Lecture 1: Entropy, convexity, and matrix scaling
CSE 599S: Entropy optimality, Winter 2016
Instructor: James R. Lee **Last updated:** January 24, 2016

1 Entropy

Since this course is about entropy maximization, it makes sense to start by defining *entropy.* For quite a while, it will suffice to work over a finite (and non-empty) probability space Ω. A *probability mass function* $p : \Omega \to [0,1]$ is any function satisfying $\sum_{x\in\Omega} p(x) = 1$. One defines the *Shannon entropy of* p as the value

$$H(p) \stackrel{\text{def}}{=} -\sum_{x\in\Omega} p(x)\log p(x)\,.$$

(Note that we take $0\log 0 = \lim_{u\to 0} u\log u = 0$.)

As usual, we will abuse notation in the following way: If X is a random variable with law p, we will also use $H(X)$ to denote this value, and refer to it as the *entropy of* X. Here, and in the entire course, log denotes the natural logarithm. One often interprets $H(X)$ as representing the amount of "uncertainty" in the value of the random variable X.

The maximum entropy distribution on Ω. Denote by $\Delta_\Omega \subseteq \mathbb{R}^\Omega$ the set of all probability mass functions on Ω. One can consider the optimization:

$$\text{maximize} \quad \{H(p) : p \in \Delta_\Omega\}\,. \tag{1.1}$$

To see that this optimization has a unique solution, observe two facts:

1. The set Δ_Ω is a compact, convex set.
2. The function $p \mapsto H(p)$ is *strictly* concave on Δ_Ω. This is the rather intuitive property that entropy increases under averaging: For $p, q \in \Delta_\Omega$ and $\lambda \in [0,1]$,

$$H\left(\lambda p + (1-\lambda)q\right) \geqslant \lambda H(p) + (1-\lambda)H(q)\,.$$

   Moreover, if $p \neq q$, then the inequality is strict.

   A proof: The map $u \mapsto -u\log u$ is strictly concave on $[0,1]$; this follows from the fact that its derivative $-(1+\log u)$ is strictly decreasing on $[0,1]$. Now, a sum of concave functions is concave, so we conclude that H is concave. Moreover, if $p \neq q$, then they differ in some coordinate; strict concavity of the map $u \mapsto -u\log u$ applied to that coordinate yields strong concavity of H.

A strictly concave function on a convex set has at most one maximum. Since our set Δ_Ω is compact and H is continuous, H achieves a maximum on Δ_Ω, hence our optimization (1.1) has a unique solution.

Of course, the optimizer is readily apparent: It is given by the uniform distribution $\mu(x) = \frac{1}{|\Omega|}$ for all $x \in \Omega$. And we have $H(\mu) = \log|\Omega|$. To see that μ is indeed the optimizer of (1.1), consider

any $p \neq \mu$. There must exist $x, y \in \Omega$ such that $p(x) < p(y)$. For $t > 0$, define the function $p_t = p + t(\mathbf{1}_x - \mathbf{1}_y)$.[1] Note that $p_t \in \Delta_\Omega$ for $t \leqslant p(y)$.

But:

$$\frac{d}{dt} H(p_t)\Big|_{t=0} = \log \frac{p(y)}{p(x)} > 0\,,$$

so p was not a global maximizer.

This phenomenon of "checking for improvement in every allowable direction" is a necessary and sufficient condition for optimality more generally.

Exercise (1 point) 1.1. Suppose that $C \subseteq \mathbb{R}^n$ is a closed and convex set and $f : C \to \mathbb{R}$ is a continuously differentiable convex function on C. Then x^* is a global minimizer of f on C if and only if it holds that for every $x \in C$,

$$\langle \nabla f(x^*), x - x^* \rangle \geqslant 0\,.$$

Use this fact to justify (2.4) below.

1.1 Relative entropy

Given two probability mass functions $p, q \in \Delta_\Omega$, one defines the *relative entropy of p with respect to q* (also called the Kullback-Leibler divergence) by

$$D(p \,\|\, q) \stackrel{\text{def}}{=} \sum_{x \in \Omega} p(x) \log \frac{p(x)}{q(x)}\,.$$

If there is an $x \in \Omega$ such that $p(x) > q(x) = 0$, we set $D(p \,\|\, q) = \infty$.

This quantity is often thought of in the following context: q is a prior probability distribution (representing, say, the assumed state of the world), and p is the posterior distribution (after one has learned something by interacting with the world). In this case, $D(p \,\|\, q)$ represents the amount of information gained. Another operational definition: $D(p \,\|\, q)$ is the expected number of extra bits needed to encode a sample from p given a code that was optimized for q.

Hypothesis testing. I have always liked the hypothesis testing interpretation coming from Sanov's theorem. Suppose we are given n i.i.d. samples $x_1, x_2, \ldots, x_n \in \Omega$ all chosen according to p or all chosen according to q (we will generally represent the corresponding product measures by p^n and q^n, respectively).

Our goal is to design a $\{0, 1\}$-valued hypothesis tester T (this is actually a family of tests—one for every n) such that

1. We always accept the null hypothesis asymptotically: $q^n\left(T(x_1, \ldots, x_n) = 1\right) \to 1$ as $n \to \infty$.
2. And we make the false positive probability $\mathrm{err}(T, n) = p^n\left(T(x_1, \ldots, x_n) = 1\right)$ as small as possible.

Define $\alpha^*(T) = \liminf_{n\to\infty} \frac{-\log \mathrm{err}(T,n)}{n}$. The idea is that T accepts a p^n sample with probability like $e^{-\alpha^*(T)n}$ as $n \to \infty$ (if we ignore second-order terms in the exponent).

[1] We use $\mathbf{1}_z$ to denote the indicator function that takes value one at $z \in \Omega$ and 0 elsewhere.

Exercise (2 points) 1.2. For every p and q with $D(p \,\|\, q) < \infty$, it holds that

$$\sup_T \alpha^*(T) = D(p \,\|\, q)\,.$$

In other words, the relative entropy captures the optimal one-sided hypothesis testing error for testing i.i.d. samples from p vs. q.

[Hint: To prove this, one should use the maximum likelihood test that classifies a sample $x = (x_1, \ldots, x_n)$ based on whether $p^n(x) > q^n(x)$.]

1.2 Properties of the relative entropy

First, note that if μ is the uniform distribution on Ω, then for any $p \in \Delta_\Omega$,

$$D(p \,\|\, \mu) = \log|\Omega| - H(p)\,.$$

Thus in this case, one might think of $D(p \,\|\, \mu)$ as the "entropy deficit" (with respect to the uniform measure).

In general, $D(p \,\|\, q)$ is now a *strictly convex* function of p on Δ_Ω. This is verified just as we verified that the Shannon entropy is strictly concave. In particular, we can use this to conclude that the relative entropy is always non-negative: $D(p \,\|\, q) \geqslant 0$ for all $p, q \in \Delta_\Omega$, with equality if and only if $p = q$. Consider the optimization: $\min\{D(p \,\|\, q) : p \in \Delta_\Omega\}$. A strictly convex function takes a unique minimum value on Δ_Ω, and it is achieved when $p = q$, where $D(q \,\|\, q) = 0$. (Again, this can be verified by moving on the line between p and q when $p \neq q$.)

A stronger convexity property is true, though we will not need it until later.

Exercise (1 point) 1.3. Prove that the relative entropy is *jointly convex* in its arguments: For every four distributions $p, \hat{p}, q, \hat{q} \in \Delta_\Omega$ and every $\lambda \in [0,1]$, it holds that

$$D\left(\lambda p + (1-\lambda)\hat{p} \,\|\, \lambda q + (1-\lambda)\hat{q}\right) \leqslant \lambda D(p \,\|\, q) + (1-\lambda) D(\hat{p} \,\|\, \hat{q})\,. \tag{1.2}$$

There are a number of ways to prove this; you should do it using convex analysis, as follows. A basic fact: If one has a twice differentiable function $f : I \to \mathbb{R}$ for some interval $I \subseteq \mathbb{R}$, then f is convex on I if and only if its second derivative is non-negative on I. Generalizing this to n dimensions is a little more involved because there are many more directions in which one has to test convexity.

Consider a closed convex set $C \subseteq \mathbb{R}^n$ and $f : C \to \mathbb{R}$. If f has continuous second-order partial derivatives on C, then f is convex on C if and only if its *Hessian matrix* is positive semi-definite on the interior of C.

The Hessian $\mathsf{Hess}(f) : C \to \mathbb{R}^{n \times n}$ assigns to every point $x \in C$ the matrix given by

$$\mathsf{Hess}(f)(x_1, \ldots, x_n)_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f(x)\,.$$

In symbols, one can write $\mathsf{Hess}(f)_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$. It is the matrix of the second-order partial derivatives of f. Since we have assumed the second-order partial derivatives are continuous, $\mathsf{Hess}(f)$ is a symmetric matrix, therefore all its eigenvalues are real.

You should prove (1.2) in two steps: First reduce it to proving convexity of a two-variate function $g : [0,1]^2 \to \mathbb{R}$, and then prove that g is convex by computing $\mathsf{Hess}(g)$ and showing its eigenvalues are non-negative.

Note that the preceding exercise follows intuitively if you solve Exercise 1.2 first: It is clearly easier to hypothesis test whether the samples $(x_1, \ldots, x_n)$ come either from p-or-q or from p'-or-q' then if each sample could be from either pair (formally, a hypothesis tester for the harder problem yields a tester for the easier problem).

2 Matrix scaling

Now we come to our first application: The matrix scaling problem. We are given two $n \times n$ square matrices with non-negative entries $X, T \in \mathbb{R}_+^{n \times n}$. Our goal is to multiply the rows and columns of the input matrix $X = (x_{ij})$ by positive numbers so that the resulting matrix has the same row and column sums as the target matrix $T = (t_{ij})$. A particulary important case is when $t_{ij} = 1/n$ for all $i, j \in [n]$, in which case we are asking for a scaling of X that is doubly-stochastic.

Equivalently, we are trying to find non-negative diagonal matrices D_1, D_2 so that $D_1 X D_2$ and T have the same row and column sums. Let us call such a matrix a *Sinkhorn scaling of* X.

Theorem 2.1 (Sinkhorn scaling)**.** *If it holds that* $x_{ij} = 0 \iff t_{ij} = 0$ *for all* $i, j \in [n]$*, then such a Sinkhorn scaling exists.*

We will prove this using entropy maximization or, more precisely, relative entropy minimization. First, by a global rescaling, we may assume that $\sum_{i,j} t_{ij} = \sum_{i,j} x_{ij} = 1$.

Denote by $\mathcal{U} \subseteq \mathbb{R}^{n \times n}$ the convex set of all matrices satisfying for all $i, j \in [n]$ the constraints:

1. (non-negativity) $y_{ij} \geqslant 0$ and $y_{ij} = 0$ if $x_{ij} = 0$
2. (column sums equal) $\sum_i y_{ij} = \sum_i t_{ij}$
3. (row sums equal) $\sum_j y_{ij} = \sum_j t_{ij}$.

Note that since $y_{ij} = 0$ when $x_{ij} = 0$, such variables do not actually play a role in $\mathcal{U}$.

Condition (2), together with the fact that $\sum_{i,j} t_{ij} = 1$, implies that $\sum_{i,j} y_{ij} = 1$ for all $Y = (y_{ij}) \in \mathcal{U}$. In particular, we can think about the members of $\mathcal{U}$ as probability distributions on $[n] \times [n]$. This leads to our optimization problem:

$$\text{minimize} \quad \{D(Y \,\|\, X) : Y \in \mathcal{U}\}\,. \tag{2.1}$$

Note that we are minimizing the relative entropy over a closed, convex set of probability measures. It's also clearly the case that $\mathcal{U}$ is non-empty: It contains the target matrix T! Thus (2.1) has a unique optimal solution. What is perhaps more surprising (and we are in store for more such surprises) is that the optimal solution Y^* will be precisely of the form $D_1 X D_2$ for some non-negative diagonal matrices D_1 and D_2, yielding a proof of Theorem 2.1.

Remark 2.2. While Theorem 2.1 might seem restrictive, note that given an input matrix X and desired row and column sums $r_1, \ldots, r_n \geqslant 0$ and $c_1, \ldots, c_n \geqslant 0$, we can consider the polytope $\mathcal{U}$ where the constraints in (ii) and (iii) are replaced by $\sum_i y_{ij} = c_j$ and $\sum_j y_{ij} = r_i$, respectively. Our analysis will show that the optimal solution to (2.1) finds a Sinkhorn scaling of X with the prescribed row and column sums *whenever such a scaling exists.*

Suppose first that $r_1, \ldots, r_n, c_1, \ldots, c_n > 0$ are all strictly positive. Then if X admits a Sinkhorn scaling $Z = D_1 X D_2$ with the prescribed row and column sums, it admits one with D_1, D_2 strictly positive. In that case, we can simply take our target to be $T = Z$. Observe that our feasible region $\mathcal{U}$ does not depend on the target, only on the row/column sums.

If sum prescribed row/column sums are zero, then the constraints in (ii) and (iii) combined with the non-negativity constraints will force possibly other y_{ij} values equal to zero, and the same analysis applies. (The key fact we will need later for strong duality to hold is *strict feasibly* of $\mathcal{U}$ in the sense of Slater's condition, and the presence of identically zero y_{ij} variables does not affect this.)

2.1 The Lagrangian and optimality conditions

The set $\mathcal{U}$ defined above is not just convex; it is actually a polyhedron: A finite-dimensional set of points satisfying a set of linear inequalities. The theory of minimality for convex functions on polyhedra is very rich. One of the most useful techniques involves relaxing the hard constraints given by $\mathcal{U}$ to obtain an unconstrained optimization problem: This is the method of Lagrangian multipliers.

We introduce $2n$ unconstrained dual variables $\{\alpha_i, \beta_j : i, j \in [n]\} \subseteq \mathbb{R}$ and consider the function

$$\Lambda(y; \alpha, \beta) = \sum_{ij} y_{ij} \log \frac{y_{ij}}{x_{ij}} + \sum_i \alpha_i \sum_j (t_{ij} - y_{ij}) + \sum_j \beta_j \sum_i (t_{ij} - y_{ij}) \,.$$

Recalling that Y^* is the unique optimal solution to (2.1), observe that

$$D(Y^* \,\|\, X) = \min_{y \geqslant 0} \max_{\alpha, \beta} \Lambda(y; \alpha, \beta) \,,$$

since if we choose any $y \notin \mathcal{U}$, the inner maximum will be ∞, and for $y \in \mathcal{U}$, we have $\Lambda(y; \alpha, \beta) = \sum_{ij} y_{ij} \log \frac{y_{ij}}{x_{ij}}$.

When *strong duality* holds, we can actually reverse the max and min to obtain

$$D(Y^* \,\|\, X) = \max_{\alpha, \beta} \min_{y \geqslant 0} \Lambda(y; \alpha, \beta) \,. \tag{2.2}$$

Slater's condition tells us that if $\mathcal{U}$ contains a *strictly feasible* point, then strong duality holds. In our setting, this corresponds to a point of $\mathcal{U}$ for which all the inequality constraints $y_{ij} \geqslant 0$ are strict inequalities, and T provides such a point since whenever y_{ij} is an actual variable (and not merely the constant 0), we have $t_{ij} > 0$. Moreover, by our assumption that $x_{ij} = 0 \implies t_{ij} = 0 \implies y_{ij} = 0$, we see that for the dual optimization (2.2) is bounded below, implying that a primal-dual optimal solution (Y^*, α^*, β^*) exists.

For concreteness, at the end of this section, we state Slater's condition and strong duality in general terms.

Continuing, strong duality tells us that here exist values for the dual variables α^* and β^* such that

$$D(Y^* \,\|\, X) = \min_{y \geqslant 0} \Lambda(y; \alpha^*, \beta^*) \,. \tag{2.3}$$

Let us assume, for the moment, that $y^*_{ij} > 0$ whenever $t_{ij} > 0$. Later, we will show that this assumption is valid. In this case, the positive orthant $\mathbb{R}^{n \times n}_+$ contains a neighborhood of Y^*, and therefore by Exercise 1.1, we should have

$$\nabla_y \, \Lambda(Y^*; \alpha^*, \beta^*) = 0 \,. \tag{2.4}$$

Computing the derivative, this implies that for every $i, j \in [n]$ with $t_{ij} > 0$, we have

$$1 + \log y^*_{ij} - \alpha^*_i - \beta^*_j - \log x_{ij} = 0\,.$$

Rearranging yields

$$y^*_{ij} = x_{ij} e^{\alpha^*_i + \beta^*_j - 1}\,.$$

Thus we have obtained our goal: Y^* has the same row and column sums as T, and it is also obtained from X by multiplying the rows and columns by non-negative weights!

We are left to show that our assumption $y^*_{ij} > 0$ (whenever $t_{ij} > 0$) is true. This is a consequence of a slightly more general phenomenon.

Lemma 2.3 (Franklin-Lorenz)**.** *Consider $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, and assume that $At = b$ for some $t \in \mathbb{R}^n$ with $t > 0$ (pointwise). If $x \in \mathbb{R}^n$ satisfies $x > 0$ and $\sum_{i=1}^n x_i = 1$, then the optimization*

$$\textit{minimize} \quad \left\{ \sum_{i=1}^n y_i \log \frac{y_i}{x_i} : Ay = b, y \geqslant 0, \sum_{i=1}^n y_i = 1 \right\}$$

has a unique minimizer y^* with $y^* > 0$.

Proof. As we have already seen, the existence of a unique optimizer follows from strict convexity, continuity of the objective function, and the fact that the domain is compact and convex. The interesting content of the lemma is that $y^* > 0$ pointwise.

Suppose this is not the case and let $I = \{i : y^*_i = 0\}$. Consider for $\lambda \in [0, 1]$, the feasible point

$$y_\lambda = (1 - \lambda)y^* + \lambda t\,,$$

and let us calculate

$$\frac{d}{d\lambda} D(y_\lambda \,\|\, x) = \sum_{i=1}^n \left(1 + \log \frac{(1-\lambda)y^*_i + \lambda t_i}{x_i}\right)(t_i - y^*_i)\,.$$

For $i \notin I$, the corresponding term is bounded as $\lambda \to 0$, but for $i \in I$, each term goes to $-\infty$. Therefore every neighborhood of y^* contains a point with lower objective value, contradicting our assumption that y^* is minimal. □

Remark 2.4. One might wonder whether the condition $t_{ij} = 0 \implies x_{ij} = 0$ is really needed in Theorem 2.1. We used it only for the purpose of establishing strong duality via Slater's condition. But a simple example shows where we can go wrong without it:

$$T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad X = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

It is manifestly impossible to find a doubly-stochastic Sinkhorn scaling of X.

The principle of maximum entropy. This is our first experience with the "principle of maximum entropy." Roughly, this principle states that if we are given a set of constraints on a probability distribution, then the "best" (simplest?) distribution that fits the data will be the one of maximum entropy. The idea here is to avoid overfitting: To obtain a distribution that contains no artifacts beyond those implied by the constraints. A more refined version might say that we minimize the

relative entropy to a "background" measure so as to obtain a solution that is as simple as possible with respect to that measure (recall how we minimize $D(Y \,\|\, X)$ where X is our target).

And indeed, this is what happened: Our optimal solution was a simple perturbation of X (obtained by multiplying the rows and columns by positive weights). A more fleshed out version of this principle might state that the simplicity of our solution is related to the simplicity of the constraints; since the constraints involve only rows or columns, so did the relationship between Y^* and X. Still, the efficacy of this method remains a mystery to some extent, and it would be good to have a more principled explanation.

2.2 Slater's condition and strong duality

Theorem 2.5. *Suppose that for $i = 0, 1, 2, \ldots, m$, the function $f_i : \mathcal{D}_i \to \mathbb{R}$ is convex on its domain $\mathcal{D}_i \subseteq \mathbb{R}^n$. Let $\mathcal{D} = \bigcap_{i=0}^m \mathcal{D}_i$. Consider $A \in \mathbb{R}^{n \times k}$ and $b \in \mathbb{R}^k$, and the convex optimization problem:*

$$\begin{array}{lrl} \textit{minimize} & f_0(x) & \\ \textit{subject to} & f_i(x) \leqslant 0 & i = 1, 2, \ldots, m \\ & Ax = b & \\ & x \in \mathcal{D} & \end{array}$$

Slater's condition: *There exists a point $\hat{x} \in \operatorname{relint}(\mathcal{D})$ such that $f_i(\hat{x}) < 0$ for $i = 1, 2, \ldots, m$, and $A\hat{x} = b$.*

If Slater's condition holds, then the duality gap is zero. Moreover, if the dual value is finite, then it is attained.

Concretely, consider dual variables $\alpha \in \mathbb{R}^m_+$ and $\beta \in \mathbb{R}^k$, and the Lagrangian

$$\Lambda(x; \alpha, \beta) = f_0(x) + \sum_{i=1}^m \alpha_i f_i(x) + \langle \beta, Ax - b \rangle .$$

Under Slater's condition, we have

$$\min_{x \in \mathcal{D}} \max_{\alpha \geqslant 0, \beta} \Lambda(x; \alpha, \beta) = \max_{\alpha \geqslant 0, \beta} \min_{x \in \mathcal{D}} \Lambda(x; \alpha, \beta) . \tag{2.5}$$

Moreover, if (2.5) is finite, then there exists a triple (x^*, α^*, β^*) that achieves the optimum, i.e. an optimal primal-dual solution.

Remark 2.6. The *interior* of a set $\mathcal{D} \subseteq \mathbb{R}^n$ is the set of points in $\mathcal{D}$ that have an open neighborhood also contained in $\mathcal{D}$. The *relative interior* discussed above is a slightly more sophisticated concept defined by

$$\operatorname{relint}(\mathcal{D}) = \{x \in \mathcal{D} : \exists \varepsilon > 0, B(x, \varepsilon) \cap \operatorname{aff}(\mathcal{D}) \subseteq \mathcal{D}\} ,$$

where $B(x, \varepsilon)$ is the ball of radius ε around x in $\mathbb{R}^n$ and

$$\operatorname{aff}(\mathcal{D}) = \left\{ \sum_{i=1}^k c_i x_i : k > 0, x_i \in \mathcal{D}, c_i \in \mathbb{R}, \sum_{i=1}^k c_i = 1 \right\}$$

is the *affine hull* of $\mathcal{D}$.

The motivation for the relative interior comes from considering, say, a line segment in the plane. The interior of the segment is empty, while the relative interior consists of the entire segment except for its endpoints.