Why Sampling?

Chapter 12 of KF textbook.

Very often we need to sample instances from a joint distribution $P(\mathbf{y}_1, \ldots, y_n | \mathbf{x})$ or $P(\mathbf{x} = x_1 \ldots, x_n)$ or $P(x_1, \ldots x_r | x_{r+1} = E_{r+1}, \ldots x_n = E_n)$. Here are some scenarios:

- During training we need to solve an intractable inference.
- We need to show a diverse set of outputs to a user instead of just the most likely value. Example: in translation.
- We need to calculate expected value of some arbitrary function $f(\mathbf{x})$ under distribution $P(\mathbf{x})$. What is the expected number of times that adjacent positions have the same label for a given $\mathbf{x}$?

Motivation: Inference Deep Language Models

- Generate sample sentences
- Generate questions
- Expected distribution of first word for sentences ending with '?'.

Motivation: Inference from VAEs

- Fix values of some of the outputs and generate most likely values of others — application missing value imputation.

$P(x) \quad x \equiv x_1, \underbrace{x_2 \ldots x_n}$

Sampling to approximate expected value of a function under $P(\mathbf{x})$

$\mathcal{X} \equiv$ space of x

eg: space of all possible sentences.

$f(x) \to R$

$$E_{P(x)}[f(x)] = \sum_{x \in \mathcal{X}} f(x)\, P(x)$$

x is discrete.

$$\int f(x)\, P(x)\, dx$$

$\mathcal{X} \leftarrow$ large

integral cannot be computed in closed form.

Approximate using sampling

$x^1, x^2, \ldots x^M \sim P(x)$

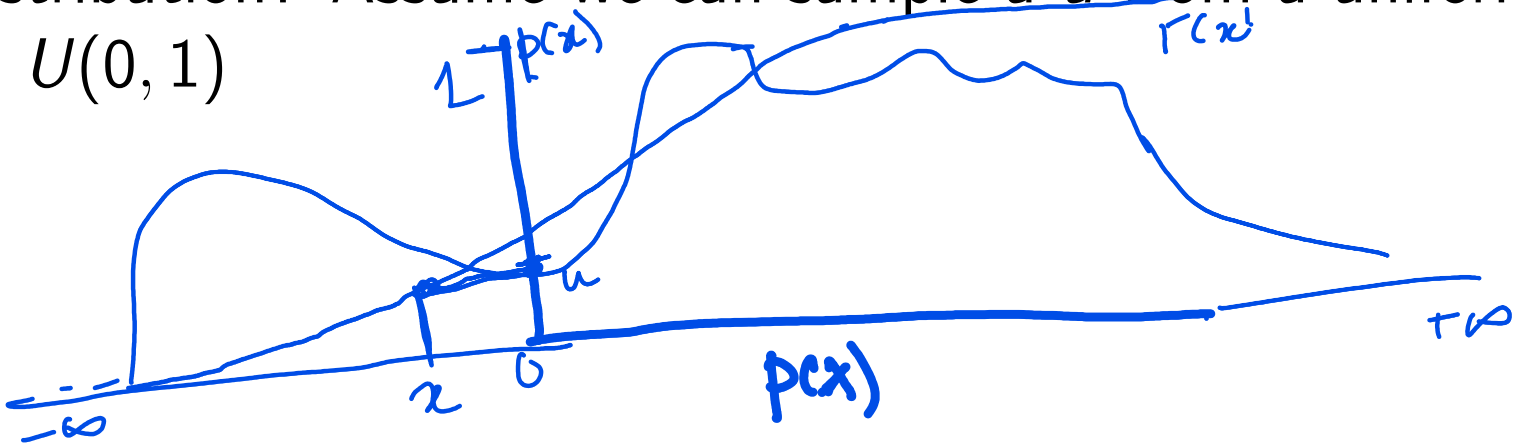
$$\approx \frac{1}{M} \sum_{i=1}^{M} f(x^i)$$

$M \to \infty \Rightarrow$ this approximation will match exact expected value.

Basics: Sampling scalar distributions

$x \in \mathbb{R}$

Let $p(x)$ be a distribution. How do we draw M samples $x^1, \ldots, x^M$ from the distribution? Assume we can sample a u from a uniform distribution $U(0,1)$



Let $F(x)$ be cumulative distribution of $p(x)$.

For $i = 1 \ldots M$

1. Sample $u_i \sim U(0,1)$
2. Find $x^i = F^{-1}(u_i)$

Basics: Sampling from multinomial distributions

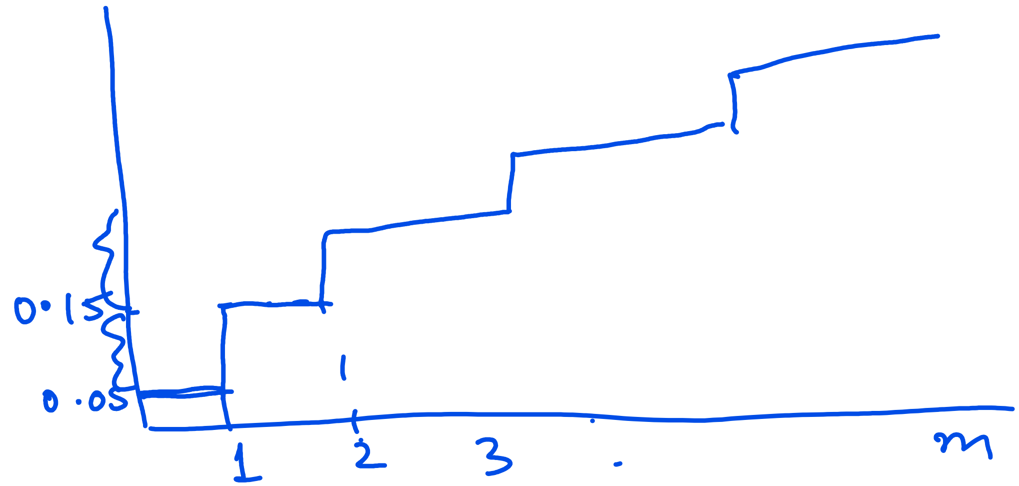
x is discrete, $x \in \{1, \ldots m\}$
$p(x) \sim Mult(p_1, \ldots, p_m)$

$u_i \overset{N}{\sim} [0, 1]$

if u_i is between $\sum_{j=0}^{k-1} p_j = \alpha_k$

and $\alpha_k + p_k$

then outp k



$p = 0.05$, $p = 0.1$

Consistent samples

As $M \to \infty$, the fraction of times in the sample that we encounter a sample in an interval $[x, x + \Delta)$ would be proportional to the true probability of that interval in $p(x)$ i.e. $F(x+\Delta) - F(x)$

How to sample multivariate distributions?

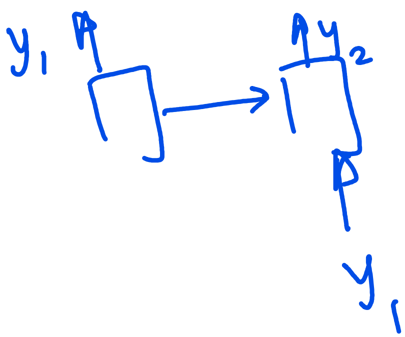
Option 1: Factorize the distribution as a Bayesian network and perform forward sampling.

eg: from a auto-regressive language model.

Assume

$$P(\vec{x} = x_1, x_2 \ldots x_n) = \prod_{j=1}^{n} P(x_j \mid Pa(x_j))$$

$$P(\vec{y} = y_1 \ldots y_n \mid x)$$



$$y_j \sim P(y_j \mid \underbrace{y_1 \ldots y_{j-1}}_{s_j}, x)$$

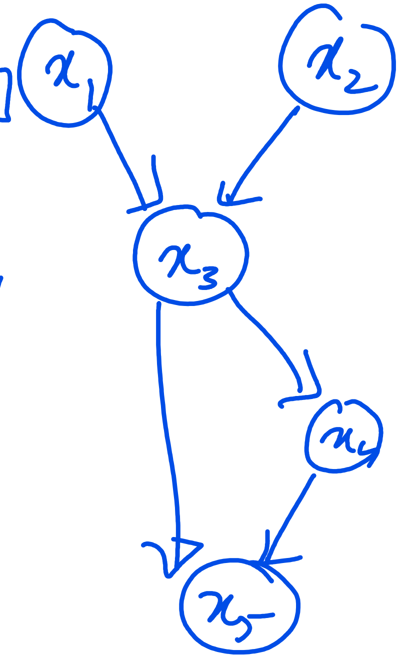
$$= P(y_j \mid s_j, x) = \text{Softmax over words.}$$

Example.

Forward Sampling Algorithm for BN

Let $x_1, x_2, \ldots x_n$ be topologically sorted as per the BN ~~graph~~ graph

$x_i \sim [1, \ldots m]$



for $i = 1$ to M

  $\xi^i = [0 \ldots 0]$

  for $j = 1$ to n /* topological order */

    $\xi^i_j \sim P(x_j \mid Pa(x_j) = \xi^i_{Pa(x_j)})$

Return $\xi^1, \xi^2, \ldots \xi^M$

Importance Sampling

- When '!' is the last token, what is the probability of x_1 being 'what'? In forward sampling most of the sampled sentences would be wasted since they would not end with '!'.
- Complete missing attribute in a VAE network for object generation. Forward sampling would not match given values most of the time.
- In general: importance sampling is useful when it is hard to sample from $P(\mathbf{x})$ or to lower the error in computation of expected value of a function.

$$E[f(\mathbf{x})] = \sum_{\mathbf{x}} P(\mathbf{x}) f(\mathbf{x})$$

where $f(\mathbf{x})$ is zero for many $\mathbf{x}$. Example, rare combinations.
Importance sampling — sample from the *important* regions.

Estimation with importance sampling.

Proposal distribution: $Q(\mathbf{x})$ from which it is easy to generate samples. Designing a good proposal distribution is an 'art' and problem-dependent.
Example $Q(x)$ for the LM task: a reverse LM.

Estimation with importance sampling.

- Get M samples S_Q from $Q(\mathbf{x})$: $\mathbf{x}^1, \ldots, \mathbf{x}^M$
- If we use these samples to estimate $E[f(\mathbf{x})]$, the estimate is not consistent.

$\hat{\mu}(S_Q^M) = \frac{1}{M}\sum_{i=1}^{M} f(x^i)$ ... lt $M \to \infty$ $\frac{1}{M}\sum_{i=1}^{M} f(x^i) \to E_{Q(x)}[f(x)]$

$\neq E_{P(x)}[f(x)]$

$\forall f$ unless $P(x) = Q(x)$

- How to use S_Q to get a consistent estimate of $E[f(\mathbf{x})]$?

Estimation with importance sampling.

$$E_P[f(x)] = \sum_{x \in \mathcal{X}} P(x) f(x) = \sum_{x \in \mathcal{X}} f(x) \underbrace{\frac{P(x)}{Q(x)}}_{w(x)} Q(x)$$

$$= E_{Q(x)}[f(x) w(x)]$$

$$\approx \frac{1}{M} \sum_{i=1}^{M} f(x^i) w(x^i)$$

where $x^1, x^2 \ldots x^M \sim Q(x)$

$$w(x^i) = \frac{P(x^i)}{Q(x^i)}$$

Estimation with importance sampling.

Given M, $Q(x)$, $P(x)$

for $i = 1$ to M

$x^i \leftarrow$ sample from $Q(x)$

$w^i \leftarrow \frac{P(x^i)}{Q(x^i)}$

Return $\boxed{\{(x^i, w^i)\}}$ $i = 1$ to M.

$E_p[f(x)] \approx \frac{1}{M}\sum_{i=1}^{M} f(x^i)\, w^i$

Limitation: If $P(x)$ cannot be computed efficiently eg: in CRF with large tree-widths calculating $w(x^i)$ is not tractable.

Normalized importance sampling.

Let $P(x) = \dfrac{\tilde{P}(x)}{Z}$ ← un-normalized probability

Z ← intractable normalizer.

$$E_{P(x)}[f(x)] = \frac{1}{Z}\sum_{x \in \mathcal{X}} f(x) \frac{\tilde{P}(x)}{Q(x)} Q(x) = \frac{1}{Z}\sum_{x \in \mathcal{X}} f(x)\tilde{w}(x) Q(x)$$

$$\approx \frac{1}{Z}\frac{1}{M}\sum_{x \in S_Q^M} f(x)\tilde{w}(x)$$

where S_Q^M = set of M samples from $Q(x)$

How to compute Z?

$$Z = \sum_{x \in \mathcal{X}} \tilde{P}(x) \quad \text{[by definition]}.$$

$$= \sum_{x} \frac{\tilde{P}(x)}{Q(x)} Q(x) = \sum_{x \in \mathcal{X}} \tilde{w}(x) Q(x) \approx \frac{1}{M}\sum_{x \in S_Q^M} [\tilde{w}(x)]$$