1 Overview

We present a complete derivation of maximum entropy. At a high level, we are assuming the first law of thermodynamics, conservation of energy, and we are trying to prove the second law of thermodynamics, entropy maximization, from information theory and vice versa. While trying to prove the second law of thermodynamics, we provide derivations of free energy from an assumption of information theory, specifically KL-divergence minimization relative to a Boltzmann distribution. We may include Boltzmann's derivation of entropy over discrete distributions.

Conversely, starting the first two laws of thermodynamics, we derive the form of KL-divergence over a Boltzmann distribution using Lagrange multipliers. This provides some mathematical insight into temperature and an interpretation of KL-divergence with a fixed temperature as the definition of the dual function. Furthermore, strong duality allows us to optimize the temperature / dual variables directly through the sum of exponents function, which relates to the normalization function.

Generalizing the problem to having multiple expected statistic constraints generalizes this derivation to multiple sufficient statistics, and the dual function defines the maximum entropy distribution.

Next, we note that KL-divergence can be defined as the Bregman divergence between the negative entropy function and the sum of exponents, but KL-divergence over any exponential family can also be defined through the normalization function or through its convex conjugate. This suggests that the convex conjugate of the normalization function should be related to the entropy of the exponential family distribution.

We note some differences, specifically between optimization over natural parameters of the exponential family compared to the optimization over probability (density) functions.

2 Information Theory and KL-divergence to maximum entropy and free energy

We start with a few thermodynamic assumptions. The first law of thermodynamics assumes that energy is conserved. Moreover, we have an intuitive notion that our non-equilibrium probability distributions will eventually converge to the equilibrium distribution, which happens to be the boltzmann distribution of the energy function.

Explicitly, we have some energy function $\varepsilon(x)$ for particles at the state x, and the particles have a probability density $q(x)$.

Then our first assumption is that average energy is the same for all distributions under consideration:

$$\int q(x)\varepsilon(x)dx = \text{constant} \tag{1}$$

A second assumption is that our distributions evolve over time towards the equilbrium distribution, and we happen to know that the equilibrium distribution is the Boltzmann distribution of the energy function, $p_{BM}(x|T) = \frac{1}{Z}e^{-\frac{\varepsilon(x)}{T}}$

$$
\begin{aligned}
q^{(t)}(x) &\to p_{BM}(x|T) \\
\lim_{t\to\infty} KL\left(q^{(t)}(x) \| p_{BM}(x|T)\right) &= 0
\end{aligned}
\tag{2}
$$

In physics we assume that the energy function is defined to be the Hamiltonian (or in quantum mechanics our expectations are actually a Hamiltonian operator acting on the distribution) and the distribution is updated through Hamiltonian dynamics. Arguably, in machine learning we omit this assumption and we are allowed to apply any optimization procedure to converge to the equilibrium distribution.

Without quite knowing what the Boltzmann distribution is, let us derive useful concepts and properties from this KL-divergence:

$$
\begin{aligned}
KL\left(q(x) \| p_{BM}(x|T)\right) &= \int q(x)\log q(x)dx - \int q(x)\log p(x|T)dx \\
&= -H\left(q(x)\right) - \int q(x)\left(-\log Z - \frac{1}{T}\varepsilon(x)\right)dx \\
&= -H\left(q(x)\right) + \log Z + \frac{1}{T}\int q(x)\varepsilon(x)dx \\
7 = -H\left(q(x)\right) + \log Z + \frac{1}{T}\bar{\varepsilon}
\end{aligned}
\tag{3}
$$

where $H\left(q(x)\right) = -\int q(x)\log q(x)dx$ is the entropy of q, and $\bar{\varepsilon}$ is the average energy, which is assumed to be constant.

Since KL is a divergence meassure, we know the useful property that it is non-negative an equal to 0 when the distributions match:

$$
\begin{aligned}
-H\left(q(x)\right) + \log Z + \frac{1}{T}\bar{\varepsilon} &\geq 0 \\
T\log Z &\geq TH\left(q(x)\right) - \bar{\varepsilon}
\end{aligned}
\tag{4}
$$

The right hand side is known as the Helmholtz free energy, and it is always lower than $T\log Z$, which is the Gibbs free energy. It is exactly equal when $q(x) = p_{bm}(x|T)$. On the computational side, it may be difficult to calculate this log partition function, so we may seek to use a much nicer distribution for which we can analyically calculate or estimate the entropy and the expected energy, and this concept of variational approximation is used in physics and variational inference.

When expected energy and temperature are fixed, since $\log Z$ of the Boltzmann distribution is fixed, we can minimize KL-divergence by maximizing entropy. And as we said before, the Bolzmann distribution happens to be the distribution which minimizes KL-divergence to itself and maximizes entropy relative to the expected energy and temperature constraints.

We may ask why is it important that we use KL-divergence to define convergence of distributions? We could conceivably use other definitions of convergence of distributions, such as convergence in probability, convergence in mean square, or convergence almost surely. Justifying this question is somewhat challenging, and perhaps somewhat circular - we are basically assuming entropy and information are important in order to justify KL-divergence.

3 Convex Optimization to KL-divergence

In the previous section, we concluded that $\min_{q \in \mathcal{Q}} KL\left(q(x) \| p_{BM}(x|T)\right)$ derives the optimization problem

$$\begin{aligned} \min_{q \in \mathcal{Q}} \quad & \int q(x) \log q(x) dx \\ \text{subject to} \quad & \int q(x) \varepsilon(x) dx = \text{constant} \end{aligned} \tag{5}$$

We now assume that entropy maximization is important, and we try to solve this constrained optimization problem using Lagrange multipliers.

$$\mathcal{L}\left(q, \lambda, \nu\right) = \int q(x) \log q(x) dx + \lambda \left(\int q(x) \varepsilon dx - c \right) + \nu \left(\int q(x) dx - 1 \right) \tag{6}$$

For given dual variables λ, ν, we define the dual function as an optimization over $\mathcal{Q}$ of the Lagrangian:

$$g(\lambda, \nu) := \min_{q \in \mathcal{Q}} \mathcal{L}\left(q, \lambda, \nu\right) \tag{7}$$

It turns out that when we have linear constraints and a convex function, we can solve this function through the convex conjugate.

Let $f : X \to R$ be a function. Then the convex conjugate $f^* : X^* \to R$ is defined as

$$f^*(x^*) = \sup_{x \in X} \left(x^*(x) - f(x)\right) \tag{8}$$

The common interpretation of the convex conjugate is that given lines of slope x^*, the convex conjugate is bias term that pushes the line to be the closest underestimate of the function. [1]

Rearranging the expression for the dual function:

$$\begin{aligned} & \min_{q \in \mathcal{Q}} \mathcal{L}(q, \lambda, \eta) \\ & \qquad = \min_{q \in \mathcal{Q}} f(q) + \lambda^\intercal \left(Aq - C\right) + \nu^\intercal \left(Bq - d\right) \\ & \qquad = \min_{q \in \mathcal{Q}} f(q) + \left(A^\intercal \lambda + B^\intercal \nu)\right)^\intercal q + \lambda^\intercal C + \nu^\intercal D \\ & \qquad = -\max_{q \in \mathcal{Q}} \left(\left(-A^\intercal \lambda - B^\intercal \nu\right) q - f(q)\right) \\ & \qquad = -f^*\left(-A^\intercal \lambda - B^\intercal \nu\right) \end{aligned} \tag{9}$$

[1] There is a more complicated geometric interpretation in terms of epigraphs and the so-called bias trick, where the epigraph dimension and the bias dimension switch roles in dual space.

For the negative entropy function, it turns out the convex conjugate is the sum of exponents:

$$\begin{aligned} f(q) &= \int q(x)\log q(x) \\ f^*(q^*) &= \max_{q\in Q}\left(q^*(q) - \int q(x)\log q(x)dx\right) \\ &= \max_{q\in Q}\left(\int q(x)(q^*(x) - \log q(x))dx\right) \end{aligned} \tag{10}$$

If we are unconstrained over our dual space of distributions, we can solve this using calculus (of variations):

$$\begin{aligned} &\frac{\partial}{\partial q(x')}\int q(x)\left(q^*(x) - \log q(x)\right)dx) \\ &= \frac{\partial}{\partial q(x')} q(x')\left(q^*(x') - \log q(x')\right) - q(x')\frac{\partial}{\partial q(x')}\log q(x') \\ &= q^*(x) - \log q(x') - 1 \end{aligned} \tag{11}$$

Note that we assume we apply any constraints about our distributions $q(x)$, such as that they normalize to 1 and have non-negative measures in the constraints. We solve for the optimal distribution by setting the partial derivatives equal to 0:

$$\begin{aligned} 0 &= q^*(x) - \log q_{opt}(x) - 1 \\ q^*(x) &= \log q_{opt}(x) + 1 \\ q_{opt}(x) &= e^{q^*(x)-1} \end{aligned} \tag{12}$$

This optimal distribution is the point where the linear approximation of the convex function is tight, that is

$$f^*(q^*) = \int q^*(x)q(x)dx - f(q) \tag{13}$$

For the negative function, we have an analytic solution for the convex conjugate:

$$\begin{aligned} f^*(q^*) &= \int q_{opt}(x)\left(q^*(x) - \log q_{opt}(x)\right)dx \\ f^*(q^*) &= \int q_{opt}(x) 1 dx \\ &= \frac{1}{e}\int e^{q^*(x)}dx \end{aligned} \tag{14}$$

At first glance, this may seem trivial, since probability distributions always integrate to 1. However, $q_{opt}(x)$ are unconstrained density functions and have no such requirements for normalization [2]. So we see that it is the normalization value for possibly unnormalized distributions.

With an analytic solution to the convex conjugate, we now have an analytic form for the dual function, $g(\lambda, \nu)$:

$$g(\lambda,\nu) = -\lambda^\intercal C - \nu^\intercal D - \int e^{-A(x)^\intercal \lambda - \nu^\intercal B - 1} \tag{15}$$

For our entropy maximization problem, we have one equality constraint for the energy and one equality constraint for the distributions to be normalizable:

$$\begin{aligned} \min_{q \in \mathcal{Q}} \quad & \int q(x)\log q(x)dx \\ \text{subject to} \quad & \int q(x)\varepsilon(x)dx = c \\ & \int q(x)dx = 1 \\ g(\lambda,\nu) &= -\lambda c - \nu - \int e^{-\lambda\varepsilon(x) - \nu - 1}dx \\ &= -\lambda c - \nu - e^{-\nu-1}\int e^{-\lambda\varepsilon(x)}dx \end{aligned} \tag{16}$$

Taking a step back, we notice that if we treated $\lambda = \frac{1}{T}$ and ν having some relation to the normalization of our q distributions, our Lagrangian is very similar to the KL-divergence with the Boltzmann distribution with temperature T.

$$\begin{aligned} \mathcal{L}(q,\lambda,\nu) &= -H(q) + \lambda\int q(x)\varepsilon(x)dx + \nu\int q(x)dx - \lambda c - \nu \\ &= KL\left(q(x)\|g(-\lambda)e^{-\lambda\varepsilon(x)}\right) - \log\int e^{-\lambda\varepsilon(x)}dx + \nu\int q(x)dx - \lambda c - \nu \end{aligned} \tag{17}$$

We see that our dual variables defines the natural parameters for an exponential family distribution with sufficient statistic $\varepsilon(x)$, and it can be interpreted as the inverse of the temperature. Namely, if we are to attempt to scale our quantities back to the expected energy, we would divide both the entropy and the free energy by λ, which would be equivalent to multiplying both by the temperature.

3.1 Maximum Entropy Distributions

Delving further into the interpretation of the dual distribution, we have

$$\begin{aligned} g(\lambda,\nu) &= -\lambda^\intercal C - \nu^\intercal D - \int q_{opt|\lambda,\nu}(x)dx \\ &\geq \int q(x)\left(q^*(-\lambda^\intercal C - \nu^\intercal D) - \log q(x)\right)dx \end{aligned} \tag{18}$$

[2] although by definition they must have positive densities

We have exact equality when $q(x) = e^{-A^\intercal\lambda - B^\intercal\nu - 1}$, and this is called Fermat's equality:

$$f^*(q^*) + f(q_{opt}) = q^*(q) \tag{19}$$

In this specific case, we have

$$H(q_{opt}(x)) = -\lambda^\intercal\bar{\varepsilon} - \nu - e^{-\nu-1}\int e^{-\varepsilon(x)^\intercal\lambda}dx \tag{20}$$

We note the similarity to Helmholtz free energy:

$$H(q_{opt}(x)) + \lambda^\intercal\bar{\varepsilon} \geq -\nu - e^{-\nu-1}\int e^{-\varepsilon(x)^\intercal\lambda}dx \tag{21}$$

¡++¿

This is a normalizable distribution when

$$\begin{aligned}\int e^{-\varepsilon(x)^\intercal\lambda}dx &= e^{\nu+1}\\ \nu + 1 &= \log\int e^{-\varepsilon(x)^\intercal\lambda}dx\end{aligned} \tag{22}$$

And we see that $\nu + 1$ defines the log normalizer of our Boltzmann distribution, so for a given λ, we are solving a minimization problem over (unnormalized distributions), with optimal solution given by the (unnormalized) Boltzmann distribution.

This means that our original problem, of KL minimization of a Boltzmann distribution, is actually solving the minimization problem defining that defines our dual function. Moreover, we can avoid having to solve this minimization problem by simply defining the Boltzmann distribution, and we can avoid solving for the entropy and lagrangian constraints since we already have an expression for the dual function $g(\lambda, \nu)$.

Moreover, the dual function is a linear offset of the normalization function / exponential error function, and instead of trying to optimize over all possible distributions relative to energy constraints, we can instead optimize the dual problem of optimizing exponential error functions over the much smaller space of natural parameters of the Boltzmann distributions / maximum entropy distributions.

Fermat's equality

Comparing the dual function and the Helmholtz free energy:

$$\begin{aligned}g(\lambda, \nu) &= -\lambda^\intercal C - \nu^\intercal D - e^{-\nu-1}\int e^{-\lambda^\intercal\varepsilon(x)}dx\\ &\geq \int q(x)\left(q^*(x) - \log q(x)\right)dx\\ &= -\int q(x)\left(\varepsilon^\intercal(x)\lambda + \nu + 1\right)dx - H(q(x))\\ &= H(q(x)) - \int q(x)\left(\nu + 1\right)dx - \lambda^\intercal\int\varepsilon(x)q(x)dx\end{aligned} \tag{23}$$

When constraints are satisfied (which admittedly are suboptimal solutions to the dual problem), we have

$$g(\lambda, \nu) \geq H(q(x)) - \lambda^{\intercal}\bar{\varepsilon} - (\nu + 1)\int q(x)dx \tag{24}$$

which is the Helmholtz free energy $-e^{\nu-1}\int q(x)dx$. So the Helmholtz free energy is an underestimate of the dual function, and it happens to be the negative of the normalization factor for the Boltzmann distribution with natural parameter λ.

$$\lambda\cdot\text{Helmholtz free energy}-(\nu+1)\int q(x)dx \leq -\lambda^{\intercal}\bar{\varepsilon}-\nu-e^{-\nu-1}\int \tilde{p}_{BM|\lambda}(x)dx \leq H(q(x)) \tag{25}$$

The benefit of the previous derivations for multiple linear constraints allows us to generalize from constraining a single expected energy to arbitrary linear constraints. In particular, we could constrain several statistics to have a certain expected value, and the change would result in a exponential family distribution over those statistics, with a separate natural parameter for each statistic.

We can include indicator functions over certain measurable sets, and constrain the expectation to 0, to constrain our distributions from having density outside of certain regions, which is a standard example of prior information.

4 Bregman Divergences and Exponential Families

We note that the convex conjugate of the negative entropy function is the sum of exponents or the normalization function for the maximum entropy exponential family. We further note that if we define the Bregman divergence over the entropy function, it defines KL-divergence. In fact, it generalizes to a divergence of unnormalized distributions.

It turns out that when the log normalizer of an exponential family is used to define a Bregman diverge, the result is the KL-divergence. Furthermore, the convex conjugate of the log normalizer also defines KL-divergence.

$$\begin{aligned}\frac{\partial}{\partial p(x')}&\int p(x)\log p(x)dx \\ &= \log p(x') + 1\end{aligned} \tag{26}$$

for $p(x) \neq 0$

Then if we define the Bregman divergence:

$$\begin{aligned}B_H\left(p(x)\|q(x)\right) &= \int p(x)\log p(x)dx - \int q(x)\log q(x)dx - \langle \log q(x) + 1, p(x) - q(x)\rangle \\ &= \int p(x)\log\frac{p(x)}{q(x)}dx + \int p(x)dx - \int q(x)dx\end{aligned} \tag{27}$$

For an exponential family:

$$\begin{aligned} G(\eta) &= \log \int h(x) e^{u(x)^\intercal \eta} dx \\ \nabla_\eta G(\eta) &= \int u(x) h(x) e^{u(x)^\intercal \eta - G(\eta)} dx \\ &= E_{x \sim p(x|\eta)} \left[u(x) \right] \end{aligned} \tag{28}$$

$$\begin{aligned} B_G\left(p(x|\eta), p(x|\eta')\right) &= \log \int h(x) e^{u(x)^\intercal \eta} dx - \log \int h(x) e^{u(x)^\intercal \eta'} dx - \left(\int p(x|\eta') u(x) \right)^\intercal (\eta - \eta') \\ &= G(\eta) - G(\eta') - \int h(x) u(x)^\intercal e^{u(x)^\intercal \eta' - G(\eta')} (\eta - \eta') \\ &= \log \frac{}{<++>} <++> \end{aligned} \tag{29}$$