The Exponential Family of Distributions

$$p(x) = h(x)\, e^{\,\theta^\top T(x) - A(\theta)}$$

θ	vector of parameters
$T(x)$	vector of “suf£cient statistics”
$A(\theta)$	cumulant generating function
$h(x)$	

Key point: x and θ only “mix” in $e^{\,\theta^T T(x)}$

The Exponential Family of Distributions

$$p(x) = h(x)\, e^{\theta^\top T(x) - A(\theta)}$$

To get a normalized distribution, for any θ

$$\int p(x)\, dx = e^{-A(\theta)} \int h(x)\, e^{\theta^\top T(x)}\, dx = 1$$

so

$$e^{A(\theta)} = \int h(x)\, e^{\theta^\top T(x)}\, dx,$$

i.e., when $T(x) = x$, $A(\theta)$ is the $\log$ of Laplace transform of $h(x)$.

Examples

Gaussian	$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\| x-\mu \|^2/(2\sigma^2)}$	$x \in \mathbb{R}$
Bernoulli	$p(x) = \alpha^x (1-\alpha)^{1-x}$	$x \in \{0,1\}$
Binomial	$p(x) = \binom{n}{x} \alpha^x (1-\alpha)^{n-x}$	$x \in \{0,1,2,\ldots,n\}$
Multinomial	$p(x) = \frac{n!}{x_1!x_2!\ldots x_n!} \prod_{i=1}^n \alpha_i^{x_i}$	$x_i \in \{0,1,2,\ldots,n\}$, $\sum_i x_i = n$
Exponential	$p(x) = \lambda e^{-\lambda x}$	$x \in \mathbb{R}^+$
Poisson	$p(x) = \frac{e^{-\lambda}}{x!} \lambda^x$	$x \in \{0,1,2,\ldots\}$
Dirichlet	$p(x) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i x_i^{\alpha_i - 1}$	$x_i \in [0,1]$, $\sum_i x_i = 1$

(don't need to memorize these except for Gaussian)

Natural Parameter form for Bernoulli

$$p(x) = h(x)\, e^{\,\theta^\top T(x) - A(\theta)}$$

$$\begin{aligned}
p(x) &= \alpha^x\,(1-\alpha)^{1-x} \\
&= \exp\left[\,\log\!\left(\alpha^x\,(1-\alpha)^{1-x}\right)\,\right] \\
&= \exp\left[\,x\log\alpha + (1-x)\log(1-\alpha)\,\right] \\
&= \exp\left[\,x\log\frac{\alpha}{1-\alpha} + \log(1-\alpha)\right] \\
&= \exp\left[\,x\,\theta - \log\left(1+e^\theta\right)\,\right]
\end{aligned}$$

so

$$T(x) = x \qquad\qquad \theta = \log\frac{\alpha}{1-\alpha} \qquad\qquad A(\theta) = \log\left(1+e^\theta\right)$$

Natural Parameter Form for Gaussian

$$
\begin{aligned}
p(x) &= \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/(2\sigma^2)} \\
&= \frac{1}{\sqrt{2\pi}}\, \exp\left(-\log\sigma - \frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} - \frac{\mu^2}{2\sigma^2} \right) \\
&= \underbrace{\frac{1}{\sqrt{2\pi}}}_{h(x)}\, \exp\big(\theta^\top T(x) - \underbrace{\log\sigma - \mu^2/(2\sigma^2)}_{A(\theta)}\big)
\end{aligned}
$$

where

$$
T(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \qquad \theta = \begin{pmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{pmatrix} \qquad
\begin{aligned} A(\theta) &= \tfrac{\mu^2}{2\sigma^2} + \log\sigma \\ &= -\tfrac{[\theta]_1^2}{4[\theta]_2} - \tfrac{1}{2}\log\left(-2[\theta]_2\right) \end{aligned}
$$

Natural Parameter Form for Multivariate Gaussian

$$p(x) = h(x)\, e^{\theta^\top T(x) - A(\theta)}$$

$$p(x) = \frac{1}{(2\pi)^{D/2}\,|\,\Sigma\,|^{1/2}}\; e^{-(x-\mu)\Sigma^{-1}(x-\mu)/2}$$

$$h(x) = (2\pi)^{-D/2} \qquad T(x) = \begin{pmatrix} x \\ x\,x^\top \end{pmatrix} \qquad \theta = \begin{pmatrix} \Sigma^{-1}\mu \\ -\frac{1}{2}\Sigma^{-1} \end{pmatrix}$$

The £rst derivative of $A(\theta)$

$$A(\theta) = \log \underbrace{\left[\int h(x)\, e^{\theta^\top T(x)}\, dx \right]}_{Q(\theta)}$$

$$\begin{aligned}
\frac{dA(\theta)}{d\theta} &= \frac{1}{Q(\theta)} \frac{dQ(\theta)}{d\theta} = \frac{Q'(\theta)}{Q(\theta)} \\
&= \frac{\int h(x)\, e^{\theta^\top T(x)}\, T(x)\, dx}{\int h(x)\, e^{\theta^\top T(x)}\, dx} \\
&= \frac{\int h(x)\, e^{\theta^\top T(x) - A(\theta)}\, T(x)\, dx}{\int h(x)\, e^{\theta^\top T(x) - A(\theta)}\, dx} \\
&= \mathsf{E}_{p_\theta}\left[T(x) \right].
\end{aligned}$$

The second derivative of $A(\theta)$

$$A(\theta) = \underbrace{\log\left[\int h(x)\, e^{\theta^\top T(x)}\, dx\right]}_{Q(\theta)}$$

$$\begin{aligned}
\frac{dA(\theta)}{d\theta} &= \frac{d}{d\theta}\left[\frac{Q'(\theta)}{Q(\theta)}\right] = \frac{d}{d\theta}\left[Q'(\theta)\frac{1}{Q(\theta)}\right] = \frac{Q''(\theta)}{Q(\theta)} - \frac{\left(Q'(\theta)\right)^2}{\left(Q(\theta)\right)^2} \\
&= \frac{\int h(x)\, e^{\theta^\top T(x)}\, T^2(x)\, dx}{\int h(x)\, e^{\theta^\top T(x)}\, dx} - \left(\mathsf{E}_{p_\theta}\left[T(x)\right]\right)^2 \\
&= \frac{\int h(x)\, e^{\theta^\top T(x) - A(\theta)}\, T^2(x)\, dx}{\int h(x)\, e^{\theta^\top T(x) - A(\theta)}\, dx} - \left(\mathsf{E}_{p_\theta}\left[T(x)\right]\right)^2 \\
&= \mathsf{E}_{p_\theta}\left[T^2(x)\right] - \left(\mathsf{E}_{p_\theta}\left[T(x)\right]\right)^2 = \mathsf{Cov}_{p_\theta}\left[T(x)\right] \succeq 0.
\end{aligned}$$

$\Longrightarrow A(\theta)$ is convex. ($\succeq$ means positive de£nite)

Maximum Likelihood

$$\ell(\theta) = \sum_{i=1}^{N} \log p\,(\,x_i\,|\,\theta\,) = \sum_{i=1}^{N} \Big[\log h(x_i) + T(x_i) - A(\theta) \Big]$$

To £nd maxmimum likelihood solution

$$\ell'(\theta) = \left[\sum_{i=1}^{N} \theta^T T(x_i) \right] - N A'(\theta)$$

So ML solution satis£es

$$A'(\hat{\theta}_{ML}) = \frac{1}{N} \sum_{i=1}^{N} T(x_i) = 0$$

(is $\hat{\theta}_{\mathrm{ML}}$ a consistent estimator then ?)
Suf£cient statistics $\frac{1}{N}\sum_{i=1}^{N} T(x_i)$ summarize data.
When can't do this analytically: convexity $\Longrightarrow$ unique global ML solution for θ.

Products

Products of E-family distributions are E-family distributions

$$\left(h(x)\, e^{\theta_1^T T(x) - A(\theta_1)}\right) \times \left(h(x)\, e^{\theta_2^T T(x) - A(\theta_2)}\right) = \\ \tilde{h}(x)\, e^{(\theta_1 + \theta_2) T(x) - \tilde{A}(\theta_1, \theta_2)}$$

but might not have a nice parametric form any more.

But the product of two Gaussians is always a Gaussian.

Conjugate Priors in Bayesian Statistics

$$p\,(\,\theta\,|\,x\,) = \frac{p\,(\,x\,|\,\theta\,)\;p(\theta)}{\int p\,(\,x\,|\,\theta\,)\;p(\theta)\,d\theta}$$

Note: denominator not a function of $\theta \Rightarrow$ just normalizing term

$$\underbrace{p(\theta)}_{\text{parametric}} \quad\longrightarrow\quad \underbrace{p\,(\,x\,|\,\theta\,)}_{\text{parametric}}\,p(\theta) \quad\longrightarrow\quad p\,(\,\theta\,|\,x\,) \propto \underbrace{p\,(\,x\,|\,\theta\,)\,p(\theta)}_{\text{mess?}}$$

Conjugacy: require $p(\theta)$ and $p\,(\,\theta\,|\,x\,)$ to be of the same form. E.g.

$$\underbrace{p(\theta)}_{\text{Dirichlet}} \quad\longrightarrow\quad \underbrace{p\,(\,x\,|\,\theta\,)}_{\text{Multinomial}}\,p(\theta) \quad\longrightarrow\quad \underbrace{p\,(\,\theta\,|\,x\,)}_{\text{Dirichlet}}$$

$p(\theta)$ and $p\,(\,x\,|\,\theta\,)$ are then called **conjugate distributions.**

Example: Dirichlet and Multinomial

$$p(\theta) \;=\; \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma\left(\alpha_i\right)} \prod_i \theta^{\alpha_i - 1} \qquad \text{Dirichlet in } \theta \qquad \Gamma(x) = (x-1)!$$

$$p\,(\,x \,|\, \theta\,) \;=\; \frac{\left(\sum_i x_i\right)!}{x_1! x_2! \ldots x_n!} \prod_{i=1}^{n} \theta_i^{x_i} \qquad \text{Multinomial in } x$$

$$p\,(\,\theta \,|\, x\,) \;\propto\; p\,(\,\theta \,|\, x\,)\, p(\theta) \;=\; \text{junk} \times \prod_i \theta_i^{x_i + \alpha_i - 1}$$

which is again Dirichlet, so we must have

$$p\,(\,\theta \,|\, x\,) = \frac{\Gamma\left(\sum_i \alpha_i + x_i\right)}{\prod_i \Gamma\left(\alpha_i + x_i\right)} \prod_i \theta_i^{x_i + \alpha_i - 1}.$$

Remember pseudocount of 1? That was just a Dirichlet prior.

Conjugate Pairs

Prior		Conditional	
Gaussian	$e^{-\| \mu-\mu_0 \|^2/(2\sigma^2)}$	Gaussian	$e^{-\| x-\mu \|^2/(2\sigma^2)}$
Beta	$\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\, \alpha^{r-1}\, (2-\alpha)^{s-1}$	Bernoulli	$\alpha^x\, (1-\alpha)^{1-x}$
Dirichlet	$\frac{\Gamma(\sum \alpha_i)}{\prod \Gamma(\alpha_i)} \prod \theta_i^{\alpha_i - 1}$	Multinomial	$\frac{(\sum x_i)!}{\prod x_i!} \prod \theta_i^{x_i}$
Inv. Wishart		Gaussian (cov)	

Note: Conjugacy is mutual, e.g.

Dirichlet $\longrightarrow$ Multinomial $\longrightarrow$ Dirichlet

Multinomial $\longrightarrow$ Dirichlet $\longrightarrow$ Multinomial