CS-598: Deep Generative and Dynamical Models **Spring 2023**

Topic 2: Energy Based Models (EBMs)

Lecturer: Arindam Banerjee *Scribe: Arindam Banerjee*

2.1 Estimating Boltzmann Distributions

2.1.1 Boltzmann (Gibbs) Distributions

The density/mass function is given by

$$p(x) = \frac{e^{-\beta E(x)}}{Z_\beta} \tag{2.1}$$

where the partition function (normalization constant) is

$$Z_\beta = \sum_{x'} e^{-\beta E(x')} \tag{2.2}$$

where $E(x)$ is the energy of configuration x.

From the statistical physics perspective, the coefficient $\beta \geq 0$ is represented as

$$\beta := \frac{1}{k_\beta T} \propto \frac{1}{T} \tag{2.3}$$

where k_β is the Boltzman constant and $T \geq 0$ is the absolute thermodynamic temperature in Kelvins. When $T \to \infty$, then $\beta \to 0$, then $p(x)$ is the uniform distribution on X. When $T \to 0$, then $\beta \to \infty$, then the support of $p(x)$ are the minima of $E(x)$.

The Boltzmann distribution can be derived as the maximum (Shannon) entropy distribution under the constraint of a given expected value of the energy function, i.e., $\mathbb{E}[E(X)] = \mu$.

Proposition 1. *Under the constraint $\mathbb{E}[E(X)] = \mu$, the maximum entropy distribution is given by $p(x) \propto e^{-\beta E(x)}$.*

Proof. Recall that Shannon entropy is given by $H(X) = -\sum_x p(x) \log p(x)$. The constrained optimization problem of maximizing entropy can be equivalently posed on as minimizing $-H(X)$ posed as:

$$\begin{aligned} \min_p \ & \sum_x p(x) \log p(x) \\ \text{s.t.} \ & \sum_x p(x) E(x) = \mu \\ & \sum_x p(x) = 1 \ , \quad p(x) \geq 0 \ . \end{aligned} \tag{2.4}$$

The Lagrangian is given by

$$L(p, \lambda_1, \lambda_2, \gamma) = \sum_x p(x) \log p(x) + \beta \left(\sum_x p(x) E(x) - \mu \right) + \lambda \left(\sum_x p(x) - 1 \right) \ ,$$

where β, λ are Lagrange multipliers corresponding to the equality constraints.[1] The first order optimality conditions give:[•]

[ab: $\log_2$ vs ln i entropy]

$$\frac{\partial L}{\partial p(x)} = 0 \implies \log p(x) + 1 + \beta E(x) + \lambda = 0 \implies p(x) = e^{-\beta E(x) - \lambda - 1} \tag{2.5}$$

$$\frac{\partial L}{\partial \beta} = 0 \implies \sum_x p(x) E(x) = \mu \ , \tag{2.6}$$

$$\frac{\partial L}{\partial \lambda} = 0 \implies \sum_x p(x) = 1 \ . \tag{2.7}$$

Plugging in (2.5) in (2.7), we get

$$\sum_x e^{-\beta E(x) - \lambda - 1} = 1 \implies e^{\lambda + 1} = \sum_x e^{-\beta E(x)}$$

which implies

$$p(x) = \frac{e^{-\beta E(x)}}{\sum_{x'} e^{-\beta E(x')}} \ ,$$

the desired form of the Boltzmann distribution. □

[ab: β gets determined by (2.6).]

[abb: move free + internal energy up, show gradient form of (2.6), show how to estimate β]

The *free energy* is defined as:

$$F_\beta := -\frac{1}{\beta} \log Z_\beta = -\frac{1}{\beta} \log \sum_x e^{-\beta E(x)} \ . \tag{2.8}$$

The expected energy $\mathbb{E}[E(X)]$ is also referred to as the *internal energy* of the model. Now note that the entropy

$$\begin{aligned} H(X) &= -\sum_x p(x) \log p(x) \\ &= -\sum_x p(x)(-\beta E(x) - \log Z) \\ &= \beta \sum_x p(x) E(x) + \log Z \\ &= \beta \left[\mathbb{E}[E(X)] - F_\beta \right] \ , \end{aligned}$$

where F_β is the free energy.

Remark (Entropy and Effect of β)**.** *Note that the equation for estimating β from* (2.6) *is*

$$\begin{aligned} &\sum e^{-\beta E(x)} E(x) = \mu \sum_x e^{-\beta E(x)} \\ \Rightarrow \qquad & H(X) = -\beta F_\beta + \beta \mathbb{E}[E(X)] \end{aligned}$$

[ab: unclear]

[1] Note that we do not explicitly introduce a Lagrange multiplier for the inequality constraints $p(x) \geq 0$ because the $\log p(x)$ in the objective would ensure $p(x) \geq 0$. For a more formal treatment, please see Topic xx notes on Optimization.

2.1.2 Parametric Boltzman Distributions

In general, the energy function $E(x)$ is unknown and one starts with a parametric form $E(x;\theta)$ where $\theta \in \mathbb{R}^p$ are a set of unknown parameters. The corresponding parametric Boltzmann distribution is given by

$$p(x;\theta) = \frac{e^{-E(x;\theta)}}{Z(\theta)} , \qquad Z(\theta) = \sum_x e^{-E(x;\theta)} . \tag{2.9}$$

Note that we are not maintaining β explicitly in (2.9) and in the sequel (unless explicitly stated) since one can always include such scaling directly in $E(x;\theta)$.

Given a set of n samples $\{x_i, i \in [n]\}$ drawn i.i.d. from $p(x;\theta)$, the core problem of interest is to do maximum likelihood estimation of θ. The log-likelihood of the observed samples is given by

$$L(\theta) = \frac{1}{n}\sum_{i=1}^n \log p_\theta(x_i) = -\frac{1}{n}\sum_{i=1}^n E(x_i;\theta) - \log Z(\theta) . \tag{2.10}$$

Minimizing the log-likelihood is typically done using a suitable variant of gradient descent. The gradient of the log-likelihood is given by

$$\nabla_\theta L(\theta) = -\frac{1}{n}\sum_{i=1}^n \nabla_\theta E(x_i;\theta) - \nabla_\theta \log Z(\theta) . \tag{2.11}$$

Now note that

$$\begin{aligned}
\nabla_\theta \log Z(\theta) &= \frac{1}{Z(\theta)}\nabla_\theta Z(\theta) \\
&= \frac{1}{Z(\theta)}\nabla_\theta \sum_x e^{-E(x;\theta)} \\
&= \frac{1}{Z(\theta)}\sum_x \nabla_\theta e^{-E(x;\theta)} \\
&= -\frac{1}{Z(\theta)}\sum_x e^{-E(x;\theta)}\nabla_\theta E(x;\theta) \\
&= -\sum_x p(x;\theta)\nabla_\theta E(x;\theta) \\
&= -\mathbb{E}_{p(\cdot;\theta)}[\nabla_\theta E(X;\theta)] .
\end{aligned}$$

Plugging the expression for $\nabla_\theta \log Z(\theta)$ in (2.11), we get

$$\nabla_\theta L(\theta) = \mathbb{E}_{p(\cdot;\theta)}[\nabla_\theta E(X;\theta)] - \frac{1}{n}\sum_{i=1}^n \nabla_\theta E(x_i;\theta) . \tag{2.12}$$

Thus, the gradient of the log-likelihood is the difference between the expected gradient of the energy function and the observed sample averaged gradient of the energy function.

The central challenge in computing the gradient of the log-likelihood in (2.12) is that the expectation in $\mathbb{E}_{p(\cdot;\theta)}[\nabla_\theta E(X;\theta)]$ is typically intractable. The approach which has been widely studied in the literature is to draw a set of samples $\{\tilde{x}_i, i \in \tilde{n}\}$ from $p(x;\theta)$, and replace the expectation with the empirical average over these samples, so that

$$\nabla_\theta L(\theta) = \mathbb{E}_{p(\cdot;\theta)}[\nabla_\theta E(X;\theta)] - \frac{1}{n}\sum_{i=1}^n \nabla_\theta E(x_i;\theta) \tag{2.13}$$

$$\approx \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \nabla_\theta E(\tilde{x}_i;\theta) - \frac{1}{n}\sum_{i=1}^n \nabla_\theta E(x_i;\theta) . \tag{2.14}$$

The above gradients enable a suitable stochastic gradient descent approach for estimating θ.[2] The core challenge reduces to drawing a set of samples $\{\tilde{x}_i, i \in \tilde{n}\}$ from $p(x;\theta)$, which is done using a suitable Markov chain Monte Carlo (MCMC) method.[3]

2.1.3 Boltzmann Mahines

We will shortly see examples of such parametric energy functions.

[2]For details, please see Topic xx notes on Optimization.
[3]For details, please see Topic xx notes on Inference.

2.2 Variational Inference

The joint distribution of a latent variable model (LVM) is given by

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) \ , \tag{2.15}$$

where $\mathbf{x}$ denotes the observed variable, $\mathbf{z}$ denotes the latent variable, and $\boldsymbol{\theta}$ denotes the parameters.

There are two key related challenges associated with inference (and estimation) in LVMs: first, accurately computing the marginal distribution

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \int_{\mathbf{z}} p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) d\mathbf{z} \ , \tag{2.16}$$

and second, accurately computing the posterior distribution

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{x})} \ . \tag{2.17}$$

The intractability of computing the marginal $p_{\boldsymbol{\theta}}(\mathbf{x})$ also makes estimating θ difficult. The challenge applies to both (a) point estimates, e.g., maximum likelihood estimation (MLE) by maximizing $\log p_{\boldsymbol{\theta}}(\mathbf{x})$ over $\boldsymbol{\theta}$, as well as (b) Bayesian estimation which, for a given prior $p(\boldsymbol{\theta})$, needs to compute $p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}'} p(\mathbf{x}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')d\boldsymbol{\theta}'}$. For the purposes of this exposition, we will focus on the MLE.

In Variational Inference (VI), one constructs a distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ with parameters $\boldsymbol{\phi}$ with the goal of approximating the true posterior, i.e.,

$$q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \approx p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) \ . \tag{2.18}$$

Definitions. In the literature, $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ is referred to as the *inference model* or the *recognition model*. In some contexts, $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ is referred to as the *encoder*. The parameters $\boldsymbol{\phi}$ for the inference model is usually referred to as *variational parameters* to contrast ϕ with the *model parameters* $\boldsymbol{\theta}$. Further, the conditional model $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ is referred to as the *generative model*.

2.3 Evidence Lower Bound (ELBO)

Given any inference model $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, we have

$$\begin{aligned}
\log p_{\boldsymbol{\theta}}(\mathbf{x}) &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x})\right] \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log\left(\frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})}\right)\right] \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log\left(\frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\frac{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})}\right)\right] \\
&= \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log\left(\frac{p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\right)\right]}_{\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x})\ \text{(ELBO)}} + \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log\left(\frac{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})}\right)\right]}_{D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))} \ .
\end{aligned}$$

The first term is referred to as the *evidence lower bound (ELBO)* or the *variational lower bound*:

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) - \log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})] \ . \tag{2.19}$$

The second term is the Kullback Leibler (KL) divergence between the approximate and the true posterior distributions. Since the KL-divergence is non-negative, i.e.,

$$D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log\left(\frac{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})}\right)\right] \geq 0 \ , \tag{2.20}$$

the ELBO forms a lower bound to the log-likelihood:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) + D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) \tag{2.21}$$

$$\geq \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) \ . \tag{2.22}$$

For Variational Inference (VI), for a given $\boldsymbol{\theta}$, one focuses on maximizing the ELBO $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x})$ w.r.t. the variational parameters $\boldsymbol{\phi}$ for a suitable choice of an inference model $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. Further, maximum likelihood estimation of $\boldsymbol{\theta}$ gets translated to maximizing the lower bound ELBO $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x})$ w.r.t. both the variational parameters $\boldsymbol{\phi}$ and the model parameters $\boldsymbol{\theta}$.

The core idea in VI is thus to convert the *high-dimensional integration* $\int_z p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})d\mathbf{z}$ to a (lower bound) *maximizing problem* $\max_{\boldsymbol{\phi}} \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x})$ which provides an approximate solution, in particular a lower bound, to the original high-dimensional integration problem. The quality of the approximation is determined by how well $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ approximates the true posterior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$. If $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$, then $D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})) = 0$, and $\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x})$, i.e., the ELBO exactly computes the log-likelihood. More generally, the goal is to find a inference model $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ which

- is flexible and approximates $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ accurately for a suitable choice of $\boldsymbol{\phi}$,
- is easy to optimize w.r.t. the variational parameters $\boldsymbol{\phi}$, and
- is easy to sample from.

The need for the ease of sampling above is to allow sample approximations of true expectations, and will be made clear in the sequel.

Two Interpretations of the ELBO. There are two common ways of interpreting the ELBO. The first is in terms of entropy regularized joint likelihood modeling:

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})] + \mathbb{H}[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})] \ , \tag{2.23}$$

where $\mathbb{H}[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[-\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})]$ is the Shannon entropy of $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. The first term is high when the approximate posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ assigns more probability mass to regions where the joint distribution $p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})$ is high. The second term encourages $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ to be more spread out so as to have high entropy.

The second interpretation is in terms of the reconstruction accuracy while regularizing the approximate posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ to stay close to the true prior $p(\mathbf{z})$:

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z})) \ . \tag{2.24}$$

The first term is the reconstruction accuracy. Viewing $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ as a stochastic encoder $\mathbf{x} \mapsto \mathbf{z}$ and $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ as a stochastic decoder $\mathbf{z} \mapsto \mathbf{x}$, the reconstruction accuracy is high when the encoder $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ puts more probability mass on latent $\mathbf{z}$ which when decoded by $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ yields high probability (e.g., observed) $\mathbf{x}$. The second term encourages the posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ to stay close to the true marginal $p_{\boldsymbol{\theta}}(\mathbf{z})$, and can be viewed as a regularizer. The regularization term has been used to explain the tendency of such latent models to prune out latent dimensions [Burda et al., Hoffman et al.].

2.4 Sharpening the ELBO

2.5 Optimizing the ELBO