CSC412/2506 Winter 2020: Probabilistic Learning and Reasoning

Tianyu Du

March 15, 2020

Contents

1 Introduction

2 Probabilistic Models

Definition 2.1. Given an i.i.d. dataset $\mathcal{D}$, the log-likelihood of θ is defined as

$$\ell(\theta; \mathcal{D}) = \sum_{i=1}^{N} \log p(x^{(i)}|\theta) \tag{2.1}$$

Definition 2.2. A **statistic** is a deterministic function of a set of random variables.

Definition 2.3. $T(X)$ is a **sufficient statistic** for random variable X if

$$T(x^{(1)}) = T(x^{(2)}) \implies L(\theta; x^{(1)}) = L(\theta; x^{(2)}) \quad \forall\theta \tag{2.2}$$

equivalently,

$$P(\theta|T(X)) = P(\theta|X) \tag{2.3}$$

3 Directed Graphical Models

3.1 Decision Theory

3.2 Latent Variables

Complete data case Let $\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^{N}$ denote the dataset, then

$$\ell(\theta; \mathcal{D}) = \sum_{i=1}^{N} \log p\left(x^{(i)}, y^{(i)}|\theta\right) \tag{3.1}$$

Partially observed dataset Let

$$\mathcal{D}^c = \{(x^{(i)}, y^{(i)}) : i \in \mathcal{C}\} \subseteq \mathcal{D} \tag{3.2}$$

denote the part of dataset with observed labels, and let

$$\mathcal{D}^m = \{(x^{(i)}) : i \in \mathcal{M}\} \subseteq \mathcal{D} \tag{3.3}$$

denote the set of observations without labels. Note that $\mathcal{C} \cup \mathcal{M} = \{1, 2, \cdots, N\}$.
Then the log likelihood is

$$\ell(\theta; \mathcal{D}) = \sum_{c\in\mathcal{C}} \log p\left(x^c, y^c|\theta\right) + \sum_{m\in\mathcal{M}} \log p\left(x^m|\theta\right) \tag{3.4}$$

$$= \sum_{c\in\mathcal{C}} \log p\left(x^c, y^c|\theta\right) + \sum_{m\in\mathcal{M}} \log \sum_{y} p\left(x^m, y|\theta\right) \tag{3.5}$$

Inference with Latent Variables Let z denote the latent variable, then

$$p(y|x) = \sum_{z} p(y|x, z)p(z) \tag{3.6}$$

3.3 Mixture Models

Inference using mixture models Let $\Theta = \{\theta_z\} \cup \{\theta_1, \theta_2, \cdots, \theta_K\}$. Where θ_z quantifies the distribution of z, and θ_k for each $k \in \{1, 2, \cdots, K\}$ denote the set of parameters describing $p(x|z = k)$.

$$p(x|\Theta) = \sum_{k=1}^{K} p(x, z = k|\Theta) \tag{3.7}$$

$$= \sum_{k=1}^{K} p(z = k|\Theta) p(x|z = k, \Theta) \tag{3.8}$$

$$= \sum_{k=1}^{K} p\left(z = k|\theta_z\right) p\left(x|z = k, \theta_k\right) \tag{3.9}$$

Posterior probabilities / responsibilities

$$p\left(z = k|x, \theta_z\right) = \frac{p\left(x|z = k, \theta_k\right) p\left(z = k|\theta_z\right)}{p(x|\Theta)} \tag{3.10}$$

$$= \frac{p\left(x|z = k, \theta_k\right) p\left(z = k|\theta_z\right)}{\sum_j p(x, z = j|\Theta)} \tag{3.11}$$

$$= \frac{p\left(x|z = k, \theta_k\right) p\left(z = k|\theta_z\right)}{\sum_j p\left(z = j|\theta_z\right) p\left(x|z = j, \theta_j\right)} \tag{3.12}$$

Gaussian mixture models

$$p(x|\theta) = \sum_k \alpha_k \mathcal{N}\left(x|\mu_k, \Sigma_k\right) \tag{3.13}$$

$$\log\left(x_1, x_2, \ldots, x_N\right)|\theta) = \sum_n \log \sum_k \alpha_k \mathcal{N}(x^{(n)}|\mu_k, \Sigma_k) \tag{3.14}$$

$$p(z = k|x, \theta) = \frac{\alpha_k \mathcal{N}\left(x|\mu_k, \Sigma_k\right)}{\sum_j \alpha_j \mathcal{N}\left(x|\mu_j, \Sigma_j\right)} \tag{3.15}$$

Mixtures of experts

$$p(y|x, \theta) = \sum_{k=1}^{K} p\left(z = k|x, \theta_z\right) p\left(y|z = k, x, \theta_K\right) \tag{3.16}$$

$$= \sum_{k=1}^{K} \alpha_k\left(x|\theta_z\right) p_k\left(y|x, \theta_k\right) \tag{3.17}$$

4 Exact Inference

Notation 4.1. Let X denote the set of all random variables in the model, and

1. X_E = The observed evidence;
2. X_F = The unobserved variable we want to infer;
3. $X_R = X - \{X_F, X_E\}$ = Remaining variables, extraneous to query.

The model defines the joint distribution of all random variables:

$$p(X_E, X_F, X_R) \tag{4.1}$$

Definition 4.1. The joint distribution over evidence and subject of inference is

$$p\left(X_F, X_E\right) = \sum_{X_R} p\left(X_F, X_E, X_R\right) \tag{4.2}$$

Definition 4.2. The conditional probability distribution for inference given evidence is

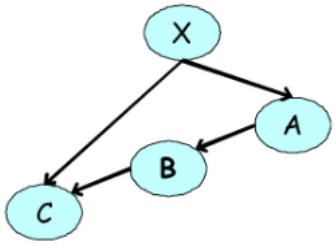
$$p\left(X_F | X_E\right) = \frac{p\left(X_F, X_E\right)}{p\left(X_E\right)} = \frac{p\left(X_F, X_E\right)}{\sum_{X_F} p\left(X_F, X_E\right)} \tag{4.3}$$

Definition 4.3. The distribution of evidence can be computed as

$$p\left(X_E\right) = \sum_{X_F, X_R} p\left(X_F, X_E, X_R\right) \tag{4.4}$$

4.1 Variable Elimination

4.2 Intermediate Factors



$$p(A, B, C) = \sum_X p(X)p(A|X)p(B|A)p(C|B, X) \tag{4.5}$$

$$= p(B|A) \underbrace{\sum_X p(X)p(A|X)p(C|B, X)}_{\text{unnormalized}} \tag{4.6}$$

Definition 4.4. A **factor** ϕ describes the local relation between random variables, meanwhile, $\int d\phi$ is not necessarily one.

Remark 4.1. Let $X_\ell \subseteq X$ be a group of local random variables, then $p(X_\ell)$ is automatically a factor $\phi(X_\ell)$.

$$p(A,B,C) = \sum_X \underbrace{p(X)p(A|X)p(B|A)p(C|B,X)}_{\text{from graphical representation}} \tag{4.7}$$

$$= \sum_X \underbrace{\phi(X)\phi(A,X)\phi(A,B)\phi(X,B,C)}_{\text{factor representation}} \tag{4.8}$$

$$= \phi(A,B) \sum_X \phi(X)\phi(A,X)\phi(X,B,C) \tag{4.9}$$

$$= \phi(A,B)\ \underbrace{\tau(A,B,C)}_{\text{another factor}} \tag{4.10}$$

4.3 Sum-Product Inference

Theorem 4.1. Consider a graphical model with random variables $X = Y \cup Z$. For an random variable Y in a directed or undirected model, $P(Y)$ can be computed using the **sum-product**

$$\tau(Y) = \sum_z \prod_{\phi \in \Phi} \phi(Scope[\phi] \cap Z, Scope[\phi] \cap Y) \tag{4.11}$$

where Φ is a set of factors.

Remark 4.2. For directed models,

$$\Phi = \{\phi_{x_i}\}_{i=1}^N = \{p\left(x_i|\text{ parents }(x_i)\right)\}_{i=1}^N \tag{4.12}$$

4.4 Complexity of Variable Elimination Ordering

Theorem 4.2. The complexity of the variable elimination algorithm is

$$\mathcal{O}(mk^{N_{max}}) \tag{4.13}$$

where

(i) m is the number of initial factors $|\Phi|$;

(ii) k is the number of states each random variable takes, assumed to be equal;

(iii) N_i is the number of random variables within each summation;

(iv) $N_{max} = \max_i N_i$.

5 Message passing, Hidden Markov Models, and Sampling

5.1 Message Passing (Computing All Marginals)

Notation 5.1. Let T denote the set of edges in a tree. For a node i, let $N(i)$ denote the set of its neighbours.

The factor of all random variables can be computed following

$$P\left(X_{1:n}\right)=\frac{1}{Z}\underbrace{\left[\prod_{i=1}^{n}\phi\left(x_i\right)\right]}_{\text{prior factors}}\underbrace{\prod_{(i,j)\in T}\phi_{i,j}\left(x_i,x_j\right)}_{\text{local factors}} \tag{5.1}$$

Definition 5.1. The **message** sent from variable j to $i \in N(j)$ is

$$m_{j\to i}\left(x_i\right)=\sum_{x_j}\left[\phi_j\left(x_j\right)\phi_{ij}\left(x_i,x_j\right)\prod_{k\in N(j)\neq i}m_{k\to j}\left(x_j\right)\right] \tag{5.2}$$

Algorithm 5.1 (Belief Propagation Algorithm)**.** Given a tree, inference on an arbitrary node $p(x_i)$ can be computed following:

1. Choose root r arbitrarily;
2. Pass messages from leaves to r;
3. Pass messages from r to leaves;
4. Compute inference

$$p\left(x_i\right)\propto\phi_i\left(x_i\right)\prod_{j\in N(i)}m_{j\to i}\left(x_i\right) \tag{5.3}$$

5.2 Markov Chains

Using chain rule of probability:

$$p\left(x_{1:T}\right)=\prod_{t=1}^{T}p\left(x_t|x_{t-1},\ldots,x_1\right) \tag{5.4}$$

Definition 5.2. A Markov chain is said to be **first-order** if

$$p\left(x_t|x_{1:t-1}\right)=p\left(x_t|x_{t-1}\right) \tag{5.5}$$

Simplification Therefore, for all first-order Markov chains, the full joint distribution can be reduced to

$$p\left(x_{1:T}\right)=\prod_{t=1}^{T}p\left(x_t|x_{t-1}\right) \tag{5.6}$$

Definition 5.3. A Markov chain is at m-order if

$$p\left(x_t|x_{1:t-1}\right)=p\left(x_t|x_{t-m:t-1}\right) \tag{5.7}$$

Definition 5.4. A Markov chain is said to be **homogenous** (i.e., stationary) if

$$p\left(x_t|x_{t-1}\right)=p\left(x_{t+k}|x_{t-1+k}\right)\quad\forall t,k \tag{5.8}$$

Parameterization Assume the random variable X_t takes k states, further suppose the chain is time homogenous. Then characterizing the transition probability

$$p(x_t|x_{t-1}, x_{t-2}, \cdots, x_{t-m}) \tag{5.9}$$

requires $(k-1)k^m$ parameters.

5.3 Hidden Markov Models

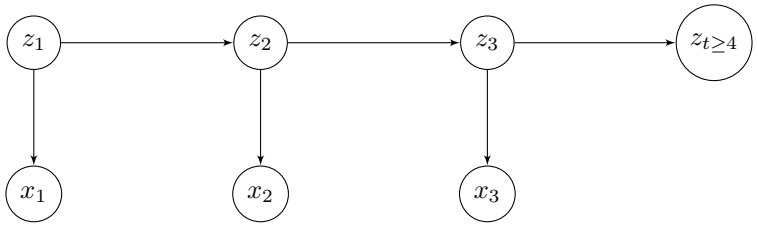


Figure 5.1: Hidden Markov Model

Joint distribution Following the conventional expansion

$$p(x_{1:T}, z_{1:T}) = p(z_1)\prod_{t=2}^{T} p(z_t|z_{t-1})\prod_{t=1}^{T} p(x_t|z_t) \tag{5.10}$$

Parameterization Assuming the HMM is homogenous with order 1, then the set of parameters Φ consists of

(i) Initial distribution of $p(z_1)$: $k-1$ parameters;

(ii) Transition distribution of $p(z_{t+1}|z_t)$: $(k-1)k$ parameters;

(iii) Emission distribution of $p(x_t|z_t)$: $(k-1)k$ parameters.

5.4 Forward-backward Algorithm

Smoothing compute posterior over past hidden state

$$p(z_\tau|x_{1:t}) \ s.t. \ 1 < \tau < t \tag{5.11}$$

Filtering compute posterior over current hidden state

$$p(z_t|x_{1:t}) \tag{5.12}$$

Prediction compute posterior over future hidden state

$$p(z_\tau|x_{1:t}) \ s.t. \ \tau > t \tag{5.13}$$

5.5 Procedure of Smoothing (Forward-backward Algorithm)

$$
\begin{aligned}
p(z_t|x_{1:T}) &\propto p(x_{1:T}, z_t) && (5.14)\\
&= p(z_t, x_{1:t})p(x_{t+1:T}|z_t, x_{1:t}) && (5.15)\\
&= p(z_t, x_{1:t})p(x_{t+1:T}|z_t) && (5.16)\\
&= p(z_t|x_{1:t})p(x_{1:t})p(x_{t+1:T}|z_t) && (5.17)\\
&\propto p(z_t|x_{1:t})p(x_{t+1:T}|z_t) && (5.18)
\end{aligned}
$$

Forward filtering (encoding) Define

$$\alpha_t(z_t) := p(z_t|x_{1:t}) \tag{5.19}$$

Then

$$
\begin{aligned}
p(z_t|x_{1:t}) &\propto p(z_t, x_{1:t}) && (5.20)\\
&= \sum_{z_{t-1}=1}^{k} p(z_{t-1}, z_t, x_{1:t}) && (5.21)\\
&= \sum_{z_{t-1}=1}^{k} p(x_t|z_{t-1}, z_t, x_{1:t-1})p(z_t|z_{t-1}, x_{1:t-1})p(z_{t-1}, x_{1:t-1}) && (5.22)\\
&= \sum_{z_{t-1}=1}^{k} p(x_t|z_t)p(z_t|z_{t-1})p(z_{t-1}, x_{1:t-1}) && (5.23)\\
&= p(x_t|z_t) \sum_{z_{t-1}=1}^{k} p(z_t|z_{t-1})p(z_{t-1}, x_{1:t-1}) && (5.24)\\
& && (5.25)
\end{aligned}
$$

Therefore, we have the forward recursion:

$$\alpha_t(z_t) = p(x_t|z_t) \sum_{z_{t-1}=1}^{k} p(z_t|z_{t-1})\alpha_{t-1}(z_{t-1}) \tag{5.26}$$

$$\alpha_1(z_1) = p(x_1, z_1) \tag{5.27}$$

Backward filtering (decoding) Define

$$\beta_t(z_t) = p(x_{t+1:T}|z_t) \tag{5.28}$$

Then,

$$p(x_{t+1:T}|z_t) = \sum_{z_{k+1}=1}^{k} p(x_{t+1:T}, z_{t+1}|z_t) \tag{5.29}$$

$$= \sum_{z_{k+1}=1}^{k} p(x_{t+1}, x_{t+2:T}, z_{t+1}|z_t) \tag{5.30}$$

$$= \sum_{z_{k+1}=1}^{k} p(x_{t+2:T}, z_{t+1}|z_t, x_{t+1})p(x_{t+1}|z_t) \tag{5.31}$$

$$= \sum_{z_{k+1}=1}^{k} p(x_{t+2:T}, z_{t+1}|z_t, x_{t+1})p(x_{t+1}|z_t) \tag{5.32}$$

$$= \sum_{z_{k+1}=1}^{k} p(x_{t+2:T}, z_{t+1}|z_t)p(x_{t+1}|z_t) \tag{5.33}$$

$$= \sum_{z_{k+1}=1}^{k} p(x_{t+2:T}|z_t, z_{t+1})p(z_{t+1}|z_t)p(x_{t+1}|z_t) \tag{5.34}$$

$$= \sum_{z_{k+1}=1}^{k} p(x_{t+2:T}|z_{t+1})p(z_{t+1}|z_t)p(x_{t+1}|z_t) \tag{5.35}$$

$$= \sum_{z_{k+1}=1}^{k} \beta_{t+1}(z_{t+1})p(z_{t+1}|z_t)p(x_{t+1}|z_t) \tag{5.36}$$

5.6 Sampling

Problem 1 Generate samples

$$\{x^{(r)}\}_{r=1}^{R} \sim p(x) \tag{5.37}$$

Problem 2 Estimate expectations of functions $f(x)$ taking random variable $x \sim p(x)$.

$$\mathbb{E}_{x\sim p(x)} f(x) = \int f(x)p(x)\ dx \tag{5.38}$$

$$\approx \hat{E} = \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}) \tag{5.39}$$

Ancestral sampling sampling in a topological order. At each step, sample from any conditional distribution that you haven't visited yet, whose parents have all been sampled. This procedure will always start with the nodes that have no parents (i.e., a root).

Generating marginal samples for nodes $Y \subseteq X$:

(i) Construct $p(X) = \prod_{x_i \in X} p(x_i|\text{parent}[x_i])$;

(ii) Sample $(x_1, x_2, \cdots, x_N) \sim p(X)$;

(iii) Ignore $x_i \notin Y$.

Generating conditional samples for nodes $Y \subseteq Xt$ conditioned on $Z \subseteq X$:

Definition 5.5. The **simple Monte Carlo** estimator $\hat{\Phi}$ is defined as

$$\frac{1}{R}\sum_{r=1}^{R} f\left(x^{(r)}\right) = \hat{E} \approx E = \mathop{\mathbb{E}}_{x\sim p(x)}[f(x)] \tag{5.40}$$

Proposition 5.1. If the sample $\{x^{(r)}\}_{r=1}^{R}$ are generated from $p(x)$, then $\hat{E}$ is an unbiased estimator of E.

Proof.

$$\begin{aligned}
\mathbb{E}[\hat{E}]_{x\sim p\left(\{x^{(i)}\}_{r=1}^{R}\right)} &= \mathbb{E}\left[\frac{1}{R}\sum_{r=1}^{R} f\left(x^{(r)}\right)\right] && (5.41)\\
&= \frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\left[f(x^{(r)})\right] && (5.42)\\
&= \frac{1}{R}\sum_{r=1}^{R}\mathbb{E}_{x\sim p(x)}[f(x)] && (5.43)\\
&= \frac{R}{R}\mathbb{E}_{x\sim p(x)}[f(x)] && (5.44)\\
&= E && (5.45)
\end{aligned}$$

■

Proposition 5.2. As the number of samples of R increases, the variance of $\hat{E}$ will decrease proportional to $\frac{1}{R}$.

Proof.

$$\begin{aligned}
Var[\hat{E}] &= Var[\frac{1}{R}\sum_{r=1}^{R} f(x^{(r)})] && (5.46)\\
&= \frac{1}{R^2}Var[\sum_{r=1}^{R} f(x^{(r)})] && (5.47)\\
&= \frac{1}{R^2}\sum_{r=1}^{R} Var[f(x^{(r)})] && (5.48)\\
&= \frac{1}{R^2}RVar[f(x)] && (5.49)\\
&= \frac{1}{R}Var[f(x)] && (5.50)
\end{aligned}$$

■

6 True Skill

Let $z_i \in \mathbb{R}$ denote player's skill such that

$$z_i \sim \mathcal{N}(0,1) \tag{6.1}$$

And skills of players are independent such that

$$p(z_{1:n}) = \prod_{i=1}^{n} p(z_i) \tag{6.2}$$

Therefore, the joint distribution of skills can be written as a multivariate Gaussian distribution.
Let X_j player A beats player B, then

$$p(\text{A beat B}|z_A, z_B) = \sigma(z_A - z_B) \tag{6.3}$$

where $\sigma = 1/(1 + \exp(z))$.
Then,

$$p(z_1, z_2|\text{A beat B}) \propto p(z_1, z_2)p(\text{A beat B}|z_1, z_2) \tag{6.4}$$

Question 1 What's the chance that player A is better than player B in game $\mathcal{G}$?

$$p(z_A > z_B|\mathcal{G}) = \mathbb{E}_{p(z_A, z_B|\mathcal{G})}\mathbb{1}\{z_A > z_B\} \tag{6.5}$$

$$\approx \frac{1}{K}\sum_{i=1}^{K} \mathbb{1}\{z_A > z_B\} \quad z_A, z_B \sim p(z_A, z_B|\text{data}) \tag{6.6}$$

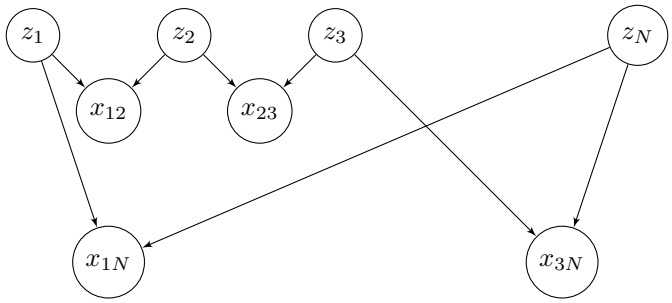


Figure 6.1: Structure of Games

Approximate Inferences We have $p(z), p(x|z)$ and x (data).
Wish to compute

$$q(z|x) \approx p(z|x) \tag{6.7}$$

Intractability

$$p(z_1|x) = \int_{\mathbb{R}} p(z_1, z_2|x)\ dz_2 \tag{6.8}$$

$$= \frac{\int_{\mathbb{R}} p(z_1, z_2, x)\ dz_2}{\iint_{\mathbb{R}^2} p(z_1, z_2, x)\ dz_2\ dz_1} \tag{6.9}$$

6.1 Variational Inferences

General Idea of Variational / Approximate Inferences

(i) Introduce a tractable **variational family** $q_\varphi(z|x)$ parameterized using φ, for example:

$$q_\varphi(z|x) = \mathcal{N}(z|\mu_\varphi, \Sigma_\varphi) \tag{6.10}$$

(ii) Define distance (not necessary a metric) between $p(z|x)$ and $q(z|x)$.

(iii) Optimize φ to minimize distance, that is,

$$\mathbb{E}_{z\sim p(z|x)}[f(z)] \approx \mathbb{E}_{z\sim q(z|x)}[f(z)] \tag{6.11}$$

6.2 Kullback–Leibler Divergence

Definition 6.1. Let q and p be density functions of two distributions Q and P, then the **KL divergence** between Q and P is defined as

$$KL(q||p) := \mathbb{E}_q\left(\log\frac{q}{p}\right) \tag{6.12}$$

$$= \int q(z)\left(\log q(z) - \log p(z)\right)\, dz \tag{6.13}$$

Properties of KL divergence

$$KL(q||p) \geq 0 \ \forall p, q \in \Delta(\mathcal{X}) \tag{6.14}$$

$$KL(q||p) = 0 \iff p = q \tag{6.15}$$

$$KL(q||p) \neq KL(p||q) \text{ in general} \tag{6.16}$$

Reverse-KL Information Projection while comparing two distributions, $D_{KL}(q||p) = \mathbb{E}_q \log\frac{q}{p}$ penalizes (i.e., becomes large) when there are points such that $q \gg p$.

Forward-KL Moment Projection $D_{KL}(p||q)$ penalizes points when $p \gg q$.

6.3 Stochastic Variational Inference

Evidence Lower Bound Suppose we are trying to approximate $p(z|x)$ using variational family $q_\varphi(z|x)$. Define

$$D_{KL}(q_\varphi(z|x)||p(z|x)) = \mathbb{E}_{q_\varphi(z|x)}\left[\log q_\varphi(z|x) - \log p(z|x)\right] \tag{6.17}$$

$$= \mathbb{E}_{q_\varphi(z|x)}\left[\log q_\varphi(z|x) - \log\frac{p(z,x)}{p(x)}\right] \tag{6.18}$$

$$= \mathbb{E}_{q_\varphi(z|x)}\left[\log q_\varphi(z|x) - \log p(z,x) + \log p(x)\right] \tag{6.19}$$

$$= \mathbb{E}_{q_\varphi(z|x)}\left[\log q_\varphi(z|x) - \log p(z,x)\right] + \mathbb{E}_{q_\varphi(z|x)}\left[\log p(x)\right] \tag{6.20}$$

$$= \underbrace{\mathbb{E}_{q_\varphi(z|x)}\left[\log q_\varphi(z|x) - \log p(z,x)\right]}_{-\mathcal{L}(\varphi;x)} + \underbrace{\log p(x)}_{\perp\varphi} \tag{6.21}$$

where $\mathcal{L}(\varphi; x)$ is often referred to as the evidence/empirical lower bound (ELBO). Therefore,

$$D_{KL}(q_\varphi(z|x)||p(z|x)) = -\mathcal{L}(\varphi; x) + \log p(x) \tag{6.22}$$

$$\implies \mathcal{L}(\varphi; x) + D_{KL}(q_\varphi(z|x)||p(z|x)) = \log p(x) \quad (\dagger) \tag{6.23}$$

$$\implies \mathcal{L}(\varphi; x) \leq \log p(x) \tag{6.24}$$

Hence, $\mathcal{L}(\varphi; x)$ is called ELBO because it is a lower bound of log-likelihood of the evidence x. By (†), maximizing ELBO and minimizing KL divergence are equivalent.

Re-parameterization Tricks In order to maximize ELBO, we want to compute the gradient of ELBO with respect to φ:

$$\nabla_\varphi \mathcal{L}(\varphi; x) = \nabla_\varphi \mathbb{E}_{z\sim q_\varphi(z|x)}[\log p(z, x) - \log q_\varphi(z|x)] \tag{6.25}$$

Interpretation of ELBO ELBO can be further rewritten as

. (6.26)

Pathwise Gradient More generally, let f be an arbitrary function, in our case, $f(z) = \log q_\varphi(z|x) - p(z, x)$. The gradient is

$$\nabla_\varphi \mathbb{E}_{q_\varphi(z)} f(z) = \nabla_\varphi \int q_\varphi(z) f(z)\ dz \tag{6.27}$$

$$= \int \nabla_\varphi q_\varphi(z) f(z)\ dz\ (\dagger) \text{ assuming continuity} \tag{6.28}$$

But (†) is not an expectation, so we cannot use Monte Carlo.
We can exchange differential operator and expectation only if the expectation is independent from φ.
So we need to factor out the randomness in f from q_φ, and put it into a parameterless form. That is, we have to re-parameterize using another random variable ε such that the expectation does not depend on φ.
Find $p(\varepsilon)$ and $T(\varepsilon, \varphi)$ such that

$$\begin{cases} \varepsilon \sim p(\varepsilon) \\ z = T(\varepsilon, \varphi) \end{cases} \implies z \sim q_\varphi(z) \tag{6.29}$$

Example 6.1. For instance, if we are trying to re-parameterize $q_{\varphi\equiv(\mu,\sigma)} = \mathcal{N}(\mu, \sigma^2)$ using a standard Gaussian noise $\mathcal{N}(0, 1)$.

$$\varepsilon \sim p(\varepsilon) = \mathcal{N}(0, 1) \tag{6.30}$$

$$z = T(\varepsilon, \varphi) = \sigma\varepsilon + \mu \tag{6.31}$$

so that $z \sim \mathcal{N}(\mu, \sigma^2)$.

Using the re-parameterization trick, the gradient can be computed as

$$\nabla_\varphi \mathcal{L}(\varphi; x) = \nabla_\varphi \mathbb{E}_{z\sim q_\varphi(z|x)}[\log p(z, x) - \log q_\varphi(z|x)] \tag{6.32}$$

$$= \nabla_\varphi \mathbb{E}_{z\sim q_\varphi(z|x)}[\log p(T(\varepsilon, \varphi), x) - \log q_\varphi(T(\varepsilon, \varphi)|x)] \tag{6.33}$$

$$= \nabla_\varphi \mathbb{E}_{\varepsilon\sim p(\varepsilon)}[\log p(T(\varepsilon, \varphi), x) - \log q_\varphi(T(\varepsilon, \varphi)|x)] \tag{6.34}$$

where the last step holds because the only source of randomness in the expression comes from ε. Since the expectation is no longer depending on φ, and we can therefore exchange the order of differentiation and expectation:

$$\nabla_\varphi \mathcal{L}(\varphi; x) = \mathbb{E}_{\varepsilon \sim p(\varepsilon)} \nabla_\varphi [\log p(T(\varepsilon, \varphi), x) - \log q_\varphi(T(\varepsilon, \varphi)|x)] \tag{6.35}$$

Stochastic/Automatic Differentiation Variational Inference As long as we code up

$$T(\epsilon, \varphi), \log p(x|z), \log p(z), \log q_\varphi(z|x) \tag{6.36}$$

We can use automatic differentiation techniques to compute gradient

$$\nabla_\varphi [\log p(T(\varepsilon, \varphi), x) - \log q_\varphi(T(\varepsilon, \varphi)|x) \tag{6.37}$$

Then estimate the expectation of gradient by Monte Carlo:

$$\mathbb{E}_{p(\varepsilon)} \nabla_\varphi \left[\log p(x, T(\varepsilon, \varphi)) - \log q_\varphi(T(\varepsilon, \varphi)|x)\right] \approx \frac{1}{K} \sum_{i=1}^{K} \nabla_\varphi \left[\log p(x, T(\varepsilon, \varphi)) - \log q_\varphi(T(\varepsilon, \varphi)|x)\right] \tag{6.38}$$

Interpretation of ELBO

Reinforce / Score Function Estimation Another method to estimate $-\nabla_\varphi \mathbb{E}_{q_\varphi(z|x)} L(\varphi)$.
Want

$$\nabla_\varphi \mathbb{E}_{q_\varphi(z)}[f(z)] \tag{6.39}$$

$$= \nabla_\varphi \int q_\varphi(z) f(z)\ dz \tag{6.40}$$

$$= \int \nabla_\varphi q_\varphi(z) f(z)\ dz \tag{6.41}$$

$$= \int f(z) \nabla_\varphi q_\varphi(z)\ dz \tag{6.42}$$

Note that

$$\nabla_\varphi \log q_\varphi(z) = \frac{\nabla_\varphi q_\varphi(z)}{q_\varphi(z)} \tag{6.43}$$

Therefore,

$$\int f(z) \nabla_\varphi q_\varphi(z)\ dz \tag{6.44}$$

$$= \int f(z) q_\varphi(z) \nabla_\varphi \log q_\varphi(z)\ dz \tag{6.45}$$

$$= \mathbb{E}_{q_\varphi(z)}[f(z) \nabla_\varphi \log q_\varphi(z)] \tag{6.46}$$

Mean-Field Variational Inference

$$q(z_1, \cdots, z_N|x) = \prod_{i=1}^{N} q(z_i|x) \tag{6.47}$$

Jensen's Inequality

Theorem 6.1. If f is convex, then

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(x)] \tag{6.48}$$

Using Jensen's inequality,

$$\log p_\theta(x) = \log \int p_\theta(x,z)\ dz \tag{6.49}$$

$$= \log \int p_\theta(x,z) \frac{q_\varphi(z|x)}{q_\varphi(z|x)}\ dz \tag{6.50}$$

$$= \log \mathbb{E}_{q_\varphi(z|x)} \int \frac{p_\theta(x,z)}{q_\varphi(z|x)}\ dz \tag{6.51}$$

$$\geq \mathbb{E}_{q_\varphi(z|x)} \int \log \frac{p_\theta(x,z)}{q_\varphi(z|x)}\ dz \tag{6.52}$$

$$\equiv ELBO(\theta, \varphi) \tag{6.53}$$

7 Sampling and Monte Carlo Methods

Problem With a target distribution

$$p(x) \tag{7.1}$$

and a function

$$f(x) \tag{7.2}$$

Wish to compute

$$\mathbb{E}_{x\sim p}[f] \equiv \int f(x)p(x)\ dx \tag{7.3}$$

Note that $\int_{p(x)}$ is rarely analytic, so we have to approximate it. However, drawing samples from $p(x)$ directly is infeasible since we cannot evaluate $p(x)$. Fortunately, we can evaluate $\tilde{p}(x) = Zp(x)$ without knowing Z.

7.1 Monte Carlo

Sampling

$$\{x_i\}_{i=1}^R \sim p(x) \tag{7.4}$$

Simple MC estimator

$$\Phi = \mathbb{E}_{x\sim p(x)}[f] \approx \hat{\Phi} = \frac{1}{R}\sum_{i=1}^{R} f(x_i) \tag{7.5}$$

Therefore, Monte Carlo turns estimation problem into a sampling problem. Note that MC estimator is unbiased and variance shrinks proportional to $\frac{1}{R}$.
Further, the variance is independent of dimensionality.

Note that sampling from high dimensions is still hard. Why? So far, we've assumed

$$p(x) = \frac{\tilde{p}(x)}{Z} = \frac{\tilde{p}(x)}{\int d\tilde{p}(x)} \tag{7.6}$$

$$p(x) \propto \tilde{p}(x) \tag{7.7}$$

Aside No that $\tilde{p}$ is the joint distribution of $q(pred; data)$. In this case, p could be the likelihood $q(pred|data) = \frac{q(pred,data)}{\int_{pred} dq(pred,data)}$. $Z = p(data)$ is often referred to as the **evidence**.

Why is sampling $x \sim p$ hard? Firstly, in most cases, we don't know the exact value of Z, and computing Z by integrating $\tilde{p}$ requires prohibitive computational power. Event if we know Z, we wish so sample from $\mathcal{X}$ where $p(x)$ is large. But for most distributions, we can't do exactly without enumerating all possible x, which is challenging when x is in high dimensional spaces.

Easy Distributions? For some families of distributions like Gaussian distribution, sampling from them can be easy.

7.1.1 Lattice Discretization (Bad idea!)

Algorithm 7.1 (Lattice Discretization)**.**

(i) Discretize $\mathcal{X}$ into finitely many uniformly spaced points $\{x_i\}$. The collection of points form a lattice of $\mathcal{X}$.

(ii) Evaluate $\tilde{p}(x_i)$ on each point x_i.

(iii) Estimate normalizing constant $Z \approx \sum_i \tilde{p}(x_i)$.

(iv) Estimate probability using $p(x) \approx \frac{\tilde{p}(x)}{Z}$.

Using lattice to sample is a bad idea, if the lattice is too coarse, the estimated distribution is biased. If the lattice is too dense, then we're wasting computational resources. Further, the number of points required as well as computational cost grow exponentially as the dimension $\mathcal{X}$ increases.

7.2 Monte Carlo Methods for Sampling

7.2.1 Uniform Sampling

7.2.2 Importance Sampling

Importance Weight Sampler Sampling from p is hard, we sample x^r from a simpler distribution q (e.g., $p \sim \mathcal{N}$) instead. Weight each sample from q by $w_r = \frac{\tilde{p}(x^r)}{q(x^r)}$. Recall that we know how to evaluate the unnormalized density $\tilde{p}$.

See more on the next section

7.2.3 Rejection Sampling

Rejection Sampling Assume

$$\tilde{p}(x) = p(x)Z_p \;\; Z_p \in \mathbb{R} \tag{7.8}$$

$$\tilde{q}(x) = q(x)Z_q \;\; Z_q \in \mathbb{R} \tag{7.9}$$

and we can evaluate $\tilde{q}$ cheaply and we can sample from q easily. Further, assume

$$\exists\ c\ s.t.\ c\tilde{q}(x) > \tilde{p}(x) \quad \forall x \tag{7.10}$$

Algorithm 7.2 (Rejection Sampling)**.**

(i) Generate two random numbers (x, u).

  (a) Sample $x \sim q$;

  (b) Sample $u \sim Unif[0, c\tilde{q}(x)]$;

(ii) Evaluate $\tilde{p}(x)$

  (a) if $u > \tilde{p}(x)$, reject this sample;

  (b) Otherwise x is accepted and added to our set of samples $\{x^{(r)}\}$.

Intuition Note that all samples are independently sampled from an identical distribution. Therefore, probability for a particular value of x to be added to the sample is

$$Prob(x) = q(x)Prob(Accept|x) \tag{7.11}$$

$$= q(x)\frac{\tilde{p}(x)}{c\tilde{q}(x)} \tag{7.12}$$

$$= \frac{\tilde{p}(x)}{Z_q c} \propto p(x) \tag{7.13}$$

Limitations For rejection sampling, if the two distributions p and q are not similar, c have to be large so that $c\tilde{q}$ covers $\tilde{q}$ everywhere in $\mathcal{X}$. A large value fo c makes acceptance rare.

Limitations The rejection sampling requires $p \approx q$. Otherwise, the C required is large. When p and q are on D dimensions, then $C \in \mathcal{O}(\exp(\sqrt{D}))$. The acceptance rate is $\frac{1}{C}$ in one dimension, and the rate reduces exponentially in dimension.

$$\mathbb{E}_p[f] = \int f(x)p(x)\ dx \tag{7.14}$$

Since integral is a linear operator, when integrant is large, the expectation is large as well.

7.2.4 Metropolis-Hastings Method

Typical set Because expectation is a linear operator, all samples that contribute significantly to the $\mathbb{E}$ come from the typical set.

Check the definition of typical set.

Markov Chain Monte Carlo (MCMC) Stochastically explore the typical set. We need a Markov transition distribution, from which we can sample x_{t+1}:

$$x_{t+1} \sim T(x'|x_t) \tag{7.15}$$

Proposition 7.1. Properties of T 1) p is invariant under T:

$$p(x) = \int T(x|x')p(x')\ dx' \tag{7.16}$$

2) Ergodic: $p^{(t)} \to p(x)$.

Metropolis Given background distribution p, we only observe $\tilde{p}$ but not p. Define T as

(i) Proposal distribution q, which is not necessarily similar to p.

$$x' \sim q(x'|x_t) \tag{7.17}$$

(ii) Accept x' as x_{t+1}?

$$a := \frac{\tilde{p}(x')q(x_t|x')}{\tilde{p}(x_t)q(x'|x_t)} \tag{7.18}$$

(a) If $a \geq 1$ accept $x_{t+1} \leftarrow x'$.

(b) If $a < 1$, accept $x_{t+1} \leftarrow x'$ with probability a; otherwise, reject and set $x_{t+1} \leftarrow x_t$.

Random Walk Metropolis

$$q(x'|x_t) = \mathcal{N}(x'|x_t, \sigma^2) \tag{7.19}$$

Hamiltonian Monte Carlo

$$x \to (x, v) \tag{7.20}$$

$$q(x) \to q(x, v) = q(x)q(v|x) \tag{7.21}$$

$$H(x, v) = -\log(q(x, v)) = -\overbrace{\log(q(x))}^{U:PE} - \overbrace{\log(q(v|x))}^{K:KE} \tag{7.22}$$

7.3 Tutorial on Sampling

Problem Say one wishes to sample

$$x \sim p(x) = \frac{\tilde{p}(x)}{Z} \tag{7.23}$$

such that

$$\int p(x)\ dx = 1 \tag{7.24}$$

$$\int \tilde{p}(x)\ dx = Z \tag{7.25}$$

$$\tilde{p}(x), p(x) \geq 0 \tag{7.26}$$

Goal For an arbitrary function ϕ, we wish to estimate

$$\mathbb{E}_{x\sim p(x)}\phi(x) = \int p(x)\phi(x)\ dx \tag{7.27}$$

by sampling $x \sim p(x)$.

Challenge We only observe the potential $\tilde{p}(x)$ but not the actual probability $p(x)$. In addition, marginalizing the potential (i.e., computing $\int_x \tilde{p}(x)\ dx$) is prohibitive, so that one may not reconstruct the probability $p(x)$ directly.

Cheap Sampling We may assume sampling from a certain class of distributions to be cheap. For instance, we can sample from a Gaussian distribution cheaply:

$$x \sim \mathcal{N}(\mu, \sigma^2) \tag{7.28}$$

Idea For most of sampling techniques covered in this course, we are sampling from an "auxiliary potential", say $\tilde{q}(x)$, from which we can sample with ease. Then, we estimate $p(x)$ via sampling result from $\tilde{q}$ and our knowledge on $\tilde{p}$.

7.3.1 Importance Sampling

Problem Suppose

$$\tilde{p}(x) = Z_p p(x) \tag{7.29}$$
$$\tilde{q}(x) = Z_q q(x) \tag{7.30}$$

where the normalizing constants are unknown. And we wish to estimate

$$\mathbb{E}_{x\sim p(x)}\phi(x) \tag{7.31}$$

for some function ϕ.

Solution We may take $\tilde{q}(x) = \mathcal{N}(\mu, \sigma^2)$, in this case, $Z_q = 1$.

[?] Do we know both $\tilde{q}$ and Z_q or we only know q?

Algorithm 7.3 (Importance Sampling)**.**

(i) Sample $X^{1:R} \sim Z_q q(x) = \tilde{q}(x)$.

(ii) Estimate expectation under q: $\mathbb{E}_q(\phi(x)) \approx \frac{1}{R}\sum_{i=1}^{R}\phi(x_i)$.

(iii) Weight each sample using importance weights $\tilde{w}_i = \frac{\tilde{p}(x_i)}{\tilde{q}(x_i)}$.

Intuition For one particular sample x, if $\tilde{p}(x) > \tilde{q}(x)$, $\tilde{p}$ is underrepresented while sampling using $\tilde{q}$, so we assign this sample with larger weight w.

Estimation Recall that the ultimate goal of sampling exercises is to estimate $\mathbb{E}_p\phi(x)$:

$$\begin{aligned}
\mathbb{E}_p\phi(x) &= \int p(x)\phi(x)\ dx && (7.32)\\
&= \int q(x)\frac{p(x)}{q(x)}\phi(x)\ dx && (7.33)\\
&= \mathbb{E}_q\frac{p(x)}{q(x)}\phi(x) && (7.34)\\
&\approx \frac{1}{R}\sum_{i=1}^{R}\underbrace{\frac{p(x_i)}{q(x_i)}}_{(\dagger)}\phi(x_i) \text{ where } x_i \sim q(x) && (7.35)
\end{aligned}$$

Note that we only have the unnormalized version of (†), therefore,

$$... = \frac{1}{R}\sum_{i=1}^{R}\frac{\tilde{p}(x_i)/Z_p}{\tilde{q}(x_i)/Z_q}\phi(x_i) \tag{7.36}$$

$$= \frac{Z_q}{Z_p}\frac{1}{R}\sum_{i=1}^{R}\frac{\tilde{p}(x_i)}{\tilde{q}(x_i)}\phi(x_i) \tag{7.37}$$

$$= \frac{Z_q}{Z_p}\frac{1}{R}\sum_{i=1}^{R}\tilde{w}_i\phi(x_i) \quad (\dagger\dagger) \tag{7.38}$$

Lemma 7.1.

$$\sum_{i=1}^{R}\tilde{w}_i = \sum_{i=1}^{R}\frac{\tilde{p}(x_i)}{\tilde{q}(x_i)} = \sum_{i=1}^{R}\frac{\tilde{p}(x_i)\frac{Z_q}{Z_p}}{\tilde{q}(x_i)\frac{Z_q}{Z_p}} \text{ where } x_i \sim q(x) \tag{7.39}$$

$$= \sum_{i=1}^{R}\frac{p(x_i)Z_q}{\tilde{q}(x_i)\frac{Z_q}{Z_p}} = \sum_{i=1}^{R}\frac{p(x_i)Z_q}{q(x_i)Z_q\frac{Z_q}{Z_p}} \tag{7.40}$$

$$= \sum_{i=1}^{R}\frac{p(x_i)}{q(x_i)\frac{Z_q}{Z_p}} = \frac{Z_p}{Z_q}\sum_{i=1}^{R}\frac{p(x_i)}{q(x_i)} \tag{7.41}$$

$$= R\frac{Z_p}{Z_q}\frac{1}{R}\sum_{i=1}^{R}\frac{p(x_i)}{q(x_i)} \tag{7.42}$$

$$\overset{R\to\infty}{=} R\frac{Z_p}{Z_q}\mathbb{E}_q\frac{p(x)}{q(x)} \tag{7.43}$$

$$= R\frac{Z_p}{Z_q}\int_{\mathbb{R}} q(x)\frac{p(x)}{q(x)}\,dx = R\frac{Z_p}{Z_q} \tag{7.44}$$

Therefore,

$$\frac{Z_p}{Z_q} = \frac{1}{R}\sum_{i=1}^{R}\tilde{w}_i \tag{7.45}$$

Hence,

$$\mathbb{E}_p\phi(x) \approx (\dagger\dagger) = \frac{\sum_{i=1}^{R}\tilde{w}_i\phi(x_i)}{\sum_{i=1}^{R}\tilde{w}_i} \tag{7.46}$$