Notes on maximum entropy

Daniel Hsu

February 17, 2025

1 Maximum entropy

- Problem setting
  - Suppose you want to model an unknown distribution over a (finite) set $\mathcal{X}$ (e.g., the set of all English words).
  - Let distribution q_0 be the "default model" you would pick absent any other information (e.g., q_0 = uniform over $\mathcal{X}$).
  - Then, you measure some "features" of the distribution.
    - You get the average (i.e., expected) values of n "feature functions"

$$T_i \colon \mathcal{X} \to \mathbb{R}, \quad i = 1, \dots, n$$

      where expectation is with respect to the unknown distribution.
      - For example:

$$\begin{aligned} T_1(x) &= \mathbb{1}\{x \text{ ends in a vowel}\} \\ T_2(x) &= \text{number of characters in } x \\ &\vdots \end{aligned}$$

    - Let b_i denote the average value of T_i.
      - Note: In typical applications, you won't have the actual average value of T_i, but rather just some noisy version of it. For example, we may have an i.i.d. sample $x^1, \dots, x^m$, and we obtain

$$b_i = \frac{1}{m} \sum_{j=1}^{m} T_i(x^j), \quad i = 1, \dots, n.$$

        Let us ignore this detail for now.
  - The default model (a.k.a. base measure) q_0 is not necessarily consistent with these measurements.
  - What model should you pick?

- **Maximum entropy (maxent) principle:** Choose model to be as close to q_0 as possible while being consistent with the measurements:

$$\begin{aligned} \min_{p \in \Delta} \quad & \mathrm{RE}(p, q_0) \\ \text{s.t.} \quad & \mathbb{E}_{X \sim p}[T_i(X)] = b_i, \quad i = 1, \ldots, n. \end{aligned}$$

  Here, minimization is over probability distributions Δ over $\mathcal{X}$, and $\mathrm{RE}(p, q)$ denotes the *relative entropy* (or *Kullback-Leibler (KL) divergence*) from p to q:

$$\mathrm{RE}(p, q) := \sum_{x \in \mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} = \mathbb{E}_{X \sim p}\left[\ln \frac{p(X)}{q(X)}\right].$$

  Note: Shannon used the term "relative entropy" for something else.

  - Why the name "maximum entropy"? If q_0 is the uniform distribution, then

$$\mathrm{RE}(p, q_0) = \sum_{x \in \mathcal{X}} p(x) \ln \frac{p(x)}{1/|\mathcal{X}|} = -\mathrm{H}(p) + \ln|\mathcal{X}|$$

    where

$$\mathrm{H}(p) := -\sum_{x \in \mathcal{X}} p(x) \ln p(x)$$

    is the *(Shannon) entropy* of p (switching from logarithm base-2 to natural logarithm). So minimizing $\mathrm{RE}(p, q_0)$ is the same as maximizing $\mathrm{H}(p)$.
  - Important property of $p \mapsto \mathrm{RE}(p, q_0)$: (strict) convexity.

- **Theorem:** If maxent problem is feasible, then (for almost all measurement values $b_1, \ldots, b_n$) solution has the following form:

$$p_\lambda(x) = \frac{1}{Z(\lambda)} \exp\left(\sum_{i=1}^n \lambda_i T_i(x)\right) q_0(x), \quad x \in \mathcal{X}$$

  for some $\lambda = (\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n$. Here, $Z(\lambda)$ is the normalization factor

$$Z(\lambda) := \sum_{x \in \mathcal{X}} \exp\left(\sum_{i=1}^n \lambda_i T_i(x)\right) q_0(x)$$

  that ensures p_λ is a valid probability distribution.

  - This parametric form for a probability distribution is called a *Gibbs distribution* or *Boltzmann distribution*. Also related to *exponential families*.
  - Example: If $T_1(x) = x$ ends in a vowel and $\lambda_1 = -2.10$, then a word that ends in a vowel is $\exp(-2.10) \approx 0.12$ as likely (according to p_λ) as one that does not.

- Some notation:

- Convenient to collect all T_i into a vector-valued function $T\colon \mathcal{X} \to \mathbb{R}^n$.
- Write $\lambda \cdot T$ for the function $x \mapsto \sum_{i=1}^n \lambda_i T_i(x)$.
- Also write $p[f] = \sum_{x \in \mathcal{X}} p(x) f(x) = \mathbb{E}_{X \sim p}[f(X)]$ for any distribution p on $\mathcal{X}$.
- Maxent problem is

$$\begin{aligned} \min_{p \in \Delta} \quad & \mathrm{RE}(p, q_0) \\ \text{s.t.} \quad & p[T] = b \end{aligned}$$

  where $b = (b_1, \ldots, b_n)$.
- A feasible maxent solution has the form

$$p_\lambda(x) = \frac{1}{Z(\lambda)} \exp(\lambda \cdot T(x)) q_0(x), \quad x \in \mathcal{X}$$

  with

$$Z(\lambda) := q_0[\exp(\lambda \cdot T)].$$

- Geometric picture
  - Constraints $p[T] = b$ define an affine subset of Δ

$$\mathcal{P} = \mathcal{P}(T, b) = \{p \in \Delta : p[T] = b\}.$$

  - Maxent = $p \in \mathcal{P}$ that minimizes $\mathrm{RE}(p, q_0)$: *information projection* of q_0 onto $\mathcal{P}$
  - The Gibbs distributions

$$\mathcal{Q} = \mathcal{Q}(T, q_0) = \{p_\lambda : \lambda \in \mathbb{R}^n\}$$

    form a nonlinear subset of Δ.
  - It turns out if $\mathcal{P} \neq \emptyset$, then $|\mathcal{P} \cap \overline{\mathcal{Q}}| = 1$, where $\overline{\mathcal{Q}}$ is the closure of $\mathcal{Q}$ (i.e., may need to consider sequences of λ's)

- Use *method of Lagrange multipliers* to deal with constraints $p[T] = b$.
  - Lagrange multipliers: $\lambda = (\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n$
  - Lagrangian function

$$\begin{aligned} \mathcal{L}(p, \lambda) &= \mathrm{RE}(p, q_0) - \lambda \cdot (p[T] - b) \\ &= \mathrm{RE}(p, q_0) - p[\lambda \cdot T] + \lambda \cdot b \end{aligned}$$

  - Maxent problem is equivalent to

$$\min_{p \in \Delta} \sup_{\lambda \in \mathbb{R}^n} \mathcal{L}(p, \lambda).$$

  - Properties of Lagrangian:
    - Lagrangian is convex in p and linear in λ.
    - Domain for p is convex and compact; domain for λ is convex.

* Therefore we can switch order of min and sup:

$$\min_{p\in\Delta}\sup_{\lambda\in\mathbb{R}^n}\mathcal{L}(p,\lambda)=\sup_{\lambda\in\mathbb{R}^n}\min_{p\in\Delta}\mathcal{L}(p,\lambda).$$

The function $\lambda\mapsto\min_{p\in\Delta}\mathcal{L}(p,\lambda)$ is called the *dual (objective) function.*

– **Donsker-Varadhan inequality:** For any function $f\colon\mathcal{X}\to\mathbb{R}$,

$$\mathrm{RE}(p,q)\geq p[f]-\ln q[\exp(f)].$$

(Special case of Fenchel-Young inequality from convex analysis.)

– Claim: For each fixed $\lambda\in\mathbb{R}^n$, function $p\mapsto\mathcal{L}(p,\lambda)$ is minimized by p_λ.

* For any $p\in\Delta$, Donsker-Varadhan (with $f(x)=\lambda\cdot T(x)$) implies

$$\begin{aligned}\mathcal{L}(p,\lambda)&=\mathrm{RE}(p,q_0)-p[\lambda\cdot T]+\lambda\cdot b\\&\geq p[\lambda\cdot T]-\ln q_0[\exp(\lambda\cdot T)]-p[\lambda\cdot T]+\lambda\cdot b\\&=-\ln Z(\lambda)+\lambda\cdot b.\end{aligned}$$

* For $p=p_\lambda$,

$$\begin{aligned}\mathcal{L}(p_\lambda,\lambda)&=\mathrm{RE}(p_\lambda,q_0)-p_\lambda[\lambda\cdot T]+\lambda\cdot b\\&=\sum_{x\in\mathcal{X}}p_\lambda(x)\ln\frac{\exp(\lambda\cdot T(x))}{Z(\lambda)}-p_\lambda[\lambda\cdot T]+\lambda\cdot b\\&=p_\lambda[\lambda\cdot T]-\ln Z(\lambda)-p_\lambda[\lambda\cdot T]-\lambda\cdot b\\&=-\ln Z(\lambda)+\lambda\cdot b.\end{aligned}$$

* So

$$\min_{p\in\Delta}\mathcal{L}(p,\lambda)=\mathcal{L}(p_\lambda,\lambda).$$

– So dual function is $\lambda\mapsto\mathcal{L}(p_\lambda,\lambda)$.

– If $\lambda^\star\in\mathbb{R}^n$ achieves

$$\mathcal{L}(p_{\lambda^\star},\lambda^\star)=\sup_{\lambda\in\mathbb{R}^n}\mathcal{L}(p_\lambda,\lambda)=\sup_{\lambda\in\mathbb{R}^n}-\ln Z(\lambda)+\lambda\cdot b.$$

then $p_{\lambda^\star}$ is maxent solution.

- Connection to **maximum likelihood principle** for Gibbs distributions $\{p_\lambda:\lambda\in\mathbb{R}^n\}$

– Suppose we obtain b as the empirical average of T over data set $x^1,\ldots,x^m$:

$$b=\frac{1}{m}\sum_{j=1}^{m}T(x^j).$$

– Consider the log-likelihood of p_λ where the data set is treated as an i.i.d. sample:

$$\begin{aligned}\ln \prod_{j=1}^{m} p_\lambda(x^j) = \sum_{j=1}^{m} \ln p_\lambda(x^j) &= \sum_{j=1}^{m} \ln\left(\frac{1}{Z(\lambda)} \exp(\lambda \cdot T(x^j)) q_0(x^j)\right) \\ &= -m \ln Z(\lambda) + \lambda \cdot \sum_{j=1}^{m} T(x^j) + \sum_{j=1}^{m} \ln q_0(x^j) \\ &= m(-\ln Z(\lambda) + \lambda \cdot b) + \sum_{j=1}^{m} \ln q_0(x^j).\end{aligned}$$

– Maximizing log-likelihood = maximizing dual function.

– Annoyingly, the log-likelihood does not always have a maximizer.

* $\mathcal{X} = \{\mathrm{a}, \mathrm{b}, \mathrm{c}\}$
* $q_0(\mathrm{a}) = q_0(\mathrm{b}) = q_0(\mathrm{c}) = 1/3$
* $T_1(x) = \mathbb{1}\{x = \mathrm{a}\}$, $T_2(x) = \mathbb{1}\{x = \mathrm{b}\}$, $T_3(x) = \mathbb{1}\{x = \mathrm{c}\}$
* $b = (0, 1/3, 2/3)$
* Log-likelihood

$$\lambda \mapsto -\ln\left(\frac{\exp(\lambda_1)}{3} + \frac{\exp(\lambda_2)}{3} + \frac{\exp(\lambda_3)}{3}\right) + \frac{\lambda_2}{3} + \frac{2\lambda_3}{3}$$

does not attain supremum at any λ

- Proof of Donsker-Varadhan inequality $\mathrm{RE}(p, q) \geq p[f] - \ln q[\exp(f)]$:

– If $\mathrm{RE}(p, q) = \infty$ (i.e., there exists x such that $q(x) = 0$ but $p(x) > 0$), then trivially true.

– So assume $p \ll q$ (i.e., no such x as above).

– Consider any $f \colon \mathcal{X} \to \mathbb{R}$, and observe that $\exp(f)$ is strictly positive function.

– Define q' to be the distribution proportional to $\exp(f)q$, i.e.,

$$q' := \frac{\exp(f)}{Z} q, \quad Z := q[\exp(f)].$$

– Then $q \ll q'$, and hence $p \ll q'$, so

$$\frac{p}{q'} = Z \exp(-f) \frac{p}{q} < \infty.$$

– Can write

$$\begin{aligned}\mathrm{RE}(p, q') = p\left[\ln \frac{p}{q'}\right] &= p\left[\ln Z - f + \ln \frac{p}{q}\right] \\ &= \ln Z - p[f] + p\left[\ln \frac{p}{q}\right] \\ &= \ln q[\exp(f)] - p[f] + \mathrm{RE}(p, q).\end{aligned}$$

– On the other hand, by Gibbs' inequality, $\mathrm{RE}(p, q') \geq 0$.

– Conclude

$$\ln q[\exp(f)] - p[f] + \mathrm{RE}(p, q) \geq 0.$$

2 Log partition function

- Consider Gibbs distributions $\mathcal{Q} = \mathcal{Q}(T, q_0) = \{p_\lambda : \lambda \in \mathbb{R}^n\}$ generated by $T\colon \mathcal{X} \to \mathbb{R}^n$ and $q_0 \in \Delta$:
$$p_\lambda(x) = \frac{1}{Z(\lambda)} \exp(\lambda \cdot T(x)) q_0(x), \quad x \in \mathcal{X}.$$
- "Normalization constant" (as function of λ)
$$Z(\lambda) = \sum_{x \in \mathcal{X}} \exp(\lambda \cdot T(x)) q_0(x)$$
is also called *partition function*.
  - May also interpret as *moment generation function* for random vector $Y := T(X)$ where $X \sim q_0$.
- Main object of interest: *log partition function* $G(\lambda) := \ln Z(\lambda)$.
  - G is convex: for any $\lambda^0, \lambda^1 \in \mathbb{R}^n$ and $\alpha \in (0,1)$,
$$\begin{aligned}
&G(\alpha\lambda^0 + (1-\alpha)\lambda^1) \\
&= \ln\left(\sum_{x \in \mathcal{X}} \exp\big((\alpha\lambda^0 + (1-\alpha)\lambda^1) \cdot T(x)\big) q_0(x)\right) \\
&= \ln\left(\sum_{x \in \mathcal{X}} \exp\big((\alpha\lambda^0) \cdot T(x)\big) \exp\big(((1-\alpha)\lambda^1) \cdot T(x)\big) q_0(x)\right) \\
&\le \ln\left(\left[\sum_{x \in \mathcal{X}} \exp\big((\alpha\lambda^0) \cdot T(x)\big)^{\frac{1}{\alpha}} q_0(x)\right]^{\alpha} \left[\sum_{x \in \mathcal{X}} \exp\big(((1-\alpha)\lambda^1) \cdot T(x)\big)^{\frac{1}{1-\alpha}} q_0(x)\right]^{1-\alpha}\right) \\
&= \alpha G(\lambda^0) + (1-\alpha) G(\lambda^1).
\end{aligned}$$
    Key step uses *Hölder's inequality*.
  - When is G strictly convex?
    - Key step based on Hölder's inequality holds with equality for $\lambda^0 \neq \lambda^1$ if and only if
$$\exp(\lambda^0 \cdot T) = c \exp(\lambda^1 \cdot T)$$
(as functions) for some constant $c > 0$ on the support of q_0.
    - This is equivalent to
$$(\lambda^0 - \lambda^1) \cdot T = \ln(c).$$
    - This means there is a non-trivial linear combination of $T_1, \ldots, T_n$ that results in a constant function.
    - Conclusion: G is strictly convex if and only if $T_1, \ldots, T_n$ are *affinely independent* (on the support of q_0).
      - Affine independence: If $\sum_{i=1}^n \lambda_i T_i$ is constant, then $\lambda_1 = \cdots = \lambda_n = 0$.

- The gradient $\nabla G\colon \mathbb{R}^n \to \mathbb{R}^n$ maps the λ parameter to the mean of T under p_λ:

$$\begin{aligned}\nabla G(\lambda) &= \frac{1}{Z(\lambda)} \sum_{x \in \mathcal{X}} \exp(\lambda \cdot T(x)) T(x) q_0(x) \\ &= p_\lambda[T].\end{aligned}$$

- If G is strictly convex, then ∇G is 1-to-1.
  - Consider any distinct $\lambda^0, \lambda^1 \in \mathbb{R}^n$, and let $\lambda(t) = (1-t)\lambda^0 + t\lambda^1$ for $t \in [0,1]$ specify the line segment between λ^0 and λ^1.
  - Strict convexity of G implies strict convexity of $G(\lambda(t))$.
  - Direct computation shows

$$\begin{aligned}\left\{\frac{\mathrm{d}}{\mathrm{d}t} G(\lambda(t))\right\}\Big|_{t=0} &= (\lambda^1 - \lambda^0) \cdot p_{\lambda^0}[T] = (\lambda^1 - \lambda^0) \cdot \nabla G(\lambda^0), \\ \left\{\frac{\mathrm{d}}{\mathrm{d}t} G(\lambda(t))\right\}\Big|_{t=1} &= (\lambda^1 - \lambda^0) \cdot p_{\lambda^1}[T] = (\lambda^1 - \lambda^0) \cdot \nabla G(\lambda^1).\end{aligned}$$

  - Strict convexity of $G(\lambda(t))$ implies $\frac{\mathrm{d}}{\mathrm{d}t} G(\lambda(t))$ is strictly increasing in t.
  - Hence

$$(\lambda^1 - \lambda^0) \cdot \nabla G(\lambda^0) \neq (\lambda^1 - \lambda^0) \cdot \nabla G(\lambda^1).$$

    Since $\lambda^1 - \lambda^0 \neq 0$, this implies

$$\nabla G(\lambda^0) \neq \nabla G(\lambda^1).$$

  - (In fact, converse statement is also true.)
- Let $\mathcal{M} := \{b : \exists p \ll q_0 \centerdot p[T] = b\}$ be the set of possible "T-means".
  - It can be shown that the image of ∇G

$$\{\nabla G(\lambda) : \lambda \in \mathbb{R}^n\}$$

    is equal to the interior $\mathcal{M}^\circ$ of $\mathcal{M}$ (i.e., all of $\mathcal{M}$ except the boundary points).
  - Proof: see Theorem 3.3 in Wainwright and Jordan's "Graphical Models ..." FnTML monograph
- Upshot: Two ways to parameterize the Gibbs distributions.
  - "Natural parameterization": $\lambda \in \mathbb{R}^n$.
  - "Mean parameterization": mean of T.
  - ∇G is the link between these parameter spaces (or interiors thereof).

- Similar for general exponential families (where q_0 may be a general σ-finite measure on a measure space).
  - Example: Poisson distribution $X \sim \mathrm{Poi}(\mu)$ for $\mu > 0$

* Probability mass function and mean:

$$\Pr(X = x) = \frac{\mu^x \exp(-\mu)}{x!}, \quad x \in \mathbb{N}_0$$
$$\mathbb{E}[X] = \mu.$$

* Let $q_0(x) = 1/x!$, $T(x) = x$, so

$$p_\lambda(x) = \exp(\lambda x - G(\lambda)) q_0(x)$$

where

$$G(\lambda) = \ln\left(\sum_{x=0}^{\infty} \exp(\lambda)^x \frac{1}{x!}\right) = \ln(\exp(\exp(\lambda))) = \exp(\lambda)$$
$$\frac{\mathrm{d}G}{\mathrm{d}\lambda}(\lambda) = \exp(\lambda) = p_\lambda[x].$$

* Link between natural parameter λ and "mean parameter" μ:

$$\mu = \exp(\lambda).$$

– Example: unit variance normal distribution $X \sim \mathrm{N}(\mu, 1)$ for $\mu \in \mathbb{R}$

* Probability density function and mean:

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right), \quad x \in \mathbb{R}$$
$$\mathbb{E}[X] = \mu.$$

* Let $q_0(x) = \exp(-x^2/2)/\sqrt{2\pi}$, $T(x) = x$, so

$$p_\lambda(x) = \exp(\lambda x - G(\lambda)) q_0(x)$$

where

$$G(\lambda) = \ln \int_{-\infty}^{\infty} \exp(\lambda x) \frac{\exp(-x^2/2)}{\sqrt{2\pi}} \,\mathrm{d}x = \ln \exp(\lambda^2/2) = \frac{\lambda^2}{2}$$
$$\frac{\mathrm{d}G}{\mathrm{d}\lambda}(\lambda) = \lambda.$$

* Link between natural parameter λ and "mean parameter" μ:

$$\mu = \lambda.$$

– Example: multinomial distribution $(X_1, \ldots, X_k) \sim \mathrm{Mult}(N; \pi_1, \ldots, \pi_k)$

* Probability mass function and mean:

$$\Pr((X_1, \ldots, X_k) = (x_1, \ldots, x_k)) = \binom{N}{x_1, \ldots, x_k} \prod_{i=1}^{k} \pi_i^{x_i}$$
$$\mathbb{E}[X_i] = N\pi_i.$$

∗ Let $q_0(x_1, \ldots, x_k) = \binom{N}{x_1, \ldots, x_k}$, $T_i(x_1, \ldots, x_k) = x_i$ for $i \in \{1, \ldots, k-1\}$, so

$$p_\lambda(x) = \exp\left(\sum_{i=1}^{k-1} \lambda_i x_i - G(\lambda)\right) q_0(x)$$

where

$$\begin{aligned} G(\lambda) &= \ln \sum_{x_1+\cdots+x_k=N} \binom{N}{x_1, \ldots, x_k} \exp(\lambda_1)^{x_1} \cdots \exp(\lambda_{k-1})^{x_{k-1}} 1^{x_k} \\ &= \ln(\exp(\lambda_1) + \cdots + \exp(\lambda_{k-1}) + 1)^N \\ &= N \ln(\exp(\lambda_1) + \cdots + \exp(\lambda_{k-1}) + 1) \\ \frac{\mathrm{d}G}{\mathrm{d}\lambda_i}(\lambda) &= N \frac{\exp(\lambda_i)}{\exp(\lambda_1) + \cdots + \exp(\lambda_{k-1}) + 1}, \quad i \in \{1, \ldots, k-1\}. \end{aligned}$$

∗ Link between natural parameters λ and "mean parameters":

$$N\pi_i = N \frac{\exp(\lambda_i)}{\exp(\lambda_1) + \cdots + \exp(\lambda_{k-1}) + 1}, \quad i \in \{1, \ldots, k-1\}.$$

∗ Note that if we had used k feature functions, $T_i(x_1, \ldots, x_k) = x_i$ for all $i \in \{1, \ldots, k\}$, then they would not be affinely independent: we would have $\sum_{i=1}^k T_i \equiv N$.

∗ Try to invert ∇G: given $b = (b_1, \ldots, b_{k-1}) \in \mathbb{R}_+^{k-1}$ (where $b_1 + \cdots + b_{k-1} \leq N$),

$$(\nabla G)^{-1}(b)_i = \ln \frac{b_i}{N - (b_1 + \cdots + b_{k-1})}.$$

This fails at the "boundary" where some $b_i \in \{0, N\}$, but works everywhere else.

3 Information geometry

- Solution $p^\star$ to maxent problem

$$\begin{aligned} \min_{p \in \Delta} \quad & \mathrm{RE}(p, q_0) \\ \text{s.t.} \quad & p[T] = b \end{aligned}$$

is *information projection* of base measure q_0 onto $\mathcal{P}(T, b) = \{p \in \Delta : p[T] = b\}$.

- In fact, for any other $p \in \mathcal{P}(T, b)$, we have

$$\mathrm{RE}(p, q_0) = \mathrm{RE}(p, p^\star) + \mathrm{RE}(p^\star, q_0).$$

This is the *Pythagorean theorem* for relative entropy.

– Proof is especially simple when $p^\star = p_\lambda$ for some $\lambda \in \mathbb{R}^n$:

$$\begin{aligned}
\mathrm{RE}(p, q_0) - \mathrm{RE}(p_\lambda, q_0) &= \mathrm{RE}(p, q_0) - p_\lambda\left[\ln \frac{p_\lambda}{q_0}\right] \\
&= \mathrm{RE}(p, q_0) - p_\lambda[\lambda \cdot T - \ln Z(\lambda)] \\
&= \mathrm{RE}(p, q_0) - p[\lambda \cdot T - \ln Z(\lambda)] \\
&= \mathrm{RE}(p, q_0) - p\left[\ln \frac{p_\lambda}{q_0}\right] \\
&= p\left[\ln \frac{p}{q_0} - \ln \frac{p_\lambda}{q_0}\right] \\
&= p\left[\ln \frac{p}{p_\lambda}\right] \\
&= \mathrm{RE}(p, p_\lambda).
\end{aligned}$$

– General version: For any $\mathcal{P} \subseteq \Delta$ closed and convex, any $q_0 \in \Delta$, the information projection $p^\star := \arg\min_{p \in \mathcal{P}} \mathrm{RE}(p, q_0)$ of q_0 onto $\mathcal{P}$ satisfies

$$\mathrm{RE}(p, q_0) \geq \mathrm{RE}(p, p^\star) + \mathrm{RE}(p^\star, q_0), \quad p \in \mathcal{P},$$

with equality if $\mathcal{P}$ is an affine set.

- Relative entropy of a Gibbs distribution p_{λ^0} from another p_λ:

$$\begin{aligned}
\mathrm{RE}(p_{\lambda^0}, p_\lambda) &= p_{\lambda^0}\left[\ln \frac{p_{\lambda^0}}{p_\lambda}\right] \\
&= p_{\lambda^0}\left[\ln \frac{\exp(\lambda^0 \cdot T - G(\lambda^0))}{\exp(\lambda \cdot T - G(\lambda))}\right] \\
&= p_{\lambda^0}\left[(\lambda^0 - \lambda) \cdot T - G(\lambda^0) + G(\lambda)\right] \\
&= G(\lambda) - \big(G(\lambda^0) + (\lambda - \lambda^0) \cdot p_{\lambda^0}[T]\big) \\
&= G(\lambda) - \big(G(\lambda^0) + (\lambda - \lambda^0) \cdot \nabla G(\lambda^0)\big).
\end{aligned}$$

– Difference between G and its affine approximation at λ^0.

– Since G is convex, this difference is always non-negative.

– Gap is called *Bregman divergence* $B_G(\lambda, \lambda^0)$ generated by G.
(Requires convexity of G, and differentiability of G at the second argument of B_G.)

- Can express Gibbs distribution p_λ in terms of a Bregman divergence

– $p_\lambda(x)$ should be a function of x (or $T(x)$), but B_G is a divergence for comparing natural parameters λ

– There is a Bregman divergence $B_F \colon \mathcal{M} \times \mathcal{M}^\circ \to \mathbb{R}_+$

$$B_F(\mu, \mu^0) = F(\mu) - \big(F(\mu^0) + (\mu - \mu^0) \cdot \nabla F(\mu^0)\big)$$

corresponding to another convex function $F\colon \mathcal{M} \to \mathbb{R}$ such that

$$p_\lambda(x) = \exp(-B_F(T(x), \nabla G(\lambda))q_F(x)$$

where q_F is a different base measure

- So what is this function F? (And what is q_F?)

- Convex duality: convex functions come in pairs G, G^*

$$G^*(\mu) := \sup_{\lambda \in \mathbb{R}^n} \lambda \cdot \mu - G(\lambda)$$

G^* is the *convex conjugate* (or *Fenchel conjugate*) of G

- G^* is supremum of affine functions (indexed by λ), so G^* is convex (even if G isn't convex!)
- Example: $G(\lambda) = \frac{1}{2}\|\lambda\|_2^2$

$$\begin{aligned} G^*(\mu) &= \sup_{\lambda \in \mathbb{R}^n} \lambda \cdot \mu - \frac{1}{2}\|\lambda\|_2^2 \\ &= \sup_{\lambda \in \mathbb{R}^n} -\frac{1}{2}\|\lambda - \mu\|_2^2 + \frac{1}{2}\|\mu\|_2^2 \\ &= \frac{1}{2}\|\mu\|_2^2 \end{aligned}$$

where the supremum is achieved by $\lambda = \mu$

- Example: $G(\lambda) = \ln(1 + \exp(\lambda))$

$$\begin{aligned} G^*(\mu) &= \sup_{\lambda \in \mathbb{R}^n} \lambda\mu - \ln(1 + \exp(\lambda)) \\ &= \mu \ln \mu + (1 - \mu) \ln(1 - \mu) \end{aligned}$$

where the supremum is achieved by $\lambda = \ln \frac{\mu}{1-\mu}$

- Example: $G(\lambda) = \ln(1 + \exp(\lambda_1) + \cdots + \exp(\lambda_n))$

$$G^*(\mu) = \sum_{i=1}^n \mu_i \ln \mu_i + \left(1 - \sum_{i=1}^n \mu_i\right) \ln\left(1 - \sum_{i=1}^n \mu_i\right)$$

- In general, assuming differentiability of G,

$$G^*(\mu) = \sup_{\lambda \in \mathbb{R}^n} \lambda \cdot \mu - G(\lambda)$$

has supremum "achieved" by λ satisfying $\mu = \nabla G(\lambda)$

- Letting $g = \nabla G$, for $\mu \in g(\mathbb{R}^n)$, we have

$$G^*(\mu) = g^{-1}(\mu) \cdot \mu - G(g^{-1}(\mu)).$$

– What is gradient of G^* (at $\mu \in g(\mathbb{R}^n)$)? Let J denote the Jacobian map for g^{-1}, so

$$\begin{aligned}\nabla G^*(\mu) &= J(\mu)\mu + g^{-1}(\mu) - J(\mu)\nabla G(g^{-1}(\mu)) \\ &= J(\mu)\mu + g^{-1}(\mu) - J(\mu)g(g^{-1}(\mu)) \\ &= g^{-1}(\mu).\end{aligned}$$

- Let $F := G^*$, so $\nabla F = g^{-1}$,

$$F(g(\lambda)) = \lambda \cdot g(\lambda) - G(\lambda),$$

and

$$\begin{aligned}p_\lambda(x) &= \exp(\lambda \cdot T(x) - G(\lambda))q_0(\lambda) \\ &= \exp(\lambda \cdot T(x) - (\lambda \cdot g(\lambda) - F(g(\lambda))))q_0(\lambda) \\ &= \exp(\lambda \cdot (T(x) - g(\lambda)) + F(g(\lambda)))q_0(\lambda) \\ &= \exp(\nabla F(g(\lambda)) \cdot (T(x) - g(\lambda)) + F(g(\lambda)))q_0(x) \\ &= \exp(-B_F(T(x), g(\lambda)))\underbrace{\exp(F(T(x)))q_0(x)}_{q_F(x)}.\end{aligned}$$

- Log-likelihood of Gibbs distribution parameter λ given data $x^1, \ldots, x^m \in \mathcal{X}$:

$$\sum_{j=1}^m \log p_\lambda(x^j) = -\sum_{j=1}^m B_F(T(x^j), g(\lambda)) + \text{terms not involving } \lambda$$

So MLE can be interpreted as minimizing a sum of Bregman divergences over the data.