CSE6243: Advanced Machine Learning **Fall 2023**

Lecture 8: Density Parametrization

Lecturer: Bo Dai *Scribes: Bharat Goyal, Chetan Reddy*

Note: *LaTeX template courtesy of UC Berkeley EECS Department.*

8.1 Recap

During previous lectures we discussed *convex optimization* and *sampling*, which are techniques used in supervised, unsupervised, and reinforcement learning. Those techniques tell us **how to learn**, but we now want to formulate **what we want to learn**. We want to parameterize the distributions or densities by determining $p(x)$. We will also look at figuring out $p(\theta)$, $p(x|\theta)$, and $p(\theta|x)$.

8.2 New Content: Distributions (Density) Parametrization

8.2.1 Exponential Family of Distributions

The probability distribution is in the *canonical parameterization*:

$$p(x) = h(x) \exp(\eta^T T(x) - A(\eta)) \tag{8.1}$$

$A(\eta)$ is the *partition function* which satisfies the following condition to ensure that $p(x)$ is a valid distribution:

$$A(\eta) = \log \left(\int h(x) \exp(\eta^T T(x)) dx \right) \tag{8.2}$$

Typically, it's easier to determine these terms by converting $p(x)$ to the format above and then comparing terms.

Properties of Exponential Distributions

Property 1 (Convexity). $A(\eta)$ is convex with respect to η.
Proof: For convexity we need to show that $A(\lambda\eta_1 + (1-\lambda)\eta_2) \leq \lambda A(\eta_1) + (1-\lambda)A(\eta_2)$, i.e.,

$$\begin{aligned}
\exp(A(\lambda\eta_1 + (1-\lambda)\eta_2)) &= \int h(x)\exp\Big((\lambda\eta_1 + (1-\lambda)\eta_2)^T T(x)\Big)dx \\
&= \int \Big(h(x)^\lambda \exp\Big(\lambda\eta_1^T T(x)\Big)\Big)\Big(h(x)^{1-\lambda}\exp\Big((1-\lambda)\eta_2^T T(x)\Big)\Big)\,dx \\
&= \int \Big(h(x)\exp\Big(\eta_1^T T(x)\Big)\Big)^\lambda \Big(h(x)\exp\Big(\eta_2^T T(x)\Big)\Big)^{1-\lambda} dx \\
&\leq \left(\int \Big(h(x)\exp\Big(\eta_1^T T(x)\Big)\Big)dx\right)^\lambda \left(\int h(x)\exp\Big(\eta_2^T T(x)\Big)dx\right)^{1-\lambda}
\end{aligned}$$

The last step of the simplification relies on Holder's identity with $p = \frac{1}{\lambda}$ and $q = \frac{1}{1-\lambda}$:

$$\int f(x)g(x)dx \leq \left(\int f(x)^p dx\right)^{\frac{1}{p}} \left(\int g(x)^q dx\right)^{\frac{1}{q}}$$

Upon taking the log of the LHS and RHS above and applying the definition of $A(\eta)$, we get:

$$A(\lambda\eta_1 + (1-\lambda)\eta_2) \leq \lambda A(\eta_1) + (1-\lambda)A(\eta_2)$$

■

Property 2 (First-order derivatives). The first derivative of $A(\eta)$ w.r.t. η is the expected value of $T(x)$, i.e.,

$$\frac{\partial A(\eta)}{\partial \eta} = \mathbb{E}_{p(x)}[T(x)]$$

Property 3 (Second-order derivatives). The second derivative of $A(\eta)$ w.r.t. η is variance of $T(x)$, i.e.,

$$\frac{\partial^2 A(\eta)}{\partial^2 \eta} = \mathbb{E}_{p(x)}[T^2(x)] - \mathbb{E}_{p(x)}[T(x)]^2$$

Basic Distributions

Bernoulli Distribution. This distribution is discrete and deals with variables that can just take on 2 values ($x \in \{0, 1\}$):

$$p(x) = \pi^x(1-\pi)^{1-x}$$

The Bernoulli distribution is in fact a part of the exponential family:

$$\begin{aligned}
p(x) &= \exp\Big(x\log(\pi)\Big)\ \exp\Big((1-x)\log(1-\pi)\Big) \\
&= \exp\Big(x\log(\pi) + \log(1-\pi) - x\log(1-\pi)\Big) \\
&= \exp\left(x\log\Big(\frac{\pi}{1-\pi}\Big) + \log(1-\pi)\right)
\end{aligned}$$

On comparing with 8.1 we can see that

$$h(x) = 1, \eta = \log\Big(\frac{\pi}{1-\pi}\Big), T(x) = x, A(\eta) = -\log(1-\pi)$$

Gaussian Distribution. This is the standard normal distribution that is often used for the purpose of simulating noise. The pdf $p(x)$ is:

$$\begin{aligned} p(x) &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)\exp\left(\frac{-x^2+2x\mu}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)\exp\left(\begin{bmatrix}\frac{\mu}{\sigma^2} & \frac{-1}{2\sigma^2}\end{bmatrix}\begin{bmatrix}x\\x^2\end{bmatrix} - \frac{\mu^2}{2\sigma^2}\right) \end{aligned}$$

On comparing this with 8.1, we get:

$$h(x) = \frac{1}{\sigma\sqrt{2\pi}}, \eta = \begin{bmatrix}\frac{\mu}{\sigma^2}\\ \frac{-1}{2\sigma^2}\end{bmatrix}, T(x) = \begin{bmatrix}x\\x^2\end{bmatrix}, A(\eta) = \frac{\mu^2}{2\sigma^2}$$

Poisson Distribution.

$$\begin{aligned} p_\lambda(x) &= \frac{\lambda^x \exp(-\lambda)}{x!} \\ &= \frac{\exp(x\log\lambda - \lambda)}{x!} \end{aligned} \tag{8.3}$$

On this with 8.1 we can see that

$$h(x) = \frac{1}{x!}, \eta = \log(\lambda), T(x) = x, A(\eta) = \lambda$$

8.2.2 Energy-based Models (EBM)

Energy-based models are more general than the exponential family. They are of the following form:

$$p(x) = \exp(f(x) - A(f)) \tag{8.4}$$

A is the partition function and takes in the function f as input. It is of the form:

$$A(f) = \log \int \exp(f(x))dx \tag{8.5}$$

This normalizes $p(x)$ such that $\int p(x)dx = 1$,

The function f in energy-based models could be a neural network.

Ising Model. The Ising Model is a specific example of energy-based models that is widely used in image denoising tasks. x is a set of discrete variables such that

$$x = [x_1, x_2, \ldots, x_k], x_i \in \{-1, 1\}$$

The probability distribution is then proportional to the following exponential quantity:

$$p(x) \propto \exp(-\mu^T x - x^T W x) \tag{8.6}$$

Specifically it is of the form:

$$p(x) = \exp(-\mu^T x - x^T W x - A(u, w)) \tag{8.7}$$

The partition function is given by the following:

$$A(u, w) = \log \left[\sum_{x \in \{+1,-1\}^k} \exp(-\mu^T x - x^T W x) \right] \tag{8.8}$$

8.2.3 Latent Variable Model

The latent variable model is a *stochastic* model of the following form:

$$p(x) = \int p(x|z)p(z)dz \tag{8.9}$$

where $p(x|z)$ and $p(z)$ are both distributions from the exponential family. This model maps the distribution $p(z)$ to the distribution $p(x)$.

Gaussian Mixture Models. Gaussian mixture models are a type of latent variable models. The marginal distribution over z is given by:

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{8.10}$$

The condition distribution $p(x|z)$ is given by:

$$p(x|z) = \mathcal{N}(\mu_k, \sigma_k^2) \tag{8.11}$$

Latent Dirichlet Allocation (LDA). LDA is another example of a latent variable model that is used in NLP tasks such as corpus generation.

Diffusion Model. The diffusion model is a special case of latent variable model that follows a Markov chain. Since it is a Markov chain, each state is dependently on only the previous state. Thus, if we have a final state x, beginning state z_o, and intermediate states $z_1, z_2, \ldots, z_{k-1}, z_k$, the probability distribution is given by:

$$p(x) = \int p(x|z_k)p(z_k|z_{k-1}) \ldots p(z_1|z_o)p(z_0)dz_0^k \tag{8.12}$$

8.2.4 Normalizing Flow Model

The normalizing flow is a *deterministic* model that allows us to map a simple distribution to a more complex one. We have latent variable Z and observed variable X and a function f such that $f(Z) = X$. This function f must be invertible so there must be a function g such that $g(f(z)) = z$.

We sample $z \sim p(z)$, then we have $x = f(z)$. We can then use change of variables to calculate the distribution $q(x)$:

$$q(x) = p(f^{-1}(x)) * \left| \det \frac{\partial f^{-1}(x)}{\partial z} \right| \tag{8.13}$$

8.2.5 Autoregressive Model (ARM)

Autoregressive models use all of the previous time steps in order to calculate the probability of the current state. The joint distribution of states $x_1, x_2, \ldots x_k$ is given by:

$$p(x_1, x_2, \ldots, x_k) \ = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \ldots p(x_k|x_{<k}) \tag{8.14}$$