# Homework 2

Deep Learning 2024 Spring

Due 11:59pm, 2024/4/13

## 1 True or False

**Problem 1.** Generative models can be used to both classify and generate images.

**Problem 2.** We can train an energy-based model without knowing the explicit density function (or normalizing factor $Z$).

**Problem 3.** Stochastic Gradient MCMC is designed to solve the optimization problem $\arg\max_\theta \mathbb{P}(\theta|\mathbf{X})$, where $\theta$ is the collection of parameters and $\mathbf{X}$ represents data.

## 2 Q&A

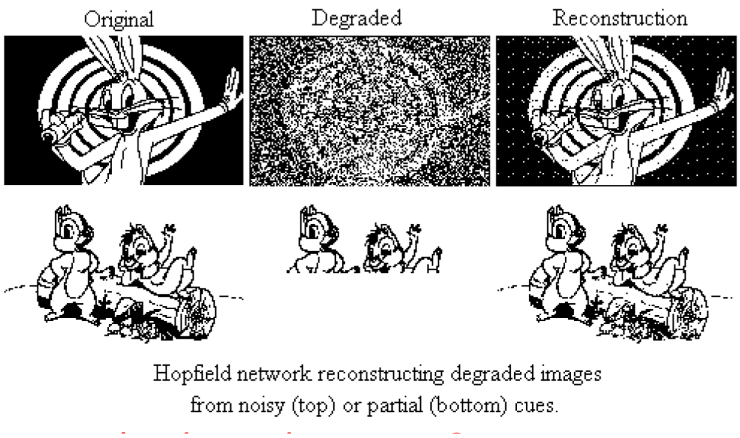


Figure 1: Noisy image (top row) and masked image (bottom row).

**Problem 4.** (Hopfield Network) Answer the following questions about the Hopfield network.

1. Figure 1 shows two types of degraded images: noisy image and masked image. Design an appropriate process to retrieve stored patterns using the Hopfield network for each case respectively.

   (Hint: The unmasked part of a masked image is the same as ground truth. By contrast, most pixels of a noisy image are different from the ground truth.)

2. (Redundancy) Although a Hopfield network explicitly stores only N patterns, redundancy in its state space allows it to represent many more configurations. Suppose the Hebbian learning rule is given by $W = \frac{1}{N}\sum_p y_p y_p^T$. We want to use the Hebbian learning rule to construct a Hopfield network. Prove that with $N$ orthogonal patterns $y_p$ ($y_p$ is a $N$-dim vector) for $p = 1, \dots, N$, the Hopfield network can memorize all $2^N$ patterns, in the sense that each of these patterns corresponds to a local minimum of the network's energy function.

**Problem 5.** (Boltzman Machine) Consider a fully connected Boltzman machine. We remark the visible units as $v$, the hidden units as $h$, and all units $y = (v, h)$. The joint probability of $v$ and $h$ is given by

$$\mathbb{P}(v,h) = \frac{\exp(y^T W y)}{\sum_{y'} \exp(y'^T W y')}. \tag{1}$$

And the marginal probability of $v$ is given by

$$\mathbb{P}(v) = \sum_h \mathbb{P}(v,h). \tag{2}$$

We aim to maximize the log-likelihood, and the loss is given by

$$L(W) = -\frac{1}{|P|}\sum_{v\in P} \log \mathbb{P}(v). \tag{3}$$

Prove that the gradient of Eq. (3) has the following form:

$$\nabla_W \text{loss}(W) = -\frac{1}{|P|}\sum_{v\in P}\left(\mathbb{E}_{h|v}\left[yy^T\right] - \mathbb{E}_{y'}\left[y'y'^T\right]\right).$$

**Problem 6.** (Gaussian RBM) Consider a restricted Boltzman machine with a single hidden layer and the following energy function $\mathcal{E}_{W,b} : \mathbb{R}^{N_h+N_v} \to \mathbb{R}$:

$$\mathcal{E}_{W,b}(v,h) = \frac{1}{2}(v-b)^T(v-b) - v^T W h$$

where $W$, $b$ are trainable parameters, $v$ is visible continuous-value units (i.e., $v \in \mathbb{R}^{N_v}$), and $h$ is hidden discrete-value units (i.e., $h \in \{-1,1\}^{N_h}$).

1. Derive the conditional distribution $\mathbb{P}(v|h)$.

2. Derive the gradient of $b$ if we train this model based on the maximum log-likelihood principle. (Hint: Your answer should contain the form of an expectation.)

**Problem 7.** (Undirected Probabilistic Model) A **graphical model** or **probabilistic graphical model** or **structured probabilistic model** is a probabilistic model for which a graph expresses the conditional

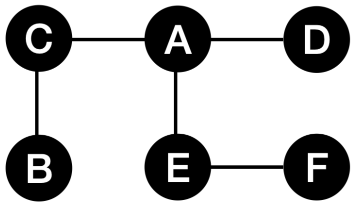


Figure 2: An example of undirected probabilistic model.

dependence structure between random variables. [1] In an **undirected graphical model**, an edge implies dependence between the corresponding random variables. Figure 2 shows an example, where the joint probability distribution can be factorized as

$$\mathbb{P}(A, B, C, D, E, F) = \frac{1}{Z} f_{AD}(A, D) f_{AC}(A, C) f_{AE}(A, E) f_{BC}(B, C) f_{EF}(E, F)$$

for some non-negative functions $f_{AB}$, $f_{AC}$, $f_{AD}$, $f_{AE}$, $f_{BC}$, and $f_{EF}$, and a normalizing factor $Z$ (also called partition function).

1. Are $D$ and $F$ independent?

2. Write down the unnormalized conditional distribution $\mathbb{P}(B, E|A)$. "unnormalized" means you can omit the normalizing factor. Are $B$ and $E$ independent given $A$?

3. We model $\mathbb{P}(A, B, C, D, E, F)$ as a Boltzman distribution

$$\mathbb{P}(A, B, C, D, E, F) \propto \exp(-\mathcal{E}(A, B, C, D, E, F))$$

where $\mathcal{E}$ is the energy function. Show that the energy function can be expressed by the following factorization:

$$\mathcal{E}(A, B, C, D, E, F) = \mathcal{E}_{AC}(A, C) + \mathcal{E}_{AD}(A, D) + \mathcal{E}_{AE}(A, E) + \mathcal{E}_{BC}(B, C) + \mathcal{E}_{EF}(E, F)$$

**Problem 8.** (Importance Sampling) $\mathsf{x}$ is a random variable. Given target distribution $p(\mathbf{x})$ and target random variable $\mathbf{y} = f(\mathbf{x})$, importance sampling gives an estimator of $\mathbb{E}[\mathbf{y}]$ from a proposal distribution $q(\mathbf{x})$:

$$\mathbb{E}_{\mathbf{x}\sim p}\left[f(\mathbf{x})\right] = \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{p(\mathbf{x})}{q(\mathbf{x})} f(\mathbf{x})\right] \approx \frac{1}{N} \sum_{x\sim q(\cdot)} \frac{p(x)}{q(x)} f(x).$$

Prove that when $q$ has the following form,

$$q^{\star}(\mathbf{x}) \propto p(\mathbf{x}) |f(\mathbf{x})|$$

the variance of this estimator can be minimized.

**Problem 9.** (Markov Chain Monte Carlo)

[1] https://en.wikipedia.org/wiki/Graphical_model

1. Prove random-walk Metropolis-Hasting sampling (i.e., $\mathbf{s}' \leftarrow \mathbf{s} +$ Gaussian noise) is a valid MCMC algorithm, i.e., it constructs a Markov chain which is ergodic and satisfies the detailed balance property.

2. Prove that Gibbs sampling is a special case of Metropolis-Hasting sampling, and that the acceptance rate of Gibbs sampling (i.e., $\alpha(\mathbf{s} \to \mathbf{s}')$) is 1.

   Here we consider the following 2-step Gibbs proposal: (1) randomly sample a coordinate index $i$; (2) sample coordinate $\mathbf{s}_i$ from the coordinate proposal $q(\mathbf{s}_i \to \mathbf{s}'_i) = p(\mathbf{s}'_i | \mathbf{s}_{j \neq i})$.

3. (**Optional Question**) In fact, Gibbs sampling is typically implemented in a *cyclic fashion*, i.e., running posterior sampling in a fixed order over all the dimensions. Prove that cyclic Gibbs sampling yields the same stationary distribution as random-order Gibbs sampling in the above question, as long as the Markov chain can access all states under the fixed ordering.