Approximate Inference: Sampling Methods (1)

Piyush Rai

Probabilistic Machine Learning (CS772A)

September 14, 2017

Approximate Inference

- Some typical inference tasks

Approximate Inference

- Some typical inference tasks
  - Compute a (possibly intractable) posterior distribution over unknowns θ

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}$$

Approximate Inference

- Some typical inference tasks
  - Compute a (possibly intractable) posterior distribution over unknowns θ

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}$$

  - Compute a difficult expectation of a random quantity w.r.t. a distribution (an integral), e.g.,

Approximate Inference

- Some typical inference tasks
  - Compute a (possibly intractable) posterior distribution over unknowns θ
$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}$$
  - Compute a difficult expectation of a random quantity w.r.t. a distribution (an integral), e.g.,
    - The posterior predictive (an expectation w.r.t the posterior over θ)
$$p(\mathcal{D}^{new}|\mathcal{D}) = \int p(\mathcal{D}^{new}|\theta)p(\theta|\mathcal{D})d\theta$$

Approximate Inference

- Some typical inference tasks
  - Compute a (possibly intractable) posterior distribution over unknowns θ

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}$$

  - Compute a difficult expectation of a random quantity w.r.t. a distribution (an integral), e.g.,
    - The posterior predictive (an expectation w.r.t the posterior over θ)

$$p(\mathcal{D}^{new}|\mathcal{D}) = \int p(\mathcal{D}^{new}|\theta)p(\theta|\mathcal{D})d\theta = \mathbb{E}_{p(\theta|\mathcal{D})}[p(\mathcal{D}^{new}|\theta)]$$

Approximate Inference

- Some typical inference tasks
  - Compute a (possibly intractable) posterior distribution over unknowns θ

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}$$

  - Compute a difficult expectation of a random quantity w.r.t. a distribution (an integral), e.g.,
    - The posterior predictive (an expectation w.r.t the posterior over θ)

$$p(\mathcal{D}^{new}|\mathcal{D}) = \int p(\mathcal{D}^{new}|\theta)p(\theta|\mathcal{D})d\theta = \mathbb{E}_{p(\theta|\mathcal{D})}[p(\mathcal{D}^{new}|\theta)]$$

    - The marginal likelihood or "evidence" (an expectation over the prior)

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\theta)p(\theta|m)d\theta$$

Approximate Inference

- Some typical inference tasks
  - Compute a (possibly intractable) posterior distribution over unknowns θ

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}$$

  - Compute a difficult expectation of a random quantity w.r.t. a distribution (an integral), e.g.,
    - The posterior predictive (an expectation w.r.t the posterior over θ)

$$p(\mathcal{D}^{new}|\mathcal{D}) = \int p(\mathcal{D}^{new}|\theta)p(\theta|\mathcal{D})d\theta = \mathbb{E}_{p(\theta|\mathcal{D})}[p(\mathcal{D}^{new}|\theta)]$$

    - The marginal likelihood or "evidence" (an expectation over the prior)

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\theta)p(\theta|m)d\theta = \mathbb{E}_{p(\theta|m)}[p(\mathcal{D}|\theta)]$$

Approximate Inference

- Some typical inference tasks
  - Compute a (possibly intractable) posterior distribution over unknowns θ

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}$$

  - Compute a difficult expectation of a random quantity w.r.t. a distribution (an integral), e.g.,
    - The posterior predictive (an expectation w.r.t the posterior over θ)

$$p(\mathcal{D}^{new}|\mathcal{D}) = \int p(\mathcal{D}^{new}|\theta)p(\theta|\mathcal{D})d\theta = \mathbb{E}_{p(\theta|\mathcal{D})}[p(\mathcal{D}^{new}|\theta)]$$

    - The marginal likelihood or "evidence" (an expectation over the prior)

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\theta)p(\theta|m)d\theta = \mathbb{E}_{p(\theta|m)}[p(\mathcal{D}|\theta)]$$

    - The expected complete data log-likelihood needed for doing MLE/MAP in LVMs (recall EM)

$$\text{Exp-CLL} = \int p(\boldsymbol{z}|\theta, \boldsymbol{x})p(\boldsymbol{x}, \boldsymbol{z}|\theta)d\boldsymbol{z}$$

Approximate Inference

- Some typical inference tasks
  - Compute a (possibly intractable) posterior distribution over unknowns θ
  $$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}$$
  - Compute a difficult expectation of a random quantity w.r.t. a distribution (an integral), e.g.,
    - The posterior predictive (an expectation w.r.t the posterior over θ)
    $$p(\mathcal{D}^{new}|\mathcal{D}) = \int p(\mathcal{D}^{new}|\theta)p(\theta|\mathcal{D})d\theta = \mathbb{E}_{p(\theta|\mathcal{D})}[p(\mathcal{D}^{new}|\theta)]$$
    - The marginal likelihood or "evidence" (an expectation over the prior)
    $$p(\mathcal{D}|m) = \int p(\mathcal{D}|\theta)p(\theta|m)d\theta = \mathbb{E}_{p(\theta|m)}[p(\mathcal{D}|\theta)]$$
    - The expected complete data log-likelihood needed for doing MLE/MAP in LVMs (recall EM)
    $$\text{Exp-CLL} = \int p(\boldsymbol{z}|\theta,\boldsymbol{x})p(\boldsymbol{x},\boldsymbol{z}|\theta)d\boldsymbol{z} = \mathbb{E}_{p(\boldsymbol{z}|\theta,\boldsymbol{x})}[p(\boldsymbol{x},\boldsymbol{z}|\theta)]$$

Approximate Inference

- Some typical inference tasks
  - Compute a (possibly intractable) posterior distribution over unknowns θ

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}$$

  - Compute a difficult expectation of a random quantity w.r.t. a distribution (an integral), e.g.,
    - The posterior predictive (an expectation w.r.t the posterior over θ)

$$p(\mathcal{D}^{new}|\mathcal{D}) = \int p(\mathcal{D}^{new}|\theta)p(\theta|\mathcal{D})d\theta = \mathbb{E}_{p(\theta|\mathcal{D})}[p(\mathcal{D}^{new}|\theta)]$$

    - The marginal likelihood or "evidence" (an expectation over the prior)

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\theta)p(\theta|m)d\theta = \mathbb{E}_{p(\theta|m)}[p(\mathcal{D}|\theta)]$$

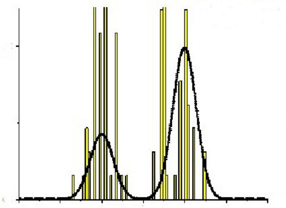
    - The expected complete data log-likelihood needed for doing MLE/MAP in LVMs (recall EM)

$$\text{Exp-CLL} = \int p(\boldsymbol{z}|\theta, \boldsymbol{x})p(\boldsymbol{x}, \boldsymbol{z}|\theta)d\boldsymbol{z} = \mathbb{E}_{p(\boldsymbol{z}|\theta,\boldsymbol{x})}[p(\boldsymbol{x}, \boldsymbol{z}|\theta)]$$

- Sampling methods provide a general way to (approximately) solve these problems

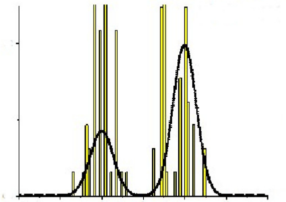
The Basic Idea

- Can approximate a distribution using a set of randomly drawn samples from it



The Basic Idea

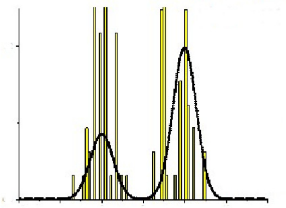
- Can approximate a distribution using a set of randomly drawn samples from it



- But how does it help to approximate a "difficult" distribution (e.g., an intractable posterior)?

The Basic Idea

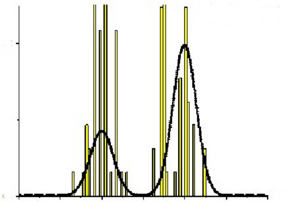
- Can approximate a distribution using a set of randomly drawn samples from it



- But how does it help to approximate a "difficult" distribution (e.g., an intractable posterior)?
  - To approximate, we need samples. Can we get these samples if the distribution itself is difficult?

The Basic Idea

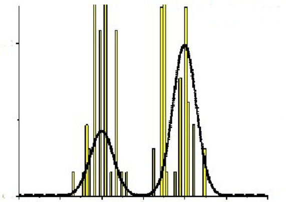
- Can approximate a distribution using a set of randomly drawn samples from it



- But how does it help to approximate a “difficult” distribution (e.g., an intractable posterior)?
  - To approximate, we need samples. Can we get these samples if the distribution itself is difficult?
  - Yes, sampling techniques that we will look at enable us do that!

The Basic Idea

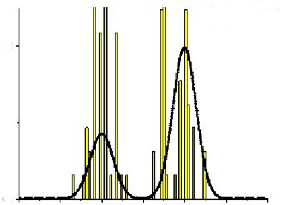
- Can approximate a distribution using a set of randomly drawn samples from it



- But how does it help to approximate a "difficult" distribution (e.g., an intractable posterior)?
  - To approximate, we need samples. Can we get these samples if the distribution itself is difficult?
  - Yes, sampling techniques that we will look at enable us do that!
- Once generated, samples can be used to approximate anything that depends on the distrib.

The Basic Idea

- Can approximate a distribution using a set of randomly drawn samples from it



- But how does it help to approximate a "difficult" distribution (e.g., an intractable posterior)?
  - To approximate, we need samples. Can we get these samples if the distribution itself is difficult?
  - Yes, sampling techniques that we will look at enable us do that!
- Once generated, samples can be used to approximate anything that depends on the distrib., e.g.,

$$p(\mathcal{D}^{new}|\mathcal{D}) = \int p(\mathcal{D}^{new}|\theta)p(\theta|\mathcal{D})d\theta = \mathbb{E}_{p(\theta|\mathcal{D})}[p(\mathcal{D}^{new}|\theta)]$$

Empirical Distribution

- Given L "points" $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$

Empirical Distribution

- Given L "points" $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$, the empirical distribution of these points is defined as

Empirical Distribution

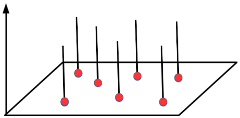
- Given L "points" $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$, the empirical distribution of these points is defined as

$$p_L(A) = \sum_{\ell=1}^{L} w_\ell \delta_{\boldsymbol{z}^{(\ell)}}(A)$$

Empirical Distribution

- Given L "points" $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$, the empirical distribution of these points is defined as

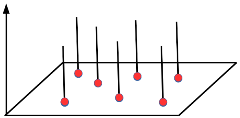
$$p_L(A) = \sum_{\ell=1}^{L} w_\ell \delta_{\boldsymbol{z}^{(\ell)}}(A)$$



Empirical Distribution

- Given L "points" $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$, the empirical distribution of these points is defined as

$$p_L(A) = \sum_{\ell=1}^{L} w_\ell \delta_{\boldsymbol{z}^{(\ell)}}(A)$$

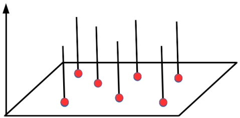


- Here $w_1, \ldots, w_L$ are weights that sum to 1, i.e., $\sum_{\ell=1}^{L} w_\ell = 1$ (for uniform weights, $w_\ell = 1/L$)

Empirical Distribution

- Given L "points" $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$, the empirical distribution of these points is defined as

$$p_L(A) = \sum_{\ell=1}^{L} w_\ell \delta_{\boldsymbol{z}^{(\ell)}}(A)$$



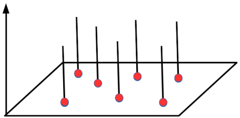
- Here $w_1, \ldots, w_L$ are weights that sum to 1, i.e., $\sum_{\ell=1}^{L} w_\ell = 1$ (for uniform weights, $w_\ell = 1/L$)
- Here $\delta_{\boldsymbol{z}}(A)$ denotes the Dirac distribution defined as

$$\delta_{\boldsymbol{z}}(A) = \begin{cases} 0 & \text{if} \quad \boldsymbol{z} \notin A \\ 1 & \text{if} \quad \boldsymbol{z} \in A \end{cases}$$

Empirical Distribution

- Given L "points" $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$, the empirical distribution of these points is defined as

$$p_L(A) = \sum_{\ell=1}^{L} w_\ell \delta_{\boldsymbol{z}^{(\ell)}}(A)$$



- Here $w_1, \ldots, w_L$ are weights that sum to 1, i.e., $\sum_{\ell=1}^{L} w_\ell = 1$ (for uniform weights, $w_\ell = 1/L$)
- Here $\delta_{\boldsymbol{z}}(A)$ denotes the Dirac distribution defined as

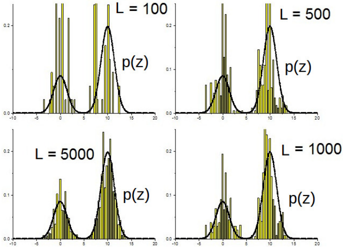
$$\delta_{\boldsymbol{z}}(A) = \begin{cases} 0 & \text{if} \quad \boldsymbol{z} \notin A \\ 1 & \text{if} \quad \boldsymbol{z} \in A \end{cases}$$

- $p_L(A)$ is a discrete distribution with finite support $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$ (can think of it as a histogram)

Empirical Approximation of a Probability Distribution

- Can approximate a probability distribution $p(\boldsymbol{z})$ by L random samples $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$ from $p(\boldsymbol{z})$

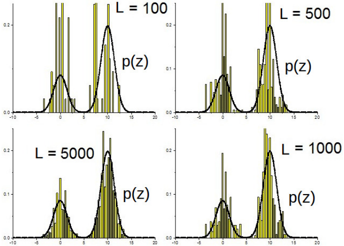
$$p(\boldsymbol{z}) \approx p_L(\boldsymbol{z}) = \sum_{\ell=1}^{L} w_n \delta_{\boldsymbol{z}^{(\ell)}}(\boldsymbol{z})$$



Empirical Approximation of a Probability Distribution

- Can approximate a probability distribution $p(\boldsymbol{z})$ by L random samples $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$ from $p(\boldsymbol{z})$

$$p(\boldsymbol{z}) \approx p_L(\boldsymbol{z}) = \sum_{\ell=1}^{L} w_n \delta_{\boldsymbol{z}^{(\ell)}}(\boldsymbol{z})$$

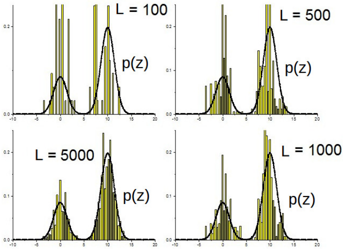


- Note: Height of each bar represents the weight of that sample

Empirical Approximation of a Probability Distribution

- Can approximate a probability distribution $p(\boldsymbol{z})$ by L random samples $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$ from $p(\boldsymbol{z})$

$$p(\boldsymbol{z}) \approx p_L(\boldsymbol{z}) = \sum_{\ell=1}^{L} w_n \delta_{\boldsymbol{z}^{(\ell)}}(\boldsymbol{z})$$



- Note: Height of each bar represents the weight of that sample
- How can we generate random samples from a distribution (both simple and "difficult" ones)?

Sampling from Distributions: Some Basic Methods

Transformation Methods: Inv-CDF Method

- Suppose we have a random sample x from an easy distribution $p(x)$ (e.g., uniform)

Transformation Methods: Inv-CDF Method

- Suppose we have a random sample x from an easy distribution $p(x)$ (e.g., uniform)
- Want to generate a random sample z from a "less easy" but still standard distribution $p(z)$

Transformation Methods: Inv-CDF Method

- Suppose we have a random sample x from an easy distribution $p(x)$ (e.g., uniform)
- Want to generate a random sample z from a "less easy" but still standard distribution $p(z)$
- Change of variables rule allows us to write

$$p(z) = p(x)\left|\frac{\partial x}{\partial z}\right|$$

Transformation Methods: Inv-CDF Method

- Suppose we have a random sample x from an easy distribution $p(x)$ (e.g., uniform)
- Want to generate a random sample z from a "less easy" but still standard distribution $p(z)$
- Change of variables rule allows us to write

$$p(z) = p(x)\left|\frac{\partial x}{\partial z}\right|$$

- Suppose $p(x)$ is uniform distribution.

Transformation Methods: Inv-CDF Method

- Suppose we have a random sample x from an easy distribution $p(x)$ (e.g., uniform)
- Want to generate a random sample z from a "less easy" but still standard distribution $p(z)$
- Change of variables rule allows us to write

$$p(z) = p(x)\left|\frac{\partial x}{\partial z}\right|$$

- Suppose $p(x)$ is uniform distribution. Integrating both sides

Transformation Methods: Inv-CDF Method

- Suppose we have a random sample x from an easy distribution $p(x)$ (e.g., uniform)
- Want to generate a random sample z from a "less easy" but still standard distribution $p(z)$
- Change of variables rule allows us to write

$$p(z) = p(x)\left|\frac{\partial x}{\partial z}\right|$$

- Suppose $p(x)$ is uniform distribution. Integrating both sides

$$x = \int_{-\infty}^{\hat{z}} p(z)dz = h(\hat{z})$$

Transformation Methods: Inv-CDF Method

- Suppose we have a random sample x from an easy distribution $p(x)$ (e.g., uniform)
- Want to generate a random sample z from a "less easy" but still standard distribution $p(z)$
- Change of variables rule allows us to write

$$p(z) = p(x)\left|\frac{\partial x}{\partial z}\right|$$

- Suppose $p(x)$ is uniform distribution. Integrating both sides

$$x = \int_{-\infty}^{\hat{z}} p(z)dz = h(\hat{z})$$

where $h(z)$ is the cumulative distribution function (CDF) of the distribution $p(z)$

Transformation Methods: Inv-CDF Method

- Suppose we have a random sample x from an easy distribution $p(x)$ (e.g., uniform)
- Want to generate a random sample z from a "less easy" but still standard distribution $p(z)$
- Change of variables rule allows us to write

$$p(z) = p(x)\left|\frac{\partial x}{\partial z}\right|$$

- Suppose $p(x)$ is uniform distribution. Integrating both sides

$$x = \int_{-\infty}^{\hat{z}} p(z)dz = h(\hat{z})$$

  where $h(z)$ is the cumulative distribution function (CDF) of the distribution $p(z)$

- Can thus generate a random sample z from $p(z)$ by first drawing $u \sim \text{Unif}(0,1)$

Transformation Methods: Inv-CDF Method

- Suppose we have a random sample x from an easy distribution $p(x)$ (e.g., uniform)
- Want to generate a random sample z from a "less easy" but still standard distribution $p(z)$
- Change of variables rule allows us to write

$$p(z) = p(x)\left|\frac{\partial x}{\partial z}\right|$$

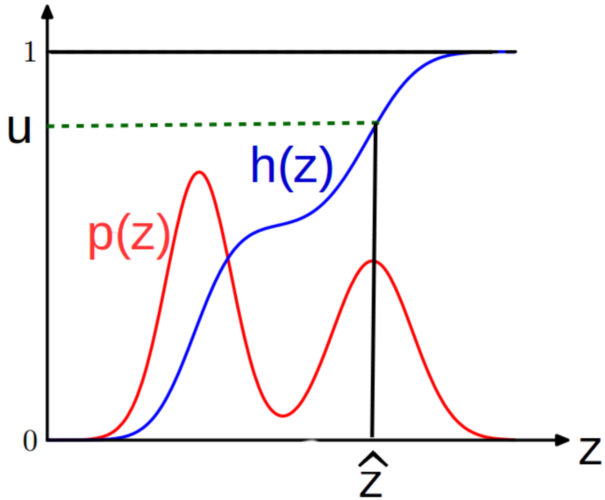
- Suppose $p(x)$ is uniform distribution. Integrating both sides

$$x = \int_{-\infty}^{\hat{z}} p(z)dz = h(\hat{z})$$

  where $h(z)$ is the cumulative distribution function (CDF) of the distribution $p(z)$
- Can thus generate a random sample z from $p(z)$ by first drawing $u \sim \text{Unif}(0,1)$ and then doing

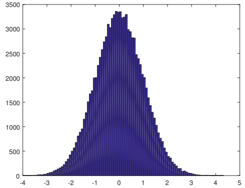
$$\boxed{\hat{z} = h^{-1}(u)} \qquad \text{(Inverse CDF method)}$$

Transformation Methods: Inv-CDF Method



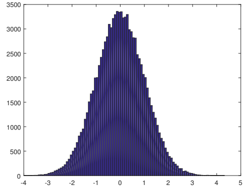
Transformation Methods: Inv-CDF Method

- Try in MATLAB: u = rand(100000,1);
  - z = icdf('normal',u,0,1); hist(z,100);

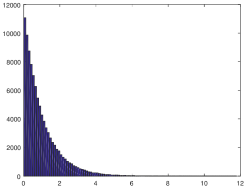


Transformation Methods: Inv-CDF Method

- Try in MATLAB: u = rand(100000,1);
  - z = icdf('normal',u,0,1); hist(z,100);



  - z = icdf('exp',u,1); hist(z,100);



Transformation Methods: Box-Muller Method

- A method to generate samples from two-dim spherical Gaussian

Transformation Methods: Box-Muller Method

- A method to generate samples from two-dim spherical Gaussian
- Assume x_1, x_2 drawn from Unif$(-1, +1)$ and only consider those pairs for which $x_1^2 + x_2^2 \leq 1$

Transformation Methods: Box-Muller Method

- A method to generate samples from two-dim spherical Gaussian
- Assume x_1, x_2 drawn from $\text{Unif}(-1, +1)$ and only consider those pairs for which $x_1^2 + x_2^2 \leq 1$
  - Note that $p(x_1, x_2) = \frac{1}{\pi}\mathbb{I}[x_1^2 + x_2^2 \leq 1]$ (uniform distribution of the points in the unit radius circle)

Transformation Methods: Box-Muller Method

- A method to generate samples from two-dim spherical Gaussian
- Assume x_1, x_2 drawn from $\text{Unif}(-1, +1)$ and only consider those pairs for which $x_1^2 + x_2^2 \leq 1$
  - Note that $p(x_1, x_2) = \frac{1}{\pi}\mathbb{I}[x_1^2 + x_2^2 \leq 1]$ (uniform distribution of the points in the unit radius circle)
- Let's define $r^2 = x_1^2 + x_2^2$

Transformation Methods: Box-Muller Method

- A method to generate samples from two-dim spherical Gaussian
- Assume x_1, x_2 drawn from $\text{Unif}(-1, +1)$ and only consider those pairs for which $x_1^2 + x_2^2 \leq 1$
  - Note that $p(x_1, x_2) = \frac{1}{\pi}\mathbb{I}[x_1^2 + x_2^2 \leq 1]$ (uniform distribution of the points in the unit radius circle)
- Let's define $r^2 = x_1^2 + x_2^2$, and define

$$z_1 = x_1 \left(\frac{-2 \ln r^2}{r^2} \right)^{1/2} \quad \text{and} \quad z_2 = x_2 \left(\frac{-2 \ln r^2}{r^2} \right)^{1/2}$$

Transformation Methods: Box-Muller Method

- A method to generate samples from two-dim spherical Gaussian
- Assume x_1, x_2 drawn from Unif$(-1, +1)$ and only consider those pairs for which $x_1^2 + x_2^2 \leq 1$
  - Note that $p(x_1, x_2) = \frac{1}{\pi}\mathbb{I}[x_1^2 + x_2^2 \leq 1]$ (uniform distribution of the points in the unit radius circle)
- Let's define $r^2 = x_1^2 + x_2^2$, and define

$$z_1 = x_1 \left(\frac{-2\ln r^2}{r^2}\right)^{1/2} \quad \text{and} \quad z_2 = x_2 \left(\frac{-2\ln r^2}{r^2}\right)^{1/2}$$

- The joint distribution of (z_1, z_2) is given by

$$p(z_1, z_2) = p(x_1, x_2)\left|\frac{\partial(x_1, x_2)}{\partial(z_1, z_2)}\right|$$

Transformation Methods: Box-Muller Method

- A method to generate samples from two-dim spherical Gaussian
- Assume x_1, x_2 drawn from Unif$(-1, +1)$ and only consider those pairs for which $x_1^2 + x_2^2 \leq 1$
  - Note that $p(x_1, x_2) = \frac{1}{\pi}\mathbb{I}[x_1^2 + x_2^2 \leq 1]$ (uniform distribution of the points in the unit radius circle)
- Let's define $r^2 = x_1^2 + x_2^2$, and define

$$z_1 = x_1 \left(\frac{-2\ln r^2}{r^2}\right)^{1/2} \quad \text{and} \quad z_2 = x_2 \left(\frac{-2\ln r^2}{r^2}\right)^{1/2}$$

- The joint distribution of (z_1, z_2) is given by

$$\begin{aligned} p(z_1, z_2) &= p(x_1, x_2)\left|\frac{\partial(x_1, x_2)}{\partial(z_1, z_2)}\right| \\ &= \left[\frac{1}{\sqrt{2\pi}}\exp(-z_1^2/2)\right]\left[\frac{1}{\sqrt{2\pi}}\exp(-z_2^2/2)\right] \end{aligned}$$

Transformation Methods: Box-Muller Method

- A method to generate samples from two-dim spherical Gaussian
- Assume x_1, x_2 drawn from Unif$(-1, +1)$ and only consider those pairs for which $x_1^2 + x_2^2 \leq 1$
  - Note that $p(x_1, x_2) = \frac{1}{\pi}\mathbb{I}[x_1^2 + x_2^2 \leq 1]$ (uniform distribution of the points in the unit radius circle)
- Let's define $r^2 = x_1^2 + x_2^2$, and define

$$z_1 = x_1 \left(\frac{-2\ln r^2}{r^2}\right)^{1/2} \quad \text{and} \quad z_2 = x_2 \left(\frac{-2\ln r^2}{r^2}\right)^{1/2}$$

- The joint distribution of (z_1, z_2) is given by

$$\begin{aligned} p(z_1, z_2) &= p(x_1, x_2)\left|\frac{\partial(x_1, x_2)}{\partial(z_1, z_2)}\right| \\ &= \left[\frac{1}{\sqrt{2\pi}}\exp(-z_1^2/2)\right]\left[\frac{1}{\sqrt{2\pi}}\exp(-z_2^2/2)\right] \end{aligned}$$

which is a two-dim Gaussian with with zero mean and identity covariance

Transformation Methods: Other Examples and Limitations

- Suppose $x \sim \mathcal{N}(0,1)$ then

$$z = \mu + \sigma x$$

Transformation Methods: Other Examples and Limitations

- Suppose $x \sim \mathcal{N}(0,1)$ then
$$z = \mu + \sigma x \sim \mathcal{N}(\mu, \sigma^2)$$
because $\mathbb{E}[z] = \mu$ and $\text{cov}[z] = \sigma^2$ (linear transformation rule of Gaussians)

Transformation Methods: Other Examples and Limitations

- Suppose $x \sim \mathcal{N}(0,1)$ then
$$z = \mu + \sigma x \sim \mathcal{N}(\mu, \sigma^2)$$
because $\mathbb{E}[z] = \mu$ and $\text{cov}[z] = \sigma^2$ (linear transformation rule of Gaussians)
- Suppose $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$ be the Cholesky decomposition of $\boldsymbol{\Sigma}$
$$\boldsymbol{z} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{x}$$

Transformation Methods: Other Examples and Limitations

- Suppose $x \sim \mathcal{N}(0,1)$ then
$$z = \mu + \sigma x \sim \mathcal{N}(\mu, \sigma^2)$$
because $\mathbb{E}[z] = \mu$ and $\text{cov}[z] = \sigma^2$ (linear transformation rule of Gaussians)
- Suppose $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$ be the Cholesky decomposition of $\boldsymbol{\Sigma}$
$$\boldsymbol{z} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Transformation Methods: Other Examples and Limitations

- Suppose $x \sim \mathcal{N}(0,1)$ then
$$z = \mu + \sigma x \sim \mathcal{N}(\mu, \sigma^2)$$
because $\mathbb{E}[z] = \mu$ and $\text{cov}[z] = \sigma^2$ (linear transformation rule of Gaussians)
- Suppose $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$ be the Cholesky decomposition of $\boldsymbol{\Sigma}$
$$\boldsymbol{z} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
- Transformation Methods are simple but have limitations

Transformation Methods: Other Examples and Limitations

- Suppose $x \sim \mathcal{N}(0,1)$ then
$$z = \mu + \sigma x \sim \mathcal{N}(\mu, \sigma^2)$$
because $\mathbb{E}[z] = \mu$ and $\text{cov}[z] = \sigma^2$ (linear transformation rule of Gaussians)
- Suppose $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$ be the Cholesky decomposition of $\boldsymbol{\Sigma}$
$$\boldsymbol{z} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
- Transformation Methods are simple but have limitations
  - Mostly limited to standard distributions and/or distributions with vert few variables

Transformation Methods: Other Examples and Limitations

- Suppose $x \sim \mathcal{N}(0,1)$ then
$$z = \mu + \sigma x \sim \mathcal{N}(\mu, \sigma^2)$$
because $\mathbb{E}[z] = \mu$ and $\text{cov}[z] = \sigma^2$ (linear transformation rule of Gaussians)
- Suppose $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$ be the Cholesky decomposition of $\boldsymbol{\Sigma}$
$$\boldsymbol{z} = \boldsymbol{\mu} + \mathbf{L}\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
- Transformation Methods are simple but have limitations
    - Mostly limited to standard distributions and/or distributions with vert few variables
- Useful nevertheless..

Rejection Sampling

- Want to sample from $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$.

Rejection Sampling

- Want to sample from $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$. Suppose we can only evaluate the numerator $\tilde{p}(\boldsymbol{z})$ at any $\boldsymbol{z}$

Rejection Sampling

- Want to sample from $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$. Suppose we can only evaluate the numerator $\tilde{p}(\boldsymbol{z})$ at any $\boldsymbol{z}$
- Suppose we have a proposal distribution $q(\boldsymbol{z})$ that we can generate samples from, and

$$Mq(\boldsymbol{z}) \geq \tilde{p}(\boldsymbol{z}) \qquad \forall \boldsymbol{z} \quad \text{(where } M > 0 \text{ is some const.)}$$

Rejection Sampling

- Want to sample from $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$. Suppose we can only evaluate the numerator $\tilde{p}(\boldsymbol{z})$ at any $\boldsymbol{z}$
- Suppose we have a proposal distribution $q(\boldsymbol{z})$ that we can generate samples from, and

$$Mq(\boldsymbol{z}) \geq \tilde{p}(\boldsymbol{z}) \qquad \forall \boldsymbol{z} \quad \text{(where } M > 0 \text{ is some const.)}$$

- Basic idea: Generate samples from the proposal $q(\boldsymbol{z})$

Rejection Sampling

- Want to sample from $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$. Suppose we can only evaluate the numerator $\tilde{p}(\boldsymbol{z})$ at any $\boldsymbol{z}$
- Suppose we have a proposal distribution $q(\boldsymbol{z})$ that we can generate samples from, and

$$Mq(\boldsymbol{z}) \geq \tilde{p}(\boldsymbol{z}) \qquad \forall \boldsymbol{z} \quad (\text{where } M > 0 \text{ is some const.})$$

- Basic idea: Generate samples from the proposal $q(\boldsymbol{z})$ and accept/reject based on some condition

Rejection Sampling

- Want to sample from $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$. Suppose we can only evaluate the numerator $\tilde{p}(\boldsymbol{z})$ at any $\boldsymbol{z}$
- Suppose we have a proposal distribution $q(\boldsymbol{z})$ that we can generate samples from, and

$$Mq(\boldsymbol{z}) \geq \tilde{p}(\boldsymbol{z}) \qquad \forall \boldsymbol{z} \quad \text{(where } M > 0 \text{ is some const.)}$$

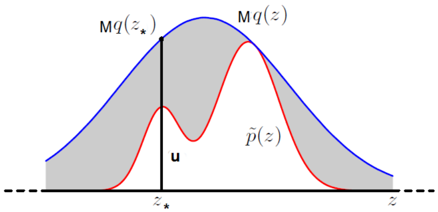
- Basic idea: Generate samples from the proposal $q(\boldsymbol{z})$ and accept/reject based on some condition
  1. Sample an r.v. $\boldsymbol{z}_*$ from $q(\boldsymbol{z})$

Rejection Sampling

- Want to sample from $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$. Suppose we can only evaluate the numerator $\tilde{p}(\boldsymbol{z})$ at any $\boldsymbol{z}$
- Suppose we have a proposal distribution $q(\boldsymbol{z})$ that we can generate samples from, and

$$Mq(\boldsymbol{z}) \geq \tilde{p}(\boldsymbol{z}) \qquad \forall \boldsymbol{z} \quad (\text{where } M > 0 \text{ is some const.})$$

- Basic idea: Generate samples from the proposal $q(\boldsymbol{z})$ and accept/reject based on some condition
  1. Sample an r.v. $\boldsymbol{z}_*$ from $q(\boldsymbol{z})$
  2. Sampling a uniform r.v. $u \sim \text{Unif}[0, Mq(\boldsymbol{z}_*)]$

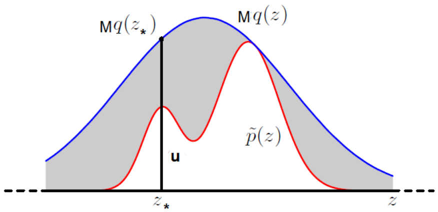


Rejection Sampling

- Want to sample from $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$. Suppose we can only evaluate the numerator $\tilde{p}(\boldsymbol{z})$ at any $\boldsymbol{z}$
- Suppose we have a proposal distribution $q(\boldsymbol{z})$ that we can generate samples from, and

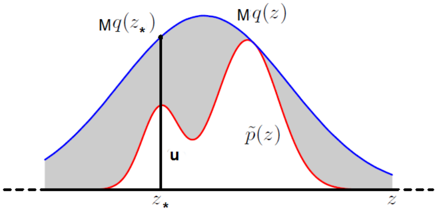
$$Mq(\boldsymbol{z}) \geq \tilde{p}(\boldsymbol{z}) \qquad \forall \boldsymbol{z} \quad \text{(where } M > 0 \text{ is some const.)}$$

- Basic idea: Generate samples from the proposal $q(\boldsymbol{z})$ and accept/reject based on some condition
    1. Sample an r.v. $\boldsymbol{z}_*$ from $q(\boldsymbol{z})$
    2. Sampling a uniform r.v. $u \sim \text{Unif}[0, Mq(\boldsymbol{z}_*)]$



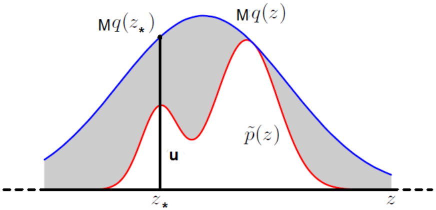
    3. If $u \leq \tilde{p}(\boldsymbol{z}_*)$ then accept $\boldsymbol{z}_*$ else reject

Rejection Sampling



- Why $z \sim q(z)$ + accept/reject rule is equivalent to $z \sim p(z)$?

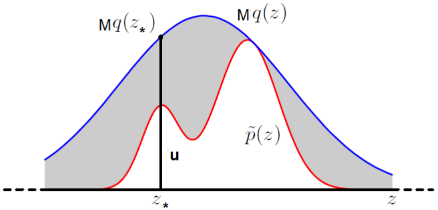
Rejection Sampling



- Why $\boldsymbol{z} \sim q(\boldsymbol{z})$ + accept/reject rule is equivalent to $\boldsymbol{z} \sim p(\boldsymbol{z})$?
- Let's look at the pdf of $\boldsymbol{z}$ *conditioned* on its acceptance, i.e., $p(\boldsymbol{z}|\text{accept})$

$$p(\text{accept}|\boldsymbol{z})$$

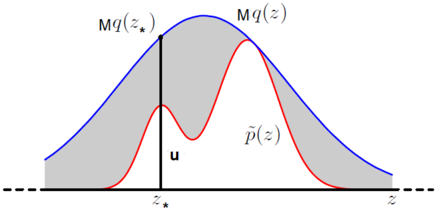
Rejection Sampling



- Why $\boldsymbol{z} \sim q(\boldsymbol{z})$ + accept/reject rule is equivalent to $\boldsymbol{z} \sim p(\boldsymbol{z})$?
- Let's look at the pdf of $\boldsymbol{z}$ *conditioned* on its acceptance, i.e., $p(\boldsymbol{z}|\text{accept})$

$$p(\text{accept}|\boldsymbol{z}) \quad = \quad \int_0^{\tilde{p}(\boldsymbol{z})} \frac{1}{Mq(\boldsymbol{z})} du = \frac{\tilde{p}(\boldsymbol{z})}{Mq(\boldsymbol{z})}$$

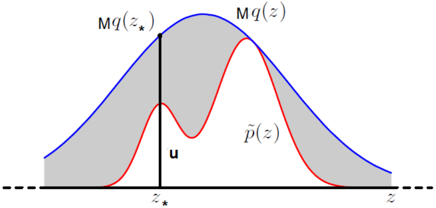
Rejection Sampling



- Why $\boldsymbol{z} \sim q(\boldsymbol{z})$ + accept/reject rule is equivalent to $\boldsymbol{z} \sim p(\boldsymbol{z})$?
- Let's look at the pdf of $\boldsymbol{z}$ *conditioned* on its acceptance, i.e., $p(\boldsymbol{z}|\text{accept})$

$$
\begin{aligned}
p(\text{accept}|\boldsymbol{z}) &= \int_0^{\tilde{p}(\boldsymbol{z})} \frac{1}{Mq(\boldsymbol{z})} du = \frac{\tilde{p}(\boldsymbol{z})}{Mq(\boldsymbol{z})} \\
p(\boldsymbol{z}, \text{accept}) &= q(\boldsymbol{z})p(\text{accept}|\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{M}
\end{aligned}
$$

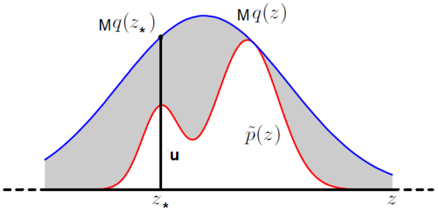
Rejection Sampling



- Why $\boldsymbol{z} \sim q(\boldsymbol{z})$ + accept/reject rule is equivalent to $\boldsymbol{z} \sim p(\boldsymbol{z})$?
- Let's look at the pdf of $\boldsymbol{z}$ *conditioned* on its acceptance, i.e., $p(\boldsymbol{z}|\text{accept})$

$$
\begin{aligned}
p(\text{accept}|\boldsymbol{z}) &= \int_0^{\tilde{p}(\boldsymbol{z})} \frac{1}{Mq(\boldsymbol{z})} du = \frac{\tilde{p}(\boldsymbol{z})}{Mq(\boldsymbol{z})} \\
p(\boldsymbol{z}, \text{accept}) &= q(\boldsymbol{z}) p(\text{accept}|\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{M} \\
p(\text{accept}) &= \int \frac{\tilde{p}(\boldsymbol{z})}{M} d\boldsymbol{z} = \frac{Z_p}{M}
\end{aligned}
$$

Rejection Sampling



- Why $\boldsymbol{z} \sim q(\boldsymbol{z})$ + accept/reject rule is equivalent to $\boldsymbol{z} \sim p(\boldsymbol{z})$?
- Let's look at the pdf of $\boldsymbol{z}$ *conditioned* on its acceptance, i.e., $p(\boldsymbol{z}|\text{accept})$

$$
\begin{aligned}
p(\text{accept}|\boldsymbol{z}) &= \int_0^{\tilde{p}(\boldsymbol{z})} \frac{1}{Mq(\boldsymbol{z})} du = \frac{\tilde{p}(\boldsymbol{z})}{Mq(\boldsymbol{z})} \\
p(\boldsymbol{z}, \text{accept}) &= q(\boldsymbol{z}) p(\text{accept}|\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{M} \\
p(\text{accept}) &= \int \frac{\tilde{p}(\boldsymbol{z})}{M} d\boldsymbol{z} = \frac{Z_p}{M} \\
p(\boldsymbol{z}|\text{accept}) &= \frac{p(\boldsymbol{z}, \text{accept})}{p(\text{accept})} = \frac{\tilde{p}(\boldsymbol{z})}{Z_p} = p(\boldsymbol{z})
\end{aligned}
$$

Sampling Methods for Approximately Computing Expectations

Computing Expectations via Monte Carlo Sampling

- Often we are interested in computing expectations of the form

$$\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$

  where $f(\boldsymbol{z})$ is some function of a random variable $\boldsymbol{z} \in p(\boldsymbol{z})$

Computing Expectations via Monte Carlo Sampling

- Often we are interested in computing expectations of the form

$$\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$

  where $f(\boldsymbol{z})$ is some function of a random variable $\boldsymbol{z} \in p(\boldsymbol{z})$

- A simple approximation scheme: Monte Carlo integration

Computing Expectations via Monte Carlo Sampling

- Often we are interested in computing expectations of the form

$$\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$

  where $f(\boldsymbol{z})$ is some function of a random variable $\boldsymbol{z} \in p(\boldsymbol{z})$

- A simple approximation scheme: Monte Carlo integration
- Suppose we can generate L independent samples from $p(\boldsymbol{z})$: $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim p(\boldsymbol{z})$

Computing Expectations via Monte Carlo Sampling

- Often we are interested in computing expectations of the form

$$\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$

  where $f(\boldsymbol{z})$ is some function of a random variable $\boldsymbol{z} \in p(\boldsymbol{z})$

- A simple approximation scheme: Monte Carlo integration
- Suppose we can generate L independent samples from $p(\boldsymbol{z})$: $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim p(\boldsymbol{z})$
- Monte-Carlo approximation replaces the expectation by an empirical average

$$\hat{f} \approx \frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)})$$

Computing Expectations via Monte Carlo Sampling

- Often we are interested in computing expectations of the form

$$\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$

  where $f(\boldsymbol{z})$ is some function of a random variable $\boldsymbol{z} \in p(\boldsymbol{z})$

- A simple approximation scheme: Monte Carlo integration
- Suppose we can generate L independent samples from $p(\boldsymbol{z})$: $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim p(\boldsymbol{z})$
- Monte-Carlo approximation replaces the expectation by an empirical average

$$\hat{f} \approx \frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)})$$

- Since the samples are independent of each other, can show the following

$$\mathbb{E}[\hat{f}] = \mathbb{E}[f] \qquad \text{and} \qquad \text{var}[\hat{f}] = \frac{1}{L}\text{var}[f] = \frac{1}{L}\mathbb{E}[(f - \mathbb{E}[f])^2]$$

Computing Expectations via Monte Carlo Sampling

- Often we are interested in computing expectations of the form

$$\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$

where $f(\boldsymbol{z})$ is some function of a random variable $\boldsymbol{z} \in p(\boldsymbol{z})$

- A simple approximation scheme: Monte Carlo integration
- Suppose we can generate L independent samples from $p(\boldsymbol{z})$: $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim p(\boldsymbol{z})$
- Monte-Carlo approximation replaces the expectation by an empirical average

$$\hat{f} \approx \frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)})$$

- Since the samples are independent of each other, can show the following

$$\mathbb{E}[\hat{f}] = \mathbb{E}[f] \qquad \text{and} \qquad \text{var}[\hat{f}] = \frac{1}{L}\text{var}[f] = \frac{1}{L}\mathbb{E}[(f - \mathbb{E}[f])^2]$$

- Note that the variance in the estimate of expectation gets better as L increases

Computing Expectations via Importance Sampling

- Using Monte Carlo integration, we approximated the expectation $\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$ as

$$\hat{f} = \frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)}) \qquad \text{where} \quad \{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim p(\boldsymbol{z})$$

- We assumed we know how to generate samples from $p(\boldsymbol{z})$. What if we don't?

Computing Expectations via Importance Sampling

- Using Monte Carlo integration, we approximated the expectation $\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$ as

$$\hat{f} = \frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)}) \qquad \text{where} \quad \{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim p(\boldsymbol{z})$$

- We assumed we know how to generate samples from $p(\boldsymbol{z})$. What if we don't?
- Suppose we can generate samples from a "proposal" distribution $q(\boldsymbol{z})$, i.e., $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$

Computing Expectations via Importance Sampling

- Using Monte Carlo integration, we approximated the expectation $\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$ as

$$\hat{f} = \frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)}) \qquad \text{where} \quad \{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim p(\boldsymbol{z})$$

- We assumed we know how to generate samples from $p(\boldsymbol{z})$. What if we don't?
- Suppose we can generate samples from a "proposal" distribution $q(\boldsymbol{z})$, i.e., $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$
- Suppose we can evaluate $p(\boldsymbol{z})$ at any given $\boldsymbol{z}$

Computing Expectations via Importance Sampling

- Using Monte Carlo integration, we approximated the expectation $\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$ as

$$\hat{f} = \frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)}) \qquad \text{where} \quad \{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim p(\boldsymbol{z})$$

- We assumed we know how to generate samples from $p(\boldsymbol{z})$. What if we don't?
- Suppose we can generate samples from a "proposal" distribution $q(\boldsymbol{z})$, i.e., $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$
- Suppose we can evaluate $p(\boldsymbol{z})$ at any given $\boldsymbol{z}$
- We can now approximate the original expectation as

$$\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$

Computing Expectations via Importance Sampling

- Using Monte Carlo integration, we approximated the expectation $\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$ as

$$\hat{f} = \frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)}) \qquad \text{where} \quad \{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim p(\boldsymbol{z})$$

- We assumed we know how to generate samples from $p(\boldsymbol{z})$. What if we don't?
- Suppose we can generate samples from a "proposal" distribution $q(\boldsymbol{z})$, i.e., $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$
- Suppose we can evaluate $p(\boldsymbol{z})$ at any given $\boldsymbol{z}$
- We can now approximate the original expectation as

$$\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} = \int f(\boldsymbol{z})\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}q(\boldsymbol{z})d\boldsymbol{z}$$

Computing Expectations via Importance Sampling

- Using Monte Carlo integration, we approximated the expectation $\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$ as

$$\hat{f} = \frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)}) \qquad \text{where} \quad \{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim p(\boldsymbol{z})$$

- We assumed we know how to generate samples from $p(\boldsymbol{z})$. What if we don't?
- Suppose we can generate samples from a "proposal" distribution $q(\boldsymbol{z})$, i.e., $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$
- Suppose we can evaluate $p(\boldsymbol{z})$ at any given $\boldsymbol{z}$
- We can now approximate the original expectation as

$$\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} = \int f(\boldsymbol{z})\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}q(\boldsymbol{z})d\boldsymbol{z} \approx \frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)})\frac{p(\boldsymbol{z}^{(\ell)})}{q(\boldsymbol{z}^{(\ell)})}$$

Computing Expectations via Importance Sampling

- Using Monte Carlo integration, we approximated the expectation $\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$ as

$$\hat{f} = \frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)}) \qquad \text{where} \quad \{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim p(\boldsymbol{z})$$

- We assumed we know how to generate samples from $p(\boldsymbol{z})$. What if we don't?
- Suppose we can generate samples from a "proposal" distribution $q(\boldsymbol{z})$, i.e., $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$
- Suppose we can evaluate $p(\boldsymbol{z})$ at any given $\boldsymbol{z}$
- We can now approximate the original expectation as

$$\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} = \int f(\boldsymbol{z})\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}q(\boldsymbol{z})d\boldsymbol{z} \approx \frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)})\frac{p(\boldsymbol{z}^{(\ell)})}{q(\boldsymbol{z}^{(\ell)})}$$

- This is basically "weighted" Monte Carlo integration

Computing Expectations via Importance Sampling

- Using Monte Carlo integration, we approximated the expectation $\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$ as

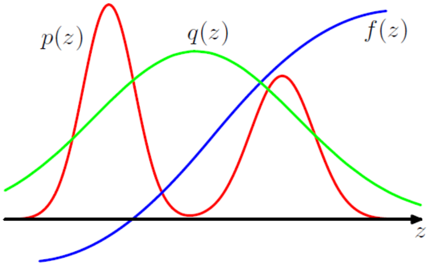
$$\hat{f} = \frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)}) \qquad \text{where} \quad \{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim p(\boldsymbol{z})$$

- We assumed we know how to generate samples from $p(\boldsymbol{z})$. What if we don't?
- Suppose we can generate samples from a "proposal" distribution $q(\boldsymbol{z})$, i.e., $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$
- Suppose we can evaluate $p(\boldsymbol{z})$ at any given $\boldsymbol{z}$
- We can now approximate the original expectation as

$$\mathbb{E}[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} = \int f(\boldsymbol{z})\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}q(\boldsymbol{z})d\boldsymbol{z} \approx \frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)})\frac{p(\boldsymbol{z}^{(\ell)})}{q(\boldsymbol{z}^{(\ell)})}$$

- This is basically "weighted" Monte Carlo integration
  - $w_\ell = \frac{p(\boldsymbol{z}^{(\ell)})}{q(\boldsymbol{z}^{(\ell)})}$ denotes the importance weight of each sample $\boldsymbol{z}^{(\ell)}$

Computing Expectations via Importance Sampling



$$\mathbb{E}[f] \approx \frac{1}{L}\sum_{\ell=1}^{L} f(z^{(\ell)})\frac{p(z^{(\ell)})}{q(z^{(\ell)})}$$

Computing Expectations via Importance Sampling

- What if we can't even evaluate $p(\boldsymbol{z})$ but can only do so up to proportionality constant
    - Suppose $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$ and only $\tilde{p}(\boldsymbol{z})$ can be evaluated

Computing Expectations via Importance Sampling

- What if we can't even evaluate $p(\boldsymbol{z})$ but can only do so up to proportionality constant
    - Suppose $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$ and only $\tilde{p}(\boldsymbol{z})$ can be evaluated
- Let's also write our proposal distribution in the same form: $q(\boldsymbol{z}) = \frac{\tilde{q}(\boldsymbol{z})}{Z_q}$

Computing Expectations via Importance Sampling

- What if we can't even evaluate $p(\boldsymbol{z})$ but can only do so up to proportionality constant
  - Suppose $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$ and only $\tilde{p}(\boldsymbol{z})$ can be evaluated
- Let's also write our proposal distribution in the same form: $q(\boldsymbol{z}) = \frac{\tilde{q}(\boldsymbol{z})}{Z_q}$
- Given $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$, the I.S. based expectation will be

$$\mathbb{E}[f] = \int f(\boldsymbol{z}) \frac{p(\boldsymbol{z})}{q(\boldsymbol{z})} q(\boldsymbol{z}) d\boldsymbol{z}$$

Computing Expectations via Importance Sampling

- What if we can't even evaluate $p(\boldsymbol{z})$ but can only do so up to proportionality constant
  - Suppose $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$ and only $\tilde{p}(\boldsymbol{z})$ can be evaluated
- Let's also write our proposal distribution in the same form: $q(\boldsymbol{z}) = \frac{\tilde{q}(\boldsymbol{z})}{Z_q}$
- Given $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$, the I.S. based expectation will be

$$\mathbb{E}[f] = \int f(\boldsymbol{z}) \frac{p(\boldsymbol{z})}{q(\boldsymbol{z})} q(\boldsymbol{z}) d\boldsymbol{z} = \int f(\boldsymbol{z}) \frac{\tilde{p}(\boldsymbol{z})/Z_p}{\tilde{q}(\boldsymbol{z})/Z_q} q(\boldsymbol{z}) d\boldsymbol{z}$$

Computing Expectations via Importance Sampling

- What if we can't even evaluate $p(\boldsymbol{z})$ but can only do so up to proportionality constant
  - Suppose $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$ and only $\tilde{p}(\boldsymbol{z})$ can be evaluated
- Let's also write our proposal distribution in the same form: $q(\boldsymbol{z}) = \frac{\tilde{q}(\boldsymbol{z})}{Z_q}$
- Given $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$, the I.S. based expectation will be

$$\mathbb{E}[f] = \int f(\boldsymbol{z})\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}q(\boldsymbol{z})d\boldsymbol{z} = \int f(\boldsymbol{z})\frac{\tilde{p}(\boldsymbol{z})/Z_p}{\tilde{q}(\boldsymbol{z})/Z_q}q(\boldsymbol{z})d\boldsymbol{z} = \frac{Z_q}{Z_p}\int f(\boldsymbol{z})\frac{\tilde{p}(\boldsymbol{z})}{\tilde{q}(\boldsymbol{z})}q(\boldsymbol{z})$$

Computing Expectations via Importance Sampling

- What if we can't even evaluate $p(\boldsymbol{z})$ but can only do so up to proportionality constant
  - Suppose $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$ and only $\tilde{p}(\boldsymbol{z})$ can be evaluated
- Let's also write our proposal distribution in the same form: $q(\boldsymbol{z}) = \frac{\tilde{q}(\boldsymbol{z})}{Z_q}$
- Given $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$, the I.S. based expectation will be

$$\mathbb{E}[f] = \int f(\boldsymbol{z})\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}q(\boldsymbol{z})d\boldsymbol{z} = \int f(\boldsymbol{z})\frac{\tilde{p}(\boldsymbol{z})/Z_p}{\tilde{q}(\boldsymbol{z})/Z_q}q(\boldsymbol{z})d\boldsymbol{z} = \frac{Z_q}{Z_p}\int f(\boldsymbol{z})\frac{\tilde{p}(\boldsymbol{z})}{\tilde{q}(\boldsymbol{z})}q(\boldsymbol{z}) \approx \frac{Z_q}{Z_p}\frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)})\frac{\tilde{p}(\boldsymbol{z}^{(\ell)})}{\tilde{q}(\boldsymbol{z}^{(\ell)})}$$

Computing Expectations via Importance Sampling

- What if we can't even evaluate $p(\boldsymbol{z})$ but can only do so up to proportionality constant
  - Suppose $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$ and only $\tilde{p}(\boldsymbol{z})$ can be evaluated
- Let's also write our proposal distribution in the same form: $q(\boldsymbol{z}) = \frac{\tilde{q}(\boldsymbol{z})}{Z_q}$
- Given $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$, the I.S. based expectation will be

$$\mathbb{E}[f] = \int f(\boldsymbol{z}) \frac{p(\boldsymbol{z})}{q(\boldsymbol{z})} q(\boldsymbol{z}) d\boldsymbol{z} = \int f(\boldsymbol{z}) \frac{\tilde{p}(\boldsymbol{z})/Z_p}{\tilde{q}(\boldsymbol{z})/Z_q} q(\boldsymbol{z}) d\boldsymbol{z} = \frac{Z_q}{Z_p} \int f(\boldsymbol{z}) \frac{\tilde{p}(\boldsymbol{z})}{\tilde{q}(\boldsymbol{z})} q(\boldsymbol{z}) \approx \frac{Z_q}{Z_p} \frac{1}{L} \sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)}) \frac{\tilde{p}(\boldsymbol{z}^{(\ell)})}{\tilde{q}(\boldsymbol{z}^{(\ell)})}$$

- Define $\frac{\tilde{p}(\boldsymbol{z}^{(\ell)})}{\tilde{q}(\boldsymbol{z}^{(\ell)})} = \tilde{r}_\ell$.

Computing Expectations via Importance Sampling

- What if we can't even evaluate $p(\boldsymbol{z})$ but can only do so up to proportionality constant
  - Suppose $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$ and only $\tilde{p}(\boldsymbol{z})$ can be evaluated
- Let's also write our proposal distribution in the same form: $q(\boldsymbol{z}) = \frac{\tilde{q}(\boldsymbol{z})}{Z_q}$
- Given $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$, the I.S. based expectation will be

$$\mathbb{E}[f] = \int f(\boldsymbol{z})\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}q(\boldsymbol{z})d\boldsymbol{z} = \int f(\boldsymbol{z})\frac{\tilde{p}(\boldsymbol{z})/Z_p}{\tilde{q}(\boldsymbol{z})/Z_q}q(\boldsymbol{z})d\boldsymbol{z} = \frac{Z_q}{Z_p}\int f(\boldsymbol{z})\frac{\tilde{p}(\boldsymbol{z})}{\tilde{q}(\boldsymbol{z})}q(\boldsymbol{z}) \approx \frac{Z_q}{Z_p}\frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)})\frac{\tilde{p}(\boldsymbol{z}^{(\ell)})}{\tilde{q}(\boldsymbol{z}^{(\ell)})}$$

- Define $\frac{\tilde{p}(\boldsymbol{z}^{(\ell)})}{\tilde{q}(\boldsymbol{z}^{(\ell)})} = \tilde{r}_\ell$. We also need $\frac{Z_q}{Z_p}$.

Computing Expectations via Importance Sampling

- What if we can't even evaluate $p(\boldsymbol{z})$ but can only do so up to proportionality constant
  - Suppose $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$ and only $\tilde{p}(\boldsymbol{z})$ can be evaluated
- Let's also write our proposal distribution in the same form: $q(\boldsymbol{z}) = \frac{\tilde{q}(\boldsymbol{z})}{Z_q}$
- Given $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$, the I.S. based expectation will be

$$\mathbb{E}[f] = \int f(\boldsymbol{z})\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}q(\boldsymbol{z})d\boldsymbol{z} = \int f(\boldsymbol{z})\frac{\tilde{p}(\boldsymbol{z})/Z_p}{\tilde{q}(\boldsymbol{z})/Z_q}q(\boldsymbol{z})d\boldsymbol{z} = \frac{Z_q}{Z_p}\int f(\boldsymbol{z})\frac{\tilde{p}(\boldsymbol{z})}{\tilde{q}(\boldsymbol{z})}q(\boldsymbol{z}) \approx \frac{Z_q}{Z_p}\frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)})\frac{\tilde{p}(\boldsymbol{z}^{(\ell)})}{\tilde{q}(\boldsymbol{z}^{(\ell)})}$$

- Define $\frac{\tilde{p}(\boldsymbol{z}^{(\ell)})}{\tilde{q}(\boldsymbol{z}^{(\ell)})} = \tilde{r}_\ell$. We also need $\frac{Z_q}{Z_p}$. Note that $\frac{Z_p}{Z_q} \approx \frac{1}{L}\sum_{m=1}^{L}\tilde{r}_m$ (exercise: verify yourself)

Computing Expectations via Importance Sampling

- What if we can't even evaluate $p(\boldsymbol{z})$ but can only do so up to proportionality constant
  - Suppose $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$ and only $\tilde{p}(\boldsymbol{z})$ can be evaluated
- Let's also write our proposal distribution in the same form: $q(\boldsymbol{z}) = \frac{\tilde{q}(\boldsymbol{z})}{Z_q}$
- Given $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$, the I.S. based expectation will be

$$\mathbb{E}[f] = \int f(\boldsymbol{z})\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}q(\boldsymbol{z})d\boldsymbol{z} = \int f(\boldsymbol{z})\frac{\tilde{p}(\boldsymbol{z})/Z_p}{\tilde{q}(\boldsymbol{z})/Z_q}q(\boldsymbol{z})d\boldsymbol{z} = \frac{Z_q}{Z_p}\int f(\boldsymbol{z})\frac{\tilde{p}(\boldsymbol{z})}{\tilde{q}(\boldsymbol{z})}q(\boldsymbol{z}) \approx \frac{Z_q}{Z_p}\frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)})\frac{\tilde{p}(\boldsymbol{z}^{(\ell)})}{\tilde{q}(\boldsymbol{z}^{(\ell)})}$$

- Define $\frac{\tilde{p}(\boldsymbol{z}^{(\ell)})}{\tilde{q}(\boldsymbol{z}^{(\ell)})} = \tilde{r}_\ell$. We also need $\frac{Z_q}{Z_p}$. Note that $\frac{Z_p}{Z_q} \approx \frac{1}{L}\sum_{m=1}^{L}\tilde{r}_m$ (exercise: verify yourself)
- Plugging these back into the expression of $\mathbb{E}[f]$, we have

Computing Expectations via Importance Sampling

- What if we can't even evaluate $p(\boldsymbol{z})$ but can only do so up to proportionality constant
  - Suppose $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$ and only $\tilde{p}(\boldsymbol{z})$ can be evaluated
- Let's also write our proposal distribution in the same form: $q(\boldsymbol{z}) = \frac{\tilde{q}(\boldsymbol{z})}{Z_q}$
- Given $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$, the I.S. based expectation will be

$$\mathbb{E}[f] = \int f(\boldsymbol{z})\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}q(\boldsymbol{z})d\boldsymbol{z} = \int f(\boldsymbol{z})\frac{\tilde{p}(\boldsymbol{z})/Z_p}{\tilde{q}(\boldsymbol{z})/Z_q}q(\boldsymbol{z})d\boldsymbol{z} = \frac{Z_q}{Z_p}\int f(\boldsymbol{z})\frac{\tilde{p}(\boldsymbol{z})}{\tilde{q}(\boldsymbol{z})}q(\boldsymbol{z}) \approx \frac{Z_q}{Z_p}\frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)})\frac{\tilde{p}(\boldsymbol{z}^{(\ell)})}{\tilde{q}(\boldsymbol{z}^{(\ell)})}$$

- Define $\frac{\tilde{p}(\boldsymbol{z}^{(\ell)})}{\tilde{q}(\boldsymbol{z}^{(\ell)})} = \tilde{r}_\ell$. We also need $\frac{Z_q}{Z_p}$. Note that $\frac{Z_p}{Z_q} \approx \frac{1}{L}\sum_{m=1}^{L}\tilde{r}_m$ (exercise: verify yourself)
- Plugging these back into the expression of $\mathbb{E}[f]$, we have

$$\mathbb{E}[f] \approx \sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)})\frac{\tilde{r}_\ell}{\sum_{m=1}^{L}\tilde{r}_m}$$

Computing Expectations via Importance Sampling

- What if we can't even evaluate $p(\boldsymbol{z})$ but can only do so up to proportionality constant
  - Suppose $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$ and only $\tilde{p}(\boldsymbol{z})$ can be evaluated
- Let's also write our proposal distribution in the same form: $q(\boldsymbol{z}) = \frac{\tilde{q}(\boldsymbol{z})}{Z_q}$
- Given $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$, the I.S. based expectation will be

$$\mathbb{E}[f] = \int f(\boldsymbol{z}) \frac{p(\boldsymbol{z})}{q(\boldsymbol{z})} q(\boldsymbol{z}) d\boldsymbol{z} = \int f(\boldsymbol{z}) \frac{\tilde{p}(\boldsymbol{z})/Z_p}{\tilde{q}(\boldsymbol{z})/Z_q} q(\boldsymbol{z}) d\boldsymbol{z} = \frac{Z_q}{Z_p} \int f(\boldsymbol{z}) \frac{\tilde{p}(\boldsymbol{z})}{\tilde{q}(\boldsymbol{z})} q(\boldsymbol{z}) \approx \frac{Z_q}{Z_p} \frac{1}{L} \sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)}) \frac{\tilde{p}(\boldsymbol{z}^{(\ell)})}{\tilde{q}(\boldsymbol{z}^{(\ell)})}$$

- Define $\frac{\tilde{p}(\boldsymbol{z}^{(\ell)})}{\tilde{q}(\boldsymbol{z}^{(\ell)})} = \tilde{r}_\ell$. We also need $\frac{Z_q}{Z_p}$. Note that $\frac{Z_p}{Z_q} \approx \frac{1}{L} \sum_{m=1}^{L} \tilde{r}_m$ (exercise: verify yourself)
- Plugging these back into the expression of $\mathbb{E}[f]$, we have

$$\mathbb{E}[f] \approx \sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)}) \frac{\tilde{r}_\ell}{\sum_{m=1}^{L} \tilde{r}_m} = \sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)}) w_\ell$$

An Application: Monte Carlo EM

- Recall that the EM algorithm requires computing the expected CLL

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\Theta^{old})}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]$$

An Application: Monte Carlo EM

- Recall that the EM algorithm requires computing the expected CLL

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\Theta^{old})}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]$$

- If the above expectation is intractable, one way is to replace it by a Monte Carlo approximation

An Application: Monte Carlo EM

- Recall that the EM algorithm requires computing the expected CLL

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\Theta^{old})}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]$$

- If the above expectation is intractable, one way is to replace it by a Monte Carlo approximation
  - Draw L samples $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(L)}$ from the posterior $p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$

An Application: Monte Carlo EM

- Recall that the EM algorithm requires computing the expected CLL

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\Theta^{old})}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]$$

- If the above expectation is intractable, one way is to replace it by a Monte Carlo approximation
  - Draw L samples $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(L)}$ from the posterior $p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$
  - Use Monte Carlo approximation of the expected CLL as

$$\mathcal{Q}(\Theta, \Theta^{old}) \approx \frac{1}{L} \sum_{\ell=1}^{L} \log p(\mathbf{X}, \mathbf{Z}^{(\ell)}|\Theta)$$

An Application: Monte Carlo EM

- Recall that the EM algorithm requires computing the expected CLL

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\Theta^{old})}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]$$

- If the above expectation is intractable, one way is to replace it by a Monte Carlo approximation
  - Draw L samples $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(L)}$ from the posterior $p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$
  - Use Monte Carlo approximation of the expected CLL as

$$\mathcal{Q}(\Theta, \Theta^{old}) \approx \frac{1}{L}\sum_{\ell=1}^{L} \log p(\mathbf{X}, \mathbf{Z}^{(\ell)}|\Theta)$$

  - Do MLE on this approximation

Summary and Comments

- Monte Carlo, Rejection Sampling, and Importance Sampling are conceptually simple methods

Summary and Comments

- Monte Carlo, Rejection Sampling, and Importance Sampling are conceptually simple methods
- Usually work well for “small” problems (with a few variables)

Summary and Comments

- Monte Carlo, Rejection Sampling, and Importance Sampling are conceptually simple methods
- Usually work well for "small" problems (with a few variables)
- However, these methods suffer in high-dimensions ($\boldsymbol{z}$ consists of lots of r.v.'s)

Summary and Comments

- Monte Carlo, Rejection Sampling, and Importance Sampling are conceptually simple methods
- Usually work well for "small" problems (with a few variables)
- However, these methods suffer in high-dimensions ($\boldsymbol{z}$ consists of lots of r.v.'s)
- Also note that MC and IS are not really "sampling" methods (only compute expectations)

Summary and Comments

- Monte Carlo, Rejection Sampling, and Importance Sampling are conceptually simple methods
- Usually work well for "small" problems (with a few variables)
- However, these methods suffer in high-dimensions ($\boldsymbol{z}$ consists of lots of r.v.'s)
- Also note that MC and IS are not really "sampling" methods (only compute expectations)
- For R.S. and I.S., it may be difficult to come up with good proposal distributions

Summary and Comments

- Monte Carlo, Rejection Sampling, and Importance Sampling are conceptually simple methods
- Usually work well for "small" problems (with a few variables)
- However, these methods suffer in high-dimensions ($\boldsymbol{z}$ consists of lots of r.v.'s)
- Also note that MC and IS are not really "sampling" methods (only compute expectations)
- For R.S. and I.S., it may be difficult to come up with good proposal distributions
- More sophisticated sampling algorithms needed, e.g., MCMC methods (after the mid-sem)