THE EXPONENTIAL FAMILY

COMP 4680/8650
Justin Domke
comp8650@anu.edu.au

Reading: Murphy, §9.1-9.2.4, 9.2.6

Definition

An exponential family is a set of distributions

$$p(x|\theta) = \frac{1}{Z(\theta)} h(x) \exp\left(\theta^T \phi(x)\right)$$
$$= h(x) \exp\left(\theta^T \phi(x) - A(\theta)\right)$$

parameterized by $\theta \in \Theta \subseteq \mathbb{R}^d$.

$Z(\theta) = \int h(x) \exp(\theta^T \phi(x)) dx$ if continuous and
$Z(\theta) = \sum_x h(x) \exp(\theta^T \phi(x))$ if discrete.

$A(\theta) = \log Z(\theta)$ is the “log-partition function”.

Here, “zustandssumme” = “sum over states”

Why care about the Exponential Family?

- Gaussian, Poisson, Exponential, and (our motivation) Markov Random Fields are all exponential families.

- Share many properties, best understood through exponential family:

  - Maximum-likelihood, and connections to moment matching.
  - Idea of “sufficient statistics”.
  - Maximum entropy interpretation.
  - Basis for variational inference.

Bernoulli Distribution

- For binary x, Bernoulli is written

$$\text{Ber}(x|\mu) = \mu^x(1-\mu)^{1-x}$$

- We can write this as

$$\text{Ber}(x|\theta) = \exp(\phi(x)^T\theta - A(\theta))$$

$$\phi(x) = [\mathbb{I}(x=0), \mathbb{I}(x=1)]$$

- Same distribution when $\theta = [\log(\mu), \log(1-\mu)]$

- What happens if we add a constant to θ ?

Gaussian Distribution

- Typically written as $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$
- We can write this as $\mathcal{N}(x|\mu,\sigma^2) = \exp(\theta^T\phi(x) - A(\theta))$ by defining $\phi(x) = [x, x^2]$.
  - Same distribution obtained when $\theta = [\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}]$.
  - What happens if $\theta_2 > 0$?
    - Must constrain $\theta \in \Theta = \{\theta \in \mathbb{R}^2 | \theta_2 < 0\}$.
  - Incidentally, $A(\theta) = -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2) - \frac{1}{2}\log(2\pi)$.

Ising Model

- We wrote this before as

$$p(x|b,w) = \frac{1}{Z}\prod_{s\in V}\exp(b_s x_s)\prod_{(s,t)\in E}\exp(w_{st}x_s x_t)$$

- Write instead as

$$p(x|\theta) = \exp(\theta^T\phi(x) - A(\theta))$$

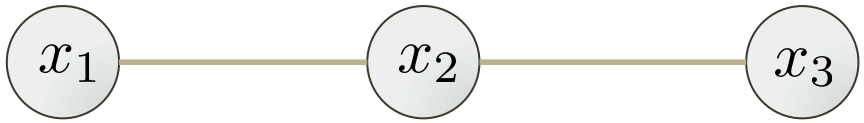
$$\phi(x) = \{x_s|s\in V\}\cup\{x_s x_t|(s,t)\in E\}$$

- Same distribution when

$$\theta = [b_1, b_2, ..., b_n, w_{1,2}, ..., w_{n-1,n}]$$

Three node Ising Model Example

$$p(x|b,w) = \frac{1}{Z}\exp(b_1 x_1)\exp(b_2 x_2)\exp(b_3 x_3)$$
$$\times \exp(w_{1,2} x_1 x_2)\exp(w_{2,3} x_2 x_3)$$



- Equivalent to

$$p(x|\theta) = \exp(\theta^T \phi(x) - A(\theta))$$

$$\phi(x) = [x_1,\ x_2,\ x_3,\ x_1 x_2,\ x_2 x_3]$$

Markov Random Field

- Typically written as

$$p(x|\theta) = \frac{1}{Z(\theta)} \prod_{c \in \mathcal{C}} \psi_c(x_c|\theta_c)$$

- Re-write as

$$p(x|\theta) = \exp(\theta^T \phi(x) - A(\theta))$$

$$\phi(x) = \{\mathbb{I}(x_c = x_c^*) | c \in \mathcal{C}, \text{all possible } x_c^*\}$$

Four node MRF Example

- Assume x is binary

$$p(x|\theta) = \frac{1}{Z(\theta)}\psi(x_{1,2}|\theta_{1,2})\psi(x_{2,3}|\theta_{2,3})\psi(x_{3,4}|\theta_{3,4})$$

- Equivalent to $p(x|\theta) = \exp(\theta^T\phi(x) - A(\theta))$ with

$$\phi(x) = [\mathbb{I}(x_1 = 0, x_2 = 0), \mathbb{I}(x_1 = 0, x_2 = 1), \mathbb{I}(x_1 = 1, x_2 = 0), ..., \mathbb{I}(x_3 = 1, x_4 = 1)]$$

- Question: How long is $\phi(x)$?

- Question: Is $\phi(x)$ redundant?

Derivatives of A

$$p(x|\theta) = h(x) \exp\left(\theta^T \phi(x) - A(\theta)\right)$$

- A seemingly "magical" property of A is that it's derivatives are the cumulants of the distribution.

$$\frac{dA}{d\theta} = \mathbb{E}_p[\phi(x)] \qquad \frac{dA}{d\theta d\theta^T} = \mathbb{V}_p[\phi(x)]$$

- People often derive these properties for particular distributions.
  - You shouldn't make this mistake!

Proof that $\frac{dA}{d\theta} = \mathbb{E}_p[\phi(x)]$

$$
\begin{aligned}
\frac{dA}{d\theta} &= \frac{d}{d\theta} \log \int_x h(x) \exp(\theta^T \phi(x)) dx \\
&= \frac{1}{\int_x h(x) \exp(\theta^T \phi(X)) dx} \frac{d}{d\theta} \int_x h(x) \exp(\theta^T \phi(x)) dx \\
&= \frac{1}{Z(\theta)} \frac{d}{d\theta} \int_x h(x) \exp(\theta^T \phi(x)) dx \\
&= \frac{1}{Z(\theta)} \int_x h(x) \frac{d}{d\theta} \exp(\theta^T \phi(x)) dx \\
&= \frac{1}{Z(\theta)} \int_x h(x) \exp(\theta^T \phi(x)) \frac{d}{d\theta} \theta^T \phi(x) dx \\
&= \frac{1}{Z(\theta)} \int_x h(x) \exp(\theta^T \phi(x)) \phi(x) dx \\
&= \int_x p(x|\theta) \phi(X) dx
\end{aligned}
$$

Proof that $\frac{dA}{d\theta d\theta^T} = \mathbb{V}_p[\phi(x)]$

$$
\begin{aligned}
\frac{dA}{d\theta d\theta^T} &= \frac{d}{d\theta^T}\frac{d}{d\theta}A(\theta) \\
&= \frac{d}{d\theta^T}\int_x p(x|\theta)\phi(x)dx \\
&= \int_x \frac{d}{d\theta^T}p(x|\theta)\phi(x)dx \\
&= \int_x p(x|\theta)\phi(x)\frac{d}{d\theta^T}\left(\theta^T\phi(x) - A(\theta)\right)dx \\
&= \int_x p(x|\theta)\phi(x)\left(\phi(x)^T - \mathbb{E}_p[\phi(X)]^T\right)dx \\
&= \int_x p(x|\theta)\phi(x)\phi(x)^T - \int_x p(x|\theta)\phi(x)\mathbb{E}_p[\phi(X)]^T dx \\
&= \mathbb{E}_p[\phi(X)\phi(X)^T] - \mathbb{E}_p[\phi(X)]\mathbb{E}_p[\phi(X)]^T
\end{aligned}
$$

Maximum Likelihood Learning

- Given $x^1, x^2, ..., x^D$, we want to solve

$$\arg\max_\theta L(\theta), \quad L(\theta) = \frac{1}{D} \sum_{d=1}^{D} \log p(x^d|\theta)$$

- Simple approach: Gradient descent. Repeatedly set

$$\theta \leftarrow \theta + \lambda \frac{dL}{d\theta}$$

Maximum Likelihood Learning

$$
\begin{aligned}
\frac{dL}{d\theta} &= \frac{1}{D}\sum_{d=1}^{D}\frac{d}{d\theta}\log p(x^d|\theta) \\
&= \frac{1}{D}\sum_{d=1}^{D}\frac{d}{d\theta}(\theta^T\phi(x^d) - A(\theta)) \\
&= \frac{1}{D}\sum_{d=1}^{D}\phi(x^d) - \mathbb{E}_p[\phi(X)] \\
&= \hat{\mathbb{E}}[\phi(X)] - \mathbb{E}_p[\phi(X)]
\end{aligned}
$$

Maximum Likelihood Learning

- Given $x^1, x^2, ..., x^D$, we want to solve

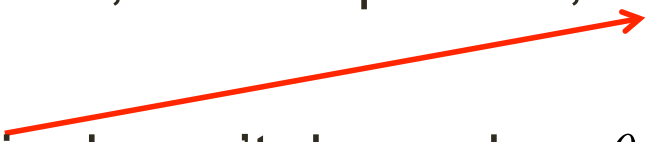
$$\arg\max_\theta L(\theta), \quad L(\theta) = \frac{1}{D}\sum_{d=1}^{D} \log p(x^d|\theta)$$

- Simple approach: Gradient descent. Repeatedly set

$$\theta \leftarrow \theta + \lambda \frac{dL}{d\theta}$$

$$\frac{dL}{d\theta} = \hat{\mathbb{E}}[\phi(X)] - \mathbb{E}_p[\phi(X)]$$

  - Notice, at the optimum, $\hat{\mathbb{E}}[\phi(X)] = \mathbb{E}_p[\phi(X)]$.



  - This doesn't depend on θ. Called a sufficient statistic.

Maximum Likelihood Learning

- Given $x^1, x^2, ..., x^D$, we want to solve

$$\arg\max_\theta L(\theta), \quad L(\theta) = \frac{1}{D}\sum_{d=1}^{D} \log p(x^d|\theta)$$

- Simple approach: Gradient descent. Repeatedly set

$$\theta \leftarrow \theta + \lambda \frac{dL}{d\theta}$$

$$\frac{dL}{d\theta} = \hat{\mathbb{E}}[\phi(X)] - \mathbb{E}_p[\phi(X)]$$

- For MRFs, if you can compute marginals, then:
  - You can compute $\mathbb{E}_p[\phi(X)]$.
  - So, you can compute $dL/d\theta$.
  - So, you can do maximum likelihood learning.

- If you remember one thing from this lecture let it be this!

Natural vs. Moment Parameters

- Given an exponential family

$$p(x|\theta) = \exp(\theta^T \phi(x) - A(\theta))$$

the distribution can be specified either by θ (the "natural" parameters) or by specifying

$$\mu = \mathbb{E}[\phi(X)]$$

(the "moment" parameters.)

If ϕ is minimal, then mapping between θ and μ is one-to-one. However, even if redundant, the distribution $p(x|\theta)$ is the same.

Where does the Exponential Family Come from?

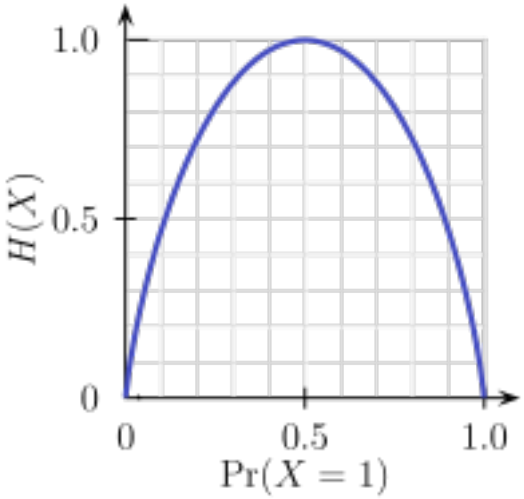
- One answer: We have lots of distributions, let's look for some common structure.

- Another answer: They are the maximum entropy distributions.

- Recall

$$H(p) = -\int_x p(x) \log p(x) dx$$

$$H(p) = -\sum_x p(x) \log p(x)$$



Maximum Entropy Interpretation

- Suppose we want to find

$$p = \arg\max_p H(p) \quad \text{s.t. } \mathbb{E}_p[f(X)] = F$$

then p is an exponential family.

Proof:

$$J(p,\lambda) = -\sum_x p(x)\log p(x) + \lambda_0(1 - \sum_x p(x))$$

$$= +\sum_k \lambda_k(F_k - \sum_x p(x) f_k(x))$$

$$0 = \frac{dJ}{dp(x)} = -(\log p(x) + 1) - \lambda_0 - \sum_k f_k(x)$$

$$\log p(x) = -1 - \lambda_0 - \sum_k \lambda_k f_k(x)$$

$$p(x) \propto \exp(-\sum_k \lambda_k f_k(x))$$

What does this tell us about graphical models?

- If you want to do maximum likelihood learning, you need the derivative of the log-partition function.
- This gradient is just a vector of marginals.
- At the ML solution, $\mathbb{E}_p[\phi(X)]$ will be equal to $\hat{\mathbb{E}}[\phi(X)]$. Thus, for MRFs, the mean values of all clique marginals are sufficient statistics.
- An MRF is the maximum entropy distribution with given marginals for each clique in the graph.

What to take home

- Definition of an exponential family.
- Bernoulli / Gaussians / MRFs as exponential families.
- Derivatives of log-partition function give cumulants.
- How to do maximum likelihood learning using gradient of log-partition function.
- Natural parameters vs. moment parameters and (relatedly) sufficient statistics