Lecture 2 — September 24

Lecturer: Lester Mackey *Scribe: Stephen Bates and Andy Tsao*

2.1 Recap

Last time, we set out on a quest to develop optimal inference procedures and, along the way, encountered an important pair of assertions: not all data is relevant, and irrelevant data can only increase risk and hence impair performance. This led us to introduce a notion of lossless data compression (**sufficiency**): T is sufficient for $\mathcal{P}$ with $X \sim \mathbb{P}_\theta \in \mathcal{P}$ if $X \mid T(X)$ is independent of θ. How far can we take this idea? At what point does compression impair performance? These are questions of **optimal data reduction**.

While we will develop general answers to these questions in this lecture and the next, we can often say much more in the context of specific modeling choices. With this in mind, let's consider an especially important class of models known as the exponential family models.

2.2 Exponential Families

Definition 1. The model $\{\mathbb{P}_\theta : \theta \in \Omega\}$ forms an **s-dimensional exponential family** if each $\mathbb{P}_\theta$ has density of the form:

$$p(x;\theta) = \exp\left(\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - B(\theta)\right) h(x)$$

- $\eta_i(\theta) \in \mathbb{R}$ are called the **natural parameters**.
- $T_i(x) \in \mathbb{R}$ are its **sufficient statistics**, which follows from **NFFC**.
- $B(\theta)$ is the log-partition function because it is the logarithm of a normalization factor:

$$B(\theta) = \log\left(\int \exp\left(\sum_{i=1}^{s} \eta_i(\theta) T_i(x)\right) h(x) d\mu(x)\right) \in \mathbb{R}$$

- $h(x) \in \mathbb{R}$: base measure.

Exponential families are of particular interest to us, because many common distributions are exponential families (e.g., Normal, Binomial, and Poisson), and exponential families are closely linked to the notion of sufficiency and the notions of optimal data reduction.

Example 1: Exponential Distribution: $\mathcal{P} = \{\text{Exp}(\theta) : \theta > 0\}$
The densities takes the form

$$p(x;\theta) = \theta e^{-\theta x} \mathbb{I}\left[x \geq 0\right] = \exp\left(-\theta x + \log(\theta)\right) \mathbb{I}[x \geq 0]$$

yielding a 1-dimensional exponential family with

- $\eta_i(\theta) = -\theta$
- $T_i(x) = x$
- $B(\theta) = -\log(\theta)$
- $h(x) = \mathbb{I}[x \geq 0]$.

Notice that there is an ambiguity in the choice of η_i and T_i, i.e., the negative sign could have been attributed to either one. This is a general property of exponential families: their parameterization is not unique.

Example 2: Beta Distribution: $\mathcal{P} = \{\text{Beta}(\alpha, \beta) : \alpha, \beta > 0\}$, $\theta = (\alpha, \beta)$

The densities take the form

$$\begin{aligned} p(x;\theta) &= x^{\alpha-1}(1-x)^{\beta-1}\mathbb{I}[x \in (0,1)]\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \\ &= \exp\left((\alpha-1)\log(x) + (\beta-1)\log(1-x) + \log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)\right)\mathbb{I}[x \in (0,1)] \end{aligned}$$

yielding a 2-dimensional exponential family with

- $\eta_1(\theta) = \alpha - 1$, $\eta_2(\theta) = \beta - 1$
- $T = (T_1, T_2) \quad \text{for} \quad T_1(x) = \log(x)$, $T_2(x) = \log(1-x)$.
- $B(\theta) = -\log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)$
- $h(x) = \mathbb{I}[x \in (0,1)]$

Similar to the ambiguity in Example 1, here we could have written $p(x;\theta)$ as

$$p(x;\theta) = \exp\left(\alpha\log(x) + \beta\log(1-x) + \log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)\right)\frac{1}{x(1-x)}\mathbb{I}[x \in (0,1)]$$

which changes the natural parameters so that the new $\eta_1(\theta) = \alpha$, $\eta_2(\theta) = \beta$, and $h(x) = \frac{\mathbb{I}[x\in(0,1)]}{x(1-x)}$.

Definition 2. An exponential family is in **canonical form** when the density has the form

$$p(x;\eta) = \exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right) h(x)$$

This parameterizes the density in terms of the **natural parameters** η instead of θ.

For a given base measure h and collection of sufficient statistics $\{T_i(x)\}$, only some values of η will give rise to valid, normalizable densities.

Definition 3. The set of all valid natural parameters Θ is called the **natural parameter space**: for each $\eta \in \Theta$, there exists a normalizing constant $A(\eta)$ such that $\int p(x,\eta) = 1$. Equivalently,

$$\Theta = \left\{\eta : 0 < \int \exp\left(\sum_{i=1}^{s} \eta_i T_i(x)\right) h(x) d\mu(x) < \infty\right\}.$$

Thus for any canonical exponential family, $\mathcal{P} = \{\mathbb{P}_\eta : \eta \in H\}$, we have $H \subseteq \Theta$. We note in passing that Θ is always a convex set.

2.2.1 Reducing the dimension

There are two cases when the superficial dimension of an s-dimensional exponential family $\mathcal{P} = \{\mathbb{P}_\eta : \eta \in H\}$ can be reduced.

Case 1: The $T_i(x)$'s satisfy an affine equality constraint $\forall x \in \mathcal{X}$.

Example:
$X \sim \text{Exp}(\eta_1, \eta_2)$, $p(x;\eta_1,\eta_2) = \exp\left(-\eta_1 x - \eta_2 x + \log(\eta_1+\eta_2)\right)\mathbb{I}[x \geq 0]$
Here, $T_1(x) = T_2(x) = x$, i.e. they're linearly dependent. We can collapse (η_1, η_2) to $\eta_1 + \eta_2$ and write the l.h.s. as $\exp\left(-(\eta_1+\eta_2)x + \log(\eta_1+\eta_2)\right)\mathbb{I}[x \geq 0]$.

Case 1 typically yields unidentifiability.

Definition 4. If $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Omega\}$, then θ is **unidentifiable** if for two parameters $\theta_1 \neq \theta_2$, $\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2}$.

In the above example, $p(x;\eta_1 + a, \eta_2 - a) = p(x;\eta_1,\eta_2)$, for any $a < \eta_2$.

Case 2: The η_i's satisfy an affine equality constraint for all $\eta \in H$.

Example:

$$\begin{aligned} p(x;\eta) &\propto \exp\left(\eta_1 x + \eta_2 x^2\right) \quad \text{for all} \quad (\eta_1,\eta_2) \quad \text{satisfying} \quad \eta_1+\eta_2 = 1 \\ &= \exp\left(\eta_1(x-x^2) + x^2\right) \end{aligned}$$

In either case, it is possible to transform the s-dimensional exponential family into an exponential family of smaller dimension. When neither case holds, we call an exponential family **minimal**:

Definition 5. A canonical exponential family $\mathcal{P} = \{\mathbb{P}_\eta : \eta \in H\}$ is **minimal** if

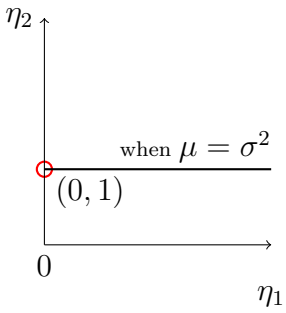
- $\sum_{i=1}^s \lambda_i T_i(x) = \lambda_0, \forall x \in \mathcal{X} \Rightarrow \lambda_i = 0, \forall i \in \{0, \ldots, s\}$ (no affine T_i equality constraints)
- $\sum_{i=1}^s \lambda_i \eta_i = \lambda_0, \forall \eta \in H \Rightarrow \lambda_i = 0, \forall i \in \{0, \ldots, s\}$ (no affine η_i equality constraints).

We will be considering two classes of minimal exponential families.

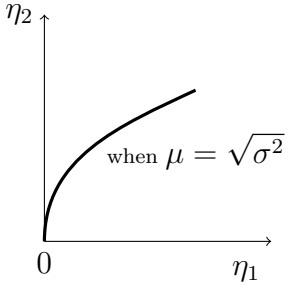
Definition 6. Suppose $\mathcal{P} = \{\mathbb{P}_\eta : \eta \in H\}$ is an s-dimensional minimal exponential family. If H contains an open s-dimensional rectangle, then $\mathcal{P}$ is called **full-rank**. Otherwise, $\mathcal{P}$ is **curved**. In curved exponential families, the η_i's are related in a non-linear way.

To summarize, we've defined three types of exponential families. We illustrate them below using the **normal distribution** $N(\mu, \sigma^2)$, where $\eta_1 = \frac{1}{2\sigma^2}, \eta_2 = \frac{\mu}{\sigma^2}, T_1(x) = -x^2, T_2(x) = x$:

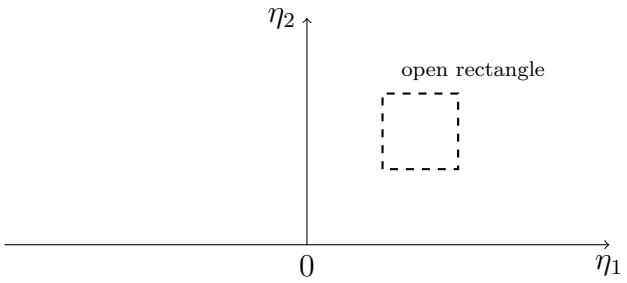
- Non-minimal (so that the dimension can be reduced): e.g., when $\mu = \sigma^2, \eta_1 = \frac{1}{2\sigma^2}, \eta_2 = 1$.



- Minimal & Curved: e.g., $\mu = \sqrt{\sigma^2}$, so $\eta_1 = \frac{1}{2\sigma^2}, \eta_2 = \frac{1}{\sqrt{\sigma^2}}, \eta_2^2 = 2\eta_1$.



- Minimal & Full-Rank: e.g., no extra constraint, where the natural parameter space is $(0, +\infty) \times \mathbb{R}$.



All three normal examples are superficially 2-dimensional, but the exponential family dimension is only irreducible in the last two cases.

2.2.2 Properties of Exponential Families

Let us consider a few important properties of exponential families.

Property 1: If $X_1, ..., X_n \overset{\text{iid}}{\sim} p(x;\theta) = \exp\left(\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - B(\theta)\right) h(x)$, then

$$p(x_1, ..., x_n; \theta) = \exp\left(\sum_{i=1}^{s} \eta_i(\theta) \sum_{j=1}^{n} T_i(x_j) - nB(\theta)\right) \prod_{j=1}^{n} h(x_j).$$

By the Neyman-Fisher factorization criterion, $(\sum_j T_1(x_j), ..., \sum_j T_s(x_j))$ is therefore a sufficient statistic, and hence exponential family data is exceptionally compressible: we can find an s-dimensional sufficient statistic for any sample size!

Property 2: If f is integrable and $\eta \in \Theta$, then $G(f,\eta) = \int f(x) \exp\left(\sum_{i=1}^{s} \eta_i T_i(x)\right) h(x) d\mu(x)$ is infinitely differentiable w.r.t. η and the derivatives can be obtained by differentiating under the integral sign. (Proof: See TSH 2.7.1 based on the dominated convergence theorem.)

Example 3: Moments of $T_i's$
Take $f(x) = 1$, then

$$G(f,\eta) \triangleq \int \exp\left(\sum_{j=1}^{s} \eta_j T_j(x)\right) h(x) d\mu(x) = \exp\left(A(\eta)\right)$$

$$\frac{\partial G(f,\eta)}{\partial \eta_i} = \int T_i(x) \exp\left(\sum_{j=1}^{s} \eta_j T_j(x)\right) h(x) d\mu(x) = \frac{\partial A(\eta)}{\partial \eta_i} \exp\left(A(\eta)\right)$$

$$\frac{\partial A(\eta)}{\partial \eta_i} = \int T_i(x) \exp\left(\sum_{j=1}^{s} \eta_j T_j(x) - A(\eta)\right) h(x) d\mu(x) = \mathbb{E}_\eta[T_i(x)].$$

so we can compute the means of the sufficient statistics by taking partial derivatives of the log-partition function! We can in fact compute all of the moments of the sufficient statistics in a similar manner. For example,

$$\frac{\partial^2 A(\eta)}{\partial \eta_i \partial \eta_j} = \mathrm{Cov}_\eta(T_i(x), T_j(x)).$$

2.3 Optimal Data Reduction via Minimal Sufficiency

Now, let's return to our initial question of optimal data reduction. We begin by defining a function of the data that cannot be reduced without sacrificing information about the model.

Definition: A sufficient statistic T is **minimal** if for every sufficient statistic T', T is a function of T'. Equivalently, T is minimal if for every sufficient statistic T', $T(x) = T(y)$ whenever $T'(x) = T'(y)$.

A minimal sufficient statistic represents the maximal (and hence optimal) lossless compression of our data. The following theorem provides a straightforward way to derive or check for minimal sufficient statistics:

Theorem 1: Let $p(x;\theta)$ be a density of X (w.r.t. μ). A statistic T is minimal sufficient if for every $x, y \in \mathcal{X}$, there exists $c_{x,y}$ independent of θ such that $p(x;\theta) = c_{x,y}p(y;\theta) \iff T(x) = T(y)$.

We will prove this theorem next time.