COMPSCI 688: Probabilistic Graphical Models

Lecture 12: Learning in Exponential Families

Dan Sheldon

Manning College of Information and Computer Sciences
University of Massachusetts Amherst

Partially based on materials by Benjamin M. Marlin (marlin@cs.umass.edu) and Justin Domke (domke@cs.umass.edu)

Exponential Families

Exponential Families

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

$$p_\theta(x) = h(x)\exp(\theta^\top T(x) - A(\theta))$$

- θ: "(natural) parameters"
- $T(x)$: "sufficient statistics"
- $A(\theta)$: "log-partition function"
- $h(x)$: "base measure" (we'll usually ignore)

Interpretation $(h(x) = 1)$

$$p_\theta(x) = \exp(\theta^\top T(x) - A(\theta))$$

- $\theta^\top T(x)$ is a real-valued "score" (positive or negative), defined in terms of "features" $T(x)$ and parameters θ
- $\exp(\theta^\top T(x))$ is an unnormalized probability
- The log-partition function $A(\theta) = \log Z(\theta)$ ensures normalization

$$p_\theta(x) = \frac{\exp(\theta^\top T(x))}{\exp(A(\theta))}, \quad A(\theta) = \log Z(\theta) = \log \int \exp(\theta^\top T(x))dx$$

- Valid parameters are the ones for which the integral for $A(\theta)$ is finite.

Applications and Importance

- We can get *many* different families of distributions by selecting different "features" $T(x)$ for a variable x in some sample space:
  - Bernoulli, Binomial, Multinomial, Beta, Gaussian, Poisson, MRFs, ...
- There is a general theory that covers learning and other properties of all of these distributions!
- A good trick to seeing that a distribution belongs to an exponential family is to match its log-density to

$$\log p_\theta(x) = \log h(x) + \theta^\top T(x) - A(\theta)$$

Preview: Graphical Models

For some intuition why exponential families could be relevant for graphical models, observe that the unnormalized probability factors over "simpler" functions, just like graphical models:

$$\exp(\theta^\top T(x)) = \exp \sum_i \theta_i T_i(x) = \prod_i \exp(\theta_i T_i(x))$$

(Think: what could $T(x)$ look like to recover a graphical model?)

Example: Bernoulli Distribution

The Bernoulli distribution with parameter $\mu \in [0, 1]$ has density (pmf)

$$p_\mu(x) = \begin{cases} \mu & x = 1 \\ 1 - \mu & x = 0 \end{cases}$$

One way to write the log-density is

$$\log p_\mu(x) = \mathbb{I}[x = 1] \log \mu + \mathbb{I}[x = 0] \log(1 - \mu)$$

To match this to an exponential family

$$\log p_\theta(x) = \log h(x) + \theta^\top T(x) - A(\theta),$$

Review: Bernoulli Distribution

To match this to an exponential family $\log p_\theta(x) = \log h(x) + \theta^\top T(x) - A(\theta)$, take

- $h(x) = 1$
- $T(x) = (\mathbb{I}[x = 1], \mathbb{I}[x = 0])$
- $\theta = (\log \mu, \log(1 - \mu))$
- $\exp(\theta^\top T(x)) = \begin{cases} e^{\theta_1} & x = 1 \\ e^{\theta_2} & x = 0 \end{cases}$
- $A(\theta) = \log(e^{\theta_1} + e^{\theta_2})$
- It's easy to check that $A(\theta) = 0$ when $\theta = (\log \mu, \log(1 - \mu))$

Example: Bernoulli, Single Parameter

We can also write the Bernoulli as a single-parameter exponential family. Rewrite the log-density as

$$\log p_\mu(x) = \log(1 - \mu) + x \log \frac{\mu}{1 - \mu}$$

Review: Bernoulli, Single Parameter

- $h(x) = 1$
- $T(x) = \mathbb{I}[x = 1] = x$
- $\theta = \log \frac{\mu}{1-\mu}$
- $\exp(\theta^\top x) = \begin{cases} e^{\theta} & x = 1 \\ 1 & x = 0 \end{cases}$
- $A(\theta) = \log(1 + e^\theta)$
- It's easy to check that $\log(1 + e^\theta) = -\log(1 - \mu)$ when $\theta = \log \frac{\mu}{1-\mu}$

Example: Normal Distribution

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2\sigma^2}(x - \mu)^2\right)$$

Review: Normal Distribution

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-\tfrac{1}{2\sigma^2}(x-\mu)^2\big)$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-\tfrac{1}{2\sigma^2}(x^2 - 2x\mu + \mu^2)\big)$$

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

- $h(x) = 1$
- $T(x) = (x^2, x)$
- $\theta = (\frac{-1}{2\sigma^2}, \frac{\mu}{\sigma^2})$
- $A(\theta) = \log \int \exp(x^2\theta_1 + x\theta_2)dx = \ldots = \frac{\mu^2}{2\sigma^2} + \log(\sqrt{2\pi\sigma^2})$

Note: we need $\theta_1 < 0$; why?

Properties of Exponential Families

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 $T(X)$.

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

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

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

$$\frac{\partial}{\partial\theta} \log \sum_x \exp(\theta^\top T(x)) = \frac{1}{\sum_x \exp(\theta^\top T(x))} \frac{\partial}{\partial\theta} \sum_x \exp(\theta^\top T(x))$$
$$= \frac{1}{Z(\theta)} \sum_x \exp(\theta^\top T(x)) \frac{\partial}{\partial\theta} \theta^\top T(x)$$
$$= \sum_x \frac{\exp(\theta^\top T(x))}{Z(\theta)} \cdot T(x)$$
$$= \sum_x p_\theta(x) \cdot T(x)$$
$$= \mathbb{E}_{p_\theta}[T(X)]$$

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

$$\boxed{\frac{\partial^2}{\partial\theta\,\partial\theta^\top} A(\theta) = \mathrm{Var}_{p_\theta}[T(X)]}$$

Notation: $\frac{\partial^2}{\partial\theta\,\partial\theta^\top} A(\theta)$ 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 $\implies$ Hessian is PSD $\implies$ A convex

Learning in Exponential Families

Log-Likelihood

The average log-likelihood in an exponential family is

$$\begin{aligned}
\mathcal{L}(\theta) &= \frac{1}{N}\sum_{n=1}^{N} \log p_\theta(x^{(n)}) \\
&= \frac{1}{N}\sum_{n=1}^{N} \left(\theta^\top T(x^{(n)})\right) - A(\theta) + \frac{1}{N}\sum_{n=1}^{N} \log h(x^{(n)}) \\
&= \theta^\top \Big(\underbrace{\frac{1}{N}\sum_{n=1}^{N} T(x^{(n)})}_{\text{"sufficient statistics"}} \Big) - A(\theta) + \text{const}
\end{aligned}$$

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

Moment-Matching

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

$$\begin{aligned}
\frac{\partial}{\partial\theta}\mathcal{L}(\theta) &= \frac{\partial}{\partial\theta}\left(\theta^\top \Big(\frac{1}{N}\sum_{n=1}^{N} T(x^{(n)})\Big) - A(\theta)\right) \\
&= \frac{1}{N}\sum_{n=1}^{N} T(x^{(n)}) - \mathbb{E}_{p_\theta}[T(X)] = 0
\end{aligned}$$

$\implies$ at maximum-likelihood parameters, we have the *moment-matching conditions*:

$$\boxed{\mathbb{E}_{p_\theta}[T(X)] = \frac{1}{N}\sum_{n=1}^{N} T(x^{(n)}) =: \hat{\mathbb{E}}[T(X]}$$

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

Concavity of Log-Likelihood

$$\mathcal{L}(\theta) = \underbrace{\theta^\top \Big(\frac{1}{N} \sum_{n=1}^{N} T(x^{(n)}) \Big)}_{\text{linear in } \theta} - \underbrace{A(\theta)}_{\text{convex}} + \text{const}$$

The log-likelihood is concave

$\implies$ every zero-gradient point is a global optimum

$\implies$ the moment-matching conditions are necessary and sufficient for optimality

Summary So Far

- $p_\theta(x) = h(x) \exp(\theta^\top T(\mathbf{x}) - A(\theta))$
- Bernoulli, normal, Poisson, MRF, ...
- First property: $\frac{\partial}{\partial \theta} A(\theta) = \mathbb{E}_{p_\theta}[T(X)]$
- Second property: $\frac{\partial^2}{\partial \theta \, \partial \theta^\top} A(\theta) = \mathrm{Var}_{p_\theta}[T(X)]$
- Likelihood: $\mathcal{L}(\theta) = \theta^\top \overline{T} - A(\theta) + \text{const}$ where $\overline{T} = \frac{1}{N} \sum_{n=1}^{N} T(x^{(n)})$ are the average sufficient statistics over the data
- $\mathcal{L}(\theta)$ is concave
- Moment-matching conditions are necessary and sufficient for parameters θ to maximize the likelihood: $\mathbb{E}_{p_\theta}[T(X)] = \overline{T} = \hat{\mathbb{E}}[T(X)]$

Pairwise MRFs as an Exponential Family

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

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

Pairwise MRFs as an Exponential Family: Review

The log-density is

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

$$\begin{aligned} &= \log \phi_{1,2}(0,0) \cdot \mathbb{I}[x_1 = 0, x_2 = 0] + \log \phi_{1,2}(0,1) \cdot \mathbb{I}[x_1 = 0, x_2 = 1] \\ &\quad + \log \phi_{1,2}(1,0) \cdot \mathbb{I}[x_1 = 1, x_2 = 0] + \log \phi_{1,2}(1,1) \cdot \mathbb{I}[x_1 = 1, x_2 = 1] \\ &+ \log \phi_{2,3}(0,0) \cdot \mathbb{I}[x_2 = 0, x_3 = 0] + \dots \\ &+ \log \phi_{3,4}(0,0) \cdot \mathbb{I}[x_3 = 0, x_4 = 0] + \dots \\ &- \log Z \end{aligned}$$

This is an exponential family with

$$\begin{aligned} T(\mathbf{x}) = \Big(&\mathbb{I}[x_1 = 0, x_2 = 0], \quad \dots, \quad \mathbb{I}[x_1 = 1, x_2 = 1], \\ &\mathbb{I}[x_2 = 0, x_3 = 0], \quad \dots, \quad \mathbb{I}[x_2 = 1, x_3 = 1], \\ &\mathbb{I}[x_3 = 0, x_4 = 0], \quad \dots, \quad \mathbb{I}[x_3 = 1, x_4 = 1]\Big) \end{aligned}$$

$$T(\mathbf{x}) = \Big(\mathbb{I}[x_i = a, x_j = b]\Big)_{(i,j)\in E,\, a\in \mathrm{Val}(X_i),\, b\in \mathrm{Val}(X_j)}$$

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

$$\log p_\theta(\mathbf{x}) = \theta^\top \mathbf{x} - A(\theta) = \Big(\sum_{(i,j)\in E} \sum_{a\in \mathrm{Val}(X_i)} \sum_{b\in \mathrm{Val}(X_j)} \theta_{ij}^{ab} \cdot \mathbb{I}[x_i = a, x_j = b]\Big) - A(\theta)$$

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:

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

$$\mathbb{E}_{p_\theta}\left[\mathbb{I}[X_i = a, X_j = b]\right] = \hat{\mathbb{E}}[\mathbb{I}[X_i = a, X_j = b]] \quad \forall (i,j) \in E, a, b,$$

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

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

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 $\mathbb{E}_{p_\theta}[X] = \mu$ and $\mathbb{E}_{p_\theta}[X^2] = \mu^2 + \sigma^2$.

Moment-matching says the max-likelihood parameters satisfy:

$$\begin{aligned} \mathbb{E}_{p_\theta}[X] = \hat{\mathbb{E}}[X] \quad &\Longrightarrow \quad \mu = \hat{\mathbb{E}}[X] \\ \mathbb{E}_{p_\theta}[X^2] = \hat{\mathbb{E}}[X^2] \quad &\Longrightarrow \quad \mu^2 + \sigma^2 = \hat{\mathbb{E}}[X^2] \\ &\Longrightarrow \quad \sigma^2 = \hat{\mathbb{E}}[X^2] - \mu^2 \end{aligned}$$

We can easily solve for the maximum-likelihood μ, σ^2.