An Introduction to Maximum Entropy Methods

Marcos Costa Santos Carreira

Global Uncertainty Reading Group

20-Apr-2023

Contents

A new season

This time there's not much previous material

- Nassim's paper with Helyette and Donald Geman
- papers and books on Maximum Entropy

Let's work it out as we go

Definitions

- Discrete (always non-negative):

$$h(p) = -\sum_{i=1}^{n} p_i \cdot \log(p_i)$$

- Continuous (Differential Entropy, can be negative):

$$h(p) = -\int_I p(x) \cdot \log(p(x))$$

Information gain

- There is a "right" choice/answer
- We want to reduce uncertainty with the lower effort (number of features) possible
- We want the decision chain that reduces entropy by the most according to some cost function (average number of questions, max number of questions, weighted cost of asking questions, etc.)

Wordle

- One choice of known words, test words reduce set of possibilities
- Easier to look at the maximum information gain for each round
- Best test word (specially on the 3rd round) might not be among the possible answers
- Useful for familiarization with these concepts (program a game solver is more fun than playing it afterwards)

Principle of Maximum Entropy

- Choose between (families of) possible densities
- Constraints (mean, variance, ...)
- The answer
  - Follows these constraints
  - And has the maximum possible entropy

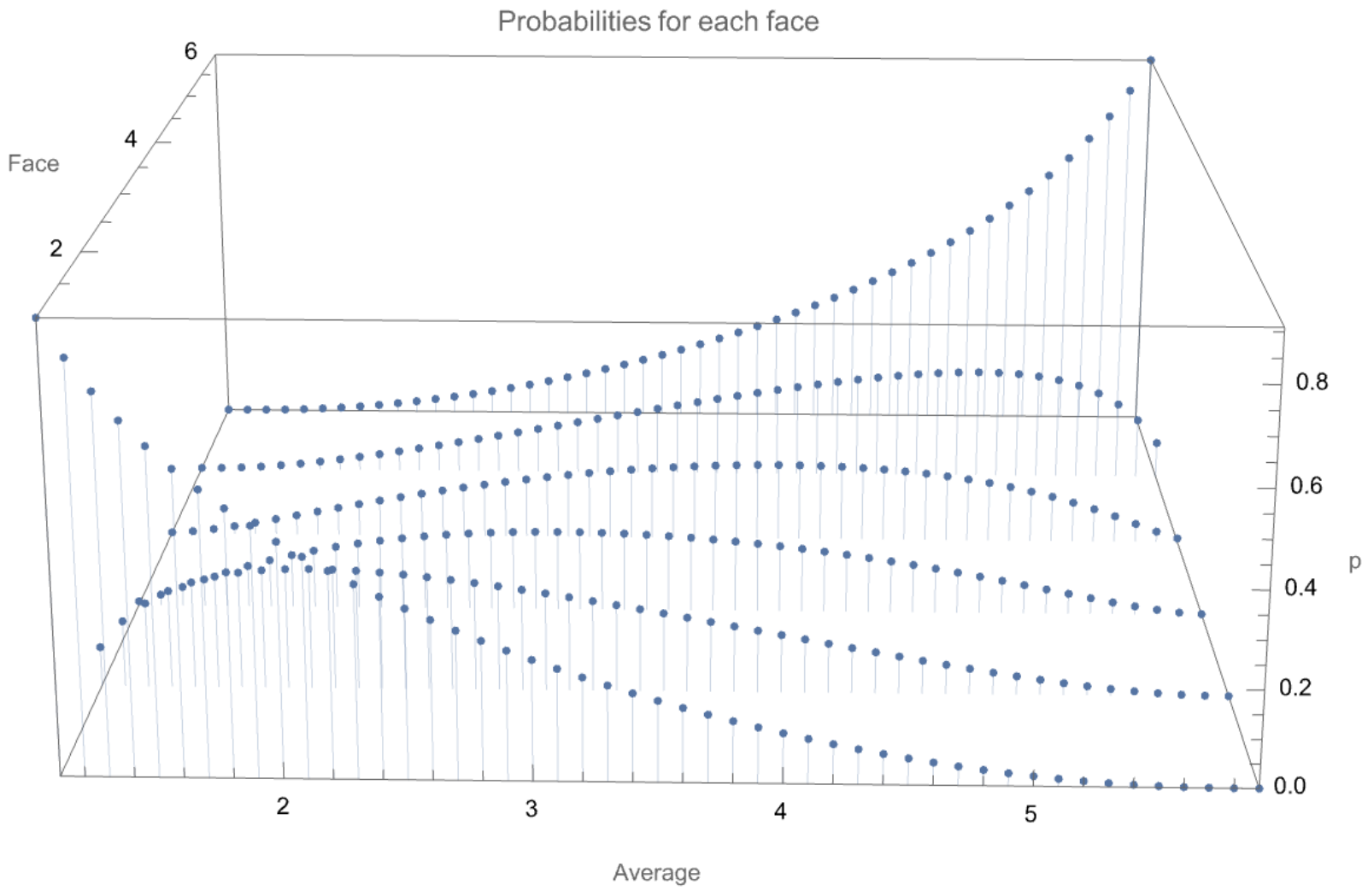
Brandeis Dice (Jaynes)

- Repeated throws of a die that might be weighted
- Given the average x of a large number of throws, what is the expected probability for each number on the next toss?

Mathematica code

```
H[x_] := Maximize[{-Sum[(p_i Log[p_i]), {i, 1, 6}], Sum[(p_i), {i, 1, 6}] == 1, Sum[(i p_i), {i, 1, 6}] == x}, Table[p_i, {i, 1, 6}]];

H[4.5]

{1.61359, {p_1 → 0.054354, p_2 → 0.0787725, p_3 → 0.114161, p_4 → 0.165447, p_5 → 0.239774, p_6 → 0.347492}}

N[H[3]]

{1.74851, {p_1 → 0.246782, p_2 → 0.20724, p_3 → 0.174034, p_4 → 0.146148, p_5 → 0.122731, p_6 → 0.103065}}
```

Plot


Probabilities for each face
Face
6
4
2
2
3
4
5
Average
0.8
0.6
0.4
0.2
0.0
p

Two approaches (Conrad)

- Given a distribution q, which constraints / extra information on p are needed
- Maximize the entropy with Lagrange multipliers and find the (family of) distributions

First approach

- Look at the integral

$$-\int_I p(x) \cdot \log(q(x))$$

- Substitute the desired distribution in place of q (e.g. $q(x) = \frac{\exp(-x/\lambda)}{\lambda}$):

$$-\int_0^\infty [p(x) \cdot \log(q(x))]\, dx = \log(\lambda) + \frac{1}{\lambda} \cdot \int_0^\infty [x \cdot p(x)]\, dx$$

- The last term is the mean of p

Examples

- From Conrad

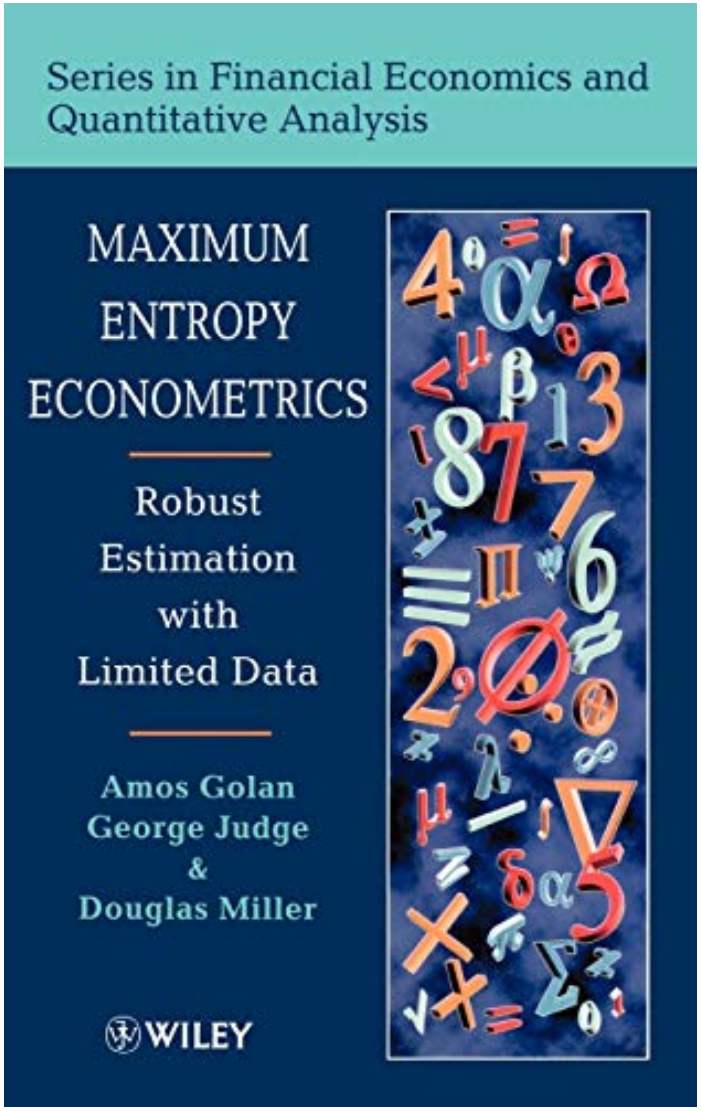
Distribution	Domain	Fixed Value
Uniform	Finite Set	None
Normal with mean μ	$\mathbf{R}$	$\int_{\mathbf{R}}(x-\mu)^2 p(x)\,\mathrm{d}x$
Exponential	$(0,\infty)$	$\int_0^\infty xp(x)\,\mathrm{d}x$

TABLE 1. Extra constraints

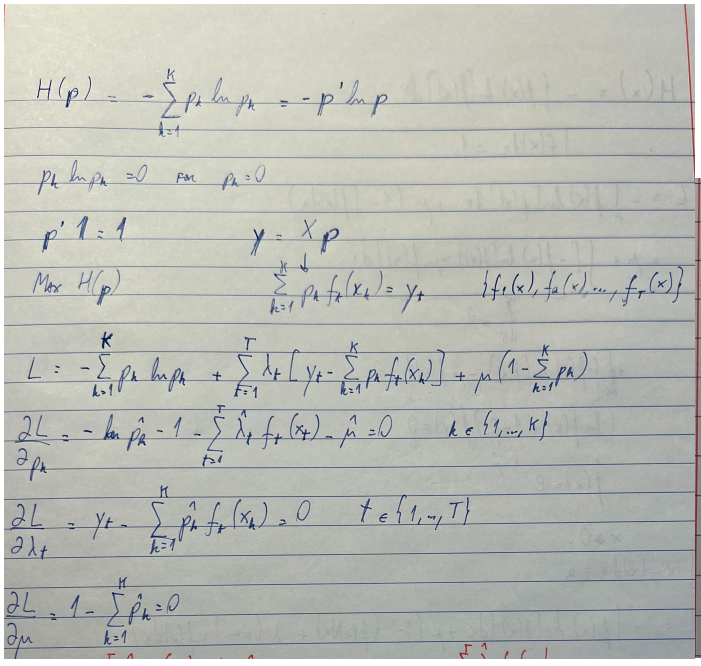
- Wikipedia: `https://en.wikipedia.org/wiki/Maximum_entropy_probability_distribution`

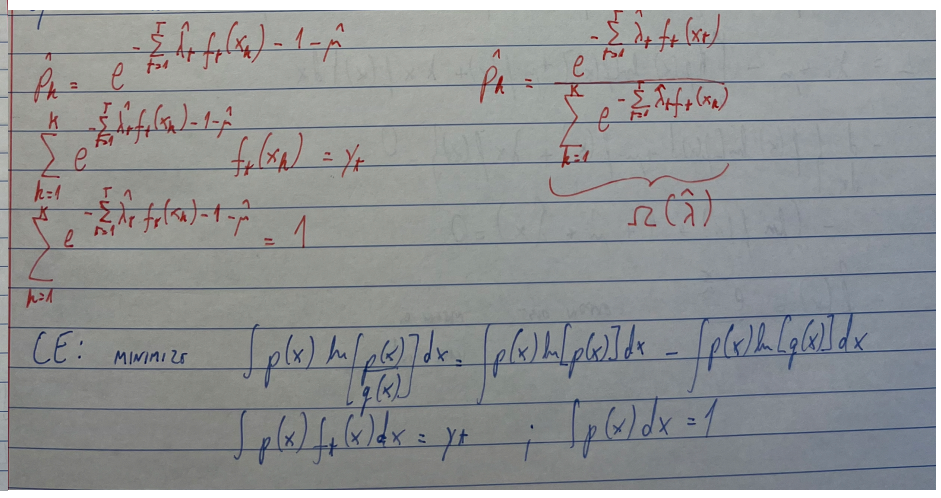
Lagrange multipliers

- This is the approach of Golan:

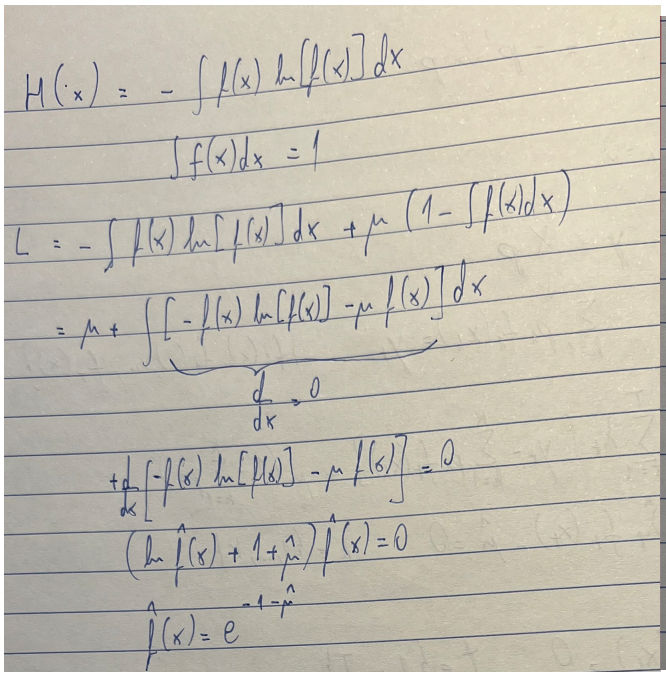


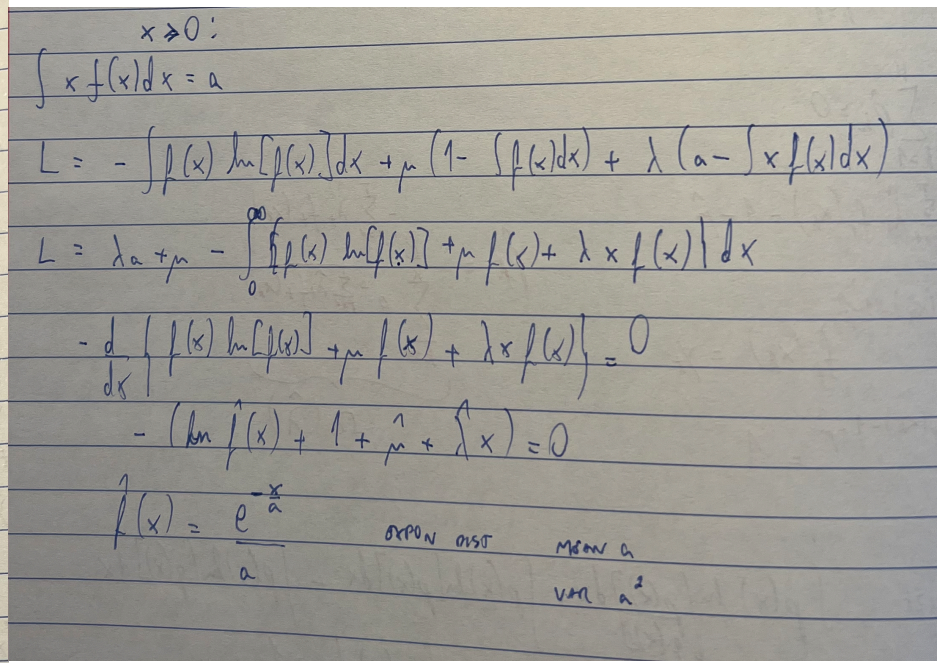
Lagrange multipliers





Lagrange multipliers





Missing data

- How can we estimation failure rates if we have never met one?
- Can we have enough flexibility in Bayesian priors (ie, not constrained by the properties of a beta distribution)?
- The probability p of success has a probability distribution - we want it to be the maximum entropy

Nassim's choice

- For MaxEnt q must be equal to 1/2:

```
Refine[CDF[BinomialDistribution[n, p], x], 0 ≤ x < n && x ∈ Integers]
BetaRegularized[1 - p, n - x, 1 + x]
f[p_] := BetaRegularized[1 - p, n - x, 1 + x];
InverseFunction[f][q]
```

InverseFunction: Inverse functions are being used. Values may be lost for multivalued inverses.

```
1 - InverseBetaRegularized[q, n - x, 1 + x]
InverseBetaRegularized[1/2, n - x, 1 + x] // TraditionalForm
```

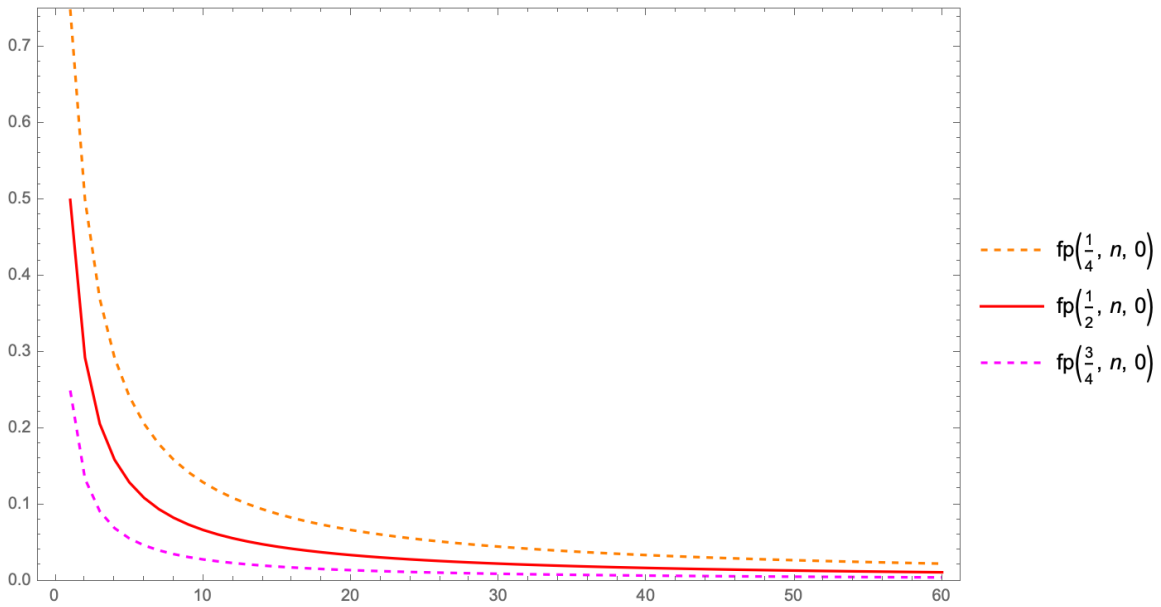
raditionalForm=

$$I^{-1}_{\frac{1}{2}}(n-x, x+1)$$

```
fp[α_, n_, x_] := 1 - InverseBetaRegularized[α, n - x, 1 + x];
```

Plot

- Bands for different values of q:

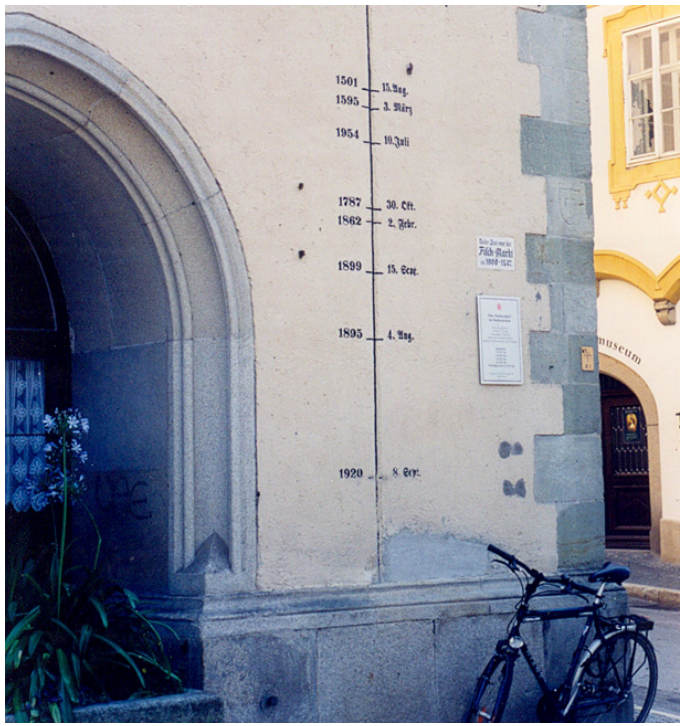


Other possibilities

- Bayesian methods with grids (non-parametric)?

High water marks

- If you want to build a house, you don't ask what's the average level of the river (or the variance, etc.)
- You look for high water marks and for how long the region has been inhabited



Geman, Geman and Taleb (2014)

- VaR Constraints
- Add one or some of the following constraints:
  - Global Mean
  - Absolute Mean
  - Power Laws for the tail
- Multi-Period?

Other possibilities

- Price of OTM options
- Multi-period (the Joseph problem)

Blind corners

- Two words no one likes to read: “Previously uncorrelated”
- Bring in the insurers!
  - NFL games cancellation
- 60-40 portfolio in 2022

To do

- Work through basic distributions with the 2 methods (Exponential, Gaussian, StudentT)
- Look more closely at Geman, Geman and Taleb (Var or CVaR?)
- Think about time series (Is the data censored? Is regime switching good?)
- The doctor problem: Can non-parametrical Bayesian methods work?
- Mutual Information
- Cross entropy
- Multivariate

What this talk was about anyway?

An introduction to maximum entropy methods

- Understand entropy and use it in practice
- Understand properties of distributions and which maximum entropy constraint they relate to
- Understand the value of the absence of a failure

Books, papers, posts

- "Maximum Ignorance Probability, with application to surgery's error rates" (Taleb's blog)
- Geman, Geman and Taleb (2014), "Tail Risk Constraints and Maximum Entropy" arXiv:1412.7647v1
- Golan, Judge and Miller (1996), "Maximum Entropy Econometrics: Robust Estimation with Limited Data"
- Keith Conrad: "Probability Distributions and Maximum Entropy" `https://kconrad.math.uconn.edu/blurbs/analysis/entropypost.pdf`