CSE6243: Advanced Machine Learning **Fall 2024**

Lecture 7: Basic, Acceptance & Rejection, and Importance Sampling

Lecturer: Bo Dai *Scribes: Max Zhang, Dongquan Shen*

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

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

7.1 Recap

7.1.1 Exponential Family Examples

1. Bernoulli:

$$p(x) = \pi^x (1-\pi)^{1-x} \quad \text{where } \pi \in [0,1]$$

We can rewrite as:

$$p(x) = \exp\left(x \log(\frac{\pi}{1-\pi}) + \log(1-\pi)\right)$$

We can see that Bernoulli fits Exponential Family canonical form where:

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

2. Gaussian:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

We can rewrite as:

$$p(x) = \exp\left(-\frac{x^2 - 2x\mu + \mu^2}{2\sigma^2} - \log(\sqrt{2\pi\sigma})\right)$$

We can see that Gaussian fits the Exponential Family canonical form where:

$$T(x) = [x^2, x], \quad h(x) = 1, \quad \eta = [-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}], \quad A(\eta) = -\frac{\mu^2}{2\sigma^2} - \log(\sqrt{2\pi\sigma^2})$$

7.1.2 Properties of Log-Partition Function A(η)

1. Convexity of A(η):

$$A(t\eta_1 + (1-t\eta_2) \leq tA(\eta_1) + (1-t)A(\eta_2)$$

2. Gradient of A(η):

$$\frac{dA(\eta)}{d\eta} = \frac{1}{Q(\eta)}\frac{\partial Q(\eta)}{\partial \eta} = \mathbb{E}_{P_\eta(n)}[T(x)]$$

7.2 New Content

7.2.1 Frequentist Learning

The frequentist views θ as some unknown parameter. We don't know what θ is, but we can formulate a hypothesis for a potential θ, acquire some data $D = \{x_i, y_i\}_{i=1}^n$, and determine the probability of observing this data if the hypothesis was true. The method of frequentist inference typically involves solving some regression. For example, suppose we were doing linear least squares, so we are trying to find

$$\theta^* = \arg\min_\theta \frac{1}{n} \sum_{i=1}^n \left\| y_i - \theta^{\mathrm{T}} x_i \right\|_2^2 + \lambda \|\theta\|_2^2$$

$\lambda\|\theta\|_2^2$ is a regularization term that is used to force the solution to be unique when your system might otherwise be undertermined (i.e. you have more entries in θ than data points).

This is an optimization problem, and we could solve it with some method such as maximum likelihood estimation (MLE). If we were using MLE, then we want to find the θ that maximizes the probability of observing all data points, which would be

$$\prod_{i=1}^n p(y_i \mid x_i, \theta)$$

We apply the monotonic function $\log(x)$ to the objective to turn it from products to sums, and we negate it to flip from a maximization to minimization problem, yielding the objective

$$-\log \prod_{i=1}^n p(y_i \mid x_i, \theta) = -\sum_{i=1}^n \log p(y_i \mid x_i, \theta)$$

And given some x_{test}, we can predict the corresponding y_{test} as

$$y_{test} = (\theta^*)^{\mathrm{T}} x_{test}$$

7.2.2 Bayesian Learning

The Bayesian views θ not simply as a parameter we are seeking, but a random variable that follows some probability distribution. The method of Bayesian inference is as follows:

- We begin with some prior distribution $p(\theta)$ we believe in. This could start as some uninformative prior.
- We are presented with new evidence $D = \{x_i, y_i\}_{i=1}^n$. Our evidence will inform our new beliefs via some likelihood, which represents how likely it was to witness this data assuming our prior. The ith data point has likelihood $p(y_i \mid x_i, \theta)$. The collective likelihood would therefore be $\prod_{i=1}^n p(y_i \mid x_i, \theta)$.
- We account for the marginal likelihood representing the probability of observing this evidence in general. This can be computed as summing up the probability of observing this evidence for a certain parameter over all parameters, which would be $\int \prod_{i=1}^n p(y_i \mid x_i, \theta) p(\theta)\, d\theta$.

Plugging all of this into Bayes' rule tells us how to update the posterior distribution via Bayesian inference. Specifically, the posterior probability for a particular θ would be

$$p(\theta \mid \{x_i, y_i\}_{i=1}^n) = \frac{\prod_{i=1}^n p(y_i \mid x_i, \theta)p(\theta)}{\int \prod_{i=1}^n p(y_i \mid x_i, \theta)p(\theta)\, d\theta}$$

And thus given some x_{test}, we can predict the corresponding y_{test} as

$$\mathbb{E}[y \mid x_{test}, D] = \int y \cdot p(y \mid x_{test}, D)\, dy = \iint p(y \mid x_{test}, \theta)p(\theta \mid D)\, d\theta\, dy$$

Notably, this double integral $\iint p(y \mid x_{test}, \theta)p(\theta \mid D)\, d\theta\, dy$ can be very computationally intensive or outright intractable to compute. This motivates the need for sampling, where we will sample a few points from the distribution $p(\theta)$ to approximate this integral rather than computing it directly.

7.2.3 Monte Carlo Approximation

When discussing sampling, let us consider a more general problem: given some arbitrary probability distribution $p(x)$ and some arbitrary function $f(x)$, we want to approximate

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

The idea is we want to sample points $\{x_i\}_{i=1}^k \sim p(x)$ such that $\frac{1}{k}\sum_{i=1}^k f(x_i) \approx \mathbb{E}_p[f(x)]$. This is known as a **Monte-Carlo approximation**.

For it to be a good sample, we need it to adhere to the following properties:

1. **It is an unbiased estimator.** That is,

$$\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^k f(x_i)\right] = \frac{1}{k}\sum_{i=1}^k \mathbb{E}[f(x_i)] = \frac{1}{k}\sum_{i=1}^k \mathbb{E}_p[f(x)] = \frac{1}{k} \cdot k \cdot \mathbb{E}_p[f(x)] = \mathbb{E}_p[f(x)]$$

   When we say that $\mathbb{E}[f(x_i)] = \mathbb{E}_p[f(x)]$, we are necessitating that the x_i are *independent and identically distributed* random variables.

2. **It adheres to the *law of large numbers*.** That is,

$$\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^k f(x_i) \overset{a.s.}{=} \mathbb{E}_p[f(x)]$$

   The $\overset{a.s.}{=}$ refers to "almost surely". More formally, we require that

$$P\left(\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^k f(x_i) - \mathbb{E}_p[f(x)] = 0\right) = 1$$

3. **The variance adheres to the following:**

$$\text{Var}\left[\frac{1}{k}\sum_{i=1}^k f(x_i)\right] = \frac{1}{k^2}\text{Var}\left[\sum_{i=1}^k f(x_i)\right] = \frac{1}{k^2}\sum_{i=1}^k \text{Var}[f(x_i)] = \frac{1}{k^2} \cdot k \cdot \text{Var}[f(x)] = \frac{1}{k}\text{Var}[f(x)]$$

   Intuitively, this is saying that as k grows large, the variance tends to 0, so our approximation is more accurate with more samples. And again, we are necessitating that the x_i are *independent and identically distributed* random variables.

Note: *For all of the sampling methods discussed, we suppose we have some way of sampling from a uniform distribution $U[0,1]$. This could be via a pseudorandom number generator for example.*

7.2.4 Inverse Probability Transform

Suppose we have access to both the *cumulative distribution function* $F(z) = \int_{-\infty}^{z} p(x)\,dx$ and its inverse function $F^{-1}(z)$ for the probability distribution $p(x)$ from which we want to sample. The steps of the sampling algorithm are then as follows:

1. Sample $\mu \sim U[0,1]$.
2. We claim that $F^{-1}(\mu) \sim p(x)$.

Proof: To see why this is true, note that $P(F^{-1}(\mu) \leq z) = P(F(F^{-1}(\mu)) \leq F(z)) = P(\mu \leq F(z))$ as $F(z)$ is a monotonically increasing function (by virtue of it being a CDF). And $P(\mu \leq F(z)) = F(z)$. So this CDF is exactly the CDF of $p(x)$, so we are indeed sampling from $p(x)$. ■

Example: $p(x) = \lambda \exp(-\lambda x)$

The CDF of this distribution is $F(z) = \int_{-\infty}^{z} p(x) = -\exp(-\lambda z)$. Its inverse is $-\frac{\log(-z)}{\lambda}$. Both were computable so we can apply the inverse probability transform method.

Unfortunately, there are very limited cases in the real world where we can derive a closed form for both the CDF and its inverse of some distribution, so we will seek better methods.

7.2.5 Acceptance-Rejection Sampling

Suppose we have a distribution $p(x)$ that is very difficult to sample. We can use Acceptance-Rejection Sampling to approximate $p(x)$. The algorithm is as follows:

1. We select a distribution $q(x)$ that is easy to sample from and ideally approximates the shape of $p(x)$ that satisfies $p(x) \leq Cq(x)$, where C is some constant.
2. Generate random $y \sim q(y)$
3. Generate random $\mu \sim U[0,1]$. If $\mu \leq \frac{p(y)}{Cq(y)}$, we accept. Otherwise, we repeat from step 2.

To garner some intuition behind this, consider throwing darts at a dartboard with its bottom edge being the x-axis and its top edge described by $Cq(y)$. We would like to sample darts uniformly randomly from this dartboard. Specifically, we sample its horizontal and vertical positions independently. Its horizontal position simply comes from sampling $y \sim q(y)$ as we want more bias towards points on the x-axis with more area above it. To determine its vertical position, we sample $\mu \sim U[0,1]$ to determine the proportion of the distance from the bottom to top edge of the dartboard it lands at, and we keep this dart if it lands in the proportion described by $p(y)$ (i.e. $\frac{p(y)}{Cq(y)}$). Since we are uniformly sampling from the area of $Cq(y)$ and discarding all darts outside $p(y)$, we are also uniformly sampling from the area of $p(y)$ and therefore sampling $p(y)$ itself.

Proof:

We will now prove that all samples we "accept" will always be a part of our actual distribution. That is,

$$P(Y = i \,|\, Y \text{ is accepted}) = P(X = i)$$

To prove this, we will use Bayes Rule.

$$P(Y = i \,|\, Y \text{ is accepted}) = \frac{P(Y = i\,,\, Y \text{ is accepted})}{P(Y \text{ is accepted})}$$

For the numerator:

$$P(Y = i\,,\, Y \text{ is accepted}) = P(Y = i)P(Y \text{ is accepted}) = q_i \frac{P_i}{Cq_i} = \frac{P_i}{C}$$

where q_i is our sample $P(Y = i)$ and $\frac{P_i}{Cq_i}$ is our sample acceptance criteria.

For the denominator:

$$P(Y \text{ is accepted}) = \sum_{i=1}^{k} P(Y = i\,,\, Y \text{ is accepted}) = \sum_{i=1}^{k} \frac{P_i}{C} = \frac{1}{C}$$

Plugging back to the original equation:

$$P(Y = i \,|\, Y \text{ is accepted}) = \frac{P(Y = i\,,\, Y \text{ is accepted})}{P(Y \text{ is accepted})} = \frac{\frac{P_i}{C}}{\frac{1}{C}} = P_i = P(X = i)$$

■

Note. One issue is that because we need $Cq(x)$ to be above all points in $p(x)$, it can force C to be very large, which can waste a lot of time sampling, since a large C makes it difficult to accept our sample per our acceptance condition $\mu \leq \frac{p(y)}{Cq(y)}$.

Example. For $D = \{X_i, Y_i\}_{i=1}^{n}$, we want to calculate:

$$p(\theta|D) = \prod_{i=1}^{n} \frac{p(Y_i|X_i, \theta)\pi(\theta)}{Z(D)}$$

where

$$Z(D) = \int \prod_{i=1}^{n} p(Y_i|X_i, \theta)\pi(\theta)d\theta$$

We propose that

$$q(\theta) = \pi(\theta)$$

We want to calculate C in our Acceptance-Rejection Sampling approximation criteria $p(x) \leq Cq(x)$:

$$C \geq \frac{p(\theta|D)}{q(\theta)} = \frac{\prod_{i=1}^{n} p(Y_i|X_i, \theta)\pi(\theta)}{Z(D)\pi(\theta)} = \frac{\prod_{i=1}^{n} p(Y_i|X_i, \theta)}{Z(D)}$$

However, because Z(D) contains an intractable integral, we will modify our proposal to work around this:

$$q(\theta) = \frac{\pi(\theta)}{Z(D)}$$

Now we have:

$$C \geq \frac{p(\theta|D)}{q(\theta)} = \frac{\prod_{i=1}^{n} p(Y_i|X_i, \theta)\pi(\theta)}{Z(D)\frac{\pi(\theta)}{Z(D)}} = \prod_{i=1}^{n} p(Y_i|X_i, \theta)$$

Thus,

$$C = \max_{\theta} \prod_{i=1}^{n} p(Y_i | X_i, \theta)$$

Note. In order to find the most optimal C, we need to solve an optimization problem. This makes finding the most optimal C using this approach redundant because if we are already able to solve the optimization, then sampling is not necessary.

7.2.6 Importance Sampling

Suppose we have some other distribution $q(x)$ from which we know how to sample. We can use this distribution to sample $p(x)$ as follows:

$$\mathbb{E}_p[f(x)] = \int f(x)p(x)\,dx = \int f(x)p(x)\frac{q(x)}{q(x)}\,dx = \mathbb{E}_q\left[f(x)\frac{p(x)}{q(x)}\right]$$

Of course, we can't just pick an arbitrary $q(x)$ and expect it to work well. For it to be a Monte Carlo approximation, we need to adhere to the variance property:

$$\text{Var}\left[\frac{1}{k}\sum_{i=1}^{k} f(x_i)\frac{p(x_i)}{q(x_i)}\right] = \frac{1}{k}\text{Var}\left[f(x)\frac{p(x)}{q(x)}\right]$$

Let's use this to derive the most optimal $q(x)$. Apply the definition of variance $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$:

$$\begin{aligned}
\frac{1}{k}\text{Var}\left[f(x)\frac{p(x)}{q(x)}\right] &= \frac{1}{k}\left(\mathbb{E}_q\left[f(x)^2\frac{p(x)^2}{q(x)^2}\right] - \mathbb{E}_q\left[f(x)\frac{p(x)}{q(x)}\right]^2\right) \\
&= \frac{1}{k}\left(\int f(x)^2\frac{p(x)^2}{q(x)}\,dx - \left(\int f(x)p(x)\,dx\right)^2\right) \\
&= \frac{1}{k}\left(\int f(x)^2\frac{p(x)^2}{q(x)}\,dx - \mathbb{E}_p[f(x)]^2\right)
\end{aligned}$$

Next, since $q(x)$ is a probability distribution, $\int q(x)\,dx = 1$. So we can multiply $\int f(x)^2\frac{p(x)^2}{q(x)}\,dx$ by $\int q(x)\,dx$ in the equality. Then consider applying the Cauchy-Schwarz inequality $|\langle u, v\rangle|^2 \leq \langle u, u\rangle \cdot \langle v, v\rangle$ in reverse with $u = f(x)\frac{p(x)}{\sqrt{q(x)}}$ and $v = \sqrt{q(x)}$ to conclude

$$\begin{aligned}
\frac{1}{k}\left(\int f(x)^2\frac{p(x)^2}{q(x)}\,dx - \mathbb{E}_p[f(x)]^2\right) &= \frac{1}{k}\left(\int f(x)^2\frac{p(x)^2}{q(x)}\,dx \cdot \int q(x)\,dx - \mathbb{E}_p[f(x)]^2\right) \\
&\geq \frac{1}{k}\left(\left(\int f(x)p(x)\,dx\right)^2 - \mathbb{E}_p[f(x)]^2\right) \\
&= \frac{1}{k}\left(\mathbb{E}_p[f(x)]^2 - \mathbb{E}_p[f(x)]^2\right) \\
&= 0
\end{aligned}$$

To maximize accuracy, we want to minimize variance, so we want to choose $q(x)$ such that this inequality becomes an equality. This will guarantee the variance drops to the optimal 0. So we have

$$\int f(x)^2 \frac{p(x)^2}{q(x)}\,dx = \left(\int f(x)p(x)\,dx\right)^2 \implies q(x) = \frac{f(x)p(x)}{\int f(x)p(x)\,dx}$$

Unfortunately, computing the optimal $q(x)$ requires knowledge of $\int f(x)p(x)\,dx = \mathbb{E}_p[f(x)]$ which was what we were trying to solve for in the first place. So we rarely do this in practice.