Statistical Optimization: Lecture 5

Optimization for Exponential Family

Zijian Guo

Zhejiang University
Center for data science

March 24, 2026

Exponential Family

- Many common statistical models belong to the exponential family, including:
  - Binomial
  - Poisson
  - Normal
  - Multinomial
- We begin with the **one-parameter** case and then extend the framework to the **multi-parameter** case.
- Key properties of exponential family: the log-likelihood is convex.

One-Parameter Exponential Family: Definition

Definition The class of distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is said to belong to the **one-parameter exponential family** if there exist real-valued functions $\eta(\theta)$, $B(\theta)$, $T(x)$, and $h(x)$ such that

$$p(x \mid \theta) = h(x) \exp\big(\eta(\theta)\, T(x) - B(\theta)\big). \tag{1}$$

Common support property: If (1) holds and $h(x)$ does not depend on θ, then the support $\{x : p(x \mid \theta) > 0\}$ does not depend on θ. Indeed, $\exp(\eta(\theta)T(x) - B(\theta)) > 0$ always, so positivity of $p(x \mid \theta)$ is determined by $h(x)$.

Example: Bernoulli

If $X \sim \text{Bernoulli}(\theta)$ with $0 < \theta < 1$, then

$$\mathbb{P}(X = x \mid \theta) = \theta^x (1-\theta)^{1-x} = (1-\theta) \exp\left(x \log \frac{\theta}{1-\theta} \right), \qquad x = 0, 1. \tag{2}$$

Thus $h(x) = 1$, $T(x) = x$, $\eta(\theta) = \log \frac{\theta}{1-\theta}$, and

$$B(\theta) = -\log(1-\theta).$$

Example: Binomial

If $X \sim \text{Binomial}(n, \theta)$ with $0 < \theta < 1$, then

$$\mathbb{P}(X = x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} = \binom{n}{x} (1-\theta)^n \exp\left(x \log \frac{\theta}{1-\theta} \right), \qquad x = 0, 1, \ldots, n. \tag{3}$$

Thus $h(x) = \binom{n}{x}$, $T(x) = x$, $\eta(\theta) = \log \frac{\theta}{1-\theta}$, and

$$B(\theta) = -n \log(1-\theta).$$

Example: Poisson

If $X \sim \mathrm{Poisson}(\theta)$ with $\theta > 0$, then

$$\mathbb{P}(X = x \mid \theta) = \frac{\theta^x e^{-\theta}}{x!} = \frac{1}{x!} \exp\big(x \log \theta - \theta\big), \qquad x = 0, 1, 2, \ldots \tag{4}$$

Thus $h(x) = 1/x!$, $T(x) = x$, $\eta(\theta) = \log \theta$, and $B(\theta) = \theta$.

Canonical One-Parameter Exponential Family

With the **natural parameter** $\eta = \eta(\theta)$, we define the **canonical one-parameter exponential family**

$$p(x \mid \eta) = h(x) \exp\big(\eta\, T(x) - A(\eta)\big), \tag{5}$$

where the **log-partition function** $A(\eta)$ is determined by normalization:

$$A(\eta) = \log \left(\int h(x) \exp(\eta T(x))\, dx \right) \tag{6}$$

in the continuous case, and the integral is replaced by a sum in the discrete case.

Define the **natural parameter space**

$$\Xi = \{\eta : \ A(\eta) < \infty\}.$$

Example: Bernoulli in Canonical Form

For $X \sim \text{Bernoulli}(\theta)$, we have

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x}, \qquad x \in \{0, 1\}.$$

Rewrite it as

$$p(x \mid \theta) = \exp\left(x \log \frac{\theta}{1-\theta} + \log(1-\theta) \right).$$

Let $\eta = \log \frac{\theta}{1-\theta}$. Then $\theta = \frac{e^\eta}{1+e^\eta}, 1-\theta = \frac{1}{1+e^\eta}$. So the density can be written in canonical form as

$$p(x \mid \eta) = \exp(x\eta - \log(1+e^\eta)).$$

Hence $h(x) = 1,\ T(x) = x,\ A(\eta) = \log(1+e^\eta),\ \Xi = \mathbb{R}$.

Example: Poisson in Canonical Form

For $X \sim \mathrm{Poisson}(\theta)$, let $\eta = \log \theta$. Then

$$p(x \mid \eta) = \frac{1}{x!} \exp\left(x\eta - e^{\eta}\right), \qquad A(\eta) = e^{\eta}, \qquad \Xi = \mathbb{R}.$$

Maximum Likelihood Estimation (MLE) and Convexity

Let $X_1, \ldots, X_n$ be i.i.d. from (5). The likelihood is

$$L(\eta) = \prod_{i=1}^{n} p(x_i \mid \eta), \qquad \ell(\eta) = \log L(\eta) = \sum_{i=1}^{n} \log p(x_i \mid \eta).$$

$$\begin{aligned} -\ell(\eta) &= -\sum_{i=1}^{n} \big(\log h(x_i) + \eta T(x_i) - A(\eta)\big) \\ &= -\sum_{i=1}^{n} \log h(x_i) \;-\; \eta \sum_{i=1}^{n} T(x_i) \;+\; nA(\eta). \end{aligned} \tag{7}$$

Dropping the constant $-\sum_i \log h(x_i)$, the MLE is equivalent to

$$\min_{\eta \in \Xi} \Big\{ nA(\eta) - \eta \sum_{i=1}^{n} T(x_i) \Big\}. \tag{8}$$

Thus, if $A(\eta)$ is convex, then (8) is a convex optimization problem.

MGF and Moments $\Rightarrow$ Convexity

Theorem(MGF and moments) Consider the canonical one-parameter exponential family (5). Assume η is an interior point of Ξ. Then for all s in a neighborhood of 0 such that $\eta + s \in \Xi$, the moment generating function (MGF) of $T(X)$ exists and

$$M(s) = \mathbb{E}_\eta\left[e^{sT(X)}\right] = \exp\big(A(\eta + s) - A(\eta)\big). \tag{9}$$

Moreover,

$$\mathbb{E}_\eta[T(X)] = A'(\eta), \qquad \mathrm{Var}_\eta(T(X)) = A''(\eta) \ \geq 0, \tag{10}$$

so $A(\eta)$ is convex on Ξ.

Proof

Using (5),

$$M(s) = \int e^{sT(x)}\, h(x) \exp(\eta T(x) - A(\eta))\, dx = e^{-A(\eta)} \int h(x) \exp\big((\eta + s)T(x)\big)\, dx.$$

By (6), the integral equals $\exp(A(\eta + s))$ (whenever $\eta + s \in \Xi$), hence (9) follows. Differentiating (9) at $s = 0$ gives

$$M'(0) = \mathbb{E}_\eta[T(X)] = A'(\eta), \qquad M''(0) = \mathbb{E}_\eta[T(X)^2] = A''(\eta) + \big(A'(\eta)\big)^2,$$

so $\mathrm{Var}_\eta(T(X)) = M''(0) - M'(0)^2 = A''(\eta) \geq 0$.

Example: Strict Convexity in the Bernoulli Family

For the Bernoulli exponential family, with $x \in \{0, 1\}$,

$$p(x \mid \eta) = \exp\big(x\eta - A(\eta)\big), \qquad \Xi = \mathbb{R},$$

where

$$A(\eta) = \log(1 + e^{\eta}).$$

Its first derivative is

$$A'(\eta) = \frac{e^{\eta}}{1 + e^{\eta}}.$$

Its second derivative is

$$A''(\eta) = \frac{e^{\eta}}{(1 + e^{\eta})^2} > 0, \qquad \forall \eta \in \mathbb{R}.$$

Therefore, $A : \Xi \to \mathbb{R}$ is strictly convex.

Multi-parameter Exponential Family

Definition The family $\{P_\theta : \theta \in \Theta\}$ is said to be a k-parameter exponential family if there exist real-valued functions $\eta_j(\theta)$, $B(\theta)$, $T_j(x)$, and $h(x)$ ($j = 1, \ldots, k$) such that

$$p(x \mid \theta) = h(x) \exp\left(\sum_{j=1}^{k} \eta_j(\theta)\, T_j(x) - B(\theta) \right). \tag{11}$$

Example: Normal distribution

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$. The joint density is

$$\begin{aligned} p(x \mid \mu, \sigma^2) &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i - \frac{n\mu^2}{2\sigma^2}\right\}. \end{aligned} \tag{12}$$

Thus it is a 2-parameter exponential family with sufficient statistics $T_1(x) = \sum_{i=1}^n x_i$ and $T_2(x) = \sum_{i=1}^n x_i^2$.

Example: Multinomial (counts form)

Let $Y = (Y_1, \ldots, Y_k) \sim \mathrm{Multinomial}(n; p_1, \ldots, p_k)$ with $\sum_{i=1}^k Y_i = n$ and $\sum_{i=1}^k p_i = 1$. Then

$$\mathbb{P}(Y = y \mid p) = \frac{n!}{\prod_{i=1}^k y_i!} \prod_{i=1}^k p_i^{y_i}.$$

Take $\eta_i = \log(p_i/p_k)$ for $i = 1, \ldots, k-1$ and $T_i(y) = y_i$. Then this is a canonical $(k-1)$-parameter exponential family with base measure $h(y) = \frac{n!}{\prod_{i=1}^k y_i!}$ and

$$A(\eta) = n \log \left(1 + \sum_{i=1}^{k-1} e^{\eta_i} \right).$$

Canonical Multi-parameter Form

Similarly, the canonical form is obtained by setting $\eta = (\eta_1(\theta), \ldots, \eta_k(\theta))$:

$$p(x \mid \eta) = h(x) \exp\big(\langle \eta, T(x) \rangle - A(\eta)\big), \tag{13}$$

where $T(x) = (T_1(x), \ldots, T_k(x))$ and

$$A(\eta) = \log \left(\int h(x) \exp(\langle \eta, T(x) \rangle)\, dx \right), \tag{14}$$

with sum/integral depending on the model.

Define the **natural parameter space** $\Xi = \{\eta : \ A(\eta) < \infty\}$.

Basic Properties in the Multi-parameter Case

Theorem For the canonical multi-parameter exponential family (13):

(a) Ξ is convex;

(b) $A : \Xi \to \mathbb{R}$ is convex;

(c) If η_0 is an interior point of Ξ, then for any s with $\eta_0 + s \in \Xi$,

$$\mathbb{E}_{\eta_0}\left[e^{\langle s, T(X)\rangle}\right] = \exp\big(A(\eta_0 + s) - A(\eta_0)\big),$$

and (whenever differentiation is justified)

$$\nabla A(\eta) = \mathbb{E}_\eta[T(X)], \qquad \nabla^2 A(\eta) = \mathrm{Cov}_\eta(T(X)) \succeq 0.$$

Lemma: Hölder's Inequality

Lemma(Hölder's Inequality) For $u(x), v(x), h(x) \geq 0$ and $r, s > 1$ with $\frac{1}{r} + \frac{1}{s} = 1$,

$$\int u(x)v(x)h(x)\,dx \leq \left(\int u(x)^r h(x)\,dx\right)^{1/r} \left(\int v(x)^s h(x)\,dx\right)^{1/s}.$$

Remark:(Cauchy-Schwarz). The special case $r = s = 2$ gives the Cauchy-Schwarz inequality.

Proof of (a)–(b) (1/2)

Let $\eta_1, \eta_2 \in \Xi$ and $\alpha \in [0,1]$. Consider

$$\exp(A(\alpha\eta_1 + (1-\alpha)\eta_2)) = \int h(x) \exp\big(\langle \alpha\eta_1 + (1-\alpha)\eta_2,\ T(x)\rangle\big)\, dx.$$

Rewrite the integrand as

$$h(x) \underbrace{\exp(\langle \eta_1, T(x)\rangle)^{\alpha}}_{u(x)} \underbrace{\exp(\langle \eta_2, T(x)\rangle)^{1-\alpha}}_{v(x)}.$$

Apply Hölder with $r = 1/\alpha$ and $s = 1/(1-\alpha)$ (interpret endpoints by continuity):

$$\exp(A(\alpha\eta_1+(1-\alpha)\eta_2)) \le \left(\int h(x) \exp(\langle \eta_1, T(x)\rangle)\, dx\right)^{\alpha} \left(\int h(x) \exp(\langle \eta_2, T(x)\rangle)\, dx\right)^{1-\alpha}.$$

Proof of (a)-(b) (2/2)

Taking logs yields

$$A(\alpha\eta_1 + (1-\alpha)\eta_2) \leq \alpha A(\eta_1) + (1-\alpha)A(\eta_2).$$

Thus A is convex on Ξ.

Since the right-hand side is finite, the left-hand side is finite as well. Hence

$$\alpha\eta_1 + (1-\alpha)\eta_2 \in \Xi,$$

so Ξ is convex.

Proof of (c) (1/3)

Fix an interior point $\eta_0 \in \Xi$. For any $s \in \mathbb{R}^k$ such that $\eta_0 + s \in \Xi$, using (13), we have

$$\mathbb{E}_{\eta_0}\left[e^{\langle s, T(X)\rangle}\right] = \int e^{\langle s, T(x)\rangle}\, h(x) \exp\big(\langle \eta_0, T(x)\rangle - A(\eta_0)\big)\, dx.$$

Hence

$$\mathbb{E}_{\eta_0}\left[e^{\langle s, T(X)\rangle}\right] = e^{-A(\eta_0)} \int h(x) \exp\big(\langle \eta_0 + s, T(x)\rangle\big)\, dx.$$

By the definition of $A(\eta)$ in (14),

$$\int h(x) \exp\big(\langle \eta_0 + s, T(x)\rangle\big)\, dx = \exp\big(A(\eta_0 + s)\big),$$

whenever $\eta_0 + s \in \Xi$. Therefore,

$$\mathbb{E}_{\eta_0}\left[e^{\langle s, T(X)\rangle}\right] = \exp\big(A(\eta_0 + s) - A(\eta_0)\big).$$

Proof of (c) (2/3)

Now write

$$A(\eta) = \log \int h(x) \exp\big(\langle \eta, T(x) \rangle\big)\, dx.$$

Whenever differentiation under the integral sign is justified, for each $j = 1, \ldots, k$,

$$\frac{\partial A(\eta)}{\partial \eta_j} = \frac{\int T_j(x)\, h(x) \exp\big(\langle \eta, T(x) \rangle\big)\, dx}{\int h(x) \exp\big(\langle \eta, T(x) \rangle\big)\, dx} = \mathbb{E}_\eta[T_j(X)].$$

Thus,

$$\nabla A(\eta) = \mathbb{E}_\eta[T(X)].$$

Proof of (c) (3/3)

Differentiating once more, for $j, \ell = 1, \dots, k$,

$$\frac{\partial^2 A(\eta)}{\partial \eta_j \, \partial \eta_\ell} = \mathbb{E}_\eta[T_j(X) T_\ell(X)] - \mathbb{E}_\eta[T_j(X)] \, \mathbb{E}_\eta[T_\ell(X)].$$

Hence

$$\nabla^2 A(\eta) = \mathrm{Cov}_\eta(T(X)).$$

In particular, for any $v \in \mathbb{R}^k$,

$$v^\top \nabla^2 A(\eta) v = \mathrm{Var}_\eta\big(v^\top T(X)\big) \geq 0,$$

so

$$\nabla^2 A(\eta) = \mathrm{Cov}_\eta(T(X)) \succeq 0.$$