Review: Exponential Family

\[ T(x) = (T_1(x), T_2(x), \ldots, T_k(x)) \]

An exponential family is a set of distributions with densities of the form

\[ p(x) = h(x) \exp(\theta^T T(x) - A(\theta)) \]

- \( \theta \): “(natural) parameters”
- \( T(x) \): “sufficient statistics”
- \( A(\theta) \): “log-partition function”
- \( h(x) \): “base measure” (usually identically one for us)

Covers many families: Bernoulli, Multinomial, Poisson, Gaussian, MRFs, …
Properties of Log-Partition Function

The log-partition function $A(\theta)$ has two critical properties that relate its derivatives to moments (expectations) of the sufficient statistics $T(X)$.

$$\text{derivative of } A(\theta) \iff \mathbb{E} [\text{function of } T(X)]$$

First Derivative of $A(\theta) \equiv$ First Moment of $T(X)$

$$A(\theta) = \log Z(\theta)$$

$$\frac{\partial}{\partial \theta} A(\theta) = \mathbb{E}_{p_\theta}[T(X)]$$

**Proof**: (assume $h(x) \equiv 1$)

$$\frac{2}{\theta} \log \sum_x \exp(\theta^T T(x)) = \frac{1}{\mathbb{E}_x} \exp(\theta^T T(x)) \frac{\partial}{\partial \theta} \sum_x \exp(\theta^T T(x))$$

$$= \mathbb{E}_x \exp(\theta^T T(x)) \frac{\partial}{\partial \theta} (\exp(\theta^T T(x)))$$

$$= \sum_x \frac{\partial}{\partial \theta} \exp(\theta^T T(x)) = \frac{\partial}{\partial \theta} \log Z(\theta), \quad T(x)$$

$$= \sum_x p_\theta(x) T(x)$$

$$= \mathbb{E}_{p_\theta}[T(X)]$$

Second Derivative of $A(\theta) \equiv$ Second Moment of $T(X)$

$$T(x) = (T_1(x), \ldots, T_d(x))$$

$$\text{Hessian} = \frac{\partial^2}{\partial \theta \partial \theta^T} A(\theta) = \text{Var}_{p_\theta}[T(X)] = \sum$$

**Notation**: $\frac{\partial^2}{\partial \theta \partial \theta^T} A(\theta)$ is the Hessian matrix of $A(\theta)$. The $(i, j)$th entry is $\frac{\partial^2}{\partial \theta_i \partial \theta_j} A(\theta)$.

**Proof**: algebra

**Important consequence**: $A(\theta)$ is convex

- Variance is PSD $\iff$ Hessian is PSD $\iff$ $A(\theta)$ is convex
Log-Likelihood

The average log-likelihood in an exponential family is

\[ L(\theta) = \frac{1}{N} \sum_{n=1}^{N} \log p(x^{(n)}) = \frac{1}{N} \sum_{n=1}^{N} \left( \theta^T T(x^{(n)}) - A(\theta) \right) \]

\[ = \frac{1}{N} \sum_{n=1}^{N} T(x^{(n)}) - A(\theta) + \frac{1}{N} \sum_{n=1}^{N} \log h(x^{(n)}) \]

\[ \text{sufficient statistics} \]

- All we need to know about the data for estimation is the average value of \( T(x^{(n)}) \), i.e., the "sufficient statistics"

Concavity of Log-Likelihood

\[ \frac{\partial^2}{\partial \theta \partial \theta^T} L(\theta) = -\frac{\partial^2}{\partial \theta \partial \theta^T} A(\theta) \]

The log-likelihood is concave

- every zero-gradient point is a global optimum
- the moment-matching conditions are necessary and sufficient for optimality

Moment-Matching

At the maximum-likelihood parameters, \( \frac{\partial}{\partial \theta} L(\theta) = 0 \)

\[ 0 = \frac{\partial}{\partial \theta} \left( \theta^T T(x^{(n)}) - A(\theta) + \text{const} \right) \]

\[ = \frac{1}{N} \sum_{n=1}^{N} T(x^{(n)}) - E_{p_{\theta}}[T(X)] \]

\[ \implies \text{at maximum-likelihood parameters, we have the moment-matching conditions:} \]

\[ E_{p_{\theta}}[T(X)] = \frac{1}{N} \sum_{n=1}^{N} T(x^{(n)}) = E[T(X)] \]

- "model expectation equals data expectation"
- sometimes we can easily solve for the maximum-likelihood parameters; other times numerical routines are needed

Summary So Far

- \( p_{\theta}(x) = h(x) \exp(\theta^T T(x) - A(\theta)) \)
- Bernoulli, normal, Poisson, MRF, . . .
- First property: \( \frac{\partial}{\partial \theta} A(\theta) = E_{p_{\theta}}[T(X)] \)
- Second property: \( \frac{\partial^2}{\partial \theta \partial \theta^T} A(\theta) = \text{Var}_{p_{\theta}}[T(X)] \)
- Likelihood: \( L(\theta) = \theta^T T - A(\theta) + \text{const} \) where \( T = \frac{1}{N} \sum_{n=1}^{N} T(x^{(n)}) \) are the average sufficient statistics over the data
- \( L(\theta) \) is concave
- Moment-matching conditions are necessary and sufficient for parameters \( \theta \) to maximize the likelihood: \( E_{p_{\theta}}[T(X)] = T = E[T(X)] \)
Pairwise MRFs as an Exponential Family

Consider the chain model on $x_1, x_2, x_3, x_4 \in \{0, 1\}$:

$$p(x) = \frac{\phi_{1,2}(x_1, x_2) \phi_{2,3}(x_2, x_3) \phi_{3,4}(x_3, x_4)}{Z}$$

The log-density is

$$\log p(x) = \log \phi_{1,2}(x_1, x_2) + \log \phi_{2,3}(x_2, x_3) + \log \phi_{3,4}(x_3, x_4) - \log Z$$

This is an exponential family with

$$T(x) = \left\{ I[x_1 = a, x_2 = b] \right\}_{(i,j) \in E, a \in \text{Val}(X_i), b \in \text{Val}(X_j)}$$

$$\theta = \left( \theta^{ab}_{ij} \right)_{(i,j) \in E, a \in \text{Val}(X_i), b \in \text{Val}(X_j)}$$

$$\log p_{\theta}(x) = \theta^T x - A(\theta) = \left( \sum_{(i,j) \in E} \sum_{a \in \text{Val}(X_i)} \sum_{b \in \text{Val}(X_j)} \theta^{ab}_{ij} \cdot I[x_i = a, x_j = b] \right) - A(\theta)$$

The final three lines are accurate for general pairwise MRFs.
Moment-Matching for Pairwise-MRFs

If we apply the moment-matching conditions to pairwise MRFs, we recover our previous result. At the maximum-likelihood parameters:

\[ E_{p^{\theta}}[T(X)] = \hat{E}[T(X)], \]

\[ E_{p^{\theta}}[I|X_i = a, X_j = b] = \hat{E}[I|X_i = a, X_j = b] \quad \forall (i, j) \in E, a, b, \]

\[ P_{p^{\theta}}(X_i = a, X_j = b) = \frac{\#(X_i = a, X_j = b)}{N} \quad \forall (i, j) \in E, a, b, \]

(we still have to solve for \( \theta \) numerically; recall that the RHS minus the LHS is the gradient of \( L(\theta) \))

Moment-Matching for Gaussians

For a normal distribution, we had \( T(x) = (x^2, x) \)

\[ \log p_{\mu, \sigma^2}(x) = x^2 \cdot \frac{-1}{2\sigma^2} + x \cdot \frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \log(\sqrt{2\pi\sigma^2}) \]

We know \( E_{p^{\theta}}[X] = \mu \) and \( E_{p^{\theta}}[X^2] = \mu^2 + \sigma^2 \).

Moment-matching says the max-likelihood parameters satisfy:

\[ E_{p^{\theta}}[X] = \bar{E}[X] \quad \implies \quad \mu = \bar{E}[X] \]

\[ E_{p^{\theta}}[X^2] = \bar{E}[X^2] \quad \implies \quad \mu^2 + \sigma^2 = \bar{E}[X^2] \]

\[ \implies \quad \sigma^2 = \bar{E}[X^2] - \mu^2 \]

We can easily solve for the maximum-likelihood \( \mu, \sigma^2 \).