Different Scenarios :

$$Z_p = \int \tilde{p}(x)\,dx$$

i.) Given x, compute $p(x)$, or $\tilde{p}(x)$, $p(x) = \dfrac{\tilde{p}(x)}{Z_p}$

ii) Sample from a distribution : Given $p(x)$, or more often $\tilde{p}(x)$, generate samples $x_i \sim p(x)$, $i = 1, \dots, n$

- Evaluate statistics / moments : $\mathbb{E}_{x \sim p(x)}[f(x)] = \int f(x)p(x)\,dx$

⊛ : Challenges arise when x is high-dimensional, $p(x)$ of $f(x)$ are complication.

e.g. Bayesian inference : $\underbrace{p(\theta|\mathcal{D})}_{\hookrightarrow \text{draw } \theta \sim p(\theta|\mathcal{D})} \propto p(\mathcal{D}|\theta)p(\theta)$

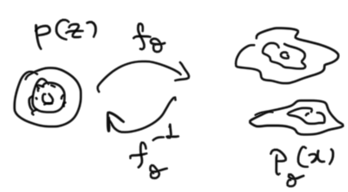
$$\underbrace{p(f_\theta(x^*)\,|\,x, y)}_{\text{predictive posterior}} = \int p(y|x,\theta)\,p(\theta|x,y)\,d\theta$$

$$= \underbrace{\mathbb{E}_{\theta \sim p(\theta|\mathcal{D})}\left[p(y|x,\theta)\right]}$$

Recall :

- Normalizing flow : given some samples x_i, estimate $p_\theta(x)$

$$\underbrace{p_\theta(x)} = \underbrace{p(z)}_{\substack{\text{Gaussian} \\ \uparrow}} \,|\det \nabla_x f_\theta(x)| = p(f_\theta^{-1}(x))\,|\det \nabla_x f_\theta$$



$z \sim \mathcal{N}(0, I)$, $x = f_\theta(z)$

- Variational inference :

$$p(x) \approx q_{\varphi}(x) \overset{M.F.}{=} \prod_{i=1}^{d} \mathcal{N}(x_i \mid \mu_i, \sigma_i^2)$$

$$\boxed{\varphi^* = \arg\min_{\varphi} \mathbb{KL}[q_{\theta}(x) \| p(x)]}$$

Examples :

i.) $$\mathbb{E}_{p \in \mathbb{R}}[h(p)] \approx \frac{1}{S} \sum_{i=1}^{S} h(p_i) \xrightarrow{S \to +\infty}$$

ii) Making predictions using an ML model :

$$p(f(x^*) \mid \mathcal{D}) = \mathbb{E}_{\theta \sim p(\theta \mid \mathcal{D})} [p(f(x^*) \mid \mathcal{D}, \theta)]$$

$$\approx \frac{1}{n} \sum p(f(x^*) \mid \mathcal{D}, \theta_i),$$

$$\theta_i \overset{i.i.d.}{\sim} p(\theta \mid \mathcal{D})$$

Why sampling :

i.) Allows us to compute expectations in high-dimensions $\left\{ \begin{array}{l} \text{estimate statisti} \\ \text{posterior inferen} \end{array} \right.$

$$p(X \in A) = \mathbb{E}[\mathbb{1}_{\{X \in A\}}], \quad \mathbb{1}_{\{X \in A\}} = \begin{cases} 1, & \text{if } X \in \\ 0, & \text{if } X \notin \end{cases}$$

ii) Allows us to compute intractable sums or integrals

Pros : 1.) Easy to implement / understand

2.) General purpose

3.) Well understood assymptotical theoretical guarantees.

3.) Well ...
(but they may be inefficient in practice)

Cons: 1.) They're too simple, ... often used innapprotiately.

2.) They tend to slow compared to deterministic approaches

3.) It can be difficult to assess their performance.

Monte Carlo Approximation :

Goal: Approximate an expectation using samples :

$$\mathbb{E}_{x \sim p(x)} [f(x)], \quad x \in \mathbb{R}^d .$$

Definition: If $x_1, x_2, \dots, x_n \overset{i.i.d}{\sim} p(x)$, then :

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} f(x_i)$$: a basic Monte Carlo (MC estimator

$$\approx \int f(x) p(x) dx$$

Remarks :

1.) $\mathbb{E}[\hat{\mu}_n] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[f(x_i)] \xrightarrow{n \to +\infty} \mathbb{E}[f(x)]$

hence $\hat{\mu}_n$ is an unbiased estimator

2.) $\hat{\mu}_n \xrightarrow[n \to +\infty]{P} \mathbb{E}[f(x)]$, $\forall \varepsilon > 0$, $P(|\hat{\mu}_n - \mathbb{E}[f(x)]| < \varepsilon)$ - as $n \to$

hence $\hat{\mu}_n$ is a consistent estimator

$(x_i \overset{i.i.d.}{\sim} p(x))$

3.) $\mathbb{V}ar[\hat{\mu}_n] = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{V}ar[f(x_i)] = \frac{1}{n} \underbrace{\mathbb{V}ar[f(x)]}_{\text{unknown}} \xrightarrow[n \to +\infty]{} 0$

$$\text{std}[\hat{\mu}_n] \sim C \cdot \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$

... d!

does not depend on ...

⊛ Practical limitation:

MC approximation relies on the fact that we can efficiently sample $x \sim p(x)$.

Importance Sampling (not a sampling metho...

i.) Use I.S. to approximate $\mathbb{E}_{x\sim p(x)}[f(x)]$ when sampling fro... $p(x)$ is not tractabl...

ii) Use I.S. to improve the accuracy/convergence rate of M.C. even if we could efficiently sample from $p(x)$

Setup: Assume that $p(x)$ is a density, i.e. $\int p(x)\,dx = 1$

$$\mathbb{E}_{x\sim p(x)}[f(x)] = \int f(x)p(x)\,dx \overset{\text{I.S.}}{=} \int \left[f(x)\frac{p(x)}{q(x)}\right] q(x)\,dx$$

$$= \mathbb{E}_{x\sim q(x)}\left[f(x)\frac{p(x)}{q(x)}\right] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)\,w(x_i),$$

$$\underbrace{w(x) = \frac{p(x)}{q(x)}}_{\text{importance weight}}, \qquad x_i \overset{\text{i.i.d.}}{\sim} q(x)$$

$q(x) \to$ proposal distribution.

$\begin{cases} \text{i.) Easy to evaluate.} \\ \text{ii.) Easy to sample from.} \end{cases}$ e.g. Gau...

Remarks:

- $\mathbb{E}\left[\hat{\mu}_n^{IS}\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{x\sim q(x)}\left[f(x_i)\frac{p(x_i)}{q(x_i)}\right] = \frac{1}{n}\sum_{i=1}^{n}\int f(x_i)\frac{p(x_i)}{q(x_i)}q(x_i)...$

$x \sim q(x)$

$= \hat{\mu}_n^{m.c.}$ which is unbiased.

- $\mathrm{Var}\left[\hat{\mu}_n^{IS}\right] = \frac{1}{n} \mathrm{Var}\left[f(x)\frac{p(x)}{q(x)}\right]$ $C \cdot \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$

could be made smaller by appropriately choosing $q(x)$

In fact, it is straightforward to solve for the optimal $q^*(x)$:

$$\boxed{\boxed{q^*(x)} = \min_q \mathrm{Var}\left[\hat{\mu}_n^{IS}\right]}$$

How to choose a good importance sampling proposal distrib

Example: $x \in \mathcal{X}$: investment

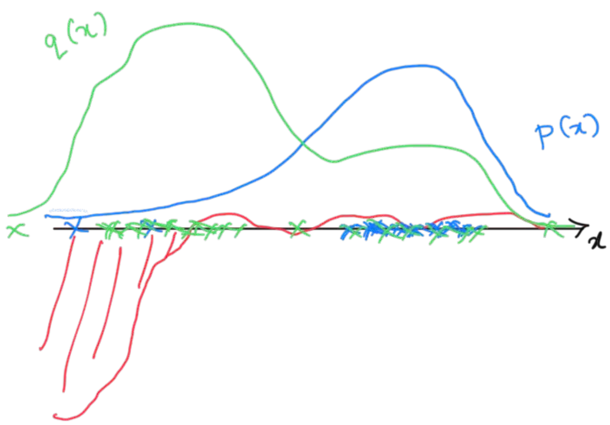
$f(x)$: return

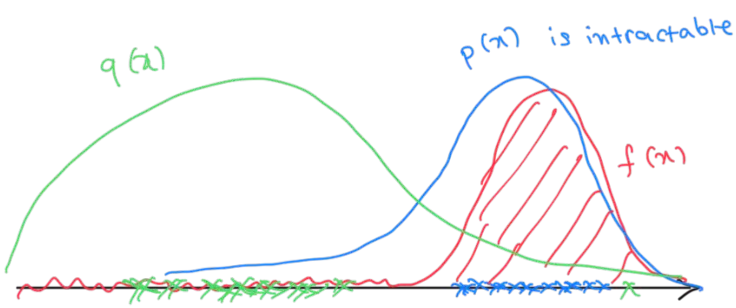
, Approximate the expected retur

$$\mathbb{E}_{x \sim p(x)}\left[f(x)\right] = \int f(x)p(x)\,dx$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} f(x_i) \quad x_i \overset{i.i.d}{\sim} p(x)$$

$$\boxed{\approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)\,w(x_i), \quad x_i \sim q(x)}$$





Sampling from a "wrong" $q(x)$ can make things go horribly wrong!

Choose $q(x)$ to be large when $|f(x)|p(x)$ is large.

Caution: • Choosing $q(x)$ in high-dimensions is very difficulty.

• It is difficult to assess how good or bad our estimator.