Capacity and Maximum Entropy Distributions

Polynomial Capacity: Theory, Applications, Generalizations

Jonathan Leake

Technische Universität Berlin

January 28th, 2020

Outline

1. Motivation
   - Optimization and Rounding
   - Inferring distributions
   - Examples and applications
2. Entropy and capacity
   - Probability notation
   - Entropy and relative entropy
   - Relative entropy and capacity
3. Computational questions for “max entropy” distributions
   - Computing max entropy distributions
   - Sampling max entropy distributions

Outline

1. Motivation
   - Optimization and Rounding
   - Inferring distributions
   - Examples and applications

2. Entropy and capacity
   - Probability notation
   - Entropy and relative entropy
   - Relative entropy and capacity

3. Computational questions for "max entropy" distributions
   - Computing max entropy distributions
   - Sampling max entropy distributions

Rounding

Goal: Optimize a function over some discrete / non-convex set.

Problem: Discrete / non-convex optimization can be hard.

E.g.: Consider trying to minimize a convex function over the lattice points of some convex polytope (e.g., the hypercube).

E.g.: Consider trying to minimize a convex function over the boundary points of the unit sphere, or over the rank-one matrices contained in the boundary of the PSD cone.

One strategy:

1. Optimize over the convex hull via convex optimization.
2. "Round" the optimal point to some point in the original set.
3. Hope / prove that the resulting point is close to of the optimum.

Next question: How should we "round"?

Inferring distributions

Goal: Given a sequence of data points from some unknown distribution on a known support set in $\mathbb{R}^n$, infer the underlying distribution.

Problem: Many possibilities for the distribution...

Sanity check: Probably the mean of the data points should be equal to the expectation of the distribution.

One possible solution: Choose the distribution which assumes no "extra information" beyond the mean of the data points.

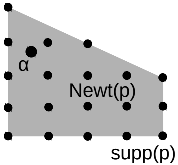
Next question: What does "extra information" mean?

Answer: Entropy maximizing distribution.

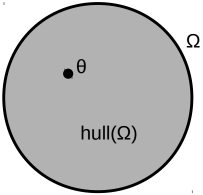
Side note: Sampling from such a distribution is a way to "round" the expectation (a point in the convex hull) to a point in the support set.

Pictures / examples

Discrete example: For a polynomial p, try to optimize over its support.



Continuous example: Try to infer a distribution on the unit circle with expectation θ.



Motivating applications

Some applications / connections without description:

- Isotropic constant [Klartag '06]
- Matrix Bingham distributions [Khatri-Mardia '77]
- Interior point methods [Bubeck-Eldan '15]
- Barycentric quantum entropy [Slater '99]

Quantum entropy: Continuous distribution over all pure states (rank-1 projections) which optimizes

$$\mathcal{H}_q(A) = \inf_{\mathbb{E}[\mu]=A} \int \mu(X) \log \mu(X) d\nu(X)$$

where A is a density matrix (PSD, trace $= 1$) and ν is Haar measure.

Entropic barrier function of a convex body $K \subset \mathbb{R}^n$:

$$B_K(\boldsymbol{v}) = \sup_{\boldsymbol{y} \in \mathbb{R}^n} \left[\langle \boldsymbol{y}, \boldsymbol{v} \rangle - \log \int_K e^{\langle \boldsymbol{y}, \boldsymbol{x} \rangle} d\boldsymbol{x} \right].$$

Question: How are these related? How is all this related to capacity?

Outline

Probability basics

For us, a **distribution** μ is a positive measure on some support set $\text{supp}(\mu) \subset \mathbb{R}^n$, such that the total measure of μ is finite.

Some notation:

- For each $S \subset \mathbb{R}^n$, define $\mu(S)$ to be the measure of the set S.
- We define the set $\text{hull}(\mu)$ to be the convex hull of $\text{supp}(\mu)$.
- A **probability distribution** has total measure is 1.
- The **expectation** is defined as usual: $\mathbb{E}[\mu] = \frac{\int x d\mu(x)}{\int d\mu(x)}$.
- The expectation is always a point in $\text{hull}(\mu)$.
- **Recall:** If μ is a discrete probability distribution on the degree vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ in the support of a polynomial p, then we can construct

$$p(\boldsymbol{x}) := \sum_{i=1}^{m} p_i \boldsymbol{x}^{\boldsymbol{v}_i} \quad \text{where} \quad p_i = \mu(\boldsymbol{v}_i),$$

  such that $\mathbb{E}[\mu] = \nabla p(\mathbf{1})$.

Entropy

Discrete definition: Given a discrete probability distribution μ, we define

$$\mathcal{H}(\mu) := -\sum_{x \in \mathrm{supp}(\mu)} \mu(x) \log \mu(x).$$

Some facts:

- $\mathcal{H}(\mu) > 0$ since $\mu(x) \in [0, 1]$.
- For a given support, maximized when μ is the uniform distribution. **Why?** Because $x \log x$ is convex.
- $\exp\left[-\mathcal{H}(\mu)\right] = \prod_{x \in \mathrm{supp}(\mu)} \mu(x)^{\mu(x)} \implies$ capacity?

Problem: What about continuous definition?

$$\mathcal{H}(\mu) := -\int_{\mathrm{supp}(\mu)} \mu(x) \log \mu(x) dx.$$

For continuous, $\mu(x) = 0$ for all x. $\implies$ Not a good definition.

Fix? Entropy of a density function with respect to some base measure.

Relative entropy (Kullback-Leibler divergence)

Discrete definition: Given a discrete probability distribution μ, and a base measure ν such that $\mathrm{supp}(\mu) \subseteq \mathrm{supp}(\nu)$, we define

$$\mathrm{D}_{\mathrm{KL}}(\mu \| \nu) := \sum_{x \in \mathrm{supp}(\nu)} \left[\frac{\mu(x)}{\nu(x)} \log \frac{\mu(x)}{\nu(x)} \right] \nu(x).$$

Continuous (general) definition: Given a base measure ν and a probability density function (pdf) ϕ, we can construct a probability measure via $\mu := \phi \cdot \nu \implies \int f(x) d\mu(x) = \int f(x)\phi(x) d\nu(x)$. Then:

$$\mathrm{D}_{\mathrm{KL}}(\mu \| \nu) := \int \phi(x) \log \phi(x) d\nu(x).$$

Some facts:

- $\mathrm{D}_{\mathrm{KL}}(\mu \| \nu) \geq 0$. (This is less clear, since $\phi \not\leq 1$.)
- For a given ν, D_{KL} is **minimized** when $\phi \equiv 1$.

Another name for μ of the above form is **absolutely continuous** w.r.t. ν.

What is relative entropy?

Last slide: $\mathrm{D_{KL}}(\mu \| \nu) := \int \phi(x) \log \phi(x) d\nu(x)$ for $\mu = \phi \cdot \nu$.

Let's consider the discrete case on support set $\mathcal{S}$, with the base measure being $\nu(x) = 1$ for all x (unnormalized uniform distribution). Then:

$$\mathrm{D_{KL}}(\mu \| \nu) = \sum_x \left[\frac{\mu(x)}{\nu(x)} \log \frac{\mu(x)}{\nu(x)} \right] \nu(x) = \sum_x \mu(x) \log \mu(x) = -\mathcal{H}(\mu).$$

Some thoughts:

- Entropy = negative relative entropy w.r.t. uniform measure.
- Generalizes entropy to continuous / other discrete base measures.
- In particular, max entropy = min relative entropy.
- $\mathrm{D_{KL}}$ is a measure of "closeness" of two distributions.

But: What does this have to do with capacity?

Relative entropy and capacity

Recall capacity: For $p(\boldsymbol{x}) = \sum_{\boldsymbol{v} \in \mathcal{S}} \nu(\boldsymbol{v}) \boldsymbol{x}^{\boldsymbol{v}}$ for some distribution ν:

$$\mathrm{Cap}_{\boldsymbol{\alpha}}(p) = \inf_{\boldsymbol{x} > 0} \frac{p(\boldsymbol{x})}{\boldsymbol{x}^{\boldsymbol{\alpha}}} = \exp \inf_{\boldsymbol{y} \in \mathbb{R}^n} \Big[\log p(e^{\boldsymbol{y}}) - \langle \boldsymbol{y}, \boldsymbol{\alpha} \rangle \Big].$$

This is essentially the **convex** Fenchel dual of $\log p(e^{\boldsymbol{y}})$.

This is the Lagrangian dual of a **minimum relative entropy program:**

$$\inf_{\substack{\mathrm{supp}(\mu) \subset \mathrm{supp}(\nu) \\ \mathbb{E}[\mu] = \boldsymbol{\alpha}}} \mathrm{D}_{\mathrm{KL}}(\mu \| \nu).$$

Further: Strong duality holds, which means the optimal values are equal:

$$-\log \mathrm{Cap}_{\boldsymbol{\alpha}}(p) = \inf_{\substack{\mathrm{supp}(\mu) \subset \mathrm{supp}(\nu) \\ \mathbb{E}[\mu] = \boldsymbol{\alpha}}} \mathrm{D}_{\mathrm{KL}}(\mu \| \nu).$$

Proof of all this: Not enlightening. This is essentially a folklore result.

Relative entropy and capacity

Last slide: For $p(\boldsymbol{x}) = \sum_{\boldsymbol{v} \in \mathcal{S}} \nu(\boldsymbol{v}) \boldsymbol{x}^{\boldsymbol{v}}$, we have

$$-\log \mathrm{Cap}_{\boldsymbol{\alpha}}(p) = \inf_{\substack{\mathrm{supp}(\mu) \subset \mathrm{supp}(\nu) \\ \mathbb{E}[\mu] = \boldsymbol{\alpha}}} \mathrm{D}_{\mathrm{KL}}(\mu \| \nu),$$

Further from strong duality: The $\boldsymbol{x} = e^{\boldsymbol{y}}$ which optimizes capacity also optimizes the entropy. **What does this mean?**

$$\mu_{\mathrm{opt}}(\boldsymbol{v}) = \boldsymbol{x}^{\boldsymbol{v}} \nu(\boldsymbol{v}) = e^{\langle \boldsymbol{y}, \boldsymbol{v} \rangle} \nu(\boldsymbol{v}).$$

That is: Min. rel. entropy distributions are always log-linear scalings of ν.

In particular: We have some facts about capacity we proved before.

- Up to $-\log$, the capacity of a polynomial is the entropy of a discrete distribution with support $\mathcal{S}$ and expectation $\boldsymbol{\alpha}$.
- Automatic: $\mathbb{E}[\nu] = \boldsymbol{\alpha}$ iff $-\log \mathrm{Cap}_{\boldsymbol{\alpha}}(p) = 0$ iff $\mathrm{Cap}_{\boldsymbol{\alpha}}(p) = 1$.
- Automatic: $-\log \mathrm{Cap}_{\boldsymbol{\alpha}}(p) \geq 0 \implies \mathrm{Cap}_{\boldsymbol{\alpha}}(p) \in [0, 1]$.
- Automatic: $\boldsymbol{\alpha} \in \mathrm{Newt}(p)$ iff $-\log \mathrm{Cap}_{\boldsymbol{\alpha}}(p)$ is finite iff $\mathrm{Cap}_{\boldsymbol{\alpha}}(p) > 0$.

Capacity for continuous distributions

Let's use the relative entropy framework to generalize capacity. Given a measure ν on a support set $\text{supp}(\nu) \subset \mathbb{R}^n$ and some $\theta \in \text{hull}(\nu)$, consider:

$$\inf_{\substack{\mu=\phi\cdot\nu \\ \mathbb{E}[\mu]=\theta}} \mathrm{D}_{\mathrm{KL}}(\mu\|\nu).$$

The Lagrangian dual of this is the same, with strong duality:

$$\inf_{\substack{\mu=\phi\cdot\nu \\ \mathbb{E}[\mu]=\theta}} \mathrm{D}_{\mathrm{KL}}(\mu\|\nu) = -\inf_{\boldsymbol{y}\in\mathbb{R}^n} \left[\log \int e^{\langle \boldsymbol{y},\boldsymbol{v}\rangle} d\nu(\boldsymbol{v}) - \langle \boldsymbol{y}, \boldsymbol{\theta}\rangle\right].$$

From strong duality: We similarly obtain $\phi_{\text{opt}}(\boldsymbol{v}) = e^{\langle \boldsymbol{y}_{\text{opt}},\boldsymbol{v}\rangle}$.

Note: When $\nu \sim p$ is discrete, we have $\log \int e^{\langle \boldsymbol{y},\boldsymbol{x}\rangle} d\nu(x) = \log p(e^{\boldsymbol{y}})$.

This gives an obvious definition for **continuous capacity**:

$$\mathrm{Cap}_{\boldsymbol{\theta}}(\nu) := \inf_{\boldsymbol{y}\in\mathbb{R}^n} \frac{\int e^{\langle \boldsymbol{y},\boldsymbol{v}\rangle} d\nu(\boldsymbol{v})}{e^{\langle \boldsymbol{y},\boldsymbol{\theta}\rangle}} \quad \Longrightarrow \quad \text{log-}\textbf{convex}\text{ program.}$$

Again, when $\nu \sim p$ this is the usual capacity.

Outline

1. Motivation
   - Optimization and Rounding
   - Inferring distributions
   - Examples and applications
2. Entropy and capacity
   - Probability notation
   - Entropy and relative entropy
   - Relative entropy and capacity
3. Computational questions for “max entropy” distributions
   - Computing max entropy distributions
   - Sampling max entropy distributions

Computing max entropy distributions

Recall: Given measure ν and $\boldsymbol{\theta} \in \text{hull}(\nu) \subset \mathbb{R}^n$, we want to optimize:

$$\inf_{\substack{\mu=\phi\cdot\nu \\ \mathbb{E}[\mu]=\boldsymbol{\theta}}} \mathsf{D}_{\mathsf{KL}}(\mu\|\nu) = \inf_{\substack{\mu=\phi\cdot\nu \\ \mathbb{E}[\mu]=\boldsymbol{\theta}}} \int \phi(\boldsymbol{x}) \log \phi(\boldsymbol{x}) d\nu(\boldsymbol{x}).$$

Problem: The domain is potentially infinite dimensional. Even in the discrete case, the domain is exponentially dimensional in the dimension.

Solution: Solve capacity formulation instead:

$$\inf_{\substack{\mu=\phi\cdot\nu \\ \mathbb{E}[\mu]=\theta}} \mathsf{D}_{\mathsf{KL}}(\mu\|\nu) = -\log \inf_{\boldsymbol{y}\in\mathbb{R}^n} \frac{\int e^{\langle \boldsymbol{y},\boldsymbol{v}\rangle} d\nu(\boldsymbol{v})}{e^{\langle \boldsymbol{y},\boldsymbol{\theta}\rangle}} \quad \text{and} \quad \phi_{\text{opt}}(\boldsymbol{v}) = e^{\langle \boldsymbol{y}_{\text{opt}},\boldsymbol{v}\rangle}.$$

That is: Computing optimal $\boldsymbol{y}$ for capacity, gives optimal density ϕ.

This is now an n-dimensional convex optimization problem, and so we can use ellipsoid method to approximate $\boldsymbol{y}_{\text{opt}}$.

[Singh-Vishnoi '15]: Discrete case. **[L-Vishnoi '20]:** Continuous case.

Computing continuous max entropy distributions

Last slide: We can use ellipsoid method to optimize

$$-\inf_{\boldsymbol{y}\in\mathbb{R}^n}\left[\log\int e^{\langle \boldsymbol{y},\boldsymbol{v}\rangle}\,d\nu(\boldsymbol{v}) - \langle \boldsymbol{y},\boldsymbol{\theta}\rangle\right].$$

To use this, we need oracle access to the objective and its gradient.

Discrete case: $\log\int e^{\langle \boldsymbol{y},\boldsymbol{v}\rangle}\,d\nu(\boldsymbol{v}) = \log p(e^{\boldsymbol{y}}) \implies$ oracle for p and ∇p.

E.g.: Spanning tree polynomials, and other "self-reducible" classes.

Continuous case: When can we compute $\log\int e^{\langle \boldsymbol{y},\boldsymbol{v}\rangle}\,d\nu(\boldsymbol{v})$?

E.g.: Let $\operatorname{supp}(\nu)$ be an orbit of Hermitian H under conjugation by $U(n)$, with ν the Haar measure induced by $U(n)$. **HCIZ formula:**

$$\int e^{\langle Y,UHU^*\rangle}\,dU = \prod_{k=1}^{n-1} k!\cdot\frac{\det(\exp[\lambda_i(Y)\cdot\lambda_j(H)]_{i,j=1}^n)}{\prod_{i<j}[\lambda_i(Y)-\lambda_j(Y)]\cdot[\lambda_i(H)-\lambda_j(H)]}.$$

Since $\int e^{\langle Y,UHU^*\rangle}\,dU = \int e^{\langle Y,X\rangle}\,d\nu(X)$, this is exactly what we want. Up to details, this includes the rank-1 projections case ($H = \operatorname{diag}(1,0,\ldots,0)$).

Sampling max entropy distributions

Want: To sample from $e^{\langle \boldsymbol{y}, \boldsymbol{v} \rangle} d\nu(\boldsymbol{v})$, given a particular $\boldsymbol{y}$.

Discrete case: Approximate counting $\iff$ approximate sampling [Jerrum-Valiant-Vazirani '86]. (Sampling for free.)

[Singh-Vishnoi '15]: Approx. counting $\iff$ max-entropy computation.

(I think all this requires "self-reducibility" structure.)

What about the continuous case? Let's try rank-1 projections.

- It is not clear how to sample from a manifold according to this density.
- **Fun fact:** $\text{diag}(\boldsymbol{v}\boldsymbol{v}^*) = (|v_1|^2, \ldots, |v_n|^2) \in \Delta_n$. That is, diag maps Hermitian rank-1 projections onto the standard simplex.
- **More fun fact:** The pushforward of the Haar measure is Lebesgue measure restricted to Δ_n. (Duistermaat-Heckman, for example).
- **Super fun fact:** Pushforward of $e^{\langle D, X \rangle} d\nu(X)$ is $e^{\langle \text{diag}(D), \boldsymbol{x} \rangle} dx$.
- **That is:** Correspondence of max-entropy distributions via diag.

Corollary: Sample Δ_n uniformly by sampling unit vectors uniformly.

Sampling in the rank-one case

Last slide: Pushforward of $e^{\langle D,X\rangle} d\nu(X)$ is $e^{\langle \mathrm{diag}(D),\boldsymbol{x}\rangle} dx \implies$ Correspondence of max-entropy distributions via diag.

Log-linear density on Δ_n: Standard machinery for sampling in this case (e.g. [Lovász-Vempala '06]), so we can sample from Δ_n.

Next question: How do we go back to rank-1 projections?

Let's compute the fiber of a given $\boldsymbol{x} \in \Delta_n$:

$$\mathrm{diag}^{-1}(\boldsymbol{x}) = \{\boldsymbol{v}\boldsymbol{v}^* : \mathrm{diag}(\boldsymbol{v}\boldsymbol{v}^*) = \boldsymbol{x}\} = \mathbb{T}^n \cdot \sqrt{\boldsymbol{x}}\sqrt{\boldsymbol{x}}^\top.$$

That is: $\boldsymbol{v} = (e^{i\theta_1}\sqrt{x_1}, \ldots, e^{i\theta_n}\sqrt{x_n})$. Since D is diagonal, we then have

$$e^{\langle D,\boldsymbol{v}\boldsymbol{v}^*\rangle} = e^{\langle \mathrm{diag}(e^{-i\boldsymbol{\theta}})D\,\mathrm{diag}(e^{i\boldsymbol{\theta}}),\sqrt{\boldsymbol{x}}\sqrt{\boldsymbol{x}}^\top\rangle} = e^{\langle D,\sqrt{\boldsymbol{x}}\sqrt{\boldsymbol{x}}^\top\rangle}$$

That is: The max entropy distribution is uniform on fibers.

Therefore: To construct a rank-1 projection sample from a simplex sample, we just need to uniformly sample from $\mathbb{T}^n$ and multiply by $\sqrt{\boldsymbol{x}}$.

Sampling in the case of other Hermitian orbits

Last two slides: How to sample for $\text{supp}(\nu) = U(n) \cdot \text{diag}(1, 0, \dots, 0)$.

Question: What about when Hermitian matrix H is more interesting?

First problem: Applying diag to general Hermitian orbits does **not** push forward the Haar measure to the uniform measure on a polytope.

However: Consider the more complex **Rayleigh map** $\mathcal{R}$, which maps Hermitian M to $(R_{i,j})_{i \leq j}$ where:

$$R_{i,j} := i^{\text{th}} \text{ largest eigenvalue of the leading principal } j \times j \text{ submatrix.}$$

Cauchy interlacing theorem: $R_{i,j+1} \geq R_{i,j} \geq R_{i+1,j+1}$ for valid i, j.

For fixed $R_{\bullet,n} = \text{eig}(H)$, these inequalities cut out the **Gelfand-Tsetlin polytope** associated to H, called $\text{GT}(H)$.

Ultra fun fact: The map $\mathcal{R}$ maps the $U(n)$ orbit of H onto $\text{GT}(H)$, and the pushforward of the Haar measure is Lebesgue on $\text{GT}(H)$, and of course also max entropy distributions.

Sampling and the GT polytope

Last slide: Rayleigh map pushes forward max entropy distributions to max entropy distributions on the GT polytope.

Like before: We can sample from the log-linear density on the GT polytope using standard techniques.

Going back to the Hermitian orbit: Similar type of argument, where fibers of certain "refined" Rayleigh maps are uniform.

Bonus: Simplex case fits into this more general case:

$$
\begin{array}{rccccccc}
R_{\bullet,n} := & 1 & & 0 & 0 & 0 & \cdots & 0 \\
R_{\bullet,n-1} := & & R_{1,n-1} & & 0 & 0 & \cdots & 0 \\
R_{\bullet,n-2} := & & & R_{1,n-2} & & 0 & \cdots & 0 \\
 & & & & \ddots & & &
\end{array}
$$

Successive differences $(1 - R_{1,n-1})$, $(R_{1,n-1} - R_{1,n-2}), \ldots, (R_{1,1})$ sum to 1 thus giving a point of Δ_n.