Lecture 3

Outline

1) Review
2) Exponential families
3) Differential identities
4) MGF
Exponential Families

An $s$-parameter exponential family is a family

$$ P = \{ p_\eta : \eta \in \Xi \} $$

with densities $p_\eta$ wrt a common measure $\mu$ on $X$ such that $X$ not nec. in $\mathbb{R}^n$, of the form

$$ p_\eta(x) = e^{\eta' T(x) - A(\eta)} h(x) \quad \text{where} $$

- $T : X \rightarrow \mathbb{R}^s$ sufficient statistic
- $h : X \rightarrow \mathbb{R}$ carrier/base density
- $\eta \in \Xi \subseteq \mathbb{R}^s$ natural parameter
- $A : \mathbb{R}^s \rightarrow \mathbb{R}$ log-partition function or normal/Bayes constant

Note: The function $A(\cdot)$ is totally determined by $h$ and $T$, since we must always have

$$ \int_X p_\eta d\mu = 1 \quad \forall \eta.$$ 

$$ \Rightarrow A(\eta) = \log \left[ \int_X e^{\eta' T(x)} h(x) d\mu(x) \right] \leq \infty $$
The natural parameter space is the set of all \( \eta \) that give us normalizable \( p_x \)

\[
\Xi_1 = \{ \eta : A(\eta) < \infty \}
\]

Note \( \Xi_1 \), determined by \( T, h, \mu \)

We could take \( \Xi \subseteq \Xi_1 \) if we wanted

\( A(\eta) \) is always convex

\[ \Rightarrow \quad \Xi_1 \) is convex \quad (HW1 Prob. 2) \]

**Example: Poisson**

\[
X \sim \text{Pois}(\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x = 0, 1, \ldots
\]

\[
p_\lambda(x) = \exp \left\{ (\log \lambda) x - \lambda \right\} \frac{1}{x!}
\]

\[
\eta(\lambda) = \log \lambda \quad T(x) = x
\]

\[
A(\lambda) = \lambda = e^{\eta} \quad h(x) = \frac{1}{x!}
\]
Differential Identities

Write \( e^{A(\eta)} = \int e^{\eta' T(x)} h(x) \, d\mu(x) \) \((\star)\)

We can derive lots of useful identities by differentiating \((\star)\) on both sides, pulling derivative inside \(\int\) \([\text{not always allowed}]\)

Keener Thm 2.4 for \( f : X \to \mathbb{R} \) let

\[
\Xi_f = \{ \eta \in \mathbb{R}^s : \int_{X} |f| e^{\eta^T} h \, d\mu < \infty \}
\]

Then \( g(\eta) = \int_{X} f e^{\eta^T} h \, d\mu \) has cts partial derivatives of all orders for \( \eta \in \Xi_f^0 \), & we can get them by differentiating under the \(\int\) sign.

\( \Rightarrow \) on \( \Xi_f^0 \), \( A(\eta) \) has all partial derivatives

Differentiate once:

\[
\frac{\partial}{\partial \eta_j} e^{A(\eta)} = \frac{\partial}{\partial \eta_j} \int e^{\eta^T T(x)} h(x) \, d\mu(x)
\]

\[
e^{A(\eta)} \frac{\partial A}{\partial \eta_j} (\eta) = \int T_j(x) e^{\eta^T T(x)} - A(\eta) h(x) \, d\mu(x)
\]

\( \Rightarrow \) \[
\frac{\partial A}{\partial \eta_j} (\eta) = \mathbb{E}_\eta[T_j(X)]
\]

\( \nabla A(\eta) = \mathbb{E}_\eta[T(X)] \)
Diff twice:
\[
\frac{\partial^2}{\partial \eta_j \partial \eta_k} e^{A(\eta)} = \frac{\partial^2}{\partial \eta_j \partial \eta_k} \int e^{z^T h} dm
\]

\[
e^{A(\eta)} \left( \frac{\partial^2 A}{\partial \eta_j \partial \eta_k} + \frac{\partial A}{\partial \eta_j} \frac{\partial A}{\partial \eta_k} \right) = \int T_j T_k e^{z^T h} \frac{e^{z^T A(\eta)}}{E[T_j T_k]} \]

\[
\frac{\partial^2 A}{\partial \eta_j \partial \eta_k} (\eta) = \text{Cov}_{\eta} (T_j, T_k)
\]

\[

\nabla^2 A(\eta) = \text{Var}_{\eta} (T(x)) \in \mathbb{R}^{5 \times 5}
\]

Example: Poisson: \( T(x) = X \), \( A(\eta) = e^\eta (= \lambda) \)
\[
E[X] = \frac{d}{d\eta} e^\eta = e^\eta = \lambda
\]
\[
\text{Var}_\eta (X) = \frac{d^2}{d\eta^2} e^\eta = e^2 = \lambda
\]

NB: We would get wrong answer by differentiating wrt \( \lambda \).
Moment-generating function

We can get $k^{th}$ order moments of $T(X)$ by

1) Differentiating $(*)$ $k$ times, then
2) Dividing by $e^{A(z)}$

That is because $M^T_{x}(u) = e^{A(z+u)} - A(z)$ is the moment-generating function (mgf) of $T(X)$ when $X \sim P_x$

$$M^T_{x}(u) = \mathbb{E}_x \left[ e^{u^T T(x)} \right]$$

$$= \int e^{u^T \pi - A(\pi)} h d\mu$$

$$= e^{A(z+u)} - A(z) \int e^{(z+u)^T - A(z+u)} h d\mu$$

Useful for

- finding moments
- finding dist. of sums of indep. RVs

Cumulant-generating function

$$K^T_{x}(u) = \log M^T_{x}(u) = A(z+u) - A(z) \quad (A \text{ is sometimes called cgf})$$
Other Parameterizations

Sometimes it is more convenient to use a different parameterization:

\[ p_\theta(x) = e^{\gamma(\theta)' T(x) - B(\theta) h(x)} \]
\[ B(\theta) = A(\gamma(\theta)) \]

Many, many examples, sometimes requires massaging to see that they are exp. fam.s:

**Ex: Normal** \( X \sim N(\mu, \sigma^2) \) \( \mu \in \mathbb{R} \) \( \sigma^2 > 0 \)

Let \( \theta = (\mu, \sigma^2) \)

\[ p_\theta(x) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

\[ = \exp \left\{ \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \log \left( 2\pi \sigma^2 \right) \right\} \]

\[ \gamma(\theta) = \left( \frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right) \]
\[ T(x) = \left( x, x^2 \right) \]
\[ h(x) = 1 \]
\[ B(\theta) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2} \log(2\pi \sigma^2) \]

Natural parameterization

\[ p_\gamma(x) = e^{\gamma'(x)^2} - A(\gamma) \]

\[ A(\gamma) = \frac{\gamma^2}{4\gamma^2} + \frac{1}{\gamma^2} \log\left( -\frac{\gamma}{\gamma^2} \right) \]
More examples

**Binomial**

\[ X \sim \text{Binom}(\gamma, \theta) \]

\[
p_\theta(x) = \theta^x (1-\theta)^{n-x} \begin{pmatrix} n \end{pmatrix} \quad x = 0, \ldots, n
\]

\[
= \left(\frac{\theta}{1-\theta}\right)^x (1-\theta)^n \begin{pmatrix} n \end{pmatrix}
\]

\[
= \exp \left\{ \log\left(\frac{\theta}{1-\theta}\right) \cdot x + n \log(1-\theta) \right\} \begin{pmatrix} n \end{pmatrix}
\]

\[
\gamma(\theta) = \log\left(\frac{\theta}{1-\theta}\right) \quad \text{"log odds ratio"}
\]

**Beta**

\[ X \sim \text{Beta}(\alpha, \beta) \]

\[
p_{\alpha,\beta}(x) = x^{\alpha-1} (1-x)^{\beta-1} / B(\alpha,\beta)
\]

\[
= \exp \left\{ \alpha \log x + \beta \log(1-x) - \log B(\alpha,\beta) \right\} \frac{1}{x(1-x)}
\]

\[
\xi = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \quad T(x) = \begin{pmatrix} \log x \\ \log(1-x) \end{pmatrix} \quad h(x) = \frac{1}{x(1-x)}
\]

Practically everything else on wikipedia too:

Beta, Gamma, Multinom., Dirichlet, Pareto, Wishart...
Interpretation: Exponential tilting

Can think of \( p_\xi(x) = e^{\xi'T(x) - A(\xi)} h(x) \) as an exponential tilt of the carrier \( h(x) \)

1) Start with carrier \( h(x) \)
2) Multiply by \( e^{\xi'T(x)} \)
3) Re-normalize by \( e^{-A(\xi)} \)

\( T(x) = (T_1(x), \ldots, T_s(x)) \) gives linear space of directions in which we can tilt \( h(x) \)

\( \Xi_1 \) = all tilts after which normalization is possible

\[ \Rightarrow \text{Decomposition into } \xi, T, h, A \text{ very non-unique} \]

1) Only \( \text{span}(T_1, \ldots, T_s) \) matters
2) Could absorb \( h \) into \( \mu \) \( (d\mu(x) = h \, d\mu(x)) \)
   \( (wlog \, h(x) \equiv 1 \text{ if we want}) \)
3) Can add constant to \( T(x) \)
Repeated Sampling

Suppose \( X_1, \ldots, X_n \overset{iid}{\sim} \rho_{\xi}(x) = e^{\zeta^T T(x) - A(\xi) h(x)} \)

Then \( X = (X_1, \ldots, X_n) \) comes from a closely related family

\[
\rho_{\xi}(x) = \prod_{i=1}^{\hat{n}} e^{\zeta^T T(X_i) - A(\xi) h(X_i)}
= \exp \left\{ \zeta^T \sum_{i=1}^{\hat{n}} T(X_i) - \hat{n} A(\xi) \right\} \prod_{i=1}^{\hat{n}} h(X_i)
\]

Important property!

This means \( \sum_{i=1}^{\hat{n}} T(X_i) \in \mathbb{R}^S \) is an effective summary of a potentially very large sample \( X \in \mathcal{X} \)

We will often analyze \( T(X) \) as a proxy for the whole data set.
**Distribution of \( T(X) \)**

Suppose \( X \sim p_\xi(x) = e^{\frac{\xi' T(x) - A(\xi)}{\omega} \text{ wrt } \mu} \) (\( \omega \log h \equiv 1 \))

Then \( T(X) \sim q_\xi(t) = e^{\frac{\xi' t - A(\xi)}{\omega}} \text{ wrt } \nu, \)

where \( \nu \) is the measure \( \mu \) "pushed forward" through \( T : X \to \mathbb{R}^s \)

\( \nu(B) = \mu(\{x : T(x) \in B\}) \)

\[
\mathbb{P}_\xi(T(X) \in B) = \sum_8 \mathbb{1}_B(T(x)) e^{\frac{\xi' T(x) - A(\xi)}{\omega}} d\mu(x)
\]

\[
= \sum_8 \mathbb{1}_B(t) e^{\frac{\xi' t - A(\xi)}{\omega}} d\nu(t)
\]

**Simplest in discrete case:** (drop \( h \equiv 1 \) assumption)

\[
\mathbb{P}_\xi(T(X) = t) = \sum_{x : T(x) = t} \mathbb{1}_B(T(x)) e^{\frac{\xi' T(x) - A(\xi)}{\omega}} h(x) \mathbb{1}_B(\{x\})
\]

\[
= e^{\frac{\xi' t - A(\xi)}{\omega}} \sum_{x : T(x) = t} h(x) \mathbb{1}_B(\{x\})
\]

\[
\nu(\{t\})
\]