Deep Learning:
Homework 3: Energy-Based Models

Due on 03/28/2021

Xinyi Zhao

Problem 1

1.1 Energy Based Models Intuition

(a) **Answer:**
Instead of trying to classify x's to y's, we would like to use Energy-based model to predict if a certain pair of (x, y) fit together or not. Or in other words, find a y compatible with x.
Energy-based model constructs an energy function to measure the compatibility between input x and output y. The lower the energy, the better the compatibility. In the inference step, given x, each y in the corresponding space has a scalar-valued energy. It is possible that there are multiple values of y, such that their energy measures are the smallest, therefore, all of them are good output corresponding to the given x.

(b) **Answer:**
Energy-based models have no requirement for normalization, which is required by probabilistic models which output probabilities, to avoid additional problems associated with estimating or calculating normalization constant factors (e.g., the denominator in Gibbs-Boltzmann distribution), which further allows more flexible modeling architecture.
In probabilistic models, you basically don't have the choice of the objective function you're going to minimize, and you have to stay true to the probabilistic framework in the sense that every object you manipulate has to be a normalized distribution. Energy-based models give you way more choices in how you handle the model, way more choices of how you train it, and what objective function you use.

(c) **Answer:**
We can convert energy function to conditional probability by using Gibbs-Boltzmann distribution.

$$\begin{aligned}
P(y, z|x) &= \frac{exp(-\beta E(x, y, z))}{\int_y \int_z exp(-\beta E(x, y, z))} \\
P(y|x) &= \int_z P(y, z|x) \\
P(y|x) &= \frac{\int_z exp(-\beta E(x, y, z))}{\int_y \int_z exp(-\beta E(x, y, z))} \\
&= \frac{exp(-\beta F_W(x, y))}{\int_y exp(-\beta F_W(x, y))}
\end{aligned}$$

(d) **Answer:**
The loss function is a function that measures the distance between the expected value

and the actual value of a model. In an energy-based model, it is used in training step, and its role is to evaluate the quality of a particular energy function on training set by measuring the discrepancy between its output and given true output.
The energy function is used in inference step, and its role is to measure the compatibility between input x and output y.

(e) **Answer:**
Yes, for example, energy loss.

$$L_{energy}(Y^i, E(W, \mathcal{Y}, X^i)) = E(W, Y^i, X^i)$$

(f) **Answer:**
By only feeding positive examples into the training step, the algorithm does not know where to push the energy up, resulting flat energy surface everywhere. It is equivalent to say all the y's in its corresponding space are equally compatible with a given x.

(g) **Answer:**

(1) Push down of the energy of data points, push up everywhere else: Max likelihood

(2) Push down of the energy of data points, push up on chosen locations: max likelihood with MC/MMC/HMC

(3) Train a function that maps points off the data manifold to points on the data manifold: denoising auto-encoder

(h) **Answer:**
For example, hinge.

$$\mathcal{L}_{hinge}(x, y, W) = [F_W(x, y) – F_W(x, \hat{y}) + m(y, \hat{y})]^+$$

$m(y, \hat{y})$ is the margin and $(x, \hat{y})$ is negative example. It pushes down on $F_W(x, \hat{y})$, while pulling up on $F_W(x, y)$.

Problem 2

1.2 Negative log-likelihood loss

(i) **Answer:**

$$P(y|x) = \frac{exp(-\beta F_W(x,y))}{\sum_{y'=1}^{n} exp(-\beta F_W(x,y'))}$$

(ii) **Answer:**

$$\begin{aligned}
\mathcal{L}_{nll}(x,y,W) &= \frac{1}{\beta}[-\log P(y|x)] \\
&= \frac{1}{\beta}[-\log \frac{exp(-\beta F_W(x,y))}{\sum_{y'=1}^{n} exp(-\beta F_W(x,y'))}] \\
&= \frac{1}{\beta}[-\log exp(-\beta F_W(x,y)) + \log \sum_{y'=1}^{n} exp(-\beta F_W(x,y'))] \\
&= F_w(x,y) + \frac{1}{\beta}\log \sum_{y'=1}^{n} exp(-\beta F_W(x,y'))
\end{aligned}$$

(iii) **Answer:**

$$\begin{aligned}
\frac{\partial \mathcal{L}_{nll}}{\partial W} &= \frac{\partial F_w(x,y)}{\partial W} + \frac{1}{\beta}\frac{\partial \log \sum_{y'=1}^{n} exp(-\beta F_W(x,y')}{\partial W} \\
&= \frac{\partial F_w(x,y)}{\partial W} + \frac{1}{\beta}\frac{1}{\sum_{y'=1}^{n} exp(-\beta F_W(x,y'))}\sum_{y'=1}^{n} -\beta\frac{\partial F_W(x,y')}{\partial W} exp(-\beta F_W(x,y')) \\
&= \frac{\partial F_w(x,y)}{\partial W} - \sum_{y'=1}^{n}\frac{\partial F_W(x,y')}{\partial W}\frac{exp(-\beta F_W(x,y'))}{\sum_{y'=1}^{n} exp(-\beta F_W(x,y'))} \\
&= \frac{\partial F_w(x,y)}{\partial W} - \sum_{y'=1}^{n} P(y'|x)\frac{\partial F_W(x,y')}{\partial W}
\end{aligned}$$

Why: The gradient of this loss function is generally very complex and hence computing, estimating or approximating the integral is very intractable. When n is very large (there is a large amount of y'), or say the cardinality of y is high, then the second term is hard to compute because of the summation and derivative.

How: This intractability can be partially solved by approximation, for example, by using MC/MCMC/HMC/CD, we can take a $\hat{y}$ sampled from $P(y|x)$, such that

$$\frac{\partial \mathcal{L}_{nll}}{\partial W} = \frac{\partial F_w(x,y)}{\partial W} - \frac{\partial F_w(x,\hat{y})}{\partial W}$$

In this way, this we cant get around this intractability.

Problem 2 continued on next page. . .

(iv) **Answer:**

The first part of the negative log-likelihood loss is the energy of the correct example, and the second part is the weighted aggregation of both correct example and incorrect examples. When the loss function is being minimized, the energy of the correct example will be greatly pushed down in the first term $F_w(x, y)$, and in the second term $\frac{1}{\beta} \log \sum_{y'=1}^{n} exp(-\beta F_W(x, y'))$, because of the $-\beta$, all $F_w(x, y')$ will be greatly pushed up. Therefore, without any hard boundary, the energy of the correct example is pushed up to negative infinity, at the meantime, all the incorrect examples are being pushed up without any hard boundary, resulting in positive infinity outcome, resulting in an energy surface with really sharp edges in case of continuous y.

Problem 3

1.3 Comparing Contrastive Loss Functions

(a) **Answer:**

$$\begin{aligned}
\frac{\partial \mathcal{L}_{simple}(x,y,\bar{y},W)}{\partial W} &= \frac{\partial [F_w(x,y)]^+}{\partial W} + \frac{\partial [m - F_w(x,\bar{y})]^+}{\partial W} \\
&= \mathbb{1}[F_w(x,y) \geq 0] \times \frac{\partial F_w(x,y)}{\partial W} \\
&\quad + \mathbb{1}[(m - F_w(x,\bar{y})) \geq 0] \times \frac{\partial (m - F_w(x,\bar{y}))}{\partial W} \\
&= \mathbb{1}[F_w(x,y) \geq 0] \times \frac{\partial F_w(x,y)}{\partial W} \\
&\quad - \mathbb{1}[(m - F_w(x,\bar{y})) \geq 0] \times \frac{\partial F_w(x,\bar{y})}{\partial W}
\end{aligned}$$

Where $\mathbb{1}[true] = 1, \mathbb{1}[false] = 0$.

$$\frac{\partial \mathcal{L}_{simple}(x,y,\bar{y},W)}{\partial W} = \begin{cases}
\frac{\partial F_w(x,y)}{\partial W} - \frac{\partial F_w(x,\bar{y})}{\partial W} & F_w(x,y) \geq 0, F_w(x,\bar{y}) \leq m \\
\frac{\partial F_w(x,y)}{\partial W} & F_w(x,y) \geq 0, F_w(x,\bar{y}) > m \\
-\frac{\partial F_w(x,\bar{y})}{\partial W} & F_w(x,y) < 0, F_w(x,y) \leq m \\
0 & otherwise
\end{cases}$$

(b) **Answer:**

$$\begin{aligned}
\frac{\partial \mathcal{L}_{hinge}(x,y,\bar{y},W)}{\partial W} &= \frac{\partial [F_w(x,y) - F_w(x,\bar{y}) + m]^+}{\partial W} \\
&= \mathbb{1}[(F_w(x,y) - F_w(x,\bar{y}) + m) \geq 0] \times (\frac{\partial F_w(x,y)}{\partial W} - \frac{\partial F_w(x,\bar{y})}{\partial W})
\end{aligned}$$

Where $\mathbb{1}[true] = 1, \mathbb{1}[false] = 0$.

$$\frac{\partial \mathcal{L}_{hinge}(x,y,\bar{y},W)}{\partial W} = \begin{cases}
\frac{\partial F_w(x,y)}{\partial W} - \frac{\partial F_w(x,\bar{y})}{\partial W} & F_w(x,y) - F_w(x,\bar{y}) \geq -m \\
0 & otherwise
\end{cases}$$

(c) **Answer:**

$$
\begin{aligned}
\frac{\partial \mathcal{L}_{square-square}(x,y,\bar{y},W)}{\partial W} &= \frac{\partial([F_w(x,y)]^+)^2}{\partial W} + \frac{\partial([m - F_w(x,\bar{y})]^+)^2}{\partial W} \\
&= 2[F_w(x,y)]^+ \frac{\partial [F_w(x,y)]^+}{\partial W} + 2[m - F_w(x,\bar{y})]^+ \frac{\partial [m - F_w(x,\bar{y})]^+}{\partial W} \\
&= \mathbb{1}[F_w(x,y) \geq 0] \times 2[F_w(x,y)]^+ \frac{\partial F_w(x,y)}{\partial W} \\
&\quad - \mathbb{1}[(m - F_w(x,\bar{y})) \geq 0] \times 2[m - F_w(x,\bar{y})]^+ \frac{\partial F_w(x,\bar{y})}{\partial W} \\
&= \mathbb{1}[F_w(x,y) \geq 0] \times 2F_w(x,y) \frac{\partial F_w(x,y)}{\partial W} \\
&\quad - \mathbb{1}[(m - F_w(x,\bar{y})) \geq 0] \times 2[m - F_w(x,\bar{y})] \frac{\partial F_w(x,\bar{y})}{\partial W}
\end{aligned}
$$

Where $\mathbb{1}[true] = 1, \mathbb{1}[false] = 0$.

$$
\frac{\partial \mathcal{L}_{square-square}}{\partial W} = \begin{cases}
2F_w(x,y)\frac{\partial F_w(x,y)}{\partial W} - 2[m - F_w(x,\bar{y})]\frac{\partial F_w(x,\bar{y})}{\partial W} & F_w(x,y) \geq 0, F_w(x,\bar{y}) \leq m \\
2F_w(x,y)\frac{\partial F_w(x,y)}{\partial W} & F_w(x,y) \geq 0, F_w(x,\bar{y}) > m \\
-2[m - F_w(x,\bar{y})]\frac{\partial F_w(x,\bar{y})}{\partial W} & F_w(x,y) < 0, F_w(x,y) \leq m \\
0 & otherwise
\end{cases}
$$

(d) **Answer:**

i The second term in NLL loss is the aggregated energy corresponding to all the y's, while simple, hinge, square-square loss has only one specific incorrect $\bar{y}$, which means NLL loss pushes up all incorrect answers, while other three only pushes up well-chosen other points.

ii In hinge loss, margin develops an energy gap between correct and incorrect answers, which makes inference easier. The reason to take positive part of $[F_w(x,y) - F_w(x,\bar{y}) + m]$ is that we want to minimize the loss at least whenever $F_w(x,y) - F_w(x,\bar{y}) + m > 0$, which is $F_w(x,\bar{y}) - F_w(x,y) < m$.

When $F_w(x,\bar{y}) - F_w(x,y) \geq m$, $F_w(x,y) - F_w(x,\bar{y}) + m \leq 0$, then we do not need to minimize the loss anymore.

iii Simple and square-square losses penalize large energy of the correct answer and smaller than m energy of incorrect answers separately, while hinge loss only depends on the gap between them. Therefore, simple and square-square losses push that $F_w(x, y) \leq 0$ and $F_w(x, \bar{y}) \geq m$, while the hinge loss pushes $F_w(x, \bar{y}) - F_w(x, y) \geq m$.

Square-square loss penalizes both terms quadratically (stronger), while simple loss penalizes both terms linearly. I would prefer to use square-square loss when higher penalty is intended to be assigned to larger energy deviation, which constrains the energy of correct answers closer to zero, and the energy of incorrect answers closer to the margin m. Otherwise I would use simple loss.