Outline

1) Review
2) Exponential families

Review

Model \( \mathcal{P} = \{ p_\theta : \theta \in \Theta \} \)

Observe \( X \sim p_\theta \), guess \( g(\theta) \)

Loss \( L(\theta, d) \), e.g. \( (d - g(\theta))^2 \)

Risk \( R(\theta; \delta(\cdot)) = \mathbb{E}_\theta[L(\theta, \delta(x))] \)
   
   e.g. \( \text{MSE}(\theta; \delta) = \mathbb{E}_\theta[(\delta(x) - g(\theta))^2] \)

Generally not possible to find uniform best estimator \( \delta \), but we can either
1) Summarize \( R \) by a scalar \( \delta \) (or \( \delta \))
2) Constrain choice of \( \delta \) (e.g. unbiased)

Ex. fam.

\[
\rho_\eta(x) = e^{\eta^T (x) - A(\eta)} h(x) \quad \text{wrt.} \quad \mu
\]

\( \mathcal{E}_1 = \{ \eta : A(\eta) = \log \int e^{\eta^T h(x)} \, dx < \infty \} \)
Exponential Families

An $s$-parameter exponential family is a family $\mathcal{F} = \{ P_\eta : \eta \in \mathcal{E} \}$ with densities $P_\eta$ wrt a common measure $\mu$ on $X$ such that

$$P_\eta(x) = e^{\eta' T(x) - A(\eta)} h(x),$$

where

$T : X \to \mathbb{R}^s$ is the sufficient statistic,

$h : X \to \mathbb{R}$ is the carrier/base density,

$\eta \in \mathcal{E} \subseteq \mathbb{R}^s$ is the natural parameter,

$A : \mathbb{R}^s \to \mathbb{R}$ is the cumulant-generating function (cgf) or normalizing constant.

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

$$\int_X P_\eta \, d\mu = 1, \quad \forall \eta.$$

$$\Rightarrow A(\eta) = \log \left[ \int_X e^{\eta' T(x)} h(x) \, d\mu(x) \right]$$

$P_\eta$ is normalizable iff $A(\eta) < \infty$. 
The natural parameter space is the set of all allowable (normalizable) \( \eta \):

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

We say \( \eta \) is in canonical form if \( \mathcal{E} = \mathcal{E} \).

Note \( \mathcal{E} \) determined by \( T, h, M \).

We could take \( \mathcal{E} \neq \mathcal{E} \), if we wanted.

\( A(\eta) \) is convex \( \Rightarrow \mathcal{E} \), is convex (HW1 Prob. 2).

Sometimes it is more convenient to use a different parameterization:

\[
\rho_\theta(x) = e^{\gamma(\theta)^\top T(x) - B(\theta)} h(x) \\
B(\theta) = A(\eta(\theta))
\]

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

Ex. 2.2: \( X \sim N(m, \sigma^2) \) \( m \in \mathbb{R} \) \( \sigma^2 > 0 \)

Let \( \Theta = (m, \sigma^2) \)

\[
\rho_\theta(x) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{- (m-x)^2 / 2\sigma^2}
\]

\[
= \exp \left\{ \frac{m}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 - \frac{m^2}{2\sigma^2} - \frac{1}{2} \log (2\pi \sigma^2) \right\}
\]
\( \eta(\theta) = \left( -\frac{1}{2\sigma^2} \right) \quad T(x) = (x^2) \quad h(x) = 1 \)

\[ B(\theta) = \frac{m^2}{2\sigma^2} + \frac{1}{2} \log \left( 2\pi \sigma^2 \right) \]

In terms of \( \tilde{\eta} \):

\[ \rho_{\tilde{\eta}}(x) = e^{\tilde{\eta}'(x^2)} - A(\tilde{\eta}) \quad ; \quad A(\tilde{\eta}) = \frac{-\tilde{\eta}^2}{\gamma^2} + \frac{1}{2} \log \left( \frac{-\pi \gamma^2}{\eta^2} \right) \]

It can be useful to analogize \( \{ \rho_{\tilde{\eta}} : \tilde{\eta} \in \Xi \} \)
to an \( s \)-dim. hyperplane in \( \infty \)-dimensional log-density space:

Assume \( \text{wlog} \quad h(x) > 0 \quad \forall x \)

(o.w. we could truncate to \( \tilde{\mathcal{X}} = \{ x \in \mathcal{X} : h(x) > 0 \} \)

Let \( f_{\tilde{\eta}}(x) = \log h(x) + \tilde{\eta}' T(x) \)

\( e^f_{\tilde{\eta}}(x) \) is always a density, but not always normalizable.

\( \Xi_1 = \{ \tilde{\eta} : \int e^f_{\tilde{\eta}} d\mu < \infty \} \)

Then \( \rho_{\tilde{\eta}}(x) = e^{f_{\tilde{\eta}}(x)}/\int e^{f_{\tilde{\eta}}(x)} d\mu(x) \)

\[ = e^{\tilde{\eta}' T(x)} h(x)/e^{A(\tilde{\eta})} \]

for \( \tilde{\eta} \in \Xi \subseteq \Xi_1 = \{ \eta : e^f_{\eta} \text{ normable} \} \)
Functional form of $\rho_n$ is in many ways not unique: for example we could always

a) Reformulate so $h(x) = 1$:

   $m \rightarrow \tilde{m}$ with $\frac{dm}{d\tilde{m}} = h$

   $h \rightarrow \tilde{h}(x) \equiv 1$

b) Re-parameterize so $0 \in \Xi$:

   Take some $\eta_0 \in \Xi$

   $\eta \rightarrow \tilde{\eta} = \eta - \eta_0$

   $h \rightarrow \tilde{h}(x) = \rho_\eta(x)$

   $A \rightarrow \tilde{A}(\tilde{\eta}) = A(\eta_0 + \tilde{\eta}) - A(\eta_0)$

c) For $c \in \mathbb{R}^s$, invertible $D \in \mathbb{R}^{s \times s}$, change sufficient statistic to

   $T \rightarrow \tilde{T}(x) = c + DT(x)$

   $\eta \rightarrow \tilde{\eta} = (D^{-1})' \eta$

   $A \rightarrow \tilde{A}$ as appropriate

etc...

Interp:

- Start with a carrier density $h(x)$
- Apply exponential tilt:
  1) multiply by $e^{\eta \cdot \tau(x)}$
  2) re-normalize to prob. density
Ex. fam in canonical form is all normalizable tilts of $h(x)$ using linear combos of $T_1(x), \ldots, T_k(x)$

Ex. 2.3:  $X_1, \ldots, X_n \overset{iid}{\sim} N(m, \sigma^2)$

$$
\rho_\theta(x) = \prod_{i=1}^{n} \rho_\theta(x_i)
$$

$$
= \exp \left\{ \sum_{i=1}^{n} \frac{m}{\sigma^2} x_i - \frac{1}{2\sigma^2} x_i^2 - \left( \frac{m}{2\sigma^2} + \frac{1}{2} \log(2\pi\sigma^2) \right) \right\}
$$

$$
= \exp \left\{ \frac{m}{\sigma^2} \sum x_i - \frac{1}{2\sigma^2} \sum x_i^2 - n \left( \ldots \right) \right\}
$$

$$
\eta(\theta) = \begin{pmatrix} m/\sigma^2 \\ -\frac{1}{2\sigma^2} \end{pmatrix} \quad T(x) = \begin{pmatrix} \sum x_i \\ \sum x_i^2 \end{pmatrix}
$$

$$
B(m, \sigma^2) = nB^{(1)}(m, \sigma^2)
$$

No accident these came out so similar:
an iid sample from an exp fam is also an exp fam

Suppose  $X_1, \ldots, X_n \overset{iid}{\sim} \rho_\theta(x) = e^{\mathbf{\eta}' T(x) - A(\mathbf{\eta})} h(x)$

Then  $X \sim \prod_{i=1}^{n} e^{\mathbf{\eta}' T(x_i) - A(\mathbf{\eta})} h(x_i)$

$$
= \exp \left\{ \mathbf{\eta}' \sum T(x_i) - n A(\mathbf{\eta}) \right\} \prod_{i} h(x_i)
$$

suff. stat.  \quad \text{cgf}  \quad \text{carrier}
The sufficient statistic $T(X)$ follows a related exp fam too:

Suppose $X \in \mathcal{X}$, $T(X) \in \mathcal{Y} = T(\mathcal{X})$, $\mu$ base measure (wlog $h = 1$)

For $B \in \mathcal{Y}$ define $\nu(B) = \mu(T^{-1}(B))$
($\nu$ called "push-forward" measure)

Then $T(X) \sim q_\eta(t) = e^{\gamma t - A(\eta)}$ wrt $\nu$

Discrete case:

$$P_\eta(T(X) \in B) = \sum_{x : T(x) \in B} e^{\gamma T(x) - A(\eta)} \mu(\{x\})$$

$$= \sum_{t \in B} \sum_{x : T(x) = t} e^{\gamma t - A(\eta)} \mu(\{x\})$$

$$= \sum_{t \in B} e^{\gamma t - A(\eta)} \nu(\{t\})$$

More generally, $\nu$ satisfies

$$\int f(T(x)) \, d\mu(x) = \int f(t) \, d\nu(t), \quad \forall \text{ ("nice") } f$$

$$P_\eta(T(X) \in B) = \int_{\mathcal{X}} 1\{T(x) \in B\} e^{\gamma T(x) - A(\eta)} \, d\mu(x)$$

$$= \int_{\tau} 1\{t \in B\} e^{\gamma t - A(\eta)} \, d\nu(t)$$

$$\Rightarrow T \sim e^{\gamma t - A(\eta)} \text{ density wrt } \nu$$
More examples

**Binomial** \( X \sim \text{Binom}(\eta, \theta) \)

\[
\rho_\theta(x) = \theta^x (1-\theta)^{\eta-x} \binom{\eta}{x}
\]

\[
= \left(\frac{\theta}{1-\theta}\right)^x (1-\theta)^{\eta-x} \binom{\eta}{x}
\]

\[
= \exp \left\{ \log \left(\frac{\theta}{1-\theta}\right) \cdot x + \eta \log(1-\theta) \right\} \binom{\eta}{x}
\]

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

**Poisson** \( X \sim \text{Pois}(\lambda) = \frac{x^\lambda e^{-\lambda}}{x!} \quad x \in \mathbb{N} \)

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

\[
\eta(\lambda) = \log \lambda
\]

Practically everything else on wikipedia too:

Beta, Gamma, Multinom., Dirichlet, Pareto, Wishart...
Differential Identities

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

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

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

\[ \Xi f = \{ \eta \in \mathbb{R}^s : \int \! |f e^{\eta^T} h \, d\mu < \infty \} \]

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

\[ \Rightarrow \text{on } \Xi \overset{0}{}, \ A(\eta) \text{ has all partial derivatives} \]

Differentiate once:

\[ \frac{\partial}{\partial \eta_i} e^{A(\eta)} = \frac{\partial}{\partial \eta_i} \int e^{\eta' T(x)} h(x) \, d\mu(x) \]

\[ e^{A(\eta)} \frac{\partial A}{\partial \eta_i} (\eta) = \int \! T_j(x) e^{\eta' T(x)} h(x) \, d\mu(x) \]

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

\[ \nabla A(\eta) = \mathbb{E}_\eta [T(X)] \]
Diff twice:
\[
\frac{\partial^2}{\partial \gamma_i \partial \gamma_k} e^{A(\gamma)} = \frac{\partial^2}{\partial \gamma_i \partial \gamma_k} \int e^{\gamma' \mathbf{T}} h d\mu
\]
\[
e^{-A(\gamma)} \left( \frac{\partial^2 A}{\partial \gamma_i \partial \gamma_k} + \frac{\partial A}{\partial \gamma_i} \frac{\partial A}{\partial \gamma_k} \right) = \frac{\sum T_i T_k e^{\gamma' \mathbf{T}} - A(\gamma)}{E[T_i T_k]}
\]
\[
\frac{\partial^2 A}{\partial \gamma_i \partial \gamma_k} = \text{Cov}_\gamma (T_i, T_k)
\]
\[
\nabla^2 A(\gamma) = \text{Var}_\gamma (T(x)) \in \mathbb{R}^{5\times 5}
\]

Example: Poisson: \( T(x) = X \), \( \gamma(x) = \log \lambda \)
\[
B(\lambda) = \lambda \Rightarrow A(\gamma) = e^\gamma
\]
\[
E[X] = \frac{d}{d\gamma} e^\gamma = e^\gamma = \lambda
\]
\[
\text{Var}_\gamma (X) = \frac{d^2}{d\gamma^2} e^\gamma = e^\gamma = \lambda
\]
Moment-generating function

Differentiating \((*)\) repeatedly we get

\[ e^{-A(\eta)} \frac{\partial^{k_1 + \ldots + k_s}}{\partial \eta_1^{k_1} \ldots \partial \eta_s^{k_s}} (e^{A(\eta)}) = \mathbb{E}_{\tilde{\eta}} \left[ T_1^{k_1} \ldots T_s^{k_s} \right] \]

That is because \( M^{\tau(x)}(\eta) = e^{A(\eta + u) - A(\eta)} \)

is the moment-generating function \((\text{mgf})\)

of \( \tau(x) \) when \( X \sim \mu \)

\[ M^{\tau(x)}(\eta) = \mathbb{E}_{\tilde{\eta}} \left[ e^{u'\tau(x)} \right] = \int e^{u'T_{\tilde{\eta}} - A(\eta)} \, h_{\mu} \, dm \]

\[ = e^{A(\eta + u) - A(\eta)} \int e^{(\eta + u)'T - A(\eta + u)} \, h_{\mu} \, dm = 1 \]