin is obviously biased to show HEADS when tossed. Both of these models that have been singled out for further discussion are shown boxed in the table.

18.3 Definition of Information Entropy

Where does the notion of entropy come into play? We begin with the definition of information entropy for a discrete probability distribution,

$$\text{Information entropy} = - \sum_{i=1}^n Q_i \ln Q_i \quad (18.1)$$

For our current example this simplifies to,

$$\text{Information entropy} = -(Q_1 \ln Q_1 + Q_2 \ln Q_2)$$

The MEP algorithm attempts to do two things simultaneously. It is not to chew gum and walk at the same time. It tries to satisfy all of the constraints that have been specified under a particular model, *and* maximize the information entropy of the resulting probability distribution.

We started off by using the phrase *information entropy* to distinguish this more general concept from the single word *entropy* as used in its classical physical meaning within the discipline of thermodynamics. From now on, though, we shall often just use entropy with the implied caveat that it refers to the information entropy of a probability distribution.

We labeled the model of a fair coin as \mathcal{M}_A . How well did the MEP accomplish these tasks for model \mathcal{M}_A ? There was only one constraint to satisfy, since by setting $\lambda = 0$, the constraint function $F(X = x_i)$ was rendered ineffective. There is, however, always an implicit constraint function lurking in the background, and that was called the *universal constraint*.

It simply enforces the fundamental requirement that the probabilities must sum to 1. This was satisfied because,

$$Q_1 + Q_2 = 1/2 + 1/2 = 1$$

There is nothing mysterious about the universal constraint. It is formally described in exactly the same way as any other constraint. The universal constraint $F_0(X = x_i)$ is mapped to 1 for every statement in the state space. The average of this constraint function is $\langle F_0 \rangle = 1$.

$$\langle F_0 \rangle = \sum_{i=1}^n F_0(X = x_i) Q_i$$

$$1 = Q_1 + Q_2 + \dots + Q_n$$

Secondly, this assignment under model \mathcal{M}_A must have the *maximum* entropy of *any* probability assignment that also satisfies the constraints. Remember that we are talking here about a model which inserts information only about the universal constraint. For example, $Q_1 = 3/4$ and $Q_2 = 1/4$ is another assignment that also satisfies the universal constraint that $Q_1 + Q_2 = 1$.

But the entropy of the MEP assignments $Q_1 = Q_2 = 1/2$ under \mathcal{M}_A is,

$$-(1/2 \ln 1/2 + 1/2 \ln 1/2) = 0.6931$$

while the entropy of the non-MEP assignments $Q_1 = 3/4, Q_2 = 1/4$ is,

$$-(3/4 \ln 3/4 + 1/4 \ln 1/4) = 0.5623$$

So this MEP assignment of $Q_1 = Q_2 = 1/2$ does possess greater entropy than any other alternative assignment that also satisfies the universal constraint. No matter which legitimate assignment you make to Q_1 and Q_2 , the MEP assignment lives up to its name. It always has the greater entropy. Try to find another assignment that satisfies the constraint $Q_1 + Q_2 = 1$, and has a higher entropy than 0.6931. It can't be done.

If you move further away from $Q_1 = Q_2 = 1/2$, say, to an assignment of $Q_1 = 0.1$ and $Q_2 = 0.9$, then the entropy of this completely legitimate assignment is,

$$\text{Information entropy} = -(0.1 \ln 0.1 + 0.9 \ln 0.9) = 0.3251$$

which is pretty obviously going in the wrong direction if you are trying to overtake the MEP assignment. Reversing the assignment to $Q_1 = 0.9$ and $Q_2 = 0.1$ leads to the same entropy value.

How about sneaking up on the MEP assignment with $Q_1 = 0.51$ and $Q_2 = 0.49$?

$$\text{Information entropy} = -(0.51 \ln 0.51 + 0.49 \ln 0.49) = 0.6929$$

Try as you might, you won't be able to find an assignment that beats the MEP assignment. Intuitively, the MEP assignment tries to spread out the Q_i probabilities as evenly as possible while still satisfying the constraints. The assignment under model \mathcal{M}_A of $Q_1 = Q_2 = 1/2$ is seen to be the smoothest, most even, most spread out assignment that satisfies the one constraint in effect, that one constraint being, of course, the universal constraint.

In all future discussions, the presence of the universal constraint is taken for granted. Thus, from now on, we feel free to talk about, say, the *single* constraint $\langle F \rangle$, all the while implicitly acknowledging that there is that ever present universal constraint lurking in the background as well.

The very same discussion applies when more information has been inserted by a particular model. Consider the model \mathcal{M}_B of the biased coin instead of the model \mathcal{M}_A of the fair coin. There are now two constraints to be satisfied, the universal constraint, and $\langle F \rangle = +1.00$. The resulting assignment of $Q_1 = 3/4$ and $Q_2 = 1/4$ *does* have the maximum entropy of any assignment that satisfies these *two* constraints.

The information entropy of model \mathcal{M}_B was calculated above as 0.5623. Now this value must be, by definition of the MEP algorithm, the very highest value of the information entropy of any possible legitimate assignment that satisfies all of the constraints that have been specified. While true, it turns out to be a moot point because there is only *one* assignment that can satisfy both constraints. Unlike the first case, there are no other legitimate numerical assignments that can simultaneously satisfy both assignments, and yet possess a lower entropy.

18.4 The Number of Constraints

If you are starting to think that we have used a sledgehammer to smash a pea, I would agree. If I ask you to assign a probability to a fair coin, you would naturally come up with the assignment that $P(\text{HEADS})=P(\text{TAILS})=1/2$. One would feel this to be correct without resorting to the MEP.

However, the utility of the MEP comes into play when we need to assign n probabilities when n is much larger than 2. In addition, if you are given $n - 1$ constraints for the n different ways that the event could occur at each trial, you also don't need the MEP to figure out the probabilities (the implicit presence of the universal constraint once again). The MEP does come into its own when the number of constraints, labeled as m , is much smaller in number than $n - 1$.

With only m constraints postulated under a particular model for the n different ways that the event could occur at each trial, some remaining ambiguity about how to assign legitimate probabilities is always present. For example, if $n = 3$, and the one constraint $Q_1 = 0.4$ is in place, then Q_2 could have a legitimate assignment of 0.4. From the universal constraint, Q_3 must then equal 0.2.

Or, Q_2 could equal 0.1 forcing Q_3 to equal 0.5, and so on for an infinite number of possibilities. The MEP is used when $m < n - 1$ to resolve any remaining ambiguities. In this case, the MEP resolves the ambiguity of the infinite number of possibilities by assigning $Q_2 = Q_3 = 0.3$. The entropy of this assignment, $-\sum_{i=1}^3 Q_i \ln Q_i$, is higher than any other assignment that also satisfies the constraint of $Q_1 = 0.4$.

18.5 Probability of Future Events

The remarks in this section are intended to ignite a debate over one of the most fundamentally confusing concepts that arises when the MEP finally meets up with inference. As mentioned in my *Apologia*, the debate is not over some arcane piece of mathematics, but simply whether a coherent line of thought has been followed.

The MEP's sole focus is on the statements in the state space. The probability for future frequency counts is a *calculation* based on the *previous* assignments as made by the MEP for the probabilities of the statements in the state space. These two operations are conceptually quite distinct, but unending confusion swirls around these two notions.

I shall give my best effort to dispel these clouds of confusion. But I know full well that such a hope is doomed to be dashed. The Biblical phrase pops to mind, "There is none so blind as he who will not see."

One source of the confusion stems from the fact that "entropic-like expressions" seem to spring up everywhere once the general notion of MEP has been released upon the world. We will give two introductory examples in this section.

A previous example, foreshadowing the assignment algorithm content of this Volume, was presented in section 14.3.1 of Volume I. Also, several exercises were worked out in Chapter Seventeen broaching the technical details involved in the distinction between an MEP assignment and future events.

By way of a refresher, look at what the formal rules for probability manipulation say about repeated flips of the coin. In Volume I, we left the derivation in the form of the Q_i because we could go no further at that point of the discussion. Since we now have the MEP assignment for the Q_i , we can substitute into the prediction formula for repeated trials.

Recapitulating, what is an IP's state of knowledge about some number of *future* trials involving the coin? We will show the expressions in terms of N and N_i instead of the correct M and M_i so as to make every one comfortable.

$$P(N_1, N_2) = \int_0^1 W(N) q_1^{N_1} q_2^{N_2} P(q_i) dq_i \quad (18.2)$$

If just one model is selected, then the probability for all of the models $P(M_k)$ reduces down to the Dirac δ function $\delta(q_i - Q_i)$. We then have the very special case of the binomial distribution.

$$P(N_1, N_2 | \mathcal{M}_k) = W(N) Q_1^{N_1} Q_2^{N_2} \quad (18.3)$$

This can be specialized even further when the MEP assigns $Q_i = 1/n$ under this one model which now is seen to be the “fair” model. The probability becomes a direct function of the multiplicity factor and the total number of points in the *sample space*,

$$P(N_1, N_2 | \mathcal{M}_k) = \frac{W(N)}{n^N} \quad (18.4)$$

This is the extreme polar opposite of Laplace's stated condition of an insufficient reason¹ for the probability of a cause. As the argument now goes, there is an extremely strong reason for adopting one model, and one model only, which assigns the numerical values of Q_i to the probabilities for HEADS and TAILS.

But there are other solutions to Equation (18.2). If all models assigning numerical values are considered to be of equal value, then the above binomial probability distribution is NOT the outcome. Instead, Laplace's argument prevails, and,

$$P(N_1, N_2) = \frac{1}{N+1} \quad (18.5)$$

¹According to the historians, Laplace never actually used a phrase similar to “insufficient reason.” It was apparently coined by Bernoulli and attributed to Laplace by later critics such as Boole, Venn, and Fisher as counterpoint to Leibniz's philosophical usage of the phrase “sufficient reason.”

On the other hand, if only those models which assign $Q_1 = 1$ or 0 are accorded equal value ($\alpha_i \rightarrow 0$ in the Dirichlet distribution), then,

$$P(N_1 = N, N_2 = 0) = P(N_1 = 0, N_2 = N) = 1/2 \quad (18.6)$$

We have to be cognizant of the fact that the binomial distribution is just one solution to the general formula in Equation (18.2). Prior to substituting the MEP assignments for Q_1 and Q_2 , take the logarithmic transform of the probability for the future frequency counts N_1 and N_2 based on the one model,

$$\begin{aligned} \ln [P(N_1, N_2 | \mathcal{M}_k)] &= \ln [W(N) Q_1^{N_1} Q_2^{N_2}] \\ &= \ln [W(N)] + N_1 \ln Q_1 + N_2 \ln Q_2 \end{aligned}$$

Because of the exponential nature of MEP formula, we can substitute in the log transforms for Q_1 and Q_2 , and then divide through by N ,

$$\begin{aligned} \frac{\ln [P(N_1, N_2)]}{N} &= \frac{\ln [W(N)]}{N} + \lambda \left[\frac{N_1}{N} F(X = x_1) + \frac{N_2}{N} F(X = x_2) \right] - \ln Z \\ \text{Let } \bar{F} &= \sum_{i=1}^n \frac{N_i}{N} F(X = x_i) \\ \frac{\ln [P(N_1, N_2)]}{N} &= \frac{\ln [W(N)]}{N} + \lambda \bar{F} - \ln Z \end{aligned} \quad (18.7)$$

After carrying out the divisions by N , we see that we have created a *sample* average based on future frequency counts. This operation is labeled as \bar{F} , and appears in the expression for the probability of future frequency counts as opposed to the informational average $\langle F \rangle$ appearing in the MEP assigned value for a single observation $P(X = x_i | \mathcal{M}_k)$. The right hand side of Equation (18.7) *looks* similar to expressions developed within the MEP formalism, especially when the first term approaches information entropy as N increases. Both $\ln [W(N)/N]$ and $\lambda \bar{F}$ are related to entropy expressions, so we call it an *entropic-like* expression.

We now have a version for the probability of the future frequency counts,

$$\begin{aligned} P(N_1, N_2 | \mathcal{M}_k) &= e^{N \times \text{Entropic-like expression}} \\ &= \exp \left[N \times \left(\ln \left[\frac{W(N)}{N} \right] + \lambda \bar{F} - \ln Z \right) \right] \end{aligned} \quad (18.8)$$

18.5.1 Example 1

As a quick check on this formula, does it provide the right answer for the probability of seeing two HEADS and two TAILS in four future flips of a fair coin? Adopting the binomial distribution as appropriate under one model, with that one model being the “fair” model, we know that this probability $P(N_1 = 2, N_2 = 2 | \mathcal{M}_k) = 3/8$.

The MEP formula will provide the numerical assignment of $Q_1 = Q_2 = 1/2$ based on $\langle F \rangle = 1.5$. Thus $Z = 2$ and $\lambda = 0$. In $N = 4$ future coin flips, the probability for seeing two HEADS and two TAILS is,

$$\begin{aligned} P(N_1 = 2, N_2 = 2 | \mathcal{M}_k) &= \exp \left[4 \times \left(\left[\frac{\ln W(4)}{4} \right] + (0 \times 1.5) - \ln 2 \right) \right] \\ &= 0.375 \end{aligned}$$

18.5.2 Example 2

Let’s look at the same example of repeated coin flips, but now ask what happens when N is large. Suppose $N = 10,000$, and we want the probability of 6,000 HEADS and 4,000 TAILS in 10,000 future coin flips. What are the perspectives taken from someone oriented to a frequentist viewpoint versus someone with a perspective anchored by the MEP?

From the MEP perspective, we focus on the dimension of the state space. This has not changed from the first example, and so remains at $n = 2$. Even though the problem seems altered when viewed from the perspective of someone with a frequentist perspective with its much larger N and N_i , its essential character has not changed from someone looking at it from the MEP perspective.

Working within the MEP perspective, some model \mathcal{M}_k will assign legitimate numerical values to Q_1 and Q_2 between 0 and 1 inclusive. In Example 1 above, the model had a parameter $\lambda = 0$, and constraint function vector of $F(X = x_i) = (1, 2)$. From the MEP perspective, it’s quite allowable to change the constraint function vector to, say, $(0, \epsilon)$ as was done in Exercise 17.7.7.

There an assigned numerical value to a probability for HEADS under some model was,

$$P(X = x_1 | \mathcal{M}_k) \equiv Q_1 = \frac{1}{1 + e^{\lambda \epsilon}}$$

Suppose that the parameter λ is changed to $-\frac{1}{T}$, so that under this new model,

$$P(X = x_1 | \mathcal{M}_k) \equiv Q_1 = \frac{1}{1 + e^{-\epsilon/T}}$$

It is clear that the partition function must be,

$$Z(T) = 1 + e^{-\epsilon/T}$$

If the information in this new model \mathcal{M}_k specifies, say, $\langle F \rangle = 5$, where we are still using a value of $\epsilon = 50$, then $Q_1 = 0.90$ and $Q_2 = 0.10$. The value of the dual parameter is $\lambda = -0.0439$ with T then equal to $T = 22.76$. Check that,

$$P(\text{HEADS} \mid \mathcal{M}_k) \equiv Q_1 = \frac{1}{1 + e^{-50/22.76}} = 0.90$$

The MEP has accomplished its objective with this assignment to the Q_i . There is nothing more for it to do. Everything else from this point forward is a consequence of the formal manipulation rules. Therefore, we will use Equation (18.8) to find $P(N_1 = 2, N_2 = 2 \mid \mathcal{M}_k)$.

Pay special attention to the fact that the information inserted by the IP into the distribution was defined as $\langle F \rangle = 5$ or $T = 22.76$. Neither of these parameters had anything to do with any data. However, we do use \bar{F} , the averages taken with respect to the data *if these were to occur*, in the calculation of the future occurrence of HEADS and TAILS.

Also, pay special attention to the fact that in deriving the Q_i via the MEP, we did not have to refer to the multiplicity factor $W(N)$. This is just another way of saying that the information used by the MEP did not involve the data. However, once again, from the frequentist perspective in the calculation of the future occurrence of HEADS and TAILS, the multiplicity factor explicitly appears.

A confusing feature, as mentioned before, is that while we note the distinction between $\langle F \rangle$ and \bar{F} , together with the necessity of taking account of, or ignoring, the multiplicity factor, $\ln Z$ is common to both perspectives. From the frequentist perspective, N does play a role in the multiplication via $N \ln Z$, but this is hidden away again when the division by N takes place.

In any case, the calculation of $P(N_1, N_2 \mid \mathcal{M}_k)$ must result in the same answer whether it is approached from either perspective. We would rightly suspect that the probability for an event specifying exactly 6000 HEADS and 4000 TAILS, especially when the model assigns $Q_1 = 0.90$ and $Q_2 = 0.10$ must be very small.

$$\begin{aligned} P(N_1 = 6000, N_2 = 4000) &= \exp \left[N \times \left(\left[\frac{\ln W(N)}{N} \right] + \lambda \bar{F} - \ln Z \right) \right] \\ &= 1.65 \times 10^{-1354} \end{aligned}$$

This number certainly qualifies as a small probability. This answer with all of the numerical details filled in is worked out in Exercise 18.7.12.

This is the correct answer. There can be no quibbling over it. This is the correct probability for 10,000 coin flips *that have not yet taken place* when conditioned on the truth of one model assigning numerical values of $Q_1 = 0.90$ and $Q_2 = 0.10$. But the confusion, of course, is at the conceptual level.

After mulling over this result for a while, the person with the frequentist viewpoint would object. “No, no, no, this is not what I meant,” they would protest. “I

want to assign the probability based on the observed data of 6000 HEADS and 4000 TAILS. I want to end up with a model with probabilities very close to $Q_1 = 0.60$ and $Q_2 = 0.40$.” From the frequentist viewpoint and the different conceptualization, the data for them have already taken place; they are not future observations.

In general, what the frequentist wants to do is *start* from a different state space involving N rather than n . Then, he wants to maximize the multiplicity factor $W(N)$ instead of the information entropy. Finally, the constraints are on the total N and the averages with respect to the already observed data \bar{F} rather than constraints stipulating probabilities summing to 1 and expectations of constraint functions $\langle F \rangle$ with respect to the desired probability distribution.

From the MEP perspective, the response would be, “Oh, you want the probability for either the next event or many next events as conditioned on a vast number of already witnessed events. That is a different probability calculation than the one conducted above. But one which, again following the formal manipulation rules, I can carry out.”

What the frequentist is searching for, in the eyes of one who recognizes the distinction between the formal manipulation rules and the MEP, is the probability for seeing any number of future frequency counts as conditioned on a *known* number of past frequency counts. This probability is found by once again applying the formal rules.

This derivation was carried out in Volume I where the predictive formula,

$$P(M_1, M_2, \dots, M_n | N_1, N_2, \dots, N_n)$$

specifies the probability of M_1, \dots, M_n future frequency counts when conditioned on N_1, \dots, N_n past frequency counts.

In our current example, our protagonist for the frequentist position would be much happier with Laplace’s *Rule of Succession*. The probability for seeing HEADS on the very next coin toss after having observed 6000 HEADS and 4000 TAILS in the previous 10,000 tosses is,

$$P(M_1 = 1, M_2 = 0 | N_1 = 6000, N_2 = 4000) = \frac{N_1 + 1}{N + n} = \frac{6001}{10002}$$

The irony here is that this (correct) answer is arrived at by first apportioning equal credence to every conceivable model assigning numerical values to HEADS and TAILS. Laplace placed that uniform distribution over the models because of “insufficient reason” to believe in any one model to the exclusion of all the others.

Thus, *every* MEP assignment is averaged over. This averaging allows the data to re-order all of the models from their initial order of uniformity. Therefore, the “Bayesian approach,” or should we say Laplace’s approach, does a far better job of instantiating our intuitive sense of frequencies in relation to probabilities than the straightforward frequentist approach!

18.6 Connections to the Literature

I highly recommend Baierlein's **Atoms and Information Theory** ([2], Chapters 2 and 3), for a very similar treatment highlighting the conceptual distinction between probabilities and frequencies. As a physicist trying to explain statistical mechanics, he was very much influenced, as I was, by Jaynes's writings.

He also emphasizes, as I rarely see elsewhere, that entropy is a quantitative measure of *missing information*. According to Baierlein, the notation Z for the sum over states derives from Planck, while the label "partition function" comes from the British physicists Charles Galton Darwin (the grandson of the more famous Charles), and Ralph H. Fowler *circa* the 1920s.

Shannon's publication [31] of his **Theory of Mathematical Communication** in the late 1940s ushered in the era of Information Theory. The tale is often told that Shannon asked John von Neumann, the eminent mathematician, physicist, computer scientist, and, let's face it, general all around Renaissance Man, what he should call his new information measure. Reportedly, von Neumann immediately saw its similarity to the formula in statistical mechanics, and suggested calling it "entropy." Von Neumann, tongue firmly inserted in cheek, told Shannon: Nobody understands what entropy means, so by employing this impressive word in your arguments you will be sure to win every time.

Jaynes and others have commented that maybe in the long run things would have turned out better if Shannon had ignored von Neumann's light-hearted quip, and chosen another name. But that didn't happen, and we are stuck with the attendant confusion between physicists who want to use the word in its original ontological sense, and those who understand that information entropy refers to the epistemological state of an information processor.

It is interesting that Shannon decided to declare that there were three properties that an information measure should possess. This is where we see the wonderful and delightful manifestation of a creative act. What I mean by this praise is that it is not so much the actual information entropy formula in itself, but rather the mystery of why Shannon should stumble upon these particular properties in order to uniquely characterize information.

The first two properties as given by Shannon [31, pg. 49] seem, after the fact, relatively simple, straightforward, and quite acceptable, even if we ourselves might never have thought of them. These first two properties are,

1. H should be continuous in the p_i .
2. If all the p_i are equal, $p_i = 1/n$, then H should be a monotonic increasing function of n . With equally likely events there is more choice, or uncertainty when there are more possible events.

We have already seen examples where the information measure H continuously (and smoothly) varies as the Q_i are changed ever so slightly. As far as the second property is concerned, if the dimension of the state space increases from $n = 2$ to $n = 3$ to $n = 4$, the fair model in each case assigns $Q_i = 1/2, 1/3$, or $1/4$. The information entropy $-\sum_{i=1}^n Q_i \ln Q_i$ in each case increases from 0.693147 to 1.09861 to 1.38629. The increase in each case is from $\ln 2$ to $\ln 3$ to $\ln 4$, or, in general, to $\ln n$.

The third property is not as immediately clear. And everybody has to work out some examples to “convince” themselves of its efficacy. This third property that Shannon said an information measure should possess is:

3. If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H .

We will repeat Shannon’s own example of this third property. Consider a model for a state space of dimension $n = 3$ where the assignments are,

$$Q_1 = 1/2, Q_2 = 1/3, \text{ and } Q_3 = 1/6$$

The information entropy of this assignment is,

$$-\sum_{i=1}^n Q_i \ln Q_i = -[(1/2 \ln 1/2) + (1/3 \ln 1/3) + (1/6 \ln 1/6)] = 1.0114$$

Here is the tricky part. Shannon says to break down this original assignment into two assignments. The original state space is reduced from $n = 3$ to $n = 2$ for the new assignments with a sum to compensate for the reduction in the state space. We now have an assignment of $Q_1 = 1/2$ and $Q_2 = 1/2$ with its information entropy as $H(1/2, 1/2) = 0.693147$.

The original $Q_2 = 1/3$ and $Q_3 = 1/6$ assignment is changed into $1/2 \times (2/3, 1/3)$. This new assignment to a state space of $n = 2$ has an information entropy of $H(2/3, 1/3) = 0.636514$.

The weighted sum of these two separate information entropies should then equal the original information entropy for the assignments in the $n = 3$ state space,

$$\begin{aligned} H(Q_1, Q_2, Q_3) &= 1.0114 \\ H(Q_1, Q_2, Q_3) &= H(1/2, 1/2) + 1/2 H(2/3, 1/3) \\ &= 0.693147 + (1/2 \times 0.636514) \\ &= 1.0114 \end{aligned}$$

From these three properties, Shannon derives, in his Appendix 2, the information entropy expression for which he became so famous.

I would like to draw attention to the fact that nowhere in Shannon's axiomatic derivation of information entropy is there any mention of data playing a role. Nor did Shannon see any need to make mention of any extraneous "prior probability" or "prior measure" that was distinct and separate from the Q_i . These remarks foreshadow the coming firestorm when Solomon Kullback's *relative entropy* enters the fray.

Jaynes [14], in an appendix to his first paper, also discusses Shannon's three axioms. He suggests "amount of uncertainty" as a term synonymous for the phrase "missing information" which, in my opinion, has tended to create, shall we say, some uncertainty.

18.7 Solved Exercises for Chapter Eighteen

Exercise 18.7.1: Calculate the probability assignment for HEADS using the MEP formula under model \mathcal{M}_B mentioned in section 18.2.1.

Solution to Exercise 18.7.1

The short form for the numerical assignment to the probability for the i^{th} statement in the state space under some given model is,

$$Q_i = P(X = x_i | \mathcal{M}_k)$$

For our current exercise, this can be written as,

$$Q_1 = P(\text{HEADS} | \mathcal{M}_B)$$

where the information inserted into the probability distribution under \mathcal{M}_B is that $\langle F \rangle = 1$. The new vector containing the constraint function mappings from the statements in the state space is $(3, -5)$.

The MEP formula to calculate Q_i with just one constraint function is then,

$$Q_i = \frac{\exp [\lambda F(X = x_i)]}{Z(\lambda)}$$

where the normalization factor is the partition function $Z(\lambda)$,

$$Z(\lambda) = \sum_{i=1}^{n=2} \exp [\lambda F(X = x_i)]$$

When the parameter $\langle F \rangle = 1$, the dual parameter $\lambda = 0.1373265$. The numerical assignment for a probability of HEADS is,

$$\begin{aligned} P(\text{HEADS} | \mathcal{M}_B) &= \frac{\exp [0.1373265 \times 3]}{Z(\lambda)} \\ &= \frac{1.5098}{Z(\lambda)} \\ Z(\lambda) &= \exp [0.1373265 \times 3] + \exp [0.1373265 \times -5] \\ &= 2.0131 \\ P(\text{HEADS} | \mathcal{M}_B) &= \frac{1.5098}{2.0131} \\ &= 0.75 \end{aligned}$$

Exercise 18.7.2: Use the *Mathematica* function `Solve[]` (Appendix A) to find the value of the Lagrange parameter in the previous exercise.

Solution to Exercise 18.7.2

The MEP formula tells us that if we want Q_1 to have a value of 0.75 under model \mathcal{M}_B , then λ is going to satisfy,

$$Q_1 = \frac{e^{3\lambda}}{e^{3\lambda} + e^{-5\lambda}} = 0.75$$

Have *Mathematica* evaluate,

```
Solve[Exp[3 λ] / (Exp[3 λ] + Exp[-5 λ]) == .75, λ]
```

to find that $\lambda = 0.1373265$.

Exercise 18.7.3: Use the same *Mathematica* function to find the value of the Lagrange multiplier that would result in the assignment of $Q_1 = 0.47$ in Exercise 17.7.5.

Solution to Exercise 18.7.3

In that exercise, the value of the Lagrange multiplier was given as $\lambda = -1$, and the problem was to find all four Q_i . Q_1 was calculated to be the assignment 0.483535. It seems plausible that if we want to find $Q_1 = 0.47$, then λ should be close to -1 .

Turn the problem around, and ask *Mathematica* what value of the Lagrange multiplier results in $Q_1 = 0.47$,

```
Solve[Exp[λ] / (Exp[λ] + Exp[λ] + Exp[4 λ] + Exp[5 λ]) == .47, λ]
```

and it reports back that $\lambda = -0.808911$.

Exercise 18.7.4: In section 18.2, what is the information entropy of the probability assignment under a model where the information is that the constraint function average is $\langle F \rangle = -4$?

Solution to Exercise 18.7.4

To orient ourselves, and to make sure we don't blunder, refer back to Table 18.1 to see that the assigned probability for HEADS under this condition must be slightly greater than 0.119. $Q_1 = 0.119$ was the assignment when $\langle F \rangle = -4.04$. If the parameter $\langle F \rangle = -4$, then the dual parameter $\lambda = -0.243239$.

The MEP formula provides us with an assignment of $Q_1 = 0.125$ and $Q_2 = 0.875$. The numerical value of the probability for HEADS under this model is $1/8$, and the numerical value of the probability for TAILS is $7/8$.

To make sure, double-check that the constraint function average is indeed -4 .

$$\begin{aligned}\langle F \rangle &= \sum_{i=1}^2 F(X = x_i) Q_i \\ &= (3 \times 1/8) + (-5 \times 7/8) \\ &= -4\end{aligned}$$

With these assignments under the specified model in hand, we can calculate the information entropy as,

$$H(Q_i) = -(1/8 \ln 1/8 + 7/8 \ln 7/8) = 0.37677$$

Exercise 18.7.5: Make a general comment on the relationship between the dual parameters from the results in Table 18.1.

Solution to Exercise 18.7.5

There is a strong non-linear relationship between the dual parameters. To change from an assignment of $Q_1 = 0.018$ to $Q_1 = 0$, the constraint function average $\langle F \rangle$ changes from -4.86 to -5.00 . To accomplish the same objective, the Lagrange multiplier λ must undergo a vast change from -0.500 to $-\infty$.

Exercise 18.7.6: What is the probability expression for seeing TAILS, followed by two HEADS in three future flips of the coin?

Solution to Exercise 18.7.6

Using the same convention as in Volume I where a subscript t attached to X indicates the trial number, the joint probability expression for $P(X_t)$ where $t = 1, 2, 3$ is written as,

$$P(X_1 = x_2, X_2 = x_1, X_3 = x_1)$$

Exercise 18.7.7: What do the formal manipulation rules for probability have to say about this expression?

Solution to Exercise 18.7.7

Since probability inherits the **Commutativity** axiom from Boolean Algebra, the statements can be conveniently reordered. The probability for seeing this future event is a marginal probability over the model space. Consider first the summation

over \mathcal{M} distinct models by the **Sum Rule**,

$$P(X_3, X_2, X_1) = \sum_{k=1}^{\mathcal{M}} P(X_3, X_2, X_1, \mathcal{M}_k)$$

Then, use the **Product Rule**,

$$P(X_3, X_2, X_1) = \sum_{k=1}^{\mathcal{M}} P(X_3 | X_2, X_1, \mathcal{M}_k) \times P(X_2 | X_1, \mathcal{M}_k) \times P(X_1 | \mathcal{M}_k) \times P(\mathcal{M}_k)$$

When a model \mathcal{M}_k is given as true to the right of the conditioned-upon symbol, then the probability of the statement to the left of the conditioned-upon symbol is independent of everything except for the model.

Thus, the probability at any trial is independent of HEADS or TAILS at any previous trial, and dependent only on the model. The expression for the joint probability becomes,

$$P(X_3, X_2, X_1) = \sum_{k=1}^{\mathcal{M}} P(X_3 | \mathcal{M}_k) \times P(X_2 | \mathcal{M}_k) \times P(X_1 | \mathcal{M}_k) \times P(\mathcal{M}_k)$$

The short form for $P(X_t = x_i | \mathcal{M}_k)$ is Q_i , so,

$$P(X_3 = x_1, X_2 = x_1, X_1 = x_2) = \sum_{k=1}^{\mathcal{M}} [Q_1 \times Q_1 \times Q_2 \times P(\mathcal{M}_k)]$$

Exercise 18.7.8: What information can be used to select three models to average over in the last exercise?

Solution to Exercise 18.7.8

The new wrinkle afforded by our exposure to the MEP algorithm is a rationale for why the probability for statements are assigned particular numerical values. Suppose we use the constraint function defined in this Chapter for the coin toss. Let the information under the three different models, \mathcal{M}_C , \mathcal{M}_A , and \mathcal{M}_B , be the respective constraint function averages $\langle F \rangle_C = -3$, $\langle F \rangle_A = -1$, and $\langle F \rangle_B = +1$.

Referring back to Table 18.1, we can see that the numerical assignments to the probability for HEADS and TAILS under these three models are,

$$Q_1^C = 0.25 \quad \text{and} \quad Q_2^C = 0.75$$

$$Q_1^A = 0.50 \quad \text{and} \quad Q_2^A = 0.50$$

$$Q_1^B = 0.75 \quad \text{and} \quad Q_2^B = 0.25$$

If the IP is operating under total ignorance about the cause of a coin toss, then it must assign equal probability to all three models. Thus,

$$P(\mathcal{M}_A) = P(\mathcal{M}_B) = P(\mathcal{M}_C) = 1/3$$

The probability for the future event of a TAILS on the first toss followed by two HEADS on the second and third toss using the formula just developed, $P(X_1, X_2, X_3) = \sum_{k=1}^3 Q_1^2 Q_2 P(\mathcal{M}_k)$, is therefore,

$$\begin{aligned} P(X_1 = x_2, X_2 = x_1, X_3 = x_1) &= (0.50 \times 0.50 \times 0.50) \times 1/3 + \\ &\quad (0.75 \times 0.75 \times 0.25) \times 1/3 + \\ &\quad (0.25 \times 0.25 \times 0.75) \times 1/3 \\ &= 0.104167 \end{aligned}$$

Exercise 18.7.9: What must be certain in this coin tossing example?

Solution to Exercise 18.7.9

Whatever must be certain must have a probability of 1. We just calculated the probability for a TAILS followed by two HEADS. There are four distinct categories that break down what must happen in three flips of the coin.

These four categories are: (1) no HEADS and three TAILS, (2) one HEADS and two TAILS, (3) two HEADS and one TAILS, and (4) three HEADS and no TAILS. Thus, the probabilities for each of these four categories must add up to 1.

We just calculated one possibility under category (3). There are a total of three possibilities as found by the multiplicity factor for category (3). The other two possibilities are that the one TAILS occurs not on the first toss of the coin, but on the second, or that the one TAILS occurs not on the first toss of the coin, but on the third. In any case, the probability for all three of these possibilities in category (3) is $3 \times 0.104167 = 0.3125$.

Category (2) follows the same pattern interchanging TAILS for HEADS. This category also has a probability of $3 \times 0.104167 = 0.3125$. Categories (1) and (4) can happen in only one way. The calculation for $P(X_1 = x_2, X_2 = x_2, X_3 = x_2)$ and for $P(X_1 = x_1, X_2 = x_1, X_3 = x_1)$, when conducted in exactly the same way as the previous exercise, results in a probability of 0.1875 for both of these categories.

Adding up the probabilities for all four categories,

$$0.1875 + 0.3125 + 0.3125 + 0.1875 = 1$$

We are certain that something from one of these categories must happen, thus confirming that we have covered all the bases of what must happen in three tosses of the coin.

Exercise 18.7.10: Numerically verify the entropic formula Equation (18.8) as it was derived in section 18.5.

Solution to Exercise 18.7.10

Building on the results of the last few exercises and Volume I, we can say that the probability for any number of future frequency counts of HEADS and TAILS in N future tosses of the coin is encapsulated within the prior predictive formula,

$$P(N_1, N_2) = \int_0^1 P(N_1, N_2 | q_1, q_2) P(q_1, q_2) dq_1$$

For the sake of the numerical example, we will accept the hypothesis that the whole coin tossing mechanism can be captured by one very precise set of numerical assignments. Thus, there is only one model that is averaged over, and the probability for this one model is $P(q_1, q_2) \equiv \delta(q_1 - Q_1, q_2 - Q_2)$.

This might be justified by extensive past observations combined with a physical examination of the coin and the manner in which the coin will be tossed. The consequence flowing from this very strong assumption behind the causal nature of the coin tossing scenario is the binomial distribution for future frequency counts of HEADS and TAILS,

$$P(N_1, N_2 | \mathcal{M}_A) = W(N) Q_1^{N_1} Q_2^{N_2}$$

For the promised numerical example, let's ask: What is the probability for seeing three HEADS and six TAILS in the next nine tosses of the coin? The answer provided by the above binomial predictive formula when the coin is fair is,

$$P(N_1 = 3, N_2 = 6 | \mathcal{M}_A) = W(9) \times Q_1^3 Q_2^6 = \frac{9!}{3! 6!} (1/2)^3 (1/2)^6 = 0.1641$$

Contrast this answer with the probability for the same frequency count if the IP were “completely uninformed” or “totally ignorant” about the causes for a HEADS or TAILS showing up. Using Laplace's *Rule of Succession*, we find that,

$$P(N_1 = 3, N_2 = 6) = \frac{1}{N+1} = \frac{1}{10}$$

Now, instead of a fair coin assignment under model \mathcal{M}_A , suppose that the one model that captures the IP's strong knowledge about the causal mechanism for HEADS or TAILS is another assignment from the MEP algorithm. This model \mathcal{M}_B inserts the information that the average of the constraint function is $\langle F \rangle = +1$. In other words,

$$P(X = x_1 | \mathcal{M}_B) \equiv P(\text{HEADS}) \equiv Q_1 = 0.75$$

$$P(X = x_2 | \mathcal{M}_B) \equiv P(\text{TAILS}) \equiv Q_2 = 0.25$$

The probability for the future frequency count of three HEADS and six TAILS is then,

$$\begin{aligned} P(N_1 = 3, N_2 = 6 | \mathcal{M}_B) &= \frac{9!}{3! 6!} (0.75)^3 (0.25)^6 \\ &= 84 \times 0.421875 \times 0.0000244 \\ &= 0.008652 \end{aligned}$$

Now, let's verify numerically whether the “entropic-like” formula,

$$P(N_1, N_2 | \mathcal{M}_B) = \exp \left[N \left(\left[\frac{\ln W(N)}{N} \right] + \lambda \bar{F} - \ln Z \right) \right]$$

provides the same answer.

$$\begin{aligned} \frac{\ln W(N)}{N} &= \frac{\ln 84}{9} \\ &= 0.492313 \\ \lambda &= 0.137327 \\ \bar{F} &= \frac{(3 \times 3) + (6 \times -5)}{9} \\ &= -2.333 \\ \lambda \bar{F} &= -0.32043 \\ \ln Z &= \ln [e^{(0.137327 \times 3)} + e^{(0.137327 \times -5)}] \\ &= 0.699662 \\ P(N_1 = 3, N_2 = 6 | \mathcal{M}_B) &= \exp [9 \times (0.492313 - 0.32043 - 0.699662)] \\ &= 0.008652 \end{aligned}$$

The two answers for the probability of the future frequency counts for HEADS and TAILS jibe. Our confidence is bolstered that the symbolic derivation is correct.

Exercise 18.7.11: Revert to the original mapping for the coin tossing scenario. What is the information entropy of the distribution under a model with parameter $\lambda = -0.405465$?

Solution to Exercise 18.7.11

With the original mapping of statements to numbers,

$$F(X = x_1) \equiv F(\text{“HEADS”}) = 1$$

and,

$$F(X = x_2) \equiv F(\text{"TAILS"}) = 2$$

the partition function becomes,

$$Z(\lambda) = \sum_{i=1}^2 e^{\lambda F(X=x_i)} = e^\lambda + e^{2\lambda} = e^{-0.405465} + e^{2 \times (-0.405465)} = 1.1111$$

Then, the numerical assignments to the probabilities are,

$$\begin{aligned} Q_1 &= \frac{\exp [\lambda F(X = x_1)]}{Z(\lambda)} \\ &= \frac{0.6667}{1.1111} \\ &= 0.60 \\ Q_2 &= \frac{\exp [\lambda F(X = x_2)]}{Z(\lambda)} \\ &= \frac{0.4444}{1.1111} \\ &= 0.40 \end{aligned}$$

The dual parameter then equals $\langle F \rangle = (1 \times 0.60) + (2 \times 0.40) = 1.40$. The information entropy of this assignment is,

$$H_{max}(Q_i) = -[(0.60 \ln 0.60) + (0.40 \ln 0.40)] = 0.67301$$

Notice that $\ln Z - \lambda \langle F \rangle$ is equal to the information entropy calculated for the assignment,

$$\begin{aligned} \ln Z &= 0.105351 \\ \lambda \langle F \rangle &= -0.405465 \times 1.40 \\ \ln Z - \lambda \langle F \rangle &= 0.67301 \end{aligned}$$

Exercise 18.7.12: Finish finding the answer for the probability of 10,000 future coin tosses introduced as Example 2 in section 18.5.2.

Solution to Exercise 18.7.12

We will use Equation (18.8) to calculate the probability of seeing 6000 HEADS and 4000 TAILS in 10,000 future coin tosses,

$$P(N_1 = 6000, N_2 = 4000 | \mathcal{M}_k) = \exp \left[N \times \left(\left[\frac{\ln W(N)}{N} \right] + \lambda \bar{F} - \ln Z \right) \right]$$

Break down the entropic expression on the right hand side into manageable chunks by first finding the term involving the multiplicity factor,

$$\begin{aligned}\frac{\ln W(N)}{N} &= \ln \left(\frac{10000!}{6000! 4000!} \right) / 10000 \\ &= 0.6725\end{aligned}$$

The second term involves the sample averages with respect to the future data,

$$\begin{aligned}\lambda \bar{F} &= -0.0439 \left[\left(\frac{6000}{10000} \times 0 \right) + \left(\frac{4000}{10000} \times 50 \right) \right] \\ &= -0.8789\end{aligned}$$

The final term involving the partition function is,

$$\begin{aligned}\ln Z &= \ln [1 + \exp(-\epsilon/T)] \\ &= \ln [1 + \exp(-50/22.76)] \\ &= 0.1054\end{aligned}$$

Putting it all back together,

$$\begin{aligned}P(N_1 = 6000, N_2 = 4000 | \mathcal{M}_k) &= \exp [10,000 \times (0.6725 - 0.8789 - 0.1054)] \\ &= 1.65 \times 10^{-1354}\end{aligned}$$

Remember that this was for an assignment under a model \mathcal{M}_k where $Q_1 = 0.90$ and $Q_2 = 0.10$.

Chapter 19

The MEP and Models for Rolling Dice

19.1 Introduction

We are still in the early stages of coming to grips with how the MEP algorithm is implemented. We are advancing at a very measured pace in order to place the MEP within the larger context of probability and inferencing.

The coin tossing example was a first attempt at introducing the MEP formula and its numerical consequences. This exercise was certainly instructive, but since there were only two possibilities at each trial, HEADS or TAILS, the advantage of the MEP assignment over your intuition was slight to non-existent.

We would like to continue our gradual introduction of the MEP formula as a way to find the Q_i . Let's increase the dimension of the state space from the rather low dimension of $n = 2$ of the coin tossing experiment.

Probability theory had its origins in games of chance like cards, dice, and coin tosses. So we are in the good company of some illustrious predecessors like Fermat, Pascal, Bernoulli, and Laplace when we use these games as examples of how to assign probabilities.

In this Chapter, the mental image of rolling a die will serve as a convenient framework for extending the MEP formalism. We said that n represents the number of ways a recordable, or measurable, event could happen at a single trial. We could just as well think of this recordable event as a data point after it has taken place.

By rolling an ordinary die, we increase the number of statements in the state space from two to six. The IP stipulates that only one of six things can happen every time the die is rolled; the face with the ONE spot turns up, the face with the TWO spot turns up, and so on.

This state space has been defined by the IP to have dimension $n = 6$. If, in fact, something else could happen that would be measured, or observed, or otherwise would serve as a category in which to place the outcome of a trial, then it behooves the IP to redefine the state space.

Let's take this opportunity to introduce a little bit more of the notation that will be needed in the future. If we imagine that we will be tossing a coin, or rolling a die, more than once, then we employ N to indicate these N tosses, or rolls. In general, we say that we have conducted N trials.

Also, in the last Chapter, we considered only three models. The first model was a model for a fair coin, while the other two models incorporated information for a coin with a specific bias to come up HEADS.

In the present Chapter, where we switch to thinking about rolling a die, we will once again resort to three models. The first model is the ubiquitous model for a "fair" die. The other two models capture a departure from fairness by contemplating some physical properties of dice.

Each of these latter two models will introduce a constraint that takes the die further away from being a fair die. Specifically, the models will consider, in turn, no constraints, one constraint, and finally, two constraints. For all three models $m < n - 1$, so we need some additional principle to resolve the remaining ambiguity after satisfying the m constraints.

The MEP algorithm is the technique we recommend for resolving this ambiguity. It permits us to assign probabilities that satisfy all the constraints postulated under a given model. Moreover, it introduces no new constraints that the model did not mention.

It does this by maximizing all missing information, (the information *not* specified by the model), after inserting the information from the constraints (the information that *was* specified by the model). Remember that when we talk about the number of constraints, the universal constraint that all probabilities must sum to 1 is always present, even though we don't include it in m .

19.2 The Fair Model and Two Competing Models

We know that the MEP and the constraints specified under a model will dictate what values the Q_i take on. It is always instructive to begin with a model that assumes that the die is fair. The fair model assigns a value of $1/n$ to all Q_i . In the coin tossing problem with $n = 2$, the fair model was,

$$Q_1 = Q_2 = 1/2$$

In the current die rolling problem with $n = 6$, the fair model is,

$$Q_1 = Q_2 = \dots = Q_6 = 1/6$$

This model, and let's call it model \mathcal{M}_A , reflects the physics of a die which is a perfect cube with a center of gravity not favoring any face, and not afflicted with any other imperfection. Also, the die will not be rolled by someone who has special skills at favoring one face over another.

What kind of physical characteristics would be reflected in some competing models? Consider the way a die might be constructed. A certain amount of material is excavated from each face of the die for the indentations that indicate the number of spots on that die face. More material is excavated from the SIX face than from the ONE face, thereby making the ONE face slightly heavier than the SIX face which appears on its opposite side. (The spots on the opposite faces of a die add up to 7.)

Suppose that the paint applied to the indentations do not compensate for the difference in lost material. The same description about the relative amount of material carved from a die face holds true to a lesser degree for the TWO–FIVE faces and the THREE–FOUR faces.

This effect will displace the center of gravity slightly, and make the die depart from the fairness which we incorporated into model \mathcal{M}_A . We would expect that this physical effect would eventually have some sort of impact in the frequency data if the die were tossed often enough. We want to capture this specific physical effect that causes a departure from fairness in a second model \mathcal{M}_B . The Q_i values under \mathcal{M}_B will be different than the $Q_i = 1/6$ values under \mathcal{M}_A .

Another possibility that arises from pondering the construction of the die is that when the die faces were cut on a milling machine, one axis might be slightly longer or smaller than the other two axes. Greatly exaggerated, this effect would make the shape of the die more in the form of a brick than a cube.

If we were to toss a brick, we know that it is more likely to land “flat,” rather than “standing up.” Again, this physical effect, no matter how small it might be, would eventually show up in frequency data if the die were rolled often enough. Model \mathcal{M}_C will incorporate this potential deviation from a perfect cube, as well as the deviation already described under model \mathcal{M}_B . The Q_i values for model \mathcal{M}_C will, therefore, be different than under both model \mathcal{M}_A and model \mathcal{M}_B .

19.3 The MEP Assignments in Rolling the Die

We begin by slightly generalizing the MEP formula in Equations (17.1) and (17.2) that appeared in the opening Chapter. There will be as many terms for the exponent in the numerator as there are number of constraints. Since we foresee that the most complex model, \mathcal{M}_C , will have $m = 2$ constraints, the MEP template is written as,

$$Q_i = \frac{\exp [\lambda_1 F_1(X = x_i) + \lambda_2 F_2(X = x_i)]}{Z(\lambda_1, \lambda_2)} \quad (19.1)$$

If a constraint is not operating, then its associated parameter value is zero. Therefore, for model \mathcal{M}_A both λ_1 and λ_2 equal 0. For model \mathcal{M}_B , λ_1 has a non-zero value, and only λ_2 is equal to zero. For the most complex model, model \mathcal{M}_C , both λ_1 and λ_2 will take on non-zero values.

In order to arrive at the assignment values under each model without undue digressions, and to maintain an overall sense of what we are about, the Lagrange multipliers, the constraint functions, and the mean values for the constraint functions, will be specified up front without further comment. After seeing what kind of Q_i values emerge from these three MEP models, we will go back and provide a bit more justification for these choices.

19.3.1 Assignment under the fair model

The MEP assignment to the six Q_i probabilities under model \mathcal{M}_A is quite easy to calculate. Since this is the model for the fair die, there are no constraints operating. Therefore, $\lambda_1 = \lambda_2 = 0$. Equation (19.1) provides us with,

$$\begin{aligned} Q_i &= \frac{\exp [\lambda_1 F_1(X = x_i) + \lambda_2 F_2(X = x_i)]}{Z(\lambda_1, \lambda_2)} \\ &= \frac{\exp [(0 \times F_1(X = x_i)) + (0 \times F_2(X = x_i))] }{Z(\lambda_1, \lambda_2)} \\ &= \frac{\exp [0]}{Z(\lambda_1, \lambda_2)} \\ &= \frac{1}{Z(\lambda_1, \lambda_2)} \end{aligned}$$

Q_1 through Q_6 are all assigned the same probability. The assigned value doesn't depend on either constraint function.

The denominator $Z(\lambda_1, \lambda_2)$ is the partition function. It is calculated as,

$$Z(\lambda_1, \lambda_2) = \sum_{i=1}^6 \exp [\lambda_1 F_1(X = x_i) + \lambda_2 F_2(X = x_i)] = \sum_{i=1}^6 \exp [0] = 6$$

If the assignments must satisfy the universal constraint of summing to 1, then we see that the normalizing factor must equal 6. For a model \mathcal{M}_A that postulates a fair die, we obtain the eminently reasonable result that all six faces are assigned the probability of 1/6.

This is analogous to the $n = 2$ situation when $P(\text{HEADS}) = P(\text{TAILS}) = 1/2$. It seems a safe conjecture that the MEP assignment for a situation with no constraints, let's call it the "fair" case, will assign a probability of $1/n$ when the event can happen in n ways.

19.3.2 An unbalanced die

Now on to model \mathcal{M}_B . Here we are interested in the implications of a die that deviates from fairness. Therefore, its assigned probabilities will deviate from 1/6. The first parameter, λ_1 , will take on a non-zero value. Now the constraint function $F_1(X = x_i)$ does matter.

Let the first constraint function be defined by the vector,

$$F_1(X = x_i) = (1, 2, 3, 4, 5, 6)$$

This specification of the constraint function is just the first part of inserting information into a probability distribution. The second part is to specify some particular value of λ_1 , and then live with the resulting dual value of the constraint function average. Alternatively, the IP might specify the constraint function average, and then find out what value of λ_1 is needed to satisfy that particular constraint function average.

We will do the latter and specify the constraint function average as the information to be inserted into the probability distribution under model \mathcal{M}_B . Now, if we were to specify $\langle F_1 \rangle = 3.5$ as that information, then the Lagrange multiplier for the first constraint would be $\lambda_1 = 0$. We would be right back to the same assignment of $Q_i = 1/6$ as under the fair model. It's easy to verify that,

$$\langle F_1 \rangle = \sum_{i=1}^6 F_1(X = x_i) Q_i = (1 \times 1/6) + \cdots + (6 \times 1/6) = 3.5$$

So, it's clear that to obtain an assignment different than the one emanating from the fair model, the IP will have to specify the constraint function average as something other than 3.5. Also, from the last Chapter, we know that the feasible values for $\langle F_1 \rangle$ must lie in the range from 1 to 6. And we know immediately, without ever resorting to the MEP algorithm, that if the IP were to set $\langle F_1 \rangle = 1$, then $Q_1 = 1$ and the remaining $Q_i = 0$. Likewise, if the IP were to set $\langle F_1 \rangle = 6$, then $Q_6 = 1$ and the remaining $Q_i = 0$.

We will assume that our die is not *that* unbalanced, and seek some intermediate value for $\langle F_1 \rangle$. How about setting $\langle F_1 \rangle = 4$? This is different than 3.5, so we won't wind up with the fair model assignments again. Specifying this piece of information under model \mathcal{M}_B should bias the outcomes towards the higher spots since the average is higher than 3.5. We would expect the converse to hold if we were to specify that the information was a constraint function average below 3.5.

So for the numerical example, let's set $\langle F_1 \rangle = 4$. The Lagrange multiplier must adjust to $\lambda_1 = 0.17463$ to satisfy this constraint. Table 19.1 at the top of the next page presents the details of the MEP computation and the assigned probabilities.

The first column shows the $n = 6$ possible occurrences at each trial, that is, the number of spots showing face up after the die was rolled. The second column gives the functional association of $F_1(X = x_i)$ to each value of $(X = x_i)$.

Table 19.1: *The details of the MEP computation for the die problem with only one constraint. This is model \mathcal{M}_B .*

Statement	$F_1(X = x_i)$	$\exp [\lambda_1 F_1(X = x_i)]$	Q_i	$\langle F_1 \rangle$
ONE	1	1.1908	0.1031	0.1031
TWO	2	1.4180	0.1227	0.2454
THREE	3	1.6886	0.1462	0.4386
FOUR	4	2.0108	0.1740	0.6960
FIVE	5	2.3944	0.2072	1.0360
SIX	6	2.8513	0.2468	1.4808
Sums		11.5539	1.0000	4.0000

The first column contains the six statements in the state space. $(X = x_i)$ is the abbreviated form for the statement, “The die was rolled and x_i spots showed face up.”. Pay very close attention to the fact that the first column lists the *statements* in the state space, while the second column is a mapping from the statements to numbers. The third column computes the numerator of Equation (19.1). Since λ_2 is equal to zero for model \mathcal{M}_B , only the first term is shown.

The sum at the bottom of the third column is the value of the partition function $Z(\lambda_1)$. Notice that this normalizing factor has shifted from a value of 6 under \mathcal{M}_A to $Z(\lambda_1) = 11.5539$ under \mathcal{M}_B .

The next-to-the-last column contains the numerical assignments, the Q_i , for the probabilities of each face. This is the number in the third column divided by the partition function. The sum of this column must equal 1.

The last column shows the mean value of the first constraint. λ_1 was adjusted until a value was found which made $\langle F_1 \rangle$ equal to 4. This is the value of $\lambda_1 = 0.17463$ as given above. If $\langle F_1 \rangle = 3.5$, then the die is *modeled* as fair. But specifying $\langle F_1 \rangle = 4$ as the value for the mean of the constraint gives rise to the Q_i that reflect a *model* for a physical bias in the construction of the die.

The probabilities for the fair die would all be $Q_i = 0.1667$ in decimal format. The assigned probabilities under \mathcal{M}_B obviously differ from the fair die assignment. Notice that the probability is lowest for the ONE spot, and increases monotonically until it reaches a maximum for the SIX spot.

The information inserted under model \mathcal{M}_B , that is, $\langle F_1 \rangle = 4$, or alternatively, $\lambda_1 = 0.17463$, led to a numerical assignment for the probabilities of all n statements in the state space. Such an assignment was carried out completely independently of any data ensuing from any number of rolls of the die.

19.3.3 An unbalanced and ill-constructed die

The assigned probabilities under \mathcal{M}_C are explained in exactly the same manner. The only change is the inclusion of a second constraint function which takes the form of the vector,

$$F_2(X = x_i) = (+1, +1, -2, -2, +1, +1)$$

To repeat, these constraint functions are arbitrary mappings from *statements* in the state space to numbers. “Arbitrary” is perhaps too strong an adjective since Jaynes told us that such a mapping is where “the physics comes in.”

Table 19.2 is just like Table 19.1 except for the inclusion of the second constraint. An application of the MEP algorithm calculated the values of the Lagrange multipliers as $\lambda_1 = 0.155216$ and $\lambda_2 = 0.149669$. As we might have expected, under the constraints of \mathcal{M}_C the die is assigned numerical values that differ substantially from a fair die. However, the way in which it is unfair is different than what we saw under \mathcal{M}_B .

Table 19.2: *The details of the MEP computation for the die problem with two constraints. This is model \mathcal{M}_C .*

Statement	F_1	F_2	$\exp [\lambda_1 F_1 + \lambda_2 F_2]$	Q_i	$\langle F_1 \rangle$	$\langle F_2 \rangle$
ONE	1	+1	1.3565	0.1236	0.1236	+0.1236
TWO	2	+1	1.5842	0.1444	0.2887	+0.1444
THREE	3	-2	1.1809	0.1076	0.3228	-0.2152
FOUR	4	-2	1.3792	0.1257	0.5028	-0.2514
FIVE	5	+1	2.5238	0.2300	1.1500	+0.2300
SIX	6	+1	2.9475	0.2687	1.6121	+0.2687
Sums			10.9722	1.0000	4.0000	0.3000

This second constraint function is meant to embody the hypothesis that the THREE-FOUR axis of the cube does not have the same length as the other two axes. The milling machine which cut the cubical die from the surrounding amorphous material was not aligned properly, thus allowing this dimension to be slightly larger than the other two dimensions. The final product is a somewhat “brick-like” die.

Let’s suppose that model \mathcal{M}_C retains the old information about the unbalanced nature of the die that was reflected in model \mathcal{M}_B , as well as incorporating this new physical imperfection. Even though model \mathcal{M}_C now consists of $m = 2$ constraints, there is still some remaining ambiguity about the numerical assignments. The MEP is relied upon to resolve this remaining ambiguity.

Once again, information could be inserted into the probability distribution by varying the two parameters λ_1 and λ_2 , or, alternatively, by varying the two dual parameters, $\langle F_1 \rangle$ and $\langle F_2 \rangle$. Let's continue to specify the constraint function averages in order to see how the Lagrange multipliers have to adjust in order to reproduce these settings. Keep $\langle F_1 \rangle = 4$ and set $\langle F_2 \rangle = 0.3$ to impose some effect from a lengthened THREE–FOUR axis.

The probabilities assigned to the THREE and FOUR faces are depressed relative to the other four faces. This is due to the presumed physical effect of lengthening this axis.

However, we can still pick out within this pattern the overall trend under \mathcal{M}_B of an increasing probability as the number of spots increases. Presumably, this is due to the physical effect of shifting the center of gravity through the unequal removal of material to form the spots, and thereby creating some imbalance in the die. Both physical imperfections are operating under model \mathcal{M}_C to produce numerical assignments that accurately reflect the information inserted by the IP.

19.4 Updating the Three Models

At this juncture, we have used the MEP to successfully assign numerical values to the probabilities of the six faces of the die. Three different sets of assignments were made under the assumption that three different models were true. That is, the probabilities were assigned under three different assumptions about the physical construction of the die. The point is that the MEP assignments insert information from tentatively entertained causal theories.

Under model \mathcal{M}_A , the die was assumed to be fair and $Q_i = 1/6$. The next two models assumed the die was “loaded,” that is, the physical construction of the die caused the probabilities to deviate from $Q_i = 1/6$. Model \mathcal{M}_B postulated that the center of gravity was moved away from the exact center of the cube. Model \mathcal{M}_C assumed not only this effect on the center of gravity, but, in addition, postulated a shortening or lengthening of one of the three dimensions of the cube.

We do not choose to differentiate among the three models at the outset. We are convinced by Laplace's argument that, in a state of total ignorance, an IP is not able to assess the relative merits of what might be causing the spots to show up. We place each model on an equal footing by assigning,

$$P(\mathcal{M}_A) = P(\mathcal{M}_B) = P(\mathcal{M}_C) = 1/3.$$

There are two ways we could change our opinion about these models. One way would be through direct physical measurements of the center of gravity and the dimensions of the cube. The other way would be to roll the die a large number of times to see if these supposed physical properties showed up in the frequency data.

From our discussion of Bayes's Theorem in Volume I, we know that the models get updated by conditioning on the data \mathcal{D} ,

$$P(\mathcal{M}_k | \mathcal{D}) = \frac{P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k)}{P(\mathcal{D})}$$

For example, if there were N previous rolls of the die, and the number of spots showing was accurately recorded at each trial, then the probability for the data consisting of these past frequency counts when conditioned on a specific model is,

$$P(\mathcal{D} | \mathcal{M}_k) = W(N) Q_1^{N_1} Q_2^{N_2} \cdots Q_n^{N_n}$$

The data are given to us without the actual specific ordering, that is, it is given to us in the form that N_1 ONEs were observed, N_2 TWOs were observed, and so on. Thus, the multiplicity factor $W(N)$ is also required.

19.5 Information Entropy Given the Models

There is something extra to pay attention to above and beyond what we have already discussed as the primary output from the MEP algorithm. We notice an interesting trend in the progression of the information entropy computed for each of these three models.

Here is the entropy for the assignments under all three models,

$$H(Q_i | \mathcal{M}_A) = 1.7918$$

$$H(Q_i | \mathcal{M}_B) = 1.7485$$

$$H(Q_i | \mathcal{M}_C) = 1.7296$$

The entropy is decreasing with each new model. Each new model was in some sense more complex than its predecessor in that another constraint function was added. We went from $m = 0$ to $m = 1$ to $m = 2$, (always remembering that the universal constraint has already been counted), as we progressed from model \mathcal{M}_A to model \mathcal{M}_B to model \mathcal{M}_C .

Thus, each new piece of information arriving with each new model acted to decrease the information entropy. This makes perfect sense. If the information entropy is some kind of measure of the amount of missing information, then by adding information with each new model, we are removing some of that missing information. It stands to reason that the entropy should decrease.

If the IP adds enough information so that there is no missing information, then the entropy collapses all the way down to zero. The entropy is zero when one statement in the state space has a probability of 1, and all the other statements have a probability of 0. The IP is absolutely certain what will happen if a trial is conducted.

19.6 The Missing Information in an Assignment

Model \mathcal{M}_C , our most complex model so far, possesses, by definition, the maximum entropy of *any* probability assignment that also satisfies the two constraints. Its entropy is,

$$\begin{aligned} H(Q_i \mid \mathcal{M}_C) &= - \sum_{i=1}^6 Q_i^C \ln Q_i^C \\ &= 1.7296 \end{aligned}$$

It is impossible to find any other assignment that satisfies the constraints, and has a higher value of the entropy. If you do assign probabilities under another model to satisfy the constraints, then you can be sure that such a distribution will have a lower entropy. It follows, therefore, that you have inserted extra information under that model. If you were not aware of what this information was, you have just proposed a model which you don't fully understand.

For example, there exists another legitimate assignment of probabilities which looks similar to the probabilities assigned under model \mathcal{M}_C as shown previously in Table 19.2. This new assignment is shown below in Table 19.3.

Table 19.3: *Another legitimate assignment of probabilities to the six faces of the die similar to model \mathcal{M}_C , but which contains extra information above and beyond what is in model \mathcal{M}_C . Its entropy therefore must be lower than model \mathcal{M}_C 's.*

Statement	$F_1(X = x_i)$	$F_2(X = x_i)$	Q_i	$\langle F_1 \rangle$	$\langle F_2 \rangle$
ONE	1	+1	0.1394	+0.1394	+0.1394
TWO	2	+1	0.1424	+0.2848	+0.1424
THREE	3	-2	0.1154	+0.3462	-0.2308
FOUR	4	-2	0.1179	+0.4716	-0.2358
FIVE	5	+1	0.1516	+0.7580	+0.1516
SIX	6	+1	0.3333	+2.0000	+0.3333
Sums			1.0000	4.0000	0.3000

This assignment also satisfies the two constraints and the universal constraint as did model \mathcal{M}_C . Check the sums under the final three columns of Table 19.3 to verify this.

Qualitatively, this new assignment also follows the pattern we remarked on with model \mathcal{M}_C . There is a general upward progression in probability from the ONE face to the SIX face due to the first constraint, and a marked depression in this trend for the THREE and FOUR faces due to the second constraint.

What's wrong with this assignment? Well, there's nothing *wrong* with this assignment. It's just that it contains extra information not in model \mathcal{M}_C . How do we know that it contains extra information?

Calculate this assignment's entropy as,

$$\begin{aligned} H(Q_i | \mathcal{M}_D) &= -\sum_{i=1}^6 Q_i^D \ln Q_i^D \\ &= 1.7057 \end{aligned}$$

The entropy for this assignment, under a model we'll call \mathcal{M}_D , at 1.7057, is, as claimed, smaller by some 0.0139 than model \mathcal{M}_C 's entropy of 1.7296. *There must exist extra information in this assignment that is missing in the assignment under model \mathcal{M}_C .*

Assignments that contain information that has not been consciously inserted under a model are unclear. The MEP explicitly guards against this outcome.

What exactly is that extra information in model \mathcal{M}_D ? A clue is provided in the fact that Table 19.3 did not show the column where the numerator and the partition function were calculated. That is because, in fact, model \mathcal{M}_D possessed a third constraint shown in the vector below,

$$F_3(X = x_i) = (-1, -1, -1, -1, -1, +5)$$

with the information defined as $\langle F_3 \rangle = 1.00$.

Model \mathcal{M}_D mirrors an hypothesis about the physical construction of the die that includes all the previously mentioned imperfections, but now includes a third imperfection that favors the SIX face. The numerator under model \mathcal{M}_D would look like,

$$\exp [\lambda_1 F_1(X = x_i) + \lambda_2 F_2(X = x_i) + \lambda_3 F_3(X = x_i)]$$

The parameters for this model \mathcal{M}_D were found to be $\lambda_3 = 0.1278$, with λ_1 and λ_2 re-adjusted to values of 0.0210 and 0.0768 to satisfy the previous constraints of $\langle F_1 \rangle = 4.00$ and $\langle F_2 \rangle = 0.30$. The MEP assignment that resulted by including this third constraint resulted in the Q_i values shown in Table 19.3. So now we know the extra information resident in model \mathcal{M}_D that wasn't in \mathcal{M}_C .

It seems, therefore, that we have to explicitly separate out this assignment in Table 19.3 as a separate model. We have to add it as a fourth model \mathcal{M}_D to the three we have already proposed. The point is simply that all of our models must clearly indicate what information they are inserting into a probability distribution.

As a final comment, don't be fooled by small differences in the entropies between any two models. Very small differences in entropy can result in quite different assignments to probability distributions.

19.7 Constraint Functions and Physical Reasoning

This section is included because Jaynes took great pains in his dice rolling examples to link up the abstract nature of constraint functions and Lagrange multipliers with some physical reasoning.

The first constraint told us that there was a shift in the probability towards the faces with the larger number of spots to fall face up. This is caused by a subtle shift in the center of gravity toward the *lower* number of spots.

As the die is tumbling across the table, the ever present effect of gravity is being exerted on all six faces of the die. If the center of gravity is moved towards the ONE spot, for example, then gravity directed towards the center of the earth, will tend to pull ever so slightly more on the ONE spot, causing it to fall face down on the table more often after a toss.

This slight tendency means that the SIX spot, opposite the ONE spot, will fall face up more often than one sixth of the time. If gravity were equally affecting the ONE and SIX spots, then these two faces should turn up equally often (within sampling variation) over a large number of tosses.

What could cause such a slight shift in the center of gravity towards the ONE spot? The excavation of material from one side of a perfect cube to create six indentations causes an imbalance. More material is removed from the side chosen for the SIX spot than from the opposite side showing the ONE spot.

Presumably lesser material was removed from the side of the cube to create just one indentation. Remember that the opposite faces of the die add up to seven. As far as we know, the differing amounts of material removed are not balanced off by other compensating physical changes to the die. Under this model, the physical effect due to material removal is not compensated by anything at all.

This is also true to a lesser degree for the TWO–FIVE dimension and the THREE–FOUR spot dimension. There is an imbalance due to the five extra spots on the ONE–SIX dimension, an imbalance due to the three extra spots on the TWO–FIVE dimension, and an imbalance due to the one extra spot on the THREE–FOUR dimension.

So the center of gravity has been shifted by some small amount from the exact center of the cube to some new point within the cube. The faces with the ONE, TWO, and THREE spots are all “heavier” than the faces with the FOUR, FIVE and SIX spots.

We could have set up a constraint function vector,

$$F_{1*}(X = x_i) = (-2.5, -1.5, -0.5, +0.5, +1.5, +2.5)$$

in order to capture this physical characteristic of the die. There is a difference of 5 between $F_{1*}(X = x_1) = -2.5$ and $F_{1*}(X = x_6) = 2.5$, a difference of 3 between

$F_{1^*}(X = x_2) = -1.5$ and $F_{1^*}(X = x_5) = 1.5$, and a difference of 1 between $F_{1^*}(X = x_3) = -0.5$ and $F_{1^*}(X = x_4) = 0.5$. These effects all have to cancel each other out because, for example, whatever increase the SIX face receives must be equally compensated for in the decrease to the ONE face. So for a fair die, we would insert the information that $\langle F_{1^*} \rangle = 0$.

Upon reflection, we observe that this constraint function of the observables is the same as the original $F_1(X = x_i)$ if we subtract the constraint average of 3.5. Whether $F_{1^*}(X = x_i)$ or $F_1(X = x_i)$ is used as the constraint function is irrelevant as long as the same information in the constraint average is specified by $\langle F_1 \rangle = 3.5$ or $\langle F_{1^*} \rangle = 0$.

The Q_i that make,

$$\sum_{i=1}^6 F_{1^*}(X = x_i) Q_i = 0$$

are the same Q_i that made,

$$\sum_{i=1}^6 F_1(X = x_i) Q_i = 3.5$$

So, the first constraint function of model \mathcal{M}_B tried to capture this physical feature of an imperfect die. The second constraint function included in model \mathcal{M}_C also tried to capture a different feasible physical imperfection.

If a solid object that is a perfect cube is rolled on a flat surface, there is no propensity for one side to fall face up in preference to another side because the three dimensions of the cube are equal. This is a characteristic of a fair die.

However, if one of the axes connecting two faces of the die is longer than another, then the spots that are on these two faces of this longer axis will tend to fall face up less often. If the object is shaped more like a brick than a cube, then when the brick is rolled it will tend to land on the face of one of the shorter dimensions rather than standing on its end, the longer dimension.

If the THREE–FOUR axis were longer than the ONE–SIX or TWO–FIVE axes, then the THREE and FOUR faces would land face up less often than the other faces. Model \mathcal{M}_C incorporated a second constraint function to mirror this physical attribute.

$$F_2(X = x_i) = (+1, +1, -2, -2, +1, +1)$$

The functional values assigned to $(X = x_3)$ and $(X = x_4)$ are equal and negative to reflect the tentatively entertained hypothesis of greater length on the axis where these faces are inscribed. The functional values assigned to $(X = x_1)$, $(X = x_2)$, $(X = x_5)$, and $(X = x_6)$ are equal and positive because their lengths are equal, and because they must also share the increase given to them from $(X = x_3)$ and $(X = x_4)$.

19.8 Connections to the Literature

Over the course of his forty plus years of writing about the MEP, Jaynes would often bring up examples involving dice. The earliest mention was perhaps at the beginning of his 1962 Brandeis Lectures [16]. I speculate that this was due to Boltzmann's use of dice to illustrate his ideas concerning probability in statistical mechanics.

Jaynes's objective here, as he said, was mainly to show that plausible *qualitative* reasoning involving data about many rolls of a die forced one very close to the answer arrived at through a quantitative analysis based on the MEP.

Unfortunately, it was in explaining this example of rolling the die and MEP that Jaynes sowed the seeds of confusion which reverberate down to the present day. In 1957, Jaynes was correctly telling us that *information* was defined as the probabilistic expectation (average) of anything that was a function of an observable. In other words, it was an expectation of a function, the constraint function, with respect to Jaynes's p_i . In our notation, it was the probabilistic expectation,

$$\langle F \rangle = \sum_{i=1}^n F(X = x_i) Q_i$$

Thus, in 1957 Jaynes was writing down the correct expression for information as $\langle f(x) \rangle = \sum_{i=1}^n p_i f(x_i)$. There was no mention that these constraint function averages were based on any actual *data*!

But in 1962 [16, pg. 41], in his opening salvo explaining the general formalism of the MEP as it related to the die tossing scenario, Jaynes invites us to consider that,

A die has been tossed a very large number N of times, and we are told that the average number of spots up per toss was not 3.5, as we might expect from an honest die, but 4.5. Translate this information into a probability assignment $P_n, n = 1, 2, \dots, 6$ for the n -th face to come up on the next toss.

Information now has been re-defined (incorrectly) as some observed average over N flips of the die. The formal manipulation rules of probability allow no equivocation here. These rules tell us exactly how the data, that is, the previous N observed rolls of the die, must be processed in order to calculate the probability for the $(N + 1)^{st}$ roll. The probability for the *next* face to come up is,

$$P(X_{N+1} = x_i | X_1, X_2, \dots, X_N) = \sum_{k=1}^{\mathcal{M}} P(X_{N+1} = x_i | \mathcal{M}_k) P(\mathcal{M}_k | \mathcal{D})$$

We have much more to say about the resulting confusion in the next Chapter.

Later in 1978, he presented a much more interesting example of a dice problem in his stunning summary of the status of the MEP [18, pp. 258–268],

Where Do We Stand on Maximum Entropy?

All of the material in this Chapter is an elaboration on the example that Jaynes labeled as Wolf's Dice Data. It remains to this day one of the best explanations linking the abstract mathematics of the MEP formula with some quite acceptable intuitive constraints on the physical characteristics of dice.

My goal in this Chapter was simply to explain in detail the expansion of the MEP formula for the situation dealing with $n = 6$. Repeating Jaynes's extensive preliminary work in solving the *Wolf's Dice Data* problem turns out to be an excellent exercise in applying the MEP. Nevertheless, in this Chapter, I never did reach the point where the actual data about Wolf's dice entered into the picture.

But Jaynes did continue on with the data analysis with some curious and, to my mind, very revealing consequences. The most important aspect from my perspective was Jaynes's insistence that the MEP be considered as an alternative to the conventional orthodox hypothesis testing one would have carried out at that time.

The way he explained this alternative approach was not only quite intriguing, but intuitively compelling as well. He asked us to compare an initial MEP assignment to the actual normed frequency counts. These initial assignments from some small set of hypotheses will likely include very few constraints. Our minds are naturally attuned to adopt Ockham's razor at the outset.

In this Chapter, I emulated this procedure by picking three models with $m = 0$, $m = 1$, and finally $m = 2$ constraints. Jaynes said to compare your simplest model with the actual data as normed frequency counts. Most likely, there will be some glaring discrepancies between the two. Instead of being discouraged by these discrepancies, we should use them as clues for additional constraints that might be lurking in the physical phenomenon.

Create a second MEP assignment with a model incorporating the information from that suspected constraint. Compare this new MEP assignment with the data once again. Hopefully, the discrepancies have been somewhat reduced.

But the “fit to the data” may still not be to your liking. So repeat the procedure. Look at the pattern of discrepancies between the normed frequency counts and your latest MEP assignment. Figure out a new constraint in addition to your already existing ones in yet another model. These revised MEP assignments should get you even closer to the actual data.

Where does one stop in such an iterative procedure in revamping the MEP to match the data? An orthodox test developed by Karl Pearson around 1900 was one way of answering this question. The so-called **chi square test** had been widely used in fitting models to data, and it is still quite popular to this day.

Jaynes showed that by the chi-square criterion, each new MEP assignment reflected an improved probability distribution which was a closer fit to the observed normed frequency counts. In other words, as more and more cogent information was brought to bear on the probability assignments, the match improved between an IP's "state of knowledge" and actual reality.

However, what is most attractive about this procedure is the notion that the IP can iteratively "subtract off" clear causal signals in an additive manner until all that remains is some sort of "random noise." This "subtracting off of a signal component from the noisy data" is operationally implemented by successive MEP assignments with models of increasing m . The final Chapter in this Volume takes an initial look at implementing such a concept.

Jaynes is quite clear on this matter [18, pg. 264],

But now I stress still another time what the principle is really telling us:
a statistically significant deviation is evidence of a new physical constraint; and the nature of the deviation gives us a clue as to what that constraint is.
[Emphasis in the original]. After subtracting off, by maximum entropy, the deviation attributable to the first constraint, the nature of the most important remaining one is revealed. Indeed, from a glance at the deviations ... the answer leaps out at us; Wolf's die was slightly "prolate," the (3-4) dimension being greater than the (2-5) and (1-6) ones.

More than that, Jaynes stated that an even better criterion than the chi-square was available: this was an adaptation of Kullback's relative entropy measure. Now this opens up a whole new way of thinking about testing statistical hypotheses.

I just wanted to broach this fascinating topic here since it did appear in Jaynes's analysis of *Wolf's Dice Data*. It is so tempting and juicy a topic, though, that I cannot possibly leave it hanging here in its current undeveloped state. Personally, I had been indoctrinated into and used the chi-square test in nearly all of my own statistical analyses without ever questioning its fundamental rationale. Like most things in the orthodox approach, it was the accepted mysterious "recipe" which one was advised to ignore at one's peril.

Thus, to exorcise my own demons, I will engage in an extensive discussion of the chi-square test from Jaynes's MEP perspective at an appropriate time and place.

19.9 Solved Exercises for Chapter Nineteen

Exercise 19.9.1: Show the detailed calculation involved in the probability assignment for seeing a FOUR under model \mathcal{M}_B .

Solution to Exercise 19.9.1

When conditioned on the information in statement \mathcal{M}_B , the MEP formula for calculating the numerical assignment to a probability for seeing the FOUR spot face up after rolling a die is,

$$\begin{aligned} Q_4 &\equiv P(\text{FOUR} | \mathcal{M}_B) \\ &= \frac{\exp[\lambda_1 F_1(X = x_4)]}{Z(\lambda_1)} \\ &= \frac{\exp[0.17463 \times 4]}{\sum_{i=1}^6 \exp[\lambda F(X = x_i)]} \\ &= \frac{2.0108}{11.5539} \\ &= 0.1740 \end{aligned}$$

Exercise 19.9.2: Show the detailed calculation involved in the probability assignment for seeing a THREE under model \mathcal{M}_C .

Solution to Exercise 19.9.2

When conditioned on the information in statement \mathcal{M}_C , the MEP formula for calculating the numerical assignment to a probability for seeing the THREE spot face up after rolling a die is,

$$\begin{aligned} Q_3 &\equiv P(\text{THREE} | \mathcal{M}_C) \\ &= \frac{\exp[\lambda_1 F_1(X = x_3) + \lambda_2 F_2(X = x_3)]}{Z(\lambda_1, \lambda_2)} \\ &= \frac{\exp[(0.155216 \times 3) + (0.149669 \times (-2))]}{\sum_{i=1}^6 \exp[\sum_{j=1}^2 \lambda_j F_j(X = x_i)]} \\ &= \frac{1.1809}{10.9722} \\ &= 0.1076 \end{aligned}$$

Exercise 19.9.3: Show the detailed calculation involved in the probability assignment for seeing a **SIX** under model \mathcal{M}_D .

Solution to Exercise 19.9.3

When conditioned on the information in statement \mathcal{M}_D , the MEP formula for calculating the numerical assignment to a probability for seeing the **SIX** spot face up after rolling a die is,

$$\begin{aligned} Q_6 &\equiv P(\text{SIX} \mid \mathcal{M}_D) \\ &= \frac{\exp[\lambda_1 F_1(X = x_6) + \lambda_2 F_2(X = x_6) + \lambda_3 F_3(X = x_6)]}{Z(\lambda_1, \lambda_2, \lambda_3)} \\ &= \frac{\exp[(0.0210 \times 6) + (0.0768 \times 1) + (0.1278 \times 5)]}{\sum_{i=1}^6 \exp[\sum_{j=1}^3 \lambda_j F_j(X = x_i)]} \\ &= \frac{2.3209}{6.9628} \\ &= 0.3333 \end{aligned}$$

Exercise 19.9.4: Which of the four models in the die rolling scenario incorporates the most amount of missing information, and which the least amount of missing information?

Solution to Exercise 19.9.4

Inspect the information entropy for the assignments provided under each of the four models. Since information entropy is a quantitative measure of the amount of missing information, the model with the highest entropy is the model with most missing information. This is model \mathcal{M}_A . The model with the least amount of missing information is model \mathcal{M}_D .

Exercise 19.9.5: How much missing information is in a model which assigns a numerical value of 1 to seeing a **TWO** spot and 0 to the other five faces?

Solution to Exercise 19.9.5

The information entropy in this assignment is,

$$H(Q_i \mid \mathcal{M}_k) = - \sum_{k=1}^6 Q_i \ln Q_i = 0$$

There is no missing information in this assignment. The IP is certain, under this model, that the die will show a **TWO** when rolled.

Exercise 19.9.6: What is the probability that this model is true after a ONE spot is observed after the die has been rolled?

Solution to Exercise 19.9.6

The probability is 0 that this model is true after a ONE has been observed. It is certain that the statement made by this model is FALSE. The statement made by this model is something like, “The correct assignment of probabilities is $Q_2 = 1$ with all other $Q_i = 0$.”.

Exercise 19.9.7: What is the probability that the very first roll of the die shows a THREE?

Solution to Exercise 19.9.7

Think of the very first roll of the die as, in fact, the *next* roll of the die if it has never been rolled before. So what do the formal rules of probability theory tell us is the answer?

Conceptually, it all depends on what stance the IP takes with regard to its knowledge about the causal nature of the dice rolls. If the IP agrees with Laplace (as I do), then the answer is,

$$\begin{aligned} P(N_1 = 0, N_2 = 0, N_3 = 1, N_4 = 0, N_5 = 0, N_6 = 0) &= \frac{(n-1)! N!}{(N+n-1)!} \\ &= \frac{5! 1!}{(1+6-1)!} \\ &= 1/6 \end{aligned}$$

If the IP prefers the binomial distribution with a belief in a “fair” die (a highly suspect position), then the answer is,

$$\begin{aligned} P(N_1 = 0, \dots, N_3 = 1, \dots, N_6 = 0 | \mathcal{M}_k) &= W(N) Q_1^{N_1} Q_2^{N_2} Q_3^{N_3} Q_4^{N_4} Q_5^{N_5} Q_6^{N_6} \\ &= 1 \times Q_3 \\ &= 1/6 \end{aligned}$$

So we see here the source of so much confusion. The same answer results for the very first throw whether you adopt a stance of total ignorance, or the polar opposite one of very strong knowledge about the causal nature of dice!

Exercise 19.9.8: What is the probability that the first two rolls of the die show a THREE and a SIX?

Solution to Exercise 19.9.8

Conceptually, it all depends on what stance the IP takes with regard to its knowledge about the causal nature of the dice rolls. If the IP agrees with Laplace about the only sensible position to take regarding the relative standing of the causes prior to knowing anything about the die, or the manner in which it will be rolled, then the answer is,

$$\begin{aligned} P(N_1 = 0, N_2 = 0, N_3 = 1, N_4 = 0, N_5 = 0, N_6 = 1) &= \frac{(n-1)! N!}{(N+n-1)!} \\ &= \frac{5! 2!}{(2+6-1)!} \\ &= 1/21 \end{aligned}$$

If the IP prefers the binomial distribution with a belief in a “fair” die, then it is adopting a position directly opposed to Laplace’s advice. Indeed, the IP now possesses a very definite opinion that there exists but one cause for the die’s outcome, and the answer is,

$$\begin{aligned} P(N_1 = 0, \dots, N_3 = 1, \dots, N_6 = 1 | \mathcal{M}_A) &= W(N) Q_1^{N_1} Q_2^{N_2} Q_3^{N_3} Q_4^{N_4} Q_5^{N_5} Q_6^{N_6} \\ &= 2 \times Q_3 \times Q_6 \\ &= 1/18 \end{aligned}$$

The probability for seeing two SIXES on the first two trials is still 1/21 under Laplace’s reasoning, but is 1/36 under the binomial distribution reasoning. The probability for *any* two future frequency counts remains the same at 1/21 when the IP is totally ignorant about the dice. This is because there are only twenty one possible frequency counts when $N = 2$. (Refer back to section 13.2 of Volume I.)

On the other hand, when the IP adopts one model to the exclusion of all others about the causal nature of the dice (and here that one model is the model for a fair die), then there are six frequency counts from the total of twenty one where the probability is 1/36, and fifteen frequency counts from the total of twenty one where the probability is 1/18. Checking, we see that $(6 \times 1/36) + (15 \times 1/18) = 1$. By resorting to a correct probabilistic analysis of repeated trials, the source of the confusion is dispelled (except for those “none so blind as will not see”).

Exercise 19.9.9: Use the “entropic–like” formula from the last Chapter to find the probability for two **FOURS** and a **FIVE** in the next three rolls of the fair die.

Solution to Exercise 19.9.9

Let’s adopt a less zealous posture with regard to the IP’s state of knowledge about the causal nature of the die. The IP is assured that the die is actually as “fair” as any die could be based on extensive physical measurements of the die about to be rolled. Furthermore, precautions will be taken to ensure that the die is rolled as “randomly” as possible.

Therefore,

$$\begin{aligned} P(\dots, N_4 = 2, N_5 = 1, N_6 = 0 | \mathcal{M}_A) &= \exp \left[N \left(\left[\frac{\ln W(N)}{N} \right] + \lambda \bar{F} - \ln Z \right) \right] \\ &= \exp \left[3 \left(\left[\frac{\ln W(3)}{3} \right] + (0 \times \bar{F}) - \ln 6 \right) \right] \\ &= 0.013889 \end{aligned}$$

Double–checking this answer with the multinomial distribution,

$$\begin{aligned} P(\dots, N_4 = 2, N_5 = 1, N_6 = 0 | \mathcal{M}_A) &= W(N) Q_1^0 Q_2^0 Q_3^0 Q_4^2 Q_5^1 Q_6^0 \\ &= 3 \times (1/6)^2 \times 1/6 \\ &= 3/216 \\ &= 0.013889 \end{aligned}$$

Exercise 19.9.10: An IP refuses to accept the claim that the die is fair. It will only be convinced by actual observations of the die, the more the better. What is the probability for the same situation as interpreted by this skeptical IP?

Solution to Exercise 19.9.10

A formula was derived in Volume I to deal with future frequency counts when conditioned on known past frequency counts. This formula was a generalization of Laplace’s *Rule of Succession*.

Suppose that the skeptical IP, who is appropriately named Thomas, were willing to update his state of knowledge based on 600 rolls of the suspect die. We will revert

to the notation used in Volume I, where M_i was used for the future frequency counts in question, and N_i for the known frequency counts of all six spots after the die has been rolled 600 times. Thus, the N_i are the observed data. For a completely ridiculous set of data, but one with an obvious agenda, say that all six spots turned up equally often.

We can specify the following for this problem. The dimension of the state space is $n = 6$. $M = 3$ with $M_4 = 2$, $M_5 = 1$, and the other $M_i = 0$. Thus, $\sum_{i=1}^6 M_i = M$. $N = 600$ with all six $N_i = 100$. We can insert these specifics into the general formula,

$$P(M_1, M_2, \dots, M_n | N_1, N_2, \dots, N_n) = C \times \frac{\prod_{i=1}^n (M_i + N_i)!}{\prod_{i=1}^n M_i!}$$

to find the probability for seeing two FOURS and a FIVE in the *next* three rolls of the die.

The constant term C is calculated first, followed by the second term containing the product of the factorials.

$$\begin{aligned} C &= \frac{M! (N + n - 1)!}{N_1! \times \dots \times N_n! (M + N + n - 1)!} \\ &= \frac{3! (600 + 6 - 1)!}{(100! \times \dots \times 100!) (3 + 600 + 6 - 1)!} \\ &= \frac{3! 605!}{(100! \times \dots \times 100!) 608!} \\ \frac{\prod_{i=1}^n (M_i + N_i)!}{\prod_{i=1}^n M_i!} &= \frac{100! \times 100! \times 100! \times 102! \times 101! \times 100!}{0! \times 0! \times 0! \times 2! \times 1! \times 0!} \\ C \times \frac{\prod_{i=1}^n (M_i + N_i)!}{\prod_{i=1}^n M_i!} &= \frac{3!}{100! \times \dots \times 100! \times 606 \times 607 \times 608} \times \\ &\quad \frac{100! \times 100! \times 100! \times 102! \times 101! \times 100!}{2!} \\ &= \frac{3 \times 102 \times 101 \times 101}{606 \times 607 \times 608} \\ &= 0.013957 \end{aligned}$$

In the end, after these tedious calculations, we have the probability of seeing two FOURS and a FIVE in the *next* three rolls of the die when conditioned on all of the data in 600 previous rolls,

$$P(M_1 = 0, M_2 = 0, M_3 = 0, M_4 = 2, M_5 = 1, M_6 = 0 | \mathcal{D}) = 0.013957$$

A large amount of data, and moreover data supporting the hypothesis of a fair die to the greatest extent possible, has forced the skeptical IP *almost* to the position of the first IP (compare 0.013957 with 0.013889). This IP believed the claim that the die was fair based solely on its physical characteristics.

Because of the data, the skeptical IP holds a measure of the degree of belief about the truth of seeing two FOURs and a FIVE in the next three rolls nearly identical to that of the non-doubting IP. Any lingering residual influence from adopting Laplace's insufficient reason to believe exclusively in a fair die, or any other model for that fact, at the very beginning of the inference has dissipated into thin air after 600 rolls of the die.

The other important lesson from this exercise is that Thomas did NOT have to rely on the MEP. He effectively averaged over *all* possible numerical assignments for the future frequency counts. Whether any one particular assignment to the Q_i came from the MEP algorithm, or the Oracle at Delphi, made no difference whatsoever to Thomas because *every* possible legitimate numerical assignment to the six Q_i was averaged over. The actual origin of the numerical assignments was irrelevant.

The IP who believed in the fair die, and therefore adopted just the one single operative model for assignment to the Q_i , did NOT have to use the MEP algorithm either. Nonetheless, it is very convenient, whether the IP adopts a single (or a few) model(s) at the beginning, or is forced by the data to adopt a single model (or a few) at the end, to have an informational rationale for those few models. And that is exactly what the MEP provides to the IP.

Exercise 19.9.11: Suppose that the model space for the dice consists of just the four models discussed in this Chapter. How does the fair model get updated if the die is rolled once and a SIX is observed?

Solution to Exercise 19.9.11

We have a model space composed of four models, \mathcal{M}_A through \mathcal{M}_D . Each has a probability of 1/4 prior to any data. If the die is rolled once and a SIX is observed, then the probability for the fair model is updated to,

$$\begin{aligned} P(\mathcal{M}_A | \mathcal{D}) &= \frac{P(\mathcal{D} | \mathcal{M}_A) P(\mathcal{M}_A)}{P(\mathcal{D})} \\ &= \frac{P(\mathcal{D} | \mathcal{M}_A) P(\mathcal{M}_A)}{\sum_{k=1}^4 P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k)} \\ &= \frac{0.1667 \times 1/4}{(0.1667 \times 1/4) + (0.2468 \times 1/4) + (0.2687 \times 1/4) + (0.3333 \times 1/4)} \\ &= 0.164128 \end{aligned}$$

The degree of belief in the fair model started out at,

$$P(\mathcal{M}_A) = 0.25$$

Now, the degree of belief is lessened slightly to,

$$P(\mathcal{M}_A | \mathcal{D}) = 0.164128$$

because the other three models have all assigned a higher probability to observing a SIX.

Exercise 19.9.12: Speculate on how a frequentist would like to use the MEP formalism?

Solution to Exercise 19.9.12

Suppose that the die has been rolled $N = 6,000$ times. The IP is to work from the given “information” that the average number of spots over these 6,000 rolls is, say, 4.5. What are the “most probable” N_i ?

For the frequentist, the pre-eminent role is played by the multiplicity factor $W(N)$. This is to be maximized subject to two constraints. The first constraint is that the sum of the N_i must equal N , while the second constraint is that the *sample* average,

$$\overline{F}(X = x_i) = \sum_{i=1}^6 \frac{N_i}{N} \times F(X = x_i)$$

must equal 4.5. It is implicit for the frequentist that $F(X = x_i) = (1, 2, 3, 4, 5, 6)$.

The largest value the multiplicity factor could ever achieve would of course be realized when all six $N_i = 1,000$.

$$W(N) = \frac{6000!}{1000! \times \dots \times 1000!}$$

This satisfies the first constraint that the N_i sum to 6,000, but unfortunately this choice for the N_i yields,

$$\overline{F} = \frac{\sum_{i=1}^6 N_i F(X = x_i)}{N} = \frac{21000}{6000} = 3.5$$

for the sample average.

It is decided that the N_i must be adjusted in some fashion to satisfy this second constraint. The frequency counts of $N_1 = 1000$, $N_5 = 4000$, $N_6 = 1000$ with the three remaining N_i all equal to 0 would seem to do the trick. The first constraint is satisfied, $\sum_{i=1}^6 N_i = N$, and the second constraint is satisfied as well,

$$[(1000 \times 1) + (0 \times 2) + (0 \times 3) + (0 \times 4) + (4000 \times 5) + (1000 \times 6)]/6000 = 4.5$$

However, we don't know whether the multiplicity factor for these frequency counts is also the largest one that could be found. There might be other frequency counts that also satisfy these two constraints, but which could happen in a greater number of ways.

In fact, the frequency counts of,

$$N_1 = 326$$

$$N_2 = 473$$

$$N_3 = 684$$

$$N_4 = 993$$

$$N_5 = 1440$$

$$N_6 = 2084$$

have the largest multiplicity factor while still managing to satisfy both constraints.

Compare,

$$\frac{6000!}{1000! 0! 0! 4000! 1000!} = 9.06 \times 10^{2256}$$

to,

$$\frac{6000!}{326! 473! 684! 993! 1440! 2084!} = 5.75 \times 10^{4195}$$

Two remarks are in order here. First of all, *Mathematica* did not choke when asked to compute these enormous numbers, and secondly, it is clear why we choose to work with $\frac{\ln W(N)}{N}$ which yield for the two cases the values 0.866139 and 1.61018.

Exercise 19.9.13: What do the formal rules of probability theory say if the frequency counts in the previous exercise were actually data?

Solution to Exercise 19.9.13

The IP seeks the proper degree of belief for, say, the statement, "The *next* roll of the die will result in a THREE." The probability for this statement is conditioned on another statement, "The die was rolled 6,000 times and a ONE appeared 326 times, . . . , and a SIX appeared 2084 times." In other words, the degree of belief for that statement about which the IP is not certain is conditioned on that statement which is certain, the known data.

We have derived the formula based on the formal manipulation rules for this situation as captured in the expression $P(M_1, M_2, \dots, M_n | \mathcal{D})$.

$$P(M_1 = 0, \dots, M_3 = 1, \dots, M_6 = 0 | \mathcal{D})$$

where $\mathcal{D} \equiv (N_1 = 326, N_2 = 473, N_3 = 684, N_4 = 993, N_5 = 1440, N_6 = 2084)$

$$\begin{aligned} &= C \times \frac{\prod_{i=1}^6 (M_i + N_i)!}{\prod_{i=1}^6 M_i!} \\ C &= \frac{M! (N + n - 1)!}{(\prod_{i=1}^6 N_i!) (M + N + n - 1)!} \\ &= \frac{1! 6005!}{326! 473! 684! 993! 1440! 2084! (1 + 6000 + 6 - 1)!} \\ &= \frac{1}{326! 473! 684! 993! 1440! 2084! 6006} \\ \frac{\prod_{i=1}^6 (M_i + N_i)!}{\prod_{i=1}^6 M_i!} &= \frac{326! 473! 685! 993! 1440! 2084!}{0! 0! 1! 0! 0! 0!} \\ P(M_1 = 0, \dots, M_3 = 1, \dots, M_6 = 0 | \mathcal{D}) &= \\ &= \frac{1}{326! 473! 684! 993! 1440! 2084! 6006} \times 326! 473! 685! 993! 1440! 2084! \\ &= \frac{685}{6006} \\ &= \frac{N_3 + 1}{N + n} \\ &= 0.114053 \end{aligned}$$

Note that from the probabilistic perspective, the IP does NOT want the “most probable” frequency counts, but rather a probability for some event conditioned on known frequency counts.

The ironic thing here is that arriving at this correct answer, the IP did NOT have to create any particular probability distribution by resorting to the MEP. All probability distributions were taken into account. And they were taken into account by relying on Laplace’s *Principle of Insufficient Reason* to assign a uniform distribution over model space. As Jaynes said, Laplace did a better job of showing how frequencies are used within probability theory than the frequentists ever did!

With ten times as much data where $N_3 = 6850$ and $N = 60000$, this probability becomes,

$$P(M_1 = 0, \dots, M_3 = 1, \dots, M_6 = 0 | \mathcal{D}) = \frac{6851}{60006} = 0.114172$$

This particular number for the probability becomes very interesting when Jaynes’s dice problem is revisited in the next Chapter.

Exercise 19.9.14: From the point of view of statistical mechanics, provide a very detailed accounting of the numbers involved when discussing an “assembly of systems.”

Solution to Exercise 19.9.14

For an easy numerical example, consider the “assembly of systems” to consist of $N = 3$ dice. The dice (the systems) are distinguished by being differently colored, but all three dice are rolled at the same time. The total energy E of this assembly is the sum of the spots on the three upturned die faces. For some assembly, the total energy is specified to be $E = 9$. Statistical mechanics wants to find the most probable occupancy numbers of the energy states of this assembly.

We now discuss all of the various counts that ensue in the resolution to this inferential problem. The dimension of the state space is $n = 6$, while the number of trials is $N = 3$. The total number of elementary points in the sample space is $n^N = 6^3 = 216$. The total number of possible frequency counts (and this is the same as the total number of distinct contingency tables) is,

$$u = \frac{(N + n - 1)!}{N! (n - 1)!} = \frac{(3 + 6 - 1)!}{3! 5!} = 56$$

Now, the total number of elementary points in the sample space is a sum over the multiplicity factors for these 56 possible contingency tables,

$$n^N = \sum_{j=1}^u W_j(N) = \sum_{j=1}^{56} W_j(3) = 216$$

The upper limit $u = 56$ in this summation can be broken down into three classes each with its own unvarying multiplicity factor for the number of trials $N = 3$. Any contingency table can assume one of the three general patterns,

Class 1. 3+0+0+0+0+0

Class 2. 2+1+0+0+0+0

Class 3. 1+1+1+0+0+0

indicating, in turn, all three dice with same face up, two dice with the same face up and the third die different, and finally all three dice with a different face up.

The decomposition of the total number of contingency tables $u = 56$ is then 6 possibilities for the first class of contingency tables, 30 for the second, and 20 for the third. These numbers are found from the combinatorial formula presented in Volume I,

$$\frac{n!}{r_z! r_s! r_d! r_t!}$$

For example, the 30 possibilities for the second class of contingency tables where two dice show the same face and the third die a different face are found by,

$$\frac{n!}{r_z! r_s! r_d! r_t!} = \frac{6!}{4! 1! 1! 0!} = 30$$

Furthermore, each one of these 30 contingency tables can happen in 3 different ways when the distinguishability of the three differently colored dice is taken into account. This is calculated by the multiplicity formula as,

$$W(N) = \frac{N!}{N_1! N_2! \cdots N_6!} = \frac{3!}{2! 1! 0! 0! 0! 0!} = 3$$

For example, take the case where one THREE and two SIXes appear. This belongs to the second class of 30 possibilities with two dice the same and the third different. Switching to Feller's notation for elementary points in the sample space, one THREE and two SIXes can happen in three different ways:

1. $\{- | - | \mathbf{g} | - | - | \mathbf{br}\}$, the green die showed the THREE, while the blue and red die showed a SIX,
2. $\{- | - | \mathbf{b} | - | - | \mathbf{gr}\}$, the blue die showed the THREE, while the green and red die showed a SIX
3. $\{- | - | \mathbf{r} | - | - | \mathbf{gb}\}$, the red die showed the THREE, while the green and blue die showed a SIX.

Thus, the decomposition of the total number of possible frequency counts times their multiplicity factors yields the total number of elementary points in the sample space.

$$n^N = \sum_{j=1}^{56} W_j(N) = (6 \times 1) + (30 \times 3) + (20 \times 6) = 216$$

The total energy of the assembly can range from $E = 3$ to $E = 18$. For example, an assembly of one THREE and two SIXes has a total energy of $E = 15$. It was given that the total energy of the assembly was, in fact, $E = 9$.

There are 6 frequency counts from the grand total of 56 that satisfy the two constraints that $N = 3$ and $E = 9$. One, THREE THREE THREE, is from the first class of frequency counts where all three dice show the same face. Two, ONE FOUR FOUR and TWO TWO FIVE, are from the second class where two faces are the same and the third different. Three, ONE TWO SIX, ONE THREE FIVE, TWO THREE FOUR, are from the third class where all three dice show a different face.

Which of these acceptable candidates can happen in the greatest number of ways? The last three frequency counts, shown here as contingency tables,

1	1	0	0	0	1	1	0	1	0	1	0	0	1	1	1	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

can each happen in six different ways. Therefore, statistical mechanics selects these three by the criterion of possessing the maximum multiplicity factor of 6. Notice that a single “most probable distribution” could not be picked out.

In contrast, the MEP calculates the probability Q_i for each energy state for a single die. $F(X = x_i) \equiv E_i \equiv i$. If a model were to insert the information that the constraint function average was the same as the average energy,

$$\langle F \rangle \equiv \overline{E} = E/N = 3$$

then,

$$Q_i = \frac{e^{\lambda F(X=x_i)}}{\sum_{i=1}^6 e^{\lambda F(X=x_i)}}$$

with the partition function Z equal to,

$$Z = e^\lambda + e^{2\lambda} + e^{3\lambda} + e^{4\lambda} + e^{5\lambda} + e^{6\lambda}$$

The formal rules of probability then take over *after* the MEP assignment. The probability for any one of the 56 contingency tables can then be calculated from,

$$P(N_1, N_2, \dots, N_6 | \mathcal{M}_k) = \exp \left[N \left(\left[\frac{\ln W(N)}{N} \right] + \lambda \overline{F} - \ln Z \right) \right]$$

Take as an example the frequency count ONE TWO SIX picked out by statistical mechanics. Then, with the sample average \overline{F} determined by,

$$\begin{aligned} \overline{F} &= [N_1/N \times F(X = x_1)] + [N_2/N \times F(X = x_2)] + \dots + [N_6/N \times F(X = x_6)] \\ &= (1/3 \times 1) + (1/3 \times 2) + \dots + (1/3 \times 6) \\ &= 3 \end{aligned}$$

and with the values for $\lambda = -0.17479$ and $Z = 3.401$ found by the MEP algorithm, we have,

$$\begin{aligned} P(N_1 = 1, N_2 = 1, \dots, N_6 = 1 | \mathcal{M}_k) &= \exp \left[3 \left(\frac{\ln 6}{3} + 3\lambda - \ln Z \right) \right] \\ &= \exp [3 \times (0.5973 + (3 \times -0.17479) - 1.2241)] \\ &= 0.0316 \end{aligned}$$

The MEP, of course, also finds numerical assignments under this model. The three relevant assignments for a ONE, a TWO, and a SIX are, $Q_1 = 0.247$, $Q_2 = 0.207$ and $Q_6 = 0.103$. This corresponds to,

$$P(N_1 = 1, N_2 = 1, \dots, N_6 = 1 | \mathcal{M}_k) = 6 \times 0.247 \times 0.207 \times 0.103 = 0.0316$$

The important thing here is that the IP has actually calculated a probability about a ONE TWO SIX. It has refrained from some vague assertion about it being, or not being, the “most probable distribution.” The IP has a quantitative degree of belief in the statement that this particular frequency count is TRUE, or equally well, a degree of belief about *any* one of the 56 frequency counts.

If everything is working correctly, the degree of belief about seeing the assembly of the three systems as THREE THREE THREE, a statement that does satisfy the total energy, should be smaller than the above assembly of the three systems as ONE TWO SIX. The MEP formula found that $Q_3 = 0.174$.

$$P(N_1 = 0, N_2 = 0, N_3 = 3, \dots, | \mathcal{M}_k) = 1 \times (0.174)^3 = 0.0053$$

It would seem that the calculation of these degrees of belief would be extremely useful to the IP when it wants to discover on which collection of frequency counts it could be reasonably certain. Thus, it would have to accumulate the probability over some set of frequency counts until the probability reached some criterion defining a level of practical certainty.

Chapter 20

Information and Data

20.1 Introduction

With the elementary examples of these first three Chapters, I have been trying to show how the MEP *should* be applied. But like everything else in this world, any attempted explanation gains in power when it also includes what should *not* be done. The world is awash with a gross misunderstanding of what the MEP is, as well as what it can not be.

I have tried to make it abundantly clear that the MEP is an excellent algorithm for assigning legitimate numerical values to probabilities for the n joint statements in the state space. These numerical values reflect the information inserted by the IP into the distribution under some model. However, it is absolutely critical that it be understood that these assignments are made only for the n statements in the state space.

No further usage of the MEP algorithm is required elsewhere up the inferential chain. In particular, the probability for the $(N+1)^{st}$ toss of a die is NOT found by applying the MEP algorithm to data from the previous N tosses. The probability for the $(N+1)^{st}$ toss is *calculated* from the formal manipulation rules of probability theory. The theory dictates that *all* assignments, irrespective of their origin, must be averaged over if the IP is ignorant of the cause for a die face to appear.

The MEP algorithm is responsible for only one assignment task. Then its job is finished. Once an assignment has been made under some model, it remains in effect for all future trials.

The critical remarks begun in the last Chapter which were directed against one version of Jaynes's explanation of the general formalism of the MEP demand something more from me. There exists so much attendant confusion surrounding *information* and *data* that it threatens to undermine any coherent understanding of the MEP.

On many occasions, Jaynes tells us that the *data* is the information being inserted into a probability distribution. On other occasions, he tells us that the information one should use in the MEP are the averages of arbitrary functions. Which is it?

It is time to disentangle, once and for all, the confusion surrounding this aspect of the MEP. Recall from Volume I, and the opening Chapters here in Volume II, that one of the most fundamental and critical conceptual notions in inference is the distinction between **probabilities** and **frequencies**. A useful and concise phrase to hang our hats on is a philosophical one that probabilities are epistemological, while frequencies are ontological.

The very same distinction applies to **information** and **data**. Information is an epistemological concept, while data is an ontological concept. Entropy, which is a quantitative measure of the amount of *missing information*, must perforce be an epistemological concept.

Because Jaynes liked to roll the dice so often in his examples illustrating the MEP, we will stick to that scenario for our numerical examples in this Chapter. As a gentle introduction before my harsher critical remarks, we will duplicate Jaynes's answer for the dice problem as he constructed it in his Brandeis lectures.

Jaynes had no equal when it came to promoting and explaining the rationale behind the principles of Maximum Entropy. My personal admiration for him is unqualified. I consider him to be a true genius.

Nonetheless, he has caused untold confusion in the minds of many with some aspects of his explanation of maximum entropy. Especially egregious is some of the language he employed, which, I have no doubt, is the source of unending miscomprehension. What I hope the commentary in this Chapter accomplishes is to pinpoint where Jaynes was right, but more importantly, where he led us astray.

After his two seminal articles in 1957 introducing the notion that information entropy should be of fundamental concern when solving inferential problems, the issue was revisited in 1962 in the lectures given at Brandeis University. As he says, to begin the lectures he merely wanted to illustrate some basic *qualitative* features of maximum entropy.

He decided to use an example of a weighted or “biased” die to sketch out the basic characteristics that a maximum entropy assignment must possess. After these preliminary remarks, he launched into what he considered to be the much more serious Physics content of his talks. He never did get around to presenting the complete quantitative maximum entropy solution to the “Brandeis dice problem.”

He rectified that oversight in 1976 in that extremely important article [18] mentioned in the last Chapter. Jaynes was himself surprised that his little introductory problem gained world wide notoriety, and which became, as he says, a *cause célèbre*. OK, Jaynes said, I will show you the full quantitative maximum entropy solution to this biased die scenario that I sketched out for you a more than a decade earlier.

20.2 The MEP, Data, and Information

One goal we have set for ourselves is to confirm the validity of Jaynes's numerical results of the Brandeis dice problem as presented in his 1976 article,

Where Do We Stand on Maximum Entropy?

We do so by using a different mathematical tactic than Jaynes relied on. Since the fundamental underlying rationale is still that of the maximum entropy principle, my route to the results duplicate Jaynes's findings exactly. From this perspective, we are simply presenting one more illustration of the MEP algorithm in action.

After showing where Jaynes was right, I go into very specific detail on where he was wrong. This is the topic where he unfortunately contributed to the massive, and seemingly irredeemable, confusion surrounding the maximum entropy principle. This confusion echoes down to the present day. In a nutshell, Jaynes persisted in using the ontological word "data" interchangeably with the epistemological word "information" when explaining the general formalism underlying the MEP.

There is one fundamental conceptual notion that I will harp on constantly. It is a quite straightforward principle, but one which very few people seem to comprehend. Because it is so important, I beg the reader's indulgence for raising my voice on this matter:

**THE ACTUALLY OBSERVED MEASUREMENTS FROM
AN EXPERIMENT, NAMELY THE DATA, OR THE
AVERAGE OF ANY DATA, OR ANY FUNCTION OF THE
DATA, HAVE ABSOLUTELY NOTHING, AND LET ME
REPEAT, NOTHING TO DO WITH THE INFORMATION
INSERTED INTO A PROBABILITY DISTRIBUTION BY
THE MAXIMUM ENTROPY ALGORITHM.**

20.3 Validating Jaynes's Numerical Results

The solution to any inferential problem must always begin with a definition of the state space. The state space consists of all the statements in the particular problem to which a probability will be assigned. Usually, these statements are joint statements, but the Brandeis die scenario is so elementary that all of the statements in the state space are simple statements.

The statements in the state space for the Brandeis die scenario are of the following familiar form, "The die showed a ONE spot.", or, "The die showed a TWO spot." through the final statement, "The die showed a SIX spot." Thus, in this problem, the state space has dimension of $n = 6$.

A probability must be attached to each one of these six statements in the state space to express the IP's degree of belief that this statement is TRUE. When the IP gets around to actually assigning legitimate numerical values to these probabilities, these assignments must be conditioned on some information. In problems to do with making inferences, we generally agree to label this information as a model with the accompanying notation that the k^{th} model is written as \mathcal{M}_k .

Continue to use the generic notation ($X = x_i$) to stand for the statement that X was observed to have the value of x_i . Thus, we write the probability assigned to the statement, "The die showed a THREE spot." as conditioned on the truth of the information inserted by some model as,

$$P(X = x_3 | \mathcal{M}_k) \equiv$$

$$P(\text{"The die showed a THREE spot."} | \text{"Information in Jaynes's model."})$$

Notice especially that I have emphasized that what appears to the left and to the right of the conditioned upon symbol are statements.

In order to make it easier on the reader, we will make sure to give the translation back and forth between our preferred notation and the notation that Jaynes used. For example, Jaynes uses p_i instead of our more expanded notation,

$$p_i \equiv Q_i \equiv P(X = x_i | \mathcal{M}_k)$$

Only one constraint function was utilized in Jaynes's illustrative model. Therefore, with $m = 1$ and $n = 6$, additional freedom exists to construct four more constraint functions under more complicated models. But to duplicate Jaynes's findings we will keep his model that set $k = m = 1$. The one constraint function is,

$$f_k(x_i) = f_1(x_i) \equiv F(X = x_i) = (1, 2, 3, 4, 5, 6)$$

This is just an arbitrary mapping from the set of statements to numbers. In a way, Jaynes could have told us far more if he had chosen a different mapping. Most people don't think about this subtlety. They automatically pass over to assuming that the constraint function must be the same as the actual data. Not so.

The information inserted into the probability distribution under Jaynes's model SHOULD BE the constraint function average, or expectation, of $f_1(x_i)$,

$$\sum_{i=1}^n p_i f_1(x_i) = \langle f_1(x) \rangle \equiv \sum_{i=1}^6 Q_i F(X = x_i) = \langle F \rangle = 4.5 \quad (20.1)$$

With this specification of the information, we have arrived at the heart of the confusion! We will elaborate on this fundamental point further on down the road, but Jaynes clearly and unmistakably gives the reader the impression that this value of 4.5 is DATA!

The die has been tossed N times, and this value of 4.5 is the average value of the observed tosses. Jaynes clearly says that the maximum entropy solution was to be thought of as applying to the *very next toss of the die after these N tosses had been made*.

Nothing could be further from the truth! The information inserted by this model is the constraint function average, and, as this implies, it must be the average of the $f_1(x_i)$, or, in other words, the average of (1, 2, 3, 4, 5, 6) with respect to the numerical assignment p_i that we are trying to find. It may lie anywhere in the range from 1 to 6.

These arbitrary numbers (1, 2, 3, 4, 5, 6) have nothing to do with the possible fact that a FIVE may have been observed on the first toss of the die, and a FOUR may have been observed on the second toss of the die leading to an average of 4.5 over two tosses. Conceptually, the average of a constraint function is NOT the average of any data, but there is nothing to prevent a data average from exactly matching a constraint function average!

Moreover, the numerical assignment to a probability under the information in some model is apropos for the first roll, for the second roll, or for the ten millionth roll of the die. Remember that,

$$P(X_N, \dots, X_2, X_1 | \mathcal{M}_k) = P(X_1 | \mathcal{M}_k) \times P(X_2 | \mathcal{M}_k) \times \dots \times P(X_N | \mathcal{M}_k)$$

An essential feature of the formal manipulation rules is that the probability at each trial depends only on the model, and not on any past data. The MEP assignment is NOT for the $(N + 1)^{st}$ roll after observing N rolls.

It is clearly obvious that the minimum value of $\langle F \rangle$ must be 1, and the maximum value of $\langle F \rangle$ must be 6. Otherwise, any number between these two anchor points can be used as information by a model. 4.5 is certainly a legitimate value for $\langle F \rangle$ as is 4.6, 1.9639, 5.999, and so on. In fact, every conceivable value of $\langle F \rangle$ must be used to generate every conceivable numerical assignment to the probabilities for the statements in the state space when the integration over model space is carried out.

Despite all of this carping, the MEP formula is definitely useful for assigning numerical values to the six statements in the state space under a model which has inserted the information that $\langle F \rangle = 4.5$.

$$Q_i \equiv P(X = x_i | \mathcal{M}_k) = \frac{e^{\lambda F(X=x_i)}}{Z(\lambda)}$$

Jaynes, in the course of deriving the maximum entropy formula, chooses the expression where the Lagrange multipliers are the negative of mine, so his formula looks like this,

$$p_i = \frac{e^{-\lambda f(x_i)}}{Z(\lambda)}$$

But this is of no consequence. Numerically, my Lagrange multipliers are the negative of his, so we come out at the same place in the end.

Jaynes proceeded to solve for the relevant values through an important consequence of the MEP formalism that relates the partition function and the Lagrange multiplier to the constraint function average,

$$-\frac{\partial \ln Z(\lambda)}{\partial \lambda} = 4.5$$

and found the following numerical values through an unnecessarily complicated root finding procedure for a polynomial,

$$\lambda = -0.37105$$

$$Z = 26.66365$$

$$p_i = (0.05435, 0.07877, 0.11416, 0.16545, 0.23977, 0.34749)$$

$$S = 1.61358$$

S is another symbol that is sometimes used, instead of our version of information entropy $H_{max}(Q_i | \mathcal{M}_k)$, to quantify the amount of missing information in any particular numerical assignment. By definition, this value must be the largest value possible of any numerical assignment that satisfies the same information.

Let's double-check these results Jaynes arrived at by using our version of the MEP algorithm. For p_4 , Jaynes has an assignment of 0.16545. p_4 is Jaynes's notation for the numerical assignment to the probability of a FOUR spot.

In our notation, it is,

$$p_4 \equiv Q_4 \equiv P(X = x_4 | \mathcal{M}_k)$$

As already mentioned, the value for the Lagrange multiplier found by our method is the negative of Jaynes's, so $\lambda = +0.37105$. The normalization factor $Z(\lambda)$ is,

$$\begin{aligned} Z(\lambda) &= \sum_{i=1}^6 e^{\lambda F(X=x_i)} \\ &= e^{\lambda \times 1} + e^{\lambda \times 2} + e^{\lambda \times 3} + e^{\lambda \times 4} + e^{\lambda \times 5} + e^{\lambda \times 6} \\ &= e^{.37105 \times 1} + e^{.37105 \times 2} + e^{.37105 \times 3} + e^{.37105 \times 4} + e^{.37105 \times 5} + e^{.37105 \times 6} \\ &= 26.66365 \end{aligned}$$

and so,

$$\begin{aligned} Q_4 &= \frac{e^{\lambda \times 4}}{Z(\lambda)} \\ &= \frac{e^{.37105 \times 4}}{26.66365} \\ &= 0.16545 \end{aligned}$$

It can be checked that the remaining Q_i also match Jaynes's p_i .

To confirm Jaynes's numerical results, I used the completely independent tactic of minimizing the following function¹

$$S(\langle F \rangle) = \ln Z(\lambda) - (\lambda \times \langle F \rangle) \quad (20.2)$$

keeping $\langle F \rangle$ fixed at 4.5 while allowing λ to vary. *Mathematica* found that the minimum of this function occurred at 1.61358 at a $\lambda = 0.37105$. Happily then, the rest of the computations for $Z(\lambda)$ and the Q_i , although carried out independently of Jaynes's calculations, must now also match Jaynes's results.

So we have confirmed down to every detail Jaynes's numerical results in his example of how to assign numerical values to the probabilities for the faces of the die under some specific model. So what is the problem? That is taken up in earnest in the next section.

20.4 Where Jaynes Was Wrong

Return to the original 1962 Brandeis lecture [16, pg. 41]. Let's repeat what Jaynes said while introducing the qualitative features of the maximum entropy solution, but now with an emphasis on the incriminating terms.

A die has been tossed a very large number N of times, and we are told that the average number of spots up per toss was not 3.5, as we might expect from an honest die, but 4.5. Translate this *information* into a probability assignment $P_n, n = 1, 2, \dots, 6$ for the n -th face to come up *on the next toss*. [my emphasis]

Let's sneak up on Jaynes's prescription as given above by looking at special cases where N is small. Consider the expression for the probability of seeing, say, a ONE on the next toss after two tosses have already been made,

$$P(X_3 = x_1 | X_1 = x_4, X_2 = x_5)$$

where the data consist of just two past observations. The first toss showed a FOUR and the second toss showed a FIVE.

What follows is conceptually very important: whatever appears inside the probability operator, $P(\dots | \dots)$, MUST BE STATEMENTS! Both expressions $(X_3 = x_1)$ and $(X_1 = x_4, X_2 = x_5)$ must refer to statements, and not to numbers. This is the very reason why constraint functions, the $F_j(X = x_i)$, appear in the MEP formalism. Those absolutely essential numbers, not statements, required for our computations must eventually show up.

You cannot literally form a “data average” because the data are statements, not numbers. The conditioning statements $(X_1 = x_4, X_2 = x_5)$ mean, “A FOUR was

¹This alternative mathematical technique for the MEP algorithm follows from the use of the Legendre transformation. This will be explained later, although once again, Jaynes gave all of the essential details.

observed on the first roll and a FIVE was observed on the second roll.” The only way you can get a number like 4.5 is through $F(X = x_4) = 4$ and $F(X = x_5) = 5$ where numbers are mapped onto the statements through the constraint functions.

We should not even be calling such things conceptual errors. It really is a matter of not adhering to the definitions of the rules of the game. Probabilities are quantitative measure of the degree of belief that STATEMENTS are true.

Therefore, by definition, statements must appear both to the left and right of the conditioned upon symbol. When we write something like,

$$P(X_3 = x_1 | \mathcal{D})$$

where $\mathcal{D} \equiv (X_1 = x_4, X_2 = x_5)$, the IP is expressing its degree of belief that the *statement* that a ONE will appear on the next toss is true, assuming that the data is true, that is, assuming that the *statement* that a FOUR was observed on the first trial, and the *statement* that a FIVE was observed on the second trial is true.

If we take Jaynes literally at his word, he wants us to form the probability distribution for, say,

$$P(X_{N+1} = x_i | \mathcal{D}) \equiv P(X_{N+1} = x_i | X_1 = x_j, X_2 = x_k, \dots, X_N = x_l)$$

to find the probability for the i^{th} face to come up on the $(N+1)^{st}$ toss after having observed the results from N previous rolls of the die.

The formal manipulation rules of probability tell us what we must do in this case. Repeating the same kind of argument as in Volume I,

$$\begin{aligned} P(X_{N+1}, \mathcal{M}_k, \mathcal{D}) &= P(X_{N+1} | \mathcal{M}_k, \mathcal{D}) P(\mathcal{M}_k | \mathcal{D}) P(\mathcal{D}) && \textbf{Product Rule Step 1} \\ \frac{P(X_{N+1}, \mathcal{M}_k, \mathcal{D})}{P(\mathcal{D})} &= P(X_{N+1} | \mathcal{M}_k, \mathcal{D}) P(\mathcal{M}_k | \mathcal{D}) && \textbf{Division Step 2} \\ \frac{P(X_{N+1}, \mathcal{M}_k, \mathcal{D})}{P(\mathcal{D})} &= P(X_{N+1} | \mathcal{M}_k) P(\mathcal{M}_k | \mathcal{D}) \text{ Assignment by model only Step 3} \\ \frac{P(X_{N+1}, \mathcal{M}_k, \mathcal{D})}{P(\mathcal{D})} &= P(X_{N+1}, \mathcal{M}_k | \mathcal{D}) && \textbf{Bayes's Theorem Step 4} \\ \sum_{k=1}^{\mathcal{M}} P(X_{N+1}, \mathcal{M}_k | \mathcal{D}) &= P(X_{N+1} | \mathcal{D}) && \textbf{Sum Rule Step 5} \\ P(X_{N+1} = x_i | \mathcal{D}) &= \sum_{k=1}^{\mathcal{M}} P(X_{N+1} = x_i | \mathcal{M}_k) P(\mathcal{M}_k | \mathcal{D}) \text{ From Step 3 Step 6} && (20.3) \end{aligned}$$

Scrutinize the simplest possible case. The die has been rolled $N = 2$ times where a FIVE showed up on the first roll, and a FOUR showed up on the second. For the sake of Jaynes’s argument, we will temporarily ignore the conceptual problem just discussed concerning a statement and a number.

Thus, we'll say that this outcome satisfies the constraint that the average of the data was 4.5. Would Jaynes then have us believe that the probability of a ONE on the third toss (the next toss) is, in fact, $p_1 = 0.05435$ as based on these data, and not on the model which used a constraint function average?

The correct solution must flow from the fundamental principles of probability theory as alluded to above and encapsulated in Equation (20.3). Therefore, the correct solution to finding the probability of ONE on the third toss given that a FIVE was observed on the first toss and a FOUR on the second toss is,

$$P(X_3 = x_1 | X_1 = x_5, X_2 = x_4) = \sum_{k=1}^{\mathcal{M}} P(X_3 = x_1 | \mathcal{M}_k) P(\mathcal{M}_k | \mathcal{D})$$

Relying upon the *Rule of Succession*, we find that the probability of seeing a ONE on the next roll of the die based on these data is, in fact,

$$P(X_3 = \text{ONE} | \mathcal{D}) = \frac{N_1 + 1}{N + n} = \frac{1}{2 + 6} = \frac{1}{8} = 0.125$$

The fourth exercise in this Chapter looks at an approximation from first principles where only three models, instead of all possible models, are used. The answer found by relying upon these three models is $P(X_3 = \text{ONE}) = 0.13$.

There is a glaring discrepancy here. We do not get the probability of,

$$p_1 = 0.05435$$

that Jaynes wants us to get. Of course, we couldn't possibly. We are talking about two different things. One is $p_1 \equiv Q_i \equiv P(X = x_1 | \mathcal{M}_k)$, and the other is $P(X_3 = x_1 | \mathcal{D})$. We are talking of the proverbial apples and oranges here!

There are other incongruities. How would you ever update the probability assignment of the one model \mathcal{M}_1 that we have been calling Jaynes's model if you started rolling the die and observed the outcomes? At what point would you tell someone that they could now employ the MEP algorithm because N was "large enough?" What if you don't envision your data as being the average of anything, but you still want to make legitimate numerical assignments to probabilities by following the MEP prescription?

Let there be no misunderstanding. It is quite true that the eventual model that is supported to the greatest extent after N becomes extremely large converges on that model containing the *information* most closely matching the *data* in N rolls.

But this is a consequence of Equation (20.3)! It has nothing to do with the MEP, or its role in the initial assignment of numerical values under some model. In fact, the initial distribution of the models must have uniformly covered every possible numerical assignment to the probabilities in the state space for this fortuitous convergence to have taken place. One model, like Jaynes's model specifying 4.5 in the Brandeis die scenario, won't do the trick.

[Eqn. on page 96](#)

20.5 The Probability of the Data

This is a good place to review the formal manipulation rules as they apply to finding the probability of the data. In the overall prediction equation for a future event, we always have the term representing the probability of the k^{th} model conditioned on the data, $P(\mathcal{M}_k | \mathcal{D})$.

By Bayes's Theorem, this updated probability for a model is,

$$\begin{aligned} P(\mathcal{M}_k | \mathcal{D}) &= \frac{P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k)}{P(\mathcal{D})} \\ &= \frac{P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k)}{\sum_{k=1}^{\mathcal{M}} P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k)} \end{aligned}$$

When we transition to continuous model space, and under the assumption that the IP is completely uninformed, the probability of the data becomes,

$$\begin{aligned} P(\mathcal{D}) &= \sum_{k=1}^{\mathcal{M}} P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k) \\ &= \int \cdot \int_{\sum q_i=1} P(N_1, N_2, \dots, N_n | q_1, q_2, \dots, q_n) P(q_1, q_2, \dots, q_n) dq_i \\ &= \frac{N! (n-1)!}{(N+n-1)!} \end{aligned}$$

For example, when the die has been rolled twice and the results observed as data, the probability of any possible data has the same value of,

$$P(N_1, N_2, \dots, N_6) = \frac{2! 5!}{(2+6-1)!} = \frac{1}{21}$$

There are a total of 21 possible frequency counts, or contingency tables, and all of these have the same probability of 1/21. The probability of data consisting of two SIXes is 1/21. The probability of data consisting of a THREE and a FOUR is 1/21.

20.6 Large N Examples

The crucial formula for solving the problem the way Jaynes actually posed it is Laplace's *Rule of Succession*. Fortunately, the *Rule of Succession* is derived by following the formal manipulation rules of probability theory. Simply stated, the formula is,

$$P(i^{th} \text{ spot on next throw} | \mathcal{D}) = \frac{N_i + 1}{N + n} \quad (20.4)$$

Jaynes wants us to assume that the die has been tossed a “large number of times” as designated by N . The average of the data, that is, the average of the large number of tosses is the “information.” Let $N = 600$ represent a large number of tosses of the die. The FOUR spot showed up 300 times and the FIVE spot showed up 300 times. We have $\sum_{i=1}^6 N_i = N = 600$ with the data average $N^{-1} \sum_{i=1}^6 N_i \times f_1(x_i) = 4.5$.

The probability for a FOUR on the next toss given these data,

$$P(M_1 = 0, \dots, M_4 = 1, \dots, M_6 = 0 | N_1 = 0, N_2 = 0, N_3 = 0, N_4 = 300, N_5 = 300, N_6 = 0)$$

that is, the probability of a FOUR on the 601st toss of the die, is correctly calculated as,

$$\begin{aligned} P(X_{601} = x_4 | \mathcal{D}) &= \frac{301}{606} \\ &= 0.4967 \\ &\approx 1/2 \end{aligned}$$

This answer is eminently reasonable, and it might be accompanied by an equally reasonable physical supposition that the “die” is a perfectly symmetrical and balanced cube with FOUR spots on three of the faces and FIVE spots on the remaining three faces. The models, out of all the models initially considered, that are supported to the greatest extent by these data are those that make the assignments close to,

$$P(X = x_4 | \mathcal{M}_k) = P(X = x_5 | \mathcal{M}_k) \approx 1/2$$

with the other four assignments very close to 0.

Consider another example with large N . The die has been tossed an even larger number of times, say $N = 10,000$, with now a different observed outcome of,

$$N_1 = 543$$

$$N_2 = 787$$

$$N_3 = 1142$$

$$N_4 = 1655$$

$$N_5 = 2398$$

$$N_6 = 3475$$

The sample average of the data is 4.5. Conditioned on these data, the probability of a FOUR on the next roll is,

$$P(X_{10,001} = x_4 | \mathcal{D}) = \frac{1656}{10006} \approx 0.1655$$

Of course, a different set of models is supported by these data. These models happen to be very close to Jaynes's MEP model. Compare with $Q_4 = 0.16545$. But, and this is an extremely critical point to understand without any qualifications, this correct answer to the problem as Jaynes posed it was arrived at by the application of the formal rules of probability manipulation. Moreover, it was arrived at by considering all models on an equal basis.

At no point in this derivation did any one MEP assignment enter (or you could say that every conceivable MEP assignment was included). The derivation involved an integration over all conceivable assignments to the Q_i , and not just one conceivable assignment. The correct answer could not possibly have been found otherwise.

20.7 The Ensuing Confusion

Here is a fun exercise for anyone afflicted with Asperger's Syndrome. Start calculating exactly what Jaynes told you to do to find the probability of the next toss after being told only that the average of N tosses is some number, here 4.5. I gave an example earlier for $N = 2$. Here is the solution for $N = 4$.

The sample space consists of $n^N = 6^4 = 1296$ distinct occurrences for four rolls of the die. But only 80 out of these 1296 satisfy the constraint that the average of the observed spots on the four rolls must equal 4.5.

Table 20.1 lists these eight patterns that satisfy the constraint, together with their associated multiplicity factors. As mentioned, the multiplicity factors for the relevant patterns force a sum of 80 elementary points out of the total of 1296 points in the sample space.

Table 20.1: The eight patterns of data that produce a data average of 4.5 in four rolls of a die. The number of ways that each pattern could occur is also shown.

Number	Pattern	Multiplicity Factor
1	SIX SIX FIVE ONE	12
2	SIX SIX FOUR TWO	12
3	SIX SIX THREE THREE	6
4	SIX FIVE FIVE TWO	12
5	SIX FIVE FOUR THREE	24
6	SIX FOUR FOUR FOUR	4
7	FIVE FIVE FIVE THREE	4
8	FIVE FIVE FOUR FOUR	6
<i>Sum</i>		80

Jaynes wants us to focus on that pattern with the maximum multiplicity factor. This is the pattern number 5 of SIX, FIVE, FOUR, THREE with 24 possible ways of occurring. One of these possible 24 ways might have been FOUR on first roll, SIX on second roll, FIVE on third roll and THREE on the last roll.

We now go through exactly the same (correct) routine where we set out the solution for $N = 2$. We can calculate the probability for any spot to occur on the yet unobserved fifth roll. What is the probability to see the FOUR spot come up on the fifth roll of the die?

To answer this, we must average the predictions for the fifth roll under every single conceivable model. This average is taken with respect to the posterior probability for all such models considered. This is Equation (20.3) once again.

$$P(X_5 = x_4 | N_1 = 0, N_2 = 0, N_3 = 1, N_4 = 1, N_5 = 1, N_6 = 1) = \frac{N_4 + 1}{N + n} = 2/10$$

You may be able see what will eventually happen. As N increases, and as we pick out the data that satisfies the given data average, and has the maximum multiplicity factor, we are eventually led to a very large posterior probability for the model which assigns the Q_i according to the information in Jaynes's model.

But the assignment,

$$P(X_{N+1} = x_i | \mathcal{M}_k)$$

in the first term of the prediction equation, Equation (20.3), must take place prior to, and independently of, any data. This is, of course, just the MEP assignment.

This averaging of N data points is conceptually and computationally different from an initial MEP assignment based on information in constraint functions and constraint function averages. Especially so when we are told to wrap our minds around the curious and extraneous fact that we are picking out *some subset of the potential data that might have occurred*.

This is a flagrant example of orthodox statistical thinking about data's role in inference. It is not in any manner whatsoever how a Bayesian thinks about how data are processed!

Bayesians take the data as they come without any pre-conditions. There is no discussion about particular subsets of potential data that might exhibit particular characteristics. Data do NOT have to follow any maximum *multiplicity factor* law.

On the other hand, a probability distribution may be forced, as we demand within the MEP formalism, to follow a maximum *information entropy* law. This is just another example of the fundamental conceptual distinction between ontology and epistemology.

20.8 Connections to the Literature

Because of Jaynes's sometimes faulty verbal explanation of the MEP with regard to the distinction between information and data, there subsequently arose unending misapprehension and confusion about the MEP. This unfortunate confusion persists to the present day.

A couple of pages after introducing the motivating die scenario (we are still at the Brandeis lectures in 1962), Jaynes [16, pg. 45] begins the formal explanation of the Maximum Entropy Principle with this,

The problem is to find the probability assignment $p_i = p(x_i)$ which satisfies *the given data*:

$$p_i \geq 0$$

$$\sum_{i=1}^n p_i = 1$$

$$\sum_{i=1}^n p_i f_k(x_i) = \langle f_k(x) \rangle = F_k \quad k = 1, 2, \dots, m$$

I'm sure that at this point Jaynes would have agreed that he had used very misleading language in calling these constraints, *the given data*. This is information about the p_i and arbitrary constraint functions, and therefore cannot possibly be information coming from the data. In writing $\sum_{i=1}^n p_i f_k(x_i)$, Jaynes obviously intends for us to interpret the constraint function average as an expectation value with respect to the p_i , the MEP distribution, and NOT the sample average of N observed data.

It seems to me he was thinking of the word *data* in some very general sense as anything that might constrain a problem, but he should have internally censored that word *data* from ever making its way onto the page. From the above, even he would have to admit that the p_i , the $f_k(x_i)$, and the $\langle f_k(x) \rangle$ cannot possibly be data, that is, the actual observed measurement of the outcome of the die toss.

I surmise that Jaynes was in a serious dilemma. He was desperately trying to find some rationale, some mode of language, that would appeal to the frequentist and data oriented mind set that controlled the statistical world of the mid-20th Century. Recall that the Brandeis lecture is a product of Jaynes's thinking that took place during the 1950s.

He knew in his heart that simply tossing the Maximum Entropy Principle into this crowd would have meant instant derision and rejection. So, again I speculate that Jaynes relied on this faulty explanation to placate his eventual critics who would attack him for not being "objectively" anchored to the data.

Disappointingly, he hasn't changed a thing in his 1976 paper [18, pg. 244] when he revisits the Brandeis die problem for the full numerical solution.

When a die is tossed, the number of spots up can have any value i in $1 \leq i \leq 6$. Suppose a die has been tossed N times and we are told only that the average number of spots up was not 3.5 as we might expect from an "honest" die but 4.5. Given this information, *and nothing else*, what probability should we assign to i spots on the next toss?

And finally, even in his book published in 2003 [23, pg. 360], there is this language that muddies the issue even further.

... we were using the F_k and $\langle f_k \rangle$ to stand for the same *number*. They are equal because we specified that the expectation values $\{\langle f_1 \rangle, \dots, \langle f_m \rangle\}$ are to be set equal to the given data $\{F_1, \dots, F_m\}$. When we want to emphasize that these quantities are expectation values over the canonical distribution [the MEP distribution], we will use the notation $\langle f_k \rangle$. When we want to emphasize that they are the given data, we will call them F_k .

This gives the distinct impression that the MEP provides an option for defining information in two distinct cases; one where it could be defined as data, and a second where it is not defined as data!

On the other hand, there are many other places in Jaynes's writings where he does impart the clear message that he wants the reader to think about the MEP in the correct fashion. He emphasizes the abstract nature of the constraint functions (but leans towards implying that this is where the Physics enters), together with the associated constraint functions averages as just given arbitrary numbers. And he calls this (correctly) the *information*, and doesn't mention anything about the *data* being the information.

For example, one has to go no further than Jaynes's own discussion of his MEP solution to the Brandeis die scenario in [18, pg. 244] that we recapitulated in section 20.4.

Now, what does this result mean? In the first place, it is a distribution ... on a space of only six points, the sample space S of a single trial. Therefore, our result as it stands is only a means of describing a state of knowledge about the outcomes of a single trial. It represents a state of knowledge in which one has only (1) the enumeration of the six possibilities; and (2) the mean value constraint ...; *and no other information*. The distribution is "maximally noncommittal" with respect to all other matters; it is as uniform (by the criterion of the Shannon information measure) as it can get without violating the given constraint.

I couldn't agree more fully if one understands that what Jaynes calls the "sample space of a single trial" is synonymous with the *state space*. In this commentary on the MEP solution, there is no mention of any data. As a matter of fact, he says, quite correctly, that the MEP distribution is "maximally non-committal" with respect to

all other matters. I interpret “all other matters” literally to mean any future data points. Again, as Jaynes correctly says here, we arrived at the MEP assignment by invoking Shannon’s information measure, defined with respect to the p_i , not by considering future data that might conform to a maximum multiplicity factor.

And Jaynes goes on in the same paper [18, pg. 245] to make my argument in this Chapter even more forcefully!

Any probability distribution over some sample space S enables us to make statements about (i.e., assign probabilities to) propositions or events defined within that space. It does not – and by its very nature cannot – make statements about any event lying outside that space. Therefore, our maximum-entropy distribution does not, and cannot, make any statement about *frequencies*.

Though the MEP assignment can only refer to the n statements in the state space, by constructing the joint probability for one specific sequence of the data \mathbb{S} and a model, the rules for manipulating the probability symbols yield,

$$\begin{aligned} P(\mathbb{S}, \mathcal{M}_k) &\equiv P(X_N, X_{N-1}, \dots, X_2, X_1, \mathcal{M}_k) \\ &= P(X_N | \dots, X_2, X_1, \mathcal{M}_k) \times \dots \times P(X_2 | X_1, \mathcal{M}_k) \times P(X_1 | \mathcal{M}_k) \times P(\mathcal{M}_k) \\ &= P(X_N | \mathcal{M}_k) \times \dots \times P(X_2 | \mathcal{M}_k) \times P(X_1 | \mathcal{M}_k) \times P(\mathcal{M}_k) \\ &= \prod_{i=1}^n Q_i^{N_i} \times P(\mathcal{M}_k) \\ P(\mathcal{D}, \mathcal{M}_k) &= W(N) \prod_{i=1}^n Q_i^{N_i} \times P(\mathcal{M}_k) \end{aligned}$$

where the Q_i are the MEP assignments under model \mathcal{M}_k .

Again, I appeal to Jaynes’s own words [18, pg. 245],

There is indeed a connection between a *probability* p_i in space S and a frequency ... in [sample space]; but we are justified in using only those connections *which are deducible from the mathematical rules of probability theory*.

So the justifiable connection between an MEP assignment to a probability Q_i in the *state space* and the probability for frequency counts as compound events in *sample space* must follow “from the mathematical rules of probability theory,” as in,

$$P(\mathcal{D}) \equiv P(N_1, N_2, \dots, N_n) = \int \cdot \int_{\sum q_i=1} W(N) \prod_{i=1}^n q_i^{N_i} \times P(q_i) dq_i$$

My final quote from the same paper [18, pg. 227] refers to the original inspiration from Boltzmann. This quote takes place in the context where Jaynes is advising us on the correct conceptual stance to take with regard to probabilities assigned via the MEP and actual frequency counts.

In Boltzmann's "method of the most probable distribution," we have already the essential *mathematical* content of the Principle of Maximum Entropy. But in spite of the conventional name, it did not really involve probability. Boltzmann was not trying to calculate a probability distribution; he was estimating some physically real occupation numbers N_k , by a criterion (value of W) that counts the number of physical possibilities; a definite number that has nothing to do with anybody's state of knowledge.

Again, I agree wholeheartedly. Apparently, I cannot resolve my cognitive dissonance between explanations of this stark clarity with equally clear explanations of the diametrically opposite nature.

But only someone who has read everything Jaynes has ever written on this topic multiple times would ever be aware of all of this conflicting and confusing language. There is no doubt in my mind whatsoever that every critic of the MEP has taken to heart Jaynes's version in which he conflates information with data.

And it is quite easy not only to sympathize with their objections, but to agree with them completely when you realize how they might have conceptualized the MEP based on these conflicting explanations. There is simply no way you can bring the Maximum Entropy Principle, Bayes's Theorem, the data, and model reevaluation together in some consistent fashion if you take Jaynes literally at his (sometimes) word that *data is information*.

To my mind, the most egregious of Jaynes's papers emphasizing this insidious "data is information" concept while trying to elucidate the MEP is his [19],

Concentration of Distributions at Entropy Maxima

The following quote appears in this article on page 320,

A random experiment has n possible results at each trial; thus in N trials there are n^N conceivable outcomes . . . Each outcome yields a set of sample numbers $\{N_i\}$ and frequencies $\{f_i = N_i/N, 1 \leq i \leq n\}$, with an entropy

$$H(f_1 \cdots f_n) = - \sum_{i=1}^n f_i \log f_i \quad (1)$$

. . . We examine the combinatorial basis for using—and the consequences of failing to use—the entropy (1) as a criterion for estimating the $\{f_i\}$.

First off, any kind of language intimating that "an estimation of frequencies" will be performed is a non-starter. Frequencies are ontological facts; there can be no "estimating" going on.

Secondly, there is no sense of any “missing information” associated with frequency counts; you either have them or you don’t. Therefore, at the conceptual level, there can be no entropy defined for frequency counts. Entropy can only be defined where it makes sense that “missing information” can be quantitatively measured, and that occurs only for a probability distribution.

Data, that is, the known frequency counts N_i , do have the absolutely essential role of modifying model space. But it DOES NOT follow that the data have to conform to the maximum entropy principle, or to any other principle for that matter. Hopefully, the data conform only to a reflection of reality.

In one way of thinking about it, data and information are diametrically opposed to one another! At the outset, the IP is setting up all conceivable numerical assignments through the auspices of those epistemological concepts of information and entropy. At the same time, though, the IP holds the entrenched wish that any subsequent data should DESTROY almost every single one of those carefully constructed information entropy assignments.

A probability distribution, on the other hand, as a repository for information about the statements in the state space can easily be seen to require something that will minimize all of the missing information. To wit, a probability distribution must maximize the entropy. But, and this is critical, a concept such as the MEP is NOT required to hew to reality at all. It must be honest only to its own internal demands to output numerical assignments that accurately reflect inserted information.

When definite information in the form of constraint function averages, together with missing information in the form of entropy, are combined within the MEP to output, say, a numerical assignment of 0.95 to the probability for HEADS, there is no concomitant restriction that such a procedure must reflect any aspect of reality. That’s a job for the data.

Inserted information in the form of constraint function averages, together with the complementary idea of missing information, MUST follow a maximum entropy principle in order to exclude any extraneous, unwanted information sneaking into a probability distribution. But data can follow any kind of sample average that reflects reality, and consequently will totally ignore entropy. It must ignore entropy because data are ontological facts and not, like probability distributions, repositories of epistemological information!

Finally, if you relied upon Jaynes’s explanation, how would you update the model space when data arrived subsequent to that used to set up the MEP? The flaw is that it seems Jaynes wants you to arrive at *one, definite* model right at the outset by depending on some bogus blending of the MEP with some original set of data.

Was $N = 600$ judged “large enough” to establish the MEP assignment? What happens when the die is rolled the 601st time, the 602nd time, and so on?

For the formal manipulation rules to work properly in updating model space we rely upon,

$$P(\mathcal{M}_k | \mathcal{D}) = \frac{P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k)}{\sum_{k=1}^{\mathcal{M}} P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k)}$$

You don't want to place your faith in just *one* model at the beginning before any data, or, for that matter, at any point along the data collection process. You always want to start out by considering a huge number \mathcal{M} of models, and then let whatever data, whenever it becomes available, inexorably winnow down that initial huge space of models.

The only way the IP can definitively discover which models are supported is by letting the above equation work on each and every piece of data. It doesn't matter to the formal rules whether there is one piece of data or a million. All of the models undergo an appropriate re-orientation without the IP having to worry about when or whether N "is large enough."

20.9 Solved Exercises for Chapter Twenty

Exercise 20.9.1: Review the derivation in Volume I for the probability of future frequency counts conditioned on the known data. Highlight the step at which the probability of the data enters the derivation.

Solution to Exercise 20.9.1

Start off with the expression for the probability of the future frequency counts conditioned on the known data, where $\mathcal{D} \equiv \{N_1, N_2, \dots, N_n\}$,

$$P(M_1, M_2, \dots, M_n | \mathcal{D}) = \int \cdot \int_{\sum q_i=1} P(M_1, M_2, \dots, M_n | \mathcal{M}_k) P(\mathcal{M}_k | \mathcal{D}) dq_i$$

Substitute for $P(\mathcal{M}_k | \mathcal{D})$ the result from Bayes's Theorem,

$$P(M_1, M_2, \dots, M_n | \mathcal{D}) = \frac{\int \cdot \int P(M_1, M_2, \dots, M_n | \mathcal{M}_k) P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k) dq_i}{P(\mathcal{D})}$$

Here, we highlight the step where we substitute in the denominator the result for the probability of the data $P(\mathcal{D})$ as presented in section 20.5,

$$\begin{aligned} P(\dots, M_i, \dots | \mathcal{D}) &= \frac{\int \cdot \int P(M_1, M_2, \dots, M_n | \mathcal{M}_k) P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k) dq_i}{\frac{N!(n-1)!}{(N+n-1)!}} \\ &= \frac{(N+n-1)!}{N!(n-1)!} \times \int \cdot \int P(M_1, M_2, \dots, M_n | \mathcal{M}_k) P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k) dq_i \\ &= \frac{(N+n-1)!}{N!(n-1)!} \times W(M) \times W(N) \times C_D \times \int \cdot \int q_1^{M_1+N_1} \cdots q_n^{M_n+N_n} dq_i \\ &= \frac{(N+n-1)!}{N!(n-1)!} \times \frac{M!}{M_1! M_2! \cdots M_n!} \times \frac{N!}{N_1! N_2! \cdots N_n!} \times (n-1)! \times \\ &\quad \frac{(M_1+N_1)! (M_2+N_2)! \cdots (M_n+N_n)!}{(M+N+n-1)!} \\ &= (N+n-1)! \times \frac{M!}{M_1! M_2! \cdots M_n!} \times \frac{1}{N_1! N_2! \cdots N_n!} \times \\ &\quad \frac{(M_1+N_1)! (M_2+N_2)! \cdots (M_n+N_n)!}{(M+N+n-1)!} \\ &= \frac{M! (N+n-1)!}{(\prod_{i=1}^n N_i!) (M+N+n-1)!} \times \frac{\prod_{i=1}^n (M_i+N_i)!}{\prod_{i=1}^n M_i!} \\ &= C \times \frac{\prod_{i=1}^n (M_i+N_i)!}{\prod_{i=1}^n M_i!} \end{aligned}$$

Exercise 20.9.2: From this result in the previous exercise, derive Laplace's *Rule of Succession* that gives the probability for the very next occurrence of the i^{th} statement.

Solution to Exercise 20.9.2

There are still N previous data points, but the IP is inquiring after just one future frequency count. When $M = M_i = 1$, Laplace's *Rule of Succession* is,

$$\begin{aligned} P(M_1 = 0, M_2 = 0, \dots, M_i = 1, \dots, M_n = 0 | N_1, N_2, \dots, N_n) \\ = \frac{(N+n-1)!}{\prod_{i=1}^n N_i! (N+n)!} \times (N_1! N_2! \dots N_i + 1! \dots N_n!) \\ = \frac{N_i + 1}{N + n} \end{aligned}$$

Exercise 20.9.3: Verify the answer to the problem in section 20.4 by using the prediction formula for future frequency counts.

Solution to Exercise 20.9.3

We need find only the probability for the *next* toss of the die. Thus, the future frequency count is simply $M = 1$. Since we want to find the probability that the next roll of the die is a ONE, we let $M_1 = 1$ and the remaining $M_i = 0$.

$$M_1 = 1$$

$$M_2 = 0$$

$$M_3 = 0$$

$$M_4 = 0$$

$$M_5 = 0$$

$$M_6 = 0$$

where $\sum_{i=1}^n M_i = M = 1$.

This probability is conditioned on the known data, in this case, two previous tosses showing a FOUR and a FIVE. The die was rolled twice in the past, so $N = 2$. The data, that is, the actually observed frequency counts in these two previous rolls

of the die, were,

$$N_1 = 0$$

$$N_2 = 0$$

$$N_3 = 0$$

$$N_4 = 1$$

$$N_5 = 1$$

$$N_6 = 0$$

where $\sum_{i=1}^n N_i = N = 2$.

In addition, we wanted these data to have an average of 4.5 spots, realizing full well that this “average of the data” is a frequency weighted sum over the one given constraint function,

$$\frac{\sum_{i=1}^6 N_i F(X = x_i)}{N} = \frac{(0 \times 1) + \cdots + (1 \times 4) + (1 \times 5) + \cdots}{2} = 4.5$$

The formal manipulation rules provide us with a formula that integrates over all conceivable legitimate numerical assignments to the six statements in the state space,

$$P(M_1, \dots, M_6 | N_1, \dots, N_6) = C \times \frac{\prod_{i=1}^n (M_i + N_i)!}{\prod_{i=1}^n M_i!}$$

We break down the calculation into two parts. First, we calculate the non-constant term, the second term, in the formula. Since the M_i and the $M_i + N_i$ are only ever 0 or 1, this term works out to 1.

$$\begin{aligned} \frac{\prod_{i=1}^6 (M_i + N_i)!}{\prod_{i=1}^6 M_i!} &= \frac{1! 0! 0! 1! 1! 0!}{1! 0! 0! 0! 0! 0!} \\ &= 1 \end{aligned}$$

Second, we calculate the constant term C , the first term in the formula.

$$\begin{aligned}
 C &= \frac{M! (N + n - 1)!}{(\prod_{i=1}^6 N_i!) (M + N + n - 1)!} \\
 &= \frac{1! (2 + 6 - 1)!}{0! 0! 1! 1! 0! (1 + 2 + 6 - 1)!} \\
 &= \frac{7!}{8!} \\
 &= \frac{1}{8} \\
 &= 0.125
 \end{aligned}$$

Thus, the answer provided by the formal manipulation rules for the probability of seeing a ONE on the next roll of the die given that we have already seen a FOUR and a FIVE is **NOT** Jaynes's MEP solution for the assigned numerical value for seeing a ONE on any trial, $p_1 = 0.05435$, given a particular model where $\langle F \rangle = 4.5$, but rather,

$$P(X_3 = \text{ONE} | \mathcal{D}) = 1/8$$

Exercise 20.9.4: Solve the problem in section 20.4 when only three models are used.

Solution to Exercise 20.9.4

Suppose that we restrict ourselves to three models for an easy numerical example that makes the point. Model \mathcal{M}_1 is Jaynes's model, model \mathcal{M}_2 inserts the information that $\langle F \rangle = 3.5$, in other words, it is the model for the “honest” die, while the third and final model \mathcal{M}_3 inserts the information that $\langle F \rangle = 2.5$. The numerical assignment to the probabilities under this last model, as you might expect, are the mirror image of Jaynes's model assignment.

Thus, the probability for a ONE spot to appear at the third toss under each of these three models is,

$$P(X_3 = x_1 | \mathcal{M}_1) = 0.05435$$

$$P(X_3 = x_1 | \mathcal{M}_2) = 0.16667$$

$$P(X_3 = x_1 | \mathcal{M}_3) = 0.34749$$

The already observed data of the first two tosses only have an impact on the second term $P(\mathcal{M}_k | \mathcal{D})$ in the prediction formula of Equation (20.3). The probability of a

FIVE on the first toss and a FOUR on the second toss under each model is,

$$P(X_1 = x_5, X_2 = x_4 | \mathcal{M}_1) = 0.23977 \times 0.16545 = 0.03967$$

$$P(X_1 = x_5, X_2 = x_4 | \mathcal{M}_2) = 0.16667 \times 0.16667 = 0.02778$$

$$P(X_1 = x_5, X_2 = x_4 | \mathcal{M}_3) = 0.07877 \times 0.11416 = 0.00899$$

Because,

$$P(\mathcal{M}_k | \mathcal{D}) = \frac{P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k)}{\sum_{k=1}^3 P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k)}$$

we can compute the posterior probability for each of the three models as,

$$P(\mathcal{M}_1 | \mathcal{D}) \approx 0.52$$

$$P(\mathcal{M}_2 | \mathcal{D}) \approx 0.36$$

$$P(\mathcal{M}_3 | \mathcal{D}) \approx 0.12$$

with the correct probability for ONE on the third toss calculated as,

$$\begin{aligned} P(X_3 = x_1 | X_1 = x_5, X_2 = x_4) &\approx \sum_{k=1}^3 P(X_3 = x_1 | \mathcal{M}_k) P(\mathcal{M}_k | \mathcal{D}) \\ &\approx (0.05435 \times 0.52) + (0.16667 \times 0.36) + (0.34749 \times 0.12) \\ &\approx 0.13 \end{aligned}$$

At the outset, we are in a state of total ignorance about which of the three models is correct. We have not seen the die, nor have we observed it being rolled. Therefore, in such a state of total ignorance, we must assign equal probabilities,

$$P(\mathcal{M}_1) = P(\mathcal{M}_2) = P(\mathcal{M}_3) = 1/3$$

To do otherwise would be to admit that, in fact, we do know something about the relative standing of the three models. The complete argument says that in a state of total ignorance we must use the completely uninformed prior probability distribution for the models in model space. Every single conceivable numerical assignment to the probability for the six statements in the state space is made by these models. No one of them is considered better or worse than any other.

The correct probability for seeing a ONE on the third roll after having seen a FIVE and a FOUR on the first two rolls is not $p_1 = 0.05435$, but rather, as we calculated in the last exercise,

$$P(X_3 = \text{ONE} | \mathcal{D}) = 0.125$$

This probability was closely approximated by 0.13 using just three models. We arrived at this close approximation to the correct answer that averages over *all* models by averaging over only *three* symmetrically chosen models.

Intuitively, this result makes perfect sense. The actual occurrence of a FIVE and a FOUR obviously favors model \mathcal{M}_1 to some extent, less so for the fair die, while the third model, which assigns low probabilities to high spots, must be the least favored of all. However, these data can not exclude all of the other competing models. They do not leave \mathcal{M}_1 as the sole survivor.

Bayes's Theorem tells us quantitatively to what extent the three models are reordered. Jaynes's model is better than the honest die model, and about four times better than the mirror image assignment in model \mathcal{M}_3 . The two observations, that is, the data, reordered our initial ignorance about the models that all three models were on the same footing.

However, when the probabilities for a ONE on the third toss are weighted by the posterior probabilities for the models, this reweighting brings down the ONE to about 0.13, not 0.05435. The numerical assignment of 0.05435 under Jaynes's model was weighted properly by 0.52, the numerical assignment of 1/6 under the fair model was weighted properly by 0.36, and, finally, the numerical assignment of 0.34749 under the third model was also weighted properly by 0.12.

Even more dramatically, in order to solve the inferential problem of the $(N+1)^{\text{st}}$ roll conditioned on knowing what happened on the previous N rolls, we must have first assigned numerical values to the probability for a face under some model as the first term in Equation (20.3) clearly demands,

$$P(X_{N+1} = x_i \mid \mathcal{M}_k)$$

and, of course, this is where the MEP must enter the picture.

As indicated by the notation, such an assignment is not conditioned on any previous data, but rather on the information resident in some model \mathcal{M}_k . This information can only be a constraint function average, and not the average of any data.

Exercise 20.9.5: Consider the case shown in Table 20.1 where the data consisted of four throws of the die. Demonstrate how widely varying the probability for the next throw can be for different data even when the sample average is 4.5 for the different data.

Solution to Exercise 20.9.5

Table 20.1 listed all of the possible cases where the sample average of the data was 4.5. What is the probability of a FOUR on the next throw given some of the different data that might have happened? Pick out four data sets where no FOURS

through three FOURS was part of the data. Let these four data sets be labeled as,

$$\begin{aligned}\mathcal{D}_1 &= \text{ SIX SIX FIVE ONE} \\ \mathcal{D}_2 &= \text{ SIX FIVE FOUR THREE} \\ \mathcal{D}_3 &= \text{ FIVE FIVE FOUR FOUR} \\ \mathcal{D}_4 &= \text{ SIX FOUR FOUR FOUR}\end{aligned}$$

The probability of a FOUR on the next throw can range from 1/10 through 4/10 depending on the particular data set even though the sample average for all four data sets was 4.5.

$$\begin{aligned}P(X_5 = \text{FOUR} | \mathcal{D}_1) &= \frac{1}{10} \\ P(X_5 = \text{FOUR} | \mathcal{D}_2) &= \frac{2}{10} \\ P(X_5 = \text{FOUR} | \mathcal{D}_3) &= \frac{3}{10} \\ P(X_5 = \text{FOUR} | \mathcal{D}_4) &= \frac{4}{10}\end{aligned}$$

The *data* are ontological in nature; they ARE NOT FORCED by anything called maximum multiplicity to fall into the pattern of data represented by \mathcal{D}_2 .

On the other hand, constraint functions and their averages CAN BE FORCED to follow maximum information entropy because *probabilities* are epistemological. Thus, Q_4 is equal to 0.16545 once and for all under a model that imposes the information that $\langle F \rangle = 4.5$. This assignment does not and can not change due to the vicissitudes of the data. What will change due to the vicissitudes of the data is $P(\mathcal{M}_k | \mathcal{D})$. The degree of belief in the truth of the model that inserted the information $\langle F \rangle = 4.5$ will change, not the assignment to the statement given the truth of the model.

Exercise 20.9.6: Analyze the data from four successive throws of one die, or four dice thrown at the same time, from the combinatorial perspective of the sample space.

Solution to Exercise 20.9.6

This exercise is a review of Feller's definition of the sample space. Since $n = 6$ and $N = 4$, the number of elementary points in the sample space is $n^N = 6^4 = 1296$.

Use Feller's abstract description involving N balls, labeled **a**, **b**, **c**, **d**, placed into n cells. Figure 20.1 indicates that the first throw, **a**, was a FOUR; the second

throw, **b**, was a THREE; the third throw, **c**, was a ONE; and the fourth throw, **d**, was a SIX. Or, the red die showed a FOUR, the green die a THREE, the yellow die a ONE, and the blue die a SIX if four differently colored dice were thrown simultaneously.

There are $n = 6$ "cells" and $N = 4$ "balls"					
ONE	TWO	THREE	FOUR	FIVE	SIX
C	*	b	a	*	d
3rd trial yellow die		2nd trial green die	1st trial red die		4th trial blue die

Figure 20.1: Feller's abstract representation for an elementary point in the sample space as it applies to data for the dice scenario.

This is a description of one of the 1296 elementary points in the sample space. It is called a simple event. This is the most detail we can provide about an event in the sample space.

The 1296 elementary points fall into five classes defined by the break down of the possible sums for $N = 4$.

Class 1. $4 + 0 + 0 + 0 + 0 + 0$

Class 2. $3 + 1 + 0 + 0 + 0 + 0$

Class 3. $2 + 2 + 0 + 0 + 0 + 0$

Class 4. $2 + 1 + 1 + 0 + 0 + 0$

Class 5. $1 + 1 + 1 + 1 + 0 + 0$

There must be a total of,

$$u = \frac{(N+n-1)!}{N! (n-1)!} = \frac{(4+6-1)!}{4! 5!} = 126$$

possible frequency counts, or contingency tables, for data consisting of four throws of one die, or four dice thrown together.

The number of different frequency counts falling into each of the above five classes must equal 126. There are 6 different contingency tables in class 1, 30 in class 2, 15 in class 3, 60 in class 4, and 15 in class 5, for the required total of,

$$6 + 30 + 15 + 60 + 15 = 126$$

These numbers are calculated by the combinatorial formula,

$$\frac{n!}{r_z! r_s! \cdots r_N!}$$

where r_z stands for the repetitions of zero counts, r_s for the repetitions of single counts, and so on. For example, the number of different contingency tables in class 4 is calculated as,

$$\frac{n!}{r_z! r_s! \cdots r_N!} = \frac{6!}{3! 2! 1! 0! 0!} = 60$$

Furthermore, each one of the 126 different frequency counts can happen in a number of different ways depending on the multiplicity factor. That is, we might have data looking like this contingency table $\boxed{1} \boxed{0} \boxed{1} \boxed{0} \boxed{0} \boxed{2}$ belonging to one of the 60 different varieties of class 4. These data can happen in,

$$W(N) = \frac{N!}{N_1! N_2! N_3! N_4! N_5! N_6!} = \frac{4!}{1! 0! 1! 0! 0! 2!} = 12$$

12 different ways when we take account of the individual distinguishability of either the trial number of one die thrown at different times, or the color of the die if four dice are thrown at the same time.

When we total up the number of different ways due to distinguishability for each different way that the frequency counts might occur within each class, we must have accounted for every single elementary point in the sample space.

$$n^N = \sum_{j=1}^{u=126} W_j(N) = (6 \times 1) + (30 \times 4) + (15 \times 6) + (60 \times 12) + (15 \times 24) = 1296$$

The elementary point shown in Figure 20.1 was one of the 24 possible configurations with one ball in one cell, that is, the detailed accounting of either the trial number or the die color for each observation. It was also one of the 15 frequency counts consisting of four different die faces in the four throws of one die, or one throw of four dice. If just one model, say, the “fair” model, is adopted, then the probability for the data that four different faces appear is 360/1296. However, if the IP is completely uninformed about the causal nature of this whole dice throwing business, the probability for these data is less probable at 15/126.

Exercise 20.9.7: Show the detailed calculation involved in the probability assignment for seeing a FOUR after a large number of previous rolls of a die.

Solution to Exercise 20.9.7

In section 20.6, the probability for seeing a FOUR after 600 rolls of the die was said to be approximately 1/2. The general formula for the probability of any number of future frequency counts given some past data was derived in Volume I. It has already been used extensively to date in this Volume. We will use that formula once

again to find the required probability.

$$P(M_1, M_2, \dots, M_n | N_1, N_2, \dots, N_n) = C \times \frac{\prod_{i=1}^n (M_i + N_i)!}{\prod_{i=1}^n M_i!}$$

The dimension of the state space is $n = 6$. We want to find the probability for the next roll of the die, so $M = 1$. We want to find the probability for a FOUR, so the future frequency counts are,

$$M_1 = 0$$

$$M_2 = 0$$

$$M_3 = 0$$

$$M_4 = 1$$

$$M_5 = 0$$

$$M_6 = 0$$

where $\sum_{i=1}^n M_i = M$. The die was rolled 600 times in the past, so $N = 600$. The data, that is, the actually observed frequency counts in these 600 rolls of the die, were,

$$N_1 = 0$$

$$N_2 = 0$$

$$N_3 = 0$$

$$N_4 = 300$$

$$N_5 = 300$$

$$N_6 = 0$$

where $\sum_{i=1}^n N_i = N$. In addition, we wanted these data to have a sample average of $\bar{F} = 4.5$ spots.

$$\bar{F} = \frac{\sum_{i=1}^6 N_i F(X = x_i)}{N} = \frac{(300 \times 4) + (300 \times 5)}{600} = 4.5$$

We break down the calculation into two parts. First, we calculate the non-constant term, the second term, in the formula. Second, we calculate the constant

term C , the first term in the formula.

$$\begin{aligned}
 \frac{\prod_{i=1}^n (M_i + N_i)!}{\prod_{i=1}^n M_i!} &= \frac{0! 0! 0! 301! 300! 0!}{0! 0! 0! 1! 0! 0!} \\
 &= 301! \times 300! \\
 C &= \frac{M! (N + n - 1)!}{(\prod_{i=1}^6 N_i!) (M + N + n - 1)!} \\
 &= \frac{1! (600 + 6 - 1)!}{0! 0! 0! 300! 300! 0! (1 + 600 + 6 - 1)!} \\
 &= \frac{605!}{300! 300! 606!} \\
 &= \frac{1}{300! 300! 606} \\
 C \times \frac{\prod_{i=1}^n (M_i + N_i)!}{\prod_{i=1}^n M_i!} &= \frac{301! \times 300!}{300! \times 300! \times 606} \\
 &= \frac{301}{606} \\
 &= \frac{N_4 + 1}{N + n} \\
 &\approx 0.4967
 \end{aligned}$$

Thus, as claimed,

$$P(X_{601} = x_4 | \mathcal{D}) \approx 1/2$$

It is easy to see that,

$$\frac{N_4 + 1}{N + 6} \rightarrow 1/2 \text{ as } N \rightarrow \infty$$

For example after $N = 6,000,000$ rolls of the die,

$$P(X_{6,000,001} = x_4 | \mathcal{D}) = \frac{3,000,001}{6,000,006} = 0.499999667$$

Exercise 20.9.8: What is the probability for seeing a ONE on the next roll of the die after an enormous amount of data that has a sample average of 4.5?

Solution to Exercise 20.9.8

Following up on the solution in the previous exercise, the probability for seeing a ONE after 6,000,000 rolls have yielded 3,000,000 FOURS and 3,000,000 FIVES is,

$$\begin{aligned} P(\text{ONE} \mid \mathcal{D}) &= \frac{N_1 + 1}{N + n} \\ &= \frac{1}{6,000,006} \\ &= 1.67 \times 10^{-7} \end{aligned}$$

Once again, it is quite clear that such a probability, while adhering strictly to the guidelines laid down by Jaynes, cannot possibly be the MEP assignment for a probability of a ONE, $p_1 = 0.05435$, under a model where $\langle F \rangle = 4.5$.

Exercise 20.9.9: When Jaynes solved the Brandeis die problem, he mentioned one of the most important consequences of the MEP. This is the relationship involving the partition function, the Lagrange multiplier, and the constraint function average. Refresh your memory concerning some basic definitions for derivatives.

Solution to Exercise 20.9.9

Jaynes points out that, in general, the MEP formalism permits one to establish an interesting relationship involving the partition function, the Lagrange multipliers, and the associated constraint function average,²

$$\frac{\partial \ln Z}{\partial \lambda_j} = \langle F_j \rangle$$

We will look at several numerical examples of this important feature of the MEP at various points throughout this Volume.

To begin, go back to the basics and look at one of the definitions for a derivative. The central difference numerical approximation for the first derivative in a generic notation is,

$$f'(a) \approx \frac{f(a+h) - f(a-h)}{2h}$$

Let the log of the partition function $\ln Z(\lambda)$ be this function $f(a)$ with one argument $a = \lambda$. Since we would like to examine this function as λ varies, let h

²Remember that my λ is the negative of Jaynes's λ .

stand for small changes in λ . Thus, we set up the numerical approximation for the partial differentiation as,

$$\frac{\partial \ln Z}{\partial \lambda} \approx \frac{f(a+h) - f(a-h)}{2h}$$

To see how close the approximation is to the actual constraint function average specified as the information in the model, let the increment h be the small value $h = 0.001$. Now, $f(a+h) \equiv \ln Z(\lambda + 0.001)$. We found out earlier in section 20.3 that $\lambda = 0.37105$. The log of this slightly altered partition function is,

$$\begin{aligned} Z(\lambda + 0.001) &= \sum_{i=1}^6 e^{(\lambda+0.001) \times F(X=x_i)} \\ &= e^{.37205 \times 1} + e^{.37205 \times 2} + e^{.37205 \times 3} + \cdots + e^{.37205 \times 6} \\ \ln Z(\lambda + 0.001) &= 3.28781 \end{aligned}$$

and likewise,

$$\begin{aligned} Z(\lambda - 0.001) &= \sum_{i=1}^6 e^{(\lambda-0.001) \times F(X=x_i)} \\ &= e^{.37005 \times 1} + e^{.37005 \times 2} + e^{.37005 \times 3} + \cdots + e^{.37005 \times 6} \\ \ln Z(\lambda - 0.001) &= 3.27881 \end{aligned}$$

with the result that,

$$\frac{\partial \ln Z}{\partial \lambda} \approx \frac{3.28781 - 3.27881}{0.002} = \frac{0.009}{0.002} = 4.5$$

Since the information in Jaynes's model specified that the mathematical expectation of the constraint function was $\langle F \rangle = 4.5$, we do have a numerical confirmation in this case that,

$$\frac{\partial \ln Z}{\partial \lambda} = \langle F \rangle$$

Exercise 20.9.10: Show all of the possibilities for satisfying the sample average constraint of 4.5 in five tosses of the die.

Solution to Exercise 20.9.10

There are no possibilities of forming a sum over frequency counts multiplied by the spots which add up to 22.5,

$$\frac{\sum_{i=1}^6 N_i F(X=x_i)}{5} = 4.5$$

Exercise 20.9.11: Discuss the confusion centering around the definition of the state space when two different information based MEP assignments are made in the coin toss scenario.

Solution to Exercise 20.9.11

Let's simplify the computational effort by looking at $M = 4$ future coin tosses. For this coin tossing scenario, the MEP assignment is always to a state space of dimension $n = 2$.

Under the fair model, the information inserted is that $\langle F \rangle = 1.5$, or, equivalently, that $\lambda = 0$. The constraint function vector is $F(X = x_i) = (1, 2)$. The partition function is $Z = 2$. The numerical assignment to the probability for either HEADS or TAILS is $1/2$. The information entropy of this particular assignment is,

$$H_{\max}(Q_i) = \ln 2 = 0.693147$$

Suppose that the future event of $M_1 = 2$ HEADS and $M_2 = 2$ TAILS is of interest. The probability conditioned on the fair model for seeing two HEADS and two TAILS in four future coin flips is $3/8$. The probability for any eventuality in four future coin tosses can be calculated just as easily.

There is no need to call on the MEP for any of these later calculations. The MEP has performed its designated role by assigning legitimate numerical values to the probabilities to the two statements in the state space. Its services are no longer required.

But, someone may protest, we are nonetheless dealing with a probability distribution $P(M_1, M_2)$ over possible events. Couldn't it too be said to have an information entropy? It does. There are five possible frequency counts,

$$\text{Number of contingency tables} = \frac{(M+n-1)!}{M! (n-1)!} = \frac{(4+2-1)!}{4! 1!} = 5$$

These five events are (1) no HEADS, four TAILS, (2) one HEADS, three TAILS, ... (5) four HEADS, no TAILS. We write this as,

$$(1) M_1 = 0, M_2 = 4, (2) M_1 = 1, M_2 = 3, \dots, (5) M_1 = 4, M_2 = 0$$

In other words, there are now five statements in a *new* state space with dimension $n = 5$. This new state space has an information entropy of,

$$\begin{aligned} H^*(p_i) &= - \sum_{i=1}^5 p_i \ln p_i \\ &= - [(1/16 \ln 1/16) + (4/16 \ln 4/16) + (6/16 \ln 6/16) + \\ &\quad (4/16 \ln 4/16) + (1/16 \ln 1/16)] \\ &= 1.40753 \end{aligned}$$

To reproduce this as if it were an MEP assignment, set up the constraint function vector as $F^*(X = x_i) = (0, 1, 1.29248, 1, 0)$ with the information under the model specifying $\langle F \rangle^* = 0.9847$ with the accompanying Lagrange multiplier of $\lambda^* = \ln 4 = 1.38629$. The partition function is $Z^* = 16$. For example,

$$\begin{aligned}
 p_3 &\equiv P(M_1 = 2, M_2 = 2) \\
 P(M_1 = 2, M_2 = 2) &= \frac{\exp [\lambda^* F^*(X = x_3)]}{Z^*(\lambda)} \\
 Z^* &= \exp [1.38629 \times 0] + \exp [1.38629 \times 1] + \\
 &\quad \exp [1.38629 \times 1.29248] + \exp [1.38629 \times 1] + [1.38629 \times 0] \\
 &= 1 + 4 + 6 + 4 + 1 \\
 &= 16 \\
 p_3 &= \frac{\exp [1.38629 \times 1.29248]}{16} \\
 &= \frac{3}{8}
 \end{aligned}$$

Calculate the constraint function average that was the information inserted into the distribution as,

$$\sum_{i=1}^5 F^*(X = x_i) p_i = (0 \times 1/16) + (1 \times 4/16) + (1.29248 \times 6/16) + (1 \times 4/16) + (0 \times 1/16) = 0.9847$$

Compare this approach to the “entropic-like” formula in the coin toss scenario which has the original $\lambda = 0$ and $Z = 2$,

$$\begin{aligned}
 \ln P(M_1 = 2, M_2 = 2) &= M \times [\ln W(M)/M + \lambda \bar{F} - \ln Z] \\
 &= \ln W(M) - M \ln Z \\
 &= \ln 6 - 4 \ln 2 \\
 &= -0.980829 \\
 \exp [-0.980829] &= 0.375 \\
 P(M_1 = 2, M_2 = 2) &= 0.375
 \end{aligned}$$

This illustrates that the new partition function Z_{new} for the new p_i had to be $M \ln Z_{old} = 4 \ln 2 = 16$ involving the original partition function Z from the original MEP assignment in the $n = 2$ state space.

Repeating the obvious, the MEP is simply not required to find any numerical assignment to the five statements about frequency counts in the new state space. But as was illustrated in this exercise, it can be done because we are nonetheless still considering an assignment for a probability.

The spaces are, however, completely different. The correct space at which we apply the MEP is the state space of dimension $n = 2$. It is neither the data space, nor the future frequency data space of dimension $M = 5$.

Once the MEP has been applied at the appropriate level of the two statements in the coin toss problem, all further probabilities are *calculated*, not assigned by any new application of the MEP.

Exercise 20.9.12: Jaynes says that the data average of 4.5 *is all you know*. How would a simple semantic clarification have eliminated all the ensuing confusion?

Solution to Exercise 20.9.12

I imagine that a rebuttal might be made that I have misconstrued what Jaynes was trying to do. The data average is all the IP knows, he said. He didn't say that any actual data points N_1, N_2, \dots, N_n were known.

But this is too vague to be of any use as I tried to explain with my examples of different data sets all adhering to the prescription of the same data average. In addition, you don't know whether N is "large enough," or to what extent you can eliminate any competing models.

And that is the real crux of the matter. The whole impetus seems to be finding right at the outset one model based simply on a data average. This is just too ill-defined to be of any use, or better yet, is correctly subsumed under the MEP and maximum likelihood. (See Chapter Twenty Three ahead for a logistic regression example.)

It would have been better, and would have exposed the flaw in not proposing the full panoply of models, for Jaynes to have said that, "If all we know is a data average of 4.5, then use the MEP to find the numerical assignments under one model that inserts the information that a constraint function average is 4.5, *and then use that as the only model under consideration for all further inferences.*" And rather than saying that the probability was being found for the *next* toss, he would have to say that it was being found for *any* toss.

Of course, the IP could never change the probabilities assigned to the six faces of the die under this one definitive model no matter how much more data it might accumulate after initially setting it up based on those unknown N_i . Otherwise, the IP would be forced into the rather uncomfortable position of trying to justify some “subjective” weight to attach to a model about some vague data average.

All of this difficulty is completely circumvented if the IP just follows the rules of the game. Let all possible numerical assignments be present at the beginning before any data, and then let any subsequent data pare down the model space. We don’t have any rules about how to process a data average into any of these initial assignments; we only have rules about constraint function averages.

Exercise 20.9.13: Consider some sort of weighted average for a future event based on the known data average.

Solution to Exercise 20.9.13

If the actual frequencies N_1, N_2, \dots, N_n were not known as the data, but only the data average, then perhaps an average based on the multiplicity factors for all acceptable frequencies could be calculated. For example, refer back to the eight possible frequencies in Table 20.1, satisfying the data average of 4.5 from $N = 4$ data points. Take the prediction for the next roll of the die under each data set and weight it according to its multiplicity factor,

$$\begin{aligned} P(X_5 = x_4 \mid \mathcal{D}) &= \sum_{l=1}^8 P(X_5 = x_4 \mid \mathcal{D}_l) \times w_l \\ &= \frac{1}{80} \times [(1/10 \times 12) + (2/10 \times 12) + (1/10 \times 6) + (1/10 \times 12) + \\ &\quad (2/10 \times 24) + (4/10 \times 4) + (1/10 \times 4) + (3/10 \times 6)] \\ &= 0.175 \end{aligned}$$

But this remains an *ad hoc* procedure, with no justification from the formal rules. It is an average over potential data.

The main sticking point however, once again, is that reality in the form of the data does not have to conform to some combinatorial principle like, “the data must have sorted itself out in the maximum number of ways it could have happened.” It’s just possible that the die was constructed such that it had three FOUR faces and three FIVE faces.

Chapter 21

The MEP and the Kangaroos

21.1 Introduction

After this initial excursion into coin tossing and dice rolling, let's revisit our old friends from Volume I, the beer-drinking kangaroos. Now, though, we can look at this scenario afresh from our new MEP perspective. After the somewhat negative tone of the last Chapter, we are now better aware of what we can, and cannot, expect from the MEP formalism.

One of the more fascinating aspects of assigning numerical values through the auspices of the MEP is the role of marginal probabilities. Suppose that the IP has constructed joint statements to define an interesting state space for some inferential problem. Then, the marginal probabilities are automatically right there at the forefront. It would be a shame if they were not to play a central role for models. Moreover, marginal probabilities are easily seen to be constraint function averages, and thus may clearly assume the role of information.

More specifically, and as promised in Volume I, what we want to begin here in this Chapter is a first look at defining independence and correlations. The ever patient kangaroos will serve as our working example to illustrate the salient points.

Casting back, we recall that the kangaroo state space referred to joint statements about hand and beer preference. By restricting ourselves to two possible observations for both hand preference and beer preference, we kept to our objective of a small state space, here obviously of dimension $n = 4$.

By allowing the state space to reference joint statements, we give ourselves the liberty of exploring the notion of predictor, or explanatory variables. Thus, observing a kangaroo's hand preference might permit us to modify our state of knowledge about its beer preference. We are trying to bring Bayes's Theorem back into the fray by talking about the effect of explanatory variables.

Up till now, whether tossing coins or rolling dice, no other explanatory statements were included within the state space to help us make better inferences. Everything revolved around setting up models involving only the appearance of HEADS or TAILS, or ONEs through SIXes. Anything that might have helped us to explain why a HEADS or a THREE was measured or observed was simply ignored. Jaynes finessed this issue with the loaded dice by forming clever constraint functions.

If we are going to develop the idea of explaining things with predictor variables, we must do so from an informational perspective. This means that the models proposed by the IP must include constraint functions, as well as their averages, involving the joint statements in the state space.

We are leading up to the notion of an association between statements in the state space. More technically, we are going to introduce the ideas of independence and correlation. Then, we will explore how this notion can be implemented within the MEP algorithm.

21.2 Contingency Tables

Repeating the mantra from Volume I, it is very important to distinguish between joint probability tables and contingency tables. Contingency tables contain the actually observed counts of kangaroos when categorized according to four traits.

Contingency tables therefore are a way of displaying any already observed data, or any potential future data. We are using contingency tables here in this latter sense, not to show any already existing data, but rather to show occupancy counts that might occur in future observations.

Joint probability tables, on the other hand, are used to display the numerical values assigned to the probabilities for the four joint statements. Joint probability tables will eventually be discussed when we get around to some examples of the numerical assignments as made by several interesting models.

To refresh our memories, we are supposing that $M = 16$ kangaroos will be observed sometime in the future. Each kangaroo will be placed into one of the four cells of the contingency table according to its beer and hand preference.

The total number of possible contingency tables, or stated equivalently, the total number of possible frequency counts for the sixteen kangaroos, is given by the combinatorial formula used extensively in Volume I,

$$\frac{(M + n - 1)!}{M! (n - 1)!} = 969$$

Thus, there will be 969 different contingency tables, no more and no less, for our particular numerical example involving $n = 4$ and $M = 16$. One of these possible 969 contingency tables is shown at the top of the next page as Figure 21.1.

	B	\bar{B}	
H	9 Cell 1	3 Cell 2	12
\bar{H}	3 Cell 3	1 Cell 4	4
	12	4	16

Figure 21.1: An example of a contingency table where $M = 16$ kangaroos will be categorized according to beer and hand preference. BH refers to a right handed Foster's drinker, $\bar{B}H$ to a right handed Corona drinker, $B\bar{H}$ to a left handed Foster's drinker, while $\bar{B}\bar{H}$ refers to a left handed Corona drinker.

This contingency table shows the 16 kangaroos as they might be categorized, sometime in the future, according to beer and hand preference. There are 9 BH kangaroos, 3 $\bar{B}H$ kangaroos, 3 $B\bar{H}$ kangaroos, and 1 $\bar{B}\bar{H}$ kangaroo. The marginal totals for hand preference, beer preference, and the overall number of kangaroos are part of the contingency table. As before, within the text, it is easier to refer to contingency tables in the following condensed format as $[9 \boxed{3} \boxed{3} \boxed{1}]$.

21.3 The MEP Algorithm and Different Models

In this section, we detail how models that the IP might specify underlie the numerical assignments to the probabilities for all four statements in the kangaroo state space. To accomplish this task, we employ Jaynes's Maximum Entropy Principle (MEP). This technique is merely one way of going about the task of making numerical assignments. We cannot claim that the MEP algorithm has any special status in this regard.

Having said that, anyone who has used the MEP algorithm comes to accept it as an unsurpassed tool for incorporating information into a probability distribution. At the same time, we invoke the insurance policy preventing unwanted information inadvertently sneaking into the distribution. It is an elegant, disciplined, and rather beautiful solution to the problem of assigning numerical values to probabilities when an IP's state of knowledge must be constrained by the information resident in some model.

We have now arrived at that juncture where it is necessary to explain how the probability for any future frequency counts, namely $P(M_1, M_2, M_3, M_4)$, gets computed if, in fact, the IP wants to examine just one particular model.

In the past, we dealt with this issue by integrating over all the assignments made by every conceivable model,

$$P(M_1, M_2, M_3, M_4) = \int \cdot \int_{\sum q_i=1} W(M) q_1^{M_1} q_2^{M_2} q_3^{M_3} q_4^{M_4} P(q_1, q_2, q_3, q_4) dq_i$$

Thus, we didn't have to wonder about how any one model actually dictated a numerical assignment. It was simply enough to know that a legitimate assignment had been carried out by some means or another.

When just one particular model is being considered, the probability over the models collapses to a Dirac delta function. The probability for any future frequency counts then simplifies to the multinomial distribution,

$$P(M_1, M_2, M_3, M_4) = W(M) Q_1^{M_1} Q_2^{M_2} Q_3^{M_3} Q_4^{M_4}$$

where the Q_i are the numerical assignments under that one model.

21.3.1 Solving with two constraints

In the original kangaroo problem as presented in the literature, two pieces of information were inserted into the probability distribution. Under such a model \mathcal{M}_k , there will be two constraint functions, together with their associated averages. The two parameters are, as usual, represented by two Lagrange multipliers.

Set $m = 2$ so that the MEP formula looks like,

$$P(X = x_i | \mathcal{M}_k) \equiv Q_i = \frac{\exp [\lambda_1 F_1(X = x_i) + \lambda_2 F_2(X = x_i)]}{Z(\lambda_1, \lambda_2)} \quad (21.1)$$

with the partition function,

$$Z(\lambda_1, \lambda_2) = \sum_{i=1}^4 \exp [\lambda_1 F_1(X = x_i) + \lambda_2 F_2(X = x_i)] \quad (21.2)$$

In order to suppress the undue proliferation of confusing notation, we keep one statement label X , and allow it to take on values x_1 through x_4 to stand for BH , \overline{BH} , $B\overline{H}$, and $\overline{B}\overline{H}$. Otherwise, we would have to adopt a more cumbersome notation like $(A = a_1, B = b_2)$ to refer to, for example, the statement $B\overline{H}$.

We will eventually construct models where $m = 0, 1, 2$, or 3 pieces of information are inserted into the probability distribution. It is easier when considering cases in this generality to go ahead and define three constraint functions right at the start, but with the realization that some or all of the associated Lagrange multipliers may be set to 0 under some models.

Thus, in effect, we will be either excluding or including information according to whether the Lagrange multiplier is 0 or not. For the time being, however, we will solve the problem as described above.

What do the first two constraint functions look like? The information used in the literature was to assert that 3/4 of the kangaroos are right handed and, in addition, that 3/4 of the kangaroos drink Foster's. We see that we can address the goal posed in the **Introduction** of using marginal probabilities.

In fact, the information to be used in the MEP *is* in the form of marginal probabilities. It is interesting and quite helpful that simply setting the constraint functions at appropriate values of 0s and 1s allows the IP to insert information about these marginal probabilities into the joint probability table.

The marginal probability,

$$P(X = x_1 | \mathcal{M}_k) + P(X = x_2 | \mathcal{M}_k)$$

that is, the probability for being right-handed, is easily read straight from the joint probability table as $Q_1 + Q_2$. Likewise, the marginal probability,

$$P(X = x_1 | \mathcal{M}_k) + P(X = x_3 | \mathcal{M}_k)$$

that is, the probability for being a Foster's drinker, is just as easily read from the joint probability table as $Q_1 + Q_3$.

Thus, if the first constraint function is set up as $F_1(X = x_i) = (1, 1, 0, 0)$ then,

$$\begin{aligned} \langle F_1 \rangle &= \sum_{i=1}^4 F_1(X = x_i) Q_i \\ &= (1 \times Q_1) + (1 \times Q_2) + (0 \times Q_3) + (0 \times Q_4) \\ &= Q_1 + Q_2 \end{aligned}$$

If the second constraint function is set up as $F_2(X = x_i) = (1, 0, 1, 0)$ then,

$$\begin{aligned} \langle F_2 \rangle &= \sum_{i=1}^4 F_2(X = x_i) Q_i \\ &= (1 \times Q_1) + (0 \times Q_2) + (1 \times Q_3) + (0 \times Q_4) \\ &= Q_1 + Q_3 \end{aligned}$$

The information, then, is that the two parameters, that is, the two constraint function averages, assume the values of $\langle F_1 \rangle = 3/4$ and $\langle F_2 \rangle = 3/4$. This identifies one specific model \mathcal{M}_k , or alternatively, the dual parameters λ_1 and λ_2 , which will make the numerical assignments to the generic q_1 through q_4 as the definite values Q_1 through Q_4 given the information in this model.

It is now an easy matter to bring in the MEP algorithm in the form of Equations (21.1) and (21.2) to find the numerical assignment for all four cells in the joint probability table. These particular numerical assignments emanate from a model \mathcal{M}_k that wants to insert the information that $\langle F_1 \rangle = 3/4$ and $\langle F_2 \rangle = 3/4$.

For example, the numerical assignment to the first cell,

$$Q_1 \equiv P(X = x_1 | \mathcal{M}_k)$$

works out to,

$$\begin{aligned} Q_1 &= \frac{\exp [\lambda_1 F_1(X = x_1) + \lambda_2 F_2(X = x_1)]}{Z(\lambda_1, \lambda_2)} \\ &= \frac{\exp [(\lambda_1 \times 1) + (\lambda_2 \times 1)]}{Z(\lambda_1, \lambda_2)} \\ &= \frac{\exp [\lambda_1 + \lambda_2]}{Z(\lambda_1, \lambda_2)} \end{aligned}$$

It turns out that in order to satisfy the constraints, $\lambda_1 = \lambda_2 = 1.098612$. The numerator for Q_1 is then calculated as,

$$\text{Numerator of } Q_1 = \exp [\lambda_1 + \lambda_2] = \exp [1.098612 + 1.098612] = 9$$

The partition function, $Z(\lambda_1, \lambda_2)$, is the sum over all four numerators calculated in the same way. It is equal to 16. The numerical value assigned to the probability for the joint statement in cell 1 of the joint probability table by model \mathcal{M}_k is 9/16.

The calculations for all four numerical assignments Q_i under this model are summarized in Table 21.1 at the top of the next page. The bottom row presents the relevant sums. The first sum shown is the partition function.

$$Z(\lambda_1, \lambda_2) = \sum_{i=1}^4 \exp [\lambda_1 F_1(X = x_i) + \lambda_2 F_2(X = x_i)] = 16$$

The sum of all four assignments must equal 1, and, indeed, $\sum_{i=1}^4 Q_i = 1$. We have referred to this as the universal constraint, and indicated it in the same manner as any other constraint, $F_0(X = x_i) = (1, 1, 1, 1)$, and $\langle F_0 \rangle = 1$. The individual components making up the two constraint averages are shown explicitly in the last two columns. Their sums form the required expectations of $\langle F_1 \rangle = 3/4$ and $\langle F_2 \rangle = 3/4$.

Table 21.1: *The MEP calculations for a specific model inserting information about two marginal probabilities. The Lagrange multipliers for this model are found to have the values $\lambda_1 = \lambda_2 = 1.098612$.*

Cell i	Numerator	Q_i	$F_1(x_i)$	$F_2(x_i)$	$\langle F_1 \rangle$	$\langle F_2 \rangle$
1	$\exp[\lambda_1 + \lambda_2] = 9$	0.5625	1	1	0.5625	0.5625
2	$\exp[\lambda_1] = 3$	0.1875	1	0	0.1875	0.0000
3	$\exp[\lambda_2] = 3$	0.1875	0	1	0.0000	0.1875
4	$\exp[0] = 1$	0.0625	0	0	0.0000	0.0000
Sums	16	1.0000			0.7500	0.7500

21.3.2 Conceptual point

I will continue to point out the following fact *ad nauseam* to the reader because it seems to be the source of the major conceptual error in using the MEP. **Nowhere in this whole process of assigning numerical values was it necessary to refer to any observed data!**

The IP is free to insert whatever information it wants in the guise of the constraint functions and their averages. The algorithm assigns the numerical values by relying solely on these two ingredients, followed by the determination of the Lagrange multipliers. Any observed data, and there aren't any at this point, are completely irrelevant to the algorithm.

Where the observed data do play a vital role is in updating the state of knowledge about the relative status of the models after conditioning on the known data. What we have developed so far using the MEP is $P(X = x_i | \mathcal{M}_k)$. Obviously, any data \mathcal{D} do not appear in this expression. The only place where \mathcal{D} will appear is in the expression $P(\mathcal{M}_k | \mathcal{D})$.

21.3.3 More models and different information

These numerical assignments, as just calculated by the MEP algorithm, are placed into the joint probability table of Figure 21.2. These particular values ensue from the information inserted by one specific model. Let's relabel this model as \mathcal{M}_3 with a view to its eventual order within all of the models discussed.

Consider a few other models where, as we mentioned earlier, m takes on the values from 0 through 3. The case where $m = 0$ is easy to calculate. This is the situation where none of the three constraints plays a role in the assignment since all three Lagrange multipliers will be set to 0. In a sense, the IP is inserting the minimum amount of information because none of the constraint functions are active.

	B	\bar{B}	
H	$P(X=x_1)$ 9/16 Q_1	$P(X=x_2)$ 3/16 Q_2	3/4
\bar{H}	$P(X=x_3)$ 3/16 Q_3	$P(X=x_4)$ 1/16 Q_4	1/4
	3/4	1/4	1

Figure 21.2: A joint probability table for the kangaroo scenario under a specific model assigning numerical values to the probability of all four cells.

The numerator for each cell will have the same value of $\exp[0] = 1$. The denominator, the partition function, is the sum over all four of these values, or,

$$Z(\lambda_1, \lambda_2, \lambda_3) = \sum_{i=1}^4 \exp [\lambda_1 F_1(x_i) + \lambda_2 F_2(x_i) + \lambda_3 F_3(x_i)] = 4$$

Each q_i then has a numerical assignment of $Q_i = 1/4$ under this model. Let's call this model \mathcal{M}_1 .

If we were operating under the same rubric of model \mathcal{M}_3 , we might have specified the two constraint function averages as $\langle F_1 \rangle = 0.5$ and $\langle F_2 \rangle = 0.5$. One of the neat features of the MEP algorithm is that it processes this redundant information without protest, returning the answer that λ_1 and λ_2 are both equal to 0.

This model, and its associated numerical assignment of $1/n$ as a probability for each statement in the state space of dimension n , is sometimes erroneously referred to as *the* maximum entropy distribution. This is obviously a misconception, of course. It is merely the numerical assignment of maximum entropy when the IP wishes to insert no information into the distribution and, therefore, just one of the infinite number of legitimate numerical assignments which might be made under the MEP.

Move up to the case where $m = 1$. This will be a new model \mathcal{M}_2 . Suppose that only one constraint function is defined. This single constraint function and its average permit the marginal probability for being right handed to be inserted as information. That is, the IP wants to observe what the ramifications are when the information that a kangaroo is right handed has probability $3/4$ is inserted into a state of knowledge. This is the only information included by the model *with no other information whatsoever included*.

That emphasized final phrase may not provide much of an emotional impact, but it is the very heart and soul of the MEP. It ensures the IP of an iron-clad guarantee that no lurking, hidden information is going to sneak into the distribution.

Thus, $F_1(X = x_i) = (1, 1, 0, 0)$, and the constraint average is set to $\langle F_1 \rangle = 3/4$ as before. Now what are the numerical assignments from model \mathcal{M}_2 ? The value of one Lagrange multiplier λ_1 must be discovered that causes the constraint average to be satisfied.

That value is $\lambda_1 = 1.096812$. Thus, both Q_1 and Q_2 have a numerator of $\exp[1.096812] = 3$. Both Q_3 and Q_4 have numerators of $\exp[0] = 1$. Summing over all four cells, the denominator has the value of $Z(\lambda_1) = 8$. The numerical assignments to a probability for the four joint statements making up the state space under this specific model are,

$$Q_1 = 3/8, Q_2 = 3/8, Q_3 = 1/8, \text{ and } Q_4 = 1/8$$

It is easy to verify that the one constraint average as well as the universal constraint are both satisfied.

Maximizing the entropy has the effect of spreading out the available probability as smoothly as possible, while still satisfying all of the available information. This effect is evident in this model because no information was specified about the marginal probabilities for beer preference. Thus, we have the consequence that $Q_1 + Q_3 = Q_2 + Q_4 = 1/2$. Or, in other words, the probability that a kangaroo prefers Foster's is the same as the probability that it prefers Corona.

At the outset, we investigated model \mathcal{M}_3 where $m = 2$, so we will skip directly to the final model \mathcal{M}_4 with $m = 3$. Now the IP will make use of all three constraint functions. But we have only defined the first two constraint functions at this point, so it behooves us to define $F_3(X = x_i)$. Define $F_3(X = x_i) = (1, 0, 0, 0)$ with the constraint average simply as $\langle F_3 \rangle = Q_1$.

As alluded to before, if there are $m = n - 1$ constraints, then entropy is not needed to solve an underdetermined problem. The constraints by themselves are sufficient to determine the numerical assignments as is easy to see in these small problems.

For example, if the information in this latest model \mathcal{M}_4 incorporates the same information about the first two constraints that $\langle F_1 \rangle = \langle F_2 \rangle = 0.75$, and then adds further information that $\langle F_3 \rangle = 0.70$, then it is clear that,

$$Q_1 = 0.70, Q_2 = 0.05, Q_3 = 0.05 \text{ and } Q_4 = 0.20$$

This results from trying to satisfy all three constraints (four, if one wants to count the universal constraint). Also, it is clear that the constraint function averages have to be feasible in the sense that if $\langle F_1 \rangle = 0.75$, then $\langle F_3 \rangle \leq 0.75$.

But this doesn't imply that the MEP algorithm can't be used in exactly the same way as we have done in the past, even though this latest model is not an

underdetermined problem. Everything proceeds in exactly the same manner with the numerical optimization routine finding the values for three Lagrange multipliers. They are found to have the values of,

$$\lambda_1 = -1.38629$$

$$\lambda_2 = -1.38629$$

$$\lambda_3 = +4.02535$$

The numerator of Q_1 is,

$$\exp [\lambda_1 + \lambda_2 + \lambda_3] = 3.5$$

and the calculated value of the partition function is,

$$Z(\lambda_1, \lambda_2, \lambda_3) = 5$$

so we see that the MEP does return the correct numerical assignment of $Q_1 = 0.70$ for the probability of the joint statement,

$$P(X = x_1 | \mathcal{M}_4) \equiv P(\text{"A kangaroo is right handed and prefers Foster's."})$$

under this final model.

21.3.4 Correlations

This final model is important because it allows us to bring up the main goal of this Chapter, the topic of *correlations*. When discussing the first three models, some commentators make a big deal of the fact that the MEP has introduced no correlations, as if that were the *raison d'être* for the MEP. It is correct, but irrelevant to any larger purpose, that the MEP has introduced no correlations under those particular models.

Another model, like our final model, *has* introduced a correlation. In fact, the MEP can incorporate the information for correlations of any strength. The MEP wouldn't be of much utility if the IP couldn't use it to operationally implement models where associations between variables were not of primary interest.

The product rule tells us that a probability for any joint statement, $P(AB)$, can be decomposed into $P(A|B) \times P(B)$. A and B are independent if the probability for A does not depend on the known value of B , that is, if $P(AB) = P(A) \times P(B)$.

This was certainly true for the first model investigated where the entry in each cell of the joint probability table was the product of the appropriate marginal probabilities. For example,

$$P(BH | \mathcal{M}_3) = P(B | \mathcal{M}_3) \times P(H | \mathcal{M}_3) = 3/4 \times 3/4 = 9/16$$

If you know for a fact that a kangaroo is right handed, that doesn't change your state of knowledge that it prefers Foster's compared to your original state of knowledge before you knew its hand preference.

In other words, when you condition on the assumed truth that the kangaroo is right handed, Bayes's Theorem tells you that,

$$P(B | H, \mathcal{M}_3) = \frac{P(X = x_1 | \mathcal{M}_3)}{P(X = x_1 | \mathcal{M}_3) + P(X = x_2 | \mathcal{M}_3)} = \frac{9/16}{9/16 + 3/16} = 3/4$$

But the original probability that a kangaroo prefers Foster's was 3/4,

$$P(B | \mathcal{M}_3) \equiv P(X = x_1 | \mathcal{M}_3) + P(X = x_3 | \mathcal{M}_3) = Q_1 + Q_3 = 3/4$$

so nothing is gained when conditioning on the known hand preference. There is no correlation between hand preference and beer preference under this model.

On the other hand, in the final model, independence does not hold. Knowing which hand the kangaroo prefers does make a difference. Now,

$$P(B | H, \mathcal{M}_4) = \frac{P(X = x_1 | \mathcal{M}_4)}{P(X = x_1 | \mathcal{M}_4) + P(X = x_2 | \mathcal{M}_4)} = \frac{0.70}{0.70 + 0.05} = 0.93$$

If you know that the kangaroo is right handed, then it is much more likely that it prefers Foster's. A correlation between hand preference and beer preference has been inserted by model \mathcal{M}_4 .

It is very important to keep in mind that independence, or for that matter, any degree of correlation between hand preference and beer preference, was simply the consequence of allowing the IP to consciously and voluntarily insert information into the joint probability distribution over the state space. The correlation, or lack of a correlation, as discussed so far has absolutely nothing to do with any data.

If data WERE to be gathered, then its job would be to comment on how to change the state of knowledge about all of the models being considered. It might turn out that one of the models postulating a mild correlation between the traits was supported more by the given data than other models postulating independence or strong correlations.

But that exercise is completely independent from the MEP's initial goal of providing legitimate numerical values to probabilities for the joint statements in the state space based on the information in a model.

21.4 Probability for Future Kangaroos

We have gone through the exercise of finding some numerical assignments under four different models by relying upon the MEP. Let's move on to the ultimate goal of assessing the IP's proper state of knowledge about future frequency counts.

From our work in Volume I, we know that the probability for the future frequency counts is the average of the assignments under each specific model with respect to the degree of belief in each of these models.

$$P(M_1, M_2, M_3, M_4) = \sum_{k=1}^{\mathcal{M}} P(M_1, M_2, M_3, M_4 | \mathcal{M}_k) P(\mathcal{M}_k)$$

We derived the interesting consequences of this manipulation rule template for the continuous space of assignments by integrating with respect to a Dirichlet distribution representing the IP's degree of belief in the model's assignment.

If the IP averages over only one model, then it has effectively employed the Dirac delta function to arrive at,

$$P(M_1, M_2, M_3, M_4) = W(M) Q_1^{M_1} Q_2^{M_2} Q_3^{M_3} Q_4^{M_4}$$

Suppose that one model were model \mathcal{M}_3 . Then, referring back to the occupancy numbers given as an example back in section 21.2, the IP's degree of belief that these frequency counts will actually come about in some future tally of sixteen kangaroos is,

$$\begin{aligned} P(M_1 = 9, M_2 = 3, M_3 = 3, M_4 = 1) &= W(M) Q_1^{M_1} Q_2^{M_2} Q_3^{M_3} Q_4^{M_4} \\ &= W(16) \times (0.5625)^9 \times (0.1875)^3 \times (0.1875)^3 \times (0.0625)^1 \\ &= \frac{16!}{9! 3! 3! 1!} \times (0.5625)^9 \times (0.1875)^3 \times (0.1875)^3 \times .0625 \\ &= 0.0245 \end{aligned}$$

If that one model were instead model \mathcal{M}_1 where all four $Q_i = 1/4$, then the prediction equation simplifies even further to,

$$\begin{aligned} P(M_1 = 9, M_2 = 3, M_3 = 3, M_4 = 1) &= \frac{W(M)}{n^M} \\ &= \frac{1,601,600}{4,294,967,296} \\ &= 0.000373 \end{aligned}$$

The IP does not believe that this frequency count is as likely under the fair model as under the first model. And, of course, that makes complete intuitive sense as well.

Under the single model \mathcal{M}_4 , where a strong correlation between hand preference and beer preference was introduced, the degree of belief switches to frequency counts

like these $\boxed{11\ 1\ 0\ 3}$, $\boxed{12\ 0\ 0\ 4}$, $\boxed{13\ 0\ 0\ 3}$, $\boxed{14\ 0\ 0\ 2}$. Together, these four future frequency counts comprise almost 15% of the total probability.

Contrary to these situations where the IP knows a lot, what if the IP is operating under a state of total ignorance? Then it has to lend equal weight to every single model making every conceivable numerical assignment to the probabilities for the four statements about the kangaroo's traits.

When the predictions from all these models are averaged,

$$P(M_1, M_2, M_3, M_4) = \frac{M! (n-1)!}{(M+n-1)!} = \frac{1}{969} = 0.001032$$

Every possible contingency table has the same probability. The IP believes that $\boxed{16\ 0\ 0\ 0}$ is just as likely as $\boxed{4\ 4\ 4\ 4}$. In other words, all sixteen kangaroos might have the same hand-beer preference, or they just as likely might be evenly distributed over all four hand-beer preferences. The IP, operating under such a state of ignorance, just cannot make any predictions that distinguish future frequency counts as it could under more definitive knowledge.

Compare the four future contingency tables just mentioned under the correlation model that showed high frequency counts for right-handed Foster's drinkers. The IP had a high degree of belief that these and very similar contingency tables would be seen in any future count of sixteen kangaroos. When the IP is saddled with complete ignorance, these four contingency tables now comprise a total probability of only $4 \times 0.001032 = 0.00413$ instead of a probability of about 0.15 under the more informed state of knowledge.

In the exercises, we will examine a few particular cases of model averaging. We will explicitly average over the four models discussed in this Chapter. We will, as well, examine the consequences of adjusting the α_i parameters in the Dirichlet distribution.

21.5 Connections to the Literature

The kangaroo scenario was discussed by Jaynes [22] his paper,

Monkeys, Kangaroos, and N.

Chapter Fifteen of Volume I [3] was given over exclusively to a detailed discussion of what it meant for an IP to be “completely uninformed.” We also went deeply into the issue of the different “spaces,” and the reciprocal uncertainty relationships that are the source of so much confusion. I relied heavily on Jaynes’s brilliant analysis and mathematical expressions in this paper for my exposition.

However, I cryptically alluded to some misgivings over Jaynes’s presentation. I said that further clarification over what I had in mind would have to wait until we had sufficient exposure to the MEP algorithm. Well, now that time has come.

As mentioned before in Volume I, we will skip over Jaynes's introductory material. We start with his section entitled, KANGAROOS, starting on page 34. There we see a *joint probability table* for the four joint statements involving a kangaroo's beer and hand preference. The marginal probabilities of 3/4 for beer preference and hand preference are shown as well.¹

An initial frisson of unanticipated anxiety intrudes itself early on when Jaynes says,

Gull and Skilling ... find the remarkable result that if the solution is to be found by maximizing some quantity, entropy is uniquely determined as the only choice that will not introduce spurious correlations [into the joint probability table], for which there is no evidence in the data. The maximum entropy solution is then advocated on grounds of logical consistency rather than multiplicity.

As mentioned earlier, the MEP algorithm has no problem with either models of independence, or models proposing any degree of association among the joint statements. So it is somewhat confusing to the reader to bring up this irrelevant point about "spurious correlations." More disturbing is the implication that the desired information under a particular model is conflated with some data. And finally, "any," not "the," maximum entropy solution is advocated neither on the grounds of logical consistency, nor on the grounds of multiplicity, but rather on the grounds that it does not introduce any extraneous information into a probability distribution.

But let's pretend to ignore all of that and move forward. One would then expect to see Jaynes present a standard MEP approach to filling in the four cells of the joint probability table with numerical assignments under some model, similar to what was done in section 21.3. But shockingly, at this juncture, Jaynes says:

But, kangaroos being indivisible, it is required also that the entries have the form $p(i, j) = N(i, j)/N$ with $N(i, j)$ integers, where N is the number of kangaroos. So for any finite N there are a finite number of integer solutions $N(i, j)$.

I used the adjective *shocking* because Jaynes is clearly defining the probabilities in the joint probability table in terms of frequency counts!

In case there was any doubt over what he had in mind, he presents two examples. In the first example, he supposes that the number of kangaroos is four. In this case, he says, there are only two solutions, $\boxed{2 \ 1 \ 1 \ 0}$, and $\boxed{3 \ 0 \ 0 \ 1}$.

In the second example, he supposes that the number of kangaroos is sixteen. Then, there are five integer solutions, (1) $\boxed{8 \ 4 \ 4 \ 0}$, (2) $\boxed{9 \ 3 \ 3 \ 1}$, (3) $\boxed{10 \ 2 \ 2 \ 2}$, (4) $\boxed{11 \ 1 \ 1 \ 3}$, and (5) $\boxed{12 \ 0 \ 0 \ 4}$.²

¹Notice that I changed the order of these traits in my version.

²There is a calculational typo in the percentages presented for these five solutions.

These integer solutions have apparently been chosen because the frequency counts adhere to the marginal probabilities of $3/4$ as specified when the problem was set up. But this is hogwash! These so-called “integer solutions” are, in fact, contingency tables. They are not joint probability tables. And they certainly are NOT any MEP solution for the given information.

Would Jaynes have us believe that the integer solution $[12 \boxed{0} 0 \boxed{4}]$ equivalent to, $p(i, j) = N(i, j)/N = (3/4, 0, 0, 1/4)$, is an acceptable MEP solution for the given information? This is nonsense. The information entropy of this assignment does not possess the maximum possible entropy of all assignments satisfying the constraints.

As we have discovered in our slow and careful exposition of the MEP algorithm, the numerical assignments to the probabilities for the four joint statements HAVE NOTHING WHATSOEVER TO DO WITH FREQUENCY COUNTS.

It *is* perfectly legitimate for an IP to assert a model that inserts information into a probability distribution. If the IP wants to insert information about marginal probabilities, suitable constraint functions and constraint function averages may be constructed, just as we did in our example of model \mathcal{M}_3 in section 21.3. The resulting MEP assignments were found to be $Q_i = (9/16, 3/16, 3/16, 1/16)$.

There are NOT five integer solutions. These five so-called solutions are just five possible instances from the overall total of 969 possible contingency tables, or frequency counts. Any one of these 969 frequency counts might occur in some future sample of sixteen kangaroos.

If we saw one of Jaynes’s five integer solutions, say, $[9 \boxed{3} 3 \boxed{1}]$, as *data*, then what would happen is that we would modify our degree of belief in all of the models under consideration. This would be accomplished through the usual manipulation rules elaborated on in Volume I, and in the Chapters presented so far in Volume II.

We know intuitively what would happen. Models that produced assignments like $Q_i = (9/16, 3/16, 3/16, 1/16)$ would be strongly supported, while other models that produced assignments like $Q_i = (0, 1, 0, 0)$ would be eliminated. It is also clear that $Q_i = (0.5624, 0.1876, 0.1876, 0.0624)$, a perfectly legitimate MEP assignment under some other model, and clearly not one of the integer solutions, would also be strongly supported by the observed data.

At this point, you are saying to yourself, “It is simply inconceivable that Ed Jaynes, of all people, would make such a blunder!” But, sure enough, right after presenting the five integer solutions, he states that “the maximum entropy solution comprises nearly two-thirds of the feasible set.” where the “maximum entropy solution” is the integer solution $[9 \boxed{3} 3 \boxed{1}]$.

But this “maximum entropy solution” is still being defined in terms of frequency counts; it is a contingency table and not a joint probability table. More seriously, this “maximum entropy solution” was found, not by the MEP algorithm, but rather by calculating the maximum *multiplicity factor* for each integer solution. No matter what the most committed apologist might offer up as an excuse, this is definitely confusing to any reader trying to understand the quintessential nature of the MEP.

After some long reflection, it becomes clear that Jaynes's purpose here is to examine the case where the number of kangaroos approaches infinity. And the point of doing that is to look at what happens to the multiplicity factor as M and the M_i get larger and larger. Jaynes tried to repeatedly justify the MEP with this kind of appeal to the multiplicity factor as M approaches infinity.

We will repeat in detail later on in this Volume the mathematical consequences of trying to equate the multiplicity factor with information entropy. Historically, this is how Boltzmann derived his entropy expressions. And, as I endeavor to show in Chapter Twenty Seven, Erwin Schrödinger employed exactly the same tactics in recapitulating Boltzmann's results.

The approach by both Boltzmann and Schrödinger was thoroughly imbued through a frequentist mind-set. Neither one of these great physicists grasped the fundamental conceptual notion of probability as an epistemological degree of belief.

My own opinion is that Jaynes, as a long and serious student of both Boltzmann and Schrödinger, probably unconsciously absorbed this *idée fixe* that probability and entropy *had* to be defined in terms of frequencies. Even though he knew better than anyone that probabilities and information entropy were fundamentally epistemological concepts, he never gave up trying to justify the MEP with explanations like the one we see expounded in this paper.

And it is a great shame because the MEP stands on its own as a simple and beautiful algorithm just as Jaynes first described it in 1957. There is absolutely no need to try to justify it with fallacious appeals to contingency tables, or data approaching infinite numbers. In the end, I think that these alternative explanations dragging in the multiplicity factor confused far more people than it helped. It is still confusing people to the present day.

As a final gripe, Jaynes constantly referred to the MEP as a suitable tool for constructing "priors." This kind of language has been the source of much confusion as well. The MEP is used to assign numerical values to the n joint statements in the state space.

This is not any kind of "prior" probability, but clearly an assignment conditioned on the information within some model. It is NOT an assignment "prior" to any data, but completely independent of any data. The only thing which might deserve the label of a "prior" probability is $P(\mathcal{M}_k)$ where an assignment is made to models "prior" to knowledge of any data.

21.6 Solved Exercises for Chapter Twenty One

Exercise 21.6.1: What are the minimum and maximum values for Q_1 ?

Solution to Exercise 21.6.1

Suppose we restrict ourselves to the class of models where the marginal probabilities for beer preference and hand preference are both fixed at 3/4. In other words, all models will contain the information $\langle F_1 \rangle = \langle F_2 \rangle = 0.75$. In a correlation model with $m = 3$, the information about Q_1 specifies a particular model. $Q_1 = 0.50$ is the minimum value and $Q_1 = 0.75$ is the maximum value. Otherwise, we would be violating one or both of the first two constraints.

Exercise 21.6.2: What are the implications of a model with $Q_1 = 0.50$?

Solution to Exercise 21.6.2

This is the strongest negative correlation possible under the restricted class of models. Bayes's Theorem tells us that,

$$P(B | H) = \frac{P(BH)}{P(H)} = \frac{0.50}{0.75} = 2/3$$

If beer preference were independent of hand preference, then the probability for Foster's would be 3/4. But under the correlational information in this model, the probability for preferring Foster's is depressed to 2/3 when it is known that the kangaroo is right-handed.

A kangaroo is certain to prefer Foster's if it known that the kangaroo is left-handed because,

$$P(B | \bar{H}) = \frac{P(B\bar{H})}{P(\bar{H})} = \frac{0.25}{0.25} = 1$$

It is impossible for a kangaroo to prefer Corona if it left-handed because this model placed a 0 into cell 4 of the joint probability table. This is yet another example of probability theory generalizing deduction.

Exercise 21.6.3: What are the implications of a model with $Q_1 = 0.75$?

Solution to Exercise 21.6.3

This is seen to be the strongest positive correlation possible under the restricted class of models. Bayes's Theorem tells us that,

$$P(B | H) = \frac{P(BH)}{P(H)} = \frac{0.75}{0.75} = 1$$

If beer preference were independent of hand preference, then the probability for Foster's would be $3/4$. But under the correlational information in this model, it is certain that a kangaroo prefers Foster's when it is known that the kangaroo is right-handed.

Likewise, it is certain that a kangaroo prefers Corona if it is left-handed,

$$P(\overline{B} | \overline{H}) = \frac{P(\overline{B} \overline{H})}{P(\overline{H})} = \frac{0.25}{0.25} = 1$$

It is impossible for a kangaroo to prefer Corona if it is right-handed and impossible for a kangaroo to prefer Foster's if it is left-handed because this model placed 0s into cells 2 and 3 of the joint probability table. Yet again another example of probability theory generalizing deduction.

Exercise 21.6.4: How does the MEP formula arrive at either of these two models?

Solution to Exercise 21.6.4

Of course, for a state space of dimension $n = 4$, under models with $m = 3$ constraints together with the universal constraint, there is no residual ambiguity for the MEP to resolve. The MEP formula can be ignored. If $Q_1 = 0.50$, then $Q_2 = 0.25$ in order to satisfy the first constraint. If $Q_1 = 0.50$, then $Q_3 = 0.25$ in order to satisfy the second constraint. Finally, $Q_4 = 0$ to satisfy the universal constraint.

Since Q_4 will be assigned a 0, we know that there will always be some numerical considerations in any computer solution. When *Mathematica* is asked to use the MEP algorithm to solve for this assignment, it arrives at the above answer within machine precision by reporting that $Q_4 = 2.32 \times 10^{-16}$. The Lagrange multipliers are given as,

$$\lambda_1 = +34.6128$$

$$\lambda_2 = +34.6128$$

$$\lambda_3 = -33.9196$$

One must admire the contortions that the MEP formula has undergone in arriving at these results. Examining the details also serves as a valuable refresher on the vector matrix multiplications going on behind the scenes.

The assigned probability for the first statement in the state space under this most extreme negative correlational model must be $Q_1 = 0.50$. This is computed

by the formula as,

$$\begin{aligned} P(X = x_1 | \mathcal{M}_k) &= \frac{e^{\lambda_1 + \lambda_2 + \lambda_3}}{e^{\lambda_1 + \lambda_2 + \lambda_3} + e^{\lambda_1} + e^{\lambda_2} + e^0} \\ &= \frac{e^{35.3060}}{e^{35.3060} + e^{34.6128} + e^{34.6128} + 1} \\ &\approx 1/2 \end{aligned}$$

Likewise, the zero probability for the last statement in the state space is approximated by,

$$\begin{aligned} P(X = x_4 | \mathcal{M}_k) &= \frac{1}{e^{\lambda_1 + \lambda_2 + \lambda_3} + e^{\lambda_1} + e^{\lambda_2} + e^0} \\ &= \frac{1}{e^{35.3060} + e^{34.6128} + e^{34.6128} + 1} \\ &\approx 0 \end{aligned}$$

In the program, the partition function is formed by multiplying the constraint function matrix (a three row by four column matrix) by the Lagrange multiplier vector (a one row by three column vector), resulting in a one row by four column vector. The conformal nature of the multiplication is $(1 \times 3) \times (3 \times 4) = 1 \times 4$. (See Appendix B for more detail on vector-matrix multiplication.) This vector-matrix multiplication shown below is the argument to the exponential function.

$$(\lambda_1 \quad \lambda_2 \quad \lambda_3) \cdot \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} = (\lambda_1 + \lambda_2 + \lambda_3 \quad \lambda_1 \quad \lambda_2 \quad 0)$$

Essentially, the same things can be said of the positive correlated model. Again for a state space of dimension $n = 4$, under models with $m = 3$ constraints together with the universal constraint, there is no residual ambiguity for the MEP to resolve. The MEP formula can be ignored. If $Q_1 = 0.75$, then $Q_2 = 0$ in order to satisfy the first constraint. If $Q_1 = 0.75$, then $Q_3 = 0$ in order to satisfy the second constraint. Finally, $Q_4 = 0.25$ to satisfy the universal constraint.

Mathematica reports back the same assignment with the two zeroes on the order of 10^{-16} . The Lagrange multipliers for this model are,

$$\lambda_1 = -34.7643$$

$$\lambda_2 = -34.0780$$

$$\lambda_3 = +69.9409$$

The assigned probability for the first statement in the state space under this most extreme positive correlational model must be $Q_1 = 0.75$. This is computed

by the formula as,

$$\begin{aligned}
 P(X = x_1 | \mathcal{M}_k) &= \frac{e^{\lambda_1 + \lambda_2 + \lambda_3}}{e^{\lambda_1 + \lambda_2 + \lambda_3} + e^{\lambda_1} + e^{\lambda_2} + e^0} \\
 &= \frac{e^{1.0986}}{e^{1.0986} + e^{-34.7643} + e^{-34.0780} + 1} \\
 &\approx 3/4
 \end{aligned}$$

Exercise 21.6.5: What is the probability of observing any frequency count for the 16 kangaroos if the IP is “completely uninformed?”

Solution to Exercise 21.6.5

If the IP is “completely uninformed,” we take this to mean exactly as Laplace explained it. The IP knows nothing about what “causes” the kangaroos to appear in a particular cell of the contingency table.

Therefore, the distribution of the degree of belief concerning these putative causes, that is, the models assigning the Q_i as numerical values, must be uniformly distributed. $P(\mathcal{M}_k)$ is distributed as a Dirichlet distribution with its α_i parameters all equal to 1.

As a consequence,

$$P(M_1, M_2, M_3, M_4) = \frac{M! (n-1)!}{(M+n-1)!} = \frac{1}{969}$$

All possible future frequency counts, or contingency tables, possess the same probability. Whether you ask about the probability that all 16 kangaroos are left-handed Corona drinkers, $P(M_1 = 0, M_2 = 0, M_3 = 0, M_4 = 16) = 1/969$, or the probability that all 16 kangaroos are evenly distributed amongst the four traits, $P(M_1 = 4, M_2 = 4, M_3 = 4, M_4 = 4) = 1/969$, the degree of belief that the statement detailing these counts is true is the same for all such statements.

This is what probability theory tells you is the inevitable ramifications of being “totally ignorant” about the causal nature of the kangaroo’s traits. It is the outcome of possessing “insufficient reason” to believe in any one cause, that is, to believe in the assignment under any one model, to the exclusion of all other causes and the assignments under all other models. It boils down to a matter of principle.

Exercise 21.6.6: What is the probability of observing the frequency count of kangaroos in the contingency table of Figure 21.1 when averaging over just four models?

Solution to Exercise 21.6.6

Average over the four models discussed in this Chapter. Furthermore, suppose that the IP spreads the degree of belief evenly among these four models. Thus, we will be calculating the average,

$$\begin{aligned}
 P(M_1 = 9, M_2 = 3, M_3 = 3, M_4 = 1) &= \\
 \sum_{k=1}^4 P(M_1 = 9, M_2 = 3, M_3 = 3, M_4 = 1 | \mathcal{M}_k) P(\mathcal{M}_k) & \\
 = \sum_{k=1}^4 P(M_1 = 9, M_2 = 3, M_3 = 3, M_4 = 1 | \mathcal{M}_k) \times 1/4 & \\
 = 1/4 \times \left[\sum_{k=1}^4 P(M_1 = 9, M_2 = 3, M_3 = 3, M_4 = 1 | \mathcal{M}_k) \right] &
 \end{aligned}$$

The probability of these frequency counts has already been calculated for two of the models, \mathcal{M}_1 and \mathcal{M}_3 . After performing the same calculations for the assignments under the other two models, the average is found to be,

$$\begin{aligned}
 P(M_1 = 9, M_2 = 3, M_3 = 3, M_4 = 1) &= \\
 1/4 \times (0.000373 + 0.003024 + 0.024521 + 0.000202) &= 0.00703
 \end{aligned}$$

Exercise 21.6.7: What observations can you make about the influence of the α_i parameters in the averaging process to find the probability of the future frequency counts?

Solution to Exercise 21.6.7

An extensive numerical investigation of the influence of the α_i parameters of the Dirichlet distribution was conducted in Chapter Fifteen of Volume I. A uniform distribution over model space was captured by setting all four α_i parameters equal to 1. We have mentioned many times that averaging with respect to this distribution results in the same probability of 1/969 to every one of the 969 possible frequency counts.

If the α_i parameters are permitted to advance in lockstep to ever larger and larger numbers, then the probability for any one of the contingency tables is a strict function of its multiplicity factor. In other words, it is as if a single model for the Q_i were averaged over, or averaging over a Dirac delta function as the probability distribution for the model space.

For example, if all four $\alpha_i = 4000$, then the probability for the contingency table $\boxed{4} \boxed{4} \boxed{4} \boxed{4}$ calculated as 0.014661 by the *Mathematica* program in Appendix E of Volume I is very, very close to,

$$\begin{aligned} P(M_1 = 4, M_2 = 4, M_3 = 4, M_4 = 4) &= \frac{W(M)}{n^M} \\ &= \frac{63,063,000}{4^{16}} \\ &= \frac{63,063,000}{4,294,967,296} \\ &= 0.014683 \end{aligned}$$

as calculated under the average for one (fair) model where the $Q_i = 1/4$. As the α_i are allowed to increase, the probability will approach arbitrarily close to 0.014683.

Finally, suppose we let the α_i parameters change so as to promote the one model, model \mathcal{M}_3 , where $Q_1 = 9/16$, $Q_2 = 3/16$, $Q_3 = 3/16$, and $Q_4 = 1/16$. Table 21.2 shows that the probability for the future frequency count of $\boxed{9} \boxed{3} \boxed{3} \boxed{1}$ increases to what it would be under model \mathcal{M}_3 .

Table 21.2: How the probability for a future frequency count of $\boxed{9} \boxed{3} \boxed{3} \boxed{1}$ changes as the α_i parameters increase in the Dirichlet distribution capturing the probability of the models.

α_1	α_2	α_3	α_4	\mathcal{A}	$P(\boxed{9} \boxed{3} \boxed{3} \boxed{1})$
1	1	1	1	4	1.03×10^{-3}
9	3	3	1	16	8.09×10^{-3}
90	30	30	10	160	2.12×10^{-2}
900	300	300	100	1600	2.42×10^{-2}
∞	∞	∞	∞	∞	2.45×10^{-2}

The calculated probability for the future frequency counts is the result of an average with respect to the probability distribution for the models. This probability distribution in turn is being controlled by the α_i parameters in the Dirichlet

distribution. The stunning, but very satisfying conclusion to all of this, is that the α_i parameters are acting *as if they were virtual observations!*

If it were not obvious at this point, it can be easily shown that we arrive at the very same probability for this frequency count by starting out with a uniform distribution over model space. Then, let the data drive us to the same conclusion. We would calculate the probability for any one of the 969 frequency counts as $P(M_1, M_2, M_3, M_4 | \mathcal{D})$.

As data, let the total number of kangaroos sampled be $N = 1600$, and in the pattern designed to maximally support model \mathcal{M}_3 , let the actual observed counts in the four cells of the contingency table be $N_1 = 900$, $N_2 = 300$, $N_3 = 300$, and $N_4 = 100$. We find that, as expected, the probability for the future frequency count $(M_1 = 9, M_2 = 3, M_3 = 3, M_4 = 1)$,

$$P(M_1, \dots, M_4 | N_1 = 900, N_2 = 300, N_3 = 300, N_4 = 100)$$

works out to,

$$\begin{aligned} P(M_1, \dots, M_4 | \mathcal{D}) &= C \times \frac{\prod_{i=1}^4 (M_i + N_i)!}{\prod_{i=1}^4 M_i!} \\ C &= \frac{M! (N + n - 1)!}{N_1! N_2! N_3! N_4! (M + N + n - 1)!} \\ &= \frac{16! (1600 + 4 - 1)!}{900! 300! 300! 100! (16 + 1600 + 4 - 1)!} \\ &= \frac{16! 1603!}{900! 300! 300! 100! 1619!} \\ \frac{\prod_{i=1}^4 (M_i + N_i)!}{\prod_{i=1}^4 M_i!} &= \frac{909! 303! 303! 101!}{9! 3! 3! 1!} \\ P(M_1, \dots, M_4 | \mathcal{D}) &= \frac{16! 1603!}{900! 300! 300! 100! 1619!} \times \frac{909! 303! 303! 101!}{9! 3! 3! 1!} \\ &= 0.0242 \end{aligned}$$

Compare this probability with the already calculated binomial probability under model \mathcal{M}_3 of,

$$P(M_1 = 9, M_2 = 3, M_3 = 3, M_4 = 1 | \mathcal{M}_3) = 0.0245$$

So, in the end, satisfactorily, the probability for any particular future frequency count is the same whether we set up one model with a high probability in the averaging process, or set up all models with the same probability, and then let a massive amount of data reorient their relative status through the averaging process.

Chapter 22

MEP Models and Correlation

22.1 Introduction

You are looking forward to a pleasant evening out at the pub with your friends, the kangaroos. You're talking about the upcoming festivities with Leonora when you decide to have some fun. You've noticed that Oscar has sandy colored fur and is right handed.

“Say Leonora, I’ll bet you that Oscar orders Foster’s tonight. You’ll give me a dollar if I’m right, and to make it interesting, I’ll give you four dollars if I’m wrong. By the way, I also have a pretty good idea of what beer you will ask for as well, but I’m not going to tell you because I know you’ll do the opposite just to spite me.”

Our objective here is to illustrate how the MEP provides a way of understanding how correlations among variables may arise. We proceed from our general principle that any such relationship among variables must be represented as information under a model. We will use the words *relationship*, *association*, and *correlation* interchangeably to express the fundamental idea of some dependency relationship among the variables in question.

The notion of correlations is central to all inferential problems. We introduced the idea in the last Chapter as we looked at various models that took us further and further away from strict independence between hand and beer preference. These models compelled us to look more closely at some sort of relationship between hand and beer preference.

This association among the kangaroo’s traits manifested itself through Bayes’s Theorem. Thankfully, this effort at introducing correlations turned out to be just another simple numerical example illustrating how the MEP algorithm works for assigning probabilities to joint statements.

As we have already learned, conditioning on statements that are independent of the statements we'd like to predict doesn't help us out one bit. Productive scientific models *must* hypothesize some degree of relationship, or dependency, between statements in the state space before any interesting causal explanations can arise. The MEP takes center stage as an IP's chief ally in thinking about, and then forming, these kind of dependent relationships among putative causal variables.

As always, we frame the discussion in terms of an inference; a state of knowledge about some future frequency count of the kangaroos described by one of the available joint statements. To make the kangaroo scenario somewhat more interesting, and yet keep the problem as simple as possible, one more binary statement is added to our already familiar beer and hand preference statements. We now include a kangaroo's fur color as a third trait.

Suppose that data are available from past frequency counts so that we can refresh our memories as to their impact on model reorientation. Our primary goal will be to acquire a state of knowledge about one trait of the very next kangaroo. This inference will be conditioned on already existing knowledge of the two remaining traits, as well as all of the past data. When we say, "the very next kangaroo," we mean any kangaroo that hasn't already been incorporated into the data base.

To alleviate the abstractness, let's use language that calls *beer preference* a personality trait, with *fur color* and *hand preference* physical traits. We will discuss models that insert information about correlations between the personality trait and the physical traits. Then, we will be able to assess how various hypothesized causal explanations involving hand preference and fur color impact an IP's state of knowledge about beer preference.

The IP will want to assess its state of knowledge about the beer preference of the very next kangaroo *after* having observed its fur color and hand preference. This assessment will be based on previous measurements where all three traits were measured together. This state of knowledge will be the probability of beer preference as computed by Bayes's Theorem averaged over all models once they have been re-ordered by the data.

22.2 Setting up the Problem

The immediate goal is to improve our familiarity with correlational models. What are the consequences for inferences about statements that are correlated with some number of supposed causal factors?

Suppose then, for the sake of this inferential problem, that the IP is curious about the personality traits of kangaroos. The IP wonders if a personality induced preference for Foster's beer is really influenced by the two physical traits of fur color and hand preference. More generally, the IP wonders whether it is a reasonable scientific question to assert that some causal factors in the form of readily observable physical characteristics of the kangaroos do underpin latent personality traits.

We will switch now to a more convenient notation than used in Volume I where beer preference (B) can be measured as either Foster's or Corona, and hand preference (H) is either right-handed or left-handed. The new variable of fur color (F) is categorized as either sandy or beige.

The most direct way of writing Bayes's Theorem for our desired inference about the personality trait conditioned on knowledge of the two physical traits, as well as conditioned on all of the past observations is,

$$\begin{aligned} P(B | H, F, \mathcal{D}) &= \frac{P(B, H, F | \mathcal{D})}{P(H, F | \mathcal{D})} \\ &= \frac{P(B, H, F | \mathcal{D})}{P(B, H, F | \mathcal{D}) + P(\overline{B}, H, F | \mathcal{D})} \end{aligned} \quad (22.1)$$

where the B and H notation indicate statements about beer and hand preference. To indicate statements about fur color, let $F \equiv$ "The kangaroo has sandy fur color." and $\overline{F} \equiv$ "The kangaroo has beige fur color."

It seems like a trivial observation, but it turns out to be critical that we keep the probabilities on the right hand side of Bayes's Theorem as *joint* probabilities. What is the point of doing that? The reason is that the MEP algorithm is designed to find numerical assignments for joint statements.

22.2.1 Predicting future frequency counts

In Volume I, one of our major achievements was to derive a general prediction formula for assessing a state of knowledge about future frequency counts when conditioned on past frequency counts. In other words, an IP could make improved inferences about some future event after some data had been collected.

The formula has usually been presented in these formats,

$$P(M_1, M_2, \dots, M_i, \dots, M_n | N_1, N_2, \dots, N_i, \dots, N_n) \equiv P(M_1, \dots, M_n | \mathcal{D}) \quad (22.2)$$

The solution for the probability in Equation (22.2) was derived from the formal manipulation rules of probability theory. It was not concerned with the assignments emanating from any one particular model. During the derivation, an integration was performed over all possible numerical assignments. All models therefore contributed to the IP's state of knowledge about the future frequency counts.

The way we have set up the inferential problem, we want to find the probability that the *next* kangaroo possesses a certain combination of traits. Thus, in this case, we are always trying to solve the simple problem where the future frequency count is expressed in our notation as $M = 1$. One of the M_i will equal 1, while all the others must equal 0.

To solve Bayes's Theorem in Equation (22.1), we will need to find the two joint probabilities appearing in the numerator and denominator,

$$P(B_{N+1}, H_{N+1}, F_{N+1} | \mathcal{D}) \equiv P(M_1 = 1, \dots, M_8 = 0 | \mathcal{D}) \quad (22.3)$$

and,

$$P(\bar{B}_{N+1}, H_{N+1}, F_{N+1} | \mathcal{D}) \equiv P(M_1 = 0, \dots, M_5 = 1, \dots, M_8 = 0 | \mathcal{D}) \quad (22.4)$$

The relevant future frequency count of $M = 1, M_1 = 1$ (with the remaining $M_i = 0$) resides in cell 1 of the eight cell contingency table. The relevant future frequency count of $M = 1, M_5 = 1$ (with the remaining $M_i = 0$) resides in cell 5 of the contingency table.

The positions of the cells will be obvious once we sketch the layout of the joint probability table and the contingency table. Figure 22.1 shows a re-organized eight cell joint probability table with the two physical traits together in each sub 2×2 table.

		B	\bar{B}
		H	\bar{H}
		Cell 1 Q ₁ $P(X=x_1)$	Cell 2 Q ₂ $P(X=x_2)$
F	\bar{F}	Cell 3 Q ₃ $P(X=x_3)$	Cell 4 Q ₄ $P(X=x_4)$
		Cell 5 Q ₅ $P(X=x_5)$	Cell 6 Q ₆ $P(X=x_6)$
		Cell 7 Q ₇ $P(X=x_7)$	Cell 8 Q ₈ $P(X=x_8)$

Figure 22.1: An eight cell joint probability table for one personality trait and two physical traits in the kangaroo scenario.

22.2.2 The state space

We added one binary variable of fur color to the two already existing binary variables of beer and hand preference. So, the dimension of the state space for this new kangaroo scenario jumps up to $n = 8$. The maximum number of constraint functions in the most complex models will then be $m = n - 1 = 7$.

We will again take advantage of constraint functions that are marginal probabilities. Thus, independent models will involve $m = 0$ through $m = 3$ constraint functions. Any correlations will have to appear in models with $m = 4, 5, 6$, or 7 constraint functions.

The joint probability table is concerned with eight joint statements. The first is, “A kangaroo prefers Foster’s beer, uses its right hand, and has sandy fur color.”, while the eighth and last joint statement is, “A kangaroo prefers Corona beer, uses its left hand, and has beige fur color.”

As usual, some model \mathcal{M}_k will assign legitimate numerical values to the probabilities for each one of these eight joint statements. The MEP algorithm is always relied upon as the operational way that the IP inserts the desired information from model \mathcal{M}_k into a numerical assignment for the probabilities.

22.2.3 The contingency table

The data are shown in a contingency table as past frequency counts for some number N of kangaroos. These N kangaroos have been correctly placed into one of the available eight categories with N_1 kangaroos in cell 1 of the contingency table, N_2 kangaroos in cell 2, and so on with N_8 kangaroos in the last cell. The sum of all these past frequency counts in each cell of the contingency table must, of course, sum up to all of the data collected, $\sum_{i=1}^n N_i = N$.

For the sake of a numerical example, suppose that $N = 100$ kangaroos have been observed in the past. All three traits have been duly recorded on each of these 100 kangaroos. It turned out that 40 right handed kangaroos with sandy fur color preferred Foster’s beer. Thus, the value of $N_1 = 40$ was placed into cell 1 of the contingency table.

Similarly, all of the data as shown below in Figure 22.2 was placed into the proper cell of the contingency table. As a check, we see that $\sum_{i=1}^n N_i = N = 100$. The marginal *sums* for the traits of preferring Foster’s, being right handed, and having sandy fur color are also boxed. I have obviously manipulated these frequency counts in order to make a point.

		B	\bar{B}		
		H	\bar{H}		
		Cell 1	Cell 2		
F	40	10			
	18	7			
		58	17	75	
		50	F	H	\bar{H}
				Cell 5	Cell 6
				5	5
\bar{F}	25	25	\bar{F}	Cell 7	Cell 8
				12	3
		17	8	15	40
		75	25	25	
				100	

Figure 22.2: An eight cell contingency table containing the data from past observations of the physical and personality traits of $N = 100$ kangaroos.

22.2.4 Models using marginal probabilities

The assignments from the MEP algorithm will arise computationally from specifying no constraint functions ($m = 0$), all the way through seven constraint functions ($m = 7$), together with their associated constraint function averages. Once we set up the most complex models with $m = 7$ constraint functions, there are no more degrees of freedom. Recall that the universal constraint forcing the numerical assignments to sum to 1 is always implicit.

Thus, the MEP algorithm would, strictly speaking, not be needed to find the numerical assignments for the most complex models with $m = 7$. There is no remaining ambiguity to resolve by maximum entropy after all of the constraints have been satisfied. Nonetheless, the MEP algorithm still works perfectly fine in this case, and we will continue to use it even when $m = 7$.

We have already learned that any model inserting information about marginal probabilities for each variable will result in independence when Bayes's Theorem is used to make inferences. Therefore, we will not really be all that interested in those models with $m = 0$ through $m = 3$ constraint functions enforcing the marginal probabilities for beer preference, hand preference, and fur color.

The IP is especially interested in whether the physical traits might be among the causal factors influencing the personality trait. The only way the IP can investigate models involving a dependency between causal factors and the variable of interest is to postulate models with $m = 4$ through $m = 7$ constraint functions.

The truly fascinating thing here about the MEP approach is the very easy way in which the constraint functions and their associated averages are formed. The constraint functions will be vectors consisting solely of 1s and 0s. In this way, values for various marginal probabilities can be inserted as the information in a model because marginal probabilities will literally be the averages that are required for these simple constraint functions.

22.3 Correlations Via MEP Models

Since we are focusing on the MEP in this Volume, we deal first with various theoretical hypotheses that a scientist might propose for predicted variables and causal explanatory variables. Here, the scientist as IP wants to make an inference about a personality trait when the causal variables are observable physical traits.

The only way that causal variables are going to have an impact in any scientific explanation is if they possess some association with the variable to be predicted. The quantitative way of expressing these dependencies is by introducing correlations through the information in models \mathcal{M}_k . We will now go into some detail as to how this is done.

We will first discuss some models where *no* correlation exists. For these models, the predicted variable of beer preference is independent of the putative causal factors of fur color and hand preference. These will be any of the models constructed from the first $m = 0, 1, 2$, or 3 constraint functions.

Next, we will get to the interesting part. When we construct models with $m = 4, 5, 6$, or 7 constraint functions, we will be creating dependencies between various variables. As we shall examine in detail later, models with four, five, or six constraint functions will introduce correlations in the form of *double interactions* between two variables, while models with seven constraint functions will introduce a more complicated *triple interaction* among all three variables.

It should always be kept in mind that in all this discussion about various models, the IP is exploring a purely speculative realm divorced from the reality of any actual observations. The data will eventually play a crucial role in our final inferences. But preliminary to that reorientation of the model space by the data, the scientist must engage in a sweeping consideration of all manner of models reflective of all sorts of causal explanations.

This is exactly the activity the IP is engaged in when it postulates models with double and triple interactions. These kinds of models express an interest in investigating causal explanations where 1) beer preference and hand preference are correlated, 2) where beer preference and fur color are correlated, 3) where fur color and hand preference are correlated, and, finally, 4) where all three variables of beer preference, hand preference, and fur color are correlated.

22.3.1 Independent models

We are quite familiar with the model employing no constraints (except for the universal constraint). When $m = 0$, all $Q_i = 1/8$. Bayes's Theorem then tells us that the probability for any inference will be $1/2$. Obviously, any such inference for the predicted variable will be independent of all causal explanatory variables.

Then there are those models with $m = 1, 2$, or 3 constraints, but which are also independent models. These models do, in fact, incorporate information about the marginal probabilities for beer preference, fur color, and hand preference. These marginal probabilities might very well be different from the value of $1/2$ provided by the fair model. Let's examine one of these independent models with $m = 3$ constraint functions as it inserts information about the marginal probabilities into the joint probabilities for all three variables, B , H , and F .

As with our previous dealings with the kangaroos, the marginal probabilities for beer and hand preference will be kept at $3/4$. Suppose now that this model introduces the information that the marginal probability for sandy fur color is 0.60 with, of course the marginal probability for beige fur color at 0.40 .

We now can say that the first constraint function involving beer preference is represented by the vector of 1s and 0s,

$$F_1(X = x_i) = (1, 1, 1, 1, 0, 0, 0, 0)$$

and its average is,

$$\langle F_1 \rangle = \sum_{i=1}^8 F_1(X = x_i) Q_i = Q_1 + Q_2 + Q_3 + Q_4 = 0.75$$

By definition, this is the marginal probability for B given the way we have structured the joint probability table.

The other two constraints follow the same pattern. The marginal probability for hand preference consists of cells 1, 3, 5, and 7 in the joint probability table. Thus, the second constraint function is,

$$F_2(X = x_i) = (1, 0, 1, 0, 1, 0, 1, 0)$$

and its average is,

$$\langle F_2 \rangle = \sum_{i=1}^8 F_2(X = x_i) Q_i = Q_1 + Q_3 + Q_5 + Q_7 = 0.75$$

The marginal probability for fur color consists of cells 1, 2, 5, and 6 in the joint probability table. Thus, the third, and final, constraint function is,

$$F_3(X = x_i) = (1, 1, 0, 0, 1, 1, 0, 0)$$

and its average is,

$$\langle F_3 \rangle = \sum_{i=1}^8 F_3(X = x_i) Q_i = Q_1 + Q_2 + Q_5 + Q_6 = 0.60$$

We put the MEP algorithm to work to find the numerical assignments that satisfy all three of these constraints, and, at the same time, has the maximum entropy of any probability distribution that might also happen to satisfy the constraints. In this way, we are assured that only the information in this model has been inserted into the resulting distribution. No other unwanted information can make its way into the distribution when the MEP algorithm is offering its protection.

The numerical assignments to all eight probabilities in the joint probability table under this new model must be different when compared to the “fair” model that assigned all cells a value of $1/8$. In fact, it turns out that $Q_1 = 0.3375$ and $Q_5 = 0.1125$ under this new model.

These are the two joint probabilities that Bayes’s Theorem needs to compute a state of knowledge about beer preference when conditioned on knowledge of fur color and hand preference.

$$P(B | H, F, \mathcal{M}_k) = \frac{Q_1}{Q_1 + Q_5} = \frac{0.3375}{0.3375 + 0.1125} = 0.75$$

More generally, Bayes's Theorem would be called upon in the same manner when the IP is trying to make any inference about a personality trait when conditioned on knowledge of physical traits.

This result clearly indicates why this model is still a model that does not have any correlations. The probability for preferring Foster's remains at a value of 0.75, even when conditioned on knowledge that the kangaroo is sandy colored and right handed. The probability has not budged from the original marginal probability for preferring Foster's.

This is not surprising since this model did not incorporate any information about a relationship between the variable to be predicted, namely, beer preference, and any of the causal variables, namely, fur color and hand preference. The putative explanatory variables are at a loss to change the IP's original state of knowledge. In other words,

$$P(B | H, F, \mathcal{M}_k) \equiv P(B | \mathcal{M}_k) = 0.75$$

and the state of knowledge about B is independent of both H and F .

Here, once again, is the MEP formula for calculating the numerical assignments just presented for this independent model with $m = 3$ constraints.

$$P(X = x_i | \mathcal{M}_k) = \frac{\exp [\sum_{j=1}^m \lambda_j F_j(X = x_i)]}{Z(\lambda_1, \lambda_2, \lambda_3)} \quad (22.5)$$

As a numerical example, use the MEP formula in Equation (22.5) to calculate the Q_1 and Q_5 assignments needed for Bayes's Theorem. Notice that the subscript i attached to the joint statement ($X = x_i$) will be fixed, in the first case, at $i = 1$ and, in the second case, at $i = 5$. Since the model has $m = 3$ constraints, the subscript j attached to both the Lagrange multipliers and the constraint functions will take on the values 1, 2, and 3 as the summation proceeds.

Thus, to calculate Q_1 we have,

$$Q_1 \equiv P(X = x_1 | \mathcal{M}_k) = \frac{\exp [\sum_{j=1}^3 \lambda_j F_j(X = x_1)]}{Z(\lambda_1, \lambda_2, \lambda_3)}$$

The sum in the argument to the exponent will use the first element in vector F_1 , the first element in vector F_2 , and the first element in vector F_3 . The first element in each of these three constraint functions is 1.

$$\begin{aligned} Q_1 &= \exp [(\lambda_1 \times 1) + (\lambda_2 \times 1) + (\lambda_3 \times 1)] / Z \\ &= \exp [\lambda_1 + \lambda_2 + \lambda_3] / Z \end{aligned}$$

The MEP algorithm returned the values of the Lagrange multipliers as,

$$\lambda_1 = 1.098612$$

$$\lambda_2 = 1.098612$$

$$\lambda_3 = 0.405465$$

and the value of the partition function is $Z = 40$. Substituting these values, we find the numerical assignment Q_1 by,

$$\begin{aligned} Q_1 &= \exp [\lambda_1 + \lambda_2 + \lambda_3] / Z \\ &= \exp [1.098612 + 1.098612 + 0.405465] / 40 \\ &= 0.3375 \end{aligned}$$

In exactly the same manner, we have,

$$Q_5 \equiv P(X = x_5 | \mathcal{M}_k) = \frac{\exp [\sum_{j=1}^3 \lambda_j F_j(X = x_5)]}{Z(\lambda_1, \lambda_2, \lambda_3)}$$

The sum in the argument to the exponent will use the fifth element in vector F_1 , the fifth element in vector F_2 , and the fifth element in vector F_3 . The fifth element in F_1 is 0, while the fifth element in F_2 and F_3 is 1.

$$\begin{aligned} Q_5 &= \frac{\exp [(\lambda_1 \times 0) + (\lambda_2 \times 1) + (\lambda_3 \times 1)]}{Z(\lambda_1, \lambda_2, \lambda_3)} \\ &= \frac{\exp [\lambda_2 + \lambda_3]}{Z(\lambda_1, \lambda_2, \lambda_3)} \\ &= \frac{\exp [1.098612 + 0.405465]}{40} \\ &= 0.1125 \end{aligned}$$

22.3.2 Models with double interactions

The IP can only obtain models that show some dependency between beer preference and the two explanatory variables if it moves up to consider more complex models with $m = 4, 5, 6$, or 7 constraints. Models with $m = 4$ constraints include one double interaction, models with $m = 5$ constraints include two double interactions, and models with $m = 6$ constraints include all three double interactions.

In all of these more complex models, we are retaining the first three constraints that incorporate the information about the marginal probabilities for beer preference, hand preference, and fur color, the so-called “main effects.” The most complex models of all, the ones with $m = 7$ constraints, would incorporate all three main effects, all three double interactions, and finally the one possible triple interaction.

Let’s now take a look at our first model that will exhibit a dependency between the personality trait and the physical traits. That is to say, we are now, after all the lead up, finally examining some correlational models.

The first class of correlational models will have $m = 4$ constraint functions and their associated constraint function averages. Keep the first three constraint

functions and their averages as discussed above for the independent models the same, but add this double interaction between beer preference and handedness. The fourth constraint function is a vector of 1s and 0s,

$$F_4(X = x_i) = (1, 0, 1, 0, 0, 0, 0, 0)$$

with a constraint function average of,

$$\langle F_4 \rangle = \sum_{i=1}^8 F_4(X = x_i) Q_i = Q_1 + Q_3 = 0.60$$

The constraint function can be generated mechanically just by multiplying the two relevant constraint functions F_1 and F_2 . Or, it can be thought about as just another marginal probability, this time a marginal probability for BH .

The number of joint probabilities involved in this more refined marginal probability must drop from four to now two joint probabilities. Here, the two joint probabilities Q_1 and Q_3 make up the marginal probability for BH . Inspection of cells 1 and 3 of the joint probability table will confirm this. Cell 1 is $P(B, H, F)$ and cell 3 is $P(B, H, \bar{F})$, so,

$$P(B, H, F) + P(B, H, \bar{F}) = P(B, H) = 0.60$$

The MEP algorithm assigns the numerical values of $Q_1 = 0.36$ and $Q_5 = 0.09$ under this new model. Do we see a change in the IP's state of knowledge about beer preference when conditioned on hand preference and fur color under a model that introduced correlations between beer preference and handedness?

$$P(B | H, F, \mathcal{M}_k) = \frac{0.36}{0.36 + 0.09} = 0.80$$

Now the sought for goal of a dependency between the personality trait and the physical traits has been achieved. The probability for beer preference did change from 0.75 upwards to 0.80 when these explanatory variables were taken into account. The IP has a greater degree of belief than it did before that this sandy fur colored right handed kangaroo prefers to drink Foster's.

It is very, very important to constantly keep in mind that these probabilities are simply the ones that ensue by assuming a particular model to be true. They may have nothing to do with reality.

At the starting point of our inferences, we have to seriously entertain every single conceivable model on an equal basis. But after we have collected data on the traits of kangaroos, we must modify our degree of belief in the relative standing of all the models. Some models will sink into oblivion when judged against the data, others will rise in prominence.

Pose this question: Is this model which is incorporating a particular correlation between the predicted variable and the explanatory variables one of those that we now believe in more firmly? The data will tell us the answer to that question.

22.3.3 Models with triple interactions

We conclude our brief sampling of independent and dependent models by looking at a more complicated model with $m = 7$ constraint functions. This model will not only insert information about 1) the three marginal probabilities for B , H , and F , but also insert information about 2) the three marginal probabilities of BH , HF , and BF , and 3) one triple interaction involving a marginal probability for BHF .

Recall that the marginal probabilities for the three main effects involved four cells of the joint probability table, while the marginal probabilities for the three double interactions involved only two cells. It follows that the marginal probability for the triple interaction is not a marginal probability after all, but merely a specification for just one cell in the joint probability table.

Thus, the final constraint function is a vector of 1s and 0s looking like,

$$F_7(X = x_i) = (1, 0, 0, 0, 0, 0, 0, 0)$$

with an average,

$$\langle F_7 \rangle = \sum_{i=1}^8 F_7(X = x_i) Q_i = Q_1 = 0.40$$

In the course of implementing the triple interaction, this correlational model is making a assignment to Q_1 directly.

When the MEP algorithm calculates the numerical assignments to all eight cells of the joint probability table under this highly correlated model, it becomes clear what is going on. The numerical assignments under the modeled relationships are,

$$Q_1 = 0.40, Q_2 = 0.10, Q_3 = 0.18, Q_4 = 0.07$$

and,

$$Q_5 = 0.05, Q_6 = 0.05, Q_7 = 0.12, Q_8 = 0.03$$

Do a few quick plausibility checks on this assignment. Do all eight assignments sum to 1? Is the marginal probability for Foster's beer preference $3/4$? Is the marginal probability for sandy fur color 0.60? Is the BF double interaction, the marginal probability for Foster's beer preference and sandy fur color over the right and left hand assignments, equal to $1/2$?

This model provides the strongest inference yet about beer preference.

$$P(B | H, F, \mathcal{M}_k) = \frac{Q_1}{Q_1 + Q_5} = \frac{0.40}{0.40 + 0.05} = 0.8889$$

Under this model, the IP is beginning to approach certainty in its degree of belief that the very next kangaroo, and say that kangaroo happens to be Oscar, will prefer Foster's over Corona. Just by looking at Oscar, we can ascertain that he is a right handed sandy colored kangaroo. If this model were to be highly supported by the data, then that bet might be won after all.

22.4 The Probability Based on all Models

The data were manufactured by me to closely match what one would expect to see under this last model when 100 kangaroos are sampled. I am the puppet master pulling the strings so that this last correlational model, together with the many other models making very similar assignments, will be hugely supported by the data. The vast majority of the models, especially the fair model and all the rest of the independent models, will not be supported by these data. These models will therefore have a negligible influence on the resulting inferences.

Using the predictive formula from Volume I, the correct state of knowledge about the next kangaroo having the joint traits of preferring Foster's beer, having sandy fur color, and using its right hand, is the probability of a future frequency count in cell 1 of the contingency table.

$$P(M_1 = 1, M_2 = 0, \dots, M_8 = 0 | \mathcal{D}) = \frac{M! (N + n - 1)!}{N_1! N_2! \dots N_8! (M + N + n - 1)!} \times \frac{\prod_{i=1}^8 (M_i + N_i)!}{\prod_{i=1}^8 M_i!}$$

Substituting the values of $M = 1$, $M_1 = 1$, $N = 100$, and $n = 8$ together with the observed data \mathcal{D} of $N_1 = 40, N_2 = 10, \dots, N_8 = 3$,

$$\begin{aligned} P(M_1 = 1, \dots, M_8 = 0 | \mathcal{D}) &= \frac{1! (100 + 8 - 1)!}{40! 10! \dots 3! (1 + 100 + 8 - 1)!} \times \frac{41! \times 10! \times \dots \times 3!}{1! \times 0! \times \dots \times 0!} \\ &= \frac{107!}{108!} \times 41 \\ &= \frac{41}{108} \end{aligned}$$

In just the same manner, the IP determines that its state of knowledge about the next kangaroo having the joint traits of preferring Corona beer, having sandy fur color, and using its right hand, is the probability of a future frequency count in cell 5 of the contingency table.

$$P(M_1 = 0, \dots, M_5 = 1, \dots, M_8 = 0 | \mathcal{D}) = \frac{6}{108}$$

With these joint probabilities in our possession, we can calculate from Bayes's Theorem the conditional probability that any new kangaroo not in the data base will

prefer Foster's beer when given that it is a sandy colored, right handed kangaroo.

$$\begin{aligned}
 P(B | H, F, \mathcal{D}) &= \frac{P(B, H, F | \mathcal{D})}{P(B, H, F | \mathcal{D}) + P(\bar{B}, H, F | \mathcal{D})} \\
 &= \frac{41/108}{41/108 + 6/108} \\
 &= 0.8723
 \end{aligned}$$

In the end, we see that this correct probability for a kangaroo to prefer Foster's is just slightly less than the probability recently calculated,

$$P(B | H, F, \mathcal{M}_k) = 0.8889$$

under the single model with $m = 7$ constraints, that is, the model implementing all of the interactions where the information in the model matched the normed frequency counts in the contingency table,

This slightly lessened degree of belief is due to the fact that the prediction formula takes into account every single conceivable numerical assignment, and not just the assignment under one model. There was a minor downward adjustment in the probability due to this action.

But the obvious point is that the vast majority of all these other models made a negligible contribution to the overall average after the data were taken into account. The only models which had any significant contribution to the average were the ones very much like our final correlational model.

22.5 Connections to the Literature

A curious and enigmatic Chapter occurs near the end of Jaynes's flawed *magnum opus* [23],

Probability Theory: The Logic of Science

This Chapter in Jaynes's book, Chapter 18, possesses the mysterious title, "The A_p distribution and rule of succession".

One of these days I am going to give a complete exegesis of the confusing and contradictory contents of this Chapter. It is my firm belief that not one in a hundred really understands what Jaynes was trying to communicate in these pages. But right now all I can do is to remark that it is here where Jaynes acknowledges Laplace's *Rule of Succession* for the important role that it deserves in making inferences.

Moreover, Jaynes clearly states that Laplace's reasoning is the correct way to think about that universally misunderstood relationship between probabilities and

frequencies. He devotes some space, but not nearly enough, to the very tricky question of how the technical aspects of the integration over all of model space can be accomplished by using Laplace transforms.

After writing down the expression for the integration of all q_i values over model space, Jaynes tells us [23, pg. 569],

Direct evaluation of this would be rather messy because all integrations after the first would be between limits that would have to be worked out; so let's use the following trick. Firstly, take the Laplace transform ...

As I mentioned in Volume I, we all are the lucky recipients of the contributions from many people who, in the past, took on the task of solving difficult problems. The happy consequence is that our lives are made much easier. One of those difficult mathematical tasks was finding an analytical solution to a multidimensional integral where the limits of the integration are changing. In the integration over model space, the integration limits are constrained by the fact that the sum of the q_i must be fixed at 1.

This is sometimes expressed as an integration “over the simplex.” To say the least, possessing an analytical solution for this “Dirichlet integral” considerably simplified the derivation of the probability for future frequency counts.

In section 22.4, and also more explicitly in Exercise 22.6.5, a relatively simple and intuitively compelling formula was used to find the probability of an event on the next trial when past data were available. But I only accepted this result for what it was after carefully reading Jaynes's explanation of the *Rule of Succession*.

And when I present these formulas, I am only copying what Jaynes himself directly mentions as an important by-product of the general rule. Jaynes asks us: What would Laplace's general rule say for the specific case of predicting some next event when given knowledge of how things had worked out in previous observations? The formula is,¹

$$P(i^{\text{th}} \text{ statement on next trial} | \text{data}) = \frac{N_i + 1}{N + n}$$

In the end, we see that Laplace makes better sense of frequencies than all of the puzzled contortions of those who vehemently demanded that probabilities be *defined* as frequencies. As Jaynes [23, pg. 576] so wryly puts it,

So I don't see how even the most ardent advocates of the frequency theory of probability can damn the rule of succession without thereby damning his own procedures; after all polemics, there remains the simple fact that in his own procedures he is doing exactly what Laplace's rule of succession tells him to do. Indeed, to define probability in terms of frequency is equivalent to saying that the rule of succession is the *only* rule which can be used for converting observational data into probability assignments.

¹Chapter 18, pg. 571, Equation (18.44)

But I think the most important feature in Jaynes's derivation (where he was now following in the footsteps of what Bayes and Laplace had done earlier), was to adopt a *uniform probability distribution over model space*. Unfortunately, his notation and all of his preceding explanations about the " A_p distribution" leading up to the juncture where the uniform distribution is applied are so obscure, that it passes by unnoticed. And Jaynes does not see fit to make much of it either.

I could be wrong here, but in my opinion Jaynes compounded the confusion with his penchant for presenting his probability of frequency formulas in terms of combinatorial expressions. I believe this format was distracting to some, and may have prevented some significant general patterns from being noticed.

For example, in his Equation (18.24), Jaynes presents the following formula for calculating the probability of future frequency counts based on past data for the "coin tossing scenario," in other words, where the dimension of the state space is $n = 2$.

$$P(M_m | N_n) = \frac{\binom{n+m}{n} \binom{N+M-n-m}{N-n}}{\binom{N+M+1}{M}}$$

I doubt that the casual reader would perceive that this formula with its combinatorial expressions is exactly the same as my formula. The psychological difficulty is that, in not knowing that the formulas are really the same, a reader would uncritically take the position that Jaynes's derivation and mine must differ in essential aspects.

In an exercise, I show all of the details in establishing the equivalence between my formulas and Jaynes's combinatorial formulas. In the final Chapter to this Volume, I return to this topic of what actually is in Jaynes's Chapter 18 through a detailed annotation of his section 5.

22.6 Solved Exercises for Chapter Twenty Two

Exercise 22.6.1: What is the value of the partition function for the “fair” model?

Solution to Exercise 22.6.1

In this Chapter’s kangaroo scenario, the fair model assigned all $Q_i = 1/8$. This was the very first independent model discussed. The calculation of the partition function, $Z(\lambda_1, \dots, \lambda_7)$, is especially simple. Since $n = 8$ and $m = 7$, with all $\lambda_j = 0$, and adopting the convenient short-cut of $F(X = x_i) \equiv F(x_i)$,

$$\begin{aligned} Z(\lambda_1, \dots, \lambda_7) &= \sum_{i=1}^n \exp \left[\sum_{j=1}^m \lambda_j F_j(x_i) \right] \\ &= \sum_{i=1}^8 \exp \left[\sum_{j=1}^7 \lambda_j F_j(x_i) \right] \\ &= \sum_{i=1}^8 \exp [\lambda_1 F_1(x_i) + \lambda_2 F_2(x_i) + \dots + \lambda_7 F_7(x_i)] \\ &= \sum_{i=1}^8 \exp [(0 \times F_1(x_i)) + (0 \times F_2(x_i)) + \dots + (0 \times F_7(x_i))] \\ &= \sum_{i=1}^8 \exp [0] \\ &= 8 \end{aligned}$$

The denominator for all Q_i under the fair model is $Z(\lambda_1, \dots, \lambda_7) = 8$, while the numerator for each Q_i is $\exp [0] = 1$.

Exercise 22.6.2: Considering the entire model space, what models exhibit the most drastic correlations?

Solution to Exercise 22.6.2

Consider the eight models that assign $Q_i = 1$, while the remaining seven probabilities must all equal 0. For example, take the model that assigns $Q_1 = 1$ with the remaining Q_2 through Q_8 equal to 0. All inferences conducted through Bayes’s Theorem will be certainties. If a kangaroo has sandy colored fur and is right handed, then it is certain that it prefers Foster’s. Of course, under this model, kangaroos other than sandy colored, right-handed Foster’s drinkers do not exist.

Exercise 22.6.3: Suppose that the model space is restricted to the eight models just mentioned in the last exercise which exhibit extreme correlations. One data point is collected. What happens to the model space?

Solution to Exercise 22.6.3

Suppose the one data point is that a right handed beige colored kangaroo was observed to prefer Corona. Thus, $N = 1$ and $N_7 = 1$. Seven models are immediately rejected, and one model is retained. The one model that was kept assigned $Q_7 = 1$ and all other $Q_i = 0$; the seven models that were rejected assigned $Q_i = 1$ for the i^{th} joint statement other than, “The kangaroo prefers Corona beer and is right handed and has beige colored fur.”

Exercise 22.6.4: What happens if a second observation is made that confirms the existence of a left handed kangaroo with sandy colored fur who prefers Foster’s?

Solution to Exercise 22.6.4

The IP is in trouble. The original model space consisted of just eight models. Seven of these were eliminated by the first data point, and only one model remained. However, the second data point is not compatible with the one remaining model. The moral is clear. It behooves the IP to start out with *all* conceivable models in the model space, and then let the data winnow out the supported models. The IP never finds itself in the precarious position of having too few models.

Exercise 22.6.5: What does the probability for any number of future frequency counts given some past data reduce down to when the IP is interested in just the next trial?

Solution to Exercise 22.6.5

Suppose that the IP is interested in the probability for one future occurrence of the i^{th} statement when conditioned on some past data, $\mathcal{D} = \{N_1, N_2, \dots, N_n\}$.

$$P(M_1 = 0, M_2 = 0, \dots, M_i = 1, \dots, M_n = 0 \mid \mathcal{D}) =$$

$$\frac{M! (N + n - 1)!}{N_1! N_2! \cdots N_n! (M + N + n - 1)!} \times \frac{\prod_{i=1}^n (M_i + N_i)!}{\prod_{i=1}^n M_i!}$$

When $M_i = 1$ and $M = 1$, this equation simplifies to,

$$\begin{aligned} P(M_1 = 0, M_2 = 0, \dots, M_i = 1, \dots, M_n = 0 | \mathcal{D}) \\ = \frac{(N + n - 1)!}{N_1! N_2! \dots N_i! \dots N_n! (N + n)!} \times N_1! N_2! \dots N_{i-1}! N_i + 1! \dots N_n! \\ = \frac{N_i + 1}{N + n} \end{aligned}$$

For example, in the final section the probability for the next kangaroo to be a right handed, sandy fur colored Foster's drinker (cell 1 in the joint statement space) was found to be,

$$P(M_1 = 1, \dots, M_8 = 0 | \mathcal{D}) = \frac{N_1 + 1}{N + n} = \frac{41}{108}$$

because the past data had $N_1 = 40$ from the total sample of $N = 100$. The dimension of the state space was $n = 8$. The probability for the next kangaroo to be a right handed, sandy fur colored Corona drinker (cell 5 in the joint statement space) was found to be,

$$P(M_1 = 0, \dots, M_5 = 1, \dots, M_8 = 0 | \mathcal{D}) = \frac{N_5 + 1}{N + n} = \frac{6}{108}$$

because the past data had $N_5 = 5$ from the total sample of $N = 100$.

Exercise 22.6.6: Consider another model of the $m = 4$ class, but now with an association between beer preference and fur color.

Solution to Exercise 22.6.6

The first model that exhibited an association between a personality trait and a physical trait was the $m = 4$ model in section 22.3.2. This model inserted information about a correlation between beer preference and handedness, with a resulting impact on the state of knowledge about beer preference.

Another model, consisting also of $m = 4$ constraint functions together with their averages defined as the parameters of the model, now replaces that original correlation with a correlation between beer preference and fur color. The constraint function for the marginal probability of BF is,

$$F_5(X = x_i) \equiv (1, 1, 0, 0, 0, 0, 0, 0)$$

The marginal probability for BF is composed of $Q_1 + Q_2$,

$$P(BF) = P(BHF) + P(B\overline{HF})$$

The constraint function average is,

$$\langle F_5 \rangle = \sum_{i=1}^8 F_5(X = x_i) Q_i = Q_1 + Q_2 = 0.50$$

Figure 22.3 is the joint probability table with all numerical assignments filled in under this model establishing a correlation between beer preference and fur color. If the information about the correlation in this model is that $\langle F_5 \rangle = 0.50$, then the numerical assignments are $Q_1 = 0.375$ and $Q_5 = 0.075$. Bayes's Theorem reports back that the state of knowledge has changed to $P(B | H, F, M_k) = 0.8333$.

		B		\bar{B}			
		H	\bar{H}	H	\bar{H}		
		Cell 1	Cell 2	Cell 5	Cell 6		
F	H	.3750	.1250	.0750	.0250	.10	.60
	\bar{F}	.1875	.0625	.1125	.0375	.15	
		.5625	.1875	.75	.1875	.0625	.25
				.75			
						Universal	1.00

Figure 22.3: The joint probability table under a model implementing a correlation between beer preference and fur color (double boxed). The three marginal probabilities (single boxed) are also included in the model.

Exercise 22.6.7: Discuss in detail a model with information from $m = 5$ constraint functions.

Solution to Exercise 22.6.7

The correlational models of the $m = 4$ class contained information about all three “main effects” and one “double interaction.” A model from the $m = 5$ class would contain information about the three “main effects” as before, but would consider correlations involving *two* double interactions.

Pick the BH and the BF marginal probabilities as the information about some relationship between beer preference and handedness, together with information about beer preference and fur color. We have already specified what the constraint functions look like for both of these double interactions. It is only a matter of deciding what information the constraint function averages are going to convey.

Of course, the exponential expression in the numerator of the MEP formula for determining the Q_i is now going to include five terms,

$$Q_i = \frac{\exp [\lambda_1 F_1(x_i) + \lambda_2 F_2(x_i) + \lambda_3 F_3(x_i) + \lambda_4 F_4(x_i) + \lambda_5 F_5(x_i)]}{Z(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)}$$

If the constraint function averages are chosen to be $\langle F_4 \rangle = 0.58$ and $\langle F_5 \rangle = 0.50$, with the first three constraint functions averages remaining at their pre-set values of 0.75, 0.75 and 0.60, then the algorithm finds the numerical assignments under this particular model shown in the joint probability table below in Figure 22.4.

		B		\bar{B}			
		H	\bar{H}	H	\bar{H}		
		Cell 1 F	Cell 2 \bar{F}	Cell 5 F	Cell 6 \bar{F}		
		.3867	.1133	.0680	.0320		
		.1933	.0567	.1020	.0480		
		.58	.17	.75	.17	.08	.25
				.75			
						Universal	1.00

Figure 22.4: *The joint probability table under a model implementing correlations between beer preference and hand preference as well as between beer preference and fur color.*

With the strength of the relationship between beer preference, hand preference, and fur color established under this model, the IP's state of knowledge about the preference for Foster's when it is already known that the kangaroo is right-handed and sandy colored, is,

$$P(B | H, F, \mathcal{M}_k) = \frac{Q_1}{Q_1 + Q_5} = \frac{0.3867}{0.3867 + 0.0680} = 0.8505$$

Exercise 22.6.8: Show that Jaynes's combinatorial expressions for the probability of future frequency counts and my formula are exactly the same.

Solution to Exercise 22.6.8

First, as a valuable refresher, recapitulate the argument for finding the probability of future frequency counts based on past data where the dimension of the state space is only $n = 2$.

The notation for the number of *future* frequency counts of the first statement in the state space is M_1 , and M_2 is the notation for the number of *future* frequency counts of the second statement in the state space. M is the total number of future frequency counts where, of course, $M_1 + M_2 = M$.

In like manner, N_1 is the notation for the number of *past* frequency counts of the first statement in the state space, and N_2 is the notation for the number of *past* frequency counts of the second statement in the state space. Together, $N_1 + N_2 = N$ where N is the total number of data points.

Write the probability $P(M_1, M_2, \mathcal{M}_k, N_1, N_2)$ over the entire complement of joint statements. We are taking advantage of probability's **Commutativity** and **Associativity** properties to order the statements in such a way so that they can be decomposed according to the **Product Rule** as,

$$P(M_1, M_2, \mathcal{M}_k, N_1, N_2) = P(M_1, M_2 | \mathcal{M}_k, N_1, N_2) \times P(\mathcal{M}_k | N_1, N_2) \times P(N_1, N_2)$$

The probability for the future frequency counts depends only on the model \mathcal{M}_k making the assignments and is independent of any past data N_1, N_2 . The previous equation can thus be simplified to,

$$P(M_1, M_2, \mathcal{M}_k, N_1, N_2) = P(M_1, M_2 | \mathcal{M}_k) \times P(\mathcal{M}_k | N_1, N_2) \times P(N_1, N_2)$$

Then use the **Sum Rule** to marginalize over all the models in model space,

$$P(M_1, M_2, N_1, N_2) = \int_0^1 P(M_1, M_2 | \mathcal{M}_k) \times P(\mathcal{M}_k | N_1, N_2) \times P(N_1, N_2) dq$$

Since $n = 2$, we have just the single integration over q , the numerical assignment to the probability for the first statement in the state space. Call upon a third formal manipulation rule, **Bayes's Theorem**, to see that,

$$\begin{aligned} P(M_1, M_2 | N_1, N_2) &= \frac{P(M_1, M_2, N_1, N_2)}{P(N_1, N_2)} \\ &= \int_0^1 P(M_1, M_2 | \mathcal{M}_k) \times P(\mathcal{M}_k | N_1, N_2) dq \end{aligned}$$

Focus on the second term under the integration. This is $P(\mathcal{M}_k | \mathcal{D})$. We relied on Bayes's Theorem to convert this into the stalwart ubiquitous Bayesian equation,

$$P(\mathcal{M}_k | \mathcal{D}) = \frac{P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k)}{P(\mathcal{D})}$$

We remind the reader that we have extensively discussed the reason why,

$$P(\mathcal{M}_k) = 1 \text{ and } P(N_1, N_2) \equiv P(\mathcal{D}) = \frac{1}{N+1}$$

when the IP is “totally ignorant” about the causes involved in the binary event. This was the repercussion of taking $P(\mathcal{M}_k)$ to have a “flat” distribution. This is the step that Bayes took, the step that Laplace took, the step that Jaynes took, and the step that I take as well. See the next exercise for more detail on $P(\mathcal{M}_k)$.

So now we have the current situation reflected in,

$$\begin{aligned} P(M_1, M_2 | N_1, N_2) &= \int_0^1 P(M_1, M_2 | \mathcal{M}_k) \times \frac{P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k)}{P(\mathcal{D})} dq \\ &= \int_0^1 P(M_1, M_2 | \mathcal{M}_k) (N+1) P(\mathcal{D} | \mathcal{M}_k) dq \end{aligned}$$

Substitute for the third term $P(\mathcal{D} | \mathcal{M}_k)$. This is the binomial distribution when the data are conditioned on a given model making assignments to q and $(1-q)$, and where we see the multiplicity factor $W(N)$ make its appearance.

$$P(M_1, M_2 | N_1, N_2) = \int_0^1 P(M_1, M_2 | \mathcal{M}_k) (N+1) W(N) q^{N_1} (1-q)^{N_2} dq$$

The first term follows the same pattern, and the second multiplicity factor $W(M)$ shows up,

$$P(M_1, M_2 | N_1, N_2) = \int_0^1 W(M) q^{M_1} (1-q)^{M_2} (N+1) W(N) q^{N_1} (1-q)^{N_2} dq$$

Pull out the three terms from under the integral that do not depend on q ,

$$P(M_1, M_2 | N_1, N_2) = W(M) W(N) (N+1) \int_0^1 q^{M_1} (1-q)^{M_2} q^{N_1} (1-q)^{N_2} dq$$

Add the exponents in the multiplication of the q and $(1-q)$,

$$P(M_1, M_2 | N_1, N_2) = W(M) W(N) (N+1) \int_0^1 q^{M_1+N_1} (1-q)^{M_2+N_2} dq$$

Here is where we acknowledge that aforementioned debt to those who found an analytical solution for the integral,

$$\int_0^1 q^{M_1+N_1} (1-q)^{M_2+N_2} dq = \frac{(M_1+N_1)! (M_2+N_2)!}{(M_1+N_1+M_2+N_2+1)!}$$

This is the integration involving the *beta integral* appearing often in probability theory. Jaynes and mathematicians refer to it as an Eulerian integral of the first kind after the renowned 18th Century Swiss mathematician Leonhard Euler. After making the substitution $M_1 + M_2 = M$ and $N_1 + N_2 = N$, we have,

$$P(M_1, M_2 | N_1, N_2) = W(M) W(N) (N+1) \frac{(M_1+N_1)! (M_2+N_2)!}{(M+N+1)!}$$

Substitute the factorial expressions for both multiplicity factors,

$$P(M_1, M_2 | N_1, N_2) = \frac{M!}{M_1! M_2!} \frac{N!}{N_1! N_2!} (N+1) \frac{(M_1 + N_1)! (M_2 + N_2)!}{(M+N+1)!}$$

and then carry out the multiplication by $(N+1)$,

$$P(M_1, M_2 | N_1, N_2) = \frac{M!}{M_1! M_2!} \frac{(N+1)!}{N_1! N_2!} \frac{(M_1 + N_1)! (M_2 + N_2)!}{(M+N+1)!}$$

Finally, collect the constant terms together into C , and then use the product symbol for the multiplication involved wherever the M_i occur,

$$P(M_1, M_2 | N_1, N_2) = C \times \frac{\prod_{i=1}^2 (M_i + N_i)!}{\prod_{i=1}^2 M_i!}$$

The constant term C involves everything that is not subject to change when finding the probability $P(M_1, M_2 | N_1, N_2)$,

$$C = \frac{M! (N+1)!}{N_1! N_2! (M+N+1)!}$$

Jaynes's wrote this same probability as just derived as his Equation (18.24) [23],

$$P(M_m | N_n) = \frac{\binom{n+m}{n} \binom{N+M-n-m}{N-n}}{\binom{N+M+1}{M}}$$

where his n corresponds to my N_1 and his m to my M_1 . The N and the M are the same for both of us. The combinatorial expression written in the $\binom{x}{y}$ notation is expressed with factorials as,

$$\binom{x}{y} = \frac{x!}{(x-y)! y!}$$

Take each such combinatorial expression in Jaynes's Equation (18.24), the two in the numerator and the one in the denominator, and convert them to factorial form,

$$\begin{aligned}
\binom{n+m}{n} &= \binom{N_1 + M_1}{N_1} \\
&= \frac{(M_1 + N_1)!}{N_1! M_1!} \\
\binom{N+M-n-m}{N-n} &= \binom{N+M-N_1-M_1}{N-N_1} \\
&= \frac{(N+M-N_1-M_1)!}{(N-N_1)! (M-M_1)!} \\
&= \frac{(N_2 + M_2)!}{N_2! M_2!} \\
\binom{n+m}{n} \binom{N+M-n-m}{N-n} &= \frac{(M_1 + N_1)!}{N_1! M_1!} \frac{(N_2 + M_2)!}{N_2! M_2!} \\
\binom{N+M+1}{M} &= \frac{(M+N+1)!}{M! (N+1)!} \\
\frac{\binom{n+m}{n} \binom{N+M-n-m}{N-n}}{\binom{N+M+1}{M}} &= \frac{\frac{(M_1 + N_1)!}{N_1! M_1!} \frac{(N_2 + M_2)!}{N_2! M_2!}}{\frac{(M+N+1)!}{M! (N+1)!}} \\
&= \frac{M! (N+1)! (M_1 + N_1)! (M_2 + N_2)!}{N_1! M_1! N_2! M_2! (M+N+1)!}
\end{aligned}$$

Now just re-arrange this expression so that it corresponds with my final expression,

$$\begin{aligned}
P(M_1, M_2 | N_1, N_2) &= \frac{M!}{M_1! M_2!} \frac{(N+1)!}{N_1! N_2!} \frac{(M_1 + N_1)! (M_2 + N_2)!}{(M+N+1)!} \\
&= C \times \frac{\prod_{i=1}^2 (M_i + N_i)!}{\prod_{i=1}^2 M_i!} \\
&= \frac{M! (N+1)!}{N_1! N_2! (M+N+1)!} \times \frac{\prod_{i=1}^2 (M_i + N_i)!}{\prod_{i=1}^2 M_i!}
\end{aligned}$$

It took us a long time, but we eventually reached our objective. We proved that Jaynes's final formula, using his combinatorial expressions, for predicting future frequency counts when conditioned on past frequency counts was exactly the same as my formula using the factorials. And, repeating myself, that equality came about because we both used the uniform distribution over model space.

Exercise 22.6.9: Use the formal manipulation rules in a slightly different manner to reach the same conclusion as in the previous exercise.

Solution to Exercise 22.6.9

Write the expression for the full joint probability of the future and past frequency counts together with the models in a different order as $P(M_1, M_2, N_1, N_2, \mathcal{M}_k)$. The two joint probability expressions must be equivalent to each other,

$$P(M_1, M_2, \mathcal{M}_k, N_1, N_2) \equiv P(M_1, M_2, N_1, N_2, \mathcal{M}_k)$$

Use the **Product Rule** to change this latter joint expression into a conditional expression,

$$P(M_1, M_2, N_1, N_2, \mathcal{M}_k) = P(M_1, M_2 | \mathcal{M}_k) P(N_1, N_2 | \mathcal{M}_k) P(\mathcal{M}_k)$$

Substituting the binomial distributions as dependent on a model \mathcal{M}_k , we have,

$$P(M_1, M_2, N_1, N_2, \mathcal{M}_k) = W(M) q^{M_1} (1-q)^{M_2} W(N) q^{N_1} (1-q)^{N_2} P(\mathcal{M}_k)$$

Insert the Dirichlet distribution with parameters $\alpha - 1$ and $\beta - 1$ for $P(\mathcal{M}_k)$,

$$\begin{aligned} P(M_1, M_2, N_1, N_2, \mathcal{M}_k) &= \\ W(M) q^{M_1} (1-q)^{M_2} W(N) q^{N_1} (1-q)^{N_2} C_{Beta} q^{\alpha-1} (1-q)^{\beta-1} \end{aligned}$$

Collect the terms into,

$$\begin{aligned} P(M_1, M_2, N_1, N_2, \mathcal{M}_k) &= \\ W(M) \times W(N) \times C_{Beta} \times q^{M_1+N_1+\alpha-1} (1-q)^{M_2+N_2+\beta-1} \end{aligned}$$

Marginalize over the models,

$$\begin{aligned} P(M_1, M_2, N_1, N_2) &= \\ \int_0^1 W(M) \times W(N) \times C_{Beta} \times q^{M_1+N_1+\alpha-1} (1-q)^{M_2+N_2+\beta-1} dq &= \\ W(M) \times W(N) \times C_{Beta} \times \int_0^1 q^{M_1+N_1+\alpha-1} (1-q)^{M_2+N_2+\beta-1} dq &= \\ W(M) \times W(N) \times C_{Beta} \times \frac{\Gamma(M_1 + N_1 + \alpha) \Gamma(M_2 + N_2 + \beta)}{\Gamma(M_1 + N_1 + \alpha + M_2 + N_2 + \beta)} \end{aligned}$$

To implement the uniform distribution over model space, the two parameters in the Dirichlet distribution (Beta distribution) are set to $\alpha = 1$ and $\beta = 1$, leading to,

$$\begin{aligned} P(M_1, M_2, N_1, N_2) &= \\ W(M) \times W(N) \times C_{Beta} \times \frac{(M_1 + N_1)! (M_2 + N_2)!}{\Gamma(M + N + 2)} &= \\ W(M) \times W(N) \times C_{Beta} \times \frac{(M_1 + N_1)! (M_2 + N_2)!}{(M + N + 1)!} \end{aligned}$$

The constant term for the Beta distribution implementing the uniform distribution over model space is,

$$C_{Beta} = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} = \frac{\Gamma(2)}{\Gamma(1) \Gamma(1)} = \frac{1!}{0! 0!} = 1$$

leaving us with,

$$P(M_1, M_2, N_1, N_2) = W(M) \times W(N) \times \frac{(M_1 + N_1)! \times (M_2 + N_2)!}{(M + N + 1)!}$$

The expression on the left hand side is the numerator in Bayes's Theorem,

$$P(M_1, M_2 | N_1, N_2) = \frac{P(M_1, M_2, N_1, N_2)}{P(N_1, N_2)}$$

The denominator is $P(N_1, N_2) \equiv P(\mathcal{D})$. It can be found using exactly the same techniques just shown for the denominator. Utilizing those simple, yet very powerful, formal manipulation rules we call the **Sum Rule** and the **Product Rule**, we decompose $P(\mathcal{D})$ into,

$$P(\mathcal{D}) = \int_0^1 P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k) dq$$

Substitute the binomial distribution for $P(\mathcal{D} | \mathcal{M}_k)$ and the Beta distribution for $P(\mathcal{M}_k)$ to arrive at,

$$P(N_1, N_2) = \int_0^1 W(N) q^{N_1} (1 - q)^{N_2} C_{Beta} q^{\alpha-1} (1 - q)^{\beta-1} dq$$

Perform the usual procedures of pulling out the constant terms that do not depend on q ,

$$\begin{aligned} P(N_1, N_2) &= W(N) \times C_{Beta} \times \int_0^1 q^{N_1+\alpha-1} (1 - q)^{N_2+\beta-1} dq \\ \int_0^1 q^{N_1+\alpha-1} (1 - q)^{N_2+\beta-1} dq &= \frac{N_1! N_2!}{(N+1)!} \\ P(N_1, N_2) &= \frac{N!}{N_1! N_2!} \times 1 \times \frac{N_1! N_2!}{(N+1)!} \\ &= \frac{1}{N+1} \end{aligned}$$

Substitute this value for the denominator in Bayes's Theorem to find that,

$$\begin{aligned}
 P(M_1, M_2 | N_1, N_2) &= \frac{P(M_1, M_2, N_1, N_2)}{P(N_1, N_2)} \\
 &= \frac{W(M) W(N) \frac{(M_1+N_1)! (M_2+N_2)!}{(M+N+1)!}}{\frac{1}{N+1}} \\
 &= W(M) W(N) (N+1) \frac{(M_1+N_1)! (M_2+N_2)!}{(M+N+1)!} \\
 &= \frac{M!}{M_1! M_2!} \frac{N!}{N_1! N_2!} (N+1) \frac{(M_1+N_1)! (M_2+N_2)!}{(M+N+1)!} \\
 &= \frac{M!}{M_1! M_2!} \frac{(N+1)!}{N_1! N_2!} \frac{(M_1+N_1)! (M_2+N_2)!}{(M+N+1)!} \\
 &= \frac{M! (N+1)!}{N_1! N_2! (M+N+1)!} \times \frac{\prod_{i=1}^2 (M_i + N_i)!}{\prod_{i=1}^2 M_i!}
 \end{aligned}$$

Having arrived at this expression, we can easily see that we are right back to the same expression for the probability of the future frequency counts as derived through a slightly different application of the formal rules in the previous exercise.

The probability for general n is,

$$P(M_1, M_2, \dots, M_n | N_1, N_2, \dots, N_n) =$$

$$\frac{M! (N+n-1)!}{N_1! N_2! \dots N_n! (M+N+n-1)!} \times \frac{\prod_{i=1}^n (M_i + N_i)!}{\prod_{i=1}^n M_i!}$$

See Jaynes's Equation (18.42) or Equation (18.43) [23] for a confirmation of this general formula after a suitable re-arrangement and change in notation.

Chapter 23

Logistic Regression

23.1 Introduction

Logistic regression is a well-known and widely used data analytic procedure. It seems to have found a prominent place in the tool kit of social scientists, and enjoys a particular popularity in medical research. The manufactured example used in this Chapter will draw its inspiration from curiosity about whether coronary heart disease may depend on age and the presence of a C-reactive protein.

Despite logistic regression's popularity, a compelling rationale for its equations is lacking in the conventional explanations. Like all orthodox statistical equations, formulas appear like my veritable *deus ex machina* to the surprise and befuddlement of the user.

However, if one approaches logistic regression from a combined Bayesian and Maximum Entropy Principle viewpoint, the rationale for the equations is relatively simple and direct. That perspective is presented here and proceeds in two major steps.

First, Bayes's Theorem is used to express the desired inference. Our proverbial information processor wishes to acquire a state of knowledge concerning a binary variable, the presence or absence of coronary heart disease, given knowledge of some number of predictor variables, here the age of the patient and the result of a lab test concerning the C-reactive protein. The terms in Bayes's Theorem are then just rearranged in order to prepare for the logistic equation.

Second, numerical values are assigned to the *joint probabilities* that appear in Bayes's theorem by conditioning on some model. All such models insert information via Jaynes's Maximum Entropy Principle. The logistic regression equation then appears after a few more simple steps.

A detailed numerical example accompanies the theoretical exposition. A very valuable feature of this approach will be the demonstration that a combined Bayesian and MEP approach is able to reproduce the *maximum likelihood* result returned by all conventional statistical software programs as *the* solution to this typical logistic regression problem.

The other goal of this effort is to debunk some common misconceptions about the role of this combined Bayesian/MEP approach to inferencing. Stated quite bluntly, this ignorance manifests itself by denying that such an easy, straightforward solution to the logistic regression problem actually exists. We shall attempt to make some modest headway against the headwinds of such ignorance by demonstrating that exactly the opposite is true.

23.2 Setting Up the Problem Conventionally

As mentioned in the **Introduction**, logistic regression is a data analysis technique employed mainly by social scientists, economists, and medical researchers. It also appears in a somewhat opaque fashion in the explanation of neural networks and nonlinear classification. In all of these disciplines, logistic regression is presented as a way to find the probability of some binary variable as a function of some number of predictor variables. The predictor variables may be continuous or discrete in nature.

Our example will start by examining the conventional presentation for, say, the probability of the presence of coronary heart disease (CHD). Such a probability is conventionally modeled by a logistic regression equation as,

$$P(\text{CHD}) = \frac{1}{1 + \exp(-Y)} \quad (23.1)$$

where Y is the regression equation containing the parameters and the predictor variables. Equation (23.1) is called the “logistic sigmoid” function.

The regression equation Y looks like this,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q \quad (23.2)$$

where the β_i are the regression parameters, and the X_i are the coding variables for the discrete values taken on by the predictor variables.

For the sake of an easy illustrative example, let’s suppose that there are two predictor variables, the patient’s age and a result from a laboratory test. In the standard treatment, no rationale is given as to why the particular distribution in Equation (23.1) should be used, or why Equation (23.2) should be involved. Invariably, no explanation is proffered as to their origins. As is typical of the conventional approach, Equations (23.1) and (23.2) are presented out of the blue as a *fait accompli* unsupported by any underlying fundamental principles.

Coronary heart disease (CHD), as the variable of interest, must take on only two values, present or absent, so that logistic regression applies. Suppose that the first predictor variable, the patient's age (AGE) is a discrete variable with three categories, less than 40, between 40 and 60, or over 60. The second predictor variable, the result of the lab test on the amount of C-reactive protein (TEST), is also a discrete variable with three categories, Low, Medium, or High.

With this specification of the problem, the regression equation looks like,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \quad (23.3)$$

A conventional statistical software program solving this logistic regression problem, (the data will be given later on), will return the maximum likelihood estimates, $\hat{\beta}_j$, for the regression parameters as,

$$\hat{\beta}_0 = -0.620$$

$$\hat{\beta}_1 = -0.723$$

$$\hat{\beta}_2 = -0.505$$

$$\hat{\beta}_3 = -0.740$$

$$\hat{\beta}_4 = -0.522$$

The instructions for the software program will tell you to code the X_i as "dummy" coding variables in order to represent the predictor variables of AGE and TEST. This is just one more example of the lack of transparency inherent in the conventional treatment. We try to avoid all such mysteries in our preferred approach. These coding variables are listed in Table 23.1 below.

Table 23.1: The "dummy" coding variables used for the two predictor variables in a conventional logistic regression program.

AGE		
X_1	X_2	Levels
1	0	UNDER 40
0	1	40 TO 60
0	0	OVER 60

TEST		
X_3	X_4	Levels
1	0	LOW
0	1	MEDIUM
0	0	HIGH

For the particular settings of the predictor variables of UNDER 40 and MEDIUM, that is, $X_1 = 1$, $X_2 = 0$ for AGE, and $X_3 = 0$, $X_4 = 1$ for TEST,

$$\begin{aligned} Y &= -0.620 + (-0.723 \times 1) + (-0.505 \times 0) + (-0.740 \times 0) + (-0.522 \times 1) \\ &= -1.865 \end{aligned}$$

Thus, the probability for the presence of coronary heart disease as a function of a patient who is under 40 years of age with a lab result showing a medium level of the C-reactive protein is,

$$\begin{aligned} P(\text{CHD}) &= \frac{1}{1 + \exp(-Y)} \\ &= \frac{1}{1 + \exp(1.865)} \\ &= 0.1341 \end{aligned}$$

We have an answer, but do we understand how we arrived at that answer? I am afraid that we do not.

To wrap up this section on the conventional set up, let's see if the probabilities for coronary heart disease move in the expected direction. Take the two extremes where we would expect the lowest probability and the highest probability. We would expect the lowest probability of coronary heart disease for a patient younger than 40 years of age with a test result of Low. For this case, $X_1 = 1$, $X_2 = 0$ for AGE, and $X_3 = 1$, $X_4 = 0$ for TEST,

$$\begin{aligned} Y &= -0.620 + (-0.723 \times 1) + (-0.505 \times 0) + (-0.740 \times 1) + (-0.522 \times 0) \\ &= -2.083 \end{aligned}$$

yielding a probability of,

$$\begin{aligned} P(\text{CHD}) &= \frac{1}{1 + \exp(-Y)} \\ &= \frac{1}{1 + \exp(2.083)} \\ &= 0.1108 \end{aligned}$$

Now examine the other extreme where we would expect the highest probability. We would expect the highest probability of coronary heart disease for a patient older than 60 years of age with a test result of High. For this case, $X_1 = 0$, $X_2 = 0$ for AGE, and $X_3 = 0$, $X_4 = 0$ for TEST,

$$\begin{aligned} Y &= -0.620 + (-0.723 \times 0) + (-0.505 \times 0) + (-0.740 \times 0) + (-0.522 \times 0) \\ &= -0.620 \end{aligned}$$

yielding a probability of,

$$\begin{aligned} P(\text{CHD}) &= \frac{1}{1 + \exp(-Y)} \\ &= \frac{1}{1 + \exp(0.620)} \\ &= 0.3498 \end{aligned}$$

23.3 Setting Up the Problem via Bayes and MEP

An easy explanation exists for the origin and derivation of logistic regression if one approaches the problem from a Bayesian and MEP perspective. Divide up the problem into two stages.

The first stage applies the formal manipulation rules of probability theory. In other words, we set up the problem for a solution by Bayes's Theorem. The second stage follows by assigning numerical values to the probabilities according to the MEP formalism. The MEP assumes the existence of a model that inserts some desired information in order to arrive at the numerical values.

Before we begin, we must construct the state space. This is the space of all *joint statements* as defined in the problem. A probability operator is wrapped around any statement in the state space to indicate an information processor's degree of belief that the statement is true.

For example, the first joint statement in the state space is, "The patient has coronary heart disease, is under 40 years old, and received a low test result." The second joint statement in the state space is, "The patient has coronary heart disease, is between the ages of 40 and 60, and received a low test result.", while the final joint statement in the state space is, "The patient does not have coronary heart disease, is over 60 years old, and received a high test result." The dimension of the state space is based on $n = 2 \times 3 \times 3 = 18$ joint statements.

In subsequent sections, both a joint probability table and a contingency table will be presented. Given the dimension of the state space, both of these tables will consist of $n = 18$ cells.

In the first stage, the product and sum rules are called on to manipulate the probability for the joint statements so that the desired probability for CHD can be calculated as a function of the patient's age and lab results. Both of these predictor variables are assumed known.

This, of course, results in the simplest form of Bayes's Theorem,

$$P(\text{CHD} | \text{AGE, TEST}) = \frac{P(\text{CHD, AGE, TEST})}{P(\text{AGE, TEST})} \quad (23.4)$$

Since the criterion variable CHD consists of only two values, the denominator in Equation (23.4) gets expanded by the **Sum Rule** into the sum of two joint probabilities,

$$P(\text{CHD=Present} \mid \text{AGE, TEST}) = \frac{P(\text{CHD=Present, AGE, TEST})}{P(\text{CHD=Present, AGE, TEST}) + P(\text{CHD=Absent, AGE, TEST})} \quad (23.5)$$

With this result we have reached the end of the formal manipulation stage. We choose to leave the probabilities on the right hand side of Equation (23.5) as joint probabilities rather than transforming them to the typical *likelihood times prior* format one usually sees in the presentation of Bayes's Theorem. This is because the MEP algorithm makes numerical assignments to *probabilities of joint statements*.

The rules of probability theory have done their job. We accept at this first level that the answer given in Equation (23.5) must be correct.

However, Bayes's Theorem could care less what numerical values are inserted for the joint probabilities on the right hand side of Equation (23.5) as long as they are legitimate probabilities. It will handle all such legitimate probabilities with equal aplomb. Data analysis can proceed no further unless there is some method for assigning numerical values to the joint probabilities. This is where the MEP enters the picture.

As explained by Jaynes, the MEP is a disciplined method for inserting desired information into a probability distribution such that only the information explicitly mentioned is included, while all other unwanted information is excluded. Numerical values are assigned by a model employing constraint functions and the averages of those functions.

Many models ranging from the simple to the complex might be proposed as tentative working hypotheses for this numerical assignment. There can be no one true probability assignment, but only assignments conditioned on the information inserted by the tentatively entertained models. The worth of these models will ultimately be judged by how closely they predict the actual data that were obtained.

Once we have an MEP model that assigns numerical values, (call it model \mathcal{M}_k and then insert it as supposedly true to the right of the conditioned upon symbol), we can write Bayes's Theorem for the logistic regression set up in Equation (23.5) as,

$$P(\text{CHD} \mid \text{AGE, TEST}, \mathcal{M}_k) = \frac{Q_i}{Q_i + Q_j} \quad (23.6)$$

where Q_i and Q_j are a convenient short notation for the joint probabilities appearing on the right hand side of Bayes's Theorem. This notation indicates that Q_i and Q_j have definite numerical values arising from the MEP.

The MEP provides a formula for computing the numerical values for Q_i and Q_j as,

$$Q_i = \frac{\exp [\lambda_1 F_1(x_i) + \lambda_2 F_2(x_i) + \cdots + \lambda_m F_m(x_i)]}{Z(\lambda_1, \lambda_2, \dots, \lambda_m)} \quad (23.7)$$

and,

$$Q_j = \frac{\exp [\lambda_1 F_1(x_j) + \lambda_2 F_2(x_j) + \cdots + \lambda_m F_m(x_j)]}{Z(\lambda_1, \lambda_2, \dots, \lambda_m)} \quad (23.8)$$

Bayes's Theorem can be conveniently rearranged in preparation for the logistic regression by dividing the numerator and denominator of Equation (23.6) by Q_i .

$$P(\text{CHD} | \text{AGE}, \text{TEST}, \mathcal{M}_k) = \frac{1}{1 + \frac{Q_j}{Q_i}} \quad (23.9)$$

The division of Q_j by Q_i simplifies to,

$$\frac{Q_j}{Q_i} = \exp(-Y) \quad (23.10)$$

where Y is,

$$Y = \sum_{l=1}^m \lambda_l [F_l(x_i) - F_l(x_j)] \quad (23.11)$$

The regression parameters are identified with the Lagrange multipliers, and the coding variables with the difference in constraint functions. Thus, in the final step,

$$P(\text{CHD} | \text{AGE}, \text{TEST}, \mathcal{M}_k) = \frac{1}{1 + \exp(-Y)} \quad (23.12)$$

and the equivalence with the logistic regression equation is proved.

A straightforward application of the **Sum and Product Rules** for finding the updated probability of a binary variable conditioned on some number of predictor variables, followed by a numerical assignment according to an MEP model, has led to an easy derivation and explanation of the logistic regression equation. Since these are the underlying fundamental principles we would apply for *any* inferencing problem, we see that logistic regression just happens to be a specific example of a general data analysis procedure.

Only the Y term requires some further discussion because of its involvement with the Lagrange multipliers λ_l and constraint functions $F_l(x_i)$ and $F_l(x_j)$. These can be matched directly to the regression parameters and coding variables in the standard treatment of logistic regression. A numerical example is presented next that shows this correspondence in detail.

We then update the probabilities for all the MEP models that were proposed by finding out how well the various models fitted the observed data. The final step is to make a prediction about the probability of coronary heart disease for some new patient based on the patient's known age, lab results, and the data from some previous sample of patients.

23.4 Numerical Example

In working out a numerical example, it helps to orient yourself to the locations of the assigned numerical values Q_i within a joint probability table. See Figure 23.1 for a sketch of such a joint probability table with the appropriate Q_i filled in for each cell.

Present			Absent				
	Under 40	40 to 60	Over 60		Under 40	40 to 60	Over 60
Low	Q_1	Q_2	Q_3	Low	Q_{10}	Q_{11}	Q_{12}
Med	Q_4	Q_5	Q_6	Med	Q_{13}	Q_{14}	Q_{15}
High	Q_7	Q_8	Q_9	High	Q_{16}	Q_{17}	Q_{18}

Figure 23.1: A $2 \times 3 \times 3$ joint probability table for the logistic regression example. The Q_i notation stands for numerical values assigned to the joint probabilities by an MEP model.

As already mentioned, there are two levels for CHD, Present or Absent, three levels for the first predictor variable of AGE, Under 40, 40 to 60, or Over 60, and three levels for the second predictor variable of TEST, Low, Medium, or High. There are thus a total of $n = 18$ cells in the joint probability table.

A Q_i value is placed into each one of these 18 cells to indicate the numerical value of the joint probability indexed by that cell. For example, Q_4 is the numerical value as assigned by some MEP model \mathcal{M}_k for the probability of the joint occurrence of CHD=Present, AGE=Under 40, and TEST=Medium.

We want to find the probability for the presence of coronary heart disease given knowledge of a C-reactive protein test and the age of the patient. Suppose that, just like the first example under the conventional approach, the particular patient we are interested in obtained a Medium score on the diagnostic test and is under 40 years of age. By specifying some MEP model, we can assign numerical values to all 18 Q_i values.

Bayes's Theorem then tell us that,

$$\begin{aligned} P(\text{CHD}= \text{Present} \mid \text{AGE}= \text{Under 40}, \text{TEST}= \text{Medium}, \mathcal{M}_k) &= \frac{Q_4}{Q_4 + Q_{13}} \\ &= \frac{1}{1 + Q_{13}/Q_4} \end{aligned}$$

For the sake of the numerical example to follow, suppose that the model \mathcal{M}_k consists of nine constraint functions $F_1(x_i)$ through $F_9(x_i)$ and their associated constraint

function averages. As we shall discover in subsequent sections, this particular model inserts information about the marginal probabilities for CHD, AGE, and TEST. It also inserts information about two double interactions, the first between CHD and AGE; the second between CHD and TEST. This is equivalent to tentatively entertaining a model where both AGE and TEST influence the probability for CHD.

From Equation (23.11), and the fact that we are using a model where $m = 9$, we know that,

$$\begin{aligned} Y &= \sum_{l=1}^9 \lambda_l [F_l(x_4) - F_l(x_{13})] \\ &= \lambda_1 [F_1(x_4) - F_1(x_{13})] + \cdots + \lambda_9 [F_9(x_4) - F_9(x_{13})] \\ &= \lambda_1 (1 - 0) + \lambda_2 (1 - 1) + \cdots + \lambda_6 (1 - 0) + \cdots + \lambda_9 (1 - 0) \\ &= \lambda_1 + \lambda_6 + \lambda_9 \end{aligned}$$

The functional values of each $F_l(x_i)$ were pulled from Table 23.2 as shown in the next section. The values of the Lagrange multipliers are found by the MEP algorithm software to be,

$$\lambda_1 = -0.620$$

$$\lambda_6 = -0.723$$

$$\lambda_9 = -0.522$$

From this result and Equation (23.10),

$$\begin{aligned} \lambda_1 + \lambda_6 + \lambda_9 &= -1.865 \\ \frac{Q_{13}}{Q_4} &= \exp(-Y) \\ &= \exp(1.865) \end{aligned}$$

Therefore, by Equation (23.12) the probability of coronary heart disease for this patient is,

$$\begin{aligned} P(\text{CHD}=Present \mid \text{AGE}=Under 40, \text{TEST}=Medium, \mathcal{M}_k) &= \frac{1}{1 + \exp(1.865)} \\ &= 0.1341 \end{aligned}$$

which is exactly the same result from the conventional analysis, but now we know the rationale for how we arrived at this answer.

To seal the deal, we'll look at one more example from the Bayesian and MEP approach. In the previous section, the conventional maximum likelihood solution found that the highest probability for CHD occurred when the patient had a High level of C-reactive protein and was over 60 years old. We know that the proper way of making the correct inference is to set up Bayes's Theorem as,

$$\begin{aligned} P(\text{CHD}= \text{Present} \mid \text{AGE}= \text{Over 60}, \text{TEST}= \text{High}, \mathcal{M}_k) &= \frac{Q_9}{Q_9 + Q_{18}} \\ &= \frac{1}{1 + Q_{18}/Q_9} \end{aligned}$$

Since we are using the same model to assign numerical values to these new joint probabilities, we have,

$$\begin{aligned} Y &= \sum_{l=1}^9 \lambda_l [F_l(x_9) - F_l(x_{18})] \\ &= \lambda_1 [F_1(x_9) - F_1(x_{18})] + \dots + \lambda_9 [F_9(x_9) - F_9(x_{18})] \\ &= \lambda_1 (1 - 0) \\ &= \lambda_1 \end{aligned}$$

All nine constraint functions had a value of 0 for $F_l(x_i)$ with the one exception of $F_1(x_9) = 1$. The probability of coronary heart disease for this new patient is then,

$$\begin{aligned} P(\text{CHD}= \text{Present} \mid \text{AGE}= \text{Over 60}, \text{TEST}= \text{High}, \mathcal{M}_k) &= \frac{1}{1 + \exp(0.620)} \\ &= 0.3498 \end{aligned}$$

which matches the result found previously.

23.5 Details of the MEP Models

What kind of information does the MEP insert into a joint probability distribution via some model? In this section, we closely examine all of the details.

Since the dimension of the state space has been defined as $n = 18$, there can be a maximum of $m = 17$ constraint functions and their associated averages. Some of the models are quite simple, while others can express complicated relationships among all the variables.

Our main goal is to form models that can implement relationships (associations, correlations) between the criterion variable of CHD and the two predictor variables

of AGE and TEST. In the numerical example of the last section, the model illustrated there implemented a relationship between CHD and both predictor variables.

Tables 23.2 and 23.3, shown over the next two pages, list the values of all 17 constraint functions, $F_1(x_i)$ through $F_{17}(x_i)$, as defined for this logistic regression problem. An average value of each constraint function is given in the bottom row of each table as the information specified by some model. Together with the universal constraint, these 17 constraint functions are a just-determined decomposition of the joint probability over the space of the 18 cells.

So, in specifying an MEP model \mathcal{M}_k that assigns numerical values to the 18 cells of joint probability table, the m appearing in Equations (23.7), (23.8), and (23.11) could range from $m = 0$ to $m = n - 1 = 17$. Whenever $m < n - 1$, we have an underdetermined problem together with an inherent ambiguity in how to assign values. The MEP serves the very important role of resolving this ambiguity about the assignment.

The MEP inserts the information from these m constraint functions, and the accompanying constraint function averages specified by the model into a joint probability distribution. *It inserts no other extraneous information; only that information explicitly identified.* This is the insurance policy that we take out when we rely on the MEP to assign numerical values.

For ease of presentation, let A stand for the binary criterion variable of CHD, let B stand for the first predictor variable of AGE, and let C stand for the second predictor variable of TEST. The constraint functions $F_1(x_i)$ through $F_{17}(x_i)$ code for the marginal probabilities of the A , B , and C variables, as well as for all possible double interactions, AB , AC , and BC , which are also finer grained marginal probabilities, and finally for the one triple interaction, ABC . These marginal probabilities and single cell probabilities constitute the $m = 17$ constraint function averages.

As an example, look at the second column of Table 23.2 showing the constraint function $F_1(x_i)$ where i runs from $i = 1$ to 18. This codes for the marginal probability of A , or CHD. Thus, $F_1(x_i) = 1$ for $i = 1$ to 9 and $F_1(x_i) = 0$ for $i = 10$ to 18. The average of this first constraint function is specified by a model as $\langle F_1 \rangle = 0.20$.

Therefore,

$$\sum_{i=1}^{18} F_1(x_i) Q_i = \langle F_1 \rangle$$

$$Q_1 + Q_2 + \cdots + Q_9 = 0.20$$

and the marginal probability for A is inserted by this model as the one piece of information (along with, of course, the universal constraint) into a joint distribution. Refer back to Figure 23.1 to see that $\sum_{i=1}^9 Q_i$ is the marginal probability for the first level of variable A , the presence of coronary heart disease. The probability for the second level of A for this particular model is automatically determined by the sum rule as 0.80.

Table 23.2: *The first nine of the seventeen constraint functions for the $2 \times 3 \times 3$ joint probability table. These are the marginal probability constraints for A, B, and C, together with the constraints for the AB and AC interactions.*

Cell <i>i</i>	<i>A</i>	<i>B</i> ₁	<i>B</i> ₂	<i>C</i> ₁	<i>C</i> ₂	<i>AB</i> ₁	<i>AB</i> ₂	<i>AC</i> ₁	<i>AC</i> ₂
	<i>F</i> ₁	<i>F</i> ₂	<i>F</i> ₃	<i>F</i> ₄	<i>F</i> ₅	<i>F</i> ₆	<i>F</i> ₇	<i>F</i> ₈	<i>F</i> ₉
1	1	1	0	1	0	1	0	1	0
2	1	0	1	1	0	0	1	1	0
3	1	0	0	1	0	0	0	1	0
4	1	1	0	0	1	1	0	0	1
5	1	0	1	0	1	0	1	0	1
6	1	0	0	0	1	0	0	0	1
7	1	1	0	0	0	1	0	0	0
8	1	0	1	0	0	0	1	0	0
9	1	0	0	0	0	0	0	0	0
10	0	1	0	1	0	0	0	0	0
11	0	0	1	1	0	0	0	0	0
12	0	0	0	1	0	0	0	0	0
13	0	1	0	0	1	0	0	0	0
14	0	0	1	0	1	0	0	0	0
15	0	0	0	0	1	0	0	0	0
16	0	1	0	0	0	0	0	0	0
17	0	0	1	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0
$\langle F_l \rangle$	0.20	0.20	0.50	0.33	0.33	0.03	0.09	0.05	0.06

A similar explanation holds for the constraint functions $F_2(x_i)$ through $F_5(x_i)$. These code for the marginal probabilities of *B* and *C*, or AGE and TEST. The only difference is that since AGE and TEST exist at three levels, two constraint functions are required for each of these predictor variables. By constructing the constraint functions $F_1(x_i)$ through $F_5(x_i)$ and their associated averages, we have generated models that insert information into a joint probability distribution about the marginal probabilities for *A*, *B*, and *C*, or CHD, AGE, and TEST.

However, it turns out that all such models that use $F_1(x_i)$ through $F_5(x_i)$ exhibit independence between the criterion variable *A* and the predictor variables *B* and *C*. We have not built any constraint functions that allow for a relationship between *A* and *B* or *A* and *C*, so that the probability for *A* can change when conditioned on *B* and/or *C*.

By constructing constraint functions $F_6(x_i)$ and $F_7(x_i)$, we specify models that allow for such an association between *A* and *B*, and functions $F_8(x_i)$ and $F_9(x_i)$

Table 23.3: *The last eight of the seventeen constraints for the $2 \times 3 \times 3$ joint probability table. These are the BC and ABC interactions.*

Cell <i>i</i>	B_1C_1 F_{10}	B_1C_2 F_{11}	B_2C_1 F_{12}	B_2C_2 F_{13}	AB_1C_1 F_{14}	AB_1C_2 F_{15}	AB_2C_1 F_{16}	AB_2C_2 F_{17}
1	1	0	0	0	1	0	0	0
2	0	0	1	0	0	0	1	0
3	0	0	0	0	0	0	0	0
4	0	1	0	0	0	1	0	0
5	0	0	0	1	0	0	0	1
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0
10	1	0	0	0	0	0	0	0
11	0	0	1	0	0	0	0	0
12	0	0	0	0	0	0	0	0
13	0	1	0	0	0	0	0	0
14	0	0	0	1	0	0	0	0
15	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0
$\langle F_l \rangle$	0.068	0.067	0.167	0.167	0.008	0.009	0.022	0.027

do the same for A and C . The constraint functions $F_{10}(x_i)$ through $F_{13}(x_i)$ permit models to be considered that also express an association between B and C . This takes care of all the double interactions. Finally, the last four constraint functions and their averages, $F_{14}(x_i)$ through $F_{17}(x_i)$, capture a relationship among all three variables at once, or what we call the triple interaction ABC .

The interaction constraints are found by multiplying the separate constraints that make up the interaction. For example, the AB_1 interaction, $F_6(x_i)$, is found from the separate A and B constraint functions as in $F_6(x_i) = F_1(x_i) \times F_2(x_i)$. Referring once again to the actual values of $F_6(x_i)$ in Table 2, we see that,

$$\sum_{i=1}^{18} F_6(x_i) Q_i = \langle F_6 \rangle$$

$$Q_1 + Q_4 + Q_7 = 0.03$$

This is the marginal probability for CHD=Present and AGE=Under 40 over all three levels of TEST.

Remember that the constraint function averages as shown in Tables 23.2 and 23.3 implement specific models with specific information. Certainly it is true that other models could be specified that set, say, $\langle F_1 \rangle = 0.25$ or $\langle F_{17} \rangle = 0.03$. Different numerical assignments to all 18 joint probabilities would then ensue under the changed information in these different models.

The reason that these particular values as shown in Tables 23.2 and 23.3 were chosen as the constraint function averages is because they match the actual data that were observed. This is the explanation for why the Bayesian/MEP procedure could duplicate the classical maximum likelihood solution.

Nonetheless, the most important lesson in this entire Chapter is:

**INFORMATION IS NOT DATA
AND
DATA IS NOT INFORMATION**

Information, as defined within the MEP, consists of the constraint functions constructed over the n statements in the state space, taken together with the average (probabilistic expectation) of those constraint functions. Strictly speaking, this definition of information **has absolutely nothing to do with any data that might subsequently be recorded**. So, do not be confused by the fact that the information in some models (mathematical expectations of constraint functions) are allowed to match normed frequency counts from actual data (as they do in this illustrative example).

The **ONLY** role (albeit an extremely important role) of data within the Bayesian approach is to update the relative status of all the models. This is seen to be a concept that is completely independent from how numerical assignments were made to all n statements in the state space from ONE particular model (the MEP process).

That is the reason why I constantly try to explain to people that the relationship between the Bayesian and MEP approaches is one that is both complementary and orthogonal. The manipulation rules, like Bayes's Theorem, take care of telling us how the probability for models must be updated when conditioned on data. It doesn't care one iota about the particular numerical values for probabilities.

On the other hand, the MEP algorithm doesn't know anything about the formal manipulation rules of probability. It only knows about the information resident in models, and how this information dictates the resulting numerical values to assign to joint probabilities. It isn't aware of and doesn't care about the probability for models, or the probability for data when it is carrying out this task of numerical assignment.

The next section presents the (artificial) data for this logistic regression problem. It discusses the orthogonal exercise of how models get updated when conditioned on the known data.

23.6 The Data

No mention has yet been made of any actual data. This is because data have not been required for any part of the argument. The probability of CHD can be calculated quite readily by relying solely on some feasible model for the numerical assignment of the joint probabilities. We have seen these calculations in the previous sections.

Those who experience some discomfort at seeing the data appear so late on the scene, harbor, I suspect, a lingering attachment to probability defined as a frequency. When the data are not around to anchor their true feelings about a probability, they start to become uncomfortable.

The only role of the data, as stressed in the last section, is to update the probabilities for all the models under consideration. To that end, Figure 23.2 presents some frequency counts placed into the cells of the contingency table from an hypothetical experiment.

Present			Absent						
Under 40	40 to 60	Over 60	Under 40	40 to 60	Over 60	50	60	90	
Low	8	22	20	60	145	78	283	333	
Med	9	27	24	58	140	75	273	333	
High	13	41	36	52	125	67	244	334	
	30	90	80	170	410	220	800		
			200	500	300			1,000	

Figure 23.2: *Observed frequency counts from an hypothetical experiment involving $N = 1,000$ patients. Each patient was placed into one and only one of the 18 cells of the contingency table depending upon the joint occurrence of CHD, AGE, and TEST.*

$N = 1,000$ patients were examined for the presence of coronary heart disease. Their age and results of the diagnostic test for C-reactive protein were recorded as well. Each patient was classified into one, and only one, of the 18 cells of the table. There are $N_1 = 8$ patients in cell 1, $N_2 = 22$ patients in cell 2, . . . , and $N_{18} = 67$ patients in the final cell. The total frequency count over all eighteen cells of the contingency table must, of course, equal 1,000, or $\sum_{i=1}^n N_i = N$.

Various marginal totals are also shown. For example, the sum of the frequency counts over the first nine cells is equal to 200. These are the 200 patients who had coronary heart disease. This will impact the main effect of CHD as implemented by $F_1(x_i)$. The sum over the first three cells, 50, is where the interaction of CHD with the Low level of TEST takes place. These data will impact models with the AC_1 double interaction as implemented by the constraint function $F_8(x_i)$.

These data will be given the notation of \mathcal{D} with the acknowledgment that there is no uncertainty attached to the counts in the cells when these have been accurately recorded. These are the same data that were used as input to the conventional statistical software solving for the maximum likelihood solution.

We can now discern the origin of the mysterious constraint function averages, $\langle F_l \rangle$, that appeared in the models. For example, the average for the first constraint function was set by some model at $\langle F_1 \rangle = 0.20$ as shown in the bottom row of Table 23.2. The marginal frequency count for the presence of CHD was 200, so this particular constraint average selected by the model happened to be the same as the normed frequency of $200/1000 = 0.20$.

Other models, as mentioned earlier, might very well choose to set $\langle F_1 \rangle = 0.19$ or $\langle F_1 \rangle = 0.21$, or any other feasible value for that matter. But if we set $\langle F_1 \rangle$ to the marginal frequency of CHD as revealed by the actual empirical data, we are doing within the MEP the same thing as what the conventional approach does when it picks a maximum likelihood estimate.

Jaynes had demonstrated this very interesting fact about maximum likelihood (but ignored as usual) in [18, pp. 270–271]. This also clears up the coincidence that the Lagrange multipliers and the maximum likelihood estimates of the regression parameters had the same values, as was discovered in the numerical example.

We have derived and extensively reviewed the formula for finding the probability of the *next* occurrence of a joint statement when conditioned on some amount of previous data. The most important conceptual part of that derivation was the fact that an integration was carried out over all conceivable numerical assignments in model space.

We apply that formula here to find the probability of CHD conditioned on AGE and TEST when *all* models have been taken into account, and not conditioned on just *one* model as in all previous examples. It turns out to be an amazingly simple prescription. Skipping over all of the details, and using the generic notation,

$$P(A_{N+1} | B_{N+1}, C_{N+1}, \mathcal{D}) = \frac{N_i + 1}{N_i + 1 + N_j + 1} \quad (23.13)$$

Applying Equation (23.13) to one of our numerical examples, we can find the probability that the *next* patient has CHD when it is known that the patient is under 40 years old, and has a medium C-reactive protein level. Such an inference would be based as well on the results from all 1,000 previous patients. This probability would be written out in full as,

$$P(\text{CHD}_{N+1} = \text{Present} | \text{AGE}_{N+1} = \text{Under 40}, \text{TEST}_{N+1} = \text{Medium}, \text{All 1000 data points})$$

The relevant frequency counts are then $N_i = N_4 = 9$ and $N_j = N_{13} = 58$. The probability that the next patient has CHD given these predictor variables and given,

as well, all of the past data is,

$$\begin{aligned}
 P(\text{CHD}_{N+1} | \text{AGE}_{N+1}, \text{TEST}_{N+1}, \mathcal{D}) &= \frac{N_4 + 1}{N_4 + 1 + N_{13} + 1} \\
 &= \frac{10}{10 + 59} \\
 &= 0.1449
 \end{aligned}$$

This is just a little higher than the value of 0.1341 as calculated under the one model that inserted information about the two double interactions, the first between CHD and AGE, and the second between CHD and TEST with $m = 9$. By including the contribution from every single model, the above probability is the correct one.

Obviously, the overwhelming majority of models are not supported by the data and they contribute almost nothing to the final outcome. There are, however, some more complicated models including the triple interactions which are somewhat supported by the data. When these are included they do tend to raise the probability slightly from 0.1341 to 0.1449.

The important point to be made here, heeding our earlier mantra, is that the DATA served to reorient model space. It had absolutely nothing to do with any initial assignment of numerical values based on the INFORMATION inserted by any one individual model.

**INFORMATION IS NOT DATA
AND
DATA IS NOT INFORMATION**

The seeds of confusion have been sown by those people who, either consciously or unconsciously, choose to confound the DATA of normed frequency counts with the INFORMATION from constraint function averages. The constraint function averages only happen to match these normed frequency counts because we wanted to duplicate the conventional maximum likelihood solution to logistic regression. Moreover, if we do focus attention on that model that matches the data, it still is only *one* model. In the end, that special maximum likelihood solution gets averaged with all of the other models.

The maximum likelihood solution is the same as the one INFORMATION model that is most strongly supported by the DATA. As the amount of data becomes larger and larger, then the correct Bayesian/MEP solution and the maximum likelihood solution will eventually converge to the same solution.

23.7 Connections to the Literature

I would just like to point out that the approach outlined here is simple and straightforward when compared to the unnecessarily complicated treatment usually given in our standard textbooks. For example, consider the discussion in [10, pp. 82–86] of an introductory logistic regression involving a bioassay problem. They begin a journey extending over five pages into esoteric areas that completely miss the main point of a Bayesian data analysis.

The inference in the bioassay problem has to do with the dosage of a drug and its effect on animal mortality. There are four increasing dosage levels administered to a total of 20 animals. Each dosage level is administered to five animals, and the number of animals that died in response to that amount of drug is recorded. It turned out from this experiment that: no animals died at the lowest dose, one animal died at the next higher level, three animals died at the next higher level, and finally all five animals died at the highest drug dosage.

What is the probability that the *next* animal will die when exposed to, say, the strongest dose of the drug? The answer is 6/7. That answer is found in the following standard fashion we have been promoting all along.

The state space has dimension $n = 8$ because the only observable statements are the eight joint statements, “The animal died at the lowest dosage.” through “The animal lived at the highest dosage.” There are data from the experiment with $N = 20$ with $N_1 = 0$, (no animals died at the lowest dosage) through $N_7 = 5$ (all animals died at the highest dosage), with, of course, $N_8 = 0$.

The inference centers on the degree of belief that the next animal exposed to any dosage level will die. Thus, $M = 1$, and since the problem states that we are interested in the probability of dying at the highest drug dose, we will have to eventually set $M_7 = 1$ and $M_8 = 1$. This follows from Bayes’s Theorem,

$$P(A = \text{Dead} \mid B = \text{Highest dose}, \mathcal{D}) = \frac{P(M_7 = 1 \mid \mathcal{D})}{P(M_7 = 1 \mid \mathcal{D}) + P(M_8 = 1 \mid \mathcal{D})}$$

Apply the familiar predictive formula for the probability of the future event $M_7 = 1$ when conditioned on one causal factor of dosage level with all of the past data explicitly listed,

$$P(M_1 = 0, \dots, M_7 = 1, M_8 = 0 \mid N_1 = 0, N_2 = 5, N_3 = 1, N_4 = 4, N_5 = 3, N_6 = 2, N_7 = 5, N_8 = 0)$$

Because $M = 1$, (see Exercise 22.6.5), the predictive formula reduces to provide our answer,

$$\begin{aligned}
P(M_1 = 0, \dots, M_7 = 1, M_8 = 0 | \mathcal{D}) &= \frac{N_7 + 1}{N + n} \\
&= \frac{6}{28} \\
P(M_1 = 0, \dots, M_7 = 0, M_8 = 1 | \mathcal{D}) &= \frac{N_8 + 1}{N + n} \\
&= \frac{1}{28} \\
P(A = \text{Dead} | B = \text{Highest dose}, \mathcal{D}) &= \frac{\frac{6}{28}}{\frac{6}{28} + \frac{1}{28}} \\
&= \frac{6}{7}
\end{aligned}$$

Here is another glaring example of how Bayesian and MEP concepts get mangled. I quote from the author of a well-regarded textbook.¹

We now turn to a Bayesian treatment of logistic regression. Exact Bayesian inference for logistic regression is intractable. In particular, evaluation of the posterior distribution would require normalization of the product of a prior distribution and a likelihood function that itself comprises a product of logistic sigmoid functions, one for every data point. Evaluation of the predictive distribution is similarly intractable.

Well, indeed! I hope to have given the reader in this Chapter some sense of just how “intractable” my combined Bayesian and MEP approach was for logistic regression. The techniques that Laplace introduced 200 years ago apparently are not good enough for modern sensibilities.

¹Christopher M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006, pg. 217.

23.8 Solved Exercises for Chapter Twenty Three

Exercise 23.8.1: What does a model incorporating just the main effects look like?

Solution to Exercise 23.8.1

A model for the main effects will incorporate information about the marginal probabilities for presence of coronary heart disease, the under 40 and the 40 to 60 age year groups, and the Low and Medium test results. This information is contained in the constraint function averages $\langle F_1 \rangle$ through $\langle F_5 \rangle$. Thus, $m = 5$ and the argument to the exponential in the numerator of the MEP formula will be $\sum_{j=1}^5 \lambda_j F_j(X = x_i)$.

Exercise 23.8.2: What does a model incorporating all of the main effects with just the one double interaction of CHD with TEST look like?

Solution to Exercise 23.8.2

The main effects information stays the same as discussed in the first exercise. The double interaction of CHD with TEST is represented by the AC_1 and AC_2 interactions. Thus, the constraint function averages of $\langle F_8 \rangle$ and $\langle F_9 \rangle$ are added to the constraint function averages $\langle F_1 \rangle$ through $\langle F_5 \rangle$. Thus, $m = 7$ for this class of models.

Exercise 23.8.3: What is an IP's state of knowledge about the presence of cardiac heart disease given knowledge that a patient's age is 65 and the test returned a result classified as medium? Present the answer in the logistic regression format.

Solution to Exercise 23.8.3

This exercise is merely a refresher of the basic concepts and notation as introduced in this Chapter. The IP's state of knowledge is calculated from Bayes's Theorem. In terms of the generic statements A , B , and C , the formal manipulation rules state that the updated state of knowledge about statement A being true given that both statements B and C are true, is,

$$P(A | BC) = \frac{P(ABC)}{P(BC)} = \frac{P(ABC)}{P(ABC) + P(\overline{ABC})}$$

This kind of notation is applicable when a statement has only two possible measurements. If, as is the situation here, two of the statements can assume three possible measurements, then to be strictly consistent we should break down each one of these statements further. It easiest to visualize this as a tree structure as presented in Figure 23.3 at the top of the next page.

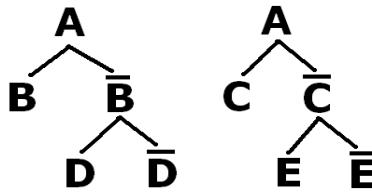


Figure 23.3: Tree structure breakdown for the generic statement notation in the logistic regression example.

Statement A concerning the presence of coronary heart disease needs no change since there are only two possible measurements. However, statements B and C have three possible measurements. If statement B reads, “Patient is under 40 years of age.”, then \bar{B} covers the cases where the patient is not under 40 years old, or, the two remaining possible measurements for AGE, label them as statement D , “Patient is between 40 and 60 years old.” and \bar{D} , “Patient is over 60 years old.” Likewise, if statement C reads, “Patient has Low test result.”, then \bar{C} covers the cases where the patient did not receive a Low test result, or, the two remaining possible measurements for TEST, label them as statement E , “Patient has Medium test result.” and \bar{E} , “Patient has High test result.”

In our current inferential scenario, this becomes,

$$P(A\bar{D}E) \equiv P(\text{CHD=Present}, \text{AGE=Over 60}, \text{TEST=Medium})$$

$$P(\bar{A}\bar{D}E) \equiv P(\text{CHD=Absent}, \text{AGE=Over 60}, \text{TEST=Medium})$$

When numerical assignments are made to these probabilities by some model, then they are represented by Q_i and Q_j . Given the way that the joint probability table has been laid out, $i = 6$ and $j = 15$.

$$P(\text{CHD=Present}, \text{AGE=Over 60}, \text{TEST=Medium} | \mathcal{M}_k) = Q_6$$

$$P(\text{CHD=Absent}, \text{AGE=Over 60}, \text{TEST=Medium} | \mathcal{M}_k) = Q_{15}$$

For the purposes of a logistic regression, Bayes's Theorem now looks like,

$$P(\text{CHD=Present} | \text{AGE=Over 60}, \text{TEST=Medium}, \mathcal{M}_k) = \frac{Q_6}{Q_6 + Q_{15}}$$

$$P(\text{CHD=Present} | \text{AGE=Over 60}, \text{TEST=Medium}, \mathcal{M}_k) = \frac{1}{1 + \frac{Q_{15}}{Q_6}}$$

Exercise 23.8.4: Show how the generic notation gets translated into the correct number of cases.

Solution to Exercise 23.8.4

We know that there are nine possible Q_i values that start with $P(A\star\star)$. The other nine are $P(\bar{A}\star\star)$.

$$\begin{aligned} P(ABC) &\rightarrow P(ABC) \\ P(AB\bar{C}) &\rightarrow P(ABE) \\ P(AB\bar{C}) &\rightarrow P(AB\bar{E}) \\ P(A\bar{B}C) &\rightarrow P(ADC) \\ P(A\bar{B}C) &\rightarrow P(A\bar{D}C) \\ P(A\bar{B}\bar{C}) &\rightarrow P(ADE) \\ P(A\bar{B}\bar{C}) &\rightarrow P(A\bar{D}E) \\ P(A\bar{B}\bar{C}) &\rightarrow P(A\bar{D}\bar{E}) \\ P(A\bar{B}\bar{C}) &\rightarrow P(A\bar{D}\bar{E}) \end{aligned}$$

Take, for example, $P(A\bar{B}C)$ which is the probability for the joint statement that a patient has coronary heart disease, is not under 40 years old, and has a Low test result. This breaks down into the probabilities, $P(ADC)$ and $P(A\bar{D}C)$. $P(ADC)$ is the probability for the joint statement that a patient has coronary heart disease, is between the ages of 40 to 60, and has a Low test result. $P(A\bar{D}C)$ is the probability for the joint statement that a patient has coronary heart disease, is over the age of 60, and has a Low test result.

Exercise 23.8.5: Supposing that the model in Exercise 23.8.2 were in effect, what does the probability in Exercise 23.8.3 work out to?

Solution to Exercise 23.8.5

The model examined in Exercise 23.8.2 included the information from both the main effects and a single double interaction involving CHD and TEST. The information in this model will determine the numerical assignments to Q_6 and Q_{15} .

Rather than going through Equation (23.11), take advantage of the heuristic available in Table 23.2. To find Q_6 , read across the row for $i = 6$ under the columns of $\langle F_1 \rangle$ through $\langle F_5 \rangle$ (main effects) plus $\langle F_8 \rangle$ and $\langle F_9 \rangle$ (CHD by TEST interaction).

The 1s and 0s in the table indicate that the numerator of Q_6 is $e^{\lambda_1 + \lambda_5 + \lambda_9}$. Doing the same thing for Q_{15} , the numerator of Q_{15} is e^{λ_5} .

Thus,

$$\begin{aligned} P(\text{CHD=Present} \mid \text{AGE=Over 60, TEST=Medium}, \mathcal{M}_k) &= \frac{1}{1 + \frac{Q_{15}}{Q_6}} \\ \frac{Q_{15}}{Q_6} &= e^{-\lambda_1 - \lambda_9} \\ P(\text{CHD=Present} \mid \text{AGE=Over 60, TEST=Medium}, \mathcal{M}_k) &= \frac{1}{1 + e^{-\lambda_1 - \lambda_9}} \\ &= \frac{1}{1 + e^{0.620 + 0.522}} \\ &= 0.24 \end{aligned}$$

As a variation on the same theme, where do Q_8 and Q_{17} come into play? Since Q_8 is the numerical assignment under this model to the probability for the joint statement that a patient has coronary heart disease, is between the ages of 40 and 60, and has a High test result, we seem to be looking for the probability that a patient has coronary heart disease given that the patient possesses these particular characteristics. The logistic regression version of Bayes's Theorem is then,

$$\begin{aligned} P(\text{CHD=Present} \mid \text{AGE}=40 \text{ to } 60, \text{TEST}=High, \mathcal{M}_k) &= \frac{1}{1 + \frac{Q_{17}}{Q_8}} \\ \frac{Q_{17}}{Q_8} &= e^{-\lambda_1} \\ P(\text{CHD=Present} \mid \text{AGE}=40 \text{ to } 60, \text{TEST}=High, \mathcal{M}_k) &= \frac{1}{1 + e^{-\lambda_1}} \\ &= \frac{1}{1 + e^{0.620}} \\ &= 0.3498 \end{aligned}$$

Using the heuristic short cut, scan across the relevant columns under this model for row $i = 8$. λ_1 and λ_3 will be in the numerator for Q_8 . Recall that F_6 and F_7 are not in the model. Scanning across the same relevant columns under this model for row $i = 17$, λ_3 will be in the numerator for Q_{17} . The subtraction dictated by the division yields the above result.

Exercise 23.8.6: Prepare a table listing Q_i for the presence of coronary heart disease together with Q_{i+9} for the absence of coronary heart disease. What is the pattern in the respective numerators for the model incorporating the information from just the main effects?

Solution to Exercise 23.8.6

Table 23.4 below lists the Q_i and Q_{i+9} together with their respective numerators. For clarity, only the Lagrange multiplier arguments to the exponential are shown. The final column shows that the probability for coronary heart disease will always be 0.3498 for all settings of the explanatory variables under this main effects model.

$$P(\text{CHD} | \text{AGE}, \text{TEST}, \mathcal{M}_k) = \frac{1}{1 + \exp(-\lambda_1)} = 0.3498$$

Table 23.4: *The pattern of Lagrange multipliers in the numerators of the assigned numerical values to the two relevant probabilities in Bayes's Theorem under a main effects model.*

Q_i	Numerator	Q_{i+9}	Numerator	Division
Q_1	$\lambda_1 + \lambda_2 + \lambda_4$	Q_{10}	$\lambda_2 + \lambda_4$	$-\lambda_1$
Q_2	$\lambda_1 + \lambda_3 + \lambda_4$	Q_{11}	$\lambda_3 + \lambda_4$	$-\lambda_1$
Q_3	$\lambda_1 + \lambda_4$	Q_{12}	λ_4	$-\lambda_1$
Q_4	$\lambda_1 + \lambda_2 + \lambda_5$	Q_{13}	$\lambda_2 + \lambda_5$	$-\lambda_1$
Q_5	$\lambda_1 + \lambda_3 + \lambda_5$	Q_{14}	$\lambda_3 + \lambda_5$	$-\lambda_1$
Q_6	$\lambda_1 + \lambda_5$	Q_{15}	λ_5	$-\lambda_1$
Q_7	$\lambda_1 + \lambda_2$	Q_{16}	λ_2	$-\lambda_1$
Q_8	$\lambda_1 + \lambda_3$	Q_{17}	λ_3	$-\lambda_1$
Q_9	λ_1	Q_{18}	0	$-\lambda_1$

Exercise 23.8.7: Do models that include information about the main effects together with double interactions of test and age impact the probability of coronary heart disease?

Solution to Exercise 23.8.7

All such models leave the updated probability of coronary heart disease at 0.3498. A model would have to include some interaction of coronary heart disease with one or more of the explanatory variables in order for the probability to change from 0.3498.

Suppose we pick a model with information from $\langle F_1 \rangle$ through $\langle F_5 \rangle$ (main effects) plus the information from $\langle F_{10} \rangle$ through $\langle F_{13} \rangle$ (AGE by TEST interactions). This model has $m = 9$ constraint functions when compared to the $m = 5$ constraint functions of the main effects model. Let the explanatory variables assume values of AGE=Under 40 and TEST=High. We need the numerical values Q_7 and Q_{16} .

To find Q_7 , read across the row for $i = 7$ under the columns of F_1 through F_5 (main effects) plus F_{10} through F_{13} (AGE by TEST interactions). The 1s and 0s indicate that the numerator of Q_7 is $e^{\lambda_1 + \lambda_2}$. Doing the same thing for Q_{16} , we find that the numerator of Q_{16} is e^{λ_2} .

Thus,

$$\begin{aligned} P(\text{CHD}=Present \mid \text{AGE}=Under 40, \text{TEST}=High, \mathcal{M}_k) &= \frac{1}{1 + \frac{Q_{16}}{Q_7}} \\ \frac{Q_{16}}{Q_7} &= e^{-\lambda_1} \\ P(\text{CHD}=Present \mid \text{AGE}=Under 40, \text{TEST}=High, \mathcal{M}_k) &= \frac{1}{1 + e^{-\lambda_1}} \\ &= \frac{1}{1 + e^{0.620}} \\ &= 0.3498 \end{aligned}$$

This little example illustrates the level of causal detail that can be squeezed from various models.

Exercise 23.8.8: Expand a little on the MEP aspects of the bioassay scenario in section 23.7.

Solution to Exercise 23.8.8

After the black-box answer of 6/7 has been found, examination of some MEP models would be called for. A generic joint probability table composed of eight cells together with the marginal probabilities for death and dosage level under some model might look like this. The actual data are also shown in parentheses.

	Dead	Alive	
DL1	Cell 1 Q_1 (0)	Cell 2 Q_2 (5)	$Q_1 + Q_2$ (5)
DL2	Cell 3 Q_3 (1)	Cell 4 Q_4 (4)	$Q_3 + Q_4$ (5)
DL3	Cell 5 Q_5 (3)	Cell 6 Q_6 (2)	$Q_5 + Q_6$ (5)
DL4	Cell 7 Q_7 (5)	Cell 8 Q_8 (0)	$Q_7 + Q_8$ (5)
	$Q_1 + Q_3 + Q_5 + Q_7$	$Q_2 + Q_4 + Q_6 + Q_8$	

The marginal probabilities for dosage levels are constrained to reflect the experimental fact that an equal number of animals were exposed at each dosage level. The information for lethality might be to set the marginal probabilities equal.

But models with information about just the marginal probabilities for death and dosage level are not interesting because they don't include any information about the interaction between dosage and death. So we can imagine examining MEP models with the constraint function averages $\langle F_1 \rangle$ through $\langle F_4 \rangle$ to reflect the above marginal probabilities for death and dosage, but where information in the form of constraint function averages $\langle F_5 \rangle$ through $\langle F_7 \rangle$ must also be brought into play to capture *interactions* between dosage and death. Thus, set $\langle F_5 \rangle = Q_1$, $\langle F_6 \rangle = Q_3$, and $\langle F_7 \rangle = Q_5$. Imagine, then, one model with the following information (heavily influenced by having already seen the solution in section 23.7),

$$\langle F_1 \rangle = 0.50$$

$$\langle F_2 \rangle = 0.25$$

$$\langle F_3 \rangle = 0.25$$

$$\langle F_4 \rangle = 0.25$$

$$\langle F_5 \rangle = 1/28$$

$$\langle F_6 \rangle = 3/28$$

$$\langle F_7 \rangle = 4/28$$

The MEP algorithm will find the numerical assignments, $Q_i \equiv P(X = x_i | \mathcal{M}_k)$, under this model, as shown below,

	Dead	Alive	
DL1	0.0357	0.2143	0.25
DL2	0.1071	0.1429	0.25
DL3	0.1429	0.1071	0.25
DL4	0.2143	0.0357	0.25
	0.50	0.50	

For example, $Q_7 = e^{\lambda_1} / Z = 0.2143 \approx 6/28$. The probability of death at the highest drug dosage under this particular model is,

$$P(A = \text{Death} | B = \text{DL4}, \mathcal{M}_k) = \frac{Q_7}{Q_7 + Q_8} = 0.857143$$

Compare this to the answer of $6/7 = 0.857143$ when *all* models are averaged over.

Chapter 24

The Legendre Transformation

24.1 Introduction

We have gained a certain familiarity, and dare I say also, a certain level of comfort with the MEP formula over these last several Chapters. But, at this juncture, let me pose this question: Does the MEP formula, in fact, produce an assignment with the maximum entropy?

Despite the apparent similarity of this question to the old riddle of, “Who is buried in Grant’s tomb?”, it is worthwhile to once again verify that the maximum entropy algorithm does indeed live up to its name.

Does the information entropy of the probability distribution assigned via the MEP, as calculated by our usual formula, possess the greatest possible entropy of *any* assignment that also satisfies the constraints?

To answer this question, we shall prove that the information entropy, when calculated for any arbitrarily selected assignment procedure, (that is, any assignment not emanating from the MEP algorithm), is always less than or equal to the entropy from the MEP formula. If, in fact, this arbitrary assignment happened to equal the information entropy from the MEP assignment, then that arbitrary assignment *is* the MEP assignment.

Let us state at the outset what we want to prove. We want to show that the information entropy for any arbitrary discrete probability distribution that is not an MEP assignment is less than the information entropy of the assignment provided by the MEP. For simplicity in the notation, we will restrict ourselves to the situation of one extra constraint above and beyond the universal constraint.

In the course of proving this assertion, we come across another very important distinguishing feature of the MEP formalism. This ubiquitous characteristic is called the **Legendre Transformation**.

It so happens that this concept is more difficult to explain effectively than you might expect for the disciplines of mathematics, physics, and statistical mechanics. But given the groundwork we have built up, the Legendre transformation is much easier to grasp when placed within the context of making numerical assignments via the MEP.

For Jaynes, the Legendre transformation seemed to be a way of viewing the reciprocal relationship between Shannon's information entropy when considered as a function of the constraint function averages *versus* the partition function in the MEP formula when considered as a function of the Lagrange multipliers.

For others, the emphasis seemed to be on the relationship between the role played by the constraint function averages and the Lagrange multipliers as dual coordinate functions. More specifically, the Legendre transformation provided an explicit formula for how to transform between one coordinate representation and its dual coordinate representation.

Admittedly, why a discussion about the relatively arcane role of the Legendre transformation needs to be brought up in the first place seems a rather forced issue. But it turns out that the Legendre transformation forms the basis for a very good computational algorithm for finding the values of the Lagrange multipliers. The ability to calculate the Q_i under any specified model follows immediately. In addition, the Legendre transformation algorithm will not only provide us with the all important values of the parameters of the model, it will also return at the same time the maximum value of the information entropy.

Now it will come as no surprise to the reader that my derivation here slavishly follows the one Jaynes presented. I add a bit more detail, and some explanatory remarks, but I cannot in good conscience claim anything other than copying Jaynes's exposition. My only addition of worth is to provide some *Mathematica* code that offer a clear computational definition of the Legendre transformation as it relates to implementing the MEP algorithm.

24.2 Jaynes's Proof

Start out with a state space of dimension n , and then label an arbitrary probability assignment to the statements in this state space as a_1, a_2, \dots, a_n . We eventually will want to prove that the information entropy of this assignment satisfies the following inequality,

$$H(a_i) \leq \ln Z(\lambda) - \lambda \langle F \rangle \quad (24.1)$$

Notice that we didn't condition the information entropy of the a_i on the assumed truth of some model \mathcal{M}_k as we have done when using the MEP algorithm. This was to emphasize that the a_i may not be derived from the same rationale.

To establish Equation (24.1), we rely on a bit of a mathematical trick, or, if you prefer, a lucky mathematical observation. This observation concerns an inequality relating $\ln x$ to $(1 - 1/x)$. The relationship, $\ln x \geq (1 - 1/x)$, is shown numerically in Table 24.1 below.

Table 24.1: *A numerical illustration of the fact that $\ln x \geq (1 - \frac{1}{x})$.*

x	$\ln x$	$(1 - \frac{1}{x})$	$\ln x - (1 - \frac{1}{x})$
0	<i>undefined</i>	<i>undefined</i>	<i>undefined</i>
0.1	-2.3026	-9.0000	+6.6974
0.5	-0.6931	-1.0000	+0.3069
0.9	-0.1054	-0.1111	+0.0058
1.0	0	0	0
1.1	0.0953	0.0909	+0.0044
10.0	2.3026	0.9000	+1.4026

As x moves away from 0 in the positive direction, $\ln x$ is always found to be greater than $(1 - 1/x)$ except at one value of x . That one value is $x = 1$. If we exploit this relationship within the entropy definition, we can begin the argument that the entropy of the Q_i derived from the MEP is the maximum entropy of any assignment.

Temporarily detach your thinking about numerical assignments ensuing from the MEP algorithm. As an abstract thought exercise, consider two discrete probability distributions, a_i and b_i , not (at first) connected to any Q_i . If we form the following expression involving these two probability distributions,

$$\sum_{i=1}^n a_i \ln \left(\frac{a_i}{b_i} \right)$$

then we can take advantage of the relationship just demonstrated in the table above between $\ln x$ and $(1 - 1/x)$. Let $x = (a_i/b_i)$ and write,

$$\ln \left(\frac{a_i}{b_i} \right) \geq 1 - \frac{b_i}{a_i}$$

$$\sum_{i=1}^n a_i \ln \left(\frac{a_i}{b_i} \right) \geq \sum_{i=1}^n a_i \left(1 - \frac{b_i}{a_i} \right)$$

The right hand side of this inequality happens to equal zero if the a_i and b_i are probability distributions. This is easily shown in the following few lines,

$$\begin{aligned} \sum_{i=1}^n a_i \left(1 - \frac{b_i}{a_i}\right) &= \sum_{i=1}^n \left(a_i - \frac{a_i b_i}{a_i}\right) \\ &= \sum_{i=1}^n a_i - \sum_{i=1}^n b_i \\ \sum_{i=1}^n a_i &= 1 \\ \sum_{i=1}^n b_i &= 1 \\ \sum_{i=1}^n a_i \left(1 - \frac{b_i}{a_i}\right) &= 0 \end{aligned}$$

since both a_i and b_i are probability distributions. This leaves us with,

$$\sum_{i=1}^n a_i \ln \left(\frac{a_i}{b_i}\right) \geq 0$$

The left hand side is expanded into,

$$\begin{aligned} \sum_{i=1}^n a_i (\ln a_i - \ln b_i) &\geq 0 \\ \sum_{i=1}^n a_i \ln a_i - \sum_{i=1}^n a_i \ln b_i &\geq 0 \\ \sum_{i=1}^n a_i \ln a_i &\geq \sum_{i=1}^n a_i \ln b_i \end{aligned} \tag{24.2}$$

At this juncture, let's imagine that the probability distribution, b_i , is, in fact, one that has been assigned according to the MEP. The b_i would then be calculated according to the standard MEP formula,

$$b_i = \frac{e^{\lambda F(X = x_i)}}{Z(\lambda)} \tag{24.3}$$

The $\ln b_i$ then work out to,

$$\begin{aligned}\ln b_i &= \ln \left[\frac{e^{\lambda F(X = x_i)}}{Z(\lambda)} \right] \\ &= \ln \left[e^{\lambda F(X = x_i)} \right] - \ln Z(\lambda) \\ &= \lambda F(X = x_i) - \ln Z(\lambda)\end{aligned}$$

Plugging this result for $\ln b_i$ into the right hand side of Equation (24.2) we have,

$$\begin{aligned}\sum_{i=1}^n a_i \ln a_i &\geq \sum_{i=1}^n a_i [\lambda F(X = x_i) - \ln Z(\lambda)] \\ &\geq \sum_{i=1}^n a_i \lambda F(X = x_i) - \sum_{i=1}^n a_i \ln Z(\lambda) \\ &\geq \sum_{i=1}^n a_i \lambda F(X = x_i) - \ln Z(\lambda) \sum_{i=1}^n a_i \\ &\geq \lambda \sum_{i=1}^n a_i F(X = x_i) - \ln Z(\lambda) \\ &\geq \lambda \langle F \rangle - \ln Z(\lambda)\end{aligned}$$

Keep in mind that, although we said that a_i was an arbitrary probability distribution, it still is able to produce a constraint function average. Therefore,

$$\sum_{i=1}^n a_i F(X = x_i) = \sum_{i=1}^n F(X = x_i) a_i = \langle F \rangle$$

All we have to do to form the information entropy of the arbitrary distribution a_i is to put a negative sign in front of the left hand side of the latest derivation. Because we put in the minus sign, we reverse the inequality,

$$\begin{aligned}-\sum_{i=1}^n a_i \ln a_i &\leq \ln Z(\lambda) - \lambda \langle F \rangle \\ H(a_i) &\leq H(\text{MEP assignment}) \\ H(\text{MEP assignment}) &\geq H(a_i)\end{aligned}$$

We have just proved what was stated as our original goal. The entropy of the MEP assignment is always going to be greater than, or equal to, any other arbitrary

distribution a_i . Referring back to $\ln x \geq (1 - 1/x)$, we observe that equality is achieved only when $x = 1$, that is, when $a_i = b_i$.

In the derivation above, the b_i were replaced with the MEP assignment. The MEP assignment is always going to have the largest entropy of any assignment that satisfies the information in a model. Therefore, we can safely assert that,

$$H_{max}(Q_i) = \ln Z(\lambda) - \lambda \langle F \rangle$$

or, in the general case of m constraints,

$$H_{max}(Q_i) = \ln Z(\lambda_1, \lambda_2, \dots, \lambda_m) - \sum_{j=1}^m \lambda_j \langle F_j \rangle$$

24.3 A Mathematical Definition

Having established this much in the previous section, we now want to see what relationship this probability and information entropy inspired derivation has with the **Legendre transformation**.

Here is the easiest general mathematical definition of the Legendre transformation I could find,

$$g(p) = \min_x [f(x) - xp] \quad (24.4)$$

where $g(p)$ is the Legendre transform of $f(x)$. It involves only one variable x .

The goal is to express some original function $f(x)$ in terms of a new function with a different argument. This new function with a different argument is $g(p)$. The argument of g is p and the argument of f is x . The key mathematical ingredient is a relationship involving the total differential of f with p ,

$$df = \frac{\partial f}{\partial x} dx = p(x) dx \text{ so that } p(x) = \frac{\partial f}{\partial x} \quad (24.5)$$

The idea behind the notation in Equation (24.4),

$$g(p) = \min_x [f(x) - xp]$$

is that we try to find the minimum of $f(x) - xp$ while letting x vary, but p must remain fixed.

24.3.1 Application to the MEP formalism

All this abstraction centers around reciprocal arguments in two different functions. It is easy to identify these two relevant functions as they exist within the MEP formalism. One function is the information entropy with the constraint function average as its argument, $H(\langle F \rangle)$. The other function is the log of the partition function with the Lagrange multiplier as its argument, $\ln Z(\lambda)$.

24.3.2 Pattern matching to the mathematical formula

The function $\ln Z(\lambda)$ is analogous to $f(x)$, where λ is the argument analogous to x . The function $H(\langle F \rangle)$ is analogous to $g(p)$ with $\langle F \rangle$ the argument analogous to p . The argument $\langle F \rangle$ to the function g is indeed the derivative of the original function f with respect to its argument x ,

$$p = \frac{\partial f}{\partial x} = \frac{\partial \ln Z}{\partial \lambda} = \langle F \rangle \quad (24.6)$$

The **Legendre transformation** of $f(x)$ to $g(p)$ is then,

$$g(p) = \min_x [f(x) - xp] \quad (24.7)$$

$$H(\langle F \rangle) = \min_{\lambda} [\ln Z(\lambda) - \lambda \langle F \rangle] \quad (24.8)$$

In Equation (24.8) we are minimizing the right hand side while varying λ . However, a constant value for $\langle F \rangle$ must be selected, and then stuck to during the minimization procedure. Thus, after the minimization has taken place we may say that,

$$H_{max}(\langle F \rangle) = - \sum_{i=1}^n Q_i \ln Q_i \quad (24.9)$$

is the Legendre transform of $\ln Z(\lambda)$.

24.3.3 Numerical example

I will illustrate the above development with the simplest possible numerical example to nail down the relationship between the notation given above, and the concepts involved in applying the Legendre transform as an MEP algorithm.

Consider the canonical coin flip where the state space has dimension $n = 2$. The two possible results are “HEADS” and “TAILS.” Arbitrarily set the constraint function mapping these statements to numbers as,

$$F(\text{“HEADS”}) = 1 \text{ and } F(\text{“TAILS”}) = 2 \quad (24.10)$$

Then the partition function is,

$$Z(\lambda) = \exp(\lambda) + \exp(2\lambda) \quad (24.11)$$

and the assigned numerical value to the probability for the two results under some given model \mathcal{M}_k is,

$$P(\text{HEADS} | \mathcal{M}_k) = \frac{\exp(\lambda)}{\exp(\lambda) + \exp(2\lambda)} \quad (24.12)$$

$$P(\text{TAILS} | \mathcal{M}_k) = \frac{\exp(2\lambda)}{\exp(\lambda) + \exp(2\lambda)} \quad (24.13)$$

The constraint function average $\langle F \rangle$ is,

$$\langle F \rangle = \left[1 \times \frac{\exp(\lambda)}{\exp(\lambda) + \exp(2\lambda)} \right] + \left[2 \times \frac{\exp(2\lambda)}{\exp(\lambda) + \exp(2\lambda)} \right] \quad (24.14)$$

$$= \frac{\exp(\lambda) + 2\exp(2\lambda)}{\exp(\lambda) + \exp(2\lambda)} \quad (24.15)$$

Specify that the information under model \mathcal{M}_k is $\langle F \rangle = 1.5$.

Now we are ready for the minimization procedure. Minimize the expression,

$$\ln Z(\lambda) - \lambda \langle F \rangle \quad (24.16)$$

as λ is varied and $\langle F \rangle$ is held fixed at 1.5. This expression reaches its minimum when $\lambda = 0$ at a value of 0.693147 and $Z(\lambda) = 2$. We see that the correct value of the Lagrange multiplier has been found for this model. Into the bargain, we have also computed the information entropy of the numerical assignment under this model.

The above procedure is an iterative procedure. During the course of the minimization, the Lagrange multiplier might very well have reached, say, $\lambda = 1$, with the result that $\ln Z = 2.313$ and $\lambda \langle F \rangle = 1 \times 1.5$. The value returned for the information entropy is then 0.813. Obviously at this stage, the minimum value of 0.693147 for the information entropy has not been achieved. But the minimum will eventually be achieved when λ is tried at a value of 0.

Do not confuse this minimization procedure with the fact that the returned minimum value of 0.693147 is the *maximum* information entropy for the assignment satisfying the constraint.

The constraint function average does equal 1.5 when $\lambda = 0$. Substitute this value for λ back into Equation (24.15).

$$\langle F \rangle = \frac{\exp(0) + 2\exp(2 \times 0)}{\exp(0) + \exp(2 \times 0)} = \frac{1+2}{1+1} = 1.5$$

The assigned numerical value to the probabilities for HEADS and TAILS must, of course, equal 1/2 when substituting $\lambda = 0$ back into Equations (24.12) and (24.13).

Double-check that,

$$H_{max}(\langle F \rangle = 1.5) = - \sum_{i=1}^n Q_i \ln Q_i = -[1/2 \ln(1/2) + 1/2 \ln(1/2)] = 0.693147$$

the same as found during the minimization of Equation (24.16).

24.4 Connections to the Literature

My derivation as presented in section 24.2 was taken fairly directly from Jaynes's Brandeis Lectures [16, pp. 46–47].¹ Here, Jaynes informs us that a Legendre transformation has been explicitly carried out [16, pg. 48].

The functions $[\ln Z(\lambda_1, \dots, \lambda_m)]$ and $[H(\langle F_1 \rangle, \dots, \langle F_m \rangle)]$ are equivalent in the sense that each gives full information about the probability distribution; indeed $[H_{\max}(Q_i) = \ln Z - \sum_{j=1}^m \lambda_j \langle F_j \rangle]$ is just the Legendre transformation that takes us from one representative function to the other.

Essentially the same derivation was presented in Chapter 11 of Jaynes's book [23]. Jaynes viewed this approach as complementary to the traditional derivation of the MEP formula as a variational problem. Jaynes felt that if you flipped back and forth between the variational approach and this Legendre transformation approach, all the loopholes were covered [23, pg. 358]. The variational derivation is discussed in the next Chapter.

This is the rigorous proof, which is independent of the things that might happen if we try to do it as a variational problem. The argument is, as we see, strong just where the variational argument is weak. On the other hand, this argument is weak where the variational argument is strong, because we had to pull the answer out of a hat in writing [Equation (24.3)]. We had to know the answer before we could prove it. If you have both arguments, side by side, then you have the whole story.

The treatment by Jaynes of the Legendre transformation, as repeated here, reflected his realization that traditional thermodynamic concepts existed at a higher, more abstract level as part of the whole process of making inferences. Today, the Legendre transformation still seems to be discussed from within a strictly physics orientation as part of thermodynamics. More than 50 years after Jaynes pointed out its relationship with the MEP, I have never seen it debated and discussed from that more important and more general perspective of inferencing.

The best explanation that I could find for the Legendre transformation from within its origin in thermodynamics, and thus well outside of the context taken here within the MEP formalism, is given in [12, pp. 87–91]. I borrowed their notation as well for its straightforward simplicity. After a geometrical introduction, the Legendre transform of a function $f(x)$ is defined as,

$$g(p) = f(x) - xp$$

with,

$$p = \frac{\partial f}{\partial x}$$

¹There is a minor typo in Jaynes's Equation (10) where $Z(\lambda_1, \dots, \lambda_n)$ should read as $Z(\lambda_1, \dots, \lambda_m)$.

An introductory example for the function $f(x) = x^2$ is presented, and the details of the Legendre transformation process are set out. The key insight is the recognition of the interchangeability of an inverse of the differentiation of a function with respect to one argument with the other argument. If the inverse of the partial differentiation of f with respect to p exists, then it is equal to x ,

$$x = \left(\frac{\partial f}{\partial p} \right)^{-1}$$

This opens up another path to Jaynes's elaboration of the many reciprocal, or dual, relationships within the MEP formalism.

For example, if we start with the above inverse expression as provided by the Legendre transformation, and apply it within the MEP formalism,

$$\left(\frac{\partial f}{\partial p} \right)^{-1} = x$$

$$\left(\frac{\partial \ln Z}{\partial \langle F \rangle} \right)^{-1} = \lambda$$

$$\frac{\partial H}{\partial \langle F \rangle} = \lambda$$

$$\left(\frac{\partial \ln Z}{\partial \langle F \rangle} \right)^{-1} = \frac{\partial H}{\partial \langle F \rangle}$$

Volume III will go into more detail about these reciprocal, or dual, relationships within the MEP formalism from the standpoint of *Information Geometry*.

24.5 Solved Exercises for Chapter Twenty Four

Exercise 24.5.1: Confirm numerically for a small problem that $\sum_{i=1}^n a_i \ln a_i$ is greater than $\sum_{i=1}^n a_i \ln b_i$ where the b_i represent the MEP assignment while the a_i are an arbitrary but still legitimate assignment.

Solution to Exercise 24.5.1

Choose the small state space where $n = 3$. The one constraint function defined over the statements in the state space is the vector $(1, 2, 3)$. The information inserted under a model \mathcal{M}_k is $\langle F \rangle = 2.1$. A legitimate assignment a_i might be $(0.2, 0.5, 0.3)$. The sum of these assignments is equal to 1, and the average of the constraint function, $\sum_{i=1}^3 F(X = x_i) a_i$, is,

$$\langle F \rangle = (1 \times 0.2) + (2 \times 0.5) + (3 \times 0.3) = 2.1$$

The negative value of the information entropy for the a_i assignment is,

$$\sum_{i=1}^3 a_i \ln a_i = (0.2 \ln 0.2) + (0.5 \ln 0.5) + (0.3 \ln 0.3) = -1.0297$$

Since the b_i emanate from the MEP formula, the comparison sum is,

$$\begin{aligned} \ln b_i &= \lambda \times F(X = x_i) - \ln Z \\ \sum_{i=1}^3 a_i \ln b_i &= [0.2 \times (\lambda \times 1 - \ln Z)] + [0.5 \times (\lambda \times 2 - \ln Z)] + [0.3 \times (\lambda \times 3 - \ln Z)] \\ &= 2.1\lambda - \ln Z \\ &= (2.1 \times 0.150566) - 1.40729 \\ &= -1.0911 \end{aligned}$$

The value of the Lagrange multiplier was found by the MEP algorithm to have a value of $\lambda = 0.150566$, while the log of the partition function $\ln Z(\lambda)$ was found to equal 1.40729. We have numerically confirmed that since $-1.0297 > -1.0911$, $\sum_{i=1}^n a_i \ln a_i > \sum_{i=1}^n a_i \ln b_i$.

The MEP assignment is,

$$b_i \equiv Q_i = (0.2846, 0.3308, 0.3846)$$

in contrast to the non-MEP assignment of,

$$a_i = (0.2, 0.5, 0.3)$$

This assignment also satisfies the two constraints that the assignments must sum to 1, and the average of the constraint function must equal 2.1. The vital additional feature of this b_i assignment is that it possesses greater information entropy than the a_i assignment. In fact, this MEP assignment of b_i possesses the largest possible information entropy of any assignment which satisfies the constraints.

Of course, $-\sum_{i=1}^n a_i \ln a_i = 1.0297$ is the information entropy of our arbitrary legitimate assignment. The information entropy of the MEP assignment, $-\sum_{i=1}^n b_i \ln b_i = 1.0911$, is the largest entropy attainable for any assignment that satisfies the constraint function average.

Exercise 24.5.2: Use *Mathematica* to calculate the results in the first exercise.

Solution to Exercise 24.5.2

The detailed description of the Legendre transformation's role within the MEP algorithm is presented in Appendix C. Here we pick out relevant pieces of the more general program to show the computations involved, and how we arrived at the values used in the first exercise.

Set the constraint function with $\mathbf{cf} = \{1, 2, 3\}$. The partition function Z is calculated by $\mathbf{z} = \mathbf{Total}[\mathbf{Exp}[\lambda \mathbf{cf}]]$. This partition function is correctly evaluated by *Mathematica* as,

$$Z(\lambda) = e^\lambda + e^{2\lambda} + e^{3\lambda}$$

We now harness the power of *Mathematica* to easily solve for the Legendre transformation. The built-in *Mathematica* function, $\mathbf{NMinimize}[f, x]$, will find the minimum of the function f with respect to the argument x . $\mathbf{NMinimize}[f, x]$ is used to implement the Legendre transformation,

$$H_{max}(Q_i) = \ln Z(\lambda) - \lambda \langle F \rangle$$

The function f is $\ln Z(\lambda) - \lambda \langle F \rangle$ and the argument that is varying during the minimization procedure is λ . The argument that is fixed at 2.1 is $\langle F \rangle$. Thus, we ask *Mathematica* to evaluate,

```
solution = NMinimize[Log[z] - 2.1 λ, λ]
```

Mathematica returns this answer, $\{1.0911, \{\lambda \rightarrow 0.150566\}\}$. The first element in the list is the minimum value that is achieved. This is the information entropy of the MEP assignment as established in our derivation of the Legendre transformation. The second element is a list of the arguments x that brought about this minimum value for the function $f(x)$. Thus, we have the solution for the Lagrange multiplier λ that implements the MEP assignment.

The expression $\lambda \rightarrow 0.150566$ in the inner list is a syntactic alternative for $\mathbf{Rule}[\lambda, 0.150566]$. We want to pick this rule out of the solution list so we can substitute it for the expression $Z(\lambda)$ as well as $\ln Z(\lambda)$. $\mathbf{Rest}[\mathbf{solution}]$ will

do the job of selecting the last element in **solution**. Now that we have the rule **Rule[λ , 0.150556]** for the value of λ , we can use it in the calculation of $Z(\lambda)$ through,

```
ReplaceAll[z, Rest[solution]]
```

Mathematica evaluates this as {4.08486} which is the value for the partition function $Z(\lambda)$ obtained by substituting $\lambda = 0.150556$ into,

$$Z(\lambda) = e^\lambda + e^{2\lambda} + e^{3\lambda} = 4.08486$$

This, of course, is the value of the denominator in the MEP formula. Before we go ahead and calculate the three values for the numerator needed to find the Q_i under this model, calculate $\ln Z(\lambda)$. This follows on directly from what we have just accomplished,

```
Log[ReplaceAll[z, Rest[solution]]]
```

This evaluates to {1.40729}, the value we used for $\ln Z(\lambda)$ in the first exercise.

Now on to the finish. Retrieve the value of λ so that it can be used to calculate the numerator in the MEP formula.

```
ReplaceAll[λ, Rest[solution]]
```

This returns a list {0.150556}.

Recall that the numerator in the MEP formula looks like,

$$\exp [\lambda \times F(X = x_i)]$$

Mathematica will do all three computations for the numerator at once, and collect them into a list with,

```
numerator = Exp[(First[ReplaceAll[λ, Rest[solution]]) cf]
```

First [...] is required to pick λ out from the list because we can't have a one element vector {0.150556} multiplying the three element vector in **cf**. With this proviso, the list {1.16249, 1.35139, 1.57098} is returned for the three values of the numerator. All that remains to be done is to divide by the partition function in the denominator, a value already found.

```
denominator = ReplaceAll[z, Rest[solution]]
```

The numerical assignment for the Q_i is then,

```
qi = numerator/First[denominator]
```

The list containing the three Q_i assignments is $\{0.284586, 0.330829, 0.384586\}$. Double-check that the information entropy of this assignment is correct by,

```
- Total[qi Log[qi]]
```

which returns the answer of 1.0911.

Exercise 24.5.3: Confirm these results using the Jaynes die scenario of Chapter Nineteen.

Solution to Exercise 24.5.3

The state space for the die problem is $n = 6$. Consider models where the information from *two* constraint function averages is to be inserted into a probability distribution over the statements in the state space. The motivation for the constraint functions given by Jaynes in his die scenario were physical constraints on the center of gravity and differing lengths of the three axes of the cube.

The first constraint function defined over the statements in the state space is the vector $(1, 2, 3, 4, 5, 6)$. This constraint function attempts to capture the physical effect of the center of gravity. The second constraint function defined over the statements in the state space is the vector $(-1, -1, +2, +2, -1, -1)$. This constraint function attempts to capture the physical effects of a die that is not a perfect cube. Here the supposition under the model concerns a *shorter*, not a longer, axis connecting the THREE and FOUR spots. Notice that this is a different model than in Chapter Nineteen. The information inserted under model \mathcal{M}_k is that $\langle F_1 \rangle = 4$ and $\langle F_2 \rangle = 1$.

A legitimate, but non-MEP assignment a_i might be $(0, 0, 1/3, 1/3, 1/3, 0)$. The sum of these assignments is equal to 1, and the averages for the two constraint functions are,

$$\begin{aligned}\langle F_1 \rangle &= (1 \times 0) + (2 \times 0) + (3 \times 1/3) + (4 \times 1/3) + (5 \times 1/3) + (6 \times 0) \\ &= 4\end{aligned}$$

$$\begin{aligned}\langle F_2 \rangle &= (-1 \times 0) + (-1 \times 0) + (2 \times 1/3) + (2 \times 1/3) + (-1 \times 1/3) + (-1 \times 0) \\ &= 1\end{aligned}$$

The information entropy for this a_i assignment is,

$$\begin{aligned}- \sum_{i=1}^6 a_i \ln a_i &= \\ - [(0 \ln 0) + (0 \ln 0) + (1/3 \ln 1/3) + (1/3 \ln 1/3) + (1/3 \ln 1/3) + (0 \ln 0)] \\ &= 1.0986\end{aligned}$$

If this assignment a_i is not the MEP assignment, then there must be another assignment b_i which also satisfies the constraint function averages, that is, contains the same information as the a_i assignment, but possesses more missing information than a_i . Therefore, the b_i probability distribution will have an information entropy greater than 1.0986.

Employ the Legendre transformation to find this assignment with the maximum entropy $H_{max}(Q_i)$. Minimize the following function,

$$\ln Z(\lambda_1, \lambda_2) - (\lambda_1 \langle F_1 \rangle + \lambda_2 \langle F_2 \rangle)$$

by varying the two Lagrange multipliers λ_1 and λ_2 while keeping the two constraint function averages fixed at $\langle F_1 \rangle = 4$ and $\langle F_2 \rangle = 1$.

The MEP algorithm finds that the minimum of 1.4767 occurs for a model with the parameter values of $\lambda_1 = 0.359707$ and $\lambda_2 = 0.541805$. The MEP assignment is then,

$$b_i \equiv Q_i = (0.0263, 0.0376, 0.2740, 0.3927, 0.1107, 0.1587)$$

This MEP assignment has an information entropy of 1.4767 which must be greater than a_i 's information entropy of 1.0986. The a_i assignment contains more information than just the two stated constraint function averages; to wit, it asserts that three assignments are zero. The b_i assignment does not contain any more information other than the two stated constraint function averages. The b_i assignment maximized the amount of missing information in the distribution because it maximized the information entropy while satisfying all of the constraints.

There is no other assignment a_i different from b_i that can have greater entropy than what b_i possesses. Therefore, every such non-MEP assignment a_i must contain hidden information other than what was stated under model M_k .

The MEP assignment makes intuitive sense as well. This model postulated a physical influence changing the center of gravity such that the probability increased as the spots progressed from ONE through SIX. A more powerful physical influence was the shortened THREE-FOUR dimension of the cube which greatly raised the probability for the THREE and FOUR spots in conjunction with the changing center of gravity. The fact that both the λ_1 and λ_2 parameters had non-zero values indicated that both constraint functions would have an impact on the assigned probability.

Exercise 24.5.4: Employ the same tactics to analyze the correlational model in the simpler kangaroo scenario.

Solution to Exercise 24.5.4

The simpler kangaroo scenario had a state space of $n = 4$. The most complex models could therefore consist of $m = 3$ constraint functions and their averages. One

complex model that was examined in Chapter Twenty One looked at the correlation resulting in an assignment of $Q_i = (0.70, 0.05, 0.05, 0.20)$.

The constraint functions formed marginal probabilities for beer preference, hand preference, and the one double interaction of hand and beer preference. Collect these three vectors into a constraint matrix for *Mathematica* so that we now have,

$$\mathbf{cm} = \{\{1, 1, 0, 0\}, \{1, 0, 1, 0\}, \{1, 0, 0, 0\}\}$$

The four numerators for the Q_i under this model are then,

$$e^{\lambda_1 + \lambda_2 + \lambda_3}, e^{\lambda_1}, e^{\lambda_2}, 1$$

as found by `Exp[Dot[lambda, cm]]`. The partition function $Z(\lambda_1, \lambda_2, \lambda_3)$ serving as the denominator for these four numerators is,

$$Z(\lambda_1, \lambda_2, \lambda_3) = e^{\lambda_1 + \lambda_2 + \lambda_3} + e^{\lambda_1} + e^{\lambda_2} + 1$$

found by `Total[Exp[Dot[lambda, cm]]]`.

Relying upon the Legendre transformation, the Lagrange multipliers are found to equal,

$$\lambda_1 = -1.38629$$

$$\lambda_2 = -1.38629$$

$$\lambda_3 = +4.02535$$

through `NMinimize[Log[z]-Dot[lambda, {0.75, 0.75, 0.70}], lambda]`.

This *Mathematica* code is simply an instantiation of the Legendre transformation formula,

$$H(\langle F \rangle) = \min_{\lambda} \left[\ln Z(\lambda_1, \dots, \lambda_m) - \sum_{j=1}^m \lambda_j \langle F_j \rangle \right]$$

When *Mathematica* finds this solution, it also gives us the information entropy of the resulting numerical assignment as,

$$H_{max}(Q_i) = - \sum_i^n Q_i \ln Q_i = 0.871133$$

This value is, of course, the maximum information entropy of any assignment satisfying the three marginal probabilities specified in the model.

The actual value of the partition function after the Lagrange multipliers have been found is then $Z(\lambda_1, \lambda_2, \lambda_3) = 5$. The actual values of the four numerators become $(3.5, 0.25, 0.25, 1)$ with the numerical assignment to the four probabilities the aforementioned $Q_i = (0.70, 0.05, 0.05, 0.20)$.

Exercise 24.5.5: Make use of the Legendre transformation algorithm to implement the MEP formula that finds the assignments for a model using information about all three double interactions in the three variable kangaroo scenario.

Solution to Exercise 24.5.5

Conceptually, everything follows exactly as before. We just have more details to keep track of.

In Chapter Twenty Two, we looked at several models for the three variable kangaroo scenario, but the model requested in this exercise was not one of them. If we include all three double interactions, we are sampling from the class of $m = 6$ models.

These models include information about the marginal probabilities of beer preference, hand preference, and fur color as well as the marginal probabilities defining all three double interactions. These models *exclude* any information about the triple interaction involving beer preference, hand preference, and fur color. This is the explicit reason why the MEP algorithm is employed. All missing information, and thus information from a triple interaction, must perforce be missing from such a model.

Since the dimension of the state space has now been raised to $n = 8$, the vectors representing each of the six constraint functions for the marginal probabilities must consist of eight elements. For example, the vector for beer preference $F_1(X = x_i)$ is $(1, 1, 1, 1, 0, 0, 0, 0)$.

This arrangement is dictated by the arbitrary way in which the joint probability table was constructed. Since the marginal probability for beer preference consists of the assignments in the first four cells of the joint probability table, $Q_1 + Q_2 + Q_3 + Q_4$, any information in the form of a constraint function average about the marginal probability for beer preference $\langle F_1 \rangle$ is,

$$\langle F_1 \rangle = \sum_{i=1}^8 F_1(X = x_i) Q_i$$

$F_1(X = x_i)$ must then be the vector shown above. The marginal probabilities for hand preference and fur color, $F_2(X = x_i)$ and $F_3(X = x_i)$, are defined similarly.

A double interaction is still a marginal probability, but one consisting of only two probabilities instead of the four probabilities as in the main effects above. For example, the vector for the double interaction of beer preference and hand preference $F_4(X = x_i)$ is $(1, 0, 1, 0, 0, 0, 0, 0)$.

The information about a double interaction involving beer preference and hand preference is an average reflected in the marginal probability $Q_1 + Q_3$. The marginal probabilities for the remaining two double interactions, $F_5(X = x_i)$ and $F_6(X = x_i)$, are defined similarly.

These six constraint functions forming marginal probabilities for beer preference, hand preference, fur color, together with all three double interactions are collected into the constraint matrix,

$$\text{cm} = \{\{1, 1, 1, 1, 0, 0, 0, 0\}, \{1, 0, 1, 0, 1, 0, 1, 0\}, \{1, 1, 0, 0, 1, 1, 0, 0\}, \\ \{1, 0, 1, 0, 0, 0, 0, 0\}, \{1, 1, 0, 0, 0, 0, 0, 0\}, \{1, 0, 0, 0, 1, 0, 0, 0\}\}$$

The eight numerators for the Q_i under this model are then found, as you would expect, by the same vector matrix multiplication expression `Exp[Dot[lambda, cm]]`. For example, the numerator for Q_1 is,

$$e^{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 + \lambda_6}$$

This result can be checked by examining the numerator of the MEP formula for Q_1 .

$$\exp \left[\sum_{j=1}^6 \lambda_j F_j(X = x_1) \right]$$

Here i is fixed at 1 while j varies from 1 to 6. So, λ_1 is multiplied by $F_1(X = x_1)$, λ_2 is multiplied by $F_2(X = x_1)$... and λ_6 is multiplied by $F_6(X = x_1)$. Inspection reveals that a 1 is present for all $F_j(X = x_1)$, confirming the above result.

What is the numerator for Q_5 ? Now i is fixed at 5. Check the values for all six $F_j(X = x_5)$. Only $F_2(X = x_5)$, $F_3(X = x_5)$, and $F_6(X = x_5)$ have a 1, while the other three constraint functions have a 0 for this joint statement. Thus, the numerator for Q_5 is,

$$e^{\lambda_2 + \lambda_3 + \lambda_6}$$

Of course, *Mathematica* saves us from the effort of having to descend down into this level of detail. The partition function $Z(\lambda_1, \dots, \lambda_6)$ is also calculated in the same way as before.

Looking inside the Legendre transformation formula once again,

$$H(\langle F \rangle) = \min_{\lambda} \left[\ln Z(\lambda_1, \dots, \lambda_m) - \sum_{j=1}^m \lambda_j \langle F_j \rangle \right]$$

we see that we now require $m = 6$ constraint function averages $\langle F_j \rangle$ to insert into the program. Let the first three marginal probabilities be the same as all of the other models discussed so far,

$$\langle F_1 \rangle = 0.75$$

$$\langle F_2 \rangle = 0.75$$

$$\langle F_3 \rangle = 0.60$$

For a specific model from the class of $m = 6$, choose the information setting the marginal probabilities for the three double interactions at,

$$\langle F_4 \rangle = 0.60$$

$$\langle F_5 \rangle = 0.50$$

$$\langle F_6 \rangle = 0.40$$

The MEP algorithm works its magic to find the six values for the Lagrange multipliers. The numerical assignment for the probabilities of the eight joint statements under this model is given in Figure 24.1 below.

		B		\bar{B}			
		Cell 1 + Cell 5					
		.40					
F	H	.365	.135	F	Cell 5 .035	Cell 6 .065	.10 .60
	\bar{H}	.235	.015		Cell 7 .115	Cell 8 .035	.15 .40
		.50	.25				
		.60	.15				
		.75	.75				
		.15	.10				
		.25	.25				
						1.00	

Figure 24.1: The numerical assignments to all eight cells of the joint probability table for the enhanced kangaroo scenario under a model incorporating all three double interactions.

Notice especially that all six constraint function averages, as well as the universal constraint, are satisfied by this assignment. The three main effects are enclosed with a single box, and the three double interactions are enclosed in a double box.

For example, the constraint function average of $\langle F_6 \rangle = 0.40$ was inserted as information about the marginal probability of the HF double interaction under this model, that is, the information about the hand preference–fur color double interaction. By the **Sum Rule**,

$$P(HF) = P(BHF) + P(\bar{B}HF) = \text{Cell 1} + \text{Cell 5} = Q_1 + Q_5 = 0.40 = \langle F_6 \rangle$$

Remember that the constraint function vector was $F_6(X = x_i) = (1, 0, 0, 0, 1, 0, 0, 0)$.

This model resulted in an information entropy of $H_{max}(Q_i) = 1.70259$. The information entropy under this model has been reduced from the maximum possible

value of $H_{max}(Q_i) = \ln 8 = 2.07944$ under the fair model where all $Q_i = 1/8$. This reflects the fact that there was far more missing information under the fair model than this current model incorporating information about the main effects and all three double interactions.

Exercise 24.5.6: What is a model with the least amount of missing information?

Solution to Exercise 24.5.6

Every time we added a constraint function to a newer, more complicated model, the information entropy decreased. In the last exercise, we saw that the information entropy decreased from $H_{max} = 2.07944$ to $H_{max} = 1.70259$ when the model transitioned from $m = 0$ to $m = 6$.

The absolute minimum value of the maximum information entropy is 0 when one of the statements is assigned a probability of 1, and the remaining statements are assigned a probability of 0. Since the entropy, the quantitative measure of the amount of missing information, is now $H_{max} = 0$, this model has the least amount of missing information. Or, putting it bluntly, there is *no* missing information in this model.

Exercise 24.5.7: What happens to the information entropy as the active information becomes more constraining?

Solution to Exercise 24.5.7

For a start, investigate what happens when a model with $m = 1$ sets the marginal probability for beer preference at 0.99. $H_{max} = 1.4423$. The next model with $m = 2$ sets, in addition, the marginal probability for hand preference at 0.99. $H_{max} = 0.80515$. The next model with $m = 3$ sets, in addition, the marginal probability for fur color at 0.99. $H_{max} = 0.16801$.

The trend is clear. The information entropy is approaching 0. The amount of active information in each succeeding model is decreasing the total amount of missing information. The numerical assignment to Q_1 by each succeeding model changes from 0.2475 to 0.4901 to 0.9703.

The trend will continue as the model with $m = 4$ sets, in addition to the three previous marginal probabilities, the marginal probability for the beer preference–hand preference double interaction at 0.99. $H_{max} = 0.112$, and $Q_1 = 0.9801$.

By now, it should be clear that, under this latest model $Q_3 = 0.0099$, because the BH double interaction,

$$P(BH | \mathcal{M}_k) = P(BHF | \mathcal{M}_k) + P(BH\bar{F} | \mathcal{M}_k)$$

is a marginal probability represented by the sum of the assignments in cells 1 and 3, and the information in this model is that $\langle F_4 \rangle = 0.99 = Q_1 + Q_3$.

Chapter 25

Deriving the Maximum Entropy Principle

25.1 Introduction

In these opening Chapters of Volume II, we were suitably content just to introduce the MEP formula for the sole intent of assigning numerical values to probabilities. An initial rationale for the formula was hinted at by showing that it maximized the information entropy of the sought for probability distribution. In some sense, the resulting probability distribution also had to satisfy some constraints under a given model. This goal might be viewed as even more important than maximizing the information entropy. But we offered no deeper justification for the MEP algorithm.

We discussed the fact that when using the phrase, “maximizing the entropy,” we might just as well be saying that we are maximizing the “missing information” that is left over after specifying the “desired information” in the constraint functions and constraint function averages. In our view, it was better to see the formula in action through some examples with easily understood state spaces before trying to derive it from its variational foundation in calculus.

Thus, we have examined the small state spaces involved in coin tossing, rolling a die, and the traits of beer-drinking kangaroos. We have extended its legitimacy to inferential problems in logistic regression, and will attempt to do the same shortly for statistical thermodynamics.

In each such case, we saw how the MEP algorithm worked to provide numerical assignments to probabilities for the statements in the state space. Each different model, characterized by either of its dual parameters, namely the Lagrange multipliers or the constraint function averages, performed the essential goal of inserting information into a probability distribution.

If the constraints did not, in and of themselves, allow the probabilities to be determined, then any remaining ambiguity in the assignments had to be adjudicated by appeal to the MEP. That is, if two possible distributions both satisfied the constraints, then we chose the one that possessed the greater information entropy measure. By acting in this way, the MEP offered us a guarantee that no extraneous information had snuck into the distribution unnoticed.

The style adopted in this Chapter is defended in the *Apologia*. That is my way of saying that there is nothing fun in this Chapter. It is essentially boring and tedious, with a languorous story line filled in with piddling details. Unfortunately, it is also a prime example of the knowledge uncertainty principle.

By that I mean you *may* accept on faith the formula that assigns a probability to the Q_i . You save yourself the time spent concentrating on the derivation of this fact. Or, you can be a skeptic and demand to be shown the proof at the cost of following a long and tedious trail with lots of annoying details.

25.2 Overview of Constrained Optimization

The MEP formula originates within a sub-discipline of calculus that concerns itself with minimizing or maximizing the output of a function. An interesting complication to this basic objective is that in attempting to maximize or minimize a function, the function itself is subject to constraints. Usually, this refined goal is shortened to the phrase *constrained optimization*.

For our specific concerns here, what we would like to maximize is the information entropy function, in other words, the missing information. The constraints must eventually reflect the desired information as they will appear in the Q_i assignments.

For example, in Chapter Nineteen we talked about capturing some information about various physical imperfections of a die. Thus, the basic premise of the MEP approach is to incorporate this desired information within some model, as was done in model \mathcal{M}_C for the die. This particular model inserted the following information:

- 1. Universal constraint:** the Q_i must sum to 1 and,
- 2. First constraint:** the center of gravity, $\langle F_1(X = x_i) \rangle$, was displaced when compared to a fair die, and,
- 3. Second constraint:** the length along one axis of the cube, $\langle F_2(X = x_i) \rangle$, was different than for a fair die.

Now, we begin a long, slow, and careful derivation of the general MEP formula,

$$Q_i = \frac{\exp [\sum_{j=1}^m \lambda_j F_j(X = x_i)]}{Z(\lambda_1, \lambda_2, \dots, \lambda_m)} \quad (25.1)$$

from its initial formulation as a maximization problem in calculus to its embodiment as Equation (25.1).

25.3 The MEP and its Origin in Calculus

To come to grips with the MEP, we have to consult that part of mathematics that provides very general theorems for finding the maxima and minima of functions. Among many other things, calculus deals with the problem of finding the particular values of the function's arguments where the function attains a maximum or minimum.

In an introductory exposition, the student is shown how to find the maximum or minimum of a function of one variable. This concept is later generalized to functions of many variables. Even this insight is generalized further by modifying where an unconstrained maximum or minimum would be found by introducing side conditions that restrict the values the arguments can take on.

The essential idea is that we have a function of several variables, and we want to maximize this function. However, we have additional information about these same several variables in the form of side conditions, so in actual fact we are faced with a constrained optimization problem. First, we present this problem in the generic notation of variational calculus. Then, we match it up with our information processing notation.

Consider the simplest of all multivariate problems, namely, a function f with just two arguments, x and y . The function that is to be maximized is then $f(x, y)$. There exists an accompanying side condition written as $g(x, y) = 0$. We skip the preliminaries, and proceed directly to the solution of this constrained optimization problem. This solution is known as the method of Lagrange multipliers, or sometimes as the method of undetermined multipliers.

The generic solution looks like this,

$$\nabla \mathbf{f} + \lambda \nabla \mathbf{g} = \mathbf{0} \quad (25.2)$$

where λ is the same constant that was presented earlier as the Lagrange multiplier. Equation (25.2) is a vector equation. To reinforce this distinction \mathbf{f} , \mathbf{g} , and $\mathbf{0}$ have been made bold face.

∇ is the symbol¹ that indicates the gradient of a function. It is clear then, by the placement of ∇ operator in the above equation, that we must specify the gradients of both functions \mathbf{f} and \mathbf{g} in Equation (25.2). The gradient is defined as a vector with elements consisting of the partial derivatives of the function with respect to its variables.

¹The symbol was introduced by Sir William Hamilton and is sometimes called the *nabla*, apparently after an ancient Assyrian harp whose shape it resembles. More commonly, it is called the *del* operator.

The gradients of \mathbf{f} and \mathbf{g} are therefore expressed as,

$$\nabla \mathbf{f} \equiv \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right) \quad (25.3)$$

$$\nabla \mathbf{g} \equiv \left(\frac{\partial g}{\partial x}, \frac{\partial g}{\partial y} \right) \quad (25.4)$$

Although Equation (25.2) is compactly represented in vector notation, we need to expand this vector equation into a set of two scalar equations,

$$\frac{\partial f}{\partial x} + \lambda \frac{\partial g}{\partial x} = 0 \quad (25.5)$$

$$\frac{\partial f}{\partial y} + \lambda \frac{\partial g}{\partial y} = 0 \quad (25.6)$$

At this point we have two equations, but three unknowns, x , y , and λ . If we had three equations for these three unknowns, then we could solve for the unknowns using the standard techniques for solving homogeneous simultaneous equations. The third equation that comes to our rescue is the side condition, $g(x, y) = 0$.

The MEP's problem is to maximize the information entropy function, a function that involves arguments called Q_i . The solution to this maximization problem is constrained by the fact that the Q_i must conform to the information inserted by the IP under some model. One piece of information that is included automatically in every single model is the universal constraint.

Because we started off our explanation with a function of two variables, we will make the correspondence to a state space with dimension $n = 2$. The IP can use this general prescription from calculus encapsulated into Equations (25.5) and (25.6) to solve the information processing problem of assigning the numerical values Q_1 and Q_2 to the probabilities for statements in this (very small) state space.

25.4 The Derivation

25.4.1 Matching things up

It is now a convenient time to match-up the MEP problem with the calculus solution to a constrained optimization problem. Since we started with the simplest possible example of two variables, we shall be forced, in the beginning, to restrict ourselves to a two-dimensional state space where $n = 2$.

The function $f(x, y)$ is the objective function to be maximized, so it matches up with the information entropy function,

$$f(x, y) \equiv H(q_1, q_2) = - \sum_{i=1}^n q_i \ln q_i$$

If x and y are the two variables in the function to be maximized, then they match up with q_1 and q_2 . We switch back to the q_i notation to reinforce the fact that the MEP has not yet found the actual Q_i .

The one side condition of $g(x, y) = 0$ matches up with the requirement that q_1 and q_2 must add up to one. We use the information processing arguments q_1 and q_2 with $G(q_i) = q_1 + q_2 - 1 = 0$ for the side condition,

$$g(x, y) = 0 \equiv G(q_1, q_2) = q_1 + q_2 - 1$$

These match-ups are summarized for your convenience in Table 25.1 below.

Table 25.1: *The match-ups between the calculus problem of maximizing a function subject to side conditions and finding a distribution with maximum entropy subject to constraints.*

Description	Calculus	Description	MEP Formula
argument 1	x	Probability for statement 1	q_1
argument 2	y	Probability for statement 2	q_2
Function maximized	$f(x, y)$	Information Entropy	$H(q_i) = -\sum_{i=1}^2 q_i \ln q_i$
Side Condition	$g(x, y) = 0$	Constraint	$G(q_i) = \sum_{i=1}^2 q_i - 1 = 0$
Lagrange multiplier	λ	Parameter	λ

25.4.2 Substituting the MEP concepts

Substituting the MEP notation into the method of Lagrange multipliers shown as Equations (25.5) and (25.6), followed by introducing the constraint equation, results in the following set of three equations with three unknowns,

$$\frac{\partial H}{\partial q_1} + \lambda \frac{\partial G}{\partial q_1} = 0 \quad (25.7)$$

$$\frac{\partial H}{\partial q_2} + \lambda \frac{\partial G}{\partial q_2} = 0 \quad (25.8)$$

$$(q_1 + q_2) - 1 = 0 \quad (25.9)$$

Let's now begin to solve these three equations. Focus first on the partial derivative of the information entropy with respect to q_1 .

$$\frac{\partial H}{\partial q_1} = \frac{\partial [-(q_1 \ln q_1 + q_2 \ln q_2)]}{\partial q_1} \quad (25.10)$$

$$= -(\ln q_1 + 1) \quad (25.11)$$

In Equation (25.10), we merely inserted the definition of information entropy for the q_i of a two dimensional state space. Since we are taking the partial derivative with respect to q_1 , we can treat $q_2 \ln q_2$ as a constant. Then find the derivative of this simpler expression,

$$\frac{d [-(q_1 \ln q_1 + k)]}{dq_1}$$

with the result shown as Equation (25.11).

The sub-step in the derivation is perhaps deserving of a little more attention. This step involves taking the derivative of the product $q_1 \ln q_1$. The rules for derivatives tell us that products like this can be handled by taking the first term in the product, q_1 , times the derivative of the other term in the product, $\ln q_1$, and adding this to the second term, $\ln q_1$, times the derivative of the first term, q_1 .

Cast into the appropriate mathematical garb, these words are expressed symbolically as,

$$\frac{d(uv)}{dq_1} = u \frac{dv}{dq_1} + v \frac{du}{dq_1} \quad (25.12)$$

where,

$$u = q_1$$

$$v = \ln q_1$$

$$\frac{dv}{dq_1} = \frac{1}{dq_1}$$

$$\frac{du}{dq_1} = 1$$

From this rule for the derivative of a product we see that,

$$\begin{aligned} \frac{\partial H}{\partial q_1} &= \frac{d(uv)}{dq_1} \\ &= - \left[\left(q_1 \times \frac{1}{q_1} \right) + (\ln q_1 \times 1) \right] \\ &= -(1 + \ln q_1) \end{aligned}$$

From symmetry considerations it must be that,

$$\frac{\partial H}{\partial q_2} = -(1 + \ln q_2)$$

The partial derivatives for the constraint on the sum of the q_i are even easier,

$$\begin{aligned}\frac{\partial G}{\partial q_1} &= \frac{\partial (q_1 + q_2 - 1)}{\partial q_1} \\ &= \frac{d (q_1 + k)}{dq_1} \\ &= 1 \\ \frac{\partial G}{\partial q_2} &= 1\end{aligned}$$

Having found these partial derivatives, the three very general equations, Equations (25.7), (25.8), and (25.9), now look like,

$$-(\ln q_1 + 1) + \lambda = 0 \quad (25.13)$$

$$-(\ln q_2 + 1) + \lambda = 0 \quad (25.14)$$

$$(q_1 + q_2) - 1 = 0 \quad (25.15)$$

These equations can be solved for the three unknowns, q_1 , q_2 , and λ , by elementary techniques.

To eliminate λ , subtract Equation (25.14) from Equation (25.13),

$$\begin{aligned}-\ln q_1 - 1 + \lambda + \ln q_2 + 1 - \lambda &= 0 \\ \ln q_1 &= \ln q_2 \\ q_1 &= q_2\end{aligned}$$

Substitute this finding into Equation (25.15),

$$\begin{aligned}(q_1 + q_2) - 1 &= 0 \\ 2q_1 - 1 &= 0 \\ q_1 &= \frac{1}{2}\end{aligned}$$

Since we know from our previous result that $q_1 = q_2$, q_2 must also equal $1/2$.

Now we can solve for the third and final unknown, λ , by substituting the known value of q_1 into Equation (25.13).

$$\begin{aligned} -[\ln(1/2) + 1] + \lambda &= 0 \\ \lambda &= 1 + \ln(1/2) \\ &= 1 + \ln 1 - \ln 2 \\ &= 1 - \ln 2 \end{aligned}$$

An alternative solution to these equations is also helpful for what lies ahead. Equations (25.13) to (25.15) are written down once more,

$$\begin{aligned} -(\ln q_1 + 1) + \lambda &= 0 \\ -(\ln q_2 + 1) + \lambda &= 0 \\ (q_1 + q_2) - 1 &= 0 \end{aligned}$$

but this time our objective is to isolate q_1 and q_2 as functions of λ . Then, we will substitute into the constraint equation.

$$\begin{aligned} \ln q_1 &= \lambda - 1 \\ \ln q_2 &= \lambda - 1 \\ q_1 &= \exp(\lambda - 1) \\ q_2 &= \exp(\lambda - 1) \end{aligned}$$

At this point we substitute these exponential expressions for q_1 and q_2 into the constraint equation that says these two terms have to add to one, Equation (25.15),

$$\begin{aligned} [\exp(\lambda - 1) + \exp(\lambda - 1)] - 1 &= 0 \\ \exp(\lambda - 1) + \exp(\lambda - 1) &= 1 \\ 2 \exp(\lambda - 1) &= 1 \\ \exp(\lambda - 1) &= \frac{1}{2} \\ \lambda - 1 &= \ln\left(\frac{1}{2}\right) \\ \lambda &= 1 - \ln 2 \end{aligned}$$

Having solved for λ we can now substitute this known value to solve for q_1 and q_2 .

$$\begin{aligned} q_1 &= \exp(\lambda - 1) \\ &= \exp(1 + \ln(1/2) - 1) \\ &= \exp[\ln(1/2)] \\ &= 1/2 \\ q_2 &= 1/2 \end{aligned}$$

25.4.3 Generalizing to a larger state space

We can generalize these results in one direction by increasing the dimension of the state space that we are considering. Let's do this by increasing the state space to three. We thereby increase the number of unknowns, q_1 , q_2 , q_3 , and λ , to four, but we keep pace by adding another equation. Since we now have four simultaneous equations for the four unknowns, we can once again solve for the value of the unknowns.

Spelling this out for the $n = 3$ case follows close on the heels of the previous exposition,

$$\begin{aligned} -(\ln q_1 + 1) + \lambda &= 0 \\ -(\ln q_2 + 1) + \lambda &= 0 \\ -(\ln q_3 + 1) + \lambda &= 0 \\ (q_1 + q_2 + q_3) - 1 &= 0 \end{aligned}$$

Employing the second technique for solving these four simultaneous equations, we attempt to isolate q_1 , q_2 , and q_3 as exponential functions of the Lagrange multiplier with the following result,

$$\begin{aligned} q_1 &= \exp(\lambda - 1) \\ q_2 &= \exp(\lambda - 1) \\ q_3 &= \exp(\lambda - 1) \end{aligned}$$

Just as before, we now want to plug these q_i as found above into the fundamental axiom of probability theory that says they must all add up to one,

$$\exp(\lambda - 1) + \exp(\lambda - 1) + \exp(\lambda - 1) = 1$$

Now it is a set of easy manipulations to arrive at,

$$\begin{aligned}
 3 \exp(\lambda - 1) &= 1 \\
 \exp(\lambda - 1) &= \frac{1}{3} \\
 \lambda - 1 &= \ln\left(\frac{1}{3}\right) \\
 \lambda &= 1 + \ln\left(\frac{1}{3}\right) \\
 &= 1 + \ln 1 - \ln 3 \\
 &= 1 - \ln 3
 \end{aligned}$$

Having found $\lambda - 1$, q_1 can be calculated,

$$\begin{aligned}
 q_1 &= \exp[\ln(1/3)] \\
 &= 1/3
 \end{aligned}$$

Since the q_i are all equal, they now have been assigned the numerical values,

$$Q_i = (1/3, 1/3, 1/3)$$

It's not that great of a leap to generalize to the case of n statements in the state space in order to write down the expressions for Q_i and λ ,

$$Q_i = \frac{1}{n}$$

$$\text{and } \lambda = 1 - \ln n$$

What we have been doing so far has been instructive in terms of translating the general prescription from calculus. It told us how to find the maximum of a function of several variables subject to side conditions so that we could maximize the information entropy subject to the simple constraint that probabilities must add up to one.

It has also been instructive to go through the practical steps of solving simultaneous linear equations, but certainly the discovery that $Q_i = 1/n$ could not be considered as some profound insight. And if it ended there, we would not be spending any time at all on the maximum entropy principle.

25.5 State Space of $n = 3$ with Two Constraints

We generalized in one direction by increasing the number of statements in the state space, but kept the number of constraints at one. We now generalize in the other direction by allowing extra constraints.

You might wonder what adding an extra constraint does to the master formula expressed in Equation (25.2). Adding an extra constraint function does not materially change the pattern of the mathematical formula. Equation (25.2) transforms into,

$$\nabla \mathbf{H} + \lambda_0 \nabla \mathbf{G}_0 + \lambda_1 \nabla \mathbf{G}_1 = \mathbf{0} \quad (25.16)$$

The universal constraint originally written as \mathbf{G} is relabeled as \mathbf{G}_0 , while the new constraint is shown as \mathbf{G}_1 .

$H(Q_i | \mathcal{M}_k)$, the information entropy function, remains as the objective function that we want to maximize. This function is maximized by varying the arguments to the function, the Q_i . \mathbf{G}_0 is the universal constraint, as just discussed in the last section, where,

$$\mathbf{G}_0(q_i) \equiv \sum_{i=1}^n q_i - 1 = 0$$

In the expanded set-up, \mathbf{G}_1 is introduced as a “second” constraint (really the first given our implicit acknowledgment of the universal constraint) concerning the average value of an observable. This constraint, quite familiar by now, is,

$$\mathbf{G}_1(q_i) \equiv \sum_{i=1}^n [F_1(X = x_i) q_i] - \langle F_1 \rangle = 0$$

We now have two constants, λ_0 and λ_1 , as Lagrangian multipliers for these two constraints. It's easily seen that the universal constraint follows the same form,

$$\sum_{i=1}^n [F_0(X = x_i) q_i] - \langle F_0 \rangle = 0 \equiv \mathbf{G}_0(q_i) \equiv \sum_{i=1}^n q_i - 1 = 0$$

25.5.1 The simultaneous equations

As a numerical example, let's say that we are considering an assignment to the probabilities for the statements in a three dimensional state space. To keep things simple, assume that the constraint function has been defined as,

$$F_1(X = x_i) = (1, 2, 3)$$

The IP would like, therefore, to find Q_1 , Q_2 , and Q_3 using the MEP formula. The IP will insert some information into the probability distribution, with the attendant cautionary language that this information is being inserted under the auspices of some model \mathcal{M}_k . This particular model will tentatively entertain the hypothesis that the average value of the one constraint function $F(X = x_i)$ is 2.5, or,

$$\langle F_1(X = x_i) \rangle = 2.5$$

We start out by setting up the template dictated by Equation (25.16). Notice that we have five unknowns, but fortunately this is balanced off with five equations. The five unknowns are the three numerical assignments to the probabilities for the three statements in the state space, q_1, q_2, q_3 , and the two parameters, the Lagrange multipliers λ_0 and λ_1 .

Expand the template in Equation (25.16), and add the two constraint equations to see these five equations,

$$\frac{\partial H}{\partial q_1} + \lambda_0 \frac{\partial G_0}{\partial q_1} + \lambda_1 \frac{\partial G_1}{\partial q_1} = 0$$

$$\frac{\partial H}{\partial q_2} + \lambda_0 \frac{\partial G_0}{\partial q_2} + \lambda_1 \frac{\partial G_1}{\partial q_2} = 0$$

$$\frac{\partial H}{\partial q_3} + \lambda_0 \frac{\partial G_0}{\partial q_3} + \lambda_1 \frac{\partial G_1}{\partial q_3} = 0$$

$$q_1 + q_2 + q_3 = 1$$

$$(1 \times q_1) + (2 \times q_2) + (3 \times q_3) = 2.5$$

We have already found the partial derivatives for the information entropy,

$$\frac{\partial H}{\partial q_i} = -(\ln q_i + 1)$$

and the partial derivatives of the universal constraint with respect to the q_i ,

$$\frac{\partial G_0}{\partial q_i} = 1$$

We only have to find the partial derivatives of the first constraint with respect to the q_i .

In order to visually unclutter the equations, let x_i stand for $F(X = x_i)$. The partial derivatives are not difficult to find,

$$\frac{\partial G_1}{\partial q_1} = \frac{\partial (x_1 q_1 + x_2 q_2 + x_3 q_3)}{\partial q_1}$$

$$= \frac{d(x_1 q_1 + k)}{dq_1}$$

$$= x_1$$

$$\frac{\partial G_1}{\partial q_2} = x_2$$

$$\frac{\partial G_1}{\partial q_3} = x_3$$

25.5.2 Solving for the Q_i

We are now in a position to set up the five simultaneous equations and solve for the specific assignments Q_i .

$$-(\ln q_1 + 1) + \lambda_0 + \lambda_1 x_1 = 0 \quad (25.17)$$

$$-(\ln q_2 + 1) + \lambda_0 + \lambda_1 x_2 = 0 \quad (25.18)$$

$$-(\ln q_3 + 1) + \lambda_0 + \lambda_1 x_3 = 0 \quad (25.19)$$

$$(q_1 + q_2 + q_3) - 1 = 0 \quad (25.20)$$

$$x_1 q_1 + x_2 q_2 + x_3 q_3 - 2.5 = 0 \quad (25.21)$$

From Equation (25.17) we obtain,

$$-\ln q_1 = 1 - \lambda_0 - \lambda_1 x_1$$

$$\ln q_1 = \lambda_1 x_1 + \lambda_0 - 1$$

$$q_1 = \exp(\lambda_1 x_1 + \lambda_0 - 1)$$

In just the same way, we obtain from Equations (25.18) and (25.19),

$$q_2 = \exp(\lambda_1 x_2 + \lambda_0 - 1)$$

$$q_3 = \exp(\lambda_1 x_3 + \lambda_0 - 1)$$

At this point we do have a formula for the Q_i , but we would like to improve it by eliminating $\lambda_0 - 1$ from within the exponential expression. Our final goal is to obtain a formula with the partition function as the normalizing factor. As a by-product, we also obtain one that does not contain a term involving $\lambda_0 - 1$.

At this juncture we begin some fancy prestidigitation, so follow my hands closely as I move the pea under the shells. Move down to Equation (25.20) in order to make use of the universal constraint stating that q_1 , q_2 , and q_3 must add up to one. Substitute into this constraint equation the exponential expressions for the q_i that we have just derived,

$$\sum_{i=1}^3 q_i = \exp(\lambda_1 x_1 + \lambda_0 - 1) + \exp(\lambda_1 x_2 + \lambda_0 - 1) + \exp(\lambda_1 x_3 + \lambda_0 - 1)$$

Since $\sum_{i=1}^3 q_i = 1$, we now have,

$$\sum_{i=1}^3 \exp(\lambda_0 - 1 + \lambda_1 x_i) = 1$$

Within the parentheses of the exponential we have two terms, $(\lambda_0 - 1)$ and $\lambda_1 x_i$, that are added together. From the rule on multiplying exponentials we know that,

$$\exp(x + y) = \exp(x) \exp(y)$$

so,

$$\sum_{i=1}^3 q_i = \sum_{i=1}^3 \exp(\lambda_0 - 1 + \lambda_1 x_i) = 1$$

could be expressed as,

$$\sum_{i=1}^3 q_i = \sum_{i=1}^3 \exp(\lambda_0 - 1) \exp(\lambda_1 x_i) = 1$$

We would like to bring out the constant term, $\exp(\lambda_0 - 1)$, from the summation,

$$\exp(\lambda_0 - 1) \sum_{i=1}^3 \exp(\lambda_1 x_i) = 1$$

Now simply divide both sides of this equation by the summation term to arrive at,

$$\exp(\lambda_0 - 1) = \frac{1}{\sum_{i=1}^3 \exp(\lambda_1 x_i)} = \frac{1}{Z}$$

To finish the derivation for the Q_i without the $\lambda_0 - 1$ term, we take advantage of the first constraint. We also take the liberty of using the general n which shouldn't be cause for too much alarm.

Substitute the exponential expression for the q_i into the definition of the average, and make use of the previous manipulations on $\exp(\lambda_0 - 1)$,

$$\begin{aligned} \langle x_i \rangle &= \sum_{i=1}^n x_i q_i \\ \sum_{i=1}^n x_i \exp(\lambda_0 - 1 + \lambda_1 x_i) &= \exp(\lambda_0 - 1) \sum_{i=1}^n x_i \exp(\lambda_1 x_i) \\ \text{since } \exp(\lambda_0 - 1) &= \frac{1}{\sum_{i=1}^n \exp(\lambda_1 x_i)} \\ \sum_{i=1}^n x_i q_i &= \frac{\sum_{i=1}^n x_i \exp(\lambda_1 x_i)}{\sum_{i=1}^n \exp(\lambda_1 x_i)} \\ \text{therefore } q_i &= \frac{\exp(\lambda_1 x_i)}{\sum_{i=1}^n \exp(\lambda_1 x_i)} \end{aligned}$$

We see that by eliminating the $\exp(\lambda_0 - 1)$ term, we have arrived at a simpler version of the general MEP algorithm formula. If the denominator is given the traditional notation from its historical origins in statistical thermodynamics,

$$Z(\lambda_1) = \sum_{i=1}^n \exp(\lambda_1 x_i)$$

then,

$$q_i = \frac{\exp(\lambda_1 x_i)}{\sum_{i=1}^n \exp(\lambda_1 x_i)} \text{ or } Q_i = \frac{e^{[\lambda F(X=x_i)]}}{Z(\lambda)} \quad (25.22)$$

25.5.3 Brute force calculation of the Lagrange multiplier

This section is a welcome respite from all the symbol shuffling just performed. It serves to reinforce the main lessons through a numerical example. Basically, we employ a crude method for zeroing in on the value of the Lagrange multiplier in a simple numerical assignment problem for a three dimensional state space. This technique is simply a grid search over the values of λ_1 . We are attempting to narrow in on the correct value of λ_1 through a finer and finer grained search.

Here is what we would like the MEP algorithm to do for us. We have a state space of dimension $n = 3$, consisting of three quite abstract statements,

Statement 1. “The object is measured as 1.”

Statement 2. “The object is measured as 2.”

Statement 3. “The object is measured as 3.”

Each of these three statements is either TRUE or FALSE. It is anticipated that the IP will be engaged in making inferences, rather than making deductions, about these statements. So, the information processor will wrap the probability operator around each statement to indicate a degree of belief that a particular statement is TRUE. Thus, we have the symbolism $P(X = x_i)$ to represent an (epistemological) state of knowledge about each of the three statements in the state space.

When we measure the object, that measurement must turn out to be one, and only one, of the values 1, 2, or 3. This is the mutually exclusive and exhaustive requirement so often mentioned.

As a precursor to any further inferential computations, we require a legitimate numerical assignment to $P(X = x_i)$. Since there can be no one true numerical assignment for a probability, but only a numerical assignment for a probability conditioned on whatever information is assumed, we write $P(X = x_i | \mathcal{M}_k)$ in order to emphasize this basic fundamental concept. Some model \mathcal{M}_k will insert information into a probability distribution with the happy consequence that numerical assignments *can* be made for the probabilities of all three statements.

The MEP formula is perfectly designed to insert information into a probability distribution, with the added assurance that whatever materializes from its employment must be a legitimate numerical assignment. It accomplishes this important task by maximizing the *missing information* that still remains in a probability distribution *after* having incorporated whatever positive information has been specifically requested by the IP. This positive information takes the usual form of a constraint function average.

So, for our little example, the constraint function has been defined as a mapping from the statements in the state space to the following three numbers,

$$F_1(X = x_i) = (1, 2, 3)$$

The information under some particular model is that the average of this constraint function is $\langle F_1 \rangle = 2.5$.

Approaching the problem from the standard calculus standpoint of finding the extrema for some objective function when subject to side conditions, we found that the missing information assumes its maximum possible value by using this formula for the Q_i ,

$$Q_i \equiv P(X = x_i | \mathcal{M}_k) = \frac{\exp [\lambda_1 F_1(X = x_i)]}{\sum_{i=1}^3 \exp [\lambda_1 F_1(X = x_i)]}$$

As mentioned before, the information can be given in one of two ways. It can be given by specifying the average of the constraint function, or by specifying the value of the Lagrange multiplier. The specification of either one of these dual parameters immediately determines its partner. Since the IP chose to insert information by giving the value of the parameter $\langle F_1 \rangle = 2.5$, the value of λ_1 is thereby determined.

The numbers in Table 25.2 at the top of the next page illustrate a grid search for λ_1 in our current problem. The table is divided into six columns with each set of two columns reflecting a finer grid search.

The first two columns show the initial coarse grained search for λ_1 conducted in increments of 1. Inspection of the table reveals that λ_1 has been examined from values ranging from +5 to -5. The number in the second column tells us whether we have conformed to the desired constraint function average,

$$\sum_{i=1}^n F_1(X = x_i) P(X = x_i | \mathcal{M}_k) = \langle F_1 \rangle$$

$$\sum_{i=1}^3 x_i Q_i = 2.5$$

for the value of λ_1 specified in the first column.

Table 25.2: A grid search calculation for the value of λ_1 under some model so that a numerical assignment can be made for the probabilities of three statements in the state space.

λ_1	$\langle F_1 \rangle$	λ_1	$\langle F_1 \rangle$	λ_1	$\langle F_1 \rangle$
5	2.9932	1.0	2.5752	0.90	2.5310
4	2.9814	0.9	2.5310	0.89	2.5264
3	2.9480	0.8	2.4833	0.88	2.5218
2	2.8509	0.7	2.4322	0.87	2.5171
1	2.5752	0.6	2.3777	0.86	2.5124
0	2.0000	0.5	2.3202	0.85	2.5076
-1	1.4248	0.4	2.2598	0.84	2.5028
-2	1.1491	0.3	2.1971	0.83	2.4980
-3	1.0520	0.2	2.1325	0.82	2.4932
-4	1.0186	0.1	2.0666	0.81	2.4883
-5	1.0068	0.0	2.0000	0.80	2.4833

Take $\lambda_1 = 1$ as given in the fifth row of the first column for a numerical example. $\sum x_i Q_i$ for this particular value of λ_1 works out to be,

$$\begin{aligned}
\sum_{i=1}^3 x_i Q_i &= \frac{\sum_{i=1}^3 x_i \exp(\lambda_1 x_i)}{\sum_{i=1}^3 \exp(\lambda_1 x_i)} \\
&= \frac{x_1 \exp(\lambda_1 x_1) + x_2 \exp(\lambda_1 x_2) + x_3 \exp(\lambda_1 x_3)}{\exp(\lambda_1 x_1) + \exp(\lambda_1 x_2) + \exp(\lambda_1 x_3)} \\
&= \frac{[1 \times \exp(1 \times 1)] + [2 \times \exp(1 \times 2)] + [3 \times \exp(1 \times 3)]}{\exp(1 \times 1) + \exp(1 \times 2) + \exp(1 \times 3)} \\
&= 2.5752
\end{aligned}$$

This value of $\lambda_1 = 1$ is therefore too large since the expectation constraint must equal 2.5. When we try $\lambda_1 = 0$ in the next row down we find that the calculated constraint is equal to 2.0, which is too low. So we know that the correct value for λ_1 lies somewhere between $\lambda_1 = 1$ and $\lambda_1 = 0$.

The next two columns (columns three and four) show the grid search between $\lambda_1 = 1$ and $\lambda_1 = 0$ in the finer grained increments of 0.1. From these two columns we can bracket the value of λ_1 as somewhere between $\lambda_1 = 0.9$ and $\lambda_1 = 0.8$.

Carrying this procedure one step further in the final two columns of Table 25.2, where the search is conducted in the even finer increments of 0.01, we are able to narrow λ_1 down to values that lie somewhere between $\lambda_1 = 0.84$ and $\lambda_1 = 0.83$. In fact, we stop at the next finer increment of 0.001 (not shown in the table) to settle on the value of $\lambda_1 = 0.834$ where $\langle F_1 \rangle$ does indeed equal 2.5.

Before we plug this value of λ_1 into the exponential expression for each Q_i , notice this fact evident from the first two columns of the table. As λ_1 becomes larger and larger, $\sum F_1(X = x_i) Q_i$ approaches the value of 3. This happens because of the boundary condition where, if the Q_i have the assignment of $Q_i = (0, 0, 1)$,

$$\sum_{i=1}^3 F_1(X = x_i) Q_i \equiv \langle F_1 \rangle = (1 \times 0) + (2 \times 0) + (3 \times 1) = 3$$

At the other extreme, as λ_1 becomes smaller and smaller, the value of the constraint function average approaches 1. This is because of the boundary condition where the probability is entirely concentrated on Q_1 as opposed to Q_3 , $Q_i = (1, 0, 0)$,

$$\sum_{i=1}^3 F_1(X = x_i) Q_i \equiv \langle F_1 \rangle = (1 \times 1) + (2 \times 0) + (3 \times 0) = 1$$

Now this line of thinking naturally leads one to wonder about the case where the probability for an observation is concentrated entirely on a measurement of 2. The average of the constraint function then is obviously 2. But this agrees with $\lambda_1 = 0$ as can be seen from the table.

In other words, if $\lambda_1 = 0$, then the constraint reflecting information about the average value of an observable is not in effect. Would the assignment of $Q_i = (0, 1, 0)$ be made in this case if $\langle F_1 \rangle = 2$?

No, it would not because $Q_i = (1/3, 1/3, 1/3)$ also satisfies $\langle F_1 \rangle = 2$. However, and most critically, this particular numerical assignment maximizes the information entropy $H(Q_i | \mathcal{M}_k)$. The assignment $Q_i = (0, 1, 0)$ has a lower entropy, namely 0, than any other assignment to the Q_i .

So what we are calling the maximum entropy assignment must be,

$$Q_i = (1/3, 1/3, 1/3)$$

as opposed to,

$$Q_i = (0, 1, 0)$$

even though both satisfy the information specified. The former assignment has the most missing information. That was why it was chosen. The latter assignment actually has no missing information as reflected in its information entropy value of 0. It is certain that the object will be measured as 2.

Table 25.3 repeats the first two columns of Table 25.2, and tacks on an additional column showing the calculated values of Q_i for the corresponding λ_1 value. The correct value of $\lambda_1 = 0.834$ is inserted into this table to record the Q_i assignment based on the MEP algorithm. The model with a parameter setting of $\lambda_1 = 0.834$ is synonymous with the model with the dual parameter setting of $\langle F_1 \rangle = 2.5$.

Table 25.3: *The MEP algorithm assigns numerical values for probabilities when given information reflected in the Lagrange multiplier.*

λ_1	$\langle F_1 \rangle$	(Q_1, Q_2, Q_3)
$+\infty$	3.0000	(0, 0, 1)
+5.000	2.9932	(0.000, 0.007, 0.993)
+4.000	2.9814	(0.000, 0.018, 0.982)
+3.000	2.9480	(0.002, 0.047, 0.950)
+2.000	2.8509	(0.016, 0.117, 0.867)
+1.000	2.5752	(0.090, 0.245, 0.665)
+0.834	2.5000	(0.116, 0.268, 0.616)
0.000	2.0000	(1/3, 1/3, 1/3)
-1.000	1.4248	(0.665, 0.245, 0.090)
-2.000	1.1491	(0.867, 0.117, 0.016)
-3.000	1.0520	(0.950, 0.047, 0.002)
-4.000	1.0186	(0.982, 0.018, 0.000)
-5.000	1.0068	(0.993, 0.007, 0.000)
$-\infty$	1.0000	(1, 0, 0)

This assignment of $Q_i = (0.116, 0.268, 0.616)$, found by the MEP formula for the information in this model, has the highest possible value of the information entropy when compared with any other possible assignment that might also satisfy the constraints. An additional row has been added at the top and bottom of the table to show that λ_1 would have to approach either $+\infty$ or $-\infty$ to reach the boundary conditions for the assignment.

Notice that when λ_1 is large and positive, the variation in the Q_i is not very “smooth” and there is a concentration on Q_3 that diminishes as λ_1 becomes smaller. As λ_1 approaches zero (no constraints other than normalization), the Q_i become “smoother,” or “more spread out.” A symmetrical “less spread out” effect occurs as λ_1 becomes increasingly negative with a corresponding concentration on Q_1 .

25.6 Final Formula

To wrap things up, it should not come as too much of a surprise that the general formula from calculus for finding those arguments that maximize the function for any number of constraint functions is,

$$\nabla \mathbf{H} + \sum_{j=0}^m \lambda_j \nabla \mathbf{G}_j = \mathbf{0} \quad (25.23)$$

with the stipulation that $m \leq n - 1$.

Reflect upon the pattern of equations, developed in the previous sections, as the state space, n , and the number of constraint functions, m , increased. There is an inescapable induction.

There will be n equations of the type,

$$-(\ln q_i + 1) + \lambda_0 + \lambda_1 F_1(X = x_i) + \lambda_2 F_2(X = x_i) + \cdots + \lambda_m F_m(X = x_i) = 0$$

together with m equations of the type,

$$\sum_{i=1}^n F_j(X = x_i) q_i - \langle F_j \rangle = 0$$

and one equation of the type,

$$\sum_{i=1}^n q_i - 1 = 0$$

Solving these $n + m + 1$ equations leads to the numerical assignments Q_i that maximize the information entropy,

$$\begin{aligned} Q_i &= \exp [(\lambda_0 - 1) + \lambda_1 F_1(X = x_i) + \lambda_2 F_2(X = x_i) + \cdots + \lambda_m F_m(X = x_i)] \\ \exp [\lambda_0 - 1] &= \frac{1}{Z} \\ &= \frac{\exp [\sum_{j=1}^m \lambda_j F_j(X = x_i)]}{Z(\lambda_1, \lambda_2, \dots, \lambda_m)} \end{aligned}$$

where the partition function in the denominator is,

$$Z(\lambda_1, \lambda_2, \dots, \lambda_m) = \sum_{i=1}^n \exp \left[\sum_{j=1}^m \lambda_j F_j(X = x_i) \right]$$

In all of this notation, it is important to remember that Q_i is really an abbreviated form for a numerical assignment to the probability for the i^{th} statement in the state space under a specific model \mathcal{M}_k ,

$$Q_i \equiv P(X = x_i | \mathcal{M}_k) = \frac{e^{\sum_{j=1}^m \lambda_j F_j(x_i)}}{Z(\lambda_1, \dots, \lambda_m)}$$

25.7 Connections to the Literature

As the reader might have expected, much of this material can be found in Jaynes's writings, albeit scattered over many years, and in many different papers with varying degrees of intended interpretation. In Jaynes's defense (should he need one), he was writing, first and foremost, to fellow physicists whose mathematical background was steeped in statistical mechanics. In his later years, he tried reaching out to a more ecumenical audience by emphasizing the consequences of the MEP for inference. Obviously, I have tried to present his arguments slanted towards this inferential perspective.

Jaynes [18, pp. 240–243] and [23, pp. 355–356] presents the mathematical formulas for the MEP in a concise format. These are the formulas which we have labored to spell out in greater detail throughout this Chapter. We mention in passing that Equation (B4) [18, pg. 242] is the motivation for Exercise 25.8.15. The quantitative solution to the loaded die problem of Chapter Nineteen follows a few sentences later [18, pp. 243–244].

Jaynes [19, pp. 320–321] has this commentary concerning the nature of the constraints that enter into the formalism of the MEP.

... in a real application one will wish, if possible, to choose the constraint matrix ... so that the resulting quantities [our $\langle F_j(X = x_i) \rangle$] represent systematic physical influences, real or conjectured, (for example, eccentric position of the center of gravity of a die), which constrain the frequencies to be different from the uniform distribution of absolute maximum entropy [our $H_{max}(Q_i) = \ln n$].

Jaynes [20, 21] can be read for further insight into all the matters discussed so far. If the reader does refer back to Jaynes's original work, he or she will often find that a minus sign appears in front of the λ s in the MEP formulas. This is because Jaynes sets up the variational problem in a different manner than I have chosen.

A detailed exposition of some of the mathematical steps that precede the MEP formula can also be found in [9]. As a final non-Jaynesian attribution, the notation used in calculus and presented in section 25.2 for finding the maximum of a function via the method of Lagrange multipliers is taken from the multivariate calculus text of Kaplan and Lewis [27, pp. 974–975].

You can pick up almost any text on advanced calculus or numerical optimization to find, with varying degrees of clarity, a discussion on constrained optimization by the method of Lagrange multipliers. The only other reference I have ever found that specifically mentions its application to information entropy is in a book on numerical optimization by Fletcher [8, pp. 222–223].

In passing, you may try your hand at implementing an MEP algorithm from the outline skeleton code given in Fletcher. I did so and used it successfully for many years before I found that the *Mathematica* implementation of the Legendre transformation gave me the same results with much less effort.

25.8 Solved Exercises for Chapter Twenty Five

Exercise 25.8.1: This exercise and the next are typical problems that calculus texts use to introduce the method of Lagrange multipliers. Find the point on a circle closest to the point $(x, y) = (1, 2)$. The circle has a radius of 5 with the origin at $(0, 0)$.

Solution to Exercise 25.8.1

The distance from any point (x, y) that we might consider from the given fixed point $(1, 2)$ is defined by,

$$d = \sqrt{(x - 1)^2 + (y - 2)^2}$$

To make the determination of the derivatives easier, we'll use the squared distance,

$$f(x, y) = (x - 1)^2 + (y - 2)^2$$

as the objective function. This is a function of two variables for which we want to find either the minimum or the maximum subject to some constraint. Here, we want to find the *minimum* distance. If we wanted to find the maximum of $f(x, y)$ without worrying about any constraints, then we could just let x and y get bigger and bigger. For example,

$$\begin{aligned} f(10, 20) &= (10 - 1)^2 + (20 - 2)^2 \\ &= 405 \end{aligned}$$

$$\begin{aligned} f(100, 200) &= (100 - 1)^2 + (200 - 2)^2 \\ &= 49,005 \end{aligned}$$

Or, if we were interested in the minimum of $f(x, y)$, again without being concerned about any constraints, we could let x get arbitrarily close to 1 and y arbitrarily close to 2.

$$\begin{aligned} f(1.1, 2.1) &= (1.1 - 1)^2 + (2.1 - 2)^2 \\ &= 0.02 \\ f(1.01, 2.01) &= (1.01 - 1)^2 + (2.01 - 2)^2 \\ &= 0.0002 \end{aligned}$$

Nevertheless, we do have a constraint on the possible values we can plug into $f(x, y)$, and this constraint says that any values for x and y must lie on a circle with a radius of 5. Let us write the equation for a circle centered at the point $(0, 0)$ as the constraint function,

$$g(x, y) = x^2 + y^2 = 25$$

Such constraints we write in the alternative form of,

$$g(x, y) = x^2 + y^2 - 25 = 0$$

Now only two numbers x and y whose squares add up to 25 can be substituted into $f(x, y)$. The easiest choice to try would be $(x, y) = (3, 4)$. This choice does satisfy the constraint equation $g(x, y)$. Substituting into $f(x, y)$ gives a squared distance of 8.

Is there another point which satisfies the constraint of being on the circle, but gives a smaller value for $f(x, y)$? If $(x, y) = (2\sqrt{2}, \sqrt{17})$ then this point also satisfies the constraint of being on the circle and, when plugged into $f(x, y)$, gives a squared distance of 7.85 from the origin, smaller than our previous point $(x, y) = (3, 4)$. Since we want to find the point closest to $(1, 2)$, this last choice of $(x, y) = (2\sqrt{2}, \sqrt{17})$ is better than $(x, y) = (3, 4)$.

We could continue to search in this fashion, finding better and better points that satisfied the constraint equation, that is, a point on the circle, but also resulted in a smaller value for $f(x, y)$. But we have the analytical expression of Equation (25.2) that will find the answer for us without the search.

$$\nabla f + \lambda \nabla g = \mathbf{0}$$

$$\nabla f \equiv \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right)$$

$$\nabla g \equiv \left(\frac{\partial g}{\partial x}, \frac{\partial g}{\partial y} \right)$$

$$\frac{\partial f}{\partial x} = 2(x - 1)$$

$$\frac{\partial f}{\partial y} = 2(y - 2)$$

$$\frac{\partial g}{\partial x} = 2x$$

$$\frac{\partial g}{\partial y} = 2y$$

Equations (25.5) and (25.6) expanded the vector notation $\nabla f + \lambda \nabla g = \mathbf{0}$ into,

$$\frac{\partial f}{\partial x} + \lambda \frac{\partial g}{\partial x} = 0$$

$$\frac{\partial f}{\partial y} + \lambda \frac{\partial g}{\partial y} = 0$$

Substitute the values just found for the partial differentiations,

$$2(x - 1) + \lambda 2x = 0$$

$$2(y - 2) + \lambda 2y = 0$$

$$x = \frac{1}{1 + \lambda}$$

$$y = \frac{2}{1 + \lambda}$$

$$y = 2x$$

Now substitute this value for y into the constraint equation,

$$x^2 + (2x)^2 = 25$$

$$x^2 + 4x^2 = 25$$

$$5x^2 = 25$$

$$x = \pm\sqrt{5}$$

$$y = \pm 2\sqrt{5}$$

as four possible points (x, y) that minimize $f(x, y)$ and are also on the circle. By trying each point, or by intuition that the closest point must be positive for both x and y , we discover that $(x, y) = (\sqrt{5}, 2\sqrt{5})$ gives $f(x, y) = 7.64$, and is therefore the point on the circle with radius 5 that is closest to $(1, 2)$.

Exercise 25.8.2: What are the dimensions of a rectangle with perimeter of length p that give the largest area?

Solution to Exercise 25.8.2

For the sake of a numerical example, let's take the given fixed perimeter of a rectangle to be 100. The function that we want to maximize is the area of the rectangle $f(x, y) = xy$. The constraint equation is $g(x, y) = 2x + 2y = 100$. Now if $x = 40$ and $y = 10$, the constraint equation of the perimeter length is satisfied and the area of the rectangle is $f(40, 10) = 400$.

If we let $x = 30$ and $y = 20$ we continue to satisfy the constraint equation of perimeter length, but we can now see that our previous choice was not the best since $f(30, 20) = 600$, obviously a larger value for the area of the rectangle than 400.

To find the values for x and y that maximize the area without going through a trial and error search, we again employ the formal solution.

$$\frac{\partial f}{\partial x} + \lambda \frac{\partial g}{\partial x} = 0$$

$$\frac{\partial f}{\partial y} + \lambda \frac{\partial g}{\partial y} = 0$$

$$y + 2\lambda = 0$$

$$x + 2\lambda = 0$$

$$y = x$$

Substituting into the constraint equation yields,

$$2x + 2y = 100$$

$$2x + 2x = 100$$

$$4x = 100$$

$$x = 25$$

If x and y equal 25 the constraint equation is satisfied and $f(25, 25) = 625$ which is the maximum value for the area. This exercise with Lagrange multipliers merely confirms what we knew from geometry that in the class of rectangles, it is the square that has the largest area for a given fixed value of the perimeter.

Exercise 25.8.3: Write out the generic pattern of equations for solving the variational problem as dictated by the method of Lagrange multipliers for three variables and two constraints.

Solution to Exercise 25.8.3

This is an exercise in using abstract notation suitable for any variational problem. The objective function that we wish to maximize is $f(x, y, z)$ where f has three arguments x , y , and z . The first constraint function is $g(x, y, z) = 0$ and the second constraint function is $h(x, y, z) = 0$. Label the Lagrange multiplier for the first constraint as λ and the Lagrange multiplier for the second constraint as μ . There are three unknowns x , y , and z plus two more unknowns λ and μ .

Thus, we will set up and solve these five equations,

$$\frac{\partial f}{\partial x} + \lambda \frac{\partial g}{\partial x} + \mu \frac{\partial h}{\partial x} = 0$$

$$\frac{\partial f}{\partial y} + \lambda \frac{\partial g}{\partial y} + \mu \frac{\partial h}{\partial y} = 0$$

$$\frac{\partial f}{\partial z} + \lambda \frac{\partial g}{\partial z} + \mu \frac{\partial h}{\partial z} = 0$$

$$g(x, y, z) = 0$$

$$h(x, y, z) = 0$$

Our information entropy–centric notation for these first three equations was shown in Equation (25.16) as,

$$\nabla \mathbf{H} + \lambda_0 \nabla \mathbf{G}_0 + \lambda_1 \nabla \mathbf{G}_1 = \mathbf{0}$$

Exercise 25.8.4: Here are some technical exercises that serve as a precursor to Equation (25.2).

Solution to Exercise 25.8.4

The total differential df of a function $f(x, y)$ is defined by,

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy$$

Thus, the total differential for the information entropy $H(q_i)$ objective function is expressed as,

$$dH = \frac{\partial H}{\partial q_1} dq_1 + \frac{\partial H}{\partial q_2} dq_2 + \cdots + \frac{\partial H}{\partial q_n} dq_n$$

In the same way, the total differential of any constraint function $G(q_i)$ is,

$$dG = \frac{\partial G}{\partial q_1} dq_1 + \frac{\partial G}{\partial q_2} dq_2 + \cdots + \frac{\partial G}{\partial q_n} dq_n$$

Exercise 25.8.5: How does *Mathematica* express the total differential?

Solution to Exercise 25.8.5

The total differential for the information entropy of a state space of dimension of $n = 3$ would be,

$$dH = \frac{\partial H}{\partial q_1} dq_1 + \frac{\partial H}{\partial q_2} dq_2 + \frac{\partial H}{\partial q_3} dq_3$$

We know that,

$$\frac{\partial H}{\partial q_i} = -(\ln q_i + 1)$$

so,

$$dH = -(\ln q_1 + 1) dq_1 - (\ln q_2 + 1) dq_2 - (\ln q_3 + 1) dq_3$$

The *Mathematica* built-in function for a total differential is `Dt[arg]` where here *arg* is the function to be optimized. Letting *Mathematica* evaluate,

$$\text{Dt}[-(q1 \text{ Log}[q1] + q2 \text{ Log}[q2] + q3 \text{ Log}[q3])]$$

results in,

$$-\text{Dt}[q1] (1+\text{Log}[q1]) - \text{Dt}[q2] (1+\text{Log}[q2]) - \text{Dt}[q3] (1+\text{Log}[q3])$$

which matches the total differential dH given above.

Exercise 25.8.6: What is a basic concept ingrained about maximization from a study of calculus?

Solution to Exercise 25.8.6

Despite all of the many things we may have forgotten from our earnest attempts at calculus, one concept always seemed very clear cut. The maximum of some function occurred at that point where the slope was equal to zero. Hearkening back to this remembrance, what are the consequences of letting the total differential df equal 0? Not only that, but form the total differential expression $d(f + \lambda g) = 0$ where the objective function and the constraint function have been combined.

Look at a generic template for two variables,

$$\begin{aligned} d(f + \lambda g) &= \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + \lambda \left(\frac{\partial g}{\partial x} dx + \frac{\partial g}{\partial y} dy \right) \\ &= \left(\frac{\partial f}{\partial x} + \lambda \frac{\partial g}{\partial x} \right) dx + \left(\frac{\partial f}{\partial y} + \lambda \frac{\partial g}{\partial y} \right) dy \\ d(f + \lambda g) &= 0 \end{aligned}$$

Since the coefficients of dx and dy must both equal 0, we see that we want to set up the two equations,

$$\frac{\partial f}{\partial x} + \lambda \frac{\partial g}{\partial x} = 0$$

$$\frac{\partial f}{\partial y} + \lambda \frac{\partial g}{\partial y} = 0$$

which are the two equations Equations (25.5) and (25.6).

Exercise 25.8.7: Apply the Lagrange multiplier method for finding the maximum of a function subject to constraints of the coin tossing scenario.

Solution to Exercise 25.8.7

Imagine that either one coin will be tossed N times, or that N coins have been tossed once. Let's use the language of saying that we have a “assembly” of N “systems” with the typical state space for each system of $n = 2$. Thus, our goal is to calculate the probability for the two frequency counts N_1 and N_2 , where N_1 is the number of HEADS that will turn up, and N_2 is the number of TAILS that will turn up.

Instead of the information entropy, let the objective function in this case be the multiplicity factor $W(N)$. This objective function is subject to two constraints. The first constraint is that N_1 and N_2 must add up to N . The second constraint is that the “total energy” over the assembly of N coins must equal some specified number.

The multiplicity factor for the series of coin tosses is easily obtained as,

$$W(N) = \frac{N!}{N_1! N_2!}$$

Since N is fixed, the multiplicity factor can be maximized by minimizing the denominator. Take the log transform of the denominator followed by the Stirling approximation for factorials,

$$\begin{aligned} \ln(N_1! \times N_2!) &= \ln N_1! + \ln N_2! \\ \ln N_1! + \ln N_2! &= N_1 \ln N_1 - N_1 + N_2 \ln N_2 - N_2 \\ &= \sum_{i=1}^n N_i \ln N_i - \sum_{i=1}^n N_i \\ &= \sum_{i=1}^n N_i \ln N_i - N \end{aligned}$$

Thus, we set up the objective function $f(N_i)$ as $\sum_{i=1}^n N_i \ln N_i - N$. The partial derivative of f with respect to its argument is,

$$\frac{\partial f}{\partial N_i} = 1 + \ln N_i$$

This objective function is subject to the two side conditions of the first and second constraint functions. The first constraint function g is,

$$g(N_i) = N - \sum_{i=1}^n N_i = 0$$

The partial derivative of g with respect to its argument is,

$$\frac{\partial g}{\partial N_i} = -1$$

The second constraint function h is,

$$h(N_i) = E - \sum_{i=1}^n N_i E_i = 0$$

The partial derivative of h with respect to its argument is,

$$\frac{\partial h}{\partial N_i} = -E_i$$

The Lagrange method of undetermined multipliers establishes a template for this problem of,

$$\frac{\partial f}{\partial N_i} + \lambda_1 \frac{\partial g}{\partial N_i} + \lambda_2 \frac{\partial h}{\partial N_i} = 0$$

After substituting the partial derivatives found above, we have,

$$\begin{aligned} (1 + \ln N_i) + \lambda_1(-1) + \lambda_2(-E_i) &= 0 \\ \ln N_i &= \lambda_1 - 1 + \lambda_2 E_i \\ N_i &= e^{\lambda_1 - 1 + \lambda_2 E_i} \\ &= e^{\lambda_1 - 1} e^{\lambda_2 E_i} \end{aligned}$$

Some further manipulations bring us to an expression for the normed frequency counts,

$$\begin{aligned} \sum_{i=1}^n N_i &= N \\ \sum_{i=1}^n N_i &= \sum_{i=1}^n e^{\lambda_1 - 1} e^{\lambda_2 E_i} \\ &= e^{\lambda_1 - 1} \sum_{i=1}^n e^{\lambda_2 E_i} \\ e^{\lambda_1 - 1} \sum_{i=1}^n e^{\lambda_2 E_i} &= N \\ \frac{N_i}{N} &= \frac{e^{\lambda_1 - 1} e^{\lambda_2 E_i}}{e^{\lambda_1 - 1} \sum_{i=1}^n e^{\lambda_2 E_i}} \\ \frac{N_i}{N} &= \frac{e^{\lambda_2 E_i}}{\sum_{i=1}^n e^{\lambda_2 E_i}} \end{aligned}$$

Exercise 25.8.8: What kind of motivation for the argument given in the last exercise do you see in the literature?

Solution to Exercise 25.8.8

This formula for the frequency counts as found by applying the method of Lagrange multipliers is identified with the Boltzmann distribution in statistical mechanics. The assembly of N coins each with only two possible outcomes E_i is a stand-in for a assembly of N molecules each with a possible energy E_i , where each system is a molecule.

The total number of molecules is fixed at N with a frequency count of N_i molecules possessing energy E_i . This serves as the first constraint. The total energy over the physical system is specified as a second constraint.

The multiplicity factor, not the information entropy, is maximized subject to the two constraints of keeping both N and the total energy E fixed. The Lagrange multiplier λ_2 is related to Boltzmann's constant and the temperature. The sum over the n possible energy states provides the partition function. N is always assumed to be a very large number on the order of, say, 6×10^{23} .

Exercise 25.8.9: Provide a numerical example to illustrate Exercise 25.8.7.

Solution to Exercise 25.8.9

Let the total number of coins tossed be $N = 100$. The constraint function mapping values to the two statements in the state space are $E_1 = 1$ and $E_2 = 2$. The constraint function of total energy is specified as $E = N_1 E_1 + N_2 E_2 = 175$. The normed frequency count for the number of HEADS is then calculated by the above formula as,

$$\begin{aligned}\frac{N_1}{N} &= \frac{e^{\lambda_2 E_1}}{\sum_{i=1}^n e^{\lambda_2 E_i}} \\ &= \frac{e^{\lambda_2 E_1}}{e^{\lambda_2 E_1} + e^{\lambda_2 E_2}} \\ &= \frac{e^{\lambda_2}}{e^{\lambda_2} + e^{2\lambda_2}}\end{aligned}$$

Suppose $\lambda_2 = 0$, then

$$\frac{N_1}{N} = \frac{e^0}{e^0 + e^0} = \frac{1}{2}$$

and the number of HEADS and TAILS would both be equal to 50.

But does this result satisfy the second constraint?

$$\sum_{i=1}^n N_i E_i = E = 175$$

It does not as,

$$N_1 E_1 + N_2 E_2 = E = 150$$

A little trial and error with N_1 and N_2 reveals that the number of HEADS must equal $N_1 = 25$ with the number of TAILS $N_2 = 75$ to satisfy both the first and second constraints.

Ask *Mathematica* to solve this equation,

```
Solve[Exp[lambda]/(Exp[lambda]+Exp[2 lambda])==.25, lambda]
```

to find the proper value for λ_2 . The evaluation returns $\{\{\text{lambda} \rightarrow 1.09861\}\}$.

As a final check, substitute the formula for $\frac{N_i}{N}$ in the calculation of the sample average energy \bar{E} ,

$$\begin{aligned} \frac{N_1}{N} &= \frac{\exp[\lambda_2 \times 1]}{\exp[\lambda_2 \times 1] + \exp[\lambda_2 \times 2]} \\ \frac{N_2}{N} &= \frac{\exp[\lambda_2 \times 2]}{\exp[\lambda_2 \times 1] + \exp[\lambda_2 \times 2]} \end{aligned}$$

The sample average energy \bar{E} is then found by,

$$\begin{aligned} \bar{E} &= \left(\frac{N_1}{N} \times E_1 \right) + \left(\frac{N_2}{N} \times E_2 \right) \\ &= \frac{(\exp[\lambda_2 \times 1] \times 1) + (\exp[\lambda_2 \times 2] \times 2)}{\exp[\lambda_2 \times 1] + \exp[\lambda_2 \times 2]} \\ &= \frac{\exp[\lambda_2] + 2\exp[2\lambda_2]}{\exp[\lambda_2] + \exp[2\lambda_2]} \\ &= \frac{\exp[1.09861] + 2\exp[2 \times 1.09861]}{\exp[1.09861] + \exp[2 \times 1.09861]} \\ &= 1.75 \end{aligned}$$

with the total energy E ,

$$\begin{aligned} E &= N \left[\left(\frac{N_1}{N} \times E_1 \right) + \left(\frac{N_2}{N} \times E_2 \right) \right] \\ &= 100 \times 1.75 \\ &= 175 \end{aligned}$$

Exercise 25.8.10: Comment on some apparent deficiencies in a critique of this traditional approach to the Boltzmann equation.

Solution to Exercise 25.8.10

We first broached this topic back in Exercises 17.7.9 and 17.7.10. The MEP *should* be employed in order to make numerical assignments, under some model, to the n probabilities of the joint statements in the state space. The information entropy is the objective function to be maximized, not the multiplicity function. The mathematical expectation of any constraint functions should be used to insert information into the distribution of probabilities over the statements in the state space, and not some sample average as obtained from data.

N_1 and N_2 are frequency counts. Making use of the MEP formula to determine frequency counts just doesn't make any sense at all. They are ontological entities. They either are known, or will be known, and called data. The MEP should be used only for epistemological entities, for example, assigning a numerical value to a probability for seeing so many HEADS and TAILS in N tosses of the coin, $P(N_1, N_2)$. This represents an epistemological concept for the IP, in other words, a state of knowledge about frequency counts.

Exercise 25.8.11: Provide the correct treatment for the generic coin tossing scenario from a probabilistic standpoint.

Solution to Exercise 25.8.11

The first question to ask is: What probability are we trying to find? Are we trying to find $P(X = x_i | \mathcal{M}_k)$, or $P(N_1, N_2 | \mathcal{M}_k)$, or $P(N_1, N_2)$, or $P(M_1, M_2 | N_1, N_2)$, or $P(M_1, M_2 | N_1, N_2)$ as N_1 and N_2 get progressively larger, that is, as more and more data are accumulated?

The MEP is used solely to make the numerical assignments,

$$Q_i \equiv P(X = x_i | \mathcal{M}_k)$$

given the information supplied by model \mathcal{M}_k . Thus, for this coin tossing scenario, under the model as discussed in this Chapter where the constraint function is,

$$F(X = x_i) = (1, 2)$$

and the constraint function average is $\langle F \rangle = 1.75$, the assignments to HEADS and TAILS are,

$$\begin{aligned}
 Q_1 &= \frac{e^\lambda}{e^\lambda + e^{2\lambda}} \\
 &= \frac{e^{1.09861}}{e^{1.09861} + e^{2(1.09861)}} \\
 &= 0.25 \\
 Q_2 &= \frac{e^{2\lambda}}{e^\lambda + e^{2\lambda}} \\
 &= \frac{e^{2(1.09861)}}{e^{1.09861} + e^{2(1.09861)}} \\
 &= 0.75
 \end{aligned}$$

The rules of probability then tell us what $P(N_1 = 25, N_2 = 75 | \mathcal{M}_k)$ must be. There is no need to rely upon the MEP at this stage by invoking the multiplicity factor as the objective function to be maximized, and sample averages from the data as constraints.

$$\begin{aligned}
 P(N_1 = 25, N_2 = 75 | \mathcal{M}_k) &= \frac{N!}{N_1! N_2!} Q_1^{N_1} Q_2^{N_2} \\
 &= \frac{100!}{25! 75!} (0.25)^{25} (0.75)^{75} \\
 &= 0.0918
 \end{aligned}$$

There is another expression for this probability taken from Exercise 17.7.9,

$$P(N_1, N_2 | \mathcal{M}_k) = \exp \left[N \left(\frac{\ln W(N)}{N} + \lambda \bar{E} - \ln Z \right) \right]$$

that shows how the multiplicity factor, together with sample averages, do make an appearance, but only as a natural consequence from the first expression.

So the probability for seeing $N_1 = 25$ HEADS and $N_2 = 75$ TAILS under this model is,

$$P(N_1 = 25, N_2 = 75 | \mathcal{M}_k) = \exp \left[N \left(\frac{\ln W(N)}{N} + (\lambda \times \bar{E}) - \ln Z \right) \right]$$

$$\begin{aligned} \frac{\ln W(N)}{N} &= \ln \left(\frac{100!}{25! 75!} \right) / 100 \\ &= 0.538454 \end{aligned}$$

$$\begin{aligned} \lambda \times \bar{E} &= 1.09861 \times 1.75 \\ &= 1.92577 \end{aligned}$$

$$\begin{aligned} \ln Z &= \ln [\exp(\lambda) + \exp(2\lambda)] \\ &= 2.4849 \end{aligned}$$

$$\begin{aligned} \frac{\ln W(N)}{N} + (\lambda \times \bar{E}) - \ln Z &= 0.538454 + 1.92577 - 2.4849 \\ &= -0.0238815 \end{aligned}$$

$$\begin{aligned} P(N_1 = 25, N_2 = 75 | \mathcal{M}_k) &= \exp [100 \times (-0.0238815)] \\ &= 0.0918 \end{aligned}$$

The parameter λ and the log of the partition function $\ln Z$ have these particular values because they arose in generating the Q_i from the MEP formula. Obviously, these same Q_i appear in $P(N_1, N_2 | \mathcal{M}_k)$.

If the IP has not settled on one exclusive model, but exists in a “state of complete ignorance” about the cause of HEADS or TAILS showing up, then it will average over models making every conceivable assignment. This probability is written as $P(N_1, N_2)$ as opposed to $P(N_1, N_2 | \mathcal{M}_k)$ in order to indicate a marginal probability which integrated over all assignments in the region from 0 to 1.

From Volume I, and Laplace’s *Rule of Succession*,

$$P(N_1 = 25, N_2 = 75) = \frac{N! (n-1)!}{(N+n-1)!} = \frac{100! 1!}{(102-1)!} = \frac{1}{101} = \frac{1}{N+1}$$

The probability for seeing any frequency count adding up to $N = 100$ has the same probability of $1/101$.

Now we need to switch to the disambiguated notation of M_i and N_i . The M_i refer to *future* frequency counts, while the N_i refer to *past* frequency counts. We should have been writing $P(M_1, M_2 | \mathcal{M}_k)$ and $P(M_1, M_2)$. It only makes sense to attach a probability to future frequency counts because the N_i , the data points, are

known with certainty. Thus, the IP will condition on the known data when trying to find the probability of something that has not yet happened.

We could first examine the case where the IP is interested in making an inference about the very next occurrence of HEADS or TAILS after the already known results from the first 100 tosses. The probability for seeing TAILS on the 101st toss is, again relying upon formulas developed in Volume I,

$$P(M_1 = 0, M_2 = 1 \mid N_1 = 25, N_2 = 75) = \frac{N_2 + 1}{N + n} = \frac{76}{102}$$

If, in fact, the IP wants the probability for the next 100 tosses based on what has already happened in the initial 100 tosses, then,

$$\begin{aligned} P(M_1 = 25, M_2 = 75 \mid N_1 = 25, N_2 = 75) &= C \times \frac{\prod_{i=1}^2 (M_i + N_i)!}{\prod_{i=1}^2 M_i!} \\ C &= \frac{M! (N + n - 1)!}{N_1! N_2! (M + N + n - 1)!} \\ &= \frac{100! (100 + 2 - 1)!}{25! 75! (100 + 100 + 2 - 1)!} \\ \frac{\prod_{i=1}^2 (M_i + N_i)!}{\prod_{i=1}^2 M_i!} &= \frac{(25 + 25)! (75 + 75)!}{25! 75!} \\ P(M_1 = 25, M_2 = 75 \mid N_1 = 25, N_2 = 75) &= 0.0651 \end{aligned}$$

A smaller probability for this particular future frequency count of HEADS and TAILS makes intuitive sense when compared to,

$$P(M_1 = 25, M_2 = 75 \mid \mathcal{M}_k) = 0.0918$$

because it is based on a limited amount of data. Assuming model \mathcal{M}_k true is equivalent to having seen an infinite amount of data that supports $Q_1 = 0.25$ and $Q_2 = 0.75$.

We can watch the probability for 25 HEADS and 75 TAILS in the next 100 tosses approach a probability of 0.0918 as the amount of data supporting model \mathcal{M}_k increases.

$$P(M_1 = 25, M_2 = 75 \mid N_1 = 25, N_2 = 75) = 0.0651$$

$$P(M_1 = 25, M_2 = 75 \mid N_1 = 50, N_2 = 150) = 0.0750$$

$$P(M_1 = 25, M_2 = 75 \mid N_1 = 300, N_2 = 900) = 0.0882$$

$$P(M_1 = 25, M_2 = 75 \mid N_1 = 600, N_2 = 1800) = 0.0899$$

$$P(M_1 = 25, M_2 = 75 \mid N_1 \rightarrow \infty, N_2 \rightarrow \infty) = 0.0918$$

Exercise 25.8.12: What is the value for λ if the dimension of the state space is $n = 10$, and the universal constraint is the only constraint that is imposed by the model?

Solution to Exercise 25.8.12

If the only constraint imposed is the universal constraint, then use the notation λ_0 for the Lagrange multiplier. From the development in section 25.4.3,

$$\lambda_0 = 1 - \ln n$$

$$n = 10$$

$$\lambda_0 = 1 - \ln 10$$

$$\lambda_0 = -1.302585$$

Exercise 25.8.13: What is the numerical assignment for the probability of the ten statements in the state space of the previous exercise?

Solution to Exercise 25.8.13

Since the universal constraint is the only information inserted into the distribution of probability assignments over the ten statements in the state space, we know first of all that,

$$Q_1 = Q_2 = \dots = Q_{10}$$

Applying the initial development of the MEP formula in section 25.4.3 when the universal constraint was the only constraint on the arguments to information entropy, the numerical assignment to all ten probabilities was,

$$\begin{aligned} Q_i &= \exp(\lambda_0 - 1) \\ &= \exp(-1.302585 - 1) \\ &= \exp(-2.302585) \\ &= 0.10 \end{aligned}$$

with the obvious consequences that,

$$\frac{1}{\exp(\lambda_0 - 1)} = \exp(1 - \lambda_0) = 10 = Z \text{ and } Q_i = \frac{\exp(0)}{Z} = \frac{1}{10}$$

Any further constraints are not being applied, so all further $\lambda_j = 0$.

Exercise 25.8.14: Use one of the standard numerical approximations to the derivative to verify that $\frac{\partial H}{\partial q_1} = -(\ln q_1 + 1)$ at the particular assignment of $Q_1 = 0.50$.

Solution to Exercise 25.8.14

Let's take the central difference formula,

$$f'(a) \approx \frac{f(a+h) - f(a-h)}{2h}$$

as one definition for a numerical derivative. Label the i^{th} term in the information entropy formula with $f(q_i) = -(q_i \ln q_i)$. Then,

$$\begin{aligned} f(q_i) &= -(q_i \ln q_i) \\ f'(q_i) &\approx \frac{[-(q_i + h) \ln (q_i + h)] - [-(q_i - h) \ln (q_i - h)]}{2h} \end{aligned}$$

$$\text{Let } h = 0.01$$

$$\begin{aligned} f'(0.50) &\approx \frac{-(0.51 \ln 0.51) - (-0.49 \ln 0.49)}{2 \times 0.01} \\ &= \frac{0.3434 - 0.3495}{0.02} \\ &= -0.30679 \end{aligned}$$

Using the exact symbolic formula, the derivative is calculated as,

$$\begin{aligned} \frac{d [-(q_i \ln q_i + k)]}{dq_i} &= -(\ln q_i + 1) \\ &= -0.306853 \end{aligned}$$

The *Mathematica* expression which evaluates to the same answer is,

```
ReplaceAll[D[-(q Log[q]), q], Rule[q, .5]]
```

Exercise 25.8.15: Use the same standard numerical approximation for the partial derivative to compute the constraint function average for the example in section 25.5.3.

Solution to Exercise 25.8.15

Using the same numerical approximation for the partial derivative as in the previous exercise, we have,

$$\begin{aligned}
 f'(a) &\approx \frac{f(a+h) - f(a-h)}{2h} \\
 f'(a) &= f'[\ln Z(\lambda_1)] \\
 \text{with } h &= 0.001 \\
 Z(\lambda_1 + 0.001) &= e^{0.835 \times 1} + e^{0.835 \times 2} + e^{0.835 \times 3} \\
 &= 2.3048 + 5.3122 + 12.2346 \\
 &= 19.8605 \\
 \ln [Z(\lambda_1 + 0.001)] &= 2.9887 \\
 Z(\lambda_1 - 0.001) &= e^{0.833 \times 1} + e^{0.833 \times 2} + e^{0.833 \times 3} \\
 &= 2.3002 + 5.2910 + 12.1703 \\
 &= 19.7615 \\
 \ln [Z(\lambda_1 - 0.001)] &= 2.9837 \\
 \frac{\ln [Z(\lambda_1 + 0.001)] - \ln [Z(\lambda_1 - 0.001)]}{2 \times 0.001} &= \frac{2.9887 - 2.9837}{0.002} \\
 &= 2.5
 \end{aligned}$$

The constraint function average for this problem was given as $\langle F_1 \rangle = 2.5$ confirming the above numerical result. The result just found in this particular case is true in general,

$$\frac{\partial [\ln Z(\lambda_1, \lambda_2, \dots, \lambda_m)]}{\partial \lambda_j} = \langle F_j \rangle$$

where $\langle F_j \rangle$ stands for the average value of any j th constraint.

Exercise 25.8.16: Verify the MEP formula for the $n = 4$ kangaroo scenario from first principles using the method of Lagrange multipliers.

Solution to Exercise 25.8.16

To orient ourselves, refer back to the solution found by an application of the MEP formula in Chapter Twenty One. Suppose we concentrate on the $m = 2$ case. The solution shown in Table 21.1 had the numerical assignments,

$$Q_i = (9/16, 3/16, 3/16, 1/16)$$

under the model with the two constraints on the marginal probabilities for hand and beer preference. We should be able to duplicate this result from first principles as outlined in this Chapter.

The objective function to be maximized is the information entropy,

$$H(q_i) = - \sum_{i=1}^4 q_i \ln q_i$$

with

$$\frac{\partial H}{\partial q_i} = -(\ln q_i + 1)$$

The universal constraint is,

$$G_0(q_i) = \sum_{i=1}^4 q_i = 1$$

with

$$\frac{\partial G_0}{\partial q_i} = 1$$

The first constraint on the marginal probability for hand preference is,

$$G_1(q_i) = q_1 + q_2 = 0.75$$

with

$$\frac{\partial G_1}{\partial q_1} = 1 \text{ and } \frac{\partial G_1}{\partial q_2} = 1$$

The second constraint on the marginal probability for beer preference is,

$$G_2(q_i) = q_1 + q_3 = 0.75$$

with

$$\frac{\partial G_2}{\partial q_1} = 1 \text{ and } \frac{\partial G_2}{\partial q_2} = 0$$

Following the template provided by Equation (25.23), we are presented with the following four equations,

$$-(\ln q_1 + 1) + \lambda_0(1) + \lambda_1(1) + \lambda_2(1) = 0$$

$$-(\ln q_2 + 1) + \lambda_0(1) + \lambda_1(1) + \lambda_2(0) = 0$$

$$-(\ln q_3 + 1) + \lambda_0(1) + \lambda_1(0) + \lambda_2(1) = 0$$

$$-(\ln q_4 + 1) + \lambda_0(1) + \lambda_1(0) + \lambda_2(0) = 0$$

From the first equation we have,

$$-\ln q_1 = 1 - \lambda_0 - \lambda_1 - \lambda_2$$

$$\ln q_1 = \lambda_0 - 1 + \lambda_1 + \lambda_2$$

$$q_1 = \exp(\lambda_0 - 1) \exp(\lambda_1 + \lambda_2)$$

In a similar manner we have,

$$q_2 = \exp(\lambda_0 - 1) \exp(\lambda_1)$$

$$q_3 = \exp(\lambda_0 - 1) \exp(\lambda_2)$$

$$q_4 = \exp(\lambda_0 - 1) \exp(0)$$

The partition function $Z(\lambda_1, \lambda_2)$ is found from the universal constraint by bringing out the constant factor $e^{\lambda_0 - 1}$ when the four assignments above are summed,

$$\sum_{i=1}^4 q_i = \exp(\lambda_0 - 1) [\exp(\lambda_1 + \lambda_2) + \exp(\lambda_1) + \exp(\lambda_2) + 1]$$

Since we have that,

$$\sum_{i=1}^4 q_i \equiv q_1 + q_2 + q_3 + q_4 = 1$$

we can finish up with,

$$\begin{aligned} \exp(\lambda_1 + \lambda_2) + \exp(\lambda_1) + \exp(\lambda_2) + 1 &= \frac{1}{\exp(\lambda_0 - 1)} \\ &= Z(\lambda_1, \lambda_2) \end{aligned}$$

Thus, we have confirmed the MEP formula as used in Chapter Twenty One on the

simplified kangaroo scenario under a model with two constraints with,

$$Q_1 = \frac{\exp(\lambda_1 + \lambda_2)}{Z(\lambda_1, \lambda_2)}$$

$$Q_2 = \frac{\exp(\lambda_1)}{Z(\lambda_1, \lambda_2)}$$

$$Q_3 = \frac{\exp(\lambda_2)}{Z(\lambda_1, \lambda_2)}$$

$$Q_4 = \frac{1}{Z(\lambda_1, \lambda_2)}$$

$$Z(\lambda_1, \lambda_2) = \exp(\lambda_1 + \lambda_2) + \exp(\lambda_1) + \exp(\lambda_2) + 1$$

To find the actual values of the two Lagrange multipliers, apply what was shown about the Legendre transformation,

$$H_{max}(Q_i) = \min_{\lambda_1, \lambda_2} [\ln Z(\lambda_1, \lambda_2) - \sum_{j=1}^2 \lambda_j \langle F_j \rangle]$$

which responds with the results,

$$H_{max}(Q_i) = 1.12467 \text{ at } \lambda_1 = \lambda_2 = 1.098612 \text{ with } Z = 16$$

which in turn produce our MEP assignments under this model of,

$$Q_i = (0.5625, 0.1875, 0.1875, 0.0625)$$

Exercise 25.8.17: Exploit the pattern evident in the last exercise to write down the assignments for the $n = 8$ enhanced kangaroo scenario under a model with constraints on the marginal probabilities for all three variables.

Solution to Exercise 25.8.17

The pattern for the numerator of each Q_i is to look at the constraint function values as the coefficients for the λ s. For example, write down the numerator for Q_1 by inspecting the constraint function values appearing as the first element in each of the three relevant constraint function vectors. These are all 1, so the numerator for Q_1 is $e^{\lambda_1 + \lambda_2 + \lambda_3}$.

In like manner, write down the numerator for Q_6 as e^{λ_3} because the sixth constraint function values for the first three constraint functions are 0, 0, and 1. Here are all eight numerators in the enhanced kangaroo scenario,

$$\text{Numerator of } Q_1 = e^{\lambda_1 + \lambda_2 + \lambda_3}$$

$$\text{Numerator of } Q_2 = e^{\lambda_1 + \lambda_3}$$

$$\text{Numerator of } Q_3 = e^{\lambda_1 + \lambda_2}$$

$$\text{Numerator of } Q_4 = e^{\lambda_1}$$

$$\text{Numerator of } Q_5 = e^{\lambda_2 + \lambda_3}$$

$$\text{Numerator of } Q_6 = e^{\lambda_3}$$

$$\text{Numerator of } Q_7 = e^{\lambda_2}$$

$$\text{Numerator of } Q_8 = e^0$$

The partition function is the sum of these eight numerators so,

$$Z(\lambda_1, \lambda_2, \lambda_3) = e^{\lambda_1 + \lambda_2 + \lambda_3} + e^{\lambda_1 + \lambda_3} + e^{\lambda_1 + \lambda_2} + e^{\lambda_1} + e^{\lambda_2 + \lambda_3} + e^{\lambda_3} + e^{\lambda_2} + 1$$

Exercise 25.8.18: How would *Mathematica* do all of this for you?

Solution to Exercise 25.8.18

The summation and multiplication of the constraint functions by the Lagrange multipliers,

$$\text{Numerator of } Q_i = \exp \left[\sum_{j=1}^m \lambda_j F_j(X = x_i) \right]$$

is taken care of by `Dot[lambda, cm]`, which is then followed by,

`Exp[Dot[lambda, cm]]`

for the list of each Q_i numerator. The partition function is then,

`Total[Exp[Dot[lambda, cm]]]`

`lambda` was set up as the list `List[λ1, λ2, λ3]` while `cm` contained the first three elements from the full constraint matrix array for the $n = 8$ kangaroo scenario,

`{ { 1, 1, 1, 1, 0, 0, 0, 0 }, { 1, 0, 1, 0, 1, 0, 1, 0 }, { 1, 1, 0, 0, 1, 1, 0, 0 } }`

Chapter 26

Statistical Mechanics and the MEP Formula

26.1 Introduction

I was very surprised, here in Volume II, to find myself including two Chapters discussing statistical mechanics. This is a highly developed and complicated area of physics best left to the more qualified subject matter experts.

Jaynes, as a card carrying physicist, had introduced his MEP within the context of statistical mechanics where, of course, it very naturally belonged. He wanted to bring to his fellow physicists's attention the fact that the MEP was not simply a very handy, albeit niche, mathematical technique developed by thermodynamicists in the 19th Century.

Jaynes had first presented the maximum entropy principle in a paper entitled, **Information Theory and Statistical Mechanics**. His motivation to develop the MEP sprang from a quite evident source.

It was here that he had pointed out that information entropy was actually a vital component necessary for making *inferences*, as opposed to thermodynamic entropy's role as some kind of physical measurement related to, say, an ideal gas. The status of this information entropy was then, (at least to Jaynes's mind), elevated to a loftier consideration than its previous historical development within thermodynamics had heretofore indicated to everyone else in the Physics community.

That circumscribed involvement of the MEP with probability and inferencing was certainly the only thing I had intended to elaborate upon. On a whim, and because Jaynes seemed to refer to him fairly often, I picked up Erwin Schrödinger's 1944 lecture notes delivered a couple of years before Shannon had gotten around to his information entropy work. I thought to myself that there was no way I was going to understand anything by Schrödinger on statistical mechanics.

But I *had* immersed myself in the mathematics of the MEP. So when I started reading Schrödinger, and managed to overcome the initial frustration of translating his idiosyncratic notation into Jaynes's MEP notation, everything was crystal clear. Familiarity with the maximum entropy principle was literally the key to unlocking Schrödinger's physics. It would have been impossible for me to understand word one had I not already been familiar with the MEP as a tool for making numerical assignments to probabilities.

The historical origins of the MEP go back to those famous 19th Century physicists like James Clerk Maxwell, Ludwig Boltzmann, and Josiah Willard Gibbs who first gave us our current classical understanding of thermodynamics and statistical mechanics. In the next Chapter, we shall delve into Erwin Schrödinger's updating of Boltzmann's statistical thermodynamics. As mentioned, this task is made less daunting by our present grasp of the MEP.

As our primary numerical example, we shall feature Boltzmann's famous probability expression for a molecule to possess a certain energy. It is a bit of a happy surprise that Boltzmann's canonical distribution, if first approached from the MEP perspective, is so very straightforward.

Boltzmann's equation comes to us "pre-packaged," so to speak, courtesy of the MEP algorithm. The interesting exercise then becomes matching up the relevant physics with the components of the MEP formula. We have to admit that our version of the MEP was derived from an information perspective, not from physics.

When explained from a strict physics standpoint, it tends to be a lot more involved. Your life becomes a lot easier if you want to think about Boltzmann's probability distribution simply as a curious practical example of a numerical assignment to probabilities in a state space. How to define a state space that captures the essence of the physics then becomes the overriding problem.

Rather than letting the intricacies of statistical mechanics drive the search for probabilities, it is far easier to proceed if, instead, we adopt the mind set that the information resident in some model assigns these probabilities. This is the general perspective which we shall always take when applying the MEP algorithm to any inferential scenario, including statistical mechanics.

More interesting, however, is the allusion the example state space makes to thermodynamics and statistical mechanics. While puzzling over thermodynamics and statistical mechanics, Jaynes realized, in the end, that the MEP was really a general tool helpful in making inferences, as opposed to just being a solution to a physics puzzle.

In the reverse direction in which these things are usually approached, we are adopting the stance that initially looks at statistical mechanics as simply another example of the MEP formula rather than the other way around. One is then more or less obligated to attempt a matching of the generic components of the MEP formalism with the physics embodied within statistical mechanics.

Thus, the generic mapping from statements in the state space to numbers gets translated into a mapping from some fundamental ontological state of an actual object to a physical quantity like energy. The Lagrange multiplier, as a parameter modifying the constraint function, gets translated into Boltzmann's constant and temperature. The objects of inferential interest are no longer coins, dice, kangaroos, or students, but now become atoms, molecules, electrons, photons, and oscillators.

However, despite the many valiant efforts by many informed sources, there do remain residual unexplained curiosities that seem to me, at least, to be glibly passed over in this Procrustean bed that the MEP makes for statistical mechanics.

We will talk about state spaces consisting of statements referring to fundamental physical notions like molecules, energy levels, and temperature. However, if thought about as just another inferential problem where we employ the MEP formula, this kind of introduction to statistical mechanics is surprisingly easy given everything we've done so far.

After we have launched into a simple example of statistical mechanics, a couple of ramifications ensuing from the MEP formula are derived. First, we present a fundamental relationship between the partition function and the constraint function average.

Then, we go into a review of Jaynes's proof that the MEP formula turns into a Legendre transformation. This insight provides a convenient algorithm for computing the parameters as Lagrange multipliers. As you might have come to expect by now, I merely flesh out Jaynes's mathematical proof with some additional steps that make it somewhat more palatable for the beginner.

This is a quantitative demonstration that the probability assigned via the MEP imputes the least amount of extraneous information above and beyond what we want to put into it. Once again, the emphasis is placed on inference and information even though the numerical example involves a fundamental physics problem.

26.2 The Transition to Statistical Mechanics

The kind of introduction to statistical mechanics pursued here could be considered as complementary to the one taken by a typical physics textbook. Rather than leaping immediately into the physics of the situation, we prefer to treat statistical mechanics more from the standpoint of information processing. As a consequence, things remain a little bit more abstract and less physical, but hopefully with a bias towards inferencing as opposed to deduction.

We now possess a good working familiarity with the MEP formula. An IP uses it to assign numerical values to probabilities of joint statements as conditioned on the information resident in some model. In a generic notation, the MEP formula for one constraint is written as,

$$P(X = x_i | \mathcal{M}_k) = \frac{\exp [\lambda F(X = x_i)]}{\sum_{i=1}^n \exp [\lambda F(X = x_i)]} \text{ or as } \frac{e^{\lambda F(x_i)}}{Z(\lambda)} \quad (26.1)$$

In a notation appropriate for further explorations in statistical mechanics, the MEP formula might be altered to look like this,

$$P(E_i | \mathcal{M}_k) = \frac{e^{-E_i/kT}}{Z(T)} \quad (26.2)$$

where the mapping from the statements to numbers is a mapping to an energy $F(x_i) \equiv E_i$. The model's one parameter becomes a function of the temperature T and Boltzmann's constant k , so that $\lambda \equiv -(1/kT)$. In this form, we have Boltzmann's equation.

As mentioned already in the **Introduction**, adopt the reverse stance that might be taken in a physics textbook. Try to think about statistical mechanics as a *direct* application of the MEP as it has been developed so far.

Therefore, Boltzmann's canonical distribution in Equation (26.2) is just another example of the MEP formula assigning a numerical value to the probability of some statement ($X = x_i$). Here such a statement must refer to a physical state, for example, the energy, of a *single* gas molecule given the information about the temperature as inserted by this *one* model.

Suppose we want the probability for the occupation numbers N_1, N_2, \dots, N_n for the n possible states over some number of trials N , or, in this case, N total molecules. The formal manipulation rules then result in a probability for these occupation numbers of,

$$P(N_1, N_2, \dots, N_n | \mathcal{M}_k) = W(N) Q_1^{N_1} Q_2^{N_2} \cdots Q_n^{N_n} \quad (26.3)$$

But where the confusion enters in a major way is that statistical mechanics wants an *immediate* determination of the single most probable set of occupation numbers via Equation (26.2). This approach is opposed to determining the *probability* for *any* set of occupation numbers via Equation (26.3) by first using Equation (26.2) for a single molecule, and then applying the formal rules for many trials.

It strives to accomplish this objective not through the familiar procedures of the MEP formalism, but rather by immediately assuming an extremely large N , and then maximizing the multiplicity factor subject to constraints on total energy defined by the sample N_i .

The critical details separating these two alternative ways of thinking about the Boltzmann distribution are investigated in the next Chapter. There, Schrödinger repeats Boltzmann's derivation by employing an alternative rationale for an MEP-like formula. Analogous to our previous discussions about the conceptual divide separating information and data, statistical mechanics wants an MEP formula NOT in the way we have derived it, but rather based on frequency counts. So, by looking closely at Schrödinger's explanation, we might be able to discover where things go awry.

26.2.1 Setting up an easy scenario

We want to strip down what could be a very complicated statistical mechanics problem to its barest essentials. To that end, suppose that every molecule can be categorized into one, and only one, of two physical states. The two physical states are defined by the x -coordinate of the molecule as it exists inside a box of length 20 cm. If the x -coordinate of the molecule in the box is between 0 and 10 cm, then it is in state 1. If the x -coordinate is between 10 cm and 20 cm, then it is in state 2.

We want to adhere strictly to the tenets of the MEP formalism at this critical juncture of transitioning to statistical mechanics. Thus, there is an information entropy expression, namely,

$$H(Q_i) = - \sum_{i=1}^n Q_i \ln Q_i$$

which is to be maximized subject to side conditions. Carefully note, however, that there isn't anything called a multiplicity factor $W(N)$ to be maximized. A multiplicity factor can only appear when counting up events over some number of trials N . We have no N at this point, and therefore no $W(N)$.

Likewise, within the MEP formalism we have side conditions looking like,

$$\langle F_j \rangle = \sum_{i=1}^n F_j(X = x_i) Q_i$$

Again, there are no expressions involving frequency counts that define a sample average,

$$\overline{F} = \frac{N_i}{N} \times F(X = x_i)$$

that appear in the MEP formula,

$$P(X = x_i | \mathcal{M}_k) \equiv Q_i = \frac{e^{\lambda F(X=x_i)}}{\sum_{i=1}^n e^{\lambda F(X=x_i)}}$$

Furthermore, we also take the stance that there can be no involvement of N , the N_i , $W(N)$, or expressions like,

$$\overline{E} = \frac{N_i}{N} \times E_i$$

in the transition of our generic MEP formula over to statistical mechanics,

$$P(E_i | k, T) = \frac{e^{-\frac{E_i}{kT}}}{\sum_{i=1}^n e^{-\frac{E_i}{kT}}}$$

We have our two statements constituting the $n = 2$ dimensional state space:

1. $(X = x_1) \equiv$ “The molecule’s x -coordinate in the box is $0 \text{ cm} < x \leq 10 \text{ cm}$.”
2. $(X = x_2) \equiv$ “The molecule’s x -coordinate in the box is $10 \text{ cm} < x \leq 20 \text{ cm}$.”

The function mapping these two statements to a physical energy is $F(X = x_1) = 0$ and $F(X = x_2) = \epsilon$.

Surprisingly, in statistical mechanics we can get by with just one constraint function and one associated constraint function average. The mapping takes any statement in the state space to some number representing the energy as captured in $F(X = x_i) = E_i$. The information as inserted under some model \mathcal{M}_k is the specification of the average energy $\langle E \rangle$, or its dual parameter, the temperature T .

Suppose that the information inserted under some model \mathcal{M}_k is $\langle E \rangle = \epsilon/2$. Then, $Q_1 = Q_2 = 1/2$. Everything works out to our desired objective of constructing an easy statistical mechanics scenario that is exactly like the coin tossing scenario examined in the first two Chapters. More importantly, the construction of this statistical mechanics scenario follows the tenets of the MEP formalism to the letter.

Of course, there is no difficulty whatsoever if attention now shifts to asking about the probability for future frequency counts N_1, N_2, \dots, N_n .¹ What is the probability for observing 10 molecules in the box with an x -coordinate somewhere less than 10 cm, with the remaining 10 molecules in the box with an x -coordinate greater 10 cm?

$$\begin{aligned} P(N_1 = 10, N_2 = 10 \mid \mathcal{M}_k) &= W(N) Q_1^{N_1} Q_2^{N_2} \\ &= W(20) (1/2)^{10} (1/2)^{10} \\ &= 0.176197 \end{aligned}$$

Here is the appropriate place for the appearance of N , $W(N)$, and the N_i .

But if you condition on just one model, you are stuck with it forever! No amount of data can ever change the probabilities for future frequency counts. And we know from our extensive discussions in Chapter Fifteen in Volume I, conditioning on one model as was done above, is exactly the same as averaging over model space explicit in the familiar formal manipulation rule,

$$P(N_1, N_2, \dots, N_n) = \int \cdot \int P(N_1, N_2, \dots, N_n \mid q_1, q_2, \dots, q_n) P(q_1, q_2, \dots, q_n) dq_i$$

¹My apologies for constantly shifting between the notation N_i and M_i for future frequency counts. I have established that M_i is my defined notation for future frequency counts with N_i for already observed past frequency counts, the data. But the M_i notation is so jarring to most people used to seeing N_i that I am forced to be inconsistent.

The expression $P(q_1, q_2, \dots, q_n)$ is the probability for the models. It is represented analytically by a Dirichlet distribution with its α_i parameters. When the two α_i parameters happen to approach extremely large numbers, the integration returns the familiar binomial distribution for the probability of the frequency counts.

In fact, we have stressed the fact that the more correct analysis is to always examine $P(N_1, N_2, \dots, N_n)$, and not just the probability conditioned on one model. We suggest that the IP should calculate the probability of the future frequency counts by averaging over all possible models. Nonetheless, if $\alpha_1 = \alpha_2 = 1,000,000$, for example, then we have essentially the same result as above, the difference being only 0.176197 versus 0.176196.

But the probability for the future frequency counts will change as the specification of the α_i parameters in the Dirichlet distribution affect the model probabilities. For example, when $\alpha_1 = \alpha_2 = 1$, reflecting complete ignorance about the model space, then the probability for any future frequency count, whether it be $N_1 = 0, N_2 = 20$, or $N_1 = N_2 = 10$, or any one of the 21 possible frequency counts, is always,

$$P(N_1, N_2) = \frac{N! (n-1)!}{(N+n-1)!} = 1/21$$

The even more bizarre, but still absolutely correct result after applying the formal rules, occurs when both α_1 and α_2 approach 0 in the Dirichlet distribution for models. Then all molecules cluster with equal probability in either the lowest energy state or the highest energy state,

$$P(N_1 = 0, N_2 = 20) = P(N_1 = 20, N_2 = 0) \approx 1/2$$

Ultimately, though, as empirical scientists, we desire that the probability of future frequency counts (derived, as always, through the formal manipulation rules) be driven exclusively from the consideration of any known data. Thus, we have the important formula,

$$P(M_1, M_2, \dots, M_n | N_1, N_2, \dots, N_n) = C \times \frac{\prod_{i=1}^n (M_i + N_i)!}{\prod_{i=1}^n M_i!}$$

where now we have reverted to the correct notation where the M_i are the future frequency counts and the N_i are the data.

Continuing with the same numerical example, we want the probability for seeing 10 future molecules in the first half of the box, and 10 future molecules in the second half of the box. This probability is conditioned on the fact that we have already collected data on 20 molecules where the split was an even 10 and 10 molecules in both sides of the box,

$$P(M_1 = 10, M_2 = 10 | N_1 = 10, N_2 = 10) = 0.126834$$

This quantitative result makes perfect sense as it based on data consisting of a relatively small sample size of only $N = 20$. Certainly the available data do support the idea that $Q_1 = Q_2 = 1/2$.

The above probability of 0.126834 *should* be less than the probability of 0.176197. The IP is less certain about the truth of 10 molecules in each side of the box because the inference is based on a limited amount of data. The larger probability is based on an “infinite” amount of virtual data. In other words, a single model is adopted whose probability is represented by the Dirac δ -function, $P(\mathcal{M}_k) = \delta(q_1 - 1/2, q_2 - 1/2)$.

26.2.2 Another numerical example of Boltzmann’s equation

Let’s move on from this bare bones example to something just a little bit more interesting. Since a molecule in the box could be located within three dimensional space with x , y , and z coordinates, consider extending the dimension of the state from $n = 2$ to $n = 8$. The volume of the box is now 20 cm \times 20 cm \times 20 cm with statements about a molecule being located in the first 10 cm or the second 10 cm for each of the three dimensions.

The mapping from each of the eight statements to an energy takes place as,

$$F(X = x_i) = E_i$$

with $E_1 = 0$, $E_2 = 1$, and so on. Each E_i will be divided by $-(kT)$ where the information inserted under some specific model is that $kT = 2$. In all of this, we have been scrupulously following the MEP formalism. As a consequence then, we have that,

$$P(E_i | k, T) = \frac{e^{-\frac{E_i}{kT}}}{\sum_{i=1}^8 e^{-\frac{E_i}{kT}}} \quad (26.4)$$

This equation just derived so effortlessly simply as an example of the MEP, is, of course, Boltzmann’s probability distribution for the energy levels of a molecule.

Table 26.1 at the top of the next page contains the details for an example, using Equation (26.4), that calculates the numerical assignment for the probability of a single molecule to possess one of these eight energy levels. We are supposing that the information inserted under some model is that $kT = 2$.

Setting $kT = 2$ dictates that the average energy is $\langle E \rangle = 1.3922$. This seemingly arbitrary and unmotivated number for the average energy corresponds to its dual parameter setting of $\lambda = -\frac{1}{kT} = -\frac{1}{2}$. Here, we choose a convenient setting for one parameter, which then forces the dual parameter to adjust to an inconvenient number. The point is that we are justified in saying that the information inserted under the model could be expressed by T , or just as well by an average energy $\langle E \rangle$.

The eight energy levels, E_1 through E_8 , are listed in the first column. This is the mapping, $F(X = x_i)$, from the statements in the state space to numbers. The second column gives each energy level multiplied by the Lagrange multiplier.

Table 26.1: Using the MEP formula to assign probabilities to energy levels in a statistical mechanics problem where $kT = 2$.

E_i	$-(E_i/kT)$	$e^{-E_i/kT}$	Q_i	$Q_i \times E_i$
0	0.0	1.0000	0.4008	0.0000
1	-0.5	0.6065	0.2431	0.2431
2	-1.0	0.3679	0.1475	0.2949
3	-1.5	0.2231	0.0894	0.2683
4	-2.0	0.1353	0.0542	0.2170
5	-2.5	0.0821	0.0329	0.1645
6	-3.0	0.0498	0.0200	0.1197
7	-3.5	0.0302	0.0121	0.0847
$Z = \sum_{i=1}^8 e^{-E_i/kT} = 2.4949$		$\sum Q_i = 1.00$	$\langle E \rangle = 1.3922$	

The third column contains the value of the numerator in the MEP formula for each Q_i . The partition function Z is the sum of all eight terms in the third column. It is shown as the first entry in the last row. When each value in the third column is divided by this sum, we have the Q_i as shown in the fourth column.

Notice first and foremost that the Q_i sum to 1, and that the average energy is 1.3922. This assignment satisfies the two constraints as given in the problem, and possesses the additional feature that it has the maximum entropy of *any* assignment that satisfies these two constraints.

The information entropy is a quantitative measure of the amount of missing information. Our purpose in maximizing the entropy was to maximize the amount of missing information, and in the process, to ensure that no additional information other than the information about the average energy, or the temperature, made its way into the probability assignment.

The Q_i progress downward in the typical exponential fashion that we have come to expect with only one Lagrange multiplier. Since λ was negative, the larger probabilities occur at the lower energy levels.

The final column computes the individual components, $E_i \times Q_i$, of the average energy $\langle E \rangle$. In these numerical exercises, we choose to pick a convenient Lagrange multiplier like $\lambda = -\frac{1}{2}$, and then live with whatever average energy this produces.

Again, this is the dual parameter flexibility inherent in the MEP where either the average energy, or the temperature can be given as the constraint. It is just easier for presentation purposes to select a nice round temperature, and then claim to know the average energy was 1.3922 instead of the other way around.

Now that we have established a numerical assignment for the energy probabilities that a single molecule might possess, we are ready to examine the probability for any number N of identical molecules. How many possible occupancy counts are there?

There were 21 possible counts for the two dimensional state space and 20 molecules. In our larger state space with, say, $N = 200$ molecules, there are an enormous number of possibilities, nearly three trillion, in fact, as revealed by the calculation,

$$\text{Possible occupancy numbers} = \frac{(N + n - 1)!}{N! (n - 1)!} = \frac{207!}{200! 7!} \approx 2.9 \times 10^{12}$$

This number, as large as it is, is dwarfed by the size of the sample space which weighs in at about,

$$n^N = 8^{200} \approx 4.1 \times 10^{180}$$

elementary points. We could, in principle, account for every one of these elementary points as we have done in the past for the dice and kangaroos, by multiplying each possible occupancy number by its multiplicity factor.

We will take a look at the probabilities for a few of these possible occupancy counts from the vast number available to us. Assume that it takes energy to shove a molecule to the right, to the back, and to the top of the box. Intuition here will gibe with actual quantitative calculations of the probabilities.

The probability for seeing all 200 molecules in the right, back, upper sector of the volume of the box, that is, the probability that all 200 molecules possess the highest possible energy, is infinitesimally small,

$$P(N_1 = 0, \dots, N_8 = 200 | \mathcal{M}_k) = \frac{200!}{0! 0! \dots 200!} \times (0.4008)^0 \times \dots \times (0.0121)^{200} \approx 3.6 \times 10^{-384}$$

The probability for seeing all 200 molecules in the left, front, lower sector of the volume of the box, that is, the probability that all 200 molecules possess the lowest possible energy, has much greater probability, but is also infinitesimally small at a probability about 10^{-80} .

As you might expect, the largest probabilities are centered within that class of possible occupancy counts that look something like,

$$P(N_1 = 80, \dots, N_8 = 2 | \mathcal{M}_k) = \frac{200!}{80! \dots 2!} \times (0.4008)^{80} \times \dots \times (0.0121)^2 \approx 5.51 \times 10^{-7}$$

Examine Table 26.2 below, and Table 26.3 on the next page, for a few representative frequency counts of the 200 molecules at each of the eight energy levels. The first table lists four occupation numbers that have a very high relative probability. The count in the first column is the most probable given this particular model. But the probabilities for these four quite similar frequency counts all hover around 10^{-7} . The sample average energy \bar{E} , as based on the particular occupation numbers, and the probability for the j^{th} sampling, $P(N_i^j)$, are computed in the last two rows.

Table 26.2: *The first set of a few of the close to three trillion possible frequency counts of 200 molecules at eight energy levels. These counts all have a relatively high probability of occurring.*

i	N_i^1	N_i^2	N_i^3	N_i^4
1	80	79	84	83
2	49	50	52	51
3	29	28	26	25
4	18	18	15	16
5	11	12	8	9
6	7	6	7	7
7	4	5	5	6
8	2	2	3	3
<i>Sum</i>	200	200	200	200
\bar{E}	1.390	1.410	1.335	1.385
$P(N_i^j)$	$\approx 5.51 \times 10^{-7}$	$\approx 4.04 \times 10^{-7}$	$\approx 1.56 \times 10^{-7}$	$\approx 1.38 \times 10^{-7}$

The second table lists another four occupation numbers that all have very low relative probability. They are presented in the order of their decreasing probability. Our already computed frequency count of 200 molecules all occupying the highest energy level has by far the lowest probability under this model.

Were an average of the sample energies \bar{E} to be taken, and weighted by their respective relative probabilities, it looks as if this value would be close to the model's specified mathematical expectation of the energy $\langle E \rangle = 1.3922$. We have sample energies in the first set like 1.39, 1.41, 1.335, and 1.385 which are weighted highly. Then we have sample energies in the second set like 1.225, 1.5, 0, and 7 which would hardly be counted at all.

Table 26.3: The second set of a few of the close to three trillion possible frequency counts of 200 molecules at eight energy levels. These counts all have a relatively low probability of occurring.

i	N_i^5	N_i^6	N_i^7	N_i^8
1	70	50	200	0
2	60	50	0	0
3	40	50	0	0
4	15	50	0	0
5	15	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	200
<i>Sum</i>	200	200	200	200
\bar{E}	1.225	1.500	0	7
$P(N_i^j)$	$\approx 10^{-12}$	$\approx 10^{-28}$	$\approx 10^{-80}$	$\approx 10^{-384}$

As a very rough crude estimate,

$$\langle \bar{E} \rangle \approx \frac{(1.39 \times 5) + (1.410 \times 4) + (1.385 \times 1) + (1.335 \times 1) + \cdots + (7 \times 0)}{11} = 1.3918$$

Moreover, the dispersion around this average over the highly probable frequency counts would not be very great. Packaging this argument up, we would conjecture that as N gets larger and larger, the average of the energy computed from the occupation numbers with respect to $P(N_1, N_2, \dots, N_n)$ would approach the average energy specified by the MEP model,

$$\lim_{N \rightarrow \infty} \langle \bar{E} \rangle_{P(N_1, N_2, \dots, N_n)} \rightarrow \langle E \rangle_{Q_i}$$

More precisely,

$$\begin{aligned} \lim_{N \rightarrow \infty} \sum_{j=1}^{\frac{(N+n-1)!}{(n-1)! N!}} \left[P(N_1^j, N_2^j, \dots, N_n^j) \times \sum_{i=1}^n \left(\frac{N_i}{N} \times E_i \right) \right] &\rightarrow \sum_{i=1}^n E_i Q_i \\ \lim_{N \rightarrow \infty} \sum_{j=1}^{\frac{(N+n-1)!}{(n-1)! N!}} \left[\frac{N!}{N_1^j N_2^j \dots N_n^j} Q_1^{N_1^j} \dots Q_n^{N_n^j} \times \sum_{i=1}^n \left(\frac{N_i}{N} \times E_i \right) \right] &\rightarrow \langle E \rangle \end{aligned}$$

26.2.3 Dimensionless numbers

There is one sticking point that we have to clear up in emphasizing the informational approach to Boltzmann's distribution. The generic argument to the exponential function in the MEP formula, $\lambda F(X = x_i)$, must be a dimensionless number. But the physics will naturally introduce numbers involving lengths, energy, time, cycles per second, temperature, and so on.

Fortunately, a dimensional analysis reveals that these units of measurement all cancel out in the end. Suppose that the energy E_i appearing in the numerator of the MEP formula, $e^{-E_i/kT}$, is measured in *ergs*. The temperature, T , measured in *degrees Kelvin*, appears in the denominator.

Now comes the crucial part. In what measurements is Boltzmann's constant k defined? It is given in *ergs per degree K*. So, T cancels out in the denominator, leaving just *ergs* in both numerator and denominator which then cancel out leaving us with the dimensionless number required.

26.3 Information and Data “Reconciled”

Permit me, at this point in our journey, to mitigate the “tempest in a teapot” that I generated in Chapters Twenty and Twenty One over my own enforced conceptual distinction between *information* and *data*. In doing so, let me also moderate my rather intemperate tirade against Jaynes in misinterpreting the distinction between data and information.

Conceptually, and from a very strict adherence to the rules of the game as we have set them up, everything I said about the confusion swirling over data and information is correct. However, with this foray into statistical mechanics, we do see how data and information, with some voluntary agreement to blur over the conceptual niceties, might harmoniously co-exist.

In statistical mechanics, just *one* model for assigning numerical values to the probabilities of energies is assumed true right at the outset. That is, the immediate acceptance of the Boltzmann distribution is not questioned. The information in that one model is the specification of an average energy $\langle E \rangle$ or, dually, the specification of the temperature T .

No observations are made on the energy levels for a huge number of molecules, no dice are rolled 20,000 times, no kangaroos are sampled to discover a “good” model from all possible models. The measurements do not revolve about any N_i , but rather about a very precise measurement of T , or \overline{E} , and this, in fact, is thought of as the *data*!

But adhering to our fundamental MEP principles, the temperature T must be a *parameter*, not *data*! Therefore, when T is measured, it is assumed in statistical

mechanics that the number of molecules distribute themselves over the energy levels such that the sample energy equals this T and that, furthermore, the molecules distribute themselves where they possess the greatest number of ways of satisfying the sample energy.

It might have already occurred to you when first discussing the coins and the dice, whether it was really necessary to go to all that trouble of tossing a coin or rolling a die an enormous number of times. Couldn't you just *physically* examine the coin or the dice in question, and then determine whether they possessed any peculiarities in manufacture or mode of tossing that would eliminate all that labor? So, rather than roll a die 20,000 times to see whether the model incorporating information about a displaced center of gravity and unequal lengths of the axes of the cube is supported by the data, simply go ahead and make precise measurements of these physical characteristics.

That's exactly the same stance statistical mechanics takes. There are no measurements of the varying energy levels of very many molecules with a tabulation of occupancy counts. The temperature T is measured, and then on the basis of the enormous value of N for the number of molecules in a volume of gas, finding probabilities based on $\langle E \rangle$ via Boltzmann's equation are going to be the same as the probabilities based on $\langle \bar{E} \rangle$ averaged with respect to $P(N_1^j, N_2^j, \dots, N_n^j)$.

But typical of the confusion rampant in this field, this distinction is not made clear. To make matters worse, making a precise measurement of the temperature as *data* is confused with a tentative specification of an average energy under one model as *information*. Furthermore, information is never measured, but a temperature is! Can the temperature then be considered as a parameter as it surely is within the MEP formalism?

We have to clearly settle the question of whether we are gathering data as defined by the N_i to reduce the model space where the α_i started out equal at 1, or whether the α_i started out at some extremely large number leading to a Dirac δ function for just one model. Here we have just another situation of the insidious conflation of the concepts of data and information. In the end, I am simply making a heartfelt plea for *disambiguation*!

A partial summing up of any lessons learned concerning the distinction between the formal tenets of the MEP and statistical mechanics must focus on this issue: The MEP serves the role of inserting information into a probability distribution by some model for the *sole purpose of tentative consideration*. All conceivable models are envisioned as variations on this theme of tentative consideration, and all models are treated on an equal basis at the outset. In contrast, in statistical mechanics *one* MEP model is judged by physical considerations to be the *only* logical candidate for a model. Tentative models and winnowing of the model space by data are seen in a completely different light within statistical mechanics.

The idea of “data reconciliation” is kindred in spirit to statistical mechanics. The core idea is that the data are matched to the information in the constraint functions of just one model. One example discussed how maximum likelihood was related to the MEP in logistic regression (Chapter Twenty Three). The rationale for this linkage between core Bayesian notions and maximum likelihood was explained by Jaynes, and will be discussed shortly in the next Chapter (Exercise 27.7.20).

26.3.1 More entropic-like formulas

If we were to take \bar{E} , or T , as the known data, then setting,

$$\lim_{N \rightarrow \infty} \langle \bar{E} \rangle_{P(N_1, N_2, \dots, N_n)} \rightarrow \langle E \rangle_{Q_i}$$

is perfectly acceptable whenever there is that implicit assumption that $N \rightarrow \infty$. Remember that back in the exercises for Chapter Seventeen, an alternative formula for the probability of the frequency counts when conditioned on one model was developed as,

$$P(N_1, N_2, \dots, N_n | \mathcal{M}_k) = \exp \left\{ N \left[\left(\frac{\ln W(N)}{N} \right) + \lambda \bar{F} - \ln Z \right] \right\}$$

As we shall see in Exercise 26.6.16, the $\frac{\ln W(N)}{N}$ term can be replaced by an entropic-like equivalent $H(f_1, f_2, \dots, f_n) \equiv H_1$. Add and subtract the term $\lambda \langle F \rangle$ in order to get to our familiar $H_{max}(Q_i) = \ln Z - \lambda \langle F \rangle \equiv H_0$. Now we have,

$$\begin{aligned} P(N_1, N_2, \dots, N_n | \mathcal{M}_k) &\approx \exp \{N [H(f_1, f_2, \dots, f_n) - \ln Z + \lambda \langle F \rangle - \lambda \langle F \rangle + \lambda \bar{F}] \} \\ &\approx \exp [N \times (H_1 - H_0 - \lambda \langle F \rangle + \lambda \bar{F})] \end{aligned}$$

As our numerical investigations have borne out, as $N \rightarrow \infty$ we would expect that the difference $\lambda(\bar{F} - \langle F \rangle)$ would approach 0. So, in the end, we are left with two entropy terms H_1 and H_0 that must be getting closer and closer in value.

$$P(N_1, N_2, \dots, N_n | \mathcal{M}_k) \rightarrow \exp [N \times (H_1 - H_0)]$$

$$P(N_1, N_2, \dots, N_n | \mathcal{M}_k) \rightarrow 1 \quad \text{Conjecture}$$

But clearly this must be wrong. As $N \rightarrow \infty$, the growth in the possible number of frequency counts N_1, N_2, \dots, N_n must be enormous. The probability for any one of these as N becomes increasingly large must get smaller and smaller. Thus, the only conclusion is that the approximation afforded by the frequency entropy $H(f_1, f_2, \dots, f_n)$ to the function of the multiplicity factor is very dangerous when multiplied by extremely large N . Exercise 26.6.17 presents a numerical example of what typically happens to the probability of the expected frequency count as N becomes increasingly large.

26.4 Interesting MEP Relationships

Now that we have looked directly at simple examples of Boltzmann's equation, there are a couple more very interesting relationships inherent in the MEP that should be discussed in detail. The first of these is the fact that the partial derivative of the logarithmic transform of the partition function with respect to the Lagrange multiplier is equal to the constraint function average associated with that multiplier.

The equation, in this case, is actually easier to comprehend than the verbose description just given. At first, we shall not prove this relationship symbolically, but just give a numerical example that clearly shows what it is all about.

The second relationship is the so-called Legendre transformation discussed in Chapter Twenty Four. These relationships inherent in the MEP were brought up time and time again whenever Jaynes had the chance for another go around in trying to explain the MEP. They also prominently appear, but in a much more obscure manner, within statistical mechanics. We will try to illustrate this difference in comprehensibility in the next Chapter when we discuss Schrödinger's approach.

26.4.1 Partition function and constraint function average

The mathematical equation encapsulating the verbose description given above is,

$$\frac{\partial \ln Z(\lambda)}{\partial \lambda} = \langle F \rangle \quad (26.5)$$

The partition function, while seemingly a bit player in its role as simply a normalizing factor for probability assignments, is actually a key element within the physical interpretation. Partial derivatives of various orders of Z play a central role.

Consider the partition function for a model with just one constraint,

$$Z(\lambda) = \sum_{i=1}^n \exp [\lambda F(X = x_i)]$$

And, to make our numerical example even easier still, we shall consider a state space of dimension $n = 2$. Suppose that the constraint function mapping the statement $(X = x_1)$ is $F(X = x_1) = 2$, and the constraint function mapping the statement $(X = x_2)$ is $F(X = x_2) = 4$.

Now, we have the two numerical assignments under some model, where λ is left unspecified, as,

$$\begin{aligned} Q_1 &= \frac{e^{2\lambda}}{e^{2\lambda} + e^{4\lambda}} \\ Q_2 &= \frac{e^{4\lambda}}{e^{2\lambda} + e^{4\lambda}} \end{aligned}$$

The constraint function average is by definition,

$$\begin{aligned}\langle F \rangle &= F(X = x_1) Q_1 + F(X = x_2) Q_2 \\ &= \left[2 \times \frac{e^{2\lambda}}{e^{2\lambda} + e^{4\lambda}} \right] + \left[4 \times \frac{e^{4\lambda}}{e^{2\lambda} + e^{4\lambda}} \right] \\ &= \frac{2e^{2\lambda} + 4e^{4\lambda}}{e^{2\lambda} + e^{4\lambda}}\end{aligned}\tag{26.6}$$

But if you now ask *Mathematica* to do the hard work involved in the partial differentiation represented by $\frac{\partial \ln Z(\lambda)}{\partial \lambda}$ you will find that,

```
D[Log[Total[Exp[\lambda {2,4}]]], \lambda]
```

evaluates to the answer just provided for $\langle F \rangle$ in Equation (26.6). This is an initial numerical confirmation that indeed,

$$\frac{\partial \ln Z(\lambda)}{\partial \lambda} = \langle F \rangle$$

Upcoming exercises, as well as ones already worked on in previous Chapters, present further confirmation verifying this relationship.

Here is the more general statement of this relationship. The partial derivative of the log of the partition function with respect to the j^{th} Lagrange multiplier is equal to the j^{th} constraint function average.

$$\frac{\partial \ln Z(\lambda_1, \dots, \lambda_j, \dots, \lambda_m)}{\partial \lambda_j} = \langle F_j \rangle\tag{26.7}$$

26.4.2 Information entropy of the MEP assignment

We have profitably employed the MEP algorithm to find numerical assignments for the probabilities required in inferences involved in tossing coins, rolling dice, assessing traits of kangaroos, and now the energy levels of molecules constituting an abstract kind of gas. We asked the question of what happens to the information entropy when more and more constraints are added.

In Chapter Nineteen dealing with the dice, the MEP algorithm led to the particularly intuitive assignment of $Q_i = 1/6$ when the only constraint was the universal constraint. The information entropy for this assignment of the Q_i , under model \mathcal{M}_A of a fair die, worked out to,

$$H(Q_i | \mathcal{M}_A) = \ln n = \ln 6 = 1.7918$$

When compared to the information entropy in the model of an unfair die, say, the most unfair model we examined, model \mathcal{M}_D , the entropy dropped to a value of,

$$H(Q_i | \mathcal{M}_D) = 1.7057$$

We are certainly entitled to make the empirical observation that the entropy of the assigned distribution with no constraints, model \mathcal{M}_A , is higher than the entropy of the assigned distribution with three constraints, model \mathcal{M}_D .

Is this true in general? What if we continue to add more constraints? Does the entropy of the distribution get smaller and smaller when the numerical assignment made via the MEP is enforcing more constraints?

We would like to have a formula that tells us the information entropy, not in terms of the usual formula,

$$H(Q_i | \mathcal{M}_k) = - \sum_{i=1}^n Q_i \ln Q_i$$

but rather as a function of the significant components of the MEP formula.

We will see that adding any information to that already specified by the IP would reduce the information entropy. The logical end of this argument is to arrive at an assignment where there is no missing information, the information entropy of the probability distribution is 0, and probabilities of 1 and 0s represent certainty about the statements in the state space.

For example, in the first case where we looked at just one constraint beyond the universal constraint, we discovered in Chapter Twenty Four that the information entropy of the MEP assignment to the Q_i can be expressed alternatively using the Legendre transformation as,

$$H_{max}(Q_i | \mathcal{M}_k) = \min_{\lambda} [\ln Z(\lambda) - \lambda \langle F \rangle]$$

This gives us a relationship where we can examine what happens to the information entropy as the partition function, the Lagrange multiplier, and the average of the constraint function change.

26.4.3 Information entropy emphasizing the MEP formula

Any standard explanation of the Lagrange multiplier method of finding the extrema of a function subject to side constraints will always emphasize the fact that many potential solutions will be found, and they have to be checked to see which of the candidates are actual solutions. Jaynes included the following alternative derivation as complementary to the variational argument as given in Chapter Twenty Five so that, considering the two derivations side by side, we could see that the MEP formula was a compelling solution.

To begin, write down the definition of information entropy for the Q_i ,

$$H(Q_i | \mathcal{M}_k) = - \sum_{i=1}^n Q_i \ln Q_i$$

As we investigated in some detail in the last Chapter, the MEP formula provides us with a means for calculating the Q_i with one piece of information,

$$Q_i = \frac{e^{\lambda F(X = x_i)}}{\sum_{i=1}^n e^{\lambda F(X = x_i)}}$$

Take the logarithmic transform of the Q_i so that it can be substituted into the information entropy formula. Carrying out such a transform on the MEP definition of the Q_i leaves us with,

$$\begin{aligned}\ln Q_i &= \ln \left[\frac{e^{\lambda F(X = x_i)}}{\sum_{i=1}^n e^{\lambda F(X = x_i)}} \right] \\ &= \ln \left[\frac{e^{\lambda F(X = x_i)}}{Z(\lambda)} \right] \\ &= \ln [e^{\lambda F(X = x_i)}] - \ln [Z(\lambda)] \\ &= \lambda F(X = x_i) - \ln [Z(\lambda)]\end{aligned}$$

Having derived this expression for $\ln Q_i$, we substitute it into the information entropy definition,

$$H(Q_i | \mathcal{M}_k) = - \sum_{i=1}^n Q_i \ln Q_i$$

with the result,

$$H(Q_i | \mathcal{M}_k) = - \sum_{i=1}^n Q_i [\lambda F(X = x_i) - \ln Z(\lambda)]$$

The next basic step is to multiply the term in brackets by the Q_i and then distribute the summation sign. The indices on the summation symbol are omitted in subsequent steps.

$$\begin{aligned}H(Q_i | \mathcal{M}_k) &= - \left[\sum Q_i [\lambda F(X = x_i) - \ln Z(\lambda)] \right] \\ &= - \left[\sum Q_i \lambda F(X = x_i) - \sum Q_i \ln Z(\lambda) \right]\end{aligned}$$

Turn your attention to the second term in the last expression,

$$\sum Q_i \ln Z(\lambda)$$

Since $\ln Z(\lambda)$ is a constant, we can bring it outside the summation sign. But then we recognize that $\sum Q_i = 1$ leaving us with,

$$\begin{aligned}\sum Q_i \lambda F(X = x_i) - \sum Q_i \ln Z(\lambda) &= \sum Q_i \lambda F(X = x_i) - \ln Z(\lambda) \sum Q_i \\ &= \sum Q_i \lambda F(X = x_i) - \ln Z(\lambda)\end{aligned}$$

We are deep within the bowels of the tiresome minutiae dictated by the knowledge uncertainty principle as mentioned in my *Apologia* for this Volume. If you have remained with me this far, I assume you desire to make a trade of your time and concentration to avoid an excessive reliance on faith for your knowledge.

At this point, we switch attention to the first term. Since λ is also a constant within a summation, it too can be brought out. After exchanging the two terms the result is,

$$H(Q_i | \mathcal{M}_k) = -[-\ln Z(\lambda) + \lambda \sum F(X = x_i) Q_i]$$

The expression that is multiplied by λ , namely, $\sum F(X = x_i) Q_i$, is just the constraint function average, $\langle F \rangle$. We substitute this into the above expression, and take care of the minus signs,

$$H(Q_i | \mathcal{M}_k) = \ln Z(\lambda) - \lambda \langle F \rangle \quad (26.8)$$

The generalization to m constraints is,

$$H_{max}(Q_i | \mathcal{M}_k) = \min_{\lambda_j} [\ln Z(\lambda_1, \lambda_2, \dots, \lambda_m) - \sum_{j=1}^m \lambda_j \langle F_j \rangle] \quad (26.9)$$

We have finally accomplished one of our goals. We have a formula that tells us what the entropy of a distribution is in terms of the partition function, the Lagrange multipliers, and the constraint function averages that were supplied as part of the information. If all this symbolic shuffling leaves you a little uneasy, then work out some of the exercises to allay your fears.

26.4.4 Another derivation of the same result

Here is another way to derive the same result. Earlier in Chapter Twenty Five, we derived the partition function in the MEP assignment for the Q_i by eliminating $\lambda_0 - 1$. But if we return to the original expression for the Q_i that includes $\lambda_0 - 1$, we have,

$$Q_i = e^{\lambda_0 - 1 + \lambda_1 F_1(X = x_i)}$$

The log transform yields,

$$\ln Q_i = \lambda_0 - 1 + \lambda_1 F_1(X = x_i)$$

As a result, we have for the entropy of the Q_i ,

$$-\sum_{i=1}^n Q_i \ln Q_i = -\sum_{i=1}^n [Q_i (\lambda_0 - 1 + \lambda_1 F_1(X = x_i))]$$

Multiplying the expression in parentheses by Q_i results in,

$$-\sum_{i=1}^n Q_i \ln Q_i = -\sum_{i=1}^n [Q_i \lambda_0 - Q_i + Q_i \lambda_1 F_1(X = x_i)]$$

Next, distribute the summation sign across all the terms in this equation, paying attention to the leading minus sign,

$$-\sum_{i=1}^n Q_i \ln Q_i = -\sum_{i=1}^n Q_i \lambda_0 + \sum_{i=1}^n Q_i - \sum_{i=1}^n Q_i \lambda_1 F_1(X = x_i)$$

We now carry out three easy operations on the derivation as it now stands:

1. Bring out the constant, λ_0 , from inside the summation of the first term,
2. Substitute 1 for the second term, and,
3. Bring the constant, λ_1 , outside the summation sign in the third term.

After doing this, we have at this stage,

$$-\sum_{i=1}^n Q_i \ln Q_i = -\lambda_0 \sum_{i=1}^n Q_i + 1 - \lambda_1 \sum_{i=1}^n F_1(X = x_i) Q_i$$

At this point in the derivation, things are looking fairly easy as we again substitute 1 for $\sum Q_i$ in the first term, and plug in the notation for the constraint function average in the third term,

$$-\sum_{i=1}^n Q_i \ln Q_i = -\lambda_0 + 1 - \lambda_1 \langle F_1 \rangle \quad (26.10)$$

We have nearly demonstrated the equality of this derivation where $\lambda_0 - 1$ is included, with the previous derivation that absorbed it into the partition function. We merely have to work on the $-\lambda_0 + 1$ term and show that it is equal to $\ln Z(\lambda_1)$ for the derivation to be complete.

In the last Chapter, $\lambda_0 - 1$ was manipulated out of the MEP assignment for the Q_i by creating $Z(\lambda_1)$. Here is the very similar approach showing that $\ln Z = -\lambda_0 + 1$,

$$\begin{aligned} \exp(\lambda_0 - 1) &= \frac{1}{\sum_{i=1}^n \exp[\lambda_1 F_1(X = x_i)]} \\ \exp(\lambda_0 - 1)^{-1} &= \sum_{i=1}^n \exp[\lambda_1 F_1(X = x_i)] \\ \exp(\lambda_0 - 1)^{-1} &= \exp(-\lambda_0 + 1) \\ \exp(-\lambda_0 + 1) &= \sum_{i=1}^n \exp[\lambda_1 F_1(X = x_i)] \\ \exp(-\lambda_0 + 1) &= Z(\lambda_1) \\ \ln Z(\lambda_1) &= -\lambda_0 + 1 \end{aligned}$$

For the final step, substitute this result just found into Equation (26.10),

$$H(Q_i | \mathcal{M}_k) = \ln Z(\lambda_1) - \lambda_1 \langle F_1 \rangle$$

to verify that this is the same formula² previously derived.

26.5 Connections to the Literature

The equation computing the probability of an ideal gas molecule to occupy the i^{th} energy cell is one of the most famous in all of physics. It is called the *Boltzmann distribution* after one of the 19th Century's most well-known physicists, Ludwig Eduard Boltzmann. Born in 1844, Boltzmann was Austrian by nationality.

He is credited by Jaynes with being the first physicist to tackle the problem of describing a *state of knowledge* with a probability assignment. Harold Jeffreys was perhaps the second most famous personage, after Boltzmann, to emphasize this central epistemological role of probability.

As is well known even to many non-physicists, Boltzmann was a tragic figure in the history of science. Some of his commentators have claimed that Boltzmann's suicide in 1906 was due to his failure to convince his contemporaries of the correctness of his viewpoint based on the then unobservable atoms.

Now, however, one sees more nuanced accounts that Boltzmann, in fact, was highly lauded by his peers. He seems to have been respected far and wide, even for his then more contentious pronouncements on statistical mechanics. It is more likely that he hanged himself after suffering from a severe attack of an inherent depression, not directly related to his scientific struggles.

Experimental justification for his life's work was forthcoming shortly after his death. The sadness surrounding the lack of appreciation for what later was recognized as true genius reminds one of Boltzmann's contemporary, the impressionist artist Vincent Van Gogh. Boltzmann's famous expression for entropy is carved on his tombstone.

Jaynes [14, pg. 9] tells us,

... the theory of maximum entropy inference is identical in mathematical form with the rules of calculation provided by statistical mechanics. ... Then if we know only the average energy $\langle E \rangle$, the maximum entropy probabilities of the levels E_i are given by a special case of [the MEP formula for assignment of numerical values to probabilities] which we recognize as the Boltzmann distribution.

²Some may wonder, (except for those seeing these things for the very first time), why I often go to the bother of showing every single small step in a derivation. I have been strongly influenced by Wolfram's wonderful explanation of theorem proving where single symbolic substitutions, applied across the board from a restricted set of rules that describe all allowable substitutions, are the paramount idea behind theorem proofs. You can see Wolfram emphasizing such a concept over traditional mathematical theorem proving in his multiway substitution systems.

Ruhla [29], Chapter 5, contains an excellent elementary introduction to the Boltzmann distribution. My treatment, as already mentioned, can perhaps be seen as complementary to any standard presentation in the physics literature.

But the advantage of studying the MEP is that the Boltzmann distribution arrives, as I mentioned above, already “pre-packaged.” The student can concentrate on the overriding physical picture without having to worry about the derivation. He or she already has it in hand, before the physics intrudes, as simply the assignment that the MEP algorithm would make strictly on the grounds of information and probability.

One of the most perplexing aspects of the Boltzmann distribution (at least to physicists) is that it seems to ignore the dynamics in any physical situation. Indeed, this is its major strength because the dynamics are impossible to follow for the number of molecules involved. Jaynes call this seeming problem “getting something for nothing.”

As a matter of fact, the principle of maximum entropy was attacked for the very same reason. When using the MEP, it seemed that inference was being built upon a foundation of sand that exploited ignorance rather than knowledge, and this did not set well with many. Jaynes [18, pp. 226–227] presents a convincing answer to these criticisms which we now repeat for its power and clarity.

From Boltzmann’s reasoning, then, we get a very unexpected and nontrivial dynamical prediction by an analysis that, seemingly, ignores the dynamics altogether! This is only the first of many such examples where it appears that we are “getting something for nothing,” the answer coming too easily to believe. Poincaré, in his essays on “Science and Method,” felt this paradox very keenly, and wondered how by exploiting our ignorance we can make correct predictions in a few lines of calculation, that would be quite impossible to obtain if we attempted a detailed calculation of the 10^{23} individual trajectories.

It requires very deep thought to understand why we are not, in this argument and others to come, getting something for nothing. In fact, Boltzmann’s argument *does* take the dynamics into account, but in a very efficient manner. Information about the dynamics entered his equations at two places. (1) the conservation of total energy; and (2) the fact that he defined his cells in terms of phase volume, which is conserved in the dynamical motion (Liouville’s theorem). The fact that this was enough to predict the correct spatial and velocity distribution on the molecules shows that the millions of intricate dynamical details that were not taken into account, *were actually irrelevant to the predictions, and would have canceled out anyways if he had taken the trouble to calculate them.*

Boltzmann’s reasoning was super-efficient; far more so than he ever realized. Whether by luck or by inspiration, he put into his equations *only* the dynamical information that happened to be relevant to the questions he was asking. Obviously, it would be of some importance to discover the secret of how this came about, and to understand it so well that we can exploit it in other problems.

If we can learn how to recognize and remove irrelevant information at the beginning of a problem, we shall be spared having to carry out immense calculations, only to discover at the end that practically everything we calculated was irrelevant to the question we were asking. And that is actually what we are after by applying Information Theory ... Boltzmann was asking only questions about *experimentally reproducible equilibrium properties*.

[Emphasis in the original.]

Because of its overwhelming conceptual importance, I think it necessary to repeat Jaynes's compelling plea that one should try to decipher the fundamental power of Boltzmann's arguments.

If we can learn how to recognize and remove irrelevant information at the beginning of a problem, we shall be spared having to carry out immense calculations, only to discover at the end that practically everything we calculated was irrelevant to the question we were asking.

This comment provided my own personal motivation for believing that complicated ontological systems are best approached, not through fundamental physics, but rather through information and inference. This difficult goal was first broached in Volume I where we discussed our basic inability to predict the *detailed* behavior of Wolfram's cellular automata into the far future, and by implication, our failure to predict the future *detailed* behavior of any ontological system, to include the ultimate evolution of our Universe.

Had Boltzmann bought into the *diktat* that his equation must include explicit reference to dynamical details, we would never have been the beneficiary of the insight that occurs for reproducible phenomena. Perhaps we would all now be working on computer programs to simulate the dynamical details of 10,000 atoms over 10^{-8} seconds to improve on our last decade's advancements to simulate 1,000 atoms over 10^{-9} seconds.

Frank Tipler [33] has his own idiosyncratic rationale for justifying the existence of the Omega Point. From my perspective, it would be an anodyne if some inferential prediction based on Boltzmann's tactic of ignoring inconsequential details in exchange for a few powerful averages could also lead to the prediction of something akin to the Omega Point.

We don't have to look to the far future to seek some solace within Boltzmann's lucky conjectures. Understanding the far past also has need of statistical mechanics. Weinberg [34] provides an exciting motivation for why a true lover of knowledge should study statistical mechanics. In writing about the extreme early history of the origin of the Universe, he emphasizes the role of statistical mechanics of thermal equilibrium in determining what now constitutes our world and everything that it contains.

The point is that during the whole of the first second the universe was presumably in a state of thermal equilibrium, in which the numbers and distributions of all particles, even neutrinos, were determined by the laws of statistical mechanics not by the details of their prior history. . . . As far as we know, nothing that we can observe depends upon the history of the universe prior to that time. . . . It is as if a dinner were prepared with great care—the freshest ingredients, the most carefully chosen spices, the finest wines—and then thrown all together in a great pot to boil for a few hours. It would be difficult for even the most discriminating diner to know what he was being served.

I have already mentioned several times that Jaynes drew his original inspiration from statistical mechanics in order to relate the MEP formula to probability. But he was led to speculate that the technique applied as well to all areas outside physics where inferences depending upon missing information were required. He was then led to generalize probability theory as the all-encompassing logic of scientific inference.

As Jaynes [16, pg. 50] says,

All of the above relations . . . are elementary consequences of maximizing the information entropy subject to constraints on average values of certain quantities. Although they bear a strong resemblance to the rules of calculation provided by statistical mechanics, they make no reference to physics, and, therefore, must apply equally well to any problem, in or out of physics, where the situation can be described by (1) enumerating a discrete set of possibilities and by (2) specifying average values of various quantities.

One of the constant misperceptions about the MEP is that it may not give the “right answer” on the first application. This may be true; but this doesn’t indicate a defect in the principle, only a gap in our state of knowledge brought about by insufficient information. Therefore, rather than being a defect, the MEP’s wrong answer points to the lacunae in our information. It is a signal that we must search for better information.

As Jaynes [16, pp. 50–51] so aptly points out,

. . . the maximum entropy probability assignment [my Q_i] cannot be regarded as describing any objectively existing state of affairs; it is only a means of describing a state of knowledge in a way that is “maximally non-committal” by a certain criterion. The above equations then represent simply the best predictions that we are able to make on the given information. We are not entitled to assert that the predictions must be “right,” only that to make any better ones, we should need more information than was given.

Jaynes often mentioned how Gibbs’s use of the maximum entropy principle led to “wrong answers” in physics. But at the turn of the 19th century, these wrong answers were the stimuli pointing to the development of the quantum theory a quarter of a century later.

26.6 Solved Exercises for Chapter Twenty Six

The first ten exercises are inspired by the explanation of Boltzmann's distribution function given in Chapter 5 of Charles Ruhla's book [29], *The Physics of Chance*. At the appropriate juncture, we point out where the conventional wisdom about probability goes awry.

Exercise 26.6.1: Examine the situation of 20 molecules in a box divided by a partition.

Solution to Exercise 26.6.1

For an introductory numerical example to Boltzmann's equation, Ruhla considers $N = 20$ molecules that can be found in either the left or right hand side of a box. Thus, we are led to a state space, or a very simple "phase space" of dimension $n = 2$. The two statements in this state space are:

$$(X = x_1) \equiv \text{"A molecule is in the left hand side of the box."}$$

and

$$(X = x_2) \equiv \text{"A molecule is in the right hand side of the box."}$$

Consider the problem from the standpoint of the sample space. There are a total of $n^N = 2^{20} = 1,048,576$ elementary points in this sample space. One of these elementary points in the sample space might be that two molecules labeled as **a** and **b** are in the left side of the box, while the remaining eighteen molecules labeled as **c** through **t** are in the right side of the box.

Exercise 26.6.2: Employ Feller's abstract definition of a sample space as discussed in Volume I to set up Ruhla's discussion of the Boltzmann distribution.

Solution to Exercise 26.6.2

There are N molecules of an "ideal gas." There are a total of n small subdivisions in phase space. Each one of the N molecules is a "another trial" just like another toss of the coin, another roll of the dice, or another kangaroo.

The N molecules are Feller's r balls, and the phase space constitute Feller's n cells. Extending the discussion from the $n = 2$ "phase space" to an n -dimensional space, there is a cell 1 which might contain two balls labeled as **d** and **e**. This corresponds to two molecules in the first cell of the phase space. There is a cell 2 containing one ball labeled as **a**. This corresponds to one molecule in the second cell of the phase space. Much further down the line, there is a cell i with three balls labeled as **b**, **c**, **z**. This corresponds to three molecules in the i^{th} cell of the phase space. Finally, the last cell is the n^{th} cell which is empty. No molecules occupy

this phase space cell. There may be many different cells in the phase space that correspond to the same *energy* depending on how the energy, or, more generally, a Hamiltonian is defined.

There are a total of n^N elementary points in this sample space. Suppose N is of the order of, say, 10^{24} . The total number of cells in phase space is an equally enormous number, but the specification is usually made that $N > n$ so that we may think of several molecules as potentially co-existing in the same cell of the phase space. Although we can imagine and construct any single elementary point in this sample space, the totality of elementary points in this sample space is beyond our comprehension.

Exercise 26.6.3: What is an elementary point of the sample space called in statistical mechanics?

Solution to Exercise 26.6.3

Returning to the numerical example in the first exercise, Ruhla looks at an elementary point in the sample space where, say, the five specific molecules **b**, **c**, **i**, **l**, **n** are in the left hand side of the box, and the remaining fifteen molecules **a**, **d**, **e**, **f**, **g**, **h**, **j**, **k**, **m**, **o**, **p**, **q**, **r**, **s**, **t** are in the right side of the box. This is another elementary point from the total of 1,048,576 available. This specific configuration of the labeled molecules in their respective sides of the box is called a *micro-state*.

Statistical mechanics uses language which says something like, “An assembly of N systems, each of which can exist in n quantum states, can be characterized by a micro-state and a macro-state.” The number of micro-states is the same as the number of elementary points in a sample space.

Exercise 26.6.4: What are aggregates of elementary points of the sample space called in statistical mechanics?

Solution to Exercise 26.6.4

This elementary point in the previous exercise has five molecules in cell 1 and fifteen molecules in cell 2. Specifying an occupation number of $N_1 = 5$ and $N_2 = 15$ where $N_1 + N_2 = N$ identifies a *macro-state*.

There are many elementary points which have five molecules in cell 1 and fifteen molecules in cell 2 if the particular labeling of the molecules is voluntarily ignored. The multiplicity factor counts up the exact number of micro-states constituting a macro-state. In this example, there are 15,504 micro-states constituting this macro-state, (an aggregation of 15,504 elementary points of the sample space), each with five molecules on the left side and fifteen molecules in the right side of the box.

$$\begin{aligned} W(N) &= \frac{20!}{5! 15!} \\ &= 15,504 \end{aligned}$$

Exercise 26.6.5: How many macro-states are there?

Solution to Exercise 26.6.5

The formula for counting up the total number of macro-states was developed in Volume I. Applying it here, we find that there are just 21 macro-states.

$$\begin{aligned} \text{Number of macro-states} &= \frac{(N + n - 1)!}{N! (n - 1)!} \\ &= \frac{(20 + 2 - 1)!}{20! 1!} \\ &= 21 \end{aligned}$$

In Volume I, we emphasized that this formula counted up the total of all possible frequency counts for the n cells of the contingency table, or all possible occupation numbers, N_1, N_2 , where in this case $N_1 + N_2 = N = 20$.

Exercise 26.6.6: What is the probability for the macro-state with five molecules in the left side and fifteen molecules in the right side of the box?

Solution to Exercise 26.6.6

Ruhla presents the usual analysis relying on the binomial distribution.

$$\begin{aligned} P(N_1 = 5, N_2 = 15 | \mathcal{M}_k) &= W(N) Q_1^{N_1} Q_2^{N_2} \\ &= \frac{20!}{5! 15!} (1/2)^5 (1/2)^{15} \\ &= 0.014786 \end{aligned}$$

But one of our major goals in Volume I was to make quite clear that such a result depends upon adopting just *one* model. The IP must have the strongest possible conviction that out of all possible causes for the molecules to be in one or the other sides of the box, only one cause, and one cause alone must be operating. This one cause is represented by the one model where,

$$P(X = x_1 | \mathcal{M}_k) = P(X = x_2 | \mathcal{M}_k) \equiv Q_1 = Q_2 = 1/2$$

If, on the other hand, the IP is in a state of complete ignorance, then as we learned in Volume I, and made mention of many times here in Volume II, the probability for five molecules on the left side and fifteen on the right side of the box is not found by calculating with the binomial distribution as Ruhla suggests.

When all models must be averaged over, or, in other words, all possible causes must be taken into account (Laplace's *Principle of Insufficient Reason* once again), the correct state of knowledge is instead,

$$P(N_1 = 5, N_2 = 15) = \frac{N! (n-1)!}{(N+n-1)!} = \frac{1}{21}$$

This probability for the occupation numbers is slightly less than 0.05, and quite different from the probability of 0.014786 when the binomial model is adopted. The IP's state of knowledge when it is completely ignorant about the causes is the same for every possible occupation number of the molecules. The probability for no molecules in the left side and twenty molecules in the right side is the same as an even division of ten molecules in both halves of the box.

Exercise 26.6.7: Use the rationale of the MEP to assign a value of 1/2 to the probability that a molecule is in the left hand side of the box.

Solution to Exercise 26.6.7

The IP will insert information under some model \mathcal{M}_k in the form of a constraint function average. Suppose that a mapping from the statement, "The molecule is in the left hand side of the box." is $F(X = x_1) = 1$, while the mapping from the statement, "The molecule is in the right hand side of the box." is $F(X = x_2) = 3$.

The information under this model is specified by the average of the constraint functions as $\langle F \rangle = 2$, or,

$$F(X = x_1) Q_1 + F(X = x_2) Q_2 = 2$$

One possible legitimate numerical assignment that satisfies the universal constraint is $Q_1 = Q_2 = 1/2$. This assignment also satisfies the given constraint function average under the model \mathcal{M}_k ,

$$(1 \times 1/2) + (3 \times 1/2) = 2$$

Finally, this particular assignment of,

$$P(X = x_1 | \mathcal{M}_k) = P(X = x_2 | \mathcal{M}_k) = 1/2$$

has the maximum information entropy,

$$-(1/2 \ln 1/2 + 1/2 \ln 1/2) = \ln n = 0.6931$$

of *any legitimate assignment* that also might happen to satisfy these two constraint as well. The amount of *missing information* in this model has been maximized, thus there is no hidden information in this assignment of which the IP was unaware.

Exercise 26.6.8: What if different information had been specified?***Solution to Exercise 26.6.8***

Suppose that another model had inserted the information that $\langle F \rangle = 1.5$. Then an assignment of $Q_1 = 3/4$ and $Q_2 = 1/4$,

$$(1 \times 3/4) + (1 \times 1/4) = 1$$

$$(1 \times 3/4) + (3 \times 1/4) = 1.5$$

satisfies the two constraint function averages. Of course, it is not really necessary to check that this assignment has the maximum entropy of any assignment that satisfies the constraints. There are $m = 2$ constraint function averages and the dimension of the state space is also $n = 2$, so there is no remaining ambiguity in the assignment for maximum entropy to resolve.

Exercise 26.6.9: What if the box had been divided into three partitions?***Solution to Exercise 26.6.9***

Then, in this case, the dimension of the state space becomes $n = 3$. Change the mapping to,

$$F(X = x_1) = 2$$

$$F(X = x_2) = 4$$

$$F(X = x_3) = 6$$

Specify just $m = 1$ piece of information that $\langle F \rangle = 4$. The assignment of $Q_1 = 0$, $Q_2 = 1$, and $Q_3 = 0$ would seem to fit the bill. This is an assignment of legitimate numerical values to probabilities. It satisfies the universal constraint that all the assigned probabilities must sum to 1. It also satisfies the informational constraint that the constraint function average must equal 4.

However, there is just one thing wrong with this assignment. Since $m < n$, we require the resulting assignment to possess the maximum entropy of any assignment that might also happen to satisfy the information. There may be many legitimate assignments that satisfy the constraint function averages. So the IP requires some additional principle to resolve this ambiguity. That additional principle specifies the best assignment as the one which has maximum information entropy.

Our current tentative assignment is woefully lacking in this regard because its information entropy is 0. The assignment of $Q_1 = 1/3$, $Q_2 = 1/3$, and $Q_3 = 1/3$ is the MEP assignment because it also satisfies all of the information under the model while possessing the maximum entropy of $\ln n = 1.0986$. No other assignment could possibly have a greater information entropy than this one.

Exercise 26.6.10: What if the box had been divided into four partitions?***Solution to Exercise 26.6.10***

The dimension of the state space becomes $n = 4$. Even though our mental picture is now of a box divided into four partitions, with N_1 molecules in the first partition, N_2 molecules in the second partition, N_3 molecules in the third partition, and N_4 molecules in the last partition, remember that we are concerned only with a very abstract picture of a state space where the statements in the state space describe some measurable physical property of a molecule. Here, the measurable physical property is whether the x -coordinate of the molecule falls within the boundaries set by the location of the partitions of the box.

Change the mapping of the statements in this state space to the numbers,

$$F(X = x_1) = 1$$

$$F(X = x_2) = 2$$

$$F(X = x_3) = 3$$

$$F(X = x_4) = 4$$

In addition, change the number of molecules from $N = 20$ to $N = 16$. We have the same numbers involved in this sample space as with the kangaroo sample space. We may take advantage of all the detailed calculations performed in Volume I for this exercise.

It doesn't make any difference, at our level of abstraction, whether the totality of the n statements are referring to position coordinates of molecules, the energy of a molecule, or the beer drinking traits of a kangaroo. It is only necessary within our schema to assert a statement which attaches a property to a trial. The schema applies equally well whether the trials refer to, say, N kangaroos observed for beer and hand preference, or measuring some property of N molecules.

However, in preparation for statistical mechanics, we will say that the mapping represents the energy E_i of a molecule. Instead of forming a constraint function average where, by definition, such an average must be defined in terms of the Q_i , form now the weighted sum of the constraint function by the frequency counts N_i . Call E the "total energy."

$$E = \sum_{i=1}^4 N_i \times F(X = x_i)$$

and say that one model inserts the information that the total energy is $E = 40$.

What is one possible integer frequency count of the molecules possessing these four energies that satisfy the total energy? If there are two molecules with energy E_1 , four molecules with energy E_2 , ten molecules with energy E_3 , and finally no molecules with the highest energy E_4 , we have,

$$E = (2 \times 1) + (4 \times 2) + (10 \times 3) + (0 \times 4) = 40$$

We can see that this frequency count for the molecules does satisfy the constraint that $\sum_{i=1}^4 N_i = N = 16$.

Here is the crucial point vital to our later discussion. The multiplicity factor for this frequency count is,

$$W(N) = \frac{16!}{2! 4! 10! 0!} = 120,120$$

and represents the number of different ways to find two distinct molecules with energy E_1 , four distinct molecules with energy E_2 , and so on.

However, there happens to exist another frequency count that also satisfies the two constraints of $\sum_{i=1}^4 N_i = 16$ and $E = 40$. This frequency count is where there are four molecules at each of the four energy levels, that is, $N_1 = 4$, $N_2 = 4$, $N_3 = 4$, and $N_4 = 4$. The multiplicity factor for this frequency count is,

$$W(N) = \frac{16!}{4! 4! 4! 4!} = 63,063,000$$

There are vastly more ways for the 16 molecules to distribute themselves over the energy levels in this latter configuration when compared to the former. This is called “the most probable distribution of the molecules.” If many integer counts can satisfy the constraints, then the one with the maximum multiplicity factor is selected. This is the essence of almost every statistical mechanics derivation for the Boltzmann equation, starting with Boltzmann himself.

But it is wrong conceptually. You never need to go to the trouble of finding any specific set of frequency counts satisfying these objectives by relying on the MEP. You are quite happy to deal with *any* data that might turn up because it gets handled through $P(\mathcal{M}_k | \mathcal{D})$.

Exercise 26.6.11: Can *Mathematica* calculate the multiplicity function for you?

Solution to Exercise 26.6.11

Submit the expression **Multinomial[args]** for evaluation where *args* are the n integer counts of interest. The n *args* must sum to N .

For example, to calculate the multiplicity factor for the first configuration of molecules, **Multinomial[2,4,10,0]** returns the answer 120,120. For the maximum multiplicity factor, **Multinomial[4,4,4,4]** returns the answer 63,063,000.

Exercise 26.6.12: Might there be some profound conceptual differences between Q_i and $P(N_1, N_2, \dots, N_n)$?

Solution to Exercise 26.6.12

Indeed, there are. Q_i is $P(X = x_i | \mathcal{M}_k)$. It represents a numerical assignment to a probability for a statement in the state space when conditioned on the truth of the information asserted under some model. $P(N_1, N_2, \dots, N_n)$ represents a state of knowledge about N_1, N_2, \dots, N_n frequency counts over N trials.

It does depend upon the Q_i through the formal manipulation relationship,

$$\begin{aligned} P(N_1, N_2, \dots, N_n) &= \sum_{k=1}^{\mathcal{M}} P(N_1, N_2, \dots, N_n | \mathcal{M}_k) P(\mathcal{M}_k) \\ &= W(N) \sum_{k=1}^{\mathcal{M}} Q_1^{N_1} \times Q_2^{N_2} \times \dots \times Q_n^{N_n} P(\mathcal{M}_k) \end{aligned}$$

but obviously the two are not the same. When the summation transitions to an integration over all model space, we have from Volume I,

$$P(N_1, N_2, \dots, N_n) = W(N) \times C_D \times \int \cdot \int_{\sum q_i=1} q_1^{N_1} \cdots q_n^{N_n} \times q_1^{\alpha_1-1} \cdots q_n^{\alpha_n-1} dq_i$$

Here is another profound conceptual difference. Even though the MEP algorithm is the highly preferred methodology for calculating the Q_i because we know what information the model is inserting into the assignments, legitimate numerical values to probabilities may be assigned by any means whatsoever. Since the probability for the frequency counts $P(N_1, N_2, \dots, N_n)$ results from an integration *over all conceivable numerical assignments*, the MEP formula under any individual model,

$$P(X = x_1 | \mathcal{M}_k) \equiv Q_i = \frac{\exp [\sum_{j=1}^m \lambda_j F_j(X = x_i)]}{Z(\lambda_1, \dots, \lambda_j, \dots, \lambda_m)}$$

need never appear! Because all conceivable assignments are eventually involved in the integration, calculating any specific assignment via the MEP is superfluous!

Exercise 26.6.13: What is the amazingly simple formula for finding the probability for a molecule to have a given energy level conditioned on a large amount of data?

Solution to Exercise 26.6.13

As a numerical example, we will continue to refer to the eight energy level Boltzmann equation in section 26.2. We are concerned here with the probability for the energy

of the *next* molecule, where $M_i = 1$ and $\sum_{i=1}^n M_i = 1$, when conditioned on a large amount of data, N_1, \dots, N_8 , in other words, known energy levels for N molecules. We hazard the opinion that this is really the probability that people are searching for with so much effort.

Applying the formal manipulation rules to this example, we have,

$$P(M_1 = 0, \dots, M_i = 1, \dots, M_8 = 0 | N_1, \dots, N_8) = \frac{N_i + 1}{N + 8}$$

The one thing that I want to emphasize above all else in this probability formula is that it was derived by assuming a uniform distribution over all models. Therefore, all conceivable MEP assignments were taken into account by the amplification of the integration shown above in Exercise 26.6.12 when the known data were added to the mix. The form of Boltzmann's equation as an exemplar of the MEP formula is completely irrelevant! It was not required in any way, shape, or form in the derivation of this (correct) prediction formula.

The deep implication is that the importance of the MEP formula only becomes apparent *after the fact!* Only after we have established some probability for future frequency counts by the above prediction formula, do we become curious as to what kind of information in some model would lead to such frequencies. The MEP formula then is very helpful in teasing out what information (temperature or average energy) is the most likely candidate.

Exercise 26.6.14: Reproduce the full formula for the probability of any number of future frequency counts conditioned on past data. Then show how the simple formula in the last exercise comes about for some *next* occurrence.

Solution to Exercise 26.6.14

Earlier in Chapter Nineteen, Exercise 19.9.13, the full formula for the probability of future frequency counts when conditioned on past frequency counts was presented as,

$$\begin{aligned} P(M_1, M_2, \dots, M_n | N_1, N_2, \dots, N_n) &= C \times \frac{\prod_{i=1}^n (M_i + N_i)!}{\prod_{i=1}^n M_i!} \\ &= \frac{M! (N + n - 1)!}{N_1! \cdots N_n! (M + N + n - 1)!} \times \frac{\prod_{i=1}^n (M_i + N_i)!}{\prod_{i=1}^n M_i!} \end{aligned}$$

Substituting $M = 1$, $M_i = 1$, all other $M_j = 0$, and $n = 8$,

$$\begin{aligned}
 P(M_1, M_2, \dots, M_8 | N_1, N_2, \dots, N_8) &= \frac{1! (N+8-1)!}{N_1! \times \dots \times N_8! (1+N+8-1)!} \times \\
 &\quad \frac{N_1! \times \dots \times (N_i+1)! \times \dots \times N_8!}{1} \\
 &= \frac{(N+7)!}{(N+8)!} \times (N_i+1) \\
 &= \frac{N_i+1}{N+8}
 \end{aligned}$$

Exercise 26.6.15: Provide a numerical example for this last exercise.

Solution to Exercise 26.6.15

Take the number of molecules whose eight energy levels have been measured as the “large” number $N = 1,000,000$. Here is a list of the data over all eight energy levels.

$$N_1 = 400,810$$

$$N_2 = 243,104$$

$$N_3 = 147,450$$

$$N_4 = 89,433$$

$$N_5 = 54,244$$

$$N_6 = 32,901$$

$$N_7 = 19,955$$

$$N_8 = 12,103$$

The data said that 32,901 of these 1,000,000 molecules were in energy state 6. Thus, the probability that the *next* molecule to be measured will be in energy state 6 given these data is,

$$\begin{aligned}
 P(M_1 = 0, M_2 = 0, \dots, M_6 = 1, \dots, M_8 = 0 \mid \mathcal{D}) &= \frac{N_i + 1}{N + 8} \\
 &= \frac{32901 + 1}{1000000 + 8} \\
 &= 0.032902
 \end{aligned}$$

Comparing this probability for the next molecule to be energy state 6 based on an extensive amount of data with Q_6 , the probability for a molecule to be in energy state 6 based solely on the information in one model, we have 0.032902 versus 0.032901.

These data were constructed to match the expectation of the frequency counts, $N_i = N \times Q_i$, stemming from the Boltzmann equation for a temperature of $kT = 2$. The derivation finding the probability for the next molecule to be in energy state 6 DID NOT rely on the one model making this specific assignment, but took all models making numerical assignments into account on an equal basis. However, *after the fact*, we can use the MEP formula to strongly implicate a model with the information $kT = 2$.

Exercise 26.6.16: Explore the connection between the multiplicity factor and information entropy.

Solution to Exercise 26.6.16

What we are going to do next is to recapitulate a well-known manipulation of the multiplicity factor. The goal is to see how such a manipulation could be connected to information entropy. The derivation will be done in much more detail than you would see elsewhere.

The relationship between the end result and entropy was utilized extensively by Boltzmann, Planck, Schrödinger, Jaynes, Ruhla, and countless others. We will see in the next Chapter how important the multiplicity factor was to Schrödinger's development of statistical thermodynamics. Nonetheless, it remains a source of unending confusion to this day.

The multiplicity factor $W(N)$ can be approximated by an important formula as N grows increasingly large. In fact, we investigate what happens to,

$$\frac{\ln [W(N)]}{N} \text{ as } N \rightarrow \infty$$

To get us in the right frame of mind, we note two elementary facts concerning operations with the natural log function.

1. Multiplication is done by addition.
2. Division is done by subtraction.

Applying these two rules allows us to immediately transform the multiplicity factor into,

$$\begin{aligned}\ln [W(N)] &= \ln \left[\frac{N!}{N_1! N_2! \cdots N_n!} \right] \\ &= \ln N! - (\ln N_1! + \ln N_2! + \cdots + \ln N_n!) \\ &= \ln N! - \sum_{i=1}^n \ln N_i!\end{aligned}$$

The next step is to get rid of the factorial expressions because we are going to deal with large N . There is an approximation³ to $\ln N!$ for large N . In its simplest form, Stirling's approximation for $\ln N!$ is,

$$\ln N! \approx N \ln N - N$$

Substituting in Stirling's approximation for $\ln N!$ and $\ln N_i!$, we have,

$$\ln [W(N)] \approx N \ln N - N - \sum_{i=1}^n (N_i \ln N_i - N_i)$$

From this point on, everything on the right hand side of the equations is an approximation based on this version of Stirling's formula. Distribute the summation sign across the terms within the parentheses to arrive at,

$$\ln [W(N)] \approx N \ln N - N - \left(\sum_{i=1}^n N_i \ln N_i - \sum_{i=1}^n N_i \right)$$

Since,

$$\sum_{i=1}^n N_i = N$$

we can substitute for the very last term resulting in,

$$\ln [W(N)] \approx N \ln N - N - \sum_{i=1}^n N_i \ln N_i + N$$

At this stage, the two N s conveniently cancel, leaving us with,

$$\ln [W(N)] \approx N \ln N - \sum_{i=1}^n N_i \ln N_i$$

³called so after the Scottish mathematician James Stirling who studied infinite series and discovered this approximation in 1730. Some historians dispute this and claim that Stirling's formula was first discovered by Abraham DeMoivre (1667-1754), a French Huguenot who spent his professional life in England and was an early and significant contributor to probability theory.

We now turn our attention to $\ln N_i$ in the summation term of the above equation. If we divide through an expression by N we should compensate for that division by multiplying by N . Since this is done within a logarithmic expression we have,

$$\begin{aligned} N_i &= \frac{N_i}{N} \times N \\ \ln N_i &= \ln N_i - \ln N + \ln N \\ &= \ln\left(\frac{N_i}{N}\right) + \ln N \end{aligned}$$

After substituting for $\ln N_i$ we have,

$$\ln [W(N)] \approx N \ln N - \sum_{i=1}^n N_i \left[\ln\left(\frac{N_i}{N}\right) + \ln N \right]$$

Multiplying the term in brackets by N_i results in,

$$\ln [W(N)] \approx N \ln N - \sum_{i=1}^n \left[N_i \ln\left(\frac{N_i}{N}\right) + N_i \ln N \right]$$

Now distribute the summation sign across the two terms inside the brackets. Notice how the last term becomes negative.

$$\ln [W(N)] \approx N \ln N - \sum_{i=1}^n N_i \ln\left(\frac{N_i}{N}\right) - \sum_{i=1}^n N_i \ln N$$

Examining this last term reveals that,

$$\sum_{i=1}^n N_i \ln N = N \ln N$$

The third term again conveniently cancels out the first term and we are left with,

$$\ln [W(N)] \approx - \sum_{i=1}^n N_i \ln\left(\frac{N_i}{N}\right)$$

All that remains to do is divide both sides by N , and we have nearly reached our goal,

$$\frac{\ln [W(N)]}{N} \approx - \sum_{i=1}^n \frac{N_i}{N} \ln\left(\frac{N_i}{N}\right)$$

We now think of the fraction N_i/N as a normed frequency and label it as f_i . Substituting f_i for N_i/N into the result above yields,

$$\frac{\ln [W(N)]}{N} \text{ as } N \rightarrow \infty \approx - \sum_{i=1}^n f_i \ln f_i$$

where the expression on the right hand side appears quite similar to information entropy. The remark about $N \rightarrow \infty$ is included because the accuracy of the Stirling approximation used in the derivation improves as N gets larger and larger.

In order to distinguish this entropy-like expression based on normed frequency counts from the information entropy based on the numerical assignment from some model, we will employ the notation of,

$$H_1 \equiv H(f_i) = - \sum_{i=1}^n f_i \ln f_i$$

as opposed to,

$$H_0 \equiv H_{max}(Q_i | \mathcal{M}_k) = - \sum_{i=1}^n Q_i \ln Q_i$$

Exercise 26.6.17: Conduct a numerical experiment in order to examine what happens to the probability of the “most probable distribution.”

Solution to Exercise 26.6.17

Consider the situation where we look at an ever increasing number of molecules in the box starting at $N = 200$, and ending with $N = 1,000,000$. Each column after the first in Table 26.4 at the top of the next page is the frequency count based on the Q_i from the model where $kT = 2$, that is, it is the “most probable distribution.”

Thus, for a million molecules, it is expected under this model that 400,810 molecules will be in the lower left front portion of the box where each molecule has the lowest energy, 243,104 molecules will be in the lower left back portion of the box where each molecule has the next lowest energy, and so on, up to 12,103 molecules in the upper back right portion of the box where each molecule has the highest energy.

However, the probability for each succeeding set of occupation numbers, even though they represent the “most probable distribution,” is decreasing rapidly as it must. This probability, even though it is a maximum, must decrease given the ever increasing N because of the huge explosion in $\frac{(N+n-1)!}{N! (n-1)!}$, the total number of possible frequency counts, or macro-states. The sample average energy \bar{E} is approaching the model’s parameter expected energy $\langle E \rangle = 1.392235$ as N increases.

Rows at the bottom of the table list the information entropy measure based on the frequency counts using the approximation for $\frac{\ln W(N)}{N} \approx H(f_1, f_2, \dots, f_n)$ as $N \rightarrow \infty$ as derived in the previous exercise, the exact value for this function of the multiplicity factor, and the unchanging value of the maximum information entropy.

You might reasonably expect that, based on the analysis in section 26.3 as $N \rightarrow \infty$ and the approximation afforded by $H(f_i)$ to the multiplicity factor function gets better and better, $P(N_i^j)$ might start to approach 1.

Table 26.4: A numerical experiment looking at the probabilities for the specific frequency count found as the expectation based on the MEP's Q_i .

i	N_i^1	N_i^2	N_i^3	N_i^4
1	80	8016	160324	400810
2	49	4862	97242	243104
3	29	2949	58980	147450
4	18	1789	35773	89433
5	11	1085	21698	54244
6	7	658	13160	32901
7	4	399	7982	19955
8	2	242	4841	12103
<i>Sum</i>	200	20000	400000	1000000
\bar{E}	1.390000	1.392250	1.392230	1.392235
$P(N_i^j)$	5.51×10^{-7}	6.04×10^{-14}	1.69×10^{-18}	6.83×10^{-20}
$H(f_i)$	1.60897	1.61039	1.61038	1.61038
$\frac{\ln W(N)}{N}$	1.53721	1.60887	1.61028	1.61034
$H_{\max}(Q_i)$	1.61038	1.61038	1.61038	1.61038

But this is clearly wrong as the above numerical experiment shows. The reality of the huge increase in the total number of possible frequency counts is beginning to sink in. $P(N_i^j)$ must approach 0 as $N \rightarrow \infty$. The problem arises in the error of the approximation $H(f_i)$ to $\frac{\ln W(N)}{N}$ brought about by the ever increasing N . The error in the approximation when multiplied by the ever increasing huge number N eventually wins out.

For example, take the case where $N = 1,000,000$. The exact value of the information entropy to eight decimal places is $H_{\max}(Q_i) = 1.61038428$. The comparable measure based on the frequency counts,

$$N_1 = 400810, N_2 = 243104, \dots, N_8 = 12103$$

is the mis-named “frequency entropy,”

$$\begin{aligned} H(f_i) &= - \sum_{i=1}^8 \frac{N_i}{N} \ln \frac{N_i}{N} \\ &= - \left[\frac{400810}{1000000} \times \ln \left(\frac{400810}{1000000} \right) + \dots + \frac{12103}{1000000} \times \ln \left(\frac{12103}{1000000} \right) \right] \\ &= 1.61038418 \end{aligned}$$

The value of the multiplicity factor function to eight decimal places is however,

$$\frac{\ln W(N)}{N} = 1.61034005$$

So, while the difference between $H(f_i) - H_{max}(Q_i)$ is only,

$$1.61038418 - 1.61038428 = -0.00000010$$

the true difference $\frac{\ln W(N)}{N} - H(f_i)$ is really much greater at,

$$1.61034005 - 1.61038418 = -0.00004413$$

When multiplied by $N = 1,000,000$, this error in the approximation reflects the difference between the correct probability of, (when $\lambda(\bar{F} - \langle F \rangle) \approx 0$),

$$P(N_1, N_2, \dots, N_8 | \mathcal{M}_k) \approx \exp \{N [\ln W(N)/N - H_{max}(Q_i)]\} \approx 6.83 \times 10^{-20}$$

and an incorrect probability based on the approximation used to derive $H(f_i)$ of,

$$P(N_1, N_2, \dots, N_8 | \mathcal{M}_k) \approx \exp \{N [(H(f_i) - H_{max}(Q_i))]\} \approx 0.90$$

Exercise 26.6.18: Use a familiar numerical approximation for the partial derivative to confirm the relationship between the partition function, the Lagrange multiplier, and the constraint function average.

Solution to Exercise 26.6.18

The following results are part of the MEP formalism, and can be used to check for consistency with the example presented in section 26.2.2. The partial derivative of the log of the partition function with respect to the j^{th} Lagrange multiplier is equal to the j^{th} constraint function average.

$$\frac{\partial \ln Z(\lambda_1, \dots, \lambda_j, \dots, \lambda_m)}{\partial \lambda_j} = \langle F_j \rangle$$

Since we have just the one constraint function, $F(X = x_i) \equiv E_i$, with its average of $\langle E \rangle = 1.3922$ specified as the information,

$$\frac{d \ln Z(\lambda)}{d \lambda} = \langle E \rangle = 1.3922$$

First, we show the central difference numerical approximation to the first derivative as,

$$\frac{d \ln Z}{d \lambda} \approx \frac{\ln Z(\lambda + \delta) - \ln Z(\lambda - \delta)}{2\delta}$$

The value for $\lambda \equiv -\frac{1}{kT}$ was determined when a temperature of $kT = 2$ was set, leading to $\lambda = -0.5$. If a delta of $\delta = 0.01$ is added to and subtracted from this value, then,

$$\lambda + 0.01 = -0.49$$

$$\lambda - 0.01 = -0.51$$

This leads to new $kT = 2.04082$ and $kT = 1.96078$.

The logs of the new partition functions calculated at these Lagrange multipliers are,

$$\ln Z(-0.49) = 0.928326$$

$$\ln Z(-0.51) = 0.900477$$

yielding within this accuracy,

$$\frac{d \ln Z}{d \lambda} \approx \frac{0.928326 - 0.900477}{0.02} \approx 1.3924$$

Exercise 26.6.19: Construct a numerical example with dice illustrating the fundamental difference between the correct MEP approach and the way that statistical mechanics does it.

Solution to Exercise 26.6.19

The basic confusion swirls around whether it is the state space with dimension n that is the focus of interest, or whether instead the focus should be on the sample space over N systems.

If, as in statistical mechanics, it is the sample space over N systems that is considered relevant, then attention must turn to those combinatorial formulas involving the total number of elementary points in the sample space n^N , the total number of possible frequency counts, $\frac{(N+n-1)!}{N!(n-1)!}$, and the number of ways for each frequency count to happen, the multiplicity factor $W(N)$.

Suppose that a die can exist in one of $n = 6$ “quantum states.” The statement ($X = x_i$) is, “The die exists in the i^{th} quantum state.” The function mapping the quantum state to an energy is $F(X = x_i) = i$. We are interested in an assembly of $N = 10$ dice. The total number of elementary points in this sample space is $n^N = 6^{10} = 60,466,176$. The total number u of possible frequency counts is,

$$\text{Number of frequency counts} = u = \frac{(N+n-1)!}{N!(n-1)!} = \frac{(10+6-1)!}{10!5!} = 3,003$$

leading to,

$$n^N = \sum_{j=1}^{u=3003} W_j(N) = 60,466,176$$

The total energy of the assembly of systems is measured as $E = 28$ where the possible range of the total energy is between $E = 10$ and $E = 60$. Figure out the most probable distribution of counts, that is, the j^{th} frequency count from the 3003 possible satisfying the constraints of total energy, and, in addition, possessing the maximum multiplicity factor as,

$$N_1 = 2, N_2 = 4, N_3 = 1, N_4 = 1, N_5 = 1, N_6 = 1$$

$$E = \sum_{i=1}^6 N_i E_i = (2 \times 1) + (4 \times 2) + \cdots + (1 \times 6) = 28$$

$$W_j(N) = \frac{10!}{2! 4! 1! 1! 1! 1!} = 75,600$$

After finding these frequency counts, we are essentially finished from the statistical mechanics point of view. We did not find any Q_i , but the frequentist would be more or less compelled to give an “estimate” of $Q_1 = 0.20, Q_2 = 0.40$, and so on.

The MEP point of view wants to first find the Q_i from Boltzmann’s “canonical distribution,”

$$Q_i \equiv P(X = x_i | \mathcal{M}_k) = \frac{e^{\lambda F(X=x_i)}}{Z(\lambda)}$$

If a model inserts the information $\langle F \rangle = 2.8$ in order to match the constraint of the total energy, then the Q_i are readily found by the MEP algorithm. Then, *the probability of any possible frequency count*, compared to just one “most probable distribution,” can be calculated by the formula,

$$P(N_1, N_2, \dots, N_6 | \mathcal{M}_k) = \exp \left[N \left(\frac{\ln W(N)}{N} + \lambda \bar{F} - \ln Z \right) \right]$$

The probability for,

$$P(N_1 = 2, N_2 = 4, N_3 = 1, N_4 = 1, N_5 = 1, N_6 = 1 | \mathcal{M}_k) = 0.002941$$

will then be the largest probability of any of the 3003 possible frequency counts.

Exercise 26.6.20: In preparation for what is to come later in Schrödinger’s analysis, prepare some numerical examples comparing the log transform of the *maximum* value of the multiplicity factor when compared to the log transform of the *sum* of all the values of the multiplicity factor.

Solution to Exercise 26.6.20

We will carry out the induction through numerical examples of sums over $W_j(N)$, the multiplicity factor for the j^{th} frequency count. Start with $N = 4$ systems, that is, by examining the sum over the multiplicity factors associated with each of the five frequency counts.

Let's write this sum as $\sum_{j=1}^u W_j(N)$ where we would like to have a formula for u , the upper limit for the index j on each multiplicity factor. We already know from previous discussions here and in Volume I, that this sum is equal to the total number of elementary points in the sample space,

$$\sum_{j=1}^u W_j(N) = n^N$$

where n , of course, refers to the dimension of the state space.

The formula for the upper limit in the sum over all the multiplicity factors is our familiar formula for the total number of distinct frequency counts,

$$u = \frac{(N+n-1)!}{N! (n-1)!}$$

Table 26.5 below is a detailed listing of all u multiplicity factors for $N = 4$ through $N = 8$ for $n = 2$. The maximum multiplicity factor is indicated by an asterisk. We are going to compare the log of the sum of all the multiplicity factors divided by N versus the log of the one maximum multiplicity factor divided by N .

Table 26.5: *The beginning of an induction comparing the log maximum multiplicity factor versus the log of the sum of all the multiplicity factors.*

j	$N = 4$	$N = 5$	$N = 6$	$N = 7$	$N = 8$
1	1	1	1	1	1
2	4	5	6	7	8
3	6*	10	15	21	28
4	4	10*	20*	35	56
5	1	5	15	35*	70*
6	*	1	6	21	56
7	*	*	1	7	28
8	*	*	*	1	8
9	*	*	*	*	1
$\sum_{j=1}^u W_j(N) = n^N$					
	16	32	64	128	256

Now, looking at $N = 8$, it is obvious that in the first case,

$$\frac{\ln [\sum W(N)]}{N} = \frac{N \ln n}{N} = \ln n = \ln 2 = 0.693147$$

and in the second case,

$$\frac{\ln [W(N)_{max}]}{N} = \frac{\ln 70}{8} = 0.531062$$

If we examine the log of the single multiplicity factor that is the maximum divided by N as $N \rightarrow \infty$, will it approach 0.693147? The numerical results confirming this conjecture are shown in Table 26.6 below.

Table 26.6: A numerical induction comparing the sum of all multiplicity factors versus the maximum multiplicity factor.

N	$\frac{\ln \sum W(N)}{N}$	$\frac{\ln W(N)_{max}}{N}$	Difference
8	0.693147	0.531062	0.162085
10	0.693147	0.552943	0.140204
100	0.693147	0.667838	0.025309
1000	0.693147	0.689467	0.003680
10000	0.693147	0.692664	0.000483
10000000	0.693147	0.693146	0.000001

Exercise 26.6.21: How would *Mathematica* help you produce the probabilities for the frequency counts appearing in Tables 26.2 and 26.3 in section 26.2.2?

Solution to Exercise 26.6.21

Create a function with two arguments to compute the probability for any occupancy numbers of 200 ideal gas molecules. The function will be called,

```
probabilityOfFrequencies[arg1, arg2]
```

and its two arguments will be two lists with the first list containing the MEP assigned probabilities for the statements in the state space, and the second list containing the desired occupancy numbers.

```
probabilityOfFrequencies[qList, Ni_List] :=
Module[{total, energy},
total = Total[Ni];
energy = N[Total[Table[Ni[[i]] (i-1), {i, 1, 8}]]/total, 4];
{total, energy, Apply[Multinomial, Ni] Apply[Times, Power[q, Ni]]}]
```

The interesting part of the code occurs in the last output line where,

$$P(N_1, N_2, \dots, N_n | \mathcal{M}_k) = W(N) Q_1^{N_1} Q_2^{N_2} \cdots Q_n^{N_n}$$

is computed. The multiplicity factor $W(N)$ is `Apply[Multinomial, Ni]` where the built-in `Apply` function replaces the `Head` of the list of frequency counts in `Ni`, which obviously must be `List`, with a new `Head` which is `Multinomial`.

So now we have **Multinomial**[N_1, N_2, \dots, N_n] which is in the correct format to compute $W(N)$. The same thing happens where **Apply** replaces the **List** of the **Power[q,Ni]** with **Times**. So now we have **Times[Power[q,Ni]]** which is in the correct format to perform all of the requisite multiplications involving the terms $Q_i^{N_i}$.

The function would be called to evaluate, for example, the frequency counts N_i^1 in Table 26.2,

```
probabilityOfFrequencies[  
  { .4008, .2431, .1475, .0894, .0542, .0329, .02, .0121 },  
  { 80, 49, 29, 18, 11, 7, 4, 2 }  
 ]
```

It returns a list showing that the sum of the frequency counts is indeed $N = 200$, the average energy for this particular frequency count is 1.39, and the probability for this particular frequency count is $P(N_1, N_2, \dots, N_8 | \mathcal{M}_k) \approx 5.51 \times 10^{-7}$.

Chapter 27

Schrödinger's Statistical Thermodynamics

27.1 Introduction

What is the payoff for the effort invested in learning about the MEP? It might surprise you to find out that it is now possible to read and understand (almost) everything Erwin Schrödinger, one of the more well-known physicists of all time, and one of the founding fathers of quantum mechanics, wrote in his introductory Chapters to his famous set of lecture notes entitled *Statistical Thermodynamics*.

The lectures were given in 1944 in Ireland at the Dublin Institute for Advanced Studies after Schrödinger had fled the Nazis. It would not be unfair to say that Jaynes, as well as succeeding generations of physicists, were heavily influenced by reading Schrödinger. One may surmise that the MEP algorithm owes something to Schrödinger's thermodynamic explanations. It is Schrödinger's rationale concerning the interplay among frequencies and probabilities which we are now about to summarize and comment upon in this Chapter.

We alluded somewhat cryptically in the last Chapter to Boltzmann's, as well as Schrödinger's, and countless others, seeming predilection for an ontological derivation of statistical thermodynamics formulas. We promised that we would revisit the fundamental conceptual divide that separates the epistemological viewpoint from the ontological perspective.

We are addressing that confusing interface that exists between the informational and the observed. The epistemological, informational domain of the MEP algorithm applies to the n statements in the state space, while the ontological, frequentist domain applies to the N observations in the sample space. Probabilities change when information changes, frequency counts do not.

We offer just a fleeting glance at some of the connections between Schrödinger's treatment of statistical thermodynamics and the MEP algorithm. Basically, it only amounts to a kind of listing where a match-up is immediately evident. My sole objective is to point out that Schrödinger's equations, seemingly quite arcane and mysterious upon first exposure, turn out to be easily absorbed after making some notational and conceptual mappings from the MEP algorithm.

To begin, one cannot resist the temptation to comment on the implications lurking behind the title to Schrödinger's Chapter II, "The Method of the Most Probable Distribution." In reading this title, one can feel, at some subliminal level, the strong influence of the frequentist (ontological) mind set as opposed to the (correct) epistemological mind set that considers probabilities as a state of knowledge. Remarks to the effect that Schrödinger hated Max Born's probabilistic interpretation of his wave equation are now better appreciated.

Schrödinger was an unabashed frequentist in his thermodynamical explanations. Thus, it is very important to note right at the outset that there exists this fundamental conceptual difference in the way Schrödinger derives his equations and the way we have presented the MEP algorithm. It cannot be glossed over or ignored.

One sees Schrödinger's approach mindlessly repeated today without further comment on the conceptual divide. But, as mentioned, we will blithely ignore all of that for the time being because we simply want to show the parallels between the MEP algorithm and Schrödinger's thermodynamic equations.

In the next few sections, we translate Schrödinger's thoughts into our preferred language. It is too boring to constantly interrupt the reader by saying that when Schrödinger says so-and-so, we mean thus-and-thus in the language that we have developed so far. At the end, in the **Connections to the Literature** section, we will extensively quote Schrödinger to explicitly compare his language and ours.

To illustrate where we are heading in this Chapter, suppose you had come across one of Schrödinger's complicated looking equations concerning the Fermi oscillator. Now, like me, you might not have the faintest idea of what a "Fermi oscillator" might be, but given your understanding of the MEP algorithm, everything in Schrödinger's formula is quite easily translated over into concepts that we do understand.

27.2 Schrödinger's Chapter II

How did Schrödinger introduce the problem? Fortunately, right at the outset, he presents us with a familiar and comfortable scenario. We are not initially discouraged from accompanying him further down the road.

He says at the outset that there is really only one fundamental question in thermodynamics: Determine the disposition of N "identical systems" over some "states" in which these systems can find themselves, together with a constraint that must be satisfied.

So he could be talking about the potential disposition of N identical kangaroos with four traits when there exist certain constraints on beer and hand preference. Of course, he is not, in fact, thinking about kangaroos, but rather about atoms, molecules, photons, or electrons existing in some quantum state. Occasionally, physicists find it easier to talk about a more abstract entity called an “oscillator” that subsumes these actual physical entities.

The constraint is a fixed total energy for “an assembly” of the N systems. By “determining the disposition,” he is thinking about how many molecules, atoms, or oscillators, distribute themselves over the various energy levels. The “assembly over N identical systems” is, for Schrödinger, the ONE macroscopic physical system under consideration.

Adding to our confusion is the historical homage to Gibbs where Schrödinger then embeds these assemblies into an imagined “ensemble.” This imaginary operation was absolutely vital to his frequency interpretation of probability. It is only when thinking about how frequently an assembly of the N systems occurs within this imaginary ensemble can one posit a probability. But for us, this is taken care of immediately without recourse to the fantasy of an “ensemble,” by simply computing $P(N_1, N_2, \dots, N_n)$.

There are n states in his state space. Each state corresponds to an energy level. There is an occupation number $a_1, a_2, \dots, a_l, \dots, a_n$ indicating how many atoms, molecules, photons, electrons, or oscillators reside in the l^{th} energy state.

If not completely obvious at this point that Schrödinger has been tightly focused on frequencies, and not on probabilities, his derivation of the analog to the MEP formula dispels all doubt. He proceeds to set the objective function that is to be maximized as the multiplicity factor $W(N)$. The two constraints are (1) that the frequency counts sum to N , and (2) that the frequency count at each energy level multiplied by the energy should equal the overall total energy.

The solution of the resulting constrained optimization problem that Schrödinger has set up is solved by the same method of Lagrange multipliers we used to find the MEP formula. The form of Schrödinger’s equation is again quite familiar even though his motivation from the frequency perspective was entirely different.

We relied on information entropy as the objective function to be maximized in order to obtain numerical assignments to *distribution of probabilities*. Schrödinger relied on frequency counts, actual or anticipated, in order to obtain *distribution of counts*. The MEP algorithm relies upon information as mathematical expectations over arbitrary mappings from the state space, while Schrödinger links his fate to the actual frequency counts, and the sum of the individual energies as multiplied by these counts.

He indicates that the “sum over states” will be very important and tells us, as hopefully we would expect a native German speaker to do, that the abbreviation Z for this term comes from the German word, *Zustandssumme*. The English translation for *Zustand* is condition, or property, or state.

27.3 Understanding Schrödinger's Equations

As mentioned before, my personal motivation for presenting this material gelled when I realized that Schrödinger's complicated equations, which made little sense to me upon first exposure, were the same equations, or implications thereof, that I had already worked out for the MEP algorithm.¹ Each of Schrödinger's equations succumbed in turn to a pattern matching exercise with the MEP template.

Here is an example of what I mean. After the development of his formula for occupancy numbers, Schrödinger continues on, and eventually writes down his Equation (2.13), which looks like this,

$$d(F + U\mu) = \mu \left(dU - \frac{1}{N} \sum_l a_l d\epsilon_l \right)$$

Now this expression, it must be admitted, is a bit obscure. Depending on your level of persistence, you might never find out what it is referring to. However, if you have just come from studying the MEP formula simply as a practical means for assigning numerical values to probabilities as we have done, then there is some hope.

The first bit of pattern matching comes from trying to make sense of the left hand side of the equation. First, identify F with $\ln Z$ as Schrödinger has just defined $\log \sum_l e^{-\mu\epsilon_l} = F$. U is Schrödinger's term for the average energy, so we immediately translate this to $\langle F \rangle$. μ is Schrödinger's version for the one Lagrange multiplier. Already we are making great headway because we see that the left hand side of the equation translates to an expression $d(\ln Z - \lambda \langle F \rangle)$.²

But $\ln Z - \lambda \langle F \rangle$ is the Legendre transformation first studied in Chapter Twenty Four. It provides us with the maximum information entropy of our assignment under some model with parameters λ and $\langle F \rangle$. So Schrödinger's Equation (2.13) has something to do with the total differential of the maximum entropy. Knowing this much gives us the courage to delve further.

Now, we will list some of Schrödinger's equations as they appear in Chapter II, and remark on their parallels to the MEP formula. In his Equation (2.6), Schrödinger writes for the average energy U ,

$$U = \frac{\sum \epsilon_l e^{-\mu\epsilon_l}}{\sum e^{-\mu\epsilon_l}}$$

Pattern matching to our familiar templates, U , the average energy, must be analogous to the average $\langle F \rangle$ of the constraint function $F(X = x_i)$ in turn defined as an energy level ϵ_l . $e^{-\mu\epsilon_l}$ is the numerator of the MEP assignment with the Lagrange multiplier $-\mu$, while $\sum e^{-\mu\epsilon_l}$ is the normalizing factor, the partition

¹When I say "I" accomplished something, I mean, of course, that Jaynes had accomplished it.

²Unfortunately, I am using two different notations for the differential throughout the book. The typography for the differential changes to either dx or dx depending on trying to match up with whomever I am currently referring to.

function $Z(\lambda)$. Together they form the numerical assignment to the probability for the statement that an atom, molecule, photon, or oscillator has energy level ϵ_l .

Well, that's the way I would say it. For Schrödinger, it was used instead to calculate the “most probable” occupation number a_l , (my N_i^j), with the greatest probability.

Thus, if we match up our Q_i with Schrödinger's expression,

$$Q_i \equiv \frac{e^{-\mu\epsilon_l}}{\sum e^{-\mu\epsilon_l}}$$

we find that the MEP parallel is,

$$U = \langle F \rangle = \sum_{i=1}^n F(X = x_i) Q_i$$

Schrödinger writes another expression for the average energy,

$$U = -\frac{\partial \log \sum e^{-\mu\epsilon_l}}{\partial \mu}$$

for which the MEP parallel is,

$$\langle F \rangle = \frac{\partial \ln Z(\lambda)}{\partial \lambda}$$

Another example of an obscure and unmotivated equation that appears mid-way through Schrödinger's explanation (bottom of page 12) is,

$$G = \log \sum_l e^{-\mu\epsilon_l} - \mu \frac{\partial \log \sum_l e^{-\mu\epsilon_l}}{\partial \mu}$$

which, after some pattern matching to the MEP algorithm is nothing more than the Legendre transformation,

$$H_{max}(Q_i) = \ln Z(\lambda) - \lambda \langle F \rangle$$

In his Equation (2.20) repeated below, he presents this relationship,

$$\Psi = S - \frac{U}{T}$$

S is thermodynamic entropy, U is the mean energy, and T is the temperature. Schrödinger tells us something pretty important: “We have thus obtained a general prescription – applicable to all cases – … for obtaining the thermodynamics of a system from its mechanics.”

But within the MEP formalism, we could have derived Schrödinger's expression in analogous fashion, with no accompanying mystery as to where everything comes from.

Start with the Legendre transformation. Then substitute for the model's single parameter as statistical mechanics would have us do it as we saw in the last Chapter, $\lambda \equiv -(1/kT)$.

$$\begin{aligned} H &= \ln Z - \lambda \langle F \rangle \\ \ln Z &= H + \lambda \langle F \rangle \\ \ln Z &= H - \frac{1}{kT} \langle F \rangle \\ k \ln Z &= k \left(H - \frac{\langle F \rangle}{kT} \right) \\ &= kH - \frac{\langle F \rangle}{T} \end{aligned}$$

Making the following equivalencies between our notation and Schrödinger's, we then arrive at Schrödinger's Equation (2.20),

$$\begin{aligned} k \ln Z &\equiv \Psi \\ kH &\equiv S \\ \langle F \rangle &\equiv U \\ \Psi &= S - \frac{U}{T} \end{aligned}$$

And to put this all to rest, here is Schrödinger's final difficult and obscure equation of Chapter II, Equation (2.23), which he calls a "well-known formula in general thermodynamics,"

$$-\frac{\partial \log \sum e^{-\mu \epsilon_l}}{\partial \mu} = T^2 \frac{\partial \Psi}{\partial T}$$

However, this equation turns out not to be so obscure and difficult after all, if one just manipulates the equation as if it had arrived from the MEP algorithm. Pattern matching once again with our MEP formula,

$$\begin{aligned}
-\frac{\partial \log \sum e^{-\mu \epsilon_l}}{\partial \mu} &\equiv \frac{\partial \ln \sum_{i=1}^n e^{\lambda F(X=x_i)}}{\partial \lambda} \\
\frac{\partial \ln \sum_{i=1}^n e^{\lambda F(X=x_i)}}{\partial \lambda} &= \frac{\partial \ln Z}{\partial \lambda} \\
\frac{\partial \ln Z}{\partial \lambda} &= \langle F \rangle \\
\Psi &\equiv kH - \frac{\langle F \rangle}{T} \\
\frac{\partial \Psi}{\partial T} &= \frac{\langle F \rangle}{T^2} \\
T^2 \frac{\partial \Psi}{\partial T} &= \langle F \rangle \\
-\frac{\partial \log \sum e^{-\mu \epsilon_l}}{\partial \mu} &= T^2 \frac{\partial \Psi}{\partial T}
\end{aligned}$$

27.4 Schrödinger's Examples

In his Chapter IV, Schrödinger presents three examples of the various thermodynamic equations as they were developed in his Chapter II. For reasons that will always escape me, some authors present their examples in order of *decreasing* difficulty. Schrödinger is guilty of this here as his three examples are in, order, (a) an ideal monatomic gas, (b) Planck's oscillator, and (c) the Fermi oscillator.³

We shall discuss the last two of Schrödinger's examples, but in reverse order. His final example, the Fermi oscillator, being the easiest of all, will be discussed first. The emphasis will be on how straight forward the solutions are when examined in the context of the MEP formalism. What I find fascinating is to see exactly how he treats the state space in these examples, but without commenting on the importance (or lack thereof) this has for statistical mechanics.

Schrödinger's objective in all of these examples is to find his Ψ function where $\Psi \equiv k \ln Z$. Then his fundamental thermodynamic equation, Equation (2.23), can be applied. But how are the state space and n defined?

If you are going to calculate the partition function, you surely need a clear definition of what n refers to. It is at this juncture that physicists are not doing

³Obviously, these oscillators take their names from two of the most famous physicists of the 20th Century, Enrico Fermi and Max Planck. Schrödinger uses an "apostrophe s" to designate *Planck's oscillator*, but does not similarly call the other oscillator *Fermi's oscillator*.

a very good job at explaining things clearly. Pay attention to the changing status of the state space and n as we work through Schrödinger's examples. If you are like me, you will find yourself floundering for a firm foot-hold on the ever-changing status of the state space.

27.4.1 Fermi oscillator

With Schrödinger's exposition of the *Fermi oscillator*, we can add another interesting case of an $n = 2$ dimensional state space. This one doesn't rely on the frivolous game of tossing coins. Nonetheless, all of the generic MEP formulas designed to deal with the coin tossing problem are directly applicable. Schrödinger discussed the *Fermi oscillator* as a straightforward example of his initial theoretical development of statistical thermodynamics as developed in his Chapter II.

Our objective is to emphasize that the generic MEP formulas that we developed for the humbler coin tossing problem still apply. So we start out by looking at the Fermi oscillator in exactly the same way as we began Volume II, that is, by looking at how the MEP algorithm makes a numerical assignment to the probability for HEADS and TAILS.

Then, we can substitute the unique notation derived from the status of the problem as a *physics* problem. The mapping from the statements in the state space are mappings to the energy levels, $F(X = x_i) \equiv \epsilon_i$. There is, as well, the special interpretation of the model parameter λ as the conjunction of Boltzmann's constant k and the temperature T . We will transition our generic notation as we applied it in the coin tossing scenario over to the Fermi oscillator.

Suppose some actual physical object has only two energy levels, 0 and ϵ . Thus, the mapping is represented by $F(X = x_1) = 0$ and $F(X = x_2) = \epsilon$. Given that this is a thermodynamics problem, the Lagrange multiplier λ is always interpreted as $-\frac{1}{kT}$.

The numerical assignment to the probability for each of the two statements in the state space then turns out to be an easy exercise in applying the MEP formula. First, we write the MEP formula out in its generic format using our standard shortcut of $Q_i \equiv P(X = x_i | \mathcal{M}_k)$.

$$\begin{aligned} Q_1 &= \frac{e^{\lambda F(X = x_1)}}{Z} \\ Q_2 &= \frac{e^{\lambda F(X = x_2)}}{Z} \end{aligned}$$

The constraint function takes the first statement, $(X = x_1) \equiv$ "The electron is in the lowest energy state.", where now the first statement is translated into the analogous physics statement, and maps it to $F(X = x_1) = 0$. It takes the second statement, $(X = x_2) \equiv$ "The electron is not in the lowest energy state." and maps

it to $F(X = x_2) = \epsilon$. With these definitions, the numerical assignments to the probability for each statement follow the MEP formula,

$$\begin{aligned} Q_1 &= \frac{e^{\lambda F(X = x_1)}}{\sum_{i=1}^2 e^{\lambda F(X = x_i)}} \\ Q_2 &= \frac{e^{\lambda F(X = x_2)}}{\sum_{i=1}^2 e^{\lambda F(X = x_i)}} \\ Q_1 &= \frac{e^{\lambda \times 0}}{e^{\lambda F(X = x_1)} + e^{\lambda F(X = x_2)}} \\ Q_2 &= \frac{e^{\lambda \times \epsilon}}{e^{\lambda F(X = x_1)} + e^{\lambda F(X = x_2)}} \\ Q_1 &= \frac{e^0}{e^0 + e^{\lambda \epsilon}} \\ Q_2 &= \frac{e^{\lambda \epsilon}}{e^0 + e^{\lambda \epsilon}} \end{aligned}$$

With the substitution of the thermodynamic interpretation for the one parameter of the model, we have,

$$\begin{aligned} Q_1 &= \frac{1}{1 + e^{\lambda \epsilon}} \\ Q_2 &= \frac{e^{\lambda \epsilon}}{1 + e^{\lambda \epsilon}} \\ Q_1 &= \frac{1}{1 + e^{-\frac{\epsilon}{kT}}} \\ Q_2 &= \frac{e^{-\frac{\epsilon}{kT}}}{1 + e^{-\frac{\epsilon}{kT}}} \end{aligned}$$

From this we can easily discern that the partition function must be,

$$Z = 1 + e^{-\frac{\epsilon}{kT}}$$

So, Schrödinger has his $\Psi \equiv k \ln Z$, and then utilizes his Equation (2.23),

$$U = T^2 \frac{\partial \Psi}{\partial T}$$

which we just derived ourselves in section 27.3 by mapping to the MEP notation.

Writing out the expressions for $k \ln Z$ and the mean energy U ,

$$\begin{aligned}\Psi &= k \ln Z \\ &= k \ln [1 + e^{-\frac{\epsilon}{kT}}] \\ \frac{\partial \Psi}{\partial T} &= \frac{\epsilon e^{-\frac{\epsilon}{kT}}}{(1 + e^{-\frac{\epsilon}{kT}}) T^2} \\ U &= T^2 \frac{\partial \Psi}{\partial T} \\ U &= T^2 \frac{\epsilon e^{-\frac{\epsilon}{kT}}}{(1 + e^{-\frac{\epsilon}{kT}}) T^2} \\ &= \frac{\epsilon e^{-\frac{\epsilon}{kT}}}{1 + e^{-\frac{\epsilon}{kT}}}\end{aligned}$$

At this point, we can double-check that Schrödinger's answer for the average energy corresponds to the answer obtained from first principles. By definition, the average energy U is the constraint function average,

$$\begin{aligned}\langle F \rangle &= F(X = x_1) Q_1 + F(X = x_2) Q_2 \\ &= \left(0 \times \frac{1}{1 + e^{\lambda\epsilon}}\right) + \left(\epsilon \times \frac{e^{\lambda\epsilon}}{1 + e^{\lambda\epsilon}}\right) \\ &= \frac{\epsilon e^{\lambda\epsilon}}{1 + e^{\lambda\epsilon}} \\ \langle F \rangle &= \frac{\epsilon e^{-\frac{\epsilon}{kT}}}{1 + e^{-\frac{\epsilon}{kT}}}\end{aligned}$$

But Schrödinger actually reports the average energy in this format,

$$U = \frac{\epsilon}{e^{\frac{\epsilon}{kT}} + 1}$$

There is just one more adjustment to get the expression arrived at above into the form as reported by Schrödinger.

$$\begin{aligned}\frac{e^{-\frac{\epsilon}{kT}}}{1 + e^{-\frac{\epsilon}{kT}}} \times \frac{e^{\frac{\epsilon}{kT}}}{e^{\frac{\epsilon}{kT}}} &= \frac{1}{e^{\frac{\epsilon}{kT}} + 1} \\ U &= \epsilon \times \frac{1}{e^{\frac{\epsilon}{kT}} + 1} \\ U &= \frac{\epsilon}{e^{\frac{\epsilon}{kT}} + 1}\end{aligned}$$

This relatively easy derivation for the Fermi oscillator as motivated by matching up with the MEP algorithm for the coin tossing scenario should be contrasted with the rather more difficult derivation one usually sees in physics textbooks that lack such a motivation.

Notice how the partition function was defined here for the Fermi oscillator as a sum over just two statements. In the next example, the partition function is a sum over an “infinite” number of statements concerning the energy level of the particle that it said to possess it.

27.4.2 Planck's oscillator

We now sketch out the salient aspects of Schrödinger's second, and somewhat harder example, *Planck's oscillator*. This formula is very important in the history of physics because it eventually appears in Planck's derivation of the correct formula for black-body radiation. This marked the beginning of the quantum world view.

Before, when we were discussing the Fermi oscillator, there existed just two possibilities: no energy, or one tiny “packet” of energy ϵ . Now, however, Planck's oscillator, instead of being restricted to just two energy levels of 0 and ϵ , may take on a whole panoply of energy levels. These is an energy level consisting of no energy packets, an energy level consisting of one packet of energy, an energy level consisting two packets of energy, and so on, up to an energy level consisting of an “infinite” number of these tiny energy packets.

The final solution is framed in terms of the relevant physical quantities of energy, temperature, Planck's constant, and Boltzmann's constant. Thus, the tiny little packets of energy are defined in *ergs* as $\epsilon \equiv h\nu$. This is Planck's constant h in *erg seconds* times the frequency ν in *cycles per second*.

And, the Lagrange multiplier λ , as we have seen before, is interpreted in thermodynamics as the physical quantities k and T . This is Boltzmann's constant k in *ergs per degree Kelvin* times temperature T in *degrees Kelvin*. When we divide $h\nu$ by kT we are then dividing *ergs* by *ergs* and so have a dimensionless number, which is just what we want.

But this insertion of actual physical quantities can be delayed until the end. It is more transparent to stick to our more general notation in the derivation. The critical partition function Z , which was so easily formed for the Fermi oscillator, now looks like this infinite sum,

$$Z = \sum_{i=0}^{n \rightarrow \infty} e^{\lambda i \epsilon} = e^{\lambda(0\epsilon)} + e^{\lambda(1\epsilon)} + e^{\lambda(2\epsilon)} + \dots + e^{\lambda(n\epsilon)} \text{ as } n \rightarrow \infty$$

It appears that the state space ($X = x_i$) becomes infinite, and grudgingly, for the first time, we abandon our finite state space. The mapping from the statements in the state space to energy is $F(X = x_i) \equiv i\epsilon$. We now include a statement ($X = x_0$).

The i^{th} statement reads that a particle possesses i little packets of energy ϵ . When we multiply the energy by λ , that is, multiply by $-\frac{1}{kT}$, we obtain our dimensionless number as the argument to the exponential, $\lambda F(X = x_i) \equiv -(i h \nu)/(k T)$.

In order to have a solution, the partition function consisting of an infinite sum must converge to some finite value. To show convergence, retain the parameter as λ before translating it over into its thermodynamic aspect as $\lambda \equiv -\frac{1}{kT}$. Fortunately, this infinite sum for the partition function turns out to have a simple analytical solution (see Exercise 27.7.7),

$$Z = \frac{1}{1 - e^{\lambda\epsilon}}$$

Since we have now found the partition function Z , the average energy U can be expressed analogously to the constraint function average $\langle F \rangle$. To find $\langle F \rangle$ directly from the MEP formula carry out the partial differentiation,

$$\langle F \rangle = \frac{\partial \ln Z}{\partial \lambda} \equiv \frac{\epsilon e^{\lambda\epsilon}}{1 - e^{\lambda\epsilon}}$$

This answer can be transformed to,

$$U = \frac{\epsilon}{e^{-\lambda\epsilon} - 1}$$

The final step is to insert the physical quantities of energy $\epsilon \equiv h\nu$ together with the model parameter $\lambda \equiv -(1/kT)$ into this equation,

$$U = \frac{h\nu}{e^{h\nu/kT} - 1}$$

Thus, the core idea of forming the partition function, and then taking the partial derivative of the log transform with respect to the temperature parameter remains central. Schrödinger finds the average energy U from his expression involving the partial derivative of the log of the partition function, just as was done for the Fermi oscillator. Thus, $\Psi = k \ln Z$ and $U = T^2 \frac{\partial \Psi}{\partial T}$ still play the pivotal roles.

The difficulties that Schrödinger pays attention to in his example of Planck's oscillator are more to do with mathematical technicalities rather than with the overarching application of the MEP algorithm. We will look at those technicalities in an exercise.

But notice the change in the state space from $n = 2$ to an infinite n . The “particles” can possess any number of these tiny little packets of energy. And the transition from one state space for the Fermi oscillator to the next for Planck's oscillator is passed over without commentary. I suspect that the main psychological reason for this is Schrödinger's, and everyone else's, inordinate attention directed at the occupancy counts as the definition of what a probability must mean.

27.5 Thermodynamic Example Revisited

We are going to revisit the simplified thermodynamics example discussed in the last Chapter. How would Schrödinger explain such an example given his tutorial in Chapter II? It must center around his fundamental background concept that out of the total number N of molecules, all of which possess some particular energy level, a_l of these occur at the l^{th} energy level. His a_l and our N_i^j are both frequency counts.

Like Boltzmann and Planck before him, Schrödinger could not free himself from an ontological mind-set. He thought that the problem must be cast in terms of actual frequency counts. Then, to arrive at his desired goal, these frequency counts were allowed to approach an infinite number.

These famous men, along with many others to follow, could never cross that great conceptual divide to the point where they could think about probabilities from an epistemological mind-set. Probabilities representing a degree of belief in the truth of some statement, and consequently an extension of logical reasoning, was a bridge too far. Apparently, an Information Processor inserting information under some model into a probability distribution was just too alien of a concept. Hard core physicists did not indulge in such folly.

Then, for these reasons, Schrödinger did not maximize the *information entropy* expression,

$$H(Q_i | \mathcal{M}_k) = - \sum_{i=1}^n Q_i \ln Q_i$$

in his derivation as we did in arriving at the MEP formula. Instead, he maximized the log of the multiplicity factor, $\ln W(N)$, and set up the constraints in terms of frequencies as well. Remember also that this is four years before Claude Shannon, and some thirteen years before Jaynes.

He begins by saying the problem concerns “ N identical systems.” But there is an immediate problem if one interprets the phrase “identical systems” literally. It wouldn’t make any sense to construct the multiplicity factor if each of the N systems couldn’t be individually distinguished. Presumably, he means the N items are all of one type, like atoms, molecules, electrons, photons, *etc.* with the possibility that they may not, for physical reasons, be distinguishable after all. This ramification could then be investigated later on.

Now, any one of these molecules could be in one of some number of states. Thus, at a minimum, each “identical system” *could* be distinguished by being in one of these states. He was unhappy with such a description because, from the perspective of quantum mechanics, “to adopt such a view was to think along severely classical lines.” If not obvious by now, we do indeed in all of our applications think along these “severely classical lines.”

By this we mean that, contrary to quantum mechanics, every measurement results in the object being in one, and only one, definite state. Our development so far does not allow us any option on this score.

Then Schrödinger matched up the l^{th} state with an energy level, ϵ_l , where a_1 of the N identical systems had energy ϵ_1 , a_2 of the N identical systems had energy ϵ_2 , \dots , and a_n of the N identical systems had energy ϵ_n . These a_l were called the *occupation numbers*, as indeed they are.

Next, he presents the multiplicity factor $W(N)$ which he suspiciously labels with the notation of P ,⁴

$$P = \frac{N!}{a_1! a_2!, \dots, a_l!, \dots}$$

to count up the number of different ways that out of the total of N systems, a_1 could be in energy state ϵ_1 , and so on. Of course, the a_l had to sum to N , and the total amount of energy was $E = \sum_l \epsilon_l a_l$.

Thus, he maximized the objective function $\ln P$ as opposed to any kind of information entropy function. This objective function was subject to the two frequency constraints just mentioned above, $\sum_{l=1}^N a_l = N$, and $E = \sum_l \epsilon_l a_l$ as opposed to the probabilistic constraints in the MEP algorithm. The resulting equations are nonetheless parallel to the ones issuing from the MEP formula. At this juncture, we require a numerical example to see where we are with all of this.

27.5.1 The example involving 200 molecules

Suppose that there are $N = 200$ hydrogen molecules H_2 constituting an “ideal gas.” These are to serve as Schrödinger’s N identical systems. Let the total energy be $E = 500$, so that the average energy with respect to N is $U = E/N = 2.5$. The analog within the MEP is to specify the mathematical expectation of the mapping with respect to a probability distribution, $\langle E \rangle = 2.5$.

We start out by trying to slowly implement Schrödinger’s plan. But here we want to transition to the notation we have been using all along. First and foremost, we have to find the occupation numbers N_1, N_2, \dots, N_8 that satisfy the two constraints concerning the total number of molecules and the total energy.

In other words, only those N_i such that $\sum_{i=1}^8 N_i = 200$ and $\sum_{i=1}^8 N_i E_i = 500$ are acceptable. Then, *after* finding an acceptable set of such occupation numbers, we try to find which one of these acceptable sets has the largest multiplicity factor.

There is an insight here that applies equally well to the MEP algorithm. In explaining the MEP, the emphasis always seems to get maneuvered onto the idea of maximizing the information entropy.

⁴Just as Boltzmann labeled it as W for *Wahrscheinlichkeit*.

Actually, things are closer to the truth if this traditional emphasis is reversed. First and foremost, it is the constraints that must be satisfied, or, in other words, the explicitly specified information must find its way into the probability distribution. *After that primary goal has been accomplished*, we want an assurance that no other unwanted extraneous information has insidiously snuck into the distribution. That, of course, is guaranteed by maximizing the quantitative measure of missing information, the information entropy.

The beauty of the MEP algorithm is that it accomplishes both of these tasks simultaneously without us having to worry about either one separately. But, to repeat myself, it *is* instructive to separate out these two tasks, as we are now going to do with Schrödinger's frequency counts.

Let's look at some candidate occupation numbers. As with the example in the last Chapter, let the energy levels be $E_1 = 0, E_2 = 1$, and so on. A gas with 100 molecules in energy state 3, 100 molecules in energy state 4, and no molecules in any other energy state would be an acceptable set.

Thus, we have $N = N_3 + N_4 = 200$, $E = (100 \times 2) + (100 \times 3) = 500$, and $U = \frac{E}{N} = 2.5$. This set of occupation numbers does satisfy all of the constraints, but has the relatively low multiplicity factor of,

$$W(N) = \frac{N!}{N_1! N_2! \cdots N_8!} = \frac{200!}{0! 0! 100! 100! \cdots 0!} = 9.05 \times 10^{58}$$

A gas not as concentrated in these two energy states with, say, 50 molecules in four consecutive energy states, and none in the remaining energy states is also acceptable,

$$N = N_2 + N_3 + N_4 + N_5 = 200$$

The total energy is,

$$E = (50 \times 1) + (50 \times 2) + (50 \times 3) + (50 \times 4) = 500$$

Moreover, this set of occupation numbers is to be preferred over the first set because it can happen in so many more ways,

$$W(N) = \frac{N!}{N_1! N_2! \cdots N_8!} = \frac{200!}{0! 50! 50! 50! 50! \cdots 0!} = 9.22 \times 10^{116}$$

Take this process down the road one more step. The set of occupation numbers $\{40, 40, 30, 30, 20, 20, 10, 10\}$ is also an acceptable set by the above criteria. What makes this set of occupation numbers even better than the first two is that the multiplicity factor,

$$W(N) = \frac{N!}{N_1! N_2! \cdots N_8!} = \frac{200!}{40! 40! 30! 30! 20! 20! 10! 10!} = 2.16 \times 10^{164}$$

is enormously larger than the first two candidates.

OK, what is the set of occupation numbers that satisfies the two constraints involving N and U , and also has the largest multiplicity factor? If the set of acceptable occupation numbers is $N_i^j = \{45, 37, 31, 25, 20, 17, 14, 11\}$, $N = 200$, $U = 2.5$, with a multiplicity factor of,

$$W(N) = \frac{N!}{N_1! N_2! \cdots N_8!} = \frac{200!}{45! 37! 31! 25! 20! 17! 14! 11!} = 1.25 \times 10^{165}$$

This set of occupation numbers satisfies the two constraints, and, moreover, can happen in more ways than any other set. Thus, it is the preferred solution. Or, as Schrödinger chooses to phrase it, we have just found the “most probable distribution” N_i^j of occupation numbers.

27.5.2 The MEP solution

In this set of occupation numbers that is the solution to Schrödinger’s problem, we see the typical geometric progression just as we did when examining the biased die. What does the parallel solution look like with the MEP algorithm? The same kind of table as constructed from Boltzmann’s distribution in Chapter Twenty Six is shown below as Table 27.1. The only change is the addition of an extra column.

Table 27.1: *Using the MEP algorithm to assign probabilities to the eight energy states in the state space. The expected value (rounded) for each occupancy number based on the Q_i are shown in the last column. Compare with the frequency counts from Schrödinger’s “most probable distribution.”*

E_i	$-(E_i/T)$	$e^{-E_i/T}$	Q_i	$Q_i \times E_i$	$200 \times Q_i$
0	-0.000	1.000	0.2263	0.0000	45
1	-0.198	0.820	0.1855	0.1855	37
2	-0.397	0.673	0.1521	0.3043	31
3	-0.595	0.551	0.1247	0.3742	25
4	-0.794	0.452	0.1023	0.4091	20
5	-0.992	0.371	0.0839	0.4194	17
6	-1.190	0.304	0.0688	0.4127	14
7	-1.389	0.249	0.0564	0.3948	11
$Z(T) = 4.4198$		$\sum Q_i = 1.00$	$\langle E \rangle = 2.50$	$N = 200$	

The dual parameter for the model used to assign these probabilities was selected as $\langle E \rangle = 2.5$. The temperature parameter will adjust accordingly. The temperature

parameter adjusts to $T \approx 5.03863 \equiv \lambda \approx -0.20$. This makes sense since the average energy was set to be $\langle E \rangle = 1.3922$ at a temperature of $T = 2 \equiv \lambda = -0.50$ in the last Chapter. Since the temperature has been raised, the probability for occupying the higher energy states will also be raised. The MEP formula found the numerical assignment to the distribution of probabilities over the eight energy states that satisfied the information in the model about the average energy, or equivalently, the information incorporated by the temperature.

Within the MEP formalism, we actually compute the *probability* for any set of frequency counts given the stated model. Thus, the parallel to Schrödinger's development is to compute $P(N_i^j)$ to see whether they are in fact the “most probable distribution of frequency counts.”

$$\begin{aligned} P(N_i^j) &\equiv P(N_1 = 45, N_2 = 37, \dots, N_8 = 11 \mid \mathcal{M}_k \text{ sets } \langle E \rangle = 2.5) \\ &= W(N) Q_1^{N_1} \times \dots \times Q_8^{N_8} \\ &\approx 8.32 \times 10^{-8} \end{aligned}$$

The probability of getting a slightly different set of frequency counts under this model must be smaller than the probability just calculated. For example, the probability of getting one more molecule in the lowest energy state E_1 offset by one less molecule in E_2 is,

$$P(N_1 = 46, N_2 = 36, \dots, N_8 = 11 \mid \mathcal{M}_k \text{ sets } \langle E \rangle = 2.5) \approx 8.16 \times 10^{-8}$$

The confusion starts to dissipate ever so slightly with the following realization. A constraint function average $\langle F \rangle$ within the MEP formalism will pick, out of the enormous number of sets of frequency counts possible in the sample space, *one* with a unique feature. That one set of frequency counts, as just seen in the last example, is the one with the greatest probability.

On the other hand, Schrödinger will find just one set of frequency counts, the “most probable distribution,” by setting an accurately measured temperature in Boltzmann's distribution. The two answers will match when out of all the models setting the value of $\langle F \rangle$, with no pretense that $\langle F \rangle$ has been measured, the one model most supported by the data N_i is equivalent to a frequency count dictated by a measured temperature where there is no pretense that any data N_i have been gathered.

Thus, there is a completely opposite sense in which these two problems are approached. In a typical problem, not one of the statistical mechanics variety, one thinks of a huge panoply of models with the data having the necessary role of winnowing out most of the models and supporting just a few. One doesn't know what the actual constraint functions and Lagrange multipliers are, and even less how they would be accurately measured.

In statistical mechanics, on the other hand, one *starts* with that single model, based on the Boltzmann distribution, which is assumed true. Then, from the physics of the situation, the constraint function and constraint function average can be accurately measured. No data gathering in the form of actual molecule counts at each energy level is ever attempted because the physics has already told you the answer to this question.

Moreover, in the MEP approach there is never any need to participate in the exercise, as was just done under Schrödinger's approach, of puzzling out the exact finite N_i that satisfy the constraints. This is one of the unnecessary reasons why in Schrödinger's approach it is always emphasized that $N \rightarrow \infty$. This feature is never part of the MEP approach because it is completely irrelevant.

Jaynes's die scenario is very interesting in this regard because it stands at the intersection of these two contrasting approaches. One could adopt the standard MEP approach as I have done, (and as Jaynes suggests at the beginning that he might be following), and let a massive number of frequency counts from actual dice rolls pare down the space of models. Or, one could adopt the statistical mechanics approach, (as Jaynes hints at later), and forgo all that effort by simply *measuring* the displacement of the center of gravity and the length of the axes of the cube.

But the danger here is the danger prevalent whenever one adopts a single model. There is never any chance that subsequent data will revise that model. You are stuck with it forever. Most likely, it is the definition of the state space that will change.

At the expense of conducting all those trials to obtain the N_i , and giving up the economy of the single measurement of the constraint function average, one makes the trade-off that an unsuspected model might, contrary to all expectations, be supported. That is why the formal rules always retain all of the models at the outset before any data have trickled in, and also why all of the models are treated with the same regard before any of the data appear.

27.6 Connections to the Literature

Everything here is an elaboration on Schrödinger's presentation in the opening Chapters I, II, III, and IV of his *Statistical Thermodynamics* [30]. We fulfill that promise made earlier of using Schrödinger's own words and notation, even if we must endure the tedium of constantly cross-matching his version with our notation.

What is truly remarkable is that the word *probability* is scarcely mentioned in his derivations. Nor does the concept appear prominently in any of his explanations of statistical thermodynamics. And on the very few occasions when it is mentioned, it is done so begrudgingly with an apparent haste to move on quickly.

His development is couched entirely in terms of frequencies. Only when these frequencies are allowed to approach infinity does Schrödinger tacitly assume that you and he are in agreement that, well, we have really addressed all we need to know about probability at this point.

It came as a profound shock to me when I first came across in some recounting of the early history of quantum mechanics that Schrödinger did not at all like Max Born's probabilistic interpretation of his wave function. To Schrödinger, the wave function was real, not something to be reduced to a probability.

I was aware that Schrödinger had always been very much wedded to a physical, ontological, and ultimately deductive interpretation of quantum mechanics. But so were many of his other equally prominent peers. It never hit home that the fundamental concept of probability as an epistemological concept was anathema to Schrödinger until I saw how he reasoned about statistical mechanics in these opening Chapters of *Statistical Thermodynamics*.

The occurrence of the word “probable” in the title of Chapter II, “The Method of the Most Probable Distribution,” refers to the distribution of the actual count of the number of molecules that one would expect to see, not to something nonsensical like “the most probable distribution of probabilities.” You end up with *one* distribution of *counts*, not a distribution of probabilities over the state space. As we have seen in other places within this Volume, it is a counting formula, the multiplicity factor, that is maximized, instead of any kind of information entropy involving a probability.

It seems clear to me that Schrödinger, like many others before him and after him, was irrevocably locked into a certain mind-set about probability. This was an entrenched disposition from which he never seemed able to escape.

Let there be no doubt that the above observation is in no way a reflection on Schrödinger's well-deserved reputation as a co-founder of quantum mechanics, his brilliant mathematical capabilities, or his reward of the Nobel prize. In fact, Schrödinger did a much better job of explaining, in a very concise and direct manner, some of the mathematical manipulations employed in the “MEP algorithm” than many of its professed admirers are able to do.

Let's begin with Schrödinger (Chapter I, page 1) setting up the fundamental problem in statistical mechanics,

There is essentially only one problem in statistical thermodynamics ... to determine the distribution of an assembly of N identical systems over the possible states in which this assembly can find itself ...

and later (Chapter II, page 5),

We are faced with an assembly of N identical systems. We describe the nature of any of them by enumerating its possible states, which we label as $1, 2, 3, 4, \dots, l, \dots$

We would not quibble with any of this as it is exactly the way we have introduced our abstract systems. We also assume an “assembly of N identical systems,” whether these “identical systems” be coins, dice, kangaroos, students, or molecules. Each one of them can only be described by one mutually exclusive and exhaustive joint statement ($X = x_i$) which we label with an index i where i runs from 1 to n .

He then shows us a revealing little table reproduced below as Table 27.2.

Table 27.2: Schrödinger’s table that sets up the canonical problem in statistical mechanics.

State No.	1	2	3	...	l	...
Energy	ϵ_1	ϵ_2	ϵ_3	...	ϵ_l	...
Occupation No.	a_1	a_2	a_3	...	a_l	...

His *State No.* l is our statement ($X = x_i$). His *Energy*, ϵ_l , is our mapping from a statement in the state space to a number, $F(X = x_i)$. His *Occupation No.* a_l is our frequency count N_i .

So, $\sum_l a_l = N$ is $\sum_{i=1}^n N_i = N$. Schrödinger defines the total energy as $\sum_l \epsilon_l a_l = E$ which, for us, becomes $\sum_{i=1}^n N_i F(X = x_i)$, a summation over the frequencies with which the various mappings from the statements occur. But this total energy is not a mathematical expectation, or, in other words, an average with respect to a probability distribution. He presents the multiplicity factor as,

$$P = \frac{N!}{a_1! a_2! a_3! \dots a_l! \dots}$$

which for us is,

$$W(N) = \frac{N!}{N_1! N_2! N_3! \dots N_n!}$$

Schrödinger then uses the variational mathematics of the Lagrange multiplier method in exactly the same way as explained in Chapter Twenty Five to derive his equations analogous to our MEP equations. It is important here to emphasize that in doing this, Schrödinger set the multiplicity factor as his objective function, not the information entropy. He set up his two constraint functions as $\sum_l a_l = N$ and $\sum_l \epsilon_l a_l = E$, not as $\sum_{i=1}^n Q_i = 1$ and $\sum_{i=1}^n F(X = x_i) Q_i = \langle F \rangle$.

In his Equation (2.4) he seeks the maximum of,

$$\log P - \lambda \sum_l a_l - \mu \sum_l \epsilon_l a_l$$

as contrasted with our,

$$-\sum_{i=1}^n Q_i \ln Q_i + \lambda_0 \sum_{i=1}^n Q_i + \lambda_1 \sum_{i=1}^n F(X = x_i) Q_i$$

He finishes his derivation with familiar looking expressions. But they are still thoroughly steeped in the frequency mind-set, with expressions like those appearing in his Equation (2.6),

$$\frac{a_l}{N} = \frac{e^{-\mu\epsilon_l}}{\sum e^{-\mu\epsilon_l}}$$

or,

$$a_l = N \times \frac{e^{-\mu\epsilon_l}}{\sum e^{-\mu\epsilon_l}}$$

where $a_l/N \equiv N_i/N$ is analogous to Q_i ,

$$Q_i = \frac{e^{\lambda F(X=x_i)}}{\sum_{i=1}^n e^{\lambda F(X=x_i)}}$$

and sample average energy,

$$\frac{E}{N} = U = \frac{\sum \epsilon_l e^{-\mu\epsilon_l}}{\sum e^{-\mu\epsilon_l}}$$

is analogous to our constraint function average,

$$\langle F \rangle = \sum_{i=1}^n F(X = x_i) Q_i = \frac{\sum_{i=1}^n F(X = x_i) e^{\lambda F(X=x_i)}}{\sum_{i=1}^n e^{\lambda F(X=x_i)}}$$

Schrödinger tells us he has achieved his goal, and here it is very important to note that it consists in finding definite occupation numbers for each energy level, the a_l , and NOT any kind of probability expression like $P(N_1, N_2, \dots, N_n)$.

[The equation for a_l] indicates the distribution of our N systems over their energy levels. It may be said to contain, in a nutshell, the whole of thermodynamics, which hinges entirely on this basic distribution.

It is clear that Schrödinger, reasoning from a strong frequentist perspective, has found what he considers the “most probable distribution” of the occupation numbers a_1, a_2, \dots, a_l , and NOT the probability for these occupation numbers,

$$P(N_1, N_2, \dots, N_n) = \int \cdot \int_{\sum q_i=1} P(N_1, N_2, \dots, N_n | q_1, q_2, \dots, q_n) P(q_1, q_2, \dots, q_n) dq_i$$

ensuing from the formal manipulation rules of probability as an integration over the probability for all models. Also, just as the MEP yields the relationship,

$$\langle F \rangle = \frac{\partial \ln Z}{\partial \lambda}$$

Schrödinger writes down,

$$U = -\frac{\partial \log \sum e^{-\mu\epsilon_l}}{\partial \mu}$$

Some further technical details concerning these equations are explored as exercises.

My amateur armchair analysis of Schrödinger's entrenched mind-set concerning probability is characterized by an increasingly speculative extrapolation which gets harder to defend the further out I go. Nonetheless, I believe there is some merit to seeing where it leads. In this extrapolation, I bounce back and forth between the conceptual sticking points.

In the beginning, Schrödinger's mind-set is easy to discern. The problem must be couched in terms of frequencies, and not with probabilities. Thus, understandably, his thinking as a physicist must be grounded in ontological facts as opposed to epistemological states of knowledge.

Continuing on, statistical thermodynamics must be explained in terms of an average energy, $U = E/N$, and not as a mathematical expectation of a constraint function over some probability distribution written as $\langle E \rangle$. Therefore, the problem must be cast in terms of data averages; not in terms of anything so bizarre as information. There is no one in this story called an information processor. However, there is somebody who counts events, and then computes averages with respect to these frequencies.

Thus, there cannot possibly exist an entity like an information processor who forms models which insert information into probability distributions. For such a modeler (like ourselves), data serve merely to recalibrate the current relative standing among all the models being considered.

For the frequentist, no amount of data can change an opinion once one specific model has been selected. If, for physical reasons, you believe in a fair die, you must assert that the occupation numbers a_l hover around 10,000 in $N = 60,000$ rolls. You must throw that die an enormous number of times before you start to see the normed frequencies N_i/N deviating from 1/6. Then, and only then, will you start to doubt that maybe, just maybe, that die is not fair.

The frequentist calls himself a sober empiricist who will only change his mind after an overwhelming number of observations. Not indulging in the fantasy of models, what is there to bounce that vast amount of data against? So when in doubt throw the die another 60,000 times. And, by the way, you must keep all the experimental conditions exactly the same for those next 60,000 throws.

The information processor has instead a huge panoply of models. These are being whittled down inexorably as every new observation trickles in. And, of course, each new observation, since it is actively changing the information processor's state of knowledge, provides a clue as to how to change the whole experimental set-up to gather the next observation.

Finally, for the frequentist based ontological mind-set, the multinomial distribution appears for a very different reason than it does for the probabilistic, epistemologically-minded information processor. For the information processor, it appears as a natural outgrowth of the formal manipulation rules when it is necessary to transition from one observation to many observations. For the frequentist,

on the other hand, and here I would include Boltzmann, Planck, Schrödinger, and Einstein, it defines the very meaning of probability.

So Schrödinger did not maximize $-\sum_{i=1}^n Q_i \ln Q_i$ subject to constraints defined in terms of the mathematical expectation $\sum_{i=1}^n F(X = x_i) Q_i$, but maximized instead the expression $-\sum_{i=1}^n f_i \ln f_i$ subject to *data* constraints involving $\sum_{i=1}^n N_i E_i$.

But, in the end, it all boils down to the same thing, does it not? After all, we are letting $N \rightarrow \infty$ so that N_i/N must be, in the limit, the same as probabilities p_i . So whether we use,

$$-\sum_{i=1}^n Q_i \ln Q_i \text{ or } -\sum_{i=1}^n f_i \ln f_i$$

it all amounts to the same thing? Thus went the conventional frequentist argument.

We have discussed Schrödinger at this level of detail to bring to the forefront the persistent confusion swirling about frequencies versus probabilities, data versus information, the MEP formula for assigning numerical values to probabilities based on mathematical expectation versus sample averages of data.

Jaynes also tried to emphasize this fundamental conceptual distinction between a mathematical expectation of a constraint function as information within the MEP and observed sample averages as data. But, unfortunately he did this in a way where the important signal he was trying to communicate ended up buried in a lot of underlying noise.

He was eager to demonstrate [18, pg. 270] that using the sample average of a constraint function based on the data from N observations, followed by making that sample average equal to the mathematical expectation of a constraint function as the information for the MEP formula, resulted in that model using the sample average having the highest probability when compared to any other model using different information.

This very important relationship he worked out was one between the orthodox statistical concept of the maximum likelihood estimator and information as defined by the MEP. This probability relationship was examined already in Volume I as the ratio of the probabilities for any two models when conditioned on the known data. The algebraic manipulations are very much the same as those developed when working out the probability of future frequency counts when we have the MEP formula for the Q_i .

Jaynes opening gambit is the question:

Out of all the hypotheses [...] which is most strongly supported by the data D according to the Bayesian, or likelihood, criterion? To answer this, choose any particular hypothesis $H_0 \equiv \{\lambda_1^{(0)} \dots \lambda_m^{(0)}\}$ as the “null hypothesis” and test it against any other hypothesis $H \equiv \{\lambda_1 \dots \lambda_m\}$ [...] by Bayes’ theorem.

The log-likelihood ratio in favor of H over H_0 is ...

$$L = r \left[\log(Z_0/Z) + \sum_{k=1}^m (\lambda_k^{(0)} - \lambda_k) \bar{f}_k \right]$$

Jaynes does not start his derivation off, as he should have, with that “Bayesian criterion” he mentions. It is clear from the above quote that when Jaynes uses the word *hypothesis*, he means the same as what we have been calling a *model* because each hypothesis, just as for each model, is indeed indexed by specifying the set of m parameters, the Lagrange multipliers. Thus, $H \equiv \mathcal{M}_A \equiv \{\lambda_1 \dots \lambda_m\}$ and $H_0 \equiv \mathcal{M}_B \equiv \{\lambda_1^{(0)} \dots \lambda_m^{(0)}\}$. The missing steps in this derivation and a numerical example are taken up in Exercises 27.7.20 and 27.7.21.

Then, the “Bayesian criterion” that Jaynes presupposes is, in our notation, the ratio of the probabilities for any two models given that the data \mathcal{D} have been observed,

$$\begin{aligned} \frac{P(H|D)}{P(H_0|D)} &\equiv \frac{P(\mathcal{M}_A|\mathcal{D})}{P(\mathcal{M}_B|\mathcal{D})} \\ \frac{P(\mathcal{M}_A|\mathcal{D})}{P(\mathcal{M}_B|\mathcal{D})} &= \frac{P(\mathcal{D}|\mathcal{M}_A)P(\mathcal{M}_A)}{P(\mathcal{D}|\mathcal{M}_B)P(\mathcal{M}_B)} \end{aligned}$$

If we make the assumption, as we always do, and which Jaynes does as well, that the ratio of the probabilities for any two models before any data is 1, then,

$$\frac{P(\mathcal{M}_A|\mathcal{D})}{P(\mathcal{M}_B|\mathcal{D})} = \frac{P(\mathcal{D}|\mathcal{M}_A)}{P(\mathcal{D}|\mathcal{M}_B)} \equiv \frac{P(D|H)}{P(D|H_0)}$$

The “log-likelihood ratio” part of Jaynes’s explanation becomes in our notation,

$$\ln \left[\frac{P(\mathcal{D}|\mathcal{M}_A)}{P(\mathcal{D}|\mathcal{M}_B)} \right] = N \left[\ln \left(\frac{Z_B}{Z_A} \right) + \sum_{j=1}^m (\lambda_j^A - \lambda_j^B) \bar{F}(X = x_i) \right]$$

with all of the details appearing in the exercises mentioned above.

After this brief interlude with Jaynes and the maximum likelihood estimator’s relationship to the constraint function average, let’s return to Schrödinger. At the end of his Chapter IV presenting his three examples, two of which we worked out, and after his solution for the Fermi oscillator, he states that (pg. 21),

Compare this with the second relevant term on the right hand side of the last equation of the preceding section (taking there $\epsilon = h\nu$). There is just one remarkable difference in sign, ∓ 1 in the denominator. We shall see later that this constitutes the relevant difference between ‘Einstein–Bose statistics’ and ‘Fermi–Dirac statistics’.

This presents us with an opportunity to quote Feller [7, pp. 38–42] on the distinction made in physics between the so-called “statistics” traveling under the

labels of (1) Maxwell–Boltzmann statistics, (2) Bose–Einstein statistics, and (3) Fermi–Dirac statistics. I hasten to point out, notwithstanding my general acceptance of Jaynes's displeasure about Feller, that here Feller is a paragon of clarity and simplicity.

There is no better summary and comparison of the crucial probabilistic distinctions among these three definitions. Physics texts will spend a very long time dwelling on the minutiae without ever touching upon the essence as Feller does. The details of Feller's example on his page 42 illustrating that different probabilities can be assigned to the same event are worked out in Exercise 27.7.18.

The number of placements of r balls in n cells resulting in the occupancy numbers r_1, \dots, r_n is given by the [multiplicity factor]. Assuming that all n^r possible placements are equally probable, *the probability to obtain the given occupancy numbers r_1, \dots, r_n equals,*

$$\frac{r!}{r_1! r_2! \cdots r_n!} n^{-r}$$

This assignment of probabilities was used in all applications mentioned so far, and it used to be taken for granted that it is inherent to the intuitive notion of randomness. No alternative assignment has ever been suggested on probabilistic or intuitive grounds. It is therefore of considerable methodological interest that *experience* compelled physicists to replace [the above equation] by others which originally came as a shock to intuition. . . . In physics [the above equation] is known as the *Maxwell–Boltzmann distribution*. [Emphasis in the original.]

A short while later, Feller addresses the Bose–Einstein and Fermi–Dirac statistics. The appearance of these “statistics” within physics after Boltzmann was the “shock to intuition” mentioned above.

Remember that we are here concerned only with *indistinguishable* particles. We have r particles and n cells. *By Bose–Einstein statistics we mean that only distinguishable arrangements are considered and that each is assigned probability $1/A_{r,n}$ with $A_{r,n}$ defined as,*

$$A_{r,n} = \binom{n+r-1}{n-1}$$

It is shown in statistical mechanics that this assumption holds true for photons, nuclei, and atoms containing an even number of particles. To describe other particles, a third possible assignment of probabilities must be introduced. *Fermi–Dirac statistics* is based on these hypotheses: (1) *it is impossible for two or more particles to be in the same cell, and* (2) *all distinguishable arrangements satisfying the first condition have equal probabilities.* The first hypothesis requires that $r \leq n$. An arrangement is then completely described by stating which of the n cells contain a particle; and since there are r particles, the corresponding cells can be chosen in $\binom{n}{r}$ ways. Hence, with *Fermi–Dirac statistics there are in all $\binom{n}{r}$ possible arrangements, each having probability $\binom{n}{r}^{-1}$.*

The straightforward example that Feller gives next to illustrate all of this is,

Examples. (a) Let $n = 5$, $r = 3$. The arrangement $(\star | - | \star | \star | -)$ has probability $\frac{6}{125}$, $\frac{1}{35}$, or $\frac{1}{10}$ according to whether Maxwell–Boltzmann, Bose–Einstein, or Fermi–Dirac statistics is used.

As mentioned, Exercise 27.7.18 works out the details in Feller’s example of computing probabilities under the three different kind of “statistics.” Commenting on Feller’s approach also provides me with another opportunity to review some of those tedious counting formulas introduced in Chapters Thirteen and Fifteen of Volume I. The details are worked out in Exercise 27.7.19.

Finally, let me conclude with a numerical example taken from Baierlein’s book on statistical mechanics [2, pg. 288]. His example nicely illustrates some of the confusion surrounding many of these issues broached so far in this Volume.

Baierlein’s intent is to explain the idea of occupation number for a system of N bosons. The essential distinction in statistical mechanics between fermions and bosons is that any number of bosons can crowd into an energy eigenstate, while no more than one fermion can do so. Once again, we are dealing with two of the “different statistics” of statistical mechanics, the Bose–Einstein statistics developed for bosons, and the Fermi–Dirac statistics developed for fermions. Let us compare Feller’s treatment as already discussed above with Baierlein’s explanation.

Baierlein sets out the problem thusly:

We take two bosons, $N = 2$, and three single particle states: $s = 1, 2, 3$. There are six states ψ_j for this two-boson system, with a unique correspondence between them and six *sets* of occupation numbers ... Now we may compute [the average occupation number for the first state] by summing over the (six) admissible sets of occupation numbers. This is done by taking n_1 to be 0 and performing a summation over the admissible values of the other n_i ’s, then moving on to $n_1 = 1$ and doing likewise, and finally taking $n_1 = 2$ and summing over the admissible values of the remaining n_i ’s ...

Contrast this procedure with Feller’s much simpler approach. We have a state space of $n = 3$ corresponding to the three possible energy eigenstates ε_i . Suppose that the bosons are actually photons. The three statements in the state space are: (1) “A photon has an energy eigenstate of ε_1 .”, (2) A photon has an energy eigenstate of ε_2 .”, and (3) “A photon has an energy eigenstate of ε_3 .”.

The number of frequency counts N is the number of bosons (photons) in the assembly, so $N = 2$. The *sample space* then consists of $n^N = 3^2 = 9$ elementary points. The three statements about the eigenstates are Feller’s three cells, while the two photons are the two balls **a**, **b** than can be distributed among the three cells. Table 27.3 at the top of the next page shows Feller’s layout of the three cells and two balls making up all nine possible elementary points.

Table 27.3: All nine elementary points in Baierlein's sample space illustrating an example of a system with two bosons.

Elementary point	Representation	E_j
1	(a b -)	$\varepsilon_1 + \varepsilon_2 = 3\varepsilon_1$
2	(a - b)	$\varepsilon_1 + \varepsilon_3 = 4\varepsilon_1$
3	(- a b)	$\varepsilon_2 + \varepsilon_3 = 5\varepsilon_1$
4	(b a -)	$\varepsilon_2 + \varepsilon_1 = 3\varepsilon_1$
5	(b - a)	$\varepsilon_3 + \varepsilon_1 = 4\varepsilon_1$
6	(- b a)	$\varepsilon_3 + \varepsilon_2 = 5\varepsilon_1$
7	(ab - -)	$\varepsilon_1 + \varepsilon_1 = 2\varepsilon_1$
8	(- ab -)	$\varepsilon_2 + \varepsilon_2 = 4\varepsilon_1$
9	(- - ab)	$\varepsilon_3 + \varepsilon_3 = 6\varepsilon_1$

We rely on our counting formulas to disentangle Baierlein's explanation. There are only two possible partitions for the sum $N = 2$. They are, in turn, (1) $2 + 0 + 0$ and (2) $1 + 1 + 0$. There are three possibilities for the $2 + 0 + 0$ pattern from the formula,

$$\frac{n!}{r_z! r_s! r_d!} = \frac{3!}{2! 0! 1!} = 3$$

and also three possibilities for the $1 + 1 + 0$ pattern,

$$\frac{n!}{r_z! r_s! r_d!} = \frac{3!}{1! 2! 0!} = 3$$

Taken together $3 + 3 = 6$ which we know must be the same as the total number of possible contingency tables, or frequency counts,

$$u = \frac{(N + n - 1)!}{N! (n - 1)!} = \frac{4!}{2! 2!} = 6$$

The three $2 + 0 + 0$ patterns can happen in only one way,

$$W(N) = \frac{N!}{N_1! N_2! N_3!} = \frac{2!}{2! 0! 0!} = 1$$

while the three $1 + 1 + 0$ patterns can happen in two ways,

$$W(N) = \frac{N!}{N_1! N_2! N_3!} = \frac{2!}{1! 1! 0!} = 2$$

Pleasantly, this simple deconstruction by the counting formulas does get reassembled back into the correct number of elementary points in the sample space through,

$$n^N = \sum_{j=1}^{u=6} W_j(N) = 1 + 1 + 1 + 2 + 2 + 2 = (3 \times 1) + (3 \times 2) = 9$$

Since we are supposed to be using Bose–Einstein statistics for this case involving bosons, then the probability for any of the six possible frequency counts is, using Feller’s formula,

$$P(N_1, N_2, N_3) = \frac{1}{\binom{n+r-1}{n-1}} = \frac{N! (n-1)!}{(N+n-1)!} = 1/6$$

But, of course, this is exactly the same formula we would use if the IP were *totally ignorant of the causes for a boson to appear in an energy eigenstate*. The probability for all models is the same with the parameters in the Dirichlet distribution at $\alpha_i = 1$.

Physics (or Einstein), on the other hand, would say that elementary point #1 (**a** | **b** | –) is “indistinguishable” from elementary point #4 (**b** | **a** | –), and that there are “really” only three representations for the first six elementary points.

However, I don’t see any difficulty in maintaining strict adherence to the formal manipulation rules of probability while correctly claiming that,

$$P(N_1 = 1, N_2 = 1, N_3 = 0) = 1/6$$

Now, in my estimation, both elementary point #1 and elementary point #4 are still “distinguishable,” that is, photon **a** or photon **b** in energy eigenstate 1 or 2 are distinguishable with probability 1/12.

In my opinion, it is very important for Physics to address this profound conceptual gap. Is it ignorance about what causes a photon to be in an energy eigenstate (my position from strictly adhering to the formal manipulation rules of probability), or “indistinguishability” of photons (Einstein’s position)? Or, are both positions simply dual, but equivalent, ways of representing the situation?

27.7 Solved Exercises for Chapter Twenty Seven

Exercise 27.7.1: Illustrate Schrödinger's Equations (2.6) with some easy examples.

Solution to Exercise 27.7.1

Argue from the general MEP approach as discussed so far. Suppose we take a state space of $n = 2$ with constraint functions $F(X = x_1) = \varepsilon_1$ and $F(X = x_2) = \varepsilon_2$. Then the numerical assignments are,

$$Q_1 = \frac{e^{\lambda \varepsilon_1}}{e^{\lambda \varepsilon_1} + e^{\lambda \varepsilon_2}}$$

$$Q_2 = \frac{e^{\lambda \varepsilon_2}}{e^{\lambda \varepsilon_1} + e^{\lambda \varepsilon_2}}$$

The constraint function average is,

$$\begin{aligned} \langle F \rangle &= \frac{\varepsilon_1 e^{\lambda \varepsilon_1}}{e^{\lambda \varepsilon_1} + e^{\lambda \varepsilon_2}} + \frac{\varepsilon_2 e^{\lambda \varepsilon_2}}{e^{\lambda \varepsilon_1} + e^{\lambda \varepsilon_2}} \\ &= \frac{\varepsilon_1 e^{\lambda \varepsilon_1} + \varepsilon_2 e^{\lambda \varepsilon_2}}{e^{\lambda \varepsilon_1} + e^{\lambda \varepsilon_2}} \end{aligned}$$

It's not hard to advance to a general n ,

$$\langle F \rangle = \frac{\sum_{i=1}^n \varepsilon_i e^{\lambda \varepsilon_i}}{\sum_{i=1}^n e^{\lambda \varepsilon_i}}$$

or, as Schrödinger wrote it with his different notation for the model parameter λ ,

$$U = \frac{\sum \varepsilon_l e^{-\mu \varepsilon_l}}{\sum e^{-\mu \varepsilon_l}}$$

Again, from our general MEP formalism we know that,

$$\frac{\partial \ln Z}{\partial \lambda} = \langle F \rangle$$

Therefore, using Schrödinger's notation, it must be that,

$$-\frac{\partial \log \sum e^{-\mu \varepsilon_l}}{\partial \mu} = U$$

Exercise 27.7.2: Confirm these results with *Mathematica*.

Solution to Exercise 27.7.2

Consider the $n = 2$ case just examined in the first exercise. Here,

$$\ln Z = \ln (e^{\lambda \varepsilon_1} + e^{\lambda \varepsilon_2})$$

Use *Mathematica* to make the assignment,

$$\mathbf{F} = \text{Log}[\text{Exp}[\lambda \varepsilon_1] + \text{Exp}[\lambda \varepsilon_2]]$$

as Schrödinger uses this notation $\log \sum_l e^{-\mu \varepsilon_l} = F$ in his Equation (2.11).

Have *Mathematica* evaluate the partial derivative of the log of the partition function with respect to the parameter, $D[\mathbf{F}, \lambda]$. It returns the answer,

$$\frac{\varepsilon_1 e^{\lambda \varepsilon_1} + \varepsilon_2 e^{\lambda \varepsilon_2}}{e^{\lambda \varepsilon_1} + e^{\lambda \varepsilon_2}}$$

confirming the answer we arrived at from basic MEP principles in the first exercise.

Exercise 27.7.3: Relying upon *Mathematica* once again, find the partial derivative of the log of the partition function with respect to each constraint function instead of the Lagrange multiplier.

Solution to Exercise 27.7.3

If we re-define \mathbf{F} according to Schrödinger as ,

$$\mathbf{F} = \text{Log}[\text{Sum}[\text{Exp}[-\mu \varepsilon_i], \{i, 1, n\}]]$$

Ask *Mathematica* to now differentiate \mathbf{F} with respect to ε_1 and ε_2 for $n = 2$,

$$D[\mathbf{F}, \varepsilon_1] \text{ and } D[\mathbf{F}, \varepsilon_2]$$

The results are,

$$\frac{\partial F}{\partial \varepsilon_1} = -\frac{\mu e^{-\mu \varepsilon_1}}{e^{-\mu \varepsilon_1} + e^{-\mu \varepsilon_2}}$$

$$\frac{\partial F}{\partial \varepsilon_2} = -\frac{\mu e^{-\mu \varepsilon_2}}{e^{-\mu \varepsilon_1} + e^{-\mu \varepsilon_2}}$$

Because Schrödinger was wedded to the frequentist view, the occupancy counts a_1 and a_2 were interpreted as (reference his Equation (2.6)),

$$a_1 = N \times \frac{e^{-\mu \varepsilon_1}}{e^{-\mu \varepsilon_1} + e^{-\mu \varepsilon_2}}$$

$$a_2 = N \times \frac{e^{-\mu \varepsilon_2}}{e^{-\mu \varepsilon_1} + e^{-\mu \varepsilon_2}}$$

Substitute $\frac{\partial F}{\partial \varepsilon_l}$ and multiply by $-(1/\mu)$,

$$\begin{aligned} -\frac{1}{\mu} \times \left(-\frac{\mu e^{-\mu \varepsilon_1}}{e^{-\mu \varepsilon_1} + e^{-\mu \varepsilon_2}} \right) &= \frac{e^{-\mu \varepsilon_1}}{e^{-\mu \varepsilon_1} + e^{-\mu \varepsilon_2}} \\ -\frac{1}{\mu} \times \left(-\frac{\mu e^{-\mu \varepsilon_2}}{e^{-\mu \varepsilon_1} + e^{-\mu \varepsilon_2}} \right) &= \frac{e^{-\mu \varepsilon_2}}{e^{-\mu \varepsilon_1} + e^{-\mu \varepsilon_2}} \\ -\frac{N}{\mu} \times \left(-\frac{\mu e^{-\mu \varepsilon_1}}{e^{-\mu \varepsilon_1} + e^{-\mu \varepsilon_2}} \right) &= N \times \frac{e^{-\mu \varepsilon_1}}{e^{-\mu \varepsilon_1} + e^{-\mu \varepsilon_2}} \\ -\frac{N}{\mu} \times \left(-\frac{\mu e^{-\mu \varepsilon_1}}{e^{-\mu \varepsilon_1} + e^{-\mu \varepsilon_2}} \right) &= N \times \frac{e^{-\mu \varepsilon_2}}{e^{-\mu \varepsilon_1} + e^{-\mu \varepsilon_2}} \\ a_1 &= -\frac{N}{\mu} \frac{\partial \log(e^{-\mu \varepsilon_1} + e^{-\mu \varepsilon_2})}{\partial \varepsilon_1} \\ a_2 &= -\frac{N}{\mu} \frac{\partial \log(e^{-\mu \varepsilon_1} + e^{-\mu \varepsilon_2})}{\partial \varepsilon_2} \end{aligned}$$

And, so we see the truth in the second line of Schrödinger's Equation (2.6),

$$a_l = -\frac{N}{\mu} \frac{\partial F}{\partial \varepsilon_l}$$

Exercise 27.7.4: Look up the definition for the “total differential.” Find where Schrödinger uses the total differential in his exposition.

Solution to Exercise 27.7.4

Schrödinger employs the total differential without explicitly saying so on page 11, Chapter II, in developing his Equation (2.12). He says,

... let us write down, using (2.6), an undoubtedly correct mathematical relation ...

$$dF = \frac{\partial F}{\partial \mu} d\mu + \sum_l \frac{\partial F}{\partial \varepsilon_l} d\varepsilon_l$$

The definition of the total differential for a function consisting of two arguments $f(x, y)$ is,

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy$$

and for any number of arguments x_i ,

$$df = \sum_{i=1}^n \frac{\partial f}{\partial x_i} dx_i$$

In Equation (2.11), Schrödinger defines the log of the partition function as,

$$F = \log \sum_l e^{-\mu \varepsilon_l}$$

where F has as arguments μ and all of the ε_l . Thus, the total differential for the log of the partition function F is,

$$dF = \frac{\partial F}{\partial \mu} d\mu + \sum_l \frac{\partial F}{\partial \varepsilon_l} d\varepsilon_l$$

Returning to the generic MEP notation and the $n = 2$ case for an easy illustration, we have,

$$d(\ln Z) = \frac{\partial \ln Z}{\partial \lambda} d\lambda + \frac{\partial \ln Z}{\partial \varepsilon_1} d\varepsilon_1 + \frac{\partial \ln Z}{\partial \varepsilon_2} d\varepsilon_2$$

Now, the fact that the partial differentiation of the log of the partition function with respect to the model parameter results in the constraint function average is part of the MEP formalism. We have mentioned it several times already,

$$\frac{\partial \ln Z}{\partial \lambda} = \langle F \rangle$$

We have also worked out the solution for this variation on the “coin tossing scenario” as,

$$\begin{aligned} \langle F \rangle &= \sum_{i=1}^n F(X = x_i) Q_i \\ &= F(X = x_1) Q_1 + F(X = x_2) Q_2 \\ &= \left(\varepsilon_1 \times \frac{e^{\lambda \varepsilon_1}}{e^{\lambda \varepsilon_1} + e^{\lambda \varepsilon_2}} \right) + \left(\varepsilon_2 \times \frac{e^{\lambda \varepsilon_2}}{e^{\lambda \varepsilon_1} + e^{\lambda \varepsilon_2}} \right) \\ &= \frac{\varepsilon_1 e^{\lambda \varepsilon_1} + \varepsilon_2 e^{\lambda \varepsilon_2}}{e^{\lambda \varepsilon_1} + e^{\lambda \varepsilon_2}} \end{aligned}$$

Our $\langle F \rangle$ is Schrödinger’s U , so that the first term in Schrödinger’s Equation (2.12) becomes,

$$dF = -U d\mu + \sum_l \frac{\partial F}{\partial \varepsilon_l} d\varepsilon_l$$

The sum in the second term was found in Exercise 27.7.3 as,

$$-\frac{N}{\mu} \frac{\partial F}{\partial \varepsilon_l} = a_l$$

We have,

$$\begin{aligned} dF &= -U d\mu + \sum_l \frac{\partial F}{\partial \varepsilon_l} d\varepsilon_l \\ &= -U d\mu - \frac{\mu}{N} \sum_l a_l d\varepsilon_l \end{aligned}$$

which is Schrödinger’s Equation (2.12).

Exercise 27.7.5: Complete Exercise 27.7.4 by deriving Schrödinger's Equation (2.13).

Solution to Exercise 27.7.5

At the end of the last exercise, we were left having worked out the total differential of $F = \log \sum_l e^{-\mu\varepsilon_l}$ as,

$$dF = -Ud\mu - \frac{\mu}{N} \sum_l a_l d\varepsilon_l$$

But Schrödinger wants the total differential of this expression $F + U\mu$. The total differential of $U\mu$ is,

$$d(U\mu) = \mu dU + U d\mu$$

Adding this to what we already have results in,

$$\begin{aligned} d(F + U\mu) &= -Ud\mu - \frac{\mu}{N} \sum_l a_l d\varepsilon_l + \mu dU + U d\mu \\ &= -Ud\mu + U d\mu + \mu dU - \frac{\mu}{N} \sum_l a_l d\varepsilon_l \\ &= \mu dU - \frac{\mu}{N} \sum_l a_l d\varepsilon_l \\ &= \mu \left(dU - \frac{1}{N} \sum_l a_l d\varepsilon_l \right) \end{aligned}$$

This is Schrödinger's Equation (2.13) (Chapter II, pg. 11) which was briefly mentioned in section 27.3 as an example of how familiarity with the MEP formalism allows one to better understand what Schrödinger was up to. Pattern matching the left hand side of Schrödinger's notation for his Equation (2.13) with the MEP formalism, $F + U\mu \equiv \ln Z - \langle F \rangle \lambda$, we observe that Schrödinger has derived the total differential for information entropy from his frequentist view point.

Exercise 27.7.6: What does the average energy for Planck's oscillator reduce to at high temperature.

Solution to Exercise 27.7.6

We arrived at the following expression for the average energy U for *one* Planck's oscillator,

$$U = \frac{h\nu}{e^{\frac{h\nu}{kT}} - 1}$$

Planck's oscillator, like the Fermi oscillator, is an abstract entity that physicists classify as simple one-dimensional harmonic oscillators. They are the analogs to our coins, dice, kangaroos, or students. The property of interest that these oscillators

are assumed to possess is an “energy” which can be formulated physically in a number of different ways.

Einstein proposed the “quantum” concept that the energy property came in discrete bundles of $h\nu$ where h is Planck’s constant. The oscillator is visualized to be moving back and forth with some “natural” frequency ν in cycles per second. Since Planck’s constant is such a small number on the order of 10^{-27} , and if the oscillator is moving slowly, say at $\nu = 1 \text{ cps}$, then the energy bundles are very small indeed. Since a “high temperature” of, say $T \approx 1000K$, will bring down the Boltzmann constant k down to an order of magnitude of around 10^{-13} , we see that the fraction is approximately,

$$\frac{h\nu}{kT} \approx \frac{10^{-27} \times 1}{10^{-16} \times 10^3} \approx 10^{-14}$$

Thus,

$$e^{\frac{h\nu}{kT}} = e^{10^{-14}}$$

is hardly different from $e^0 = 1$.

In fact, the Taylor series expansion for the exponential function, keeping just the first term, is $e^x = 1 + x$. Substituting this into the average energy U ,

$$\begin{aligned} U &= \frac{h\nu}{e^{\frac{h\nu}{kT}} - 1} \\ &= \frac{h\nu}{1 + \frac{h\nu}{kT} - 1} \\ &= kT \end{aligned}$$

From this point it is very easy to determine what physicists call the “heat capacity” at this high temperature to be,

$$C = \frac{dU}{dT} = k$$

The rate of change of the average energy with respect to changes in the temperature is a constant value, namely, Boltzmann’s constant k .

From our perspective, we are more interested in the analogous MEP finding concerning the change in the constraint function average with respect to changes in the model parameter,

$$g = \frac{\partial \langle F \rangle}{\partial \lambda} \text{ and } g^{-1} = \frac{\partial \lambda}{\partial \langle F \rangle}$$

which is more than likely to be highly non-linear. Considering the complementary set of dual parameters, the constraint function averages and Lagrange multipliers, how much does one parameter change when a dual parameter is changing? This finding foreshadows much more to come in Volume III about the metric tensor’s role in Information Geometry.

Exercise 27.7.7: Show how the infinite sum in the partition function for Planck's oscillator is resolved.

Solution to Exercise 27.7.7

The mathematical resolution to these kinds of questions about the partition function are important for the MEP. The sum in the partition function for Planck's oscillator,

$$Z_{\text{Planck}} = e^{\lambda \times 0 \times \varepsilon} + e^{\lambda \times 1 \times \varepsilon} + e^{\lambda \times 2 \times \varepsilon} + \cdots + e^{\lambda \times n \times \varepsilon}$$

is the sum of an infinite series. It is, in fact, the sum of a geometric series,

$$\sum_{i=0}^{n \rightarrow \infty} x^i = 1 + x + x^2 + \cdots + x^n$$

which has the solution of,

$$\sum_{i=0}^{n \rightarrow \infty} x^i = \frac{1}{1-x} \text{ when } |x| < 1$$

Let $x = e^{\lambda \varepsilon}$ so that, (if $|x| < 1$, then λ must eventually be negative)

$$\begin{aligned} Z_{\text{Planck}} &= \sum_{i=0}^{n \rightarrow \infty} e^{\lambda i \varepsilon} \\ &= \sum_{i=0}^{n \rightarrow \infty} (e^{\lambda \varepsilon})^i \\ &= \sum_{i=0}^{n \rightarrow \infty} x^i \\ &= \frac{1}{1 - e^{\lambda \varepsilon}} \end{aligned}$$

Exercise 27.7.8: Compare the partition function for Planck's oscillator with that of the Fermi oscillator.

Solution to Exercise 27.7.8

Since the dimension of the state space is $n = 2$ for the Fermi oscillator as opposed to $n \rightarrow \infty$ for the Planck oscillator, the Fermi oscillator has only the first two terms in Z above where $i = 0$ and $i = 1$. Thus,

$$\begin{aligned} Z_{\text{Fermi}} &= \sum_{i=0}^1 e^{i \lambda \varepsilon} \\ &= e^{0 \times \lambda \times \varepsilon} + e^{1 \times \lambda \times \varepsilon} \\ &= 1 + e^{\lambda \varepsilon} \end{aligned}$$

as compared to,

$$Z_{\text{Planck}} = \frac{1}{1 - e^{\lambda \varepsilon}}$$

Exercise 27.7.9: Show the partition function for Planck's oscillator after the appropriate substitutions have been made.

Solution to Exercise 27.7.9

The partition function for Planck's oscillator in generic notation was shown as,

$$Z_{\text{Planck}} = \sum_{i=0}^{\infty} e^{\lambda i \varepsilon}$$

Substituting for the energy packet, $\varepsilon \equiv h \nu$, and then for the Lagrange multiplier, $\lambda \equiv -\frac{1}{kT}$, we have,

$$Z_{\text{Planck}} = \sum_{i=0}^{\infty} e^{-\frac{i h \nu}{kT}}$$

and then,

$$Z_{\text{Planck}} = \frac{1}{1 - e^{-\frac{h \nu}{kT}}}$$

Exercise 27.7.10: In preparation for the upcoming numerical exercises, use the accepted values for Planck's constant and Boltzmann's constant.

Solution to Exercise 27.7.10

Planck's constant is,

$$h = 6.62607 \times 10^{-27} \text{ ergs second}$$

For a frequency ν of an oscillator, take the following arbitrary value,

$$\nu = 5 \times 10^{12} \text{ cycles per second}$$

Thus,

$$\begin{aligned} h \nu &= (6.62607 \times 10^{-27} \text{ ergs second}) \times (5 \times 10^{12} \text{ cycles per second}) \\ &= 3.31 \times 10^{-14} \text{ ergs} \end{aligned}$$

Boltzmann's constant is,

$$k = \frac{1.38065 \times 10^{-16} \text{ ergs}}{\text{degrees Kelvin}}$$

Let the temperature be $T = 5000 \text{ K}$. Hence,

$$\begin{aligned} kT &= \left(\frac{1.38065 \times 10^{-16} \text{ ergs}}{\text{K}} \right) \times (5000 \text{ K}) \\ &= 6.90 \times 10^{-13} \text{ ergs} \end{aligned}$$

The dimensionless number appearing as the argument to the exponential function is then,

$$\begin{aligned}-\frac{h\nu}{kT} &= -\frac{3.31 \times 10^{-14} \text{ ergs}}{6.90 \times 10^{-13} \text{ ergs}} \\ &= -0.0479924\end{aligned}$$

A typical term appearing in the numerator of Boltzmann's distribution for, say, one energy packet $i = 1$, would then look like,

$$e^{-\frac{h\nu}{kT}} = e^{-0.0479924} = 0.953141$$

Exercise 27.7.11: What are the first two terms in the partition function for Planck's oscillator?

Solution to Exercise 27.7.11

We would need to calculate the numerator for no energy packets plus one energy packet, in other words, where $i = 0$ and $i = 1$ in the sum for the partition function,

$$Z_{\text{Planck}} = \sum_{i=0}^{\infty} e^{-\frac{i h \nu}{k T}}$$

We have just calculated a value of 0.953141 for one small energy packet $h\nu$ when $i = 1$. The first term, when $i = 0$ is simply,

$$e^{-\frac{0 h \nu}{k T}} = e^0 = 1$$

The sum of the first two terms in the partition function is then,

$$1 + 0.953141 = 1.953141$$

Exercise 27.7.12: Take a brute force numerical approach in examining the plausibility of the expression for the partition function of Planck's oscillator.

Solution to Exercise 27.7.12

We arrived at this expression for the partition function for Planck's oscillator,

$$Z_{\text{Planck}} = \frac{1}{1 - e^{-\frac{h\nu}{kT}}}$$

Substituting the numerical values from the previous exercise,

$$\begin{aligned} Z_{\text{Planck}} &= \frac{1}{1 - e^{-\frac{h\nu}{kT}}} \\ &= \frac{1}{1 - e^{-0.0479924}} \\ &= 21.3406 \end{aligned}$$

An approximation to the infinite sum from the direct definition of the partition function by taking just the first n terms, is,

$$Z_{\text{Planck}} \approx \sum_{i=0}^n e^{-\frac{i h \nu}{k T}} = 1 + e^{-\frac{h \nu}{k T}} + e^{-\frac{2 h \nu}{k T}} + \cdots + e^{-\frac{n h \nu}{k T}}$$

If we approximate this infinite sum with, say, the first 501 terms we have,

$$Z_{\text{Planck}} \approx \sum_{i=0}^{500} e^{-\frac{i h \nu}{k T}} = 21.3406$$

We started down this road in the previous exercise where just the first two terms resulted in a sum of 1.953141. By the time we reach an energy level consisting of 500 little energy packets, the contribution to the overall sum for the partition function is down to about 10^{-11} . Thus, we have numerically confirmed, in one case, the equivalency between the two forms for the partition function,

$$Z_{\text{Planck}} = \sum_{i=0}^{\infty} e^{-\frac{i h \nu}{k T}} \equiv \frac{1}{1 - e^{-\frac{h \nu}{k T}}}$$

Exercise 27.7.13: How did Schrödinger actually go about finding the average energy for Planck's oscillator?

Solution to Exercise 27.7.13

In his worked out example on page 20 in Chapter IV for finding the average energy for Planck's oscillator, Schrödinger takes us on a circuitous route with unnecessary mathematical diversions. One of my pet peeves falls upon those who try to explain some important fundamental concept with an anticipated enlightening example, and then proceed to further obfuscate the path to enlightenment with issues that are, at best, peripheral to their main point.

Schrödinger is guilty of this sin here by getting us involved in zero point energy and manipulations of hyperbolic functions that are not germane to the larger issue of simply showing us how to quickly arrive at the partition function to be followed by the derivatives that need to be found. Neither does he inform us that he is going to switch to thinking about Planck's oscillator as a quantum mechanical oscillator.

Schrödinger informs us, out of the blue, that for Planck's oscillator, the mapping from the statements in the state space $F(X = x_i)$ are to energy levels $F(X = x_0) \rightarrow \frac{1}{2}\varepsilon$, $F(X = x_1) \rightarrow 1\frac{1}{2}\varepsilon$, \dots , $F(X = x_l) \rightarrow l + \frac{1}{2}\varepsilon$. Where did this curious half an energy bundle, the zero point energy, come from?

In any case, the partition function Z as the sum over the infinite number of statements in this state space for Planck's oscillator becomes,

$$Z = \sum_{i=1}^n e^{\lambda F(X=x_i)} \equiv \sum_{l=0}^{\infty} e^{-\mu\varepsilon(l+\frac{1}{2})} = e^{-\mu\frac{1}{2}\varepsilon} + e^{-\mu 1\frac{1}{2}\varepsilon} + e^{-\mu 2\frac{1}{2}\varepsilon} + \dots$$

Each term in the argument to the exponential function can be broken down into a multiplication,

$$e^{(-\mu\frac{1}{2}\varepsilon)+(-\mu l\varepsilon)} = e^{-\mu\frac{1}{2}\varepsilon} \times e^{-\mu l\varepsilon}$$

Extract the common factor from inside the sum that doesn't depend on l ,

$$Z = e^{-\frac{1}{2}\mu\varepsilon} \sum_{l=0}^{\infty} e^{-l\mu\varepsilon}$$

At this point, Schrödinger makes the substitution $x = \mu\varepsilon = \mu h\nu = \frac{h\nu}{kT}$, so that we can get to the sum of the infinite series which we already know,

$$\begin{aligned} Z &= e^{-\frac{1}{2}x} \sum_{l=0}^{\infty} e^{-lx} \\ \sum_{l=0}^{\infty} e^{-lx} &= \frac{1}{1 - e^{-x}} \\ Z &= \frac{e^{-\frac{1}{2}x}}{1 - e^{-x}} \end{aligned}$$

Here is where we get bogged down by fooling around with hyperbolic function manipulations. Immediately after listing the above result, Schrödinger equates it to,

$$Z = \frac{1}{2 \sinh(\frac{1}{2}x)}$$

Searching for how *Mathematica* defines the hyperbolic sine function, we find that by evaluating `TrigToExp[Sinh[x]]`,

$$\sinh(x) = \frac{e^x - e^{-x}}{2}$$

Schrödinger must have utilized some algebraic manipulations along these lines to arrive at the partition function Z ,

$$\begin{aligned} Z &= \frac{e^{-\frac{1}{2}x}}{1 - e^{-x}} \\ \frac{e^{-\frac{1}{2}x}}{1 - e^{-x}} \times \frac{e^{\frac{1}{2}x}}{e^{\frac{1}{2}x}} &= \frac{1}{e^{\frac{1}{2}x} - e^{-\frac{1}{2}x}} \\ \sinh [(1/2) x] &= \frac{e^{\frac{1}{2}x} - e^{-\frac{1}{2}x}}{2} \\ \frac{1}{\sinh [(1/2) x]} &= \frac{2}{e^{\frac{1}{2}x} - e^{-\frac{1}{2}x}} \\ Z &= \frac{1}{2} \times \frac{1}{\sinh [(1/2) x]} \\ Z &= \frac{1}{2 \sinh [(1/2) x]} \\ Z &= \frac{1}{2} \left[\text{hyperbolic cosecant} \left(\frac{x}{2} \right) \right] \end{aligned}$$

Now that we have reproduced Schrödinger's plan of attack in more detail, we can form his Ψ function in preparation for the required eventual differentiation with respect to T to find the average energy U ,

$$\Psi = k \ln Z$$

$$x = h\nu/kT$$

$$\begin{aligned} k \ln Z &= k \ln \left[\text{hyperbolic cosecant} \left(\frac{h\nu}{2kT} \right) \right] \\ \frac{\partial \Psi}{\partial T} &= \frac{h\nu \coth \left(\frac{h\nu}{2kT} \right)}{2T^2} \\ U &= T^2 \frac{\partial \Psi}{\partial T} \\ T^2 \frac{\partial \Psi}{\partial T} &= \frac{h\nu}{2} \times \coth \left(\frac{h\nu}{2kT} \right) \end{aligned}$$

Substitute the definition of the hyperbolic cotangent, multiply by 1 to convert the exponentials, and finally substitute back $x = h\nu/kT$,

$$\begin{aligned}
U &= \frac{h\nu}{2} \times \frac{e^{1/2x} + e^{-1/2x}}{e^{1/2x} - e^{-1/2x}} \\
&= \frac{h\nu}{2} \times \frac{e^{1/2x} + e^{-1/2x}}{e^{1/2x} - e^{-1/2x}} \times \frac{e^{1/2x}}{e^{1/2x}} \\
&= \frac{h\nu}{2} \times \frac{e^x + 1}{e^x - 1} \\
&= \frac{h\nu}{2} + \frac{h\nu}{e^{h\nu/kT} - 1}
\end{aligned}$$

After Schrödinger drops the zero point energy $h\nu/2$, we have the same result as found by the relative straightforward approach in section 27.4.2 without all of the gyrations of the above derivation. Throughout, I took pity on both myself and the reader by handing over the difficult partial differentiation and the resulting simplifications to *Mathematica* which returns the answer in the blink of an eye. I would have had a very difficult time as a student attending Herr Doktor Professor Schrödinger's lectures.

Exercise 27.7.14: How did Schrödinger express the standard MEP formalism of differentiating the log of the partition function with respect to the Lagrange multiplier to find the average energy?

Solution to Exercise 27.7.14

Here is the MEP notation for finding the average of the constraint function in our standard generic notation,

$$\langle F \rangle = \frac{\partial \ln Z}{\partial \lambda}$$

Schrödinger wrote the average energy as,

$$\frac{E}{N} = U = T^2 \frac{\partial \Psi}{\partial T}$$

where $\Psi = k \ln Z$.

Exercise 27.7.15: Show the transition from Schrödinger's expression to the direct expression for the average energy of the Fermi oscillator.

Solution to Exercise 27.7.15

Schrödinger states on his page 20 that the average energy for the Fermi oscillator is expressed as,

$$U = \frac{\varepsilon}{e^{\frac{\varepsilon}{kT}} + 1}$$

In section 27.4.1, the partition function for the Fermi oscillator was found as,

$$Z_{\text{Fermi}} = 1 + \exp\left(-\frac{\varepsilon}{kT}\right)$$

Differentiate Schrödinger's Ψ function with respect to T to find,

$$\begin{aligned} \frac{\partial \Psi}{\partial T} &= \frac{\partial \{ k \ln [1 + \exp(-\frac{\varepsilon}{kT})] \}}{\partial T} \\ \frac{\partial \{ k \ln [1 + \exp(-\frac{\varepsilon}{kT})] \}}{\partial T} &= \frac{\exp(-\frac{\varepsilon}{kT}) \varepsilon}{[1 + \exp(-\frac{\varepsilon}{kT})] T^2} \\ T^2 \frac{\partial \Psi}{\partial T} &= \frac{\exp(-\frac{\varepsilon}{kT}) \varepsilon}{1 + \exp(-\frac{\varepsilon}{kT})} \end{aligned}$$

This expression is still not in the form that Schrödinger presented, so there must be some more algebraic manipulation required. Split off the exponential fraction on the right hand side of the above equation and multiply by 1,

$$\frac{\exp(-\frac{\varepsilon}{kT})}{1 + \exp(-\frac{\varepsilon}{kT})} \times \frac{\exp(\frac{\varepsilon}{kT})}{\exp(\frac{\varepsilon}{kT})} = \frac{1}{\exp(\frac{\varepsilon}{kT}) + 1}$$

to see Schrödinger's expression for the average energy,

$$U = T^2 \frac{\partial \Psi}{\partial T} = \frac{\varepsilon}{e^{\frac{\varepsilon}{kT}} + 1}$$

Working from first principles, we showed in this Chapter that,

$$\begin{aligned} \langle F \rangle &= F(X = x_1) Q_1 + F(X = x_2) Q_2 \\ &= \left(0 \times \frac{1}{1 + e^{\lambda\varepsilon}}\right) + \left(\varepsilon \times \frac{e^{\lambda\varepsilon}}{1 + e^{\lambda\varepsilon}}\right) \\ &= \frac{\varepsilon \times e^{\lambda\varepsilon}}{1 + e^{\lambda\varepsilon}} \\ \frac{\varepsilon \times e^{\lambda\varepsilon}}{1 + e^{\lambda\varepsilon}} &= \frac{\varepsilon \exp(-\frac{\varepsilon}{kT})}{1 + \exp(-\frac{\varepsilon}{kT})} \\ \frac{\varepsilon \exp(-\frac{\varepsilon}{kT})}{1 + \exp(-\frac{\varepsilon}{kT})} \times \frac{\exp(\frac{\varepsilon}{kT})}{\exp(\frac{\varepsilon}{kT})} &= \frac{\varepsilon}{e^{\frac{\varepsilon}{kT}} + 1} \end{aligned}$$

This average constraint function $\langle F \rangle$ for the generic coin toss MEP assignment matches Schrödinger's average energy U for a Fermi oscillator.

Exercise 27.7.16: Show the transition from Schrödinger's expression to the direct expression for the average energy of Planck's oscillator.

Solution to Exercise 27.7.16

Since we have already delved into the mathematical details of how Schrödinger actually arrived at his final expression for the average energy of Planck's oscillator with the complication of zero point energy, we will simply write the final result,

$$U = \frac{h\nu}{\exp(\frac{h\nu}{kT}) - 1}$$

Back in Exercise 27.7.7 we found that the partition function for Planck's oscillator was,

$$Z_{\text{Planck}} = \frac{1}{1 - e^{\lambda\varepsilon}} \equiv \frac{1}{1 - \exp(-\frac{h\nu}{kT})}$$

Following Schrödinger's directive to differentiate the log of this partition function multiplied by k with respect to the temperature, and then multiply by T^2 ,

$$\begin{aligned} \frac{\partial(k \ln Z)}{\partial T} &= \frac{\partial \left[k \ln \left(\frac{1}{1 - e^{-\frac{h\nu}{kT}}} \right) \right]}{\partial T} \\ \frac{\partial \left[k \ln \left(\frac{1}{1 - e^{-\frac{h\nu}{kT}}} \right) \right]}{\partial T} &= \frac{h\nu}{(e^{\frac{h\nu}{kT}} - 1) T^2} \\ \frac{\partial(k \ln Z)}{\partial T} T^2 &= \frac{h\nu}{\exp(h\nu/kT) - 1} \end{aligned}$$

Exercise 27.7.17: Illustrate these previous exercises with an n -sided die that acts like Planck's oscillator.

Solution to Exercise 27.7.17

As a thought exercise (*Gedanken experiment*), consider an n -sided die with n large, say, $n = 1,000$. The dimension of the state space is therefore 1,000 with the state space composed of 1,000 statements. The first statement in the state space, ($X = x_1$), might be something like, "After the die is rolled, it is in state 1."

Then, of course, $P(X = x_i)$ represents the IP's degree of belief that this statement is TRUE, "After the die is rolled, it is in the i^{th} state." Accordingly, $P(X = x_i | \mathcal{M}_k)$ is an IP's degree of belief that this statement is TRUE, "After the die is rolled, it is in the i^{th} state." when conditioned on the information inserted by model \mathcal{M}_k .

For any particular inferential application, it is necessary to define a mapping, $F(X = x_i)$, from the statements in the state space to numbers. A mapping inspired

by our recent exposure to statistical mechanics might be something like,

$$F(X = x_i) \equiv E_i \equiv (i - 1) h \nu$$

interpreted as something like, “Each of the n die faces possess a measurable trait composed from some number of elementary ‘packets’ where each such packet is determined by Planck’s constant and a frequency in cycles per second.”

The first die face, $i = 1$, does not possess any packets, the second die face, $i = 2$, has only one, and the 1000^{st} die face, $i = 1,000$, has accumulated 999 of these packets which thereby define its trait. The idea is that with increasing i , the die face is in a higher energy state, and that it is physically more difficult for the die to land on this face. It is the opposite of our previous die scenario where it was easier to land in, say, the SIX state because of the physical characteristics of the die due to changes in the center of gravity and shorter length on one of the axes.

To calculate the numerical assignment to the probability for any statement in this state space, we use the MEP formula,

$$Q_i \equiv P(X = x_i | \mathcal{M}_k) = \frac{\exp [\lambda F(X = x_i)]}{\sum_{i=1}^n \exp [\lambda F(X = x_i)]}$$

But we know the partition function will be approximated by,

$$Z_{\text{Planck}} = \frac{1}{1 - e^{-\frac{h\nu}{kT}}}$$

With this much background context under our belt, we are ready for some numerical experiments. Suppose that $h\nu$ is on the order of,

$$h\nu \approx 10^{-27} \times 10^{13} = 10^{-14}$$

and kT is on the order of,

$$kT \approx 10^{-16} \times 10^3 = 10^{-13}$$

so that,

$$\begin{aligned} \frac{h\nu}{kT} &= -\frac{10^{-14}}{10^{-13}} \\ e^{-\frac{h\nu}{kT}} &= e^{-0.1} \\ Z_{\text{Planck}} &= \frac{1}{1 - e^{-0.1}} \\ &= 10.5083 \end{aligned}$$

The normalizing constant for each Q_i is then 10.5083. Let’s conduct a check with *Mathematica* by calculating the sum of all the numerators in the Q_i . Evaluating,

$$\text{Sum}[\text{Exp}[-(i-1) . 1], \{i, 1, 1000\}]$$

confirms that $Z = \sum_{i=1}^{1000} \exp [-(i-1) \frac{h\nu}{kT}] = 10.5083$.

Some selected Q_i are,

$$\begin{aligned}
 Q_1 &= \frac{\exp\left(-\frac{0 h \nu}{k T}\right)}{Z_{\text{Planck}}} \\
 &= \frac{1}{10.5083} \\
 &= 0.0952 \\
 Q_2 &= \frac{\exp\left(-\frac{1 h \nu}{k T}\right)}{Z_{\text{Planck}}} \\
 &= \frac{0.9048}{10.5083} \\
 &= 0.0861 \\
 Q_{11} &= \frac{\exp\left(-\frac{10 h \nu}{k T}\right)}{Z_{\text{Planck}}} \\
 &= \frac{0.3679}{10.5083} \\
 &= 0.0350 \\
 Q_{500} &= \frac{\exp\left(-\frac{499 h \nu}{k T}\right)}{Z_{\text{Planck}}} \\
 &= \frac{2.12 \times 10^{-22}}{10.5083} \\
 &\approx 10^{-23}
 \end{aligned}$$

We observe the expected steady downward trend in the assigned probabilities as the die face increases. By the time the die face is at 500, the probability is down to the order of 10^{-23} , and the last die face is down to the order of 10^{-45} . The first 50 die faces accumulate about 99% of all the probability with the last 950 faces sharing the rest of the remaining probability.

Using Schrödinger's language, our "system" consists of just this *one n*-sided die. Its state can only be ascertained by looking at the number of spots on the face that lands up after it has been rolled. An "assembly" of N such systems is a listing of N rolls of this die. Schrödinger would want to find the one most probable distribution of the $a_l \equiv N_i$, the frequency counts for each die face.

The correct action, as conducted above, is to make a numerical assignment Q_i as a legitimate probability under the information contained in the model with the parameter $\lambda \equiv -\frac{1}{k T}$. Then, and only then, by invoking the formal manipulation rules of probability, NOT by invoking any MEP formula, the *probability* for any count of the die faces, $P(N_1, N_2, \dots, N_{1000})$, can be calculated.

Exercise 27.7.18: Provide the details for Feller's example illustrating the distinction among the three "statistics" prominently mentioned in statistical mechanics.

Solution to Exercise 27.7.18

Feller's number of cells n is the same as our n , and in his example first presented on pg. 336, $n = 5$. For us, n is the dimension of the state space, which consists here of five statements. Feller's number of balls r is our M , the number of future counts, and in the example $M = 3$. We are back to our preferred notation for future frequency counts. Feller labels the $r = 3$ balls as **a**, **b**, and **c**.

Feller portrays the desired occupancy numbers as $(\star | - | \star | \star | -)$ to indicate one ball in cell 1, one ball in cell 3, and one ball in cell 4. There are no balls in cells 2 and 5. In our notation, $M_1 = 1, M_2 = 0, M_3 = 1, M_4 = 1, M_5 = 0$ with $\sum_{i=1}^5 M_i = M = 3$.

The sample space consists of $n^r \equiv n^M = 5^3 = 125$ elementary points. Feller makes the claim that the probability for the occupancy numbers is $6/125$ under Maxwell–Boltzmann statistics, $1/35$ under Bose–Einstein statistics, and $1/10$ under Fermi–Dirac statistics.

Under Maxwell–Boltzmann, Feller says that all of the elementary points have an equal probability of $1/125$. A compound event must include the multiplicity factor for this occupancy pattern $(\star | - | \star | \star | -)$. If ball **a** is in cell 1, ball **b** in cell 3, and ball **c** in cell 4, this is the lowest level description that we can make for the sample space. It is one elementary point from the total of 125 elementary points in the sample space, specifically $(\mathbf{a} | - | \mathbf{b} | \mathbf{c} | -)$.

However, ball **b** could be in cell 1, ball **a** in cell 3, and ball **c** in cell 4 and still qualify for the compound event $(\star | - | \star | \star | -)$ as $(\mathbf{b} | - | \mathbf{a} | \mathbf{c} | -)$. This represents another elementary point in the sample space. There are a total of six different possibilities that satisfy the compound event, and adding up the equal probability for each of these six elementary points results in the aforementioned probability of $6/125$.

In our formulation, on the other hand, if the fair model is the only operative model, then the probability for one simple event, followed by another simple event, followed by a third simple event is equal to $(1/n) \times (1/n) \times (1/n) = n^{-M}$. Count up the number of different ways this could happen by computing the multiplicity factor,

$$W(M) = \frac{M!}{M_1! \cdots M_5!}$$

The above probability for a simple event can then be summed to form the compound event. The probability for the future frequency counts, when there is absolutely no doubt about the fair model being the true model, is,

$$P(M_1 = 1, M_2 = 0, M_3 = 1, M_4 = 1, M_5 = 0 | \mathcal{M}_k) = \frac{W(M)}{n^M} = \frac{6}{125}$$

Under Bose–Einstein statistics, Feller provides the formula,

$$\begin{aligned} A_{r,n} &= \binom{n+r-1}{r} \\ &= \frac{(n+r-1)!}{(n-1)! r!} \\ &= \frac{(5+3-1)!}{4! 3!} \\ &= 35 \end{aligned}$$

Feller says that the probability of the above occupancy pattern under Bose–Einstein statistics is,

$$\frac{1}{A_{r,n}} = \frac{1}{35}$$

In our formulation, when there is complete uncertainty about which model is true, in other words, when the IP is in the polar opposite state of knowledge about the models compared to the Maxwell–Boltzmann case, the probability for the future frequency counts is,

$$P(M_1 = 1, M_2 = 0, M_3 = 1, M_4 = 1, M_5 = 0) = \frac{(n-1)! M!}{(M+n-1)!} = \frac{1}{35}$$

Under Fermi–Dirac statistics, Feller provides the formula,

$$\frac{1}{\binom{n}{r}} = \frac{1}{\binom{5}{3}} = \frac{1}{10}$$

In our formulation, when there is some uncertainty about which models are true, and this uncertainty takes the specific form discussed in Volume I under the topic of the Jeffreys and Haldane priors for models, the probability for the future frequency counts depends on the α_i parameters in the Dirichlet distribution. When the five α_i parameters for the probability of the models are specified as,

$$\alpha_1 = 1, \alpha_2 \rightarrow 0, \alpha_3 = 1, \alpha_4 = 1, \alpha_5 \rightarrow 0$$

then the probability for the future frequency counts becomes,

$$P(M_1 = 1, M_2 = 0, M_3 = 1, M_4 = 1, M_5 = 0) = \frac{1}{10}$$

The conclusion drawn is rather startling. From our perspective, rather than resorting to some physical explanation involving “distinguishability of particles,” these various statistics arise from the IP’s state of knowledge about the model space.

Exercise 27.7.19: Use Feller's example to review the counting formulas that aid in the analysis of the sample space.

Solution to Exercise 27.7.19

In Feller's notation there are $n = 5$ cells and $r = 3$ balls. Suppose we borrow from the theme of this Chapter, and say that the $n = 5$ statements are about an electron possessing an energy level. We have $M = 3$ electrons corresponding to the $r = 3$ balls which can be placed into an energy level. The state space has dimension $n = 5$ and consists of the five statements, "An electron is in the lowest energy state." through "An electron is in the highest energy state."

The sample space consists of $n^M = 5^3 = 125$ elementary points. One of these 125 elementary points might be the statement, "Electron **b** is in the lowest energy state, electron **c** is in the third energy state, and electron **a** is in the fourth energy state." We are still using Feller's labeling of the three balls (electrons) by **a**, **b**, and **c**.

The IP would ultimately like to solve an inferential problem expressing a state of knowledge about the future frequency counts of electrons in these energy states. Three different answers were provided as worked through in the last exercise. The IP arrived at different states of knowledge about the same frequency counts when predicated on differing states of knowledge about "causal factors." This knowledge about the "causal factors" is captured by the probability for all the models in model space.

Here we just want to show the breakdown, or looking it the other way, the aggregation of elementary points to form the entire sample space. But this modest goal is very enlightening because it shows where the numbers in the probabilities under the three different statistics arise. Since $M = 3$, the sum of the future frequency counts $M = \sum_{i=1}^5 M_i$ can be broken down into three classes,

Class 1. $3 + 0 + 0 + 0 + 0 = 3$

Class 2. $2 + 1 + 0 + 0 + 0 = 3$

Class 3. $1 + 1 + 1 + 0 + 0 = 3$

One of our counting formulas tells us how many contingency tables there are in each of the above categories. For example, we know that there must be five possibilities for the first sum. The formula confirms this,

$$\text{Number of contingency tables} = \frac{n!}{r_1! \cdots r_M!} = \frac{5!}{4! 0! 0! 1!} = 5$$

Each one of these contingency tables can be formed in a number of ways by taking the individuality of the repetitions into account. Here, we assume that the electrons

are distinguishable, that is why we gave them the distinctive labels **a**, **b**, and **c**. The multiplicity factor $W(M)$ is used to count up the number of ways each contingency table could be formed when we take distinguishability of electrons into account. There is obviously only one way to pack all three electrons into a single energy level,⁵

$$W(M) = \frac{M!}{M_1! M_2! \cdots M_n!} = \frac{3!}{3! 0! 0! 0!} = 1$$

The total number of elementary points from the total of 125 accumulated so far is then $5 \times 1 = 5$. There are 120 elementary points remaining.

The breakdown continues on in the same vein by examining the next category where the sum of the frequency counts is reflected in contingency tables with two electrons in one energy state and the third in another. The counting formula tells us how many contingency tables there are of the second category. The formula returns 20 contingency tables,

$$\text{Number of contingency tables} = \frac{n!}{r_z! \cdots r_M!} = \frac{5!}{3! 1! 1! 0!} = 20$$

The multiplicity factor is,

$$W(M) = \frac{M!}{M_1! M_2! \cdots M_n!} = \frac{3!}{2! 1! 0! 0!} = 3$$

The total number of elementary points from the total of 125 accumulated so far is then $(5 \times 1) + (20 \times 3) = 65$. There are 60 elementary points remaining.

The final category consists of a sum of the frequency counts where each frequency count is one electron in each of three different energy levels.⁶ The counting formula tells us how many contingency tables there are in this final category. The formula returns 10 contingency tables,

$$\text{Number of contingency tables} = \frac{n!}{r_z! \cdots r_M!} = \frac{5!}{2! 3! 0! 0!} = 10$$

The multiplicity factor is,

$$W(M) = \frac{M!}{M_1! M_2! \cdots M_n!} = \frac{3!}{1! 1! 1! 0!} = 6$$

We have now accounted for all of the elementary points with,

$$n^M = (5 \times 1) + (20 \times 3) + (10 \times 6) = 125$$

By working through this counting exercise we are able to discern the origin of the numbers that appear in the probabilities for the requested pattern of future frequency counts. These future frequency counts in the example are,

$$M_1 = 1, M_2 = 0, M_3 = 1, M_4 = 1, M_5 = 0$$

⁵This situation is not allowed by physics because of the Pauli exclusion rule.

⁶The only situation allowed under Pauli's rule.

The number of elementary points in the sample space is 125. The multiplicity factor for these particular future frequency counts is 6. The total number of distinct contingency tables, or future frequency counts is $5 + 20 + 10 = 35$. The total number of future frequency counts satisfying the requirement that a maximum of one electron could be in any energy level is 10.

All elementary points have the same probability of $n^{-r} = \frac{1}{125}$ under Maxwell–Boltzmann. Six elementary points make up the compound event of $\boxed{1\ 0\ 1\ 1\ 0}$ so the probability for the compound event is $6/125$. All 35 contingency tables, such as $\boxed{0\ 0\ 3\ 0\ 0}$ and also $\boxed{1\ 0\ 1\ 1\ 0}$, have the same probability of $1/35$ under the Bose–Einstein statistics. There are only 10 future frequency counts, or contingency tables, with a single count at three different energy levels. One example is $\boxed{1\ 0\ 1\ 1\ 0}$. All ten have the same probability of $1/10$ under Fermi–Dirac.

But the truly remarkable fact is that we can reproduce these probabilities by looking at the probability distribution for the models, $P(\mathcal{M}_k)$. This assignment to model space must be done in any case because the formal rules demand it.

We explored this fundamental requirement, demanded by the formal manipulation rules, from a computational perspective in Chapter Fifteen, Volume I, in the context of the kangaroo scenario. The Dirichlet distribution captures the IP’s state of knowledge about any specific model being true. By adjusting the α_i parameters in the Dirichlet distribution, Dirac δ distributions, uniform distributions, Jeffreys priors, and any other state of knowledge can be implemented.

The probability under Maxwell–Boltzmann occurred when all the $\alpha_i \rightarrow \infty$ and each of the 125 elementary points in sample space had equal probability under one model. The probability under Bose–Einstein occurred when all the $\alpha_i = 1$, and every model had the same probability. The probability under Fermi–Dirac must then occur when some $\alpha_i \rightarrow 0$, and the remaining $\alpha_i = 1$.

Since the probability for the stated future frequency counts is even greater under the Fermi–Dirac statistics, there must be a setting for the α_i parameters in the probability for the models that reproduces such a probability. Certain cells with zero frequency counts are favored and the remaining cells evenly divide the total frequency counts. So by setting $\alpha_1 = \alpha_3 = \alpha_4 = 1$ and letting $\alpha_2 \rightarrow 0$ and $\alpha_5 \rightarrow 0$, the probability under Fermi–Dirac statistics can be reproduced as well.

Exercise 27.7.20: Fill in the missing steps of Jaynes’s derivation showing the relationship between a maximum likelihood estimator and the MEP formalism.

Solution to Exercise 27.7.20

In the **Connections to the Literature** section, I felt that it was time to examine Jaynes’s explanation for the traditional maximum likelihood method as viewed from the perspective of the MEP. During an inference, an IP always arrives at some point

where it wants to compare two models after some data have been observed. Thus, the log-likelihood ratio arises quite naturally even for a Bayesian,

$$\ln \left[\frac{P(\mathcal{D} | \mathcal{M}_A)}{P(\mathcal{D} | \mathcal{M}_B)} \right] = \ln [P(\mathcal{D} | \mathcal{M}_A)] - \ln [P(\mathcal{D} | \mathcal{M}_B)]$$

$$\ln [P(\mathcal{D} | \mathcal{M}_A)] = \ln [W(N) Q_{1A}^{N_1} \cdots Q_{nA}^{N_n}]$$

$$\ln [P(\mathcal{D} | \mathcal{M}_B)] = \ln [W(N) Q_{1B}^{N_1} \cdots Q_{nB}^{N_n}]$$

The multiplicity factor $W(N)$ cancels out because it is the same under the two models for the same set of data.

$$\begin{aligned} \ln [P(\mathcal{D} | \mathcal{M}_A)] - \ln [P(\mathcal{D} | \mathcal{M}_B)] &= \ln [Q_{1A}^{N_1} \cdots Q_{nA}^{N_n}] - \ln [Q_{1B}^{N_1} \cdots Q_{nB}^{N_n}] \\ &= (N_1 \ln Q_{1A} + \cdots + N_n \ln Q_{nA}) - \\ &\quad (N_1 \ln Q_{1B} + \cdots + N_n \ln Q_{nB}) \\ &= \sum_{i=1}^n N_i \ln Q_{iA} - \sum_{i=1}^n N_i \ln Q_{iB} \\ &= \sum_{i=1}^n N_i [\ln Q_{iA} - \ln Q_{iB}] \\ \ln \left[\frac{P(\mathcal{D} | \mathcal{M}_A)}{P(\mathcal{D} | \mathcal{M}_B)} \right] &= \sum_{i=1}^n N_i \ln \left[\frac{Q_{iA}}{Q_{iB}} \right] \end{aligned}$$

But with the development of the MEP in this Volume, we now have an expression for the numerical assignments Q_{iA} and Q_{iB} under the information provided by models \mathcal{M}_A and \mathcal{M}_B .

$$\ln Q_{iA} = \sum_{j=1}^m \lambda_j^A F_j(X = x_i) - \ln Z_A$$

$$\ln Q_{iB} = \sum_{j=1}^m \lambda_j^B F_j(X = x_i) - \ln Z_B$$

With the MEP at our disposal, we are able to explicitly see the role that the data, that is N , the N_i , and the observed sample averages of the constraint functions, $\bar{F}(X = x_i)$, play in re-orienting the model space. Starting with,

$$\ln \left[\frac{P(\mathcal{D} | \mathcal{M}_A)}{P(\mathcal{D} | \mathcal{M}_B)} \right] = \sum_{i=1}^n N_i \ln \left[\frac{Q_{iA}}{Q_{iB}} \right]$$

$$\begin{aligned}
\ln \left[\frac{P(\mathcal{D} | \mathcal{M}_A)}{P(\mathcal{D} | \mathcal{M}_B)} \right] &= \sum_{i=1}^n N_i \left[\left(\sum_{j=1}^m \lambda_j^A F_j(X = x_i) - \ln Z_A \right) - \left(\sum_{j=1}^m \lambda_j^B F_j(X = x_i) - \ln Z_B \right) \right] \\
&= \left(\sum_{j=1}^m \lambda_j^A \sum_{i=1}^n N_i F_j(x_i) - \sum_{i=1}^n N_i \ln Z_A \right) - \\
&\quad \left(\sum_{j=1}^m \lambda_j^B \sum_{i=1}^n N_i F_j(x_i) - \sum_{i=1}^n N_i \ln Z_B \right) \\
&= \left(\sum_{j=1}^m \lambda_j^A \sum_{i=1}^n N_i F_j(x_i) - N \ln Z_A \right) - \left(\sum_{j=1}^m \lambda_j^B \sum_{i=1}^n N_i F_j(x_i) - N \ln Z_B \right) \\
&= N \left[\sum_{j=1}^m \lambda_j^A \sum_{i=1}^n \frac{N_i}{N} F_j(x_i) - \ln Z_A - \sum_{j=1}^m \lambda_j^B \sum_{i=1}^n \frac{N_i}{N} F_j(x_i) - \ln Z_B \right] \\
&= N \left[\left(\sum_{j=1}^m \lambda_j^A \bar{F}_j(x_i) - \ln Z_A \right) - \left(\sum_{j=1}^m \lambda_j^B \bar{F}_j(x_i) - \ln Z_B \right) \right] \\
&= N \left[\sum_{j=1}^m (\lambda_j^A - \lambda_j^B) \bar{F}_j(x_i) + \ln \left(\frac{Z_B}{Z_A} \right) \right]
\end{aligned}$$

Our final result,

$$\ln \left[\frac{P(\mathcal{D} | \mathcal{M}_A)}{P(\mathcal{D} | \mathcal{M}_B)} \right] = N \left[\sum_{j=1}^m (\lambda_j^A - \lambda_j^B) \bar{F}_j(x_i) + \ln \left(\frac{Z_B}{Z_A} \right) \right]$$

is the same as Jaynes's log-likelihood ratio (see section 29.7.1 on pg. 418 as well),

$$\log \frac{P(D | H)}{P(D | H_0)} = r \left[\log(Z_0/Z) + \sum_{k=1}^m (\lambda_k^{(0)} - \lambda_k) \bar{f}_k \right]$$

remembering that Jaynes's λ_k are the negative of our λ_j . After a numerical example, we will comment on Jaynes's final important observation concerning this derivation.

This rationale given by Jaynes forms the basis of my so-called “reconciliation of the data and information” in the last Chapter. It is one easy way of accepting the fact that if you insert the maximum likelihood data as the actual information under some model right at the outset, you would end up supporting that same model anyway after the data have been processed.

Exercise 27.7.21: Use the example of 200 gas molecules to illustrate the previous exercise.

Solution to Exercise 27.7.21

Under a model \mathcal{M}_A , suppose we take as the “information” the sample average energy of $\bar{F}(X = x_i) = U = 2.5$, and make it equal to the constraint function average $\langle E \rangle$. The parameter dual to the constraint function average is the Lagrange multiplier, and this is equal to $\lambda^A = -\frac{1}{T} = -\frac{1}{5.04} = -0.1985$. The partition function under model \mathcal{M}_A is $Z_A = 4.4198$. (See Table 27.1.)

Suppose that we take as a competing model, a model \mathcal{M}_B , that has a higher temperature of $T = 10$. The comparable value for the Lagrange multiplier is $\lambda^B = -\frac{1}{T} = -0.10$, and for the partition function $Z_B = 5.79$. The data are the same under any model, thus the sample average must also remain the same at 2.5 for the two models under consideration. The question that Jaynes asked is: How much is model \mathcal{M}_A preferred over model \mathcal{M}_B ?

From the results of the last exercise we have that,

$$\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} = \exp \left\{ N \times \left[\sum_{j=1}^m (\lambda_j^A - \lambda_j^B) \bar{F}_j(x_i) + \ln \left(\frac{Z_B}{Z_A} \right) \right] \right\}$$

We have only one constraint function and one Lagrange multiplier in each model so this simplifies to,

$$\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} = \exp \left\{ N \times \left[(\lambda^A - \lambda^B) \bar{F}(x_i) + \ln \left(\frac{Z_B}{Z_A} \right) \right] \right\}$$

Substituting in the correct values,

$$\begin{aligned} \frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} &= \exp \left\{ N \times \left[(\lambda^A - \lambda^B) \bar{F}(x_i) + \ln \left(\frac{Z_B}{Z_A} \right) \right] \right\} \\ &= \exp \left\{ 200 \times \left[[(-0.1985) - (-0.10)) \times 2.5] + \ln \left(\frac{5.79}{4.42} \right) \right] \right\} \\ &\approx 100 \end{aligned}$$

Thus, the model that inserts information that matches the observed data is preferred with a weight of about 100 times over another model with differing information. This is as it should be, and we are always happy when a Bayesian prescription accords with our intuition.

Furthermore, it is clear that as the information under any model \mathcal{M}_k approaches the information under model \mathcal{M}_A , both the Lagrange multipliers λ_j^k and λ_j^A , and the partition functions Z_k and Z_A will approach each other as well. The argument to the exponential will then approach 0, and the ratio of the probabilities for the two models, when conditioned on the data, will approach 1. This is all as it should be.

Exercise 27.7.22: Confirm the approximate relative weighting for the two models in the last exercise.

Solution to Exercise 27.7.22

From first principles, it must be that,

$$P(\mathcal{D} | \mathcal{M}_A) = W(N) \prod_{i=1}^n Q_{iA}^{N_i}$$

Substituting the Q_i assignments under model \mathcal{M}_A together with the known data results in,

$$\begin{aligned} P(\mathcal{D} | \mathcal{M}_A) &= W(N) \prod_{i=1}^8 Q_{iA}^{N_i} \\ &= W(N) \times (0.2263)^{45} \times (0.1855)^{37} \times \cdots \times (0.0564)^{11} \\ &= 8.32 \times 10^{-8} \end{aligned}$$

Considering that the data, the N_i , are the same for model \mathcal{M}_B , we find that,

$$\begin{aligned} P(\mathcal{D} | \mathcal{M}_B) &= W(N) \prod_{i=1}^8 Q_{iB}^{N_i} \\ &= W(N) \times (0.1728)^{45} \times (0.1564)^{37} \times \cdots \times (0.0858)^{11} \\ &= 7.76 \times 10^{-10} \end{aligned}$$

Thus,

$$\begin{aligned} \frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} &= \frac{P(\mathcal{D} | \mathcal{M}_A) P(\mathcal{M}_A)}{P(\mathcal{D} | \mathcal{M}_B) P(\mathcal{M}_B)} \\ &= \frac{8.32 \times 10^{-8}}{7.76 \times 10^{-10}} \times 1 \\ &\approx 100 \end{aligned}$$

Of course, the multiplicity factor $W(N)$ cancels in the above ratio so it wouldn't be necessary to explicitly calculate it.

Exercise 27.7.23: What was Jaynes's point?

Solution to Exercise 27.7.23

Using the MEP in statistical inference does not necessarily conflict with established orthodox principles. Substituting the known sample averages from the data as the

information under a model will lead to that model having the maximum likelihood when compared to any other model. Refer back to the numerical example for logistic regression in Chapter Twenty Three for a confirmation of this fact.

Bayesian principles studied in Volume I as a subset of the formal manipulation rules already told us that the data would re-orient the original model space from one of equality for all models prior to any data into a relative weighting showing the importance for any model after the data. This relative weighting for the models was then used to weight the probability for any new observation conditioned on any model.

By exploiting the MEP formula for assigning numerical values to probabilities, we have available to us an explicit quantitative proof that a model based on the data must be weighted more than any other model. The use of “maximum likelihood estimators” in orthodox statistics achieves the same result that these MEP and Bayesian principles accomplish, but with far less clarity and adherence to the fundamental principles.

Although, I have to admit that, under this “data reconciliation” policy, there is something a bit unsettling about focusing on the maximum likelihood MEP model *after* the data. All MEP assignments to joint probabilities must be made *prior* to the data. What has to give in the conceptual underpinning is, shall we say, a certain “flexibility” to look at the maximum likelihood MEP model in a special light after the data.

Chapter 28

Fisher Information and Relative Entropy

28.1 Introduction

I would be remiss in my introduction to the concepts of *information* and *entropy* were I not to make mention of two popular notions. They are often spoken of in the same breath with the MEP. These two ideas are **Fisher Information** and **Relative Entropy**.

In fact, these two concepts do exhibit an interesting relationship with the MEP as it has been presented so far here in Volume II. Moreover, it is vital to disentangle the confusion that swarms over these co-mingled ideas like bees over a disturbed hive.

Unfortunately, there is no way I can do justice to these concepts in the space of one Chapter. That job has been relegated to Volume III where I go into much more detail about the role of these concepts as they exist under the overarching framework of *Information Geometry*.

It seems that historical precedence must be acceded to Fisher who, in the 1920s, started using the word *information* in a technical sense in problems of statistical inference. Later in 1945, the greater mathematical significance of Fisher information was revealed by Fisher's acolyte, the brilliant Indian statistician C. R. Rao. These ideas were crystallized within the famous Cramér–Rao theorem which, still today, occupies a central place in almost all expositions of orthodox statistical inference.

Rao saw that Fisher information could be thought of as a Riemannian metric for a differentiable manifold whose set of points consisted of probability distributions. This kind of language sparked an extremely fruitful alternative approach to probability distributions based purely on geometric considerations.

Leaning heavily on the highly developed mathematics of differential geometry, the subsequent progress of what ultimately came to be called *information geometry* provided a more appealing, and, to some, a more rigorous justification for things that we have taken for granted all along within the MEP approach.

However, it can not be emphasized too strongly that both Fisher's and Rao's motivation for invoking this new idea of information was really quite different than what drove Shannon and Jaynes. Fisher information was intimately bound up with the traditional concerns of Fisherian inference; that is, very broadly speaking, things like sufficient statistics, the minimum variance of unbiased estimators, the merit in maximum likelihood estimators, and so on.

Another distinct meaning for *information* was sprung upon the world by Solomon Kullback in the 1950s. There really is a phenomenon that is aptly captured by the word *Zeitgeist*. This "spirit of the times" shakes off the old lexicon, and drives people who resonate with it to novel ways of reconceptualizing the world. This intellectual ferment seems, curiously, to coalesce in many like minds at roughly the same historical time period.

Outside the cloistered world of Fisherian statistics, revolutionary events precipitated by World War II, Turing, Weiner, cybernetics, the dawn of the computer age, von Neumann's entropy in quantum mechanics, Brillouin's combining of science and information, together with myriad reinforcing influences had all burst upon the scene. Kullback himself had spent the war years in breaking Japanese code, while at the same time, Turing was working on Enigma at Bletchley Park. Quietly, Bayes's Theorem was enjoying a behind the scenes resurgence, even while the Fisherian world view seemed to brook no challenge in occupying center stage.

Since this ferment was contemporaneous with Shannon and Jaynes, there was a natural confusion surrounding this exciting new word *information*. As a matter of fact, all pundits agreed that we had officially entered **The Information Age**. Thus, the word *information*, and to a lesser extent *entropy*, became entangled with so many competing motivations that, even today, we are still trying to sort out the ensuing mess.

Of course, as everyone knows, Fisher and Rao would never have justified their idea of information by mentioning Bayes's Theorem. This was a taboo idea. They wouldn't have touched Bayes's Theorem with a ten foot pole.

Surprisingly, however, Kullback *did* draw inspiration from Bayes's Theorem and based his motivation for information upon it. Shannon's approach to information and his ultimate definition was, on the other hand, more of an axiomatic one; stipulating reasonable primitive assertions about what might be desired, but still not calling upon Bayes's Theorem for its justification.

This Chapter tries to sort out some of the attendant confusion surrounding all these jumbled up perceptions of information. Thus, we are attempting to clarify the information formulas associated with the names of Shannon, Jaynes, Kullback, and Fisher. As a first step, we are content to illustrate some of the technical interrelationships that exist among these different usages of the word *information*.

28.2 Fisher Information

There will be no attempt here to explain in any detail the whys and wherefores of the formulas that are presented. That will have to wait until Volume III. So expect the following to be all rather mysterious.

In differential geometry, the primitive notion is of a manifold consisting of a set of points endowed with a coordinate system. For information geometry, this set of points becomes a family of probability distributions. A metric tensor is needed for the manifold in order to define distances and angles between the points that live in the manifold.

This metric tensor is given the notation $g_{rc}(p)$. The double subscripts are a clue that the metric tensor is a matrix with the subscripts indicating a generic r^{th} row and c^{th} column entry of the matrix. The subscripts r and c are needed to avoid a clash with already established notation since typically used indices like i, j, k, m , and n have already been allocated for other purposes.

The argument to the metric tensor is given the notation p which can variously mean a point, a probability distribution, or the parameters in a model. Remember that the model is synonymous with some definite specification for the Lagrange multipliers, the λ_j parameters. We say that there are m parameters in any given model with a maximum of $m = n - 1$ parameters. Each different point, each different probability distribution, each different model operationalized by Lagrange multipliers would then have a different metric tensor associated with it.

From a purely geometric standpoint, several equivalent descriptions and notations can be given for the metric tensor. At its most abstract level, the metric tensor is an inner product over some abstract mathematical space. As Fisher surmised, for applications to probability a definition of an inner product like the following is apropos,

$$g_{rc}(p) = E \left[\frac{\partial \ln P(X = x_i | \mathcal{M}_k)}{\partial \lambda_r} \times \frac{\partial \ln P(X = x_i | \mathcal{M}_k)}{\partial \lambda_c} \right] \quad (28.1)$$

where $E [\dots]$ is the notation for the probability expectation operator. Equation (28.1) is, in fact, the definition of the **Fisher information matrix**.

Here is its relationship to the MEP formula. We know that the MEP gives us the numerical assignment for the probability of the i^{th} statement in the state space as conditioned on the information in model \mathcal{M}_k ,

$$P(X = x_i | \mathcal{M}_k) = \frac{\exp [\sum_{j=1}^m \lambda_j F_j(x_i)]}{Z(\lambda_1, \lambda_2, \dots, \lambda_m)}$$

Thus,

$$\ln [P(X = x_i | \mathcal{M}_k)] = \sum_{j=1}^m \lambda_j F_j(x_i) - \ln Z$$

The first term under the Expectation operator $E [\cdots]$ in Equation (28.1) is then,

$$\begin{aligned}\frac{\partial \ln P(X = x_i | \mathcal{M}_k)}{\partial \lambda_r} &= F_r(x_i) - \frac{\partial \ln Z}{\partial \lambda_r} \\ \frac{\partial \ln Z}{\partial \lambda_r} &= \langle F_r \rangle \\ \frac{\partial \ln P(X = x_i | \mathcal{M}_k)}{\partial \lambda_r} &= F_r(x_i) - \langle F_r \rangle\end{aligned}$$

The second term under the Expectation operator in Equation (28.1) is likewise,

$$\frac{\partial \ln P(X = x_i | \mathcal{M}_k)}{\partial \lambda_c} = F_c(x_i) - \langle F_c \rangle$$

Fisher's information matrix defined at point p then becomes,

$$g_{rc}(p) = E [(F_r(x_i) - \langle F_r \rangle) \times (F_c(x_i) - \langle F_c \rangle)] \quad (28.2)$$

But this is just the definition of the variance–covariance matrix for the constraint functions when we recall the standard statistical definitions for a variance,

$$\text{var}(X) = E [(X - \mu_X)^2]$$

and a covariance,

$$\text{covar}(X, Y) = E [(X - \mu_X) \times (Y - \mu_Y)]$$

One quick conclusion we might draw from this is that the MEP declares that the constraint functions and their *averages* constitutes the information, whereas Fisher declares that, instead, it is the variances and covariances that make up the information. It appears at first glance that Fisher demands “more” information requirements because for the $m \times m$ matrix associated with a model consisting of m parameters, there are $(m \times (m + 1))/2$ independent entries.

The covariance between constraint functions $F_r(x_i)$ and $F_c(x_i)$ is the same as the covariance between constraint functions $F_c(x_i)$ and $F_r(x_i)$. Thus, the information matrix is symmetric, and the number of independent entries is not m^2 , but rather the sum of the diagonal entries and one side of the off-diagonal entries. Thus, there are m variances and $(m \times (m - 1))/2$ covariances.

In one way of looking at it, Fisher information is really specifying information that a more complicated MEP model might also specify. The MEP model would first specify the straightforward averages of the constraint functions, followed by averages of further constraint functions that are in effect a multiplication of all combinations of the first set when adjusted by their averages.

28.2.1 Numerical example of a Fisher information matrix

We are in dire need of a simple numerical example to nail down all of the foregoing. Return to the initial kangaroo scenario in Chapter Twenty One which should serve our needs quite nicely.

This inferential problem was concerned only with a kangaroo's beer and hand preference, so we had a small state space with dimension $n = 4$. Suppose we examine the model with $m = 2$ parameters as our example.

This was the independence model where the two marginal probabilities were the only information inserted into the model making the numerical assignment for the probabilities of the four joint statements in the state space. This information in the form of marginal probabilities for beer and hand preference was $\langle F_1 \rangle = 3/4$ and $\langle F_2 \rangle = 3/4$.

The full Fisher information matrix will then consist of $m \times m = 4$ entries. But one entry is redundant and only three are needed. There will be $m = 2$ variances and just $(m(m - 1))/2 = 1$ covariance. Generically, the Fisher information matrix is expanded into the form that looks like,

$$g_{rc}(p) = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}$$

Examine the r and c subscripts to identify the appropriate row and column placements in the matrix. g_{11} goes into the first row and first column, g_{12} goes into the first row and second column, g_{21} goes into the second row and first column, and, finally, g_{22} goes into the second row and second column. Since the matrix is symmetric, $g_{12} = g_{21}$.

The diagonal entries g_{11} and g_{22} are the variances of the two constraint functions, while the off-diagonal entries, g_{12} and g_{21} , represent the same covariance between the first and second constraint functions.

The argument p is the numerical assignment to the probability distribution under the independence model. This assignment was,

$$p \equiv Q_i \equiv P(X = x_i | \mathcal{M}_k) = (9/16, 3/16, 3/16, 1/16)$$

Thus, we are going to find the metric tensor, or the Fisher information matrix, for this one specific distribution. Changing the model results in a different numerical assignment, and a different Fisher information matrix.

Let's explicitly examine how to use Equation (28.2) to calculate the entry g_{21} in the second row and first column of the Fisher information matrix. This would be the covariance between the second and first constraint function.

$$\begin{aligned}
g_{21}(p) &= E [(F_2(x_i) - \langle F_2 \rangle) \times (F_1(x_i) - \langle F_1 \rangle)] \\
&= E [(F_2(x_i) - 3/4) \times (F_1(x_i) - 3/4)] \\
&= \sum_{i=1}^{n=4} \{ [(F_2(x_i) - 3/4) \times (F_1(x_i) - 3/4)] \times P(X = x_i | \mathcal{M}_k) \} \\
&= \sum_{i=1}^{n=4} \{ [(F_2(x_i) - 3/4) \times (F_1(x_i) - 3/4)] \times Q_i \}
\end{aligned}$$

We can also fill in the known values for these two constraint functions as,

$$F_1(x_i) = (1, 1, 0, 0) \text{ and } F_2(x_i) = (1, 0, 1, 0)$$

We are in possession of everything we need for the computation. In the end, it turns out that $g_{21} = 0$. The constraint functions are independent, a not surprising conclusion.

Table 28.1 below shows all of the details.

Table 28.1: *The computational details for g_{21} , the covariance between the second and first constraint functions under the independence model, for the simplified kangaroo scenario of Chapter Twenty One.*

$F_1(x_i) - \langle F_1 \rangle$	$F_2(x_i) - \langle F_2 \rangle$	$Col 2 \times Col 1$	Q_i	$Col 3 \times Col 4$
1 - 3/4	1 - 3/4	1/16	9/16	9/256
1 - 3/4	0 - 3/4	-3/16	3/16	-9/256
0 - 3/4	1 - 3/4	-3/16	3/16	-9/256
0 - 3/4	0 - 3/4	9/16	1/16	9/256
<i>Sums</i>			1.00	0

The variances for the two constraint functions, g_{11} and g_{22} , are calculated in exactly the same way as shown in the Exercises. We can fill out the Fisher information matrix for the probability distribution under this model as,

$$g_{rc}(p) = \begin{pmatrix} 3/16 & 0 \\ 0 & 3/16 \end{pmatrix}$$

28.2.2 What is it used for?

Now that we have calculated the Fisher information matrix, what can it be used for? We cannot pursue the same line of thinking that Fisher and Rao took because that would lead us into the same abyss that statistical inference found itself in for half a Century. Instead, consider the more abstract geometrical implications.

Since, by definition, it is the metric tensor for a Riemannian manifold, it can be used to find distances between points that live in the manifold. Or, in other words, it can be used to find the “distances” between probability distributions.

We motivate this novel concept by considering the less abstract notion of distance that we are all familiar with. In the Euclidean plane, the distance between two points is given by the formula,

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (28.3)$$

In the Riemannian manifold with a metric tensor, despite the initial discomfort over the symbols, the distance between two points is defined quite similarly,

$$d = \sqrt{(\lambda^p - \lambda^q) \cdot G \cdot (\lambda^p - \lambda^q)^T} \quad (28.4)$$

$(\lambda^p - \lambda^q)$ is a row vector, so $(\lambda^p - \lambda^q)^T$, where the T indicates the matrix operation **transpose**, must be a column vector. The “central dot” notation indicates either a vector–matrix, or a vector–vector multiplication. G is the matrix $g_{rc}(p)$.

Here is an easy example illustrating how the information geometry formula for distance is just a generalization of our familiar distance in the flat plane. (x_1, y_1) are the coordinates for the first point p , while (x_2, y_2) are the coordinates for the second point q .

These coordinates are defined in relationship to their basis vectors. But these are simply the orthogonal x and y axes that we don’t take any special notice of since they are always implicitly in the Euclidean background.

In the Riemannian manifold, the points p and q are two different *probability distributions*. The coordinates for p are the Lagrange multipliers λ_1^p and λ_2^p . Take the second probability distribution, q , to be close by with coordinates displaced by a small amount Δ .

In contrast to the Euclidean case, the Riemannian manifold must take explicit account of the nature of the basis vectors. They may not happen to be orthonormal vectors which is where the metric tensor $G(p)$ plays its role.

Even though the tedium here is rising rapidly, paying close attention to the actual details in Equation (28.4) can pay dividends. We have the row vector in the first term $(\lambda^p - \lambda^q)$ multiplying a matrix G which results in a row vector. This row vector then multiplies the column vector $(\lambda^p - \lambda^q)^T$ resulting in a scalar. Fortunately, we can take the square root of a scalar.

Equation (28.3), the familiar analytical geometry formula for the distance between two points, can be re-cast into the same format as Equation (28.4). Here is an easy numerical example.

Let the coordinates for point p , (x_1, y_1) , be $(2, 3)$ in the Euclidean plane. Point q with coordinates (x_2, y_2) is at position $(5, 7)$. Then,

$$\begin{aligned} d &= \sqrt{\left(\begin{pmatrix} 5-2 & 7-3 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 5-2 \\ 7-3 \end{pmatrix} \right)} \\ &= \sqrt{\left(\begin{pmatrix} 3 & 4 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 4 \end{pmatrix} \right)} \\ &= 5 \end{aligned}$$

The distance between our two probability distributions p and q in the Riemannian manifold uses the metric tensor $G(p)$ at p just like,

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

was used as the metric tensor in the Euclidean plane.

The coordinates at p are ($\lambda_1^p = 1.098612, \lambda_2^p = 1.098612$). Remember that p is the numerical assignment resulting from the $m = 2$ independence model. q is the probability distribution that is “close by,” so let $\Delta = (-0.01, +0.01)$. The coordinates for q are then displaced by Δ to ($\lambda_1^q = 1.108612, \lambda_2^q = 1.088612$).

Then,

$$\begin{aligned} d &= \sqrt{\left(\begin{pmatrix} -0.01 & +0.01 \end{pmatrix} \cdot \begin{pmatrix} 3/16 & 0 \\ 0 & 3/16 \end{pmatrix} \cdot \begin{pmatrix} -0.01 \\ +0.01 \end{pmatrix} \right)} \\ &= 0.00612372 \end{aligned}$$

Thus, by using the Fisher information matrix in this way to calculate distances between probability distributions in a Riemannian manifold, we have found a quantitative measure for the distance between the probability distribution dictated by the original independence model \mathcal{M}_2 and some near by probability distribution as dictated by another model. This geometric interpretation ties in beautifully with Kullback’s definition of information which we take up next.

28.3 Kullback’s Divergence Measure

There is a plethora of names associated with Kullback’s notion of information. Among the more popular terms are: *relative entropy*, *cross-entropy*, *directed divergence*, *KL measure*, *mean information per observation*, *alpha projection*, and many others to be sure.

The attendant confusion surrounding the disparate uses of the word *information* is understandable because the following equation was eventually seen to be the easiest way to comprehend Kullback's analysis,

$$KL(p, q) = \sum_{i=1}^n p_i \ln \left(\frac{p_i}{q_i} \right) \quad (28.5)$$

Well, of course, this bears a striking resemblance to Shannon's entropy, and the two were immediately conflated without much further thought on the matter.

But if one goes back to Kullback's original presentation, the origin of this formula is remarkable indeed. Kullback's inspiration was not the same as Shannon's; his inspiration was Bayes's Theorem.

28.3.1 The KL measure and Bayes's Theorem

Kullback decided that the critical idea, as based on Bayes's Theorem, was the log of the likelihood ratio. This he called "the information in an observation,"

the information result[s] from the observation $X = x$, and we define the logarithm of the likelihood ratio, $\log[f_1(x)/f_2(x)]$ as the information in $X = x$ for discrimination in favor of H_1 against H_2 . [Emphasis in the original.]

When he wanted the *mean* information for discrimination in favor of H_1 against H_2 , per observation from probability distribution H_1 , he derived,

$$KL(p, q) = \sum_{i=1}^n f_1(x_i) \ln \left[\frac{f_1(x_i)}{f_2(x_i)} \right]$$

Now, there is no doubt, despite this confusing notation, that Kullback intended for $f_1(x_i)$ and $f_2(x_i)$ to be understood as likelihoods. That is, the probability for the observed data is conditioned on two models, exactly as the likelihood is usually understood within Bayes's Theorem, or orthodox statistics for that matter.

He wrote Bayes's Theorem in this form for two models,

$$P(H_i | x) = \frac{P(H_i) f_i(x)}{P(H_1) f_1(x) + P(H_2) f_2(x)}$$

By rewriting this more clearly in our preferred notation as the probability for some model, and say we choose model M_A , when conditioned on the data, we regain familiar expressions.

These preferred expressions represent, of course, yet another application of the formal manipulation rules, *i.e.*, Bayes's Theorem together with the **Product** and **Sum Rules**. It is the orthogonal exercise to everything we have been doing with the MEP.

$$\begin{aligned}
P(\mathcal{M}_A \mid \mathcal{D}) &= \frac{P(\mathcal{D}, \mathcal{M}_A)}{P(\mathcal{D})} \\
&= \frac{P(\mathcal{D} \mid \mathcal{M}_A) P(\mathcal{M}_A)}{P(\mathcal{D})} \\
&= \frac{P(\mathcal{D} \mid \mathcal{M}_A) P(\mathcal{M}_A)}{P(\mathcal{D} \mid \mathcal{M}_A) P(\mathcal{M}_A) + P(\mathcal{D} \mid \mathcal{M}_B) P(\mathcal{M}_B)}
\end{aligned}$$

Make the identification with Kullback's notation that,

$$\begin{aligned}
f_1(x) &\equiv P(\mathcal{D} \mid \mathcal{M}_A) \\
P(H_1) &\equiv P(\mathcal{M}_A) \\
P(\mathcal{D}) &\equiv P(\mathcal{D} \mid \mathcal{M}_A) P(\mathcal{M}_A) + P(\mathcal{D} \mid \mathcal{M}_B) P(\mathcal{M}_B) \\
P(\mathcal{D} \mid \mathcal{M}_A) P(\mathcal{M}_A) + P(\mathcal{D} \mid \mathcal{M}_B) P(\mathcal{M}_B) &\equiv P(H_1) f_1(x) + P(H_2) f_2(x)
\end{aligned}$$

Bayes's Theorem is the absolutely indispensable probability manipulation rule we employ to reorder the relative standing of models after some data have been observed. The log of the ratio of any two models after the data is,

$$\begin{aligned}
\frac{P(\mathcal{M}_A \mid \mathcal{D})}{P(\mathcal{M}_B \mid \mathcal{D})} &= \frac{P(\mathcal{D} \mid \mathcal{M}_A)}{P(\mathcal{D} \mid \mathcal{M}_B)} \times \frac{P(\mathcal{M}_A)}{P(\mathcal{M}_B)} \\
\ln \left[\frac{P(\mathcal{M}_A \mid \mathcal{D})}{P(\mathcal{M}_B \mid \mathcal{D})} \right] &= \ln \left[\frac{P(\mathcal{D} \mid \mathcal{M}_A)}{P(\mathcal{D} \mid \mathcal{M}_B)} \right] + \ln \left[\frac{P(\mathcal{M}_A)}{P(\mathcal{M}_B)} \right]
\end{aligned}$$

Kullback also makes it the central principle of his information measure, but in a rather peculiar variation of the above result,

$$\ln \left[\frac{P(\mathcal{D} \mid \mathcal{M}_A)}{P(\mathcal{D} \mid \mathcal{M}_B)} \right] = \ln \left[\frac{P(\mathcal{M}_A \mid \mathcal{D})}{P(\mathcal{M}_B \mid \mathcal{D})} \right] - \ln \left[\frac{P(\mathcal{M}_A)}{P(\mathcal{M}_B)} \right] \quad (28.6)$$

In his notation, this becomes,

$$\log \frac{f_1(x)}{f_2(x)} = \log \frac{P(H_1 \mid x)}{P(H_2 \mid x)} - \log \frac{P(H_1)}{P(H_2)} \quad (28.7)$$

He clearly intended that the log of the likelihood ratio,

$$\ln \left[\frac{P(\mathcal{D} \mid \mathcal{M}_A)}{P(\mathcal{D} \mid \mathcal{M}_B)} \right] \equiv \ln \left[\frac{f_1(x)}{f_2(x)} \right] \quad (28.8)$$

be understood as the core concept in his information measure, albeit that the data here is but one observation.

He gives us the interpretation that the right hand side of Equation (28.6),

$$\ln \left[\frac{P(\mathcal{M}_A \mid \mathcal{D})}{P(\mathcal{M}_B \mid \mathcal{D})} \right] - \ln \left[\frac{P(\mathcal{M}_A)}{P(\mathcal{M}_B)} \right]$$

is a measure of how much we should favor model \mathcal{M}_A over \mathcal{M}_B after *one* data point. The difference on the right hand side is the *information* an IP receives when comparing the models before the data point and after the data point.

To my way of thinking, the most important aspect of Kullback's derivation as compared to Shannon's is that it is based entirely on Bayes's Theorem, or, in other words, based entirely on the formal manipulation rules for probability. We could have just as easily written down Kullback's variation,

$$\ln \left[\frac{P(\mathcal{D} | \mathcal{M}_A)}{P(\mathcal{D} | \mathcal{M}_B)} \right] = \ln \left[\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} \right] - \ln \left[\frac{P(\mathcal{M}_A)}{P(\mathcal{M}_B)} \right]$$

back in Volume I when we were concentrating on the formal rules.

Shannon, on the other hand, set up some axioms which he thought any information measure should obey, and then derived his entropy measure. There was no need to invoke Bayes's Theorem at any point in his derivation.

Our use of the symbol \mathcal{D} is somewhat misleading. Kullback, at this juncture, is not defining his measure with respect to all of the data, but only with respect to one observation x . If, in fact, as Kullback says, the emphasis is on a single observation ($X = x_i$), then the log of the likelihood ratio for a single observation when conditioned on two models is,

$$\ln \left[\frac{f_1(x)}{f_2(x)} \right] \equiv \ln \left[\frac{P(X = x_i | \mathcal{M}_A)}{P(X = x_i | \mathcal{M}_B)} \right] \equiv \ln \left(\frac{Q_i^A}{Q_i^B} \right) \quad (28.9)$$

Kullback wants the *mean* information for discrimination in favor of model \mathcal{M}_A against model \mathcal{M}_B . So he takes the average of this expression with respect to Q_i^A to define his information measure $I(A : B)$ relating the two distributions under models \mathcal{M}_A and \mathcal{M}_B ,

$$I(A : B) = \sum_{i=1}^n Q_i^A \ln \left(\frac{Q_i^A}{Q_i^B} \right) \quad (28.10)$$

28.3.2 KL measure and maximum entropy

After all of this, there really is no disagreement between the MEP and Kullback's information measure. If we *minimize* $I(A : B)$, then we are doing the same thing as *maximizing* Shannon's entropy $H(p)$, if we recognize one major *caveat*. The distribution under model \mathcal{M}_B must be the uniform distribution under the fair model.

Trying to get a distribution under \mathcal{M}_A as "close" to the uniform distribution under \mathcal{M}_B by minimizing Kullback's measure is the same as maximizing Shannon's information entropy for the distribution under \mathcal{M}_A . And this makes sense because under Kullback's interpretation we are trying to get the distribution under \mathcal{M}_A as close as possible to the distribution with "no information."

Strictly for our own notational convenience when we delve into *Information Geometry*, and to make contact with the form typically seen, rewrite Kullback's information measure $I(A : B)$ for the discrete case as,

$$I(A : B) \equiv KL(p, q) = \sum_{i=1}^n p_i \ln \left(\frac{p_i}{q_i} \right) \quad (28.11)$$

where p and q refer to points representing probability distributions.

Furthermore, let's standardize by calling this measure *relative entropy*. It can be reworked into this format,

$$KL(p, q) = \sum_{i=1}^n p_i \ln \left(\frac{p_i}{q_i} \right) = \sum_{i=1}^n p_i \ln p_i - \sum_{i=1}^n p_i \ln q_i \quad (28.12)$$

The first term, $\sum_{i=1}^n p_i \ln p_i$, we recognize as the negative of Shannon's information entropy for p . If the distribution q is the uniform distribution over a state space with dimension n , then $q_i = 1/n$.

The relative entropy between the distribution p and the uniform distribution q is then,

$$\begin{aligned} KL(p, q) &= \sum_{i=1}^n p_i \ln p_i - \sum_{i=1}^n p_i \ln q_i \\ &= \sum_{i=1}^n p_i \ln p_i - \sum_{i=1}^n p_i \ln \left(\frac{1}{n} \right) \\ &= \sum_{i=1}^n p_i \ln p_i + \ln n \\ &= \ln n - H(p) \end{aligned}$$

This is the negative of Shannon's entropy together with the positive constant $\ln n$. Once the dimension n of the state space is established for a given inferential problem, it does not change. $\ln n$ is then a constant. So we *minimize* $KL(p, q)$, in this case, by *maximizing* $H(p)$.

28.3.3 Two extreme cases and the middle

Taking the first of two extreme cases, if p were itself the uniform distribution, then there could be no information in favor of p against q . Kullback's measure would then be zero, because $KL(p, q) = \ln n - \ln n$. The second extreme case would have p indicate a statement certain to be true. Then, Kullback's measure would be at the maximum value of $\ln n$, because $KL(p, q) = \ln n - 0$. This is the largest mean information per observation in favor of discriminating p from q .

Everything else would be between these two extremes. For example, if $n = 3$ and p were $(1/3, 1/2, 1/6)$, then Kullback's measure of information for discriminating between p and $q = (1/3, 1/3, 1/3)$ would be greater than 0, but less than $\ln 3$.

$$\begin{aligned} \sum_{i=1}^n p_i \ln p_i &= (1/3 \ln 1/3) + (1/2 \ln 1/2) + (1/6 \ln 1/6) \\ &= -1.0114 \\ \ln 3 &= 1.0986 \\ KL(p, q) &= \ln n + \sum_{i=1}^n p_i \ln p_i \\ &= 1.0986 - 1.0114 \\ &= 0.0872 \end{aligned}$$

If you minimize Kullback's information measure for discriminating some probability distribution p that must satisfy some constraints against the most uninformative distribution q that satisfies no constraints (except the universal constraint), then you are finding the MEP distribution.

Again, conceptual ideas can best be illustrated with extreme cases. Suppose, that the distribution p must satisfy the constraint average of $\langle F \rangle = 2$ where the constraint function is $F(X = x_i) = (1, 2, 3)$. Let p be $(0, 1, 0)$ which does satisfy this constraint as well as the universal constraint. Kullback's measure between p and the most uninformative distribution $q = (1/3, 1/3, 1/3)$ is $\ln 3 = 1.0986$.

But is this the minimum value of the information measure? No, it is not. If you vary the distribution p trying to find the minimum discrepancy between p and the most uninformative distribution, then you know that the minimum must occur at 0. This p would then have $\sum_{i=1}^3 p_i \ln p_i = -\ln 3$. In other words, p must possess the absolute minimum discrimination against q , or $p = (1/3, 1/3, 1/3)$. All the points p , as you vary them trying to achieve the minimum value of Kullback's measure, must always satisfy the constraints of $\langle F \rangle = 2$.

As a less extreme illustration, what is the minimum KL measure between some distribution p and the uninformative distribution q when p must satisfy the constraint of $\langle F \rangle = 2.1$? (Refer back to Exercise 24.5.1.) As a thought experiment, imagine you are exploring the space of all p distributions satisfying this constraint.

Here is one of those p distributions, $p = (0.2, 0.5, 0.3)$, satisfying both constraints and which has an information entropy of,

$$H(p_i) = - \sum_{i=1}^3 p_i \ln p_i = 1.0297$$

Thus, the separation between this distribution p and the distribution with the

maximum amount of missing information q is,

$$KL(p, q) = \ln n + \sum_{i=1}^3 p_i \ln p_i = 1.0986 - 1.0297 = 0.0689$$

Have we minimized the mean discriminability between any p and the most uninformative distribution q ? If we were to continue our search, we would eventually discover that the minimum relative entropy occurs for a distribution p ,

$$p = (0.2846, 0.3308, 0.3846)$$

This distribution also satisfies the information inserted under both constraints, but is closer to the distribution with no information. Now, for this optimal p ,

$$H(p_i) = - \sum_{i=1}^3 p_i \ln p_i = 1.0911$$

Thus, the separation, or literally, “squared distance” between these two distributions is,

$$KL(p, q) = \ln n + \sum_{i=1}^3 p_i \ln p_i = 1.0986 - 1.0911 = 0.0075$$

This is as “close” as any p can come to this particular q , a uniform distribution of $1/n$ having maximum missing information. We are very close to 0 with this best distribution p , but the relative entropy cannot actually be 0 because then p would itself be the distribution with the maximum amount of missing information $p = (1/3, 1/3, 1/3)$. Since the constraint function average of $\langle F \rangle = 2$ under this model does not satisfy the specified information of $\langle F \rangle = 2.1$, it is unacceptable.

Of course, we found this satisfactory p through the MEP. The p that had maximum missing information, that is, had maximum Shannon information entropy subject to the constraints was $p = (0.2846, 0.3308, 0.3846)$. So, in the end, you can search for that p that has maximum entropy while still satisfying the constraints. Or, in Kullback’s rather strained explanation, you can search for that distribution that has a minimum mean information for discrimination in favor of p against a uniform distribution q . In either case, you end up with the same distribution.

28.3.4 A Legendre transformation for relative entropy

The Legendre transformation was found to implement a computational algorithm for the MEP. We devoted an entire Chapter to reflect its importance. The relative entropy can be reworked into a similar expression.

Recapitulating the Legendre transformation as we pursued it in Chapter Twenty Four, but now with an emphasis on the expectation with respect to distribution p ,

$$\begin{aligned}
-\sum_{i=1}^n p_i \ln p_i &= -E_p (\ln p_i) \\
&= -E_p \left[\sum_{j=1}^m \lambda_j^p F_j(X = x_i) - \ln Z_p \right] \\
&= - \left[\sum_{j=1}^m \lambda_j^p E_p [F_j(X = x_i)] - E_p (\ln Z_p) \right] \\
&= - \left[\sum_{j=1}^m \lambda_j^p \langle F_j \rangle_p - \ln Z_p \right] \\
&= \ln Z_p - \sum_{j=1}^m \lambda_j^p \langle F_j \rangle_p \\
H_{max}(p) &= \min_{\lambda_j^p} \left[\ln Z_p - \sum_{j=1}^m \lambda_j^p \langle F_j \rangle_p \right]
\end{aligned}$$

A similar expression can be worked out for $KL(p, q)$,

$$\begin{aligned}
\sum_{i=1}^n p_i \ln \left(\frac{p_i}{q_i} \right) &= \sum_{i=1}^n p_i (\ln p_i - \ln q_i) \\
&= \sum_{i=1}^n p_i \ln p_i - \sum_{i=1}^n p_i \ln q_i \\
&= E_p \left[\sum_{j=1}^m \lambda_j^p F(X = x_i) - \ln Z_p \right] - E_p \left[\sum_{j=1}^m \lambda_j^q F(X = x_i) - \ln Z_q \right] \\
&= \left[\sum_{j=1}^m \lambda_j^p \langle F_j \rangle_p - \ln Z_p \right] - \left[\sum_{j=1}^m \lambda_j^q \langle F_j \rangle_p - \ln Z_q \right] \\
KL(p, q) &= \sum_{j=1}^m (\lambda_j^p - \lambda_j^q) \langle F_j \rangle_p + \ln \left(\frac{Z_q}{Z_p} \right)
\end{aligned}$$

28.3.5 KL measure as a distance measure

Kullback's information measure has a very direct connection with Fisher's information measure. It is, as alluded to above, a distance measure between two probability distributions p and q when these two distributions are considered abstractly as points in a Riemannian manifold. As a matter of fact, the square root of twice the KL measure between the two probability distributions is approximately the same distance as calculated through the Fisher information matrix.

Consider the same example as in section 28.2. There we calculated that the "distance" between an independence model and a near by model was 0.00612372. Calculate the distance between these two distributions using Kullback's measure,

$$\begin{aligned} KL(p, q) &= \sum_{i=1}^4 p_i \ln \left(\frac{p_i}{q_i} \right) \\ &= p_1 \ln \left(\frac{p_1}{q_1} \right) + \cdots + p_4 \ln \left(\frac{p_4}{q_4} \right) \\ &= 9/16 \ln \left(\frac{9/16}{q_1} \right) + \cdots + 1/16 \ln \left(\frac{1/16}{q_4} \right) \end{aligned}$$

The numerical assignments under the model close to independence are calculated using $\lambda_1^q = 1.108612$ and $\lambda_2^q = 1.088612$ in the MEP formula. The coordinates for q were shifted by a $\Delta = (+0.01, -0.01)$. The MEP algorithm yields the following assignment under the near by model,

$$q = (0.562489, 0.189381, 0.185651, 0.062499)$$

This can be compared with the assignment under the independence model,

$$p = (0.5625, 0.1875, 0.1875, 0.0625)$$

The computation can now be completed,

$$KL(p, q) = 0.5625 \ln \left[\frac{0.5625}{0.562489} \right] + \cdots + 0.0625 \ln \left[\frac{0.0625}{0.062499} \right]$$

$$\sqrt{2 \cdot KL(p, q)} = 0.00612372$$

Thus, both the Fisher information and Kullback information are involved in finding the distance between two probability distributions. They also provide the same answer.

28.4 Data as Information

Even in this very perfunctory, broad-brush comparison of both Fisher information and relative entropy with Shannon entropy, we detect a basic fundamental difference distinguishing them. In both Fisher information and relative entropy,

data is information!

Moreover, this information stems from some pre-existing probability distribution. Neither technique establishes an initial numerical assignment for any of the probability distributions in question. Where did those p and q distributions come from?

Shannon's measure of *missing information* has as its *raison d'être* an initial numerical assignment to what was formerly merely an abstract probability. The information consists of the constraint functions and their averages.

information is NOT data!

By maximizing Shannon's missing information measure, the positive contribution from the constraint functions and their averages is guaranteed to be the only information involved. Nowhere is there any reliance on data in this process of assigning initial numerical values. The information provided by any model is independent of the data.

28.5 Connections to the Literature

We examine some of what I consider to be peculiarities of Kullback's way of thinking about information. His take on these issues is not quite what you might expect. Moreover, his language and notation always demand careful attention to make sure that they jibe with what you think they mean. My suspicion, based on my own personal experience, is that we read his interlocutors first, and then are a bit surprised when we try to relocate this same attitude in Kullback's own writings.

It is a shame that Kullback's book [26] is such a difficult read because there are many highlights. For example, one is met with a heavy dose of measure theory right off the bat in the first few pages. I'm sure that this kind of introduction discouraged many people from further inquiry. Like many books in math, if you possess the patience to ignore some early impenetrable parts, there are often hidden delights in plain sight further down the road.

It doesn't take too long before there is an initial clash between what you thought you were going to extract from Kullback and what is actually there in his own words. Here, we are going to take a look at Kullback's very first examples and problems. Immediately, we get the sense that there may be some unresolved differences in approach lurking in the wings.

You might expect that your first exposure to what hopefully are some simple examples in his Section 4 of Chapter 1 would alleviate any lingering discomfort over really having understood his definition of information. What follows is a verbatim presentation of his first two examples, *Example 4.1* and *Example 4.2*. [26, pg. 7]

My critical comments are interspersed along the way. I was surprised to have this much trouble so early on trying to decipher Kullback's fundamental concepts.

As an extreme case, suppose that H_2 represents a set of hypotheses, one of which must be true, and that H_1 is a member of the set of hypotheses H_2 : then $P(H_2) = 1$, $P(H_2 | x) = 1$, and the right hand side of,

$$\log \frac{f_1(x)}{f_2(x)} = \log \frac{P(H_1 | x)}{P(H_2 | x)} - \log \frac{P(H_1)}{P(H_2)}[\lambda]$$

yields as the information in x in favor of H_1 , the value

$$\log P(H_1 | x) - \log P(H_1) = \log \frac{P(H_1 | x)}{P(H_1)}$$

When is this value zero?

Let our attempts at deciphering Kullback's thought processes begin with this version of the coin tossing scenario. Suppose that H_2 is a set consisting of just two models. Model \mathcal{M}_A inserts information such that $P(\text{HEADS}) = Q_1^A$ and $P(\text{TAILS}) = Q_2^A$. The second model \mathcal{M}_B inserts information such that $P(\text{HEADS}) = Q_1^B$ and $P(\text{TAILS}) = Q_2^B$. The set H_2 then consists of these two models $\{\mathcal{M}_A, \mathcal{M}_B\}$, one of which must be true.

The entire model space then consists of just these two models, $\mathcal{M} = 2$, where, by definition,

$$P(H_2) \equiv \sum_{k=1}^{\mathcal{M}} P(\mathcal{M}_k) = P(\mathcal{M}_A) + P(\mathcal{M}_B) = 1$$

H_1 is a member of the set H_2 where we let $H_1 \equiv \mathcal{M}_A$.

Kullback uses the notation x to represent *one* observation. Here, in our example, this is either HEADS or TAILS. Now, it doesn't make any difference whether the probability for these two models is conditioned on an observation x or not; they still must sum to 1. Therefore,

$$P(\mathcal{M}_A | x) + P(\mathcal{M}_B | x) = 1$$

and so it is true that $P(H_2 | x) = 1$.

For example, let the probability of HEADS ($X = x_1$) or TAILS ($X = x_2$) under each of the two models be,

$$P(X = x_1 | \mathcal{M}_A) = 0.8$$

$$P(X = x_2 | \mathcal{M}_A) = 0.2$$

$$P(X = x_1 | \mathcal{M}_B) = 0.4$$

$$P(X = x_2 | \mathcal{M}_B) = 0.6$$

with equal prior probabilities for the models $P(\mathcal{M}_A) = P(\mathcal{M}_B) = 1/2$.

By Bayes's Theorem, the probability for model \mathcal{M}_A after observing, say, one HEADS is updated from 1/2 to,

$$\begin{aligned} P(\mathcal{M}_A | X = x_1) &= \frac{P(X = x_1 | \mathcal{M}_A) P(\mathcal{M}_A)}{P(X = x_1 | \mathcal{M}_A) P(\mathcal{M}_A) + P(X = x_1 | \mathcal{M}_B) P(\mathcal{M}_B)} \\ &= \frac{0.8 \times 0.5}{(0.8 \times 0.5) + (0.4 \times 0.5)} \\ &= 2/3 \end{aligned}$$

Likewise, the probability for model \mathcal{M}_B after observing one HEADS is updated to,

$$\begin{aligned} P(\mathcal{M}_B | X = x_1) &= \frac{P(X = x_1 | \mathcal{M}_B) P(\mathcal{M}_B)}{P(X = x_1 | \mathcal{M}_B) P(\mathcal{M}_B) + P(X = x_1 | \mathcal{M}_A) P(\mathcal{M}_A)} \\ &= \frac{0.4 \times 0.5}{(0.4 \times 0.5) + (0.8 \times 0.5)} \\ &= 1/3 \end{aligned}$$

confirming that the sum of the probabilities after conditioning on the data still must equal 1,

$$P(\mathcal{M}_A | x) + P(\mathcal{M}_B | x) = 1$$

The re-orientation of the relative importance of the two models after observing HEADS as a result of the toss also makes sense since HEADS had a higher probability under model \mathcal{M}_A .

Thus, it is also true as Kullback states in his example that,

$$\begin{aligned} \ln \left[\frac{P(\mathcal{M}_A | x)}{P(\mathcal{M}_A | x) + P(\mathcal{M}_B | x)} \right] - \ln \left[\frac{P(\mathcal{M}_A)}{P(\mathcal{M}_B) + P(\mathcal{M}_B)} \right] &= \ln P(\mathcal{M}_A | x) - \ln P(\mathcal{M}_A) \\ &= \ln \left[\frac{P(\mathcal{M}_A | x)}{P(\mathcal{M}_A)} \right] \end{aligned}$$

The language Kullback employs to explain this is that the *information* in favor of model \mathcal{M}_A by observing HEADS one time is,

$$\begin{aligned} \text{Information} &\equiv \ln \left[\frac{P(\mathcal{M}_A | \text{HEADS})}{P(\mathcal{M}_A)} \right] \\ &= \ln \left[\frac{2/3}{1/2} \right] \\ &= 0.2877 \end{aligned}$$

The information would be 0 when the ratio on the right hand side is 1, or when the probability for the model conditioned on the data is the same as its prior probability. This fits in with Kullback's definition of information as the change in the probability for the models by comparing the model's posterior probability with its prior probability.

Kullback continues on in his *Example 4.1* with a special case which he wishes to exploit in his next example as the definition of entropy.

If the observation x proves that H_1 is true, that is, $P(H_1 | x) = 1$, then the information in x about H_1 is $-\log P(H_1)$ Note that when H_1 is initially of small probability the information resulting from its verification is large, whereas if its probability initially is large the information is small. Is this intuitively reasonable?

If $P(\mathcal{M}_A | \text{HEADS})$ were to equal 1, then by the **Sum Rule** we know that $P(\mathcal{M}_B | \text{HEADS})$ must equal 0. The assignment under model \mathcal{M}_A would then be $Q_1^A = 1, Q_2^A = 0$, the model of choice for a coin with two HEADS. The assignment under model \mathcal{M}_B would then be $Q_1^B = 0, Q_2^B = 1$, the model of choice for a coin with two TAILS. The appearance of HEADS rules out model \mathcal{M}_B , and leaves model \mathcal{M}_A with a probability of 1 after seeing HEADS.

$$\ln P(\mathcal{M}_A | \text{HEADS}) - \ln P(\mathcal{M}_A) = \ln 1 - \ln P(\mathcal{M}_A) = -\ln P(\mathcal{M}_A)$$

If the prior probability for the model $P(\mathcal{M}_A)$ were originally $1/2$, (implying that $P(\mathcal{M}_B) = 1/2$), then the information in HEADS for discriminating in favor of \mathcal{M}_A is $-\ln(1/2) = 0.6931$. If the prior probability for $P(\mathcal{M}_A)$ was originally 0.01, then the information in HEADS for discriminating in favor of \mathcal{M}_A is $-\ln(0.01) = 4.6052$.

This was a very long excursion to see if any inconsistencies cropped up between Kullback's notation and its interpretation within our preferred notation. So far, no such notational or conceptual inconsistencies have appeared, and apparently, when Kullback set out the basis for his information measure, he really did mean it to be a reworking of Bayes's Theorem.

Peculiarly, he presents the log likelihood ratio as the difference between the ratio of the posterior probability for two models and the ratio of the prior probability for the two models.

$$\begin{aligned} \log \frac{f_1(x)}{f_2(x)} &= \log \frac{P(H_1 | x)}{P(H_2 | x)} - \log \frac{P(H_1)}{P(H_2)} \\ \ln \left[\frac{P(\text{HEADS} | \mathcal{M}_A)}{P(\text{HEADS} | \mathcal{M}_B)} \right] &= \ln \left[\frac{P(\mathcal{M}_A | \text{HEADS})}{P(\mathcal{M}_B | \text{HEADS})} \right] - \ln \left[\frac{P(\mathcal{M}_A)}{P(\mathcal{M}_B)} \right] \end{aligned}$$

But Kullback wants the *mean* information, per observation, for discrimination in favor of model \mathcal{M}_A against model \mathcal{M}_B . This entails taking an average with respect

to $P(x | \mathcal{M}_A)$. So, multiplying the information in the form of the log likelihood ratio for each possible observation by the probability for that observation under model \mathcal{M}_A results in,

$$\left(P(\text{HEADS} | \mathcal{M}_A) \times \ln \left[\frac{P(\text{HEADS} | \mathcal{M}_A)}{P(\text{HEADS} | \mathcal{M}_B)} \right] \right) + \left(P(\text{TAILS} | \mathcal{M}_A) \times \ln \left[\frac{P(\text{TAILS} | \mathcal{M}_A)}{P(\text{TAILS} | \mathcal{M}_B)} \right] \right)$$

All of this is more easily digestible when we use our short form already defined and used extensively,

$$Q_i \equiv P(X = x_i | \mathcal{M}_k)$$

Kullback's information measure,

$$I(1 : 2) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} d\lambda(x)$$

becomes in our notation,

$$KL(p, q) = \sum_{i=1}^n p_i \ln \left(\frac{p_i}{q_i} \right) \equiv \sum_{i=1}^n Q_i^A \ln \left(\frac{Q_i^A}{Q_i^B} \right)$$

So, for example, Kullback's information measure in the coin tossing example has a value of,

$$KL(p, q) = 0.8 \ln \left(\frac{0.8}{0.4} \right) + 0.2 \ln \left(\frac{0.2}{0.6} \right) = 0.3348$$

This value is also a measure of the “separation” between the distribution under model \mathcal{M}_A and the distribution under model \mathcal{M}_B .

Take note that his information measure was derived from just *one* observation. Don't be confused by my dragging in the word “data,” even though it's true HEADS is one data point. But it is not the whole collection of data points as usually implied when the word “data” is used.

Mentioning the case where $P(H_1 | x) = 1$ in *Example 4.1* was a way to introduce the generalization in his next example, *Example 4.2*.

To carry this notion somewhat further, suppose a set of mutually exclusive and exhaustive hypotheses H_1, H_2, \dots, H_n exists and that from an observation we can infer which of the hypotheses is true. . . . Here, the mean information in an observation about the hypotheses is . . .

$$-P(H_1) \log P(H_1) - P(H_2) \log P(H_2) - \dots - P(H_n) \log P(H_n)$$

Th[is] expression . . . is also called the entropy of the H_i 's. . . .

Thus, in a “yes” or “no” selection with a probability of $1/2$ for each alternative,

$$-1/2 \log 1/2 - 1/2 \log 1/2 = \log_2 2 = 1 \text{ “bit.”}$$

When n hypotheses are equally probable, so that $P(H_i) = 1/n, i = 1, \dots, n$ we find that,

$$-\sum_{i=1}^n P(H_i) \log P(H_i) = \log n$$

Here some serious confusion intrudes upon our ongoing deciphering of Kullback's intentions. If the match-up between hypotheses and models is correct, which from his exposition of Bayes's Theorem and *Example 4.1* seems mandatory, than why is he forming a sum over the probability for models and calling it "entropy?"

Any sum over the \mathcal{M} models in model space is going to be a sum over a different index than sum over n statements in state space. For example, this distinction is reflected in expressions like these,

$$-\sum_{k=1}^{\mathcal{M}} P(\mathcal{M}_k) \ln P(\mathcal{M}_k)$$

versus,

$$-\sum_{i=1}^n P(X = x_i | \mathcal{M}_k) \ln P(X = x_i | \mathcal{M}_k) \equiv -\sum_{i=1}^n Q_i \ln Q_i$$

How can this expression,

$$-\sum_{i=1}^n P(H_i) \log P(H_i)$$

be the entropy when the entropy has been defined according to Shannon as a sum over the assigned probability to the n statements in *state space*? There is confusion over Kullback's use of the notation n and H_i . The confusion centers on which space is being discussed. Is it the state space of n statements, or the model space of \mathcal{M} model statements?

But it also true, as Kullback had pointed out in *Example 4.1*, that, when an observation x rules out all other hypotheses except one, the log likelihood expressed as the difference between the posterior probability for the model minus the prior probability for the model is indeed minus the log of the prior probability. We gave as an example observing HEADS where the one model assigned probability 1 to HEADS and the other model assigned 0 to HEADS.

$$\ln \left[\frac{P(\text{HEADS} | \mathcal{M}_A)}{P(\text{HEADS} | \mathcal{M}_B)} \right] \equiv \ln P(\mathcal{M}_A | \text{HEADS}) - \ln P(\mathcal{M}_B) = \ln \frac{1}{P(\mathcal{M}_B)} = -\ln P(\mathcal{M}_B)$$

How could it then come about that there was an equivalency between,

$$-\sum_{i=1}^n Q_i \ln Q_i$$

and,

$$-\sum_{k=1}^{\mathcal{M}} P(\mathcal{M}_k) \ln P(\mathcal{M}_k)?$$

In the statement of the problem, to reinforce this earlier notion, Kullback said that,

Or we may be dealing with an experiment for which the outcome may be one of n categories, there are no errors of observation, and there is no uncertainty about the inference of the category after making the observation.

This can only mean that, as we discussed before in the coin tossing scenario, each model assigns 1 to one of the statements and 0 to all of the remaining statements. This is the only way that a single observation could prove that one of the models was certain and the rest were ruled out.

For example, if the experiment were rolling a die, then $\mathcal{M} = 6$ models would be proposed,

$$P(X = x_i | \mathcal{M}_A) = (1, 0, 0, 0, 0, 0) \text{ through } P(X = x_i | \mathcal{M}_F) = (0, 0, 0, 0, 0, 1)$$

When a THREE is actually observed, the five models $\mathcal{M}_A, \mathcal{M}_B, \mathcal{M}_D, \mathcal{M}_E$ and \mathcal{M}_F are ruled out, while model \mathcal{M}_C has a probability $P(\mathcal{M}_C | \text{THREE}) = 1$.

Fulfilling Kullback's criterion in this manner, we have that the information in the observation is $-\ln P(\mathcal{M}_C)$, leading to the mean information about the hypotheses $-\sum_{k=1}^6 P(\mathcal{M}_k) \ln P(\mathcal{M}_k)$. If all six models have been given the same prior probability of $P(\mathcal{M}_k) = 1/6$, then according to Kullback's interpretation of the entropy of the H_i 's

$$H(Q_i) = -\sum_{k=1}^{\mathcal{M}} P(\mathcal{M}_k) \ln P(\mathcal{M}_k) = \ln 6$$

Meanwhile, returning to the Shannon entropy for the uniform distribution over the $n = 6$ statements in the state space, (the numerical assignment under *one* model, the fair model), we find, of course, that,

$$H(Q_i) = -\sum_{i=1}^6 Q_i \ln Q_i = \ln 6$$

We are left with a numerical confirmation for conflicting concepts. This ending we cannot afford to accept at face value.

The immediate resolution is that a serious confusion about fundamental concepts has been papered over by a judicious choice of exactly $\mathcal{M} = n = 6$ models. These six models also possess just the right properties and, furthermore, possess just the right prior probabilities to duplicate the correct entropy formula.

What began with a hope that Kullback's very first examples would clarify his fundamental concepts concerning information has been dashed. Careful consideration of his opening explanations instead leave one in a quite serious state of confusion! I am sorry to say that things don't get much better as one proceeds further along into his book.

28.6 Solved Exercises for Chapter Twenty Eight

Exercise 28.6.1: Consider the standard coin toss scenario. What is the variance for our usual constraint function?

Solution to Exercise 28.6.1

The definition of the variance for the one constraint function $F(X = x_i) = (1, 2)$ used in the coin tossing scenario is,

$$\text{Var}[F(X = x_i)] = E\{\[F(X = x_i) - \mu_X] \times [F(X = x_i) - \mu_X]\}$$

Any mathematical expectation must, of course, be defined with respect to some numerical assignment Q_i under a given model,

$$Q_i \equiv P(X = x_i | \mathcal{M}_k)$$

In particular, μ_X is the mathematical expectation of the constraint function with respect to the assignments under some model. I advocate utilizing the MEP formula to find these required Q_i .

Just as in Chapter Seventeen, suppose that some model inserted the information into the probability distribution that $\langle F \rangle = 1.25$. Then we know that the assignment under this model is $Q_i = (0.75, 0.25)$. With the identification of $\mu_X \equiv \langle F \rangle$, we now have all the ingredients to calculate the variance of the constraint function as,

$$\begin{aligned} \text{Var}[F(X = x_i)] &= E\{\[F(X = x_i) - \langle F \rangle] \times [F(X = x_i) - \langle F \rangle]\} \\ &= \sum_{i=1}^2 [F(X = x_i) - \langle F \rangle] \times [F(X = x_i) - \langle F \rangle] P(X = x_i | \mathcal{M}_k) \\ &= [(1 - 1.25) \times (1 - 1.25) \times 0.75] + [(2 - 1.25) \times (2 - 1.25) \times 0.25] \\ &= [(-0.25)^2 \times 0.75] + [(0.75)^2 \times 0.25] \\ &= 0.1875 \end{aligned}$$

From the perspective of Fisher information, with only one constraint function, $g_{rc}(p)$ is no longer a matrix, but just a single number $g(p)$ defined at the point p with one coordinate λ . The computational formula for $g(p)$ is the same as just worked out for the variance.

Following Equation (28.2) with only one parameter λ , and strict adherence to the MEP formalism,

$$\begin{aligned}
 g(p) &= \sum_{i=1}^2 \left[\left(\frac{\partial \ln P(X = x_i | \mathcal{M}_k)}{\partial \lambda} \right)^2 \right] \times P(X = x_i | \mathcal{M}_k) \\
 \frac{\partial \ln P(X = x_i | \mathcal{M}_k)}{\partial \lambda} &= F(x_i) - \frac{\partial \ln Z}{\partial \lambda} \\
 \frac{\partial \ln Z}{\partial \lambda} &= \langle F \rangle \\
 \left[\frac{\partial \ln P(X = x_i | \mathcal{M}_k)}{\partial \lambda} \right]^2 &= [F(x_i) - \langle F \rangle]^2 \\
 g(p) &= \sum_{i=1}^2 [F(x_i) - \langle F \rangle]^2 Q_i \\
 &= 0.1875
 \end{aligned}$$

Exercise 28.6.2: What is the “distance” between the distribution under the model in the first exercise and a nearby distribution?

Solution to Exercise 28.6.2

The model in the first exercise was defined by its parameter $\lambda^p = -1.098612$ as we saw in Chapter Seventeen. A “nearby” model would have this parameter displaced slightly to, say, $\lambda^q = -1.088612$ defining a new model and a new point q . The formula for calculating the distance between points p and q is,

$$d = \sqrt{(\lambda^p - \lambda^q) \cdot G \cdot (\lambda^p - \lambda^q)^T}$$

which for the simple case here works out to,

$$\begin{aligned}
 d &= \sqrt{(-1.098612 + 1.088612) \times g(p) \times (-1.098612 + 1.088612)} \\
 &= \sqrt{(-0.01)^2 \times 0.1875} \\
 &= 0.0044
 \end{aligned}$$

Exercise 28.6.3: Confirm this answer found in the last exercise using the Kullback measure.

Solution to Exercise 28.6.3

The Kullback measure for the separation between these two distributions is,

$$\begin{aligned} KL(p, q) &= \sum_{i=1}^2 p_i \ln \left(\frac{p_i}{q_i} \right) \\ &= 0.75 \ln \left(\frac{0.75}{q_1} \right) + 0.25 \ln \left(\frac{0.25}{q_2} \right) \end{aligned}$$

We find the numerical assignments q_1 and q_2 under the second model from the MEP formula as,

$$\begin{aligned} q_1 &= \frac{e^{\lambda F(X=x_1)}}{Z(\lambda)} \\ &= \frac{e^{(-1.088612 \times 1)}}{Z(\lambda)} \\ q_2 &= \frac{e^{\lambda F(X=x_2)}}{Z(\lambda)} \\ &= \frac{e^{(-1.088612 \times 2)}}{Z(\lambda)} \\ Z(\lambda) &= e^{(-1.088612 \times 1)} + e^{(-1.088612 \times 2)} \\ q_1 &= 0.7481 \\ q_2 &= 0.2519 \end{aligned}$$

The numerical assignments under this nearby model must be close to the original model. This is confirmed with the calculation that $q_1 = 0.7481$ and $q_2 = 0.2519$. Substituting these values into the Kullback measure,

$$\begin{aligned} KL(p, q) &= 0.75 \ln \left(\frac{0.75}{q_1} \right) + 0.25 \ln \left(\frac{0.25}{q_2} \right) \\ &= 0.75 \ln \left(\frac{0.7500}{0.7481} \right) + 0.25 \ln \left(\frac{0.2500}{0.2519} \right) \\ &= 9.56 \times 10^{-6} \end{aligned}$$

The distance between the two probability distributions as found by using Fisher information is confirmed by,

$$\sqrt{(\lambda^p - \lambda^q) \cdot G \cdot (\lambda^p - \lambda^q)^T} \equiv \sqrt{2 \times KL(p, q)} = 0.0044$$

Exercise 28.6.4: Carry out the calculations to determine the distance between the probability distribution under the correlation model in the first kangaroo scenario and another close by probability distribution.

Solution to Exercise 28.6.4

The distance will be found by using both Fisher information and the KL measure. The correlation model in the kangaroo scenario of Chapter Twenty One used $m = 3$ constraint functions and their associated averages.

Therefore, the Fisher information matrix, a metric tensor, will be a 3×3 matrix. Since this is a symmetric matrix, only the $(m \times (m + 1))/2 = 6$ entries from the full matrix consisting of 9 entries are needed. These six entries are the variances for each of the three constraint functions, together with the three covariances between functions 1 and 2, functions 1 and 3, and functions 2 and 3.

Referring back to Chapter Twenty One, section 21.3.3, these constraint functions and their expectations under the correlation model were,

$$F_1(X = x_i) = (1, 1, 0, 0) \text{ with } \langle F_1 \rangle = 0.75$$

$$F_2(X = x_i) = (1, 0, 1, 0) \text{ with } \langle F_2 \rangle = 0.75$$

$$F_3(X = x_i) = (1, 0, 0, 0) \text{ with } \langle F_3 \rangle = 0.70$$

The Fisher information matrix is calculated as the matrix,

$$G(p) = \begin{pmatrix} 0.1875 & 0.1375 & 0.1750 \\ 0.1375 & 0.1875 & 0.1750 \\ 0.1750 & 0.1750 & 0.2100 \end{pmatrix}$$

Observe that the entries $g_{13}(p) = 0.1750$ in the first row and third column and $g_{31}(p) = 0.1750$ in the third row and first column are equal due to symmetry in the information matrix. In other words, the covariance between $F_1(X = x_i)$ and $F_3(X = x_i)$ must be the same as the covariance between $F_3(X = x_i)$ and $F_1(X = x_i)$.

Table 28.2, at the top of the next page, shows all of the details for this particular entry. The last column sum is the computation of $g_{31}(p)$,

$$g_{31}(p) = \sum_{i=1}^4 \{ [(F_3(x_i) - \langle F_3 \rangle) \times (F_1(x_i) - \langle F_1 \rangle)] \times Q_i \}$$

Table 28.2: The computational details for g_{31} , the covariance between the third and first constraint functions under the correlation model in the simplified kangaroo scenario of Chapter Twenty One.

$F_3(x_i) - \langle F_3 \rangle$	$F_1(x_i) - \langle F_1 \rangle$	$Col\ 1 \times Col\ 2$	Q_i	$Col\ 3 \times Col\ 4$
1 – 0.70	1 – 0.75	0.075	0.70	0.05250
0 – 0.70	1 – 0.75	-0.175	0.05	-0.00875
0 – 0.70	0 – 0.75	0.525	0.05	0.02625
0 – 0.70	0 – 0.75	0.525	0.20	0.10500
<i>Sums</i>			1.00	0.1750

The distance between the correlation model and a nearby model where the Lagrange multipliers have been changed by $\Delta = (\lambda_j^p - \lambda_j^q) = (+0.01, -0.01, +0.01)$ works out to,

$$d = \sqrt{\Delta \cdot G \cdot \Delta} = .00556776$$

Employing Kullback's information measure for the distance between these same two probability distributions called p and q , we have,

$$KL(p, q) = \sum_{i=1}^n p_i \ln \left(\frac{p_i}{q_i} \right)$$

The p_i under the correlational model were $p_i = (0.70, 0.05, 0.05, 0.20)$ as dictated by parameters $\lambda_1^p = -1.38629$, $\lambda_2^p = -1.38629$, and $\lambda_3^p = 4.02535$. Subject these parameters to a slight displacement in order to create a nearby distribution. The displacement chosen was $\Delta = (+0.01, -0.01, +0.01)$. The parameters for the second probability distribution q were then $\lambda_1^q = -1.39629$, $\lambda_2^q = -1.37629$, and $\lambda_3^q = 4.01535$. Relying upon the MEP formula, these parameters result in the numerical assignment of $q_i = (0.6979, 0.0498, 0.0509, 0.2014)$.

Carrying out the required computations to find the distance between these two distributions yield,

$$\begin{aligned} KL(p, q) &= \sum_{i=1}^4 p_i \ln \left(\frac{p_i}{q_i} \right) \\ &= 0.70 \ln \left(\frac{0.70}{0.6979} \right) + \dots + 0.20 \ln \left(\frac{0.20}{0.2014} \right) \\ \sqrt{2KL(p, q)} &= .00557657 \end{aligned}$$

The Fisher information distance closely approximates this KL distance.

Exercise 28.6.5: Use the Kullback measure to examine the separation between the fair model for the die rolling scenario and any other model.

Solution to Exercise 28.6.5

For this inferential scenario, the dimension of the state space becomes $n = 6$. Or, using our new language from *Information Geometry*, the dimension of a Riemannian manifold is 6.

We identify a point q in this manifold with the probability distribution for a “fair” die. Thus, the assignments under this fair model must be,

$$q_i = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$$

This point q has a special status because all of its coordinates λ_j^q are 0.

Now consider another point p situated somewhere else in the manifold, and therefore possessing different coordinates. Refer to Chapter Nineteen, where we introduced Jaynes’s characterization of a physically biased die with $m = 2$ constraint functions to represent the effects of a misplaced center of gravity and a lengthening of one of the axes of the cube.

The coordinates λ_j^p under this particular model of a physical bias in the die were seen to be $\lambda_1^p = 0.155216$ and $\lambda_2^p = 0.149669$. The assignments under this model were found by the MEP as, (refer back to Table 19.2),

$$p_i = (0.1236, 0.1444, 0.1076, 0.1257, 0.2300, 0.2687)$$

The separation $d = \sqrt{2KL(p, q)}$ between these two points in the Riemannian manifold is calculated as,

$$\begin{aligned} KL(p, q) &= \sum_{i=1}^6 p_i \ln \left(\frac{p_i}{q_i} \right) \\ &= 0.1236 \ln \left(\frac{0.1236}{1/6} \right) + \dots + 0.2687 \ln \left(\frac{0.2687}{1/6} \right) \\ &= 0.0622 \\ \sqrt{2KL(p, q)} &= 0.352731 \end{aligned}$$

Exercise 28.6.6: Find the posterior probability of model \mathcal{M}_A versus model \mathcal{M}_C for the Jaynes die rolling scenario after one observation.

Solution to Exercise 28.6.6

Begin with the standard Bayesian formula as derived with the formal probability manipulation rules. This formula permits us to relate a posterior model probability

on the left hand side, to the likelihood, a prior model probability, and the data, on the right hand side,

$$P(M_k | \mathcal{D}) = \frac{P(\mathcal{D} | M_k) P(M_k)}{P(\mathcal{D})}$$

The ratio of the posterior probability for any two models, here models \mathcal{M}_A and \mathcal{M}_C from Chapter Nineteen, becomes,

$$\frac{P(M_A | \mathcal{D})}{P(M_C | \mathcal{D})} = \frac{P(\mathcal{D} | M_A) P(M_A)}{P(\mathcal{D} | M_C) P(M_C)}$$

Take the log of these ratios,

$$\ln \left[\frac{P(M_A | \mathcal{D})}{P(M_C | \mathcal{D})} \right] = \ln \left[\frac{P(\mathcal{D} | M_A)}{P(\mathcal{D} | M_C)} \right] + \ln \left[\frac{P(M_A)}{P(M_C)} \right]$$

Since the prior probabilities for any two models are always equal under Laplace's *Principle of Insufficient Reason*, the second term on the right hand side becomes zero,

$$\ln \left[\frac{P(M_A | \mathcal{D})}{P(M_C | \mathcal{D})} \right] = \ln \left[\frac{P(\mathcal{D} | M_A)}{P(\mathcal{D} | M_C)} \right]$$

But the data consist of only one observation, so,

$$\ln \left[\frac{P(M_A | X_1 = x_i)}{P(M_C | X_1 = x_i)} \right] = \ln \left[\frac{P(X_1 = x_i | M_A)}{P(X_1 = x_i | M_C)} \right]$$

Suppose that a SIX was observed after the first roll of the die.

$$\ln \left[\frac{P(M_A | X_1 = \text{SIX})}{P(M_C | X_1 = \text{SIX})} \right] = \ln \left[\frac{P(X_1 = \text{SIX} | M_A)}{P(X_1 = \text{SIX} | M_C)} \right]$$

Substituting on the right hand side for the numerical assignments to a SIX under each model, we have, as expected, a slight preference for model \mathcal{M}_C ,

$$\begin{aligned} \ln \left[\frac{P(M_A | X_1 = \text{SIX})}{P(M_C | X_1 = \text{SIX})} \right] &= \ln \left[\frac{1/6}{0.2687} \right] \\ &= -0.4776 \end{aligned}$$

We could have just as easily flipped the models and found that,

$$\ln \left[\frac{P(M_C | X_1 = \text{SIX})}{P(M_A | X_1 = \text{SIX})} \right] = 0.4776$$

leading to the quantitative re-shuffling of the former equal prior probabilities for the two models to a slight favoring of model \mathcal{M}_C ,

$$\frac{P(M_C | X_1 = \text{SIX})}{P(M_A | X_1 = \text{SIX})} = 1.6122$$

If instead of the SIX on the first roll of the die, a THREE had been observed, then model \mathcal{M}_A would have been slightly favored,

$$\frac{P(M_C | X_1 = \text{THREE})}{P(M_A | X_1 = \text{THREE})} = 0.6456$$

Construct Table 28.3 at the top of the next page to show the effect on model re-orientation for all possible observations.

Table 28.3: How the probabilities for two models are re-ordered after a single observation in Jaynes's die scenario.

Observation	log Ratio	Relative favoring of \mathcal{M}_C over \mathcal{M}_A
ONE	-0.2989	0.7416
TWO	-0.1434	0.8664
THREE	-0.4376	0.6456
FOUR	-0.2821	0.7542
FIVE	+0.3221	1.3800
SIX	+0.4776	1.6122

Exercise 28.6.7: Map the result from the last exercise into Kullback's information concepts.

Solution to Exercise 28.6.7

Kullback defines his information measure $I(1 : 2)$ as the *mean* information in favor of model \mathcal{M}_C against model \mathcal{M}_A per observation,

$$I(\mathcal{M}_C : \mathcal{M}_A) \equiv I(1 : 2) = \int f_1(x) \log \left[\frac{f_1(x)}{f_2(x)} \right] d\lambda(x)$$

with the explanation that the logarithm of the likelihood ratio,

$$\log \left[\frac{f_1(x)}{f_2(x)} \right] = \log \frac{P(H_1 | x)}{P(H_2 | x)} - \log \frac{P(H_1)}{P(H_2)}$$

is the information in one observation, or one data point $X_1 = x_i$, for discrimination in favor of model \mathcal{M}_C against model \mathcal{M}_A .

Carry out the following mappings,

$$f_1(x_i) \equiv p_i \equiv P(X = x_i | \mathcal{M}_C)$$

$$f_2(x_i) \equiv q_i \equiv P(X = x_i | \mathcal{M}_A)$$

$$\log \left[\frac{f_1(x)}{f_2(x)} \right] \equiv \ln \left(\frac{p_i}{q_i} \right) \equiv \ln \left[\frac{P(X_1 = x_i | \mathcal{M}_C)}{P(X_1 = x_i | \mathcal{M}_A)} \right]$$

$$\sum_{i=1}^n p_i \ln \left(\frac{p_i}{q_i} \right) \equiv \int f_1(x) \log \left[\frac{f_1(x)}{f_2(x)} \right] d\lambda(x)$$

Then, using the values for the log likelihood ratio as computed in Table 28.3, we must find concordance with the result in Exercise 28.6.5 that $KL(p, q) = 0.0622$,

$$\begin{aligned} KL(p, q) &= \sum_{i=1}^6 p_i \ln \left(\frac{p_i}{q_i} \right) \\ &= (0.1237 \times -0.2989) + \dots + (0.2687 \times 0.4776) \\ &= 0.0622 \end{aligned}$$

Exercise 28.6.8: Comment on whether these various informational concepts are consistent.

Solution to Exercise 28.6.8

It is now clear that Kullback's measure,

$$KL(p, q) \equiv I(1 : 2) = \int f_1(x) \log \left[\frac{f_1(x)}{f_2(x)} \right] d\lambda(x) \equiv \sum_{i=1}^n p_i \ln \left(\frac{p_i}{q_i} \right)$$

is a calculation carried out **only after the p_i and q_i have already been assigned by the MEP!**

In fact, the p_i and q_i are exactly the same as the numerical assignments Q_i under each model. But the variational calculus procedures have already been called upon within the MEP to find the p_i and the q_i . There is no further requirement to carry out some additional variational procedure with Kullback's measure as the objective function. This is an endemic and very serious conceptual confusion that is not at all appreciated in discussions that conflate Shannon entropy and Kullback's relative entropy.

Information, as defined by Kullback, is the re-orientation of model space under the auspices of Bayes's Theorem after one observation. This is NOT how information was defined within the MEP. Nonetheless, we gain access to Kullback's information for FREE using the formal manipulation rules of probability theory. New concepts like minimizing $KL(p, q)$ are extraneous to what is already present.

Any generalization of Shannon entropy such as we are considering here in,

$$KL(p, q) = \sum_{i=1}^n p_i \ln \left(\frac{p_i}{q_i} \right)$$

should not depend on the data because Shannon's definition did not depend on any data. But, in Kullback's presentation, everything centers around the re-orientation of model space pre and post-data acquisition, as in the recent examples of the die involving model \mathcal{M}_A and model \mathcal{M}_C ,

$$\frac{P(M_A | \mathcal{D})}{P(M_C | \mathcal{D})}$$

as more and more data are acquired.

As Kullback says [26, pg. 5],

The right hand side of (2.3), [this is]

$$\log \frac{f_1(x)}{f_2(x)} = \log \frac{P(H_1 | x)}{P(H_2 | x)} - \log \frac{P(H_1)}{P(H_2)}$$

is a measure of the difference between the logarithm of the odds in favor of H_1 after the observation $X = x$ and before the observation. [Emphasis added.]

Exercise 28.6.9: Try to dispel the fog of confusion by listing all of the various probability expressions we have encountered so far.

Solution to Exercise 28.6.9

In an attempt to recover our footing, it is necessary to go back to the beginning and recapitulate some basic concepts. The probability distribution for the joint statements involving what can be observed, and a model assigning numerical values as the *probabilities* for what can be observed is,

$$P(X = x_i, \mathcal{M}_k)$$

Both $(X = x_i)$ and \mathcal{M}_k are *statements* because the arguments to a probability must be statements.

The statements represented by $(X = x_i)$ are of the kind, “HEADS shows face up on tossed coin.”, or “THREE shows face up on a rolled die.”, or “A kangaroo prefers Foster’s and is left-handed.”

The statements represented by \mathcal{M}_k are of the kind, “A numerical value of 3/4 as assigned to HEADS and a value of 1/4 as assigned to TAILS is the correct assignment.”, or, “A numerical value of 1/6 as assigned to all faces of the die is the correct assignment.” or, “The numerical values under some proposed association between hand and beer preference is the correct assignment.”

By the formal manipulation rules of probability, namely, the **Sum Rule** and the **Product Rule**, we have first that,

$$P(X = x_i, \mathcal{M}_k) = P(X = x_i | \mathcal{M}_k) \times P(\mathcal{M}_k)$$

and secondly that,

$$P(X = x_i) = \sum_{k=1}^{\mathcal{M}} P(X = x_i | \mathcal{M}_k) \times P(\mathcal{M}_k)$$

The degree of belief that a joint statement $(X = x_i)$ and \mathcal{M}_k are both true, can be broken down into the probability that the statement $(X = x_i)$ is true given that the model \mathcal{M}_k actually is true times the degree of belief that model \mathcal{M}_k assigning the numerical values to $P(X = x_i)$ is true.

So, it could be that a joint statement in the coin tossing scenario,

$$P(\text{HEADS, FAIR MODEL}) = 1/2 \times 0.01 = 0.005$$

reflects the legitimate degree of belief that the following statement is true: “HEADS shows up and the fair model is correct.” There is that separate assessment concerning the degree of belief about the fair model being correct.

$P(X = x_i | \mathcal{M}_k)$ is, by definition, the same as the Q_i , the numerical assignment made by the MEP formula under the information in some model \mathcal{M}_k . It also is the same as the p_i and q_i as used in Kullback’s information measure. Kullback’s notation for $P(X = x_i | \mathcal{M}_k)$ is $f_i(x)$ and, most often H_i for \mathcal{M}_k .

However, $P(X = x_i | \mathcal{M}_k)$ is NOT data! The statement $(X = x_i)$ appears as the first argument in the template $P(\star | \star)$. The IP does not know $(X = x_i)$, but is assuming that \mathcal{M}_k is true. Thus, numerical values can be assigned by the information resident in \mathcal{M}_k as probabilities, or degrees of belief, that $(X = x_i)$ is true. The $(X = x_i)$ are *potential* data points whenever they have been observed or measured and placed to the right of the conditioned upon symbol, $P(\star | \mathcal{D})$.

Shannon entropy is,

$$H(Q_i) = - \sum_{i=1}^n Q_i \ln Q_i \equiv - \sum_{i=1}^n P(X = x_i | \mathcal{M}_k) \ln P(X = x_i | \mathcal{M}_k)$$

It is a quantitative measure of the amount of missing information in a probability distribution. Its definition is not in any way related to any data, that is, any already observed $(X = x_i)$.

Kullback’s information measure $I(1 : 2)$ is related directly to data as indicated by its derivation from Bayes’s Theorem, and Kullback’s own remarks repeated above that it represents how the models are favored pre and post-data.

$$I(1 : 2) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} d\lambda(x) = \int \log \frac{P(H_1 | x)}{P(H_2 | x)} d\mu_1(x) - \log \frac{P(H_1)}{P(H_2)}$$

We claimed above that the formal manipulation rules eventually do provide us with Kullback’s reliance on data as information. We have encountered probability expressions like $P(\mathcal{M}_k | \mathcal{D})$ comparable to Kullback’s $P(H_1 | x)$ where data \mathcal{D} are processed. The data appear in prediction equations like,

$$P(X_{N+1} = x_i | \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(X_{N+1} = x_i | \mathcal{M}_k) P(\mathcal{M}_k | \mathcal{D})$$

where an IP seeks a degree of belief about the next, as yet unknown, statement after having observed N data points, $\mathcal{D} = (X_1 = x_i), (X_2 = x_i), \dots, (X_N = x_i)$.

Finally, our ultimate probability expressions allow us to calculate the degree of belief about any number as yet unknown statements M_1, M_2, \dots, M_n when conditioned on already observed data N_1, N_2, \dots, N_n as,

$$P(M_1, M_2, \dots, M_n | N_1, N_2, \dots, N_n)$$

Kullback's definition as a mean *after* one data point has been acquired,

$$I(1 : 2) = \sum_{i=1}^n f_1(x) \log \frac{f_1(x)}{f_2(x)} \equiv \int \log \frac{P(H_1 | x)}{P(H_2 | x)} d\mu_1(x) - \log \frac{P(H_1)}{P(H_2)}$$

versus Shannon's definition as a mean of $-\ln Q_i$ *before* any data have been acquired,

$$H(Q_i) = - \sum_{i=1}^n Q_i \ln Q_i$$

are conceptually different. There is no escaping that fact!

Exercise 28.6.10: What is the “MEP formula” when relative entropy is used as the objective function instead of Shannon’s entropy?

Solution to Exercise 28.6.10

There is a mistaken impression rampant that Shannon’s entropy is not sufficiently general, and that one ought to use relative entropy instead when deriving the MEP algorithm. This is especially thought to be the case when dealing with continuous as opposed to discrete distributions.

Proceeding through the same kind of optimization technique that relies upon the method of undetermined multipliers, but this time seeking a minimum instead of the maximum, replace Shannon’s entropy with relative entropy to find this alternative MEP formula,

$$p_i \equiv P(X = x_i | \mathcal{M}_k) = \frac{q_i e^{\sum_{j=1}^m \lambda_j F_j(x_i)}}{\sum_{i=1}^n q_i e^{\sum_{j=1}^m \lambda_j F_j(x_i)}}$$

The p_i are from the probability distribution represented by point p , and the q_i are from the probability distribution represented by point q in the relative entropy expression,

$$KL(p, q) = \sum_{i=1}^n p_i \ln \left(\frac{p_i}{q_i} \right)$$

The q_i cause a great deal of consternation because they are thought to be some kind of baseline “measure” vitally necessary to the success of the MEP. But how are these q_i found from fundamental principles?

There is no mystery if one follows the MEP as outlined here in this Volume. Both p and q are initially assigned numerical values through the MEP. An IP wants only to compare them by finding out how far apart they are, after the data, so that the prediction for the next unobserved statement can be made by averaging over all models in model space.

Chapter 29

Relative Entropy and Correlational Models

29.1 Introduction

It is time to disambiguate some of the confusion surrounding the maximum entropy principle and relative entropy. As the reader knows by now, my arguments are anchored in numerical examples. The kangaroo scenario introduced in Chapter Twenty Two is the milieu for subjecting relative entropy to numerical scrutiny.

Chapter Twenty Two introduced the idea of how the MEP would deal with associations, relationships, or correlations between and among three traits possessed by the kangaroos. It flows immediately from our conceptual foundation that this kind of correlational information must be implemented within the models assigning numerical values to the probabilities for the statements in the state space.

It is very helpful to adopt a geometrical attitude towards the strictly symbolic formulas that arise when considering information and entropy. By adopting such a stance, relationships that otherwise might be rather difficult to accept are more easily visualized. There are some who choke on a disparaged *philosophical* rationale involving subjective terms like entropy and information, but who would more readily accept down to earth geometrical concepts like distances, angles, perpendicularity, and so on.

From this geometric vantage point, the relative entropy is viewed as a squared *distance* between two points in some manifold \mathcal{S} . Any two probability distributions p and q under this interpretation of relative entropy must first have been assigned numerical values by the MEP. Minimizing relative entropy should NOT be seen as some kind of alternative, more general, version of Shannon entropy. Relative entropy is brought into play only *after* the MEP has been invoked for both points p and q at the outset.

29.2 The Background Concepts

It is worthwhile to occasionally recapitulate the fundamental conceptual ideas that we now accept as the backbone of any inferential procedure. An IP has a degree of belief about the truth of some joint statement in a state space. This degree of belief held by the IP is quantified by a probability.

This probability must be a real number between 0 and 1, and the probability may very well include either of these two end points. The degree of belief that the first statement must be true, or that the second statement must be true, ..., or that the n^{th} statement must be true is a certainty; in other words, a probability of 1. The numerical value of a probability for any particular joint statement is determined by the information that has been injected into the distribution of the probability under the auspices of some model.

All of these notions centering around words and phrases like: *degrees of belief, probabilities, information, entropy, and models*, are intrinsically epistemological concepts. This means that they must be linked to conscious entities, generically called information processors, who are characterized as possessing a *state of knowledge*.

This epistemological state of the IP is distinct from any ontological state of the physical world. The IP can only be aware of its own state of knowledge about ontological facts. It can not be aware of the ultimate essence of the ontological facts themselves.¹

Complementary to the “foreground” of the information inserted by some model is the “background” of the resulting missing information. Whatever amount of information has been inserted into a probability distribution must have associated with it some complementary amount of missing information. This amount of missing information is quantified by information entropy.

The maximum entropy principle is an algorithm for inserting active foreground information into a probability distribution, while simultaneously maximizing the amount of background missing information. The ensuing numerical values from the MEP algorithm are therefore legitimate probabilities attached to the statements in the state space. These numbers reflect both the active foreground information as well as the passive background missing information.

An IP may have a degree of belief that a kangaroo prefers Foster's, is right-handed, and has sandy colored fur; this being one of the statements from the eight dimensional state space. The IP's degree of belief might be quantified by an assigned numerical value of $1/8$ for the probability that this statement is true. This particular assigned numerical value of $1/8$ is neither right nor wrong in some absolute sense. It is the proper assignment when the active foreground information under some model is simply that all the probabilities assigned to the statements in the state space must sum to 1. Whatever missing information there is has been maximized.

¹The Kantian position.

While the MEP must include active foreground information, it must also do something else. The passive background missing information is what has been maximized by Shannon's entropy measure. There is no other active information other than the requirement of summing to 1 reflected in the assignment of 1/8 to all eight statements. This is guaranteed by the fact that no other possible distribution can have an information entropy larger than this assignment.

The degree of belief captured by the value of 1/8 that a kangaroo prefers Foster's, is right-handed, and has sandy colored fur is neither right nor wrong. It is the correct assignment when the information under this one model is assumed correct.

Another legitimate assignment might be that a kangaroo is certain to prefer Corona, is left-handed, and has beige colored fur. In this case, the degree of belief that this statement is true must be 1. Once again, this assignment is neither right nor wrong in some absolute sense. It is the proper assignment under the information from some other model. By the way, in this situation, there is no missing information for this certain assignment and, appropriately, the information entropy quantifies that there is no missing information by returning 0.

If additional active information in the form of constraint function averages is specified under different models, then the MEP returns appropriate numerical values reflecting this information. For example, in the kangaroo scenario we looked not only at models enforcing independence, but also at models enforcing varying degrees of correlation among the listed traits.

When we examine these models in light of a geometric framework, as we propose to do in this Chapter, we gain new insights. These insights are geometrical in nature, and so we use language involving words like distance, angles, curves, shortest distance between two points, coordinates, and the like.

29.3 Probability Distributions as Points

The fundamental conceptual simplification afforded by adopting the geometrical viewpoint is that an entire probability distribution is represented simply as one point. Points do not exist in a vacuum; they must inhabit some kind of abstract space. The mathematical development has taken the differentiable manifold as the primary space of interest. We use language that says points as probability distributions live in a differentiable manifold with a coordinate system.

Notational simplicity is a side benefit. Entire probability distributions as points are labeled simply as p , q , r , and so on. The coordinate system for the manifold is a dual coordinate system. Fortunately, we are already familiar with such a dual coordinate system, or dual set of parameters, from our study of the MEP. These are the Lagrange multipliers, the λ_j , and their duals, the constraint function averages, the $\langle F_j \rangle$. Later on, these coordinates will assume a more abstract nature to correspond to the demands of differential geometry. Thus, the coordinate λ_j might be labeled as θ^j and the coordinate $\langle F_j \rangle$ as η_j .

Some additional structure has to be imposed on the manifold and its points. A local linearization at each point is operationally achieved by a *tangent space*. This tangent space can have an inner product defined on it. This is how we arrived at the Fisher information matrix as a metric tensor. If we have this additional structure of a metric $g_{rc}(p)$ at point p for the manifold holding our probability distributions, then the space is called a Riemannian manifold.

Despite all of the attendant mathematical complexity brought on board by adopting this advanced geometrical viewpoint, the intuitive motivation still remains quite powerful. That is, we see that the inferential issues of interest now become things like “distance” between points, angles, shortest distances, perpendicularity, Pythagorean relationships, projections, and the host of geometric analogies familiar to us from Euclidean geometry.

For the sake of some concrete numerical examples, these abstract points will be the probability distributions discussed in Chapter Twenty Two. Our motivation there was to see how the MEP implemented relationships among variables through correlational models. Now these points and the models underlying them will serve as non-trivial, but still computationally feasible, examples for the elementary ideas mentioned above which appear in *Information Geometry*.²

For the sake of the upcoming numerical examples, take point p as a distribution under the most complicated correlation model. This is a triple interaction model with $m = 7$ constraint functions. At the opposite end of the spectrum, take point u as the uniform distribution with $m = 0$ constraint functions where all $u_i = 1/8$.

The MEP formula under a given model provides us with everything we need to perform the computations as demanded by *Information Geometry*. That is, the MEP formula written as,

$$Q_i \equiv P(X = x_i | \mathcal{M}_k) = \frac{\exp [\sum_{j=1}^m \lambda_j F_j(X = x_i)]}{Z(\lambda_1, \lambda_2, \dots, \lambda_m)} \quad (29.1)$$

will permit us to investigate all of the various geometric relationships that happen to capture our interest.

The MEP formula could have just as easily been expressed as,

$$P(X = x_i | \mathcal{M}_k) = \exp \left[\sum_{j=1}^m \lambda_j F_j(X = x_i) - \ln Z(\lambda_1, \lambda_2, \dots, \lambda_m) \right] \quad (29.2)$$

When we want to refer to a specific distribution like point p we will write,

$$p \equiv P(x_i | \mathcal{M}_k) = \exp \left[\sum_{j=1}^m \lambda_j^p F_j(x_i) - \ln Z_p \right] \quad (29.3)$$

²We take this opportunity to remind the reader that Volume III is devoted entirely to this topic.

Since the log transforms of probability distributions are essential to relative entropy, the above expression simplifies to,

$$\ln p \equiv \ln [P(x_i | \mathcal{M}_k)] = \sum_{j=1}^m \lambda_j^p F_j(x_i) - \ln Z_p \quad (29.4)$$

For example, under the “no information” model, point u has all $\lambda_j^u = 0$. Thus,

$$\begin{aligned} \ln u_i &= \sum_{j=1}^m \lambda_j^u F_j(x_i) - \ln Z_u \\ &= -\ln Z_u \\ &= -\ln 8 \\ u_i &= \exp [-\ln 8] \\ &= 1/8 \end{aligned}$$

What then is the distance-like measure of the separation between this point u representing no relationship whatsoever among the variables, and a point p representing a very strong relationship among the variables? As mentioned, one model that was examined in Chapter Twenty Two was a model incorporating the three marginal probabilities, all three double interactions, and the single triple interaction. Thus, this model had correlations between BH , BF , HF , and BHF . The number of constraint functions necessary to implement all of this information was $m = n - 1 = 7$.

Employing the MEP algorithm, the numerical assignment of probabilities to all eight joint statements under this highly correlated model was,

$$p \equiv P(x_i | \mathcal{M}_k) = (0.40, 0.10, 0.18, 0.07, 0.05, 0.05, 0.12, 0.03)$$

The information entropy of this assignment was $H_{max}(p) = 1.75079$. It is the *maximum* value for the background missing information after inserting the active information from the full correlational model.

We now want to demonstrate the relative entropy relationship in another manner. The ubiquitous Legendre transformation, used in the implementation of an MEP algorithm to find the numerical assignments for any distribution, will make an appearance. Kullback's information measure could also be expressed as,

$$KL(p, u) = \sum_{i=1}^8 p_i \ln \left(\frac{p_i}{u_i} \right) = E_p [\ln p_i - \ln u_i] \quad (29.5)$$

Looking first at the expectation of $\ln p_i$ we see that it is merely,

$$E_p [\ln p_i] = -H_{max}(p)$$

The expectation, *still with respect to the distribution p*, of the log transform of the MEP formula for point u is,

$$\begin{aligned} E_p [\ln u_i] &= E_p \left[\sum_{j=1}^7 \lambda_j^u F_j(x_i) - \ln Z_u \right] \\ &= \sum_{j=1}^7 \lambda_j^u \langle F_j \rangle_p - \ln Z_u \end{aligned}$$

Now we are left with,

$$KL(p, u) = \ln Z_u - \sum_{i=1}^7 \lambda_j^u \langle F_j \rangle_p - H_{max}(p)$$

Since the Lagrange multipliers λ_j^u are all equal to zero under the model for point u ,

$$KL(p, u) = \ln Z_u - H_{max}(p)$$

Kullback's measure of the separation between points p and u is then,

$$KL(p, u) \equiv \sum_{i=1}^8 p_i \ln \left(\frac{p_i}{u_i} \right) = \ln n - H_{max}(p) = \ln 8 - 1.75079 = 0.32865$$

No other distribution satisfying p 's unique constraints can get closer than this to the uniform distribution u . The MEP assignment for p has the *minimum* distance to the MEP assignment for u .

What if a point q is different than the uniform distribution where all $\lambda_j^u = 0$? The relative entropy formula should still apply. Select for a numerical example our new point q and its distance from p . Suppose that this new distribution q is the distribution for an independence model.

In Chapter Twenty Two, using the MEP algorithm, we found that the numerical assignment under an independence model with $m = 3$ was,

$$Q_i \equiv P(x_i | \mathcal{M}_k) = (0.3375, 0.1125, 0.2250, 0.0750, 0.1125, 0.0375, 0.0750, 0.0250)$$

with $H_{max}(p) = 1.75079 < H_{max}(q) = 1.79768 < H_{max}(u) = 2.07944$.

The independence model incorporated information about the three marginal probabilities for beer preference, hand preference, and fur color. The assignments must change from the assignments $P(x_i | \mathcal{M}_k) = 1/8$ under the fair model.

Additionally, since more information was inserted under this model with $m = 3$ constraint functions versus the model with $m = 0$ constraint functions, the information entropy decreased from $H(u) = 2.07944$ to $H(q) = 1.79768$. Intuitively, you might expect that the Kullback relative entropy measure between p and q might decrease as well. It does indeed as the calculation,

$$KL(p, q) = \sum_{j=1}^7 (\lambda_j^p - \lambda_j^q) \langle F_j \rangle_p + \ln \left(\frac{Z_q}{Z_p} \right) = 0.04689$$

attests to. The details are carried out in Exercises 29.9.9 and 29.9.11.

If we were to calculate $KL(q, u)$ in the same way, we would find the following relationship,

$$\begin{aligned} KL(p, u) &= KL(q, u) + KL(p, q) \\ 0.32865 &= 0.28176 + 0.04689 \end{aligned}$$

As we will discover in an upcoming section, pqu forms a right triangle with the squared distance of each leg, $KL(q, u)$ and $KL(p, q)$, equal to the squared distance of the hypotenuse $KL(p, u)$.

29.4 Canonical Divergence

The Kullback measure, often interpreted as the difficulty of discriminating between two probability distributions, can then be viewed, perhaps more profitably, as the squared distance between the two points p and q ,

$$KL(p, q) = \sum_{i=1}^n p_i \ln \left(\frac{p_i}{q_i} \right) \quad (29.6)$$

As we have just demonstrated, this measure can be decomposed as,

$$KL(p, q) = \sum_{i=1}^n p_i \ln p_i - \sum_{i=1}^n p_i \ln q_i \quad (29.7)$$

The first term is the negative of Shannon's information entropy. The second term, invoking the MEP distribution for q , is,

$$\begin{aligned} \ln q_i &= \sum_{j=1}^m \lambda_j^q F_j(x_i) - \ln Z_q \\ \sum_{i=1}^n p_i \ln q_i &= E_p [\ln q_i] \\ E_p [\ln q_i] &= \sum_{j=1}^m \lambda_j^q \langle F_j \rangle_p - \ln Z_q \end{aligned}$$

Putting the two terms back together, we find that,

$$KL(p, q) = \ln Z_q - \sum_{j=1}^m \lambda_j^q \langle F_j \rangle_p - H_{max}(p) \quad (29.8)$$

The information geometry inspired language used here is that we are constructing a **canonical divergence** for the space holding the points representing our probability distributions. This space, a differentiable Riemannian manifold, has a **dual coordinate system** consisting of the parameters λ^q and $\langle F \rangle_p$, together with the **potentials** $\ln Z_q$ and $H_{max}(p)$.

There also exist dual forms for the Kullback measure,

$$KL(p, q) = -KL^*(q, p) = \sum_{i=1}^n q_i \ln \left(\frac{q_i}{p_i} \right)$$

with the following not too surprising symmetry exhibited,

$$KL^*(q, p) = \ln Z_p - \sum_{j=1}^m \lambda_j^p \langle F_j \rangle_q - H_{max}(q)$$

For a numerical example, let's look at two models, one with $m = 4$ constraints and the other with $m = 7$ constraints. Both of these models were discussed in Chapter Twenty Two. The first model was an example of a single double interaction. This double interaction was a correlational model incorporating information about an association between beer preference and hand preference. Thus, a fourth constraint function $F_4(X = x_i)$, and its associated expectation, $\langle F_4 \rangle = Q_1 + Q_3$, were necessary.

The second model implemented all three double interactions and the single triple interaction as well as the main effects. Thus, it was necessary to insert information in the form of constraint averages $\langle F_1 \rangle$ through $\langle F_7 \rangle$.

The second full interaction model with $m = 7$ inserted specific information in the form of a constraint function average of $\langle F_4 \rangle_p = 0.58$. This distribution has already been labeled as point p . The single double interaction model from the $m = 4$ class expressed the information in the association between beer and hand preference with a constraint function average of $\langle F_4 \rangle_q = 0.60$. This distribution will be our new point q .

We would expect that these two distributions shouldn't be that far apart. They do differ in the information about the strength of the interaction between beer preference and hand preference. That difference is reflected in the information in the constraint function averages for $\langle F_4 \rangle_q$ of 0.60 versus $\langle F_4 \rangle_p$ of 0.58. The relative entropy as a distance like measure between these two points should be fairly small.

We can now calculate the relative entropy, or the separation between these two distributions p and q by Equation (29.8),

$$KL(p, q) = \ln Z_q - \sum_{j=1}^4 \lambda_j^q \langle F_j \rangle_p - H_{max}(p)$$

Make the appropriate substitutions, and take care that $\langle F_4 \rangle_p = 0.58$,

$$H_{max}(p) = 1.75079$$

$$Z_q = 25$$

$$\ln Z_q = 3.2189$$

$$\begin{aligned} \sum_{j=1}^4 \lambda_j^q \langle F_j \rangle_p &= (0.405465 \times 0.75) + \cdots + (0.980829 \times 0.58) \\ &= 1.4204 \end{aligned}$$

$$\begin{aligned} KL(p, q) &= 3.2189 - 1.4204 - 1.75079 \\ &= 0.0477 \end{aligned}$$

to confirm that the separation between these two very similar models should be much smaller than the separation between disparate models like the full correlational model and the uniform distribution calculated earlier as $KL(p, u) = 0.32865$.

29.5 The Pythagorean Relationship

The best known relationship in Euclidean geometry is the one attributed to the Greek philosopher Pythagoras involving the squared distances of the sides of a right triangle. Interestingly, a similar Pythagorean relationship exists in *Information Geometry* involving the separation between the three points p , q , and r . Since we have already seen that the Kullback measure is something like a squared distance between two probability distributions, the following conjecture seems plausible when the conditions exist for a “right triangle.”

$$KL(p, r) = KL(p, q) + KL(q, r) \quad (29.9)$$

Thus, we have a “triangular relationship” among the three points p , q , and r in the Riemannian manifold. This Pythagorean relationship is a beautiful illustration of the efficacy of viewing abstractions through a geometrical lens. See Figure 29.1 at the top of the next page for a sketch of the three probability distributions as they exist as points in some manifold \mathcal{S}^n . Two of the distributions q and r live in a smaller dimension sub-manifold \mathcal{S}^m , while the distribution p is in \mathcal{S}^n .

It is easy to visualize (and this is again one of the great virtues for adopting a geometrical approach), that because of the Pythagorean relationship, the separation between the distributions p and r along the “hypotenuse” is always going to be greater than the separation between the distributions p and q as well as the separation between the distributions q and r because they are “the sides of the right

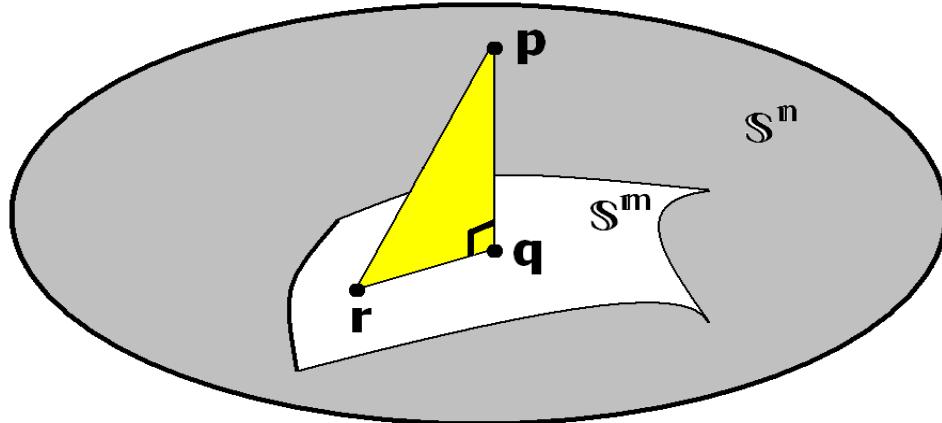


Figure 29.1: *A sketch of the information geometry characterization of the Pythagorean relationship among three points representing probability distributions.*

triangle.” Under the circumstance where point r is allowed to approach point q , the separation between p and q becomes the same as between p and r .

Then, the *minimum* distance between p and any r becomes the distance between p and q . Thus, point p is said to be **projected** onto the model space defining point q . All of the other points r as they are allowed to approach q also live in the same model space as q .

This projection to q is the closest that point p can get to simpler probability distributions in other model spaces. Let’s say that the probability distribution p incorporates the information about many interactions. Then, the phrase “simpler probability distributions” referring to q implies that the correlations among the variables involve fewer interactions.

If you are interested in the computational details involved in the Pythagorean relationship, follow Exercises 29.9.12 through 29.9.17. These exercises will serve as our numerical example of the Pythagorean relationship among three distributions p , q , and r . All three distributions implement some varying degree of correlation among the variable to predicted and the explanatory variables via the information in their respective models.

29.6 Projection of Complicated Models

Eventually, we will have to make reasonable approximations to models that are so complex that they defy our ability to compute their relevant characteristics under the MEP algorithm. Of course, this restriction doesn’t apply here with our enhanced kangaroo scenario where everything, including the most complicated correlational models, is computationally feasible.

But as we showed at the end of Volume I, trying to predict the future status of complicated ontological systems like cellular automata will very quickly tax our framework from a computational perspective. Thus, it would be very helpful if it were possible to **project** a point p representing an extremely complicated correlated model to a subspace where the models were more computationally feasible. The projection to a point q in the subspace would be the “closest” the complex model p could get to something we could contemplate working with.

We can gain an initial appreciation for what is involved here by looking at the implications of the Pythagorean relationship. So we extend the discussion to four points, p , q , r , and s inhabiting a Riemannian manifold.

The visual image is of the complicated correlational model, as represented by point p , living in a very high dimensional manifold \mathcal{S}^n . We would like to project point p down into a much smaller submanifold \mathcal{S}^m such that the distance between it, and another point q living in the submanifold is as small as possible. We should examine many other points living in the same submanifold as q . But suppose we limit it to just two other points r and s .

We can now envision looking at various triangular relationships taking three points at a time. We will first examine the triangular relationship involving p , s , and r , and then p , q , and r . The idea is that the first triangular relationship does not satisfy the Pythagorean relationship, while the second one does. Thus, we see that neither r nor s are as close to p as q is. The projection of p to q represents the best approximation of a complicated model in a higher dimensional space to a less complicated model in a lower dimensional space.

For our numerical examples, let point p once again be from the class of the most highly correlated models where $m = 7$. The points q , r , and s are models from less complicated correlational demands. Because we are already familiar with them from our previous examples, suppose that they live in the submanifold consisting of the single double interaction BH where $m = 4$. The BH interaction involved a constraint function so that the IP could construct an association involving beer and hand preference.

We want to project point p down from the $m = 7$ space to the lower subspace of $m = 4$. How well can we approximate a difficult distribution like p by more tractable distributions like q , r , and s ? What projection of p down to the subspace results in a distribution that has minimum separation from p ?

Write out the putative Pythagorean relationship for the three distributions p , r , and s ,

$$KL(p, r) = KL(p, s) + KL(s, r)$$

This relationship will hold, as our intuition from Euclidean geometry informs us, only if the triangle is a right triangle. In other words, the curve connecting p and s must be perpendicular. However, there is no guarantee that the three points p , r , and s we have arbitrarily picked out will have this feature.

We can calculate all three distances as relative entropies and look at,

$$KL(p, s) + KL(s, r) - KL(p, r)$$

to get some idea of the discrepancy of the three points p , r , and s from the Pythagorean relationship. If this calculation were to yield 0, then the relationship holds. But most likely, unless we have been extremely lucky in our arbitrary selection, there will be some departure from 0. This is a reflection of the **Law of Cosines** in ordinary geometry.

The exercises will go into all of the details, so we will concentrate on the larger picture here. The distribution p is the result of a model inserting information about all three marginal probabilities, all three double interactions, and a final triple interaction. The constraint function average for the BH interaction specified by the model for p was $\langle F_4 \rangle_p = 0.58$. Of course, constraint function averages were also specified for the remaining two double interactions BF and HF , as well as the triple interaction BHF for distribution p .

All three distributions q , r , and s reflect only the information for one double interaction BH . That is, only four constraint function averages were specified under these models. Select the models $\langle F_4 \rangle_q = 0.58$, $\langle F_4 \rangle_r = 0.56$, and $\langle F_4 \rangle_s = 0.60$ for the example. The first three constraint function averages remain the same for all models.

We might very well expect relatively large distances between p and s , between p and q , and between p and r as the relative entropy calculations reflect the separation between the highly correlated distribution p with $m = 7$ constraints and less correlated distributions with $m = 4$ constraints. On the other hand, the distances between s and q and between q and r , should be somewhat smaller as these are all models with correlation information differing only slightly on the single double interaction between beer and hand preference.

The triangle psr formed by these three distributions is more like an isosceles triangle with the two sides ps and pr almost equal. It clearly is not a right triangle and,

$$KL(p, s) + KL(s, r) - KL(p, r) \neq 0$$

The triangle formed by points p , s , and r is not characterized by the required perpendicularity.

What we have learned from this is that there must be another point q living in the submanifold of dimension $m = 4$ that is even closer to p than s and r . The discrepancy from perpendicularity for point q would then be 0, indicating that we can now fulfill the conditions for the Pythagorean relationship. The minimum distance from point p in the higher dimensional manifold to the submanifold would be the relative entropy between distribution p and distribution q . The point q in the submanifold S^4 would be the **projection** of point p in S^8 .

Look at a diagram of the triangular relationship among the four points p , q , r , and s in Figure 29.2. It would be a good guess that point q lies somewhere between points s and r if it is to be the perpendicular projection of point p to the same manifold that contains r and s .

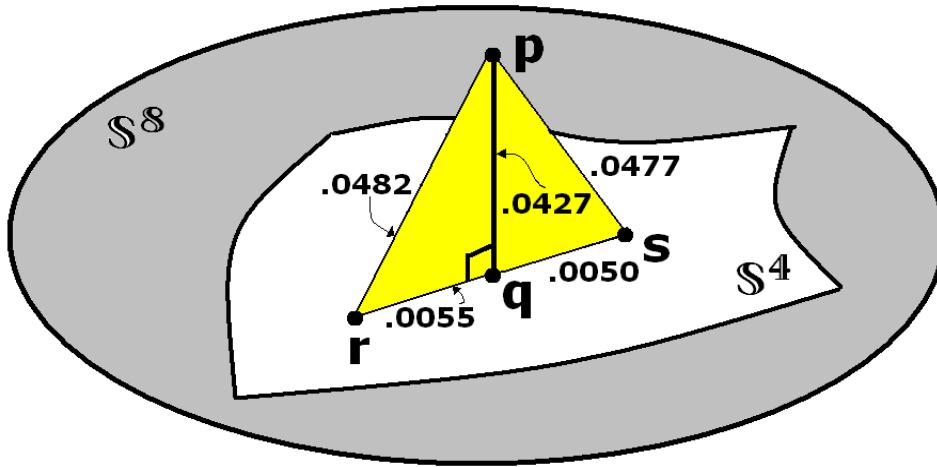


Figure 29.2: A sketch of the “triangular” relationships among the four points p , q , r and s representing probability distributions. The distribution p in S^8 is projected onto point q in S^4 .

In fact, the point q that satisfies the requirement of being the projection of p to the submanifold has the *minimum* distance from p of $KL(p, q) = 0.0427$. If the triangle pqr is a right triangle, then the squared distance between point q and point r must be,

$$\begin{aligned} KL(p, r) &= KL(p, q) + KL(q, r) \\ KL(q, r) &= 0.0482 - 0.0427 = 0.0055 \end{aligned}$$

for the Pythagorean relationship to hold.

Point q must also be a model with $m = 4$ constraint function averages, with the *BH* double interaction $\langle F_4 \rangle_q$ at some feasible value. It will perhaps come as no surprise that the best approximation to a complicated correlational model like p , when that approximation must live in the submanifold characterized solely by beer and hand preference correlations, is the distribution q with $\langle F_4 \rangle_q = 0.58$.

This constraint function average falls exactly between the constraint function averages for s and r . The minimum separation between p and a point living in a submanifold occurs where the constraint function average of the model in the submanifold matches the constraint function average of p .

29.7 What About the Data?

Where do the data gathered about the kangaroos enter into the conversation about relative entropy? So far, in calculating these geometrical properties between the various distributions in the kangaroo scenario, we have not once had recourse to the data contained in the contingency table mentioned in Chapter Twenty Two. That is because the data are not germane in implementing any of these concepts.

The relative entropy is a measure of the “distance” separating two distributions if we are thinking in geometrical terms. It can also be thought of as the relative “difficulty” in discriminating between any two models when based on an observation. But all of this is **conceptually orthogonal** to how data affect an inference.

That is not to say that the data are not important. In fact, they are extremely important. They allow the IP to update its state of knowledge about the relative status of all the models.

We also have this expression for the relative entropy between points p and q (section 28.3.4 and Exercise 29.9.9),

$$KL(p, q) = \sum_{i=1}^8 p_i \ln \left(\frac{p_i}{q_i} \right) = \sum_{j=1}^7 (\lambda_j^p - \lambda_j^q) \langle F_j \rangle_p + \ln \left(\frac{Z_q}{Z_p} \right) \quad (29.10)$$

We recognize this as very similar to the formula for the ratio of the probability of any two models after having observed some data,

$$\ln \left[\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} \right] = N \left[\sum_{j=1}^m (\lambda_j^A - \lambda_j^B) \bar{F}_j(x_i) + \ln \left(\frac{Z_B}{Z_A} \right) \right] \quad (29.11)$$

This was discussed in connection with Jaynes’s explanation of maximum likelihood in Chapter Twenty Seven. It is enlightening to comment again on the intimate relationship between Kullback’s derivation of his distance measure and fundamental probability manipulations followed by insertion of MEP assignments.

The log of the ratio of any two models conditioned on the observed data is,

$$\ln \left[\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} \right] = \ln \left[\frac{P(\mathcal{D} | \mathcal{M}_A)}{P(\mathcal{D} | \mathcal{M}_B)} \right] + \ln \left[\frac{P(\mathcal{M}_A)}{P(\mathcal{M}_B)} \right]$$

This formula was derived from the formal manipulation rules, and has nothing to do with the MEP or relative entropy.

Furthermore, if the arguments defending Laplace’s view on treating the relative standing of causes prior to any data have any merit, then the ratio of any two models prior to the data must be 1. We are then left with,

$$\ln \left[\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} \right] = \ln \left[\frac{P(\mathcal{D} | \mathcal{M}_A)}{P(\mathcal{D} | \mathcal{M}_B)} \right]$$

Make the association that model \mathcal{M}_A determines point p and model \mathcal{M}_B determines point q . Then,

$$\begin{aligned} \ln \left[\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} \right] &= \ln [P(\mathcal{M}_A | \mathcal{D})] - \ln [P(\mathcal{M}_B | \mathcal{D})] \\ \ln [P(\mathcal{M}_A | \mathcal{D})] &= \ln W(N) + \sum_{i=1}^n N_i \ln p_i \\ \ln [P(\mathcal{M}_B | \mathcal{D})] &= \ln W(N) + \sum_{i=1}^n N_i \ln q_i \\ \ln [P(\mathcal{M}_A | \mathcal{D})] - \ln [P(\mathcal{M}_B | \mathcal{D})] &= \sum_{i=1}^n N_i (\ln p_i - \ln q_i) \\ &= N \left[\sum_{i=1}^n \left(\frac{N_i}{N} \right) \ln \left(\frac{p_i}{q_i} \right) \right] \end{aligned}$$

If the numerical assignments under the model for point p were to exactly match the normed frequency counts N_i/N , we would have the result that the ratio of the probabilities for the two models conditioned on the data is a function of the relative entropy. If model \mathcal{M}_A is matched up with the maximum likelihood estimate model N_i/N and p , and model \mathcal{M}_B is matched up with q , then

$$\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} = \exp [N \times KL(p, q)] \quad (29.12)$$

I suspect though, that in this case, one would want to emphasize, as Kullback did, the role of the data in re-orienting the model space by looking at it from the Bayesian perspective as opposed to looking at it from the somewhat more abstract geometrical approach. We have achieved what I earlier labeled, somewhat flippantly, as a “reconciliation” of the conceptual error conflating data with information.

And this is precisely what we have done in the previous section where p captured all of the interactions within a complicated model. The seven constraint function averages, $\langle F_1 \rangle = 0.75$, $\langle F_2 \rangle = 0.75$, \dots , $\langle F_7 \rangle = 0.40$, all matched the data in the contingency table. Thus, whether we were thinking of p as a point separated by some “distance” from another point q as measured by the Kullback relative entropy, or whether we were thinking of the re-orientation of the posterior probabilities for two models from the Bayesian perspective after some data, computationally we were doing the same thing.

29.7.1 Maximum likelihood estimates

In the conventional statistical approach, the notion of finding “maximum likelihood estimates” of parameters is paramount. Surprisingly, no such notion is required for any inference. It is a completely superfluous concept. No “estimates” of parameters are ever required. In fact, parameters are never estimated, models are never estimated, and probabilities are never estimated.

As a consequence, the word “estimate” is banished from our vocabulary, along with words like “random,” “random variables,” “measurable spaces” together with all of the other unnecessary jargon that accompanies the conventional approach. Later on, we will discuss all of the superfluous concepts and accompanying language that have grown up like weeds to choke the flower of inference.

But it is vital to our complete understanding to discern the origin of all such concepts. This effort is aided by linking it to explanations invoking the MEP. The log likelihood of some set of data is,

$$\ln [P(\mathcal{D} | \mathcal{M}_k)] = \ln W(N) + N \left(\sum_{j=1}^m \lambda_j \bar{F}_j - \ln Z \right)$$

The maximum likelihood estimate of a parameter λ_j is conventionally defined as,

$$\frac{\partial \ln [P(\mathcal{D} | \mathcal{M}_k)]}{\partial \lambda_j} = 0$$

With this definition, we have,

$$\begin{aligned} \frac{\partial \ln [P(\mathcal{D} | \mathcal{M}_k)]}{\partial \lambda_j} &= \frac{\partial \ln [W(N) + N(\sum_{j=1}^m \lambda_j \bar{F}_j - \ln Z)]}{\partial \lambda_j} \\ &= N \left[\bar{F}_j - \frac{\partial \ln Z}{\partial \lambda_j} \right] \\ &= N\bar{F}_j - N\langle F_j \rangle \\ N(\bar{F}_j - \langle F_j \rangle) &= 0 \\ \bar{F}_j &= \langle F_j \rangle \end{aligned}$$

When the normed frequency counts constituting the data are treated just like any another model assigning numerical values to probabilities for the statements in the state space, then the sample averages for the constraint functions from the data are thought of in exactly the same manner as the expected averages from some model. This is how a “maximum likelihood estimate” is perceived from the MEP perspective.

29.8 Connections to the Literature

I won't take the trouble of mentioning the various conceptual tracks that have significantly influenced my views of *Information Geometry*. Since Volume III is given over entirely to a more in-depth discussion of all these issues that have been perfunctorily raised here, I will wait until then for my attributions.

I will have plenty of opportunity there to trace my own personal acknowledgments, and to elaborate upon the points of agreement and disagreement. Nonetheless, I should mention my basic reliance upon the long-term contributions made by the Japanese scientist Shun-ichi Amari [1] on explicating the relationship between information geometry and inference.

One minor notational item worth mentioning is that the relative entropy between two distributions p and q is sometimes called the *Kullback–Leibler measure* after its two originators. Hence the pervasive notation $KL(p, q)$ for the relative entropy. This work has an early attribution date of 1951 concurrent with Jaynes's original thoughts on the matter, but well before Jaynes's original publication.

It is a curious footnote to this history that it was Fisher's acolyte, C.R. Rao, who was apparently the first to remark upon the connections of Fisher's information measure (1922) to concepts in *Differential Geometry* (1945). Thus, the priority for the profound geometrical insights into the MEP and the Bayesian updating of models belongs not to the Bayesian branch of the family tree, but rather to the Fisherian branch!

29.9 Solved Exercises for Chapter Twenty Nine

Exercise 29.9.1: What is the largest possible Kullback distance between two distributions?

Solution to Exercise 29.9.1

If a distribution p is a distribution with no missing information, then its Shannon entropy is $H_{max}(p) = 0$. If a distribution u is a distribution with the largest possible amount of missing information, then its Shannon entropy is $H_{max}(u) = \ln n$.

The largest possible separation is the separation between the distribution with the most missing information and the distribution with no missing information. This is equal to $\Delta H = H_{max}(u) - H_{max}(p) = \ln n - 0 = \ln n$.

The Kullback measure of relative entropy, or separation, or discriminability, between these two distributions p and u is,

$$KL(p, u) = \ln Z_u - \sum_{j=1}^m \lambda_j^u \langle F_j \rangle_p - H_{max}(p)$$

A distribution p with no missing information is a distribution which is certain about one of the statements in the state space. The numerical value of 1 is assigned as the probability to one of the statements, while the remaining $n - 1$ statements have a probability of 0. The Shannon entropy of such a distribution is,

$$H_{max}(p) = -\sum_{i=1}^n Q_i \ln Q_i = -[(0 \ln 0) + (0 \ln 0) + \cdots + (1 \ln 1) + \cdots (0 \ln 0)] = 0$$

A distribution u with the maximum amount of missing information assigns a numerical value of $1/n$ to each of the n statements in the state space. The Shannon entropy of such a distribution is,

$$\begin{aligned} H_{max}(u) &= -\sum_{i=1}^n Q_i \ln Q_i \\ &= -\left[\frac{1}{n} \ln \left(\frac{1}{n} \right) + \cdots + \frac{1}{n} \ln \left(\frac{1}{n} \right) \right] \\ &= -\sum_{i=1}^n \left[\frac{1}{n} \times (\ln 1 - \ln n) \right] \\ &= -\sum_{i=1}^n \left[\frac{1}{n} \times (-\ln n) \right] \\ &= -\sum_{i=1}^n -\frac{\ln n}{n} \\ &= \ln n \end{aligned}$$

In order for the numerator of each Q_i to equal 1, the Lagrange multipliers λ_j^u must all equal 0. Of course, the partition function for this distribution u with the most amount of missing information is $Z_u = n$ and $\ln Z_u = \ln n$.

Thus, the relative entropy between p and u is,

$$\begin{aligned} KL(p, u) &= \ln Z_u - \sum_{j=1}^m \lambda_j^u \langle F_j \rangle_p - H_{\max}(p) \\ &= \ln n - \sum_{j=1}^m (0 \times \langle F_j \rangle_p) - H_{\max}(p) \\ H_{\max}(p) &= 0 \\ KL(p, u) &= \ln n \end{aligned}$$

In this case, it follows that $KL(p, u) = H_{\max}(u) = \ln n$.

Exercise 29.9.2: Take the distribution u in the above exercise and discuss what happens as the first Lagrange multiplier is allowed to increase or decrease away from its original value of 0.

Solution to Exercise 29.9.2

The numerical assignment to the probabilities for all eight joint statements in the kangaroo state space is 1/8 under the model for point u . And in order to achieve this assignment, the MEP formula had to set all of the model parameters, the Lagrange multipliers λ_j^u , equal to 0. Establish a new point q with coordinates λ_j^q .

As the first Lagrange multiplier λ_1^q is allowed to move slowly away from 0, while the remaining Lagrange multipliers remain fixed at 0, the numerical assignment to the probabilities of the joint statements must change from the uniform assignment of 1/8. Before this change, the distribution u could legitimately be said to contain “no information.” However, now that λ_1^q is taking on non-zero values, the constraint function $F_1(x_i)$ is being enforced. The marginal probability for beer preference $\langle F_1 \rangle$ is being forced to change from its former value of 1/2 that it had under u .

Information is being inserted into a new distribution q because the dual parameter of the model, the constraint function average $\langle F_1 \rangle$, is changing as the parameter of the model λ_1^q is changing. In the evocative geometrical language, we imagine that as λ_1^q is changing, it is tracing out a “coordinate curve” in the manifold.

The curve is starting out at point q and moving through the manifold along the curve to new points q^* . The constituents of this curve are themselves all probability distributions, that is, they are all points living in the same n -dimensional manifold. Only the first coordinate is changing; all of the other coordinates are fixed at 0.

Change the model to $\lambda_1^q = 0.10$. All of the numerical assignments to the probabilities of the joint statements change. The first four Q_i change to 0.131245 from their previous value of 0.1250 when $\lambda_1^q = 0$. The constraint function average changes to $\langle F_1 \rangle = 0.524979$ from 1/2.

Change the model to $\lambda_1^q = 0.20$. All of the numerical assignments to the probabilities of the joint statements change. The first four Q_i change to 0.137458 from their previous value of 0.131245 when $\lambda_1^q = 0.10$. The constraint function average changes to $\langle F_1 \rangle = 0.549834$ from 0.524979.

Change the model to $\lambda_1^q = 0.30$. All of the numerical assignments to the probabilities of the joint statements change. The first four Q_i change to 0.143611 from their previous value of 0.137458 when $\lambda_1^q = 0.20$. The constraint function average changes to $\langle F_1 \rangle = 0.574443$ from 0.549834.

You can see where this is heading. As the first Lagrange multiplier increases in value, it enforces a higher marginal probability for beer preference. By the time the curve has advanced all the way to, say, $\lambda_1^q = 4.0$, the first four $Q_i = 0.245533$. The constraint function average is all the way up to $\langle F_1 \rangle = 0.982014$. More and more “information” is being inserted into these new q^* with the complementary result that more and more missing information is being removed from each subsequent q^* .

Is this true? Check the Shannon information entropy for each new point along the curve. If more and more information is being inserted by each new point along the curve, and concomitantly there is less and less missing information at each new point, then Shannon’s information entropy should mirror this fact.

At our starting point of $\lambda_1^q = 0$, the information entropy is at our already thoroughly discussed absolute maximum value of $H(q) = \ln 8 = 2.07944$. As we start moving out along the curve, $\lambda_1^q = 0.10$, the information entropy begins dropping to $H(q^*) = 2.07819$. At the next point along the curve $\lambda_1^q = 0.20$, the information entropy continues to decrease with $H(q^*) = 2.07447$. At our ending point where the coordinate is $\lambda_1^q = 4.0$, the information entropy has dropped all the way down to $H(q^*) = 1.47639$. More and more information is definitely going into these points as we progress along the curve because our quantitative measure of the missing information is dropping all along the path.

Generally, the same thing happens if we let the first coordinate of the curve proceed in the negative direction from $\lambda_1^q = 0$ to $\lambda_1^q = -0.10$ to $\lambda_1^q = -0.20$, and so on. The first constraint function average inserting information about the marginal probability for beer preference moves from 1/2 down to lower values like 0.4750, 0.4502, and 0.0180 at $\lambda_1^q = -4.0$.

But there is an even more interesting geometrical relationship going on here. Now, consider that the fixed point p is the most complicated correlational model with information from all $m = 7$ constraint functions entering into the numerical assignments. The distance between p and any point along the q^* curve can be computed.

At first, the distance between p and the starting uniform distribution u is relatively great, but as we progress along the curve the distance between p and the new q^* begins to decrease. It reaches a minimum distance at some q^* along the curve, but then begins to increase once again. We have no trouble visualizing this geometrically. From our exposure to Euclidean geometry, we suspect that this minimum distance occurs where p is perpendicular to the q^* curve.

We will verify this numerically using our current example. We have already solved the problem for the starting point u where we found that $KL(p, u) = 0.3287$ (end of section 29.3). If the relative entropy is calculated from point p to several points along the curve as λ_1^q is increased, we find that the distance from p immediately starts decreasing with $KL(p, q^*) = 0.3049$ at $\lambda_1^q = 0.10$, $KL(p, q^*) = 0.2836$ at $\lambda_1^q = 0.20$, and so on until we reach a certain coordinate value where the distance is at a minimum. Then, as the curve continues on past this special coordinate value with ever increasing coordinate values, the distance between p and the q^* starts increasing again.

What is this special value of the Lagrange multiplier that results in the minimum separation between p and all the points q^* along the curve? It turns out to be $\lambda_1^q = 1.098612$ with the minimum distance of $KL(p, q_{Min}^*) = 0.1978$. The numerical assignment to the probabilities for the joint statements under this particular model result in a constraint function average of $\langle F_1 \rangle = 0.75$. The actual assignments are,

$$q_{Min}^* = \{0.1875, 0.1875, 0.1875, 0.1875, 0.0625, 0.0625, 0.0625, 0.0625\}$$

or,

$$q_{Min}^* = \left\{ \frac{3}{16}, \frac{3}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16} \right\}$$

As must be true for any MEP solution, these assignments satisfy all of the information in the model. They satisfy first and foremost the universal constraint of summing to 1. Next, they satisfy the information that the marginal probability for beer preference, $Q_1 + Q_2 + Q_3 + Q_4$, is,

$$P(BHF | \mathcal{M}_k) + P(B\overline{H}F | \mathcal{M}_k) + P(BH\overline{F} | \mathcal{M}_k) + P(B\overline{H}\overline{F} | \mathcal{M}_k) = 0.75$$

There is no other information under this model. We are assured of this fact through our use of the MEP algorithm. The quantitative measure of the amount of missing information is $H(q_{Min}^*) = 1.9486$. Any other distribution which might be able to satisfy these constraints and has a different numerical assignment must have an entropy lower than 1.9486. Therefore, this competing distribution has a lower amount of missing information. It must have inserted more active information than the model for q_{Min}^* .

This point q_{Min}^* , which is closest to point p , is seen to match point p 's constraint function average $\langle F_1 \rangle_p = 0.75$. This is true in general.

Exercise 29.9.3: Engage in a similar discussion about a distribution r that allows its second parameter to vary. The curve for r intersects the point q_{Min}^* .

Solution to Exercise 29.9.3

If we allow the second Lagrange multiplier to vary from 0 while keeping all the other parameters fixed at 0, then we would expect parallel results to those just found above. By varying the second parameter of any model, information is being inserted into distributions about the marginal probability for hand preference as opposed to beer preference. If we were manipulating the third parameter of any model, information would be inserted about the marginal probability for fur color.

Sure enough, as the second parameter starts to move away from 0, the numerical assignments under these new models start to change. We can observe these changes in the second constraint function average which started out at $\langle F_2 \rangle = 1/2$. Under a new model with $\lambda_2^r = 0.10$, $\langle F_2 \rangle = 0.524979$. Under a new model with $\lambda_2^r = 0.20$, $\langle F_2 \rangle = 0.549834$, and so on, just as before.

The distance from p to the various points r^* along the r curve are also the same as before, reaching a minimum distance of $KL(p, r_{Min}^*) = 0.1978$ at the model where $\lambda_2^r = 1.098612$. The constraint function average at r_{Min}^* is $\langle F_2 \rangle = 0.75$ matching the second constraint function average for p .

But even though the distance from p to r_{Min}^* is exactly the same as the distance from p to q_{Min}^* , the two distributions r_{Min}^* and q_{Min}^* are two different distributions. Therefore, there must be a separation between them. It is,

$$KL(q_{Min}^*, r_{Min}^*) = 0.2747$$

To resolve this apparent paradox, think of a globe of the earth with two lines of longitude running down from the North Pole and stopping at the equator. Both distances are the same minimum distances (great circle distances) and intersect the equator at right angles, but these two points on the equator might be separated by as much as 180 degrees at opposite points on the earth.

Exercise 29.9.4: Provide more details on the calculation of the distance between the two minimum distance points from distribution p .

Solution to Exercise 29.9.4

The two points q_{Min}^* and r_{Min}^* are themselves numerical assignments to the probabilities for the eight statements in the state space. They are two distinct models, giving rise to two distinct numerical assignments to the probabilities for the statements in the state space, $P(X = x_i | \mathcal{M}_{q_{Min}^*})$ and $P(X = x_i | \mathcal{M}_{r_{Min}^*})$.

$$P(X = x_i | \mathcal{M}_{q_{Min}^*}) = \{3/16, 3/16, 3/16, 3/16, 1/16, 1/16, 1/16, 1/16\}$$

$$P(X = x_i | \mathcal{M}_{r_{Min}^*}) = \{3/16, 1/16, 3/16, 1/16, 3/16, 1/16, 3/16, 1/16\}$$

The separation between them, and there must be a separation since they make different assignments, is given a quantitative form by the Kullback measure,

$$KL(q_{Min}^*, r_{Min}^*) = \sum_{i=1}^n q_{Min}^* \ln \left(\frac{q_{Min}^*}{r_{Min}^*} \right)$$

This formula is reworked into,

$$\begin{aligned} KL(q_{Min}^*, r_{Min}^*) &= E_{q_{Min}^*} [\ln q_{Min}^*] - E_{q_{Min}^*} [\ln r_{Min}^*] \\ &= -H(q_{Min}^*) + \ln(Z_{r_{Min}^*}) - E_{q_{Min}^*} \left[\sum_{j=1}^4 \lambda_j^r F_j(x_i) \right] \end{aligned}$$

Focus attention on the final term $E_{q_{Min}^*} [\sum_{j=1}^4 \lambda_j^r F_j(x_i)]$. The parameters for this model $\mathcal{M}_{r_{Min}^*} \lambda_j^r$, are all 0 except for $\lambda_2^r = 1.098612$. The expectation for the second constraint function $F_2(x_i)$ is taken with respect to the distribution of q_{Min}^* , that point on the curve for q where it reached the minimum distance from p . The expectation $E_{q_{Min}^*} [F_2(x_i)]$ is then explicitly,

$$\begin{aligned} E_{q_{Min}^*} [F_2(x_i)] &= \\ &(1 \times 3/16) + (0 \times 3/16) + (1 \times 3/16) + (0 \times 3/16) + \\ &(1 \times 1/16) + (0 \times 1/16) + (1 \times 1/16) + (0 \times 1/16) \\ &= 1/2 \end{aligned}$$

The calculation can now be completed,

$$\begin{aligned} KL(q_{Min}^*, r_{Min}^*) &= E_{q_{Min}^*} [\ln q_{Min}^*] - E_{q_{Min}^*} [\ln r_{Min}^*] \\ &= -H(q_{Min}^*) + \ln(Z_{r_{Min}^*}) - (1.098612 \times 0.50) \\ &= -1.94863 + \ln 16 - 0.54931 \\ &= 0.2747 \end{aligned}$$

Exercise 29.9.5: Verify a couple of the numerical assignments under model $\mathcal{M}_{q_{Min}^*}$.

Solution to Exercise 29.9.5

Select the fourth joint statement in the kangaroo state space,

$B\overline{H}\overline{F} \equiv$ “The kangaroo prefers to drink Foster’s beer, is left-handed, and has beige colored fur.”

The probability for this statement under the requested model is calculated by the MEP formula as,

$$P(X = x_4 | \mathcal{M}_{q_{Min}^*}) = \frac{\exp [\sum_{j=1}^m \lambda_j^q F_j(x_4)]}{Z(\lambda_1, \dots, \lambda_7)}$$

On the q curve all of the coordinates λ_j^q are equal to 0 except for λ_1^q . At the point on the curve where q_{Min}^* achieves the minimum distance from p , the coordinate is $\lambda_1^q = 1.098612$.

Thus,

$$\begin{aligned} P(X = x_4 | \mathcal{M}_{q_{Min}^*}) &= \frac{\exp [\lambda_1^q F_1(x_4)]}{Z(\lambda_1, \dots, \lambda_7)} \\ &= \frac{\exp [1.098612 \times 1]}{Z(\lambda_1, \dots, \lambda_7)} \\ &= \frac{3}{Z(\lambda_1, \dots, \lambda_7)} \end{aligned}$$

Since we know that the first constraint function was defined as,

$$F_1(x_i) = (1, 1, 1, 1, 0, 0, 0, 0)$$

and only λ_1^q is different than 0, the sum in the partition function must be,

$$\sum_{i=1}^8 \exp [\lambda_1^q F_1(x_i)] = \exp [1.098612 \times 1] + \dots + \exp [1.098612 \times 0] = (4 \times 3) + (4 \times 1) = 16$$

Finally, we see that the assigned numerical value to the probability for the fourth joint statement in the kangaroo state space under this particular model is,

$$P(X = x_4 | \mathcal{M}_{q_{Min}^*}) = 3/16$$

Using the same argument, the probability for fifth joint statement,

$\overline{BHF} \equiv$ “The kangaroo prefers to drink Corona beer, is right-handed, and has sandy colored fur.”

is,

$$P(X = x_5 | \mathcal{M}_{q_{Min}^*}) = 1/16$$

Exercise 29.9.6: What is the distance from the point u representing the distribution with the most missing information to arbitrary points varying the final constraint function average?

Solution to Exercise 29.9.6

We can rattle off the features for point u in our sleep. If it is to be the distribution with the maximum amount of missing information, then the model must set all of its Lagrange multipliers equal to 0. The numerical assignment is $1/n$ with a maximum information entropy of $\ln n$. Given our current enhanced kangaroo scenario, the distribution $u \equiv (1/8, \dots, 1/8)$ with an information entropy of $\ln 8 = 2.0794$.

Now consider four other points, or distributions, p , q , r , and s , whose models differ only in the information inserted by the final constraint function average. The seventh constraint function for the enhanced kangaroo scenario was defined as the vector,

$$F_7(x_i) = (1, 0, 0, 0, 0, 0, 0, 0)$$

This was the way we implemented a triple interaction involving BHF , that is, correlational models involving the specific setting for Q_1 . The constraint function average for $\langle F_7 \rangle$ then became simply Q_1 .

Point p 's model specified $\langle F_7 \rangle_p = 0.40$ with the familiar MEP assignment,

$$p = (0.40, 0.10, 0.18, 0.07, 0.05, 0.05, 0.12, 0.03)$$

If different models were to vary the dual parameter $\langle F_7 \rangle$, while keeping the remaining parameters fixed, we could arrive at, say, three other points for comparison. Let point q be the distribution that results from setting $\langle F_7 \rangle_q = 0.42$. Let point r be the distribution that results from setting $\langle F_7 \rangle_r = 0.38$. Finally, let point s be the distribution that results from setting $\langle F_7 \rangle_s = 0.384$. We will see why this model is mentioned in just a moment.

The Kullback distance measure, or relative entropy, between each of these points and u is calculated according to the formula,

$$\begin{aligned} KL(p, u) &= \sum_{i=1}^n p_i \ln \left(\frac{p_i}{u_i} \right) \\ &= \ln n - H_{max}(p) \end{aligned}$$

leading to an easy template,

$$KL(\star, u) = \ln n - H_{max}(\star)$$

Thus, the distance from u for each of the points p , q , r , and s , with the aforementioned information in the constraint function average $\langle F_7 \rangle$ changing for each point, while all of the other constraint function averages are kept the same,

$$KL(p, u) = \ln 8 - H_{max}(p)$$

$$= 0.3287$$

$$KL(q, u) = \ln 8 - H_{max}(q)$$

$$= 0.3905$$

$$KL(r, u) = \ln 8 - H_{max}(r)$$

$$= 0.3143$$

$$KL(s, u) = \ln 8 - H_{max}(s)$$

$$= 0.3136$$

All of these distributions are some distance away from the uniform distribution as they must be. They all implement highly correlated models with lots of information, or, in other words, models with very little missing information in comparison to this very special model u with no information of any kind. Point u , of course, is the consequence from the model with the maximum amount of missing information.

Point q is moving farther away from u . But point r is closer in comparison to point p . The minimum distance is achieved by point s with Q_1 assigned as 0.384. If we want to maintain all of the other information unchanged, this point s with a model specifying the changed last constraint function average, is as close as we can get to the uniform distribution.

Exercise 29.9.7: What happens in general to the Lagrange multipliers when just one constraint function average is changed?

Solution to Exercise 29.9.7

Take the previous exercise as an example. Even though just one constraint function average, $\langle F_7 \rangle$, was changing, λ_7 was not the only Lagrange multiplier that changed. All of the λ_j parameters were changing in response to whatever changes were made to $\langle F_7 \rangle$.

Exercise 29.9.8: Calculate the distance between s of section 29.6 and p . Show that it is not the same as the distance between p and s .

Solution to Exercise 29.9.8

Using Equation (29.8), the squared distance between s and p is,

$$KL(s, p) = \sum_{i=1}^n s_i \ln \left(\frac{s_i}{p_i} \right) = \ln Z_p - \sum_{j=1}^7 \lambda_j^p \langle F_j \rangle_s - H_{max}(s)$$

The calculation carried out in section 29.6 found that the squared distance between p and s was $KL(p, s) = 0.0477$.

The distribution p reflecting the correlational model with $\langle F_4 \rangle_p = 0.58$ had a partition function of $Z_p(\lambda_1, \dots, \lambda_7) = 33.33$. Through the MEP algorithm, the seven Lagrange multipliers for distribution p were found to have the values,

$$\lambda_1^p = 0.847298$$

$$\lambda_2^p = 1.386290$$

$$\lambda_3^p = 0.510826$$

$$\lambda_4^p = -0.441833$$

$$\lambda_5^p = -0.154151$$

$$\lambda_6^p = -1.386290$$

$$\lambda_7^p = 1.828130$$

The constraint function averages for distribution s were the same as distribution p for the first three constraint functions, that is, the information about the marginal probabilities for the three traits were the same for the two models. However, one of the things that made them different models was the fact that the information about the correlation between beer preference and hand preference was $\langle F_4 \rangle_s = 0.60$ for point s as opposed to $\langle F_4 \rangle_p = 0.58$ for point p .

When the information entropy was found for distribution s , again through the auspices of the MEP algorithm, it was $H_{max}(s) = 1.7789$. No other distribution like s could possess an information entropy this large because, if it did, it would by definition possess more missing information. No distribution can have more missing information than the maximum entropy distribution.

Thus,

$$\begin{aligned}
 KL(s, p) &= \ln Z_p - \sum_{j=1}^7 \lambda_j^p \langle F_j \rangle_s - H_{max}(s) \\
 &= \ln 33.333 - [(0.847298 \times 0.75) + \dots + (1.82813 \times 0.36)] - 1.7789 \\
 &= 0.0461
 \end{aligned}$$

Contrast this distance measure with $KL(p, s) = 0.0477$. It is important to note that the constraint function averages appearing in the above calculation, $\langle F_j \rangle_s$ are with respect to the distribution s . Thus, for example, $\langle F_7 \rangle_s = 0.36$, and not $\langle F_7 \rangle_p = 0.40$. Even though s has only four Lagrange multipliers, all seven constraint function averages with respect to s , $\langle F_j \rangle_s$, as required in the sum, do exist.

Exercise 29.9.9: Show another formula for the relative entropy that doesn't immediately take advantage, as done so far, of the definition for the information entropy of point p .

Solution to Exercise 29.9.9

We have been using this formula for a generic p and q ,

$$KL(p, q) = \sum_{i=1}^n p_i \ln \left(\frac{p_i}{q_i} \right) = E_p [\ln p_i - \ln q_i]$$

But if $\ln p_i$ is expanded according to the MEP formula just as we have been doing for $\ln q_i$, then instead of taking advantage of,

$$E_p [\ln p_i] = -H_{max}(p)$$

we go ahead and multiply through by p_i ,

$$\begin{aligned}
 \sum_{i=1}^n p_i \ln \left(\frac{p_i}{q_i} \right) &= \sum_{i=1}^n p_i \times \left\{ \left[\sum_{j=1}^7 \lambda_j^p F_j(x_i) - \ln Z_p \right] - \left[\sum_{j=1}^7 \lambda_j^q F_j(x_i) - \ln Z_q \right] \right\} \\
 &= \sum_{i=1}^n p_i \times \left[\sum_{j=1}^7 (\lambda_j^p - \lambda_j^q) F_j(x_i) - \ln Z_p + \ln Z_q \right] \\
 KL(p, q) &= \sum_{j=1}^7 (\lambda_j^p - \lambda_j^q) \langle F_j \rangle_p + \ln \left(\frac{Z_q}{Z_p} \right)
 \end{aligned}$$

Exercise 29.9.10: Verify this last expression by computing the distance between distribution p , the $m = 7$ correlational model that includes the information from all of the interactions, and distribution q , which is the $m = 4$ correlational model that includes only the information from the BH double interaction.

Solution to Exercise 29.9.10

The generic distribution q in the new formula is the same as distribution q in section 29.4 and distribution s in section 29.6. Taking the new formula for the relative entropy,

$$KL(p, q) = \sum_{j=1}^7 (\lambda_j^p - \lambda_j^q) \langle F_j \rangle_p + \ln \left(\frac{Z_q}{Z_p} \right)$$

as it was derived in the last exercise, and plugging in the appropriate numbers,

$$\begin{aligned} KL(p, q) &= [(0.847298 - 0.405465) \times 0.75] + [(1.38629 - 0.405465) \times 0.75] + \\ &\quad [(0.510826 - 0.405465) \times 0.60] + [(-0.441833 - 0.980829) \times 0.58] + \\ &\quad [(-0.154151 - 0) \times 0.50] + [(-1.38629 - 0) \times 0.45] + \\ &\quad [(1.82813 - 0) \times 0.40] + \ln \left(\frac{25.00}{33.33} \right) \\ &= 0.0477 \end{aligned}$$

This matches $KL(p, s)$ in section 29.6 and $KL(p, q)$ in section 29.4, both of which had $\langle F_4 \rangle = 0.60$. Note that the final three Lagrange multipliers under the model determining q are equal to 0.

Exercise 29.9.11: Use this expression to compute the squared distance between distribution p , the $m = 7$ correlational model that includes the information from all of the interactions, and a distribution q , which now we take to be the $m = 3$ independence model of section 29.3.

Solution to Exercise 29.9.11

Taking this recent formula for the relative entropy,

$$KL(p, q) = \sum_{j=1}^7 (\lambda_j^p - \lambda_j^q) \langle F_j \rangle_p + \ln \left(\frac{Z_q}{Z_p} \right)$$

and plugging in the appropriate values for the $m = 3$ independence model q ,

$$\lambda_1^q = \lambda_2^q = 1.09861, \lambda_3^q = 0.405465 \text{ and } Z_q = 40$$

$$\begin{aligned}
KL(p, q) &= [(0.847298 - 1.09861) \times 0.75] + [(1.38629 - 1.09861) \times 0.75] + \\
&\quad [(0.510826 - 0.405465) \times 0.60] + [(-0.441833 - 0) \times 0.58] + \\
&\quad [(-0.154151 - 0) \times 0.50] + [(-1.38629 - 0) \times 0.45] + \\
&\quad [(1.82813 - 0) \times 0.40] + \ln\left(\frac{40.00}{33.33}\right) \\
&= 0.04689
\end{aligned}$$

These are the computational details alluded to at the end of section 29.3. The squared distance from the more complicated model p is observed to be closer to the independence model than to another model containing the information from one double interaction.

Exercise 29.9.12: List the canonical divergence formulas that are required for the numerical solutions concerning the three points p , q , and r in section 29.6.

Solution to Exercise 29.9.12

The canonical divergence formulas relating points p , q , and r are,

$$\begin{aligned}
KL(p, q) &= \ln Z_q - \sum_{j=1}^7 \lambda_j^q \langle F_j \rangle_p - H_{max}(p) \\
KL(q, r) &= \ln Z_r - \sum_{j=1}^7 \lambda_j^r \langle F_j \rangle_q - H_{max}(q) \\
KL(p, r) &= \ln Z_r - \sum_{j=1}^7 \lambda_j^r \langle F_j \rangle_p - H_{max}(p)
\end{aligned}$$

Exercise 29.9.13: Carry out the detailed computation for $KL(p, q)$.

Solution to Exercise 29.9.13

In applying the appropriate formula for the relative entropy between p and q ,

$$KL(p, q) = \ln Z_q - \sum_{j=1}^7 \lambda_j^q \langle F_j \rangle_p - H_{max}(p)$$

the partition function Z_q for distribution q , the information entropy $H_{max}(p)$ for distribution p , the Lagrange multipliers λ_j^q for distribution q , and the constraint function averages $\langle F_j \rangle_p$ under p must be specified.

The constraint function averages under p are,

$$\langle F_1 \rangle_p = 0.75$$

$$\langle F_2 \rangle_p = 0.75$$

$$\langle F_3 \rangle_p = 0.60$$

$$\langle F_4 \rangle_p = 0.58$$

$$\langle F_5 \rangle_p = 0.50$$

$$\langle F_6 \rangle_p = 0.45$$

$$\langle F_7 \rangle_p = 0.40$$

The first three constraint function averages $\langle F_j \rangle$ are the same under all of the models discussed. These constraint function averages represent the information about the marginal probabilities for beer preference, hand preference, and fur color.

Make use of the MEP algorithm to find the four Lagrange multipliers λ_j^q and log of the partition function $\ln Z_q$. By fixing the constraint function averages under q where $\langle F_4 \rangle_q$ must match $\langle F_4 \rangle_p$,

$$\langle F_1 \rangle_q = 0.75$$

$$\langle F_2 \rangle_q = 0.75$$

$$\langle F_3 \rangle_q = 0.60$$

$$\langle F_4 \rangle_q = 0.58$$

the numerical value of the assignments for the distribution q are found to be,

$$q = \{0.348, 0.102, 0.232, 0.068, 0.102, 0.048, 0.068, 0.032\}$$

with an information entropy of $H_{max}(q) = 1.79348$. The partition function is $Z_q = 31.25$ with $\ln Z_q = \ln(31.25) = 3.442$. The Lagrange multipliers are,

$$\lambda_1^q = 0.753772$$

$$\lambda_2^q = 0.753772$$

$$\lambda_3^q = 0.405465$$

$$\lambda_4^q = 0.473458$$

We have to use the MEP algorithm a second time to find the information entropy for the assignments under p . The numerical assignments under the highly correlated model for point p have been discussed many times before,

$$p = \{0.40, 0.10, 0.18, 0.07, 0.05, 0.05, 0.12, 0.03\}$$

with $H_{max}(p) = 1.75079$. Remember that this information entropy is the lowest entropy of any distribution discussed because this model has the most number of constraints. Each piece of additional information in the form of another constraint function average reduces the amount of missing information. Therefore, we recognize that the information entropy must be reduced as well.

The squared distance between the full correlational model p and its projection to a less complicated model q , which matches p at least through the first four constraint function averages, is then, (see Figure 29.2 on pg. 415),

$$\begin{aligned} KL(p, q) &= \ln Z_q - \sum_{j=1}^4 \lambda_j^q \langle F_j \rangle_p - H_{max}(p) \\ &= 3.442 - (0.753772 \times 0.75) + \dots + (0.473458 \times 0.58) - 1.75079 \\ &= 0.0427 \end{aligned}$$

Exercise 29.9.14: Confirm that the numerical assignments for q satisfy the information under the model. Verify as well the information entropy of the resulting distribution.

Solution to Exercise 29.9.14

The numerical value of the assignments for the distribution q were found in the last exercise as,

$$q = \{0.348, 0.102, 0.232, 0.068, 0.102, 0.048, 0.068, 0.032\}$$

These assignments are listed in order for the eight cells of the joint probability table from Q_1 through Q_8 .

The marginal distribution for beer preference is $P(B | \mathcal{M}_k) = 0.75$. Assignments Q_1 through Q_4 add up to $0.348 + 0.102 + 0.232 + 0.068 = 0.75$. The marginal distribution for hand preference is $P(H | \mathcal{M}_k) = 0.75$. Assignments Q_1, Q_3, Q_5 and Q_7 add up to $0.348 + 0.232 + 0.102 + 0.068 = 0.75$. The marginal distribution for fur color is $P(F | \mathcal{M}_k) = 0.60$. Assignments Q_1, Q_2, Q_5 , and Q_6 add up to $0.348 + 0.102 + 0.102 + 0.048 = 0.60$. The marginal distribution for the BH double interaction is $P(BHF | \mathcal{M}_k) + P(BH\bar{F} | \mathcal{M}_k) = 0.58$. Assignments Q_1 and Q_3 add up to $0.348 + 0.232 = 0.58$.

The information entropy is calculated from first principles as,

$$\begin{aligned}
 H_{max}(q) &= - \sum_{i=1}^8 Q_i^q \ln Q_i^q \\
 &= - [(0.348 \ln 0.348) + (0.102 \ln 0.102) + \dots + (0.032 \ln 0.032)] \\
 &= 1.79348
 \end{aligned}$$

Exercise 29.9.15: Carry out the detailed computation for $KL(q, r)$.

Solution to Exercise 29.9.15

In applying the appropriate formula for the relative entropy between q and r ,

$$KL(q, r) = \ln Z_r - \sum_{j=1}^7 \lambda_j^r \langle F_j \rangle_q - H_{max}(q)$$

the partition function Z_r and the Lagrange multipliers λ_j^r for distribution r must be found. The last three λ_j^r are all equal to 0. The information entropy $H_{max}(q)$ and the four constraint function averages $\langle F_j \rangle_q$ under q are already known from the previous exercise.

Make use of the MEP algorithm to find the four Lagrange multipliers λ_j^r and log of the partition function $\ln Z_r$. We have a new model and a different numerical assignment to the probabilities of the eight joint statements in the state space by fixing the constraint function averages under r at,

$$\langle F_1 \rangle_r = 0.75$$

$$\langle F_2 \rangle_r = 0.75$$

$$\langle F_3 \rangle_r = 0.60$$

$$\langle F_4 \rangle_r = 0.56$$

Point r , as mentioned, has the same marginal probabilities for beer preference, hand preference, and fur color as all of the other models, but inserts different information about the correlation enforced by the double interaction BH . The numerical values of the assignment to the distribution r are found via the MEP formula,

$$r = \{0.336, 0.114, 0.224, 0.076, 0.114, 0.036, 0.076, 0.024\}$$

with an information entropy of $H_{max}(r) = 1.79759$. These assignments may be checked at your leisure to make sure that they satisfy all the constraints. The most

important one is that $Q_1 + Q_3 = 0.336 + 0.224 = 0.56$. The partition function is $Z_r = 41.67$ with $\ln Z_r = \ln(41.67) = 3.7297$. The Lagrange multipliers are,

$$\begin{aligned}\lambda_1^r &= 1.15268 \\ \lambda_2^r &= 1.15268 \\ \lambda_3^r &= 0.405465 \\ \lambda_4^r &= -0.071767\end{aligned}$$

Thus, we have, (see Figure 29.2 on pg. 415 again),

$$\begin{aligned}KL(q, r) &= \ln Z_r - \sum_{j=1}^4 \lambda_j^r \langle F_j \rangle_q - H_{max}(q) \\ &= 3.7297 - [(1.15268 \times 0.75) + \dots + (-0.071767 \times 0.58)] - 1.79348 \\ &= 0.0055\end{aligned}$$

Exercise 29.9.16: Carry out the detailed computation for $KL(p, r)$.

Solution to Exercise 29.9.16

By now the general scheme for computing the distance between any two points is apparent. For the third and final leg of the “triangle” pqr , we will find the relative entropy $K(p, r)$.

$$KL(p, r) = \ln Z_r - \sum_{j=1}^7 \lambda_j^r \langle F_j \rangle_p - H_{max}(p)$$

The partition function Z_r for distribution r , the information entropy $H_{max}(p)$ for distribution p , the Lagrange multipliers λ_j^r for distribution r , and all seven constraint function averages $\langle F_j \rangle_p$ under p must be found. But these components are already available to us from the solution to the two previous distance measures.

Thus, we have,

$$\begin{aligned}KL(p, r) &= \ln Z_r - \sum_{j=1}^7 \lambda_j^r \langle F_j \rangle_p - H_{max}(p) \\ &= 3.7297 - [(1.15268 \times 0.75) + \dots + (0 \times 0.40)] - 1.75079 \\ &= 0.0482\end{aligned}$$

Exercise 29.9.17: What can we now surmise about triangle pqr ?

Solution to Exercise 29.9.17

The triangle satisfies the Pythagorean relationship,

$$KL(p, r) = KL(p, q) + KL(q, r)$$

$$0.0482 = 0.0427 + 0.0055$$

Exercise 29.9.18: Show heuristically that point q is at the minimum distance from p by examining two points very close to q and on either side of it.

Solution to Exercise 29.9.18

We have surmised that the minimum distance from point p to any point in the submanifold where $m = 4$ is the distance from p to q , $KL(p, q) = 0.0426887$. Label a point very close to q as q^+ with constraint function average $\langle F_4 \rangle_q^+ = 0.581$, and likewise label another point very close to q , and on the other side of q , as q^- with constraint function average $\langle F_4 \rangle_q^- = 0.579$. Both of these nearby points should have a slightly larger distance from p than q .

And just like all of the previous exercises, calculate,

$$KL(p, q^+) = \ln Z_{q^+} - \sum_{j=1}^7 \lambda_j^{q^+} \langle F_j \rangle_p - H_{max}(p)$$

$$= 0.0427017$$

$$KL(p, q^-) = \ln Z_{q^-} - \sum_{j=1}^7 \lambda_j^{q^-} \langle F_j \rangle_p - H_{max}(p)$$

$$= 0.0427018$$

confirming that $KL(p, q) = 0.0426887$ has indeed found the minimum.

Exercise 29.9.19: Confirm that point s as discussed in section 29.6 is indeed further from point p than point q .

Solution to Exercise 29.9.19

Point s is the numerical assignment arising from applying the MEP to insert the information from a model asserting that the first four constraint function averages are,

$$\langle F_1 \rangle_s = 0.75$$

$$\langle F_2 \rangle_s = 0.75$$

$$\langle F_3 \rangle_s = 0.60$$

$$\langle F_4 \rangle_s = 0.60$$

The relative entropy between this distribution s and distribution p is,

$$KL(p, s) = \ln Z_s - \sum_{j=1}^7 \lambda_j^s \langle F_j \rangle_p - H_{max}(p)$$

The answer from this formula is $KL(p, s) = 0.0477$ verifying that its distance is greater than the distance from p to q of 0.0427. This answer may be checked by calculating,

$$KL(p, s) = \sum_{i=1}^8 p_i \ln \left(\frac{p_i}{s_i} \right)$$

The numerical assignments to the p_i are known from previous exercises, while the assignments to s_i from enforcing the constraints and maximizing the entropy are,

$$s_i = \{0.36, 0.09, 0.24, 0.06, 0.09, 0.06, 0.06, 0.04\}$$

Note especially that the BH constraint $\langle F_4 \rangle_s = Q_1 + Q_3 = 0.36 + 0.24 = 0.60$ is satisfied. Carrying out the calculation, we have,

$$\begin{aligned} KL(p, s) &= \sum_{i=1}^8 p_i \ln \left(\frac{p_i}{s_i} \right) \\ &= 0.40 \ln \left(\frac{0.40}{0.36} \right) + \dots + 0.03 \ln \left(\frac{0.03}{0.04} \right) \\ &= 0.0477 \end{aligned}$$

Exercise 29.9.20: Calculate the sample average for each constraint function based on the data for the enhanced kangaroo scenario.

Solution to Exercise 29.9.20

The average for any constraint function based on the actual frequency counts from the data is,

$$\bar{F}_j = \sum_{i=1}^n \frac{N_i}{N} \times F_j(x_i)$$

Substituting $n = 8$, $N = 100$, the known frequency counts N_i from each cell of the contingency table, and $j = 1$ for the first constraint function, we have,

$$\begin{aligned}\overline{F}_1 &= \sum_{i=1}^8 \frac{N_i}{N} \times F_1(x_i) \\ &= \left(\frac{40}{100} \times 1 \right) + \left(\frac{10}{100} \times 1 \right) + \cdots + \left(\frac{3}{100} \times 0 \right) \\ &= 0.75\end{aligned}$$

In like manner, the sample averages of all seven constraint functions with respect to the data presented in the contingency table of Figure 22.2 on pg. 153, are,

$$\overline{F}_1 = 0.75$$

$$\overline{F}_2 = 0.75$$

$$\overline{F}_3 = 0.60$$

$$\overline{F}_4 = 0.58$$

$$\overline{F}_5 = 0.50$$

$$\overline{F}_6 = 0.45$$

$$\overline{F}_7 = 0.40$$

The numerical assignments to the probabilities in the enhanced kangaroo state space follow from the *information* in the model for point p . But the crucial feature from the perspective of maximum likelihood estimation is that this information concerning all seven constraint function averages $\langle F_1 \rangle = 0.75$ through $\langle F_7 \rangle = 0.40$ exactly match these sample averages.

Exercise 29.9.21: How much is the model for point p favored over the model for point q when conditioned on the actual data as presented in the contingency table of Figure 22.2?

Solution to Exercise 29.9.21

Point p is the distribution based on the highly correlated model with $m = 7$, while point q is the distribution based on a single correlation between hand preference and beer preference with $m = 4$. Let p be the generic model \mathcal{M}_A and q the generic model \mathcal{M}_B .

From Equation (29.12), the ratio of the posterior probability for these two models is calculated as,

$$\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} = e^{[N \times KL(p, q)]}$$

$$N = 100$$

$$KL(p, q) = 0.0426887$$

$$\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} = e^{[100 \times 0.0426887]}$$

$$= 71.44$$

Exercise 29.9.22: What is the probability that some other kangaroo not in the data base prefers Foster's beer given that it is right handed and sandy colored?

Solution to Exercise 29.9.22

If we average over just these two points p and q , then,

$$\begin{aligned} P(B | H, F, \mathcal{D}) &= \sum_{k=1}^2 \left[\frac{P(BHF | \mathcal{M}_k)}{P(HF | \mathcal{M}_k)} \right] \times P(\mathcal{M}_k | \mathcal{D}) \\ &= \frac{\left(\frac{0.40}{0.40+0.05} \times 71.44 \right) + \left(\frac{0.348}{0.348+0.102} \times 1 \right)}{72.44} \\ &= 0.8873 \end{aligned}$$

Compare this conditional probability that some other kangaroo prefers Foster's when averaged over just two models to the probability as averaged over all models,

$$P(B | H, F, \mathcal{D}) = \frac{41}{41 + 6} = 0.8723$$

Exercise 29.9.23: What is the relative weighting of this model for point p compared to another model for a point r with very similar correlations?

Solution to Exercise 29.9.23

A model \mathcal{M}_C similar to model \mathcal{M}_A has its Lagrange multiplier parameters λ_j^C shifted slightly from those already found for model \mathcal{M}_A . For a numerical exercise, suppose that all the Lagrange multipliers for model \mathcal{M}_C are shifted downwards by some small amount like -0.01 . Thus, $\lambda_1^C = 0.837298, \dots, \lambda_7^C = 1.81813$ as compared to $\lambda_1^A = 0.847298, \dots, \lambda_7^A = 1.82813$.

This allows us to write,

$$\begin{aligned} \ln \left[\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_C | \mathcal{D})} \right] &= N \left[\sum_{j=1}^m (\lambda_j^A - \lambda_j^C) \bar{F}_j + \ln \left(\frac{Z_C}{Z_A} \right) \right] \\ \sum_{j=1}^7 (\lambda_j^A - \lambda_j^C) \bar{F}_j &= (0.01 \times 0.75) + (0.01 \times 0.75) + \dots + (0.01 \times 0.40) \\ &= 0.0403 \end{aligned}$$

The partition function Z_C must be very close to Z_A . Therefore $\ln(Z_C/Z_A)$ is going to be close to zero. In fact, since $Z_C = 32.027$ it is,

$$\ln \left(\frac{32.027}{33.333} \right) \approx -0.04$$

$$\begin{aligned} \ln \left[\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_C | \mathcal{D})} \right] &\approx 100 \times [0.0403 + (-0.04)] \\ &\approx 0.03 \\ \frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_C | \mathcal{D})} &\approx e^{0.03} \\ &\approx 1.03 \\ &= \exp[N \times KL(p, r)] \end{aligned}$$

In the end, everything makes perfect sense. The data support point p to the greatest extent possible, and therefore p is relatively highly favored over any point like q that is some distance away. But the “separation” between p and r , as measured by relative entropy $KL(p, r) \approx 0.0003$, is not that dramatic. The predictions made by the model for r are counted almost as much as those made by p .

Exercise 29.9.24: Discuss the relative weighting of models in the coin flip scenario after having observed the data from 100 previous coin tosses.

Solution to Exercise 29.9.24

This exercise is a rather lengthy and discursive tour of an elementary inferential problem. It serves as a sort of summary review of where we have arrived at this point in our journey.

It is always worthwhile to periodically go over old ground when we are engaged in absorbing new concepts. What is new here, as compared to the treatment of the same inferential problem in Volume I, is the availability of the MEP formula for the numerical assignment to the probabilities in the state space when conditioned on the information resident in some model.

In the literature, the coin tossing scenario is usually seen as an example of what is typically called “Bernoulli trials and the resulting Binomial distribution.” Even so simple an inferential problem as this generates an enormous amount of attendant confusion as we have been at pains to suggest in Volume I. Proponents of “Bayesian solutions” are no less guilty than the unenlightened. But really, shouldn’t an analysis of this situation be devoid of any mysteries?

The coin tossing scenario is an example of a state space with two statements; thus, the dimension of the state space is $n = 2$. Suppose that $N = 100$ tosses of the coin have resulted in the observed data of $N_1 = 62$ HEADS and $N_2 = 38$ TAILS.

The listing of the data is actually a joint statement something like, “TAILS was observed on trial 1, and HEADS was observed on trial 2, …, and TAILS was observed on the last trial.” Since the symbol \mathcal{D} appears inside the probability operator as, for example, it does in $P(\mathcal{M}_k | \mathcal{D})$, \mathcal{D} MUST BE a statement. Likewise, \mathcal{M}_k MUST BE a statement as well.

The specific realization of the data as $N_1 = 62$ HEADS and $N_2 = 38$ TAILS is any sequence containing 62 HEADS and 38 TAILS in 100 coin tosses. The total probability for the data would then take account of the multiplicity factor counting up the number of ways any such sequence might occur.

In my notation, any one sequence \mathbb{S} looks like,

$$\mathbb{S} = \{(X_1 = x_2) \text{ and } (X_2 = x_1) \text{ and } \dots \text{ and } (X_{100} = x_2)\}$$

with 62 ($X = x_1$) (HEADS) and 38 ($X = x_2$) (TAILS). The total number of such sequences with 62 HEADS and 38 TAILS is,

$$W(N) = \frac{N!}{N_1! N_2!} = \frac{100!}{62! 38!}$$

But in the end, we won’t have to worry about the multiplicity factor when comparing models conditioned on the data because the data are the same irrespective of what the models have to say, and the multiplicity factor cancels out.

The formal rules say that,

$$P(\mathbb{S} | \mathcal{M}_k) = \frac{P(\mathbb{S} \text{ and } \mathcal{M}_k)}{P(\mathcal{M}_k)}$$

The probability of any sequence and a model is found from an application of the

Commutativity property inherited from Boolean Algebra and the **Product Rule**,

$$P(\{(X_1 = x_2) \text{ and } (X_2 = x_1) \text{ and } \cdots \text{ and } (X_{100} = x_2)\} \text{ and } \mathcal{M}_k) =$$

$$P(X_{100} = x_2 | X_{99} = x_1, X_{98} = x_1, \dots, X_1 = x_2, \mathcal{M}_k) \times$$

$$P(X_{99} = x_1 | X_{98} = x_1, X_{97} = x_1, \dots, X_1 = x_2, \mathcal{M}_k) \times$$

⋮

$$P(X_1 = x_2 | \mathcal{M}_k) \times P(\mathcal{M}_k)$$

The numerical assignment of a probability at any trial X_t when conditioned on some model is independent of the results at all previous trials, and depends only on the model, $P(X_t = x_i | \mathcal{M}_k)$. Thus, we have,

$$\begin{aligned} P(\mathbb{S} | \mathcal{M}_k) &= \frac{P(\mathbb{S} \text{ and } \mathcal{M}_k)}{P(\mathcal{M}_k)} \\ P(\mathbb{S} | \mathcal{M}_k) &= \frac{Q_2 \times Q_1 \times \cdots Q_2 \times P(\mathcal{M}_k)}{P(\mathcal{M}_k)} \\ &= Q_1^{N_1} Q_2^{N_2} \\ P(\mathcal{D} | \mathcal{M}_k) &= \frac{N!}{N_1! N_2!} Q_1^{N_1} Q_2^{N_2} \end{aligned}$$

To answer the question as posed, choose any two models and call them model \mathcal{M}_A and model \mathcal{M}_B . Suppose we are interested in the fair model as model \mathcal{M}_A and the model that matches the data as \mathcal{M}_B . What is the relative weighting given to these two models in the averaging procedure to find the probability for HEADS or TAILS on the 101st toss of the coin?

Expand the log likelihood ratio, and prepare for the sample averages,

$$\begin{aligned} \ln \left[\frac{P(\mathcal{D} | \mathcal{M}_A)}{P(\mathcal{D} | \mathcal{M}_B)} \right] &= \sum_{i=1}^n N_i \ln Q_i^A - \sum_{i=1}^n N_i \ln Q_i^B \\ &= N \times \left(\sum_{i=1}^2 \frac{N_i}{N} \ln Q_i^A - \sum_{i=1}^2 \frac{N_i}{N} \ln Q_i^B \right) \end{aligned}$$

Here is where we invoke the MEP in order to find Q_i^A and Q_i^B . Since only $m = 1$ parameter λ is required for this case where $n = 2$,

$$\ln Q_i^A = \lambda^A F(x_i) - \ln Z_A$$

$$\ln Q_i^B = \lambda^B F(x_i) - \ln Z_B$$

Substitute into the previous equation,

$$\begin{aligned}\sum_{i=1}^n \frac{N_i}{N} \ln Q_i^A &= \lambda^A \bar{F} - \ln Z_A \\ \sum_{i=1}^n \frac{N_i}{N} \ln Q_i^B &= \lambda^B \bar{F} - \ln Z_B \\ \ln \left[\frac{P(\mathcal{D} | \mathcal{M}_A)}{P(\mathcal{D} | \mathcal{M}_B)} \right] &= N \times \left[(\lambda^A - \lambda^B) \bar{F} + \ln \left(\frac{Z_B}{Z_A} \right) \right]\end{aligned}$$

The sample average is,

$$\begin{aligned}\bar{F} &= \sum_{i=1}^2 \frac{N_i}{N} \times F(X = x_i) \\ &= \left(\frac{62}{100} \times 1 \right) + \left(\frac{38}{100} \times 2 \right) \\ &= 1.38\end{aligned}$$

We already know that $\lambda^A = 0$ and $Z_A = 2$. We solve for λ^B and Z_B by exploiting the Legendre transformation with $\bar{F} = \langle F \rangle = 1.38$,

$$H_{max}(Q_i^B) = \min_{\lambda^B} [\ln Z_B - \lambda^B \langle F \rangle]$$

The solution can be found through this *Mathematica* expression,

```
NMinimize[Log[Exp[ $\lambda$ ] + Exp[2  $\lambda$ ]] - ( $\lambda$   $\times$  1.38),  $\lambda$ ]
```

which returns $\lambda^B = -0.489548$, and $Z_B = 0.98855$.

$$\begin{aligned}\ln \left[\frac{P(\mathcal{D} | \mathcal{M}_A)}{P(\mathcal{D} | \mathcal{M}_B)} \right] &= N \times \left[(\lambda^A - \lambda^B) \bar{F} + \ln \left(\frac{Z_B}{Z_A} \right) \right] \\ &= 100 \times \left[((0 - (-0.489548)) \times 1.38 + \ln \left(\frac{0.98855}{2} \right) \right] \\ &= -2.9083 \\ \frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} &= e^{-2.9083} \\ &= 0.0546 \\ &= \exp[-N \times KL(p, q)]\end{aligned}$$

In the end, after all of the data have been accounted for, we see that the relative standing of the fair model ascribing a numerical value of 1/2 to a probability for HEADS, and 1/2 to a probability for TAILS is only about 1/20th the weight of the model ascribing a numerical value of 0.62 to a probability for HEADS and 0.38 to a probability for TAILS.

The probability for HEADS on the *next* toss of the coin, namely the 101st toss, when averaging over just these two models is,

$$P(\text{HEADS}_{N+1} | \mathcal{D}) = \frac{(1/2 \times 0.0546) + (0.62 \times 1)}{1.0546} = 0.613787$$

Compare this probability for the next toss to the probability averaged over all models as found by Laplace's *Rule of Succession*,

$$P(\text{HEADS}_{N+1} | \mathcal{D}) = \frac{N_1 + 1}{N + n} = \frac{63}{102} = 0.617647$$

Exercise 29.9.25: What is the relative weighting of a highly correlated model in the kangaroo scenario as compared to the uniform distribution under the fair model when conditioned on the data?

Solution to Exercise 29.9.25

The most highly correlated model in the kangaroo scenario that we have examined so far was the $m = 7$ model called point p with the assigned values of,

$$p \equiv Q_i^A \equiv P(X = x_i | \mathcal{M}_A) \equiv (0.40, 0.10, 0.18, 0.07, 0.05, 0.05, 0.12, 0.03)$$

This assignment matches the normed frequency counts in the data. We have been calling it the “maximum likelihood” model. The assigned values under the “fair” model, call it point u , are,

$$u \equiv Q_i^B \equiv P(X = x_i | \mathcal{M}_B) \equiv (1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8)$$

The ratio of the probability for these two models when conditioned on the data from the $N = 100$ kangaroos is (refer back to Figure 22.2),

$$\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} = \exp [N \times KL(p, u)]$$

$$KL(p, u) = 0.328654$$

$$\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} = \exp [100 \times 0.328654]$$

$$\approx 1.88 \times 10^{14}$$

This is a summary derivation skipping over all the details so that we can get to the conclusion rapidly. The details are left to the end of the exercise.

As might have been foreseen, the model incorporating information from double and triple interactions was enormously favored over a model that did not include information of any kind. The data strongly favor model \mathcal{M}_A with its supposed correlational structure as compared to model \mathcal{M}_B with its complete lack of any relationship among the variables.

A probability of 1/8 for, say, $P(B_{N+1}H_{N+1}F_{N+1} | \mathcal{M}_B)$ in cell 1 of the joint probability table as assigned under the fair model will not play any role in the averaging reflected in,

$$P(B_{N+1}H_{N+1}F_{N+1} | \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(B_{N+1}H_{N+1}F_{N+1} | \mathcal{M}_k) P(\mathcal{M}_k | \mathcal{D})$$

while a probability of 0.40 for $P(B_{N+1}H_{N+1}F_{N+1} | \mathcal{M}_A)$ as assigned under the highly correlated model will play a large role.

Here are the details that we skipped over in the summary conclusion. The log of the ratio of the probability for the two models given the data was determined in Exercise 29.9.24 to be,

$$\ln \left[\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} \right] = N \left[\sum_{j=1}^m (\lambda_j^A - \lambda_j^B) \bar{F}_j + \ln \left(\frac{Z_B}{Z_A} \right) \right]$$

The Lagrange multipliers for model \mathcal{M}_B are all $\lambda_j^B = 0$. This is the implication of saying that this model inserted no information into its distribution. The partition function Z_A for model \mathcal{M}_A is returned as part of the MEP algorithm at the same time as the values of the Lagrange multipliers. It was $Z_A = 33.333$.

So we have,

$$\ln \left[\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} \right] = 100 \left[\sum_{j=1}^7 \lambda_j^A \bar{F}_j + \ln \left(\frac{8}{33.33} \right) \right]$$

We have already calculated the seven parameters λ_j^A for model \mathcal{M}_A as,

$$\lambda_j^A = (0.847298, 1.38629, 0.510826, -0.441833, -0.154151, -1.38629, 1.82813)$$

and the sample averages of the seven constraint functions in Exercise 29.9.20.

$$\begin{aligned} \sum_{j=1}^7 \lambda_j^A \bar{F}_j &= (0.847298 \times 0.75) + (1.38629 \times 0.75) + \cdots + (1.82813 \times 0.40) \\ &= 1.755769 \end{aligned}$$

We now have,

$$\begin{aligned}
 \ln \left[\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} \right] &= 100 \left[1.755769 + \ln \left(\frac{8}{33.33} \right) \right] \\
 &= 100 [1.755769 - 1.427115] \\
 &= 100 \times 0.328654 \\
 \frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} &= \exp [100 \times 0.328654] \\
 &\approx 1.88 \times 10^{14}
 \end{aligned}$$

Exercise 29.9.26: What is the probability of two future observations consisting of a TAILS and an EDGE when there are data available from three previous tosses of a “thick” coin?

Solution to Exercise 29.9.26

This is another long exercise which summarizes many inferential lessons from times past. Its main virtue is that all of its intricate details can still be followed from a computational perspective. Ultimately, that is what we are striving for; partial confirmation of our symbolic derivations through computation of special cases.

The IP is interested in the next two coin flips, so $M = 2$. The state space consists of three statements where the canonical coin toss space has been augmented with a third statement, “The coin lands on its EDGE.” The dimension of the state space is now $n = 3$.

The future observations of a TAILS and an EDGE lead to $M_1 = 0$, $M_2 = 1$, and $M_3 = 1$ where $\sum_{i=1}^3 M_i = M$. The data consist of three previous coin flips, so $N = 3$. Suppose that two TAILS, one HEAD, and no EDGES were observed. Thus, $N = 3$ and $N_1 = 1$, $N_2 = 2$, and $N_3 = 0$.

The probability for these two future observations can always be found from the predictive distribution derived from the formal manipulation rules.

$$\begin{aligned}
 P(M_1 = 0, M_2 = 1, M_3 = 1 | N_1 = 1, N_2 = 2, N_3 = 0) &= \\
 \frac{M! (N + n - 1)!}{N_1! N_2! N_3! (M + N + n - 1)!} \times \frac{\prod_{i=1}^3 (M_i + N_i)!}{\prod_{i=1}^3 M_i!} &= \\
 \frac{2! (3 + 3 - 1)!}{1! 2! 0! (2 + 3 + 3 - 1)!} \times \frac{1! 3! 1!}{0! 1! 1!} &= \\
 \frac{5!}{7!} \times 3! &= \frac{1}{7}
 \end{aligned}$$

The probability for the future occurrence of TAILS and an EDGE when conditioned on the data is,

$$P(M_1 = 0, M_2 = 1, M_3 = 1 \mid \mathcal{D}) = \frac{1}{7}$$

The probability for this possibility on the fourth and fifth tosses of the coin conditioned on the data from the three previous tosses is $1/7$. Remember that this probability is the result of averaging over every conceivable model. So what role does the MEP play if we already have the answer?

Before we answer that question, as a sort of validity check on the sanity of the above predictive formula, compute the probability for all six possibilities on the next two coin flips. These probabilities must sum to 1, or else something is wrong. Furthermore, do some rough qualitative musings about these probabilities, admittedly after the fact, that are sensitive to any severe jolt to our intuition. Table 29.1 below shows all of the details.

Table 29.1: *Checking that the probabilities for all six possible frequency counts on the next two coin flips add up to 1.*

Possibility	Description	Future frequency counts	Probability
1	Two HEADS	2, 0, 0	1/7
2	Two TAILS	0, 2, 0	2/7
3	Two EDGES	0, 0, 2	1/21
4	One HEADS, One TAILS	1, 1, 0	2/7
5	One HEADS, One EDGE	1, 0, 1	2/21
6*	One TAILS, One EDGE	0, 1, 1	1/7
<i>Sum</i>			21/21

The highest probability events are two TAILS, or one HEADS and one TAILS. Since we saw two TAILS and a HEADS in the data this makes sense. The lowest probability event is two EDGES. Since we did not see an EDGE in the data, this also makes sense.

Again, given what transpired with the data, the probability is $5/7$ that we will not see the coin land on EDGE in the next two flips. Similar kinds of *ex post facto* reasoning seem to lend credibility to the formal symbol manipulations that resulted in our predictive formula.

When we found the probability for a TAILS and an EDGE on the next two tosses of the coin conditioned on the known data from three previous tosses of the coin, we relied on this formal manipulation formula,

$$P(M_1, M_2, M_3 | \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(M_1, M_2, M_3 | \mathcal{M}_k) P(\mathcal{M}_k | \mathcal{D})$$

Let's take a more in-depth look at just a few of the models that played a role in this average. Table 29.2 below shows six more models in addition to a model \mathcal{M}_A . Model \mathcal{M}_A is seen to be a model that matches the normed frequency counts of the data.

Table 29.2: *A few of the models that played a role in the probability for the next two tosses of the coin.*

\mathcal{M}_k	$P(X = x_i \mathcal{M}_k)$	$P(M_i \mathcal{M}_k)$	$\frac{P(\mathcal{M}_k \mathcal{D})}{P(\mathcal{M}_A \mathcal{D})}$
\mathcal{M}_A	(1/3, 2/3, 0)	0	1.0000
\mathcal{M}_B	(0.05, 0.475, 0.475)	0.45125	0.0761
\mathcal{M}_C	(0.10, 0.45, 0.45)	0.40500	0.1367
\mathcal{M}_D	(0.15, 0.425, 0.425)	0.36125	0.1829
\mathcal{M}_E	(0.20, 0.40, 0.40)	0.32000	0.2160
\mathcal{M}_F	(0.25, 0.375, 0.375)	0.28125	0.2373
\mathcal{M}_G	(0.30, 0.60, 0.10)	0.12000	0.7290

Taking the average over just these seven models results in a probability of 0.1471, a good approximation to the actual probability of $1/7 = 0.1429$,

$$\begin{aligned} P(M_1 = 0, M_2 = 1, M_3 = 1 | \mathcal{D}) &\approx \sum_{k=1}^7 P(M_1, M_2, M_3 | \mathcal{M}_k) P(\mathcal{M}_k | \mathcal{D}) \\ &\approx \frac{(0 \times 1) + (0.45125 \times 0.0761) + \cdots + (0.12 \times 0.729)}{2.578} \\ &\approx 0.1471 \end{aligned}$$

As we bring in predictions from more and more models, and weight these predictions by how much they are supported by the data, we will find that the average will approach a probability of $1/7$. Perhaps, like me, you never cease to be amazed at how much was accomplished in the symbolic prediction formula when you become aware of much computational effort it would take to duplicate it.

Let's return now to the MEP. The MEP assigns numerical values to the probabilities of the three statements by inserting information, under the guise of some model \mathcal{M}_k , into the probability distribution. For model \mathcal{M}_A we have,

$$Q_i = P(X = x_i | \mathcal{M}_A) = \frac{\exp [\sum_{j=1}^m \lambda_j^A F_j(x_i)]}{Z(\lambda_1, \dots, \lambda_m)}$$

Since $n = 3$, the maximum that m can assume is $m = 2$. All of the models discussed here will consist of $m = 2$ constraint functions and their averages. Thus, the MEP formula becomes,

$$Q_i = P(X = x_i | \mathcal{M}_A) = \frac{\exp [\lambda_1^A F_1(x_i) + \lambda_2^A F_2(x_i)]}{Z(\lambda_1, \lambda_2)}$$

Establish some arbitrary mapping for each statement in the state space to the real line. Let this mapping from statements go to just 0 or 1 by setting up the following two constraint functions,

$$F_1(X = x_i) = (1, 0, 0) \text{ and } F_2(X = x_i) = (0, 1, 0)$$

Curiously, we see that the constraint function averages are literally specifying the probability for HEADS and TAILS. And this is allowed as part of the MEP formalism. It is worthwhile to explore the idea that information can directly specify a numerical assignment to a probability for a single statement,

$$\begin{aligned} \langle F_1 \rangle &= \sum_{i=1}^3 F_1(x_i) Q_i \\ &= Q_1 \\ \langle F_2 \rangle &= \sum_{i=1}^3 F_2(x_i) Q_i \\ &= Q_2 \end{aligned}$$

Suppose model \mathcal{M}_A inserts information under the guise that $\langle F_1 \rangle = Q_1 = 1/3$ and $\langle F_2 \rangle = Q_2 = 2/3$. We'll see why in just a moment. Then, of course, under this model, Q_3 must equal zero. The likelihood for the future data when conditioned on such a model is then,

$$\begin{aligned} P(M_1 = 0, M_2 = 1, M_3 = 1 | \mathcal{M}_A) &= W(M) \times \prod_{i=1}^3 Q_i^{M_i} \\ &= \frac{2!}{0! 1! 1!} \times (1/3)^0 \times (2/3)^1 \times 0^1 \\ &= 0 \end{aligned}$$

For model \mathcal{M}_F , on the other hand, the likelihood for the future data is,

$$\begin{aligned} P(M_1 = 0, M_2 = 1, M_3 = 1 \mid \mathcal{M}_F) &= W(M) \times \prod_{i=1}^3 Q_i^{M_i} \\ &= \frac{2!}{0! 1! 1!} \times (0.25)^0 \times (0.375)^1 \times (0.375)^1 \\ &= 0.28125 \end{aligned}$$

To calculate the ratio of the probability of any model \mathcal{M}_k , after the data, to the fixed baseline maximum likelihood model \mathcal{M}_A , we have the formula,

$$\frac{P(\mathcal{M}_k \mid \mathcal{D})}{P(\mathcal{M}_A \mid \mathcal{D})} = \exp \left[N \times \left(\sum_{j=1}^m (\lambda_j^k - \lambda_j^A) \bar{F}_j + \ln \left(\frac{Z_A}{Z_k} \right) \right) \right]$$

Suppose we are interested in finding the relative contribution to the overall average from model \mathcal{M}_G . The Lagrange multipliers and the partition functions for both models, with the sample averages are,

$$\lambda_1^A = 35.0482$$

$$\lambda_2^A = 35.7414$$

$$\lambda_1^G = 1.09861$$

$$\lambda_2^G = 1.79176$$

$$Z_A = 5 \times 10^{15}$$

$$Z_G = 10$$

$$\bar{F}_1 = 1/3$$

$$\bar{F}_2 = 2/3$$

The goal of assigning a 0 to Q_3 under model \mathcal{M}_A has caused the MEP algorithm to produce large values. Substituting these above numbers into the formula, we can calculate the relative contribution of model \mathcal{M}_G as compared to the maximum likelihood model \mathcal{M}_A .

$$\begin{aligned}
\frac{P(\mathcal{M}_G | \mathcal{D})}{P(\mathcal{M}_A | \mathcal{D})} &= \exp \left[N \times \left(\sum_{j=1}^2 (\lambda_j^G - \lambda_j^A) \bar{F}_j + \ln \left(\frac{Z_A}{Z_G} \right) \right) \right] \\
\sum_{j=1}^2 (\lambda_j^G - \lambda_j^A) \bar{F}_j &= [(1.09861 - 35.0482) \times 1/3] + [(1.79176 - 35.7414) \times 2/3] \\
&= -33.949623 \\
\ln \left(\frac{Z_A}{Z_G} \right) &= \ln \left(\frac{4.99312 \times 10^{15}}{10} \right) \\
&= 33.844252 \\
\sum_{j=1}^2 (\lambda_j^G - \lambda_j^A) \bar{F}_j + \ln \left(\frac{Z_A}{Z_G} \right) &= -0.105371 \\
\frac{P(\mathcal{M}_G | \mathcal{D})}{P(\mathcal{M}_A | \mathcal{D})} &= \exp [3 \times (-0.105371)] \\
&= 0.7290
\end{aligned}$$

To end this exercise, we would like to numerically investigate a variation on the formula for the weighting factor involved in the probability of future frequency counts conditioned on the known data. This formula uses the relative entropy and, to my way of thinking, indicates the relevant context where the relative entropy, or Kullback measure, *should* appear when making inferences.

Consider, then, this compact formula for the posterior ratio of the probability for any model to the baseline, maximum likelihood model \mathcal{M}_A , where now a minus sign intrudes into the formula,

$$\frac{P(\mathcal{M}_k | \mathcal{D})}{P(\mathcal{M}_A | \mathcal{D})} = \exp [-N \times KL(p, q)]$$

From our previous perspective, the maximum likelihood model \mathcal{M}_A assigned values that matched the normed frequency counts. Its relative weight compared to itself was always going to be 1. If this baseline model was compared to any other model, then its relative weight to all other models was always going to be greater than 1.

Of course, if it is more palatable, and this is what we have done in this exercise, we can always turn this around so that the fixed baseline model of the normed frequency counts has the maximum weight of 1. All other models will then have a weighting compared to the baseline model that is less than 1. Models far from the data will have a relative weight compared to the baseline model that is small, while models that approach the data ever more closely will approach a relative weight of 1.

Models with numerical assignments close to the data as a normed frequency count, $(1/3, 2/3, 0)$, will have relative weights close to 1, while models that are distant from model \mathcal{M}_A will have relative weights much less than 1. Averaging the future observations with respect to these weights, no matter which way you want to look at it, will provide the answer to $P(M_1, M_2, M_3 | \mathcal{D})$.

To start, take the log transformation of the likelihood ratio for, say, model \mathcal{M}_E compared to model \mathcal{M}_A .

$$\begin{aligned} \ln \left[\frac{P(\mathcal{D} | \mathcal{M}_E)}{P(\mathcal{D} | \mathcal{M}_A)} \right] &= \ln [P(\mathcal{D} | \mathcal{M}_E)] - \ln [P(\mathcal{D} | \mathcal{M}_A)] \\ &= \left(\ln W(N) + \sum_{i=1}^3 N_i \ln Q_i^E \right) - \left(\ln W(N) + \sum_{i=1}^3 N_i \ln Q_i^A \right) \\ &= \sum_{i=1}^3 N_i (\ln Q_i^E - \ln Q_i^A) \\ &= \sum_{i=1}^3 N_i \ln \left(\frac{Q_i^E}{Q_i^A} \right) \\ &= N \sum_{i=1}^3 \frac{N_i}{N} \ln \left(\frac{Q_i^E}{Q_i^A} \right) \end{aligned}$$

But model \mathcal{M}_A , as the maximum likelihood model, has Q_i^A exactly the same as the normed frequency counts $f_i \equiv N_i/N$. Therefore,

$$\ln \left[\frac{P(\mathcal{D} | \mathcal{M}_E)}{P(\mathcal{D} | \mathcal{M}_A)} \right] = N \sum_{i=1}^3 Q_i^A \ln \left(\frac{Q_i^E}{Q_i^A} \right)$$

The generic notation for the Kullback distance measure is,

$$KL(p, q) = \sum_{i=1}^n p_i \ln \left(\frac{p_i}{q_i} \right)$$

but we don't quite have this at this point in the derivation. Substituting p_i for Q_i^A and q_i for Q_i^E , we have instead,

$$KL^*(p, q) = \sum_{i=1}^n p_i \ln \left(\frac{q_i}{p_i} \right)$$

But this can be fixed through,

$$-(\ln p_i - \ln q_i) = \ln \left(\frac{q_i}{p_i} \right)$$

so let $KL^*(p, q) = -KL(p, q)$. To finish up,

$$\begin{aligned}\ln \left[\frac{P(\mathcal{D} | \mathcal{M}_E)}{P(\mathcal{D} | \mathcal{M}_A)} \right] &= N \sum_{i=1}^3 Q_i^A \ln \left(\frac{Q_i^E}{Q_i^A} \right) \\ \frac{P(\mathcal{D} | \mathcal{M}_E)}{P(\mathcal{D} | \mathcal{M}_A)} &= e^{-N [KL(p, q)]} \\ \frac{P(\mathcal{M}_E | \mathcal{D})}{P(\mathcal{M}_A | \mathcal{D})} &= \frac{P(\mathcal{D} | \mathcal{M}_E)}{P(\mathcal{D} | \mathcal{M}_A)} \times \frac{P(\mathcal{M}_E)}{P(\mathcal{M}_A)} \\ \frac{P(\mathcal{M}_E | \mathcal{D})}{P(\mathcal{M}_A | \mathcal{D})} &= e^{-N [KL(p, q)]}\end{aligned}$$

This formula explicitly shows the relative entropy, and to confirm that we arrive at the same relative weighting of model \mathcal{M}_E to model \mathcal{M}_A ,

$$\begin{aligned}KL(p, q) &= \sum_{i=1}^3 Q_i^A \ln \left(\frac{Q_i^A}{Q_i^E} \right) \\ &= \left[1/3 \ln \left(\frac{1/3}{0.20} \right) \right] + \left[2/3 \ln \left(\frac{2/3}{0.40} \right) \right] + \left[0 \ln \left(\frac{0}{0.40} \right) \right] \\ &= 0.510826 \\ \frac{P(\mathcal{M}_E | \mathcal{D})}{P(\mathcal{M}_A | \mathcal{D})} &= e^{-N [KL(p, q)]} \\ &= e^{-3 \times 0.510826} \\ &= 0.2160\end{aligned}$$

In Table 29.2, this same value was calculated by the formula,

$$\frac{P(\mathcal{M}_E | \mathcal{D})}{P(\mathcal{M}_A | \mathcal{D})} = \exp \left[N \times \left(\sum_{j=1}^2 (\lambda_j^E - \lambda_j^A) \bar{F}_j + \ln \left(\frac{Z_A}{Z_E} \right) \right) \right]$$

Chapter 30

The Gaussian Distribution

30.1 Introduction

The univariate and multivariate Gaussian distributions enjoyed the status of being the very first distributions that Shannon showed had maximum entropy subject to some specified information. This still represents the “crown jewel,” so to speak, in the efforts at justifying the idea of entropy as it relates to probability. This novel concept of *information entropy* could then be perceived as actually being useful in creating probability distributions out of whole cloth.

The Gaussian distribution, then and now, occupied the supreme position as the most ubiquitous probability interface to the real world. At the time, it must have come as a bit of a shock to the statistical establishment to experience the seemingly strange logic of Shannon’s demonstration.

It struck right at the heart of the prevailing rationale that probabilities were objective entities defined by long term frequencies. How could maximizing something called *missing information* result in a probability distribution? How could a constraint function and its average over the statements in the state space result in a probability distribution? How could a procedure that did not require, or ever mention the need for any data, result in a probability distribution?

It represented such a rupture to conventional thinking that, although nobody could deny Shannon’s demonstration that invoking entropy produced a Gaussian distribution, the mind recoiled at such an attack on the foundations of probability. *It simply wasn’t objective!* Hence, the pejorative term *subjective probability* was coined to forestall the incipient cognitive dissonance.

The shock waves from Shannon’s novel derivation of the Gaussian, subsequently followed by Jaynes’s all out assault using the MEP, still reverberate down to our own times. Eventually, various people started picking off, one by one, all of the conventional probability distributions. They demonstrated that, if one were creative enough, all these distributions were, in fact, MEP distributions.

As a matter of fact, there is no probability distribution that is not an MEP distribution. My personal favorite is given center stage in Chapter Thirty One. Here, I show that the Cauchy distribution is an MEP distribution. The Cauchy distribution has always held an unsavory reputation in the orthodox world. It was more often than not described as a “pathological” distribution.

The impetus for me personally was the many times I would ask someone to give me an example of a distribution that was not an MEP distribution. Most often mentioned was the Cauchy distribution, followed in popularity by the Student t -distribution. When I protested that these distributions had been shown many times over to be MEP distributions, I was thought delusional. In any case, countered my debaters, if these proofs did happen to exist somewhere, they must be wrong.

30.2 From the Discrete to the Continuous

This is the not the first time we have had to confront the major conceptual issue of transitioning from a discrete space to a continuous space. It cropped up initially in Volume I when the transition took place from the set of \mathcal{M} discrete models to models covering all conceivable numerical assignments to probabilities. The model space had to cover the real line from 0 to 1. Lately, we have seen that Planck’s oscillator permitted the number of energy levels to increase without bound.

Nonetheless, as far as the state space was concerned, most of our thinking up to now has been focused on a finite set of statements ($X = x_i$) consisting of n statements in total. Computations of constraint function averages, as well as partition functions, were done in the form of sums where i ranged from 1 to n .

Conceptually, we would like to cover the case where the number of statements n increases to an ever larger number. Eventually, we admit that we must take the plunge and let $n \rightarrow \infty$. We want the mathematics to handle the case where any observation can be as refined as we like.

The MEP formalism has worked just fine when statements naturally form a discrete set. We encountered no difficulties when the only observations that could be made were statements concerning the truth of HEADS or TAILS, or ONE through SIX, or PREFERS FOSTER’S and RIGHT–HANDED, and so on.

But if we want to allow ourselves the freedom of making a measurement of 27.067 units on some trial, we must transition to a continuous set. We think of this measurement as a statement, “The measurement was between 27.0665 units and 27.0675 units.” or a mapping from this statement to an interval on the real line, $27.0665 \leq F(X = x_i) \leq 27.0675$. Since Newton, Leibniz, and the invention of the calculus, we have accommodated our thinking about these situations by allowing the partitioning of the $F(X = x_i)$ to become as small as desired.

Any interval, no matter how small, can be integrated over to replace the summation that was used in the discrete case. The probability attached to any statement becomes an integration over a *probability density function*. The integration over all statements, or now the integration over the x -axis, must still integrate to 1.

Formerly, when we were involved with only discrete sets, we were permitted to say things like: The probability of HEADS is 1/2 under some model, or the probability of a THREE is 1/6 under some model. But now for the continuous case, we cannot say we are finding the probability for a measurement of exactly 27 units. We must employ the language that we are finding the probability, say, of a measurement between 26 and 28 units. This probability is the integration over the probability density function between the limits of 26 and 28.

Here is the new notation for the continuous case. It will be on display in the upcoming sections devoted to the Gaussian distribution. The probability for the interval of statements between the statement that the measurement had the lower value l and the upper value u , is written as,

$$P(l \leq x \leq u | \mathcal{M}_k) = \int_l^u pdf(x) dx \quad (30.1)$$

This probability is conditioned, as usual, on the supposition that some model is true. In this Chapter, we will be concerned with models that insert information that result in Gaussian distributions.

Since the mapping from the statements is to the entire real line, the probability for all of the statements covering the interval from the lower value of $-\infty$ to the upper value of $+\infty$ is a certainty,

$$P(-\infty \leq x \leq +\infty | \mathcal{M}_k) = \int_{-\infty}^{\infty} pdf(x) dx = 1 \quad (30.2)$$

The expectation, or average, of some constraint function is defined as,

$$E [F_j(x)] = \int_{-\infty}^{\infty} F_j(x) pdf(x) dx \quad (30.3)$$

Even this notation has some ambiguity attached to it. In order to conform to conventional ways of writing these expressions, take note that wherever x appears, it cannot refer to any *statement* ($X = x$). The usage of x in these expressions implies that statements have already been mapped to the real line.

30.3 A Standard MEP Formula

Suppose we were back in our comfortable discrete universe. If we wanted some numerical assignment as the probability for the i^{th} joint statement in the state space based on a model with two constraint functions, we would write out the

template for the generic MEP formula as,

$$P(X = x_i \mid \mathcal{M}_k) = \frac{\exp [\lambda_1 F_1(X = x_i) + \lambda_2 F_2(X = x_i)]}{Z(\lambda_1, \lambda_2)} \quad (30.4)$$

The partition function would be a sum over n ,

$$Z(\lambda_1, \lambda_2) = \sum_{i=1}^n \exp [\lambda_1 F_1(X = x_i) + \lambda_2 F_2(X = x_i)] \quad (30.5)$$

This MEP template can still be used in the continuous case with the necessary modifications,

$$pdf(x \mid \mathcal{M}_k) = \frac{e^{[\lambda_1 F_1(x) + \lambda_2 F_2(x)]}}{\int_{-\infty}^{\infty} e^{[\lambda_1 F_1(x) + \lambda_2 F_2(x)]} dx} \quad (30.6)$$

This is a model with two parameters as indicated by the two Lagrange multipliers.

30.4 MEP Characterization of a Gaussian

The Gaussian distribution, as with all probability distributions arriving from the orthodox approach, is a veritable *deus ex machina*. It descends from the clouds, appears on stage, and then solves all of our problems. The audience was never to question the timely introduction of the problem solving *deus*; just as we are never to question the introduction of some pertinent probability distribution!

We encountered this once before in our explanation of logistic regression in Chapter Twenty Three. The logistic sigmoid function is plucked from the shelf with nary a hint of any rhyme nor reason of any larger significance to our problem. In the conventional approach, the origin of any probability distribution is never deemed worthy of discussion.

We would prefer to wriggle out from this uncomfortable situation. We rely upon the MEP to provide a one size fits all justification for any probability distribution. Its basic rationale expunges all the mystery of how probability distributions are created, perhaps creating a duller world for those who crave the unexplainable.

But in a world of information, probability distributions assume their rightful place as an IP's state of knowledge about statements. The MEP solves the mysteries by introducing information in the form of constraint function averages, and missing information in the form of entropy. By leaning on the MEP, we do acquire at least some sense of the origin of the Gaussian, and the cloudy realms from which it descended.

To start, we are simply going to rearrange the standard way in which the Gaussian probability density function is usually presented. Personally, I prefer to start from the given equation for the density function. Then, I show through a series of straightforward algebraic manipulations how it matches up with the standard MEP formula. There is an attendant shock value for the uninitiated.

30.4.1 An algebraic re-arrangement

This section is tedious by its very nature. But it does possess the redeeming virtue of showing that the Gaussian as it is written in its standard format, can be directly matched up with the MEP formalism in Equation (30.6).

$$\begin{aligned}
 pdf(x | \mathcal{M}_k) &= \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2\sigma^2}(x - \mu)^2 \right] \\
 (x - \mu)^2 &= x^2 - 2x\mu + \mu^2 \\
 -\frac{1}{2\sigma^2}(x - \mu)^2 &= -\frac{x^2 - 2x\mu + \mu^2}{2\sigma^2} \\
 &= -\frac{x^2}{2\sigma^2} + \frac{2x\mu}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} \\
 -\frac{1}{2\sigma^2}(x - \mu)^2 &= \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} \\
 \exp \left[-\frac{1}{2\sigma^2}(x - \mu)^2 \right] &= \exp \left[\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} \right]
 \end{aligned}$$

With one final step, we accomplish our goal of rearranging the Gaussian distribution into an alternative format where we can pattern match with the generic MEP formula,

$$\frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2\sigma^2}(x - \mu)^2 \right] = \exp \left[\left(\frac{\mu}{\sigma^2} \right) x + \left(-\frac{1}{2\sigma^2} \right) x^2 - \frac{\mu^2}{2\sigma^2} - \ln(\sqrt{2\pi}\sigma) \right] \quad (30.7)$$

In this form, it is easy to see where the first and second constraint functions $F_1(x) = x$ and $F_2(x) = x^2$ occur. The accompanying Lagrange multipliers must then be,

$$\lambda_1 = \frac{\mu}{\sigma^2} \quad (30.8)$$

$$\lambda_2 = -\frac{1}{2\sigma^2} \quad (30.9)$$

We shall find out in an exercise that the information inserted into a Gaussian distribution, namely the constraint function averages, are respectively,

$$\langle F_1 \rangle = \mu \quad (30.10)$$

$$\langle F_2 \rangle = \mu^2 + \sigma^2 \quad (30.11)$$

The generic MEP formula for two parameters and these constraint functions must have this form for a Gaussian,

$$pdf(x | \mathcal{M}_k) = \exp [\lambda_1 x + \lambda_2 x^2 - \ln Z(\lambda_1, \lambda_2)] \quad (30.12)$$

with $Z(\lambda_1, \lambda_2)$ the left over terms not involving x or x^2 , that is,

$$\ln Z(\lambda_1, \lambda_2) = \frac{\mu^2}{2\sigma^2} + \ln (\sqrt{2\pi}\sigma) \quad (30.13)$$

$$Z(\lambda_1, \lambda_2) = \exp \left(\frac{\mu^2}{2\sigma^2} \right) \sqrt{2\pi}\sigma \quad (30.14)$$

To double-check the correctness of Equation (30.14), perform the integration as indicated in Equation (30.6) to find the partition function directly. The resulting integration reveals this strange looking expression,

$$Z(\lambda_1, \lambda_2) = \int_{-\infty}^{+\infty} e^{\lambda_1 x + \lambda_2 x^2} dx = \frac{e^{-\frac{(\lambda_1)^2}{4\lambda_2}} \sqrt{\pi}}{\sqrt{-\lambda_2}} \quad (30.15)$$

This expression in Equation (30.15) is nevertheless the same as in Equation (30.14). This will be proven in Exercise 30.8.1.

30.4.2 Another conventional expression

The Gaussian distribution is thus seen to be a legitimate and fully-fledged MEP distribution. It incorporates information satisfying two constraint functions while maximizing all of the missing information not contained in these two constraint functions. The above demonstration has accomplished that much for us.

There is nothing that prevents a constant being subtracted from $F_1(x)$ to form a different constraint function. If that constant is μ , then, $F_1^*(x) = x - \mu$. The expectation of $F_1^*(x)$ is, by definition, $\langle F_1^* \rangle = E[(x - \mu)] = 0$. The second constraint function becomes $F_2^*(x) = (x - \mu)^2$ with expectation $\langle F_2^* \rangle = E[(x - \mu)^2] = \sigma^2$

The first parameter λ_1 becomes 0, the second parameter stays at $\lambda_2 = -1/2\sigma^2$, and the partition function becomes $\sqrt{2\pi}\sigma$. This MEP probability distribution is once again in the form of the Gaussian distribution,

$$\begin{aligned} pdf(x | \mathcal{M}_k) &= \frac{\exp [\lambda_2 F_2^*(x)]}{Z(\lambda_1, \lambda_2)} \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-1/2 \left(\frac{x-\mu}{\sigma} \right)^2} \end{aligned} \quad (30.16)$$

30.4.3 The Legendre transformation

Use the Legendre transformation to find the entropy of the Gaussian distribution. Since the Gaussian is an MEP distribution, this information entropy must be the

largest possible for any distribution satisfying the constraints. Thus, all *missing information* has been maximized in a Gaussian distribution. We are assured that only the information about $E[(x - \mu)] = 0$ and $E[(x - \mu)^2] = \sigma^2$ has made its way into the distribution.

In applying the Legendre transformation, start out as always by showing the information entropy with the constraint function averages as the two arguments,

$$\begin{aligned} H_{max}(\langle F_1^* \rangle, \langle F_2^* \rangle) &= \ln Z(\lambda_1, \lambda_2) - \sum_{j=1}^2 \lambda_j \langle F_j^* \rangle \\ &= \ln Z(\lambda_1, \lambda_2) - \lambda_1 \langle F_1^* \rangle - \lambda_2 \langle F_2^* \rangle \\ \langle F_1^* \rangle &= 0 \\ H_{max}(\langle F_1^* \rangle, \langle F_2^* \rangle) &= \ln Z(\lambda_1, \lambda_2) - \lambda_2 \langle F_2^* \rangle \\ &= \ln Z(\lambda_1, \lambda_2) - \left(-\frac{1}{2\sigma^2} \times \sigma^2 \right) \\ &= \ln Z(\lambda_1, \lambda_2) + 1/2 \\ &= \ln (\sqrt{2\pi} \sigma) + 1/2 \end{aligned}$$

Thus, the information entropy of any univariate Gaussian depends only σ . When different models insert information in the form of larger σ , then the information entropy of that new Gaussian increases.

All this makes perfect sense because σ measures the “spread” of the Gaussian curve. A larger σ signifies a broader distribution, higher entropy, and therefore more missing information.

Curiously, the entropy does not depend on the parameter μ , the location of the Gaussian along the x -axis. As a consequence, if two distributions have different locations but equal spread, the missing information is the same for both distributions. As already mentioned, the entropy for any Gaussian with a specified μ and σ has the largest entropy possible of any continuous distribution satisfying the two constraints $E[(x - \mu)] = 0$ and $E[(x - \mu)^2] = \sigma^2$, otherwise the Gaussian wouldn’t be an MEP distribution.

30.5 Probability under a Gaussian Model

A concrete example can now be given of calculating a probability for all statements in some interval when the model is a Gaussian. The Gaussian is an MEP model with two parameters.

The parameters come in dual forms. One form is the Lagrange multipliers λ_1 and λ_2 , while the dual form is the constraint function averages $\langle F_1 \rangle$ and $\langle F_2 \rangle$. These dual parameters express the same information under some model.

Suppose that the inferential problem involves an IP's state of knowledge about some measurement, but the IP doesn't want to place any restrictions on this quantitative observation. The state space is then allowed to have statements about any infinitely small interval over the entire x -axis from $-\infty$ to $+\infty$. What is the probability for the measurement to fall between, say, 26 and 28 units under some given model?

If we specify an MEP model with parameters $\mu = 27$ and $\sigma^2 = 9$ in order to implement a Gaussian distribution, then the probability density function under this model is,

$$pdf(x | \mathcal{M}_k) = \frac{1}{3\sqrt{2\pi}} e^{-1/2 (\frac{x-27}{3})^2}$$

The probability for the measurement to be between the given lower and upper limits of 26 and 28 is,

$$\begin{aligned} P(26 \leq x \leq 28 | \mathcal{M}_k) &= \int_{26}^{28} \frac{1}{3\sqrt{2\pi}} e^{-1/2 (\frac{x-27}{3})^2} dx \\ &= 0.2611 \end{aligned}$$

It would make sense that as the measurement interval shrinks, the probability should decrease as well. Thus, we find that,

$$P(26.99 \leq x \leq 27.01 | \mathcal{M}_k) = 0.00266$$

$$P(26.999 \leq x \leq 27.001 | \mathcal{M}_k) = 0.000266$$

The actual value of the Gaussian pdf curve at $x = 27$ is,

$$\frac{1}{3\sqrt{2\pi}} e^{-1/2 (\frac{27-27}{3})^2} = 0.132981$$

Now, of course, we realize that this number is not the probability for $x = 27$. But to approximate the integral, we erect a little rectangle with base length, say, of 0.002 with x in the center corresponding to $26.999 \leq x \leq 27.001$. The vertical dimension of the rectangle at x is 0.132981. The area of this rectangle is,

$$\text{Area of rectangle} = 0.002 \times 0.132981 = 0.000266$$

which approximates the actual value of the integral as just calculated.

It is easy to visualize that as the measurement interval is shrinking, the base of the rectangle along the x -axis is shrinking along with it. The area of the rectangle, with the vertical dimension staying fixed at the pdf value, represents the probability for this ever shrinking interval. As expected, this probability is becoming smaller and smaller. This is just another way of thinking about the state space as the dimension $n \rightarrow \infty$.

30.6 Relative Entropy Expressions

With this development of the Gaussian as an MEP distribution, it is interesting to take a look at some relative entropy expressions. Foreshadowing our preferred choice of the traditional nomenclature when we delve into *Information Geometry* in Volume III, adopt the convention that probability distributions are points in a Riemannian manifold. Let these probability distributions as points be labeled as point p , point q , point r , and so on.

The expressions for relative entropy would then look like,

$$KL(p, q) = \int_{-\infty}^{\infty} p \ln \left(\frac{p}{q} \right) dx \quad (30.17)$$

$$= E_p \left[\ln \left(\frac{p}{q} \right) \right] \quad (30.18)$$

Evaluating $\ln \left(\frac{p}{q} \right)$ then becomes important for two reasons. First, as just seen, it is involved in the relative entropy expression. Secondly, this expression is the log likelihood ratio, obviously something we need to calculate as well because of its importance in the reorientation of model space after data have been collected.

By Kullback's own definition, the point p is,

$$pdf(x | \mathcal{M}_1) \equiv pdf(x | \mu_p, \sigma_p^2) \equiv p$$

and the point q is,

$$pdf(x | \mathcal{M}_2) \equiv pdf(x | \mu_q, \sigma_q^2) \equiv q$$

For two Gaussian models,

$$\begin{aligned} \ln \left(\frac{p}{q} \right) &= \ln p - \ln q \\ &= \frac{1}{2\sigma_q^2} (x - \mu_q)^2 - \frac{1}{2\sigma_p^2} (x - \mu_p)^2 + \ln \left(\frac{\sigma_q}{\sigma_p} \right) \end{aligned} \quad (30.19)$$

This involves a somewhat lengthy derivation carried out in Exercise 30.8.5 with a numerical example in Exercise 30.8.6.

Prior to this exercise though, it is even more instructive to look at the relative entropy in Equation (30.18) when put into this format,

$$E_p \left[\ln \left(\frac{p}{q} \right) \right] = E_p [\ln p] - E_p [\ln q]$$

Concentrate on the first term,

$$E_p [\ln p] = \int_{-\infty}^{\infty} p \ln p dx = \int_{-\infty}^{\infty} pdf(x | \mathcal{M}_1) \ln [pdf(x | \mathcal{M}_1)] dx$$

which is the continuous analog to the negative of the discrete information entropy,

$$-H(Q_i) = \sum_{i=1}^n Q_i \ln Q_i$$

Since,

$$\ln p = \lambda_1^p F_1(x) + \lambda_2^p F_2(x) - \ln Z_p$$

the expectation of $\ln p$ is,

$$E_p [\ln p] = E_p [\lambda_1^p F_1(x) + \lambda_2^p F_2(x) - \ln Z_p]$$

Distributing the expectation operator across these terms results in known constraint function averages,

$$\begin{aligned} E_p [\ln p] &= E_p [\lambda_1^p F_1(x) + \lambda_2^p F_2(x) - \ln Z_p] \\ &= \lambda_1^p E_p [F_1(x)] + \lambda_2^p E_p [F_2(x)] - E_p [\ln Z_p] \\ &= \lambda_1^p \langle F_1 \rangle_p + \lambda_2^p \langle F_2 \rangle_p - \ln Z_p \end{aligned}$$

So we have,

$$-\int_{-\infty}^{\infty} p \ln p dx = \ln Z_p - \sum_{j=1}^2 \lambda_j^p \langle F_j \rangle_p$$

which we know from the Legendre transformation is the maximum information entropy of the MEP distribution.

We have already calculated the maximum entropy of the Gaussian distribution to be,

$$H_{max}(\langle F_1 \rangle, \langle F_2 \rangle) = \ln (\sqrt{2\pi} \sigma) + 1/2$$

Therefore,

$$\int_{-\infty}^{\infty} p \ln p dx = -\ln (\sqrt{2\pi} \sigma) - 1/2$$

Let's see if this is consistent by substituting back in the values for the parameters and the averages of the constraint functions for $E_p [\ln p]$,

$$\begin{aligned} \lambda_1^p \langle F_1 \rangle_p + \lambda_2^p \langle F_2 \rangle_p - \ln Z_p &= \frac{\mu_p}{\sigma_p^2} \mu_p - \frac{1}{2\sigma_p^2} (\mu_p^2 + \sigma_p^2) - \ln Z_p \\ &= \frac{\mu_p^2}{\sigma_p^2} - \frac{\mu_p^2 - \sigma_p^2}{2\sigma_p^2} - \ln Z_p \\ &= \frac{2\mu_p^2 - \mu_p^2 - \sigma_p^2}{2\sigma_p^2} - \ln Z_p \\ &= \frac{\mu_p^2 - \sigma_p^2}{2\sigma_p^2} - \ln Z_p \\ &= \frac{\mu_p^2}{2\sigma_p^2} - \frac{1}{2} - \ln Z_p \end{aligned}$$

Refer back to Equation (30.13) to find $\ln Z_p$,

$$\begin{aligned}
 E_p [\ln p] &= \frac{\mu_p^2}{2\sigma_p^2} - \frac{1}{2} - \ln Z_p \\
 \ln Z_p &= \frac{\mu_p^2}{2\sigma_p^2} + \ln(\sqrt{2\pi}\sigma_p) \\
 E_p [\ln p] &= \frac{\mu_p^2}{2\sigma_p^2} - \frac{1}{2} - \frac{\mu_p^2}{2\sigma_p^2} - \ln(\sqrt{2\pi}\sigma_p) \\
 &= -\frac{1}{2} - \ln(\sqrt{2\pi}\sigma_p) \\
 H_{max}(p) &= -E_p [\ln p] \\
 &= \ln(\sqrt{2\pi}\sigma_p) + \frac{1}{2} \tag{30.20}
 \end{aligned}$$

The same sort of general attack can be conducted on $E_p [\ln q]$. Substitute Equations (30.8) through (30.13),

$$\begin{aligned}
 -E_p [\ln q] &= -(\lambda_1^q \langle F_1 \rangle_p + \lambda_2^q \langle F_2 \rangle_p - \ln Z_q) \\
 &= -\frac{\mu_q}{\sigma_q^2} \mu_p + \frac{1}{2\sigma_q^2} (\mu_p^2 + \sigma_p^2) + \ln Z_q \\
 &= \frac{-2\mu_q\mu_p + \mu_p^2 + \sigma_p^2}{2\sigma_q^2} + \frac{\mu_q^2}{2\sigma_q^2} - \frac{\mu_q^2}{2\sigma_q^2} + \ln Z_q \\
 &= \frac{-2\mu_q\mu_p + \mu_p^2 + \mu_q^2}{2\sigma_q^2} + \frac{\sigma_p^2}{2\sigma_q^2} - \frac{\mu_q^2}{2\sigma_q^2} + \ln Z_q \\
 &= \frac{(\mu_p - \mu_q)^2}{2\sigma_q^2} + \frac{\sigma_p^2}{2\sigma_q^2} - \frac{\mu_q^2}{2\sigma_q^2} + \ln Z_q \\
 &= \frac{(\mu_p - \mu_q)^2}{2\sigma_q^2} + \frac{\sigma_p^2}{2\sigma_q^2} - \frac{\mu_q^2}{2\sigma_q^2} + \frac{\mu_q^2}{2\sigma_q^2} + \ln(\sqrt{2\pi}\sigma_q) \\
 &= \frac{(\mu_p - \mu_q)^2}{2\sigma_q^2} + \frac{\sigma_p^2}{2\sigma_q^2} + \ln(\sqrt{2\pi}\sigma_q)
 \end{aligned}$$

Put these two pieces back together (see Exercise 30.8.7) to yield as a final answer to the relative entropy,

$$KL(p, q) \equiv E_p \left[\ln \left(\frac{p}{q} \right) \right] = \ln \left(\frac{\sigma_q}{\sigma_p} \right) + \frac{(\mu_p - \mu_q)^2}{2\sigma_q^2} + \frac{\sigma_p^2}{2\sigma_q^2} - 1/2 \tag{30.21}$$

30.7 Connections to the Literature

An exhaustive set of references to the Gaussian, or Normal, distribution, would alone constitute a rather lengthy book. Therefore, I must content myself with just a couple of references more germane to our peculiar way of looking at things.

Shannon [31, pg. 87–91] discusses certain properties characterizing the entropy of continuous distributions. Here he presents his startling (to the conventional wisdom of the day) maximum entropy characterization of Gaussian distributions. But what I find more fascinating are his comments about the difference between discrete and continuous entropies.

In the discrete case the entropy measures in an absolute way the randomness of the chance variable. In the continuous case the measurement is *relative to the coordinate system*. If we change coordinates the entropy will in general change. In fact if we change to coordinates $y_1 \dots y_n$ the new entropy is given by

$$H(y) = \int \dots \int p(x_1 \dots x_n) J\left(\frac{x}{y}\right) \log p(x_1 \dots x_n) J\left(\frac{x}{y}\right) dy_1 \dots dy_n$$

where $J\left(\frac{x}{y}\right)$ is the Jacobian of the coordinate transformation. . . . In the continuous case the entropy can be considered a measure of randomness *relative to an assumed standard*, namely the coordinate system chosen with each small volume element $dx_1 \dots dx_n$ given equal weight. When we change the coordinate system the entropy in the new system measures the randomness when equal volume elements $dy_1 \dots dy_n$ in the new system are given equal weight. [Emphasis in the original.]

Thus, the seemingly universal acceptance of the perceived defect in Shannon's translation of discrete entropy over to continuous entropy seems to demand a more critical appraisal. We will reserve this topic for the next Volume.

In section 30.4.1, the approach taken to illustrate the fact that the Gaussian follows the MEP formula template was inspired by Amari's [1, pp. 85–86] very similar explanation. Accompanying Amari's commentary, there is a phrase which is in sore need of disambiguation! After writing down the standard expression for a Gaussian, he says,

This is the 2-dimensional space formed by the normal distribution . . .

But the state space for the Gaussian is, in fact, the space over the real line. Conceptually, we let the *dimension* of the state space $n \rightarrow \infty$. It is true that the dimension of the *parameter* space is only 2 because there are two Lagrange multipliers. Or, in other words, we might as well be talking about the dimension of the model space which, if you have accompanied me on this journey, is conceptually quite distinct from the state space. To my mind, those who conceptually never entertain the idea of model space when talking about probability are particularly prone to unnecessary ambiguity in this area.

30.8 Solved Exercises for Chapter Thirty

Exercise 30.8.1: Show that the integration for the partition function in Equation (30.15) works out to Equation (30.14).

Solution to Exercise 30.8.1

The parameters λ_1 and λ_2 have already been expressed in terms of the dual parameters μ and σ .

$$\lambda_1 = \frac{\mu}{\sigma^2}$$

$$\lambda_2 = -\frac{1}{2\sigma^2}$$

So the algebraic expressions appearing in the integration of Equation (30.15) can be worked out to be equivalent to Equation (30.14),

$$\begin{aligned} (\lambda_1)^2 &= \frac{\mu^2}{\sigma^4} \\ 4\lambda_2 &= -\frac{4}{2\sigma^2} \\ -\frac{(\lambda_1)^2}{4\lambda_2} &= -\frac{\frac{\mu^2}{\sigma^4}}{-\frac{4}{2\sigma^2}} \\ &= \frac{\mu^2}{2\sigma^2} \\ \exp\left[-\frac{(\lambda_1)^2}{4\lambda_2}\right] &= \exp\left[\frac{\mu^2}{2\sigma^2}\right] \\ \sqrt{-\lambda_2} &= \sqrt{-\left(-\frac{1}{2\sigma^2}\right)} \\ \frac{\sqrt{\pi}}{\sqrt{\frac{1}{2\sigma^2}}} &= \sqrt{2\pi}\sigma \\ \frac{\exp\left[-\frac{(\lambda_1)^2}{4\lambda_2}\right]\sqrt{\pi}}{\sqrt{-\lambda_2}} &= \exp\left(\frac{\mu^2}{2\sigma^2}\right)\sqrt{2\pi}\sigma \end{aligned}$$

The *Mathematica* code to implement the integration in Equation (30.15) is,

```
Integrate[Exp[\lambda_1 x + \lambda_2 x^2], \{-\infty, \infty\}, Assumptions \rightarrow \lambda_2 < 0]
```

Exercise 30.8.2: What are the constraint function averages for the Gaussian distribution when derived from the MEP perspective?

Solution to Exercise 30.8.2

Do not rely on familiarity with the Gaussian distribution, but adhere strictly to the MEP formalism in order to claim that the constraint function averages are,

$$\langle F_1 \rangle = \frac{\partial \ln Z}{\partial \lambda_1}$$

$$\langle F_2 \rangle = \frac{\partial \ln Z}{\partial \lambda_2}$$

We already know that the partition function has been evaluated as,

$$Z(\lambda_1, \lambda_2) = \int_{-\infty}^{+\infty} e^{\lambda_1 x + \lambda_2 x^2} dx = \frac{e^{-\frac{(\lambda_1)^2}{4\lambda_2}} \sqrt{\pi}}{\sqrt{-\lambda_2}}$$

Take the partial derivative of $\ln Z$ with respect to the first Lagrange multiplier,

$$\frac{\partial \ln Z}{\partial \lambda_1} = -\frac{\lambda_1}{2\lambda_2}$$

$$\lambda_1 = \frac{\mu}{\sigma^2}$$

$$\lambda_2 = -\frac{1}{2\sigma^2}$$

$$-\frac{\lambda_1}{2\lambda_2} = -\frac{\frac{\mu}{\sigma^2}}{2(-\frac{1}{2\sigma^2})}$$

$$\langle F_1 \rangle = \mu$$

The expectation of the first constraint function $F_1(x) = x$ is $\langle F_1 \rangle = \mu$. Take the partial derivative of $\ln Z$ with respect to the second Lagrange multiplier,

$$\frac{\partial \ln Z}{\partial \lambda_2} = \frac{(\lambda_1)^2 - 2\lambda_2}{4(\lambda_2)^2}$$

$$\lambda_1 = \frac{\mu}{\sigma^2}$$

$$\lambda_2 = -\frac{1}{2\sigma^2}$$

$$\frac{(\lambda_1)^2 - 2\lambda_2}{4(\lambda_2)^2} = \frac{\left(\frac{\mu}{\sigma^2}\right)^2 - 2\left(-\frac{1}{2\sigma^2}\right)}{4\left(-\frac{1}{2\sigma^2}\right)^2}$$

$$\langle F_2 \rangle = \mu^2 + \sigma^2$$

The expectation of the second constraint function $F_2(x) = x^2$ is $\langle F_2 \rangle = \mu^2 + \sigma^2$.

Exercise 30.8.3: What is the expectation of $F_2^*(x)$?

Solution to Exercise 30.8.3

In the discrete case where the state space consisted of n mutually exclusive and exhaustive statements ($X = x_i$), an expectation of some j^{th} constraint function with respect to the numerical assignments for the probability of the statements under some model was of the form,

$$\langle F_j \rangle = \sum_{i=1}^n F_j(X = x_i) Q_i$$

In transitioning to the continuous case, the expectation becomes an integration instead of a sum,

$$\langle F_j \rangle = \int_{-\infty}^{\infty} F_j(x) \text{pdf}(x | \mathcal{M}_k) dx$$

where the mapping from the n statements has gone to $F_j(x)$ with $F_j(x)$ ranging over the real line from $-\infty$ to $+\infty$.

$F_2^*(x)$ was defined as $(x - \mu)^2$. Thus, the expectation of this function with respect to the Gaussian is,

$$\begin{aligned} \langle F_2^* \rangle &= \int_{-\infty}^{\infty} F_2^*(x) \text{pdf}(x | \mathcal{M}_k) dx \\ &= \int_{-\infty}^{\infty} (x - \mu)^2 \times \frac{e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}}{\sqrt{2\pi}\sigma} dx \\ &= \sigma^2 \end{aligned}$$

Exercise 30.8.4: What is the expectation of $F_1^*(x)$?

Solution to Exercise 30.8.4

$F_1^*(x)$ was defined as $x - \mu$. Thus, the expectation of this function with respect to the Gaussian is,

$$\begin{aligned} \langle F_1^* \rangle &= \int_{-\infty}^{\infty} F_1^*(x) \text{pdf}(x | \mathcal{M}_k) dx \\ &= \int_{-\infty}^{\infty} (x - \mu) \times \frac{e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}}{\sqrt{2\pi}\sigma} dx \\ &= 0 \end{aligned}$$

Exercise 30.8.5: What is the log likelihood ratio for a single observation given the information under two different Gaussian distributions?

Solution to Exercise 30.8.5

In the past, we would have written this as,

$$\text{log-likelihood ratio} = \frac{P(X = x_i | \mathcal{M}_A)}{P(X = x_i | \mathcal{M}_B)}$$

For Gaussian distributions, we now write,

$$\text{log-likelihood ratio} = \frac{\text{pdf}(x | \mu_p, \sigma_p^2)}{\text{pdf}(x | \mu_q, \sigma_q^2)}$$

Following along from our initial exposure to *Information Geometry*, and simply for convenience's sake, let the first Gaussian distribution be called point p and the second Gaussian distribution point q . The log likelihood ratio is then,

$$\ln \left(\frac{p}{q} \right) = \ln p - \ln q$$

Employing the constraint function $F_2^*(x) = (x - \mu)^2$, the typical format for the log transform of an MEP distribution appears as,

$$\ln p = \lambda_2^p F_2^p(x) - \ln Z_p$$

$$\ln q = \lambda_2^q F_2^q(x) - \ln Z_q$$

$$\ln \left(\frac{p}{q} \right) = \lambda_2^p F_2^p(x) - \ln Z_p - (\lambda_2^q F_2^q(x) - \ln Z_q)$$

$$= \lambda_2^p F_2^p(x) - \lambda_2^q F_2^q(x) + \ln \left(\frac{Z_q}{Z_p} \right)$$

The substitution of the parameters and the constraint functions then yield,

$$\lambda_2^p = -\frac{1}{2\sigma_p^2}$$

$$\lambda_2^q = -\frac{1}{2\sigma_q^2}$$

$$F_2^p(x) = (x - \mu_p)^2$$

$$F_2^q(x) = (x - \mu_q)^2$$

$$\ln \left(\frac{p}{q} \right) = \left[-\frac{1}{2\sigma_p^2} (x - \mu_p)^2 \right] - \left[-\frac{1}{2\sigma_q^2} (x - \mu_q)^2 \right] + \ln \left(\frac{Z_q}{Z_p} \right)$$

$$= \frac{1}{2\sigma_q^2} (x - \mu_q)^2 - \frac{1}{2\sigma_p^2} (x - \mu_p)^2 + \ln \left(\frac{Z_q}{Z_p} \right)$$

Take the log transform of both partition functions,

$$\ln Z_q = 1/2 \ln 2 + 1/2 \ln \pi + \ln \sigma_q$$

$$\ln Z_p = 1/2 \ln 2 + 1/2 \ln \pi + \ln \sigma_p$$

$$\ln \left(\frac{Z_q}{Z_p} \right) = \ln \sigma_q - \ln \sigma_p$$

$$\text{log-likelihood ratio} = \frac{1}{2\sigma_q^2} (x - \mu_q)^2 - \frac{1}{2\sigma_p^2} (x - \mu_p)^2 + \ln \left(\frac{\sigma_q}{\sigma_p} \right)$$

Exercise 30.8.6: Write some *Mathematica* code to construct a numerical example confirming the formula for the log-likelihood ratio as derived in the previous exercise.

Solution to Exercise 30.8.6

Suppose that we look at the Gaussian probability density function values at $x = 26$ for one model with the information reflected by the parameters $\mu_p = 28$ and $\sigma_p^2 = 9$ as compared to another model with information of $\mu_q = 25$ and $\sigma_q^2 = 16$. According to the last exercise,

$$\begin{aligned} \text{log-likelihood ratio} &= \frac{1}{2\sigma_q^2} (x - \mu_q)^2 - \frac{1}{2\sigma_p^2} (x - \mu_p)^2 + \ln \left(\frac{\sigma_q}{\sigma_p} \right) \\ &= \left[\frac{1}{2 \times 16} \times (26 - 25)^2 \right] - \left[\frac{1}{2 \times 9} \times (26 - 28)^2 \right] + \left[\ln \left(\frac{4}{3} \right) \right] \\ &= 0.0967099 \end{aligned}$$

Confirming this result with *Mathematica*, we find that *Mathematica* evaluates the expression,

```
Log[N[PDF[NormalDistribution[28,3],26]] / 
N[PDF[NormalDistribution[25,4],26]] 
(* end Log *)]
```

as 0.0967099.

Exercise 30.8.7: Finish the derivation started in section 30.6 showing the relative entropy between two Gaussian distributions.

Solution to Exercise 30.8.7

The two expectations with respect to p appearing in the relative entropy,

$$KL(p, q) = E_p [\ln p - \ln q]$$

were found in section 30.6 as,

$$\begin{aligned} E_p [\ln p] &= -\ln(\sqrt{2\pi}\sigma_p) - 1/2 \\ -E_p [\ln q] &= \frac{(\mu_p - \mu_q)^2}{2\sigma_q^2} + \frac{\sigma_p^2}{2\sigma_q^2} + \ln(\sqrt{2\pi}\sigma_q) \\ E_p [\ln p - \ln q] &= -\ln(\sqrt{2\pi}\sigma_p) - 1/2 + \frac{(\mu_p - \mu_q)^2}{2\sigma_q^2} + \frac{\sigma_p^2}{2\sigma_q^2} + \ln(\sqrt{2\pi}\sigma_q) \\ KL(p, q) &= \ln\left(\frac{\sigma_q}{\sigma_p}\right) + \frac{(\mu_p - \mu_q)^2}{2\sigma_q^2} + \frac{\sigma_p^2}{2\sigma_q^2} - 1/2 \end{aligned}$$

Exercise 30.8.8: Confirm that a necessary condition holds for the above relative entropy.

Solution to Exercise 30.8.8

If the coordinates for the two points p and q are the same, then it follows that the distance separating them must be 0. If $\mu_p = \mu_q$ and $\sigma_p^2 = \sigma_q^2$, then,

$$\begin{aligned} KL(p, q) &= \ln\left(\frac{\sigma_q}{\sigma_p}\right) + \frac{(\mu_p - \mu_q)^2}{2\sigma_q^2} + \frac{\sigma_p^2}{2\sigma_q^2} - 1/2 \\ &= \ln 1 + 0 + 1/2 - 1/2 \\ &= 0 \end{aligned}$$

Chapter 31

The Cauchy Distribution

31.1 Introduction

I was utterly fascinated when it first dawned on me that the Cauchy distribution was just another MEP distribution. All my life I had been brainwashed by my orthodox training to think of the Cauchy distribution as some sort of strange “pathological” distribution.

I chuckled sympathetically when Jaynes recounted the story of having one of his papers on the Cauchy distribution twice rejected by a referee who basically told him there was no future in even beginning to contemplate such monstrous aberrations of probability theory.

I also had felt the sting of contempt from established statisticians when, on various occasions, I had inquired as to exactly which distributions were not MEP distributions. I was understandably curious as to the nature of these mysterious probability distributions, not having discovered any myself that were not derivable via the Maximum Entropy Principle.

Invariably, after posing this question, I would be offered a list of potential candidates such as the Cauchy and Student-*t* distributions. I interjected that the characterization of these distributions as maximum entropy distributions had already appeared many times over in the literature.

The brainwashing undergone by these individuals invariably led to the dismissal of any probability distribution via the MEP as some sort of curious and ill-conceived fad. The refusal to accept any MEP derivation was so ingrained that my protestations that these derivations existed were met with the response that even if they did, they most certainly were wrong. Shades of Galileo trying to convince the Cardinals that the moons of Jupiter could actually be seen through his telescope.

In this Chapter, I shall derive the simple Cauchy distribution and some of its generalizations via the Maximum Entropy Principle. These Cauchy distributions are thus ordinary numerical assignments to probabilities, no different than any other “non-pathological” distribution. During this enterprise we might have to change some notation, as well as delve into a little more detail than is usually done. As is typically my style, I like to place a great deal of emphasis on the **conceptual** issues involved, both before and after the mathematical details.

I begin with perhaps the most crucial **conceptual** point. What is the point of the Maximum Entropy Principle? As I have been saying all along, it happens to be a very good algorithm for assigning numerical values to probabilities according to some model. This model includes the desired information, and excludes all of the unwanted information.

An Information Processor is allowed to insert information into a probability distribution to represent a state of knowledge. The information takes the form of one or several constraint functions on the state space, together with the expected values of those constraint functions. The resulting state of knowledge is reflected in a probability distribution as calculated via the MEP formula. It is *one* possible numerical assignment to the abstract probabilities under a model that incorporates very well defined information. It is not to be thought of, in any sense, as the “true” or “correct” assignment, but merely one compatible with the prescribed information.

Any algorithm, like the MEP algorithm, that makes legitimate assignments to probabilities is completely **orthogonal** to what have been labeled as the formal manipulation rules of probability theory. They exist in different universes. The manipulation rules, like Bayes’s Theorem, have no concept of how to make numerical assignments, and could care less that they lack this capacity. Their own sense of justification centers around the fact that they are mathematical theorems which are correct for any and all legitimate numerical assignments.

Likewise, the MEP doesn’t have anything to say about abstract probabilities, and how they might be manipulated via some axiomatic system. The MEP is concerned only with assigning legitimate numerical values to probabilities. These assignments *are*, however, always subject to the formal manipulation rules. It is as if these rules were inviolable laws of nature which the MEP must adhere to, but has no way of understanding.

Think about it this way. Probability, acting as a generalization of Classical Logic and relying upon its formal rules of manipulation, will tell us what is an acceptable inference at an abstract level. *After* this much has been accomplished, the MEP is an excellent means for extracting some sense of how the degree of belief in the truth of one of the statements might be dependent upon other statements thought of as causal factors. The MEP will then motivate us to think about causal linkages amongst all the variables from the perspective of *missing information*.

31.2 MEP Approach to a Cauchy Distribution

Let's not be disingenuous and pretend that we will reveal some shocking new MEP formula for the Cauchy distribution at the end of our derivation. Everyone has seen the standard Cauchy formula, together with its various generalizations. The MEP will provide us with the same expression for the Cauchy distribution.

Nonetheless, it *is* interesting to wonder how one might arrive at the standard Cauchy expression when starting from the seemingly different template of the MEP formula. The most primitive expression for the probability density function of a Cauchy distribution looks like this,

$$pdf(x | \mathcal{M}_{Ca}) = \frac{1}{\pi} \frac{1}{1+x^2} \quad (31.1)$$

What state space have we defined for the simplest of the Cauchy distributions? Granted that we have already admitted to a probability density function, we must be dealing with a continuous, instead of a discrete, state space.

Let the state space be indicated by x , and let it extend from $-\infty < x < \infty$. So the state space we are dealing with is no different than the state space of concern to the Gaussian distribution in the last Chapter. Perhaps this is an inkling of the enormous space of potential models involved in the realm of continuous state spaces.

Remarkably, it turns out that only one constraint function, and its average, are needed as the information to derive this simplest Cauchy distribution. In the conventional orthodox approach to probability theory, the density functions are always mysterious entities with no unifying principle informing us as to how they were derived.

That's why I like to label them as *deus ex machina*. They appear on the scene fully formed and ready to solve our inferential problem, but we haven't a clue as to how they got there.

Contrary to this perplexing situation with regard to the origin of some arcane probability density function, we always know the genesis of a probability distribution or density function when we rely upon the MEP. One of the nice things about the MEP formula is that, right at the start of the numerical assignment process for some new problem, and when we are most vulnerable to being temporarily adrift, we always have the MEP formula template to guide us on our journey.

The state of knowledge about x provided by the MEP formula for just one constraint looks like this in its generic representation for a continuous state space,

$$pdf(x | \mathcal{M}_k) = \frac{e^{\lambda F(x)}}{\int_{-\infty}^{\infty} e^{\lambda F(x)} dx} \quad (31.2)$$

What turns out in the end to demand creativity is not the MEP formula itself, but rather the question of what should serve as a constraint function. The MEP

never placed any restrictions on the mapping from statements in the state space to numbers, so we enjoy complete freedom in this area. Once people realized the freedom in constructing the mapping $F(x)$, a whole new realm opened up.

Thus, let the following creative mapping $F(x)$ be inserted as the constraint function,

$$F(x) = \ln(1 + x^2) \quad (31.3)$$

The expected value of this constraint function is set at,

$$E[F(x)] = E[\ln(1 + x^2)] = \langle F \rangle = 1.38629 \quad (31.4)$$

as the information under model $\mathcal{M}_k \equiv \mathcal{M}_{Ca}$. What in the world would lead one to specify such a seemingly unmotivated number as a constraint function average?

31.3 The Derivation

Substitute the given constraint function in the generic representation,

$$pdf(x | \mathcal{M}_{Ca}) = \frac{e^{\lambda \ln(1+x^2)}}{Z(\lambda)} \quad (31.5)$$

followed by the obvious transformation in the numerator,

$$pdf(x | \mathcal{M}_{Ca}) = \frac{(1+x^2)^\lambda}{Z(\lambda)} \quad (31.6)$$

Admittedly, with one eye on our ultimate goal, let the Lagrange multiplier take on the value $\lambda = -1$ in order to satisfy the same information given in the expected value $\langle F \rangle = 1.38629$.

Then,

$$pdf(x | \mathcal{M}_{Ca}) = \frac{1}{Z(-1)} \frac{1}{1+x^2} \quad (31.7)$$

The universal constraint that every probability distribution and probability density function must satisfy is,

$$\int_{-\infty}^{\infty} pdf(x | \mathcal{M}_{Ca}) dx = \int_{-\infty}^{\infty} \frac{1}{Z(-1)} \frac{1}{1+x^2} dx = 1 \quad (31.8)$$

To find the normalization constant, take $1/Z(\lambda)$ outside the integral and multiply both sides,

$$Z(\lambda) = \int_{-\infty}^{\infty} \frac{1}{1+x^2} dx \quad (31.9)$$

Happily, this integral is elementary,

$$\int_{-\infty}^{\infty} \frac{1}{1+x^2} dx = \tan^{-1}(x) \Big|_{-\infty}^{+\infty} = \frac{\pi}{2} - (-\frac{\pi}{2}) = \pi = Z(\lambda) \quad (31.10)$$

Now the density function for this developing MEP distribution looks like,

$$pdf(x | \mathcal{M}_{Ca}) = \frac{1}{\pi} \frac{1}{1+x^2} \quad (31.11)$$

recognized as the simplest Cauchy distribution of Equation (31.1), and described as having a “location parameter” of 0 and a “scale parameter” of 1.

In the above derivation, the Lagrange multiplier was specified as the information in the model. As mentioned many times, it doesn’t make any difference which coordinate system we specify as the information, the other will adjust accordingly. So we can either specify the expected values, and then find the appropriate Lagrange multipliers, or, alternatively, we can specify the Lagrange multipliers, and then find the corresponding expected values.

The latter is what we did because we knew that we had to specify $\lambda = -1$ to in order to arrive at the desired expression. But we can now calculate the expected value of the constraint function to find the dual parameter. Starting from the direct definition of an expected value, we have,

$$E[F(x)] = \int_{-\infty}^{\infty} F(x) pdf(x | \mathcal{M}_{Ca}) dx \quad (31.12)$$

Substituting for $F(x)$, and the MEP assignment as just derived, we find that,

$$\langle F \rangle = E[F(x)] = \int_{-\infty}^{\infty} \ln(1+x^2) \times \frac{1}{\pi(1+x^2)} dx = 1.38629 \quad (31.13)$$

Hence, the origin of this mysterious number $\langle F \rangle = 1.38629$. It is the parameter dual to the Lagrange multiplier. We could have specified it as the information first, and then found the corresponding Lagrange multiplier of $\lambda = -1$.

31.4 A Geometric Motivation

Figure 31.1 at the top of the next page sketches out the trigonometry of the right triangle. From the standard picture of a right triangle inscribed in the unit circle in panel (a), we have $\tan(\theta) = \frac{y}{x}$. As an aid to memory, the tangent of angle θ is sometimes expressed as,

$$\tan(\theta) = \frac{\text{opposite side}}{\text{adjacent side}}$$

The re-orientation in panel (b) where now $\tan(\theta)$, the opposite side divided by adjacent side, is $\frac{x}{y}$. This is how we will visualize physical situations like the distance of a pendulum from its support, a machine gun randomly firing bullets, or, in the example presented later in this Chapter, detecting random flashes from a lighthouse in order to determine its location.

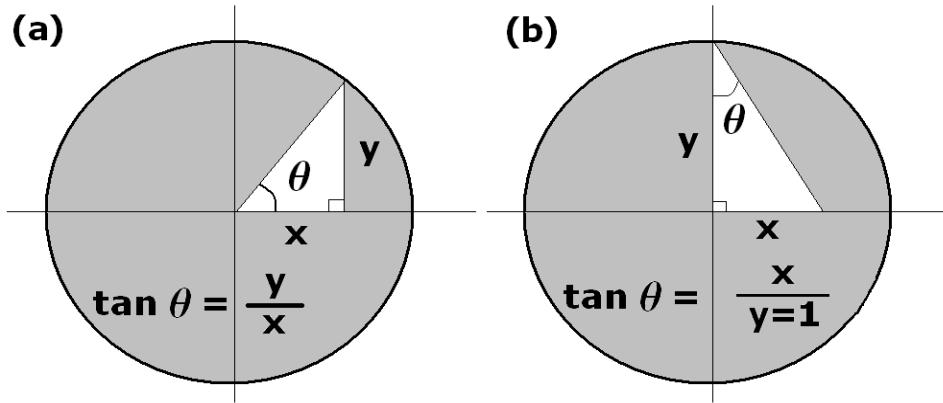


Figure 31.1: The geometry of the right triangle used to set up the physical motivation for the Cauchy distribution.

31.5 More General Cauchy Distributions

Having seen the “trick” by which the simple Cauchy distribution was generated from the MEP formula, one is tempted to generalize along a similar line. A more general Cauchy distribution, taking account of so-called “location and scale parameters,” can also be derived in the same way by the MEP formalism.

However, in a major conceptual break with the usual explanations, we think of these so-called location and scale parameters, not as *parameters of a model*, which they clearly are not, but rather as additional statements y and z which can be measured or observed just like x .

Therefore, extend the state space to three joint statements about x , y , and z . As before, x is the length opposite the angle θ , y is the length of the side adjacent to angle θ , and z is the length from the origin on the x -axis to where the y -axis intersects. We will place the completely reasonable restriction on y that it must have a positive length. We are setting things up for the geometric description of the upcoming numerical example concerning the lighthouse.

Now, we will make use of the MEP formula to derive a numerical assignment to the probability for any joint statement concerning these three lengths. As always, such an assignment is conditioned on the information placed into the probability density function by some model.

The probability density function will then be written as $pdf(x, y, z | \mathcal{M}_{Ca})$ showing the explicit conditioning on this model. We are in the continuous realm where we don’t restrict ourselves to discrete values of the lengths, but always frame the statements in terms of a probability over some interval of the lengths. As a consequence, what were formerly sums now transition into integrals.

Guided by the MEP, we know how to proceed. Introduce two constraint functions on the statements,

$$F_1(x, y, z) = \ln(y) \quad (31.14)$$

and,

$$F_2(x, y, z) = \ln[y^2 + (x - z)^2] \quad (31.15)$$

Insert them as the required information, together with their two associated Lagrange multipliers, λ_1 and λ_2 , into the MEP formula.

But for a change of pace, let's call the two Lagrange multipliers α and β . I do this to highlight the distinction between the parameters of a model and observable statements. You will not find this distinction mentioned in any of the current literature.

Now, write out the MEP formula for this situation,

$$pdf(x, y, z | \mathcal{M}_{Ca}) = \frac{e^{\alpha \ln y + \beta \ln [y^2 + (x - z)^2]}}{Z(\alpha, \beta)} \quad (31.16)$$

If the information inserted by the model takes the form of specifying that the Lagrange multipliers are set at the values $\alpha = 1$ and $\beta = -1$, then the resulting expression looks like,

$$pdf(x, y, z | \mathcal{M}_{Ca}) = \frac{1}{Z(\alpha, \beta)} \frac{y}{y^2 + (x - z)^2} \quad (31.17)$$

To find the partition function $Z(\alpha, \beta)$, integrate the pdf from $-\infty$ to $+\infty$ which must equal 1.

$$\int_{-\infty}^{+\infty} pdf(x, y, z | \mathcal{M}_{Ca}) dx = 1 \quad (31.18)$$

If the integration is performed over x , while $y > 0$ and z are considered as given, the normalizing factor is π as before.

$$\int_{-\infty}^{+\infty} \frac{y}{y^2 + (x - z)^2} dx = Z(\alpha, \beta) \quad (31.19)$$

$$= \pi \quad (31.20)$$

The pdf for the generalized Cauchy now looks like,

$$pdf(x, y, z | \mathcal{M}_{Ca} \rightarrow \alpha = 1 \wedge \beta = -1) = \frac{1}{\pi} \frac{y}{y^2 + (x - z)^2} \quad (31.21)$$

In the literature, this Cauchy density function expression as I have given it above is usually written instead as if x were conditioned on knowledge of a “scale parameter β ” and a “location parameter α ,”

$$pdf(x | \alpha, \beta) = \frac{1}{\pi} \frac{\beta}{\beta^2 + (x - \alpha)^2} \quad (31.22)$$

If $\alpha = 0$ and $\beta = 1$, the more general expression in Equation (31.22) reduces to the simpler Cauchy density of Equation (31.1).

As we shall see in a numerical example in an upcoming section where $y = 4$ and $z = 2$, the expectation of the two constraint functions given above take on the values,

$$E [F_1(x, y = 4, z = 2)] = 1.38629 \quad (31.23)$$

$$E [F_2(x, y = 4, z = 2)] = 4.15888 \quad (31.24)$$

These values are found, as before, by integrating both of these functions with respect to the probability density function,

$$pdf(x, y, z | \mathcal{M}_{Ca}) = \frac{1}{\pi} \frac{y}{y^2 + (x - z)^2} \quad (31.25)$$

as in,

$$E [F_1(x, y = 4, z = 2)] = \int_{-\infty}^{\infty} \frac{4 \ln 4}{\pi (x^2 - 4x + 20)} dx = 1.38629 \quad (31.26)$$

$$E [F_2(x, y = 4, z = 2)] = \int_{-\infty}^{\infty} \frac{4 \ln (x^2 - 4x + 20)}{\pi (x^2 - 4x + 20)} dx = 4.15888 \quad (31.27)$$

31.6 A Rationale from the Geometric Layout

This derivation of a generalized Cauchy density from the MEP formula, although undoubtedly correct, has more than a whiff of *ad hockery* to it. Is there some other supporting rationale based on how the problem was set up from a geometric standpoint? Moreover, can we arrive at the same formula by resorting to the fundamental rules of probability manipulation?

Set up a joint density function for x , y , z , and θ by writing $pdf(x, y, z, \theta)$. Use the **Product Rule** to decompose this joint expression into a product of conditional expressions where we use P instead of pdf simply to shorten the expressions,

$$P(x, y, z, \theta) = P(x | y, z, \theta) \times P(y | z, \theta) \times P(z | \theta) \times P(\theta)$$

The first term on the right hand side, the random distance x subtended by the random angle θ , certainly does depend on both the length of y , the offset distance z from the origin, and the angle θ . Let this be a function of the trigonometric relationships, so that,

$$\begin{aligned}
\tan(\theta) &= \frac{x-z}{y} \\
\tan^2(\theta) &= \frac{(x-z)^2}{y^2} \\
1 + \tan^2(\theta) &= \frac{y^2}{y^2} + \frac{(x-z)^2}{y^2} \\
&= \frac{y^2 + (x-z)^2}{y^2} \\
\frac{1}{1 + \tan^2(\theta)} &= \cos^2(\theta) \\
P(x | y, z, \theta) &\propto \cos^2(\theta) \\
P(x | y, z, \theta) &\propto \frac{y^2}{y^2 + (x-z)^2}
\end{aligned}$$

The second term, the length of y , does not depend on either z or θ , so take the probability that it assumes some value in some small interval as,

$$P(y | z, \theta) \propto \frac{1}{y}$$

Likewise, the third term, the length of the offset z is independent of the angle θ and its probability can be captured by a constant term,

$$P(z | \theta) \propto \frac{1}{k}$$

The last term concerning the probability of the angle θ was already pre-determined by the statement of the problem. There it was given that the angle θ was “random.” This is typically interpreted as again a uniform distribution for θ from -90° to $+90^\circ$ so that,

$$P(\theta) \propto \frac{1}{\pi}$$

Putting all this back together again leads to the joint pdf,

$$\begin{aligned}
P(x, y, z, \theta) &\propto \frac{y^2}{y^2 + (x-z)^2} \times \frac{1}{y} \times \frac{1}{k} \times \frac{1}{\pi} \\
&\propto \frac{1}{k \pi} \frac{y}{y^2 + (x-z)^2}
\end{aligned}$$

We will see in the numerical example of the next section that these proportionality constants eventually get absorbed into an overall partition function. Thus, even

though the above is by no means an air-tight argument, it does lend some plausibility to an otherwise *ad hoc* MEP formula when viewed from the vantage point of the geometric relationships of a right triangle.

During the above argument, the expression $\cos^2(\theta)$ popped up. As we will explore more fully in the exercises, there are some curious mathematical relationships involving the Cauchy distribution and the integral,

$$\int_0^{\pi/2} \cos^2(\theta)^m d\theta \quad (31.28)$$

31.7 The Lighthouse Example

We will elucidate the Cauchy distribution with the following inferential scenario and numerical example. See Figure 31.2 for a sketch of this inferential problem. There is a lighthouse some distance y out to sea from the seashore. The lighthouse emits a light flash “at random” which is then detected on the shoreline at location x .

The IP is located on the shore at its arbitrarily set up origin on the x -axis. The lighthouse is perpendicular to the x -axis at location z . The lighthouse then is located at coordinates (z, y) which we shall more conveniently write as (y, z) .

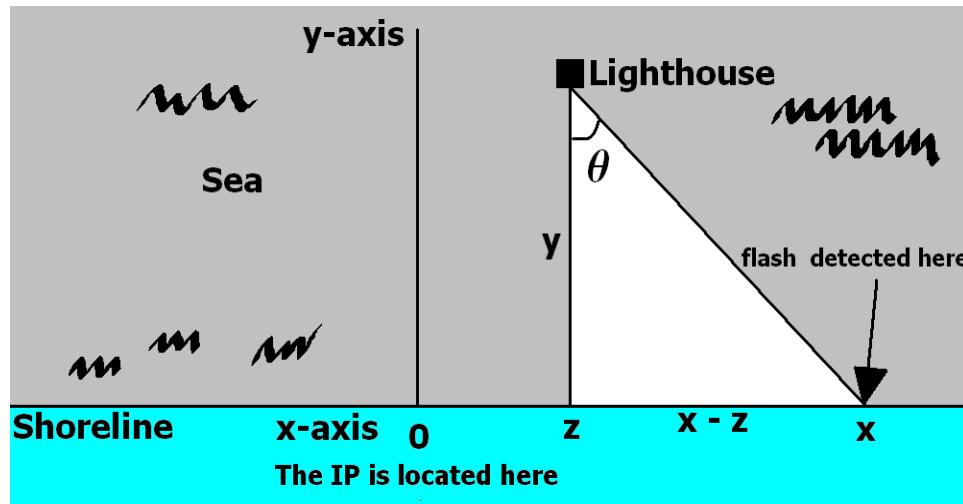


Figure 31.2: A sketch of the geometric layout for the lighthouse scenario.

A number of flashes, x_1, x_2, \dots, x_N , are detected at positions x along the shore. What is the probability density function for the location (y, z) of the lighthouse? Or, stated more bluntly, what is an IP’s degree of belief about where the lighthouse is located? How large of an interval for both y and z is required before the IP has accumulated some appreciable degree of belief about the location of the lighthouse?

As mentioned in the last section, in most descriptions, the more general Cauchy distribution is written with an “location parameter” α and a “scale parameter” β as in Equation (31.22),

$$pdf(x | \mathcal{M}_{Ca}) = \frac{1}{\pi} \frac{\beta}{\beta^2 + (x - \alpha)^2}$$

But writing the pdf in this way is actually a common conceptual error. In reality, we have statements x , y , and z representing the measurable physical coordinates of the lighthouse with z an offset to coordinate x . These are statements that are TRUE or FALSE, or alternatively thought of as observables that can be measured. Therefore, x , y , and z are all statements that can appear either to the left or right of the conditioned upon symbol when writing out a probability.

The two parameters for these models are actually α and β which appear as the Lagrange multipliers in the MEP formalism in Equation (31.16). Thus, the pdf should be written as in Equation (31.21),

$$pdf(x, y, z | \mathcal{M}_{Ca} \rightarrow \{\alpha = 1, \beta = -1\}) = \frac{1}{\pi} \frac{y}{y^2 + (x - z)^2}$$

The specific MEP model \mathcal{M}_{Ca} has set the values, $\alpha = 1$ and $\beta = -1$, and these are what really should be called the parameters of the model.

Our ultimate inferential goal is to reason in an optimal fashion. We accomplish this by generalizing how we would reason if we could reason *logically* about the truth of some statement concerning the location of the lighthouse. Unfortunately, we can not use logic in this problem to make a deduction about the lighthouse.

The best we can do is to cast it as a problem using probability. However, we can not simply set up a probability for a statement like, “The lighthouse is 4 km out to sea and 2 km to the right of where I am standing.” because we are dealing with statements about continuous observations and density functions.

Therefore, we have to phrase the probability of statements over intervals of y and z as in, for example, “The lighthouse is between 1.9 and 2.1 km to my right and between 3.9 and 4.1 km out to sea,”

$$P(3.9 < y < 4.1, 1.9 < z < 2.1)$$

which involves an integration of a density function over these y and z intervals.

For an extremely small data set like $N = 3$, and for a model adopting the Cauchy distribution, we can not expect our degree of belief in any statement to accumulate to anything close to 1, or close to certainty, except for large intervals over y and z . For example, based on just three flashes there is a low degree of belief, $P(2 < y < 12, 0 < z < 4) \approx 0.165$, that the lighthouse is located within 2 to 12 km out to sea, and within 0 to 4 km to the right of our current position, even over these rather wide intervals.

What the Information Processor is uncertain about in the lighthouse problem is the location of the lighthouse in terms of the distances y and z . The observed data are known and they are the x locations of the N flashes. For an initial numerical exercise, write out the joint distribution for the location of three flashes, x_1 , x_2 , and x_3 , the distance y of the lighthouse from shore, and the distance z of the lighthouse from our arbitrarily designated origin on the shore as $pdf(x_1, x_2, x_3, y, z)$.

The inference concerns the unknown location of the lighthouse at (y, z) when conditioned on some known data in the form of the location of the flashes as captured at the shore. Bayes's Theorem tells us that, conditioned on the appropriateness of the one Cauchy model \mathcal{M}_{Ca} we have adopted, (and it does seem appropriate since it is based on the geometry of the right triangle),

$$pdf(y, z | x_3, x_2, x_1, \mathcal{M}_{Ca}) = \frac{pdf(x_3, x_2, x_1, y, z | \mathcal{M}_{Ca})}{pdf(x_3, x_2, x_1 | \mathcal{M}_{Ca})} \quad (31.29)$$

The hallmark of the denominator in Bayes's Theorem is the summation, or, in this case, the integration, over all y and z values,

$$pdf(y, z | x_3, x_2, x_1, \mathcal{M}_{Ca}) = \frac{pdf(x_3, x_2, x_1, y, z | \mathcal{M}_{Ca})}{\int_{-\infty}^{+\infty} \int_0^{\infty} pdf(x_3, x_2, x_1, y, z | \mathcal{M}_{Ca}) dy dz} \quad (31.30)$$

Decompose the numerator with the **Commutativity** axiom, then followed by the **Product Rule**, the independence of any current x measurement from any previous x measurements, and the fact that $P(y, z | \mathcal{M}_{Ca})$ will be some constant k .

$$\begin{aligned} P(x_3, x_2, x_1, y, z, \mathcal{M}_{Ca}) &= P(x_1, x_2, x_3, y, z, \mathcal{M}_{Ca}) \\ P(x_1, x_2, x_3, y, z, \mathcal{M}_{Ca}) &= P(x_1 | x_2, x_3, y, z, \mathcal{M}_{Ca}) \times P(x_2 | x_3, y, z, \mathcal{M}_{Ca}) \times \\ &\quad P(x_3 | y, z, \mathcal{M}_{Ca}) \times P(y, z | \mathcal{M}_{Ca}) \\ &= \prod_{t=1}^3 P(x_t | y, z, \mathcal{M}_{Ca}) \times P(y, z | \mathcal{M}_{Ca}) \\ &= k \prod_{t=1}^3 P(x_t, y, z | \mathcal{M}_{Ca}) \\ &\propto \prod_{t=1}^3 \left[\frac{y}{y^2 + (x_t - z)^2} \right] \end{aligned}$$

By Bayes's Theorem we must then compute,

$$\begin{aligned} pdf(y, z | x_1, x_2, x_3, \mathcal{M}_{Ca}) &= \frac{P(x_1, x_2, x_3, y, z | \mathcal{M}_{Ca})}{P(x_1, x_2, x_3 | \mathcal{M}_{Ca})} \\ &= \frac{\prod_{t=1}^3 \frac{y}{y^2 + (x_t - z)^2}}{\int_{-\infty}^{+\infty} \int_0^{\infty} \prod_{t=1}^3 \frac{y}{y^2 + (x_t - z)^2} dy dz} \quad (31.31) \end{aligned}$$

Suppose for the sake of this numerical exercise that the actual location of the lighthouse is $z = 2$ km to the right from where we are standing at our arbitrarily selected $x = 0$ location on the shore, and $y = 4$ km out to sea. Three flashes were detected at locations on the shore of $x_1 = -6$, $x_2 = 2$, and $x_3 = 7$, all in km.

We will compute the value of the conditional pdf at various values of y and z . Then we will draw contour lines wherever equal values of $pdf(y, z | x_1, x_2, x_3, \mathcal{M}_{Ca})$ occur. The contour plot in Figure 31.3 below shows the updated state of knowledge about the location of the lighthouse based on just these three data points. The actual location of the lighthouse at $(y = 4, z = 2)$ is shown as a black rectangle in the plot. The contours showing the updated state of knowledge based on the observed data make sense given that we know the correct location of the lighthouse.

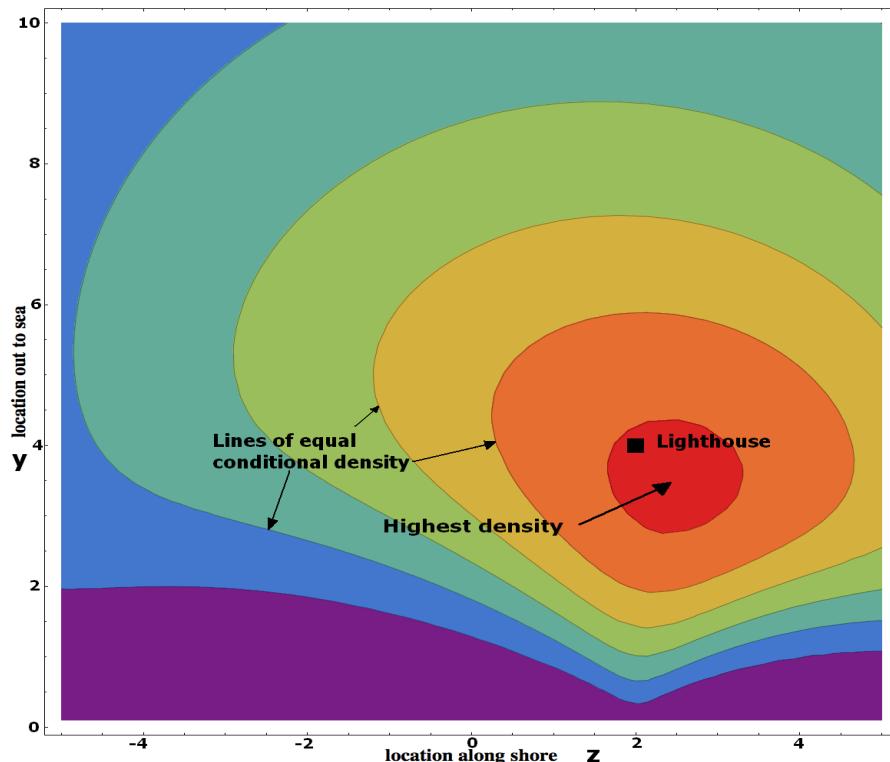


Figure 31.3: Contour plot of the unnormalized posterior distribution of y , location of lighthouse out to sea, and z , location of lighthouse along shore, given the data from three detected flashes and a Cauchy model.

But such a plot only gives us an idea of where the maximum values are located and how spread out the unnormalized probability is over y and z . It doesn't tell us what we really want to know; the degree of belief about any joint intervals $y_{lower} < y < y_{upper}$ and $z_{lower} < z < z_{upper}$ concerning how far the lighthouse is

located either to the right or left of us, and how far out to sea it is. For that, we will have to compute integrals like,

$$P(y_{lower} < y < y_{upper}, z_{lower} < z < z_{upper}) = \frac{\int_{z_l}^{z_u} \int_{y_l}^{y_u} pdf(y, z | x_1, \dots, x_N, \mathcal{M}_{Ca}) dy dz}{\int_{-\infty}^{\infty} \int_0^{\infty} pdf(y, z | x_1, \dots, x_N, \mathcal{M}_{Ca}) dy dz} \quad (31.32)$$

Exercise 31.10.16 constructs a contour plot from 50 data points, instead of just three observed locations of the flashes discussed here as an introductory example. When there are many data points, it makes sense to construct the log of the likelihood rather than the straightforward multiplication of the data.

As you might expect, with more data the IP has a higher degree of belief in any given interval for the location of the lighthouse along the shore and location out to sea. A visual inspection of the contour plot in Figure 31.6 as carried out in Exercise 31.10.16 seems to indicate that the bulk of the probability lies somewhere between 0 and 4 km along the shore and 2 to 6 km out to sea. An integration over this interval calculates the probability of the lighthouse location when conditioned on the observed location x of 50 flashes as,

$$P(2 < y < 6, 0 < z < 4) \approx 0.93$$

31.8 Conceptual Issues

The simple and general Cauchy distributions (and, by extension, the Student- t distributions as well) are derivable by the Maximum Entropy Principle. This fact had already been well established in the literature.

ALL probability distributions are derivable from the MEP, supposing a suitable creativity as to the information that needs to be inserted. In other words, we are talking about the creativity for finding all suitable constraint functions and their expected values. On this point, it is interesting to visit Cover and Thomas's textbook [6] and see what they have to say in their Exercise 12.6.

The inferential problem in this Chapter differs from others we have considered. Previously, our interest was focused first on the relative standing of the whole space of models after some data had been observed. This was necessary before making an inference about some future observation. This goal was encapsulated in the prediction formula,

$$P(x_{N+1} | x_1, x_2, \dots, x_N) = \sum_{k=1}^{\mathcal{M}} P(x_{N+1} | \mathcal{M}_k) P(\mathcal{M}_k | \mathcal{D})$$

But in adopting the single Cauchy model, we have effectively eliminated the whole concept of updating the model space. We began with only one model \mathcal{M}_{Ca} , and that dependence on a single model never changed.

Now this is interesting because in this case we did have a reasonable justification for adopting one model, namely, the geometry of the right triangle. If our knowledge of the physical causes at play are ever this profound, then we never have to consider the panoply of models as we have been doing as a matter of standard procedure up till now. Statistical mechanics seems to adopt the very same stance with regard to its constraint function as energy and the temperature as the parameter.

The inference in this problem, as illustrated by the lighthouse exercise, was about the location coordinates of the lighthouse when we didn't know where in fact it was located, but did possess the known measurements of the light flashes on the shore. The data of the flashes was never going to update any model space because there was never any space of models to begin with. We were stuck with the undeniable correctness of the Cauchy distribution as the one model. But these data certainly were indispensable when the formal manipulation rules showed us how they should be processed in conjunction with the unknown location of the lighthouse.

The inferences made in these kind of Cauchy distribution examples are particularly galling to orthodox statisticians. That is why they placed so many stumbling blocks in the path of people like Jaynes who tried to point out the appeal of the Bayesian approach.

For example, there is simply no need to bring the concept of sufficient statistics into any conversation when the MEP and the formal manipulation rules are used to make inferences. This demand for a statistic of any kind is a red herring left over from orthodox Fisherian concepts. It is particularly noxious because it doesn't make any sense, no matter how people may try to force it down your throat.

Statistics, sufficient or otherwise, are functions of the data. The creation of numerical assignments via the MEP principle has absolutely nothing to do with any data! It has everything to do with inserting information into states of knowledge through the auspices of constraint functions, Lagrange multipliers, and the expected values of constraint functions **defined over the state space**.

Any data that eventually come into play are processed in an optimal fashion by using Bayes's Theorem. As in our numerical example, any updated state of knowledge about the coordinates of the lighthouse is calculated based on known and measured position of the light flashes, which are the data in this example. The appearance of the log likelihood within the Bayesian approach comes about quite naturally without any mysterious incantations.

What I think has not been noticed is the distinction between parameters of a model and statements in a state space. What are sometimes called parameters in the formulation of problems, as, for example, the so-called θ , α , and β "parameters" in the Cauchy lighthouse scenario, should really be thought of as propositions, statements, or measurable observations about angles and distances.

31.9 Connections to the Literature

The numerical example of using the Cauchy distribution to solve the lighthouse location problem is taken directly from Gull [11] who attributes it to a Cambridge University undergraduate homework exercise. My solution presented in section 31.7 essentially follows the lead shown by Gull. Gull uses the Bayesian approach to calculate probabilities for finding first the lighthouse location along the shore and, secondly, as I did, both location along the shore and out to sea.

His contour plots look quite similar to mine. Curiously though, he never once mentions that the Cauchy distribution can be derived via the MEP despite his deep knowledge of, and experience with, this technique.

Gull also engages in a potentially misleading abuse of language by stating that the Cauchy distribution has many more “bad” data points than the Gaussian distribution. In his contour plots for $N = 100$ data points, he goes even further, and says that the bulk of the “good” data eventually overwhelm the “bad.”

Of course, there never are any “good” or “bad” data points. Some models are supported to a greater or lesser extent by whatever transpired as data. And if, as in this case when there is a compelling physical reason to trust in one model (the right triangle again), we would expect to see “bad data” (an extremely large x value) because such “bad data” must eventually occur when the random angle is greater than, say, 89° or less than -89° .

We recognize and sympathize with his acknowledged intent; we are fixated by the idea that all data are supposedly “drawn” from the Gaussian distribution. And a model like the Cauchy distribution with its (quite expected) occasional very large x values strikes us as “pathological.”

A similar attitude, I believe, also underlies the whole “black swan” phenomenon. If a sensitivity were more prevalent concerning the unbelievably huge space of models that can exist over the real line for just one variable (and I’m not referring to just the whole space of Gaussian models reached by allowing its two parameters to range freely), then people wouldn’t be quite as infatuated with “black swan” surprises.

Sivia [32] gives a nice presentation of the same lighthouse problem accompanied by drawings clearly showing the trigonometry of the right triangle leading to the Cauchy distribution. He illustrates contour plots only for statements z concerning location along the shore, but not for statements y concerning location out to sea.

Sivia also does a masterful job of explaining what a transformation of variables means for probability theory. He derives the simple Cauchy distribution as a transformation on the underlying uniform distribution over the angle θ . We will illustrate how *Mathematica* tackles this problem in Exercise 31.10.6.

My only quibble with Sivia is that he discusses the lighthouse problem in the context of “parameter estimation,” thereby incorrectly confusing, like everyone else, *statements* about location with the *parameters* of a model.

Even those who profess an inclination towards the MEP and Bayesian approach often miss the distinction between the parameters of a model and the statements in the state space. As a consequence, propositions, statements, or measurable observations such as the x , y , and z of the lighthouse example should appear to the left of the conditioned upon symbol within the probability notation, rather than to the right as part of some model.

It also pains me to say that even the *Mathematica* documentation for the Cauchy distribution refers to a “location parameter” and a “scale parameter.” If, as in my approach in this Chapter, the Cauchy distribution had been derived via the MEP, it would have been quite clear that the parameters of the model refer to something else entirely.

Finally, as alluded to above, I find it quite curious that authors with compatible feelings about inference, like Stephen Gull and Devinder Sivia, discuss the origin for the state of knowledge for the lighthouse problem, not from the more fundamental MEP principle, but rather as a typical *deus ex machina*, or an example of a change of variables. But the MEP ought to be brought in right at the start because it represents the overarching principle as to how an information processor assigns numerical values according to the information inserted by some specific model, while at the same time excluding all other information not in that model.

Feller [7] has a similar physical motivation for finding the distribution of light from a rotating mirror, bullets from a rotating machine gun, as well as many other interesting facts about both the univariate and multivariate Cauchy distribution.

Box and Tiao [4, pg. 64] present a nice numerical illustration of the Cauchy distribution together with the attendant numerical processing of the data. As you might have guessed, I admire authors who amplify a theoretical treatment with clearly worked out examples.

... a sample of $n = 5$ observations (11.4, 7.3, 9.8, 13.7, 10.6) were randomly drawn from the Cauchy distribution ... Thus, assuming little were known about θ *a priori*, the posterior distribution is approximately,

$$p(\theta | y) \approx cH(\theta) \quad -\infty < \theta < \infty$$

where

$$H(\theta) = 10^5 [1 + (7.3 - \theta)^2]^{-1} [1 + (9.8 - \theta)^2]^{-1} \dots [1 + (13.7 - \theta)^2]^{-1}$$

the factor 10^5 being a convenient multiplier and c is the normalizing constant.

Their notation for the Cauchy distribution is,

$$p(y | \theta) = \pi^{-1} [1 + (y - \theta)^2]^{-1}$$

which corresponds to our,

$$pdf(x | y = 1, z = \theta, \mathcal{M}_{Ca}) = \frac{1}{\pi} \frac{y}{y^2 + (x - z)^2}$$

Unfortunately, they also fall into the trap of calling θ a “location parameter.”

Their example presents me with an opportunity to rail against one of the most wide spread conceptual errors in statistics; one so thoroughly entrenched in the literature it seems impossible to eradicate it. This is the use of the language that asserts, as Box and Tiao do above, that data or observations are “randomly drawn” from some probability distribution. Nothing could be further from the truth!

Observations or measurements are the result of some physical process. Physical reality is not subject to any probability distribution that causes things to happen. Probability distributions are entirely epistemological creations that exist solely to aid an IP’s inferences. As such, they exist solely in the mind of the IP, never in an external reality!

Data should never be conceptualized as a random drawing from a probability distribution because any such probability distribution is “merely” the outcome of information inserted by an IP’s conscious mind. This conceptual distortion is what Jaynes was trying to convey when he talked about the *mind projection fallacy*. Just because an IP happens to possess a state of knowledge about an event does not cause that event to take place!

So to use the common phraseology that talks about data being randomly drawn from the Gaussian distribution, or the Cauchy distribution, or any probability distribution is utter nonsense! As we have been at pains to emphasize from the very beginning, any probability distribution captures some *state of knowledge* held by an IP. That epistemological state is dictated by the information resident in some model.

Now, the Cauchy distribution is very interesting with regard to these comments because the line demarcating physical reality and a model becomes blurred in just this case, no doubt adding to the already persistent confusion. This is because the *one* model adopted leading to the Cauchy distribution (and let me repeat for good measure also arrived at via the MEP formula) *is* based on the physical reality of the correct geometric representation of the right triangle.

Let’s return to Box and Tiao’s example of the Cauchy model. As Bayesians, they want the probability of the unknown “location parameter” $\theta \equiv z$ when conditioned on the known data. By Bayes’s Theorem, this is,

$$pdf(z | y = 1, x_1, \dots, x_5, \mathcal{M}_{Ca}) = \frac{pdf(x_1, \dots, x_5, z | y = 1, \mathcal{M}_{Ca})}{\int_{-\infty}^{\infty} pdf(x_1, \dots, x_5, z | y = 1, \mathcal{M}_{Ca}) dz}$$

In the numerator, we have,

$$pdf(x_1, \dots, x_5, z | y = 1, \mathcal{M}_{Ca}) \propto \prod_{t=1}^5 \frac{1}{1 + (x_t - z)^2}$$

which is Box and Tiao’s,

$$H(\theta) = 10^5 [1 + (7.3 - \theta)^2]^{-1} \times [1 + (9.8 - \theta)^2]^{-1} \times \dots \times [1 + (13.7 - \theta)^2]^{-1}$$

In the denominator, we integrate over all values for z to find the normalizing factor for correct probabilities. Box and Tiao give a detailed explanation of a traditional numerical integration routine based on Simpson's rule over the values where their $H(\theta)$ is substantially greater than 0, which turns out to be from about $\theta = 6.5$ through $\theta = 14.5$.

I have the luxury on relying upon *Mathematica* to carry out the same numerical integration in a short piece of code, and, moreover, performing the integration over the entire real line,

```
NIntegrate[( $H(\theta)$ ), {z, -∞, ∞}]
```

which returns the more accurate value for $c^{-1} = 536.326$, almost the same as the $c^{-1} = 535$ reported by Box and Tiao over their shorter interval.

With the normalizing factor,

$$\int_{-\infty}^{\infty} pdf(x_1, \dots, x_5, z | \mathcal{M}_{Ca}) dz = 536.326$$

in the denominator of Bayes's Theorem now known, it is possible to calculate the probability for any desired z interval as,

$$P(z_l < z < z_u | \mathcal{D}) = \frac{\int_{z_l}^{z_u} pdf(x_1, \dots, x_5, z | \mathcal{M}_{Ca}) dz}{\int_{-\infty}^{\infty} pdf(x_1, \dots, x_5, z | \mathcal{M}_{Ca}) dz}$$

Based on this formula, the probability that z is less than 11.5, ($z_l = -\infty, z_u = 11.5$), is 87.7%. Box and Tiao report the probability that the location parameter θ is less than 11.5 as 87.9%. Without their explicitly saying so, it is evident that the five simulated data points were taken from a Cauchy distribution where $y = 1$ and $z = 10$.

Equally as important as this numerical illustration of how data are handled within a Cauchy distribution, or really more important by far, is Box and Tiao's correct emphasis that the Bayesian approach had no difficulty reaching an inference based on a Cauchy model. The orthodox literature had long portrayed the Cauchy distribution as "pathological," partly because it did not conform to the Fisherian demands for the existence of "sufficient statistics." The Bayesian approach has no need for any such thing as "sufficient statistics." Box and Tiao reaffirm this as the reason for giving a detailed numerical example relying upon the Cauchy distribution.

Finally, we arrive at my attribution to Kapur [25]. There is a great deal of ambivalence here on my part which I haven't completely resolved. His book, on the whole, is absolutely required reading for his extensive and wide-ranging material on the MEP. In fact, my original discovery that the Cauchy distribution could be proved to be an MEP derived distribution came directly from studying Kapur.

Mitigating this praise is an unfortunate deluge of typographical errors scattered densely on almost every page. In short, the printing and/or proofreading enterprise went seriously awry somewhere along the way.

Furthermore, while in total agreement with his clear and fundamental support for the MEP, there are whole Chapters which seem completely disconnected from these basic principles. One is hard-pressed to make hide nor hair of any of it.

Jeffreys [24, pg. 76] presents this formula as an example of Pearson's hierarchy of probability distribution types, namely one of Type VII which is itself a simplification of a Type IV,

$$y = \beta^{2m-1} \frac{(m-1)!}{\sqrt{\pi}(m-3/2)!} [(x-\lambda)^2 + \beta^2]^{-m}$$

Making the association that $\beta \rightarrow y$, $\lambda \rightarrow z$, $m \rightarrow 1$, and $y \rightarrow \text{pdf}(x, y, z | \mathcal{M}_{Ca})$, we see that this complicated and unmotivated formula is, in fact, the general Cauchy distribution derived via the MEP with the information $\alpha = 1$ and $\beta = -1$,

$$\text{pdf}(x, y, z | \mathcal{M}_{Ca}) = \frac{1}{\pi} \cdot \frac{y}{y^2 + (x-z)^2}$$

My final comment is on the curious and alien ways that people think about things, and then try to explain these mysterious thought processes to others. No finer examples exist of this tortured enterprise than Kapur's derivation of the Cauchy distribution as extensively detailed later on in the exercises, and Jeffreys's use of complex numbers to relate Pearson's Type IV probability distributions to the general Cauchy distribution.

After you have spent a very long time subjecting these equations to careful and painful scrutiny, you find in the end that they are all correct. Only some strange quirk will compel you to wade through this morass. I sincerely doubt that anyone stumbling upon Jeffreys's or Kapur's explanations for, say, the Cauchy distribution, would have the fortitude to follow through to the end. But it would be a shame to abandon them at the outset because, as I mentioned earlier, there are indeed some astonishing mathematical relationships to be revealed.

31.10 Solved Exercises for Chapter Thirty One

Exercise 31.10.1: Select some values of θ somewhat uniformly between -90° and $+90^\circ$ and calculate $\tan(\theta)$ at $z = 0$ and $y = 1$ to get a feel for the range of x in the simple Cauchy distribution.

Solution to Exercise 31.10.1

At extreme angles, for example at $\theta = -89^\circ$, $x = -57.29$, or at $\theta = +89^\circ$, $x = 57.29$. But from about $\theta = -45^\circ$ through $\theta = +45^\circ$, x remains in the range from about -1 to $+1$. Look at the left side of Table 31.1 for $\tan(\theta)$ as θ increases from $\theta = 0^\circ$ through $\theta = 80^\circ$. In the right half of the table, a finer grained listing from $\theta = 81^\circ$ through $\theta = 89^\circ$ shows the rapidly increasing values for x . The relationship is symmetric from $\theta = 0^\circ$ through -90° and $\theta = 0^\circ$ through $+90^\circ$, so only positive values for θ are shown.

Table 31.1: The kind of x values to expect from a Cauchy distribution as the angle θ varies uniformly over -90° through $+90^\circ$.

θ	$\tan(\theta)$	θ	$\tan(\theta)$
0°	0.000	81°	6.314
10°	0.176	82°	7.115
20°	0.364	83°	8.144
30°	0.577	84°	9.514
40°	0.839	85°	11.430
50°	1.192	86°	14.301
60°	1.732	87°	19.081
70°	2.747	88°	28.636
80°	5.671	89°	57.290

Exercise 31.10.2: Randomly select, say, 100 x values from the simple Cauchy distribution. How does this selection stack up with what we found out above?

Solution to Exercise 31.10.2

Use the *Mathematica* built-in function **RandomVariate[]** to sample from the Cauchy distribution with “location parameter” $\alpha = 0$ and “scale parameter” $\beta = 1$. *Mathematica* will generate a list of 100 random values of x with this code,

```
RandomVariate[CauchyDistribution[0,1],100]
```

A cursory examination of this list will reveal that about half of the x values are between -1 and $+1$. Amongst all these small x values, we see the occasional $x = -14.289$, or $x = 11.985$. Such relatively large values of x should not surprise us given the uniform sampling of angles from $\theta = -90^\circ$ through $\theta = +90^\circ$.

Exercise 31.10.3: What is the exact probability that the x values will be between -1 and $+1$?

Solution to Exercise 31.10.3

To find this probability, integrate over the Cauchy probability density function between the limits of -1 and $+1$,

$$P(-1 < x < +1) = \int_{-1}^{+1} \frac{1}{\pi} \frac{1}{1+x^2} dx = 1/2$$

Exercise 31.10.4: What is the exact probability that the angle θ lies between -45° and $+45^\circ$?

Solution to Exercise 31.10.4

We have just discovered that $P(-1 < x < +1) = 1/2$. Since the angle θ varies uniformly over a total interval of 180° , an interval of 90° from -45° through $+45^\circ$ must consume $1/2$ of the total probability. To complete the picture,

$$\arctan(-1) = -45^\circ \text{ and } \arctan(+1) = +45^\circ$$

Exercise 31.10.5: Is the probability for any θ interval the same probability for the corresponding x interval?

Solution to Exercise 31.10.5

Integrate the pdf for the Cauchy distribution for an interval x corresponding to say $\theta = 5^\circ$ and $\theta = 6^\circ$. Since $\tan(5^\circ) = 0.0874887$ and $\tan(6^\circ) = 0.105104$, an integration yields,

$$P(0.0874887 < x < 0.105104) = \int_{0.0874887}^{0.105104} \frac{1}{\pi} \frac{1}{1+x^2} dx = 0.0056$$

Since the probability for all one degree intervals has to be the same since θ is distributed uniformly, an integration for an interval x corresponding to a different one degree interval, say, $\theta = 88^\circ$ through $\theta = 89^\circ$ must yield the same probability,

$$P(28.63 < x < 57.29) = \int_{28.63}^{57.29} \frac{1}{\pi} \frac{1}{1+x^2} dx = 0.0056 = \frac{1}{180}$$

Exercise 31.10.6: How does *Mathematica* deal with such probability transformations?

Solution to Exercise 31.10.6

The fundamental premise about the Cauchy distribution is that the angle θ is distributed uniformly between $-\pi/2$ and $\pi/2$. But what does this uniformity imply for the probability distribution of x when $x = \tan(\theta)$? In the last two exercises, we showed numerically how this worked out. *Mathematica* provides the transformed distribution by evaluating this code,

```
PDF[TransformedDistribution[Tan[\theta],  
Distributed[\theta, UniformDistribution[{-\pi/2,\pi/2}]]],x]
```

and returning as the answer,

$$\frac{1}{\pi(1+x^2)}$$

Exercise 31.10.7: Plot the density function for what we have labeled as the simple Cauchy distribution.

Solution to Exercise 31.10.7

Use the *Mathematica* **Plot[]** function to generate a picture of the simple Cauchy probability density function,

```
Plot[PDF[CauchyDistribution[0,1],x],{x,-4,4},  
Filling→Bottom]
```

to construct the curve appearing below in Figure 31.4. Notice that sketches of this sort are somewhat deceiving in that it isn't quite clear how much probability remains in the tails of this density function. Performing an integration just as in the

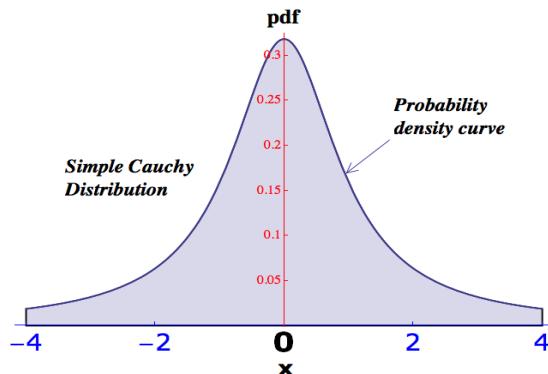


Figure 31.4: The simple Cauchy probability density function when $y = 1$ and $z = 0$.

last few exercises between the limits of $x = -4$ and $x = +4$, in other words, at those x values where the sketch disappears, we find that about 15% of the probability still remains at more extreme x values.

Exercise 31.10.8: In a random sample of, say, 100 x values, how many are greater than +4 or less than -4?

Solution to Exercise 31.10.8

In the random sample that I generated in Exercise 31.10.2, I counted up 12 instances of x values greater than +4 or less than -4. With the expectation for such an occurrence at about 15, based on the integration in the last exercise, there is no cause for alarm.

Exercise 31.10.9: What kind of x values are to be expected if, instead of $y = 1$ and $z = 0$ as in the previous exercises, now $y = 4$ and $z = 2$ as specified in the lighthouse scenario?

Solution to Exercise 31.10.9

Now,

$$\begin{aligned}\tan(\theta) &= \frac{x-z}{y} \\ &= \frac{x-2}{4} \\ x &= 4\tan(\theta) + 2\end{aligned}$$

The values that appeared in Table 31.1 now look like the ones at the top of the next page in Table 31.2.

Exercise 31.10.10: Do a quick and dirty check on the plausibility of these numbers appearing in the table above.

Solution to Exercise 31.10.10

For x values between about -30 and +30, the angle θ varied uniformly between about -82° and $+82^\circ$. This covers about $164/180 = 0.91$ of the entire θ space. The integration of the more general Cauchy pdf over this range yields a probability of,

$$P(-30 < x < +30) = \int_{-30}^{+30} \frac{1}{\pi} \frac{4}{16 + (x-2)^2} dx = 0.915$$

Table 31.2: The kind of x values to be expected from a Cauchy distribution as the angle θ varies uniformly over -90° through $+90^\circ$. The y and z locations are changed from $y = 1$ and $z = 0$ to the lighthouse scenario where $y = 4$ and $z = 2$.

θ	$4 \tan(\theta) + 2$	θ	$4 \tan(\theta) + 2$
0°	2.000	81°	27.255
10°	2.075	82°	30.462
20°	3.456	83°	34.577
30°	4.309	84°	40.058
40°	5.356	85°	47.720
50°	6.767	86°	59.203
60°	8.928	87°	78.325
70°	12.990	88°	116.545
80°	24.685	89°	231.160

Exercise 31.10.11: Use *Mathematica* to confirm that the transformed distribution of $4 \tan(\theta) + 2$ is indeed a Cauchy distribution.

Solution to Exercise 31.10.11

Mathematica will provide the transformed distribution of $4 \tan(\theta) + 2$ where the angle θ follows the uniform distribution between -90° and $+90^\circ$, or between $-(\pi/2)$ and $+(\pi/2)$, with,

```
TransformedDistribution[4 Tan[\theta] + 2,
    Distributed[\theta, UniformDistribution[{-\pi/2, \pi/2}]]]
```

Upon evaluation, the answer returned is **CauchyDistribution[2, 4]**.

If we place this code into,

```
Simplify[PDF[(* above expression *), x]]
```

Mathematica returns with,

$$\frac{4}{\pi (20 - 4x + x^2)}$$

Check that the expression for the more general Cauchy distribution for three statements x , y , and z ,

$$pdf(x, y, z | \mathcal{M}_{Ca}) = \frac{1}{\pi} \frac{y}{y^2 + (x - z)^2}$$

does work out to the *Mathematica* result,

$$pdf(x, y = 4, z = 2 | \mathcal{M}_{Ca}) = \frac{1}{\pi} \frac{4}{4^2 + (x - 2)^2} = \frac{4}{\pi (20 - 4x + x^2)}$$

Exercise 31.10.12: Plot the density function for the more general Cauchy distribution.

Solution to Exercise 31.10.12

The density function is plotted in Figure 31.5. Notice that the pdf peaks as it should at $x = 2$ since $z = 2$. Since $y = 4$, there is considerable probability for much larger x values. The plot peters out at $x = -30$ and $x = +30$.

We know that about 0.08 of the remaining probability is still out there in the tails of this pdf. When I counted up x values greater than $+30$ and less than -30 in a random sample of 100, I found 4 such values with the expectation that I would find about 8. The most extreme value in the random sample was a value of $x = +97.48$ corresponding to an angle of about $\theta = 87.5^\circ$.

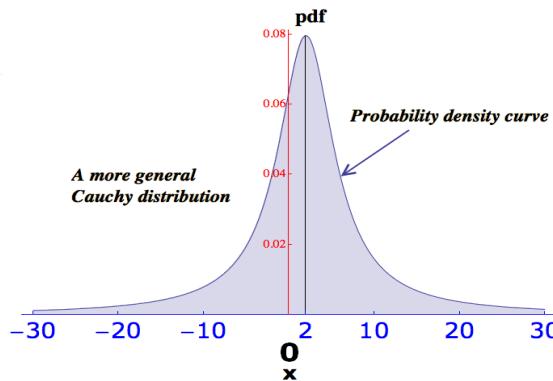


Figure 31.5: A more general Cauchy density function with $y = 4$ and $z = 2$.

Exercise 31.10.13: Calculate the unnormalized pdf for a joint statement about the location of a flash, the lighthouse's location out to sea, and its offset from our chosen origin.

Solution to Exercise 31.10.13

For a numerical example of such a calculation, suppose that we want to find the pdf for the data point $x = -6$ when we use the lighthouse scenario values of $y = 4$ and $z = 2$,

$$\begin{aligned} \text{pdf}(x = -6, y = 4, z = 2 \mid \mathcal{M}_{Ca}) &\propto \frac{y}{y^2 + (x - z)^2} \\ &\propto \frac{4}{16 + (-6 - 2)^2} \\ &\propto 0.05 \end{aligned}$$

Exercise 31.10.14: In preparation for making an inference about the location of the lighthouse, write out the *Mathematica* expression proportional to the log likelihood of one observation of the location of the flash.

Solution to Exercise 31.10.14

The log of the pdf for the location of the flash x along the seashore, the location y of the lighthouse out to sea, and the distance z of the lighthouse from the origin on the x -axis is,

$$\ln \text{pdf}(x, y, z | \mathcal{M}_{Ca}) = \ln y - \ln [y^2 + (x - z)^2]$$

A *Mathematica* function to make this calculation is,

```
lighthouse[x_, y_, z_] := Log[y] - Log[Power[y, 2] + Power[(x - z), 2]]
```

Evaluating `N[lighthouse[-6, 4, 2]]` yields -2.99573 . Place this result as the argument to the exponential and it corresponds to the answer in the last exercise,

$$e^{-2.99573} = 0.05$$

Exercise 31.10.15: Find the log likelihood for the three data points given in section 31.7.

Solution to Exercise 31.10.15

`Map[]` the above `lighthouse[]` function to three data points,

$$x_1 = -6, x_2 = 2, x_3 = 7$$

where these are the detected distances of three flashes from our current origin on the seashore,

```
Map[lighthouse[#, 4, 2] &, { -6, 2, 7 }]
```

This returns a list `{ -2.99573, -1.38629, -2.32728 }` as the log likelihood at each of the three flash locations. We want the sum of the elements in this list so we form,

```
Total[Map[lighthouse[#, 4, 2] &, { -6, 2, 7 }]]
```

which returns `-6.7093`.

We require a function to use in `ContourPlot[]` so form,

```
lighthousedata[y_, z_] :=
  Exp[Total[Map[lighthouse[#, y, z] &, { -6, 2, 7 }]]]
```

A call to `lighthousedata[4, 2]` results in `0.00121951`. This is proportional to the likelihood of the three data points and the values $y = 4$ and $z = 2$. This value would be one point in `ContourPlot[]` at “location out to sea” $y = 4$ and “location along the shore” $x = 2$.

Exercise 31.10.16: Extrapolating from the above, write more general *Mathematica* code for calculating contour plots for any set of simulated data.

Solution to Exercise 31.10.16

Let **data** contain 50 simulated data points from a Cauchy distribution described by “location parameter” **a = 2** and “scale parameter” **b = 4**. We labor to correct this conceptual error by identifying these so-called “parameters” with statements *z* and *y* concerning the location of the lighthouse along the shore and its location out to sea.

```
data = RandomVariate[CauchyDistribution[2, 4], 50];
```

We have already seen the likelihood function for one observation,

$$pdf(x, y, z, | \mathcal{M}_{Ca}) \propto \frac{y}{y^2 + (x - z)^2}$$

coded as,

```
lighthouse[x_, y_, z_] := Log[y] - Log[Power[y, 2] + Power[(x - z), 2]];
```

We require a function as an argument to **ContourPlot[]** that is the likelihood over all the data points,

```
lighthousedata[z_, y_, constant_] :=
  Exp[constant + Total[Map[lighthouse[#, y, z] &, data]]]
```

A constant was added to the **lighthousedata[]** function in order to prevent excessively small numbers. Now we finish up with all the arguments and options to **ContourPlot[]** for producing a nice visual image of the unnormalized probability for statements concerning the location along the shore and location out to sea,

```
ContourPlot[lighthousedata[z, y, 140], {z, 0, 4}, {y, 2, 6},
ColorFunction → "Rainbow",
FrameStyle → Directive[Black, Thick, Italic, 26],
FrameLabel → {Style["Location along shore", Bold, Blue, 26],
{Style["Location out to sea", Bold, Blue, 26]}},
Epilog → Rectangle[{1.9, 3.9}, {2.1, 4.1}]
(* end of ContourPlot *)]
```

As shown in Figure 31.6, the output from this program produces a plot with regions of equal likelihood marked out as elliptical contours. The plot extends from location *z* along the shore from 0 to 4 km, and location *y* out to sea from 2 to 6 km. The true location of the lighthouse at *z* = 2 and *y* = 4 km is the black rectangle. It is located very close to the region of maximum likelihood. Of course, we still don’t know the actual probability for any *z* and *y* intervals until we do the integrations.

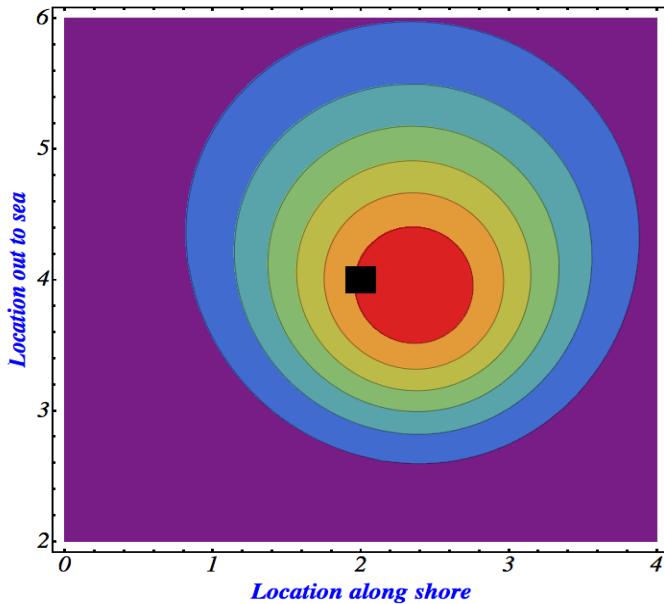


Figure 31.6: Regions of equal, but non-normalized, likelihood for the location of the lighthouse based on $N = 50$ flashes. The true location of the lighthouse is indicated by the rectangle.

Exercise 31.10.17: Show the details of the integration that calculates the probability for the location of the lighthouse.

Solution to Exercise 31.10.17

By using Equation (31.32) at the end of section 31.7, the probability for the location of the lighthouse between 0 and 4 km to the right of where we were standing, and 2 to 6 km out to sea, was said to be about 0.93 on the basis of 50 data points.

We first have to integrate **lighthousedata[z,y,140]** with z limits from $-\infty$ to ∞ and y limits from 0 to ∞ as the denominator, or normalizing factor for the numerator.

NIntegrate[lighthousedata[z,y,140], { z, -\infty, \infty }, { y, .01, \infty }]

results in an answer of 291.876.

The integration in the numerator for the interval we are interested in is,

NIntegrate[lighthousedata[z,y,140], { z, 0, 4 }, { y, 2, 6 }]

and results in an answer of 270.526. The probability is then approximately,

$$P(2 < y < 6, 0 < z < 4) \approx \frac{270.526}{291.876} \approx 0.93$$

Exercise 31.10.18: Show the mathematical consequences of Kapur's MEP derivation of the Cauchy distribution.

Solution to Exercise 31.10.18

The derivation I provided using the MEP formula for the Cauchy distribution seems to me to be the most straightforward path to the desired goal. Kapur [25, pg. 50] approached the problem differently, but from an interesting angle.

His derivation uncovers some curious mathematical relationships involving the digamma function. In the following exercises, we will need to rely heavily upon *Mathematica*'s assistance in checking Kapur's various, and as I happen to see them, particularly unmotivated formulas.

To orient ourselves for what is coming, review these trigonometric relationships from the right triangle. First, we have the fundamental pictorial definition of the cosine of an angle θ as,

$$\cos(\theta) = \frac{\text{adjacent side}}{\text{hypotenuse}}$$

Then, given the way we have drawn and labeled the lighthouse problem,

$$\cos^2(\theta) = \frac{y^2}{y^2 + (x-z)^2}$$

$$\text{If } y = 1 \text{ and } z = 0, \text{ then } \cos^2(\theta) = \frac{1}{1+x^2}$$

$$\cos^2(\theta)^{(b-1)} = \frac{1}{(1+x^2)^b}$$

$$\cos^{2b-2}(\theta) = \frac{1}{(1+x^2)^b}$$

Kapur does not integrate with respect to x ,

$$\int_{-\infty}^{\infty} \frac{1}{(1+x^2)^b} dx = Z(b)$$

to find the partition function, but chooses instead to examine the integral with respect to θ ,

$$\int_0^{\pi/2} \cos^{(2b-2)}(\theta) d\theta$$

over the limits of 0 to $\pi/2$.

Mathematica does the symbolic integration of $\cos^{2b-2}(\theta)$ between the limits of 0° and $\pi/2 = 90^\circ$,

$$\int_0^{\pi/2} \cos^{2b-2}(\theta) d\theta$$

as,

```
Integrate[Power[Cos[\theta], 2 b - 2], {\theta, 0, \pi/2}]
```

and returns the answer,

$$\frac{\sqrt{\pi} \Gamma(b - 1/2)}{2 \Gamma(b)}$$

In order for the integration to proceed, the restriction that $b > 1/2$ had to be imposed. This makes sense because one can never get past the restriction of the square root of the hypotenuse of the right triangle $\sqrt{x^2 + y^2}$.

Invoking the universal constraint, we are able to find the partition function,

$$2 \int_0^{\pi/2} \frac{1}{Z(b)} \times \cos^{2b-2}(\theta) d\theta = 1$$

$$Z(b) = \frac{\sqrt{\pi} \Gamma(b - 1/2)}{\Gamma(b)}$$

What would b have to equal in order for $Z(b) = \pi$ as it must for the simple Cauchy distribution? If $b = 1$, as we know it is for the simple Cauchy distribution, then the above result yields,

$$Z(b = 1) = \frac{\sqrt{\pi} \Gamma(b - 1/2)}{\Gamma(b)} = \frac{\sqrt{\pi} \Gamma(1/2)}{\Gamma(1)} = \frac{\sqrt{\pi} \sqrt{\pi}}{\Gamma(1)} = \pi$$

Exercise 31.10.19: Confirm that the integration over θ as well as the integration over x yield the same result.

Solution to Exercise 31.10.19

Perform the integration of,

$$\int_{-\infty}^{\infty} \frac{1}{(1+x^2)^{\lambda}} dx$$

with *Mathematica* evaluating,

```
Integrate[1/((1+x^2)^{\lambda}), {x, -\infty, \infty}, Assumptions \rightarrow Re[\lambda] > 1/2]
```

and returning the same result,

$$\frac{\sqrt{\pi} \Gamma(\lambda - 1/2)}{\Gamma(\lambda)}$$

Exercise 31.10.20: Where does Kapur go from here?

Solution to Exercise 31.10.20

Kapur examines the integration involving $\cos^2(\theta)$ raised to arbitrary power m .

$$\int_0^{\pi/2} \cos^2(\theta)^m d\theta = f(m)$$

```
Integrate[Power[Cos[\theta], 2 m], {\theta, 0, \pi/2},
Assumptions \rightarrow Re[m] > -(1/2)]
```

The solution for $f(m)$ comes out similar to the above exercise in terms of the Gamma function,

$$f(m) = \frac{\Gamma(\frac{1}{2}) \Gamma(m + \frac{1}{2})}{2 \Gamma(m + 1)}$$

Thus, we see that $m = b - 1$ and checking $m = 0, b = 1$ with $2b - 2 = 0$, we find that $f(m = 0)$,

$$\begin{aligned} f(0) &= \frac{\Gamma(\frac{1}{2}) \Gamma(\frac{1}{2})}{2 \Gamma(1)} \\ &= \frac{\sqrt{\pi} \sqrt{\pi}}{2} \\ &= \frac{\pi}{2} \end{aligned}$$

Mathematica would, of course, have to evaluate,

```
Integrate[Power[Cos[\theta], 0], {\theta, 0, \pi/2}]
```

as $\pi/2$.

Exercise 31.10.21: How does Kapur plan to relate all of this to the information used by the MEP that results in the Cauchy distribution?

Solution to Exercise 31.10.21

So far, we have looked at the consequences of the universal constraint function as one piece of information. But we have another piece of information to process via the MEP and that is the average of the constraint function,

$$F(x) = \ln(1 + x^2)$$

With this in mind, Kapur then examines the integral,

$$\int_0^{\pi/2} \cos^2(\theta)^m \ln[\cos^2(\theta)] d\theta = f'(m)$$

whose solution is curiously the first derivative of the solution to the universal constraint just found and where the digamma function $\psi(m)$ makes its appearance.

$$f'(m) = \frac{\Gamma(\frac{1}{2}) \Gamma(m + \frac{1}{2}) [\psi(m + \frac{1}{2}) - \psi(m + 1)]}{2 \Gamma(m + 1)}$$

In terms of b , this is,

$$f'(b) = \frac{\Gamma(\frac{1}{2}) \Gamma(b - \frac{1}{2}) [\psi(b - \frac{1}{2}) - \psi(b)]}{2 \Gamma(b)}$$

Exercise 31.10.22: Kapur then asserts something amazing. Confirm his assertion.

Solution to Exercise 31.10.22

Kapur says that the average of the constraint function is some number c which can be calculated as,

$$c = -\frac{f'(m)}{f(m)}$$

Sure enough, substituting for $f(m)$ and $f'(m)$ and then canceling,

$$\begin{aligned} c &= -\frac{f'(m)}{f(m)} \\ &= -\left[\frac{\frac{\Gamma(1/2) \Gamma(m+1/2)}{2\Gamma(m+1)} [\psi(m+\frac{1}{2}) - \psi(m+1)]}{\frac{\Gamma(1/2) \Gamma(m+1/2)}{2\Gamma(m+1)}} \right] \\ &= \psi(m+1) - \psi\left(m + \frac{1}{2}\right) \\ &= \psi(b) - \psi\left(b - \frac{1}{2}\right) \end{aligned}$$

Exercise 31.10.23: Convert the parameter c into a *Mathematica* function.

Solution to Exercise 31.10.23

Either of these two functions will work,

```
cb[b_] := PolyGamma[b] - PolyGamma[b - 1/2]
cm[m_] := PolyGamma[m + 1] - PolyGamma[m + 1/2]
```

cb[1] or **cm[0]** returns our infamous number of 1.38629 for the average of the constraint function $F(x) = \ln(1 + x^2)$, bringing us full circle.

Exercise 31.10.24: Use *Mathematica* to find the derivative of the log of the Gamma function.

Solution to Exercise 31.10.24

Continuing our reliance on *Mathematica* to assist us in these abstruse matters, we find that,

```
D[Log[Gamma[\lambda]], \lambda]
```

is something *Mathematica* expresses as **PolyGamma[0, λ]**, or in more traditional mathematical notation as $\psi^{(0)}(\lambda)$. These are the digamma functions $\psi(\frac{1}{2})$ and $\psi(m + 1)$ appearing above in $f'(m)$.

Exercise 31.10.25: Construct the curve showing the relationship between the parameters b and c .

Solution to Exercise 31.10.25

With the development of the function **cb[]** in Exercise 31.10.23 that finds the constraint function average as a function of b , it is straightforward to plot their relationship. Figure 31.7 presents the curve relating the parameters b and c as Kapur labeled them, or more familiarly, as the relationship between the dual parameters λ and $\langle F \rangle$.

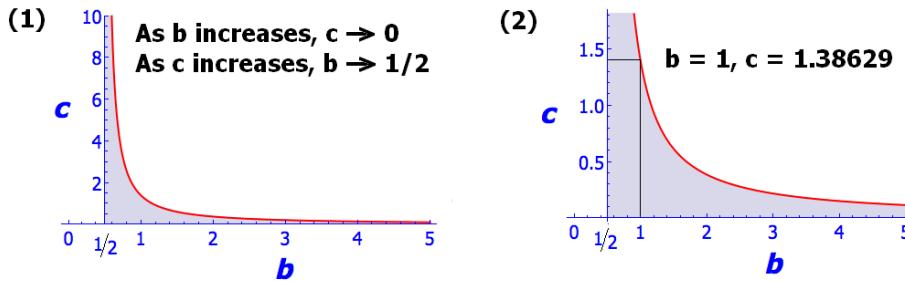


Figure 31.7: A plot of the highly non-linear relationship between Kapur's parameters b and c in the MEP derivation of the Cauchy distribution.

This relationship is seen to be highly non-linear one. In panel (1) with larger values on the c -axis, we see that as one dual parameter $b \rightarrow \infty$, the other dual parameter $c \rightarrow 0$. Likewise, as $b \rightarrow 1/2$, $c \rightarrow \infty$. In panel (2), we've blown up the central part of the curve a bit to show where our simple Cauchy distribution lies with parameters of $b = 1$ and $c = 1.38629$.

This is a characteristic of the dual parameters we have commented on before. For example, in Chapter Eighteen we remarked on the trade-off made by the Lagrange parameter in order to reach the extreme values of the constraint function average.

Exercise 31.10.26: Begin the transition from the Cauchy distribution towards the Student- t distributions.

Solution to Exercise 31.10.26

In the last exercise, we've seen that it is possible to assess the general relationship between the Lagrange multiplier and the expected value of the constraint function.

The non-linear relationship between the two parameters in the standard MEP formula was plotted for the Cauchy distribution.

To begin, examine a new point on the curve. Suppose now that we were interested in setting $b = 2$ within the MEP formalism so that we generalize the Cauchy distribution to,

$$pdf(x | \mathcal{M}_{Ca}) = \frac{1}{Z(b)} \frac{1}{(1+x^2)^2}$$

with the outcome that now the partition function is $Z(b) = \frac{\pi}{2}$ instead of the original example where $Z(b) = \pi$. The value of $b = 2$ using the formula developed shows that the constraint function average must now be,

$$\langle F \rangle \equiv c = \psi(2) - \psi(3/2) = 0.386294$$

Figure 31.8 is the plot of this generalized Cauchy distribution indexed by model parameter $b = 2$ and its dual parameter $c = 0.386294$.

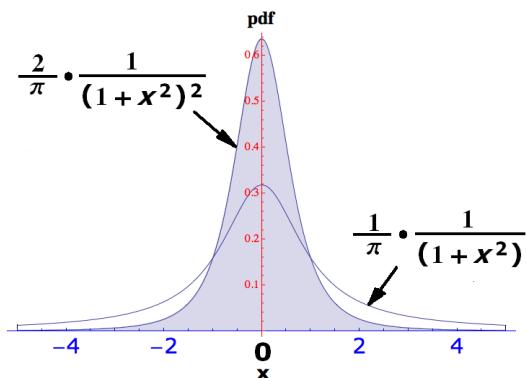


Figure 31.8: A more general Cauchy distribution that looks like the Student-t distribution.

Exercise 31.10.27: Do this exercise one more time by letting b take on another value.

Solution to Exercise 31.10.27

Figure 31.9 on the next page is the plot of a generalized Cauchy distribution indexed by model parameter $b = 3$ and its dual parameter $c = 0.219628$. The use of b and c here is just Kapur's notation for what we have been calling λ and $\langle F \rangle$. The probability density function found by the MEP formula is,

$$pdf(x | \mathcal{M}_{Ca}) = \frac{1}{Z(b)} \frac{1}{(1+x^2)^3} = \frac{8}{3\pi(1+x^2)^3}$$

Under the information in this model \mathcal{M}_{Ca} , we still are relying upon just the one constraint function $F(x) = \ln(1 + x^2)$. The expectation of this constraint function with respect to the general Cauchy distribution is the information inserted into the probability density function under this model,

$$\langle F \rangle \equiv E[F(x)] = c = \int F(x) \text{pdf}(x | \mathcal{M}_{Ca}) dx$$

Setting this parameter at a value of $c = 0.219628$ is the same as setting its dual parameter at $b = 3$ which is confirmed from **cb[3]**,

$$c = \psi(3) - \psi(2.5) = 0.219628$$

A direct integration to find the expectation of the constraint function yields,

$$c = E[F(x)] = \int_{-\infty}^{\infty} \ln(1 + x^2) \frac{8}{3\pi(1 + x^2)^3} dx = 0.219628$$

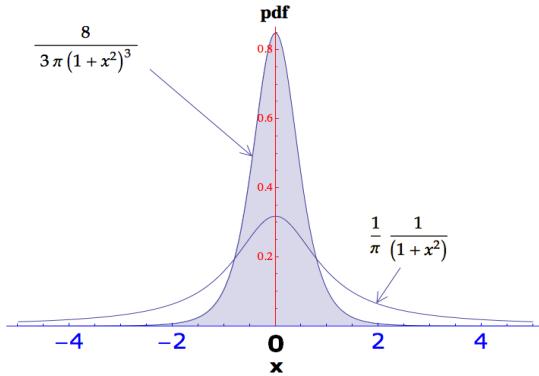


Figure 31.9: Another general Cauchy distribution.

Exercise 31.10.28: Verify the value of the partition function for the last exercise.

Solution to Exercise 31.10.28

The partition function $Z(b)$ is found for $b = 3$ as,

$$Z(b) = \int_{-\infty}^{\infty} \frac{1}{(1 + x^2)^3} dx = \frac{3\pi}{8}$$

With $b = 3$, $m = 2$, $2m = 4$, and $2b - 2 = 4$, this is the same as,

$$2 \int_0^{\pi/2} \cos^{(2b-2)}(\theta) d\theta \equiv 2 \int_0^{\pi/2} \cos^{2m}(\theta) d\theta \equiv 2 \int_0^{\pi/2} \cos^4(\theta) d\theta = \frac{3\pi}{8}$$

Chapter 32

Discovering Causal Models with the MEP

32.1 Introduction

In this final Chapter to Volume II, we would like to make inferences for a slightly more complicated problem than those we have dealt with so far. Our friends, the congenial kangaroos, have proved to be most willing companions for our various experiments. We will ask their indulgence once again to assist us in numerically illustrating our objectives.

The previous inferential problem was to predict what beer a kangaroo prefers given that we knew its hand preference and fur color. To this mix, we now add the kangaroo's intelligence. The kangaroos possess stout egos in these matters, and are perfectly amenable to being categorized in perhaps unflattering terms with regard to their intellectual prowess.

Our goal is to expand the setting that was treated in Chapter Twenty Two. There, in general terms, we wanted to leverage probability theory as a generalization of logic to predict a behavioral trait, beer preference, when conditioned on two observed physical traits of hand preference and fur color.

To augment our current repertoire of a kangaroo's physical traits which, to date, consist of only hand preference and fur color, we add a kangaroo's intelligence to the mix. Think of these now three easily measured physical traits as explanatory variables, interacting with the behavioral variable to be predicted, beer preference, within some sort of *causal model*. This scenario is a caricature, but not that far removed, of how science is actually conducted from an inferential viewpoint.

In Chapter Twenty Two, the objective was to introduce correlational models. Prediction using probability theory did not result in anything not already known if some causal association or relationship did not exist amongst some of the variables being considered.

One way to generate causal models of correlation, association, or relationships is to use the MEP. The language of main effects, together with double and triple interactions, borrowed from classic analysis of variance (ANOVA), was introduced as a way that the MEP could incorporate *information* about relationships into a probability distribution. As always, our primary reliance upon the MEP was to give us the actual numerical assignments to joint probabilities when conditioned on the differing information resident in the different correlational models under consideration.

One of our goals in this Chapter is to reverse the order in which we dissect the inferential problem. We are going to stress the importance of the *data* first and foremost rather than begin as we usually do by discussing some large space of models. Only *after* the data have been digested in the proper manner through the formal manipulation rules, and “black box” predictions made, will we delve into what might be candidates for tentative causal hypotheses behind the predictions.

Therefore, the example treated here is a data oriented problem with a substantial amount of data on all four variables. The formulas developed in Volume I, derived as they were from the formal manipulation rules, will be used to calculate a state of knowledge about the beer preference for any new kangaroo that was not part of the data base. The amazing feature of these formal rules is that they were derived as an average over *all conceivable numerical assignments*.

Therefore, one might not be too far wrong by branding the MEP as completely superfluous. Nonetheless, the MEP remains an invaluable tool. Its relevance is seen to reside not at the beginning of the inferential process in actually calculating all of these conceivable numerical assignments. It would be quite deflating to have almost every single one of these assignments eliminated by the data. Rather, the MEP is now seen as a much more useful tool *after* the data have picked out the most supported joint probabilities.

Then the question can be posed: What sort of causal models could have led to such joint probabilities? The MEP will come into its own as the IP explores what kind of information in correlational models might serve as a provocative stimulus for understanding some underlying physical causation.

Of course, all of this is done in a whimsical manner by talking about the beer preferences of kangaroos. The whimsy seeks to defuse any sense of rising anxiety that is surely swirling in the minds of many over taboo topics. The inferential problem is conducted in the context of answering the question: What beer would Oscar prefer since we know he is a kangaroo who is a right-handed beige colored idiot?

32.2 The Data and the Contingency Table

We emphasized in the **Introduction** that we were going to reverse our usual order by presenting the actual data gathered about the kangaroos and their four traits. To that end, a contingency table is laid out with the cells in the table containing the frequency counts for the particular joint statement indexed by that cell. For clarity, the marginal sums of the frequency counts are not displayed along the appropriate margins of the table, but appear instead in their own separate table.

Each cell will index a joint statement about beer preference, hand preference, fur color, and intelligence. Beer preference, hand preference, and fur color all have only two possible measurements, while intelligence is measured by placement into one of seven categories. The two possible measurements for beer preference, hand preference, and fur color are familiar from previous examples. The seven categories for kangaroo intelligence are labeled, starting from the lowest and proceeding through to the highest intelligence level, as:

Category 1. IDIOT

Category 2. DULL

Category 3. SLOW

Category 4. AVERAGE

Category 5. SHARP

Category 6. BRILLIANT

Category 7. GENIUS

Just like the joint probability table, the contingency table consists of a total of n cells where n is the dimension of the state space. Here, $n = 2 \times 2 \times 2 \times 7 = 56$. Figure 32.1, comprising the entire next page, shows the contingency table containing the data. Each of the 56 cells holds the frequency count from an overall sample of $N = 1000$ kangaroos. The next page after that shows some of the marginal sums from the contingency table. Also shown are the comparable normed frequency counts obtained upon dividing by $N = 1000$.

It is of paramount importance to always keep in mind the conceptual distinction between any marginal normed frequency count, another aspect of the *data*, and the *information* in the mathematical expectation of a constraint function. They may be exactly the same numerically, but they represent different concepts.

Rather than represent intelligence by a Gaussian distribution as you might have expected, I want to emphasize the generality of the scheme developed so far. By that I mean any kind of departure from normality can be modeled with the constraint functions and their averages. Perhaps, more interestingly, interactions between any intelligence level and beer preference can be examined individually and separately, rather than the correlation forced upon us by adopting the Gaussian distribution.

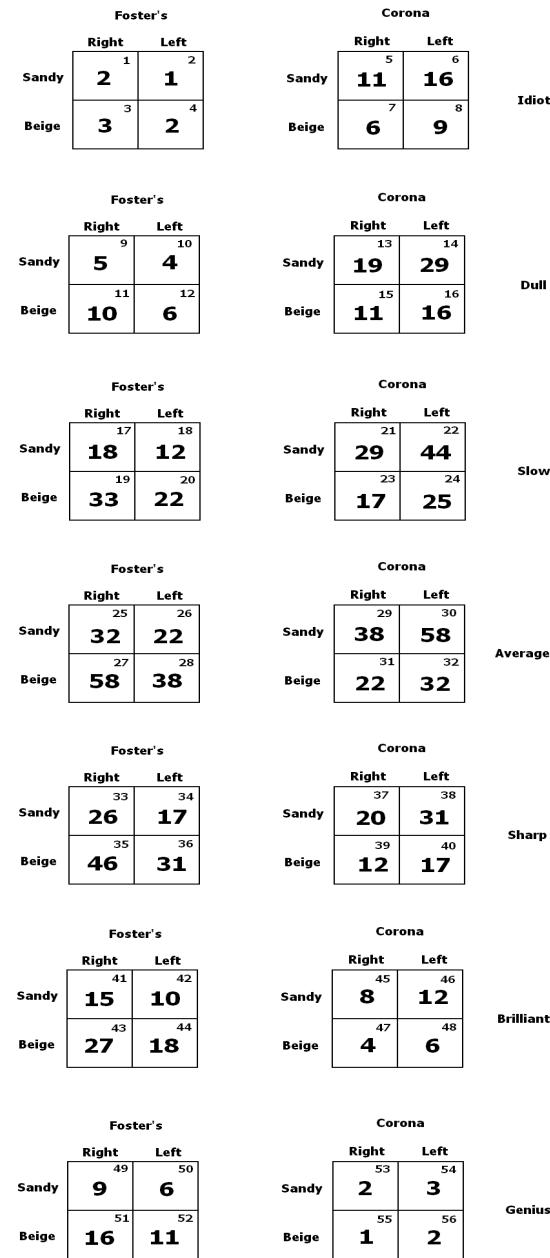


Figure 32.1: The contingency table containing the data for the numerical example of assessing a kangaroo's beer preference when conditioned on three known physical traits as well as the data. The total amount of data is $N = 1000$.

Table 32.1 below lists some of the marginal sums computed from the contingency table. The first ten entries comprise the marginal sums over the main effects. The second nine marginal sums comprise the double interactions of beer preference by hand preference, beer preference by fur color, and finally beer preference by intelligence.

Table 32.1: Some significant marginal sums from the data in the contingency table of Figure 32.1.

Marginal Sums	Frequency Counts	Normed Frequencies
Foster's drinker	500	0.50
Right handed	500	0.50
Sandy fur color	500	0.50
Idiot	50	0.05
Dull	100	0.10
Slow	200	0.20
Average	300	0.30
Sharp	200	0.20
Brilliant	100	0.10
Genius	50	0.05
Foster's and right hand	300	0.300
Foster's and sandy fur color	180	0.180
Foster's and idiot	8	0.008
Foster's and dull	25	0.025
Foster's and slow	85	0.085
Foster's and average	150	0.150
Foster's and sharp	120	0.120
Foster's and brilliant	70	0.070
Foster's and genius	42	0.042

These data are highlighted because the subsequent models to be investigated incorporate the information from these same effects. The main effects and all interactions, except for the quadruple interactions, will always be sums over appropriate joint statements, or, in other words, sums over a number of appropriate cells in the contingency table.

For example, the Foster's main effect is a sum over all levels of hand preference, fur color, and intelligence. Curiously, the quadruple interactions will not be a sum over any other statement, but will always be the frequency count in one cell. For example, the first of the six quadruple interactions, $BHFI_1$, is the frequency count in cell 1.

32.3 An Inference about the Next Kangaroo

The formal manipulation rules provide us with generic templates for all of our inferences. The most utilized generic template of all is Bayes's Theorem, which, say, for three statements A , B , and C , would be written as,

$$P(A | B, C) = \frac{P(A, B, C)}{P(B, C)} = \frac{P(A, B, C)}{P(A, B, C) + P(\bar{A}, B, C)} \quad (32.1)$$

One lesson to be learned from this way of writing Bayes's Theorem is that it is the joint probabilities we are most concerned with. That is why I emphasized that the MEP was designed to assign numerical values to joint probabilities.

Bayes's Theorem as written above in Equation (32.1) states something that is true for all probabilities in general, irrespective of whether that probability is conditioned upon some model, such as $P(A, B, C | \mathcal{M}_k)$, or conditioned on some data, such as $P(A, B, C | \mathcal{D})$.

The formal manipulation rules have provided us with the formulas to compute any inference concerning the next kangaroo's beer preference. For a first look, consider the extension of the generic Bayes's Theorem notation as it relates to the probability of the next occurrence of statement B conditioned on the known explanatory variables, statements H , F , and I .

This probability is also conditioned on the past N occurrences of these four statements, in other words, it is a probability given that we have actually collected some data \mathcal{D} . Thus, our statements will have the subscript $N + 1$ attached.

$$P(B_{N+1} | H_{N+1}, F_{N+1}, I_{N+1}, \mathcal{D}) = \frac{P(B_{N+1}, H_{N+1}, F_{N+1}, I_{N+1} | \mathcal{D})}{P(H_{N+1}, F_{N+1}, I_{N+1} | \mathcal{D})} \quad (32.2)$$

Since it is the joint probabilities we must compute, turn the denominator into,

$$P(B_{N+1} | H_{N+1}, F_{N+1}, I_{N+1}, \mathcal{D}) = \frac{P(B_{N+1}, F_{N+1}, H_{N+1}, I_{N+1} | \mathcal{D})}{P(B_{N+1}, H_{N+1}, F_{N+1}, I_{N+1} | \mathcal{D}) + P(\bar{B}_{N+1}, H_{N+1}, F_{N+1}, I_{N+1} | \mathcal{D})} \quad (32.3)$$

Actually, we already possess a very general and powerful formula for the probability of any number of future frequency counts, M_1, M_2, \dots, M_n , conditioned on the data, N_1, N_2, \dots, N_n . These data are, of course, the past frequency counts as they appear in the contingency table of Figure 32.1.

This extensively used formula, developed initially in Volume I, is written as,

$$P(M_1, M_2, \dots, M_n | N_1, N_2, \dots, N_n) = C \times \frac{\prod_{i=1}^n (M_i + N_i)!}{\prod_{i=1}^n M_i!} \quad (32.4)$$

Since we are interested in just the next kangaroo, set $M = 1$. The probability of the joint statements in the numerator and denominator of Bayes's Theorem, as they appeared above in Equation (32.3),

$$P(B_{N+1}, H_{N+1}, F_{N+1}, I_{N+1} | \mathcal{D}) \text{ and } P(\overline{B}_{N+1}, H_{N+1}, F_{N+1}, I_{N+1} | \mathcal{D})$$

will be the probability that some particular $M_i = 1$ with the remaining $M_j = 0$.

In other words, we will be picking out two particular cells from a contingency table holding the future frequency counts. For example, the joint statement,

$$B_{N+1} = \text{Foster's}, H_{N+1} = \text{Right-handed}, F_{N+1} = \text{Sandy}, I_{N+1} = \text{Idiot}$$

is cell 1, and the joint statement,

$$\overline{B}_{N+1} = \text{Corona}, H_{N+1} = \text{Right-handed}, F_{N+1} = \text{Sandy}, I_{N+1} = \text{Idiot}$$

is cell 5, given the way we have constructed the contingency table.

When $M = 1$, Equation (32.4) reduces to an especially simple form,

$$\begin{aligned} P(M_i = 1, M_j = 0 | \mathcal{D}) &= C \times \frac{\prod_{i=1}^n (M_i + N_i)!}{\prod_{i=1}^n M_i!} \\ C &= \frac{M! (N + n - 1)!}{(\prod_{i=1}^n N_i!) (M + N + n - 1)!} \\ &= \frac{(N + n - 1)!}{\prod_{i=1}^n N_i! (N + n)!} \\ &= \frac{1}{\prod_{i=1}^n N_i! (N + n)} \\ P(M_i = 1, M_j = 0 | \mathcal{D}) &= \frac{1}{\prod_{i=1}^n N_i! (N + n)} \times \frac{\prod_{i=1}^n (M_i + N_i)!}{\prod_{i=1}^n M_i!} \\ &= \frac{N_1! N_2! \cdots (N_i + 1)! \cdots N_n!}{N_1! N_2! \cdots N_i! \cdots N_n! (N + n)} \\ &= \frac{N_i + 1}{N + n} \end{aligned}$$

The probability for a future occurrence in cell 1 with $M = 1$, $M_1 = 1$, $N_1 = 2$, $N = 1000$, $n = 56$, is then,

$$P(M_1 = 1, M_j = 0 | \mathcal{D}) = \frac{3}{1056}$$

and the probability for a future occurrence in cell 5 with $N_5 = 11$ is,

$$P(M_5 = 1, M_j = 0 | \mathcal{D}) = \frac{12}{1056}$$

Plugging these values into Bayes's Theorem, we find the embarrassingly simple answer that,

$$\begin{aligned}
 P(B_{N+1} | H_{N+1}, F_{N+1}, I_{N+1}, \mathcal{D}) &= \\
 &\frac{P(B_{N+1}, F_{N+1}, H_{N+1}, I_{N+1} | \mathcal{D})}{P(B_{N+1}, H_{N+1}, F_{N+1}, I_{N+1} | \mathcal{D}) + P(\overline{B}_{N+1}, H_{N+1}, F_{N+1}, I_{N+1} | \mathcal{D})} \\
 &= \frac{\frac{3}{1056}}{\frac{3}{1056} + \frac{12}{1056}} \\
 &= \frac{1}{5}
 \end{aligned}$$

This is that “black–box” answer, referred to previously, which, while undoubtedly correct, does not provide any insight into what kind of causal model might be operating in the background to give rise to the data that were collected. This is where the usefulness of the MEP asserts itself.

We are calling upon it *after* the data have been analyzed to help discover what kinds of information are in models strongly supported by the data. And here, of course, we are only interested in models with correlations among the variables.

It must be recognized that in the derivation of the general formula in Equation (32.4), all conceivable models were averaged over. So, it was not even necessary to employ the MEP at that earlier stage to assign numerical values to probabilities. The multiple integration had to assign every conceivable numerical assignment in order to find its answer. Now, though, it *is* of some interest to investigate which MEP models survived the onslaught of the data.

But before we do that, let's summarize our inferences about the next kangaroo's beer preference given observation of its three relevant physical traits together with all of the past data. Thus, 28 probabilities for preferring Foster's will be calculated. The remaining 28 probabilities for preferring Corona would simply be one minus the probability for preferring Foster's.

Refer to Table 32.2, comprising the entire next page, for all of the details. Once again, in somewhat of a caricature of a real world scientific problem, clear cut trends are visible for these few explanatory variables. The IP discerns that the kangaroo's intelligence has a major role in increasing its preference for Foster's. The probability for preferring Foster's rises steadily as the intelligence level category increases. Or, what is the same thing, the degree of belief that some next kangaroo prefers Corona steadily increases as its intelligence level decreases.

Table 32.2: The probability for any next, that is, the $(N + 1)^{st}$ kangaroo, to prefer Foster's given observation of the three physical traits of hand preference, fur color, and intelligence. The data were collected from $N = 1000$ kangaroos.

No.	Hand	Fur Color	Intelligence	$P(B = \text{Foster's} \mid H, F, I)$
1	Right	Sandy	Idiot	0.2000
2	Right	Sandy	Dull	0.2308
3	Right	Sandy	Slow	0.3878
4	Right	Sandy	Average	0.4583
5	Right	Sandy	Sharp	0.5625
6	Right	Sandy	Brilliant	0.6400
7	Right	Sandy	Genius	0.7692
8	Left	Sandy	Idiot	0.1053
9	Left	Sandy	Dull	0.1429
10	Left	Sandy	Slow	0.2241
11	Left	Sandy	Average	0.2805
12	Left	Sandy	Sharp	0.3600
13	Left	Sandy	Brilliant	0.4583
14	Left	Sandy	Genius	0.6364
15	Right	Beige	Idiot	0.3636
16	Right	Beige	Dull	0.4783
17	Right	Beige	Slow	0.6538
18	Right	Beige	Average	0.7195
19	Right	Beige	Sharp	0.7833
20	Right	Beige	Brilliant	0.8485
21	Right	Beige	Genius	0.8947
22	Left	Beige	Idiot	0.2308
23	Left	Beige	Dull	0.2917
24	Left	Beige	Slow	0.4694
25	Left	Beige	Average	0.5417
26	Left	Beige	Sharp	0.6400
27	Left	Beige	Brilliant	0.7308
28	Left	Beige	Genius	0.8000

Within this major trend due to intelligence, there are modulations due to hand preference and fur color. The effect due to intelligence is weakest for left handed sandy colored kangaroos, and strongest for right handed beige colored kangaroos. The other two possibilities for hand preference and fur color are intermediate in the strength of the association with intelligence.

32.4 MEP Models

Now that we can compute the probability for the behavioral trait of beer preference conditioned on any observed value of the physical traits of hand preference, fur color, and intelligence in a “black box” fashion, our attention would naturally turn to some sort of causal explanation for these results. This is where we rely upon the MEP to provide us with some models that use information matching the observed data so that we can inspect what kind of information is in these models. It is hoped that the information in these “good models” might provide us with some provocative clues as to what is going on in terms of the observed relationship amongst all four variables.

So, to repeat one of the main objectives of this Chapter, we are not using the MEP at the very beginning of the inferential process to find numerical assignments issuing from a huge number of models. Instead, we are using it after the fact, that is, after the data have winnowed down the huge space of models to something manageable, to help us discover causal associations amongst all of the variables. We are searching for correlational models supported by the data.

Nevertheless, this goal does not deter us from taking a preliminary global overview of the broad class of models as inspired by an ANOVA type decomposition. There are four main effects due to B , H , F , and I with one constraint function each for B , H , and F , and six constraint functions for I for a total of nine Lagrange multipliers as parameters in these simple models.

There are six double interactions. For example, BH requires one constraint function, and FI requires six constraint functions. All together, another 21 Lagrange multipliers are needed for models incorporating double interactions.

Triple interactions, as exemplified by BHF , require one constraint function, while BHI requires six constraint functions for another 19 Lagrange multipliers in these increasingly complicated models. The final quadruple interactions require six constraint functions.

All told, adding up the Lagrange multipliers within this hierarchy of increasingly complicated models, $n - 1 = 55$ degrees of freedom are consumed,

$$m = 9 \text{ (main)} + 21 \text{ (double)} + 19 \text{ (triple)} + 6 \text{ (quadruple)} = 55$$

See Exercise 32.9.1 for a summary of this overview of ANOVA inspired models.

The information in each model, as they progress from simple to complicated, consists, of course, in the average of however many constraint functions are to be included. For a simple model incorporating just the main effects, the information is the marginal probability for each of B , H , F , and the first six levels of intelligence. Once these have been specified, the marginal probabilities for \overline{B} , \overline{H} , \overline{F} , as well as the last intelligence category, have also been determined.

Models that incorporate the following information about these above marginal probabilities,

$$P(B) = P(H) = P(F) = 1/2 \text{ and } P(I_1) = 0.05, \dots, P(I_6) = 0.10$$

will be better supported than any other purely main effects model because this information matches the actual data. Refer back to Table 32.1.

Each constraint function, $F_j(X = x_i)$, will take the form of a vector with 56 elements. All 56 elements in every constraint function vector will consist of 0s and 1s. The pattern of 0s and 1s will be dictated by however one chooses to construct the joint probability table.

For example, the constraint function $F_1(X = x_i)$ that captures the marginal probability of B , looks like,

$$\left(\underbrace{1, 1, 1, 1, 0, 0, 0, 0}_{\text{First eight statements}}, \dots, \underbrace{1, 1, 1, 1, 0, 0, 0, 0}_{\text{Last eight statements}} \right)$$

because the statement ($B = \text{Foster's}$) constitute the first four cells of each eight cell sub-table at each of the seven intelligence levels. Refer back to the contingency table in Figure 32.1 to confirm this pattern.

The information in this constraint function is the average,

$$\langle F_1 \rangle = Q_1 + Q_2 + Q_3 + Q_4 + \dots + Q_9 + Q_{10} + Q_{11} + Q_{12} + \dots + Q_{49} + Q_{50} + Q_{51} + Q_{52} = 1/2$$

The sum of the Q_i shown above is the marginal probability, $P(B | \mathcal{M}_k) = 1/2$, under this model.

For any model that includes the constraint function $F_1(X = x_i)$, the MEP formula will include a term in the numerator like,

$$Q_i = \frac{e^{\lambda_1 F_1(X=x_i) + \dots + \lambda_j F_j(X=x_i) + \dots}}{Z(\lambda)}$$

Q_{49} will include the Lagrange multiplier λ_1 in the numerator, $e^{\lambda_1 + \dots + \lambda_j + \dots}$, for example, but Q_{48} will not. The exercises will delve into much more detail about where the Lagrange parameters appear as dictated by the MEP formula.

32.5 Search for Good Models

This section gives a very much abbreviated taste of what in actuality would be a time consuming search through the space of models trying to find that class of models deserving of more thorough study. Let's quickly dispense with any examination of independence models, not because they shouldn't be examined, but, as mentioned, merely because we want to quickly home in on a class of models with correlations.

A full independence model consists of the first $m = 9$ constraint functions and their expectations, that is, the marginal probabilities for B , H , F , and six Intelligence Levels, I_1 through I_6 . But we know that the probability for preferring Foster's given any combination of the three physical traits will always remain the same at $1/2$, the same as the marginal probability for B , under such an independence model.

For the physical traits to have any impact on the behavioral trait, as evidenced by the data, we must find a correlational model. So begin searching within the space of models incorporating the double interactions.

Construct the constraint function for the first double interaction BH , the one between beer preference and hand preference, and set its expectation to match the actual data. Add it to the already existing nine constraint functions of the full independence model. This model has inserted information about a relationship between beer preference and hand preference, so the probability for preferring Foster's must change when conditioned on whether the kangaroo is right or left handed. It will not change when conditioned on any measurement for fur color or intelligence.

Sure enough, the degree of belief in some kangaroo preferring Foster's when given the hand preference does change from $1/2$. However, in a manner to be investigated in the next section, the numerical assignments from this first correlational model are not very close to the data. We must continue searching for better correlational models.

So enlarge the number of parameters from $m = 10$ to $m = 11$ for new models incorporating information about not only all the marginal probabilities for the traits themselves, but now including the effects of both hand preference in the double interaction BH , and the effects of fur color in the double interaction BF . Once again, the probabilities for preferring Foster's change in the direction you would expect.

The probabilities change conditioned on whether the new kangaroo is left or right handed, and on whether its fur is colored sandy or beige. They do not change when conditioned on the kangaroo's intelligence. But, as before, it turns out that these models are not supported by the data very well either. So the search must continue.

Plod on in the same vein by now adding in turn all six interactions of beer preference with intelligence. Models with $m = 12, 13, 14, 15, 16$ and, finally, $m = 17$ parameters are considered.

At this stage, we have correlational models that incorporate information about associations between the behavioral trait and all three of the physical traits. This final double interaction model consisting of $m = 17$ parameters produces numerical assignments in the joint probability table that are reasonably well supported by the data.

Thus ends our abbreviated search through model space. The MEP formula was used, late in the inferential process, to generate numerical assignments in order to compare them with the data. It is a fascinating, but still largely unexplored question, not only here in our toy problem, but also for real world scientific issues, as to the efficacy and meaning of even more complicated models.

These increasingly complicated models would incorporate information from triple, quadruple, and possibly even higher order interactions when more explanatory variables are included in an inferential problem. Models incorporating information about higher order interactions are invariably implicated with what in ordinary discourse we call “coincidence,” “surprise,” or “accident.” In other words, it takes a very strange and peculiar combination of just the “right” settings for all of the explanatory variables to result in a high probability for something that ordinarily would be viewed as quite mundane.

32.6 Signal Plus Noise Models

These considerations naturally lead us into that extremely fascinating topic of data analysis discussed under the heading of “signal plus noise.” I hope to be able to address this topic in great detail later on, but right now I just want to present a numerical example based on our current scenario.

Consider for the moment a less involved inferential scenario where, instead of seven levels of intelligence for the kangaroos, we lump everything into just the two categories, high and low intelligence, in order to reduce the state space from dimension $n = 56$ to $n = 16$.

Our main goal in doing this is not so much to reduce the state space, but rather to make it easier to compute all possible interactions. The ANOVA type decomposition of the degrees of freedom will look at models that have a maximum of $m = n - 1 = 15$ constraint functions. The motivation is to account for some sort of “subtle signal” buried in “background noise.” The IP would prefer to “subtract out” the obscuring effects of the noise in order to clearly perceive the signal.

If we define the noise to be all those main effects and interactions not impacting the inference, then what remains is the signal. Set up a baseline model with $m = 8$ constraint functions and their averages to represent the information in the noise. These would be the four main effects of B , H , F , I , the three double interactions of HF , HI , FI , and the triple interaction of HFI .

This leaves models using up the information in the remaining $m = 7$ constraint functions and their averages. These models would involve the three double interactions of BH , BF , and BI , the three triple interactions of BHF , BHI , and BFI , and the one quadruple interaction of $BHFI$.

Notice that what we have called the “noise” consists of completely legitimate interactions and main effects. These are no different than the equally legitimate interactions defining the “signal.”

The signal interactions achieve their special status because they change the probabilities for $P(B | H, F, I, \mathcal{D})$, while the information in the noise does not. Thus, the IP would like to “subtract out” the presently uninteresting correlations that don’t impact beer preference, in order to more clearly observe those correlations in the signal that do impact beer preference. This is merely another example of the saying, “One man’s signal is another woman’s noise.”

In the exercises, we carry out an interesting numerical example motivated by this “signal plus noise” description of data analysis. Basically, we want to discover whether there is any justification for preferring some models implementing signal plus noise over the baseline model representing pure noise.

The analysis shows that the data support evidence for a signal. The signal involves information that hand preference, fur color, and intelligence all impact the probability for beer preference when the “irrelevant noise” is subtracted out. Interestingly, these good signal models also *exclude* information from possible signal components consisting of the triple interactions and the quadruple interaction. This exclusion of unwanted information in models is a nice by-product of using the MEP.

32.7 Relative Status of Models

The primary prediction formula we are using was derived from the formal manipulation rules as an average weighted by the updated probability for all the models,

$$P(B_{N+1}, H_{N+1}, F_{N+1}, I_{N+1} | \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(B_{N+1}, H_{N+1}, F_{N+1}, I_{N+1} | \mathcal{M}_k) P(\mathcal{M}_k | \mathcal{D}) \quad (32.5)$$

We already have the “black box” answer for the left hand side of Equation (32.5). Our concern now focuses on $P(\mathcal{M}_k | \mathcal{D})$, the second term on the right hand side. Which models were heavily weighted by the data in forming the average prediction under all models?

As we’ve discussed many times before, the relative status of any two models when conditioned on the data is,

$$\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} = \frac{P(\mathcal{D} | \mathcal{M}_A)}{P(\mathcal{D} | \mathcal{M}_B)} \times \frac{P(\mathcal{M}_A)}{P(\mathcal{M}_B)} \quad (32.6)$$

When the IP begins the inferential process, that is, prior to any data, it is completely uninformed. Therefore, the ratio of any two models is always 1. When a log transform is taken, the *log likelihood ratio*, the log transform of the first term on the right hand side of Equation (32.6), is of paramount importance.

With the MEP formula at our disposal, this becomes the *entropic-like* expression using sample averages instead of constraint function averages,

$$\ln \left[\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} \right] = N \left[\sum_{j=1}^m (\lambda_j^A - \lambda_j^B) \bar{F}_j + \ln \left(\frac{Z_B}{Z_A} \right) \right] \quad (32.7)$$

studied earlier in Chapter Twenty Nine. If we want to emphasize the perspective as seen from information geometry, then we might express this formula by including the relative entropy,

$$\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} = \exp [N \times KL(p, q)] \quad (32.8)$$

Either of these equivalent formulas can be leveraged in the search for good models supported by the data. The relative status of any two models can be compared, after the data, with regard to their impact in forming the average in Equation (32.5).

The IP can fairly quickly discern that most models have an inconsequential impact, and then proceed to zoom in on that class of good models where their weighting within Equation (32.5) does play a role. Once again, we relegate these concerns to the exercises where you may investigate them at your leisure.

32.8 Connections to the Literature

In this final Chapter to Volume II, I pay homage to Jaynes's essentially correct insights into how an inference should be conducted. The most salient of these is his constant emphasis on following the formal manipulation rules of probability theory. Secondly, his refusal to buckle under to conventional wisdom by demonstrating the compelling logic behind Laplace's *Rule of Succession* is to be admired.

His insights are the linchpin not only to the formal prediction equations allowing for the *Rule of Succession*, but have served as an invaluable guide throughout my writing of these books.

The reader might reasonably question the sincerity of my appreciation of Jaynes given the many critical comments attached to the following annotation. But my remarks are easily recognizable as a common human weakness; we criticize only those we care about. It is not worth my while to criticize those unfortunates who don't have the dimmest perception of what inferencing is all about.

My comments in this section are confined to an annotated version of Jaynes's section 5, Chapter 18, entitled "An application," as it appears in [23],

Probability Theory: The Logic of Science

Jaynes's development within his section 5 are

quoted verbatim.

Whenever I wish to interject with my annotations, my commentary is indicated by **bold typeface**.

The title of Jaynes's Chapter 18 is, "The A_p distribution and rule of succession." I have wished to go on record for quite some time now with my view as to what is going on in this part of Jaynes's book. I have a strong feeling that this particular Chapter of his book lends itself to grave misunderstandings. I would like to do my best to clarify the confusion to whatever extent I am able. This section represents the first step in the full annotation of Chapter 18.

Basically, Jaynes presents this section as "an application" of the uniform distribution over model space to verify Laplace's *Rule of Succession* for the simple case where some statement can be either true or false. Jaynes generalizes the *Rule of Succession* later on in Section 18.10. Unfortunately, his rationale for the mysterious A_p distribution is so obscure that it destroys his otherwise fine presentation of the essential mathematical ingredients involved in the *Rule of Succession*.

Perhaps most astonishing of all is that most of this material was actually written in the mid-1950s! What appears in section 5 is essentially unchanged from Jaynes's Socony Mobil Oil Lectures published in February 1958. Thus, the language, the notation, and expressions that strike one as odd at first (and second) readings now become more understandable when the time frame is factored in. The realization that most of this was written in the 1950s dispels some of the head scratching over language like "inner and outer robots," and the not very well developed notation.

Pre-eminent among many confusing issues is what Jaynes mysteriously calls an " A_p distribution." Now, I would be willing to bet that not one in a thousand understands what he means by this " A_p distribution." To cut the long-winded explanation very short, everything becomes remarkably clear if one just substitutes $P(\mathcal{M}_k)$ whenever Jaynes writes or mentions this " A_p distribution."

Despite these shortcomings, Jaynes teaches us two very important lessons in his explanation of Laplace's *Rule of Succession*. First, it is directly derivable in just a few steps using the formal manipulation rules of probability theory. And secondly, a uniform distribution MUST be assigned over the space of models to arrive at the probability of future frequency counts when conditioned on past frequency counts.

The next three subsections commence with the detailed annotation. Each of my subsections arbitrarily divides up Jaynes's section 5 by the pages where the material appeared. The reader may assume that if I am not commenting on what Jaynes wrote, I agree with him.

32.8.1 Page 561

We begin with Jaynes as he sets up the ensuing discussion.

Now let's imagine that a "random" experiment is being performed. From the results of the experiment in the past, we want to do the best job we can of predicting results in the future. To make the problem a definite one, introduce the propositions:

$X \equiv$ For each trial we admit two prior hypotheses: A true, and A false.

It is best to reserve the phrase "prior hypotheses" for the model space. Adhering to the definitions I have set forth, " A true and A false" refers to a two dimensional state space with $n = 2$. This state space consists of the two statements $A = a_1$ and $A = a_2$.

This can mean variously that A was observed in one of two conditions, A can only be measured as a_1 or a_2 , A can be placed unambiguously into only one of two categories, or, A true and A false. I have usually illustrated this case with the canonical coin flip scenario where A can only be observed as HEADS or TAILS, but the Fermi oscillator can also serve as an example.

The underlying "causal mechanism" is assumed the same at every trial. This means, for example, that (1) the probability assigned to A at the n th trial does not depend on n , and (2) evidence concerning the results of past trials retains its relevance for all time; thus for predicting the outcome of trial 100, knowledge of the result of trial 1 is just as relevant as is knowledge of the result of trial 99. There is no other prior evidence.

$N_n \equiv A$ true n times in N trials in the past.

$M_m \equiv A$ true m times in M trials in the future.

For me, as it was for Laplace, the underlying "causal mechanism" is the model assigning a value of q to A true and $(1 - q)$ to A false. The value of q and $(1 - q)$ as assigned by some model, $q \equiv P(A = a_1 | \mathcal{M}_k)$, stays the same at every trial. Thus, I agree with Jaynes that, say, for the first two trials,

$$P(A_2, A_1, \mathcal{M}_k) = P(A_2 | A_1, \mathcal{M}_k) P(A_1 | \mathcal{M}_k) P(\mathcal{M}_k)$$

$$P(A_2 = a_1, A_1 = a_2, \mathcal{M}_k) = q \times (1 - q) \times P(\mathcal{M}_k)$$

And, extrapolating out to, say, 100 trials,

$$P(A_{100} = a_2, A_{99} = a_2, \dots, A_1 = a_1, \mathcal{M}_k) = (1 - q) \times (1 - q) \times \dots \times q \times P(\mathcal{M}_k)$$

In my notation, N is the total number of past trials, N_1 is the total number of times A was true in the past, and N_2 is the total number of times A was false in the past. M is the total number of future trials, M_1 is the total number of times A will be true in the future, and M_2 is the total number of times A will be false in the future.

Thus, even though we have a definite time ordering for 100 trials as in the above example leading to a multiplication of probabilities like,

$$q \times (1 - q) \times \cdots \times (1 - q)$$

the commutativity of multiplication says that we could re-arrange the order of the expression to,

$$q^{N_1} (1 - q)^{N_2}$$

When Jaynes says that the probability assigned to A at the n th trial does not depend on n , he is restricting the analysis, as I have also done up to this point, by not considering time series models where the assignment could depend on the trial number.

The verbal statement of X suffers from just the same ambiguities that we have found before, and which have caused so much trouble and controversy in the past. One of the important points we want to put across here is that we have not defined the prior information precisely until we have given, not just verbal statements, but equations which show how we have translated them into mathematics by specifying the prior probabilities to be used.

But there is nothing vague or ambiguous about statement X . It is a perfectly fine definition of the state space.

32.8.2 Page 562

In the present problem, this more precise statement of X is, as before,

$$(A_p | X) = 1 \quad 0 \leq p \leq 1 \tag{18.17}$$

with the additional understanding (part of the prior information for this particular problem) that the *same* A_p distribution is to be used for calculations pertaining to all trials.

Here, Jaynes has seriously confused us and, unfortunately, has confused himself. As mentioned at the outset, if you struggle through Jaynes's entire Chapter 18 trying to figure out what (A_p) really is, you eventually realize that it is simply $P(\mathcal{M}_k)$. So one could legitimately write out something like,

$$pdf(\mathcal{M}_k | I_0) = 1$$

to indicate the degree of belief about the models that are assigning every conceivable p (Here, I am using q). The IP's degree of belief in all of these models is the same.

In other words, the Information Processor is in a state of "total ignorance" about the causal mechanism, and uses a uniform probability density function taking on the value of 1 from $q = 0$ all the way through $q = 1$. This has nothing to do with a "more precise statement of X ," but rather everything to do with the important conceptual issue of specifying the relative standing of all the models assigning the numerical values q .

The model distribution, or (A_p) distribution, is never conditioned on the statement X . It is, in fact, just the other way around. The assignment of a probability to the statement X is always conditioned on the model as in, for example, when we write $P(A = a_1 | \mathcal{M}_k)$.

But overlooking this mistake, Jaynes is correct in telling us that in order to derive Laplace's Rule of Succession, we must adopt a uniform prior over model space. And he is also correct in emphasizing that the q and $(1 - q)$ are the same for all N trials. Eventually, all possible assignments to q are looked at during the course of the integration over q from 0 to 1.

What we are after is $P(M_m | N_n)$. Firstly, note that by many repetitions of our product and sum rules in the same way that we found Eq. (9.34), we have found the binomial distributions

$$\begin{aligned} P(N_n | A_p) &= \binom{N}{n} p^n (1-p)^{N-n} \\ P(M_m | A_p) &= \binom{M}{m} p^m (1-p)^{M-m} \end{aligned} \quad (18.18)$$

and at this point we see that, although A_p sounds like an awfully dogmatic and indefensible statement to us in the way we introduced it, this is actually the way in which probability *is* introduced in almost all present textbooks.

A better and clearer notation for $P(M_m | N_n)$ is my $P(M_1, M_2 | N_1, N_2)$ for the case we are dealing with here where the dimension of the state space is just $n = 2$. But Jaynes is correct in stating that the ultimate objective is to make an inference about the future frequency counts M_1 and M_2 as conditioned on the past frequency counts N_1 and N_2 .

Also, everything is much clearer if instead of $P(N_n | A_p)$ we write,

$$P(N_1, N_2 | \mathcal{M}_k) \equiv W(N) q^{N_1} (1-q)^{N_2}$$

Jaynes got hung up on wanting to avoid at all costs the notion of introducing a "probability of a probability." That was the origin for the A_p distribution and the strange notation eschewing the probability symbol $P(\dots)$ by using parentheses around A_p (and other distributions) to indicate the distribution of A_p as (A_p) . But if he had just realized that \mathcal{M}_k , which does exactly the same thing as he wants his A_p to do, was simply another *statement*, then he could have saved us all the anguish by just writing a probability for a statement as in $P(\mathcal{M}_k)$.

Furthermore, and most importantly, \mathcal{M}_k should no longer be viewed as an "awfully dogmatic and indefensible statement," but rather exactly what probability theory tells us to do if we want to condition a numerical assignment of probability on the information contained within some model.

One postulates that an event possesses some intrinsic, "absolute" or "physical" probability, whose numerical value we can never determine exactly. Nevertheless, no one questions that such an "absolute" probability exists. Cramér (1946, p. 154), for example, takes it as his fundamental axiom. That is just as

dogmatic a statement as our A_p ; and we think it is, in fact, just our A_p . The equations we see in current textbooks are all like the two above: whenever p appears as a *given* number, an adequate notation would show that there is an A_p hiding invisibly in the right hand side of the probability symbols.

Here Jaynes is at his best in telling us that the concept of probability belongs to epistemology and not to ontology. He forcefully reminds us that probabilities are NEVER to be thought of as the physical property of an object, but rather as the embodiment of the information as held by an Information Processor. Every probability assignment is conditioned on the information resident within some model \mathcal{M}_k “hiding invisibly in the right hand side of the probability symbols.” Notice the “recent” reference (for Jaynes writing in the 1950s) to Cramér.

Mathematically, the main functional differences between what we are doing here and what is done in current textbooks are: (1) we recognize the existence of that right hand side of *all* probabilities, whether or not an A_p is hiding in them; and (2) thanks to Cox’s theorems, we are not afraid to use Bayes’ theorem to work any proposition – including A_p – back and forth from one side of our symbols to the other. In refusing to make free use of Bayes’s theorem, orthodox writers are depriving themselves of the most powerful single principle in probability theory. When a problem of inference is studied long enough, sometimes through a string of *ad hockeries* for decades, one is always forced eventually to a conclusion that could have been derived in three lines from Bayes’ theorem. But those cases refer to “external” probabilities at the interface between the robot and the outside world: now we shall see that Bayes’ theorem is equally powerful and indispensable for manipulating “inner” probabilities.

We arrive now at the heart of the matter. Jaynes begins the mathematical derivation which culminates in Laplace’s Rule of Succession for the specific case of the coin flip. Or more correctly, the state of knowledge for the very next coin flip after having observed the results from N previous coin flips.

Now we need to find the prior probability $P(N_n | X)$. This is determined already from $(A_p | X)$, for our trick of resolving a proposition into mutually exclusive alternatives gives us

$$\begin{aligned} P(N_n | X) &= \int_0^1 dp (N_n A_p | X) \\ &= \int_0^1 dp P(N_n | A_p)(A_p | X) \\ &= \binom{N}{n} \int_0^1 dp p^n (1-p)^{N-n} \end{aligned} \tag{18.19}$$

It is necessary to translate Jaynes's faulty notation in order to better understand what is going on.

$$\begin{aligned}
 P(N_n | X) &\equiv P(N_1, N_2 | \mathcal{I}_0) \\
 P(N_1, N_2 | \mathcal{I}_0) &\equiv P(\mathcal{D} | \mathcal{I}_0) \\
 \int_0^1 dp (N_n A_p | X) &\equiv \sum_{k=1}^{\mathcal{M}} P(\mathcal{D}, \mathcal{M}_k | \mathcal{I}_0) \\
 \sum_{k=1}^{\mathcal{M}} P(\mathcal{D}, \mathcal{M}_k | \mathcal{I}_0) &\equiv \sum_{k=1}^{\mathcal{M}} P(\mathcal{D} | \mathcal{M}_k, \mathcal{I}_0) P(\mathcal{M}_k | \mathcal{I}_0) \\
 \sum_{k=1}^{\mathcal{M}} P(\mathcal{D} | \mathcal{M}_k, \mathcal{I}_0) P(\mathcal{M}_k | \mathcal{I}_0) &\equiv \int_0^1 \frac{N!}{N_1! N_2!} q^{N_1} (1-q)^{N_2} P(q) dq \\
 \int_0^1 \frac{N!}{N_1! N_2!} q^{N_1} (1-q)^{N_2} P(q) dq &\equiv \frac{N!}{N_1! N_2!} \int_0^1 q^{N_1} (1-q)^{N_2} dq
 \end{aligned}$$

I would not want to label $P(N_n | X) \equiv P(\mathcal{D} | \mathcal{I}_0)$ as a “prior probability.” It is the probability for the data, the already observed coin tosses. Or, at least call it a marginal probability as it represents a marginalization over all the models. The most important step, of course, is the last one where a uniform distribution over model space is invoked.

32.8.3 Page 563

Jaynes continues with,

The integral we have to evaluate is the complete Beta–function:

$$\int_0^1 dx x^r (1-x)^s = \frac{r! s!}{(r+s+1)!} \quad (18.20)$$

Thus, we have

$$P(N_n | X) = \begin{cases} \frac{1}{N+1} & 0 \leq n \leq N \\ 0 & N < n \end{cases} \quad (18.21)$$

In my notation, this is,

$$\begin{aligned} \int_0^1 q^{N_1} (1-q)^{N_2} dq &= \frac{N_1! N_2!}{(N_1 + N_2 + 1)!} \\ P(N_1, N_2 | \mathcal{I}_0) &= \frac{N!}{N_1! N_2!} \int_0^1 q^{N_1} (1-q)^{N_2} dq \\ &= \frac{N!}{N_1! N_2!} \frac{N_1! N_2!}{(N+1)!} \\ &= \frac{1}{N+1} \end{aligned}$$

However, just at this point Jaynes makes an appallingly bad remark! Referring to this last result, he comments,

i.e. just the uniform distribution of maximum entropy:

It is quite clear that the Maximum Entropy Principle has not made an appearance anywhere in this derivation!!! It did not explicitly appear because it was not needed. The q values were integrated from 0 to 1, therefore the Maximum Entropy Principle was entirely superfluous. All of the possible legitimate q values were already assigned during the course of the integration!

Now, if someone were to single out, say, the particular numerical assignment $q = 0.75$, and inquire as to how information in the role of a model would lead to such an assignment, then the Maximum Entropy algorithm could be brought in for the explanation.

Furthermore, the MEP is used for making numerical assignments to the probabilities for the statements in the state space and nowhere else. The probability of N_1 frequency counts for HEADS and N_2 frequency counts for TAILS as $P(N_1, N_2) = 1/(N+1)$ does NOT derive from any application of MEP, but rather from a direct application of the formal rules of probability theory.

$P(M_m | X)$ is found similarly. Now we can turn (18.18) around by Bayes' theorem:

$$(A_p | N_n) = (A_p | X) \frac{P(N_n | A_p)}{P(N_p | X)} = (N+1)P(N_n | A_p) \quad (18.22)$$

In my notation, this is,

$$P(\mathcal{M}_k | \mathcal{D}) = \frac{P(\mathcal{D} | \mathcal{M}_k, \mathcal{I}_0) P(\mathcal{M}_k | \mathcal{I}_0)}{P(\mathcal{D} | \mathcal{I}_0)}$$

where $P(\mathcal{M}_k | \mathcal{I}_0)$ is still the uniform distribution for all of the q assignments. Thus, with

$$\begin{aligned} P(N_1, N_2 | \mathcal{I}_0) &= P(\mathcal{D} | \mathcal{I}_0) \\ &= \frac{1}{N+1} \\ P(\mathcal{M}_k | \mathcal{D}) &= \frac{P(\mathcal{D} | \mathcal{M}_k, \mathcal{I}_0)}{1/(N+1)} \\ &= (N+1) P(\mathcal{D} | \mathcal{M}_k) \end{aligned}$$

There is a typo in Jaynes's version where $P(N_p | X)$ should be $P(N_n | X)$.

and so finally the desired probability is

$$P(M_m | N_n) = \int_0^1 dp (M_m A_p | N_n) = \int_0^1 dp P(M_m | A_p N_n) (A_p | N_n) \quad (18.23)$$

Since $P(M_m | A_p N_n) = P(M_m | A_p)$ by the definition of A_p , we have worked out everything in the integrand. Substituting into (18.23), we have again an Eulerian integral, and our result is

$$P(M_m | N_n) = \frac{\binom{n+m}{n} \binom{N+M-n-m}{N-n}}{\binom{N+M+1}{M}} \quad (18.24)$$

There are a lot of steps to fill in even before we can get to (18.24). The most important thing Jaynes emphasizes here is that the probability for the future frequency counts depends only on the model, and is independent of the past frequency counts. This is an important conceptual point that I have also tried to stress in my derivations. All the details in arriving at Jaynes's Equation (18.24) are presented in Exercises 32.9.5 and 32.9.6. There the equivalence between my final result and Jaynes's formula is proven. The mathematical part is essentially over at this point. Jaynes's final effort to end Section 5 is a numerical example of (18.24) for assessing the state of knowledge about the very next trial.

Note that this is not the same as the hypergeometric distribution (3.22) of sampling theory. Let's look at this result first in the special case $M = m = 1$; it then reduces to the probability of A being true in the next trial, given that it has been true n time [sic] in the previous N trials. The result is

$$P(A | N_n) = \frac{n+1}{N+2} \quad (18.25)$$

We recognize Laplace's rule of succession, which we found before and discussed briefly in terms of urn sampling in (6.29)–(6.46). Now we need to discuss it more carefully, in a wider context.

For a numerical example, Jaynes picks a simple case where the rule of succession is used to find the probability for a success on the very next trial

after having observed some number of successes in the past. Jaynes doesn't even use his own notation here which should be $P(M_1 | N_n)$.

In any case, in my notation, since we are looking at the probability for A true (a success) at the very first future trial, we have $M = 1$, $M_1 = 1$, and $M_2 = 0$. We are told that A was true N_1 times in the past and false N_2 times in the past (N_1 past successes and N_2 past failures) over a total of N past trials. These N trials constitute the observed data.

Thus, we seek to calculate $P(M_1 = 1, M_2 = 0 | N_1, N_2)$. Substitute these values into the formula to find that,

$$\begin{aligned} P(M_1 = 1, M_2 = 0 | N_1, N_2) &= \frac{M!}{M_1! M_2!} \frac{(N+1)!}{N_1! N_2!} \frac{(M_1 + N_1)! (M_2 + N_2)!}{(M + N + 1)!} \\ &= \frac{1!}{1! 0!} \frac{(N+1)!}{N_1! N_2!} \frac{(N_1 + 1)! N_2!}{(N+1+1)!} \\ &= \frac{(N+1)!}{N_1! N_2!} \frac{(N_1 + 1)! N_2!}{(N+2)!} \\ &= \frac{N_1 + 1}{N + 2} \end{aligned}$$

which is the same as Jaynes's (and Laplace's) answer.

Jaynes goes on in section 18.10 to provide a more general formula covering more than two statements in the state space. His formula is exactly the same as my Equation (32.4), the workhorse formula for this Chapter, and which was first introduced in Chapter Eleven of Volume I. To repeat, these prediction formulas follow from the formal manipulation rules of probability theory, a flat prior for the models in model space, and the desire to calculate the probability of future frequency counts from knowledge of past frequency counts, the measured data.

Jaynes's Equation (18.43), his generalization of Equation (18.24), was written by him in this form,

$$P(m_1 \cdots m_K | n_1 \cdots n_K) = \frac{\binom{n_1+m_1}{n_1} \cdots \binom{n_k+m_k}{n_k}}{\binom{N+M+K-1}{M}} \quad (18.43)$$

Exercise 32.9.6 shows that Jaynes's combinatorial formula as expressed above in his Equation (18.24) is the same as my Equation (32.4) for the case $K \equiv n = 2$.

32.9 Solved Exercises for Chapter Thirty Two

Exercise 32.9.1: Prepare a summary table showing the breakdown of $m = 55$ constraint functions for the state space of dimension $n = 56$ as inspired by an ANOVA type decomposition.

Solution to Exercise 32.9.1

Using ANOVA terminology, there are 9 main effects, 21 double interactions, 19 triple interactions, and 6 quadruple interactions for a total of $m = 9 + 21 + 19 + 6 = 55$ constraint functions, or $n - 1$ degrees of freedom (df). See Table 32.3 below.

Table 32.3: The decomposition of $m = 55$ constraint functions into main effects together with all double, triple, and quadruple interactions following a typical ANOVA schema.

ANOVA effects	Labels	Combinations	df Breakdown
Main Effects	B H F I	$\binom{4}{1} = 4$	1 1 1 6
		Total	9
Double Interactions	BH BF BI HF HI FI	$\binom{4}{2} = 6$	1 1 6 1 6 6
		Total	21
Triple Interactions	BHF BFI BHI HFI	$\binom{4}{3} = 4$	1 6 6 6
		Total	19
Quadruple Interactions	BHFI	$\binom{4}{4} = 1$	6
		Total	6
Overall Total			55

The labels are B for beer preference, H for hand preference, F for fur color, and I for intelligence. BH stands for the double interaction involving beer preference and hand preference, HFI stands for the triple interaction involving hand preference, fur color and intelligence, and $BHFI$ stands for the quadruple interaction involving beer preference, hand preference, fur color, and intelligence.

The **Combinations** column shows how many potential interactions of a given type can be formed. In the final **df Breakdown** column, since there are only two possible measurements for B , H , and F , only one constraint function is required. Since there are seven possible measurements for I , six constraint functions are required. Any interaction involving I will also require six constraint functions.

Exercise 32.9.2: Examine in detail the double interaction between beer preference and intelligence category 6. What are the implications for the numerical assignments to all eight probabilities for brilliant kangaroos?

Solution to Exercise 32.9.2

Refer back to Table 32.3 in the previous exercise. The IP would have worked its way up to $m = 17$ parameters for models that included nine parameters for the main effects, two more parameters for the BH and BF interactions, and then six more parameters for the BI interactions. The constraint function $F_{17}(X = x_i)$ would be a vector with 56 elements consisting of all 0s and 1s capturing BI_6 . The first forty elements in the vector would be 0s, followed by four 1s, at positions 41, 42, 43, and 44 with all 0s completing the vector after that.

$$F_{17}(X = x_i) = \left(\underbrace{0, 0, 0, 0, 0, 0, 0, 0, \dots}_{\text{First forty elements}}, \underbrace{1, 1, 1, 1, 0, 0, 0, 0}_{\text{elements 41 through 48}}, \underbrace{0, 0, 0, 0, 0, 0, 0, 0}_{\text{elements 49 through 56}} \right)$$

The 1s in the vector appear in the position corresponding to the four cells in the joint probability table for the joint statements about Foster's preference and being brilliant. Obviously, the information wants to say something about the numerical values Q_{41} , Q_{42} , Q_{43} , and Q_{44} . Setting $\langle F_{17} \rangle = 0.07$ incorporates information about this interaction.

Now, the overall marginal probability for the Brilliant intelligence category had already been set by the main effects information when $P(I_6) = 0.10$. So the eight cells in the joint probability table, Q_{41} through Q_{48} , can not add up to more than 0.10.

In fact, under the full independence model, the numerical assignments to Q_{41} through Q_{48} were all equal to 0.0125. No correlational model could show an influence of intelligence on beer preference unless the information in $\langle F_{17} \rangle$ changed these assignments. This is what the particular model of $\langle F_{17} \rangle = 0.07$ accomplished.

The new assignments under this model as computed by the MEP formula were $Q_{41} = 0.0151$, $Q_{42} = 0.0101$, $Q_{43} = 0.0269$, and $Q_{44} = 0.0179$. Check that,

$$\sum_{i=1}^{56} F_{17}(X = x_i) Q_i = \langle F_{17} \rangle = Q_{41} + Q_{42} + Q_{43} + Q_{44} = 0.07$$

The assignments to the other four cells Q_{45} through Q_{48} completing the Brilliant category, must then add up to 0.03 under this model because of the constraint that overall marginal probability for Brilliant must equal 0.10.

Notice that because of these constraints placed on the assignments by previous information incorporated under earlier models, the information in BI_6 changed the assignments ever so slightly. For example, Q_{41} went from 0.0125 under the independence model to 0.0151 under a model with the additional information from the BH , BF , and BI interactions.

Exercise 32.9.3: Use *Mathematica* to compute the numerators in the MEP formula for various models. Examine the patterns revealed.

Solution to Exercise 32.9.3

Construct the following *Mathematica* function,

```
numerator[m_] := Exp[Dot[Map[Subscript[\lambda, #] &, Range[m]], 
    Take[constraintmatrix, m, All]]]
```

Evaluate **numerator[1]** to find the pattern of results for the 56 numerators in the MEP formula for a model with $m = 1$ parameter. This is the very primitive model which incorporates information only about the marginal probability for beer preference.

$$\overbrace{e^{\lambda_1}, e^{\lambda_1}, e^{\lambda_1}, e^{\lambda_1}, 1, 1, 1, 1, \dots, e^{\lambda_1}, e^{\lambda_1}, e^{\lambda_1}, e^{\lambda_1}}^{\text{first 8 numerators}}, \overbrace{1, 1, 1, 1, 1, 1, 1, 1}^{\text{last 8 numerators}}$$

The constraint vector for $F_1(X = x_i)$ has the pattern,

$$(\overbrace{1, 1, 1, 1, 0, 0, 0, 0}^{\text{first eight statements}}, \dots, \overbrace{1, 1, 1, 1, 0, 0, 0, 0}^{\text{last eight statements}})$$

so as the $F(X = x_i)$ rolls through its values, the numerators for successive Q_i will look like,

$$e^{\lambda_1 F_1(X=x_1)}, e^{\lambda_1 F_1(X=x_2)}, \dots, e^{\lambda_1 F_1(X=x_7)}, e^{\lambda_1 F_1(X=x_8)}$$

or, as *Mathematica* informed us, the first eight numerators as well as the last eight numerators assume the pattern,

$$e^{\lambda_1 \times 1}, e^{\lambda_1 \times 1}, \dots, e^{\lambda_1 \times 0} e^{\lambda_1 \times 0} = e^{\lambda_1}, e^{\lambda_1}, e^{\lambda_1}, e^{\lambda_1}, 1, 1, 1, 1$$

If the information about the marginal probability for beer preference is that it equals 1/2, then $\lambda_1 = 0$. All 56 numerators will equal 1, and the partition function is $Z(\lambda_1) = 56$. Every $Q_i = 1/56$.

Exercise 32.9.4: What patterns are revealed in the next model with $m = 2$ parameters?

Solution to Exercise 32.9.4

Under a model with $m = 2$ parameters, information is included about the marginal probability for hand preference in addition to beer preference in this slightly more complicated model. Evaluating **numerator[2]** yields the repeating pattern,

$$\overbrace{e^{\lambda_1+\lambda_2}, e^{\lambda_1}, e^{\lambda_1+\lambda_2}, e^{\lambda_1}, e^{\lambda_2}, 1, e^{\lambda_2}, 1}^{\text{first 8 numerators}}, \dots, \overbrace{e^{\lambda_1+\lambda_2}, e^{\lambda_1}, e^{\lambda_1+\lambda_2}, e^{\lambda_1}, e^{\lambda_2}, 1, e^{\lambda_2}, 1}^{\text{last 8 numerators}}$$

The constraint vector for $F_2(X = x_i)$ has the pattern,

$$(\underbrace{1, 0, 1, 0, 1, 0, 1, 0}_{\text{first eight statements}}, \dots, \underbrace{1, 0, 1, 0, 1, 0, 1, 0}_{\text{last eight statements}})$$

because, given the way the joint probability table was constructed, cells 1, 3, 5, 7, \dots , 49, 51, 53, 55 are where the statements about right handed preference are located. The numerators for successive Q_i will then look like,

$$e^{\lambda_1 F_1(X=x_1) + \lambda_2 F_2(X=x_1)}, e^{\lambda_1 F_1(X=x_2) + \lambda_2 F_2(X=x_2)}, \dots$$

and substituting in the appropriate 1s and 0s from the two constraint functions,

$$e^{(\lambda_1 \times 1) + (\lambda_2 \times 1)}, e^{(\lambda_1 \times 1) + (\lambda_2 \times 0)}, \dots, e^{(\lambda_1 \times 0) + (\lambda_2 \times 0)}$$

confirming the *Mathematica* result above.

If the constraint function average for this second constraint function, like the first constraint function, is set to match the marginal probability for hand preference, then,

$$\sum_{i=1}^n F_2(X = x_i) Q_i = \langle F_2 \rangle = Q_1 + Q_3 + Q_5 + Q_7 + \dots + Q_{49} + Q_{51} + Q_{53} + Q_{55} = 1/2$$

The two parameters λ_1 and λ_2 must both equal 0 in order to satisfy these two marginal probabilities, the universal constraint, and still produce a numerical assignment that has maximum entropy.

Exercise 32.9.5: Work through the details to arrive at Jaynes's Equation (18.24).

Solution to Exercise 32.9.5

Picking up the derivation where we left it in section 32.8.3,

$$\begin{aligned}
 P(M_1, M_2 | N_1, N_2) &= \int_0^1 P(M_1, M_2, \mathcal{M}_k | N_1, N_2) dq \\
 &= \int_0^1 P(M_1, M_2 | \mathcal{M}_k, N_1, N_2) P(\mathcal{M}_k | N_1, N_2) dq \\
 &= \int_0^1 P(M_1, M_2 | \mathcal{M}_k) P(\mathcal{M}_k | N_1, N_2) dq \\
 &= \int_0^1 P(M_1, M_2 | \mathcal{M}_k) (N+1) P(\mathcal{D} | \mathcal{M}_k) dq \\
 &= \int_0^1 P(M_1, M_2 | \mathcal{M}_k) (N+1) W(N) q^{N_1} (1-q)^{N_2} dq \\
 &= \int_0^1 W(M) q^{M_1} (1-q)^{M_2} (N+1) W(N) q^{N_1} (1-q)^{N_2} dq \\
 &= W(M) W(N) (N+1) \int_0^1 q^{M_1} (1-q)^{M_2} q^{N_1} (1-q)^{N_2} dq \\
 &= W(M) W(N) (N+1) \int_0^1 q^{M_1+N_1} (1-q)^{M_2+N_2} dq \\
 &= W(M) W(N) (N+1) \frac{(M_1+N_1)! (M_2+N_2)!}{(M_1+N_1+M_2+N_2+1)!} \\
 &= W(M) W(N) (N+1) \frac{(M_1+N_1)! (M_2+N_2)!}{(M+N+1)!} \\
 P(M_1, M_2 | N_1, N_2) &= \frac{M!}{M_1! M_2!} \frac{N!}{N_1! N_2!} (N+1) \frac{(M_1+N_1)! (M_2+N_2)!}{(M+N+1)!} \\
 &= \frac{M!}{M_1! M_2!} \frac{(N+1)!}{N_1! N_2!} \frac{(M_1+N_1)! (M_2+N_2)!}{(M+N+1)!} \\
 &= \frac{M! (N+1)!}{N_1! N_2! (M+N+1)!} \times \frac{(M_1+N_1)! (M_2+N_2)!}{M_1! M_2!} \\
 &= C \times \frac{\prod_{i=1}^2 (M_i + N_i)!}{\prod_{i=1}^2 M_i!}
 \end{aligned}$$

Exercise 32.9.6: Work through the details to arrive at Jaynes's Equation (18.24) as displayed in the form of his combinatorial notation for the final result.

Solution to Exercise 32.9.6

$$\begin{aligned}
 \binom{n+m}{n} &= \binom{N_1 + M_1}{N_1} \\
 &= \frac{(M_1 + N_1)!}{N_1! M_1!} \\
 \binom{N+M-n-m}{N-n} &= \binom{N+M-N_1-M_1}{N-N_1} \\
 &= \frac{(N+M-N_1-M_1)!}{(N-N_1)! (M-M_1)!} \\
 &= \frac{(N_2 + M_2)!}{N_2! M_2!} \\
 \binom{n+m}{n} \binom{N+M-n-m}{N-n} &= \frac{(M_1 + N_1)!}{N_1! M_1!} \frac{(N_2 + M_2)!}{N_2! M_2!} \\
 \binom{N+M+1}{M} &= \frac{(M+N+1)!}{M! (N+1)!} \\
 \frac{\binom{n+m}{n} \binom{N+M-n-m}{N-n}}{\binom{N+M+1}{M}} &= \frac{\frac{(M_1 + N_1)!}{N_1! M_1!} \frac{(N_2 + M_2)!}{N_2! M_2!}}{\frac{(M+N+1)!}{M! (N+1)!}} \\
 &= \frac{M! (N+1)! (M_1 + N_1)! (M_2 + N_2)!}{N_1! M_1! N_2! M_2! (M+N+1)!} \\
 &= \frac{M! (N+1)!}{(M+N+1)! N_1! N_2!} \times \frac{(M_1 + N_1)! (M_2 + N_2)!}{M_1! M_2!}
 \end{aligned}$$

Exercise 32.9.7: Close out the annotated version of Jaynes's exposition of Laplace's Rule of Succession with one more numerical example.

Solution to Exercise 32.9.7

Jaynes ended his section 18.5 by showing that the probability of getting HEADS on the next toss of the coin after seeing N_1 HEADS and N_2 TAILS in the previous

N tosses of the coin was,

$$P(M_1 = 1, M_2 = 0 \mid N_1, N_2) = \frac{N_1 + 1}{N + 2}$$

Reflecting upon the next step from this initial example highlights some interesting situations. Suppose that, instead of the very next trial, the IP is interested in what would happen on the next *two* trials.

Three outcomes are possible: (1) Two HEADS, (2) two TAILS, and (3) one HEADS and one TAILS. We have discussed what the *Rule of Succession* says when there are no data at all and a uniform prior for model space is used. The coin has never been tossed, and under this state of total ignorance, the *Rule of Succession* leads us to a probability of $1/(M + 1) = 1/3$ for each one of these three outcomes.

Now, however, suppose that we *have* seen the coin tossed eight times, and know the results from these eight previous trials. This must change our state of knowledge somehow. We can not claim to be in a state of total ignorance any longer. The *Rule of Succession* was designed to take account of any past data and any number of future trials.

Thus, consider the state of knowledge about the future outcome of one HEADS and one TAILS given that we have observed four HEADS and four TAILS in eight previous flips of the coin. For the future frequency counts $M = 2$, $M_1 = 1$, $M_2 = 1$, and for the past frequency counts $N_1 = 4$, $N_2 = 4$, $N = 8$, we use the *Rule of Succession* to find $P(M_1 = 1, M_2 = 1 \mid N_1 = 4, N_2 = 4)$ as,

$$\begin{aligned} P(M_1, M_2 \mid N_1, N_2) &= \frac{M!}{(M + N + 1)!} \times \frac{(N + 1)!}{N_1! N_2!} \times \frac{(M_1 + N_1)! (M_2 + N_2)!}{M_1! M_2!} \\ &= \frac{2!}{11!} \times \frac{9!}{4! 4!} \times \frac{5! 5!}{1! 1!} \\ &= 5/11 \end{aligned}$$

After you think about this result for a little bit, it makes eminent sense. It's a little less than $1/2$, but greater than $1/3$. The answer would have been $1/3$ if we didn't have access to the data. It would have been $W(M) \times q(1-q) = 2 \times 1/4 = 1/2$ if there had been an infinite amount of data maintaining equality between the N_1 and N_2 frequency counts. In other words, by approaching an infinite amount of data with equality between the frequency counts would have supported the Dirac δ -function for the model space, $P(\mathcal{M}_k) = \delta(q - 1/2)$.

What about the other two outcomes? Their probabilities are both $3/11$. It's always nice to see that these three probabilities for the possible outcomes add up to 1. For two TAILS or two HEADS, the probability is less than $1/3$, but greater than $1/4$, which again makes complete sense. The IP is not in a state of total

ignorance where the probability would have been $1/3$. Instead, the IP has definitely been nudged away from $1/3$ in the direction of $P(\mathcal{M}_k) = \delta(q - 1/2)$ by the available data where the probability for either of these two outcomes would have been $1/4$.

Exercise 32.9.8: Recapitulate the derivation for the joint probability of the traits for the next kangaroo when conditioned on the data.

Solution to Exercise 32.9.8

This derivation follows closely the ones already presented in Volume I, Chapters Twelve and Fifteen by following the formal manipulation rules. But it is always a good idea to double check that these supposedly general formulas do in fact apply to specific cases when they arise. We start with Equation (32.5) in section 32.7.

$$\begin{aligned} P(B_{N+1}, H_{N+1}, F_{N+1}, I_{N+1} | \mathcal{D}) &= \\ \sum_{k=1}^{\mathcal{M}} P(B_{N+1}, H_{N+1}, F_{N+1}, I_{N+1} | \mathcal{M}_k) P(\mathcal{M}_k | \mathcal{D}) \end{aligned}$$

Any selected joint probability $P(B_{N+1}, H_{N+1}, F_{N+1}, I_{N+1} | \mathcal{M}_k)$, the first term in the summation on the right hand side, will be designated as Q_i . The more complicated expression revolves around $P(\mathcal{M}_k | \mathcal{D})$.

$$P(\mathcal{M}_k | \mathcal{D}) = \frac{P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k)}{P(\mathcal{D})}$$

$$P(\mathcal{D} | \mathcal{M}_k) = W(N) q_1^{N_1} q_2^{N_2} \cdots q_n^{N_n}$$

$$P(\mathcal{M}_k) = C_D \int \cdot \int q_1^{\alpha_1-1} q_2^{\alpha_2-1} \cdots q_n^{\alpha_n-1} dq_i$$

$$\begin{aligned} P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k) &= W(N) \times C_D \times \int \cdot \int q_1^{N_1} q_2^{N_2} \cdots q_n^{N_n} q_1^{\alpha_1-1} q_2^{\alpha_2-1} \cdots q_n^{\alpha_n-1} dq_i \\ &= W(N) \times C_D \times \int \cdot \int q_1^{N_1+\alpha_1-1} q_2^{N_2+\alpha_2-1} \cdots q_n^{N_n+\alpha_n-1} dq_i \end{aligned}$$

$$P(\mathcal{D}) = \frac{N! (n-1)!}{(N+n-1)!} \quad \text{Exercise 12.6.7, Volume I}$$

$$\frac{P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k)}{P(\mathcal{D})} = \frac{W(N) \times C_D \times \int \cdot \int q_1^{N_1+\alpha_1-1} q_2^{N_2+\alpha_2-1} \cdots q_n^{N_n+\alpha_n-1} dq_i}{\frac{N! (n-1)!}{(N+n-1)!}}$$

Now we are at,

$$\sum_{k=1}^{\mathcal{M}} P(B_{N+1}, H_{N+1}, F_{N+1}, I_{N+1} | \mathcal{M}_k) P(\mathcal{M}_k | \mathcal{D}) \equiv$$

$$\frac{W(N) \times C_D \times \int \cdots \int q_1^{N_1+\alpha_1-1} \cdots \overbrace{q_i^{N_i+1+\alpha_i-1}}^{\text{First term enters here!}} \cdots q_n^{N_n+\alpha_n-1} dq_i}{\frac{N! (n-1)!}{(N+n-1)!}}$$

With all $\alpha_i = 1$, the Dirichlet integral is equal to,

$$\int \cdots \int q_1^{N_1+\alpha_1-1} \cdots q_i^{N_i+1+\alpha_i-1} \cdots q_n^{N_n+\alpha_n-1} dq_i = \frac{N_1! N_2! \cdots (N_i+1)! \cdots N_n!}{(N+n)!}$$

Substituting,

$$\begin{aligned} & \sum_{k=1}^{\mathcal{M}} P(B_{N+1}, H_{N+1}, F_{N+1}, I_{N+1} | \mathcal{M}_k) P(\mathcal{M}_k | \mathcal{D}) \equiv \\ & \frac{W(N) \times C_D \times \frac{N_1! N_2! \cdots (N_i+1)! \cdots N_n!}{(N+n)!}}{\frac{N! (n-1)!}{(N+n-1)!}} \\ & = \frac{\frac{N!}{N_1! N_2! \cdots N_n!} \times C_D \times \frac{N_1! N_2! \cdots (N_i+1)! \cdots N_n!}{(N+n)!}}{\frac{N! (n-1)!}{(N+n-1)!}} \\ & = \frac{N! \times C_D \times \frac{N_i+1}{(N+n)!}}{\frac{N! (n-1)!}{(N+n-1)!}} \\ & = \frac{(N+n-1)!}{N! (n-1)!} \times N! \times C_D \times \frac{N_i+1}{(N+n)!} \\ & = \frac{1}{(n-1)!} \times C_D \times \frac{N_i+1}{N+n} \\ C_D & = \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\prod_{i=1}^n \Gamma(\alpha_i)} \\ C_D & = (n-1)! \end{aligned}$$

Since the IP started out completely uninformed about the relative status of the models, all $\alpha_i = 1$. The normalizing factor for the Dirichlet distribution is then easily computed.

The final result we were looking for is the amazingly and almost embarrassingly simple answer, but nonetheless an intuitively compelling one, that it is the past frequency counts that determine this probability for any joint statement concerning the next kangaroo's behavioral trait of beer preference, and three physical traits of hand preference, fur color and intelligence,

$$\begin{aligned} P(B_{N+1}, H_{N+1}, F_{N+1}, I_{N+1} \mid \mathcal{D}) &= \frac{1}{(n-1)!} \times (n-1)! \times \frac{N_i + 1}{N+n} \\ &= \frac{N_i + 1}{N+n} \end{aligned}$$

Exercise 32.9.9: What is the probability of the data?

Solution to Exercise 32.9.9

The probability of the data $P(\mathcal{D})$ appeared as the denominator in the expression,

$$P(\mathcal{M}_k \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathcal{M}_k) P(\mathcal{M}_k)}{P(\mathcal{D})}$$

This was solved as a sub-problem in the last exercise.

$$P(\mathcal{D}) = \frac{N! (n-1)!}{(N+n-1)!}$$

The data are the past frequency counts, and their probability is found in exactly the same way by following the formal manipulation rules.

$$\begin{aligned} P(N_1, N_2, \dots, N_n) &= \\ &\int \cdot \int_{\sum_{i=1}^n q_i = 1} P(N_1, N_2, \dots, N_n \mid q_1, q_2, \dots, q_n) P(q_1, q_2, \dots, q_n) dq_i \\ &= W(N) \times C_D \times \int \cdot \int q_1^{N_1+\alpha_1-1} q_2^{N_2+\alpha_2-1} \cdots q_n^{N_n+\alpha_n-1} dq_i \\ &= \frac{N!}{N_1! N_2! \cdots N_n!} \times \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\prod_{i=1}^n \Gamma(\alpha_i)} \times \frac{\prod_{i=1}^n \Gamma(N_i + \alpha_i)}{\Gamma(\sum_{i=1}^n N_i + \alpha_i)} \\ &= \frac{N!}{N_1! N_2! \cdots N_n!} \times (n-1)! \times \frac{\prod_{i=1}^n \Gamma(N_i + \alpha_i)}{\Gamma(N+n)} \\ &= \frac{N!}{N_1! N_2! \cdots N_n!} \times (n-1)! \times \frac{N_1! N_2! \cdots N_n!}{\Gamma(N+n)} \\ &= \frac{N! (n-1)!}{(N+n-1)!} \end{aligned}$$

For any possible data where both N and n are fixed, the probability of the data is a constant; therefore its cancelation is perfectly acceptable. This is no different than the extended discussion in Volume I as to why the probability of no HEADS, one HEADS, two HEADS, \dots ninety nine HEADS in the next ninety nine tosses of the coin all possess the same probability of,

$$P(N_1, N_2) = \frac{N! (n-1)!}{(N+n-1)!} = \frac{99! \times 1}{100!} = \frac{1}{N+1} = \frac{1}{100}$$

For our current numerical example, the probability for any possible observed data is the extremely small probability of,

$$P(\mathcal{D}) = \frac{1000! 55!}{1055!} = 2.8 \times 10^{-93}$$

The inverse of this number tells us that there are 3.57×10^{92} possible frequency counts, or contingency tables, with each possible contingency table having this same probability, just like in the coin toss.

Thus, we see that the likelihood ratio,

$$\frac{P(\mathcal{D} | \mathcal{M}_A)}{P(\mathcal{D} | \mathcal{M}_B)}$$

is really the crucial part of the computation for determining the relative status of the models in,

$$\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} = \frac{P(\mathcal{D} | \mathcal{M}_A)}{P(\mathcal{D} | \mathcal{M}_B)} \times \frac{P(\mathcal{M}_A)}{P(\mathcal{M}_B)}$$

Exercise 32.9.10: What is the most extreme specific application of the general result for the probability of future frequency counts given some data? By most extreme, I mean the easiest inferential scenario you can think of.

Solution to Exercise 32.9.10

The most general scenario refers to the probability that several joint statements A, B, C, \dots will occur in the future when conditioned on some known data that have taken place in the past.

$$P(A_{N+M} = a_1, B_{N+M} = b_2, C_{N+M} = c_3, \dots, Z_{N+M} = z_i | \mathcal{D}) = \\ P(M_1, M_2, \dots, M_n | N_1, N_2, \dots, N_n)$$

The “black box” answer is,

$$P(M_1, M_2, \dots, M_n | \mathcal{D}) = \frac{M! (N+n-1)!}{N_1! N_2! \dots N_n! (M+N+n-1)!} \times \frac{\prod_{i=1}^n (M_i + N_i)!}{\prod_{i=1}^n M_i!}$$

The most primitive specific application is to the case where the IP is assessing its state of knowledge about the *first* occurrence of some statement which can take on only *two* possible measurements when there are *no* data available. In addition, the IP is in a state of complete ignorance about the causes of the phenomena.

In this case, $n = 2$, $M = 1$, $M_1 = 0$ or 1 , $M_2 = 0$ or 1 , $N = 0$, $N_i = 0$, and $\alpha_i = 1$. The answer from the general formula (not from some inappropriately applied *Principle of Indifference*) is, of course,

$$\begin{aligned} P(M_1, M_2, \dots, M_n | \mathcal{D}) &= \frac{M! (N+n-1)!}{N_1! N_2! \dots N_n! (M+N+n-1)!} \times \frac{\prod_{i=1}^n (M_i + N_i)!}{\prod_{i=1}^n M_i!} \\ &= \frac{1! (0+2-1)!}{0! 0! \dots 0! (1+0+2-1)!} \times \frac{1!}{1!} \\ &= 1/2 \end{aligned}$$

Recapitulate the derivation from the beginning for the simplest possible specific application,

$$P(X_1 = x_1 | \mathcal{D}) = \sum_{k=1}^{\mathcal{M}} P(X_1 = x_1 | \mathcal{M}_k) P(\mathcal{M}_k | \mathcal{D})$$

Transitioning to an integration over the continuous model space where the q assignments are made,

$$P(X_1 = x_1 | \mathcal{D}) = \int_0^1 P(X_1 = x_1 | \mathcal{M}_k) P(\mathcal{M}_k | \mathcal{D}) dq$$

We repeat the usual series of transformations to find the answer for $M_1 = 1$,

$$\begin{aligned} &\int_0^1 P(X_1 = x_1 | \mathcal{M}_k) P(\mathcal{M}_k | \mathcal{D}) dq \\ &= \int_0^1 W(M) q_1^{M_1} q_2^{M_2} W(N) q_1^{N_1} q_2^{N_2} C_{Beta} q_1^{\alpha_1-1} q_2^{\alpha_2-1} dq_1 \\ &= W(M) \times W(N) \times C_{Beta} \times \int_0^1 q_1^{M_1} q_2^{M_2} q_1^{N_1} q_2^{N_2} q_1^{\alpha_1-1} q_2^{\alpha_2-1} dq_1 \\ &= W(1) \times W(0) \times C_{Beta} \times \int_0^1 q_1^1 q_2^0 q_1^0 q_2^0 q_1^0 q_2^0 dq_1 \\ &= W(1) \times W(0) \times C_{Beta} \times \int_0^1 q_1 dq_1 \\ &= \int_0^1 q_1 dq_1 \\ &= 1/2 \end{aligned}$$

Therefore, the probability of HEADS on the first toss of the coin when no previous data are available, and, furthermore, when the IP does not know anything about what might be causing the appearance of HEADS or TAILS must be $1/2!$ I mention this seemingly trivial example because the most widely held opinion for the answer of $1/2$ is founded on an erroneous rationale. This fallacious rationale is supported by an appeal to Laplace's **Principle of Insufficient Reason** that, since there are only two possibilities, and we are indifferent between them, the probability must be apportioned equally to the two propositions.

This is not true as the above derivation makes clear. There are two main conceptual points. First, the formal manipulation rules of probability must be followed in setting up the problem as a marginalization over the model space. Secondly, the **Principle of Insufficient Reason** comes into play with the uniform assignment over model space, that is when $P(\mathcal{M}_k)$ is distributed as a Dirichlet distribution with all $n \alpha_i$ parameters equal to 1. Refer back to Chapter Fifteen, Volume I, for the more in-depth discussion.

Exercise 32.9.11: Show the contingency table containing the data for the abbreviated kangaroo scenario.

Solution to Exercise 32.9.11

Suppose that $N = 1600$ kangaroos were measured (observed, categorized) on the four traits of beer preference, hand preference, fur color, and intelligence. The contingency table shown at the top of the next page in Figure 32.2 contains the data for the following seven exercises.

Since each of the four variables was measured as being in one of two categories, the table consists of 16 cells. Some of the marginal sums are displayed with the marginal sums for beer preference, hand preference, fur color, and intelligence shown at the bottom.

Exercise 32.9.12: What is the probability that the next kangaroo will drink Corona as his (or her) beer of preference, when that kangaroo just happens to be a beige colored, left-handed kangaroo with below average intelligence?

Solution to Exercise 32.9.12

This exercise emphasizes the primary importance of the formal manipulation rules for predicting the future. Prior to examining any model from the MEP perspective, the IP already has the means for answering any question like that posed in this exercise.

		B				
		H	\bar{H}			
		1	2			
F		85	64	149		
\bar{F}		214	117	331		
		299	181	480		
		H	\bar{H}			
		5	6			
		7	8			
F		155	16	171		
\bar{F}		138	11	149		I
		293	27	320	800	

		H	\bar{H}			
		9	10			
		11	12			
F		43	48	91		
\bar{F}		138	91	229		
		181	139	320		
		H	\bar{H}			
		13	14			
		15	16			
F		197	32	229		
\bar{F}		230	21	251		I
		427	53	480	800	

$$\begin{aligned} B &= 800 & F &= 640 \\ H &= 1200 & I &= 800 \end{aligned}$$

1600

Figure 32.2: The contingency table containing the data for a new numerical example of assessing a kangaroo's beer preference when conditioned on known physical traits. The intelligence categories have been reduced from seven in the original example to two categories in this abbreviated example.

$$P(\overline{B}_{N+1} | \overline{H}_{N+1}, \overline{F}_{N+1}, \overline{I}_{N+1}, \mathcal{D}) = \frac{N_{16} + 1}{(N_{16} + 1) + (N_{12} + 1)} = \frac{22}{22 + 92} = 0.1930$$

Referring back to the table in Figure 32.2, we see that cell 16 indexes the joint statement concerning Corona preference \overline{B} , left handedness \overline{H} , beige fur color \overline{F} , and below average intelligence \overline{I} . The probability for the next kangaroo to appear in this cell,

$$\frac{N_{16} + 1}{N + n} = \frac{22}{1616}$$

is the numerator of Bayes's Theorem. The denominator must also include the probability that the next kangaroo prefers Foster's, is left-handed, beige colored, and below average in intelligence, that is, the probability appearing in cell 12,

$$\frac{N_{12} + 1}{N + n} = \frac{92}{1616}$$

Exercise 32.9.13: In preparation for calculating any statement's probability as a numerical assignment under some model, how might some of the constraint functions be constructed?

Solution to Exercise 32.9.13

Construct the vector implementing the first constraint function $F_1(X = x_i)$. This constraint function is designed to capture the information about the marginal probability for beer preference. Since i runs from 1 through 16, there will be 16 elements in this vector.

Given the way the joint probability table was set up, we see that cells 1, 2, 3, and 4, together with cells 9, 10, 11, and 12, define the marginal probability for beer preference. Thus,

$$F_1(X = x_i) = (1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0)$$

The mathematical expectation of this constraint function is,

$$\langle F_1 \rangle = \sum_{i=1}^{16} F_1(X = x_i) Q_i = Q_1 + Q_2 + Q_3 + Q_4 + Q_9 + Q_{10} + Q_{11} + Q_{12}$$

If a model stipulates that $\langle F_1 \rangle = 1/2$, then this is the *information* inserted into a probability distribution under that particular model. The information under this model will eventually be part of the “noise” component.

Another constraint function that will be part of the noise is the *HF* interaction, the interaction capturing a relationship between hand preference and fur color. Construct the vector implementing *HF* as the fifth constraint function $F_5(X = x_i)$. This constraint function is designed to capture the information about the marginal probability for hand preference and fur color. Since i runs from 1 through 16, there will be 16 elements in this vector as well.

Given the way the joint probability table was set up, we see that cells 1, 5, 9, and 13 define the marginal probability for hand preference and fur color. Thus,

$$F_5(X = x_i) = (1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0)$$

The mathematical expectation of this constraint function is,

$$\langle F_5 \rangle = \sum_{i=1}^{16} F_5(X = x_i) Q_i = Q_1 + Q_5 + Q_9 + Q_{13}$$

If a model stipulates that $\langle F_5 \rangle = 0.30$, then this is the *information* inserted into a probability distribution under that particular model.

The noise component will be defined as a baseline model with $m = 8$ constraint functions, $F_1(X = x_i)$ through $F_8(X = x_i)$. The construction of the vectors for $F_1(X = x_i)$ and $F_5(X = x_i)$ was just demonstrated. The other six constraint functions constituting the noise component, that is, the constraint functions for *H*, *F*, *I*, *HI*, *FI*, and *HFI*, are produced in exactly the same fashion.

Exercise 32.9.14: How would a constraint function that is part of the signal component be constructed?

Solution to Exercise 32.9.14

We defined the signal component to be all those models containing information about relationships between beer preference and a physical trait. Thus, the very first signal model would include all the information from the baseline noise component with $m = 8$, plus one additional constraint function, for a total of $m = 9$ constraint functions. The constraint function $F_9(X = x_i)$ captures any association between beer preference and hand preference present in the double interaction BH .

Construct the vector implementing the BH relationship as the ninth constraint function $F_9(X = x_i)$. This constraint function is designed to capture the information about the marginal probability for beer preference and hand preference. Since i runs from 1 through 16, there will be 16 elements in this vector.

Given the way the joint probability table was set up, we see that cells 1, 3, 9, and 11 define the marginal probability for beer preference and hand preference. Thus,

$$F_9(X = x_i) = (1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0)$$

The mathematical expectation of this constraint function is,

$$\langle F_9 \rangle = \sum_{i=1}^{16} F_9(X = x_i) Q_i = Q_1 + Q_3 + Q_9 + Q_{11}$$

If a model stipulates that $\langle F_9 \rangle = 0.30$, together with all the other constraint function averages, then this is the *information* inserted into a probability distribution under that particular model that was designed to capture one kind of signal.

Exercise 32.9.15: How much is the noise model with $m = 8$ constraint functions preferred over the fair model?

Solution to Exercise 32.9.15

We have always called the “fair model” the model with no constraint functions, that is, where $m = 0$. The numerical assignments under the fair model are $Q_i = 1/n$, and in this case, $Q_i = 1/16$. We will temporarily establish this fair model as the baseline model \mathcal{M}_B . We want to compare the noise model with $m = 8$ constraint functions, (this will be model \mathcal{M}_A), to this baseline model. The constraint function averages in the noise model match the actual data averages.

The relative entropy formula for computing the ratio of the probability for model \mathcal{M}_A over model \mathcal{M}_B when conditioned on the known data is,

$$\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} = e^{[N \times KL(p, q)]}$$

Point p matches up with model \mathcal{M}_A and point q matches up with model \mathcal{M}_B . What actually gets computed is the log likelihood ratio,

$$\ln \left[\frac{P(\mathcal{D} | \mathcal{M}_A)}{P(\mathcal{D} | \mathcal{M}_B)} \right] = N \times \left[\sum_{j=1}^m (\lambda_j^A - \lambda_j^B) \bar{F}_j + \ln \left(\frac{Z_B}{Z_A} \right) \right]$$

Since the current baseline model \mathcal{M}_B has all $\lambda_j^B = 0$, $m = 8$, $Z_B = 16$, and $N = 1600$, the above expression reduces to,

$$\ln \left[\frac{P(\mathcal{D} | \mathcal{M}_A)}{P(\mathcal{D} | \mathcal{M}_B)} \right] = 1600 \times \left[\sum_{j=1}^8 \lambda_j^A \bar{F}_j + \ln \left(\frac{16}{Z_A} \right) \right]$$

Examine Table 32.4 below for the details of the calculation. From this table, we find that the argument to the exponential is,

$$1600 \times [0.7312 + \ln (16/28.5714)] = 242.228$$

leading to the ratio favoring the noise model \mathcal{M}_A over the baseline fair model \mathcal{M}_B of approximately,

$$\frac{P(\mathcal{M}_A | \mathcal{D})}{P(\mathcal{M}_B | \mathcal{D})} = e^{242.228} \approx 1.58 \times 10^{105}$$

Table 32.4: The details of the computation to find the log likelihood of the noise model versus the fair model.

j	Effect	\bar{F}_j	λ_j^A	$\lambda_j^A \times \bar{F}_j$
1	B	800/1600	0	0
2	H	1200/1600	1.189580	0.892185
3	F	640/1600	-0.336472	-0.134589
4	I	800/1600	0.133531	0.066766
5	HF	480/1600	-0.090972	-0.027292
6	HI	592/1600	-0.177983	-0.065854
7	FI	320/1600	-0.133531	-0.026706
8	HFI	240/1600	0.177983	0.026697
$\sum_{j=1}^8 \lambda_j^A \bar{F}_j$				0.7312
Z_A				28.5714

Of course, this enormous number reflects the fact that this so-called “noise” model captures a lot of relevant information supported by the data. This relevant information concerns such things as the four main effects, three double interactions, and a triple interaction. But the critical lesson is that, under this model, the probability for beer preference will not change when conditioned on the physical traits.

Exercise 32.9.16: Prepare a table showing the probability ratio for each succeeding signal plus noise model over its predecessor.

Solution to Exercise 32.9.16

Table 32.5 below contains the relevant calculations. Model \mathcal{M}_B in the final column refers to the preceding model, not to the original baseline model.

Table 32.5: Successive signal models showing the improvement over the previous baseline model. The improvement stops at $m = 12$.

m	Effect	$N \times \left[\sum \lambda_j^A \bar{F}_j + \ln \left(\frac{16}{Z_A} \right) \right]$	Δ	$\frac{P(\mathcal{M}_A \mathcal{D})}{P(\mathcal{M}_B \mathcal{D})}$
8	Noise	242.228	242.228	$\approx 10^{105}$
9	BH	343.488	101.260	$\approx 10^{43}$
10	BF	382.015	38.527	$\approx 10^{16}$
11	BI	417.814	35.799	$\approx 10^{15}$
12	BHF	417.814	0	≈ 1
13	BHI	417.814	0	≈ 1
14	BFI	417.814	0	≈ 1
15	BHFI	417.814	0	≈ 1

Exercise 32.9.17: Show the joint probability table with the numerical assignments under the noise model compared to the assignments under the best signal plus noise model.

Solution to Exercise 32.9.17

Figure 32.3 at the top of the next page presents a joint probability table with two sets of numerical assignments. The top assignment in each cell is from the noise only model with the information from $m = 8$ constraint functions. The bottom assignment in each cell is from the signal plus noise model with the information from $m = 11$ constraint functions.

What is the probability under the best signal plus noise model with $m = 11$ constraint functions for a kangaroo to prefer Corona given that it is left-handed, sandy colored, and of below average intelligence?

$$P(\bar{B} | \bar{H}, \bar{F}, \bar{I}, \mathcal{M}_{S+N}) = \frac{0.0133}{0.0133 + 0.0567} = 0.1900$$

Compare this answer with the answer calculated previously in Exercise 32.9.12 that is the result of averaging over all possible models,

$$P(\bar{B} | \bar{H}, \bar{F}, \bar{I}, \mathcal{D}) = \frac{22}{22 + 92} = 0.1930$$

		B			
		H	\bar{H}	H	\bar{H}
F		1 .0750 .0534	2 .0250 .0398	5 .0750 .0966	6 .0250 .0102
\bar{F}		3 .1100 .1335	4 .0400 .0733	7 .1100 .0865	8 .0400 .0067
				I	
				\bar{I}	
		H		\bar{H}	
F		9 .0750 .0266	10 .0250 .0302	13 .0750 .1234	14 .0250 .0198
\bar{F}		11 .1150 .0865	12 .0350 .0567	15 .1150 .1435	16 .0350 .0133

Noise

$$\begin{aligned} P(B) &= 0.50 & P(HF) &= 0.30 \\ P(H) &= 0.75 & P(HI) &= 0.37 \\ P(F) &= 0.40 & P(FI) &= 0.20 \\ P(I) &= 0.50 & P(HFI) &= 0.15 \end{aligned}$$

Signal

$$\begin{aligned} P(BH) &= 0.30 \\ P(BF) &= 0.15 \\ P(BI) &= 0.30 \end{aligned}$$

1.00

Figure 32.3: A joint probability table showing the numerical assignments under the noise model (top) and the signal plus noise model (bottom).

Somewhat surprisingly, *one* model is doing an excellent job of producing almost the same probability that resulted from taking all models into account. The reason for this remarkable performance, as you might have begun to suspect by now, is that, in fact, the data were generated by this particular signal plus noise model.

But explanations like this, while true, are also the source of much confusion. Earlier, I railed against the deceptive language asserting that “observations were drawn from some probability distribution.”

Similarly, we cannot justify any explanation like the above that represents an 180 degree turn in the logic of the situation. For all non-simulated data, in other words, when we are talking about the real world, no signal plus noise model can cause that data! Rather, such models capture the best information that an information processor has about the physical causation!

With this caveat ringing in our ears, the important fact remains that the information in our best correlational model points to some physical causation between the behavioral trait of beer preference and the three physical traits of hand preference, fur color, and intelligence. These kinds of inferences are as valid for the real world as they are for our whimsical caricatures.

Exercise 32.9.18: Use logistic regression to show why some interactions would not affect the probability for beer preference.

Solution to Exercise 32.9.18

Chapter Twenty Three was devoted to demonstrating that what is called logistic regression in the orthodox literature is simply a standard solution when viewed from the combined Bayesian and MEP approach.

We already know that the probability for any next kangaroo to prefer Foster's is simply 1/2, even when conditioned on knowledge of the three physical traits of that kangaroo, under the information in a model \mathcal{M}_* that takes account of only the main effect marginal sums.

Such a model has $m = 4$ constraint functions together with their averages. Applying Bayes's Theorem under this model shows that it is the assigned numerical values to the probabilities of the joint statements in cells 1 and 5 of the joint probability table that determine the conditional probability for beer preference when given the three causal factors of hand preference, fur color, and intelligence,

$$P(B | H, F, I, \mathcal{M}_*) = \frac{Q_1}{Q_1 + Q_5} = 1/2$$

Under a different model \mathcal{M}_{**} that now includes information about the double interaction between hand preference and fur color, which makes $m = 5$, the probability that the next kangaroo prefers Foster's does not change from 1/2. It is irrelevant that this new model might be preferred over the old model by support from the data. The fact that a relationship does exist between fur color and hand preference has no bearing on beer preference.

Consider the logistic regression approach that takes the Bayes's Theorem result and divides though by Q_1 ,

$$\begin{aligned} P(B | H, F, I, \mathcal{M}_*) &= \frac{Q_1}{Q_1 + Q_5} \\ \frac{Q_1}{Q_1 + Q_5} &= \frac{1}{1 + \exp(-Y)} \\ Y &= \lambda_1 \\ \lambda_1 &= 0 \\ P(B | H, F, I, \mathcal{M}_*) &= 1/2 \end{aligned}$$

Here are the details. The Y value is found from the MEP formula for Q_5 and Q_1 , where starting with Q_5 and $m = 4$,

$$Q_5 = \frac{e^{\lambda_1 F_1(X=x_5) + \lambda_2 F_2(X=x_5) + \lambda_3 F_3(X=x_5) + \lambda_4 F_4(X=x_5)}}{Z(\lambda_1, \dots, \lambda_4)}$$

In this case, $F_1(x_5) = 0$, $F_2(x_5) = F_3(x_5) = F_4(x_5) = 1$. Thus, we have,

$$Q_5 = \frac{e^{\lambda_2 + \lambda_3 + \lambda_4}}{Z(\lambda_1, \dots, \lambda_4)}$$

Likewise, for Q_1 ,

$$Q_1 = \frac{e^{\lambda_1 F_1(X=x_1) + \lambda_2 F_2(X=x_1) + \lambda_3 F_3(X=x_1) + \lambda_4 F_4(X=x_1)}}{Z(\lambda_1, \dots, \lambda_4)}$$

where now $F_1(x_1) = F_2(x_1) = F_3(x_1) = F_4(x_1) = 1$. Therefore,

$$Q_1 = \frac{e^{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4}}{Z(\lambda_1, \dots, \lambda_4)}$$

with the result that the ratio is,

$$\frac{Q_5}{Q_1} = e^{-\lambda_1}$$

Exactly the same kind of analysis reveals that even with the additional constraint function under model \mathcal{M}_{**} representing the double interaction HF , that extra Lagrange multiplier cancels out. Notice that since now $m = 5$, there are five constraint function values at $F_j(X = x_1)$ and $F_j(X = x_5)$ to take account of. These are,

$$F_1(x_5) = 0, F_2(x_5) = F_3(x_5) = F_4(x_5) = F_5(x_5) = 1$$

and,

$$F_1(x_1) = F_2(x_1) = F_3(x_1) = F_4(x_1) = F_5(x_1) = 1$$

with the outcome that the ratio is still,

$$\frac{Q_5}{Q_1} = e^{-\lambda_1}$$

This same pattern repeats itself all the way through models with $m = 8$, with the result that the probability that the next kangaroo prefers Foster's never changes from $1/2$, although other probabilities may be changing due to the different numerical assignments under the different models as m is changing.

However, when we reach models with $m = 9$ by including information about an association between beer preference and hand preference, the logistic regression approach shows that,

$$P(B | H, F, I, \mathcal{M}_{***}) = \frac{Q_1}{Q_1 + Q_5} \neq 1/2$$

This model \mathcal{M}_{***} contains all of the information in the previous models, as well as new information about the association in the beer preference, hand preference interaction BH . The pattern of 1s and 0s for the constraint functions $j = 1$ through $j = 9$ is,

$$F_j(X = x_5) = (0, 1, 1, 1, 1, 1, 1, 1, 0)$$

$$F_j(X = x_1) = (1, 1, 1, 1, 1, 1, 1, 1, 1)$$

The ratio does change under the inclusion of information about a correlation between beer preference and some other causal factor.

$$\begin{aligned} \frac{Q_5}{Q_1} &= e^{-\lambda_1 - \lambda_9} \\ \lambda_1 &= +1.38629 \\ \lambda_9 &= -1.79176 \\ P(B | H, F, I, \mathcal{M}_{***}) &= \frac{1}{1 + \exp(Y)} \\ &= \frac{1}{1 + 1.5} \\ &= 0.40 \end{aligned}$$

The probability that a kangaroo prefers Foster's has decreased from the value $1/2$ under this model. The Lagrange multiplier parameter λ_9 did not cancel out as had the parameters for the other interactions.

Exercise 32.9.19: Write some short, strictly utilitarian, *Mathematica* code to check the findings from the last exercise.

Solution to Exercise 32.9.19

Construct a function called **lrcheck**[\dots] with three arguments. The first argument will be the number of constraint functions in a model, and the next two arguments will be the two Q_i values appearing in Bayes's Theorem.

For example, a call to **lrcheck**[8, 5, 1] should evaluate to $e^{-\lambda_1}$. The first argument is 8 indicating that the (noise) model with the information from $m = 8$ constraint functions will be checked. The next two arguments of 5 and 1 indicate that Q_5 will be divided by Q_1 .

All parameter arguments from 1 through 8 will return $e^{-\lambda_1}$. However, when **lrcheck**[9, 5, 1] is evaluated, the answer $e^{-\lambda_1 - \lambda_9}$ is returned, verifying our hand calculations carried out in the last exercise.

```
lrcheck[m_, q1_, q2_] :=
  Module[{numlist},
    numlist = Exp[Dot[Map[Subscript[λ, #] &, Range[m]], 
      Take[constraintmatrix, m, All]]];
    numlist[[q1]] / numlist[[q2]]
    (* end of Module *)]
```

To conduct numerical experiments, embed this code into a **Manipulate []**. This is the very practical way that *Mathematica* implements Wolfram's exhortation to "explore the computational universe."

Appendix A

The Initial MEP Formula and *Mathematica*

I have included Appendices for Volume II, just as I did for Volume I, to discuss how *Mathematica*, in its role as a (very) general programming language, can implement the computations required by information processing. As we develop the MEP algorithm here in Volume II, we will be relying more and more on the computational abilities of *Mathematica* to provide the numerical answers that are required.

We begin gently with the benign computational demands of Chapter Seventeen. These are indeed minimal, and could be handled quite easily with a hand calculator. But things won't remain that way forever.

Most concepts we discuss will eventually be cast into an operational definition via *Mathematica*. It is probably best to refresh our memories of the introductory lessons in Volume I with some material that is not too difficult. A direct translation of Equation (17.1) into *Mathematica* accomplishes that objective.

The MEP formula will provide us with a legitimate numerical assignment to the probability for a statement in the state space when conditioned on the information resident in some model. This is, of course, the familiar $Q_i \equiv P(X = x_i | \mathcal{M}_k)$.

As we saw in Chapter Seventeen, the most primitive computation in the MEP algorithm occurs in the numerator where we would like to compute,

$$\text{numerator for } Q_i = e^{\lambda \times F(X=x_i)}$$

for a model with just one constraint function.

Remember that all symbolic expressions in *Mathematica* adhere to this generic syntax: `head[arg1, arg2, ...]`. This syntactic template should certainly be applicable to *Mathematica*'s own built-in expression for computing the exponential function.

The *Mathematica* syntax for implementing e^z is **Exp[z]** where **Exp** is the **head** and **z** is the one argument required by this function. z is the power to which e is raised, and as a **FullForm** expression looks like **Power[E, z]**. In our case, the exponent is $z = \lambda \times F(X = x_i)$. Do not confuse *Mathematica*'s notation z for the argument to **Exp[]** with the partition function notation $Z(\lambda_1, \dots, \lambda_m)$.

Mathematica treats everything very generally, so it assumes that z is a complex number. But we won't require that level of generality, and z will always be a real number in the MEP formula.

One of the neat things about *Mathematica* is that it will keep evaluating everything in a recursive manner until nothing more remains to be done. This means that for our situation of computing the numerator of Q_i in the MEP formula,

$$e^{\lambda \times F(X=x_i)}$$

the multiplication of the constraint function $F(X = x_i)$ by the Lagrange multiplier λ will be evaluated first. This answer will then serve as the one argument z to **Exp[z]**.

In the example of the coin toss, the constraint function assigned the value of 1 to HEADS, $F(X = x_1) = 1$, and 2 to TAILS, $F(X = x_2) = 2$. So we wind up with the expression,

$$\text{Exp}[\text{Times}[\lambda, 2]]$$

in the numerator for the numerical assignment of a probability to TAILS.

But it is just as easy for *Mathematica* to keep all the values of the constraint function in a list. So now we write **Exp[Times[λ, cf]]**, or **Exp[λ cf]** as the expression in the numerator of Equation (17.1) where the shorter syntactical form for multiplication is used. The constraint function **cf** is a vector representing the constraint function assignments at all n statements in the state space. For the coin tossing scenario, $F(X = x_1) = 1$ and $F(X = x_2) = 2$, so let the vector $(1, 2)$ represent the constraint function assignments.

Mathematica represents a vector as a list, **List[arg₁, arg₂, ...]** where the arguments are, of course, the individual elements making up the vector. We could assign this vector to a symbol by **Set[cf, List[1, 2]]**, or, even easier, through the common syntactical shortcut, **cf = {1, 2}**.

For this specific example of the coin toss with the dimension of the state space at $n = 2$, we want *Mathematica* to compute the numerator of Q_i as,

$$\text{Exp}[\lambda \text{ cf}] \equiv \text{Exp}[\lambda \{1, 2\}]$$

If we were programming in an older style programming language, we would have to think about declaring variables of a particular type, with the right number of dimensions, keeping track of the dimension indices, and so on.

But *Mathematica* does what is called *threading over lists*. So not only will it multiply both constraint function assignments by λ , it will also take the exponential function of the result of these two values, and return quite properly a list with two elements, $\{\text{Exp}[\lambda], \text{Exp}[2\lambda]\}$.

We might like to construct a function so that we can specify whatever we like for the Lagrange multiplier and the constraint function.

```
MEPnumerator[ $\lambda$ _, cfList]:=Exp[ $\lambda$  cf]
```

When we call this function with the arguments of a Lagrange multiplier of 0 and the constraint function $(1, 2)$, **MEPnumerator**[0, {1, 2}] returns the list {1, 1}. You can see that *Mathematica* kept on evaluating expressions like $\{\text{Exp}[\text{Times}[0,1]], \text{Exp}[\text{Times}[0,2]]\}$ because it could.

All we need now in order to compute the Q_i values from the MEP formula is the value of the partition function. The *Mathematica* function **Total** [arg] sums all the elements in the list taken as the argument to **Total**. We have already seen that **Exp**[λ cf] returns a list, so **Total**[**Exp**[λ cf]] will return the partition function. For example, **Total**[**Exp**[λ , {1, 2}]] returns $e^\lambda + e^{2\lambda}$.

It is interesting to examine how *Mathematica* thinks about the partition function $e^\lambda + e^{2\lambda}$ if for any reason it had to be manipulated further as a symbolic expression,

```
Plus[Power[E, \[Lambda]], Power[E, Times[2, \[Lambda]]]]
```

The next illustration is the construction of a small function to calculate Q_i for the coin tossing scenario given any model \mathcal{M}_k .

```
Qi[ $\lambda$ _, cfList]:=Module[{numerator, denominator},
  numerator = Exp[ $\lambda$  cf];
  denominator = Total[Exp[ $\lambda$  cf]];
  N[numerator/denominator]
  (* end Module *)]
```

Suppose we want to calculate Q_1 and Q_2 using the MEP formula. The IP inserts information under some model by specifying that the Lagrange multiplier (the parameter) is equal to 0. This is, of course, what we have been calling the “fair model.” This insertion of information into a probability distribution over the statements in the state space could be accomplished just as well by specifying that the constraint function average $\langle F \rangle = 1.5$.

Evaluate the *Mathematica* function just written with these arguments for the Lagrange multiplier and the constraint function. **Qi**[0, {1, 2}] will return the answer {0.5, 0.5}.

Module[] allows us to pack up the definition of our function **Qi**[] where **numerator** and **denominator** are the *local variables* used in this function. They

are given in a list at the beginning of **Module**[]. The semicolons at the end of each line suppress any output. The only output we want to see is the numerator divided by the denominator. Since **numerator** is a list, we automatically get a list back when it is divided by **denominator**. The **N**[] function forces the division to return real number values; otherwise *Mathematica* might leave them in symbolic form. The final right] closes off **Module**[].

On the left hand side, where the function **Qi**[*args*] is defined with its two arguments, we see the ubiquitous use of *pattern matching* objects. By writing λ - for the first argument to **Qi**[], we can place any variable name starting with λ in the function definition. By writing **cf_List** we are restricting the pattern match for the second argument to be strictly a **List** match. If what we place as the second argument to **Qi**[] is not a list, then the function will not evaluate. Once again, any variable named **cf** in the right hand side must also be a list.

To repeat, a syntactical form like **cf_List** is an abbreviated version for a full form expression, **Pattern**[**cf**, **Blank**[**List**]]. This full form expression reveals the underlying fundamental syntactical structure of,

```
head[arg1, arg2, ...]
```

where the second argument to **Pattern**[], here **Blank**[**List**], can be recursively defined with its own,

```
head[arg1, arg2, ...]
```

structure.

To end this Appendix, we show how seven of the ten Q_i assignments in Table 17.1 were computed. *Mathematica* users make frequent use of an iteration function called **Table**[*expression*, *list*]. The expression to be iterated over is the first argument. The second argument is a list containing the variable that is being iterated, together with the beginning and ending value of this iteration variable.

Thus, if we write,

```
Table[Qi[λ,{1, 2}], {λ, -3, 3}]
```

the first argument to **Table**[] is the function **Qi**[*args*]. This will calculate two Q_i assignments for any given Lagrange multiplier (first argument to **Qi**[]), and constraint function (second argument to **Qi**[]). The second argument supplied to **Table**[*args*] involves a list containing the iteration variables. The iteration variable is λ , which is seen to be the Lagrange multiplier argument to **Qi**[].

The starting value for the Lagrange multiplier is -3 which is incremented by 1 until the final value of $+3$ is reached. The constraint function does not change, and remains the same at $(1, 2)$. This expression will compute seven of the ten Q_i

assignments in Table 17.1. The assignment for $\lambda = -1.09861$ can be computed by a separate call to **Qi**[*args*].

In Chapter Seventeen, we relied upon a heuristic search to find the Lagrange multiplier as a parameter. Such a parameter was required under a model desiring to insert the information that the constraint function average was 1.25. This resulted in a numerical assignment of 0.75 for the probability of HEADS.

Mathematica has a rather powerful numerical routine called **Solve**[] which we can use as a black box to find out officially that $\lambda = -1.09861$ under this model. **Solve**[*eqn, var*] looks like this for our current needs,

```
Solve[Exp[λ] / (Exp[λ] + Exp[2 λ]) == .75, λ]
```

The first argument to **Solve**[] is the equation we are interested in. In other words, it is the implementation of the MEP formula for the coin tossing scenario. The double equals symbol **==** links the left and right hand sides of the equation.

What is the variable *var* in the equation being solved for? That is, what value of the model parameter λ results in the equation producing 0.75? This is the second argument to **Solve**[*eqn, var*].

Mathematica evaluates this expression and returns the answer that λ should be assigned as -1.09861 . From Table 17.1, we knew that $\lambda = -1$ produced a probability for HEADS of 0.7311, and $\lambda = -2$ produced a probability for HEADS of 0.8808. Eventually, we could have homed in on the answer that *Mathematica* provided to us immediately by employing the rather crude and time consuming trial and error approach as used in Table 17.1.

Appendix B

Manipulating the Lagrange Multiplier Parameters

Surely, the most widely used function in *Mathematica* must be **Manipulate**[*args*]. This built-in function provides the user with a way to interact dynamically with a program. The interesting outcomes depend upon variables that the user would like to change, or, in other words, would like to *manipulate*.

For our current needs, we would like to employ **Manipulate**[*args*] to see how the numerical assignments to the joint statements in the state space change as the parameters of models change. For example, in our second illustration of the MEP formula we looked at how the six numerical assignments to the die faces changed as the Lagrange multipliers were varied.

Thus, by using **Manipulate**[*args*], we would like to immediately observe the effects of different models on probability assignments. I wrote a small *Mathematica* program to find the results that were presented in Chapter Nineteen.

First of all, we need to take a global overview for the initial breakdown of the extensive complement of arguments to **Manipulate**[*args*].

Manipulate[*Function that depends on the changing variables,*
Controls that allow the user to manipulate variables,
Initialization of supporting code]

Partially fill in this skeletal outline by showing the function that will implement the MEP formula. Here, a slightly different **Qi**[*args*] function than introduced in Appendix A will specify three Lagrange multipliers for the argument,

Manipulate[**Qi**[{ $\lambda_1, \lambda_2, \lambda_3$ }],
Controls that allow the user to manipulate variables,
Initialization of supporting code]

These arguments to **Qi**[*args*] are the variables that we would like to manipulate by interacting with the controls provided by *Mathematica*. The function,

Qi[*Lagrange multipliers*]

is the heart of the program. It computes the probability assignments as they depend on the list of the three Lagrange multipliers, or equivalently, the three parameters.

Now insert that part of the code to implement the controls allowing the user to manipulate the Lagrange multipliers,

```
Manipulate [Qi[{ $\lambda_1, \lambda_2, \lambda_3$ }],  
 {{ $\lambda_1, 0, " \lambda_1 "$ }, -1, 1, 0.00001},  
 {{ $\lambda_2, 0, " \lambda_2 "$ }, -1, 1, 0.00001},  
 {{ $\lambda_3, 0, " \lambda_3 "$ }, -1, 1, 0.00001},  
 Initialization of supporting code ]
```

With this code, *Mathematica* provides the user with three “slider” controls as seen in Figure B.1 appearing at the end of this Appendix. Manipulating these controls allows any one of the three parameters to be changed. The default value of each parameter is initialized at 0, and the user can change the parameters in increments of 0.00001 between the set limits of -1 to $+1$. The default value of $\lambda_1 = \lambda_2 = \lambda_3 = 0$ is seen to provide the information for the fair model of the die where each $Q_i = 1/6$.

Let’s examine in some detail the most important piece of code. This is the function **Qi**[{ $\lambda_1, \lambda_2, \lambda_3$ }] implementing the MEP formula. It will therefore produce the numerical assignments for the probability of each of the six spots for whatever model is specified by the three arguments. These are the three Lagrange multipliers λ_1, λ_2 , and λ_3 associated with each of the three constraint functions.

In Chapter Nineteen, the three constraint functions were constructed so as to mimic a physical defect in the die. These physical defects took the die further and further away from a perfectly balanced symmetrical cube. In other words, the invocation of each constraint took the numerical assignments to the probabilities for the six faces further and further away from the fair assignment of $Q_i = 1/6$.

The constraint function vectors are represented by the three lists,

$$\begin{aligned} F_1(X = x_i) &= \{1, 2, 3, 4, 5, 6\}, \\ F_2(X = x_i) &= \{1, 1, -2, -2, 1, 1\}, \\ F_3(X = x_i) &= \{-1, -1, -1, -1, -1, 5\} \end{aligned}$$

Mathematica constructs matrices by making a list of vectors which ends up as a list of lists. We label as a constraint matrix all three constraint function vectors packaged together as a matrix, and expressed as,

$$\mathbf{cm} = \{\{1, 2, 3, 4, 5, 6\}, \{1, 1, -2, -2, 1, 1\}, \{-1, -1, -1, -1, -1, 5\}\}$$

We need to form the numerator, $\exp [\sum_{j=1}^m \lambda_j F_j(X = x_i)]$, in the formula for the Q_i . The sum is formed by a *Mathematica* built-in function with two arguments, **Dot**[*v*,*m*], where we just want to suggest that the arguments can be vectors, matrices, or tensors for that matter. The arguments for our particular situation are the vector containing the parameters, and the matrix containing all of the constraint functions,

```
Dot[lambda, cm]
```

This function performs the important matrix by vector multiplication required for the numerator. For example, what we get back from an application of **Dot** for the first part of the numerator of Q_1 is,

$$[\lambda_1 \times F_1(X = x_1)] + [\lambda_2 \times F_2(X = x_1)] + [\lambda_3 \times F_3(X = x_1)] = \lambda_1 + \lambda_2 - \lambda_3$$

Before exponentiation, this is exactly what we want.

As another example, the numerator before exponentiation for the assignment to the SIX will be $6\lambda_1 + \lambda_2 + 5\lambda_3$. If the IP specifies a model by setting the arguments to **Qi**[] through the slider controls as {0,0,0.499}, then the numerator for Q_6 under this model will have the value of $e^{5 \times 0.499} = 12.1217$.

The next steps are easy. First, apply **Exp**[] to carry out the aforementioned exponentiation in the numerator, followed by **N**[] to force an actual numerical result instead of a possibly symbolic one. At this point, we have the numerators for all of the Q_i ,

```
numerator = N[Exp[Dot[lambda, cm]]]
```

Recall the comment from the Appendices to Volume I that we always end up writing highly nested code in *Mathematica*. But the pay back is that *Mathematica* automatically returns for this very short piece of code the vector of six elements containing all six numerators. This is the result of the vector-matrix multiplication of the vector of Lagrange multipliers and the constraint function matrix.

Thus, all we need do is find the sum of these numerators for the partition function, and then divide each element in the numerator list by the partition function.

```
z = Total[numerator]
qi = numerator / z
```

It's a good idea, I think, to interrupt the explanation of the code here to discuss an always tricky issue. One must always pay a great deal of attention to the standard symbolic notation used for vector and matrix multiplications. How does the standard subscripting notation associated with these operations eventually get implemented in *Mathematica*?

Clarification is not only helpful for understanding the material in this Volume with regard to the MEP formula, but will also come into play later on in Volume III when Information Geometry is of central interest.

For the situation considered here in forming the numerator of Q_i via the MEP algorithm, we always think back to the standard rules we first learned in performing matrix-matrix multiplications. Namely, do a column by row multiplication followed by addition.

Specifically, if we want to form the matrix multiplication $\mathbf{C} = \mathbf{AB}$, we multiply every element in the j^{th} column of \mathbf{B} by every element in the i^{th} row of \mathbf{A} in order to form the single ij^{th} element in \mathbf{C} . There are a total of p elements in each row and column. The formula with all the subscripts is,

$$c_{ij} = \sum_{k=1}^p a_{ik} b_{kj}$$

In order for this to work, the matrices must be *conformal*. That is, there must be p columns in \mathbf{A} to match up with the p rows in \mathbf{B} . If there are a total of m rows in \mathbf{A} , and a total of n columns in \mathbf{B} , then matrix \mathbf{A} has dimensions of $m \times p$, with matrix \mathbf{B} having dimensions of $p \times n$. The resulting matrix product $\mathbf{C} = \mathbf{AB}$ has dimensions of $m \times n$. The first subscript in a generic matrix element like c_{ij} always refers to the row number, while the second subscript refers to the column number.

For example, suppose that \mathbf{A} is a (4×2) matrix, while \mathbf{B} is a (2×3) matrix. Therefore, since $m = 4$, $p = 2$, and $n = 3$, \mathbf{C} will have dimensions $(m \times n)$ and will be a (4×3) matrix. Any particular element of the \mathbf{C} matrix, say c_{32} , the element in the third row and second column of \mathbf{C} ($i = 3$, $j = 2$) is calculated by the above formula as,

$$c_{32} = \sum_{k=1}^2 a_{3k} b_{k2} = a_{31}b_{12} + a_{32}b_{22}$$

I mention all of this detail because it was something that one had to pay close attention to when writing any pre-*Mathematica* code. The dimensions, the double subscripting, and the summation over proper indices were details in the code that one could not ignore. However, in *Mathematica* one is relieved of all this tedious burden. Nonetheless, it is imperative to double-check that the work *Mathematica* has quietly done for you behind the scenes in these matrix multiplications is what you intended.

Matrices with a designated number of *rows* and *columns* can be created by,

```
Array[Function[{x,y}, Subscript[a,x,y]], {rows, columns}]
```

A shorter form for creating, say, the above (4×2) matrix \mathbf{A} is,

```
A = Array[a##&, {4,2}]
```

The matrix multiplication $\mathbf{C} = \mathbf{AB}$ is conveniently carried out by **Dot[A, B]**. There is no worrying about any of the indexing details. Since vectors and matrices are represented symbolically by lists, and lists of lists, *Mathematica* provides the command **MatrixForm[]** so that the output of the matrix multiplication can be viewed in the standard manner instead of the more cumbersome list notation.

Check the hand calculation above by evaluating,

```
MatrixForm[Dot[A,B]]
```

by looking at any element c_{ij} . Sure enough, the entry in the third row and second column is $a_{31}b_{12} + a_{32}b_{22}$.

In the die scenario, if you were thinking in terms of a constraint function matrix multiplying a vector with Lagrange multipliers, then something is not quite right based on the explanation just provided. The constraint function matrix, analogous to **A**, has six columns and three rows, and therefore has dimensions of (3×6) . The Lagrange multiplier vector, analogous to **B**, has one column and three rows, and therefore has dimensions (3×1) . These two matrices are not conformal.

If, instead, **A** were of dimensions (6×3) , and **B** with dimensions (3×1) , then the matrices are conformal. The resulting product is a column vector of dimensions (6×1) . In *Mathematica* we could have accomplished this by,

```
Dot[Transpose[cm], lambda]
```

where the *Mathematica* built-in function **Transpose[]** would interchange the rows and columns of its argument.

By reversing the order of the arguments as we did above in **Dot[lambda, cm]** we arrive at a correct matrix by vector multiplication. In this case **lambda** is a row vector with dimensions (1×3) and **cm** is a matrix with dimensions (3×6) . Thus, the multiplication yields a row vector of dimension (1×6) . The last element in this vector is $6\lambda_1 + \lambda_2 + 5\lambda_3$.

Putting these pieces together, the bulk of the function definition for **Qi[]** is as follows:

```
Qi[lambdaList] := Module[{cm = {{1, 2, 3, 4, 5, 6},
                           {1, 1, -2, -2, 1, 1},
                           {-1, -1, -1, -1, -1, 5}},
                           numerator, z, qi},
                           numerator = N[Exp[Dot[lambda, cm]]];
                           z = Total[numerator];
                           qi = numerator/z;
                           (* further code for presentation of results *)]
```

This function definition for **Qi[]**, as well as other supporting code, takes place within the **Initialization** option for **Manipulate[]**.

Figure B.1 presents the interactive scheme under **Manipulate[]**. It provides the user with three sliders in order to change each of the three Lagrange multipliers. Each change results in a different model. This is a rather crude representation of the actual working environment within *Mathematica* as the graphics are in color and have higher resolution.

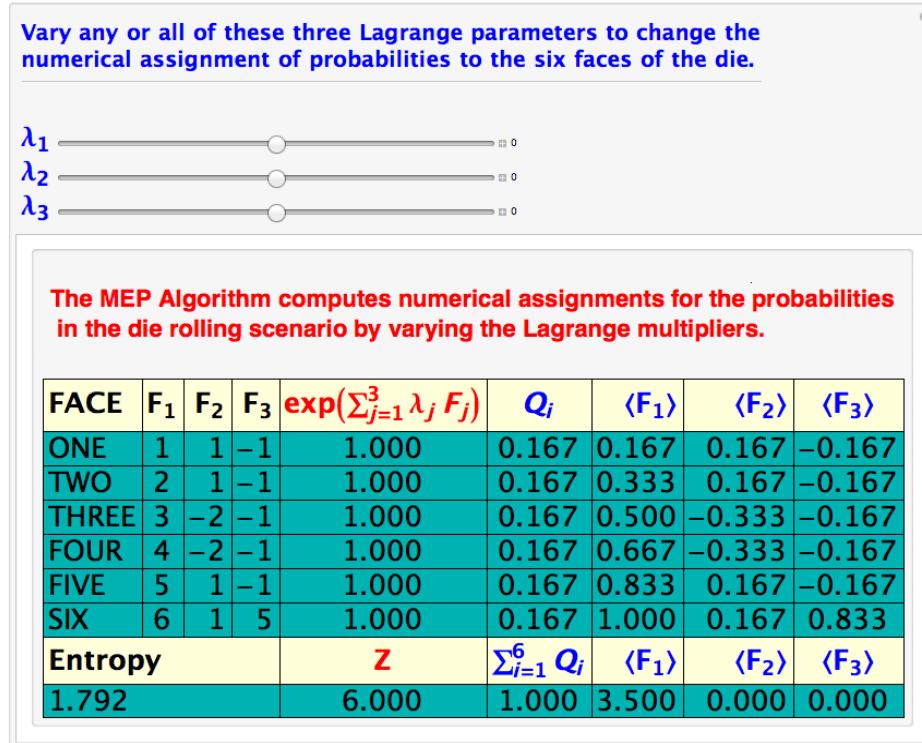


Figure B.1: Mathematica output using the **Manipulate[]** command to implement the MEP formula for the die scenario.

Appendix C

Mathematica and the Legendre Transformation

In this Appendix, we deconstruct and annotate code implementing our version of an MEP algorithm. As presented in Chapter Twenty Four, this algorithm relies heavily upon our description of the Legendre transformation as it relates to the numerical assignments made by the MEP.

The heart of the program is a single user-defined function called,

legendreMethod[*args*]

The arguments to this function are not important right now, so we will get to them later. As a point of programming etiquette, *Mathematica* asks the user to start any user-defined function with the lower case. Capital letters are reserved to begin function names that are *Mathematica* built-in functions.

To begin with the deconstruction, we will examine a single line of code in the middle of our defined function **legendreMethod[*args*]**. A lot of time is spent annotating this code because it might help in understanding the Legendre transformation as it relates to MEP assignments.

The derivation behind the rationale for using the Legendre transformation to find the numerical assignment possessing maximum entropy was already presented in Chapter Twenty Four. A numerical example illustrating *Mathematica* performing the numerical computations on a simple problem was presented in Exercise 24.5.2.

As discussed in Chapter Twenty Four, we define a function,

$$\ln Z(\lambda_1, \lambda_2, \dots, \lambda_m) - \sum_{j=1}^m \lambda_j \langle F_j \rangle$$

that we want to minimize. The function has two sets of parameters, the dual

parameters λ_j and $\langle F_j \rangle$. The Lagrange multipliers, the λ_j , are allowed to vary as we seek the minimum, while the constraint function averages, the $\langle F_j \rangle$, remain fixed at the values specified by the model \mathcal{M}_k .

This prescription gets translated into *Mathematica* as,

```
solution = NMinimize[Log[z] - Dot[lambda, cfa], lambda]
```

NMinimize[f, x] is the *Mathematica* built-in function that numerically finds the minimum value of some function f with its arguments x . Here that function as the first argument to **NMinimize**[] is,

```
Log[z] - Dot[lambda, cfa]
```

This is to be minimized by varying the parameters in **lambda**. The constraint function average parameters $\langle F_j \rangle$ are contained in the list **cfa**.

solution returns a list containing two elements. The first element in this list is the actual minimum value obtained for the function. This will be the information entropy of the MEP assignment under model \mathcal{M}_k . The second element is another list containing the values for the Lagrange multipliers that brought about the minimum value reported.

As an example of what the evaluation looks like, here is the output from Exercise 24.5.3.

```
{1.47665, {λ1 → 0.359707, λ2 → 0.541805}}
```

The first element in the **solution** list, 1.47665, is the information entropy of the MEP assignment. The second element in the **solution** list is another list containing the values of the two Lagrange multipliers that are the parameters for this model.

The inner list containing the *Mathematica* expressions for λ_1 and λ_2 ,

```
{λ1 → 0.359707, λ2 → 0.541805}
```

is a widely used *Mathematica* syntactic short-cut for rules. An expression like $\lambda_1 \rightarrow 0.359707$ appearing as the first element in this inner list is actually an instantiation of **Rule**[**lhs**, **rhs**]. The left hand side gets replaced by the right hand side under this rule. Thus, λ_1 actually is set to 0.359707 and λ_2 to 0.541805.

Anything that was formerly just a symbolic expression involving the λ_j may now receive these actual values when we use the *Mathematica* function,

```
ReplaceAll[expr, rule]
```

For example, the partition function $Z(\lambda_1, \lambda_2, \dots, \lambda_m)$ had to be evaluated earlier within the code for **legendreMethod[args]** by the expression,

```
z = Total[Exp[Dot[lambda, cm]]]
```

The answer was returned in symbolic form as,

$$Z(\lambda_1, \lambda_2) = e^{\lambda_1 - \lambda_2} + e^{2\lambda_1 - \lambda_2} + e^{3\lambda_1 + 2\lambda_2} + e^{4\lambda_1 + 2\lambda_2} + e^{5\lambda_1 - \lambda_2} + e^{6\lambda_1 - \lambda_2} \quad (\text{C.1})$$

because at that point there were no assigned values to **lambda**.

The evaluation of **z** was strictly in terms of an expression involving the symbols **lambda**. After the **solution** has been found, **ReplaceAll[args]** can be used to evaluate **z** numerically because the actual numerical values for **lambda** are now available.

The code,

```
znew = First[ReplaceAll[z, Rest[solution]]]
```

is added. We see that, when filling in the template for **ReplaceAll[expr, rule]**, **z** is the expression that will be replaced with the two rules,

$$\{\lambda_1 \rightarrow 0.359707, \lambda_2 \rightarrow 0.541805\}$$

These two rules are, in fact, at the position **Rest[solution]** in the **solution** list.

The value for **znew** is 31.7307 given **z** in Equation (C.1). This is the numerical value of the partition function $Z(\lambda_1, \lambda_2)$ and, of course, will act as the denominator when the MEP assignment is calculated. The **First[args]** is required because **znew** is returned as a list and we need to pick out the only element in this list.

The remainder of the code in **legendreMethod[args]** follows easily after all of this development. We also need the actual values for λ_1 and λ_2 to calculate the numerator in the MEP assignment.

```
lambdaNew = First[ReplaceAll[lambda, Rest[solution]]]
```

The numerator is formed from,

```
numerator = N[Exp[Dot[lambdaNew, cm]]]
```

Finally, the MEP assignment is just,

```
qi = numerator/znew
```

OK. At this point, we are finished with the deconstruction of our work-horse function `legendreMethod[args]`. Let's stitch it all back together again.

```
legendreMethod[m_Integer, cfa_List] :=
Module[ {cm, lambda, z, solution, entropy, znew, lambdanew,
         numerator, qi},
       cm = Take[constraintmatrix, m, All];
       lambda = Map[Function[x, Subscript[λ, x]], Range[m]];
       z = Total[Exp[Dot[lambda, cm]]];
       solution = NMinimize[Log[z] - Dot[lambda, cfa], lambda];
       entropy = First[solution];
       znew = First[ReplaceAll[z, Rest[solution]]];
       lambdanew = First[ReplaceAll[lambda, Rest[solution]]];
       numerator = N[Exp[Dot[lambdanew, cm]]];
       qi = numerator/znew ]
```

`legendreMethod[args]` has two arguments `m_Integer` and `cfa_List`. The first argument `m_Integer` is where we specify how many parameters are in the model M_k . The second argument `cfa_List` is where we can specify the list of constraint function averages in the model.

The creation of the symbolic expressions for the subscripted parameters,

```
lambda = Map[Function[x, Subscript[λ, x]], Range[m]]
```

is interesting in its own right as an elementary application of two very important *Mathematica* tools, `Map[args]` and `Function[args]`. The outer function `Map[args]` takes two arguments, `Map[function, expression]`, where the *expression* is a list. This built-in *Mathematica* function *maps* the function to every element in the list.

The function as the first argument to `Map[function, list]` does not have to be explicitly defined as a separate function someplace else with its own special name. It can be what *Mathematica* calls a *pure function*. Here, `Function[args]` has two arguments, the first a single dummy parameter indicated by `x`, and a body `Subscript[λ, x]` which is the unnamed function containing the dummy parameter `x` somewhere in the body.

This is a very simple function that just attaches its dummy parameter `x` as the subscript to λ . Jump back up to `Map[function, list]` and its second argument. We could have inserted a list here like `{1, 2}` and the subscripting pure function would have taken each element in this list in turn as the parameter. The evaluation would return λ_1 and λ_2 in the list `lambda`. We achieve greater generality by using the *Mathematica* built-in function `Range[args]` with an integer as its one argument. `Range[m]` will then generate subscripted λ parameters all the way up to whatever is specified in the model by `legendreMethod[m_Integer, cfa_List]`.

Some people prefer to take advantage of syntactic short-cuts wherever possible in writing *Mathematica* code. I do not.

I am of the opinion that the code just *looks* neater with a minimal dosage of short-cuts. I have seen some *Mathematica* programs where the author was so infatuated with throwing in every single possible short-cut at his disposal that the resulting program might as well been on the tomb of some ancient Egyptian pharaoh.

Here is a little example of what I mean. In the above *Mathematica* code, we needed to form the symbolic expressions for the Lagrange multipliers. Take a look at the following five expressions which progressively use more syntactic short-cuts. All five expressions evaluate to the same desired goal of $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$.

1. `Map[Function[x, Subscript[\lambda, x]], List[1, 2, 3, 4]]`
2. `Map[Function[x, Subscript[\lambda, x]], {1, 2, 3, 4}]`
3. `Map[Function[x, Subscript[\lambda, x]], Range[4]]`
4. `Function[x, Subscript[\lambda, x]] /@ Range[4]`
5. `Subscript[\lambda, #] & /@ Range[4]`

Appendix D

The Nested Structure of *Mathematica* Programs

Mathematica programs, like all software, will eventually look quite complicated. My very own pedagogical requirements demand that any explanation of a complicated *Mathematica* program proceed in a top-down manner. That is the only way one can begin to disentangle the inner workings of such a program.

Since *Mathematica* programs by their very nature tend to be written as highly nested code, it is well-nigh impossible to decipher what some program is trying to accomplish without first peeling off the outer layers of the onion before proceeding in a downward progression through successive layers to the inner core.

Unfortunately, I am never, or hardly ever, rewarded with any such top-down deciphering of a programmer's code in *Mathematica*. I would like to see *any kind* of meta-*Mathematica* code acting as a top-down explanation to accompany the actual detailed code. This meta-code would be an ever expanding revelation of more and more internal structure within the program. This request, of course, is completely independent of the need for extensive commentary code scattered generously throughout the program.

So what this Appendix does is provide you with an example, drawing upon our interest in the MEP formalism, of what I would like to see become more prevalent in this regard. I admit that I go a little overboard in this example to make sure my point is not missed.

As already mentioned, most *Mathematica* programs designed for some sort of user interaction will use the **Manipulate[]** function as the outermost function. So the very top-most meta-code would look something like,

```
Manipulate[ (* what needs to be manipulated *) ]
```

Subsequent meta-code would augment and fill in more and more of the superstructure being built up.

Manipulate[] will want as a first argument some main function, and suppose we are going to use the Legendre transformation as this main function to make numerical assignments to probabilities. This is followed by a control that allows the user to manipulate some argument in the Legendre transformation function.

```
Manipulate[
(* a Legendre transformation function *),
(* a control to manipulate an argument for this function *),
(* end of Manipulate *) ]
```

I feel that it is important in such an overview that the brackets and parentheses that close off expressions be explicitly noted. As mentioned, *Mathematica* programs are highly nested. Without such explicit aids, one is completely lost as to the beginning and ending of relevant expressions.

Another critical argument to **Manipulate[]** is the location for all the supporting code needed to implement the program. For example, where is the Legendre transformation code defined? *Mathematica* can place this within the initialization option of **Manipulate[]**. Our top-down meta-*Mathematica* code now looks like,

```
Manipulate[
(* a Legendre transformation function *),
(* a control to manipulate an argument for this function *),
Initialization:→ ( (* all supporting code defined here *)
(* end of Initialization option *) )
(* end of Manipulate *) ]
```

Drilling down into **Initialization**, we find the definition of a Legendre transformation function with its arguments,

```
Manipulate[
(* a Legendre transformation function *),
(* a control to manipulate an argument for this function *),
Initialization:→ ( (* some stuff here *),
legendreMethod[m_Integer, cfa_List] := Module[
(* compute numerical values assigned to probabilities *)
(* end of Module *) ]
(* end of Initialization option *) )
(* end of Manipulate *) ]
```

There might very well be a **Grid[]** function contained somewhere within the main working function **legendreMethod[]** in order to present some results in a table. We would like to indicate to the user where this fits in with the ever-growing structure of the program. So we show,

```
Manipulate[
(* a Legendre transformation function *),
(* a control to manipulate an argument for this function *),
Initialization:→ ( (* some stuff here *),
legendreMethod[m_Integer,cfa_List] := Module[
(* compute numerical values assigned to probabilities *)
Grid [(* results *)]
(* more stuff *)
(* end of Module *) ]
(* end of Initialization option *) )
(* end of Manipulate *) ]
```

I think you get my gist. To close out this Appendix, I fill in the top part of the structure. First, I fill in the expression that we want to interact with, followed by a control (a setter bar) by which we can change the number of constraint functions that the Legendre transform will use in making its numerical assignments to the probabilities,

```
Manipulate[
legendreMethod[m, Take[cfa, m]] ,
{{m,3,Style["Number of constraint functions", Bold, 16]},Range[7],
ControlType→ SetterBar, ControlPlacement→ Top},
Initialization:→ ( (* some stuff here *),
legendreMethod[m_Integer,cfa_List] := Module[
(* compute numerical values assigned to probabilities *)
Grid [(* results *)]
(* more stuff *)
(* end of Module *) ]
(* end of Initialization option *) )
(* end of Manipulate *) ]
```

Hopefully, if you were to inspect the full program when accompanied by some ever expanding top-down meta-code, rather than being put-off by its outward complexity, you would be encouraged to dwell a little longer on the details of the program. I know I would.

Bibliography

- [1] Amari, Shun-ichi and Nagaoka, Hiroshi. *Methods of Information Geometry*. Originally published in Japanese by Iwanami Shoten Publishers, Tokyo, 1993. Translated by D. Harada and published by Oxford University Press, 2000.
- [2] Baierlein, R. *Atoms and Information Theory*. W.H. Freeman and Company, San Francisco, CA, 1971.
- [3] Blower, David J. *Information Processing: Boolean Algebra, Classical Logic, Cellular Automata, and Probability Manipulations. Volume I*. CreateSpace, Amazon.com, 2011.
- [4] Box, G. E. P. and Tiao, G. C. *Bayesian Inference in Statistical Analysis*. Wiley Classics Library Edition. Originally published by Addison Wesley, 1973, John Wiley & Sons, 1992.
- [5] Chandler, David. *Introduction to Modern Statistical Mechanics*. Oxford University Press, New York, NY, 1987.
- [6] Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. Second Edition. John Wiley & Sons, Hoboken, NJ, 2006.
- [7] Feller, William. *An Introduction to Probability Theory and Its Applications*. Volume I, Third Edition. Volume II, Second Edition. Revised printing. John Wiley & Sons, New York, NY, 1968 and 1971.
- [8] Fletcher, Roger. *Practical Methods of Optimization*. Second Edition. John Wiley & Sons, 1987.
- [9] Fougère, P. F. Maximum Entropy Calculations on a Discrete Probability Space. In *Maximum Entropy and Bayesian Methods in Science and Engineering (Vol. 1)*, ed. by G. J. Erickson and C. R. Smith, pp. 205–234, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.
- [10] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian Data Analysis*, First CRC reprint, CRC Press, 2000.

- [11] Gull, S. F. Bayesian Inductive Inference and Maximum Entropy. In *Maximum Entropy and Bayesian Methods in Science and Engineering* (Vol. 1), ed. by Erickson, G. J. and Smith, C.R., Kluwer Academic Publishing, Dordrecht, The Netherlands, 1988.
- [12] Greiner, W., Neise, L., and Stöcker, H. *Thermodynamics and Statistical Mechanics*. Springer-Verlag, New York, NY, 1995.
- [13] Jaynes, Edwin T. *E.T. Jaynes: Papers on Probability, Statistics and Statistical Physics*. Edited by R. D. Rosenkrantz, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1983.
- [14] Jaynes, Edwin T. Information Theory and Statistical Mechanics I, *Physical Review*, 106, pp. 171–190, 1957, Chapter 2 in *Papers*.
- [15] Jaynes, Edwin T. Information Theory and Statistical Mechanics II, *Physical Review*, 108, pp. 620–630, 1957, Chapter 3 in *Papers*.
- [16] Jaynes, Edwin T. Information Theory and Statistical Mechanics, *1962 Brandeis Summer Institute in Theoretical Physics.*, ed. by K. Ford, Benjamin Cummings Publishing, 1962, Chapter 4 in *Papers*.
- [17] Jaynes, Edwin T. Prior Probabilities, *IEEE Trans. on Systems, Science and Cybernetics.*, SSC-4, 227–241, 1968, Chapter 7 in *Papers*.
- [18] Jaynes, Edwin T. Where Do We Stand on Maximum Entropy?, *The Maximum Entropy Formalism*, ed. by R. D. Levine and M. Tribus, MIT Press, Cambridge, MA, 1978, Chapter 10 in *Papers*.
- [19] Jaynes, Edwin T. Concentrations of Distributions at Entropy Maxima, *19th NBER-NSF Seminar on Bayesian Statistics.*, Montreal, October 1979, Chapter 11 in *Papers*.
- [20] Jaynes, Edwin T. On the Rationale of Maximum-Entropy Methods. *Proc. of the IEEE*, Vol. 70, No. 9, September 1982.
- [21] Jaynes, Edwin T. Where Do We Go From Here? *Maximum Entropy and Bayesian Methods in Inverse Problems*, ed. by C. Ray Smith and W. T. Grandy, pp. 21–58, D. Reidel Publishing, Dordrecht, The Netherlands, 1985.
- [22] Jaynes, Edwin T. Monkeys, Kangaroos, and N. *Maximum Entropy and Bayesian Methods in Applied Statistics*, ed. by J. H. Justice, pp. 27–58, Cambridge University Press, 1986.
- [23] Jaynes, Edwin T. *Probability Theory: The Logic of Science*. ed. by G. Larry Bretthorst, Cambridge University Press, New York, NY, 2003.
- [24] Jeffreys, Harold. *Theory of Probability*. Third Edition, Oxford University Press, 1961.

- [25] Kapur, J. N. *Maximum Entropy Models in Science and Engineering*. Revised Edition, Wiley Eastern Limited, New Delhi, India, 1993.
- [26] Kullback, S. *Information Theory and Statistics*. This Dover edition, first published in 1997, is an unabridged republication of the Dover 1968 edition which was an unabridged republication of the work originally published in 1959 by John Wiley & Sons, New York, with a new preface and corrections and additions by the author, Dover Publications, Mineola, NY, 1997.
- [27] Kaplan, W. and Lewis, D. J.. *Calculus and Linear Algebra*, Volume 2, John Wiley & Sons, New York, NY, 1971.
- [28] Moore, Walter J. *Schrödinger: life and thought*. Cambridge University Press, Cambridge, UK, 1989.
- [29] Ruhla, Charles, *The Physics of Chance*. Oxford University Press, Oxford, UK, 1989.
- [30] Schrödinger, Erwin. *Statistical Thermodynamics*. Dover Publications, Mineola, NY, 1989. A reprint of a work first published in 1946 by the Cambridge University Press, Cambridge, UK, entitled, “A Course of Seminar Lectures Delivered on January–March 1944, At the School of Theoretical Physics, Dublin Institute for Advanced Study.”
- [31] Shannon, Claude E. and Weaver, Warren. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL, 1949.
- [32] Sivia, D. S. *Data Analysis: A Bayesian Tutorial*. Oxford University Press, Oxford, UK, 1997.
- [33] Tipler, Frank J. *The Physics of Immortality: Modern Cosmology, God and the Resurrection of the Dead*, Doubleday, New York, NY, 1994.
- [34] Weinberg, Steven. *The First Three Minutes*. Basic Books, New York, NY, 1977.
- [35] Wolfram, Stephen. *A New Kind of Science*. Wolfram Media, Inc., Champaign, IL, 2002.

