2 Exponential Families and Information Inequality

- Exponential families form a basis for many statistical methodology such as generalized linear models (GLMs), undirected graphical models, etc.
- They define a broad class of distributions covering distributions such as Gaussian, Bernouilli, beta, Poisson etc.
- They also arise as the solutions of interesting optimization problems.

Definition 6. *Exponential families are defined as a collection of densities with respect to a base measure ν (either counting or Lebesgue)*

$$\mathcal{P} = \{p_\theta(x) : \theta \in \Theta\} \quad \textit{where} \quad p_\theta(x) = \exp\{\langle \theta, \phi(x)\rangle - \psi(\theta)\}p_0(x).$$

Above,

- $\theta \in \Theta \subset \mathbb{R}^d$*: Natural parameter*
- $\phi : \mathcal{X} \to \mathbb{R}^d$*: Sufficient statistics*
- $\psi : \mathbb{R} \to \mathbb{R}$*: log-partition function, cumulant generating function (CGF)*
- $p_0(x)$*: carrying density w.r.t. carrying measure $\nu(\mathrm{d}x)$ on $\mathcal{X}$. We will ignore this part mostly as it can be combined with the carrying measure ν.*

The natural parameter θ lives in a parametric space where the CGF is finite: $\Theta = \{\theta : \psi(\theta) < \infty\}$. Since p_θ is a density, we have

$$1 = \int p_\theta(x)\,\mathrm{d}\nu(x) \quad \text{and} \quad \psi(\theta) = \log\left\{\int \exp\{\langle \theta, \phi(x)\rangle\}p_0(x)\,\mathrm{d}\nu(x)\right\}.$$

Note that in this class we only consider the measure $d\nu(x)$ either as the Lebesgue measure when the random variable is continuous or as the counting measure when it is discrete.

Example. Let X be a Bernoulli random variable with mean μ, i.e., $\mathbb{P}(X = 1) = \mu$ and $\mathbb{P}(X = 0) = 1 - \mu$. We can write the probability mass function as $p_\theta(x) = \mu^x(1-\mu)^{1-x} = \exp\{x \log \mu + (1-x)\log(1-\mu)\}$ where $x \in \{0, 1\}$. One way to write the Bernoulli distribution as an exponential family is through the following natural parameter and sufficient statistic

$$\theta = \begin{bmatrix} \log \mu \\ \log(1-\mu) \end{bmatrix}, \quad \phi(x) = \begin{bmatrix} x \\ 1-x \end{bmatrix}$$

We say that an exponential family is *minimal* if there is no linear relations/constraints between the entries of the sufficient statistic and the natural parameter vectors. Notice that the above formulation is not minimal. Re-write the PMF, natural parameter, and CGF:

$$p(x) = \exp\left\{x \log \frac{\mu}{1-\mu} + \log(1-\mu)\right\} \quad \text{with}$$
$$\theta = \log \frac{\mu}{1-\mu}, \quad \psi(\theta) = \log(1 + e^\theta), \quad \mu = \frac{e^\theta}{1+e^\theta}.$$

Proposition 7. Θ *is a convex set, and $\psi(\theta)$ is a convex function.*

Proof. Θ is a convex set if for $\theta_1, \theta_2 \in \Theta$, $\theta_\lambda = \lambda\theta_1 + (1-\lambda)\theta_2 \in \Theta$, $\forall \lambda \in [0,1]$.

$$\psi(\theta) < \infty \;\; \Leftrightarrow \;\; e^{\psi(\theta)} < \infty \;\; \Leftrightarrow \;\; \int \exp\{\langle \theta, \phi(x)\rangle\} d\nu(x) < \infty$$

$$\begin{aligned}
\exp\big(\psi(\theta_\lambda)\big) &= \int \exp\{\langle \theta_\lambda, \phi(x)\rangle\} d\nu(x) = \int \Big(e^{\langle \theta_1, \phi(x)\rangle}\Big)^{\lambda} \Big(e^{\langle \theta_2, \phi(x)\rangle}\Big)^{1-\lambda} d\nu(x) \\
&\leq \left(\exp\big(\psi(\theta_1)\big) \int p_{\theta_1}(x) d\nu(x)\right)^{\lambda} \left(\exp\big(\psi(\theta_2)\big) \int p_{\theta_2}(x) d\nu(x)\right)^{1-\lambda} \\
&= \exp\big(\psi(\theta_1)\big)^{\lambda} \exp\big(\psi(\theta_2)\big)^{1-\lambda} < \infty.
\end{aligned}$$

Where the inequality is justified above from Hölder's inequality for integrals: $\int |fg| du \leq (\int |f|^p)^{1/p} (\int |g|^q)^{1/q}$, $p^{-1} + q^{-1} = 1$. This completes the proof of first part. The second part follows applying logs to both sides,

$$\psi(\theta_\lambda) \leq \lambda\psi(\theta_1) + (1-\lambda)\psi(\theta_2).$$

□

2.1 Moments of exponential families

It can be shown that the moments of the sufficient statistic associated with an exponential family can be linked to the corresponding orders of differentiation of that family's CGF.

- **Mean:** We can write

$$\begin{aligned}
1 &= \int p_\theta(x)\, \mathrm{d}\nu(x) = \int e^{\langle \theta, \phi(x)\rangle - \psi(\theta)}\, \mathrm{d}\nu(x) \;\; \text{differentiating both sides w.r.t } \theta \\
0 &= \int e^{\langle \theta, \phi(x)\rangle - \psi(\theta)} \Big(\phi(x) - \nabla\psi(\theta)\Big) d\nu(x) \\
0 &= \mathbb{E}[\phi(x)] - \nabla\psi(\theta) \underbrace{\int e^{\langle \theta, \phi(x)\rangle - \psi(\theta)} d\nu(x)}_{1} \;\; \Leftrightarrow \;\; \mathbb{E}[\phi(x)] = \nabla\psi(\theta) \coloneqq \mu
\end{aligned}$$

- **Variance:** Taking one more derivative yields that $\mathrm{Cov}(\phi(x)) = \nabla^2\psi(\theta)$.
- **Higher-order moments:** Similarly, by taking more derivatives of the above equation, we can obtain higher-order moments.

Proposition 8 (Invertibility)**.** *If ψ is strictly convex, then $\nabla\psi : \Theta \to \mathcal{M}$ is invertible for $\mathcal{M} = \{\mu : \mu = \nabla\psi(\theta) \;\; \text{for} \;\; \theta \in \Theta\}$.*

Proof. We need to show that for $\theta_1, \theta_2 \in \Theta$,

$$\theta_1 = \theta_2 \Leftrightarrow \nabla\psi(\theta_1) = \nabla\psi(\theta_2).$$

One side is trivial. For the other side, we write

$$\nabla\psi(\theta_2) = \nabla\psi(\theta_1) + \int_0^1 \nabla^2\psi(\theta_1 + \tau(\theta_2 - \theta_1))(\theta_2 - \theta_1) d\tau.$$

Suppose it was the case that $\exists \theta_1, \theta_2$, $\theta_1 \neq \theta_2$ such that $\nabla\psi(\theta_1) = \nabla\psi(\theta_2)$, then it would be that $0 = \int_0^1 \nabla^2\psi(\theta_1 + \tau(\theta_2 + \theta_1))(\theta_2 - \theta_1)d\tau$. However, because ψ is strictly convex, $\nabla^2\psi > 0$, so the previous integral must be greater than zero and therefore $\nabla\psi(\theta_1) \neq \nabla\psi(\theta_2)$. □

Since the mapping $\nabla\psi : \Theta \to \mathcal{M}$ is invertible, we can write

1. $(\nabla\psi)^{-1}(\mu) = \theta$
2. $\Sigma = \nabla^2_\theta\psi(\theta) = \nabla_\theta\nabla_\theta\psi(\theta) = \nabla_\theta\mu$ or equivalently, $\Sigma = \frac{\mathrm{d}\mu}{\mathrm{d}\theta}$ where Σ is the covariance matrix of $\phi(X)$. Intuitively (from Leibniz notation), we have $\frac{\mathrm{d}\theta}{\mathrm{d}\mu} = \Sigma^{-1}$. This can be shown using chain rule.

$$\mu = \nabla_\eta\psi(\eta) \implies \frac{d\mu}{d\mu} = \frac{d\eta}{d\mu}\nabla^2_\eta\psi(\eta) \implies \frac{d\eta}{d\mu} = \Sigma^{-1}.$$

2.2 MLE, Score, Information

In this section, we consider the basic MLE setup where we assume $\mathbf{x} = [x_1, \ldots, x_n]$ where $x_i \overset{iid}{\sim} p_\theta(x)$. Using the iid assumption, we can write the joint density as

$$\begin{aligned} p_\theta(\mathbf{x}) = \prod_{i=1}^n p_\theta(x_i) &= \exp\left\{\left\langle \theta, \sum_{i=1}^n \phi(x_i)\right\rangle - n\psi(\theta)\right\}\prod_{i=1}^n p_0(x_i) \\ &= \exp\left\{n[\langle\theta, \bar{\phi}\rangle - \psi(\theta)]\right\}p_0(\mathbf{x}), \quad \text{where} \quad \bar{\phi} = \frac{1}{n}\sum_{i=1}^n \phi(x_i). \end{aligned}$$

The corresponding log-likelihood, and score with respect to θ and μ are therefore:

- **Log-likelihood:** $\ell_\theta(\mathbf{x}) = n[\langle\theta, \bar{\phi}\rangle - \psi(\theta)] + \text{ const}$
- **Score w.r.t.** θ: $\nabla_\theta\ell_\theta(\mathbf{x}) = n[\bar{\phi} - \nabla\psi(\theta)]$
- **Score w.r.t.** μ: $\nabla_\mu\ell_\theta(\mathbf{x}) = \Sigma^{-1}n[\bar{\phi} - \nabla\psi(\theta)]$
- **Information w.r.t.** θ: $\mathcal{I}_\theta = \mathbb{E}[\nabla\ell_\theta\nabla\ell_\theta^T] = -\mathbb{E}[\nabla^2\ell_\theta] = n\Sigma$.
- **Information w.r.t.** μ: $\mathcal{I}_\mu = n\Sigma^{-1}$.

Remark. Information matrix quantifies how much information the observable statistic $\phi(\mathbf{x})$ contains about the parameter of interest.

We compute the **MLE** of natural parameter θ by solving the following equation for θ.

$$\begin{aligned} &\nabla\ell_\theta(\mathbf{x}) = 0 \Leftrightarrow \bar{\phi} = \nabla\psi(\hat{\theta}^{\mathrm{MLE}}) \quad \Leftrightarrow \\ &\hat{\theta}^{\mathrm{MLE}} = (\nabla\psi)^{-1}(\bar{\phi}) \text{ by the invertibility of } \nabla\psi. \end{aligned}$$

Similarly, we can also find the MLE for the mean μ by differentiating the log-likelihood w.r.t. μ and setting it to 0. Since we focus on strictly convex CGFs (which imply $\Sigma \succ 0$), the MLE can be computed to be $\hat{\mu}^{\mathrm{MLE}} = \bar{\phi}$. Therefore, we notice that the mapping $\nabla\psi$ also maps the MLEs.
Remark. As a side note, we can see that the score function has 0 expectation:

$$\mathbb{E}[\nabla\ell_\theta(\mathbf{x})] = \mathbb{E}[\bar{\phi}] - \nabla\psi(\theta) = \mu - \mu = 0.$$

Asymptotics of MLE: We can leverage the sample average structure of $\bar{\phi}$ and obtain its asymptotic distribution using Central Limit Theorem (CLT). That is, with a slight abuse of notation

$$\hat{\mu}^{\text{MLE}} = \bar{\phi} = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i) \approx\sim \mathcal{N}(\mu, \Sigma/n).$$

Here, it is worth noting that the distribution $\mathcal{N}$ is approximate, but the mean and the variance are exact. The correct way to state this result is

$$\sqrt{n}(\hat{\mu}^{\text{MLE}} - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma). \tag{2.1}$$

The asymptotic distribution of $\hat{\theta}^{\text{MLE}}$ requires an extra derivation. Notice that there is a non-linearity $(\nabla\psi)^{-1}$ applied to the sample average form $\bar{\phi}$. We know, by CLT, that $\bar{\phi}$ will be Gaussian, but we need a way of dealing with the non-linear function applied to it.

Proposition 9 (Delta Method)**.** *Assume that a random variable is asymptotically normal, i.e.,* $\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma)$. *Then, for a differentiable function* f*, we have*

$$\sqrt{n}(f(\hat{\mu}) - f(\mu)) \xrightarrow{d} \mathcal{N}(0, \nabla f(\mu)^\top \Sigma \nabla f(\mu)).$$

Using the Delta method on the asymptotic result obtained for $\hat{\mu}^{\text{MLE}}$ in (2.1), and also recalling that $\hat{\theta}^{\text{MLE}} = \nabla\psi^{-1}(\hat{\mu}^{\text{MLE}})$, we can write

$$\sqrt{n}(\hat{\theta}^{\text{MLE}} - \theta) \xrightarrow{d} \mathcal{N}(0, \nabla_\mu(\nabla\psi)^{-1}(\mu)^\top \Sigma \nabla_\mu(\nabla\psi)^{-1}(\mu)). \tag{2.2}$$

We can compute the quantity $\nabla_\mu(\nabla\psi)^{-1}(\mu)$ using the chain rule (left as exercise), but below we just use the Leibniz notation.

$$\nabla_\mu(\nabla\psi)^{-1}(\mu) = \frac{\mathrm{d}\theta}{\mathrm{d}\mu} = \left[\frac{\mathrm{d}\mu}{\mathrm{d}\theta}\right]^{-1} = \Sigma^{-1}.$$

Therefore, the variance term in (2.2) becomes

$$\nabla_\mu(\nabla\psi)^{-1}(\mu)^\top \Sigma \nabla_\mu(\nabla\psi)^{-1}(\mu) = \Sigma^{-1}.$$

Thus,

$$\hat{\theta}^{\text{MLE}} \approx\sim \mathcal{N}(\theta, \Sigma^{-1}/n) \tag{2.3}$$

or equivalently $\sqrt{n}(\hat{\theta}^{\text{MLE}} - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1})$. In (2.3), distribution $\mathcal{N}$, as well as the mean and the variance are approximate.

Remark. Proof for the delta method was hinted using the Taylor Series expansion of the function under consideration. This is a very handy theorem.

2.3 Information inequality

In this section, we will derive a lower bound on the variance of a generic estimator which we call as the information inequality. Later, we will use our main result here to derive the celebrated Cramer-Rao lower bound. Information inequality is very much related to the Fisher information which is where it get its name from. It is a classical concept and defines the notion of *efficiency* for estimators.

As in the previous section, suppose that we have data from an exponential family and we have a statistic of the form $T : \mathbb{R}^d \to \mathbb{R}^p$ with

$$\mathbb{E}_\theta[T(\bar{\phi})] = \xi(\theta)$$

for some differentiable function ξ.

Remark. If we have two matrices A and B, we write $A \succeq B$ if $A - B \succeq 0$, i.e., $A - B$ is positive semi-definite. This is equivalent to saying $\forall u$, $\langle u, (A-B)u\rangle \geq 0$.

Theorem 10 (Information Inequality)**.** *Variance of any estimator of the above form can be lower bounded as*

$$\mathrm{Var}(T(\bar{\phi})) \succeq \frac{1}{n}\nabla\xi(\theta)^\top \Sigma^{-1}\nabla\xi(\theta).$$

Proof. In the first step of the proof, we compute a useful expression for the $\nabla\xi$. We write

$$\begin{aligned}
\nabla\xi(\theta) &= \int \nabla p_\theta(x_1, .., x_n) T(\bar{\phi})^\top \,\mathrm{d}\nu \\
&= \int n[\bar{\phi} - \nabla\psi(\theta)] T(\bar{\phi})^\top p_\theta(x_1, .., x_n)\,\mathrm{d}\nu \\
&= n\mathbb{E}_\theta\Big[(\bar{\phi} - \nabla\psi(\theta))T(\bar{\phi})^\top\Big] \\
&= n\mathbb{E}_\theta\Big[(\bar{\phi} - \nabla\psi(\theta))(T(\bar{\phi}) - \xi(\theta))^\top\Big]
\end{aligned}$$

The first term inside the expectation is in $\mathbb{R}^d$ and the second term belongs to $\mathbb{R}^p$. Therefore, the above expectation is a $d \times p$ matrix.

Next, choose any vector $u \in \mathbb{R}^p$ and compute the quantity,

$$\begin{aligned}
\frac{1}{n}\langle \nabla\xi(\theta)u, \Sigma^{-1}\nabla\xi(\theta)u\rangle &= u^\top \mathbb{E}_\theta\Big[(T(\bar{\phi}) - \xi(\theta))(\bar{\phi} - \nabla\psi(\theta))^\top\Big]\Sigma^{-1}\nabla\xi(\theta)u \\
&= \mathbb{E}_\theta\big[\langle T(\bar{\phi}) - \xi(\theta), u\rangle\langle \bar{\phi} - \nabla\psi(\theta), \Sigma^{-1}\nabla\xi(\theta)u\rangle\big] \\
\text{(by Cauchy-Schwartz)} &\leq \mathbb{E}_\theta\Big[\langle T(\bar{\phi}) - \xi(\theta), u\rangle^2\Big]^{1/2}\mathbb{E}_\theta\Big[\langle \bar{\phi} - \nabla\psi(\theta), \Sigma^{-1}\nabla\xi(\theta)u\rangle^2\Big]^{1/2} \\
&\leq \langle u, \mathrm{Var}(T(\bar{\phi}))u\rangle^{1/2}\big[\langle \Sigma^{-1}\nabla\xi(\theta)u, \mathrm{Var}(\bar{\phi})\Sigma^{-1}\nabla\xi(\theta)u\rangle\big]^{1/2} \\
&= \langle u, \mathrm{Var}(T(\bar{\phi}))u\rangle^{1/2}\left[\frac{1}{n}\langle \nabla\xi(\theta)u, \Sigma^{-1}\nabla\xi(\theta)u\rangle\right]^{1/2}
\end{aligned}$$

where in the last step we used the fact that $\mathrm{Var}(\bar{\phi}) = \frac{1}{n}$. We notice that the second term on the last line is the square root of the left hand side. Canceling these and squaring both sides concludes the proof. □

An immediate corollary of this result is the celebrated Cramer-Rao lower bound.

Corollary 11 (Cramer-Rao Lower Bound)**.** *If $T(\bar{\phi})$ is an unbiased estimator for θ, i.e., $\mathbb{E}_\theta[T(\bar{\phi})] = \theta$, then*

$$\mathrm{Var}(T(\bar{\phi})) \succeq \frac{1}{n}\Sigma^{-1}.$$

The lower bound in the above corollary is the inverse Fisher information with respect to the parameter being estimated. That is, the bound reads $\mathrm{Var}(T(\bar{\phi})) \succeq \mathcal{I}_\theta^{-1}$.

If we were to estimate another parameter such as μ, we can derive a similar bound using the information inequality. In this case, our unbiased estimator $T(\bar{\phi})$ (for μ) has an expectation

$$\mathbb{E}_\theta[T(\bar{\phi})] = \xi(\theta) = \mu.$$

Notice that in this case $\xi = \nabla\psi$ and consequently $\nabla\xi = \nabla^2\psi = \Sigma$. Plugging this into the information inequality yields a lower bound on the variance of the estimator as

$$\mathrm{Var}(T(\bar{\phi})) \succeq \frac{1}{n}\Sigma\Sigma^{-1}\Sigma = \frac{1}{n}\Sigma = \mathcal{I}_\mu^{-1}.$$

Remarkably in this case, information inequality yields a lower bound which is again the inverse Fisher information with respect to the parameter being estimated.

Remark. Estimators that achieve Cramer-Rao lower bound are called *efficient*. For example, MLE for μ, $\bar{\phi}$, has the variance $\frac{1}{n}\Sigma$ which is the Cramer-Rao lower bound! So MLE already achieves this bound in this case. Although it is worth noting that MLE in general may not be efficient, yet it is asymptotically efficient.