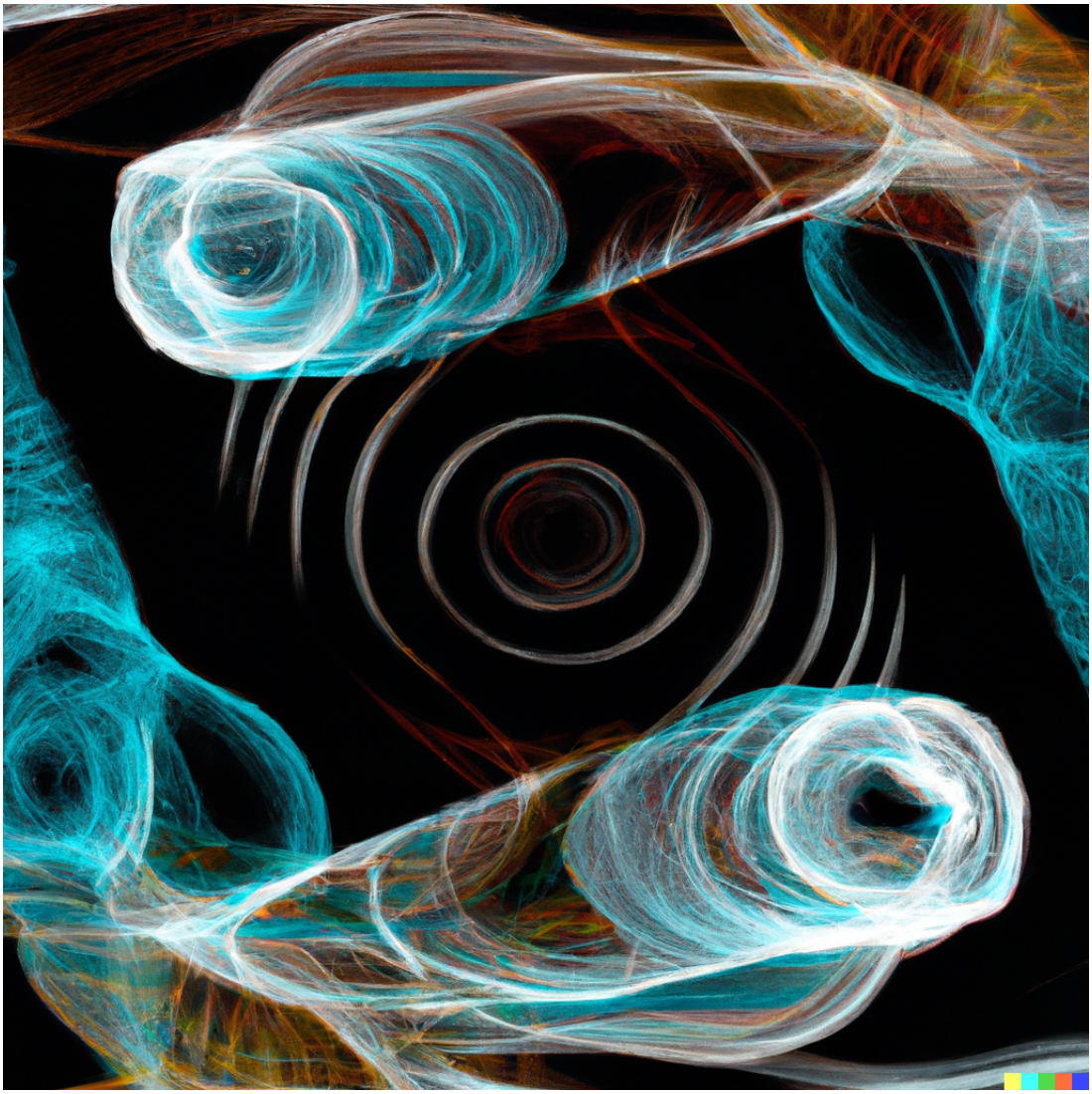
A roadmap to Message Passing methods for inference on Mixed Generalized Linear Models, with emphasis on Mixed Noiseless Phase Retrieval

Simone Maria Giancola

April 6, 2024



Thesis, Master of Science in Data Science
Advisor: Prof. Carlo Lucibello
Bocconi University, Milan, Italy

Abstract

Disclaimer This document is a writing sample of my thesis. The last two Chapters are removed because they are research oriented and not reviewed/published officially. See the Introduction for a description of contents.
The cover was generated with DALL-E 3, a software by OpenAI. The exact query was: “The essence of disordered systems in inference problems, futuristic art”.

Acknowledgements

This thesis is the culmination of 5 challenging and beautiful years, which marks the end of the happiest period of my life experience so far. I would need to write a book to properly acknowledge everyone, but will do my best on this single page.
First and foremost, I would like to thank my advisor, Carlo Lucibello. Over the span of two degrees in Milan, I was a lost student with no basics whatsoever, and he provided me with all the tools to grow as an aspiring researcher, all solely for the love of teaching and disseminating science. I would also like to stress that Luca Saglietti was equally involved in this philantropic activity, and is not my official advisor due to the fact that only one is allowed. I am greatly indebted to both, and can only wish for them the best for what the future holds in their careers. I cherish every single moment, and learnt with time that the Physical intuition and *modus pensandi* is fundamental for academic research.
Over the course of last year, I also had the wonderful opportunity to visit groups that further fostered my motivation to pursue research. In particular, I would like to thank Marco Mondelli and Bruno Loureiro. Marco was the first external Professor to support me and made me enjoy my stay in Austria, with complete freedom and very interesting questions that constitute portions of the final part of this document. Bruno is second just in time, as he was equally as stimulating and helped me learn a lot during my three months in Paris.
I am amazed by how these 4 young successful scientists share very little on their research interests and background but believe in the common thread of supporting younger students and doing beautiful science.
Staying in academia, I am also humbled to say that I was supported by many other incredible professors, that evidently love their field and never avoided giving valuable advice. Without enough space to express my gratitude, I wish to make the names of Giuseppe Savaré and Marc Mézard, that both made the final steps of my studies beautiful. At the same time, I was lucky enough to interact with many students and Post-Docs that have been fundamental. Thank you Diyuan, you are so talented and all of the final part of this work is in close collaboration of me *learning from* you. I do wish you the best. Thank you Leonardo, Simon, Tony, Yihan and Amedeo. I admire your motivation, and never would have thought I could find so many young thriving scientists. Every discussion I had with you was an occasion to learn something new.

Last but not least, I cannot even express how grateful I am to my family. You all always believed in me and my desire to do different things. Thank you to my parents, Carlo and Marianna, for the unconditional support on every aspect. Without you, I would not be here. Thank you to my brother Michele, for growing up with me and making every day less heavy on my soul. Thank you to my grandparents Lina and Luigi, I cannot even picture with words the love you gave me. Thank you my dear extended family of uncles, in Milan and away in Sicily.

Cari amici, la legge mi impone di avere spazio solo per una pagina, e io invece vi chiamerei per nome. Farò in modo di scrivervi una montagna di parole in separata sede. Per ora, sappiate che vi sono eternamente grato.

I acknowledge support from the ISTernship Summer Programme, ref. num. MPC-2023-01128, financed by ISTA, awarded by the OeAD.

List of Algorithms

List of Figures

List of Symbols

The list collects basic symbols used in the document.

Computer Science

$\mathsf{poly}(n)$ polynomial time functions in n

$O(\cdot), o(\cdot), \omega(\cdot), \Omega(\cdot)$ asymptotic notation

Physics

$\langle\rangle_\beta$	Boltzmann posterior expectation at β inverse temperature
$\mathcal{H}$	Shannon entropy
$\mathcal{Z}$	Partition Function
$\mathfrak{F}$	free energy
$\mathscr{F}$	free entropy
$\mathscr{H}$	Hamiltonian
T, β	Temperature, inverse temperature

Statistics and Mathematics

α	Alternation probability in Mixed model
δ	aspect ratio, dataset complexity
$\mathbb{E}\,[]$	expectation
$\mathscr{D}$	dataset
$\mathscr{P}(\mathscr{X})$	set of probability measures on the space $\mathscr{X}$
d	sample dimension
n	sample size

Contents

Introduction

The following document is a redaction of part of the topics I studied in the past three years. The motivations borrow largely from the Statistical Physics literature, which I at the same time loved and had trouble understanding. The elegance of the field is evident, as well as the power of the results. As a matter of fact, some techniques originally intended for the study of Physical phenomena happen to be very effective when dealing with learning systems. To this day, the correspondence amazes me. Despite the large amount of words I could spend admiring the results and the depth of the consequences, I have to stress that I am yet to be accustomed to the field. This Thesis is an attempt to narrow the gap, which hopefully has been successful, in that I took some time to tidy up a large world of publications and techniques. What follows is a brief summary of the objectives and the content of it.

Audience and Style of narrative For the first three Chapters, the ideal reader is inexperienced, and still in the need of a very detailed explanation of the steps. The reason for this choice is that I myself am very junior, and decided to strengthen my basics. The last two Chapters are more advanced and gloss over details, to get quickly to the current state of matter. The funneling from the first to the second step is necessary, but might be abrupt to the less experienced reader. For this reason, some easier/more informal topics are purposely placed in between the advanced ones, as to make the development not too hard and technical.
Whenever possible, formal statements and intuitive explanations are accompanied with all proofs that are not too technical and are at the same time instructive/fundamental. If a proof is very standard, meaning that one internet search gives the answer, it is omitted. If an advanced statement has a clear proof on a reference, preference is given to citing the exact location of the proof rather than copy pasting it on this document.

Addenda To make up for missing steps in the theory and the concepts, I have added two types of boxes:

Question which collect the main aims of the current discussion

Laws/concepts which state general ideas to keep in mind

References which provide the reader with additional sources of information, that hold more details, missing proofs, more precise definitions.

Interesting results that are not essential for the main topic and definitions with standard notations are placed in the Appendix, where the reader can justify some potentially hand-wavy sentences present in the main text.
I think it is also important to remark that this is just a portion of what I wrote down. In particular, I plan to make available a document with topics roughly around the basics of Thermodynamics and mathematical tools for Statistical Physics, which would greatly help the discussion and notation in these Chapters. As a matter of fact, it is impossible to give proper justification to all the objects and the steps required to get a working knowledge of publications of the last 20 years: the history of the field is just too long. I am also working through a deeper redaction of the important aspects of Replica Theory, which is sadly avoided here.

About the last claims Given the fact that this project is still in its primordial phase, all of the claims contained in the second part Chapter V from Section **??** are to be taken as speculations. In other words, we discuss two potential ideas, which constitute part of the research carried out while I was at IST Austria in the summer of 2023. For the sake of preserving current topics of research, the online version of this document will omit those sections. We now showcase a non-technical summary of the contents. For a research oriented review, refer to the abstract.

Informal Motivation and Outline In this thesis, we will introduce the standard way of describing the phenomenology of models falling under the umbrella of Statistics/Machine Learning which arose in the field of Statistical Physics. The main contribution is interpreting the models as *physical* and describing them in terms of phase transitions. This has great advantage in the modern setting in which the abundance of data makes a Data Science question essentially akin to a question in Thermodynamics, where the large size assumption is structural. In other words, a set of tools derived for a large number of particles is useful to confront scenarios with a large number of data-points. We will provide the reader with the essential mathematical tools to derive a starting description of the solvability of a model according to some parameters that describe the appearance of the data. The important industrial advantage is that when this is known for a given task, an ideal company can decide *ex-ante* if investing in the development of a tailored algorithm will be beneficial or not. The summarized conclusion could be stated in even one line "The available data is over/under the solvability threshold". Despite being a set of concepts still under development, it is widely agreed that the conjectures match with experimental evidence, therefore vindicating the methods.
Given the breadth and difficulty of the subject, it is also standard to focus on a sub-class of models, or even a single one. In the second part, the funnelling of the narrative eventually converges towards a single model of importance to the community: noiseless mixed phase retrieval. Summarizing, there are two main issues with it.
On the computational side, we have that the (believed to be) State of the Art Algorithm requires an exogenous initialization that cannot be achieved with a random guess. Put in simple words, another procedure is needed to warm start the best algorithm known so far. This poses the difficult question of understanding if the regime of parameters in which the former and its initializer work coincide. If so, one would like to bring forward arguments that prove it, if not, the next step would be finding a justifiable alternative, if existing.
On the pure solvability side, we would like to understand when the model parameters make the signal detectable at all, and if there are configurations such that the signal is detectable but no Algorithm can. These peculiar regions are themselves a topic of discussion, with a large body of recent literature, bringing forward various conjectural techniques.

Notation

An index of how some basic objects are drawn is in the List of Symbols. Some will be dealt with as standard, without a precise definition, but matching in notation with the introductory document in preparation. A simple search command on the main references (e.g. (**mezardSpinGlassTheory1986**; **mezardInformationPhysicsComputation2009**; **zdeborovaStatisticalPhysicsInference2016**; **krzakalaStatisticalPhysicsMethods2021**; **zhangPreciseAsymptoticsSpectral2022**)) and/or a basic course in Thermodynamics (**arovasLectureNotesThermodynamics2019**) are sufficient to recover more context.

Objects Scalars are in *italic lowercase*, vectors are in ***bold italic lowercase***, matrices are in **BOLD UPPERCASE**. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, a subset determined by some other index i of it is denoted as ∂i.
Matrices are in $\mathbb{R}^{n_1 \times n_2}$ spaces, while generally concatenations of vectors are in $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ when defining functions.

Random quantities are capitalized. The only possible confusion is that a deterministic and a random matrix are both bold uppercase. Whenever not clear from context this will be specified. For example, x is a scalar, X is a scalar r.v., $\boldsymbol{x}$ is a vector, $\boldsymbol{X}$ is a random vector, $\mathbf{X}$ is either a random vector or a random matrix. When we will want to make explicit the dependence on a vector of vectors, we will use the notation $\vec{\boldsymbol{v}}$, and state it clearly. Other notations we could potentially use are $\boldsymbol{x}_{[A]}$ when denoting a subset of $\boldsymbol{X}$ of elements that also belong to a set A, or the `Python`-like notation $\boldsymbol{X}_{[:i]}$ that means elements up to the i^{th} included, or $\boldsymbol{X}_{[i:]}$ which means elements from the i^{th} onwards.
Throughout the text, $\boldsymbol{y}, y, \mathbf{Y}$ will denote observations, while $\boldsymbol{x}, x, \mathbf{X}$ will denote signals to be recovered. The respective spaces, when possibly intended to be general, will be $\mathscr{Y}, \mathscr{X}$.

Conventions Placeholder functions are e.g. f, g, h, while important functions will be defined and assigned specific symbols. Similarly, a, b, c, k, are placeholder variables. The sample size will always be n, the feature size will always be d, apart from examples where we might need them to differ. By i we intend a generic sample, and by j a generic feature. The symbols $i \in [n]$ are to be understood as $i \in \{1, \dots, n\}$. Constants are clear from context and usually take the form of greek lowercase non bold letters, or c, C. Ground truth, when the original object to be found is known, is identified with the symbol $*$, while an object optimal wrt some criterion has symbol $\star$.
The symbol $\succeq$ is the partial order of matrices in the p.s.d. sense, namely $\mathbf{A} \succeq \mathbf{B}$ if and only if $\mathbf{A} - \mathbf{B}$ is p.s.d. and we write $\mathbf{A} - \mathbf{B} \succeq 0$.
Derivatives are denoted with the apex $'$ or classic symbols $\frac{\mathrm{d}}{\mathrm{d}x}, \nabla_x, \frac{\partial}{\partial x}, \partial_x$. When dependence of quantities is crucially on more than one variable we will use ∂ differentiation.
Norms are always denoted as $\|\cdot\|_{\cdot}$ and the normed space is clear from context. The $p = 2$ norm is denoted as $\|\cdot\|_2$, while an r norm as $\|\cdot\|_r$.
The symbol ln denotes natural logarithm, and throughout the text we use the convention $0 \log 0 = 0$ where log is in any base. Recall that $\overline{\mathbb{R}} = [-\infty, \infty] = \mathbb{R} \cup \{\pm\infty\}$.
Imaginary numbers are denoted with the symbol i. For spaces of functions, we use L^p, $\mathrm{C}^1, \mathrm{C}^\infty, \mathrm{Lip}(C)$.

Probability The space of probability measures on $\mathscr{X}$ is denoted as $\mathscr{P}(\mathscr{X})$. For $d \in \mathbb{N}$ let $\mathscr{P}_d(r)$ be the set of Borel Probability measures μ on $\mathbb{R}^d$ such that $\mu(\|\boldsymbol{x}\|_2^r) < \infty$. Equality wrt some criterion is denoted as $\doteq$. For example, equality in distribution is $\overset{\mathsf{d}}{=}$. Expectations are written as $\mathbb{E}_{\cdot}[\cdot]$ where the subscript is optionally used to make clear with respect to which randomness we are averaging. In Statistical Physics literature, the symbol $\langle\cdot\rangle_{\cdot}$ is used when averaging over the Boltzmann distribution, and this will be useful when in need of distinguishing between many integrals over randomness.

Basic Acronyms Standard abbreviations are used throughout the text. We report them here for the sake of completeness: left hand side (LHS), right hand side (RHS), without loss of generality (wlog), want to show (wts), *id est*, that is (i.e.), example given (e.g.), random variable (r.v.), with respect to (wrt),positive semi-definite (p.s.d.), with probability (w.p.), with high probability (w.h.p.), The following are equivalent (TFAE).

Acronyms

a.e. almost everywhere. 82

a.s. almost sure. 108

AE Average Error. 9

AMP Approximate Message Passing. 59

BP Belief Propagation. 45, 47

cdf cumulative distribution function. 109

CLT Central Limit Theorem. 62

d-RSB dynamic Replica Symmetry Breaking. 56

DGP Data Generating Process. 13, 60

e.g. example given. xvii

EA Edwards-Anderson. 23

EC Error Counter. 9

G-AMP Generalized AMP. 61

GC Glivenko-Cantelli. 113

GLM Generalized Linear Model. 5

i.e. *id est*, that is. xvii

LHS left hand side. xvii

MAP Maximum A Posteriori. 8

MEC Minimum Error Counter. 11

MLE Maximum Likelihood Estimator. 8

MMAE Minimum Mean Average Error. 11

MMSE Minimum Mean Squared Error. 11

OL Overlap. 9

p.s.d. positive semi-definite. xvii, 73

pdf probability distribution function. 10

PGF Probability Generating Function. 45

r-BP relaxed Belief Propagation. 62

r.v. random variable. xvi, xvii

RHS right hand side. xvii

RS Replica Symmetric. 56, 75

s-RSB static Replica Symmetry Breaking. 56

S-to-C Statistical to Computational. 16

SE Squared Error. 9

SNR Signal-to-Noise Ratio. 7

SoS Sum of Squares. 55

TAP Thouless-Anderson-Palmer. 60

TFAE The following are equivalent. xvii

VC Vapnik-Chervonenkis. 109

w.h.p. with high probability. xvii

w.p. with probability. xvii

wlog without loss of generality. xvii

wp with probability. 3

wrt with respect to. xvii

wts want to show. xvii

Chapter I

Modern Statistical Inference

In this Chapter, we review Inference with an eye on Statistical Physics, to later review Statistical Physics with an eye on Inference in Chapter II. For this reason, the two are to be dealt with jointly.
Section I.1 introduces the motivations of this thesis. After a discussion of the origins of the method, in Subsections I.1.1 and I.1.2 the mathematical framework is presented. The narrative proceeds through a summary of the Bayesian viewpoint in Section I.2, with the first branching of approaches: point estimates (Subsec. I.2.1) and the advantage of aiming for an answer in terms of distributions as far as possible in the analysis, introducing the notions of risk and decision (Subsec. I.2.2). To conclude, we give a brief justification on a part of the parameterization of our problem in Section I.3 in terms of aligning with modern, high-dimensional datasets, and explain the landscape of feasible and efficiently feasible estimation with the notion of Statistical-to-Computational gap (Subsec. I.3.1).

I.1 What is Statistical Inference?

Inference is the result of a reasoning process of building a conclusion backed by logic about an observation. The Greek Philosopher *Aristotles* (300 BC) theorized two main branches in the field: *deduction* and *induction*. The former is more rigid, and consists in evoking logical arguments that bring from true premises to a conclusion. The latter is focused on deriving universal laws from local/restricted evidence. We avoid discussing this distinction, and rather present an example and a paradigm of interest. Given the mathematical flavour, it will be increasingly formalized, with the purpose of justifying each step in a bottom-up fashion.

Statistical Physics is a field that aims to describe **macroscopic** phenomena as a direct derivation of **microscopic** physical laws in large size systems. Inference, in some sense, aims to find structure in data. Despite seemingly different, *Pierre Simone, Marquis de Laplace* (1749-1827) contributed greatly to the former, and initiated the latter, a motivation mentioned in the interesting review (**zdeborovaStatisticalPhysicsInference2016**). The most straightforward link is established considering the microscopic atoms to be representatives of the $i \in [n]$ datapoints/nodes/neurons. In today's applications, the connection is even stronger. Large sets of information to be analyzed, extracting relations between billions of items involving billions of parameters are in great accordance with techniques that were designed for systems of the size of Avogadro's number, $n = 10^{23}$. We take this as a motivating introduction to present our general framework, but stress that the connections become considerably deeper (see (**mezardInformationPhysicsComputation2009**) for a classic reference or those mentioned in (**zdeborovaStatisticalPhysicsInference2016**)). In particular, we will focus on methodologies that originally started in the spin glasses literature (**mezardSpinGlassTheory1986**). Among other things, a very interesting analogy between phase transitions and inference feasibility boundaries will be observed.

The perspective is that of a **statistician**: a person that is given a problem with a well defined set of questions that can be answered. Analyzing the various properties of the phenomenon, the statistician will ultimately derive satisfactory conclusions,

to some extent. The way in which the subject is able to access the problem is via a channel of communication. The statistician is a *receiver* of information from a sender. The architecture of the channel is known up to some point, and the original message has to be retrieved. We view this as an imaginary scenario in which a friend of a statistician is on one side, and our character is on the other, and they can only exchange information through an object that is termed *channel*. This idealistic scenario turns out to be very accurate in reality, and also comes with different interpretations. While the choice is just to give an idea of the playground in which we are placed, it bears enough character to be termed sufficient.

For ease of exposition, we will mostly make the assumption that the generating process is known up to some point, and the aim is only to recover the original formulation. In other words, we assume that we do not need to test an hypothesis on the model. With this set of information, the problem is not statistically friendly. To be clearer: it rarely makes sense to assume that no model selection is needed. However, in Physics one usually inspects the mechanics and validity of a model rather than actually deriving it, so the choice is justifiable.
Incidentally, many modern studies (above all those on neural networks) appear to be well working despite glossing over grounded details. We take this as a partial justification for the freedom to place ourselves in this scenario. Although we will not focus on Neural Networks in this document, the reasoning can be fairly extended with the general aim of describing the phenomenology of some observations.

Further References

Many nice introductions to the common aspects of Machine Learning and Statistical Physics are found in literature (**krzakalaStatisticalPhysicsInference2015**; **zdeborovaStatisticalPhysicsInference2016**; **krzakalaStatisticalPhysicsMethods2021**; **decelleIntroductionMachineLearning2022**). The collection starting with the preface (**agliariMachineLearningStatistical2020**) is also a good starting point to see current research topics at the interface. For deep learning oriented reviews, some options are (**advaniStatisticalMechanicsComplex2013**; **bahriStatisticalMechanicsDeep2020**).

I.1.1 A perspective from message retrieval

In the simplest possible setting, a statistical inference problem is formulated as follows. For a given set of signals (variables) $\mathbf{X}^* \in \mathscr{X}$ an observer has access to n observations $\mathbf{Y} \in \mathscr{Y}^{\times n}$, or more in general to a dataset $\mathscr{D}$ with n samples of the phenomenon. The task is to retrieve $\mathbf{X}^*$ from $\mathbf{Y}$, upon *knowledge* of which *relation* the two have. With knowledge and relation, we specifically mean that signals and observations can be embedded in an equation that describes how one originates from the other. The easiest understanding of this is a communication system, where:

$$\mathbf{X}^* \rightsquigarrow \boxed{\mathfrak{C}} \rightsquigarrow \mathbf{Y}, \tag{I.1.1}$$

and it is understood that $\mathfrak{C}$ is a channel to be specified.

Example I.1.2. *A sender sends a (vector) message $\boldsymbol{x}^*$, and a statistician reads on a tape $y \in \mathbb{R}$, a scalar. Upon seeing a stream of n messages, the receiver has read $\boldsymbol{y} \in \mathbb{R}^n$ communications.*

At this point, it is standard to assume that the messaging system is *unfaithful*, namely that $\mathbf{Y} \neq (\mathbf{X}^*)^{\times n}$. The reason is simple: aiming to reproduce a realistic scenario, it is common to hypothesize that the act of communicating *changed* the signal, so that the observed phenomenon of the receiver is not equal to the original one. This channel of communication, represented as a nontrivial function $\mathscr{X} \rightsquigarrow \boxed{\mathfrak{C}} \rightsquigarrow \mathscr{Y}$ is placed on purpose, and expands the formalism to potential tweakings of the signal. Without this change, the questions would be self-answering, and $n = 1$ would ensure signal retrieval.

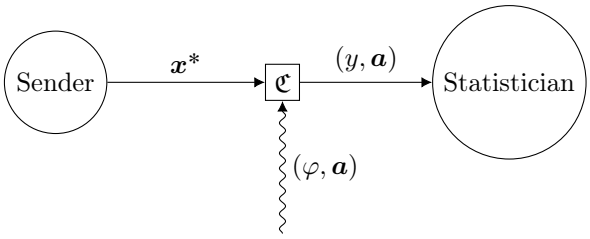


Figure I.1: Parametrized Compression Map
The scheme is a typical example of vector reconstruction from scalar observations.

Most practical applications pursue guarantees for efficient communication, which is to be understood as being able to summarize enough, potentially at a lower cost than the original message. In our mathematical formalism, this summarization is seen via a compression. Namely, the channel applies a function $\varphi : \mathscr{X} \to \mathscr{Y}$ that maps messages from the sender to the receiver. In particular, we will focus on a compression that lowers substantially the dimensionality of the message, so that $\mathscr{X} \subseteq \mathbb{R}^d$ and $\mathscr{Y} \subseteq \mathbb{R}$, which amounts to compressing by a factor of d. It is worth noticing that this is only one of the many justifications for the formalism, and one of the many scenarios. It could be that the signal and the observation have the same size, but the function corrupts differently. Often, we represent this through a scheme, which is a rule for performing compression. We see this as a specific form of $\varphi(\cdot;\cdot)$, possibly parametrized. For the sake of the narrative, these principles of compression motivate a restriction of our interests to a paradigmatic setting, found in the box below.

Vector reconstruction from scalars

In most examples, unless specified, we will have $\mathscr{Y} \subseteq \mathbb{R}$ and $\mathscr{X} \subseteq \mathbb{R}^d$. Thus, observations are scalars and independent variables are vectors. We stack n observations into a vector $\boldsymbol{y} \in \mathbb{R}^n$. In this case, the way compression is performed is typically through a function φ that among other operations applies a measurement matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, which is received by the statistician.

A diagram is Figure I.1.

Example I.1.3. *Let φ be such that $\varphi(\boldsymbol{x}) = \langle \boldsymbol{a}_i, \boldsymbol{x}\rangle$ where $\boldsymbol{a}_i \in \mathbb{R}^d$ over $i \in [n]$ observations. Letting $\mathbf{A} \in \mathbb{R}^{n\times d}$, we have that $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}^*$. This is a linear system of equations. The channel is projecting at each observation the signal according to a set of vectors $(\boldsymbol{a}_i)_{i\leq n}$. The observations in this case are $\mathscr{D} = \{(\boldsymbol{a}_i, y_i)\}_{i=1}^n$. The $\boldsymbol{a}_i$ vectors act as auxiliary information: they parametrize the act of communication.*

We narrow the ways in which these realizations manifest, focusing on a specific formal approach: *statistical inference*. An inferential process that involves statistics is described by embedding the signal-message problem with randomness. In particular, the n observations are a sample from a larger population, which admits a probability distribution. In this renewed scenario, we endow $(\boldsymbol{y}, \boldsymbol{x}^*)$ with a probability measure, and aim to draw conclusion about the latter from the former. To be able to do this, we must adopt the assumption of a specific *statistical model*.
While seemingly restrictive, this step is in some sense a generalization. Deterministic phenomena are a subset of stochastic phenomena described by unit-mass measures, i.e. events with probability (wp) 1. In addition to this, the choice is natural in two ways. Firstly, it is safe to study probability models: the risk of making mistaken assumptions is mitigated. Secondly, natural events present randomness: randomness induced by complexity[1]which accounts for what we cannot explain, and randomness induced by variability, which we can explain, but cannot decode at each time instantly. In simple terms we are considering "the process of using data analysis to infer properties of an underlying distribution of probability"

[1]this aspect is strongly linked to the probabilistic approach to thermodynamics we will discuss in another document.

(**uptonDictionaryStatistics2008**). After this second lifting, we will have to study problems where the channel acts on the signal and the process is nondeterministic according to some assumed rules. We summarize this as $\boldsymbol{Y} = \varphi \odot \boldsymbol{x}^*$, with capitals to highlight the randomness we could have, and $\odot$ denoting element-wise application of the compression function. Notice that potentially the focus has transitioned to *answering probabilistic questions* about $\boldsymbol{x}^*$, since randomness will be involved.
We further allow the relationship to be corrupted by noise. For simplicity, we will only consider *additive noise*, which amounts to adding a scalar term $\epsilon_i \in \mathbb{R}$ to the equation for each signal and observation.
To justify why this makes sense, we must resort to application arguments. The main reason is that real systems are subject to failure, and being able to describe such failure is useful whenever the theoretical construction can be made close to the concrete problem. The concept of corruption can be formalized in many ways, but always with similar descriptions. The most intuitive explanation is seen for a measurement problem. If the channel φ is any physical instrument, then many objections can be made to the validity of it. Assuming for simplicity that it does not have systematic errors[2], its calibration to a certain precision will necessarily mean that all the quantities of lower order will be (randomly) neglected.
In other words, the choice of an instrument with associated units of measure is already implying a mistake: if precision is up to centimeters, any length that is in the millimeters is by construction uncertain, and measurements are expressed as $\# \pm 1$ cm. In most cases, if the instrument is reliable enough, this noise can be taken to be random and symmetric about zero. On average no mistake is made, but eventually a mistake of overestimation or underestimation is inevitable, with equal weights. Again, this is a reasonable construction, since *a priori* it would not make sense to consider asymmetric noise, but rather admit that our instrument has a larger variance and is still symmetric. Additionally, this makes most models more amenable to mathematical analysis.

I.1.2 Statistical Model and Statistical Problem

We thus reach a general expression

$$\boldsymbol{Y} = \varphi \odot (\boldsymbol{x}^*, \mathbf{A}) + \boldsymbol{\epsilon}, \tag{I.1.4}$$

where $\mathbb{E}[\epsilon_i] = 0$ and $\mathrm{Var}[\epsilon_i] = \Delta$ for all $i \in [n]$ and $\odot$ denotes component wise application of φ. By the generality of our reasoning, it is possible to retrieve all the previous stages, or any combination, by just choosing the specific dimensions, randomness, and functions in place. To represent this, we will give a formal definition that encapsulates a large set of scenarios.

Definition I.1.5 (Statistical Model)**.** *A collection $\mathfrak{M} = (\mathscr{X}, \mathscr{Y}, \varphi, \mathbb{P}_{\boldsymbol{\epsilon}}, \mathbb{P}, n)$ that describes communication of signals from a sender to a receiver through a channel $\mathfrak{C}$. We recognize:*

- *$(\mathscr{X}, \mathscr{Y})$ the message space and the receiver space.*
- *φ the compression function*
- *$\mathbb{P}_{\boldsymbol{\epsilon}}$ the distribution of the noise vector $\boldsymbol{\epsilon} \in \mathscr{Y}^{\times n}$, assumed to factorize into $\mathbb{P}_\epsilon$*
- *$\mathbb{P}$, the collection of randomness over any of the above objects*
- *n the sample size.*

A depiction is Figure I.2.

When given a collection of n samples, the statistician is also provided partial or full knowledge about the channel , e.g. some distributions and the compression map $\mathfrak{C} = \{\varphi, \mathbb{P}_{\boldsymbol{\epsilon}}, \mathbb{P}_{\mathfrak{C}}\}$. This motivates the construction of a specific principle.

[2] e.g. it does not always add +2 since it is out of scale

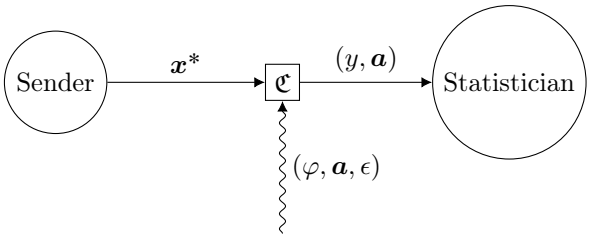


Figure I.2: Sender-Statistician diagram of a Statistical Model $\mathfrak{M}$.
The channel applies the compression map and noise. The result is that the receiver gets only the measurement used $\boldsymbol{a}$ and the final observation y. The task is recovering $\boldsymbol{x}^*$ by reusing this channel. A difficult task means that the channel has to be queried a large number of times. A task is impossible when even an infinite amount of queries does not provide a meaningful result.

Problem I.1.6 (Statistical Problem)**.** *The Statistical problem is handed to the statistician. The material is a deterministically observed tuple $\mathfrak{P} = (\mathscr{D}, \mathfrak{H})$, with which we would like to answer a question about the statistical model, that is its source. It is composed of an n-sized sample from the population, which we call dataset $\mathscr{D}$, and a hypothesis $\mathfrak{H}$. In practice, the hypothesis is just an indication of how calculations should be performed, what can be assumed, what is known, what is to be inferred.*
The statistician is in particular required to answer questions about $\boldsymbol{x}^*$.

Below, we show some starting examples of how one can describe common problems in this framework.

Example I.1.7 (Intuitive Statistical Inference problems)**.** *On the informal side, some statistical inference problems are:*

1. *recovering the original image from a blurry observation*
2. *reconstructing a corrupted message from a noisy voice line*

Example #1 could be formalized with observations $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ where $\mathbf{Y}_i = \mathbf{X}^* + \boldsymbol{\Xi}_i$. In this case all the objects are matrices, and $\mathbf{X}^*$ is the deterministic original image. The channel just amounts to adding noise $\boldsymbol{\Xi}_i$ to each observation, blurring the image.
Example #2 is seen as a general message transmission where $\boldsymbol{y}_i = \varphi(\boldsymbol{x}^*)$ and $\varphi(\cdot)$ maps to some resulting vector that (possibly) is itself non-scalar.

Example I.1.8 (Denoising)**.** *Let the statistician observe a vector $\boldsymbol{y} \in \mathbb{R}^n$ noisy observations of a signal $x^* \in \mathbb{R}$, corrupted by a Gaussian noise $\boldsymbol{\epsilon} \in \mathbb{R}^n$. The equation of the model reads:*

$$\boldsymbol{y} = \begin{bmatrix} x^* \\ \vdots \\ x^* \end{bmatrix} + \boldsymbol{\epsilon} \quad y_i = x^* + \epsilon_i \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \Delta), \quad \forall i \in [n]. \tag{I.1.9}$$

Where we see that there is no compression, and the mapping function is $\varphi(x) = x$. In this case, the variance of the noise is the only disturbing factor that does not allow to guess directly the solution.

Example I.1.10 (Classic Linear Regression)**.** *In linear regression, the model is expressed as*

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{x}^* + \boldsymbol{\epsilon} \qquad y_i = \langle \boldsymbol{a}_i, \boldsymbol{x}^* \rangle + \epsilon_i \quad \epsilon_i \sim \mathcal{N}(0, \Delta) \quad \forall i \in [n]. \tag{I.1.11}$$

The compression component here is the matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ and the noise component is a standard Gaussian $\boldsymbol{\epsilon}$, which plays the role of corruption. In the classic framework, the task is rephrased as retrieving the vector that relates a list of scalar dependent observations $\boldsymbol{y}$ and independent explanatory variables encoded in a matrix $\mathbf{A}$. The dataset is $\mathscr{D} = (\boldsymbol{y}, \mathbf{A})$.

In particular, we will focus on a paradigm of inference that generalizes many models, and a specific case that is of practical interest in many applications. We briefly outline them below.

Example I.1.12 (Generalized Linear Model)**.** *For the purpose of this document, a Generalized Linear Model (GLM) is a problem where the compression function is $\varphi(\cdot, \mathbf{A}) : \mathbb{R}^d \to \mathbb{R}$, where $\mathbf{A}$ is a possibly random matrix in $\mathbb{R}^{n\times d}$ independent of the signal and the map is applied component-wise. Unless otherwise stated, the observations are independent and identically distributed. Noise is additive.*

Example I.1.13 (Phase retrieval)**.** *A specific example of GLM is Phase retrieval, where the compression function is $\varphi(\boldsymbol{m}) = \begin{bmatrix}|m_1| & \cdots & |m_n|\end{bmatrix}^\top$, the symbol m_i representing the message from the compression $\langle \boldsymbol{a}_i, \boldsymbol{x}^* \rangle$ at each row. We further assume that observations $i \in [n]$ are independent and identically distributed and that the vectors $\boldsymbol{a}_i$ admit a distribution. Basically, the original vector is projected in different ways at each observation, its sign is removed, and further it is corrupted by white noise $\epsilon_i \sim \mathcal{N}(0, \Delta)$. The task is recovering $\boldsymbol{x}^*$ from $\mathscr{D} = (\boldsymbol{y}, \mathbf{A})$, which is an n-sized sample from the randomness of the channel. We write this succintly as:*

$$\boldsymbol{y} = |\mathbf{A}\boldsymbol{x}^*| + \boldsymbol{\epsilon}, \tag{I.1.14}$$

where $|\cdot|$ is applied row-wise and $\mathbf{A}$ is a random matrix.

Common scenarios in which the phase retrieval problem arises are related to measurements in which only the magnitude of the observation is recovered, and the orientation is *lost*. Indeed, the modulus has this exact role in the formalism of Example I.1.13. From a theoretical standpoint, it is interesting for the same reason: the phase holds a considerable amount of information to detect a message, and losing it poses a challenging task. To give an example, any phase retrieval problem for *centered* randomness sources will have expectation zero, since it is not known if the single observation came from $\langle \boldsymbol{a}_i, \boldsymbol{x}^* \rangle$ or $\langle \boldsymbol{a}_i, -\boldsymbol{x}^* \rangle$. Clearly, more observations will provide information of how this vector might place itself, but estimation might also end up in results that hold up to a *global sign change*, to be intended as *flipping all signs* in the guess of the signal. We will come back to this example continuously. Even more importantly, in Chapter V we will take a generalized formulation and analyze it in detail.

I.2 Bayesian Inference

The paradigm is not closed, since no methodology can be provided as of now. We close our mathematical model by focusing on Bayesian parametric Estimation problems of a specific kind. Formally, from a common distribution on observations $(\mathscr{D}; \mathfrak{H})$ we assume that:

- the observations admit a joint distribution $\mathbb{P}[\mathscr{D}]$
- the observations are iid, hence wlog one can study them individually
- the distribution over signal and observation is separable into a prior and a likelihood term, so that $\mathbb{P}[(\boldsymbol{X}, \mathscr{D})] = \mathbb{P}[\boldsymbol{X}]\mathbb{P}[\mathscr{D} \mid \boldsymbol{X}]$
- alternatively information about the channel $\mathfrak{C}$ is given to derive a likelihood from a prior.
- whenever disputable, probabilities and densities are well-behaved.

Placing a prior on the signals ensures that Bayes' Theorem is applicable. This choice has a big philosophical implication. A *random* truth to be recovered means that we will wish to formulate an answer that takes into account the fact that the prior is a *hypothesis*. In simple words, if we wished to retrieve a random version of $\boldsymbol{x}^*$, a prior would already be enough. Instead, we pursue an answer that is backed by information, with a more robust degree of belief/reliability.

Remark I.2.1. *Bayesian inference and its terminology are assumed to be known, a good reference for its foundations is (***definettiLogicaIncerto1989***).*

Remark I.2.2. *It is important to understand that this is almost a forced choice. Even if it were known that $\boldsymbol{x}^*$ existed, it is not clear how to retrieve it directly from $\boldsymbol{y}$. What one can do is hope to find a scheme to update a trial density that describes a current belief with subsequent observations of the phenomenon. The improper choice of assigning a prior is however largely debated in the Bayesian vs Frequentist approach to statistics (see (***cramerMathematicalMethodsStatistics1999***; **fellerIntroductionProbabilityTheory2009***; **jeffreysTheoryProbability1998***; **keynesTreatiseProbability2004***))*
In simple words, our choice allows to model a phenomenon of interest stochastically, despite it not being necessarily stochastic in nature.

Inference is then translated into answering questions about the posterior distribution:

$$\mathbb{P}\left[\boldsymbol{X} \mid \mathscr{D}\right] = \frac{\mathbb{P}\left[\mathscr{D} \mid \boldsymbol{X}\right]\mathbb{P}\left[\boldsymbol{X}\right]}{\mathbb{P}[\mathscr{D}]}, \tag{I.2.3}$$

which is effectively an update of our belief based on the observations $\mathscr{D}$. To show that the paradigm applies also to more general scenarios, we provide another standard example that involves corrupted matrices.

Example I.2.4 (Spiked Wigner Model)**.** *A canonical example of signal recovery is formalized as follows. For a true signal $\boldsymbol{x}^*$ we observe a matrix:*

$$\mathbf{Y} = \sqrt{\frac{\lambda}{n}}\boldsymbol{x}^*\boldsymbol{x}^{*\top} + \mathbf{W}, \tag{I.2.5}$$

where $W_{ij} \sim \mathcal{N}(0,1)$ for all $i \neq j$ and $W_{ii} \sim \mathcal{N}(0,\Delta)$ on the diagonal. In this case $\lambda \in \mathbb{R}$ plays the role of a Signal-to-Noise Ratio (SNR). In simple terms, it is a tunable parameter that determines how much the true signal weights wrt background noise. To different λ correspond different statistical problems. Intuitively, the higher the parameter the weaker is the corruption (i.e. the signal makes most of the matrix $\mathbf{Y}$. This is superfluous, as it could be absorbed in the noise variance with a rescaling. However, we will see why it makes sense to make it explicit when discussing hardness in Section II.6.
A Bayesian structure on the model is then specified after allowing $\boldsymbol{x}^*$ to be randomly distributed (again, for the sake of embedding the problem in the formalism). An interesting and very tractable case is when $x_i \sim \mathsf{Rad}(\pm 1)$, which amounts to each entry being ± 1 with given probabilities (see **alaouiFundamentalLimitsDetection2018***).*

The formalism allows to provide the statistician with many answers about basic models, and much inspiration about the more complicated ones. There are two main flavours. On one side, we might hope to compute the posterior and sample from it, one the other side, we might wish to find a reasonable estimator for $\boldsymbol{x}^*$ that exploits data.

Example I.2.6 (Gaussian Bayesian Posterior)**.** *Let $X \sim \mathcal{N}(0,1)$, $Y_i = X + \epsilon_i$ where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1)$. A well known result is that for a sequence of observations $\boldsymbol{y} \in \mathbb{R}^n$:*

$$\mathbb{P}\left[X \mid \boldsymbol{Y} = \boldsymbol{y}\right] = \frac{\mathbb{P}\left[\boldsymbol{y} \mid X\right]\mathbb{P}\left[X\right]}{\mathbb{P}\left[\boldsymbol{y}\right]} \propto \mathbb{P}\left[\boldsymbol{y} \mid X\right]\mathbb{P}\left[X\right] = \prod_{i=1}^{n} e^{\frac{-(y_i-x)^2}{2}} e^{\frac{-x^2}{2}}, \tag{I.2.7}$$

where we have ignored constants wrt X, used the iid assumption and expressed everything in terms of densities. Reordering some terms:

$$\dots \propto e^{-\frac{1}{2}(x^2 + \sum_i y^2 - 2x\sum_i y_i + nx^2)} \propto e^{-\frac{1}{2}[(n+1)x^2 - 2x\sum_i y_i]}. \tag{I.2.8}$$

The kernel of a normal distribution $\mathcal{N}(\mu,\sigma^2)$ is then recognized. The mean and variance have closed form expressions:

$$\mu = \frac{\sum_i y_i}{n+1} \qquad \sigma^2 = \frac{1}{n+1}. \tag{I.2.9}$$

Example I.2.10 (Noisy Rademacher Signal)**.** *A low-dimensional model similar to Exm. I.2.4 is constructed as follows. Assume the relationship is:*

$$y_i = x^* + \epsilon_i\sqrt{\Delta} \quad \epsilon_i \sim \mathcal{N}(0,\Delta), \Delta \in \mathbb{R} \qquad \forall i \in [n]. \tag{I.2.11}$$

Placing a Rademacher one dimensional prior $X \sim \mathsf{Rad}(\pm 1)$*, one can check that the posterior will have density:*

$$\mathbb{P}\left[X \mid \boldsymbol{Y} = \boldsymbol{y}\right] = \frac{1}{1 + e^{-x\sum_i \frac{y_i}{\Delta}}} = \sigma\left(x\sum_i \frac{y_i}{\Delta}\right) \tag{I.2.12}$$

Where $\sigma(\cdot)$ *is the sigmoid function.*

In general, when the posterior is easy to compute and easy to sample from, the problem of inference is solved. For reasons that we will see later, it turns out that this is not always the case, because the computation of the denominator of Bayes' rule is very hard. We call this object a **partition function**. Intuitively, the difficulty in computing it lies in the fact that for *large* ambient spaces of the signal, the term becomes a large sum, with no general expression. On the other hand, it can be shown that it is a cumulant generating function for the randomness of the problem, hence bearing the statistical information required to answer inference questions. On a different perspective, it has the role of normalizing the likelihood-prior numerator to a valid probability, establishing how different configurations partition into the space of probabilities. Clearly, for a hard to solve problem, it will be hard to compute. The very question of many fields falling into the inference incubator is finding workarounds, approximations or direct results on the partitition function.
For the moment, we avoid discussing this, and move to the most sensible next thing, which is to provide a point estimate that can describe the shape of the posterior, and more importantly, give an approximation of the signal $\boldsymbol{x}^*$. We review this method below.

Most of the cases, we will focus on parametric models, which add the further notion that the relationship of the encoding between observation and signal is by hypothesis in a space that can be parametrized. This amounts to assuming that the random distributions admit a description $\mathbb{P}[\cdot;\boldsymbol{\vartheta}]$ where $\boldsymbol{\vartheta} \in \mathbb{R}^p$. Knowledge or partial knowledge of the distributions amounts to different specifications of these parameters.

I.2.1 Point Estimates

An estimator $\widehat{\boldsymbol{x}}$ is a function that admits as input a dataset and returns a guess for $\boldsymbol{x}^*$. Mathematically:

$$\widehat{\boldsymbol{x}} : \{\mathscr{D}\} \to \mathscr{X}. \tag{I.2.13}$$

The most intuitive choices of an estimator are the Maximum Likelihood Estimator (MLE) or the Maximum A Posteriori (MAP). These have form:

$$\widehat{\boldsymbol{x}}_{\mathsf{MLE}} \coloneqq \arg\max_{\mathscr{X}} \left\{\mathbb{P}\left[\mathscr{D} \mid \boldsymbol{X}\right]\right\} \tag{I.2.14}$$

$$\widehat{\boldsymbol{x}}_{\mathsf{MAP}} \coloneqq \arg\max_{\mathscr{X}} \left\{\mathbb{P}\left[\boldsymbol{X} \mid \mathscr{D}\right]\right\}. \tag{I.2.15}$$

We claim that the two have some advantanges but present serious criticalities.

Remark I.2.16 (Pros and shortcomings of MLEs and MAPs)**.** *Recognize the following easy assertions for the MLE and the MAP:*

- *both are a point estimate. If a distribution is bimodal, they are inaccurate.*
- *the MAP and MLE are useful when the distribution is concave and first order algorithms are well-behaved*
- *if the distribution is not concave, algorithms for the MAP and the MLE converge to a sub-optimal point (a local stationary point). In other words, both depend on starting conditions*

- *they are not a distributional estimate, and thus lack flexibility*
- *the MLE is sensitive to outliers, since it relies on the distribution of observations only*
- *estimating them can be an NP-hard problem (***shimonyFindingMAPsBelief1994***;* **yiAlternatingMinimizationMixed2014***).*

Despite being a non-exhaustive list, we claim it is sufficient to justify a different approach.

Given the circumstances, a naturally arising question would be the following.

Bayesian Point Estimation

Is it possible to consider different strategies?

Both approaches treat the target as something that can be estimated pointwise. On the other hand, Bayesian principles implicitly require to treat unknowns as if they came from a distribution. This suggests postponing the search for a single answer to the latest stage possible. Leveraging the randomness, we hope that the result will be

- robust to variations of the signal
- adjusted to the peculiarities of the problem.

I.2.2 Risk-Based Approach from Decision Theory

To answer the above question, we define a measure of correctness for estimators and a systematic way to evaluate comparisons. This is carried out by constructing performance-based estimators, that are compared in terms of an error/loss function:

$$\mathcal{L} : \mathscr{X} \times \mathscr{X} \to \mathbb{R}_+ \quad (\widehat{\boldsymbol{x}}(\mathscr{D}), \boldsymbol{x}^*) \to \mathcal{L}(\widehat{\boldsymbol{x}}(\mathscr{D}), \boldsymbol{x}^*). \tag{I.2.17}$$

Notice that the loss is random given that $\mathscr{D}$ is randomly sampled. Interestingly, we wish to average the result over the randomness of observations. This ensures that our techniques will be somewhat robust to variations in the signal and adaptive to the problem by construction. To a problem with randomness, we inject further randomness and model the whole in a way such that *reliability* is encoded in a loss.

Example I.2.18 (Common Error functions)**.** *Among the most famous costs we list:*

- *the* L^2 *norm, or Squared Error (SE)* $\mathcal{L}_{\mathrm{L}^2}(\widehat{\boldsymbol{x}}(\mathscr{D}), \boldsymbol{x}^*) := \frac{1}{d} \|\widehat{\boldsymbol{x}}(\mathscr{D}) - \boldsymbol{x}^*\|_2^2$
- *the Average Error (AE), or* L^1 *norm* $\mathcal{L}_{\mathrm{L}^1}(\widehat{\boldsymbol{x}}(\mathscr{D}), \boldsymbol{x}^*) := \frac{1}{d} \|\widehat{x}(\boldsymbol{y}) - \boldsymbol{x}^*\|_1$
- *the Error Counter (EC), for discrete data,* $\mathcal{L}_{\mathsf{EC}}(\widehat{\boldsymbol{x}}(\mathscr{D}), \boldsymbol{x}^*) := \frac{1}{d} \sum_{j=1}^{d} \mathbb{1}_{\widehat{x}_j \neq x_j^*}$.

Remark I.2.19. *The* L^2 *error is a proper norm, a proper loss and it induces a Hilbert space under suitable conditions. Namely, all square integrable functions wrt a reference measure* μ *form a Hilbert space with equipped* L^2 *norm and a notion of inner product. Being Hilbert, every Cauchy sequence is convergent, and all the nice properties of Hilbert spaces are derived.*

Another performance indicator commonly encountered in Physics is the Overlap (OL) of the estimator with the signal, which is just a normalized alignment:

$$m := \frac{\langle \widehat{\boldsymbol{x}}, \boldsymbol{x}^* \rangle^2}{\|\widehat{\boldsymbol{x}}\|_2^2 \, \|\boldsymbol{x}^*\|_2^2} \in [0, 1]. \tag{I.2.20}$$

Remark I.2.21. *The overlap and the* L^2 *error are closely linked when signals and estimators have fixed norm.*

With a notion of error, we can further construct a framework that takes into account the probabilistic structure of the problem. Indeed, $\boldsymbol{x}^*$ is in principle unknown and cannot be used for a direct evaluation of the loss, which is itself random. Leveraging access to the various distributions allows to construct a strategy that is robust against the presumed randomness of the problem. If we imagine that $\boldsymbol{x}^*$ was sampled, we can represent the performance in terms of a determinstic quantity.

Definition I.2.22 (Risk)**.** *Given an estimator $\widehat{\boldsymbol{x}}$ and an error function $\mathcal{L}$, the risk is a functional that averages the loss over the joint distribution $\mathbb{P}[\cdot,\cdot]$ of signal and dataset. Namely*

$$\overline{\mathscr{R}}(\widehat{\boldsymbol{x}};\mathcal{L}) \coloneqq \mathbb{E}_{\boldsymbol{X},\mathscr{D}}\left[\mathcal{L}(\widehat{\boldsymbol{X}}(\mathscr{D}),\boldsymbol{X})\right] = \mathbb{E}_{\mathscr{D}}\left[\mathscr{R}(\widehat{\boldsymbol{X}} \mid \mathscr{D};\mathcal{L})\right] = \mathbb{E}_{\mathscr{D}}\left[\mathbb{E}_{\boldsymbol{X}|\mathscr{D};\mathcal{L}}[\mathcal{L}]\right]. \tag{I.2.23}$$

We remark that $\mathscr{D}$ is per se a random variable, and $\boldsymbol{X}^*$ is random (written as $\boldsymbol{X}$ here) since it is assumed to be sampled from a prior. By the assumed factorization of the probability, the second equality follows, with the risk now being the expectation of the loss wrt the posterior. In particular:

$$\arg\min_{\widehat{\boldsymbol{x}}} \overline{\mathscr{R}}(\widehat{\boldsymbol{x}};\mathcal{L}) = \arg\min_{\widehat{\boldsymbol{x}}} \mathbb{E}_{\mathscr{D}}\left[\mathscr{R}(\widehat{\boldsymbol{x}} \mid \mathscr{D};\mathcal{L})\right]. \tag{I.2.24}$$

Remark I.2.25. *At first sight the posterior version of the risk and the averaged version are different. An application of the towering property*[3] *saves us from this tedious distinction. Indeed, it suffices to consider the conditional risk for fixed $\mathscr{D}$. A result that holds almost surely for the posterior risk will then transfer to the risk. We see this since the almost sure result holds for any $\mathscr{D}$ with nonzero probability, so that integrating over the randomness of $\mathscr{D}$ does not impact the conclusion.*

Example I.2.26. *Inference of a signal x^* from observations $\boldsymbol{y} = \mathscr{D}$ all admitting a probability distribution function (pdf) means that the risk structure is:*

$$\begin{aligned}
\overline{\mathscr{R}}(\widehat{x};\mathcal{L}) &= \mathbb{E}_{X,\boldsymbol{Y}}\left[\mathcal{L}(\widehat{x}(\boldsymbol{Y}),X)\right] && \text{(I.2.27)}\\
&= \int_{\mathscr{Y}^{\times n}} \int_{-\infty}^{\infty} \mathcal{L}(\widehat{x}(\boldsymbol{y}),x)p(x \mid \boldsymbol{y})\,\mathrm{d}x\,p(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} && \text{(I.2.28)}\\
&= \mathbb{E}_{\boldsymbol{Y}}\left[\mathscr{R}(\widehat{x} \mid \boldsymbol{Y};\mathcal{L})\right]. && \text{(I.2.29)}
\end{aligned}$$

In general we would hope to find the best possible estimator wrt a loss. For a class of estimators $\mathscr{A}$ this will be the infimum over all estimators in the set. We name it **Bayes Action** and its respective risk the **Bayes risk**.

$$\widehat{\boldsymbol{x}}^{\star}_{\mathscr{A}}(\mathcal{L}) \coloneqq \arg\inf_{\widehat{\boldsymbol{x}}\in\mathscr{A}} \left\{\overline{\mathscr{R}}(\widehat{\boldsymbol{x}};\mathcal{L})\right\}, \quad \overline{\mathscr{R}}(\widehat{\boldsymbol{x}}^{\star}_{\mathscr{A}};\mathcal{L}) = \overline{\mathscr{R}}^{\star}(\mathcal{L}). \tag{I.2.30}$$

Remark I.2.31. *Notice that we do not focus on the performance in terms of prediction (e.g. supervised learning), but rather in recovering the original signal. This means that the principle of empirical risk minimization is not interesting.*

Remark I.2.32. *We further observe that for a finite prior, an application of Fubini's Theorem would allow to exchange the expectations for the risk. Then:*

$$\overline{\mathscr{R}}(\widehat{\boldsymbol{x}};\mathcal{L}) = \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{\mathscr{D}|\boldsymbol{X}}\left[\mathcal{L}(\widehat{\boldsymbol{x}}(\mathscr{D}),\boldsymbol{X})\right]\right] = \mathbb{E}_{\boldsymbol{X}}\left[\mathcal{R}(\widehat{\boldsymbol{x}}(\mathscr{D}),\boldsymbol{X};\mathcal{L})\right]. \tag{I.2.33}$$

*Where we have recognized the frequentist risk $\mathcal{R}(\cdot,\cdot;\cdot)$, that averages over the likelihood $\mathbb{P}[\mathscr{D} \mid \boldsymbol{X}]$. The best estimator wrt the average of frequentist risk is termed **Bayes rule**. It is explicitly dependent on the choice of the prior, but we will keep it fixed.*

In the Bayesian setting, allowing the signal to have a distribution, we find that a finite prior[4] *makes the frequentist Bayes decision rule coincide with the Bayesianist*

[3] $\mathbb{E}_{X,Y}[f(X,Y)] = \mathbb{E}_{Y}\left[\mathbb{E}_{X|Y}[f(X,Y=y)]\right]$

[4] this is true in most cases

Bayes action. Actually, this is the classical starting point for deriving our presentation of the risk. One starts from the frequentist risk. Adding a prior, the risk $\overline{\mathscr{R}}$ can be defined, and it will depend on the prior. Assuming that the conditionals can be defined and that the distributions are "nice", the posterior risk minimized almost surely will minimize the risk, as discussed previously.

Given the above discussion, we speak freely of Bayes decision rule and Bayes action, as they are the same. Precisely, the Bayes rule is obtained by taking the Bayes action for each particular $\mathscr{D}$. In the framework of inference, emphasis is placed on using only one realization of $\mathscr{D}$ to perform inference.

Another criterion for frequentist risks is the minmax criterion, by which one selects:

$$\mathcal{R}^{\star}_{\mathsf{Mm}} := \inf_{\widehat{\boldsymbol{x}}} \sup_{\boldsymbol{x}^* \in \mathscr{X}} \mathcal{R}(\widehat{\boldsymbol{x}}, \boldsymbol{x}^*; \mathcal{L}). \tag{I.2.34}$$

In words, it is the estimator that achieves the smallest maximum frequentist risk, so it is best in the worst case. We term this arg inf sup the **minimax optimal estimator** accordingly.

Remark I.2.35. *While the Bayes approach has the flavour of an average-case performance analysis, the minimax approach is meant to be worst-case. In this document, we are interested in typical behaviors, so we will keep the discussion simple for the latter. In the next statements we argue that the average-case results will still give information about the worst-case scenario.*

Definition I.2.36 (Least favorable prior)**.** *A prior $\mathbb{P}_0$ determines the notion of risk starting from the frequentist setting. Then, a prior is termed least favorable for performing a task when its risk is always higher than any other choice. Mathematically for fixed $\mathcal{L}$:*

$$\overline{\mathscr{R}}_{\mathbb{P}_0}(\widehat{\boldsymbol{x}}_{\mathbb{P}_0}; \mathcal{L}) \geqslant \overline{\mathscr{R}}_{\mathbb{P}}(\widehat{\boldsymbol{x}}_{\mathbb{P}}; \mathcal{L}) \quad \forall \mathbb{P}, \mathbb{P}_0, \tag{I.2.37}$$

with $\widehat{\boldsymbol{x}}_{\mathbb{P}_0}, \widehat{\boldsymbol{x}}_{\mathbb{P}}$ chosen according to the prior as Bayes rules, and the risk depends on the prior chosen.

Proposition I.2.38. *Fix a loss $\mathcal{L}$. If $\mathbb{P}_0$ is a prior such that its associated Bayes rule $\widehat{\boldsymbol{x}}_{\mathbb{P}_0}(\cdot)$ satisfies*

$$\overline{\mathscr{R}}_{\mathbb{P}_0}(\widehat{\boldsymbol{x}}_{\mathbb{P}_0}; \mathcal{L}) = \int \mathcal{R}(\widehat{\boldsymbol{x}}_{\mathbb{P}_0}, \boldsymbol{x}^*; \mathcal{L}) \, \mathrm{d}\mathbb{P}_0[\boldsymbol{x}^*] = \sup_{\boldsymbol{x}^*} \mathcal{R}(\widehat{\boldsymbol{x}}_{\mathbb{P}_0}, \boldsymbol{x}^*; \mathcal{L}) \tag{I.2.39}$$

then:

1. *it is minimax*
2. *if it is the unique Bayes estimator it is unique minimax*
3. *$\mathbb{P}_0$ is least favorable*

Proof. **(Claim #1)** Let $\widehat{\boldsymbol{x}} \neq \widehat{\boldsymbol{x}}_{\mathbb{P}_0}$. Then:

$$\sup_{\boldsymbol{x}^*} \mathcal{R}(\widehat{\boldsymbol{x}}_{\mathbb{P}}, \boldsymbol{x}^*; \mathcal{L}) \geqslant \int \mathcal{R}(\widehat{\boldsymbol{x}}, \boldsymbol{x}^*; \mathcal{L}) \, \mathrm{d}\mathbb{P}_0 \boldsymbol{x}^* \geqslant \int \mathcal{R}(\widehat{\boldsymbol{x}}_{\mathbb{P}}, \boldsymbol{x}^*) \, \mathrm{d}\mathbb{P}_0(\boldsymbol{x}^*) = \sup_{\boldsymbol{x}^*} \mathcal{R}(\widehat{\boldsymbol{x}}_{\mathbb{P}_0}; \mathcal{L}). \tag{I.2.40}$$

Accordingly, $\widehat{\boldsymbol{x}}_{\mathbb{P}_0}$ is minimax.
(Claim #2) Uniqueness of Bayes risk implies that $\geqslant$ above in the second inequality becomes $>$. Then the estimator is the unique one that achieves maximum risk, and is the unique minimax.
(Claim #3) Consider $\mathbb{P} \neq \mathbb{P}_0$, with associated Bayes estimator $\widehat{\boldsymbol{x}}_{\mathbb{P}}$. The risks are such that:

$$\overline{\mathscr{R}}(\widehat{\boldsymbol{x}}_{\mathbb{P}}; \mathcal{L}) = \int \mathcal{R}(\widehat{\boldsymbol{x}}_{\mathbb{P}}, \boldsymbol{x}^*) \, \mathrm{d}\mathbb{P}(\boldsymbol{x}^*) \leqslant \int \mathcal{R}(\widehat{\boldsymbol{x}}_{\mathbb{P}_0}, \boldsymbol{x}^*) \, \mathrm{d}\mathbb{P}(\boldsymbol{x}^*) \leqslant \sup_{\boldsymbol{x}^*} \overline{\mathscr{R}}_{\mathbb{P}_0}(\widehat{\boldsymbol{x}}_{\mathbb{P}_0}, \boldsymbol{x}^*). \tag{I.2.41}$$

In particular, the first inequality holds since $\widehat{\boldsymbol{x}}_{\mathbb{P}}$ is the Bayes rule wrt $\mathbb{P}$. □

By construction, the risk and the Bayes decision rule depend on the choice of $\mathcal{L}(\cdot,\cdot)$. In some cases, it is possible to derive a closed form of this minimum, as the next Proposition shows.

Proposition I.2.42 (Best estimators for different error functions)**.** *Go back to Example I.2.18, then:*

1. *the Minimum Mean Squared Error (MMSE) estimator is* $\widehat{\boldsymbol{x}}_{\mathsf{MMSE}}(\mathscr{D}) = \mathbb{E}_{\boldsymbol{X}|\mathscr{D}}[\boldsymbol{X}]$,
2. *the Minimum Mean Average Error (MMAE) estimator is* $\widehat{\boldsymbol{x}}_{\mathsf{MMAE}}(\mathscr{D}) = \mathsf{med}_{\boldsymbol{X}|\mathscr{D}}(\boldsymbol{X})$,
3. *the Minimum Error Counter (MEC) estimator is the maximum of the marginal for each entry* $j \in [d]$*, namely:*

$$\widehat{\boldsymbol{x}}_{\mathsf{MEC}}(\mathscr{D}) = \begin{bmatrix} \arg\max \mu_1(x_1 \mid \mathscr{D}) \\ \vdots \\ \arg\max \mu_d(x_d \mid \mathscr{D}) \end{bmatrix} \quad \mu_j(x_j \mid \mathscr{D}) \coloneqq \int \mathbb{P}\left[\boldsymbol{X} \mid \mathscr{D}\right] \prod_{l \neq j} \mathrm{d}x_l \quad \forall j \in [d]. \tag{I.2.43}$$

namely, it is an entry-wise MAP.

Proof. Consider Equation I.2.24, then we can minimize the posterior risk. The strategy is common to the three proofs, and consists in evaluating the gradient vector $\nabla_{\widehat{\boldsymbol{x}}}\mathscr{R} \in \mathbb{R}^d$ and setting it to zero entry-wise. This will be sufficient to find a minimum since the losses are convex in $\widehat{\boldsymbol{x}}$.
(Claim 1**)**

$$\nabla_{\widehat{\boldsymbol{x}}}\mathscr{R}(\widehat{\boldsymbol{x}} \mid \mathscr{D}; \mathsf{MSE}) = \frac{2}{d}\mathbb{E}_{\boldsymbol{X}|\mathscr{D}}\left[\widehat{\boldsymbol{x}}(\mathscr{D}) - \boldsymbol{X}\right] = \boldsymbol{0} \iff \widehat{\boldsymbol{x}}(\mathscr{D}) = \mathbb{E}_{\boldsymbol{X}|\mathscr{D}}\left[\boldsymbol{X}\right]. \tag{I.2.44}$$

In the last step, we used the fact that for fixed $\mathscr{D}$ the estimator is constant.
(Claim 2**)** Following the same fashion we first express the posterior risk

$$\nabla_{\widehat{\boldsymbol{x}}}\mathscr{R}(\widehat{\boldsymbol{x}} \mid \mathscr{D}; \mathsf{MAE}) \propto \nabla_{\widehat{\boldsymbol{x}}} \left(\mathbb{E}_{\boldsymbol{X}|\mathscr{D}}\left[\mathcal{L}(\widehat{\boldsymbol{x}}(\mathscr{D}), \boldsymbol{X})\mathbb{1}_{\widehat{\boldsymbol{x}} \geqslant \boldsymbol{X}}\right] + \mathbb{E}_{\boldsymbol{X}|\mathscr{D}}\left[\mathcal{L}(\widehat{\boldsymbol{x}}(\mathscr{D}), \boldsymbol{X})\mathbb{1}_{\widehat{\boldsymbol{x}} \geqslant \boldsymbol{X}}\right]\right) \tag{I.2.45}$$

$$= \nabla_{\widehat{\boldsymbol{x}}} \left(\int_{-\infty}^{\widehat{\boldsymbol{x}}} \widehat{\boldsymbol{x}} - \boldsymbol{x} \, \mathrm{d}\mathbb{P}\left[\boldsymbol{x}|\mathscr{D}\right] + \int_{\widehat{\boldsymbol{x}}}^{\infty} \boldsymbol{x} - \widehat{\boldsymbol{x}} \, \mathrm{d}\mathbb{P}\left[\boldsymbol{x} \mid \mathscr{D}\right] \right) \tag{I.2.46}$$

$$= \int_{-\infty}^{\widehat{\boldsymbol{x}}} \mathrm{d}\mathbb{P}\left[\boldsymbol{x}|\mathscr{D}\right] - \int_{\widehat{\boldsymbol{x}}}^{\infty} \mathrm{d}\mathbb{P}\left[\boldsymbol{x} \mid \mathscr{D}\right] = 0 \tag{I.2.47}$$

Where in the first passage we have ignored the normalization and in the second passage we have used Leibniz's Integral rule (see Subsec. A.1.1 for a discussion and references). Then, the condition implies, together with $\int_{-\infty}^{\infty} \mathbb{P}\left[\boldsymbol{x} \mid \mathscr{D}\right] = 1$ that $\boldsymbol{x}(\mathscr{D}) = \mathsf{med}_{\boldsymbol{X}|\mathscr{D}}(\boldsymbol{X})$, the median.
(Claim 3**)** The indicator function is not differentiable, so we resort to inspecting the form of the posterior risk.

$$\mathscr{R}(\widehat{\boldsymbol{x}} \mid \mathscr{D}; \mathbb{EC}) \propto \sum_{x} \mathbb{P}\left[\boldsymbol{x} \mid \mathscr{D}\right] \sum_{j=1}^{d} \mathbb{1}_{\widehat{x}_j \neq x_j} = \underbrace{\sum_{\boldsymbol{x}} \mathbb{P}\left[\boldsymbol{x} \mid \mathscr{D}\right]}_{=1} - \sum_{\boldsymbol{x}} \mathbb{P}\left[\boldsymbol{x} \mid \mathscr{D}\right] \sum_{j=1}^{d} \mathbb{1}(x_j = \widehat{x}_j). \tag{I.2.48}$$

Minimizing such function means bringing the second term as close as possible to 1. The choice of entry-wise maximums of the marginals is sound, since the weight of probability is the highest at the maximizer, and the test of the indicator is entry-by-entry. □

Remark I.2.49. *For the Error counter, the overlap and the MSE the optimal estimator depends only on the marginals when the features are iid.*

Remark I.2.50. *The expert reader might object that we justified the introduction of a loss to avoid providing point estimates, to later discuss the best estimators, which are still answers based on points. Moreover, one could argue that the MMSE estimator and the MAP coincide for distributions that are jointly Gaussian. All of these comments are right, bt we also remark that it is standard to approach problems in Machine Learning/inference with a notion of performance (a loss) and that depending on the situation one estimator or the other might be more useful. Additionally, this coincidence applies to the choice of square loss only, and we eventually find marginals. The decision theoretic approach bears more generality than the simple point estimate.*

In general, such a computation presents two hurdles:

1. the minimization is potentially difficult (avoided in Prop. I.2.42),
2. the posterior is hard to evaluate, as before (not avoidable),
3. there is no knowledge of the true distributions (avoided below).

Bayes-Optimal Inference

It is now important to remark that we are not discussing the dependence on the choice of the prior. Additionally, to derive sensible bounds, we will focus on the Bayes-Optimal setting, which provides the largest set of information to the statistician.

Bayes-Optimality

With Bayes-Optimal inference we mean that all distributions are available to the statistician: the true prior, the true likelihood and the true posterior. In other words, the Data Generating Process (DGP) is known.

Intuitively, this would give a lower bound on any other setting, since it is the best possible scenario in which the statistician knows *everything but the solution.* Upon knowledge of the data generating process, Bayes-Optimal inference can be performed, and a statistician has access to an expression of the posterior distribution $\mathbb{P}[\boldsymbol{X} \mid \mathscr{D}]$, potentially not in closed form.
The easiest instance of the non-Bayes-Optimal setting is when the true family of distributions is known, but not up to the true parameters. Assuming the parametrizations are all stored in a vector $\boldsymbol{\vartheta}_*$, the statistician knows that the randomness is from that model, but is clueless about the true value of $\boldsymbol{\vartheta}_*$ and must resort to a guess $\boldsymbol{\vartheta}$. This situation is termed **mismatched**. Different constructions can be designed, depending on how much information is made available. Adding degrees of freedom to the problem of the statistician, the task is made harder (more objects to infer). The Bayes-Optimal setting is thus the least difficult edge-case, for which conclusions about the general framework can be made. In the next subsection, we specifically describe our main scenario of interest, where Bayes-Optimal problems will be crucial to establish bounds for the general method of Bayesian inference.

Example I.2.51 (Bayes-Optimal and non-Bayes-Optimal)**.** *A Bayes-optimal problem could be knowing that* $\mathbb{P}[\cdot;\boldsymbol{\vartheta}_*] = \mathcal{N}(\mu_*, \sigma_*^2)$*, while a partially non-Bayes Optimal problem is the unknown variance/known mean analogue where* $\mathcal{N}(\mu_*, \sigma^2)$*.*

I.3 Average Case Efficiency in High-Dimensions

In modern inference problems the amount of available data is large in both directions: samples and features. This raises a huge concern in resources needed to answer even basic questions about the phenomenon under consideration. In mathematical terms, the scenario justifies observing limiting properties of a model, with $n \to \infty, d \to \infty$ but with **aspect ratio fixed** $\delta = \frac{n}{d} \in \Theta(1)$. This loosely means that we imagine that the dataset has many rows and many columns (so many that their precise number is negligible, and there will be dimensionless properties that

kick in at some limit), but that their proportion is kept fixed (so that the dimensionless properties will depend on such δ). A good interpretation is seeing δ as an indicator of complexity (equivalently of information) for each measurement. The higher it is, the least (respectively more) complex (resp., informative) the model will be. Intuitively in the most extreme example, it should be easier to infer a one dimensional vector from one million measurements ($d = 1, n = 10^6$) rather than a 1 million dimensional vector from one measurement ($d = 10^6, n = 1$).

The key observation regarding this choice is that we basically get rid of (n, d), which is useful in two ways:

- some parameters are lost
- sizes inspected will be by construction large and in accordance with the current state of affairs of applied problems.

In a real scenario, the limiting statements will be valid for large enough n with *adjustable reliability*, by the definition of limit. All will regard a Statistical Model $\mathfrak{M}$ (Def. I.1.5) and an associated Problem $\mathfrak{P}$ (Def. I.1.6). Additionally, we let $\mathfrak{P}$ be Bayes-Optimal, so that the fewest information is unknown and any real world scenario, where the source is not necessarily known, will be at best as good.

Statistical inference in these premises is broadly concerned with two questions. The statements are borrowed from (**zdeborovaStatisticalPhysicsInference2016**), and will be explained below. Both have been approached with concepts that have a long history in Statistical Physics.

Sufficient Information

From (**zdeborovaStatisticalPhysicsInference2016**).

"Under what conditions is the information contained in the observations **sufficient** for *satisfactory* **recovery** of the variables?"

Answers to this question are formulated with techniques that belong to **Statistics** and **Information Theory**. The main objective is providing a formalism by which it can be precisely stated when information about a phenomenon is enough to derive a meaningful description of it. We give two precise notions below that are the canonical starting points.

Remark I.3.1. *Notice that the overlap is different from losses as a principle, since we wish it to be high (not low like losses), the statements must be adjusted accordingly. Moreover, it is not in general true that losses behave as below, but it is also true that for sufficiently nice distributions (e.g. spherical, Gaussian) this is the case.*

Problem I.3.2 (Weak Recovery)**.** *Find a sufficient condition on $\mathfrak{P}$ in terms of its parameters such that the signal of $\mathfrak{M}$ is estimated with non-trivial success wrt a given loss. Namely, find an estimator $\widehat{\boldsymbol{x}}(\mathscr{D})$ such that there exists a constant $c < c_{\mathsf{triv}}$ for which:*

$$\lim_{n\to\infty} \mathbb{P}\left[\mathcal{L}(\widehat{\boldsymbol{x}}(\mathscr{D}), \boldsymbol{x}^*) \leqslant (c + o(1))n\right] = 1, \tag{I.3.3}$$

where c_{triv} is the performance density of a trivial estimator, to be understood as the number of errors-per-sample. In other terms, find a procedure that beats random guessing by a vanishing but nonzero[5] amount almost surely wrt the assumed randomness.

Problem I.3.4 (Strong Recovery)**.** *Find a sufficient condition on $\mathfrak{P}$ such that the signal of $\mathfrak{M}$ is estimated correctly wrt a given loss. Namely, find an estimator $\widehat{\boldsymbol{x}}(\mathscr{D})$ such that:*

$$\lim_{n\to\infty} \mathbb{P}\left[\mathcal{L}(\widehat{\boldsymbol{x}}(\mathscr{D}), \boldsymbol{x}^*) = o(n)\right] = 1. \tag{I.3.5}$$

In simple words, find an estimator that has vanishing loss at the limit. In some references, this is termed weak consistency/almost exact recovery, with strong consistency/exact recovery being 0 instead of $o(n)$.

[5] $o(1)\, n = o(n)$

Before continuing with the second set of Problems, it is crucial to observe that we are not at all restricting ourselves in terms of number of operations to perform. The question is merely a feasibility question: given a setting, find an answer within a time possibly scaling very fast in n. Realistically, a problem is solvable in two *phases*:

- when it is simply solvable,
- when it is efficiently possible to do so.

This motivates the introduction of a computationally bounded requirement for solvability.

Computational Efficiency

From (**zdeborovaStatisticalPhysicsInference2016**).

"Can the inference be done in an **algorithmically** *efficient* way? What are the **optimal** algorithms for this task?"

To inspect this problem, tools from **Computer Science** need to be considered, in the sense that we must design procedures (algorithms) that are efficient wrt the relevant size of the problem. For simplicity, we will refer to efficient solvability with $\mathsf{poly}(n)$ running time. This choice amounts to having a single distinction[6]:

- polyonomial algorithms $\mathcal{A}$
- non-polynomial procedures.

Remark I.3.6. *Despite being a binary labeling (i.e. polynomial vs non-polynomial), this is already sufficient to answer industry related tasks. A super-polynomial procedure requires times that for moderate sized n are already unfeasible. Even worse, exponential-times procedures quickly get to the point that one computation per second requires more time than the age of the universe. From a practical perspective, a complex classification of hierarchies is not needed.*

Armed with this, we can define the Problems analogous to Probs. I.3.2, I.3.4.

Problem I.3.7 (Weak efficient Recovery)**.** *Devise an estimator $\widehat{\boldsymbol{x}}$ that solves Prob. I.3.2 in efficient polynomial time.*

Problem I.3.8 (Strong efficient Recovery)**.** *Devise an estimator that solves Prob. I.3.4 in efficient polynomial time.*

To find these estimators, it might be that an explicit set of steps is to be taken. In this case, we can think of a proper algorithm $\mathcal{A}$ providing the solution. When the estimator is just the result of a single calculation, the notion of algorithm is somewhat vacuous, but one can still check the computational complexity of it in terms of the size of the problem. If the single operation has a cost that is non-polynomial in size, then the imaginary procedure is non-efficient.

Remark I.3.9. *We find three other equivalent notions of recovery (see (***reevesAllorNothingPhenomenonSparse**
guWeakRecoveryThreshold2023*) and (***abbeCommunityDetectionStochastic2022***)).*
For weak recovery, we can consider:

$$\lim_{n\to\infty} \sup \frac{1}{n}\mathbb{E}\left[\mathcal{L}(\widehat{\boldsymbol{X}}, \boldsymbol{X}^*)\right] \leqslant c_{\mathsf{triv}} - \epsilon \quad \textit{for some } \epsilon > 0, \tag{I.3.10}$$

or

$$\lim_{n\to\infty} \mathbb{P}\left[\mathcal{L}(\widehat{\boldsymbol{x}}, \boldsymbol{x}^*) \leqslant (c_{\mathsf{triv}} - \Omega(1))n\right] = 1 - o(1), \tag{I.3.11}$$

or

$$\lim_{n\to\infty} \sup \frac{\mathbb{E}\left[\mathcal{L}(\widehat{\boldsymbol{X}}(\mathscr{D}), \boldsymbol{X}^*)\right]}{\mathcal{L}(\widehat{\boldsymbol{x}}_{\mathsf{triv}}, \boldsymbol{x}^*)} < 1. \tag{I.3.12}$$

[6] in some cases, a richer collection of results can be derived by describing how algorithms behave in a finer spectrum of complexities.

All of the above encode the same notion: at the limit, the estimator must be at least slightly better than random guessing.
For strong recovery:

$$\lim_{n\to\infty} \sup \frac{1}{n}\mathbb{E}\left[\mathcal{L}(\widehat{\boldsymbol{X}}, \boldsymbol{X}^*)\right] = 0, \tag{I.3.13}$$

or

$$\lim_{n\to\infty} \mathbb{P}\left[\mathcal{L}(\widehat{\boldsymbol{x}}, \boldsymbol{x}^*) \leqslant o\left(1\right) n\right] = 1 - o\left(1\right), \tag{I.3.14}$$

or

$$\lim_{n\to\infty} \sup \frac{\mathbb{E}\left[\mathcal{L}(\widehat{\boldsymbol{X}}(\mathscr{D}), \boldsymbol{X}^*)\right]}{\mathcal{L}(\widehat{\boldsymbol{x}}_{\mathsf{triv}}, \boldsymbol{x}^*)} = 0. \tag{I.3.15}$$

*All of the above are the same notion: when norms of the vectors are bounded by the fact that $\langle \widehat{\boldsymbol{x}}(\mathscr{D}), \boldsymbol{x}^* \rangle \leqslant \|\widehat{\boldsymbol{x}}(\mathscr{D})\| \, \|\boldsymbol{x}^*\|$ the overlap is non-vanishing and well-defined (***arousOnlineStochasticGradient2021***). More importantly, in the model we inspect in Chapter V a random guess is vanishing in performance as $\sim d^{-\frac{1}{2}}$ by classic high dimensional probability arguments (see (**vershyninHighDimensionalProbabilityIntroduction20**
Moreover, the nicest formulation is the one with expectations, which in most of the cases will be expressed in terms of overlaps, requiring a trivial "flip" of all these definitions.*

The picture becomes fairly easier for specific choices of the loss. By Proposition I.2.42, we know that for the MSE the posterior expectation minimizes the posterior risk almost surely, and thus minimizes the risk. A closed form of its expression would be sufficient to evaluate the behavior of the MMSE of recovering the true signal. In this case it suffices to understand the behavior of the following limit:

$$\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{X}}\left[(\boldsymbol{X} - \mathbb{E}_{\boldsymbol{X}|\mathscr{D}}\left[\boldsymbol{X} \mid \mathscr{D}\right])^2\right]. \tag{I.3.16}$$

In general, it is not guaranteed that the conditional expectation is computable, and one might resort to approximations, leading to the side-question of quantifying how much far away in performance an estimator can get.

I.3.1 What is a Statistical-to-Computational Gap?

From the above notions, we can draw some immediate implications.

Fact I.3.17. *It trivially holds that:*

- *strong recovery implies weak recovery*
- *strong efficient recovery implies weak efficient recovery*
- *weak efficient recovery implies weak recovery.*

In an (almost) general setting, a statistical problem $\mathfrak{P}$ linked to a statistical model $\mathfrak{M}$ will depend on a set of (hyper) parameters $(n, d, \boldsymbol{\vartheta}) \in \mathbb{R}^{1\times 1\times H}$. Letting $n, d \to \infty$ means that the picture in the description of the high dimensional limit will depend on choices of $\boldsymbol{\vartheta}$, which appear in general in a Cartesian plane $\mathbb{R}^{H+1}$. The above Facts then define nested regions of the space with associated notions of hardness/solvability and algorithmic hardness/efficient solvability. Given that the three settings are in an ordered relation, the appearence is that of a matriosk: the impossible phase is contained in the hard phase which is contained in the easy phase, or viceversa, depending on the meaning of *is contained in*. To make this easier, we identify separate regions in which the problem is exclusively impossible, hard or easy, leading to a partition of the parameter space. The general picture looks like the diagram of Figure I.3. To give a concrete example, we will take the simplest case in which there is only one hyperparameter. Here $H = 0$ and by our construction the only parameter is $\delta = \frac{n}{d} \in \Theta\left(1\right)$, but in general there could be other examples, such as the SNR of the channel, which is completely analogous. A diagram of their behavior is Figure I.4.

With this in hand, it is reasonable to discuss the existance of the three phases for a given problem, as well as their size and position on the phase diagram. From

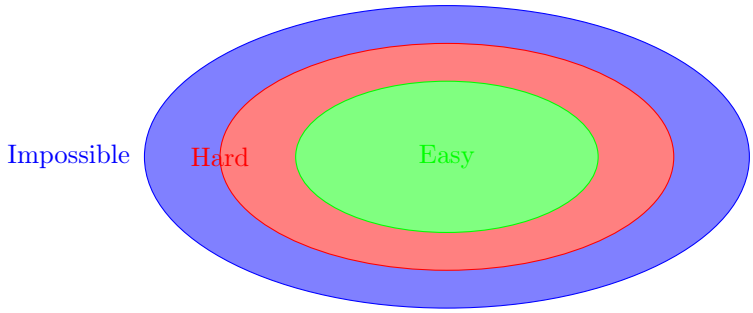


Figure I.3: Multidimensional cartoon of the phases of an inference problem
Fix a problem. Here by Impossible we mean "instances of the problem in this ring have no solution to the problem", while Hard is understood as "instances have a non-polynomial procedure that returns a solution" and Easy "instances can be solved efficiently". If we saw the structure as a matriosk, the perspective above would change to instances with efficient, polynomial, non-polynomial solution, instances with non-polynomial or polynomial solution and exclusively polynomial instances.

Figure I.4: Signal-to-Noise Ratio-like Phase Diagram
Intuitively, the SNR is a quantity that indicates how *strong* the signal magnitude is wrt to the noise of the channel. The larger it is, the more we expect the problem is easy to solve. Given this, the general phase diagram for problems that depend only on the SNR will be line split into three regions. In the first, we fail because the observations appear as if they are noise. In the second, we succeed inefficiently, because the solution exists but can only be found by e.g. exhaustive search. In the rightmost setting, the strength of the signal is large enough as to be isolated in efficient time.

a practical standpoint, the interesting phase is the Hard one. Indeed, unsolvable problems are just not interesting, while a solvable problem bears the question of the worthiness of finding its solution: if too much resources are required, one might as well ignore it. The hard phase describes a scenario in which it is not advised to aim for retrieval, despite it being possible. Whenever it exists, we speak of a Statistical to Computational (S-to-C) gap, meaning that there exists configurations in $\mathbb{R}^{H+1}$ of the parameter space by which the problem is information-theoretically (IT) solvable but not algorithmically solvable wrt its size. More generally, characterizing the whole space of parameters allows to *close* a problem, meaning that in any case, the statistician will know its basic phenomenology.

Example I.3.18 (Phase retrieval)**.** *Assume you are given a dataset $\mathscr{D}$ which comes from a channel with transmission mechanism:*

$$y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x}^* \rangle| + \epsilon_i \quad \forall i \in [n], \tag{I.3.19}$$

with $\epsilon_i \sim \mathcal{N}(0, \Delta)$ Gaussian noise. In literature, this is a Phase Retrieval problem. It is widely studied because it presents the difficulty that it is ***symmetric****: information about the sign of $\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle$ is lost in the channel. For this reason, the signal can be recovered only up to a global sign change, since $|\langle \boldsymbol{a}, \boldsymbol{b} \rangle| = |\langle \boldsymbol{a}, -\boldsymbol{b} \rangle|$ for any pair of vectors $(\boldsymbol{a}, \boldsymbol{b})$.*
From a mathematical standpoint, it is an interesting question. In some sense it is also the prototypical example of a GLM (Exm. I.1.12) that is not linear regression. If we go at the high dimensional limit where $n, d \to \infty, \delta \in \Theta(1)$, the parameters are (δ, Δ): the complexity and the noise. Their ratio could be seen as a SNR, but we will rather focus on the case where also $\Delta \to 0^+$, termed noiseless phase retrieval. Then, the phase diagram could be regarded as that of Figure I.4, with δ in place of SNR*. In (***mondelliFundamentalLimitsWeak2018***), the authors find that $\delta_{\mathsf{IT}} \geqslant \frac{1}{2}$ and $\delta_{\mathsf{alg}} = \delta_{\mathsf{IT}}$. This means that if $n > 2d$, i.e. the observations are*

twice the number of features, one can expect that at high dimensions the signal can be recovered, efficiently. In this particular case, there is no gap: as soon as it is possible to retrieve the signal, an algorithm can in polynomial time. This model will be generalized and discussed further in Chapter V.

Further References

We will exclusively focus on the mathematical formalism, but it is worth mentioning why understanding the peculiarities Phase Retrieval is crucial in applications.
From an experimental point of view, it is seen as the task of solving the Phase problem (**taylorPhaseProblem2003**), an impediment that often appears in real scenarios. Some measurement tools only give an estimate of the intensity of some quantity, neglecting its direction. In most cases, this direction is fundamental to improve the quality of the analysis. Examples include but are not limited to:

- optical systems (**fienupPhaseretrievalAlgorithmsComplicated1993**; **kristPhaseretrievalAnalysisPre1995**)
- crystallography (**hauptmanPhaseProblemNeutron2003**)
- electron crystallography (**hendersonStructurePurpleMembrane1986**; **dorsetDirectPhasingProtein1996**; **dorsetDirectPhaseDetermination1997**)

Chapter II

The Perspective of Statistical Physics

To overcome the hurdles of the methods presented earlier, many approaches are considered. The focus of this document being Statistical Physics, we have reviewed the modern setting of Statistical inference to introduce the final motivation to study techniques that were originally intended for complex systems.

This Chapter is a collection of basic results about the Statistical Physics viewpoint of Inference. In Section II.1, we will introduce the physical terminology and a good formalization. Section II.2 develops further the connections between canonical models in Thermodynamics and Learning problems, with the very peculiar notion of quenched disorder playing the key role of distinguishing types of randomness. Sections II.3 and II.5 are made to introduce the reader to two paradigms of analysis which serve for the purpose of creating a narrative and a common playground to analyze models, while Section II.4 in the middle of the two is more technical, and provides the reader with two fundamental tools used throughout the document in the background. Lastly, Section II.6 returns to discuss how much Statistical Physics and Inference are essentially the same field with two different vocabularies. At the cost of reading research in two fields, it is immediate to notice that Phase transitions are a relevant question for learning problems (Subsec. II.6.1), and that they greatly benefit from the Mathematics of objects that were formalized more in Information Theory (Subsec. II.6.2). In this last matter, it is also very important to stress that this is not a comprehensive guide to how much the fields are interconnected when asking the right questions, but rather a selection of results. Hopefully, the references cited will give more guidance to the interested reader. As a matter of fact, the connections would have required an independent, dense book on its own.

Further References

On the Philosophical approach side, the connection between Bayesian inference and Physics is in its nature a discussion topic. A classic reference is (**cousinsWhyIsnEvery1995**).

II.1 Boltzmann-like distributions

The first observation is that many (see Rem. II.1.3) posterior distributions allow for a description in terms of a Boltzmann canonical distribution. Indeed, from Bayes' Theorem:

$$\mathbb{P}[\boldsymbol{X} \mid \mathscr{D}] = \frac{\mathbb{P}[\mathscr{D} \mid \boldsymbol{X}]\,\mathbb{P}[\boldsymbol{X}]}{\mathbb{P}[\mathscr{D}]} = \frac{1}{\mathcal{Z}(\mathscr{D};\beta,\mathscr{D})} e^{-\beta \mathscr{H}(\boldsymbol{X};\mathscr{D})}, \tag{II.1.1}$$

where $\beta \in \mathbb{R}_+$ is an additional parameter. In general, one needs to place into $\mathcal{Z}(\mathscr{D};\beta)$ all terms not depending on $\boldsymbol{X}$ and into the Hamiltonian all terms dependent on $\boldsymbol{X}$.

This is a rule of thumb, by the simple identity $x = e^{\ln x}$ at $\beta = 1$ the Hamiltonian

$$\mathscr{H}(\boldsymbol{X}; \mathscr{D}) = \ln\left[\frac{1}{\mathbb{P}\left[\mathscr{D}|\boldsymbol{X}\right]\mathbb{P}\left[X\right]}\right] \tag{II.1.2}$$

is an equivalent description of the posterior. While this is just a rearrangement of the terms, it highlights in a more explicit sense that some macroscopic conclusions can be made. Among all, the explicit Partition Function is the Cumulative Generating Function (CGF) of the randomness, and in the thermodynamic limit results in closed form formulas for the average energy and any moment.

Remark II.1.3. *This choice might look arbitrary. For $\beta = 1$, it is a mathematical identity. For $\beta \neq 1$, we find a relaxation of the problem to a more general model that was largely studied in Physics, and comes with a nice set of tools to analyze it. This trick being itself a topic of discussion, we remind the reader that on top of this we took a Bayesian perspective, and most of the times Gaussian noise.*

In the next example we show how this emerges naturally for a problem of inference.

Example II.1.4 (Bridge from Bayesian Theory to Statistical Physics I)**.** *Consider the model $y_i = \varphi(\boldsymbol{x}^*) + \epsilon_i$, where $\varphi(\cdot)$ is a deterministic transformation, and noise is Gaussian with variance Δ. It might be tempting to observe that Bayes' rule takes the form:*

$$\mathbb{P}(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{\mathbb{P}(\boldsymbol{x})\mathbb{P}(\boldsymbol{y} \mid \boldsymbol{x})}{\mathbb{P}(\boldsymbol{y})} = \underbrace{\frac{1}{\mathbb{P}(\boldsymbol{y})}}_{:=\mathcal{Z}(\boldsymbol{y})} \exp\left\{\underbrace{\ln[\mathbb{P}(\boldsymbol{x})\mathbb{P}(\boldsymbol{y} \mid \boldsymbol{x})]}_{=-\beta\mathscr{H}}\right\} = \mathbb{P}_{\mathsf{Boltz},\boldsymbol{y}}(\boldsymbol{x}; \beta = 1). \tag{II.1.5}$$

*However, a slightly different parametrization is nicer for computations (***krzakalaStatisticalPhysicsMet***
By the assumption on noise and the fact that the channel maps to scalar values independently for each observation, we can factorize the likelihood as a product of Gaussians. Then, aim to completely isolate $\boldsymbol{x}$ and obtain:*

$$\mathbb{P}(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{1}{\mathbb{P}(\boldsymbol{y})}\mathbb{P}[\boldsymbol{x}]\frac{\prod_i e^{-\frac{(y_i - \varphi(\boldsymbol{x}))^2}{2\Delta}}}{\sqrt{2\pi\Delta}} = \underbrace{\frac{e^{-\frac{\sum_i y_i^2}{2\Delta}}}{(\sqrt{2\pi\Delta})^n \mathbb{P}_Y(\boldsymbol{y})}}_{:=\mathcal{Z}(\boldsymbol{y})}\mathbb{P}[\boldsymbol{x}]e^{\sum_i -\frac{\varphi(\boldsymbol{x})^2}{2\Delta} + \frac{\varphi(\boldsymbol{x})y_i}{\Delta}} \tag{II.1.6}$$

$$= \frac{\mathbb{P}(\boldsymbol{x})e^{\sum_i -\frac{\varphi(\boldsymbol{x})^2}{2\Delta} + \frac{\varphi(\boldsymbol{x})y_i}{\Delta}}}{\mathcal{Z}(\boldsymbol{y})}$$

Where we recognize the terms:

$$\mathcal{Z}(\boldsymbol{y}) = \frac{e^{-\frac{\sum_i y_i^2}{2\Delta}}}{(\sqrt{2\pi\Delta})^n \mathbb{P}[\boldsymbol{y}]} = \int \frac{e^{-\frac{\sum_i y_i^2}{2\Delta}}}{(\sqrt{2\pi\Delta})^n \mathbb{P}[\boldsymbol{y}]}\mathbb{P}[\boldsymbol{x}]\,\mathrm{d}\boldsymbol{x} = \int e^{\frac{-\varphi(\boldsymbol{x})^2}{2\Delta} + \frac{\sum_i y_i \varphi(\boldsymbol{x})}{\Delta}}\mathbb{P}[\boldsymbol{x}]\,\mathrm{d}\boldsymbol{x} \tag{II.1.7}$$

$$\beta = 1$$

$$\mathscr{H}(\boldsymbol{x}) = -\frac{\varphi(\boldsymbol{x})^2}{2\Delta} + \sum_{i=1}^{n}\frac{\sum_i y_i \varphi(\boldsymbol{x})}{\Delta}.$$

In particular, the term $\mathcal{Z}(\boldsymbol{y})$ can be read as the sum over possible configurations $\boldsymbol{x}$ of the numerator.
A little more work shows that this holds also for GLMs like those of Example I.1.12. A function $\varphi(\cdot; \mathbf{A})$ with $\mathbf{A}$ random induces a posterior:

$$\mathbb{P}\left[\boldsymbol{X} \mid \mathscr{D}\right] = \frac{\mathbb{P}\left[\mathscr{D} \mid \boldsymbol{X}\right]\mathbb{P}\left[\boldsymbol{X}\right]}{\mathbb{P}\left[\mathscr{D}\right]} = \frac{\mathbb{P}\left[(\boldsymbol{y}, \mathbf{A}) \mid \boldsymbol{X}\right]\mathbb{P}\left[\boldsymbol{X}\right]}{\mathbb{P}\left[(\boldsymbol{y}, \mathbf{A})\right]}, \tag{II.1.8}$$

where we have made the content of the dataset explicit. Given independence over the observations, the likelihood factorizes. Hence:

$$\mathbb{P}\left[y_i, \boldsymbol{a}_i \mid \boldsymbol{x}\right] = \mathbb{P}\left[y_i \mid \boldsymbol{a}_i, \boldsymbol{x}\right] \mathbb{P}\left[\boldsymbol{a}_i \mid \boldsymbol{x}\right] = \mathbb{P}\left[\boldsymbol{y}_i \mid \boldsymbol{a}_i, \boldsymbol{x}\right] \mathbb{P}\left[\boldsymbol{a}_i\right] \quad \forall i \in [n], \tag{II.1.9}$$

by independence of the compressing matrix and the signal. The representation follows by isolating $\boldsymbol{x}$, since $\mathbb{P}\left[\boldsymbol{y}_i \mid \boldsymbol{a}_i, \boldsymbol{x}\right]$ is a Gaussian density. Therefore, the partition function admits a more explicit expression as in Eqn. II.1.6:

$$\mathcal{Z}(\boldsymbol{y}, \mathbf{A}) = \frac{e^{-\frac{\sum_i y_i^2}{2\Delta}} \prod_{i=1}^n \mathbb{P}\left[\boldsymbol{a}_i\right]}{(\sqrt{2\pi\Delta})^n \mathbb{P}[\boldsymbol{y}, \mathbf{A}]} \tag{II.1.10}$$

Remark II.1.11 (Partition function as a Likelihood Ratio)**.** *It is remarkable that such a formalism is equivalent to asserting that:*

$$\mathcal{Z}(\mathscr{D}) = \frac{\mathbb{P}^{\mathsf{rand}}(\mathscr{D})}{\mathbb{P}(\mathscr{D})} \tag{II.1.12}$$

*Where random means e.g. a purely random noise $y_i \sim \mathcal{N}(0, \Delta)$ and random matrix vectors, and on the denominator we find the probability of $\mathscr{D}$ coming from the true data generating process. The observation is made in (***krzakalaStatisticalPhysicsMethods2021***), for the first part of Example II.1.4, with the second part as a natural extension. This has very interesting ties with other directions of study, focused on Likelihood ratios (***kuniskyHypothesisTestingLowdegree2020***). Comments in this matter are postponed to a future exploration.*

Remark II.1.13. *The connection with a general Bayesian problem can also be seen in the context of Graphical models, which we discuss in Chapter III.*

Example II.1.14. *We now provide a more Physical interpretation of the parametrization, inspired from the study of magnetic materials. In the simple case in which the dataset is made of independent observations that depend on a subset $\partial i \subset [d]$ of the signal entries, the posterior of the model can be safely written as:*

$$\mathbb{P}[\boldsymbol{x}|\boldsymbol{y}] = \frac{1}{\mathcal{Z}(\boldsymbol{y})} \exp\left\{\beta \sum_{i=1}^n \ln \mathbb{P}[\boldsymbol{y}_i | \{\boldsymbol{x}_j\}_{j\in\partial i}] + \beta \sum_{j=1}^d \ln \mathbb{P}[\boldsymbol{x}_j]\right\} = \frac{1}{\mathcal{Z}(\boldsymbol{y})} \exp\left\{-\beta \mathscr{H}(\boldsymbol{x}, \boldsymbol{y})\right\}. \tag{II.1.15}$$

where a uniform additional β factor was added. In a Statistical Physics model, the ***first term*** *inside the exponential would be the* ***interaction term****, and the* ***second*** *would be the* ***(local) magnetic field****. The marginals of the posterior are interpreted as* ***local magnetizations*** *of single spins at $\beta = 1$.*

We report one very important property of the β parametrization and then proceed with further comments about this construction.

Proposition II.1.16 (Low temperature minimum energy configuration convergence)**.** *Letting $p_\beta(\boldsymbol{X})$ be a canonical Boltzmann distribution with Hamiltonian $\mathscr{H}$ well-behaved (to be discussed in the proof), the following limits are true:*

1. $\lim_{\beta\to 0} p_\beta(\boldsymbol{X}) = \mathcal{U}nif(\mathscr{X})$
2. $\lim_{\beta\to\infty} p_\beta(\boldsymbol{X}) = \mathcal{U}nif(\{\boldsymbol{x}^\star = \arg\min \mathscr{H}(\boldsymbol{x})\})$
3. *in particular, the mean energy concentrates at the minimum for $\beta \to \infty$*

Proof. **(Claim** 1**)** Obvious, letting $\beta \to 0$ the weights become 1 for each $\boldsymbol{x} \in \mathscr{X}$. The partition function is the sum of the individual weights and the distribution is uniform.
(Claim 2**)** A tedious derivation shows that the Boltzmann probability distribution concentrates around the minimum for $\beta \to \infty$. Assuming that x is scalar and there

is only one minimizer $x^\star$, the steps are as follows:

$$\lim_{\beta\to\infty}\frac{1}{\mathcal{Z}(\beta)}e^{-\beta\mathscr{H}(x)} = \lim_{\beta\to\infty}\frac{e^{-\beta\mathscr{H}(x)}}{\sum\limits_{\{x'\}}e^{-\beta\mathscr{H}(x')}} = \lim_{\beta\to\infty}\frac{e^{-\beta\mathscr{H}(x)}}{\sum\limits_{\{x',x'\neq x^\star\}}e^{-\beta\mathscr{H}(x')}+e^{-\beta\mathscr{H}(x^\star)}} \tag{II.1.17}$$

$$= \lim_{\beta\to\infty}\frac{\frac{e^{-\beta\mathscr{H}(x)}}{e^{-\beta\mathscr{H}(x^\star)}}}{\frac{e^{-\beta\mathscr{H}(x^\star)}}{e^{-\beta\mathscr{H}(x^\star)}}+\frac{\sum\limits_{\{x',x'\neq x^\star\}}e^{-\beta\mathscr{H}(x')}}{e^{-\beta\mathscr{H}(x^\star)}}} \tag{II.1.18}$$

$$= \lim_{\beta\to\infty}\frac{e^{-\beta(\mathscr{H}(x)-\mathscr{H}(x^\star))}}{1+\sum\limits_{\{x',x'\neq x^\star\}}e^{-\beta(\mathscr{H}(x')-\mathscr{H}(x^\star))}} \tag{II.1.19}$$

$$= \begin{cases}1 & if\ x=x^\star\\ 0 & otherwise\end{cases}. \tag{II.1.20}$$

The general case follows by the same principle with some adjusted arguments in the algebraic tricks.
(Claim 3) Recall that $\partial_\beta\mathcal{Z}(\beta)=\langle\mathscr{H}(\boldsymbol{X})\rangle_\beta$. Taking the limits they are the same, and we can say that:

$$\lim_{\beta\to\infty}\left(\langle\mathscr{H}(\boldsymbol{X})\rangle_\beta\right) = \lim_{\beta\to\infty}\left(\frac{1}{\mathcal{Z}(\beta)}\sum_{\{\boldsymbol{x}\}}\mathscr{H}(\boldsymbol{x})e^{-\beta\mathscr{H}(\boldsymbol{x})}\right) = \min_{\{\boldsymbol{X}\}}\{\mathscr{H}(\boldsymbol{X})\}, \tag{II.1.21}$$

where we have assumed a discrete distribution. If we wanted to prove this for a continuous distribution, we would have exchanged limit and partial differentiation or limit and integral (see Subsec. A.1.2 for a discussion). In a Physical sense, this operation is always meaningful once the limit exists, since we are talking about a real object. Attempting to prove this as a mathematical statement requires more work. We can for example require that *well-behaved* above means that e.g. dominated convergence (Thm. A.1.12) applies, or assume that the conditions of Lebesgue-Vitali's Theorem (Thm. A.1.18) are valid. We avoid lengthy discussions and just imagine that this is allowed holds. For meaningful Hamiltonians, this is true. □

Remark II.1.22. *In Inference problems, we do not even require the temperature to change, so the discussion of Claim #3 in the above statement can be ignored.*

For this reason, taking the low temperature limit forces the distribution to concentrate on the *ground states*[1] of its associated Hamiltonian. This is a mild concentration phenomenon, since it is an indication that high probability events dominate the outcome of randomness. Accordingly, the MAP estimator is the ground state of the Hamiltonian, which also happens to be the $\beta\to\infty$ limit point of the degenerate distribution, As a direct consequence of Proposition II.1.16, the MAP is the estimator achieving MMSE at $\beta\to\infty$. To build a stronger connection with learning problems, we first present a perspective originated in the study of magnetic materials.

II.2 Spin Glasses and the concept of disorder

Ferromagnetic Models The simplest ferromagnetic model is the Ising model. This is described as a binary-interaction Hamiltonian for binary spins of size n placed on a general graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$. The prescribed strength of interactions is $\mathbf{J}$, the magnetic moment is μ and the local magnetic field is $\boldsymbol{h}$. We endow the configurations with the canonical ensemble distribution. Mathematically:

$$\mathscr{H}(\boldsymbol{x};\mathbf{J},\boldsymbol{h}) = -\sum_{(i,j)\in\mathcal{E}}J_{ij}x_ix_j-\mu\sum_{i=1}^n h_ix_i,\quad \mathbb{P}[\boldsymbol{x}]=\frac{e^{-\beta\mathscr{H}(\boldsymbol{x})}}{\mathcal{Z}(\beta)}\,\mathrm{d}\nu(\boldsymbol{x}), \tag{II.2.1}$$

[1] the global minima

where ν is the counting measure on $\{\pm 1\}^n$.
Even for simple cases of this model open questions resist years of research. In principle, the hardness of every question relates with how one deals with the free entropy $\mathscr{F}(\beta)$. While this model exhibits randomness, we recognize that it is not exactly equivalent to our case: the Hamiltonian has interactions and local terms, but the problem is not inherently Bayesian as the only randomness is in the *signal* $\boldsymbol{x}$. To get to an analogous example, we need to inject more probabilistic objects.
Spin glasses are systems where the Hamiltonian presents *quenched disorder* (**mezardSpinGlassTheory1986**; **mezardInformationPhysicsComputation2009**; **castellaniSpinGlassTheoryPedestrians2005**; **zdeborovaStatisticalPhysicsInference2016**; **krzakalaStatisticalPhysicsMethods2021**).
Just like in the scenario of a Bayesian problem, this disorder is encoded as a random variable that plays the role of the dataset $\mathscr{D}$. The simplest example is the Edwards-Anderson (EA) model (**edwardsTheorySpinGlasses1975**). This has Hamiltonian:

$$\mathscr{H}(\boldsymbol{x};\mathbf{J}) = -\sum_{(i,j)\in\mathcal{E}} J_{ij}x_ix_j, \quad \mathbf{J} = \{J_{ij}\}_{i,j\in\mathcal{E}} \sim \mathbb{P}_{\mathscr{D}}, \tag{II.2.2}$$

and the standard choice of the *couplings* distribution is taken to be Gaussian.
Here a distinction needs to be made. The word *quenched* refers to the timescale of variation. When computing any quantity wrt the posterior, we will first deal with a quenched probability $\mathbb{P}[\boldsymbol{X} \mid \mathscr{D}]$, and then average it out wrt the quenched disorder. The analogy is that the randomness in $\mathscr{D}$ is held fixed while the randomness over the signal $\boldsymbol{x}^*$ varies. In general, the integrals $\int \mathrm{d}\mathscr{D}, \int \mathrm{d}\boldsymbol{x}$ are exchangeable when there are *no functions in the middle*. We easily see this with an application of Jensen's inequality.

Example II.2.3. *Let $\mathcal{Z}(\mathscr{D};\beta=1) = \mathcal{Z}(\mathscr{D})$ be the partition function in the case of the GLM. Then it is a random variable with randomness in the dataset realizations, but it is also an integral per se. For any convex function $f:\mathbb{R}^d\to\mathbb{R}$:*

$$f\left(\mathbb{E}_{\mathscr{D}}\left[\mathcal{Z}(\mathscr{D})\right]\right) \leqslant \mathbb{E}_{\mathscr{D}}\left[f\left(\mathcal{Z}(\mathscr{D})\right)\right]. \tag{II.2.4}$$

As previously discussed, there are many reasons why we would like to tackle the problem of computing the partition function, the two most important ones being having an expression for the posterior and accessing the thermodynamical variables. We remark that this computation has two main hurdles:

1. it is a sum of $n\to\infty$ or in general $n\gg 1$ many exponential terms
2. it is a random realization, and one needs to be careful with taking its expectation at the right time.

While a solution to #1 will require a very long detour, it is common in Statistical Physics to ignore #2 with a judgeful twist of the problem.

Definition II.2.5 (Self-Averaging Quantity)**.** *A quantity $Q(\boldsymbol{y})$ that is dependent on the quenched disorder of a model is said to be self-averaging if in the thermodynamic limit $n\to\infty$ it is independent of the realization of disorder but is just related to its statistical properties. Mathematically:*

$$\lim_{n\to\infty} \mathbb{P}[|Q(\boldsymbol{y}) - \mathbb{E}\left[Q(\boldsymbol{y})\right]| > \epsilon] = 0 \quad \forall\epsilon>0 \tag{II.2.6}$$

Remark II.2.7. *The definition is slightly different from convergence in probability since both $Q(\boldsymbol{y})$ and $\mathbb{E}\left[Q(\boldsymbol{y})\right]$ in principle are indexed by the system size. Contrarily, convergence in probability holds for a sequence (x_n) and a fixed value x. We take this matter with some more care. The right formalism is as follows. Set $X := |Q(\boldsymbol{Y}) - \mathbb{E}\left[Q(\boldsymbol{Y})\right]|$. Then $X \xrightarrow{p} 0$. By definition of convergence in probability:*

$$\lim_{n\to\infty} \mathbb{P}\left[|X-0|>\epsilon\right] = 0 \quad \forall\epsilon>0, \tag{II.2.8}$$

which is also expressed as $\mathrm{p}\lim_{n\to\infty}|Q(\boldsymbol{Y}) - \mathbb{E}\left[Q(\boldsymbol{Y})\right]| = 0$. The equation $\lim_{n\to\infty} Q(\boldsymbol{Y}) = \lim_{n\to\infty}\mathbb{E}\left[Q(\boldsymbol{Y})\right]$ still does not make sense, as on the LHS we have a random variable and on the RHS we have a deterministic quantity, but we take it as a definition of the probability convergence above. This implicitly requires that in the limit $Q(\boldsymbol{Y})$ is not a random variable.

Remark II.2.9. *Even if a quantity $\boldsymbol{Y}$ is not self averaging, it turns out that its logarithm $\ln \boldsymbol{Y}$ often is. In these cases, it is more interesting to look at $\exp\{\ln \mathbb{E}[\boldsymbol{Y}]\}$ than $\mathbb{E}[\boldsymbol{Y}]$ itself.*

It is in general expected that the **free entropy density** is self-averaging in the thermodynamic limit. This means that:

$$\lim_{n\to\infty} \frac{1}{n} \ln[\mathcal{Z}_n(\beta, \mathscr{D})] = \lim_{n\to\infty} f_n(\beta, \mathscr{D}) = \lim_{n\to\infty} \mathbb{E}_{\mathscr{D}}[f_n(\beta, \mathscr{D})] = f(\beta), \tag{II.2.10}$$

by Definition II.2.5. Quenched disorder realizations do not impact the value of the free energy in the limit and we can use the penultimate expression throughout. Notice that we are purposely considering the free energy density as to be an intensive quantity that does not diverge with n. A way to make this work is attempting to obtain an extensive expression for the free energy, which in turn, being the log of a sum of exponential terms, means that the Hamiltonian is extensive. With this premise, it is possible to prove that the variance wrt the disorder vanishes at the $n \to \infty$ limit (e.g. it is of order $\frac{1}{\sqrt{n}}$).
One could also reason as follows. The partition function contains many random contributions, and it is not in general expected that its most probable value coincides with its mean: then $\mathcal{Z}_n(\mathscr{D})$ is not self-averaging. On the contrary, the normalized sum of independent terms tends to a Gaussian distribution by the CLT. Then if $\mathcal{Z}_n \sim e^{n f_n(\beta)}$ at large n it is the case that $\ln \mathcal{Z}_n(\mathscr{D})$ is self-averaging, with associated free entropy density:

$$f(\beta) = \lim_{n\to\infty} \frac{1}{n} \mathbb{E}_{\mathscr{D}}[\ln \mathcal{Z}_n(\mathscr{D}; \beta)]. \tag{II.2.11}$$

While this is not rigorous, it is also not crucial for the exposition. As a matter of fact, Statistical Inference and Information Theory directly study the mean wrt the disorder. For the sake of simplicity, we will take it as a granted assumption when needed. Our model of interest is also proved to be self-averaging in a large portion of cases (see (**barbierAdaptiveInterpolationMethod2018**; **barbierOptimalErrorsPhase2019**; **aubinCommitteeMachineComputational2019**; **barbierOverlapMatrixConcentration2020**)).

Wishing to compute an expression for Equation II.2.10 is the starting point of a Statistical Physics inference problem. A special case of the discussion in Exm. II.2.3 gives us the so-called *annealed* free energy approximation, which relies on the fact that $\ln(\cdot)$ is a concave function. Assuming that the probabilities have a density wrt to a reference measure (Lebesgue for simplicity):

$$\mathbb{E}_{\mathscr{D}}[\ln \mathcal{Z}(\mathscr{D})] = \mathbb{E}_{\mathscr{D}}\left[\ln \int p(\boldsymbol{x} \mid \mathscr{D}) p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}\right] \geqslant \ln\left[\mathbb{E}_{\mathscr{D}}\left[\int p(\boldsymbol{x} \mid \mathscr{D}) p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}\right]\right], \tag{II.2.12}$$

and we have found an overall lower bound to the quenched free energy, or equivalently an upper bound to the quenched free entropy[2]. In both perspectives, one can think of a blanket estimate. For low enough temperatures, the lower bound is known to be not tight in many models. We give a mathematical a physical/quantitative argument argument for this.

(Math) Jensen's Inequality is not necessarily tight.

(Phys) The free energy being extensive, the intensive density should fluctuate with decaying rate as $n \to \infty$. The precise power of the fluctuations is not important, and can be taken to be $\frac{1}{\sqrt{n}}$ for simplicity (**mezardSpinGlassTheory1986**). The partition is instead a sum of exponential terms, with potentially large fluctuations. An average over $\mathcal{Z}$ might be dominated by rare but dominant in size fluctuations.

Example II.2.13 (Non-tight Jensen's for quenched and annealed averages)**.** *The classic scenario in which quenched and annealed averages are different is very easy. Recall that the partition function is random. Let $a, b \in \mathbb{R}$ be constants. If:*

$$\mathcal{Z}(\mathscr{D}; \beta) = \begin{cases} e^{-\beta n} & w.p.\ \frac{a}{n} \\ e^{-b\beta n} & w.p.\ 1 - \frac{a}{n}, \end{cases} \tag{II.2.14}$$

[2] recall that the two differ by a minus sign.

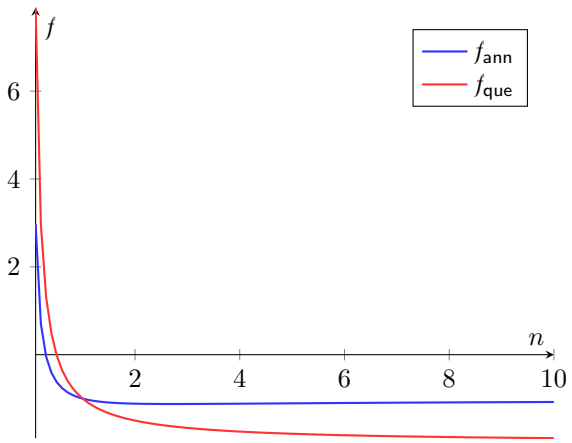


Figure II.1: Quenched and Annealed free entropy density as a function of the number of particles.
Parameters $a = 1, b = 2, \beta = 3$. Given that we consider the large size limit, what matters is the $n \geqslant 1, n \to \infty$ behavior.

with $b \neq 1$ the two free entropies differ. The quenched free entropy density is:

$$f_{\text{que}}(\beta) = \lim_{n\to\infty} \frac{1}{n\beta} \mathbb{E}\left[\ln \mathcal{Z}(\mathscr{D};\beta)\right] = \lim_{n\to\infty} \frac{1}{n\beta}\left[\frac{a}{n}(-\beta n) + \left(1-\frac{a}{n}\right)(-b\beta n)\right] = -b. \tag{II.2.15}$$

While the annealed free entropy density is:

$$\begin{aligned}
f_{\text{ann}}(\beta) &= \lim_{n\to\infty} \frac{1}{n\beta} \ln \mathbb{E}\left[\mathcal{Z}(\mathscr{D};\beta)\right] && \text{(II.2.16)}\\
&= \lim_{n\to\infty} \frac{1}{n\beta} \ln\left[\frac{a}{n} e^{-\beta n} + \left(1-\frac{a}{n}\right) e^{-b\beta n}\right] && \text{(II.2.17)}\\
&= \lim_{n\to\infty} \frac{1}{n\beta} \ln\left[e^{-\beta n}\left(\frac{a}{n} + \left(1-\frac{a}{n}\right) e^{-\beta(b-1)n}\right)\right] && \text{(II.2.18)}\\
&= -1 + \lim_{n\to\infty} \frac{1}{n\beta} \ln\left[\frac{a}{n} + \left(1-\frac{a}{n}\right) e^{-\beta(b-1)n}\right] && \text{(II.2.19)}\\
&= -1 + \lim_{n\to\infty} \frac{1}{n\beta} \ln\left[\frac{1}{n}\left(a + (n-a)\, e^{-\beta(b-1)n}\right)\right] && \text{(II.2.20)}\\
&= -1 + \lim_{n\to\infty} \frac{1}{n\beta} \ln\left[a + (n-a)\, e^{-\beta(b-1)n}\right]. && \text{(II.2.21)}
\end{aligned}$$

The function inside the logarithm is slightly more difficult to deal with. We claim that the limit tends to zero (notice that we also have a $\frac{1}{n\beta}$ term). To show this, one can do a tedious application of Hôpital's rule, and verify that the limit of the result is null. We show an example of the behaviors in Figure II.1.

On the other hand, when the timescales of disorder change and configuration change are similar, the annealed average is the correct one (**krzakalaStatisticalPhysicsInference2015**). The reason is that the motivation for which physicists introduce quenched averages is to deal with different frequencies in the stochasticity of a phenomena, and when these are made to be equal, the whole construction is not needed. From a Bayesian point of view, this is always in principle a question that is not interesting to ask. The terms *quenched* and *annealed* are indeed self-explanatory when considering Physical problems. Nevertheless, even in the theoretical setting, the annealed average provides a lower bound, and can answer interesting questions, the quickest among all providing an immediate approximation of the true CGF. We just add two properties, where the second especially is useful for optimization problems (i.e. study of ground states of a given Hamiltonian).

Proposition II.2.22 (Properties of Quenched/Annealed Free entropy)**.** *Consider a sample size n and an inverse temperature β. Let $u_n(\beta = \infty) := \frac{1}{n}\mathbb{E}_{\mathscr{D}}\left[\min_{\mathscr{X}} \mathscr{H}(\boldsymbol{x};\mathscr{D})\right]$, i.e. the ground state density at zero temperature, a minimum of the Hamiltonian by Prop. II.1.16. Then:*

1. $\frac{\mathrm{d}f_n^{(\mathsf{quench})}(\beta)}{\mathrm{d}\beta} \leqslant 0$

2. *the annealed and the quenched free entropies are convex in β (i.e. convex in $T = \frac{1}{\beta}$)*

3. $u_n(\beta = \infty) \geqslant \min_{\beta\in[0,\infty]} f_n^{(\mathsf{ann})}(\beta)$.

Proof. All statements follow by a careful inspection of (**mezardInformationPhysicsComputation2009**) where the discussion uses as subject the free energy density, working out the necessary sign flips. □

The connection with inference is made stronger when aligning it with some field-specific terminology, that elaborates further with physical arguments concepts that are implicit in a Bayesian inference problem. For this reason, we opt to argue on two macro-intepretations of how disorder is handled.

II.3 Teacher-Student Model

A related paradigm of analysis was introduced by Gardner and Derrida (**gardnerThreeUnfinishedWork**). In short, it presents a generative perspective on an inference problem with nice properties. This is to be seen in comparison and conjunction with the section on Planted models we will present later.

Definition II.3.1 (Teacher Student Scenario)**.** *Two actors are interacting:*

- *a **Teacher**, who generates $\boldsymbol{x}^*$ from a prior distribution $\mathbb{P}_{\mathsf{TP}}[\boldsymbol{x}]$, and outputs $\boldsymbol{y}$ from a likelihood $\mathbb{P}_{\mathsf{TL}}[\boldsymbol{y} \mid \boldsymbol{x}^*]$.*
- *a **Student**, receiving from the Teacher $\boldsymbol{y}$ and information about the prior and likelihood, with the aim of retrieving the $\boldsymbol{x}^*$ used in generating process.*

*Here **retrieving** is purposely sloppy, as in some cases it will require just feasibility and in others also efficiency of the method, as discussed when presenting the two main questions of inference.*
In some cases, the distributions will be fully available (Bayes-Optimal), while in others just partially. In addition to this, variants where there is a further parametrization are naturally included in this setting. While we focus on the first case, the approach is rather general.

It is also common to place the analysis in the simplest possible scenario. To do so, we further assume that the entries of the prior (i.e. the features $j \in [d]$) and the observations (i.e. the samples $i \in [n]$) are independent, leading to the expressions:

$$\mathbb{P}_{\mathsf{TP}}[\boldsymbol{x}] = \prod_{j=1}^{d} \mathbb{P}_{\mathsf{TP}}[\boldsymbol{x}_i] \quad \mathbb{P}_{\mathsf{TL}}[\boldsymbol{y} \mid \boldsymbol{x}] = \prod_{i=1}^{n} \mathbb{P}_{\mathsf{TL}}[\boldsymbol{y}_i \mid \boldsymbol{x}]. \tag{II.3.2}$$

In some cases, the observations might also depend on a subset of the features, where we might use the subset notation $\partial i \subset [d]$, to denote a neighborhood of the observation i and express the new likelihood as:

$$\mathbb{P}_{\mathsf{TL}}[\boldsymbol{y} \mid \boldsymbol{x}] = \prod_{i=1}^{n} \mathbb{P}_{\mathsf{TL}}[\boldsymbol{y}_i \mid \{\boldsymbol{x}_j\}_{j\in\partial i}]. \tag{II.3.3}$$

In models such as the Ising model (Eqn. II.2.1), site-marginals, termed local magnetizations, are an interesting object of study. It turns out that for factorizing problems these are also important for a general Bayesian construction (see Prop. I.2.42 and the remark below). We denote them with the usual literature notation:

$$\mu_j(\boldsymbol{x}_j) \coloneqq \int \mathbb{P}[\boldsymbol{x} \mid \boldsymbol{y}] \prod_{j'\neq j} d\boldsymbol{x}_{j'} \quad j \in [d]. \tag{II.3.4}$$

Having shown how disorder plays an important role in random distributions, we turn to presenting two important statements in the literature of Statistical Physics.

II.4 Nishimori, Stein and more about Bayes-Optimality

The following presentation of the Nishimori identity is borrowed from (**zdeborovaStatisticalPhysicsInference201**
and gives a quick idea of its importance.
Assume we are able to sample from the posterior $\mathbb{P}[\boldsymbol{x} \mid \boldsymbol{y}]$ of a signal $\boldsymbol{x}^*$. For three independent samples $\boldsymbol{x}, \boldsymbol{x}_1, \boldsymbol{x}_2 \sim \mathbb{P}[\boldsymbol{X} \mid \boldsymbol{y}]$ it trivially holds that a function $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ taking vector valued inputs will be such that:

$$\mathbb{E}\left[f(\boldsymbol{X}_1, \boldsymbol{X}_2)\right] = \int f(\boldsymbol{x}_1, \boldsymbol{x}_2) \underbrace{p(\boldsymbol{y})p(\boldsymbol{x}_1 \mid \boldsymbol{y})}_{=p(\boldsymbol{x}_1, \boldsymbol{y})} \underbrace{p(\boldsymbol{x}_2 \mid \boldsymbol{y})}_{p(\boldsymbol{x}_2|\boldsymbol{y})} \mathrm{d}\boldsymbol{x}_1 \, \mathrm{d}\boldsymbol{x}_2 \, \mathrm{d}\boldsymbol{y} \tag{II.4.1}$$

$$\mathbb{E}\left[f(\boldsymbol{X}^*, \boldsymbol{X})\right] = \int f(\boldsymbol{x}^*, \boldsymbol{x}) p(\boldsymbol{x}^*, \boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}^* \, \mathrm{d}\boldsymbol{x} = \int f(\boldsymbol{x}^*, \boldsymbol{x}) p(\boldsymbol{x}^*, \boldsymbol{x}, y) \, \mathrm{d}\boldsymbol{x}^* \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y} \tag{II.4.2}$$

$$= \int f(\boldsymbol{x}^*, \boldsymbol{x}) p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{x}^*) p_{\mathsf{TP}}(\boldsymbol{x}^*) p_{\mathsf{TL}}(\boldsymbol{y} \mid \boldsymbol{x}^*) \, \mathrm{d}\boldsymbol{x}^* \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y} \tag{II.4.3}$$

where the crucial difference between the two is that it makes sense to express $\boldsymbol{y}$ conditionally on $\boldsymbol{x}^*$ to reach the different formulation via the Teacher's distributions. Additionally, the assumptions on the model allow us to say that $\mathbb{P}[\boldsymbol{x} \mid \boldsymbol{x}^*, \boldsymbol{y}] = \mathbb{P}[\boldsymbol{x} \mid \boldsymbol{y}]$ by conditional independence of the posterior sampling[3]. Given this, in the Bayes-Optimal setting it will be the case that:

$$\mathbb{E}\left[f(\boldsymbol{x}_1, \boldsymbol{x}_2)\right] = \mathbb{E}\left[f(\boldsymbol{x}, \boldsymbol{x}^*)\right], \tag{II.4.4}$$

since the TP, TL densities are used in the first expectation, and the integral is over the exact same functions. This property, known in its general form as the *Nishimori identity* is fundamental for many of the concepts we will see later. It can be briefly interpreted as a trick of replacing random posterior samples with the true signal upon knowledge of the true data generating process. In the mismatched-prior/likelihood case, the same reasoning does not guarantee this general condition. The formal statement is reported in the next proposition.

Proposition II.4.5 (Nishimori Identity)**.** *Given k iid samples from the posterior distribution and a measurable integrable function $f : \mathbb{R} \times \mathbb{R}^{d \times k} \to \mathbb{R}$ we can swap one copy of $\boldsymbol{X}^{(i)}$ with the true signal $\boldsymbol{X}^*$. Mathematically, establishing the notation*

$$\left\langle f\left(\boldsymbol{Y}, \boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(k)}\right)\right\rangle_k := \int \prod_{i=1}^{k} f(\boldsymbol{Y}, \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(k)}) p(\boldsymbol{x}^{(i)} \mid \boldsymbol{Y}) \, \mathrm{d}\boldsymbol{x}^{(i)}, \tag{II.4.6}$$

where we stress that $\boldsymbol{Y}$ is random, we have that:

$$\mathbb{E}_{\boldsymbol{Y}}\left[\left\langle f\left(\boldsymbol{Y}, \boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(k)}\right)\right\rangle_k\right] = \mathbb{E}_{(\boldsymbol{Y}, \boldsymbol{X}^*)}\left[\left\langle f\left(\boldsymbol{Y}, \boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(k-1)}, \boldsymbol{X}^*\right)\right\rangle_{k-1}\right] \tag{II.4.7}$$

Proof. One has to be careful about the operations, but the identity follows after some workarounds. We start from the RHS and keep $\boldsymbol{x}^*$ to distinguish it from the others in the integrations. Expanding the outer integral we get:

$$\mathbb{E}_{\boldsymbol{X}^*, \boldsymbol{Y}}\left[\left\langle f(\boldsymbol{Y}, \boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(k-1)}, \boldsymbol{X}^*)\right\rangle_{k-1}\right] = \int p(\boldsymbol{x}^*, \boldsymbol{y}) \left\langle f(\boldsymbol{y}, \boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(k-1)}, \boldsymbol{x}^*)\right\rangle_{k-1} \mathrm{d}\boldsymbol{x}^* \, \mathrm{d}\boldsymbol{y}, \tag{II.4.8}$$

which by separating the joint density into evidence and posterior gives the expression:

$$\int p(\boldsymbol{y}) \int \left\langle f(\boldsymbol{y}, \boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(k-1)}, \boldsymbol{x}^*)\right\rangle_{k-1} p(\boldsymbol{x}^* \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{x}^* \, \mathrm{d}\boldsymbol{y}. \tag{II.4.9}$$

The next step follows from the above discussion. In principle, the Bayes-Optimal setting allows us to replace $\boldsymbol{x}^*$ with $\boldsymbol{x}^{(k)}$ since it is a dummy index of summation,

[3] recall that by construction $\boldsymbol{x}$ is sampled from the posterior, and so it depends only on the realization of the observation.

and we have access to the true likelihood and prior of generation. Then this mask on the true signal density gives the expression:

$$\int\int \left\langle f(\boldsymbol{y}, \boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(k-1)}, \boldsymbol{x}^{(k)}) \right\rangle_{k-1} p(\boldsymbol{x}^{(k)} \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{x}^{(k)} p(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}. \tag{II.4.10}$$

The final step is by definition, the k^{th} posterior is brought inside the bracket notation, while outside the expectation is wrt $\boldsymbol{Y}$. Eventually:

$$\dots = \int \left\langle f(\boldsymbol{y}, \boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(k-1)}, \boldsymbol{X}^{(k)}) \right\rangle_{k} p(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \tag{II.4.11}$$

$$= \mathbb{E}_{\boldsymbol{Y}} \left[\left\langle f(\boldsymbol{Y}, \boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(k)}) \right\rangle_{k} \right]. \tag{II.4.12}$$

□

Remark II.4.13. *While we have used the $\boldsymbol{y}$ notation, everything follows equivalently if we consider a more complicated observation such as $\mathscr{D} = (\boldsymbol{y}, \mathbf{A})$. The choice was influenced by the fact that we could make clear when $\boldsymbol{y}$ was random and when it was fixed.*

This integration trick has very interesting consequences, and different versions. We report some of them below:

Remark II.4.14 (Magnetization with Nishimori)**.** *The concept of magnetization refers to ferromagnetic materials. For our purposes, we will regard it as the average over the dataset and signal randomness of the overlap of a posterior mean (denoted with the Boltzmann notation here) and the true signal. For the moment, we ignore their norms and assume those are unity. It reads:*

$$m \coloneqq \mathbb{E}_{(\mathscr{D}, \boldsymbol{X}^*)} \left[\langle \boldsymbol{X}(\mathscr{D}) \rangle^\top \boldsymbol{X}^* \right], \tag{II.4.15}$$

where $\mathscr{D}$ is random. Letting $f(\mathscr{D}, \boldsymbol{X}, \boldsymbol{X}^*)$ to be our function, Proposition II.4.5 allows us to say:

$$m = \mathbb{E}_{(\mathscr{D}, \boldsymbol{X}^*)} \left[f(\mathscr{D}, \boldsymbol{X}, \boldsymbol{X}^*) \right] = \mathbb{E}_{\mathscr{D}} \left[\left\langle f(\boldsymbol{X}, \boldsymbol{X}', \mathscr{D}) \right\rangle \right] = \mathbb{E}_{\mathscr{D}} \left[\left\langle \boldsymbol{X}(\mathscr{D})^\top \boldsymbol{X}'(\mathscr{D}) \right\rangle \right], \tag{II.4.16}$$

where we just replaced the true signal with a posterior sample $\boldsymbol{X}'$ and replaced the dummy index. In the scalar case, it would read:

$$m = \mathbb{E}_{\mathscr{D}} \left[\langle X(\mathscr{D}) \rangle^2 \right] = \mathbb{E}_{\mathscr{D}} \left[\left\langle X^{(1)}(\mathscr{D}) X^{(2)}(\mathscr{D}) \right\rangle \right], \tag{II.4.17}$$

for two independent replicas[4] *$(x^{(1)}, x^{(2)})$. In general, this always holds by the principle that the true signal can be taken as a posterior sample in the Bayes-Optimal setting. This is intuitive, and could also be seen as an application of Bayes's Theorem.*

Lemma II.4.18 (Optimal Bayesian inference MMSE)**.** *Assume centered (mean zero) distributed signals. In Optimal Bayesian inference it holds that:*

$$\mathsf{MMSE} = \mathbb{E}_{(\boldsymbol{X}^*, \mathscr{D})} \left[(\boldsymbol{X}^* - \langle \boldsymbol{X}(\mathscr{D}) \rangle)^2 \right] = \rho - m, \quad \rho \coloneqq \mathbb{E}_{\boldsymbol{X}^*} \left[{\boldsymbol{X}^*}^\top \boldsymbol{X}^* \right]. \tag{II.4.19}$$

Where ρ can be seen as variance of a centered message, i.e. a self-overlap of the truth, and we denote the matrix square with the scalar square for conciseness in the first expression.

Proof. Denote throughout the square of matrices with the scalar square. We just need to expand the product inside the expectation to find that:

$$\mathbb{E}_{\boldsymbol{X}^*, \mathscr{D}} \left[(\boldsymbol{X}^* - \langle \boldsymbol{X}(\mathscr{D}) \rangle)^2 \right] = \mathbb{E}_{(\boldsymbol{X}^*, \mathscr{D})} \left[(\boldsymbol{X}^*)^2 \right] + \mathbb{E}_{(\boldsymbol{X}^*, \mathscr{D})} \left[\langle \boldsymbol{X}(\mathscr{D}) \rangle^2 \right] \tag{II.4.20}$$

$$- 2\mathbb{E}_{\boldsymbol{X}^*, \mathscr{D}} \left[\langle \boldsymbol{X}(\mathscr{D}) \rangle^\top \boldsymbol{X}^* \right] \tag{II.4.21}$$

$$= \rho + q - 2m = \rho + m - 2m = \rho - m, \tag{II.4.22}$$

where we used the Nishimori condition at the end of the first line. □

[4] A replica is not an independent sample but rather a copy of the original vector. In other words, this is the square norm of the estimator

Lemma II.4.18 guarantees that we only need to compute m (the overlaps), since ρ is rather easy to get, and often stated in the model assumptions. The Nishimori Identity applied to other losses follows the same principles. For example:

$$\mathsf{MEC} = \mathbb{E}_{(\mathscr{D},\boldsymbol{X}^*)}\left[\mathcal{L}(\boldsymbol{X}(\mathscr{D}), \boldsymbol{X}^*; \mathsf{EC})\right] \quad \boldsymbol{X} \overset{d}{=} \boldsymbol{X}^* \sim \mathbb{P}[\boldsymbol{X} \mid \boldsymbol{y}]. \tag{II.4.23}$$

Next, we prove very useful lemma for Gaussian distributions and some generalizations of it.

Lemma II.4.24 (Stein's Lemma (**steinEstimationMeanMultivariate1981**))**.** *Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Let g be differentiable and such that both $\mathbb{E}[(X-\mu)g(X)]$ and $\mathbb{E}[g'(X)]$ exist. Then:*

$$\mathbb{E}\left[g(X)(X-\mu)\right] = \sigma^2 \mathbb{E}\left[g'(X)\right]. \tag{II.4.25}$$

Proof. Found in (**krzakalaStatisticalPhysicsMethods2021**).
Recall the easy Gaussian identity:

$$\int_{-\infty}^{\infty} (x-\mu) e^{-\frac{(x-\mu)^2}{2\sigma^2}} \,\mathrm{d}x = -\sigma^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{II.4.26}$$

Denote the Gaussian density with $\psi(x) \coloneqq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Expressing the LHS of Equation II.4.25 we get applying integration by parts[5]:

$$\begin{aligned}
\mathbb{E}\left[g(X)(X-\mu)\right] &= \int p(x) g(x)(x-\mu)\,\mathrm{d}x = \int \underbrace{\psi(x)}_{f'}\underbrace{(x-\mu)}\,\underbrace{g(x)}_{h}\,\mathrm{d}x \\
&= \left[g(x)\int \psi(x)(x-\mu)\,\mathrm{d}x\right]\Bigg|_{-\infty}^{\infty} - \int g'(x)\left(\int \psi(z)(z-\mu)\,\mathrm{d}z\right)\mathrm{d}x \\
&= \left[g(x)(-\sigma^2)e^{\frac{-(x-\mu)^2}{2\sigma^2}}\right]\Bigg|_{-\infty}^{\infty} - \int g'(x)(-\sigma^2)e^{\frac{-(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x \qquad \text{(II.4.27)} \\
&= \left[g(x)(-\sigma^2)e^{\frac{-(x-\mu)^2}{2\sigma^2}}\right]\Bigg|_{-\infty}^{\infty} + \sigma^2\mathbb{E}[g'(X)].
\end{aligned}$$

Where in Equation II.4.27 we applied the Gaussian identity on both the $\mathrm{d}z$ integral and the integral in the first term. What is missing is proving that the first term is null, this can be realized through some additional arguments.
Inspect the product of $\psi(\cdot)g(\cdot)$ (up to a constant equal to the argument considered) at the extremes $x \to \pm\infty$ where we need to evaluate it. Recall that for finite μ (the mean of $\psi(\cdot)$) we will have that $\lim_{x\to\pm\infty}\psi(x) = 0$ and $\psi(\cdot)$ is decreasing for all $x > \mu$. It is also useful to express $g(\cdot)$ differently as:

$$g(x) = g(x_0) + \int_{x_0}^{x} g'(y)\,\mathrm{d}y. \tag{II.4.28}$$

Now, let $x_0 > \mu$ ensuring that for all $x > x_0$:

$$g(x)\psi(x) = g(x_0)\psi(x) + \psi(x)\int_{x_0}^{x} g'(y)\,\mathrm{d}y \leqslant g(x_0)\psi(x) + \int_{x_0}^{x}\psi(y)g'(y)\,\mathrm{d}y, \tag{II.4.29}$$

where in the last step we used the fact that $\psi(\cdot)$ is decreasing. Then,

$$\lim_{x\to\infty}\sup g(x)\psi(x) \leqslant \int_{x_0}^{\infty}\psi(y)g'(y)\,\mathrm{d}y \quad \forall x_0 > \mu. \tag{II.4.30}$$

Hence, for any ϵ there is an x_0 large enough such that the RHS is less than ϵ. Consequently the LHS is null. The reasoning for $-\infty$ is symmetric. Having proved that the first term is null, the result follows.

[5] $\int f'h = [fh] - \int fh'$

Another proof of this nullity follows by dominated convergence (Thm. A.1.12). It is rather easy to establish that:

$$|g'(y)|\psi(x)\mathbb{1}_{[\mu,x]}(y) \leqslant |g'(y)|\psi(y)\mathbb{1}_{[\mu,x]}(y), \quad \forall y \in \mathbb{R}. \tag{II.4.31}$$

Therefore, the RHS is integrable by assumption, while the LHS tends to zero pointwise for every y as $x \to \infty$. An application of dominated convergence then gives:

$$\lim_{x\to\infty} g(x)\psi(x) = \lim_{x\to\infty} (g(x) - g(\mu))\psi(x) = \lim_{x\to\infty} \int_{\mu}^{x} g'(y)\psi(x)\,\mathrm{d}y = 0, \tag{II.4.32}$$

where in the first step we use the fact that the density tends to zero, in the second we use the integral representation by the differentiability of $g(\cdot)$ and in the third we use the previous equation. Again, the other limit is symmetrically obtained. □

Corollary II.4.33 (Multivariate Stein's Lemma/Gaussian integration by parts)**.** *Consider a Gaussian vector $\boldsymbol{X} = (X_1, \dots, X_n) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and a function $g : \mathbb{R}^d \to \mathbb{R}$. Let all g and all its partial derivatives to be integrable with respect to the Gaussian density. It holds that:*

$$\mathbb{E}\left[g(\boldsymbol{X})(\boldsymbol{X} - \boldsymbol{\mu})\right] = \boldsymbol{\Sigma} \cdot \mathbb{E}\left[\nabla g(\boldsymbol{X})\right], \tag{II.4.34}$$

where the product on the LHS is the inner product matrix-vector, which outputs a vector in $\mathbb{R}^n$. Entry by entry, this reads:

$$\mathbb{E}\left[g(\boldsymbol{X})(X_i - \mu_i)\right] = \sum_{j=1}^{d} \Sigma_{ij}\mathbb{E}\left[\frac{\partial g(\boldsymbol{X})}{\partial X_j}\right], \quad i \in [d]. \tag{II.4.35}$$

In the bidimensional case of (X, Y) Gaussian and $g : \mathbb{R} \to \mathbb{R}$, we get the refined statement

$$\mathrm{CoV}\left[g(X), Y\right] = \mathrm{CoV}\left[X, Y\right]\mathbb{E}\left[g'(X)\right]. \tag{II.4.36}$$

Proof. We prove the statement in Eqn. II.4.35 for arbitrary i, which allows to recover the vector valued expression since they are the same, and reduces the $d = 2$ statement to a special case. To ease calculations, we let $\boldsymbol{X}$ have mean zero wlog. The general case follows by substitution. The argument is from (**panchenkoGaussianDistributionLecture2**) Consider the vector $\boldsymbol{X}$. Fix a target entry $i \in [d]$ to compute $\mathbb{E}\left[g(\boldsymbol{X})X_i\right]$, denote its variance as $\sigma^2 = \mathbb{E}\left[X_i^2\right]$. Let $\widetilde{X}_j := X_j(1 - \rho_j)$, where $\rho_j = \frac{1}{\sigma^2}\mathbb{E}\left[X_jX_i\right] = \frac{1}{\sigma^2}\mathrm{CoV}\left[X_jX_i\right]$. Then it holds that:

$$\mathbb{E}\left[\widetilde{X}_jX_i\right] = \mathbb{E}\left[X_jX_i\right] - \rho_j\sigma^2 = 0 \quad \forall j \tag{II.4.37}$$

by construction. Uncorrelated Gaussians are independent, so we can say that for $\widetilde{X} \perp\!\!\!\perp X_i$ the function $g(\boldsymbol{X}) = g(\widetilde{\boldsymbol{X}} + X_i\boldsymbol{\rho})$ can be seen as a function of X_i only when taking expectations in X_i. An application of Stein's Lemma II.4.24 gives:

$$\mathbb{E}_{X_i}\left[X_ig(\boldsymbol{X})\right] = \mathbb{E}_{X_i}\left[X_ig(\widetilde{\boldsymbol{X}} + X_i\boldsymbol{\rho})\right] = \sigma^2\mathbb{E}_{X_i}\left[\left.\frac{\partial g(\widetilde{\boldsymbol{X}} + t\boldsymbol{\rho})}{\partial t}\right|_{t=X_i}\right]. \tag{II.4.38}$$

By the assumed integrability, an application of Fubini's Theorem returns the expectation wrt the vector $\boldsymbol{X}$, which reads:

$$\mathbb{E}_{\boldsymbol{X}}\left[X_ig(\boldsymbol{X})\right] = \sigma^2\mathbb{E}_{\boldsymbol{X}}\left[\left.\frac{\partial g(\widetilde{\boldsymbol{X}} + t\boldsymbol{\rho})}{\partial t}\right|_{t=X_i}\right]. \tag{II.4.39}$$

Lastly, we unravel the derivative to get:

$$\left.\frac{\partial g(\widetilde{\boldsymbol{X}} + t\boldsymbol{\rho})}{\partial t}\right|_{t=X_i} = \sum_{j=1}^{d} \rho_j \frac{\partial g(\widetilde{\boldsymbol{X}} + X_i\boldsymbol{\rho})}{\partial X_j} = \sum_{j=1}^{d} \rho_j \frac{\partial g(\boldsymbol{X})}{\partial X_j}. \tag{II.4.40}$$

Recognizing that $\rho_j\sigma^2 = \Sigma_{ij}$ the claim follows. □

Remark II.4.41. *A fundamental requirement to apply Stein's Lemma is integrability of g and its partial derivatives with respect to the Gaussian Measure. Subexponential growth of g and its partial derivatives form a set of sufficient conditions for this by dominated convergence. Mathematically, we could require $|g(\boldsymbol{x})| \leqslant c_1 e^{\|\boldsymbol{x}\|}$ and either of the following for each $j \leqslant d$ and fixed i:*

$$\left|\frac{\partial g(\boldsymbol{x})}{\partial x_j}\right| \leqslant c_1 e^{c_2\|\boldsymbol{x}\|} \quad \textit{or} \quad \Sigma_{ij} = 0, \tag{II.4.42}$$

where $c_1, c_2 \in \mathbb{R}$ are positive constants.

In reality, an even more general statement can be proved (see Subsec. A.5).

II.5 Planted Models

Further References

More context and precise statements are found in literature. Two good sources for a first understanding are (**zdeborovaStatisticalPhysicsInference2016**) (**krzakalaStatisticalPhysicsInference2015**). See instead (**achlioptasAlgorithmicBarriersPhase2008**; **krzakalaHidingQuietSolutions2009**; **mosselStochasticBlockModels2012**) for other results.

In many cases, the idea of recovering a known signal gives a second notion of free energy approximation. This is best understood under the lenses of a classical paradigm of analysis which we will discuss at length in this subsection.

The first step is generating a signal completely at random. This means sampling $\boldsymbol{x}^*$ from a prior $\mathbb{P}[\boldsymbol{x}]$ that does not have any informative shape.

Example II.5.1. *For $\mathscr{X} = \{\pm 1\}^d$ a random sampling is accomplished with a coin toss at each $j \in [d]$. This is a d-dimensional Rademacher distribution* $\mathsf{Rad}(\pm 1)$.

Example II.5.2. *If the signal space is a unit sphere $\mathscr{X} = \mathbb{S}^{d-1} = \{\boldsymbol{x} \mid \|\boldsymbol{x}\|_2^2 = 1\}$, a uniform sample is obtained by sampling $\boldsymbol{x} \sim \mathcal{N}(0, \mathbf{I}_d)$ standard Gaussian and setting:*

$$\boldsymbol{x}^* = \begin{bmatrix} \frac{x_1}{\|\boldsymbol{x}\|_2} \\ \vdots \\ \frac{x_d}{\|\boldsymbol{x}\|_2} \end{bmatrix}. \tag{II.5.3}$$

We briefly explain why this is true. Notice that for an orthogonal matrix $\mathbf{O}$ it holds that $\mathbf{O}\boldsymbol{X} \overset{\mathsf{d}}{=} \boldsymbol{X}$, since the characteristic functions are equal. Then $\boldsymbol{X}$ is orthogonally invariant. By the same reasoning:

$$\mathbf{O}\frac{\boldsymbol{X}}{\|\mathbf{O}\boldsymbol{X}\|_2} \overset{\mathsf{d}}{=} \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|_2}, \tag{II.5.4}$$

and the vector belongs to the unit sphere.

Having a signal $\boldsymbol{x}^*$, we then generate the disorder $\mathscr{D}$ such that it is sampled from the energy-based likelihood:

$$\mathscr{D} \sim \mathbb{P}[\mathscr{D} \mid \boldsymbol{x}^*], \quad p(\mathscr{D} \mid \boldsymbol{x}^*) \propto e^{-\beta \mathscr{H}(\boldsymbol{x}^*; \mathscr{D})}, \tag{II.5.5}$$

assuming for simplicity that it admits a density. According to this sampling scenario, we can write Bayes' Theorem for the just derived likelihood:

$$\mathbb{P}\left[\mathscr{D} \mid \boldsymbol{x}^*\right] = \mathbb{P}\left[\boldsymbol{x}^* \mid \mathscr{D}\right] \frac{\mathbb{P}\left[\mathscr{D}\right]}{\mathbb{P}\left[\boldsymbol{x}^*\right]} \propto \mathbb{P}\left[\boldsymbol{x}^* \mid \mathscr{D}\right] \mathbb{P}\left[\mathscr{D}\right], \tag{II.5.6}$$

where in the last step we have used the uniformity of the prior. Then, by the proportionality of the likelihood, the posterior is up to normalization:

$$p(\boldsymbol{x}^* \mid \mathscr{D}) \propto \frac{e^{-\beta \mathscr{H}(\boldsymbol{x}^*;\mathscr{D})}}{p(\mathscr{D})} \implies p(\boldsymbol{x}^* \mid \mathscr{D}) = \frac{e^{-\beta \mathscr{H}(\boldsymbol{x}^*;\mathscr{D})}}{\mathcal{Z}(\beta,\mathscr{D})}, \quad \mathcal{Z}(\beta,\mathscr{D}) = \int e^{-\beta \mathscr{H}(\boldsymbol{x};\mathscr{D})}\,\mathrm{d}\boldsymbol{x}. \tag{II.5.7}$$

Remark II.5.8. *The planted signal $\boldsymbol{x}^*$ is a sample from the posterior distribution, which is in turn the canonical ensemble distribution at equilibrium for a given hamiltonian $\mathscr{H}(\cdot;\mathscr{D})$.*

Unfortunately, there is a very important difference with the quenched ensemble, where the disorder is a random realization of an observed phenomena. Here the disorder $\mathscr{D}$ admits a density:

$$p(\mathscr{D};\mathsf{pla}) = w(\mathscr{D};\beta)\mathcal{Z}(\beta;\mathscr{D}) \propto \mathcal{Z}(\beta;\mathscr{D}) \quad s.t. \quad \int w(\mathscr{D};\beta)\mathcal{Z}(\beta;\mathscr{D})\,\mathrm{d}\mathscr{D} = 1, \tag{II.5.9}$$

where we enforce the last equation to make it a meaningful probability density. The proportionality is obtained by Eqn. II.5.7. The weight factor w can be seen as :

$$w(\mathscr{D}) = \frac{p(\mathscr{D};\mathsf{que})}{\int p(\mathscr{D};\mathsf{que})\mathcal{Z}(\beta;\mathscr{D})\,\mathrm{d}\mathscr{D}}, \tag{II.5.10}$$

where we used another color to make it explicit that the objects are different[6]. Thanks to these relations, we eventually obtain by a combination of Eqns. II.5.9, II.5.10 that:

$$p(\mathscr{D};\mathsf{pla}) = \frac{\mathcal{Z}(\beta;\mathscr{D})}{\mathcal{Z}(\beta;\mathsf{ann})} p(\mathscr{D};\mathsf{que}), \quad \mathcal{Z}(\beta,\mathsf{ann}) := \int p(\mathscr{D};\mathsf{que})\mathcal{Z}(\beta,\mathscr{D})\,\mathrm{d}\mathscr{D}, \tag{II.5.11}$$

with all the specific names spelled out.

Remark II.5.12. *The term on the LHS is the term on the LHS of Eqn. II.5.9, the probability of the disorder in the planted distribution. The weight function is formally a reweighting of the uniform sampling of the quenched distribution by the partition function that cancels out the spiked importance of doing a likelihood sampling in the planted model. The term at the denominator is instead seen as the annealed partition since it is effectively the average over disorder of the quenched partition function. This is best seen in terms of the discussion on Jensen's inequality:* $\log(\mathcal{Z}_{\mathsf{ann}}) = \log \mathbb{E}_{\mathscr{D}}\left[\mathcal{Z}_{\mathsf{que}}\right] \leqslant \mathbb{E}_{\mathscr{D}}\left[\log \mathcal{Z}_{\mathsf{que}}\right]$.

At this point, it is mindful to ask what kind of properties are retained with this injection of a true signal. We report some of the statements of (**krzakalaStatisticalPhysicsInference20**
to give an idea.

Fact II.5.13 (Planted vs Annealed Ensemble)**.** *Consider the planted ensemble, for which randomness is wrt the posterior distribution (Eqn. II.5.7) and the annealed ensemble, that averages out disorder and signal randomness simultaneously. Fix a temperature β for both ensembles and assume that they always admit nice probabilities and densities. Then*

1. *averages over the disorder of the ensembles are equivalent*
2. *the free energies are different*

Proof. **(Claim #1)** We avoid writing dependence on β explicitly since it is the same for both ensembles by assumption. Consider an observable $\mathscr{O}(\boldsymbol{x};\beta,\mathscr{D})$. Denoting $\langle\cdot\rangle_{\mathsf{ann}}, \langle\cdot\rangle_{\mathsf{pla}}$ as the annealed and planted averages we get that:

$$\mathbb{E}_{\mathscr{D}}\left[\langle \mathscr{O}(\boldsymbol{x};\mathscr{D})\rangle_{\mathsf{pla}}\right] = \int p(\mathscr{D};\mathsf{pla}) \int \frac{1}{\mathcal{Z}(\mathscr{D})} e^{-\beta \mathscr{H}(\boldsymbol{x};\mathscr{D})}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\mathscr{D} \tag{II.5.14}$$

$$= \int \frac{\mathcal{Z}(\beta;\mathscr{D})}{\mathcal{Z}(\beta;\mathsf{ann})} p(\mathscr{D};\mathsf{que}) \int \frac{1}{\mathcal{Z}(\mathscr{D})} e^{-\beta \mathscr{H}(\boldsymbol{x};\mathscr{D})}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\mathscr{D} \tag{II.5.15}$$

[6] This is superfluous since the disorder at the denominator is integrated out but hopefully makes the explanation clearer

where we used Eqn. II.5.11. Then a simplification of the two planted partition functions leads to:

$$\mathbb{E}_{\mathscr{D}}\left[\langle \mathscr{O}(\boldsymbol{x};\mathscr{D})\rangle_{\mathsf{pla}}\right] = \int p(\mathscr{D};\mathsf{que}) \underbrace{\frac{1}{\mathcal{Z}(\beta;\mathsf{ann})} \int e^{-\beta\mathscr{H}(\boldsymbol{x};\mathscr{D})}\,\mathrm{d}\boldsymbol{x}}_{\text{annealed}}\,\mathrm{d}\mathscr{D} = \mathbb{E}_{\mathscr{D}}\left[\langle \mathscr{O}(\boldsymbol{x};\mathscr{D})\rangle_{\mathsf{ann}}\right], \tag{II.5.16}$$

where we recognized that the annealed ensemble average was recovered via the definition of annealed partition function in Eqn. II.5.11. In simpler words, we are effectively averaging at the same time over disorder and signal randomness. One could also check that this is true by observing that the joint for annealed models and planted models. The former is proportional to the energy, i.e. $p(\mathscr{D},\boldsymbol{x};\mathsf{ann}) \propto \exp\{-\beta\mathscr{H}(\boldsymbol{x};\mathscr{D})\}$. The latter by construction satisfies:

$$p(\mathscr{D},\boldsymbol{x};\mathsf{pla}) = p(\mathscr{D}\mid\boldsymbol{x};\mathsf{pla})p(\boldsymbol{x};\mathsf{pla}) \propto \exp\{-\beta\mathscr{H}(\boldsymbol{x};\mathscr{D})\}. \tag{II.5.17}$$

(Claim #2) The planted ensemble is by construction a *typical* problem instance, and we might compute its free energy. The annealed free energy is instead an average over instances (over the disorder). In principle, it is not necessary that they are the same object. □

Remark II.5.18. *As a byproduct of the second claim above, we get that the annealed free energy is not the free energy of a typical problem instance of the annealed ensemble, but rather an average estimate. Given that it is a mean, and not a representative, it can be negative. One of the simplest examples is the Random Energy Model (***derridaRandomenergyModelExactly1981***). On the contrary, the planted model exhibits a realistic sample, which cannot have negative free energy. By #1 above, it can be intepreted as a typical instance of the annealed ensemble.*

Having explained in which terms the annealed and planted ensemble are related, we now turn to establishing links between the quenched ensemble and the planted ensemble. In earlier works, this connection was derived in a very peculiar way. Consider a Hamiltonian that presents disorder in coupling variables $\mathbf{J}$, such as the EA Hamiltonian of Eqn. II.2.2. Starting from a planted ensemble a *gauge invariant* mapping of the Hamiltonian can lead to a quenched ensemble with a different distribution of the disorder. From this distribution, a relationship between the proportion of positive couplings and the inverse temperature can be established. This is nothing but an equation between two scalar quantities (j_+,β) that can be drawn on a line in $\mathbb{R}^2$, termed *Nishimori Line* (**nishimoriStatisticalPhysicsSpin2001**). On the Nishimori line, the planted ensemble is a Nishimori ensemble, which in turn is a particular type of quenched ensemble. If we assume self-averaging of the free energy densities, by Remark II.5.18, the annealed typical behaviors and quenched typical behaviors are matched. We make this explicit by equating their free energy densities in the limit[7], which implies that the planted partition $\mathcal{Z}(\beta;\mathscr{D})$ is equal to the annealed partition $\mathcal{Z}(\beta;\mathsf{ann})$ in Equation II.5.11, and we get for free that the disorder densities are the same in the planted and quenched ensemble. It turns out that this derivation, despite being very useful in determining a path to the connection between ensembles, is not necessary (**zdeborovaStatisticalPhysicsInference2016**). As detailed in (**ibaNishimoriLineBayesian1999**), there is a more general connection between Bayes-Optimal inference and the study of Spin Glasses. Recall that for a given Hamiltonian a planted configuration is effectively an equilibrium configuration of the posterior. This can be seen also in terms of the Nishimori Identity (Prop. II.4.5), which states that a sample from the posterior behaves like the planted signal under averages or in the thermodynamic limit (**zdeborovaStatisticalPhysicsInference2016**), where averaging properties start to hold. This fact, combined with Remark II.5.18 and Equation II.5.11 is already sufficient to obtain $p(\mathscr{D};\mathsf{pla}) = p(\mathscr{D};\mathsf{que})$. The final conclusion is that "generating instances from the planted ensemble is the same thing as generating from the randomly-quenched ensemble" (**zdeborovaStatisticalPhysicsInference2016**). When the free energy densities are equal as above, we refer to this equivalence as *quiet planting*.

[7] which means we discard fluctuating differences at finite size

In the context of the Teacher Student Model, the randomness of $\boldsymbol{y}$ is due to the action of the teacher. From the perspective of the student, $\boldsymbol{y}$ is held fixed, and acts as *quenched disorder*.

If we add planting, each instance of $\boldsymbol{y}_i$ is not properly iid, but is known to be generated from a specific instance $\boldsymbol{x}^*$. To understand this, we take inspiration from (**krzakalaStatisticalPhysicsInference2015**) and (**zdeborovaStatisticalPhysicsInference2016**) to present the general idea behind this. The signal $\boldsymbol{x}^*$ is essentially injected/planted by the teacher, and the quenched noisy observation $\boldsymbol{y}$ is sampled from the likelihood $\mathbb{P}_{\mathsf{TL}}[\boldsymbol{y} \mid \boldsymbol{x}^*]$, which makes it an equilibrium configuration. The power of this backwards task is best understood in terms of the classical way of exploring the phase space by Monte Carlo simulations, which from $\boldsymbol{y}$ attempt to retrieve $\boldsymbol{x}^*$ with strong difficulties in complex problems. Contrarily, a planted model does so with no effort. Following the approach in (**zdeborovaStatisticalPhysicsInference2016**), we refer to the objects as **planted disorder**, and **planted configuration**. A Teacher-Student Model will then take the self explanatory name of **planted model**.

Example II.5.19. *For the general form of the posterior in Eqn. II.1.15, we recognize that the Bayes-Optimal distribution of our model will be the $\beta = 1$ version.*

Further References

For a perspective in error-correcting codes, see (**sourlasSpinGlassesErrorCorrecting1994**; **sourlasSpinglassModelsErrorcorrecting1989**; **sourlasStatisticalMechanicsErrorCorrecting1994**), or the pedagogical presentation in the book (**nishimoriStatisticalPhysicsSpin2001**). Studies in Mathematics are found in (**jerrumLargeCliquesElude1992**; **jerrumSimulatedAnnealingGraph1993**).

The utility of this perspective is in the way planting is explicitly stated in the data generating process. Many models fall under this general construction. Some examples are found in (**zdeborovaStatisticalPhysicsInference2016**). It is however crucial to understand that a necessary requirement is that the Boltzmann distribution must be also the posterior of the inference problem. A situation in which this does not happen is when the quenched disorder is generated by auxiliary variables which are not part of the posterior-inference connection (**zdeborovaStatisticalPhysicsInference2016**).

II.6 Revisiting Hardness Concepts with Statistical Physics

Armed with all these notions, we are now ready to discuss the strategies we will oversee and those that we be avoided for the sake of time. To keep the exposition interesting, we present right away a connection between Phase Transition Types and the Phase diagram that we could encounter in an inference problem.

II.6.1 Describing the stages of inference

Much of the exploration falling under the large incubator of Statistical Physics and beyond aims to provide an understanding of the phases of a certain problem. While the concept of Phase can be somehow *loose* (**krzakalaStatisticalPhysicsMethods2021**), its utility is certainly benefiting from this. By **phase**, we will purposely intend any collection of parametrizations of a given problem that obeys the same phenomenology. In order to make this clearer, one identifies a space of parameters, say $H + 1$ real variables, and a notion of manifestation of their choice, which we call **order parameter**, i.e. an observable which can be measured. Arising in the context of Thermodynamics, the notion of Physical phase transition presumes that $n \to \infty$, and is always related to an abrupt change[8], in the order parameter of choice. The

[8] in fact: a discontinuity

most intuitive way to describe this graphically is a plane in $\mathbb{R}^{H+1}$ dimensions where the silhouette of the order parameter presents a visible change.

The careful reader will have recognized a great similarity with the discussion in Section I.3. Indeed, the notion of Statistical-to-Computational gap is nothing but a partition of the space of parameters of an inference problem into phases of inference, described by a performance indicator (say, a loss). Surprisingly, not only the methods end up being helpful in both fields, but even the questions become substantially equivalent. In this subsection, we give further space to the Theoretical terminology and notions, leaving some details to further Chapters.

Since the statistics of the object under analysis are to some level well-described by the free energy/entropy and its derivatives, it follows that a phase transition is a discontinuity of the derivative of the free energy in some order. The general definition of a phase transition occurrence is related to the free energy being non-analytic.The peculiar form of the function (logarithm of a sum of exponentials) makes it analytic for any system size n. Potential non-analytic points will appear only when $n \to \infty$. This justifies identifying phase transitions at the thermodynamic limit. One of the most common classifications of such phenomena is attributed to Ehrenfest (**ehrenfestConceptualFoundationsStatistical1960**). Two main classes are considered, based on the fact that non differentiability at some derivative implies non-analyticity.

- **1st order** when the derivative of $\mathscr{F}(\beta)$ is discontinuous. Physics intuition explains that this happens when there is phase coexistance in a region of phase space. These two stable phases lead to **metastability** at a higher free energy than the optimal one. For problems that can be observed in a Euclidean setting, it is possible to describe the regions of coexistance clearly, and refer to its boundaries as **spinodal points**. Note also that the first derivative of the free energy is an entropy term.

- **second order** when the second derivative of $\mathscr{F}(\beta)$ is discontinuous. Crucially, this also means that we can expand to first order the free energy wrt the variable that we are using to compute the derivative. In some cases, the second derivative of the free entropy is a susceptibility term, that is in connection with lengthscales of correlation of particles (for a clearer discussion, refer to (**mezardInformationPhysicsComputation2009**)).

- k^{th} **order** analogously to the two above.

Remark II.6.1 (Discontinuous free energy derivatives and thermodynamical quantities)**.** *Notice that the free energy derivative being discountinuous also makes the thermodynamically associated quantity discontinuous, as well as all the other quantities that involve it in their expressions.*

Remark II.6.2. *One might ask, why we can speak about phases and we do not have single points behaving differently? We leave this important but intuitive aspect on hold until we will present the model in Chapter V, but it could be already answered with the information at hand, and is indeed glossed over in literature. Briefly, if one chooses a reasonable order parameter, then the phase diagram will be some sort of tiling of the parameter space, where no tiles sharing the same label will be separate, making the word partition somewhat non-precise. As a consequence, there will be a clear line of equilibrium where two phases are said to coexist, which is nothing but the boundary between two regions. A graphical example is Figure II.2.*

The two types are just a description of phase transitions in general terms with some comments. Adopting a problem/question focused perspective, a different collection of phase transitions can appear, as well as different comments on it. In the next subsection, we briefly describe some methods in literature that appeared to solve questions around these topics, about the geometry of solutions, the ground states, or the shape of the free energy. The hardness landscape presents connections with these types of phase transitions. In simple words, the notion of phase transition is strongly related to the question one wants to ask and the method used, and should be thought of rather as a paradigm of description.

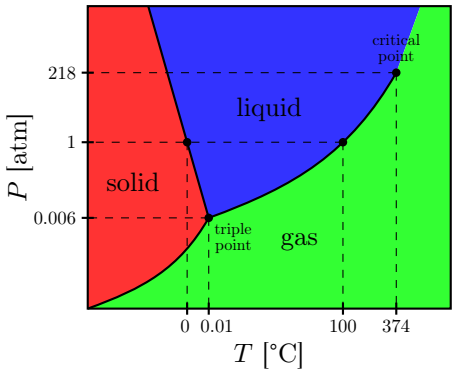


Figure II.2: A portion of the Pressure-Temperature Phase diagram of water, Source: Izaak Neutelings, tikz.net
The most classic example of phase diagram. Phase boundaries are black lines, and the $\mathbb{R}^2_+$ plane is partitioned into three phases, corresponding to the states of matter of the $\mathsf{H_2O}$ molecule.

Remark II.6.3 (A nice observation by (**zdeborovaStatisticalPhysicsInference2016**))**.** *A second order transition is often of the information theoretic type, while a first type transition is related to algorithmic hardness. The latter implies a consideration of the concept of metastable solutions, i.e. points that are not the actual target where some iterative process gets stuck for large timescales.*

II.6.2 Mutual Information and Free Energy

To give a quick comment on the generality of the method, it is worth discussing its connections with Information Theory. Recall that a Csizár divergence with associated function f defines the Csizár-mutual information of two random variables with joint measure $(X, Y) \sim \rho$ and individual measures μ, ν as:

$$\mathfrak{I}(X;Y) := \mathsf{d}_f(\rho||\mu \otimes \nu). \tag{II.6.4}$$

In particular, the choice $f(x) = x \log x$ returns the classic KL divergence. We focus on it from now on.
Aiming to show results for a particular but wide class of DGPs, we now turn to the narrative of (**barbierOverlapMatrixConcentration2020**). Assume the signal is some vector $\boldsymbol{x}^* \in \mathbb{R}^d$, or possibly a matrix. Let the prior be:

- parametric $\mathbb{P}_0[\cdot;\vartheta]$
- with bounded support $\mathsf{supp}(\mathbb{P}_0[\cdot \mid \vartheta_0])$ for all $\vartheta_0 \in \Theta_0$.

Build the observations and the signal as:

$$\boldsymbol{x}^* \sim \mathbb{P}_0[\cdot \mid \vartheta_0] \quad \boldsymbol{y} \sim \mathbb{P}_{\mathsf{out}}\left(\cdot \mid \boldsymbol{x}^*, \vartheta_{\mathsf{out}}\right), \tag{II.6.5}$$

where $\boldsymbol{y}$ is in some euclidean space of sampling and $\vartheta_{\mathsf{out}} \in \Theta_{\mathsf{out}}$. Assuming a Bayes-optimal setting, everything is known to the statistician and we have access to the full posterior:

$$\mathbb{P}\left[\boldsymbol{X}^* = \boldsymbol{x} \mid \boldsymbol{y}, \vartheta\right] = \frac{\mathbb{P}_0[\boldsymbol{x} \mid \mathscr{D}]\mathbb{P}_{\mathsf{out}}(\boldsymbol{Y} \mid \boldsymbol{x}, \vartheta_{\mathsf{out}})}{\int \mathbb{P}_{\mathsf{out}}(\boldsymbol{Y} \mid \boldsymbol{x}', \vartheta_{\mathsf{out}})\, \mathrm{d}\mathbb{P}_0[\boldsymbol{x}' \mid \mathscr{D}]} \tag{II.6.6}$$

$$= \frac{1}{\mathcal{Z}_d(\boldsymbol{y}, \vartheta)} \mathbb{P}_0(\boldsymbol{x} \mid \vartheta_0) \exp\left\{-\mathscr{H}(\boldsymbol{x}, \boldsymbol{y}, \vartheta_{\mathsf{out}})\right\}, \tag{II.6.7}$$

where $\vartheta = (\vartheta_0, \vartheta_{\mathsf{out}})$ are generic variables that can be high dimensional.

Taking the formalism of (**barbierOverlapMatrixConcentration2020**) the average free entropy in this case reads:

$$\mathscr{F} = \mathbb{E}\left[\ln \mathcal{Z}(\boldsymbol{Y}, \vartheta)\right] = \mathbb{E}\left[\int \exp\left\{-\mathscr{H}(\boldsymbol{x}, \boldsymbol{Y}, \theta_{\mathsf{out}})\right\} \mathrm{d}\mathbb{P}_0(\boldsymbol{x} \mid \theta_0)\right], \tag{II.6.8}$$

where $\mathbb{E}\,[]$ is $\mathbb{E}_{\vartheta}\left[\mathbb{E}_{\boldsymbol{X}|\theta}\left[\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X},\theta_{\mathsf{out}}}[\cdot]\right]\right]$, and averages all the quenched variables. Namely, it integrates over the randomness of $(\boldsymbol{X}, \boldsymbol{Y}, \vartheta)$. Contrarily, $\boldsymbol{x}$ inside is dynamic, and is distributed in terms of the posterior (**barbierOverlapMatrixConcentration2020**). With this construction, a nice direct connection with Information Theory objects can be established.

Proposition II.6.9. *Consider a problem as above and the expected free energy, which is minus the average free entropy. Then if the prior factorizes into equal components:*

$$\mathfrak{I}(\boldsymbol{X}, \boldsymbol{Y} \mid \vartheta) = \mathfrak{F}(\boldsymbol{Y} \mid \vartheta) - \mathcal{H}(\boldsymbol{Y} \mid \boldsymbol{X}, \vartheta). \tag{II.6.10}$$

Proof. Straightforward. Notice that the free energy, as defined, is the Shannon entropy conditional on ϑ, i.e. $\mathfrak{F}(\boldsymbol{Y} \mid \vartheta) = -\mathcal{H}(\boldsymbol{Y} \mid \vartheta)$. Then this is just a rearrangement of the classic decomposition of mutual information. □

Remark II.6.11. *In the case of a dataset which we discussed until now, we can recover this result. Recall that $\mathscr{D} = (\boldsymbol{y}, \mathbf{A})$, so $\boldsymbol{y}$ takes the positions y_i of the claim for a distribution p_{out}, while ϑ is a combination of $(\vartheta_0, \vartheta_{\mathsf{out}})$. In ϑ_0 we place whatever parameter the prior might have, while $\vartheta_{\mathsf{out}} = \mathbf{A} = \{\boldsymbol{a}_i\}_{i=1}^n$, to be taken iid row by row. For n rows, at each step we have that:*

$$y_i \overset{\text{iid}}{\sim} p_{\mathsf{out}}\left(\cdot \mid \langle \boldsymbol{a}_i, \boldsymbol{x}^* \rangle\right), \quad i \in [n], \quad \mathbb{P}_0 = p_0^{\otimes d} \tag{II.6.12}$$

so that the total likelihood is obtained by factorization $\mathbb{P}_{\mathsf{out}}(\cdot \mid \theta_{\mathsf{out}}, \boldsymbol{x}^*) = \prod_{i=1}^n p_{\mathsf{out}}(\cdot \mid \langle \boldsymbol{a}_i, \boldsymbol{x}^* \rangle)$. The results are seamlessly adapted for $\mathfrak{I}(\boldsymbol{X}, \mathscr{D} \mid \vartheta_0)$.

By the above identity, since it is often the case that the conditional entropy on the RHS is constant, we can say that the expected free entropy and the mutual information are equal up to constants:

$$\mathfrak{I}(\boldsymbol{X}, \mathscr{D} \mid \vartheta) \cong -\mathbb{E}_{\mathscr{D}}\left[\mathscr{F}(\mathscr{D})\right]. \tag{II.6.13}$$

In some cases, the connection mutual information/free entropy and the minimum MSE estimator is even more explicit, as shown in the next Theorem, borrowed from (**krzakalaStatisticalPhysicsMethods2021**). In general, the objective is the same but the relation is more convoluted. As a matter of fact, the partition function is a generator for the randomness of the problem, and the MMSE being a quantity dependent on randomness will be necessarily related to its *shape*.

Theorem II.6.14 (I − MMSE Theorem, simple case)**.** *For a single disturbed measurement:*

$$Y = X^* + \sqrt{\Delta}\epsilon \quad \epsilon \sim \mathcal{N}(0, 1) \tag{II.6.15}$$

It holds that:

$$\frac{\partial \mathscr{F}(\Delta, n)}{\partial \Delta^{-1}} = \frac{1}{2} m \tag{II.6.16}$$

$$\frac{\partial \mathfrak{I}(\Delta)}{\partial \Delta^{-1}} = \frac{1}{2}(\rho - m). \tag{II.6.17}$$

Where ρ is the variance of the signal and m is the posterior expectation, making $\rho - m$ the value of the MMSE.

Proof. Firstly, we express the free entropy as a function of the noise to take its expectation:

$$\mathscr{F} = \mathbb{E}_Y\left[\ln[\mathcal{Z}(Y)]\right] = \mathbb{E}_Y\left[\ln \int e^{-\frac{x^2}{2\Delta} + \frac{xY}{\Delta}} \mathbb{P}_X(x)\,\mathrm{d}x\right] = \mathbb{E}_{(X^*,Z)}\left[\ln \int e^{-\frac{x^2}{2\Delta} + \frac{xX^*}{\Delta} + \frac{xZ}{\sqrt{\Delta}}}\right] \mathbb{P}_X(x)\,\mathrm{d}x. \tag{II.6.18}$$

The partial derivative yields[9]:

$$\frac{\partial \mathscr{F}(\Delta, n)}{\partial \Delta^{-1}} = \mathbb{E}_{(X*,Z)}\left[\underbrace{\frac{1}{\mathcal{Z}(Y)}\int e^{-\frac{x^2}{2\Delta}+\frac{xX*}{\Delta}+\frac{xZ}{\sqrt{\Delta}}}\mathbb{P}_X(x)}_{\text{Boltzmann weight}}\left[-\frac{x^2}{2}+xX^*+\frac{xZ}{2}\sqrt{\Delta}\right]\mathrm{d}x\right] \tag{II.6.19}$$

$$= \mathbb{E}_{(X*,Z)}\left[\left\langle -\frac{X^2}{2}+XX^*+\frac{XZ}{2}\sqrt{\Delta}\right\rangle\right], \tag{II.6.20}$$

where we have used brackets to denote the posterior expectation. Rearranging terms one gets:

$$\cdots = -\frac{1}{2}\mathbb{E}_{(X*,Z)}\left[\left\langle X^2\right\rangle\right] + \mathbb{E}_{(X*,Z)}\left[X^*\left\langle X\right\rangle\right] + \frac{1}{2}\mathbb{E}_{(X*,Z)}\left[\underbrace{Z}_{\sim\mathcal{N}(0,1)}\underbrace{\sqrt{\Delta}\left\langle X\right\rangle}_{:=g(z)}\right] \tag{II.6.21}$$

$$= -\frac{1}{2}\mathbb{E}_{(X*,Z)}\left[\left\langle X^2\right\rangle\right] + \mathbb{E}_{(X*,Z)}\left[X^*\left\langle X\right\rangle\right] + \frac{1}{2}\mathbb{E}_{(X*,Z)}\left[\underbrace{\left\langle X^2\right\rangle - \left\langle X\right\rangle^2}_{=g'(z)}\right] \tag{II.6.22}$$

$$= -\frac{1}{2}\mathbb{E}_{(X*,Z)}\left[\left\langle X^2\right\rangle\right] + \mathbb{E}_{(X*,Z)}\left[\left\langle X\right\rangle^2\right] + \frac{1}{2}\mathbb{E}_{(X*,Z)}\left[\left\langle X^2\right\rangle\right] - \frac{1}{2}\mathbb{E}_{(X*,Z)}\left[\left\langle X\right\rangle^2\right]$$
$$= \frac{1}{2}m.$$

Where in the passage after Equation II.6.21 we used Nishimori's (Prop. II.4.5) for the second term and Stein's (Lem. II.4.24). Explicitly, using x to highlight the weights from the function in the expectation:

$$\left\langle X\right\rangle = \frac{1}{\mathcal{Z}(Y)}\int e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x\mathbb{P}_X(x)\,\mathrm{d}x = \frac{\left[\int e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x\right]\mathbb{P}_X(x)\,\mathrm{d}x}{\int e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}\mathbb{P}_X(x)\,\mathrm{d}x}, \tag{II.6.23}$$

and the classic differentiation rule for fractions:

$$\frac{\partial}{\partial z}\left(\frac{\mathrm{num}(z)}{\mathrm{den}(z)}\right) = \frac{[\partial_z \mathrm{num}(z)]\mathrm{den}(z) - [\partial_z \mathrm{den}(z)]\mathrm{num}(z)}{[\mathrm{den}(z)]^2} \tag{II.6.24}$$

$$\mathrm{num}(z) = \int e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x\mathbb{P}_X(x)\,\mathrm{d}x \tag{II.6.25}$$

$$\mathrm{den}(z) = \mathcal{Z}(Y) \quad y = x^* + \sqrt{\Delta}z \tag{II.6.26}$$

$$\partial_z \mathrm{num}(z) = \int e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x\underbrace{\frac{\partial}{\partial z}\left(-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}\right)}_{=\frac{x}{\sqrt{\Delta}}}\mathbb{P}_X(x)\,\mathrm{d}x \tag{II.6.27}$$

$$\partial_z \mathrm{den}(z) = \int e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}\underbrace{\frac{\partial}{\partial z}\left(-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}\right)}_{=\frac{x}{\sqrt{\Delta}}}\mathbb{P}_X(x)\,\mathrm{d}x. \tag{II.6.28}$$

Filling all the passages we plug these results inside the expression to find that:

[9] ignoring some regularity conditions which we enforce

$$\frac{\partial\sqrt{\Delta}\langle X\rangle}{\partial z} = \underbrace{\frac{\sqrt{\Delta}}{[\mathcal{Z}(Y)]^2}}_{=[\mathrm{den}(z)]^2}\Bigg\{\underbrace{\int e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}} x\frac{x}{\sqrt{\Delta}}\mathbb{P}_X(x)\,\mathrm{d}x}_{=\partial_z\mathrm{num}(z)}\underbrace{\mathcal{Z}(Y)}_{=\mathrm{den}(z)} \tag{II.6.29}$$

$$-\left[\underbrace{\int dx e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}\frac{x}{\sqrt{\Delta}}\mathbb{P}_X(x)\,\mathrm{d}x}_{=\partial_z\mathrm{den}(z)}\right]\underbrace{\int e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x\mathbb{P}_X(x)\,\mathrm{d}x}_{=\mathrm{num}(z)}\Bigg\}$$

$$= \frac{\sqrt{\Delta}}{[\mathcal{Z}(Y)]^2}\frac{\mathcal{Z}(Y)}{\sqrt{\Delta}}\int e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x^2\mathbb{P}_X(x)\,\mathrm{d}x \tag{II.6.30}$$

$$-\sqrt{\Delta}\frac{1}{\sqrt{\Delta}}\frac{\int e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x\mathbb{P}_X(x)\,\mathrm{d}x}{\mathcal{Z}(Y)}\frac{\int e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x\mathbb{P}_X(x)\,\mathrm{d}x}{\mathcal{Z}(Y)}$$

$$= \frac{1}{\mathcal{Z}(Y)}\int e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x^2\mathbb{P}_X(x)\,\mathrm{d}x - \left[\frac{\int e^{-\frac{x^2}{2\Delta}+\frac{xx^*}{\Delta}+\frac{xz}{\sqrt{\Delta}}}x\mathbb{P}_X(x)\,\mathrm{d}x}{\mathcal{Z}(Y)}\right]^2 \tag{II.6.31}$$

$$= \left\langle X^2\right\rangle - \langle X\rangle^2\,. \tag{II.6.32}$$

Therefore, the result of Equation II.6.17 can be obtained using Proposition II.6.9, the fact that the entropy of a Gaussian channel is constant at $\mathcal{H}(X\mid Y)\cong\frac{1}{2}\ln 2\pi e\Delta$ and the just proved Equation II.6.16. □

Chapter III

Message Passing Algorithms

As we briefly discussed, the description of the randomness of a problem has a peculiar interpretation with tools of Statistical Physics. We now turn to the second objective of this document which is designing a procedure that is able to retrieve signals in practice. To do so, we will present a model, a relaxed version and a practical algorithm, which is believed to be state of the art. This path goes through some yet to answer questions, but turns out to have a complete description in its final form for most models. In Chapter V, we will use it to showcase the way to tackle a specific problem.
In Section III.1 Graphical Models are presented. This visualization, despite not being universal, is very useful to provide an aesthetic and effective presentation of Message Passing Algorithms (Sec. III.2) in general and of Belief Propagation, which will be the focus of our discussion. Instead of writing down the equations directly, we take a more didactic path, building up from easy settings to a fairly general formulation. This goes through discussing the simplest types of graphs (trees, Subsec. III.2.2). Proceeding, we get into the realm of general graphs representing distributions in Section III.2.3, and report the very nice results about their representations in terms of quantities originated in Physics. As a final set of interesting topics, we derive a high-dimensional deterministic set of equations able to describe the dynamics of the Algorithm in Section III.2.4, which presents the difficulty of simulating the trajectories (Sec. III.2.5), and give yet another view of the role of phase transitions in inference problems in Section III.3. In particular, this last topic is quickly overviewed, and must be deepened with more advanced references, included in the discussion.

Notation Throughout, we index what we call *variables* by i, j, l, m, and *factors* by a, b, c, d. A graph is an object $\mathcal{G}(\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ are nodes and $\mathcal{E}$ are the edges. The symbol $\partial a = \{i_1^a, \dots, i_{k_a}^a\}$ denotes the neighbors of a given node a. A bipartite graph is a special graph where the vertex set is made of two groups $(\mathcal{V}, \mathcal{W})$ that have no intraconnections and only interconnect. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, we write $\boldsymbol{x}_{\partial a} \equiv (x_{i_1^a}, \dots, x_{i_{k_a}^a})$, which is nothing but a subset of the entries corresponding to the neighborhood of a given node.

III.1 Graphical Models

The central question of interest is computing the marginals of an inference problem. We provide two justifications for this. Recalling the discussion in the previous Chapter, and in particular Proposition I.2.42, we understand that this is crucial in determining the behavior of optimal estimators. Apart from this, it is also important as an independent question: having access to the marginals of a distribution we can reconstruct the behavior of a collection of objects that behave stochastically. The best way to visualize the utility of graphical models is in visualizing Constraint Satisfaction Problems, but their potential extends well beyond. A probability distribution can be decomposed into conditionals. For $\boldsymbol{x} \in \mathbb{R}^d$ and well

defined conditional distributions it is always true that:

$$p(\boldsymbol{x}) = p_1(x_1)p(x_2 \mid x_1) \cdot p(x_d \mid x_1, \dots, x_{d-1}). \tag{III.1.1}$$

In almost all cases[1], the properties of such distribution can be expressed with a graph structure. From now onwards, we assume that this is the case and aim to design a procedure that estimates the marginals by operating on a graph. Differently from distributions, graphs are operational structures: they allow for the construction of actions over edges/nodes.
If we saw Equation III.1.1 with the simplest graph, we would have the trivial clique with $d-1$ vertices, and edges joining each variable $j \in [d]$ to each other one. If the statistician stops here, every distribution is graphically equivalent, and the visual perspective is useless. Additionally, it is not guaranteed that the expressiveness of the graph will characterize the true distribution. This motivates the wish to find a different visualization that is advantageous computationally.
In some cases, the dependence structure of a probabilistic model simplifies, due to some conditional independence properties. Consequently, also the representative graph loses edges and becomes a peculiar descriptor.

Remark III.1.2. *Even in the simplest case where the domain is $\mathscr{X} = \{\pm 1\}^d$, the generic problem of inferring $\mu(\boldsymbol{x})$ or its marginals is hard: the space of a marginal distribution has 2^{d-1} potential realizations over which one should sum. Given the absence of structural assumptions, this is the best one can do and is clearly computationally inefficient. A starting justification for considering structured probabilities is that the generic question is too hard. On the contrary, some assumptions are just useless or too easy. For example:*

- *if the variables are completely independent $p(\boldsymbol{x}) = \prod_{j=1}^{d} p(x_j)$*
- *if the variables are split into K independent distinct clusters $p(\boldsymbol{x}) = \prod_{k=1}^{K} p(\boldsymbol{x}_k)$, and the atomic structure of the cluster makes inference be as hard as the generic problem, with $2^{k-1} < 2^{d-1}$ size this time, but still exponential.*

On the contrary, a choice of overlapping subsets of conditionals is the interesting one to inspect.

In general, graphical models are used to express the dependence structure of a probabilistic model, by explicitly showing how its density function factorizes. Each comes with its rules for visualization. In this document, we focus on factor graphical models, which operate on factor graphs. A factor graph is a bipartite graph $\mathcal{G}(\mathcal{V}, \mathcal{W}, \mathcal{E})$, where the two vertex sets $(\mathcal{V}, \mathcal{W})$ are joined with edges $\mathcal{E}$. In our notation, $\mathcal{V}$ will be the set of variable nodes, $\mathcal{W}$ the set of factor nodes. Factors are just a generalization of conditionals, which do not require to be proper densities. We will see that this is not problematic.
Let $\mathcal{X}^d = \mathscr{X}$, and denote the neighborhood of a set element according to a given topology as ∂i. The generic factorization of a generic function $g(\boldsymbol{X}) : \mathscr{X} \to \mathbb{R}$ is intended to be an equation

$$g(\boldsymbol{x}) = \prod_{a \in \mathcal{W}} \psi_a(\boldsymbol{x}_{\partial a}), \quad \psi_a : \mathcal{X}^{|\partial a|} \to \mathbb{R}. \tag{III.1.3}$$

Conversely, a function $g(\boldsymbol{x})$ with an arbitrary factorization can be represented as a factor graph, where edges join a variable i and a factor a if $i \in \partial a$. To represent a distribution $\mu(\boldsymbol{x})$ the Equation is adjusted with a normalization term $\mathcal{Z}$ that sums over the space of $\mathscr{X}$.

Remark III.1.4. *The notion of potentials slightly generalizes conditional probabilities, but is essentially the same. Notice that we are not restricting the ∂a sets to be non-overlapping. The interesting instances will be when the intersections of neighborhoods are non null and $|\partial a| \ll d$. The former condition makes the problem*

[1] This is a very delicate matter. There is a large body of literature focusing on the question of representational power of distributions in terms of graphs. For simplicity, we assume that it is always possible to do so. In future works, this topic will be dealt with in detail.

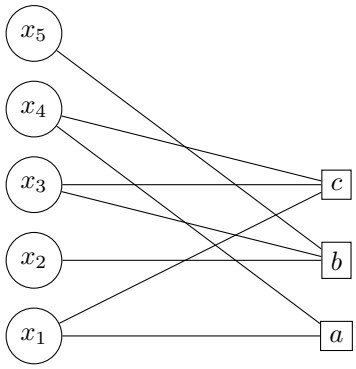


Figure III.1: Factor graph of Example III.1.5

different from the distinct cluster factorization. The latter is useful since potential-level information is non overwhelming. Indeed, the smaller the $|\partial a|$ wrt the total size d, the more local the dependences are.

Example III.1.5. *Consider a function $f : \mathbb{R}^5 \to \mathbb{R}$ which factorizes as:*

$$f(\boldsymbol{x}) = f(x_1, x_2, x_3, x_4, x_5) = f(x_1, x_4) f(x_2, x_3, x_5) f(x_1, x_3, x_4). \tag{III.1.6}$$

We recognize 3 factors, with different neighborhoods. Aligning with the notation introduced we write $f(\boldsymbol{x}) = \psi_a(\boldsymbol{x}_{\partial a})\psi_b(\boldsymbol{x}_{\partial b})\psi_c(\boldsymbol{x}_{\partial x})$. The bijection with the graph of Figure III.1 is evident.

having established a set of rules, we give a definition to our object that is tailored to the scope of this document.

Definition III.1.7 (Probalistic Graphical Model)**.** *A tuple $\mathfrak{G} = (\mu, \mathcal{G})$ where the two elements are in bijection, and are understood as follows.*
A probability density $\mu(\boldsymbol{x})$ on $\mathcal{X}^d$ that has a prescribed factorization into a product of potentials normalized by a common partition function:

$$\mu(\boldsymbol{x}) = \frac{1}{\mathcal{Z}} \prod_{a \in \mathcal{W}} \psi(\boldsymbol{x}_{\partial a}) \cong \prod_{a \in \mathcal{W}} \psi(\boldsymbol{x}_{\partial a}). \tag{III.1.8}$$

An unambiguously associated bipartite graph $\mathcal{G}(\mathcal{V}, \mathcal{W}, \mathcal{E})$ that obeys the same structure, with $|\mathcal{V}| = d, |\mathcal{W}| = n$.

In words, a graphical model is a visualization of the dependence structure of some probability density function. It turns out to be helpful as a tool for performing analysis, being a clever definition of the terms in play. With the term *graphical model*, we will precisely refer to a *factor graph graphical model*, but there are other types such as Bayesian networks, undirected graphical models, Restricted Boltzmann Machines. Its utility is just in terms of highlighting the connections. Physically, it is a proper visualization of how different *local potentials* (factors) cooperate globally at the level of the distribution. Locality is an important and intuitive concept in Physics, that bears the role of making computations easier by neglecting relatively small quantities. In many applications out of the realm of Physical models, it turns out to be adequate, and the techniques designed are passed on seamlessly. The choice of a factorizing distribution as in Equation III.1.3 is then partially justified: it is not the most general measure, but it is appropriate in a vast amount of cases. An equivalent treatment can be carried out if one is willing to consider the factor graph as a *hyper-graph*: we let the variable nodes constitute the nodes and the hyper-edges be the factor nodes. Given that a hyper-edge can join more than two endpoints, the hyper-graph perspective is as valid as the bipartite.

Example III.1.9. *Go back to Example III.1.5 and Figure III.1. The equivalent hypergraph is established by connecting variable nodes with factor edges whenever they are part of the same factor. The result is Figure III.2, where each hyperedge is colored differently and connects more than two nodes.*

Another representation with hyper-edges is Figure III.3, which is literally a bipartite graph. In other words, all are equivalent.

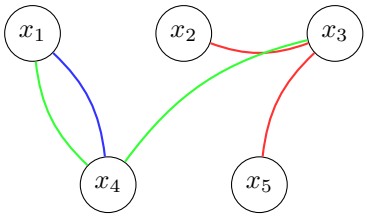


Figure III.2: Factor Hypergraph of Example III.1.9, hyperedges are colored

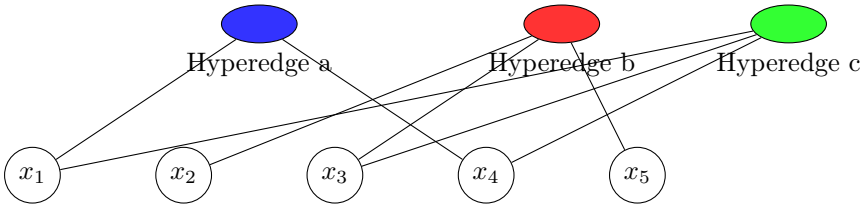


Figure III.3: Factor hypegraph of Example III.1.9

We keep the bipartite graph visualization. The study of random realizations of factor graphs is made easier by collecting objects that share a property under the same *ensemble*. Namely, for a given property, we place into its ensemble each instance of a graph satisfying that property and aim to sample uniformly from this basket. The most common/useful ones are taken from (**mezardInformationPhysicsComputation2009**) and listed below.

- **random k factor graph** $\mathbb{G}_d(k,n)$, where $k \geqslant 1$, d the number of nodes.
  To generate a sample from it, for each factor a choose a k-tuple among the $\binom{d}{k}$ options and assign it to ∂a.
- **non overlapping- k graph** $\mathbb{G}_d(k,\delta)$, where $k \geqslant 1$, n is the number of factors, d the number of nodes, δ a parameter denoting the per-variable rate of factors.
  To generate a sample from it, consider the k-tuples, a collection of $\binom{d}{k}$ elements. Each of them is a factor grouping k variables. Add it to the graph with a weighted probability $\frac{d\delta}{\binom{d}{k}}$.

Remark III.1.10. *It is important to notice that the number of edges in the second case is a random variable, since it depends on the probabilistic assignment of the edges. In expectation it will be* $\mathbb{E}[n] = \binom{d}{k}\frac{d\delta}{\binom{d}{k}} = d\delta$. *Having enumerated all the possible factor nodes among the* $\binom{d}{k}$*, every neighborhood* ∂a *is distinct.*

In the applications we will consider, the size of n, d is large, to obtain the high-dimensional thermodynamic limit. Particularly, we will let $d \to \infty, n \to \infty$ with constant *aspect ratio* $\frac{n}{d} \in \Theta(1)$. For $\mathbb{G}_d(k,n)$ this amounts to requiring that $\delta := \frac{n}{d}$ is constant. For $\mathbb{G}_d(k,\delta)$, the parameter δ needs to be constant. The reasons for seeking answers in this regime are many (recall the discussion in Chapters I, II). Firstly, it is realistic: modern Learning tasks have very wide and tall datasets, with many rows and many columns. Secondly, keeping δ fixed at complexity $\frac{n}{d}$, the number of individuals per feature we see is kept constant across pairs (n,d). This ensures that for any δ-complex distribution of problem instances, we will be able to derive descriptions that only depend on how small/large it is.

Further References

In other words, allowing the δ gap to vary is more complicated. We require it fixed to have a common phenomenology across all (n,d) size pairs. In other lines of research, vanishing dimensions are studied. For examples of "non-constant complexity" models, (**brennanReducibilityStatisticalComputationalGaps2020**) is a good starting point.

In the above ensembles, the degree of factor and variable nodes is always constant at k. If we allow this to vary, we use the local notion of *degree profile*. For a factor graph $\mathcal{G}$, we let:

- V_i be the fraction of variables with degree i,
- F_i be the fraction of factors with degree i.

The objects are unambiguous: the infinite series $\boldsymbol{V} = (V_i)_{i\geqslant 0}$ and $\boldsymbol{F} = (F_i)_{i\geqslant 0}$ are proper distributions over the non-negative integers and take non-zero values in a finite number of positions. Indeed, the former must be such that $\sum_{i\geqslant 0} V_i = d$, while the latter has condition $\sum_{i\geqslant 0} F_i = n$. A useful representation is given by the generating functions[2] of the two distributions. Being discrete, we have that:

$$G_{\boldsymbol{V}}(t) = \sum_{i\geqslant 0} V_i t^i, \quad G_{\boldsymbol{F}}(t) = \sum_{i\geqslant 0} F_i t^i. \tag{III.1.11}$$

Consequently, being that $\boldsymbol{V}, \boldsymbol{F}$ are non-zero in a finite subset of the naturals, the PGFs are finite polynomials, and the moments of the degree profiles[3] (V, F) are derived as:

$$\mathbb{E}[V] = \sum_{i\geqslant 0} V_i i = G_{\boldsymbol{V}}(1), \quad \mathbb{E}[F] = \sum_{i\geqslant 0} F_i i = G_{\boldsymbol{F}}(1). \tag{III.1.12}$$

Example III.1.13. *Consider the ensembles $\mathbb{G}_d(k,n), \mathbb{G}_d(k,\delta)$. Both are such that the degree profile of factors is fixed at k, while the degree profile of variables is a random variable.*

Allowing for more flexibility, we have enough objects to define a *degree-constrained factor graph* ensemble. This will be denoted as $\mathbb{D}_d(\boldsymbol{V}, \boldsymbol{F})$, and is understood as enforcing the distributions $(\boldsymbol{V}, \boldsymbol{F})$ to the variables and factors degree profiles. Sampling from $\mathbb{D}_d(\boldsymbol{V}, \boldsymbol{F})$ is done uniformly across graphs with this structure. Properties and a procedure for generating instances of this ensemble are outlined in (**mezardInformationPhysicsComputation2009**).

III.2 Message Passing Algorithms and Belief Propagation

For the moment, we implement the narrative of (**mezardInformationPhysicsComputation2009**), which presents important results, and aim to outline their presentation in our notation. Assume we wish to compute the marginals μ_j where the randomness of $\boldsymbol{x} \in \mathcal{X}^d$ is represented as a graphical model, and $|\mathcal{X}| < \infty$. This makes the presentation simpler. A sum over all possible configurations is clearly unfeasible, as each space of configurations, even if finite in size, multiplies, leading to a number of operations that grows exponentially in size. Precisely, we would need $|\mathcal{X}|^d$ operations, and if $d \gg 1$ this is hard to perform (recall Rem. III.1.2). Computations become tractable in graphical models that present a tree structure. The solution as a dynamic programming procedure operates constructively from the leaves to the root.
It turns out that this can be seen from the perspective of a message passing algorithm, which was found independently in many fields. Among the early studies, we recognize (**betheStatisticalTheorySuperlattices1935**; **pearlReverendBayesInference1982**; **gallagerLowdensityParitycheckCodes1962**). When speaking of Belief Propagation (BP), we precisely mean the method that originated in Artificial Intelligence, but all can be thought of as equivalent formulations.

[2] akin to the MGF, when the distribution is discrete it is useful to consider the Probability Generating Function (PGF). It is a series $G_{\boldsymbol{p}}(t) := \sum_{i\geqslant 0} p(t) t^i$ for a distribution $p(\cdot)$ and $t \in \mathbb{R}$. Analogously, it can be used to derive expressions of the moments as $\mathbb{E}[X^k] = \left(t \frac{\partial G_{\boldsymbol{p}}(t)^k}{\partial t^k} \right) \Big|_{t=1^-}$. The relationship with the MGF is easily derived as $G_{\boldsymbol{p}}(t) = M_{\boldsymbol{p}}(\ln t)$.

[3] seen as random variables arising from the distribution over nodes

III.2.1 General Framework

All the algorithms we will see can be thought of as instances of the same class of algorithms.

Definition III.2.1 (Message-Passing Algorithm Structure)**.** *For a given factor graph $(\mu, \mathcal{G}(\mathcal{V}, \mathcal{W}, \mathcal{E}))$, a general message-passing algorithm works on a modified directed bipartite multi-graph $\mathcal{G}_{\rightarrow}(\mathcal{V}, \mathcal{W}, \mathcal{E}_{\rightarrow})$ with:*

- *an alphabet $\mathscr{M}$ of messages that change over time. For each message, there is one directed edge.*
- *a set of update functions taking information from neighbors and modifying the local messages, namely:*

$$\Lambda_{i\to a} : \mathscr{M}^{|\partial i \setminus a|} \to \mathscr{M} \quad \Gamma_{a\to i} : \mathscr{M}^{|\partial a \setminus i|} \to \mathscr{M} \tag{III.2.2}$$

- *an initialization, possibly random, seen as a function from $\mathcal{E}_{\rightarrow}$ to $\mathscr{M}$ for each of the multi-edges.*
- *a decision rule, seen as a function $\widehat{\Lambda}_i$ from $\mathscr{M}^{\partial i \setminus a}$ to the space of decisions.*

Given that the graph is bipartite, we can unambiguously refer to a message and an associated edge as $\nu_{i\to a}^{(t)}, \nu_{a\to i}^{(t)} \in \mathscr{M}$. Moreover:

- *The initialization is required because we eventually want to build an algorithm.*
- *The decision rule returns a local estimate of the quantity of interest for a given site/variable i. This justifies our notation, since the domain is the available information to a given site (i.e. its neighborhood).*
- *The codomain could be any, but in most of the cases will be $\mathscr{P}(\mathcal{X})$, the set of probability distributions where the marginal lies.*
- *The hat exemplifies the notion of estimation.*

Given this rather general set of rules, we can envision many procedures:

- **Parallel procedures**, which update at the same time t all the variable and factor nodes according to their neighboring messages, and return an estimate via the decision rule at some later time $t' > t$.
- **Sequential procedures**, which are carried out similarly, with the difference that at each t only one directed edge $\vec{e} \in \mathcal{E}_{\rightarrow}$ is chosen (uniformly).
- Basically any procedure involving a choice of updates, with its advantages and disadvantages to be considered.

In particular, we will be interested in random distributions over targets, given the scenario which we described in Chapter II. This motivates the introduction of a very specific object.

Definition III.2.3 (Random Graphical Model)**.** *Given a target $\boldsymbol{x} \in \mathbb{R}^d$, a random graphical model is a random distribution with structure. We denote it as $(\mu, \mathcal{G}, \mathscr{D})$. Precisely, we take it to be defined up to normalization by interaction potentials of factors grouping variables and local potentials of variables. Denoting these as $\psi_a : \mathcal{X}^{\partial a} \to \mathbb{R}, \psi_i : \mathcal{X} \to \mathbb{R}$, we assume that the graph $\mathcal{G}$ is a randomly sampled from $\mathbb{G}_d(k, \delta)$ or $\mathbb{G}_d(k, n)$. Eventually, the distribution reads*

$$\mu(\boldsymbol{x}) = \prod_{a \in \mathcal{W}} \psi_a(\boldsymbol{x}_{\partial a}) \prod_{i \in \mathcal{V}} \psi_i(x_i). \tag{III.2.4}$$

The interaction potentials are random and sampled according to a precise scheme. Namely for any given degree k (i.e. neighborhood size) we make available a distribution over disorder $\mathbb{P}_{\widehat{\mathscr{D}}}^{[k]}$ from which we sample a k sized disorder instance $\widehat{\mathscr{D}}_k$ and set $\psi_a(\cdot) = \psi^{(k)}(\cdot, \dots, \cdot; \widehat{\mathscr{D}}_a)$ for $|\partial a| = k$.
The same is done for the variable potentials, which require only a one dimensional sample termed $\mathscr{D}_i$.

Remark III.2.5. *The random graphical model is rather complicated, but just because many objects are to be listed. We need a graph, a set of interaction potentials, a set of variable potentials, and rules to choose them randomly. The structure of the graph induces the set of possible realizations of the potentials.*

One instance of a procedure to make inference on a random graphical model is BP, which builds on an attempt to describe locally the problem.
To present it, we will take a constructive approach, and start from a Tree Graphical model that has non-random distribution.

III.2.2 Deterministic Trees

Let the distribution be:

$$\mu(\boldsymbol{x}) = \frac{1}{\mathcal{Z}} \prod_{i=1}^{M} \psi_a(\boldsymbol{x}_{\partial a}) \cong \prod_{i=1}^{M} \psi_a(\boldsymbol{x}_{\partial a}), \tag{III.2.6}$$

where the $\cong$ symbol it to be intended as *up to normalization.* The following Theorem is easily proved by induction on the number of factors.

Theorem III.2.7 (Thm. 14.2 in (**mezardInformationPhysicsComputation2009**))**.** *Consider a tree graphical model $\mathfrak{G} = (\mathcal{G}, \mu)$. Then:*

$$\mu(\boldsymbol{x}) = \prod_{a \in \mathcal{W}} \mu_a(\boldsymbol{x}_{\partial a}) \prod_{i \in \mathcal{V}} \mu_i(x_i)^{1-|\partial i|}, \tag{III.2.8}$$

where the terms on the LHS are marginals of factors and variables.

The importance of the Theorem above is that we are now motivated to find a nice expression of the variable marginals. Doing so, we will embark on a slightly longer route via to design a suitable message-passing Algorithm.

Messages are probability distributions as per Def. III.2.1. The idea is that at each t:

- $\nu_{i \to a}^{(t)}$ attempts to describe the marginal of a graphical model that removed the factor a
- $\widehat{\nu}_{a \to i}^{(t)}$ attempts to describe the marginal of a graphical model that removed all factors $b \in \partial i$ apart from a.

As per (**mezardInformationPhysicsComputation2009**), the BP equations are:

$$\nu_{j \to a}^{(t+1)} \cong \prod_{b \in \partial j \setminus a} \widehat{\nu}_{b \to j}^{(t)}(x_j) \tag{III.2.9}$$

$$\widehat{\nu}_{a \to j}^{(t)}(x_j) \cong \sum_{\boldsymbol{x}_{\partial a \setminus j}} \psi_a(\boldsymbol{x}_{\partial a}) \prod_{k \in \partial a \setminus j} \nu_{k \to a}^{(t)}(x_k), \tag{III.2.10}$$

with the conventions that:

- $\partial j \setminus a = \varnothing \implies \nu_{j \to a}(x_j)$ is uniform
- $\partial a \setminus j = \varnothing \implies \widehat{\nu}_{a \to j}(x_j) = \psi_a(x_j)$.

Given an initialization $(\boldsymbol{\nu}^{(0)}, \widehat{\boldsymbol{\nu}}^{(0)})$, we then term:

- a **BP fixed point**, a collection of $2|\mathcal{E}|$ messages (i.e. every edge of the multigraph) that satisfy Eqns. III.2.9, III.2.10 with time independence. In other terms, a collection of expressions that when iterated with the BP equations does not change.
- a **t-time estimate** of the local marginal, the product of all incoming messages from neighboring factors. Mathematically:

$$\widehat{\mu}_i(x_i) \coloneqq \nu_i^{(t)}(x_i) \cong \prod_{a \in \partial i} \widehat{\nu}_{a \to i}^{(t-1)}(x_i). \tag{III.2.11}$$

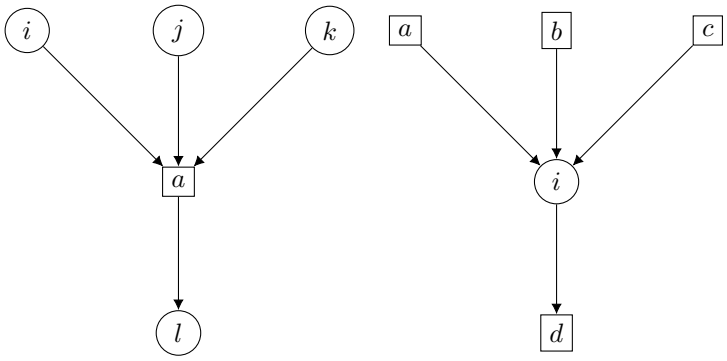


Figure III.4: Messages Cartoon

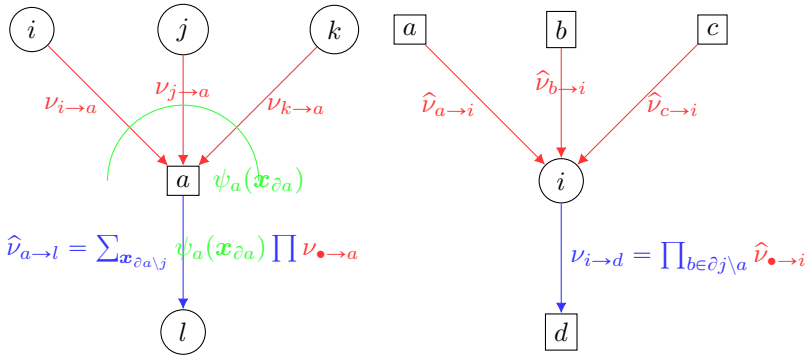


Figure III.5: Detailed Messages Cartoon
A figurative intepretation of Equations III.2.9 on the left, and III.2.10 on the right.

Algorithm 1 Belief Propagation (parallel, generic)

Inputs: graphical model $\mathfrak{G} = (\mathcal{G}, \mu)$ with associated directed graph $\mathcal{G}_{\rightarrow}$ and a a set of local potentials $\boldsymbol{\psi}$ for each node.
Hyperparameters: desired accuracy ϵ, iterations T, starting distribution p, a rule for the maximum message change calculation.

Init: let the BP messages be sampled iid from p
for $t \in \{0, \dots, T\}$ **do**
  for edges (a, j) in $\mathcal{E}_{\rightarrow}$ **do**
    compute $\widehat{\nu}_{a\to j}$ with Eqn. III.2.10
  end for
  for edges (j, a) in $\mathcal{E}_{\rightarrow}$ **do**
    compute $\nu_{j\to a}$ with Eqn. III.2.9
  end for
  $\Delta \leftarrow$ maximum message change at time t
  if $\Delta < \epsilon$ **then**
    return $(\boldsymbol{\nu}^{(t)}, \widehat{\boldsymbol{\nu}}^{(t)})$
  end if
end for
return Warning: no convergence

Accordingly, we reach the formulation of an example algorithm that operates under these principles. The choice of starting with a tree model is justified by the following Theorems, which will also serve to present the Thermodynamic counterpart of this algorithm.

Theorem III.2.12 (BP is exact on Trees)**.** *Consider a tree-graphical model with maximum distance between variables d_{max} (aka diameter). Then:*

1. *for any intialization p, Algorithm 1 converges after $< d_{max}$ iterations*
2. *the estimates built as in Eqn. III.2.11 are exact at convergence.*

Proof. Statements #1, #2 are found in (**mezardInformationPhysicsComputation2009**) □

As largely discussed in our main reference, (**mezardInformationPhysicsComputation2009**) this is not the only guarantee on Trees, which present a long list of peculiarities. We briefly outline them below to have a sufficient amount of information for the next steps.

Being able to compute marginals of single variables, joints over collections of m variables are in general hard to adapt. The updated model that conditions on one variable is:

$$\mu(\boldsymbol{x} \mid x_i = x) \cong \prod_{a=1}^{M} \psi_a(\boldsymbol{x}_{\partial a}) \mathbb{1}_{\{x_i = x\}}. \tag{III.2.13}$$

But aiming to do so in $|\mathcal{X}|^{m-1}$ possible assignments of m variables appearing as indicators is unfeasible. Fortunately, tree models admit a nicer expression also in this case (**mezardInformationPhysicsComputation2009**).
Given this shortcut, terms such as the entropy the average energy and the free energy/entropy can be computed and expressed as sums of local BP messages via a clever application of Thm. III.2.7.

Remark III.2.14. *Crucially, the structural equations for some macroscopic quantities such as [4]$(\mathscr{U}(\mu), \mathscr{S}(\mu), \mathscr{F}(\mu))$ are dependent on the distribution μ, but the results derived from BP provide expressions in terms of messages $(\boldsymbol{\nu}, \widehat{\boldsymbol{\nu}})$ at convergence or at time t as an estimate.*

More importantly than other quantities, we get an expression of the free entropy as a *functional* of the messages, which we report below for convenience.

Definition III.2.15 (Bethe Free Entropy)**.** *Consider a graphical model $\mathfrak{G}$ and its message passing graph $\mathcal{G}_{\rightarrow}$ with a set of $2|\mathcal{E}|$ messages $\vec{\boldsymbol{\nu}} := (\boldsymbol{\nu}, \widehat{\boldsymbol{\nu}})$. The Bethe Free entropy is defined as:*

$$\mathscr{F}_{\mathsf{Bet}}(\vec{\boldsymbol{\nu}}) := \sum_{a \in \mathcal{W}} \ln \mathcal{Z}_a(\vec{\boldsymbol{\nu}}) + \sum_{i \in \mathcal{V}} \ln \mathcal{Z}_i(\vec{\boldsymbol{\nu}}) - \sum_{(i,a) \in \mathcal{E}} \ln \mathcal{Z}_{ia}(\vec{\boldsymbol{\nu}}), \tag{III.2.16}$$

where

$$\mathcal{Z}_i(\vec{\boldsymbol{\nu}}) = \sum_{\boldsymbol{x}_{\partial a}} \psi_a(\boldsymbol{x}_{\partial a}) \prod_{i \in \partial a} \nu_{i \to a}(x_i) \tag{III.2.17}$$

$$\mathcal{Z}_a(\vec{\boldsymbol{\nu}}) = \sum_{x_i} \prod_{b \in \partial i} \widehat{\nu}_{b \to i}(x_i) \tag{III.2.18}$$

$$\mathcal{Z}_{ia}(\vec{\boldsymbol{\nu}}) = \sum_{x_i} \nu_{i \to a}(x_i) \widehat{\nu}_{a \to i}(x_i). \tag{III.2.19}$$

Roughly, it can be interpreted as a sum of logarithms of partition functions that refer to models that ignore part of the graph. The edges partition functions at the end are removed to avoid double counting.

Given the above Theorem and the comments, largely discussed in (**mezardInformationPhysicsComputation2**
a very important conclusion is the following.

[4] respectively, the internal energy, the canonical entropy and the free energy.

Theorem III.2.20 (Bethe Free entropy is exact on trees, Thm. 14.3 (**mezardInformationPhysicsCom**
For a graphical model $\mathfrak{G} = (\mu, \mathcal{G})$ and associated $\mathcal{G}_{\rightarrow}$ it holds $\mathscr{F}(\mu) = \mathscr{F}_{\mathsf{Bet}}(\vec{\boldsymbol{\nu}}_\star)$, for $\vec{\boldsymbol{\nu}}_\star$ the messages at convergence of Algorithm 1.

For graphical models, this gives rise to a simple sampling algorithm that builds recursively the distribution μ starting from a single marginal.

Example III.2.21. *A very important instance of tree graphical models is that of **pairwise models**, where each variable and factor node has degree 2. These can be seen as a generalization of the Ising model, and the factors are absorbed into the edges of a real graph. Namely, the structure of $\mathcal{G}$ induces a graph $\mathcal{G}_{\rightarrow}$ which is not a proper multigraph since the degree 2 restriction makes each factor join only two variables. Then, we can describe this as a simple graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ where edges (i, j) are present whenever there is a factor constraining them. This simplifies greatly the expressions of Eqns. III.2.9, III.2.10. Precisely, we can disregard one set of messages and iterate over the variable i to edge (i, j) messages, termed $i \to ij$, which suffice. The messages are then:*

$$\nu^{(t)}_{i\to ij} = \nu^{(t)}_{i\to j} \cong \prod_{l\in\partial i\setminus j} \sum_{x_l} \psi_{il}(x_i, x_j)\nu^{(t-1)}_{l\to i}(x_l). \tag{III.2.22}$$

*Also the free entropy admits a nicer expression (see (***mezardInformationPhysicsComputation2009***))*

III.2.3 Deterministic Graphs

Despite being designed for marginals of trees, Algorithm 1 is applicable to different problems and general graphs. The former deviations will be avoided for the sake of simplicity, but can be found in (**mezardInformationPhysicsComputation2009**). The latter question is more general and requires some care. An application of BP to non-tree graphical models will be distinguished as a *Loopy* BP inference procedure. Given that the main Theorems of this subsection do not hold in the general setting, the methodology might seem to fall short. There is no guarantee of convergence, of stability of the fixed points, of reliability of the estimation. In answering these questions there is no general formalism, but given the right restrictions, some conclusions can be reached. We will focus on providing a brief comment on:

- the usage of the Bethe free energy as an approximate mean-field like computation
- a deterministic descriptions of the dynamics
- numerical methods for simulations.

Remark III.2.23. *Belief Propagation and the soon to be presented Approximate Message Passing are not variational mean-field methods in the classical sense. Namely, the free energy approximation they perform is not a bound on the true value. However, under certain conditions, it can be shown that this approximation is asymptotically exact. In other words, we are not placing a blanket on the true object of interest, but rather doing a guess that might become precise in the large size limit.*

Bethe Free Entropy Working on a non-tree model makes the Bethe free energy potentially ambiguous. This is because it could be the case that messages do not represent distributions. To avoid this, we restrict the domain of $\mathscr{F}_{\mathsf{Bet}}$ to *locally consistent marginals*. For a given graphical model $\mathfrak{G}$ and associated $\mathcal{G}_{\rightarrow}$ we identify these as the set of collections of distributions over factors and variables $\vec{\boldsymbol{\rho}} = (\boldsymbol{\rho}_{\mathcal{V}}, \boldsymbol{\rho}_{\mathcal{W}}) = (\{\rho_i\}_{i\in\mathcal{V}}, \{\rho_a\}_{a\in\mathcal{W}})$ where $\rho_i \in \mathscr{P}(\mathcal{X})$ and $\rho_a \in \mathscr{P}(\mathcal{X}^{|\partial a|})$ that admit a reasonable description in terms of factor-variable bonds. Mathematically, this means:

$$\sum_{\boldsymbol{x}_{\partial a\setminus i}} \rho_a(\boldsymbol{x}_{\partial a}) = \rho_i(x_i) \quad \forall a \in \mathcal{W}, \quad \forall i \in \partial a. \tag{III.2.24}$$

Definition III.2.25 (Locally consistent marginals)**.** *For a graphical model $\mathfrak{G}$ and its directed graph $\mathcal{G}_{\rightarrow}$ the set of locally consistent marginals is referred to with the symbol $\mathsf{LOC}(\mu, \mathcal{G}, \mathcal{G}_{\rightarrow})$.*

Accordingly, a general Bethe free entropy is seen as a function

$$\mathscr{F}_{\mathsf{Bet}} : \mathsf{LOC}(\mu, \mathcal{G}, \mathcal{G}_{\rightarrow}) \to \mathbb{R} \tag{III.2.26}$$

$$\vec{\boldsymbol{\nu}} \to \mathscr{F}_{\mathsf{Bet}}(\vec{\boldsymbol{\nu}}), \tag{III.2.27}$$

with explicit expression derived in analogy to Definition III.2.15:

$$\mathscr{F}_{\mathsf{Bet}}(\vec{\boldsymbol{\rho}}; \mathsf{LOC}) = -\sum_{a \in \mathcal{W}} \rho_a(\boldsymbol{x}_{\partial a}) \ln \frac{\rho_a(\boldsymbol{x}_{\partial a})}{\psi_a(\boldsymbol{x}_{\partial a})} - \sum_{i \in \mathcal{V}} (1 - |\,\partial i|) \rho_i(x_i) \ln \rho_i(x_i). \tag{III.2.28}$$

If the statistician is willing to add the requirement that marginals are locally consistent, the now workable Bethe free entropy turns out to be a good variational object to derive properties of Algorithm 1.

Proposition III.2.29 (Properties of Bethe Free entropy)**.** *For a graphical model, the following are true statements.*

1. *if for every factor a and neighborhood ∂a it holds $\psi_a(\boldsymbol{x}_{\partial a}) > 0$ there is a bijection between the stationary points of $\mathscr{F}_{\mathsf{Bet}}(\cdot; \mathsf{LOC})$ above and the fixed points of the BP equations Eqns. III.2.9, III.2.10.*
2. *as a consequence, the condition on the potentials is sufficient for the existance of a fixed point of the BP equations.*
3. *the general notion of Bethe free entropy has the following properties:*
   (a) *any stationary point of $\mathscr{F}_{\mathsf{Bet}}(\vec{\boldsymbol{\nu}})$ is a fixed point of the BP equations*
   (b) *any stationary point of the BP equations such that $\mathscr{F}_{\mathsf{Bet}}(\vec{\boldsymbol{\nu}})$ is finite is a fixed point of the latter.*

   In particular, we recognize that #3.(b) is a partial inverse induced by the result of #1, #2, that is easily understood as an existance condition on the logarithm of Eqn. III.2.28. Moreover, the condition being not satisfied would make the graphical model vacuous (one null potential makes the distribution have weight zero on those configurations).

Remark III.2.30. *The free energy evaluated as in Eqn. III.2.28 is convex for tree graphs and graphs with one cycle. Consequently, the Proposition above proves existence and uniqueness of BP fixed points, by existence and uniqueness of the stationary point of the function. In general, $\mathscr{F}_{\mathsf{Bet}}(\cdot; \mathsf{LOC})$ is non-convex and there are no guarantees.*

In reality, when the graph is tree-like with high probability and $n \to \infty$, the BP fixed equations are still robust and able to converge to the marginals.

Example III.2.31. *In a pairwise graphical model the easy structure helps understand better this claim. The message $\nu_{i \to j}(x_i)$ represents a belief in a graph where the factor potential $\psi_{ij}(x_i, x_j)$ was removed. The expression is seen as:*

$$\nu_{i \to j}(x_i) \cong \sum_{\boldsymbol{x}_{\partial i \setminus j}} \prod_{l \in \partial i \setminus j} \psi_{il}(x_i, x_l) \mu_{\partial i \setminus j}(\boldsymbol{x}_{\partial i \setminus j}), \tag{III.2.32}$$

where $\mu_{\partial i \setminus j}(\boldsymbol{x}_{\partial i \setminus j})$ is the joint of $\partial i \setminus j$ variables for removed factors $\psi_{il}(x_i, x_l)$. Namely, we count all possible occurrences of removal of potentials across neighbors but j, to get the removal of j. To be a BP fixed point, a collection[5] $\boldsymbol{\nu}$ must satisfy Eqn. III.2.22. A sufficient condition is:

$$\mu_{\partial i \setminus j}(\boldsymbol{x}_{\partial i \setminus j}) = \prod_{l \in \partial i \setminus j} \nu_{l \to i}(x_l). \tag{III.2.33}$$

The equality is certainly true if either of the following hold for $\{x_l \,:\, l \in \partial i \setminus j\}$:

[5] recall that in pairwise models there is only one type of message.

- *independence in* $\mu_{\partial i\backslash j}(\boldsymbol{x}_{\partial i\backslash j})$
- *the marginals in* $\mu_{\partial i\backslash j}(\boldsymbol{x}_{\partial i\backslash j})$ *coincide with* $\nu_{l\to i}(x_i)$.

In tree graphical models, these are true. In absence of long range correlations and for $n\to\infty$*, they turn out to be true also for locally tree like models (***gallagerLowdensityParitycheckCodes lubyAnalysisLowDensity1998***; **richardsonIntroductionAnalysisIterative2001***). The intuition is that for a given node* i*, the children* (j,j') *are whp distant from each other, with a distance that increases with* n *and correlations that are inversely proportional to the distance. For large sizes, correlations are essentially null and distances are essentially infinite.*

III.2.4 A deterministic description for Random Graphs

For this subsection, we state a collection of assumptions that allow us to derive the ancestor property of State Evolution, which we will review in Subsection IV.2.

Assumption III.2.34 (Structural)**.** *For all* i *variables and* a *factors of the given factor graph the update functions simplify their function form in the following sense:*

(A1) updates on $i\to a$ *depend solely on* $(|\,\partial i|,\mathscr{D}_i)$*, simplifying to:*

$$\Lambda_{i\to a}:=\Lambda_{i\to a}(\{\widehat{\nu}_{b\to i}\ :\ b\in\partial i\backslash a\}):=\Lambda_{i\to a}(\{\widehat{\nu}_1,\dots,\widehat{\nu}_l;\mathscr{D}_i),\tag{III.2.35}$$

where $l=|\,\partial i|-1$ *and* $\{\widehat{\nu}_{b\to i}\ :\ b\in\partial i\backslash a\}=\{\widehat{\nu}_1,\dots,\widehat{\nu}_l\}$.

(A2) updates on $a\to i$ *depend solely on* $(|\,\partial a|,\widehat{\mathscr{D}}_a)$*, simplifying to:*

$$\Gamma_{a\to i}:=\Gamma_{a\to i}(\{\nu_{j\to b}\ :\ j\in\partial a\backslash i\})\equiv\Gamma_{a\to i}(\{\nu_1,\dots,\nu_k;\widehat{\mathscr{D}}_a),\tag{III.2.36}$$

where $k=|\,\partial a|-1$ *and* $\{\nu_{j\to b}\ :\ j\in\partial a\backslash i\}=\{\nu_1,\dots,\nu_k\}$.

(A3) The decision function obeys the same principles.

For a random graphical model $\mathfrak{G}=(\mu,\mathcal{G},\mathscr{D})$ and directed graph $\mathcal{G}_{\to}$, initialize all BP messages $(\nu^{(0)}_{i\to a},\widehat{\nu}^{(0)}_{a\to i})$ at $t=0$ as iid random variables independent of n. In the large size limit $n\to\infty$ we require some further technical assumptions to state the result.

Assumption III.2.37 (Technical)**.** *Consider a Message-Passing Algorithm as in Definition III.2.1. Let:*

(TA1) the messages live in $\mathscr{M}\subseteq\mathbb{R}^d$*, the dataset of randomness for each node be such that* $\mathsf{supp}(\mathbb{P}_{\mathscr{D}})\subseteq\mathbb{R}^d$

(TA2) the update functions $(\Lambda_{i\to a},\Gamma_{a\to i})$ *be continuous wrt the topology of* $\mathbb{R}^d$.

We further define the notion of t-directed neighborhood $\mathsf{N}_{i\to a,t}(\mathcal{G})$ of a message $i\to a$ as the subgraph of nodes that can be reached in t steps from i without using in any step (i,a). We have the following result.

Lemma III.2.38. *Let* $\mathcal{G}$ *be a random factor graph from the* $\mathbb{D}_d(\boldsymbol{V},\boldsymbol{F})$ *ensemble (see Sec. III.1). For any* (i,a) *and any* t

$$\mathsf{N}_{i\to a,t}(\mathcal{G})\overset{n\to\infty}{\underset{\mathsf{d}}{\to}}\mathbb{T}_t(\boldsymbol{V},\boldsymbol{F}),\tag{III.2.39}$$

where the d *convergence is "in distribution" and the RHS is an ensemble over trees with a well-defined structure (see (***mezardInformationPhysicsComputation2009***)).*

Convergence in distribution to an object that is independent of the edge (i,a) is sufficient to state the following.

Theorem III.2.40 (Density Evolution Recursive relation)**.** *This result is best described with a constructive proof in (***mezardInformationPhysicsComputation2009***) with a final Proposition.*
Let Assumptions III.2.34, III.2.37 hold for a random graphical model. Choose an initialization of messages from a common distribution. In other words, let $\nu_{i\to a}^{(0)} \overset{\mathsf{d}}{=} \nu^{(0)}$ *and* $\widehat{\nu}_{i\to a}^{(0)} \overset{\mathsf{d}}{=} \widehat{\nu}^{(0)}$.
For $t \geqslant 0$ *and* $(i,a) \in \mathcal{E}$ *an edge at random from* $\mathcal{G}$*, where* $\mathcal{G}$ *was sampled from* $\mathbb{G}_d(k,n)$ *or* $\mathbb{D}_d(\boldsymbol{V},\boldsymbol{F})$ *defined in Section III.1 the messages converge to well defined random variables:*

$$\nu_{i\to a}^{(t)} \underset{\mathsf{d}}{\overset{n\to\infty}{\to}} \nu^{(t)} \quad \widehat{\nu}_{a\to i}^{(t)} \underset{\mathsf{d}}{\overset{n\to\infty}{\to}} \widehat{\nu}^{(t)}. \tag{III.2.41}$$

Additionally, the random variables obey the following recursion in terms of the update equations:

$$\nu^{(t+1)} \overset{\mathsf{d}}{=} \Lambda_l(\widehat{\nu}_1^{(t)},\dots,\widehat{\nu}_{l-1}^{(t)},\widehat{\nu}_l^{(t)};\mathscr{D}), \quad \widehat{\nu}^{(t)} = \Gamma_k(\nu_1^{(t)},\dots,\nu_{k-1}^{(t)},\nu_k^{(t)};\widehat{\mathscr{D}}_k). \tag{III.2.42}$$

In particular,

- *the first* $l-1$ *and* $k-1$ *terms in the functions are copies of* $\nu^{(t)}$ *and* $\widehat{\nu}^{(t)}$*,*
- *the last terms (in blue) are integers sampled from two distributions* (λ_l,ρ_k)*, which are the degree distribution of a root node in the variables and in the factors (node i included and modulo node i respectively).*
- *the disorder(s)* $(\mathscr{D},\widehat{\mathscr{D}})$ *are distributed as* $(\mathbb{P}_{\mathscr{D}},\mathbb{P}_{\widehat{\mathscr{D}}}^{[k]})$ *respectively.*

Remark III.2.43. *Being very involved as a statement, we provide the most direct consequence in plain words. Instead of computing all the single messages, for large sizes, it can be concluded that each message has the same* ***scalar*** *distribution. Consequently, the dynamics are completely described by the iterations of* $(\nu^{(t)},\widehat{\nu}^{(t)})$*. This is a considerable dimensionality reduction from* $\omega(n)$ *computations to* $O(1)$ *computations at each time step.*

As we will later present in Section IV.2, and Chapter V, this is nothing but the primordial version of the deterministic description at the thermodynamic limit of Approximate Message Passing. Remarkably, this algorithm is proved/conjectured to be optimal for a large class of inference problems. As BP under suitable conditions presents a simplified description of its dynamics, Approximate Message Passing will. Before presenting it, we close our remarks on BP, discussing how to simulate density evolution equations. Then, we take a constructive approach and showcase the physics-inspired steps that lead to Approximate Message Passing. In a second moment, we will see it can also be defined in a completely Statistical framework, which however might lack the effort to get to its formulation.

III.2.5 Simulations

Despite their simplicity, the density evolution equations are not solvable for most tasks. We briefly outline a technique to provide numerical estimates of $(\nu,\widehat{\nu})$. Again, given the diverse set of origins of BP-type iterations, the Algorithm comes under different names. In Coding Theory, it appeared as "*sampled density evolution*" or "*Monte Carlo method*". In Statistical Physics, the term used is "*population dynamics*". All are self-explanatory of one of the principles behind the idea:

- we will approximate $(\nu,\widehat{\nu})$ with a large population of N samples
- eventually, it is a Monte Carlo sampling scheme.

Given the stochastic nature of the method, it is important to describe how to deal with the various divergent parameters and how to judge if the chain is *well-behaved*. Below, we collect some general ideas.

Remark III.2.44. *Returning the final vectors* $(\boldsymbol{\nu}_{\mathsf{emp}}^{(T)},\widehat{\boldsymbol{\nu}}_{\mathsf{emp}}^{(T)})$ *is more general. The statistician can compute various quantities dependent on the density evolution random variables. If only the value of the density evolution equations was needed, the empirical distribution of the output would have sufficed.*

Algorithm 2 Population Dynamics, generic

Also termed: Monte Carlo method/Sampled Density Evolution.
Inputs: graphical model $\mathfrak{G}$ with associated directed graph $\mathcal{G}_{\rightarrow}$.
Precisely the (ρ, λ) distributions of Thm. III.2.40.
The noise distributions of variables $\mathbb{P}_{\mathscr{D}}$ and factors $\mathbb{P}^{[k]}_{\widehat{\mathscr{D}}}$.
Hyperparameters: population size N, maximum iterations T

Init: let the initial estimates $\{\nu_i^{(0)}\}$ be sampled iid from p
for $t \in \{0, \dots, T\}$ **do**
  for $i = 1, \dots, N$ **do**
    draw $k \sim \rho$ ▷ an integer
    draw $i_{(1)}, \dots, i_{(k-1)}$ uniformly from $\{1, \dots, N\}$
    draw $\widehat{\mathscr{D}}_k \sim \mathbb{P}^{[k]}_{\widehat{\mathscr{D}}}$
    $\widehat{\nu}_i^{(t)} \leftarrow \Gamma_k(\nu_{i_{(1)}}^{(t-1)}, \dots, \nu_{i_{(k-1)}}^{(t-1)}; \widehat{\mathscr{D}}_k)$
  end for
  for $i = 1, \dots, N$ **do**
    draw $l \sim \lambda$ ▷ an integer
    draw $i_{(1)}, \dots, i_{(l-1)}$ uniformly from $\{1, \dots, N\}$
    draw $\mathscr{D} \sim \mathbb{P}_{\mathscr{D}}$
    $\nu_i^{(t)} \leftarrow \Lambda_l(\nu_{i_{(1)}}^{(t-1)}, \dots, \nu_{i_{(l-1)}}^{(t-1)}; \mathscr{D})$
  end for
end for
return $\boldsymbol{\nu}^{(T)}_{\mathsf{emp}} = \{\nu_i^{(T)}\}_{i=1}^N, \widehat{\boldsymbol{\nu}}^{(T)}_{\mathsf{emp}} = \{\widehat{\nu}_i^{(T)}\}_{i=1}^N$ ▷ as empirical distributions

Remark III.2.45. *Algorithm 2 is generic since one could have chosen a different way of storing the chain to inspect its stability.*

Important Remarks The first sampling is done according to the p distribution that was used to initialize BP in Algorithm 1. At any finite iteration T, the following limit is a consequence of the Glivenko-Cantelli Theorem (see Subsec. A.4 for a complete discussion):

$$\lim_{N\to\infty} \mathsf{emp}(\{\nu_i\}^{(T)}_{1,\dots,N}) \rightrightarrows \nu^{(T)}, \quad \lim_{N\to\infty} \mathsf{emp}(\{\widehat{\nu}_i\}^{(T)}_{1,\dots,N}) \rightrightarrows \widehat{\nu}^{(T)}, \tag{III.2.46}$$

where $\mathsf{emp}(\cdot)$ is the empirical distribution and $\rightrightarrows$ denotes uniform convergence in the argument x (Def. A.1.1). However, we also need to let the dynamics converge, which requires taking the $T \to \infty$ limit. In this case, the order of taking these operations matters.

If one takes $\lim_{T\to\infty} \lim_{N\to\infty}$, the order is not relevant for our purposes, since we reduce the problem to dealing with the previous $(\nu^{(t)}, \widehat{\nu}^{(t)})$ random variables that are assumed to be hard to solve.
On the contrary, the limit $\lim_{N\to\infty} \lim_{T\to\infty}$ needs some care. If the fixed point is unique, the time evolution will unambiguously lead to the right fixed point, and the subsequent $N \to \infty$ application will just *drive* it to the non-empirical fixed point. Removing this assumption, consider the Markov Chain of iterates to be convergent to a unique stationary distribution[6]. How can Population Dynamics reach the multiple fixed-points behavior of density evolution? Empirically, it appears that this unique convergence happens only if $T \equiv T(N)$ grows a lot faster than N. For an appropriate choice of scaling, the convergence will get to *quasi-fixed-points* that are the fixed points of Density Evolution.

To check if this scenario is verified, one can inspect averages of test functions over distributions. If the dynamics reached a quasi-fixed-point, it is expected that the averages at fixed time of an N population present stationarity up to order $O\left(\frac{1}{\sqrt{N}}\right)$ population level fluctuations (**mezardInformationPhysicsComputation2009**). In other words, we expect that averages of test functions do not exceed finite population size effects, which would mean that there is a time-dependent effect acting

[6] assume the simplest setting for this to hold

on the sampling scheme. Mathematically, for a sequence of samples after burn-in $t_\star$, evaluate:

$$\frac{1}{N}\sum_{i=1}^{N}\xi(\nu_i^{(t)}) \quad \xi : \mathscr{P}(\mathcal{X}) \to \mathbb{R}, \quad t \in \{t_\star, \dots, t_\star + K\} \tag{III.2.47}$$

and derive a judgement from the behavior of the time indexed list of expectations. Ideally, they should have *small* variance wrt N (say $O\left(\frac{1}{\sqrt{N}}\right)$ as above).
The same method can be used to derive estimates of arbitrary functions. If $\xi : \mathscr{P}(\mathcal{X})^{\otimes l} \to \mathbb{R}$, a function of iid copies of the distributions, the formula reads:

$$\widehat{\xi}^{(t)} = \frac{1}{R}\sum_{r=1}^{R}\xi(\nu_{i_{r(1)}}^{(t)}, \dots, \nu_{i_{r(l)}}^{(t)}), \tag{III.2.48}$$

where $(i_{r(1)}, \dots, i_{r(l)})$ are independent samples from the N-sized empirical estimate of the population at time t (R is a shortcut for *approximate* replicas). For good results, one typically takes $R \in \Theta(N)$ and $l \ll N$, so that the empirical bias is small and the averaging pool is large enough. In simple words, for each N population at time t one takes many samples of a function requiring a relatively small number of inputs.

Applicative power Being that the messages expressions are strongly dependent on the underlying graph and the problem structure, the general framework largely extends for specific models. Tailored results always allow for a more complete description. Some examples are in (**mezardInformationPhysicsComputation2009**), (**krzakalaStatisticalPhysicsMethods2021**), (**zdeborovaStatisticalPhysicsInference2016**). An extensive account of the main references up to 2009 is given in the closing Notes of (**mezardInformationPhysicsComputation2009**).

Further References

The Bethe free entropy can be seen as the first level of a hierarchy of approximations. This concept is nicely explained in (**yedidiaBetheFreeEnergy2001**). Recently (**weinKikuchiHierarchyTensor2019**) it was found to be connected to Sum of Squares (SoS) proofs, a technique originating in Computer Science.

III.3 Describing the measure via phases

Having explained some features of this method, we now briefly present a classification of the hardness of a problem from a different peerspective than Sec. I.3.1 and Sec. II.6.1, which however bears the same underlying ideas. We do so to actually explain that this classification will not be needed in the Bayes-Optimal setting. Despite this, we choose to discuss it for the sake of clarity. Given the difficulty and vastness of the matter, it is strongly suggested to refer to the material in the reference box below. In the future, this topic will be properly expanded in an independent document.

Further References

For a formal and longer treatment, refer to (**mezardInformationPhysicsComputation2009**) and the references in the Notes at the end of Chapter 19 there. Without being exhaustive, the concept of Replica Symmetry Breaking is also discussed in other textbooks (**mezardSpinGlassTheory1986**; **talagrandSpinGlassesChallenge2003**), reviews (**castellaniSpinGlassTheoryPedestrians2005**; **zdeborovaStatisticalPhysicsInference2016**), courses (**krzakalaStatisticalPhysicsMethods2021**).

In the previous sections, we discussed situations in which the probability measure of a certain graphical model presented a structure that allows to derive answers in an algorithmically efficient fashion. Crucially, this corresponded to the assumption of a *locally-tree-like* structure, preventing loops at short distances that (roughly) induce backward correlations between variables. Considering again a factor graph, we introduce some notions before giving an intuition of the object to focus on. With the word **cavity**, we will intend a subset of the variable nodes $\mathcal{C} \subset \mathcal{V}$, which induces a subgraph $\mathcal{G}_{\mathcal{C}} = (\mathcal{C}, \mathcal{V}_{\mathcal{C}}, \mathcal{W}_{\mathcal{C}})$ made of:

- all the variables in $\mathcal{C}$
- all the factors $a \in \mathcal{W}$ such that $\partial a \subset \mathcal{C}$
- all the edges joining factors and variables considered.

Additionally, we say that an edge $(i, a) \in \partial\mathcal{C}$ when $i \in \mathcal{C}$ and $a \notin \mathcal{W}_{\mathcal{C}}$, meaning that it is not included in the induced subgraph, but "almostt" there. As before, we interpret the collection $\{\widehat{\nu}_{a\to i}\}$ as probability measures over $\mathcal{X}$ where the edges in the directed graph $\mathcal{G}_{\to}$ can be taken to be anywhere, not just inside $\mathcal{G}_{\mathcal{C}}$.

Consistency in the BP method for inference can be loosely intended as it granting the existance of a set of messages $\{\widehat{\nu}_{a\to i}\}$ such that a generic probability distribution μ can be expressed within some accuracy by the induced factor graph over cavities. Giving more details (**mezardInformationPhysicsComputation2009**), μ is a **Bethe measure** if there exists $\{\widehat{\nu}_{a\to i}\}$ such that for almost all cavities $\mathcal{C} \equiv \mathcal{C}(n)$ such that $|\mathcal{C}|$ is bounded as $n \to \infty$ it holds that:

$$\mu_{\mathcal{C}}(\boldsymbol{x}_{\mathcal{C}}) \simeq \prod_{a\in\mathcal{W}_{\mathcal{C}}} \psi_a(\boldsymbol{x}_{\partial a}) \prod_{(i,a)\in\partial\mathcal{C}} \widehat{\nu}_{a\to i}(x_i) + \mathsf{err}(\boldsymbol{x}_{\mathcal{C}}), \tag{III.3.1}$$

the last term being an error quantity to be specified, as well as the notion of "almost all". For example, one could choose a vanishing tolerance $\epsilon \equiv \epsilon(n)$ so that the norm of the error is smaller than it, for more than $1 - \epsilon$ cavities with bounded size. Being a particular choice of probabilistic structure, it will necessarily induce some features in the Statistics of the model. For example, the authors in (**mezardInformationPhysicsComputation2009**) mention that under additional conditions on the factor potentials and loops over $\mathcal{G}$, it can be shown that the BP equations are *almost* satisfied by the messages in the above equation, with the notion of almost being again subject to an accuracy. Perhaps more importantly, it is important to notice that while a Bethe Measure requires existance of these approximate solutions, the injectivity of the mapping from quasi solutions to Bethe-measures is not guaranteed (see (**mezardInformationPhysicsComputation2009**)). Keeping this in mind, we formulate a partition of the phase diagram of a problem into the appearence of its distribution:

- Replica Symmetric (RS), when μ is a Bethe measure, or a decomposition into a finite number of Bethe measure with global symmetries (e.g. sign flip)
- dynamic Replica Symmetry Breaking (d-RSB), when μ is a Bethe measure, but there is an exponentially large in n number of them, and μ decomposes into a convex combination of them. Mathematically:

  $$\mu(\boldsymbol{x}) = \sum_m w_m \mu_m(\boldsymbol{x}), \tag{III.3.2}$$

  a sum of exponentially many in n terms with w_m exponentially small weights.
- static Replica Symmetry Breaking (s-RSB), when the same as d-RSB happens, but a finite number of the w_m take $\Theta(1)$ weight and make μ not a Bethe measure.
- full RSB, where a hierarchy appears (see reference box for some ideas).

Clearly, the performance of BP and related conclusions to be taken with it is largely influenced by the object of inference. Roughly, the inductive bias of the Algorithm is just wrongfully placed if the measure is not sufficiently Bethe Measure-like. While many solutions to this have been envisioned, we avoid discussing them

due to space constraints. In addition to this, and even more importantly for us, there is no s-RSB or further in Bayes-Optimal inference, so there can be only RS or d-RSB (**bouchaudOutEquilibriumDynamics1997**). To understand this, see (**zdeborovaStatisticalPhysicsInference2016**) for a justification with a different definition of RSB, or the references therein. Given this, BP can describe the marginals of the RS phase, and can be shown to be accurate also in the d-RSB phase, as argued in (**zdeborovaStatisticalPhysicsInference2016**), vindicating the method when everything but the signal is known. In short, despite the great power of the formalism, we do not need it for our purpose and can proceed ignoring this finer comment.

Chapter IV

Approximate Message Passing

In this Chapter, we provide an instructive derivation of an efficient and powerful approximation of Belief Propagation, which we discussed in Chapter III. Section IV.1 serves the purpose of giving an explanation of the origin of the procedure. To do so, we follow a mid-step, relaxed BP. Section IV.2 discusses the analogue of the density evolution equations (cf. III.2.4) for the newly obtained Algorithm, termed State Evolution.

IV.1 Approximate Message Passing, Physics Intuition

While the physics derivation of Approximate Message Passing (AMP) is very interesting, we restrict ourselves to the basic notions here since it is well treated in literature. The reasons are twofold. First, it is important to give a historical note on where the method originates to understand it. Secondly, the notation in this easier setting will be beneficial for when we will need it in later parts. A nice outline of the steps that led to its formulation can be found in (**zdeborovaStatisticalPhysicsInference2016**). The box below summarizes the essential historical steps.

In Chapter V we will present the recent results of (**fengUnifyingTutorialApproximate2021**; **aubinCommitteeMachineComputational2019**; **tanMixedRegressionApproximate2023**), which encompass the first discoveries and generalized formulations. For the moment, we stay on the Physics intuition side and present a quick derivation of these equations from (**zdeborovaStatisticalPhysicsInference2016**) which is a sketch of the argument in (**kabashimaPhaseTransitionsSample2016**).

Further References

For the connection with the development of compressed sensing, refer to (**zdeborovaStatisticalPhysicsInference2016**). As a technique it appeared in (**donohoMessagePassingAlgorithms2009**) for linear estimation in compressed sensing, i.e. $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}^* + \boldsymbol{\epsilon}$, in (**mezardSpaceInteractionsNeural1989**) for the perceptron and (**kabashimaCDMAMultiuserDetection2003**) for Code Division Multiple Access. The first studies on Generalized Approximate Message Passing date back to (**mezardSpaceInteractionsNeural1989**; **kabashimaBPBasedAlgorithmPerforming2004**), where it was derived for perceptrons. The widely believed conjecture that AMP is optimal among all efficient algorithms started to be inspected after the work of (**braunsteinLearningMessagePassing2006**). Surprinsingly, there are two main connections with earlier iterative procedures. In (**ranganEstimationRandomLinear2010**) it was argued that the work of (**braunsteinLearningMessagePassing2006**) is closely related to a relaxed formulation of BP, while in (**kabashimaCDMAMultiuserDetection2003**) the connection with Thouless-Anderson-Palmer (TAP) equations was established (see (**thoulessSolutionSolvableModel1977**), (**zdeborovaStatisticalPhysicsInference2016**). The popularization of AMP eventually reached its peak with its application to compressed sensing (**donohoMessagePassingAlgorithms2009**), which was followed by formal proofs of its potential (**bayatiUniversalityPolytopePhase2015**; **bayatiDynamicsMessagePassing2011**). Eventually, (**ranganGeneralizedApproximateMessage2012**) brought Generalized AMP to the attention of the research community, underlining that it could work for non-linear estimation models (i.e. GLMs). Being that the whole framework is a philosophy of approach rather than a strict set of rules, many variants of the algorithms appeared in the years. We gloss over their details and opt for presenting only one. A derivation in the style of Statistical Physics argument for Gaussian noise channels is found in (**krzakalaProbabilisticReconstructionCompressed2012**), later generalized by (**kabashimaPhaseTransitionsSample2016**)

For now, we will restrict to a self contained introduction of its origin, to later deal with the fundamental properties in our problem of interest. Every result can be derived with a "Physics-blind" approach in the form of Theorems. For the sake of simplicity, we will focus on a linear estimation problem, which is very close to the model of (**tanMixedRegressionApproximate2023**) that we will study in later Chapters. Specifically, we assume the following DGP:

$$\boldsymbol{y} = \varphi_{\mathsf{out}}\left(\mathbf{A}\boldsymbol{x}^*\right) \quad y_i = \varphi_{\mathsf{out}}\left(\langle \boldsymbol{a}_i, \boldsymbol{x}^* \rangle\right) \quad \forall i \tag{IV.1.1}$$

where φ_{out} plays the role of an output channel with potentially added noise, applied component-wise. The terms $(\boldsymbol{y}, \mathbf{A})$ are observed and φ_{out} is known. The observations are independent. Given such a setting, it is natural to introduce the likelihood of a given observation dataset. By the independence of rows we will use the notation:

$$\mathbb{P}[\boldsymbol{y} \mid \boldsymbol{z}] = \prod_{i=1}^{n} \mathbb{P}_{\mathsf{out}}[y_i|z_i] \quad z_i = \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle \tag{IV.1.2}$$

and the output probability accounts for the structure of φ_{out}. Unless otherwise stated, we also take the prior on the signal to be iid from a one-dimensional distribution. This means that $\mathbb{P}\left[\boldsymbol{X}^*\right] = \prod_{j=1}^{d} \mathbb{P}\left[X_j^*\right] = \mathbb{P}\left[X_1^*\right]^d$. Sometimes, to avoid using the star notation, and put emphasis on the sampling of the true signal, we will use the symbols $\mathbb{P}_{\boldsymbol{X}} = \mathbb{P}_X^{\otimes d}$.

Example IV.1.3 (Noiseless and Noisy phase retrieval)**.** *Let $\varphi_{\mathsf{out}}(z) = |z| + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \Delta)$, and z is the scalar inner product of the two vectors. The formulation*

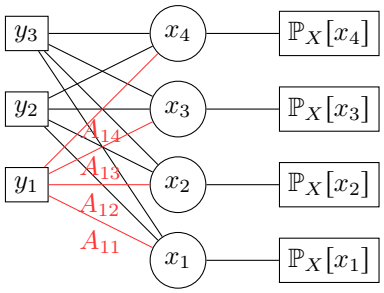


Figure IV.1: Factor graph of a linear estimation problem.
Parameters $(n, d) = (3, 4)$, showing in red only the matrix values of the first sample for readability.

is that of phase retrieval with noise. Removing ϵ, one gets noiseless phase retrieval. In the first case, the likelihood reads:

$$\mathbb{P}[\boldsymbol{y} \mid \boldsymbol{z}] = \prod_{i=1}^{n} \mathbb{P}_{\mathsf{out}}[y_i|z_i] = \frac{1}{\sqrt{2\pi\Delta}} \exp\left\{ -\frac{1}{2\Delta} \left(y_i - \sum_{j=1}^{d} A_{ij} x_j \right)^2 \right\} \tag{IV.1.4}$$

It is possible to recover a standard Bayesian-Planting approach by just placing a prior on the signal.

Approximate Message Passing's first appearances date back to studies on the perceptron (**mezardSpaceInteractionsNeural1989**). However, it was only with its introduction in the compressed sensing setting that it got the current name and found its success (**donohoMessagePassingAlgorithms2009**). For a more general formulation including applications to the model we are considering, one must however use Generalized AMP (G-AMP), which was first introduced in (**ranganGeneralizedApproximateMessag** The Statistical Physics origin of the method is attributed to the theory of TAP Equations (**thoulessSolutionSolvableModel1977**).

Remark IV.1.5. *As pointed out in (***zdeborovaStatisticalPhysicsInference2016***), TAP is for Ising Spins with pairwise interactions (i.e. a 2-body Hamiltonian with discrete Rademacher signals), while AMP is constructed to work with continuous signals and d-body interactions.*

Terminology for the AMP' Physics side Wishing to give a bird's eye view of what lead to AMP from the Physics side is worth but must be taken with great care. Behind the motivation, there are a lot of research works involved. It is useful to introduce some terminology to be sufficiently prepared for their concepts. In our setting, the prior and the observations $\boldsymbol{y}$ play the role of factors. What is variable is the vector we would like to use as estimator ($\widehat{\boldsymbol{x}}$). The matrix entries A_{ij}, specifying the interactions $y_i, \widehat{x}_j$ can be thought of as weights of the edges. Assuming a factorizing prior, a visualization is Figure IV.1. Given a graphical model of this form, it is possible to design a *Belief Propagation* (BP) Algorithm on it to approximate the posterior marginals. The iterations take the form of messages from factors to variables and viceversa. For a factorizing prior the posterior reads

$$\mathbb{P}[\boldsymbol{x} \mid \boldsymbol{y}, \mathbf{A}] = \frac{1}{\mathcal{Z}(\boldsymbol{y}, \mathbf{A})} \prod_{i=1}^{n} \mathbb{P}_{\mathsf{out}}[y_i|z_i] \prod_{j=1}^{d} \mathbb{P}_X[x_j] \quad z_i = \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle \tag{IV.1.6}$$

and the BP messages are a generalization of Eqns. III.2.9, III.2.10 (see (**zdeborovaStatisticalPhysicsInference20** for pairwise models as in Example III.2.31:

$$\nu_{j\to i}(x_j) = \frac{1}{\mathcal{Z}_{j\to i}} \mathbb{P}_X[x_i] \prod_{k\neq i} \nu_{k\to j}(x_j) \tag{IV.1.7}$$

$$\nu_{i\to j}(x_j) = \frac{1}{\mathcal{Z}_{i\to j}} \mathbb{P}_X[x_i] \int \mathbb{P}_{\mathsf{out}}(y_i|\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle) \prod_{r\neq j}^{d} \mathrm{d}[x_r \nu_{r\to i}(x_r)] \tag{IV.1.8}$$

where the integral in red is $d-1$ dimensional. For large signal vectors, this computation is intractable: the only option is perfoming numerical integrations, and no general approach is given by theory.

Remark IV.1.9 (Details about the derivation)**.** *We have effectively changed the domain of signals from discrete to continuous, getting complicated integrals. The first derivation for continuous signals is attributed to (***baronBayesianCompressiveSensing2009***). While* $\boldsymbol{y}$ *is the planted disorder, the matrix* $\mathbf{A}$ *is often described as randomly-quenched. If the matrix entries are* ***independent*** *and the variance is* $O\left(\frac{1}{d}\right)$ *then the model is a* ***mean-field spin glass****, a class of models studied in (***zdeborovaStatisticalPhysicsInference*** Bayes-Optimality and this construction ensure that the fixed point of the BP equations with minimum free energy* $\mathfrak{F}$ *will describe the exact posterior marginals, as we saw previously.*

One option to make the operations feasible is deriving a form of relaxed Belief Propagation (r-BP) (**guoAsymptoticMeanSquareOptimality2006**; **ranganEstimationRandomLi** The general idea is reducing computational overhead by just passing messages that refer to the first two moments, largely simplifying the problem. Intuitively, models that are in some sense Gaussian or close enough to it will not be harmed. We summarize the heuristic with the following observations made in (**zdeborovaStatisticalPhysicsInference20** (**krzakalaStatisticalPhysicsMethods2021**) and (**ranganEstimationRandomLinear2010**).

- the conditioning on y_i is[1] $z_i = A_{ij}x_j + \sum_{r\neq j} A_{ir}x_r$
- by the independence of the x_j, a basic application of the Central Limit Theorem (CLT) gives that:

$$\sum_{r\neq j} A_{ir}x_r \sim \mathcal{N}(\omega_{i\to j}, V_{i\to j}) \quad \omega_{i\to j} = \sum_{r\neq j} A_{ir}a_{r\to i}, \quad V_{i\to j} = \sum_{r\neq j} v_{r\to i} \tag{IV.1.10}$$

  where

$$a_{r\to i} := \int x_r \nu_{r\to i}(x_r)\,\mathrm{d}x_r \qquad v_{r\to i} := \int x_r^2 \nu_{r\to i}(x_r)\,\mathrm{d}x_r - a_{r\to i}^2 \tag{IV.1.11}$$

- ignoring the normalization, the second messages are recasted to a scalar Gaussian integral of the form:

$$\nu_{i\to j}(x_j) \cong \int \mathbb{P}_{\mathsf{out}}\left[y_i|z_i\right] \exp\left\{-\frac{1}{2V_{i\to j}}(z_i - \omega_{i\to j} - A_{ij}x_j)^2\right\}\mathrm{d}z_i \tag{IV.1.12}$$

- expand the square with collected terms $(z_i - \omega_{i\to j}, A_{ij}x_j)$
- perform a Taylor Expansion of the exponential of the second term. With a clever use of the order of the variance of A_{ij} which yields $A_{ij} \in O\left(\frac{1}{\sqrt{d}}\right)$ obtain that

$$\nu_{i\to j}(x_j) \cong \int \mathbb{P}_{\mathsf{out}}\left[y_i|z_i\right]\left(1 + A_{ij}^2x_j^2 - 2(z_i-\omega_{i\to j})A_{ij}x_j + \frac{1}{2}(z_i-\omega_{i\to j})^2A_{ij}^2x_j^2 + o\left(\frac{1}{d}\right)\right) \exp\left\{-\frac{1}{2V_{i\to j}}(z_i-\omega_{i\to j})^2\right\}\mathrm{d}z_i \tag{IV.1.13}$$

- suppressing all the indices, denote for a function $f:\mathbb{R}\to\mathbb{R}$:

$$\mathrm{d}\phi_{\mathsf{out}}^{\mathsf{BP}}(f;\omega,y,V,z) := \frac{f(z)\cdot\mathbb{P}_{\mathsf{out}}[y|z]\exp\left\{-\frac{1}{2V}(z-\omega)^2\right\}}{\int\mathbb{P}_{\mathsf{out}}[y|z]\exp\left\{-\frac{1}{2V}(z-\omega)^2\,\mathrm{d}z\right\}}\mathrm{d}z \tag{IV.1.14}$$

  which **tweaks** f according to the parametrization of the Gaussian $\times$ output measure. Accordingly, defining the **output function** as

$$g_{\mathsf{out}}(\omega,y,V) := \int \mathrm{d}\phi_{\mathsf{out}}^{\mathsf{BP}}\left(\frac{z-\omega}{V};\omega,y,V,z\right), \tag{IV.1.15}$$

[1] this choice looks arbitrary, but recall that the conditioning appears in $\nu_{i\to j}$, thus the choice of isolating ij indices.

it enjoys the property[2]:

$$\int \mathrm{d}\phi_{\mathsf{out}}^{\mathsf{BP}}\left(\left(\frac{z-\omega}{V}\right)^2;\omega,y,V,z\right)=\frac{1}{V}+\partial_\omega g_{\mathsf{out}}(\omega,y,V)+g_{\mathsf{out}}^2(\omega,y,V). \tag{IV.1.16}$$

- Eventually find that the using g_{out}, by rebringing the linearized terms in exponential form, and normalizing, the second message (Eqn. IV.1.8) reduces to the expression

$$\nu_{i\to j}(x_j)=\sqrt{\frac{A_{i\to j}}{2\pi N}}\exp\left\{-\frac{x_j^2}{2d}A_{i\to j}+B_{i\to j}\frac{x_j}{\sqrt{d}}-\frac{(B_{i\to j})^2}{2A_{i\to j}}\right\} \tag{IV.1.17}$$

$$A_{i\to j}=-\partial_\omega g_{\mathsf{out}}(\omega_{i\to j},y_i,V_{i\to j})A_{ij}^2 \quad B_{i\to j}=g_{\mathsf{out}}(\omega_{i\to j},y_i,V_{i\to j})A_{ij}, \tag{IV.1.18}$$

  where ∂_ω denotes the derivative wrt $\omega_{i\to j}$ with no ambiguity.

This procedure gives us a nice expression for the second messages, which in turn yields a compact form to the first messages (Eqn. IV.1.7)

$$\nu_{j\to i}(x_j)\propto\mathbb{P}_X[x_i]e^{-\frac{1}{2\Sigma_{j\to i}}(x_j-R_{j\to i})^2} \tag{IV.1.19}$$

where $(R_{j\to i},\Sigma_{j\to i})$ are the mean and the variance of the exponential density obtained by combining the just derived expressions. Just as in the design of $\phi_{\mathsf{out}}^{\mathsf{BP}}$ define a tweaked measure for ease of notation

$$\mathrm{d}\phi_0^{\mathsf{BP}}(f;\Sigma,R):=\frac{f(x)\cdot\mathbb{P}_X[x]e^{-\frac{1}{2\Sigma}(x-R)^2}}{\int\mathbb{P}_X[x]e^{-\frac{1}{2\Sigma}(x-R)^2}\,\mathrm{d}x}\,\mathrm{d}x \tag{IV.1.20}$$

and two *input functions*

$$f_a(\Sigma,R):=\int\mathrm{d}\phi_0^{\mathsf{BP}}(x;\Sigma,R)\qquad f_v(\Sigma,R):=\Sigma\,\partial_R f_a(\Sigma,R)=\Sigma\,\partial_R\left(\int\mathrm{d}\phi_0^{\mathsf{BP}}(x;\Sigma,R)\right). \tag{IV.1.21}$$

Remark IV.1.22. *The input functions are used to update the local mean and the local variance at the level of a node. They aggregate information from the neighbors and average over the local "Gaussianized" posterior.*

As for BP, it is possible to express this as an **asymptotically exact** iterative algorithm that returns an estimate of the means and variances of the marginals, which in our notation are $\boldsymbol{a}=(a_1,\dots,a_n)$ and $\boldsymbol{v}=(v_1,\dots,v_n)$. The update procedure is listed in Algorithm 3.

Generalized AMP G-AMP can be derived from a further relaxation of Algorithm 3 that leads to the formulation of (**ranganGeneralizedApproximateMessage2012**). It can be proved that the terms of the target node (i.e. where the message is delivered) are weakly interacting with the objects over which the iteration is done. The principle is similar to that of the derivation of the TAP equations (see (**zdeborovaStatisticalPhysicsInference201**
for a comparison). What happens is that it is possible to group all the weakly interacting terms into an **Onsager correction term** (**thoulessSolutionSolvableModel1977**). The exactness of marginals estimation with BP will then guarantee the exactness of marginals estimation of AMP up to the second moment. The key steps just require being careful with neglecting terms up to leading order. It is crucial to define the non-target-dependent versions of our iterators. Fast forwarding some computations, these are:

$$\omega_i^{(t+1)}=\sum_j A_{ij}a_{j\to i}^{(t)} \qquad V_i^{(t+1)}=\sum_j A_{ij}^2 v_{j\to i}^{(t)} \tag{IV.1.23}$$

$$\Sigma_j^{(t+1)}=\frac{1}{\sum_i A_{i\to j}^{(t+1)}} \qquad R_j^{(t+1)}=\frac{\sum_i B_{i\to j}^{(t+1)}}{\sum_i A_{i\to j}^{(t)}} \tag{IV.1.24}$$

[2] we do not prove this, but it is claimed in **zdeborovaStatisticalPhysicsInference2016**

Algorithm 3 Relaxed-Belief Propagation (r-BP)

Input: $(\boldsymbol{y}, \mathbf{A})$

Init: $a_{j\to i}^{(t=0)}, \mathfrak{v}_{j\to i}^{(t=0)}, t \leftarrow 1$

while $a_{j\to i}^{(t)}, \mathfrak{v}_{j\to i}^{(t)}$ not converged **do**

for $i \in [n], j \in [d]$ **do**

$V_{i\to j}^{(t)} \leftarrow \sum_{k\neq j} A_{ik}^2 \mathfrak{v}_{k\to i}^{(t-1)}$

$\omega_{i\to j}^{(t)} \leftarrow \sum_{k\neq j} A_{ik} a_{k\to i}^{(t-1)}$

end for

for $i \in [n], j \in [d]$ **do**

$B_{i\to j}^{(t)} \leftarrow g_{\mathsf{out}}(\omega_{i\to j}^{(t)}, y_i, V_{i\to j}^{(t)}) A_{ij}$

$A_{i\to j}^{(t)} \leftarrow -\partial_\omega g_{\mathsf{out}}(\omega_{i\to j}^{(t)}, y_i, V_{i\to j}^{(t)}) A_{ij}^2$

end for

for $i \in [n], j \in [d]$ **do**

$\Sigma_{j\to i}^{(t)} \leftarrow \dfrac{1}{\sum_{r\neq i} A_{r\to j}^{(t)}}$

$R_{j\to i} \leftarrow \Sigma_{j\to i}^{(t)} \sum_{r\neq i} B_{r\to j}^{(t)}$

end for

for $i \in [n], j \in [d]$ **do**

$a_{j\to i}^{(t)} \leftarrow f_a(\Sigma_{j\to i}^{(t)}, R_{j\to i}^{(t)})$

$\mathfrak{v}_{j\to i}^{(t)} \leftarrow f_{\mathfrak{v}}(\Sigma_{j\to i}^{(t)}, R_{j\to i}^{(t)})$

end for

$t \leftarrow t + 1$

end while

$$\boldsymbol{a}_i^{(t)} \leftarrow f_a\left(\frac{1}{\sum_{r\neq i} A_{r\to j}}, \frac{\sum_{r\neq i} B_{r\to j}}{\sum_{r\neq i} A_{r\to j}}\right)$$

$$\boldsymbol{\mathfrak{v}}_i^{(t)} \leftarrow f_{\mathfrak{v}}\left(\frac{1}{\sum_{r\neq i} A_{r\to j}}, \frac{\sum_{r\neq i} B_{r\to j}}{\sum_{r\neq i} A_{r\to j}}\right)$$

return $(\boldsymbol{a}^{(t)}, \boldsymbol{\mathfrak{v}}^{(t)})$

and one can check that the following chain of approximations can be established with $a_j^{(t)} := f_a(R_j^{(t)}, \Sigma_j^{(t)}), v_j^{(t)} := f_v(R_j^{(t)}, \Sigma_j^{(t)})$:

$$V_i^{(t+1)} \approx \sum_j A_{ij} v_j^{(t)} \tag{IV.1.25}$$

$$\Sigma_j^{(t+1)} \approx \frac{1}{-\sum_i A_{ij}^2 \, \partial_\omega g_{\text{out}}(\omega_i^{(t+1)}, y_i, V_i^{(t+1)})} \tag{IV.1.26}$$

$$R_j^{(t+1)} \approx \Sigma_j^{(t)} \left[\sum_i A_{ij} g_{\text{out}}(\omega_{i\to j}^{(t+1)}, y_i, V_{i\to j}^{(t+1)}) \right] \tag{IV.1.27}$$

$$\text{still messages on second term} \tag{IV.1.28}$$

$$g_{\text{out}}(\omega_{i\to j}^{(t+1)}, y_i, V_{i\to j}^{(t+1)}) \approx g_{\text{out}}(\omega_i^{(t+1)}, y_i, V_i^{(t+1)}) - A_{ij} a_j^{(t)} \, \partial_\omega g_{\text{out}}(\omega_i^{(t+1)}, y_i, V_i^{(t+1)}) \tag{IV.1.29}$$

$$\implies R_j^{(t+1)} \approx a_j^{(t)} + \Sigma_j^{(t+1)} \sum_i A_{ij} g_{\text{out}}(\omega_i^{(t+1)}, y_i, V_i^{(t+1)}) \tag{IV.1.30}$$

$$a_{j\to i} \approx a_j^{(t)} - g_{\text{out}}(\omega_i^{(t)}, y_i, V_i^{(t)}) A_{ij} v_j^{(t)} \tag{IV.1.31}$$

$$\omega_i^{(t+1)} = \sum_j A_{ij} a_j^{(t)} - \sum_j g_{\text{out}}(\omega_i^{(t)}, y_i, V_i^{(t)}) A_{ij}^2 v_i^{(t)} \tag{IV.1.32}$$

$$= \sum_j A_{ij} a_j^{(t)} - V_i^{(t)} g_{\text{out}}(\omega_i^{(t)}, y_i, V_i^{(t)}) \tag{IV.1.33}$$

Thanks to these expressions, we can reformulate the r-BP procedure of Alg. 3 into GAMP. The result is Algorithm 4. It is important to notice that the Onsager correction terms wrt Algorithm 3 are just those in red, and are strongly liked to those arising from the TAP equations derivation (see (**zdeborovaStatisticalPhysicsInference2016**)).

Algorithm 4 Generalized Approximate Message Passing (GAMP)

Input: $\boldsymbol{y}, \mathbf{A}$
Init: $a_j^{(t=0)}, v_j^{(t=0)}, g_{out,i}^{(t=0)} \forall i, t \leftarrow 1$

while $a_j^{(t)}, v_j^{(t)}$ not converged **do**
  for $i \in [n]$ **do**
    $V_i^{(t)} \leftarrow \sum_j A_{ij}^2 v_j^{(t-1)}$
    $\omega_i^{(t)} \leftarrow \sum_j A_{ij} a_j^{(t-1)} - V_i^{(t)} g_{out,i}^{(t-1)}$
  end for
  for $i \in [n]$ **do**
    $g_{out,i}^{(t)} \leftarrow g_{\text{out}}(\omega_i^{(t)}, y_i, V_i^{(t)})$
    $\Sigma_i^{(t)} \leftarrow \frac{1}{-\sum_i A_{ij}^2 \, \partial_\omega g_{\text{out}}(\omega_i^{(t)}, y_i, V_i^{(t)})}$
    $R_j \leftarrow a_j^{(t-1)} + \Sigma_j^{(t)} \sum_i A_{ij} g_{out,i}^{(t)}$
  end for
  for $i \in [n], j \in [d]$ **do**
    $a_j^{(t)} \leftarrow f_a(\Sigma_j^{(t)}, R_j^{(t)})$
    $v_j^{(t)} \leftarrow f_v(\Sigma_j^{(t)}, R_j^{(t)})$
  end for
  $t \leftarrow t + 1$
end while
return $\boldsymbol{a}^{(t)}, \boldsymbol{v}^{(t)} \in \mathbb{R}^d$

Remark IV.1.34. *The computational time of GAMP is $O(nd)$, only matrix operations are performed.*

Remark IV.1.35. *While this derivation sketch is general for factorizing prior and independent observations, it is worth noticing that any model will just require to work out the specific (g_{out}, f_a) functions, with all the rest being equal throughout. We*

call the former the output (denoising) function and the latter the input (denoising) function. A crucial aspect remains being able to neglect terms.

In (**krzakalaProbabilisticReconstructionCompressed2012**), it is shown that the most probable value of the parameters is found by maximizing the partition function, which in turn is equivalent to minimizing (respectively, maximizing) the **Bethe free energy**[3] (Entropy). Such function is directly related to the Belief Propagation messages, with an elegant sum of partitions across variable and factor nodes and an adjustment via the partitions of edges to avoid double counting. Mathematically:

$$\begin{aligned}\mathscr{F}_{\mathsf{Bet}} &= \sum_i \log \mathcal{Z}_i + \sum_j \log \mathcal{Z}_j - \sum_{ij} \log \mathcal{Z}_{ij} \\ \mathcal{Z}_i &= \int \frac{e^{-\frac{1}{2V_i}(\omega_i - z)^2}}{\sqrt{2\pi V_i}} \mathbb{P}_{\mathsf{out}}[y_i|z]\,\mathrm{d}z \\ \mathcal{Z}_j &= \int \prod_i \nu_{i\to j}(x_j)\mathbb{P}_0[x_j]\,\mathrm{d}x_j \\ \mathcal{Z}_{ij} &= \int \nu_{i\to j}(x_j)\nu_{j\to i}(x_j)\,\mathrm{d}x_j.\end{aligned} \tag{IV.1.36}$$

Remark IV.1.37. *This is essentially the same object we found in Definition III.2.15.*

In (**krzakalaVariationalFreeEnergies2014**) via the AMP approximations it is further shown that at the fixed points of the GAMP equations are the stationary points of the logarithm of the posterior likelihood (**krzakalaVariationalFreeEnergies2014**):

$$\mathscr{F}^{\mathsf{GAMP}}_{\mathsf{Bet}}(\{R_j\},\{\Sigma_j\},\{\omega_i\},\{a_j\},\{v_j\}) \tag{IV.1.38}$$

$$= \sum_i \log \mathcal{Z}_i - \sum_j \log \mathcal{Z}(R_j,\Sigma_j) - \sum_j \frac{v_j + (a_j - R_j)^2}{2\Sigma_j^2} - \sum_i \frac{(\omega_i - \sum_j A_{ij}a_j)^2}{2V_i} \tag{IV.1.39}$$

$$V_i = \sum_j A_{ij}^2 v_j \tag{IV.1.40}$$

$$\mathcal{Z}(R,\Sigma) = \int \mathbb{P}_0[x] e^{-\frac{1}{2\Sigma^2}(x-R)^2}\,\mathrm{d}x. \tag{IV.1.41}$$

The result is clearly similar to that of Proposition III.2.29. While this turns out to be a weak condition, since stationarity does not imply that the function is minimized, it is possible to reformulate the optimization in terms of a variational Bethe free energy that satisfies the consistency conditions between the various parameters. This construction eventually leads to a variational expression that Message Passing equations minimize:

$$\mathscr{F}^{\mathsf{GAMP}}_{\mathsf{Bet}}(\{R_j\},\{\Sigma_j\},\{\omega_i^\star\},\{a_j^\star\},\{v_j^\star\}) = \mathscr{F}^{\mathsf{var}}_{\mathsf{Bet}}(\{R_j\},\{\Sigma_j\}), \tag{IV.1.42}$$

where:

$$\begin{aligned}\mathscr{F}^{\mathsf{var}}_{\mathsf{Bet}}(\{R_j\},\{\Sigma_j\}) = &\sum_j \mathsf{d}_{\mathsf{KL}}(\phi_0^{\mathsf{BP}}(\mathbb{P}_X,\Sigma_j,R_j)||\mathbb{P}_X) + \sum_i \mathsf{d}_{\mathsf{KL}}(\phi^{\mathsf{BP}}_{\mathsf{out}}(1,\omega_i,y_i,V_i)||\mathbb{P}_{\mathsf{out}}) \\ &+ \frac{1}{2}\sum_i \log 2\pi V_i^\star + 1 + V_i^\star\,\partial_\omega g_{\mathsf{out}},\end{aligned} \tag{IV.1.43}$$

where $\mathbb{P}_X$ is the prior, $\mathbb{P}_{\mathsf{out}}$ is the probability of the output channel, and the term with KL pedix is the Kullback-Leibler divergence.

[3] an approximation of the free energy to graphs with no cycles (trees), that in many cases is asymptotically exact (**mezardInformationPhysicsComputation2009**; **krzakalaProbabilisticReconstructionCompressed2012**; **krzakalaStatisticalPhysicsMethods2021**). As we saw previously, this happens when the graph is almost cycle-less, meaning that it is locally tree-like.

Remark IV.1.44 (A thorough explanation of Equation IV.1.43)**.** *We roll out the expressions as follows. First of all the notations* $(\phi_{\mathsf{out}}^{\mathsf{BP}}, \phi_0^{BP})$ *were presented in Eqns. IV.1.14, IV.1.20 as differentials. In this case, we naturally express them as probability distributions* $\phi_{\mathsf{out}}^{\mathsf{BP}}(1, \omega_i, y_i, V_i), \phi_0^{\mathsf{BP}}(\mathbb{P}_X, \Sigma_j, R_j)$*. Surprisingly, the latter is also the optimal mean-field approximation that achieves the minimum of the Mean-Field free energy (see (***krzakalaVariationalFreeEnergies2014***)):*

$$\phi_0^{\mathsf{BP}}(\mathbb{P}_X, R_j, \Sigma_j)[x_j] = \frac{1}{\mathcal{Z}(R_j, \Sigma_j)} \mathbb{P}_X[x_j] e^{-\frac{1}{2\Sigma_j}(x_j - R_j)^2}, \tag{IV.1.45}$$

for a coordinate x_j *of interest. In the original publication (***krzakalaVariationalFreeEnergies2014***) the former has a* $\frac{1}{2\pi V_i}$ *factor added, but we just take it inside the normalization. The peculiar result is that when rewritten, the following equality is made more evident:*

$$\mathcal{Z}_i(\omega_i, y_i, V_i)\sqrt{2\pi V_i} = \int \mathbb{P}_{\mathsf{out}}[y_i \mid z] \exp -\frac{1}{2V_i}(z - \omega_i)^2 \, \mathrm{d}z, \tag{IV.1.46}$$

where $\mathcal{Z}_i$ *was defined in Equation IV.1.36 and analogously in Eqn. III.2.17. Accordingly, their Kullback-Leibler divergences greatly simplify (suppressing the indices* j, i*)*

$$-\mathsf{d}_{\mathsf{KL}}(\phi_0^{\mathsf{BP}} || \mathbb{P}_X) = \ln \mathcal{Z}(R, \Sigma) + \frac{\mathfrak{v} + (a - R)^2}{2\Sigma^2} \tag{IV.1.47}$$

$$-\mathsf{d}_{\mathsf{KL}}(\phi_{\mathsf{out}}^{\mathsf{BP}}(1, \omega, y, V) || \mathbb{P}_{\mathsf{out}}) = \log \mathcal{Z} + \frac{1}{2}\left(\log 2\pi V + 1 + V(\partial_\omega g_{\mathsf{out}} + g_{\mathsf{out}}^2)\right). \tag{IV.1.48}$$

Under suitable conditions (to be presented later, see Section V.2), the iterations of GAMP are guaranteed to converge, with conjecturally better MSE than any efficient algorithm. For a discussion of potential extensions to accomodate more sets of assumptions see (**zdeborovaStatisticalPhysicsInference2016**) and the references therein. One of the options is directly minimizing the variational expression of Eqn. IV.1.42 which slows down the dynamics but ensures that the GAMP fixed points will be local minimas.

IV.2 State evolution

The first appeareance of the term was in (**donohoMessagePassingAlgorithms2009**) for a problem of compressed sensing. Rigorous results for AMP and G-AMP were later proved in (**bayatiDynamicsMessagePassing2011**; **javanmardStateEvolutionGeneral2012**; **bayatiUniversalityPolytopePhase2015**). As a concept, it is similar to the self-consistent equations found in the theory of TAP equations with the work of Bolthausen, but is also closely related to density evolution (see (**bolthausenIterativeConstructionSolutions2014** and Subsection III.2.4). Via Statistical Physics techniques, it is derived starting from the BP message equations, which anticipate G-AMP. From there, using the fact that incoming messages are assumed to be conditionally independent, the extensive sums are made of uncorrelated terms. By another application of the CLT, the leading behavior at large sizes is Gaussian, only in terms of the mean and the variance.

A complete physical perspective is the derivation in (**krzakalaProbabilisticReconstructionCompressed2012** There, the authors show that these equations can also be independently derived via the cavity method or with the replica symmetric analysis of the model in the Bayes-Optimal setting, and give more context of the formal similarity with TAP equations. The result is that at $n \to \infty$ the dynamics are described by a **scalar recursion**, with considerable dimensionality reduction.

As a first step, we reparametrize the output distributions as a delta density over the noisy realizations of the input signals:

$$\mathbb{P}_{\mathsf{out}}[y|z] = \int \mathbb{P}_\epsilon(\epsilon) \boldsymbol{\delta}[y - \varphi_{\mathsf{out}}(z, \epsilon)] \, \mathrm{d}\epsilon, \tag{IV.2.1}$$

where φ_{out} is the output function: the channel that only adds noise at the end. The random variables $V_i = \sum_j A_{ij}^2 v_j$ concentrate around their mean (Def. II.2.5) since:

$$\mathbb{E}\left[V_i\right] = \sum_{i'} \frac{v_{i'}}{n} \quad \mathrm{Var}\left[V_i\right] \in o\left(1\right). \tag{IV.2.2}$$

Then, it must be the case that $V^{t+1} = \frac{1}{n}\sum_i v_i$. Additionally, by the BP assumptions, the variables

$$\omega_{i\to j} = \sum_{k\neq j} A_{ij} a_{j\to k} \quad z_{i\to j} = \sum_{k\neq j} A_{ij} x_j^* \tag{IV.2.3}$$

are sums of uncorrelated variables. They can be interpreted as the messages of the estimator (via the ω mean) and the ground truth signal $\boldsymbol{x}^*$. Then, in the limit, they will converge to Gaussian with variance-covariance structure:

$$\mathbb{E}\left[\omega^2\right] = \mathbb{E}\left[a^2\right] = q \quad \mathbb{E}\left[z\omega\right] = \mathbb{E}\left[x^* a\right] = m \quad \mathbb{E}\left[z^2\right] = \mathbb{E}\left[(x^*)^2\right] = \rho. \tag{IV.2.4}$$

Moving to (R_j, Σ_j), one can see that:

$$\frac{R_j}{\Sigma_j} = \sum_i B_{i\to j} = \sum_i A_{ij} g_{\mathsf{out}}(\omega_{i\to j}, y_i, V_{i\to j}) \tag{IV.2.5}$$

$$= \sum_i A_{ij} g_{\mathsf{out}}\left(\omega_{i\to j}, h\left(\sum_{k\neq j} A_{ik} x_k^*, \epsilon_i\right), V\right) + x_j^* \delta \widehat{m} \quad \widehat{m} := \mathbb{E}_{\epsilon,z,\omega}\left[\partial_\omega g_{\mathsf{out}}(\omega, h(z,\epsilon), V)\right] \tag{IV.2.6}$$

$$= \mathcal{N}(0,1)\cdot\sqrt{\delta\widehat{q}} + x_j^*\delta\widehat{m} \qquad \widehat{q} := \mathbb{E}_{\omega,z,\epsilon}\left[g_{\mathsf{out}}^2(\omega, h(z,\epsilon), V)\right], \tag{IV.2.7}$$

where $\delta = \frac{n}{d}$.
With a simpler concentration argument, we also claim that the following holds

$$\frac{1}{\Sigma} = \delta\widehat{q}. \tag{IV.2.8}$$

Eventually the expressions for q, m are reduced to:

$$q = \mathbb{E}_{x*}\left[\mathbb{E}_{R,\Sigma}\left[f_a^2(\Sigma, R)\right]\right] \quad m = \mathbb{E}_{x*}\left[\mathbb{E}_{R,\Sigma}\left[x\right]^* f_a(\Sigma, R)\right] \tag{IV.2.9}$$

where x^* is to be intended as one of the entries of $\boldsymbol{x}^*$, since they are iid.

Bayes-Optimal simplification Without additional assumptions, the State Evolutions equations are written all together as:

$$q = \mathbb{E}_{x*}\left[\mathbb{E}_{R,\Sigma}\left[f_a^2(\Sigma, R)\right]\right] \qquad m = \mathbb{E}_{x*}\left[\mathbb{E}_{R,\Sigma}\left[x\right]^* f_a(\Sigma, R)\right] \tag{IV.2.10}$$

$$\widehat{q} = \mathbb{E}_{\omega,z,\epsilon}\left[g_{\mathsf{out}}^2(\omega, h(z,\epsilon), V)\right] \qquad \widehat{m} = \mathbb{E}_{\epsilon,z,\omega}\left[\partial_\omega g_{\mathsf{out}}(\omega, h(z,\epsilon), V)\right]. \tag{IV.2.11}$$

In the Bayes-Optimal case, by the Nishimori condition, one can add that $q = m$. Unsuprisingly, it can also be proved that $\widehat{q} = \widehat{m}$ (**kabashimaBPBasedAlgorithmPerforming2004**), which is another Nishimori identity. Then, for Gaussian White Noise, the state evolution will eventually involve the MMSE in a very peculiar way (**krzakalaProbabilisticReconstruction**

$$\begin{cases} m^{(t+1)} = \int \mathbb{P}_X[x] \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\xi^2} f_a^2\left(\frac{1}{\delta\widehat{m}^{(t)}}, x + \frac{\xi}{\sqrt{\delta\widehat{m}^{(t)}}}\right) \mathrm{d}\xi\, \mathrm{d}x \\ \widehat{m}^{(t)} = \dfrac{1}{\Delta + \mathbb{E}_X\left[X^2\right] - m^{(t)}}. \end{cases} \tag{IV.2.12}$$

In general, the expression of the second iterator is more difficult (see (**zdeborovaStatisticalPhysicsInfer**
As a morale, we moved from vectorial inference to a scalar recursion that is able to describe exactly how the estimator and the true signal will overlap.

Free Energy The scalar description of the dynamics of AMP allows for a scalar expression of the Bethe free energy, which can be derived as the optimal points of the two objects of iteration or of the MSE of an estimator $E = MSE(\widehat{\boldsymbol{x}}, \boldsymbol{x}^*)$. These are formulations shown in (**zdeborovaStatisticalPhysicsInference2016**). To give greater insight into the more complicated setting of the next Chapter, we choose instead to provide the formulation of (**barbierOptimalErrorsPhase2019**). The lesson is rather simple: in the Bayes-Optimal setting the quenched free energy "splits" at the thermodynamic limit. Here by *splits* we mean that it can be expressed as the optimal configuration over two free energies related to two subproblems of inference, precisely as a sup inf of the two. The meaning of the two subproblems is analogous to the general case of Chapter V and will not be anticipated, but the result is roughly of this nature:

$$\lim_{n\to\infty} f_n = \sup_{r\geqslant 0} \inf_{q\in[0,\rho]} \psi_{\mathbb{P}_0}(r) + \delta\Psi_{\mathbb{P}_{\text{out}}}(q;\rho) - \frac{rq}{2}, \tag{IV.2.13}$$

where:

- ρ is the variance of the signal
- δ is the aspect ratio $\delta = \frac{n}{d}$
- ψ, Ψ are free entropies relative to two retrieval problems described in terms of a zero channel and an output channel.

Recognizing that r is the analogue of $\widehat{m}$, we further establish that the fixed points of the state evolution equations are in correspondence with the stationary points of this free energy, as State Evolution practically implements the following iterations:

$$\begin{cases} m^{(t+1)} = \frac{1}{2}\,\partial_r \psi_{\mathbb{P}_0}(\widehat{m}^{(t)}) \\ \widehat{m}^{(t)} = \delta\frac{1}{2}\,\partial_q \Psi_{\mathbb{P}_{\text{out}}}(m^{(t)};\rho). \end{cases} \tag{IV.2.14}$$

To obtain this result, there are many approaches. A heuristic guess is the replica method (see for example (**krzakalaProbabilisticReconstructionCompressed2012**)). Recently, the adaptive interpolation method rigorized some of the earlier guesses of the replica method (**barbierAdaptiveInterpolationMethod2018**).

Chapter V

Inference on Many Signals

In this final Chapter, we focus on a relevant model of inference, and leverage all the previous discussions to understand the phases of its solvability. To begin, Section V.1 is the bridge between the Physics literature and the Statistics literature, of which we present the formulation in Section V.2. Continuing, Section **??** is the largest portion, where we tackle the problem and its complicacies. While there are precise results for the phase diagram of spectral estimators, not everything is known for phase retrieval. Therefore, we numerically and theoretically inspect where AMP places itself and if it can be helped in all regimes by a spectral estimator. Incidentally, we also argue that the Physics and the Statistics formulation are equivalent up to a change of variable, allowing us to exploit the results from both fields.

So far, we moved from:

1. deterministic tree graph models of a discrete signal
2. deterministic graph models of a discrete signal
3. random graph models of a discrete signal
4. random graph models of a continuous signal.

We now focus on random graphical models of multiple signals on a continuous domain. Having briefly discussed the general ideas behind the passages from BP to r-BP and AMP for the simpler example, we avoid repeating them. The Statistical Physics derivation is always similar: messages are simplified to Gaussian messages that only keep the first two moments. For each stage, it is possible to derive a rigorous set of results that ignores the intuition. As a consequence, this Chapter will have a short Statistical Physics presentation of the framework and a long, yet absolutely non-exhaustive, collection of results with formal Theorems.

V.1 The Committee Machine

We place ourselves in the teacher-student scenario. This is briefly understood with the narrative of a teacher entity that generates the disorder starting from a true signal $\mathbf{X}^*$ that was sampled from a distribution. Remarkably, the signal is **a matrix**, which can be thought of as a collection of weights of a K neurons two layer neural network structure that one wishes to recover. Precisely, the teacher model reads:

$$Y_i = \varphi_{\mathsf{out}}\left(\left\{\frac{1}{\sqrt{d}}\sum_{j=1}^{d} A_{ij}X^*_{jl}\right\}_{l=1}^{K}, \lambda_i\right) \quad \text{or} \quad Y_i \sim \mathbb{P}_{\mathsf{out}}\left(\cdot \mid \left\{\frac{1}{\sqrt{d}}\sum_{j=1}^{d} A_{ij}X^*_{jl}\right\}_{l=1}^{K}\right), \tag{V.1.1}$$

where λ_i is a factor accounting for the noise of the channel. Following standard assumptions, we let $X^*_{jl} \sim \mathbb{P}_0, A_{ij} \sim \mathcal{N}(0,1)$. The teacher hands in a collection of observations $\mathscr{D} = (\boldsymbol{y}, \mathbf{A}) = \{(y_i, \boldsymbol{a}_i)\}_{i=1}^{n}$ to the student that wishes to recover $\mathbf{X}^*$. In Statistical-Physics parlance, $\mathscr{D}$ acts as quenched disorder.

Remark V.1.2. *The model is not exactly a 2-layer neural network but rather includes it as a special case. No details about the noise are given, and one could design different models. In general, it can be thought of as a complicated noisy channel that operates on a set of signals* $\mathbf{X}^*$ *in a well-defined manner.*

Remark V.1.3. *The notation* $\mathbb{P}_0$ *for the prior will largely rely on the difference in appearance between scalars, vectors and matrices. In other words,* $\mathbb{P}_0$ *is the scalar prior,* $\mathbb{P}_{\boldsymbol{X}}$ *is the vector prior, and* $\mathbb{P}_{\mathbf{X}}$ *is the matrix prior. Assuming independence across rows (samples) they all end up being products* $\otimes$ *of each other. Sometimes we will need to sample a vector* $\boldsymbol{X}^*$ *from the distribution over the prior, which we will write as* $\boldsymbol{X}^* \sim \mathbb{P}_0^{\otimes K}$*, by independence and the distribution being identical. When we will wish to emphasize that we want to include the limiting behavior, we will place an overline over the random variable. This is done when studying asymptotic results, often requiring that the actual distribution converges to a well-behaved random variable.*

To perform inference, we are clearly interested in edge cases. In the Teacher-Student setting this is exactly Bayes-Optimal inference, defined as the situation in which the student knows everything but the signal. Let us place ourselves in this scenario: it is evident that no other estimation procedure to retrieve the signal will be better; by logical contradiction, an actor with less information about the model doing better would make no sense. Additionally, the precise questions we ask will be the classic ones, namely:

Recovery

What are the choices of parameters for which it is information-theoretically impossible to learn?

Efficient Recovery

What are the choices of parameters for which it is computationally impossible to learn?

Both are clearly to be answered in the case in which most of the information is given to the student, as the answer will (informally be):

" *For this choice of parameters, no student, not even the best possible, can retrieve the signal.*"

" *For this choice of parameters, no student, not even the best possible, is able to retrieve efficiently the signal.*"

Incidentally, the concept of *choice of parameters* is something that was largely studied in Physics, due to the everlasting presence of phase transition phenomena: sharp changes of behavior that highlight precise phase boundaries such that e.g. above we have impossibility and below we have noticeable feasibility of a problem. We do not discuss much this matter since it is very vast, but just mention that the free entropy is an object that can answer these questions since its *silhouette* can be expressed as a function of some order parameter, to be understood as another object that describes unambiguously the outcome of a statistical procedure. All the previous Chapters touched upon aspects of this, hopefully in a sufficient manner.

Example V.1.4. *Consider a signal and an estimator. A natural notion of performance is the overlap* $m = \frac{\langle \hat{\boldsymbol{x}}, \boldsymbol{x}^* \rangle}{\|\hat{\boldsymbol{x}}\| \|\boldsymbol{x}^*\|} \in [-1, 1]$*. We will see that in our cases the free entropy is directly expressed as a function* $\mathscr{F}(m; \cdot)$*. More in general, a nice form of the free entropy will give access to quantities that depend on the randomness of the problem.*

Crucially, the inspection of the various phases of inference will be described with a phase diagram. We understand it as follows. For simplicity, let the parameters be over the reals. A phase diagram is a cartesian plane where for each axis a parameter is chosen. Then, each point will correspond to a problem instance, and conditioned

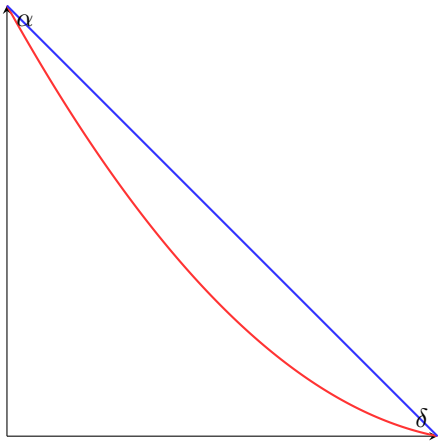


Figure V.1: Typical phase diagram
For example, impossible below the red line, algorithmically hard below the blue line (and above the red line), easy above the blue line. Increasing one parameter, the scenarios are in order. By *order* here we mean two things: (i) all the three phases or a subset of the three are explored, and (ii) this exploration by changing one parameter is always with the same appareance. Namely, by taking constants α_1, α_2 and varying δ the two slices have the same behavior. As an example, consider: impossible, hard, easy for the first α_1. Consider also the same or a subset with the same ordering for the second. It is never the case that for distinct α_1, α_2 the slices appear as hard-impossible-easy and, say, impossible-hard.

on being able to *solve* the questions about it, each point will be labelled as solvable or unsolvable in some sense. Consequently, a collection of regions where the problem is solvable or unsolvable arises.

Remark V.1.5. *It is hard to establish a priori whether the phase diagram will look as a partitioning of the space or as a checkered cartesian plane with alternating hard-easy phases. However, most if not all of the problems of inference and Physics behave like the former: the space is partitioned into phases with exact descriptions and no two phases with the same features are separate. This makes physical sense: once one varies the right parameter (in jargon, an order parameter) it should describe exactly the phases of the problem. In some sense, it is fundamental to understand its phenomenology, and its jumps determine changes of phase in a sequential way, from impossible, to easy to hard without turning back. In other words, a confused scenario such as Figure V.2 is unlikely. For example, if we let ϑ_1 be the SNR, we expect that as the signal increases in magnitude wrt noise, we will get from impossible, to easy to hard, and never back to either of the first two. Given this intuition, we expect a phase plot in the spirit of Figure V.1.*

A formula for the free entropy

A remarkable result of (**aubinCommitteeMachineComputational2019**) is a rigorous formula that describes the behavior of the model under a set of assumptions. With *describe* here we mean that a formula of the free entropy is given, where it is known that the free entropy is sufficient to recover the MMSE of the model.

Let $\mathcal{S}_K^+$ denote the space of p.s.d. matrices in $\mathbb{R}^{K\times K}$, $\mathcal{S}_K^{++}$ for positive definite and for $\mathbf{M} \in \mathcal{S}_K^+$ use $\mathcal{S}_K^+(\mathbf{M})$ for the set of matrices $\mathbf{Q}$ that are p.s.d. and such that also $\mathbf{M} - \mathbf{Q}$ is p.s.d.

Fact V.1.6. *For $K \geqslant 1$:*

1. *$\mathcal{S}_K^+$ is convex*
2. *$\mathcal{S}_K^+(\mathbf{M})$ is convex and compact for any nondegenerate $\mathbf{M} \in \mathcal{S}_K^+$*

Proof. **(Claim #1)** Convexity is trivial in $\mathcal{S}_K^+$.

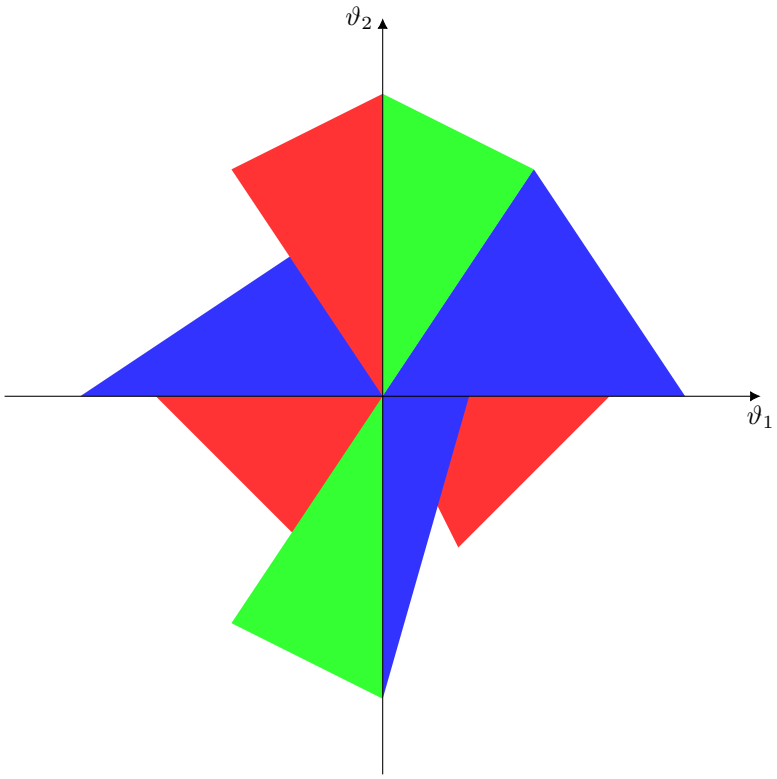


Figure V.2: A Phase diagram that is not expected
Imagine that the green, blue and red regions represent easy, hard and impossible phases. This scenario is not expected.

(Claim #2) we prove convexity by taking $\alpha \in [0,1]$ and $\mathbf{B}_1, \mathbf{B}_2 \in \mathcal{S}_K^+(\mathbf{M})$. Clearly:

$$\mathbf{M}-(\alpha\mathbf{B}_1+(1-\alpha)\mathbf{B}_2) = \alpha\mathbf{M}+(1-\alpha)\mathbf{M}-\alpha\mathbf{B}_1-(1-\alpha)\mathbf{B}_2 = \alpha(\mathbf{M}-\mathbf{B}_1)+(1-\alpha)(\mathbf{M}-\mathbf{B}_2). \tag{V.1.7}$$

By hypothesis, the matrix differences are in $\mathcal{S}_K^+$ and they satisfy the definition of p.s.d. matrices. By convexity of the set of p.s.d. matrices, the convex combination itself $\alpha\mathbf{B}_1 + (1-\alpha)\mathbf{B}_2$ is p.s.d. Consequently, the set $\mathcal{S}_K^+$ is convex.
To prove compactness, we prove boundedness and closedness of the set wrt a matrix norm that is induced by the inner product $\langle \mathbf{B}_1, \mathbf{B}_2 \rangle := \operatorname{Tr}(\mathbf{B}_1\mathbf{B}_2)$ for some matrices $\mathbf{B}_1, \mathbf{B}_2$. Let $(\mathbf{B}_n)_{n\geqslant 0} \subseteq \mathcal{S}_K^+(\mathbf{M})$. This is a consequence of the fact that (i) Matrices are in bijection with euclidean spaces, (ii) K is finite, so the space is finite-dimensional, (iii) by (i) and (ii) the Heine-Borel Theorem can be applied (Thm. A.1.4). To show closedness, we want to show that $\lim_{n\to\infty} \mathbf{B}_n \in \mathcal{S}_K^+(\mathbf{M})$. Inspecting the second condition we find that for $\boldsymbol{x} \in \mathbb{R}^K$:

$$\boldsymbol{x}^\top(\mathbf{M} - \lim_{n\to\infty} \mathbf{B}_n)\boldsymbol{x} = \boldsymbol{x}^\top\mathbf{M}\boldsymbol{x} - \boldsymbol{x}^\top \lim_{n\to\infty} \mathbf{B}_n\boldsymbol{x}. \tag{V.1.8}$$

By continuity of the quadratic form $\boldsymbol{x} \to \boldsymbol{x}^\top\mathbf{B}\boldsymbol{x}$ it holds that:

$$\lim_{n\to\infty} \boldsymbol{x}^\top\mathbf{B}_n\boldsymbol{x} = \boldsymbol{x}^\top \lim_{n\to\infty} \mathbf{B}_n\boldsymbol{x} \quad \forall \boldsymbol{x} \in \mathbb{R}^K, \tag{V.1.9}$$

and closedness is proved since the limiting element belongs to the set. For boundedness, we use the fact that each matrix is p.s.d. and also its difference wrt the pivot $\mathbf{M}$ is p.s.d. If $\mathbf{M}$ is non-degenerate (i.e. not zero or infinity) the eigenvalues are always bounded by those of $\mathbf{M}$. Using the eigendecomposition of matrices, we get that the set $\mathcal{S}_K^+(\mathbf{M})$ is bounded by $\mathbf{M}$ and $\mathbf{0}$. □

Given a model such as that of Equation V.1.1, the object of interest is the posterior. Its expression is easily derived as:

$$\mathbb{P}[\mathbf{X} \mid \mathscr{D}] = \frac{1}{\mathcal{Z}(\mathscr{D}; n)} \prod_{j=1}^{d} \mathbb{P}_0[\boldsymbol{w}_j] \prod_{i=1}^{n} \mathbb{P}_{\text{out}}\left(Y_i \mid \left\{\frac{1}{\sqrt{d}} \sum_{j=1}^{d} A_{ij} X_{jl}^*\right\}_{l=1}^{K}\right) \tag{V.1.10}$$

where $\boldsymbol{w}_j = \left[w_{j,1}, \ldots, w_{j,K}\right]^\top$. Assuming that the free entropy is self-averaging and extensive, the object of interest is the free entropy density. This is standard in Statistical Physics. We are then drawn to find a formula for:

$$f_n = \frac{1}{n}\mathbb{E}_{\mathscr{D}}\left[\ln \mathcal{Z}(\mathscr{D}; n)\right]. \tag{V.1.11}$$

In the classic proportional scaling assumption where $n, d \to \infty$ with $\delta = \frac{n}{d} \in \Theta(1)$ fixed, conjectural formulas using the replica method are confirmed by the Theorem we will showcase, which appeared in (**aubinCommitteeMachineComputational2019**). The morale of the result is that in the high-dimensional limit the true problem is expressed as a combination of two *easier* channels that jointly determine how the original model behaves in a very specific sense. For stability, we add noise to Equation V.1.1, so that:

$$Y_i = \varphi_{\mathsf{out}}\left(\left\{\frac{1}{\sqrt{d}}\sum_{j=1}^{d} A_{ij}X^*_{jl}\right\}_{l=1}^{K}, \lambda_i\right) + \epsilon_i, \quad \forall i \in [n] \tag{V.1.12}$$

where $\epsilon_i \sim \mathcal{N}(0, \Delta)$. This allows us to express the output Y in terms of a noisy channel of a Gaussian density:

$$\mathbb{P}_{\mathsf{out}}(y \mid \boldsymbol{u}) = \frac{1}{\sqrt{2\pi\Delta}}\int_{-\infty}^{\infty} e^{-\frac{1}{2\Delta}\left(y - \varphi_{\mathsf{out}}(\boldsymbol{u}, \lambda)\right)^2} \mathrm{d}\mathbb{P}[\lambda], \quad \boldsymbol{u} \in \mathbb{R}^K. \tag{V.1.13}$$

Following this adjustment, we define two *auxiliary problems* with their respective channels:

- **zero problem** where we are asked to retrieve $\boldsymbol{X}^* \sim \mathbb{P}_0^{\otimes k}$ from observations:

$$\boldsymbol{Y}_0 = \mathbf{R}^{\frac{1}{2}}\boldsymbol{X}^* + \boldsymbol{\epsilon}_0, \quad \boldsymbol{\epsilon_0} \sim \mathcal{N}(\mathbf{0}_K, \mathbf{I}_K), \quad \mathbf{R} \in \mathcal{S}_K^+, \tag{V.1.14}$$

  associated posterior density

$$P(\boldsymbol{w} \mid \boldsymbol{Y}_0) = \frac{1}{\mathcal{Z}_{\mathbb{P}_0}}\mathbb{P}_0[\boldsymbol{w}]\exp\left\{\boldsymbol{Y}_0^\top\mathbf{R}^{\frac{1}{2}}\boldsymbol{w} - \frac{1}{2}\boldsymbol{w}^\top\mathbf{R}\boldsymbol{w}\right\}, \tag{V.1.15}$$

  and free entropy:

$$\psi_{\mathbb{P}_0}(\mathbf{R}) = \mathbb{E}_{\boldsymbol{Y}_0}\left[\ln \mathcal{Z}_{\mathbb{P}_0}\right]. \tag{V.1.16}$$

- **a partial recovery from interpolation problem**, which asks to retrieve $\boldsymbol{U}^* \sim \mathcal{N}(\mathbf{0}_K, \mathbf{I}_K)$ upon knowledge of $\boldsymbol{V} \sim \mathcal{N}(\mathbf{0}_K, \mathbf{I}_K)$ and a set of observations $\widetilde{\boldsymbol{Y}}_0$ from the output channel, where its density is precisely:

$$\widetilde{\boldsymbol{Y}}_0 \sim P_{\mathsf{out}}\left(\cdot \mid \mathbf{Q}^{\frac{1}{2}}\boldsymbol{V} + (\mathbf{P} - \mathbf{Q})^{\frac{1}{2}}\boldsymbol{U}^*\right). \tag{V.1.17}$$

  Additionally, let $\mathbf{P} = \mathbb{E}\left[\boldsymbol{X}^*\boldsymbol{X}^{*\top}\right] \in \mathcal{S}_K^+$ be the correlation of the true signals for $\boldsymbol{X}^* \sim \mathbb{P}_0^{\otimes K}$ and $\mathbf{Q} \in \mathcal{S}^+(\mathbf{P})$ be an overlap matrix. The posterior density is analogous to the previous case:

$$P(\boldsymbol{u} \mid \widetilde{\boldsymbol{Y}}_0, \boldsymbol{V}) = \frac{1}{\mathcal{Z}_{P_{\mathsf{out}}}}\frac{\exp\left\{-\frac{1}{2}\boldsymbol{u}^\top\boldsymbol{u}\right\}}{(2\pi)^{\frac{K}{2}}}P_{\mathsf{out}}\left(\widetilde{\boldsymbol{Y}}_0 \mid \mathbf{Q}^{\frac{1}{2}}\boldsymbol{V} + (\mathbf{P} - \mathbf{Q})^{\frac{1}{2}}\boldsymbol{U}^*\right), \tag{V.1.18}$$

  with associated free entropy:

$$\Psi_{\mathbb{P}_{\mathsf{out}}}(\mathbf{Q}; \mathbf{P}) = \mathbb{E}_{(\widetilde{\boldsymbol{Y}}_0, \boldsymbol{V})}\left[\ln \mathcal{Z}_{\mathbb{P}_{\mathsf{out}}}\right]. \tag{V.1.19}$$

Remark V.1.20. *The partial recovery problem is more general than scalar recovery. The $\widetilde{\boldsymbol{Y}}_0$ term in our case will be a number, since we take a compression function φ_{out} that maps to scalars. The above definition can already be simplified.*

Armed with these two channels, we define the RS *free entropy density* as:

$$f_{\mathsf{RS}}(\mathbf{Q},\mathbf{R};\mathbf{P}) \coloneqq \psi_{\mathbb{P}_0}(\mathbf{R}) + \delta\Psi_{\mathbb{P}_{\mathsf{out}}}(\mathbf{Q};\mathbf{P}) - \frac{1}{2}\mathrm{Tr}\left(\mathbf{R}\mathbf{Q}\right). \tag{V.1.21}$$

The utility of this object is solely in it giving a nicer expression for the original free entropy, as we will see just after the assumptions.

Assumption V.1.22 (Assumptions for Committee result)**.** *Assume that:*

(A1) the suppport of the prior $\mathsf{supp}(\mathbb{P}_0)$ *is a bounded subset of* $\mathbb{R}^K$

(A2) the function $\varphi_{\mathsf{out}} : \mathbb{R}^K \times \mathbb{R} \to \mathbb{R}$ *is bounded, twice differentiable, and with buonded continuous first and second derivatives*

(A3) the data matrix is made of standard independent Gaussians $A_{ij} \sim \mathcal{N}(0,1)$

*(A4) the technical requirement of (***aubinCommitteeMachineComputational2019***) is satisfied*[1]

Theorem V.1.23 (Theorem 3.1 (**aubinCommitteeMachineComputational2019**))**.** *Let Assumptions V.1.22 hold. Assume the scaling* $n, d \to \infty, \delta \in \Theta(1)$*. Then the noisy model V.1.12 with associated channel found in Eqn. V.1.13 has limiting free energy density:*

$$\lim_{n\to\infty} f_n = \sup_{\mathbf{R}\in\mathcal{S}_K^+} \inf_{\mathbf{Q}\in\mathcal{S}_K^+(\mathbf{P})} f_{\mathsf{RS}}(\mathbf{Q},\mathbf{R};\mathbf{P}). \tag{V.1.24}$$

Proof. See the various comments and a sketch in (**aubinCommitteeMachineComputational2019**), while details and properties are in the appendix. The general technique is the adaptive interpolation method (**barbierAdaptiveInterpolationMethod2018**; **barbierOptimalErrorsPha**

□

Upon solving this formula, it is possible to access information theoretic results about the model of Eqn. V.1.12, and consequently for small noise $\Delta \to 0$ of that of Equation V.1.1. Namely, we let Δ be generic, and then obtain the RS free energy density of the noiseless model by sending $\Delta \to 0$ after everything else. This is justified by the fact that at each Δ, the problem is valid, and it suggests that it will be valid at null noise. In other words, the $\boldsymbol{\epsilon}$ added term is just an artifact to use the proof technique of (**aubinCommitteeMachineComputational2019**).

V.1.1 Physics-AMP

With more difficult but analogous calculations to those of Section IV.1, in (**aubinCommitteeMachineCo**
it is shown that starting from r-BP one can derive an instance of GAMP to perform inference on the model of Equation V.1.1, which we term Committee-Generalized Approximate Message Passing (C-GAMP) by analogy. As discussed before, it is conjectured to be optimal among all[2] polynomial algorithms. We report its pseudocode below and then present the deterministic description of its performance in the Bayes-Optimal case. By construction, it is a generalization of Algorithm 4 for multiple signals. To present it, we first need to give an overview of the objects involved.

The most important aspect is that we wish to infer a collection of vectors. In earlier examples, we had discussed the retrieval of only one vector. Proceeding by analogy, scalar quantities become vectors in $\mathbb{R}^K$ and vector quantities become matrices with K columns. We will then iterate over some matrix estimators $\boldsymbol{\mathcal{A}}^{(t)} \in \mathbb{R}^{d\times K}, \vec{\boldsymbol{\mathcal{V}}}^{(t)} \in \mathbb{R}^{(K\times K)\times d}$ with variances $\boldsymbol{\mathcal{V}}_j^{(t)} \in \mathcal{S}_K^+$ for each $j \in [d]$. These are seen for a single $j \in [d]$ as a vector of marginals $\boldsymbol{a}_j^{(t)} \in \mathbb{R}^K$ and an associated variance $\boldsymbol{\mathcal{V}}_j^{(t)} \in \mathcal{S}_K^+$. Accordingly, the mean is a vector, $\boldsymbol{\omega}_i \in \mathbb{R}^K$, and the variance is such that $\mathbf{V}_i \in \mathcal{S}_K^+$. Lastly, the auxiliary mean is denoted as $\boldsymbol{R}_j \in \mathbb{R}^K$ and the auxiliary variance is $\boldsymbol{\Sigma}_j \in \mathcal{S}_K^+$. Similarly, the denoising functions admit vector valued

[1] we do not provide comments on these since it is a specific quantity in that appears in the proofs.

[2] definitely a large class of

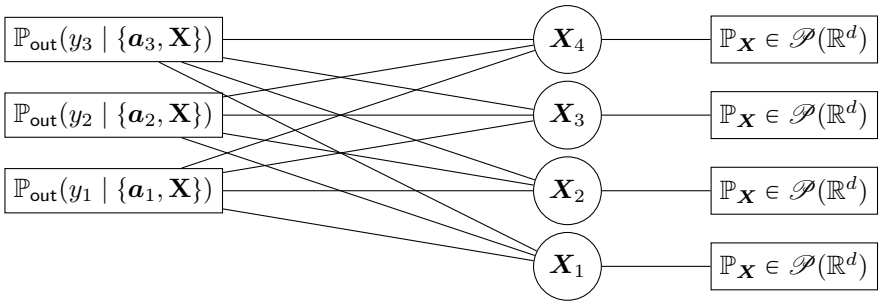


Figure V.3: Factor graph of multiple signal estimation problem.
Parameters $(n, d, L) = (3, d, 4)$. Over the edges, we can define the BP messages.
Figure inspired from (**aubinCommitteeMachineComputational2019**).

expressions. We give precise formulations as follows. The shortcuts for distributions are:

$$Q_0(\boldsymbol{X}; \boldsymbol{\Sigma}, \boldsymbol{R}) \coloneqq \frac{1}{\mathcal{Z}_{\mathbb{P}_0}} \mathbb{P}_0(\boldsymbol{X}) \exp\left\{-\frac{1}{2}\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{R}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right\} \tag{V.1.25}$$

$$[\text{prior on } \boldsymbol{X} \text{ is in } \mathbb{R}^K]$$

$$Q_{\text{out}}(\boldsymbol{z}; \boldsymbol{\omega}, y, \mathbf{V}) \coloneqq \frac{1}{\mathcal{Z}_{\mathbb{P}_{\text{out}}}} \exp\left\{-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\omega})^\top \mathbf{V}^{-1}(\boldsymbol{z} - \boldsymbol{\omega})\right\} P_{\text{out}}(y \mid \boldsymbol{z}). \tag{V.1.26}$$

These allow us to define concisely the denoising functions as expectations wrt (Q_0, Q_{out}) .

$$g_{\text{out}}(\boldsymbol{\omega}, y, \mathbf{V}) \coloneqq \partial_{\boldsymbol{\omega}} \ln \mathcal{Z}_{\mathbb{P}_{\text{out}}} = \mathbf{V}^{-1}\mathbb{E}_{Q_{\text{out}}}\left[\boldsymbol{z} - \boldsymbol{\omega}\right] \tag{V.1.27}$$

$$\partial_{\boldsymbol{\omega}} g_{\text{out}}(\boldsymbol{\omega}, y, \mathbf{V}) = \mathbf{V}^{-1}\mathbb{E}_{Q_{\text{out}}}\left[(\boldsymbol{z} - \boldsymbol{\omega})(\boldsymbol{z} - \boldsymbol{\omega})^\top\right] - \mathbf{V}^{-1} - \boldsymbol{g}_{\text{out}}\boldsymbol{g}_{\text{out}}^\top, \tag{V.1.28}$$

$$f_a(\boldsymbol{\Sigma}, \boldsymbol{R}) \coloneqq \frac{\partial}{\partial[\boldsymbol{\Sigma}^{-1}\boldsymbol{R}]}\left(\ln \mathcal{Z}_{\mathbb{P}_0}\right) = \mathbb{E}_{Q_0}\left[\boldsymbol{X}^*\right], \tag{V.1.29}$$

$$f_{\mathcal{V}}(\boldsymbol{\Sigma}, \boldsymbol{R}) \coloneqq \frac{\partial}{\partial[\boldsymbol{\Sigma}^{-1}\boldsymbol{R}]}\left(f_a(\boldsymbol{\Sigma}, \boldsymbol{R})\right) = \mathbb{E}_{Q_0}\left[\boldsymbol{X}^*\boldsymbol{X}^{*\top}\right] - \boldsymbol{f}_a\boldsymbol{f}_a^\top. \tag{V.1.30}$$

In particular, we remark that:

- the terms in red are shortcuts: they are respectively the functions of the line above, evaluated on the input of their LHS, namely:

$$\boldsymbol{g}_{\text{out}} \coloneqq g_{\text{out}}(\boldsymbol{\omega}, y, \mathbf{V}) \in \mathbb{R}^K, \quad \boldsymbol{f}_a \coloneqq f_a(\boldsymbol{\Sigma}, \boldsymbol{R}) \in \mathbb{R}^K \tag{V.1.31}$$

  whenever clear from context, we will use this notation.

- from the LHS we have definitions, and the rightmost expressions are the simplified form, which is
  - just the expectation of one of the two channels in Eqns. V.1.25, V.1.26
  - or some chain rule of them when we have an additional derivative.

- In particular, g_{out} is the mean of $\mathbf{V}^{-1}(\boldsymbol{z} - \boldsymbol{\omega})$ under Q_{out} and the denoising f_a is the mean of Q_0 (**aubinCommitteeMachineComputational2019**).

The graphical model is Figure V.3, and by relaxing the BP equations on it, we can reach a formulation of AMP for the inference problem of the committee machine (see (**aubinCommitteeMachineComputational2019**)). This can be understood as a vector valued formulation where each variable node is multidimensional $\boldsymbol{X}_l^* \in \mathbb{R}^d$, for a total of L signals. The result is Algorithm 5 below. Akin to a density evolution result, in the high dimensional limit $n, d \to \infty, \delta \in \Theta(1)$ it is possible to describe the dynamics of the Algorithm according to 4 parameters of interest by a sophisticated Law of Large Numbers. These are nothing but the overlaps of estimators and ground truths, and their respective Legendre conjugates (see footnote in the next page for more context), leading to a description in terms of 4 matrices

Algorithm 5 Committee Generalized Approximate Message Passing (C-GAMP)

Input: $\boldsymbol{y} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{n \times d}$
Init: $\boldsymbol{a}_j^{(t=1)} \in \mathbb{R}^K, \boldsymbol{\mathcal{V}}_j^{(t=1)} \in \mathcal{S}_K^+$ for all $j \in [d]$
Init: $\boldsymbol{g}_{out,i}^{(t=0)} = \mathbf{0}, \partial_{\boldsymbol{\omega}} \boldsymbol{g}_{\mathsf{out},i}^{(t=0)} \in \mathcal{S}_K^+$, for all $i \in [n]$
Init: $\boldsymbol{\Sigma}_j^{(t=0)} = \mathbf{I}_K$ for all $j \in [d]$
Init: $t \leftarrow 1$
  while $\boldsymbol{\mathcal{A}}^{(t)}, \vec{\boldsymbol{\mathcal{V}}}^{(t)}$ not converged **do**
    for $i \in [n]$ **do**
      $\mathbf{V}_i^{(t)} \leftarrow \sum_j \frac{1}{d} A_{ij}^2 \boldsymbol{\mathcal{V}}_j^{(t)}$
      $\boldsymbol{\omega}_i^{(t)} \leftarrow \sum_j \frac{1}{\sqrt{d}} A_{ij} \boldsymbol{a}_j^{(t)} - \frac{1}{d} A_{ij}^2 \left[\boldsymbol{\Sigma}_j^{(t-1)}\right]^{-1} \boldsymbol{\mathcal{V}}_j^{(t)} \boldsymbol{\Sigma}_j^{(t-1)} g_{\mathsf{out},i}^{(t-1)}$
    end for
    for $i \in [n]$ **do**
      $\boldsymbol{g}_{out,i}^{(t)} \leftarrow g_{\mathsf{out}}(\boldsymbol{\omega}_i^{(t)}, y_i, \mathbf{V}_i^{(t)})$
      $\partial_{\boldsymbol{\omega}} \boldsymbol{g}_{\mathsf{out},i}^{(t)} \leftarrow \partial_{\boldsymbol{\omega}} g_{\mathsf{out}}(\boldsymbol{\omega}_i^{(t)}, y_i, \mathbf{V}_i^{(t)})$
    end for
    for $j \in [d]$ **do**
      $\boldsymbol{\Sigma}_j^{(t)} \leftarrow -\left[\sum_i \frac{1}{n} A_{ij}^2 \, \partial_{\boldsymbol{\omega}} \boldsymbol{g}_{\mathsf{out},i}^{(t)}\right]^{-1}$
      $\boldsymbol{R}_j^{(t)} \leftarrow \boldsymbol{\Sigma}_j^{(t)} \left[\sum_{i=1}^n \frac{1}{\sqrt{d}} A_{ij} \boldsymbol{g}_{\mathsf{out},i}^{(t)} \boldsymbol{a}_j^{(t)}\right]$
    end for
    for $j \in [d]$ **do**
      $\boldsymbol{a}_j^{(t+1)} \leftarrow f_{\boldsymbol{a}}(\boldsymbol{\Sigma}_j^{(t)}, \boldsymbol{R}_j^{(t)})$
      $\boldsymbol{\mathcal{V}}_j^{(t+1)} \leftarrow f_{\mathcal{V}}(\boldsymbol{\Sigma}_j^{(t)}, \boldsymbol{R}_j^{(t)})$
    end for
    $t \leftarrow t + 1$
  end while
  return $\boldsymbol{\mathcal{A}}^{(t)}, \vec{\boldsymbol{\mathcal{V}}}^{(t)} \in \mathbb{R}^d$

in $\mathcal{S}_K^+$. A careful calculation (**aubinCommitteeMachineComputational2019**) shows that these have a precise form. Under the additional setting of Bayes-Optimal inference, two Nishimori identities reduce the number of overlap matrices to 2. Fast forwarding some computations, the iterations take form:

$$\mathbf{Q}^{(t+1)} = 2\nabla_{\mathbf{R}}\psi_{\mathbb{P}_0}(\mathbf{R}^{(t)}), \quad \mathbf{R}^{(t)} = 2\delta\nabla_{\mathbf{Q}}\Psi_{\mathbb{P}_{\text{out}}}(\mathbf{Q}^{(t)};\rho), \tag{V.1.32}$$

where $\mathbf{Q}^{(t)} := \lim_{d\to\infty}\frac{1}{d}[\widehat{\mathbf{X}}^{(t)}]^\top\mathbf{X}^*$, and the estimator is nothing but the matrix of means returned by AMP $\widehat{\mathbf{X}}^{(t)} := \mathscr{A}^{(t)}$.

Remark V.1.33. *The fixed points of the iteration in Equation V.1.32 correspond to the stationary points of the $\boldsymbol{f}_{\mathsf{RS}}$ function of Equation V.1.21. Iterating them, the dynamics will stop at a stationary point of $\boldsymbol{f}_{\mathsf{RS}}$.*

To give more context, the actual result is explained below. More details are found in (**aubinCommitteeMachineComputational2019**). For ease of notation, let $\boldsymbol{z}_i = \frac{1}{d}\mathbf{X}^*\boldsymbol{a}_i$ denote the i^{th} input we give to the channel before injecting noise/applying the map, and more importantly for any $t \geqslant 1$ define:

$$\overline{\mathcal{V}}^{(t)} := \mathbb{E}_{\mathbf{X}^*}\left[\lim_{d\to\infty}\frac{1}{d}\sum_{j=1}^{d}\mathcal{V}_j^{(t)}\right], \quad \mathbf{X}^* \sim \mathbb{P}_0^{\otimes K} \tag{V.1.34}$$

which is the limiting mean variance estimator for a given set of signals $\boldsymbol{X}^* \in \mathbb{R}^K$. Assume a Bayes-Optimal setting. For $t \geqslant 1$, in the high dimensional limit, the mean and variance components $(\boldsymbol{R}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ are found to behave as:

$$\boldsymbol{R}^{(t)} := \boldsymbol{R}^{(t)}(\boldsymbol{X}^*, \boldsymbol{\xi}) \sim \boldsymbol{X}^* + \sqrt{\widehat{\mathbf{Q}}^{(t)}}\boldsymbol{\xi}, \quad \boldsymbol{\Sigma}^{(t)} \sim [\widehat{\mathbf{Q}}^{(t)}]^{-1}, \tag{V.1.35}$$

where $\boldsymbol{X}^* \sim \mathbb{P}_0^{\otimes K}$, $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$, and the order parameter $\widehat{\mathbf{Q}}^{(t)}$ is one of the two iterators of interest[3].

Remark V.1.36. *It is important to note that the mean $\boldsymbol{R}$ and the matrix $\mathbf{R}$ are unrelated. Unfortunately, too many symbols are required. This is the only inconsistency in notation.*

In other words, the true state evolution equations read:

$$\mathbf{Q}^{(t+1)} = \mathbb{E}_{(\boldsymbol{X}^*,\boldsymbol{\xi})}\left[f_v(\boldsymbol{\Sigma}^{(t)}, \boldsymbol{R}^{(t)}(\boldsymbol{X}^*,\boldsymbol{\xi}))f_v(\boldsymbol{\Sigma}^{(t)}, \boldsymbol{R}^{(t)}(\boldsymbol{X}^*,\boldsymbol{\xi}))^\top\right] \tag{V.1.37}$$

$$\widehat{\mathbf{Q}}^{(t)} = \delta\mathbb{E}_{(\boldsymbol{\omega},\boldsymbol{z},\lambda)}\left[g_{\text{out}}(\boldsymbol{\omega}, \varphi_{\text{out}}(\boldsymbol{z},\lambda), \overline{\mathcal{V}}^{(t)})g_{\text{out}}(\boldsymbol{\omega}, \varphi_{\text{out}}(\boldsymbol{z},\lambda), \overline{\mathcal{V}}^{(t)})^\top\right], \tag{V.1.38}$$

$$(\boldsymbol{z},\boldsymbol{\omega}) \sim \mathcal{N}(\mathbf{0}, \mathbf{C}^{(t)}), \quad \mathbf{C}^{(t)} = \begin{bmatrix}\overline{\mathbf{P}} & \mathbf{Q}^{(t)} \\ \mathbf{Q}^{(t)} & \mathbf{Q}^{(t)}\end{bmatrix}$$

where $\widehat{\mathbf{Q}} = \mathbf{R}$ and $\overline{\mathbf{P}}$ is the mean covariance of the true signals, i.e.

$$\overline{\mathbf{P}} = \mathbb{E}_{\mathbf{X}^*}\left[\lim_{d\to\infty}\frac{1}{d}\sum_{j=1}^{d}\boldsymbol{X}_j^*\boldsymbol{X}_j^{*\top}\right]. \tag{V.1.39}$$

Equations V.1.37, V.1.38 are just an explicit version of Equation V.1.32.

V.2 AMP exclusively from Statistics

The presentation of the previous section had the purpose of showcasing the intuition behind AMP from a Physics perspective. This reasoning extends to a plethora of models we do not touch upon. Preference is given to general results

[3] i.e. it is the matrix $\mathbf{R}$, here the notation is chosen to be suggestive of the Physics interpretation. The matrix $\mathbf{R} = \widehat{\mathbf{Q}}$ could seen as the Field coming from the system (modulo the prior). It is the analog of the magnetization-field coupling. In the Ising model, the magnetization is Legendre conjugated with the external field, here the overlap $\mathbf{Q}$ is the Legendre conjugate of the field $\mathbf{R}$, that can be thus seen as the necessary field that coupled with the prior provides the correct overlap.

which lack the interpretation, preferring the statistics-perspective, but come with great expressivity. In particular, we will focus on presenting the formulations which are suitable for the setting of our problem, and link their statistical formulations with the Physics literature on the topic. The main reference in this case is (**tanMixedRegressionApproximate2023**), which in turn is an built upon the very useful review (**fengUnifyingTutorialApproximate2021**). For these reasons, this exposition can be thought of as a different construction with respect to the previous one, tailored for the specific case of mixed regression, that eventually leads to the same results.

Matrix GLM The case of two mixed signals naturally generalizes to the multiple signals form where $\mathbf{X}^* \equiv [\boldsymbol{x}_1^* | \dots | \boldsymbol{x}_L^*] \in \mathbb{R}^{d \times L}$, which is associated to observations $\boldsymbol{Y} \in \mathbb{R}^{n \times L_{out}}, \mathbf{A} = \{\boldsymbol{a}_i\}_{i=1}^n \in \mathbb{R}^{n \times d}$ via the output function:

$$\boldsymbol{y}_i = \varphi\left((\mathbf{X}^*)^\top \boldsymbol{a}_i, \boldsymbol{\lambda}_i\right) \quad i \in [n], \quad \varphi : \mathbb{R}^L \times \mathbb{R}^{L_\Lambda} \to \mathbb{R}^{L_{out}} \tag{V.2.1}$$

where $\{\boldsymbol{\lambda}_i\}_{i=1}^n$ are vectors of auxiliary latent random variables and $\mathbf{A} = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_n]^\top$ is the data matrix.

Example V.2.2 (Matrix GLM special case)**.** *We briefly recover the case of mixed phase retrieval with the choice:*

$$y_i = \varphi(\boldsymbol{z}_i, \boldsymbol{\lambda}_i) = \left| \langle \boldsymbol{x}_1^*, \boldsymbol{a}_i \rangle \eta_{i1} + \dots + \langle \boldsymbol{x}_L^*, \boldsymbol{a}_i \rangle \eta_{iL} \right| + \epsilon_i \tag{V.2.3}$$

$$\boldsymbol{z}_i = (\boldsymbol{X}^*)^\top \boldsymbol{a}_i \in \mathbb{R}^{L \times d}, \ \sum_{l=1}^{L} \eta_{il} = 1, \{\eta_{il}\} \in \{0,1\}^L, \quad \epsilon_i \sim \mathcal{N}(0, \Delta), \ \boldsymbol{\lambda_1} = (\{\eta_{il}\}, \epsilon_i)\,. \tag{V.2.4}$$

In simple words the output is determined by only one of the L signals at each time since the latents are all binary valued and only one is non-zero for each sample i. To recover our specific problem of interest, noiseless mixed phase retrieval, we set $L = 2$, $\eta \sim \mathcal{B}ern(\alpha)$ and $\Delta = 0$.

A large body of literature provides ways to compute tractable forms of the global Least Squares Estimator for the mixed linear regression case of Eqn. V.2.1, which in principle is an NP-Hard problem. An exposition of references is found in (**tanMixedRegressionApproximate2023**). It is also underlined how all of these methods are not able to exploit information about the mixing signals (e.g. the prior in Bayes-Optimal inference) and do not result in useful descriptions when $\frac{n}{d} = \delta \in \Theta(1)$, requiring instead that the number of samples is at least $d \log d$.

Most of the results stated in the following are strong theoretical statements of the Physics formulation. For this purpose, we introduce some specific objects that will be needed.

Definition V.2.5 (Complete Convergence)**.** *First introduced in (***hsuCompleteConvergenceLaw1947***), discussed in (***fengUnifyingTutorialApproximate2021***). For a sequence of random variables (X_n) on a euclidean space E we say $X_n \xrightarrow{c} x \in E$ when $Y_n \xrightarrow{a.s.} x$ for any sequence $(Y_n) \subset E$ such that $Y_n \overset{d}{=} X_n$ for all n.*

Fact V.2.6. *It holds that $X_n \xrightarrow{c} x \implies X_n \xrightarrow{a.s.} x$, where the latter is almost sure convergence (Def. A.3.1).*

Proof. Trivially, the sequence (X_n) is equally distributed wrt the sequence itself and assumed to be almost surely convergent to be completely convergent. □

Definition V.2.7 (Wasserstein Distance, Euclidean case)**.** *For (ν, μ) in $\mathscr{P}_d(r)$ the r-Wasserstein distance is:*

$$\mathsf{d}_{\mathsf{Wass}}^r(\mu, \nu) := \inf_{(\boldsymbol{X}, \boldsymbol{Y}): \boldsymbol{X} \sim \mu, \boldsymbol{Y} \sim \nu} \mathbb{E}\left[\|\boldsymbol{X} - \boldsymbol{Y}\|_2^r\right]^{\frac{1}{r}} \tag{V.2.8}$$

Remark V.2.9. *This is the Euclidean formulation of Wasserstein distance which can be far more general.*

Structural Assumptions while our case is non degenerate and easily falls into the framework considered, we briefly report the modeling assumptions that accomodate the result of (**tanMixedRegressionApproximate2023**):

- The data matrix $\mathbf{A}$ is assumed to be such that the rows are iid $\boldsymbol{a}_i \sim \mathcal{N}(\mathbf{0}_d, \frac{1}{n} I_d)$
- the signals matrix $\mathbf{X}^*$ is independent of the data matrix and the auxiliary variables matrix $\boldsymbol{\Lambda}$
- it holds that $\delta \in \Theta(1)$ and $n \to \infty$
- for some $r \in [2, \infty]$ there are well-behaved $(\overline{\boldsymbol{X}}, \overline{\boldsymbol{\lambda}})$ vectors respectively in $\mathbb{R}^L, \mathbb{R}^{L_\Lambda}$ satisfying:

$$\overline{\boldsymbol{X}} \sim \mathbb{P}_{\overline{\boldsymbol{X}}} \quad \overline{\boldsymbol{\lambda}} \sim \mathbb{P}_{\overline{\boldsymbol{\lambda}}} \quad \mathbb{E}\left[\overline{\boldsymbol{X}}^\top \overline{\boldsymbol{X}}\right] > 0 \quad \mathbb{E}\left[\sum_{l=1}^{L} |\overline{\boldsymbol{X}}_l|^r\right] < \infty, \quad \mathbb{E}\left[\sum_{l=1}^{L} |\overline{\boldsymbol{\lambda}}_l|^r\right] < \infty, \tag{V.2.10}$$

  such that the empirical distributions of rows of the signal and auxiliary data matrices converge completely in r-Wasserstein as $n \to \infty$ to those well behaved distributions. Namely for empirical distribution measures (ν_d, ν_n) averaging over the respective rows:

$$\mathsf{d}^r_{\mathsf{Wass}}(\nu_d(\mathbf{X}^*), \mathbb{P}_{\overline{\boldsymbol{X}}}) \xrightarrow{c} 0 \qquad \mathsf{d}^r_{\mathsf{Wass}}(\nu_n(\boldsymbol{\Lambda}), \mathbb{P}_{\overline{\boldsymbol{\lambda}}}) \xrightarrow{c} 0. \tag{V.2.11}$$

We are now ready to report the AMP iteration of (**tanMixedRegressionApproximate2023**), and its theoretical result. The iterative equations serve the purpose of estimating $\widehat{\mathbf{X}}^{(t)} \in \mathbb{R}^{d \times L}, \boldsymbol{\Theta}^{(t)} \in \mathbb{R}^{n \times L}$, where the latter estimates the matrix $\mathbf{A}\mathbf{X}^*$:

$$\boldsymbol{\Theta}^{(t)} = \mathbf{A}\widehat{\mathbf{X}} - \widehat{\mathbf{W}}^{t-1}\left(\mathbf{F}^{(t)}\right)^\top, \qquad \widehat{\mathbf{W}}^{(t)} = g_t(\boldsymbol{\Theta}^{(t)}; \boldsymbol{Y}) \qquad \mathbf{F}^{(t)} = \frac{1}{n}\sum_{j=1}^{d} f_t'(\boldsymbol{X}_j^{(t)}) \tag{V.2.12}$$

$$\mathbf{X}^{(t+1)} = \mathbf{A}^\top \widehat{\mathbf{W}}^{(t)} - \widehat{\mathbf{X}}^{(t)}\left(\mathbf{C}^{(t)}\right)^\top \qquad \widehat{\mathbf{X}}^{(t+1)} = f_{t+1}(\mathbf{X}^{(t+1)}) \qquad \mathbf{C}^{(t)} = \frac{1}{n}\sum_{i=1}^{n} g_t'(\boldsymbol{\Theta}_i^{(t)}; \boldsymbol{Y}_i) \tag{V.2.13}$$

where it is agreed that:

- the functions (g_t, f_t) are applied row wise to their inputs as:

$$g_t : \mathbb{R}^L \times \mathbb{R}^{L_{out}} \to \mathbb{R}^L \qquad g_t(\boldsymbol{\Theta}^{(t)}; \boldsymbol{Y}) = \left(g_t(\boldsymbol{\Theta}^{(t)}_{[1,:]}; \boldsymbol{Y}_1), \dots, g_t(\boldsymbol{\Theta}^{(t)}_{[n,:]}; \boldsymbol{Y}_n)\right)^\top \tag{V.2.14}$$

$$f_t : \mathbb{R}^L \to \mathbb{R}^L \qquad f_t(\mathbf{X}^{(t)}) = (f_t(\boldsymbol{X}^{(t)}_{[1,:]}), \dots, f_t(\boldsymbol{X}^{(t)}_{[d,:]}))^\top \tag{V.2.15}$$

- the matrices $\mathbf{F}^{(t)}, \mathbf{C}^{(t)}$ are Jacobians $\in \mathbb{R}^{L \times L}$ wrt the first argument.

Remark V.2.16. *Clearly, Eqns. V.2.12, V.2.13 can be cast into an algorithm. The initializations would be* $\widehat{\mathbf{X}}^{(0)}$ *and* $\widehat{\mathbf{W}}^{(-1)} = \mathbf{0} \in \mathbb{R}^{n \times L}$.

Remark V.2.17. *Iterating Eqns. V.2.12, V.2.13 for* T *times has complexity* $O(npLT)$*, since they are all matrix operations.*

Similar iteration schemes in literature are Iterative thresholding algorithms, which however lacked the peculiarity of debiasing the estimators. Here the terms $-\widehat{\mathbf{W}}^{t-1}\left(\mathbf{F}^{(t)}\right)^\top$ and $-\widehat{\mathbf{X}}^{(t)}\left(\mathbf{C}^{(t)}\right)^\top$ perform exactly this task, and are essentially akin to the Onsager correction of Chapter IV. This allows us to rexpress the complicated recursive properties of the matrix AMP in terms of a recursion of **deterministic vectors**[4] which are guaranteed to be the convergent limit of the empirical distribution of a function of the original signal distribution. In turn, given that f_t is just a function, the estimators will converge empirically in the rows after the application of f_t. This last point makes f_t be interepreted as a denoising function that might be adjusted to exploit the properties of the original signal distribution. A similar statement holds for the $\boldsymbol{\Theta}^{(t)}$ estimator.

[4] Before we had a recursion of deterministic scalars because we were estimating vectors, now we are estimating matrices.

State Evolution formally More precisely, we find that the empirical distribution of the rows of $\widehat{\mathbf{X}}^{(t)}$ and $\mathbf{\Theta}^{(t)}$ converge in a very strong sense (i.e. same of the assumptions) to the distributions of the vectors:

$$\mathbf{M}_X^{(t)}\overline{\boldsymbol{X}}+\boldsymbol{G}_X^{(t)} \quad \boldsymbol{G}_X^{(t)} \sim \mathcal{N}(\mathbf{0}, \mathbf{T}_X^{(t)}) \qquad \mathbf{M}_\Theta^{(t)}\boldsymbol{Z}+\boldsymbol{G}_\Theta^{(t)} \quad \boldsymbol{Z} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\delta}\mathbb{E}\left[\mathbf{X}\mathbf{X}^\top\right]\right) \quad \boldsymbol{G}_\Theta^{(t)} \sim \mathcal{N}(\mathbf{0}, \mathbf{T}_\Theta^{(t)}) \tag{V.2.18}$$

where the matrices $(\mathbf{M}_X^{(t)}, \mathbf{T}_X^{(t)}, \mathbf{M}_\Theta^{(t)}, \mathbf{T}_\Theta^{(t)})$ are **deterministic**, **derived recursively** and all in $\mathbb{R}^{L\times L}$. We now turn to describe how these are defined. For ease of notation, rewrite the second denoising function as:

$$\hbar : \mathbb{R}^L \times \mathbb{R}^L \times \mathbb{R}^{L_\Lambda} \to \mathbb{R}^L \qquad \hbar_t(\boldsymbol{z}, \boldsymbol{u}, \boldsymbol{v}) := g_t(\boldsymbol{u}, \varphi(\boldsymbol{z}, \boldsymbol{v})) \tag{V.2.19}$$

where it is again intended that it acts row-wise. The recursion is then defined given $\mathbf{\Sigma}^{(t)} \in \mathbb{R}^{2L\times 2L}$. The first step is sampling vectors:

$$\begin{bmatrix} \boldsymbol{Z} \\ \boldsymbol{Z}^{(t)} \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^{(\mathbf{t})}) \tag{V.2.20}$$

independent of $\overline{\boldsymbol{\lambda}} \sim \mathbb{P}_{\overline{\boldsymbol{\lambda}}} \in \mathbb{R}^{L_{out}}$. Then we evaluate the next deterministic step:

$$\mathbf{M}_X^{(t+1)} = \mathbb{E}\left[\partial_{\boldsymbol{Z}}\hbar_t(\boldsymbol{Z}, \boldsymbol{Z}^{(t)}, \overline{\boldsymbol{\lambda}})\right] \tag{V.2.21}$$

$$\mathbf{T}_X^{(t+1)} = \mathbb{E}\left[\hbar_t(\boldsymbol{Z}, \boldsymbol{Z}^{(t)}, \overline{\boldsymbol{\lambda}})\hbar_t(\boldsymbol{Z}, \boldsymbol{Z}^{(t)}, \overline{\boldsymbol{\lambda}})^\top\right], \tag{V.2.22}$$

where the derivative in the first term is the Jacobian of the first argument. Having the new mean and the new variance of the noise, the variance-covariance matrix for the following step is updated block wise as:

$$\mathbf{\Sigma}^{(t+1)} = \begin{bmatrix} \mathbf{\Sigma}_{(11)}^{(t+1)} & \mathbf{\Sigma}_{(12)}^{(t+1)} \\ \mathbf{\Sigma}_{(21)}^{(t+1)} & \mathbf{\Sigma}_{(22)}^{(t+1)} \end{bmatrix} \tag{V.2.23}$$

where:

$$\mathbf{\Sigma}_{(11)}^{(t+1)} = \frac{1}{\delta}\mathbb{E}\left[\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^\top\right] \tag{V.2.24}$$

$$\mathbf{\Sigma}_{(12)}^{(t+1)} = \mathbf{\Sigma}_{(21)}^{(t+1)} = \frac{1}{\delta}\mathbb{E}\left[\overline{\boldsymbol{X}} f_{t+1}(\mathbf{M}_X^{(t+1)}\overline{\boldsymbol{X}} + \boldsymbol{G}_X^{(t+1)})^\top\right] \tag{V.2.25}$$

$$\mathbf{\Sigma}_{(22)}^{(t+1)} = \frac{1}{\delta}\mathbb{E}\left[f_{t+1}(\mathbf{M}_X^{(t+1)}\overline{\boldsymbol{X}} + \boldsymbol{G}_X^{(t+1)}) f_{t+1}(\mathbf{M}_X^{(t+1)}\overline{\boldsymbol{X}} + \boldsymbol{G}_X^{(t+1)})^\top\right]. \tag{V.2.26}$$

Such recursion only requires a baseline $\mathbf{\Sigma}^{(0)} \in \mathbb{R}^{2L\times 2L}$. Below we provide the additional model assumptions for the results to hold.

Assumption V.2.27 (Matrix GLM AMP model conditions)**.** *Reporting the requirements of (***tanMixedRegressionApproximate2023***), we wish:*

(A1) (reasonable start) there is a convergent matrix $\mathbf{\Sigma}^{(0)}$ *and* $c_0 \in \mathbb{R}$ *such that in the thermodynamic limit with* $\delta \in \Theta(1)$ *it holds:*

$$\frac{1}{n}\begin{bmatrix} (\mathbf{X}^*)^\top\mathbf{X}^* & (\mathbf{X}^*)^\top\widehat{\mathbf{X}}^{(0)} \\ (\widehat{\mathbf{X}}^{(0)})^\top\mathbf{X}^* & (\widehat{\mathbf{X}}^{(0)})^\top\widehat{\mathbf{X}}^{(0)} \end{bmatrix} \xrightarrow{c} \mathbf{\Sigma}^{(0)} \qquad \frac{1}{d}\sum_{j=1}^{d}\sum_{l=1}^{L}|\widehat{X}_{jl}^{(0)}|^r \xrightarrow{c} c_0 \tag{V.2.28}$$

where the r *power of convergence is the same of the Wasserstein convergence of the structural assumptions. Additionally, there is a Lipschitz function* $f_0 : \mathbb{R}^L \to \mathbb{R}^L$ *such that:*

$$\frac{1}{d}\left(\widehat{\mathbf{X}}^{(0)}\right)^\top \phi(\mathbf{X}^*) \xrightarrow{c} \mathbb{E}\left[f_0(\overline{\boldsymbol{X}})\phi(\overline{\boldsymbol{X}})^\top\right] \quad \mathbf{\Sigma}_{(22)}^{(0)} - \mathbb{E}\left[f_0(\overline{\boldsymbol{X}})\phi(\overline{\boldsymbol{X}})^\top\right] \succeq 0, \tag{V.2.29}$$

for all $\phi \in \mathrm{Lip}(\mathbb{R}^L, \mathbb{R}^L)$*, where* $\mathrm{Lip}(\mathbb{R}^L, \mathbb{R}^L)$ *is the set of functions* $f : \mathbb{R}^L \to \mathbb{R}^L$ *that are Lipschitz (wrt any finite constant).*

(A2) (reasonable denoisers) for any time step, the denoising function f_{t+1} *is:*

– *non constant*

– *Lipschitz in* $\mathbb{R}^L$

– *with derivative continuous and Lebesgue almost everywhere (a.e.) (see Sec. A.3)*

while the denoising function $\hbar_t$ *(which is just* g_t*) is:*

– *Lipschitz on* $\mathbb{R}^{2L+L_\Lambda}$

– *non-null measure non-constant in the auxiliary latents, namely:*

$$\mathbb{P}_{\overline{\boldsymbol{\lambda}}}(\{\boldsymbol{v} : (\boldsymbol{z}, \boldsymbol{u}) \to \hbar_t(\boldsymbol{u}, \boldsymbol{z}, \boldsymbol{v}) \text{ non-constant}\}) > 0 \tag{V.2.30}$$

– *measure zero discontinuous derivative in the inputs of the learning process, namely:*

$$\mathbb{P}\left[(\boldsymbol{Z}^{(t)}, \overline{\boldsymbol{Y}}) \in \mathcal{D}_t\right] = 0 \quad \mathcal{D}_t \subset \mathbb{R}^{L+L_\Lambda} \tag{V.2.31}$$

where $\mathcal{D}_t$ *is the set of discontinouities of* g_t'*.*

Remark V.2.32. *Every algorithm requires an initialization. In the case of symmetric channels such as phase retrieval, the null initializaiton ends up being a stationary fixed point of the dynamics. As a consequence, in some cases the statistician needs to find a leading estimator to feed the AMP algorithm. A choice for the estimator is regarded as a head-start, i.e.* $\widehat{\mathbf{X}} = \epsilon \mathbf{X}^*$ *with* $\epsilon \in (0, 1)$ *or any other scaling such that the first estimator is mildly correlated with the ground truth. Another option is just starting with a rescaling of the all ones matrix. Both guarantee that the first Assumption is verified. Due to the symmetry of the Phase Retrieval channel, this is complicated, as a random guess will have* $\sim d^{-\frac{1}{2}}$ *overlap with the true signal, thus being null at infinite dimension. As we will discuss later, in many cases, a leading spectral estimator is used to give a warm-start.*

Definition V.2.33 (Pseudo-Lipschitz Space of functions)**.** *We define* $\text{P-Lip}_m(r, C)$ *as the set of functions* $\phi : \mathbb{R}^m \to \mathbb{R}$ *such that:*

$$|\phi(\boldsymbol{x}) - \phi(\boldsymbol{y})| \leqslant C(1 + \|\boldsymbol{x}\|_2^{r-1} + \|\boldsymbol{y}\|_2^{r-1}) \|\boldsymbol{x} - \boldsymbol{y}\|_2 \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m \tag{V.2.34}$$

Functions in $\mathbb{R}^m$ *belonging to an* r *class for some* C *are said to be* r*-Pseudo-Lipschitz.*

We now write down the main results of (**tanMixedRegressionApproximate2023**), which help us greatly in determining the properties of the AMP iteration we presented.

Theorem V.2.35 (Matrix GLM AMP, Theorem 1 of (**tanMixedRegressionApproximate2023**))**.** *The Iterations in Eqns. V.2.12, V.2.13, which constitute an AMP algorithm for Matrix GLM models formalized as in Equation V.2.1 are such that, under Assumption V.2.27, and the condition that* $\mathbf{T}_X^{(1)}$ *is positive definite, the following holds for all* $t \geqslant 0$ *as* $n, d \to \infty, \delta \in \Theta(1)$*:*

$$\sup_{\phi \in \text{P}-\text{Lip}_{2L}(r,1)} \left| \frac{1}{d} \sum_{j=1}^{d} \phi(\boldsymbol{X}_j^{(t+1)}, \boldsymbol{X}_j^*) - \mathbb{E}\left[\phi(\mathbf{M}_X^{(t+1)} \overline{\boldsymbol{X}} + \boldsymbol{G}_X^{(t+1)}, \overline{\boldsymbol{X}})\right] \right| \xrightarrow{c} 0 \tag{V.2.36}$$

$$\sup_{\phi \in \text{P}-\text{Lip}_{2L+L_\Lambda}(r,1)} \left| \frac{1}{n} \sum_{i=1}^{n} \phi(\boldsymbol{\Theta}_i^{(t)}, \boldsymbol{\Theta}_i, \boldsymbol{\lambda}_i) - \mathbb{E}\left[\phi(\mathbf{M}_\Theta^{(t)} \boldsymbol{Z} + \boldsymbol{G}_\Theta^{(t)}, \boldsymbol{Z}, \overline{\boldsymbol{\lambda}})\right] \right| \xrightarrow{c} 0 \tag{V.2.37}$$

where $\boldsymbol{\Theta}_i \coloneqq (\mathbf{X}^*)^\top \boldsymbol{a}_i \in \mathbb{R}^L$ *for each* $i \in [n]$ *and the noise terms are as before:*

$$\boldsymbol{G}_X^{(t+1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{T}_X^{(t+1)}) \perp \overline{\boldsymbol{X}} \qquad \boldsymbol{G}_\Theta^{(t)} \sim \mathcal{N}(\mathbf{0}, \mathbf{T}_\Theta^{(t)}) \perp (\boldsymbol{Z}, \overline{\boldsymbol{\lambda}}) \tag{V.2.38}$$

Proof. (**tanMixedRegressionApproximate2023**) □

Corollary V.2.39 (Wasserstein Convergence version of matrix GLM AMP, Cor. 7.21 of (**fengUnifyingTutorialApproximate2021**))**.** *Theorem V.2.35 can be restated to a Wasserstein convergence result. The former equation is equivalent to asserting that the joint empirical distribution of the rows of* $(\widehat{\mathbf{X}}_j, \mathbf{X}_j^*)$ *is c-convergent in Wasserstein distance to the joint distribution of the vectors on the RHS. A similar statement holds for the latter.*

Remark V.2.40. *The result of the Corollary-Theorem is very powerful. In the high dimensional limit, the empirical distribution of a large class of functions evaluated on the rows of the estimator and the ground truth is replaced by the same function evaluated on a sample* $\overline{\boldsymbol{X}} \sim \mathbb{P}_{\overline{\boldsymbol{X}}}$*, the limiting prior, and a* ***deterministically perturbed*** *version of the same sample, with known matrices* $(\mathbf{M}_X, \mathbf{T}_X)$ *performing the perturbation. The degree of reliability of such result is complete convergence, which is even stronger than almost sure convergence. More importantly, this convergence is pointwise in time: at teach time step t, the result holds.*

At this point one could ask what functions belong to the class of P-Lip functions. Many objects of interest satisfy the requirements of Def. V.2.33.

Example V.2.41 (Overlap Evaluation Matrix GLM)**.** *An interesting performance indicator for an estimator is the overlap. Briefly, it is the normalized alignment of the estimator with the ground truth. Take an estimator matrix* $\widehat{\mathbf{X}}$ *of the signal. Each row* $j \in [d]$ *is a feature, with L associated vectors that could be the feature "effect". Each column* $l \in [L]$ *is a signal estimator, trying to approximate the effects* $j \in [d]$ *of one of the possible vectors of e.g. mixed regression. By the result of Thm. V.2.35, the overlap of the* l^{th} *estimator is such that in the high dimensional limit*

$$\frac{\left\langle \widehat{\boldsymbol{x}}_l^{(t)}, \boldsymbol{x}_l \right\rangle^2}{\left\| \widehat{\boldsymbol{x}}_l^{(t)} \right\|_2^2 \|\boldsymbol{x}_l\|_2^2} \xrightarrow{c} \frac{\mathbb{E}\left[f_{t,l}(\overline{\boldsymbol{X}}^{(t)}) \overline{x}_l \right]^2}{\mathbb{E}\left[f_{t,l}(\overline{\boldsymbol{X}}^{(t)}) \right]^2 \mathbb{E}\left[\overline{x}_l\right]^2} \quad \forall l \in [L] \quad \overline{\boldsymbol{X}}^{(t)} \coloneqq \mathbf{M}_X^{(t)} \overline{\boldsymbol{X}} + \mathbf{G}_X^{(t)} \tag{V.2.42}$$

where $f_{t,l}$ *is the* l^{th} *component of the denoising function output and on the RHS,* $\overline{x}_l$ *is the* l^{th} *component of* $\overline{\boldsymbol{X}}$*, a scalar.*
Notice that we use the apex index t for time, the overline symbol surely means that the object is a vector, and the l index on the matrix $\widehat{\mathbf{X}}$ *selects a ground truth estimator* $\widehat{\boldsymbol{x}}_l \in \mathbb{R}^d$*. The notation at first sight is very heavy.*
In words, a column of the estimator matrix, being completely specified in empirical mean by the expectation of a well defined sample from the prior, eventually reduces to a mere evaluation of the equivalent limiting version of the overlap expressed in terms of these expectations.

Example V.2.43 (MSE evaluation Matrix GLM)**.** *As in the previous example, the MSE of the* l^{th} *ground truth estimator can be evaluated in the high dimensional limit given the specialized result:*

$$\frac{1}{d}\left\| \boldsymbol{x}_l - \widehat{\boldsymbol{x}}_l^{(t)} \right\|_2^2 \xrightarrow{c} \mathbb{E}\left[\left(\overline{x_l} - f_{t,l}(\overline{\boldsymbol{X}}^{(t)}) \right)^2 \right] \quad \forall l \in [L] \quad \overline{\boldsymbol{X}}^{(t)} \coloneqq \mathbf{M}_X^{(t)} \overline{\boldsymbol{X}} + \mathbf{G}_X^{(t)}. \tag{V.2.44}$$

The examples provide even more evidence of the strength of Theorem V.2.35. In the high dimensional limit, the iterates of the AMP estimators, which are over very tall matrices $n \to \infty$, are such that a big class of performance functions evaluated on their rows and those of the ground truth is essentially equivalent to scalar expectations and function evaluations of a tweaked prior. The **state evolution** result is captured by the matrices $(\mathbf{M}_X^{(t)}, \mathbf{T}_X^{(t)})$, which achieve this. While for vector AMP the calculations reduced to scalar state evolution in this case we have a state evolution that is in terms of vectors in $\mathbb{R}^L$.

State Evolution optimization Remarkably, the state evolution iterates just depend on the choice of the denoising functions $(f_t(\cdot), g_t(\cdot;\cdot))$. A natural question is the existance of optimal functions in closed form. This is answered in the next Proposition. First we report the intuition presented in (**tanMixedRegressionApproximate2023**).

We have two recursions[5] that describe the limiting dynamics of our estimators in a correlation plus noise design. To choose the best denoising functions, one would hope to make this noise as small as possible. This objective can be expressed mathematically with some linear algebra. First we define two versions of the iterates in State Evolution that rearrange the terms[6]:

$$\widetilde{\boldsymbol{Z}} \coloneqq \boldsymbol{Z} + \left(\mathbf{M}_{\Theta}^{(t)}\right)^{-1} \boldsymbol{G}_{\Theta}^{(t)} \quad \widetilde{\boldsymbol{X}}^{(t+1)} \coloneqq \overline{\boldsymbol{X}} + \left(\mathbf{M}_{X}^{(t+1)}\right) \boldsymbol{G}_{X}^{(t+1)} \tag{V.2.45}$$

these are just used to express concisely two *effective noise covariance matrices* (**tanMixedRegressionApproximate2023**). Indeed, we now express the magnitude of the noise at time t for the two estimates:

$$\mathbf{N}_{\Theta}^{(t)} \coloneqq \mathrm{CoV}\left[\widetilde{\boldsymbol{Z}}^{(t)} - \boldsymbol{Z}\right] = \left(\mathbf{M}_{\Theta}^{(t)}\right)^{-1} \mathbf{T}_{\Theta}^{(t)} \left(\mathbf{M}_{\Theta}^{(t)}\right)^{\top} \tag{V.2.46}$$

$$\mathbf{N}_{X}^{(t)} \coloneqq \mathrm{CoV}\left[\widetilde{\boldsymbol{X}}^{(t)} - \boldsymbol{X}\right] = \left(\mathbf{M}_{X}^{(t)}\right)^{-1} \mathbf{T}_{X}^{(t)} \left(\mathbf{M}_{X}^{(t)}\right)^{\top}. \tag{V.2.47}$$

Given such a construction, the objective has moved to a minimization of the traces of these two matrices to make them small in norm. It is worth noticing that the terms in Eqn. V.2.46 are determined by $\boldsymbol{\Sigma}^{(t)}$, which depends on f_t. To see this, observe how Eqn. V.2.18 establishes the relation, with the noise of sampling defined in Eqns. V.2.23-V.2.26, paying particular attention to Eqn. V.2.25. Similarly, those in Eqn. V.2.47 are determined by g_t. This is concluded by a quick glance at Eqn. V.2.18, and how its terms are determined in Eqns. V.2.21, V.2.22, with the function $\hbar$ constructed from g in Eqn. V.2.19.
In presence of only one signal, which is the case of the classic GLM, optimality is achieved in the Bayes-Optimal setting by maximizing the scalar SNRs $\frac{m_X^{(t)}}{\sqrt{q_X^{(t)}}}, \frac{m_\Theta^{(t)}}{\sqrt{q_\Theta^{(t)}}}$ (**ranganGeneralizedApproximateMessage2012**; **fengUnifyingTutorialApproximate2021**). This reduces to finding the denoisers such that $m = \rho = \sqrt{q}$ which lets state evolution be described by a single parameter and $\widehat{\boldsymbol{x}}^{(t)} = \rho^{(t)}\overline{\boldsymbol{x}} + \sqrt{\rho^{(t)}}\xi_t$. In the multi-signal setting of Matrix GLM we find a closed form expression which will eventually be equivalent in our model of interest due to a very interesting property.

Proposition V.2.48 (Bayes-Optimal denoisers, Prop. 2 of (**tanMixedRegressionApproximate2023**))**.** *If $(\mathbb{P}_{\overline{\boldsymbol{X}}}, \mathbb{P}_{\boldsymbol{\lambda}})$ are known, for each time step $k \geqslant 1$ it holds that:*

1. *given $(\mathbf{M}_X^{(t)}, \mathbf{T}_X^{(t)})$, the trace of the effective noise of estimating $\boldsymbol{\Theta}^{(t)}$ is minimized when the denoiser is:*

$$f_t^{\star}(\boldsymbol{s}) = \mathbb{E}\left[\overline{\boldsymbol{X}} \mid \mathbf{M}_X^{(t)}\overline{\boldsymbol{X}} + \boldsymbol{G}_X^{(t)} = \boldsymbol{s}\right] \qquad \boldsymbol{G}_X^{(t)} \perp \overline{\boldsymbol{X}} \tag{V.2.49}$$

   where $\boldsymbol{G}_X^{(t)} \sim \mathcal{N}(\mathbf{0}, \mathbf{T}_X^{(t)})$ and $\overline{\boldsymbol{X}} \sim \mathbb{P}_{\overline{\boldsymbol{X}}}$, i.e. the Bayes-Optimal condition

2. *given $(\mathbf{M}_\Theta^{(t)}, \mathbf{T}_\Theta^{(t)})$ the trace of the effective noise of estimating $\mathbf{X}^{(t)}$ is minimized when the denoiser is:*

$$g_t^{\star}(\boldsymbol{u}, y) = \mathrm{CoV}\left[\boldsymbol{Z} \mid \boldsymbol{Z}^{(t)} = \boldsymbol{u}\right]^{-1} \left(\mathbb{E}\left[\boldsymbol{Z} \mid \boldsymbol{Z}^{(t)} = \boldsymbol{u}, \overline{\boldsymbol{Y}} = \boldsymbol{y}\right] - \mathbb{E}\left[\boldsymbol{Z} \mid \boldsymbol{Z}^{(t)} = \boldsymbol{u}\right]\right) \tag{V.2.50}$$

   where

   - *$\overline{\boldsymbol{Y}} = \varphi(\boldsymbol{Z}, \overline{\boldsymbol{\lambda}})$ is random and in $\mathbb{R}^{L_{out}}$*
   - *$\boldsymbol{y}$ is a fixed realization of it*
   - *$\overline{\boldsymbol{\lambda}} \sim \mathbb{P}_{\overline{\boldsymbol{\lambda}}}$ is in $\mathbb{R}^{\Lambda}$, sampled independently of $\boldsymbol{Z}$*
   - *the vector $[\boldsymbol{Z}, \boldsymbol{Z}^{(t)}]^{\top} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{(t)}\right)$.*

Proof. Found in (**tanMixedRegressionApproximate2023**) □

[5] we ignored the result about $\boldsymbol{\Theta}$ because it is essentially equivalent
[6] if the inverse does not exist, conside the -1 as the pseudoinverse (see Sec. A.2)

Remark V.2.51. *Both results assume that the other State Evolution matrices are known for the previous time-step in the algorithm. This is required by their recursive definition, and one must consider that there will be a leading estimator* $\widehat{\mathbf{X}}$ *from which the iterations start. In particular, Bayes-Optimality is achieved time-wise, i.e. separately for each* t*.*

Appendix A

Auxiliary Results

In this Chapter, we collect definitions and statements that are mentioned in the main part of the text. They are not in a specific order. For organizational purposes, they are divided in sections.

A.1 Analysis

Definition A.1.1 (Uniform convergence)**.** *Consider a sequence of functions $(f_n)_{n\in\mathbb{N}}$ all such that $f_n : E \to \mathbb{R}$, where E is generic. The sequence is uniformly convergent to $f : E \to \mathbb{R}$ and we write $f_n \rightrightarrows f$ when for every $\epsilon > 0$ there exists $\overline{n} \in \mathbb{N}$ such that for all $n \geqslant \overline{n}$ and for all $x \in E$ it holds:*

$$|f_n(x) - f(x)| < \epsilon.$$

The convergence is uniform in the sense that it holds for the same $\overline{n}$ for all x.

Definition A.1.2 (Open Cover)**.** *For a set A an open cover is a collection of open subsets $(\mathcal{A}_i)_{i\in\mathcal{I}}$ such that the union of these sets contains A. A subcover is a subcollection of a cover of a set that still covers the set.*

Definition A.1.3 (Compact set)**.** *A set A is compact when each open cover has a finite subcover.*

Theorem A.1.4 (Heine-Borel)**.** *Let $A \subset \mathbb{R}^n$. TFAE:*

1. *A is compact*
2. *A is closed and bounded.*

A.1.1 Differentiation under Integral

It is often the case in Statistics that we stumble upon expressions such as:

$$\nabla_{\boldsymbol{x}} \int_{a(\boldsymbol{x})}^{b(\boldsymbol{x})} f(\boldsymbol{x}, t)\, \mathrm{d}t.$$

In this Subsection, we would like to state conditions that allow for the derivative to be brought inside the integral.

Example A.1.5. *In most applications, we will derive a result for $a(x) \equiv a, b(x) \equiv x$. This is simply:*

$$\frac{\mathrm{d}}{\mathrm{d}x} \int_a^x f(x,t)\, \mathrm{d}t = f(x,x) + \int_a^x \frac{\partial f(x,t)}{\partial x}\, \mathrm{d}t.$$

We now present sufficient conditions for this identity to hold.

Proposition A.1.6 (A version of Leibniz's integral rule)**.** *Let $f(x,t), \partial_x f(x,t)$ be continuous in their arguments in a region that for some x_0, x_1 includes the area:*

$$\{(x,t) : a(x) \leqslant t \leqslant b(x),\ x_0 \leqslant x \leqslant x_1\}.$$

Let the functions $a(\cdot), b(\cdot)$ be C^1 in the same slice $x_0 \leqslant x \leqslant x_1$. Then:

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_{a(x)}^{b(x)} f(x,t)dt = f(x,b(x))\frac{\mathrm{d}b(x)}{\mathrm{d}x} - f(x,a(x))\frac{\mathrm{d}a(x)}{\mathrm{d}x} + \int_{a(x)}^{b(x)} \frac{\partial f(x,t)}{\partial x}\,\mathrm{d}t.$$

Notice that this result is more general than Example A.1.5.

The proof of Prop. A.1.6 is easily found. The conditions can be relaxed if moving to a measure theoretic formulation.

Proposition A.1.7 (A stronger version of Leibniz's integral rule)**.** *Consider a measure space $(\mathscr{X}, \mathcal{F}, \mu)$. Let $A \subset \mathbb{R}$ be open. A function $f : A \times \mathscr{X} \to \mathbb{R}$ such that:*

1. *$f(a, \boldsymbol{x})$ is integrable in $\mathscr{X}$ for all $a \in A$.*
2. *$\partial_a f$ exists almost everywhere in $\mathscr{X}$ for all $a \in A$.*
3. *the partial derivative $\partial_a f$ is bounded by an integrable function $g : \mathscr{X} \to \mathbb{R}$ for all a and for almost every $\boldsymbol{x} \in \mathscr{X}$.*

Then, for any $a \in A$:

$$\frac{\mathrm{d}}{\mathrm{d}a}\int_{\mathscr{X}} f(a,\boldsymbol{x})\,\mathrm{d}\mu(\boldsymbol{x}) = \int_{\mathscr{X}} \partial_a f(a,\boldsymbol{x})\,\mathrm{d}\mu(\boldsymbol{x}).$$

Proof. **(preliminary object)** By definition of derivative, we have:

$$\frac{\partial}{\partial a} f(a,\boldsymbol{x}) = \lim_{h\to 0} \frac{f(a+h,\boldsymbol{x}) - f(a,\boldsymbol{x})}{h}.$$

Using this, consider a sequence $(h_n)_{n\in\mathbb{N}}$ such that $h_n \to 0$, and denote the sequence obtained by using $(h_n)_{n\in\mathbb{N}}$ above as $g_n(a,\boldsymbol{x})$. By #2 it holds that $\lim_{n\to\infty} g_n(a,\boldsymbol{x})$ is measurable[1].
(Mean Value Theorem and Dominated Convergence) By an application of the Mean Value Theorem and #3 one can state that[2]:

$$|g_n(a,\boldsymbol{x})| \leqslant \sup_{a\in A} |\partial_a f(a,\boldsymbol{x})| \leqslant g(\boldsymbol{x}) \quad \forall a \in A.$$

Therefore, by dominated convergence (Thm. A.1.12) applied for each $a \in A$:

$$\lim_{n\to\infty}\int_{\mathscr{X}} g_n(a,\boldsymbol{x})\mathrm{d}\mu(\boldsymbol{x}) = \int_{\mathscr{X}} \lim_{n\to\infty} g_n(a,\boldsymbol{x})\mathrm{d}\mu(\boldsymbol{x}). \tag{A.1.8}$$

On the LHS, we recognize the definition of derivative for the integral, and on the RHS the definition of integral of a derivative. Mathematically one can realize that:

$$\lim_{h_n\to 0} \frac{\int_{\mathscr{X}} f(a+h_n,\boldsymbol{x})\mathrm{d}\mu(\boldsymbol{x}) - \int_{\mathscr{X}} f(a,\boldsymbol{x})\mathrm{d}\mu(\boldsymbol{x})}{h_n} = \int_{\mathscr{X}} \partial_a f(a,\boldsymbol{x})\mathrm{d}\mu(\boldsymbol{x}),$$

which is the claim of the statement.

□

[1] Observe that is $f : \mathsf{cl}(A) \to \mathbb{R}$, where we have enlarged the definition of the function on its closure, is differentiable then the function:

$$f_n(x) \coloneqq \begin{cases} \frac{f\left(x+\frac{1}{n}\right)-f(x)}{\frac{1}{n}} & \text{if } x+\frac{1}{n} \in \mathsf{cl}(A) \\ 0 & \text{else} \end{cases}$$

then f_n is measurable and $f_n \overset{n\to\infty}{\to} f'$.

[2] The first inequality holds since by the MVT one has that $\inf_{a,b} f'(x) \leqslant \frac{f(b)-f(a)}{b-a} \leqslant \sup_{[a,b]} f'(x)$, where we use $[a,b] = [a, a+h_n] \subset \mathsf{cl}(A)$ and realize that the supremum over A will be larger for sure.

It turns out that these identities are much more general, and can be formulated in terms of distributions. We avoid reporting these formulations and just point out the reader to literature.

Further References

The topic is discussed in a general sense in (**talvilaNecessarySufficientConditions2001**). Statements in Real Analysis are found in (**follandRealAnalysisModern1999**) or (**amannAnalysisIII2009**). For a formulation in the theory of distributions, the first statements are found in (**schwartzTheorieDistributions1957**; **jonesTheoryGeneralisedFunctions1982**). Some examples and non-examples are found in (**conradDifferentiatingIntegralSign2016**).
A collection of notable statements is (**chengDifferentiationIntegralSign2010**). There it is mentioned that a clean proof for the most general one involving distributions is found in a report by the same author (**chengDifferentiationIntegralSign2010a**). This report is only found in a website that may be unstable (not uploaded by the author), but also could remain there forever. For this reason, a copy of it can be retrieved by emailing the author of this document.

Remark A.1.9. *Since the matter is in general very delicate, we will avoid discussing its conditions and assume throughout that integration and differentiation can be exchanged whenever necessary.*

A.1.2 Limiting under the integral

Throughout this subsection, assume a common measure space $(\mathscr{X}, \mathcal{F}, \mu)$.
Another common situation is the following. We are given an uncountable sequence of parametrized functions $(f_\theta(\boldsymbol{x}))_{\theta\in\mathbb{R}}$ and wish to do the following for a measurable set A:

$$\lim_{\theta\to\theta^*} \int_A f_\theta(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_A \lim_{\theta\to\theta^*} f_\theta(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{A.1.10}$$

Checking that this holds is in general hard, but some basic results in Measure Theory establish sufficient conditions.

Theorem A.1.11 (Monotone Convergence)**.** *Consider $(f_n)_{n\in\mathbb{N}}$ and $A \in \mathcal{F}$. Let $f_n : A \to \overline{\mathbb{R}_+}$ be non-negative and non-decreasing for all $\boldsymbol{x}$. Assume*[3] *$(f_n)_{n\in\mathbb{N}}$ are $(\mathcal{F}, \mathscr{B}(\overline{\mathbb{R}_+}))$ measurable and that the sequence converges pointwise to f, i.e.:*

$$f(x) \coloneqq \lim_{n\to\infty} f_n(\boldsymbol{x}).$$

Then the limit function f is $(\mathcal{F}, \mathscr{B}(\overline{\mathbb{R}_+}))$ measurable and:

$$\lim_{n\to\infty} \int_A f_n\,\mathrm{d}\mu = \int_A f\,\mathrm{d}\mu.$$

Proof. (**bogachevMeasureTheory2007**). □

Theorem A.1.12 (Dominated Convergence)**.** *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of measurable functions that converges pointwise to f. If:*

$$|f_n(\boldsymbol{x})| \leqslant g(\boldsymbol{x}) \quad \forall n, \forall \boldsymbol{x} \in \mathscr{X}$$

for some μ-integrable function g, then:

1. *the pointwise limit f is μ-integrable*
2. *it holds:*

$$\lim_{n\to\infty} \int_{\mathscr{X}} |f_n - f|\,\mathrm{d}\mu = 0$$

[3] the symbol $\mathscr{B}(\overline{\mathbb{R}_+})$ denotes the Borel sigma-algebra of the set $\overline{\mathbb{R}_+}$.

3. as a consequence of #2, it also holds that:

$$\lim_{n\to\infty}\int_{\mathscr{X}} f_n \,\mathrm{d}\mu = \int_{\mathscr{X}} f \,\mathrm{d}\mu.$$

Proof. (**bogachevMeasureTheory2007**). □

Corollary A.1.13. *The statements of Theorem A.1.11 and A.1.12 can be relaxed to μ-almost everywhere assumptions of non-decreasing and bounded respectively.*

For finite measures, a weaker condition is sufficient. This is useful when no bounding function is found. We use notions in Bogachev's notation (**bogachevMeasureTheory2007**). In particular, convergence in measure is sometimes termed global convergence. Similar statements can be found in (**follandRealAnalysisModern1999**).

Definition A.1.14 (Convergence in measure)**.** *Found in (***bogachevMeasureTheory2007***). Consider a sequence of functions $(f_n)_{n\in\mathbb{N}}$ and a function f all μ-measurable and defined on $\mathscr{X}$ taking values in $\mathbb{R}$. We say $(f_n)_{n\in\mathbb{N}}$ converges in measure to f and write $f_n \xrightarrow{\mu} f$ when for every $\epsilon > 0$:*

$$\lim_{n\to\infty} \mu\left(\{\boldsymbol{x} \in \mathscr{X} : |f_n(\boldsymbol{x}) - f(\boldsymbol{x})| \geqslant \epsilon\}\right) = 0.$$

We express this concisely by saying it is μ-convergent.

Definition A.1.15 (Uniform p-integrability)**.** *Let $p \geqslant 1$. A set of functions $F \subset \mathrm{L}^p(\mu)$ is uniformly integrable when:*

$$\lim_{H\to\infty} \sup_{f\in F} \int_{|f|^p > H} |f|^p \,\mathrm{d}\mu = 0.$$

By the definition of limit, this is equivalent to:

$$\forall \epsilon > 0,\ \exists H \equiv H(\epsilon)\ :\ \sup_{f\in F} \int_{|f|^p \geqslant H(\epsilon)} |f|^p \,\mathrm{d}\mu < \epsilon.$$

*The case for $p = 1$ is treated in (***bogachevMeasureTheory2007***).*

Definition A.1.16 (Uniform absolute continuous integrability)**.** *Let $p \geqslant 1$. A set of functions $F \subset \mathrm{L}^p(\mu)$ has uniformly absolutely continuous integrals if for every $\epsilon > 0$ there exists $\delta > 0$ such that:*

$$\mu(A) < \delta \implies \int_A |f|^p \,\mathrm{d}\mu < \epsilon \quad \forall f \in F.$$

*This is the arbitrary $p \geqslant 1$ case of (***bogachevMeasureTheory2007***).*

Proposition A.1.17. *Let μ be a finite measure and F be a family of μ-integrable functions. The collection is uniformly integrable if and only if it is bounded in L^1 and has uniformly continuous integrals.*
If μ is atomless, uniform integrability is just equivalent to uniformly continuous integrals at $p = 1$.

Proof. See (**bogachevMeasureTheory2007**). □

Theorem A.1.18 (Lebesgue-Vitali)**.** *Let μ be a finite measure and $(f_n)_{n\in\mathbb{N}} \subset \mathrm{L}^p(\mu)$ with $p \geqslant 1$ be a collection of measurable functions. TFAE:*

1. *for $f \in \mathrm{L}^p(\mu)$ it holds $f_n \xrightarrow{\mathrm{L}^p} f$*
2. *$f_n \xrightarrow{\mu} f$ and $(|f_n|^p)_{n\in\mathbb{N}}$ is uniformly integrable.*

Proof. The proof for $p = 1$ is found in (**bogachevMeasureTheory2007**). The adaptation for $p \geqslant 1$ is natural. Notice that here uniform integrability is equivalent to uniform absolute continuous integrability. □

To restate this result for non-finite measures we need to work slightly more.

Definition A.1.19 (p-Tightness)**.** *Consider a collection of functions $F = (f_n)_{n\in\mathbb{N}} \subset \mathrm{L}^p(\mu)$. We say the collection is p-tight when for every $\epsilon > 0$ there exists $A_\epsilon \in \mathcal{F}$ with finite measure such that :*

$$\sup_{f_n\in F} \int_{\mathscr{X}\setminus A_\epsilon} |f_n|^p \,\mathrm{d}\mu < \epsilon.$$

Remark A.1.20. *Tigthness is trivially verified for finite measures.*

Theorem A.1.21 (Egoroff)**.** *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of measurable functions taking values on a separable metric space. Let $A \subset \mathscr{X}$ have finite μ-measure. Assume $f_n \xrightarrow{\mu} f$ hwen restricted to A. Then for every $\epsilon > 0$ there exists $B \subset A$ measurable such that $\mu(B) < \epsilon$ and $f_n \rightrightarrows f$ on $A\setminus B$.*

Proof. See (**bogachevMeasureTheory2007**) or the handwritten notes (**torresLebesgueTheoryLecture2017**) or (**ziemerModernRealAnalysis2017**). □

Theorem A.1.22 (Lebesgue-Vitali for infinite measures)**.** *Let μ be a measure, $(f_n)_{n\in\mathbb{N}} \subset \mathrm{L}^p(\mu)$ with $p \geqslant 1$ be a collection of measurable functions and $f \in \mathrm{L}^p(\mu)$. Then the statement:*

$$f_n \xrightarrow{\mathrm{L}^p} f$$

is equivalent to the following three conditions:

1. $f_n \xrightarrow{\mu} f$
2. $(f_n)_{n\in\mathbb{N}}$ *has p-uniformly continuous integrals*
3. $(f_n)_{n\in\mathbb{N}}$ *is p-tight*

Proof. The $p = 1$ statement is found in (**bogachevMeasureTheory2007**) without proof. A reference with all the steps are the handwritten notes (**torresLebesgueTheoryLecture2017**) which follow (**ziemerModernRealAnalysis2017**), and relies on Egoroff's Thm. A.1.21. □

Fact A.1.23. *The almost everywhere version of the dominated convergence Theorem (Thm. A.1.12) is implied by Vitali Convergence.*

Proof. The sequence $(f_n)_{n\in\mathbb{N}}$ is assumed to be almost sure convergent to f, and dominated by an integrable g for each n. We want to show it is uniformly integrable and tight.
(Uniform Integrability) Fix $\epsilon > 0$. By the integrability and positivity of g, we take g_ϵ:

- measurable
- bounded
- with $\mathsf{supp}(g_\epsilon)$ finite
- such that $0 \leqslant g_\epsilon \leqslant g$ throughout
- satisfying $0 \leqslant \int_{\mathscr{X}} g \,\mathrm{d}\mu - \int_{\mathscr{X}} g_\epsilon \,\mathrm{d}\mu < \frac{\epsilon}{2}$.

Let $A \subset \mathscr{X}$ be measurable. By the positivity of $g - g_\epsilon$ we know:

$$\int_A g \,\mathrm{d}\mu - \int_A g_\epsilon \,\mathrm{d}\mu < \frac{\epsilon}{2}.$$

By the boundedness of g_ϵ there exists $M \geqslant 0$ such that $0 \leqslant g_\epsilon \leqslant M$. Combining the two facts yields:

$$\int_A g \,\mathrm{d}\mu < \int_A g_\epsilon \,\mathrm{d}\mu + \frac{\epsilon}{2} \leqslant M\mu(A) + \frac{\epsilon}{2}.$$

A choice $\delta = \frac{\epsilon}{2M}$ gives:

$$\mu(A) < \delta \implies \int_A g \,\mathrm{d}\mu < M\delta + \frac{\epsilon}{2} = \epsilon,$$

so that g is uniformly integrable by the arbitrariness of A. By assumption $|f_n| \leqslant g$, which holds also under integration and we get that $(f_n)_{n\in\mathbb{N}}$ is uniformly integrable.
(Tightness) To show that $(f_n)_{n\in\mathbb{N}}$ is tight it suffices to show g is. Fix $\epsilon > 0$ and choose again g_ϵ such that it satisfies the inequality of the previous point. By that condition, there exists a set $A \subset \mathscr{X}$ such that:

$$\mu(A) < \infty, \quad \mu(g_\epsilon \mathbb{1}_{\mathscr{X}\setminus A}) = 0.$$

Using this last identity, we can bound the value of g integrated outside of A:

$$\int_{\mathscr{X}\setminus A} g \,\mathrm{d}\mu = \int_{\mathscr{X}\setminus A} g - g_\epsilon \,\mathrm{d}\mu \leqslant \int_{\mathscr{X}} g - g_\epsilon \,\mathrm{d}\mu = \int_{\mathscr{X}} g \,\mathrm{d}\mu - \int_{\mathscr{X}} g_\epsilon \,\mathrm{d}\mu < \epsilon.$$

Then g is 1-tight. Accordingly the integral over $\mathscr{X}\setminus A$ of any f_n is bounded by that of g and $(f_n)_{n\in\mathbb{N}}$ is 1-tight. By the Vitali-convergence Theorem the limit f is integrable and it holds:

$$\lim_{n\to\infty} \int_{\mathscr{X}} f_n \,\mathrm{d}\mu = \int f \,\mathrm{d}\mu.$$

□

Remark A.1.24. *To move from a countable sequence to an uncountable sequence as we wished, it suffices to recall that the definition of $f_\theta(\boldsymbol{x}) \overset{\theta\to\theta^*}{\to} f_{\theta*}(\boldsymbol{x})$ is that for all sequences $(\theta_n)_{n\in\mathbb{N}}$ converging to θ^* the limit $f_{\theta_n}(\boldsymbol{x}) \to f_{\theta*}(\boldsymbol{x})$ holds.*

Recently, a necessary an sufficient condition for stating Equation A.1.10 was established for rather mild conditions[4] (**kamihigashiInterchangingLimitIntegral2020**).

A.2 Algebra

Let $\mathbb{K}$ be either the field of real numbers or of complex numbers.

Definition A.2.1 (Hermitian Transpose)**.** *For a matrix $\mathbf{A} \in \mathbb{K}^{n\times m}$, denote the Hermitian transpose (also, complex conjugate) with the dagger. Namely, identify with $\mathbf{A}^\dagger$ the object:*

$$(\mathbf{A}^\dagger)_{ij} \coloneqq \overline{A}_{ij},$$

where $\overline{A}_{ij} = \Re(A_{ij}) - \Im(A_{ij})$ is the complex conjugate of the element at the $(ij)^{th}$ entry.
Observe that this is equivalently written as:

$$\mathbf{A}^\dagger = \left(\overline{\mathbf{A}}\right)^\top = \overline{\mathbf{A}^\top},$$

since conjugation and transposition commute.

Definition A.2.2 (Hermitian matrix)**.** *A (square) matrix such that $\mathbf{A}^\dagger = \mathbf{A}$.*

Definition A.2.3 (Pseudoinverse)**.** *The pseudoinverse of $\mathbf{A} \in \mathbb{K}^{n\times m}$ is another linear map $\mathbf{A}^+ \in \mathbb{K}^{m\times n}$ that satisfies the following properties:*

(weak Id.) $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$

(weak Inv.) $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$

(Herm a.) $(\mathbf{A}\mathbf{A}^+)^\dagger = \mathbf{A}\mathbf{A}^+$

(Herm b.) $(\mathbf{A}^+\mathbf{A})^\dagger = \mathbf{A}^+\mathbf{A}$.

It is also termed Moore-Penrose inverse.

[4] Namely $\lim_{n\to\infty} f_n$ being integrable and the measure μ being sigma finite.

A.3 Probability facts

Definition A.3.1 (Almost Sure convergence)**.** *A sequence of random variables* $(X_n)_{n\in\mathbb{N}}$ *where* $\boldsymbol{X}_n : \Omega \to \mathscr{X}$ *on a common probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *converges almost sure (a.s.) to* $\boldsymbol{X}$ *on the same probability space if:*

$$\mathbb{P}\left[\omega \in \Omega \mid \lim_{n\to\infty} \boldsymbol{X}_n(\omega) = \boldsymbol{X}(\omega)\right] = 1.$$

In compact notation we write this as $\boldsymbol{X}_n \overset{a.s.}{\to} \boldsymbol{X}$*.*

A function is termed **Lebesgue integrable** when its Lebesgue integral is finite. A property of a probability space is said to hold **almost everyhwere** (a.e.) when it does not hold on a set that has measure zero.

A.4 Empirical Distributions

Consider a sequence of iid observations $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$, these have the same cumulative distribution function (cdf). This statistical setting was largely studied, and many results are known about the behavior of the sequence. In particular, we focus on a standard strong result that is sufficient for our purposes.

A.4.1 Vapnik-Chervonenkis Theory

Assume we are in the usual probability space $(\mathscr{X}, \mathcal{F}, \mathbb{P})$. The sigma-algebra $\mathcal{F}$ is a collection of sets. We wish to describe how this behaves. By convention, denote the intersection of a family of subsets $\mathcal{F}$ with a given set A as the object:

$$\mathcal{F} \cap A \coloneqq \{A \cap B \mid B \in \mathcal{F}\}$$

We say the class $\mathcal{F}$ **shatters** a set A if for each subset $B \subset A$ there exists $C \in \mathcal{F}$ such that $B = C \cap A$.

Fact A.4.1. *A class* $\mathcal{F}$ *shatters a set when* $\mathcal{F} \cap A = 2^A$*, the power set of* A*.*

Proof. By definition of $\mathcal{F} \cap A$. □

We implicitly make the assumption that A has finite size n, so that the size of the power set is $2^n < \infty$.

Definition A.4.2 (Shattering coefficient)**.** *Consider* $\mathcal{F}$ *seen as a set of sets and* $\mathscr{X}$ *be the ambient space. Its* n^{th} *shattering coefficient is:*

$$S_{\mathcal{F}}(n) \coloneqq \sup_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n \in \mathscr{X}} \left| \{\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \cap B,\ B \in \mathcal{F}\} \right|.$$

In words, it is the size of the most numerous set arising from a choice of n *elements inside the sample space and a set* B *from the collection of sets. It measures the size of* $A \cap \mathcal{F}$ *as a function of* $|A| = n$*. It is also termed* ***growth function****.*

Fact A.4.3. *For a collection of sets* $\mathcal{F}$ *it holds that:*

1. *the shattering coefficient is upper bounded,* $S_{\mathcal{F}}(n) \leqslant 2^n$*.*
2. *Tightness implies shattering: if* $S_{\mathcal{F}}(n) = 2^n$ *is tight then there exists* $A = \{x_1, \ldots, x_n\}$ *that is shattered by* $\mathcal{F}$ *and such that* $|A| = n$
3. *Monotonicity of non-shattering: if* $S_{\mathcal{F}}(N) < 2^N$ *for* $N > 1$ *then* $S_{\mathcal{F}}(n) < 2^n$ *is not tight for all* $n \geqslant N$*.*

Proof. **(Claim #1)** By Fct. A.4.1, any intersection of an element of $\mathcal{F}$ and A is contained in the power set, which has dimension 2^n.
(Claim #2) this is trivial, since the dimension of a shattering set is 2^n and we set the shattering coefficient to be 2^n. In other words, the collection of sets $\mathcal{F}$ is able to shatter at least one set $A \subset \mathscr{X}^n$.
(Claim #3) The shattering coefficient is solely related to the structure of $\mathcal{F}$. If it fails to be tight at $N > 1$, then it must fail to be tight for any $n \geqslant N$. □

Definition A.4.4 (Vapnik-Chervonenkis dimension)**.** *The Vapnik-Chervonenkis (VC) dimension of a collection of sets $\mathcal{F}$ is defined as:*

$$\mathsf{VC}(\mathcal{F}) \coloneqq \min_n \{n \ : \ S_{\mathcal{F}}(n) < 2^n\}.$$

In some references, one rather uses the definition with equality, which is just the same up to a ± 1:

$$\mathsf{VC}_0(\mathcal{F}) \coloneqq \max_n \{n \ : \ S_{\mathcal{F}}(n) = 2^n\} = \mathsf{VC}(\mathcal{F}) - 1.$$

In light of Fact. A.4.3(#1, #2, #3) we know that:

1. the shattering coefficient makes sense, since it is a finite number for each choice of $n \in \mathbb{N}$
2. the shattering coefficient attains its maximum when there is a shattered set
3. the shattering coefficient is such that once there exists a size with no shattered set, the collection of sets never shatters at higher dimensions.

Then, it is understood that if for any $n \in \mathbb{N}$ it is always possible to find a shattered set, then $\mathsf{VC}_0(\mathcal{F}) = \infty$, and accordingly $\mathsf{VC}(\mathcal{F}) = \infty$.

Definition A.4.5 (VC class)**.** *A collection of sets $\mathcal{F}$ such that $\mathsf{VC}(\mathcal{F}) < \infty$. Namely, a class for which there exists $N < \infty$ such that there is no shattering sets from there onwards.*

Lemma A.4.6 (Pascal's Lemma)**.** *For positive natural number n, k we have:*

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}.$$

Proof. **(Combinatorially)** The LHS is the number of ways in which n elements can be placed into k baskets. If we choose an arbitrary element i, the number of ways to place all the elements but i into k baskets is $\binom{n-1}{k}$. Similarly, the number of ways in which $n-1$ elements can be placed into $k-1$ baskets, when one is already occupied by i is $\binom{n-1}{k-1}$. The LHS is the sum of the two terms, since they encompass all possibilities.
(Algebraically) This is just a rearrangement of terms:

$$\begin{aligned}
\binom{n-1}{k} + \binom{n-1}{k-1} &= \frac{(n-1)!}{k!(n-1-k)!} + \frac{(n-1)!}{(k-1)!(n-k)!} && \text{(A.4.7)}\\
&= (n-1)!\left[\frac{1}{k!(n-k-1)!} + \frac{1}{(k-1)!(n-k)!}\right] && \text{(A.4.8)}\\
&= (n-1)!\left[\frac{n-k}{k!(n-k)!} + \frac{k}{k!(n-k)!}\right] && \text{(A.4.9)}\\
&= \frac{n!}{k!(n-k)!} = \binom{n}{k}. && \text{(A.4.10)}
\end{aligned}$$

□

Lemma A.4.11 (Sauer-Selah Lemma)**.** *Consider a family of sets $\mathcal{F}$ and a set A. If $\mathsf{VC}(\mathcal{F}) = S$ then:*

$$S_{\mathcal{F}}(n) \leqslant \sum_{s=0}^{S} \binom{n}{s} \quad \forall n.$$

Proof. The proof we show was found in (**balcanVapnikChervonenkisDimensionSauer2011**) and is interesting because it is a good exercise for induction on two parameters. Also, it highlights a different interpretation of the objects in play. Three alternative proofs can be found in (**ngoThreeProofsSauerShelah2010**).
Proceed by induction on (S, n). Let $\Xi_S(n) \coloneqq \sum_{s=0}^{S} \binom{n}{s}$.
(Base case A) For $n = 0$ and $S \in \mathbb{N}$ we have only one subset, so that $S_{\mathcal{F}}(n) \leqslant 1 = \Xi_S(0)$.

(Base case B) Let $S = 0$ and $n \in \mathbb{N}$. Then, no set can be shattered, and there is a unique assignment for each point. Hence, $S_{\mathcal{F}}(n) = 1 \leqslant \Xi_S(0)$.
(Induction hypothesis) Assume the statement $S_{\mathcal{F}}(n') \leqslant \Xi_{S'}(n')$ is true for each $n' \leqslant n, S' \leqslant S$ choice in the natural numbers with at least one of the two inequalities strictly true.
(Inductive step, and a different perspective) Each element $B \in \mathcal{F}$ of the sigma algebra can be seen as a function $f_B : \mathscr{X} \to \{0, 1\}$ which is the indicator function $f_B \coloneqq \mathbb{1}_B$. Accordingly, the sigma-algebra is a collection of binary functions (a concept class in (**balcanVapnikChervonenkisDimensionSauer2011**)). For a set $A \subset \mathscr{X}$, define the following object:

$$D_{\mathcal{F}}(A) \coloneqq \{(f_B(a_1), \ldots, f_B(a_{|A|})) \ : \ B \in \mathcal{F}\} \subseteq \{0, 1\}^n,$$

which is the set of functions that map the elements of $A = \{a_1, \ldots, a_{|A|}\}$ to a binary vector. Define a class of functions $\mathcal{G} = \{g : A \to \{0, 1\}\}$ such that $D_{\mathcal{F}}(A) = \mathcal{G}(A) = \mathcal{G}$. A set $\widetilde{A} \subseteq A$ shattered by $\mathcal{G}$ is also shattered by $\mathcal{F}$ since the constructions coincide. Hence, it necessarily holds $\mathsf{VC}(\mathcal{G}) \leqslant \mathsf{VC}(\mathcal{F})$.
(Building candidates and a representative function) The aim now is providing a sensible construction of two sets of functions $(\mathcal{G}_1, \mathcal{G}_2)$ for which we can apply the induction hypothesis. For $g \in \mathcal{G}$ there corresponds a binary labeling of the elements $A_{-1} = \{a_1, \ldots, a_{n-1}\}$ (notice we are ignoring one, hence the minus). For each of these sublabelings, we build a representative function g as follows:

1. the mapping satisfies $g : \mathcal{G} \to \mathcal{G}_1$, with $\mathcal{G}_2 = \mathcal{G} \backslash \mathcal{G}_1$ yet to specify
2. for each function $g \in \mathcal{G}_2$ there must exist a $\widetilde{g} \in \mathcal{G}_1$ such that $(g, \widetilde{g})$ agree on the set $A_{-1} = \{a_1, \ldots, a_{n-1}\}$ and disagree on the last term a_n that $(g, \widetilde{g})$ are representing.
3. We set wlog $g(a_n) = 1, \widetilde{g}(a_n) = 0$. In words $\mathcal{G}_2$ is labeling the n^{th} point as positive.

The construction ensures that:

$$|D_{\mathcal{F}}(A)| = |\mathcal{G}(A)| = |\mathcal{G}_1(A)| + |\mathcal{G}_2(A)|. \tag{A.4.12}$$

Hence, by the inclusion $\mathcal{G}_1 \subseteq \mathcal{G}$ the following bound holds $\mathsf{VC}(\mathcal{G}_1) \leqslant \mathsf{VC}(\mathcal{G}) \leqslant S$.
(subproof) We now want to show that:

$$|\mathcal{G}_1(A)| = |\mathcal{G}_1(A \backslash \{a_n\})|,$$

which means that the added number on binaries does not matter in terms of how the labeling is defined. Proceeding by double inclusion, one direction is trivial. The hard direction holds since there is no function g such that:

$$g(A_- \cup \{0\}) \in \mathcal{G}_1, \quad g(A_- \cup \{1\}) \in \mathcal{G}_1,$$

by the construction above. By the induction hypothesis, we have that

$$|\mathcal{G}_1(A)| \leqslant \Xi_S(n-1). \tag{A.4.13}$$

(back to inductive step) Consider a set C. If C is shattered by $\mathcal{G}_2$ then $C \cup \{a_n\}$ is shattered by $\mathcal{G}$. To see this notice that:

- $a_n \notin C$ since all maps $g \in \mathcal{G}_2$ are labeling a_n as positive by the convention #3 above
- necessarily $C \cup \{a_n\}$ is shattered by $\mathcal{G}$ since there is a negative labeling function $\widetilde{g} \in \mathcal{G}_1$ in the opposite set that guarantees this shattering.

Consequently:

$$\mathsf{VC}(\mathcal{G}_2) \leqslant \mathsf{VC}(\mathcal{G}) - 1 \leqslant S - 1. \tag{A.4.14}$$

Following the subproof above, we also claim the equality $|\mathcal{G}_2(A)| = |\mathcal{G}_2(A \backslash \{a_n\})|$. Hence, by this fact and Equation[5] A.4.14 we conclude analogously:

$$|\mathcal{G}_2(A)| \leqslant \Xi_{S-1}(n-1).$$

[5] crucially, notice that we are obtaining $\leqslant S - 1$ instead of $\leqslant S$ as in Equation A.4.13

Returning to the decomposition of Equation A.4.12, the steps above ensure that:

$$|D_{\mathcal{F}}(A)| \leqslant \Xi_S(n-1) + \Xi_{S-1}(n-1). \tag{A.4.15}$$

(sum simplification) What remains is to express Equation A.4.15 in terms of $\Xi_S(n)$. This is done by just rearranging terms. As a first step:

$$\Xi_S(n-1) + \Xi_{S-1}(n-1) = \sum_{s=0}^{S} \binom{n-1}{s} + \sum_{s=0}^{S-1} \binom{n-1}{s} = \binom{n-1}{0} + \sum_{s=1}^{S} \binom{n-1}{s} + \sum_{s=1}^{S} \binom{n-1}{s-1}, \tag{A.4.16}$$

where we have taken out the first term in the first sum and changed the index of summation in the second. Now using the fact that $\binom{n}{0} = \binom{n-1}{0} = 1$ for all $n \geqslant 1$ we have a term in front and two summations over the same index:

$$\Xi_S(n-1) + \Xi_{S-1}(n-1) = \binom{n}{0} + \sum_{s=1}^{S} \binom{n-1}{s} + \binom{n-1}{s-1}. \tag{A.4.17}$$

The combinatorial identity of Pascal's Lemma applied on (n, s) for each s gives the claim, with base case $s = 0$ being $\binom{n}{0}$. □

Lemma A.4.18. *Notice that for $n \geqslant S$, there is an additional bound $\sum_{s=0}^{S} \binom{n}{s} \leqslant \left(\frac{en}{S}\right)^S$, so the conclusion above is expressed in asymptotic notation as:*

$$n \geqslant \delta, \quad \mathsf{VC}(\mathcal{F}) = S \implies S_{\mathcal{F}}(n) \in O\left(n^S\right). \tag{A.4.19}$$

Proof. This is an easy combinatiorial bound. By assumption $0 \leqslant \frac{S}{n} \leqslant 1$. Thus:

$$\left(\frac{S}{n}\right)^S \sum_{s=0}^{S} \binom{n}{s} \leqslant \sum_{s=0}^{S} \left(\frac{S}{n}\right)^s \binom{n}{s} \leqslant \sum_{s=0}^{n} \left(\frac{S}{n}\right)^s \binom{n}{s} = \left(1 + \frac{S}{n}\right)^n \leqslant e^S. \tag{A.4.20}$$

Rearranging, we get the claim. □

Corollary A.4.21. *An equivalent statement is if $\mathcal{F}$ has n <u>distinct</u> elements and $|\mathcal{F}| > \sum_{s=0}^{S-1} \binom{n}{s}$ then $\mathcal{F}$ shatters a set of size S.*

Proof. We prove that if Lemma A.4.11 is $A \implies B$, here we have $\neg B \implies \neg A$. This holds by construction. The hypothesis that the VC dimension is S is paired with $\mathcal{F}$ not shattering an S-sized set. These are obviously the negation of each other. The conclusions are also in contradiction. □

A.4.2 Glivenko-Cantelli Theory

In realistic scenarios, we only have access to a finite sample from a population. If this population is not deterministic, then we will observe that the randomness is manifested through the n phenomena. If we assume that these come from the same distribution, the notion of empirical distribution function is natural.

Definition A.4.22 (Empirical Distribution function)**.** *The iid sequence $(\boldsymbol{X}_n)_{n\in\mathbb{N}}$ of n iid samples from $\mathbb{P}$ has empirical measure:*

$$\mathbb{P}_n(\cdot) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\boldsymbol{X}_i}(\cdot), \tag{A.4.23}$$

which evaluates probabilities over Borel sets $A \subset \mathcal{F}$ as:

$$\mathbb{P}_n(A) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\boldsymbol{X}_i}(A). \tag{A.4.24}$$

More in general, for a collection of measurable functions $\{f : \mathscr{X} \to \mathbb{R}\}$ we define its empirical measure as:

$$\mathbb{P}_n(f) \coloneqq \int f \, \mathrm{d}\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{X}_i). \tag{A.4.25}$$

When we wish to emphasize the empirical distribution/measure, we might also use $\mathsf{emp}(\cdot)$ for clarity (as done in Chapter III).

A classic result in Probability Theory states that the empirical distribution function of a real valued process has good convergence properties. This is the content of the Glivenko-Cantelli Theorem.

Theorem A.4.26 (Glivenko-Cantelli)**.** *If $\{X_i\}_{i=1}^n$ are iid with cdf F then:*

$$\|F(x) - F_n(x)\| = \sup_{x\in} |F(x) - F_n(x)| \xrightarrow[n\to\infty]{\text{a.s.}} 0. \tag{A.4.27}$$

In words, the empirical process converges almost surely uniformly in $\mathbb{R}$.

If one wishes to adapt this to multidimensional empirical processes, the result is more delicate. We need to establish modes of convergence and uniform convergence for measurable sets or measurable functions and link this with a geometrical result in the flavour of Vapnik-Chervonenkis, who were the first to establish the connection.

Further References

The next results are stated withour proof, as they are a mix of different statements. To check all the required steps, a clean summary with proofs and connections is given in (**caramanisUniformGlivenkoCantelli2000**), or other short notes (**kahleGlivenkoCantelliTheorem2006**; **rabanGlivenkoCantelliTheoremIntroduction**; **banerjeeEmpiricalProcessesGlivenko2010**). Some seminal works that established a network of results are (**vapnikUniformConvergenceRelative1971**; **talagrandGlivenkoCantelliProblem1987**; **dudleyUniversalDonskerClasses1987**; **dudleyUniformUniversalGlivenkoCantelli1991**). In terms of pedagogical lecture notes, some worth mentioning for an understanding and extended references are (**wellnerEmpiricalProcessesTheory2004**; **senGentleIntroductionEmpirical2022**). A book with fundamental results is (**dudleyUniformCentralLimit1999**).

For this reason, we wish to inspect the multidimensional empirical measure in relation to a set of measurable sets $\mathscr{C}$ and a set of measurable functions $\mathscr{A}$. When we express a norm wrt to either of the sets, we mean the supremum among all the choices inside. For example, $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{A}} := \sup_{f\in\mathscr{A}} |\mathbb{P}_n(f) - \mathbb{P}(f)|$, where we use the standard notation for Measure Theory[6]. For simplicity, we assume that the expectations are always finite so that we can integrate without having to check.

Remark A.4.29. *The norms are random variables since they depend on the n sized sample.*

Definition A.4.30 (Glivenko-Cantelli class)**.** *An collection of sets is a Glivenko-Cantelli (GC) class if either of the following are true:*

$$\|\mathbb{P} - \mathbb{P}_n\|_{\mathscr{C}} \xrightarrow[n\to\infty]{\text{a.s.}} \to 0 \tag{A.4.31}$$

$$\|\mathbb{P} - \mathbb{P}_n\|_{\mathscr{C}} \xrightarrow[n\to\infty]{\text{p}} \to 0 \tag{A.4.32}$$

$$\|\mathbb{P} - \mathbb{P}_n\|_{\mathscr{C}} \xrightarrow[n\to\infty]{\text{L}^1} \to 0, \tag{A.4.33}$$

where in particular it can be shown that they are equivalent.

Definition A.4.34. *A GC class of functions is analogous to the GC class above.*

Definition A.4.35 (Universal GC class)**.** *A class $\mathcal{F}$ is universal GC if it is a GC class wrt any $\mathbb{P}$ on $(\mathscr{X}, \mathcal{F})$.*

[6] namely:

$$\mu f := \mathbb{E}_\mu [f]. \tag{A.4.28}$$

Definition A.4.36 (Uniform GC class)**.** *A class $\mathcal{F}$ is uniform GC if in addition the globality wrt to any $\mathbb{P}$ on $(\mathscr{X}, \mathcal{F})$ is uniform. Mathematically:*

$$\sup_{\mathbb{P}\in\mathbb{P}(\mathscr{X},\mathcal{F})} \mathbb{E}\left[\|\mathbb{P}-\mathbb{P}_n\|_{\mathscr{C}}\right] \underset{n\to\infty}{\to} 0. \tag{A.4.37}$$

Definition A.4.38 (Suslin Space)**.** *A measurable space $(\mathscr{J}, \mathcal{J})$ such that it admits a Borel measurable function that is a surjection from a Polish space $\mathscr{P}$ onto S.*

Definition A.4.39 (Image admissible Suslin)**.** *Consider a measure space $(\mathscr{X}, \mathcal{F})$, a set $\mathscr{A}$ of functions on $\mathscr{X}$ and a map:*

$$M : \mathscr{A} \times \mathscr{X} \to \mathbb{R}. \tag{A.4.40}$$

We say M is image admissible Suslin via $(\mathscr{J}, \mathcal{J}, G)$ if

- *$(\mathscr{J}, \mathcal{J})$ is Suslin*
- *$J : \mathscr{J} \to \mathscr{X}$ is a surjection*
- *the map $(\boldsymbol{j}, \boldsymbol{x}) \to M(G(\boldsymbol{j}), \boldsymbol{x})$ is measurable on $\mathscr{J} \times \mathscr{X}$.*

We say a class is image admissible Suslin over the standard measure space if there exists a function M and a triplet satisfying the conditions.

Theorem A.4.41 (VC-GC connection)**.** *Let $\mathscr{A}$ be image admissible Suslin, TFAE:*

1. *$\mathscr{A}$ is uniform GC*
2. *$\mathscr{A}$ is VC in the sense of Def. A.4.5.*

Eventually, for uniform almost sure convergence of empirical measures, one can check that the class of test functions is a VC class, to obtain that in that precise restriction (though very wide) of space the result is valid.

A.5 More About Stein and Counting

As a general interesting fact, we quickly report a scalar converse of Stein's Lemma that can be found in (**nourdinMultivariateNormalApproximation2010**). To prove it, we first need a Lemma of independent interest. The two statements are found in (**chenFundamentalsSteinMethod2011**)

Lemma A.5.1. *Fix $z \in \mathbb{R}$. Let $\Phi(z) = \mathbb{P}[Z \leqslant z]$ be the cdf of Z. The unique bounded solution $f(x) \coloneqq f_z(x)$ to the equation:*

$$f'(x) - xf(x) = \mathbb{1}_{\{x\leqslant z\}} - \Phi(z) \tag{A.5.2}$$

is:

$$f_z(x) = \begin{cases} \sqrt{2\pi}e^{\frac{x^2}{2}}\Phi(x)[1-\Phi(z)] & x \leqslant z \\ \sqrt{2\pi}e^{\frac{x^2}{2}}\Phi(z)[1-\Phi(x)] & x > z. \end{cases} \tag{A.5.3}$$

Proof. Multiply the LHS and RHS of Eqn. A.5.2 by $e^{\frac{x^2}{2}}$. This gives:

$$(e^{\frac{x^2}{2}}f(x))' = e^{\frac{x^2}{2}}\left(\mathbb{1}_{\{x\leqslant z\}} - \Phi(z)\right). \tag{A.5.4}$$

Integrating the two over $\mathrm{d}x$ yields:

$$f_z(x) = e^{\frac{x^2}{2}}\int_{-\infty}^{x}[\mathbb{1}_{\{x\leqslant z\}} - \Phi(z)]e^{\frac{-x^2}{2}}\,\mathrm{d}x = -e^{\frac{x^2}{2}}\int_{x}^{\infty}[\mathbb{1}_{\{x\leqslant z\}} - \Phi(z)]e^{\frac{-x^2}{2}}\,\mathrm{d}x. \tag{A.5.5}$$

The last equation is equivalent to the claim. To see that $f_z(x)$ is bounded, refer to the statement in (**chenFundamentalsSteinMethod2011**) and the explanations therein.
Lastly, we claim that the solution to Eqn. A.5.2 is $f_z(x) + C$. This constant is a solution to the homogeneous equation, and thus has form $C = ce^{\frac{x^2}{2}}$ for some c constant. To have a bounded solution, we are forced to take $c = 0$. □

Proposition A.5.6 (Partial Converse of Stein's Lemma that characterizes Scalar Gaussians)**.** *A random variable X has Standard Gaussian Law if and only if:*

$$\mathbb{E}\left[f'(X) - Xf(X)\right] = 0, \tag{A.5.7}$$

for each continuous piecewise differentiable function with $\mathbb{E}\left[|f'(X)|\right] < \infty$. In particular, the direction Standard Gaussian $\implies$ Equation A.5.7 can be relaxed to "for all absolutely continuous functions with integrable derivative".

Proof. (**steinBoundErrorNormal1972**) is one of the early references. A pedagogical proof is as follows.
(relaxed direction) Let $f : \mathbb{R} \to \mathbb{R}$ be absolutely continuous with integrable derivative. If X is standard Gaussian then:

$$\mathbb{E}\left[f'(X)\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f'(x) e^{-\frac{x^2}{2}} \, \mathrm{d}x \tag{A.5.8}$$

$$= \frac{1}{\sqrt{2\pi}} \left[\int_{-\infty}^{0} f'(x) \left(\int_{-\infty}^{x} -we^{-\frac{w^2}{2}} \, \mathrm{d}w \right) \mathrm{d}x + \int_{0}^{\infty} f'(x) \left(\int_{x}^{\infty} we^{-\frac{w^2}{2}} \, \mathrm{d}w \right) \mathrm{d}x \right], \tag{A.5.9}$$

by using the integral representation of the exponential function on the two half planes. Then an application of Fubini's Theorem with the right change of variables in the integration indexes gives:

$$\mathbb{E}\left[f'(X)\right] = \frac{1}{\sqrt{2\pi}} \left[\int_{-\infty}^{0} \left(\int_{w}^{0} f'(x) \, \mathrm{d}x \right) (-w) e^{-\frac{w^2}{2}} \, \mathrm{d}w + \int_{0}^{\infty} \left(\int_{0}^{w} f'(x) \, \mathrm{d}x \right) we^{-\frac{w^2}{2}} \, \mathrm{d}w \right] \tag{A.5.10}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} [f(w) - f(0)] we^{-\frac{w^2}{2}} \, \mathrm{d}w \tag{A.5.11}$$

$$= \mathbb{E}\left[Xf(X)\right]. \tag{A.5.12}$$

(Eqn. A.5.7 $\implies$ Standard Gaussian) Consider the function $f_z(x)$ of Lemma A.5.1. It is continuous and piecewise continuously differentiable. The discussion in the proof also explains that $f_z(x)$ is bounded. Then the identity $\mathbb{E}\left[f'(X)\right] = \mathbb{E}\left[Xf(X)\right]$ holds in particular for $f_z(x)$ and we get:

$$0 = \mathbb{E}\left[f_z(X) - Xf_z(X)\right] = \mathbb{E}\left[\mathbb{1}_{\{X \leqslant z\}} - \Phi(z)\right] = \mathbb{P}(X \leqslant z) - \Phi(z), \tag{A.5.13}$$

and X has a Gaussian Law by the arbitrariness of z.

□

We introduce some notation taken from the original formulation (**isserlisFormulaProductMomentCoefficient**
and presented succintly in (**kennethtayWhatIsserlisTheorem2019**). Consider a collection $\mathcal{A} \coloneqq (a_1, \ldots, a_N)$ where possibly numbers are repeatedly sampled from the integers in $\{1, \ldots, d\}$. A vector $\boldsymbol{X} \in \mathbb{R}^d$ has product in $\mathcal{A}$ defined with the expression:

$$X_{\mathcal{A}} = \prod_{a_i \in A} X_{a_i}, \quad X_{\varnothing} \coloneqq 1. \tag{A.5.14}$$

Then $\Pi(\mathcal{A})$ denotes the set of pairings (partitions into disjoint pairs) of $\mathcal{A}$. For a pairing $\pi \in \Pi(\mathcal{A})$ we let $\mathcal{A}\backslash\pi$ be the set of indices $i \in [d]$ such that we could express the pairing as:

$$\{(a_i, a_{\pi(i)}) \mid i \in \mathcal{A}\backslash\pi\} \tag{A.5.15}$$

Remark A.5.16. *Note that $\Pi(\mathcal{A}) = \varnothing$ if $\mathcal{A}$ has odd size. Simply, there are no pairings since one element is always left out.*

Example A.5.17. *Take $(d = 4, N = 6)$. Consider a set $\mathcal{A} = \{1, 2, 3, 4, 1, 2\}$, where we used colors for repetitions. A simple pairing is placing adjacent elements as pairs. We then have $\pi = (\{1, 2\}, \{3, 4\}, \{1, 2\})$. Then $\mathcal{A}\backslash\pi = \{1, 3, 1\}$, where the map $\pi(i)$ performs $\pi(1) = 2, \pi(2) = 4, \pi(1) = 2$. Notice that once we choose a pairing π the set $\mathcal{A}\backslash\pi$ is represented by essentially the same function.*

Armed with this construction and a set $\mathcal{A}$, we can define the structure of the Hafnian:

$$\mathsf{Haf}(\boldsymbol{X};\mathcal{A}) := \sum_{\pi\in\mathcal{A}}\prod_{i\in\mathcal{A}\setminus\pi} X_{a_i}X_{a_{\pi(i)}}, \tag{A.5.18}$$

In words, for a fixed set $\mathcal{A}$, list all the possible pairings, and for each of these, make the product of the elements in the set $\mathcal{A}\setminus\pi$.

Remark A.5.19. *The Hafnian is an object connected to the notions of Determinant, Permanent and Pfaffian. We gloss over these details and just provide a statement.*

Example A.5.20. *We restrict ourselves to the simplest non trivial example. Let* $\mathcal{A} = \{1,2,3,1\}$. *Then the Hafnian is:*

$$\sum_{\pi\in\mathcal{A}}\prod_{i\in\mathcal{A}\setminus\pi} X_{a_i}X_{a_{\pi(i)}} = X_1X_1\cdot X_2X_3 + X_1X_2\cdot X_1X_3 + X_1X_2\cdot X_1X_3 \tag{A.5.21}$$

$$= X_1^2\cdot X_2X_3 + 2X_1X_2\cdot X_1X_3, \tag{A.5.22}$$

where in the first step we kept the colors, but eventually the index is the same.

In the context of Gaussian Vectors, a very nice statements about these combinations of moments can be established, notably, this is equivalent to a generalization of Stein's Lemma.

Theorem A.5.23 (Isserli's Theorem)**.** *Given a Gaussian vector* $\boldsymbol{X}\in\mathbb{R}^d$ *with zero mean, it holds:*

$$\mathbb{E}\left[X_{\mathcal{A}}\right] = \sum_{\pi\in\mathcal{A}}\prod_{i\in\mathcal{A}\setminus\pi}\mathbb{E}\left[X_{a_i}X_{a_{\pi(i)}}\right] = \mathsf{Haf}(\mathbb{E}\otimes\boldsymbol{X};\mathcal{A}), \tag{A.5.24}$$

where the tensor product symbol is used to denote element-wise expectation.
Namely, the $\mathcal{A}$-Hafnian of the Covariances in the entries is equal to the expectation of a power of the entries instructed by the same $\mathcal{A}$ choice of integers.
This result is often termed Wick's Probability formula.

Proof. (**jansonGaussianHilbertSpaces1997**). □

Remark A.5.25. *Notice that this is equivalent to the multivariate version of Stein's Lemma (Cor. II.4.33). We only prove here that Gaussian integration by parts implies Isserli's result.*
By induction on n, where $f:\mathbb{R}^n\to\mathbb{R}$ this is trivial. Taking $f(\boldsymbol{x}) = x_2\cdots x_n$ the result follows.

Remark A.5.26. *Notice that:*

1. *Taking $|\mathcal{A}|$ odd, there are no pairings, and the expectation is null.*
2. *for even size (say $2n$), there are $\frac{(2n)!}{2^n n!} = (2n-1)!!$ pairings*
3. *Given #1, #2, we immediately recover as a special case the result for moments of Centered independent Gaussians $X\sim\mathcal{N}(0,\sigma^2)$, which admit an expression:*

$$\mathbb{E}\left[X^n\right] = \begin{cases}(n-1)!!\sigma^n & n \text{ even}\\ 0 & n \text{ odd}\end{cases}. \tag{A.5.27}$$