

Sampling from Probability Distributions

CS772A: Probabilistic Machine Learning

Piyush Rai

Other Divergence Measures

- VI minimizes $KL(q||p)$ but other divergences can be minimized as well
 - Recall that VI with minimization of $KL(q||p)$ leads to underestimated variances
- A general form of divergence is Renyi's α -divergence defined as

$$D_{\alpha}^R(p(\mathbf{Z})||q(\mathbf{Z})) = \frac{1}{\alpha - 1} \log \int p(\mathbf{Z})^{\alpha} q(\mathbf{Z})^{1-\alpha} d\mathbf{Z}$$

- $KL(p||q)$ is a special case with $\alpha \rightarrow 1$ (can verify using L'Hopital rule of taking limits)
- An even more general form of divergence is f -Divergence

$$D_f(p(\mathbf{Z})||q(\mathbf{Z})) = \int q(\mathbf{Z}) f\left(\frac{p(\mathbf{Z})}{q(\mathbf{Z})}\right) d\mathbf{Z}$$

- Many recent variational inference algorithms are based on minimizing such divergences



Variational Inference: Some Comments

- Many probabilistic models nowadays rely on VI to do approx. inference
- Even mean-field with locally-conjugacy used in lots of models
 - This + SVI gives excellent scalability as well on large datasets
- Progress in various areas has made VI very popular and widely applicable
 - Stochastic Optimization (e.g., SGD)
 - Automatic Differentiation
 - Monte-Carlo gradient of ELBO
- Note: Most of these ideas apply also to [Variational EM](#)
- Many VI and advanced VI algos are implemented in probabilistic prog. packages (e.g., Tensorflow Probability, PyTorch, etc), making VI easy even for complex models
- Still a very active area of research, especially for doing VI in complex models
 - Models with discrete latent variables
 - Reducing the variance in Monte-Carlo estimate of ELBO gradients
 - More expressive variational distribution for better approximation

We covered many of the threads being explored in recent work but a lot of work still being done in this area



Sampling for Approximate Inference

- Some typical tasks that we have to solve in probabilistic/fully-Bayesian inference

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

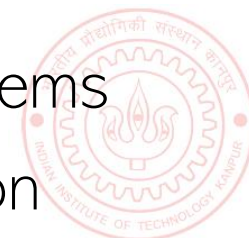
Posterior predictive distribution $\rightarrow p(\mathcal{D}^{new}|\mathcal{D}) = \int p(\mathcal{D}^{new}|\theta)p(\theta|\mathcal{D})d\theta = \mathbb{E}_{p(\theta|\mathcal{D})}[p(\mathcal{D}^{new}|\theta)]$

Needed for model selection (and in computing posterior too) \rightarrow Marginal likelihood $\rightarrow p(\mathcal{D}|m) = \int p(\mathcal{D}|\theta)p(\theta|m)d\theta = \mathbb{E}_{p(\theta|m)}[p(\mathcal{D}|\theta)]$

Needed in EM \rightarrow Expected complete data log-likelihood $\rightarrow \text{Exp-CLL} = \int p(\mathbf{z}|\theta, \mathbf{x})p(\mathbf{x}, \mathbf{z}|\theta)d\mathbf{z} = \mathbb{E}_{p(\mathbf{z}|\theta, \mathbf{x})}[p(\mathbf{x}, \mathbf{z}|\theta)]$

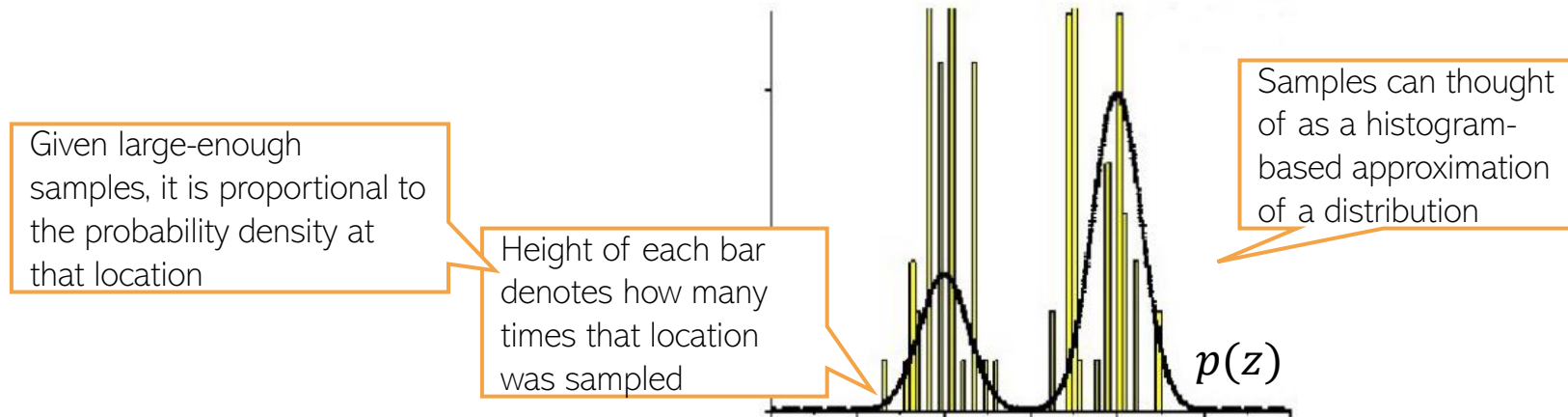
Needed in VI \rightarrow Evidence lower bound (ELBO) $\rightarrow \mathcal{L}(q) = \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{z})] - \mathbb{E}_q[\log p(\mathbf{z})]$

- Sampling methods provide a general way to (approximately) solve these problems
- More general than VI methods which only approximate the posterior distribution



Approximating a Prob. Distribution using Samples ⁵

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



- The samples can also be used for computing expectations (Monte-Carlo averaging)
- Usually straightforward to generate samples if it is a simple/standard distribution
- The interesting bit: Even if the distribution is “difficult” (e.g., an intractable posterior), it is often possible to generate random samples from such a distribution, as we will see.



The Empirical Distribution

- Sampling based approx. can be formally represented using an **empirical distribution**
- Given L points/samples $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \dots, \mathbf{z}^{(L)}$, empirical distr. defined by these is

Dirac Distribution with finite support at $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \dots, \mathbf{z}^{(L)}$

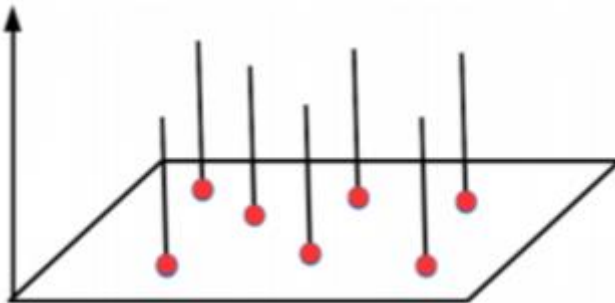
Weights sum to 1

Weight of point $\mathbf{z}^{(\ell)}$

Can think of A as being the area over which we want to evaluate the distribution

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

Dirac Distribution

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




Sampling: Some Basic Methods

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

Determinant
of Jacobian

- Most of these basic methods are based on the idea of transformation
 - Generate a random sample \mathbf{x} from a distribution $q(\mathbf{x})$ which is easy to sample from
 - Apply a transformation on \mathbf{x} to make it random sample \mathbf{z} from a complex distr $p(\mathbf{z})$

$F(z)$: CDF of $p(z)$

- Some popular examples of transformation methods

- Inverse CDF method

$$\mathbf{x} \sim \text{Unif}(0, 1) \Rightarrow \mathbf{z} = \text{Inv-CDF}_{p(\mathbf{z})}(\mathbf{x}) \sim p(\mathbf{z})$$

- Reparametrization method

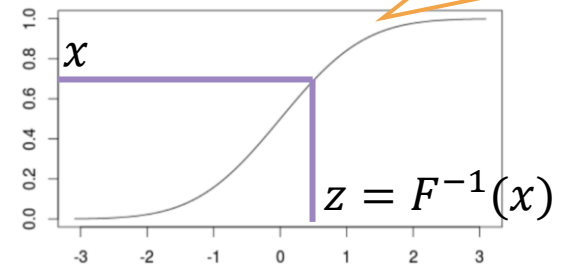
$$\mathbf{x} \sim \mathcal{N}(0, 1) \Rightarrow \mathbf{z} = \mu + \sigma \mathbf{x} \sim \mathcal{N}(\mu, \sigma^2)$$

- Box-Mueller method: Given (x_1, x_2) from $\text{Unif}(0, 1)$, generate (z_1, z_2) from $\mathcal{N}(0, \mathbf{I}_2)$

$$z_1 = \sqrt{-2 \ln x_1} \cos(2\pi x_2), \quad z_2 = \sqrt{-2 \ln x_1} \sin(2\pi x_2)$$

- Transformation Methods are simple but have limitations

- Mostly limited to standard distributions and/or distributions with very few variables



Rejection Sampling

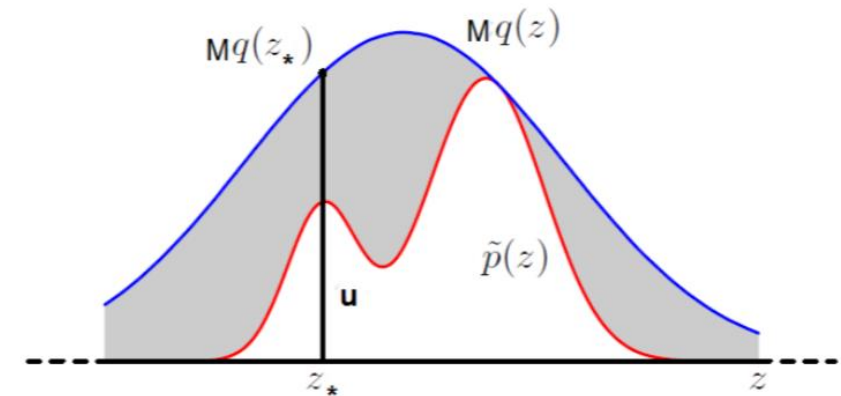
- Goal: Generate a random sample from a distribution of the form $p(z) = \frac{\tilde{p}(z)}{Z_p}$, assuming
 - We can only evaluate the value of numerator $\tilde{p}(z)$ for any z
 - The denominator (normalization constant) Z_p is intractable and we don't know its value

Should have the same support as $p(z)$

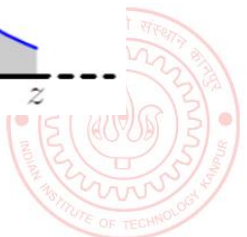
- Assume a **proposal distribution** $q(z)$ we can generate samples from, and

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

- Rejection Sampling then works as follows
 - Sample a random variable z_* from $q(z)$
 - Sampling a uniform r.v. $u \sim \text{Unif}[0, Mq(z_*)]$
 - If $u \leq \tilde{p}(z_*)$ then accept z_* , otherwise reject it



- All accepted z_* 's will be random samples from $p(z)$. Proof on next slide



Rejection Sampling

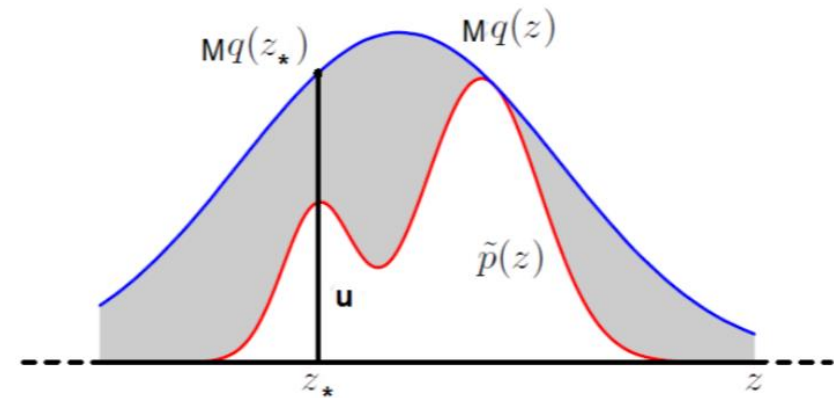
- Why $z \sim q(z)$ + accept/reject rule is equivalent to $z \sim p(z)$?
- Let's look at the pdf of the z 's that were accepted, i.e., $p(z|\text{accept})$

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

$$p(z, \text{accept}) = q(z)p(\text{accept}|z) = \frac{\tilde{p}(z)}{M}$$

$$p(\text{accept}) = \int \frac{\tilde{p}(z)}{M} dz = \frac{Z_p}{M}$$

$$p(z|\text{accept}) = \frac{p(z, \text{accept})}{p(\text{accept})} = \frac{\tilde{p}(z)}{Z_p} = p(z)$$



Computing Expectations via Monte Carlo Sampling¹⁰

- Often we are interested in computing expectations of the form

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

where $f(z)$ is some function of the random variable $z \sim p(z)$

- A simple approx. scheme to compute the above expectation: Monte Carlo integration

- Generate L independent samples from $p(z)$: $\{z^{(\ell)}\}_{\ell=1}^L \sim p(z)$
- Approximate the expectation by the following empirical average

Assuming we know how to sample from $p(z)$

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

- Since the samples are independent of each other, we can show the following (exercise)

Unbiased expectation

$$\mathbb{E}[\hat{f}] = \mathbb{E}[f]$$

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

Variance in our estimate decreases as L increases

Computing Expectations via Importance Sampling¹¹

- How to compute Monte Carlo expec. if we don't know how to sample from $p(\mathbf{z})$?
- One way is to use transformation methods or rejection sampling
- Another way is to use **Importance Sampling** (assuming $p(\mathbf{z})$ can be evaluated at least)
 - Generate L indep samples from a **proposal** $q(\mathbf{z})$ we know how sample from: $\{\mathbf{z}^{(\ell)}\}_{\ell=1}^L \sim q(\mathbf{z})$
 - Now approximate the expectation as follows

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

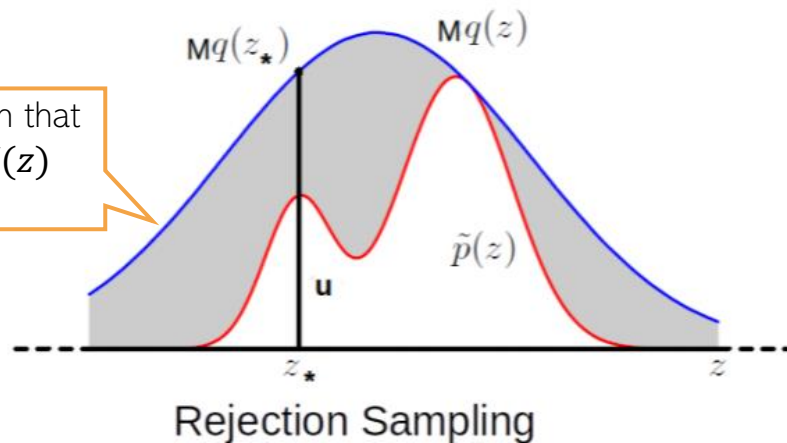
- This is basically “weighted” Monte Carlo integration
 - $w^{(\ell)} = \frac{p(\mathbf{z}^{(\ell)})}{q(\mathbf{z}^{(\ell)})}$ denotes the **importance weight** of each sample $\mathbf{z}^{(\ell)}$
- IS works even when we can only evaluate $p(\mathbf{z}) = \frac{\tilde{p}(\mathbf{z})}{Z_p}$ up to a prop. constant
- Note: Monte Carlo and Importance Sampling are NOT sampling methods!
 - These are only uses for computing expectations (approximately)

See PRML 11.1.4

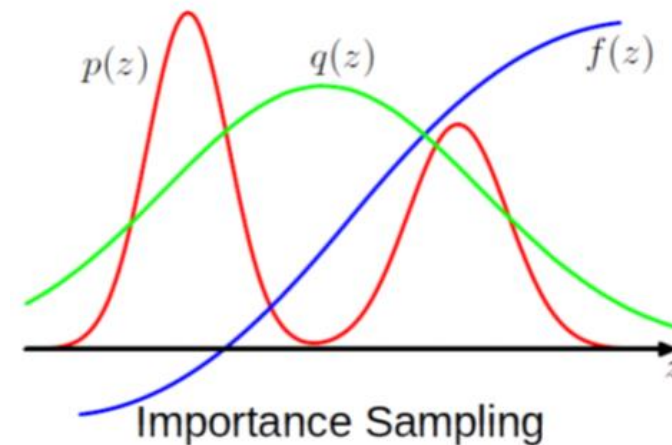


Limitations of the Basic Methods

- Transformation based methods: Usually limited to drawing from standard distributions
- Rejection Sampling and Importance Sampling: Require good proposal distributions



$q(z)$ should be such that $Mq(z)$ envelopes $\tilde{p}(z)$ everywhere

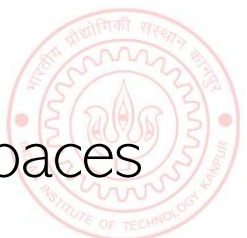


$$\mathbb{E}[f] \approx \frac{1}{L} \sum_{\ell=1}^L f(z^{(\ell)}) \frac{p(z^{(\ell)})}{q(z^{(\ell)})}$$

Ideally, would like $q(z)$ to give samples from where $p(z)$ is large or $f(z)p(z)$ is large

Difficult to guarantee so if z is high-dimensional

- In general, difficult to find good prop. distr. especially when z is high-dim
- More sophisticated sampling methods like MCMC work well in such high-dim spaces

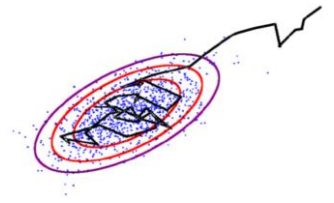


Markov Chain Monte Carlo (MCMC)

If the target is a posterior, it will be conditioned on data, i.e., $p(\mathbf{z}|\mathbf{x})$

- Goal: Generate samples from some target distribution $p(\mathbf{z}) = \frac{\tilde{p}(\mathbf{z})}{Z_p}$
- Assume we can evaluate $p(\mathbf{z})$ at least up to a proportionality constant
 - \mathbf{z} usually is high-dim
 - Means we can at least evaluate $\tilde{p}(\mathbf{z})$
- MCMC uses a **Markov Chain** which, when converged, starts giving samples from $p(\mathbf{z})$

$\mathbf{z}^{(1)} \rightarrow \mathbf{z}^{(2)} \rightarrow \mathbf{z}^{(3)} \rightarrow \dots \rightarrow \mathbf{z}^{(L-2)} \rightarrow \mathbf{z}^{(L-1)} \rightarrow \mathbf{z}^{(L)}$
 initial samples typically garbage after convergence, actual samples from $p(\mathbf{z})$

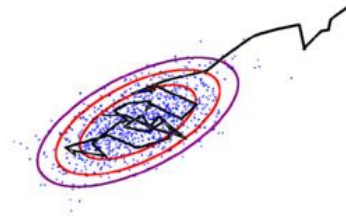


- Given current sample $\mathbf{z}^{(\ell)}$ from the chain, MCMC generates the next sample $\mathbf{z}^{(\ell+1)}$ as
 - Use a **proposal distribution** $q(\mathbf{z}|\mathbf{z}^{(\ell)})$ to generate a candidate sample \mathbf{z}_*
 - **Accept/reject** \mathbf{z}_* as the next sample based on an **acceptance criterion** (will see later)
 - If accepted, set $\mathbf{z}^{(\ell+1)} = \mathbf{z}_*$. If rejected, set $\mathbf{z}^{(\ell+1)} = \mathbf{z}^{(\ell)}$
- Important: The proposal distribution $q(\mathbf{z}|\mathbf{z}^{(\ell)})$ depends on the previous sample $\mathbf{z}^{(\ell)}$
 - Should also have the same support as $p(\mathbf{z})$



MCMC: The Basic Scheme

14



- The chain run infinitely long (i.e., upon convergence) will give ONE sample from $p(\mathbf{z})$
- But we usually require **several samples** to approximate $p(\mathbf{z})$
- This is done as follows
 - Start the chain at an initial $\mathbf{z}^{(0)}$
 - Using the proposal $q(\mathbf{z}|\mathbf{z}^{(\ell)})$, run the chain long enough, say T_1 steps
 - Discard the first $T_1 - 1$ samples (called “**burn-in**” **samples**) and take last sample $\mathbf{z}^{(T_1)}$
 - Continue from $\mathbf{z}^{(T_1)}$ up to T_2 steps, discard intermediate samples, take last sample $\mathbf{z}^{(T_2)}$
 - This discarding (called “**thinning**”) helps ensure that $\mathbf{z}^{(T_1)}$ and $\mathbf{z}^{(T_2)}$ are **uncorrelated**
 - Repeat the same for a total of S times
 - In the end, we now have S *approximately independent* samples from $p(\mathbf{z})$
- Note: Good choices for T_1 and $T_i - T_{i-1}$ (thinning gap) are usually based on heuristics

MCMC is exact in theory but approximate in practice since we can't run the chain for infinitely long in practice

Thus we say that the samples are approximately from the target distribution

Will treat it as our first sample from $p(\mathbf{z})$

Requirement for Monte Carlo approximation

