Importance Sampling

Suppose your goal is to compute

$$E[f(x)] = \int_{-\infty}^{\infty} f(x)\, p(x)\, dx$$

Chebychef inequality (see Review):

$$E[f(x)] \approx \frac{1}{N_e} \sum_{j=1}^{N_e} f(x_j) \quad , \; x_j \sim p(x)$$

This works in $N_e \to \infty$ limit, i.e.

$$\frac{1}{N_e} \Sigma f(x_j) \to E[f(x)] \text{ as } N_e \to \infty$$

In practice: N_e might be very large.

Example:
$$f(x) = \begin{cases} 0 & \text{if } x < 4 \\ 1 & \text{if } x \geq 4 \end{cases} \qquad X \sim N(0,1)$$

$$E[f(x)] = \int_{-\infty}^{\infty} f(x)\, p(x)\, dx = \int_{-\infty}^{\infty} f(x) \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}x^2)$$

$$= \int_{4}^{\infty} 1 \cdot \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}x^2)\, dx \approx 3.17 \cdot 10^{-5}$$

Monte Carlo:

$$E[f(x)] \approx \frac{1}{N_e} \Sigma f(x_j) \qquad x_j \sim N(0,1)$$

Since "most" x_j are in $(-2,2)$ (see, again, Chebychef) $f(x_j) = 0$ "most of the time".

↳ Unless N_c is large, we get

$$E[f(x)] = \frac{1}{N_c} \sum f(x_i) = \frac{1}{N} \sum 0 = 0.$$

Why? Samples $x_i \sim p(x)$ fall into region where $f(x)$ is small (zero) $\quad (-2, 2)$

Idea: focus samples into region where $f(x)$ is large.

Define:

$$E_p[f(x)] = \int f(x)\, p(x)\, dx$$

↑ which distribution we average over

Then:

$$E_p[f(x)] = \int f(x)\, p(x)\, dx$$

$$= \int f(x) \frac{p(x)}{q(x)}\, q(x)\, dx$$

↑ assuming that $q(x) = 0$ only when $p(x) = 0$, i.e. support of q includes support of p.

$$= E_q\left[f(x) \frac{p(x)}{q(x)}\right]$$

Monte Carlo: (Importance Sampling)

$$E_p[f(x)] = E_q\left[f(x)\frac{p(x)}{q(x)}\right] = \frac{1}{N_e}\sum_{j=1}^{N_e} f(x_j)\frac{p(x_j)}{q(x_j)}$$

$x_j \sim q(x_j)$

Terminology: $p(x)$ → target distribution

$q(x)$ → proposal distribution

$w(x) = \frac{p(x)}{q(x)}$ → weight (function)

Algorithm: · Draw N_e samples from proposal q

· Compute

$$E[f(x)] \approx \frac{1}{N_e}\sum_{j=1}^{N_e} f(x_j)\, w(x_j)$$

This "biases" samples to fall into region where f is large (important).

Back to example:

$$f(x) = \begin{cases} 0 & \text{if } x < 4 \\ 1 & \text{if } x \geq 4 \end{cases} \qquad p(x) = N(0,1)$$

$$q(x) = N(2,1)$$

↳ This helps and gives better estimates for "small N_e".

↳ You explore this in HW.

Question: How to pick a "good" proposal such that

$$\frac{1}{N_e}\sum_{j=1}^{N_e} f(x_j)\frac{p(x_j)}{q(x_j)} \qquad x_j \sim q(x)$$

is a "good" approximation of $E_p[f(x)]$ for a "reasonable" N_e?

Note: What "good" and "reasonable" means depends on the problem.

Recall the Review:

$$E_p[g(x)] = \frac{1}{N_e}\sum_{j=1}^{N_e} g(x_j) \qquad x_j \sim p$$

↳ error in estimate $\propto \dfrac{\sigma(g(x))}{\sqrt{N_e}}$ (see review)

$\sigma(g(x))$ is std. dev. of r.v. $g(x)$.

From Chebyshev we get that

$\eta = \frac{1}{N_e}\sum \eta_i$ η_i i.i.d.

$P(|\eta - E(\eta)| > \kappa \frac{\sigma}{\sqrt{N_e}}) \leq \frac{1}{\kappa^2}$

Applying this to importance sampling estimate:

↳ $\dfrac{\sigma\left(f(x)\frac{p(x)}{q(x)}\right)}{\sqrt{N_e}}$ is "error" in importance sampling.

⇒ we want that Std. dev. of

$$f(x)\frac{p(x)}{q(x)} \quad \text{is "small".}$$

⇒ smallest it can be is zero, what happens when

$$f(x)\frac{p(x)}{q(x)} = \text{Const.}$$

⇒ a good proposal distribution is such that

$$q(x) \propto p(x) f(x).$$

"Problem": $q(x) \geq 0$.

<u>Definition:</u> The optimal proposal distribution is

$$q^*(x) = \frac{|f(x)|\, p(x)}{E_p[|f(x)|]}$$

and minimizes $\sigma\left(f(x)\frac{p(x)}{q(x)}\right)$.

(1.) Check that q^* is a distribution.

$q^*(x) \geq 0$ ✓

$$\int_{-\infty}^{\infty} q^*(x)\,dx = \int \frac{|f(x)|\,p(x)}{E_p[|f(x)|]}\,dx = \frac{1}{E_p[|f(x)|]} E_p[|f(x)|]$$

$= 1$ ✓

(2) $\sigma\left(f(x)\frac{p(x)}{q^*(x)}\right) \leq \sigma\left(f(x)\frac{p(x)}{q(x)}\right)$

Recall: $\sigma_{IF}^2 \; Var(x)$

$Var(x) = E[x^2] - E[x]^2$

Proof:

$$\sigma^2\left(\int p/q^*\right) = Var\left(\int p/q^*\right)$$

$$= E_{q^*}\left[\left(\int p/q^*\right)^2\right] - \underbrace{E_{q^*}\left[\int p/q^*\right]^2}_{:=\mu = E[\int(x)]}$$

$$\Rightarrow \sigma_{q^*}^2\left(\int p/q^*\right) + \mu^2 = E_{q^*}\left[\left(\int p/q^*\right)^2\right]$$

$$= \int \int(x)^2 \frac{p(x)^2}{q^*(x)^2} \cancel{q^*(x)}\, dx = \underbrace{\int \frac{\int(x)^2 p(x)^2}{|\int(x)|^* \, p(x)}}\, dx \; E_p[|\int(x)|]$$

$$\int \frac{|\int(x)|^2}{|\int(x)|} p(x)\, dx = E_p[|\int(x)|]$$

$$= E_p[|\int(x)|]^2 = E_q\left[|\int(x)| \frac{p(x)}{q(x)}\right]^2$$

Some other proposal

$Var(x) = E[x^2] - E(x)^2$

$\Rightarrow E[x]^2 = E[x^2] - Var(x)$

$$\leq E_q\left[|\int(x)|^2 \frac{p^2(x)}{q^2(x)}\right] - Var\left(|\int(x)| \frac{p(x)}{q(x)}\right]$$

$$\leq E_q\left[\underbrace{|\int(x)|^2}_{=\int(x)^2} \frac{p^2(x)}{q^2(x)}\right] = \sigma_q^2\left(|\int| \frac{p}{q}\right) + \mu^2$$

$E[x^2] = \sigma^2 + E[x]^2$

$$\sigma_{q^*}^2\left(\int p/q^*\right) \leq \sigma_q^2\left(\int p/q^*\right) \quad \checkmark$$

Problems:

(i) We don't know how to draw samples from $p(x)$.
How do we draw samples from $q^*(x)$ or $p(x)|f(x)|$?

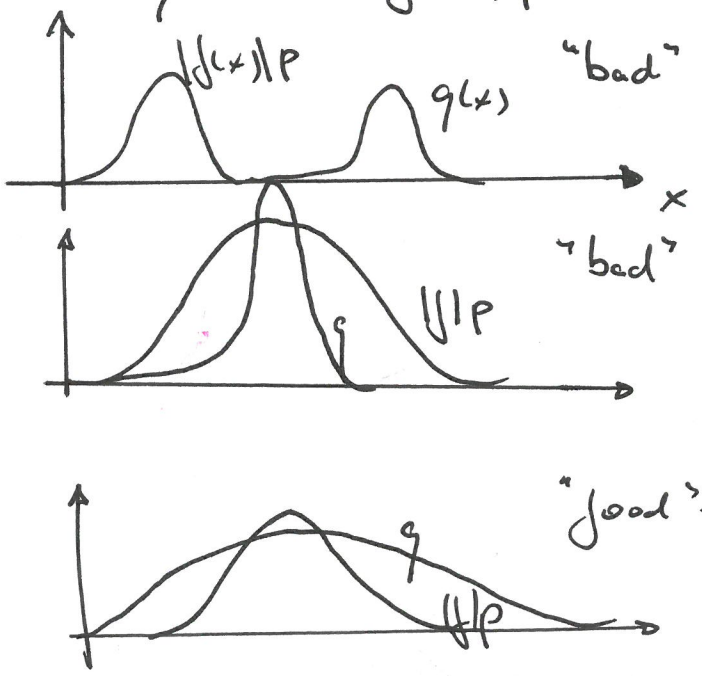
(ii) We need to know or compute $\mathbb{E}_p[|f(x)|]$.
This seems difficult.

Lesson: - optimal proposal $q^* = \frac{|f(x)|\,p(x)}{\mathbb{E}_p[|f(x)|]}$ usually hard to compute w.r.t.

- it is useful tool to figure out what a "good" proposal may look like.

(a) $q(x)$ should be large when $|f(x)|\,p(x)$ is large.
(b) avoid small q when $|f(x)|\,p(x)$ is large.

In cartoons:



Another difficulty of importance sampling

$$E_p[f(x)] = E_q\left[f(x)\frac{p(x)}{q(x)}\right] \approx \frac{1}{N_e}\sum_{j=1}^{N_e} f(x_j)\frac{p(x_j)}{q(x_j)} \qquad x_j \sim q(x).$$

weights: $w(x_j) = \dfrac{p(x_j)}{q(x_j)}$

→ we need to know $p(x_j)$.

→ we often don't.

In DA:

$p_0(x)$ → prior

$y = Hx + \eta$, $\eta \sim N(0, R)$

$$p(x|y) = \frac{p_0(x)\,p_e(y|x)}{p_y(y)}$$

We know $p(x|y)$ up to a multiplicative constant, but the constant is hard to compute.

$$p_y(y) = \int p_0(x)\,p_e(y|x)\,dx$$

another integral! $\left(\int p(x|y)\,dx = 1\right)$

Help: Self-normalized importance sampling.

$w(x_j) \propto \dfrac{p(x_j)}{q(x_j)}$ $\quad j = 1, \dots, N_e$

Note: Const. cannot depend on x_j!

Define: $\hat{w}(x_j) = w(x_j) \Big/ \sum_{j=1}^{N_e} w(x_j)$

Is independent of proportionality constant.

$$w(x_j) = C \, \hat{p}(x_j)/\hat{q}(x_j) \quad \text{where} \quad \hat{p}(x_j) \propto p(x_j), \quad \hat{q}(x_j) \propto q(x_j)$$

$$\hat{w}(x_j) = \frac{C \, \hat{p}(x_j)/\hat{q}(x_j)}{\sum C \frac{\hat{p}(x_j)}{\hat{q}(x_j)}} = \frac{\hat{p}(x_j)/\hat{q}(x_j)}{\sum_{j=1}^{N_e} \hat{p}(x_j)/\hat{q}(x_j)}$$

↝ For self-normalized importance sampling, we only need to know $q(x)$ and $p(x)$ up to a multiplicative const.

↝ This makes a feasible algorithm.

$$\boxed{E_p(f(x)) \approx \sum_{j=1}^{N_e} f(x_j) \, \hat{w}(x_j)}$$

Yet another difficulty of importance sampling.

Suppose we are interested in $p(x|y)$, and the r.v. $x|y$.

We want to compute, at least, $E[x|y]$ and $\mathrm{cov}(x|y)$

↳ $\mu = E[x] = \int x \, p(x|y) \, dx$

$$\sigma^2 = E((x-\mu)^2) = \int (x-\mu)^2 \, p(x|y) \, dx$$

↝ What is a good "general purpose" proposal to compute several expected values of r.v. $x|y$?

Spoiler alert: · there are no good general purpose proposals.
· some are "ok".

↝ To find out which ones are ok, we need to find a way to assess how good a proposal is.

Consider:

$$E_p[f(x)] \approx \sum_{j=1}^{N_e} f(x_j)\, w_j$$

w_j are "normalized" weights i.e. $\sum_{j=1}^{N_e} w_j = 1$

↝ weights do not depend on f.

↝ we can assess how good q is by looking at the weights.

Worst case: $w_j = 0$ for $j \neq k$

$w_k = 1$

↝ one weight is one, all others are zero

↝ one sample hogs all the probability, but there is no guarantee it's a good one.

↝ estimate:

$$E_p[f(x)] = \sum_{j=1}^{N_e} f(x_j)\, w(x_j) = f(x_k)$$

↝ terrible estimate based on only one sample!

Example:

$$\mu = E_p[x] = \sum x_j \omega_j = x_k$$

$$\text{Var}(x) = E_p[(x-\mu)^2] = \sum (x_j - \mu)^2 \omega_j = (x_k - \mu) = x_k - x_k = 0$$

but var. might not be zero.

How can we capture this?

Define "effective sample size":

$$N_{eff} = N_e / g$$

Idea: IS estimate with N_e samples are "as good" as estimate based on samples from target $p(x)$ with N_{eff} samples

$$\mu_p = \sum_{j=1}^{N_{eff}} f(x_j), \quad x_j \sim p$$

$$\mu_q = \sum_{j=1}^{N_e} f(x_j)\, \omega_j, \quad x_j \sim q$$

$$\therefore \mu_p \simeq \mu_q$$

We define:

$$g = \frac{E[\omega^2]}{E[\omega]^2}.$$

This is a heuristic quantity. There are other ways to define this. We don't yet know what is most suitable way to describe all this.

ρ captures well the worst and best cases.

best case : $q(x) \propto p(x)$

$\hookrightarrow \omega \propto \frac{q(x)}{p(x)} = \text{const.}$

Then $\rho = \frac{E(\omega^2)}{E(\omega)^2} = \frac{\text{Var}(\omega) + E(\omega)^2}{E(\omega)^2} = 1 + \frac{\text{var}(\omega)}{E(\omega)^2}$

$\uparrow$

$\text{Var}(x) = E(x^2) - E(x)^2$

$= 1 \quad \because \text{Var}(\omega) = 0$ and $\omega = \text{Const.}$

$\Rightarrow N_{eff} = N_e/\rho = N_e.$ ✓

Worst Case:

$\omega_1 = 0 \quad , \omega_j = 0 \; \forall j \neq 1$

Then $\rho = \frac{E(\omega^2)}{E(\omega)^2} \simeq \frac{\frac{1}{N_e}\sum \omega_j^2}{\left(\frac{1}{N_e}\sum \omega_j\right)^2} = N_e \sum \omega_j^2 = N_e$

$N_{eff} = \frac{N_e}{\rho} = 1$ ✓

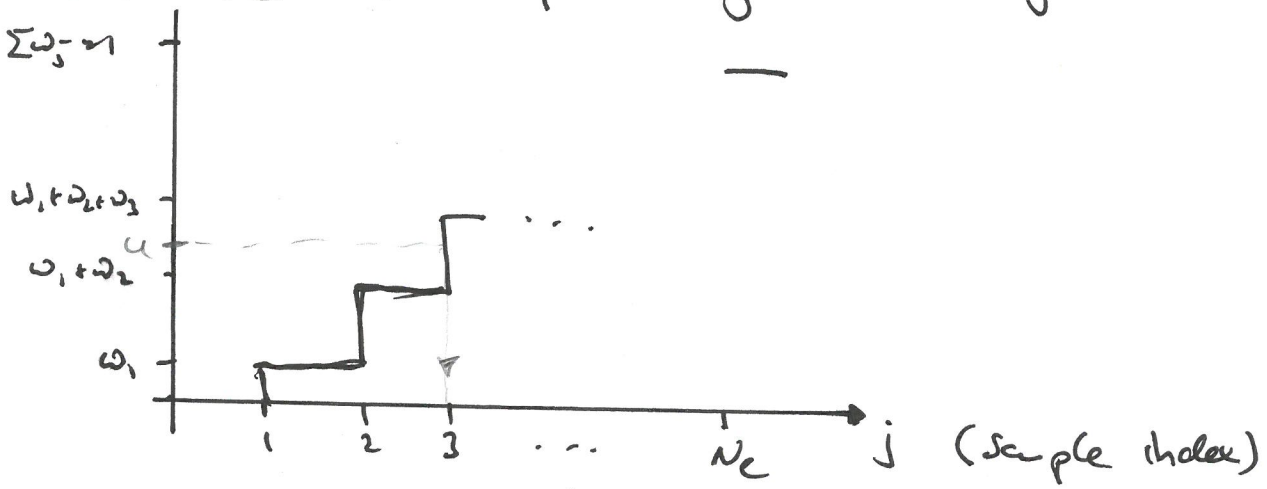
Resampling:

Idea: take a weighted ensemble $\{x_j, w_j\}$ $j = 1 \dots N_e$ and transform it into an ensemble $\{\tilde{x}_j\}$ with "equal weights" or "no weights":

$$\sum_{j=1}^{N_e} x_j w_j = \frac{1}{N_e} \sum_{j=1}^{N_e} \tilde{x}_j$$

One way of doing this is "resampling with replacement":

- Given a weighted ensemble $\{x_j, w_j\}$
- Draw the PDF (probability distribution function)



- Draw uniform number between 0 and 1
- If $\sum_{j=1}^{n-1} w_j < u < \sum_{j=1}^{n} w_j$: pick particle with index n.
- Do this N_e times.

You end up with an ensemble such that samples with large weights are repeated more often than samples with small weights.

And you have:

$$E_p[f(x)] \approx \frac{1}{N_{eff}} \sum f(x_j) \qquad x_j \sim p$$

$$\approx \sum f(x_j)\, \hat{w}_j \qquad x_j \sim q$$

$$\approx \frac{1}{N_e} \sum f(\tilde{x}_j), \qquad \tilde{x}_j \sim q$$

resampled ensemble.

⇒ You can use resampled ensemble in the same way as if you had an ensemble obtained by directly drawing from target distribution p.

⇒ Resampled ensemble can be used just like a "direct" ensemble for computing expected values

⇒ This is how importance sampling is most often used

⇒ q and N_{eff} assess how good importance sampling works in this sense (i.e. independently of $f(x)$).

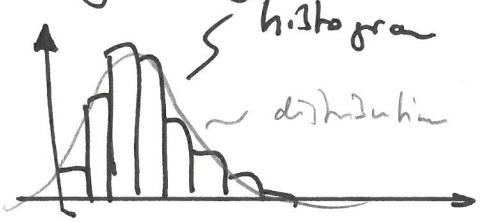
Histograms:

- Given an ensemble $x_j \sim p(x_j)$ $j = 1 \dots N_e$, e.g., obtained by importance sampling + resampling

- put samples x_j into "bins":

$$x_j \in I_{jk} \Rightarrow x_j \text{ ends up in bin } k.$$

- Draw height of bins (normalized so that integral = 1)



⇒ If you have a "good" ensemble, then histogram "looks like" posterior distribution of x.

⇒ You can think of importance sampling as a technique to draw samples whose histogram looks like (converges, as $N_e \to \infty$) the target distribution.