

# Importance sampling and Monte Carlo estimators

Ten problems on Monte Carlo estimation: the plain estimator and its  $1/\sqrt{n}$  error, the importance-sampling identity and when it fails, the variance of the weighted estimator, the zero-variance proposal, a Gaussian tail probability where a shifted proposal cuts the variance by a factor of 218, the self-normalised estimator for unnormalised targets and its finite-sample bias, the exponential growth of weight variance with dimension, the effective sample size, control variates and antithetic variates, with worked solutions and the mistakes that swap the weights, call the self-normalised estimator unbiased, use a proposal narrower than the target, pair antithetic samples for a symmetric integrand, or quote the plain Monte Carlo variance for a weighted estimate.

## BEFORE YOU START

Every expectation that cannot be done in closed form is done by averaging samples, and the whole craft is in the variance: an estimate whose error shrinks like  $1/\sqrt{n}$  is only as good as the constant in front. Importance sampling changes that constant by drawing from a distribution of one's own choosing and correcting with weights, and it can make it smaller by orders of magnitude or, with the wrong choice, infinite. These ten problems derive the plain estimator and its error, prove the importance-sampling identity and find the variance it carries, identify the proposal that drives the variance to zero, work a Gaussian tail probability where a shifted proposal is 218 times more efficient, handle targets known only up to a constant with the self-normalised estimator and quantify its bias, show why weights degenerate exponentially with dimension and how the effective sample size measures it, and end with the two cheapest variance reductions, control variates and antithetic pairs. The five mistakes at the end are the ones that return a number with no warning: weights with the numerator and denominator swapped, a self-normalised estimate called unbiased, a proposal with lighter tails than the target, antithetic pairs on a symmetric integrand, and the plain Monte Carlo variance quoted for a weighted estimate.

- The target is a distribution  $p$  and the quantity wanted is  $\mu = \mathbb{E}_p[f(X)]$ ; sums are written for a finite set of outcomes and integrals for densities, and every argument below works for both. A proposal is another distribution  $q$  with  $q(x) > 0$  wherever  $f(x)p(x) \neq 0$ . The importance weight is  $w(x) = p(x)/q(x)$ , and  $\mathbb{E}_q$ ,  $\text{Var}_q$  are expectation and variance when  $X \sim q$ .

- With  $x_1, \dots, x_n$  independent draws, the plain Monte Carlo estimator is  $\hat{\mu}_n = \frac{1}{n} \sum_i f(x_i)$  for  $x_i \sim p$ , and the importance-sampling (IS) estimator is  $\hat{\mu}_{\text{IS}} = \frac{1}{n} \sum_i w(x_i) f(x_i)$  for  $x_i \sim q$ . The variance page's rules for sums of independent variables are used throughout;  $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$  and  $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ , with correlation  $\rho = \text{Cov}(X, Y) / \sqrt{\text{Var}(X) \text{Var}(Y)}$ .
- An unnormalised target is  $\tilde{p}(x) = Zp(x)$  with  $Z = \sum_x \tilde{p}(x)$  unknown; its weights are  $\tilde{w} = \tilde{p}/q = Zw$ . The self-normalised estimator is  $\hat{\mu}_{\text{SN}} = \sum_i \tilde{w}(x_i) f(x_i) / \sum_i \tilde{w}(x_i)$ . The effective sample size of weights  $w_1, \dots, w_n$  is  $\text{ESS} = (\sum_i w_i)^2 / \sum_i w_i^2$ .
- $\varphi(x) = e^{-x^2/2} / \sqrt{2\pi}$  is the standard normal density and  $\Phi$  its distribution function, so  $\mathbb{P}(X > a) = 1 - \Phi(a)$  for  $X \sim \mathcal{N}(0, 1)$ ;  $1 - \Phi(3) \approx 1.350 \times 10^{-3}$  and  $1 - \Phi(6) \approx 9.866 \times 10^{-10}$ .  $\mathcal{N}(m, \sigma^2)$  has density  $\varphi((x - m)/\sigma)/\sigma$ , and  $\mathcal{N}(0, \sigma^2 I_d)$  is the  $d$ -dimensional Gaussian with independent coordinates (the Gaussian page).  $\mathbf{1}\{A\}$  is 1 when  $A$  holds and 0 otherwise.  $U \sim \mathcal{U}(0, 1)$  is uniform on  $[0, 1]$ , with  $\mathbb{E}[U] = \frac{1}{2}$  and  $\text{Var}(U) = \frac{1}{12}$ .

Builds on: [Variance, covariance and correlation](#), [The multivariate Gaussian: gradients and identities](#)

## Problems

**1.** Show that  $\hat{\mu}_n$  is unbiased with  $\text{Var}(\hat{\mu}_n) = \text{Var}_p(f)/n$ , so its standard error is  $\sqrt{\text{Var}_p(f)/n}$ . How many more samples are needed to halve the standard error? difficulty ·

**2.** Prove the importance-sampling identity  $\mathbb{E}_p[f(X)] = \mathbb{E}_q[w(X)f(X)]$  with  $w = p/q$ , so that  $\hat{\mu}_{\text{IS}}$  is unbiased. Where does the proof use  $q(x) > 0$  wherever  $f(x)p(x) \neq 0$ , and what happens if that fails? difficulty ·

**3.** Show that  $\text{Var}(\hat{\mu}_{\text{IS}}) = \frac{1}{n} (\mathbb{E}_q[w^2 f^2] - \mu^2)$  and that  $\mathbb{E}_q[w^2 f^2] = \mathbb{E}_p[w f^2]$ . When is the variance infinite even though the estimator is unbiased? difficulty ··

**4.** Show that  $\mathbb{E}_q[w^2 f^2] \geq (\mathbb{E}_p|f|)^2$  for every proposal  $q$ , with equality for  $q^*(x) = |f(x)|p(x)/\mathbb{E}_p|f|$ . Conclude that for  $f \geq 0$  the proposal  $q^* = fp/\mu$  gives an estimator with zero variance, and explain why it cannot be used directly. difficulty ...

**5.** Let  $X \sim \mathcal{N}(0, 1)$  and  $\mu = \mathbb{P}(X > 3) = 1 - \Phi(3)$ . (a) Compute the per-sample variance  $\text{Var}_p(\mathbf{1}\{X > 3\})$  of plain Monte Carlo and the relative standard error  $\sqrt{\text{Var}_p}/n/\mu$  for  $n = 1$ . (b) With the proposal  $q = \mathcal{N}(3, 1)$ , show that  $w(x) = e^{9/2-3x}$  and that  $\mathbb{E}_q[w^2 f^2] = e^9(1 - \Phi(6))$ . (c) Compute the IS variance per sample and the ratio of the two variances. difficulty ...

**6.** With an unnormalised target  $\tilde{p} = Zp$  and weights  $\tilde{w} = \tilde{p}/q$ , show that  $\mathbb{E}_q[\tilde{w}] = Z$  and  $\mathbb{E}_q[\tilde{w}f] = Z\mu$ , so the self-normalised estimator  $\hat{\mu}_{\text{SN}} = \sum_i \tilde{w}_i f(x_i) / \sum_i \tilde{w}_i$  converges to  $\mu$ . Show that for finite  $n$  it is biased in general, by computing  $\mathbb{E}[\hat{\mu}_{\text{SN}}]$  for  $n = 1$ . difficulty ..

**7.** Let  $p = \mathcal{N}(0, I_d)$  and  $q = \mathcal{N}(0, \sigma^2 I_d)$  with  $\sigma^2 > \frac{1}{2}$ . Show that in one dimension  $\mathbb{E}_q[w^2] = \mathbb{E}_p[w] = \frac{\sigma^2}{\sqrt{2\sigma^2 - 1}}$ , hence in  $d$  dimensions  $\mathbb{E}_q[w^2] = \left(\frac{\sigma^2}{\sqrt{2\sigma^2 - 1}}\right)^d$ , and evaluate for  $\sigma^2 = 2$  and  $d = 10, 50$ . What happens for  $\sigma^2 \leq \frac{1}{2}$ ? difficulty ...

**8.** For positive weights  $w_1, \dots, w_n$ , show that  $1 \leq \text{ESS} = \frac{(\sum_i w_i)^2}{\sum_i w_i^2} \leq n$ , with  $\text{ESS} = n$  exactly when all weights are equal and  $\text{ESS} \rightarrow 1$  when one weight dominates. Show that with normalised weights  $W_i = w_i / \sum_j w_j$ ,  $\text{ESS} = 1 / \sum_i W_i^2$ , and that for large  $n$ ,  $\text{ESS}/n \approx 1/\mathbb{E}_q[w^2]$  when  $w = p/q$  with  $p$  and  $q$  normalised. difficulty ..

**9.** Let  $g$  be a function with known mean  $\mathbb{E}_p[g] = \gamma$ . Show that the control-variate estimator  $\hat{\mu}_c = \frac{1}{n} \sum_i (f(x_i) - c(g(x_i) - \gamma))$  is unbiased for every  $c$ , that its variance is  $\frac{1}{n}(\text{Var}(f) - 2c \text{Cov}(f, g) + c^2 \text{Var}(g))$ , minimised at  $c^* = \text{Cov}(f, g) / \text{Var}(g)$  with value  $\frac{1}{n} \text{Var}(f)(1 - \rho^2)$ . Evaluate for  $f(U) = e^U$ ,  $g(U) = U$  with  $U \sim \mathcal{U}(0, 1)$ . difficulty ..

**10.** For  $U \sim \mathcal{U}(0, 1)$ , the antithetic pair estimator uses  $\frac{1}{2}(f(U) + f(1 - U))$ . Show that it is unbiased and that its variance is  $\frac{1}{2} \text{Var}(f(U))(1 + \rho_a)$  with  $\rho_a = \text{Corr}(f(U), f(1 - U))$ , so it beats two independent samples exactly when  $\rho_a < 0$ . Evaluate for  $f(u) = e^u$ . difficulty ..

## Answers

1.  $\mathbb{E}[\hat{\mu}_n] = \mu$ ,  $\text{Var}(\hat{\mu}_n) = \text{Var}_p(f)/n$ , standard error  $\sqrt{\text{Var}_p(f)/n}$ ; halving it takes four times the samples
2.  $\mathbb{E}_p[f] = \mathbb{E}_q[wf]$  and  $\hat{\mu}_{\text{IS}}$  is unbiased, provided  $q > 0$  wherever  $fp \neq 0$ ; if the condition fails, the estimator is biased by the missing mass  $\sum_{x:q(x)=0} p(x)f(x)$
3.  $\text{Var}(\hat{\mu}_{\text{IS}}) = \frac{1}{n}(\mathbb{E}_q[w^2 f^2] - \mu^2) = \frac{1}{n}(\mathbb{E}_p[wf^2] - \mu^2)$ ; it is infinite when  $\mathbb{E}_p[wf^2] = \mathbb{E}_p[f^2 p/q]$  diverges, which happens when  $q$  has lighter tails than  $f^2 p$
4.  $\mathbb{E}_q[w^2 f^2] \geq (\mathbb{E}_p[f])^2$  with equality at  $q^* \propto |f|p$ ; for  $f \geq 0$ ,  $q^* = fp/\mu$  has zero variance; it cannot be used because normalising it requires  $\mu$ , the unknown
5. (a)  $\text{Var}_p = \mu(1 - \mu) \approx 1.348 \times 10^{-3}$ , relative error  $\approx 27.2$  per sample; (b)  $w(x) = e^{9/2-3x}$  and  $\mathbb{E}_q[w^2 f^2] = e^9(1 - \Phi(6)) \approx 7.994 \times 10^{-6}$ ; (c) IS variance  $\approx 6.17 \times 10^{-6}$  per sample, 218 times smaller, relative error  $\approx 1.84$  per sample
6.  $\mathbb{E}_q[\tilde{w}] = Z$ ,  $\mathbb{E}_q[\tilde{w}f] = Z\mu$ , so  $\hat{\mu}_{\text{SN}} \rightarrow \mu$ ; but  $\mathbb{E}[\hat{\mu}_{\text{SN}}] \neq \mu$  for finite  $n$ , and for  $n = 1$  it equals  $\mathbb{E}_q[f]$
7.  $\mathbb{E}_q[w^2] = \left( \frac{\sigma^2}{\sqrt{2\sigma^2 - 1}} \right)^d$  for  $\sigma^2 > \frac{1}{2}$ ; for  $\sigma^2 = 2$  it is  $\approx 4.2$  at  $d = 10$  and  $\approx 1.3 \times 10^3$  at  $d = 50$ ; for  $\sigma^2 \leq \frac{1}{2}$  it is infinite
8.  $1 \leq \text{ESS} \leq n$ , equal to  $n$  for equal weights and near 1 for one dominant weight;  $\text{ESS} = 1 / \sum_i W_i^2$ ;  $\text{ESS} / n \approx 1 / \mathbb{E}_q[w^2]$
9.  $\hat{\mu}_c$  is unbiased for every  $c$ ;  $\text{Var}(\hat{\mu}_c) = \frac{1}{n}(\text{Var}(f) - 2c \text{Cov}(f, g) + c^2 \text{Var}(g))$ , minimised at  $c^* = \text{Cov}(f, g) / \text{Var}(g)$  with value  $\frac{1}{n} \text{Var}(f)(1 - \rho^2)$ ; for  $e^U$  with control  $U$ :  $c^* \approx 1.690$ ,  $\rho^2 \approx 0.984$ , variance  $0.2420 \rightarrow 0.0039$
10. The antithetic pair is unbiased with variance  $\frac{1}{2} \text{Var}(f(U))(1 + \rho_a)$ , better than two independent samples exactly when  $\rho_a < 0$ ; for  $e^u$ ,  $\rho_a \approx -0.968$  and the variance per pair is  $\approx 0.0039$  against  $0.1210$

## Worked solutions

### Problem 1

Show that  $\hat{\mu}_n$  is unbiased with  $\text{Var}(\hat{\mu}_n) = \text{Var}_p(f)/n$ , so its standard error is  $\sqrt{\text{Var}_p(f)/n}$ . How many more samples are needed to halve the standard error?

$$\mathbb{E}[\hat{\mu}_n] = \frac{1}{n} \sum_i \mathbb{E}_p[f(x_i)] = \frac{1}{n} \cdot n\mu = \mu.$$

Linearity; every  $x_i$  is a draw from  $p$ .

$$\text{Var}(\hat{\mu}_n) = \frac{1}{n^2} \sum_i \text{Var}_p(f(x_i)) = \frac{\text{Var}_p(f)}{n}.$$

Independence makes the variance of the sum the sum of the variances;  
 $\text{Var}(aX) = a^2 \text{Var}(X)$  with  $a = 1/n$ .

The standard error  $\sqrt{\text{Var}_p(f)/n}$  is halved when  $n$  is multiplied by 4.

$$\sqrt{1/(4n)} = \frac{1}{2} \sqrt{1/n}.$$

$$\mathbb{E}[\hat{\mu}_n] = \mu, \text{Var}(\hat{\mu}_n) = \text{Var}_p(f)/n,$$

**standard error**  $\sqrt{\text{Var}_p(f)/n}$ ; **halving it takes four times the samples**

Monte Carlo converges at the rate  $1/\sqrt{n}$  in every dimension, which is its whole appeal against quadrature and its whole weakness against anything with a closed form: each extra digit of accuracy costs a hundred times the samples. The constant  $\text{Var}_p(f)$  is the only thing that can be changed, and Problems 2 to 10 are all ways of changing it. For a probability  $\mu = \mathbb{E}_p[\mathbf{1}\{A\}]$  the variance is  $\mu(1 - \mu)$  and the relative error  $\sqrt{(1 - \mu)/(n\mu)}$ , which for a rare event is enormous; Problem 5 makes this concrete.

### Problem 2

Prove the importance-sampling identity  $\mathbb{E}_p[f(X)] = \mathbb{E}_q[w(X)f(X)]$  with  $w = p/q$ , so that  $\hat{\mu}_{\text{IS}}$  is unbiased. Where does the proof use  $q(x) > 0$  wherever  $f(x)p(x) \neq 0$ , and what happens if that fails?

$$\mathbb{E}_p[f] = \sum_x p(x)f(x) = \sum_{x:q(x)>0} q(x) \frac{p(x)}{q(x)} f(x).$$

Multiply and divide by  $q(x)$ , which is allowed only where  $q(x) > 0$ ; the terms with  $q(x) = 0$  are dropped from the sum.

|                                                                                                                                                                                                                                                                                  |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>The dropped terms have <math>f(x)p(x) = 0</math> by the support condition, so they contributed nothing.</p>                                                                                                                                                                   | <p>The condition says <math>q = 0</math> only where the summand <math>p(x)f(x)</math> is already zero.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| $\sum_{x:q(x)>0} q(x) w(x) f(x) = \mathbb{E}_q[w(X)f(X)].$                                                                                                                                                                                                                       | <p>A <math>q</math>-weighted sum over the support of <math>q</math> is an expectation under <math>q</math>.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| $\mathbb{E}[\hat{\mu}_{\text{IS}}] = \frac{1}{n} \sum_i \mathbb{E}_q[w(x_i)f(x_i)] = \mu.$                                                                                                                                                                                       | <p>Linearity with <math>x_i \sim q</math>, then step 3.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| <p><math>\mathbb{E}_p[f] = \mathbb{E}_q[wf]</math> <b>and <math>\hat{\mu}_{\text{IS}}</math> is unbiased, provided <math>q &gt; 0</math> wherever <math>fp \neq 0</math>; if the condition fails, the estimator is biased by the missing mass</b></p> $\sum_{x:q(x)=0} p(x)f(x)$ | <p>The identity is a change of measure: samples from <math>q</math> are reweighted to look like samples from <math>p</math>, and the weight is large where <math>q</math> undersamples <math>p</math> and small where it oversamples. The condition is exactly what the ELBO page's <math>w = p(x, z)/q(z)</math> and the policy-gradient page's behaviour policy need. It fails silently: a proposal that never visits a region where <math>p</math> has mass returns an estimate that is biased by everything in that region, with no large weight to signal it, because the samples that would have carried the large weight are never drawn. Unbiased is the easy part; Problem 3 is the hard part.</p> |

### Problem 3

Show that  $\text{Var}(\hat{\mu}_{\text{IS}}) = \frac{1}{n} (\mathbb{E}_q[w^2 f^2] - \mu^2)$  and that  $\mathbb{E}_q[w^2 f^2] = \mathbb{E}_p[w f^2]$ . When is the variance infinite even though the estimator is unbiased?

|                                                                                                                                 |                                                                                                                                                                              |
|---------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| $\text{Var}(\hat{\mu}_{\text{IS}}) = \frac{1}{n} \text{Var}_q(wf) = \frac{1}{n} (\mathbb{E}_q[(wf)^2] - (\mathbb{E}_q[wf])^2).$ | <p>Problem 1's argument with <math>wf</math> in place of <math>f</math> and <math>q</math> in place of <math>p</math>; the variance as second moment minus squared mean.</p> |
| $\mathbb{E}_q[wf] = \mu, \text{ so } \text{Var}(\hat{\mu}_{\text{IS}}) = \frac{1}{n} (\mathbb{E}_q[w^2 f^2] - \mu^2).$          | <p>Problem 2.</p>                                                                                                                                                            |

$$\mathbb{E}_q[w^2 f^2] = \sum_x q(x) \frac{p(x)^2}{q(x)^2} f(x)^2 = \sum_x p(x) \frac{p(x)}{q(x)} f(x)^2 = \mathbb{E}_p[w f^2].$$

Cancel one  $q$  against one of the two in  $w^2$ ; what remains is a  $p$ -weighted sum of  $w f^2$ .

$$\text{Var}(\hat{\mu}_{\text{IS}}) = \frac{1}{n} (\mathbb{E}_q[w^2 f^2] - \mu^2) = \frac{1}{n} (\mathbb{E}_p[w f^2] - \mu^2); \text{ it is infinite when } \mathbb{E}_p[w f^2] = \mathbb{E}_p[f^2 p/q] \text{ diverges, which happens when } q \text{ has lighter tails than } f^2 p$$

The variance is controlled by  $\mathbb{E}_p[w f^2]$ : the weight  $p/q$  averaged under  $p$  and amplified by  $f^2$ . Where  $q$  is much smaller than  $p$  the weight is huge and the rare samples that land there dominate the average; if  $p/q$  grows fast enough in the tails the second moment is infinite and the estimator, though unbiased, has no central limit theorem, so the sample average lurches whenever a tail sample arrives and its empirical standard error is meaningless. Problem 7 computes  $\mathbb{E}_p[w]$  for Gaussians and finds exactly this: a proposal narrower than the target by a factor of  $\sqrt{2}$  in standard deviation already gives infinite variance (Mistake 3). The rule of thumb is that  $q$  must have tails at least as heavy as  $p$ 's, and Problem 4 says what the best  $q$  is.

#### Problem 4

Show that  $\mathbb{E}_q[w^2 f^2] \geq (\mathbb{E}_p[f])^2$  for every proposal  $q$ , with equality for  $q^*(x) = |f(x)p(x)|/\mathbb{E}_p[f]$ . Conclude that for  $f \geq 0$  the proposal  $q^* = fp/\mu$  gives an estimator with zero variance, and explain why it cannot be used directly.

$$\mathbb{E}_q[w^2 f^2] = \mathbb{E}_q[(w|f|)^2] \geq (\mathbb{E}_q[w|f|])^2.$$

$f^2 = |f|^2$ ; then Jensen's inequality for the convex function  $t \mapsto t^2$ , the Jensen page's Problem 5(a): a second moment is at least the squared mean.

$$\mathbb{E}_q[w|f|] = \mathbb{E}_p[f].$$

Problem 2's identity applied to  $|f|$ .

For  $q^* = |f|p/\mathbb{E}_p|f|$ :  $w^*|f| = \frac{p}{q^*}|f| = \mathbb{E}_p|f|$ , a constant, so equality holds in step 1.

$p/q^* = \mathbb{E}_p|f|/|f|$  wherever  $f \neq 0$ , and the  $|f|$  cancels; a constant has zero Jensen gap.  $q^*$  is a valid proposal: nonnegative, sums to one, and positive wherever  $fp \neq 0$ .

For  $f \geq 0$ :  $\mathbb{E}_p|f| = \mu$ ,  $q^* = fp/\mu$ , and every sample gives  $w^*(x)f(x) = \mu$  exactly, so  $\text{Var}(\hat{\mu}_{\text{IS}}) = 0$ .

Steps 2 and 3 with  $|f| = f$ ; the estimator is the constant  $\mu$  whatever is drawn.

$\mathbb{E}_q[w^2 f^2] \geq (\mathbb{E}_p|f|)^2$  **with equality at**  
 $q^* \propto |f|p$ ; **for  $f \geq 0$ ,  $q^* = fp/\mu$  has zero variance; it cannot be used because normalising it requires  $\mu$ , the unknown**

The best proposal is not the target but the target reshaped by the integrand: sample where  $|f|p$  is large, which for a tail probability means sampling in the tail (Problem 5) and for the ELBO means sampling near the posterior. The zero-variance proposal is a statement about direction rather than a recipe, since writing it down solves the problem, but every good proposal is an approximation to it, and the bound in step 1 gives the floor that any approximation is measured against. For  $f$  taking both signs the minimum variance is  $(\mathbb{E}_p|f|)^2 - \mu^2 > 0$ , so the floor is not zero. Mistake 4 on the Jensen page is the reason  $q = p$  is usually far from optimal, which Problem 5 quantifies.

## Problem 5

Let  $X \sim \mathcal{N}(0, 1)$  and  $\mu = \mathbb{P}(X > 3) = 1 - \Phi(3)$ . (a) Compute the per-sample variance  $\text{Var}_p(\mathbf{1}\{X > 3\})$  of plain Monte Carlo and the relative standard error  $\sqrt{\text{Var}_p/n}/\mu$  for  $n = 1$ .

- (b) With the proposal  $q = \mathcal{N}(3, 1)$ , show that  $w(x) = e^{9/2-3x}$  and that  $\mathbb{E}_q[w^2 f^2] = e^9(1 - \Phi(6))$ .  
(c) Compute the IS variance per sample and the ratio of the two variances.

(a)  $f = \mathbf{1}\{X > 3\}$  has  $f^2 = f$ , so

$\text{Var}_p(f) = \mu - \mu^2 = \mu(1 - \mu) \approx 1.348 \times 10^{-3}$ , and the relative standard error at  $n = 1$  is  $\sqrt{(1 - \mu)/\mu} \approx 27.2$ .

An indicator is Bernoulli with success probability  $\mu \approx 1.350 \times 10^{-3}$ ; divide the standard error by  $\mu$ .

$$(b) w(x) = \frac{\varphi(x)}{\varphi(x-3)} = \exp\left(-\frac{x^2}{2} + \frac{(x-3)^2}{2}\right) = \exp\left(\frac{-6x+9}{2}\right) = e^{9/2-3x}.$$

The two densities share the  $1/\sqrt{2\pi}$ ; expand  $(x-3)^2 = x^2 - 6x + 9$  and the  $x^2$  cancels.

$$\mathbb{E}_q[w^2 f^2] = \int_3^\infty \frac{\varphi(x)^2}{\varphi(x-3)} dx = \int_3^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-x^2 + \frac{(x-3)^2}{2}\right) dx.$$

Problem 3's  $\mathbb{E}_q[w^2 f^2] = \int q(p/q)^2 f^2 = \int p^2/q$  over the region where  $f = 1$ ; one factor of  $\sqrt{2\pi}$  survives.

$$-x^2 + \frac{x^2 - 6x + 9}{2} = -\frac{x^2}{2} - 3x + \frac{9}{2} = -\frac{(x+3)^2}{2} + 9.$$

Complete the square:  $-\frac{1}{2}(x^2 + 6x) = -\frac{1}{2}(x+3)^2 + \frac{9}{2}$ , plus the  $\frac{9}{2}$  already there.

$$\mathbb{E}_q[w^2 f^2] = e^9 \int_3^\infty \varphi(x+3) dx = e^9 \int_6^\infty \varphi(u) du = e^9(1 - \Phi(6)) \approx 8103.1 \times 9.866 \times 10^{-10} \approx 7.994 \times 10^{-6}.$$

Substitute  $u = x + 3$ ; the integral of  $\varphi$  from 6 is the upper tail.

(c)  $\text{Var}_q(wf) = 7.994 \times 10^{-6} - \mu^2 \approx 7.994 \times 10^{-6} - 1.822 \times 10^{-6} = 6.17 \times 10^{-6}$ , and the ratio is  $1.348 \times 10^{-3}/6.17 \times 10^{-6} \approx 218$ ; the relative standard error at  $n = 1$  is  $\sqrt{6.17 \times 10^{-6}}/\mu \approx 1.84$ .

Problem 3;  $\mu^2 \approx (1.350 \times 10^{-3})^2$ .

**(a)**  $\text{Var}_p = \mu(1 - \mu) \approx 1.348 \times 10^{-3}$ ,  
**relative error**  $\approx 27.2$  **per sample; (b)**  
 $w(x) = e^{9/2-3x}$  **and**  $\mathbb{E}_q[w^2 f^2] = e^9(1 - \Phi(6)) \approx 7.994 \times 10^{-6}$ ; **(c) IS variance**  
 $\approx 6.17 \times 10^{-6}$  **per sample, 218 times**  
**smaller, relative error**  $\approx 1.84$  **per**  
**sample**

Plain Monte Carlo sees a success once in 741 draws and needs about  $n \approx 74,000$  samples for a 10% relative error; the shifted proposal lands half its draws in the tail with weights no larger than  $e^{9/2-9} = e^{-4.5}$ , and reaches the same accuracy with  $n \approx 340$ . The gain comes from sampling where  $fp$  is large, Problem 4's prescription, and the shift to the threshold rather than beyond it is the standard choice because it keeps the weights bounded on the region that matters. The check evaluates both variances by quadrature and confirms the factor of 218.

## Problem 6

With an unnormalised target  $\tilde{p} = Zp$  and weights  $\tilde{w} = \tilde{p}/q$ , show that  $\mathbb{E}_q[\tilde{w}] = Z$  and  $\mathbb{E}_q[\tilde{w}f] = Z\mu$ , so the self-normalised estimator  $\hat{\mu}_{\text{SN}} = \sum_i \tilde{w}_i f(x_i) / \sum_i \tilde{w}_i$  converges to  $\mu$ . Show that for finite  $n$  it is biased in general, by computing  $\mathbb{E}[\hat{\mu}_{\text{SN}}]$  for  $n = 1$ .

$$\mathbb{E}_q[\tilde{w}] = \sum_x q(x) \frac{Zp(x)}{q(x)} = Z \sum_x p(x) = Z. \quad \text{The } q \text{ cancels and } p \text{ sums to one.}$$

$$\mathbb{E}_q[\tilde{w}f] = Z \mathbb{E}_q[wf] = Z\mu.$$

$$\tilde{w} = Zw \text{ and Problem 2.}$$

$\hat{\mu}_{\text{SN}} = \frac{\frac{1}{n} \sum_i \tilde{w}_i f(x_i)}{\frac{1}{n} \sum_i \tilde{w}_i}$ , and as  $n \rightarrow \infty$  the numerator tends to  $Z\mu$  and the denominator to  $Z$ , so the ratio tends to  $\mu$ .

Divide top and bottom by  $n$ ; each is a plain Monte Carlo average under  $q$  and converges to its mean (the law of large numbers); the ratio of the limits is the limit of the ratio because  $Z > 0$ .

For  $n = 1$ :  $\hat{\mu}_{\text{SN}} = \frac{\tilde{w}(x_1)f(x_1)}{\tilde{w}(x_1)} = f(x_1)$   
 with  $x_1 \sim q$ , so  $\mathbb{E}[\hat{\mu}_{\text{SN}}] = \mathbb{E}_q[f] \neq \mu$  in general.

The single weight cancels; the estimate is the integrand at a draw from the proposal, not from the target.

$\mathbb{E}_q[\tilde{w}] = Z$ ,  $\mathbb{E}_q[\tilde{w}f] = Z\mu$ , **so**  $\hat{\mu}_{\text{SN}} \rightarrow \mu$ ;  
**but**  $\mathbb{E}[\hat{\mu}_{\text{SN}}] \neq \mu$  **for finite**  $n$ , **and for**  
 $n = 1$  **it equals**  $\mathbb{E}_q[f]$

Self-normalisation is what makes importance sampling usable for posteriors, energy-based models and any  $\tilde{p}$  whose normaliser is unknown: the unknown  $Z$  cancels in the ratio. The price is bias of order  $1/n$ , because the ratio of two unbiased estimates is not unbiased (the Jensen page's Mistake 3: the denominator's fluctuations do not average out through a reciprocal), and the check enumerates every pair of draws for  $n = 2$  to measure it. The bias vanishes faster than the standard error, which is  $O(1/\sqrt{n})$ , so it is harmless for large  $n$  and the variance of Problem 3, with  $f$  replaced by  $f - \mu$ , is the right measure of quality; Mistake 2 overstates the guarantee.

### Problem 7

Let  $p = \mathcal{N}(0, I_d)$  and  $q = \mathcal{N}(0, \sigma^2 I_d)$  with  $\sigma^2 > \frac{1}{2}$ . Show that in one dimension  $\mathbb{E}_q[w^2] = \mathbb{E}_p[w] = \frac{\sigma^2}{\sqrt{2\sigma^2 - 1}}$ , hence in  $d$  dimensions  $\mathbb{E}_q[w^2] = \left(\frac{\sigma^2}{\sqrt{2\sigma^2 - 1}}\right)^d$ , and evaluate for  $\sigma^2 = 2$  and  $d = 10, 50$ . What happens for  $\sigma^2 \leq \frac{1}{2}$ ?

$$\begin{aligned} \text{In one dimension, } \mathbb{E}_q[w^2] &= \int \frac{p(x)^2}{q(x)} dx = \\ &= \int \frac{(2\pi)^{-1} e^{-x^2}}{(2\pi\sigma^2)^{-1/2} e^{-x^2/(2\sigma^2)}} dx = \\ &= \frac{\sigma}{\sqrt{2\pi}} \int \exp\left(-x^2\left(1 - \frac{1}{2\sigma^2}\right)\right) dx. \end{aligned}$$

Problem 3's  $\mathbb{E}_q[w^2] = \mathbb{E}_p[w] = \int p^2/q$ ;  
 collect the constants and the exponents.

$$\begin{aligned} \int e^{-\alpha x^2} dx &= \sqrt{\pi/\alpha} \text{ for } \alpha > 0, \text{ with } \alpha = \\ 1 - \frac{1}{2\sigma^2} &= \frac{2\sigma^2 - 1}{2\sigma^2}. \end{aligned}$$

The Gaussian integral; it converges only  
 when  $\alpha > 0$ , that is  $\sigma^2 > \frac{1}{2}$ .

$$\mathbb{E}_q[w^2] = \frac{\sigma}{\sqrt{2\pi}} \sqrt{\frac{2\pi\sigma^2}{2\sigma^2 - 1}} = \frac{\sigma^2}{\sqrt{2\sigma^2 - 1}}.$$

Multiply out;  $\sqrt{\pi/\alpha} =$   
 $\sqrt{2\pi\sigma^2/(2\sigma^2 - 1)}.$

In  $d$  dimensions  $p$  and  $q$  factor over coordinates, so  $w(x) = \prod_{j=1}^d w_j(x_j)$  and  $\mathbb{E}_q[w^2] = \prod_j \mathbb{E}_q[w_j^2] = \left(\frac{\sigma^2}{\sqrt{2\sigma^2 - 1}}\right)^d$ .

Independent coordinates: the expectation of a product of independent factors is the product of the expectations (the variance page).

$\sigma^2 = 2$ :  $\frac{2}{\sqrt{3}} \approx 1.1547$ , so  $\mathbb{E}_q[w^2] \approx 4.2$  for  $d = 10$  and  $\approx 1.3 \times 10^3$  for  $d = 50$ .

$1.1547^{10} \approx 4.21$  and  $1.1547^{50} \approx 1329$ .

$\mathbb{E}_q[w^2] = \left(\frac{\sigma^2}{\sqrt{2\sigma^2 - 1}}\right)^d$  **for  $\sigma^2 > \frac{1}{2}$ ; for  $\sigma^2 = 2$  it is  $\approx 4.2$  at  $d = 10$  and  $\approx 1.3 \times 10^3$  at  $d = 50$ ; for  $\sigma^2 \leq \frac{1}{2}$  it is infinite**

The base  $\sigma^2/\sqrt{2\sigma^2 - 1}$  equals 1 only at  $\sigma^2 = 1$  and exceeds 1 on both sides, so any mismatch between proposal and target is raised to the power  $d$ : the weights' variance,  $\mathbb{E}_q[w^2] - 1$ , grows exponentially with dimension, and with it the variance of every importance-sampling estimate (Problem 3). Problem 8 turns this into an effective sample size of roughly  $n/1329$  at  $d = 50$ , so a million samples are worth about 750. Below  $\sigma^2 = \frac{1}{2}$  the proposal's tails are too light and the integral diverges: unbiased, infinite variance, Mistake 3. This is why importance sampling is a low-dimensional tool, why the ELBO page's single-sample bound is loose in high dimensions, and why Markov chain methods replace it when  $d$  is large.

## Problem 8

For positive weights  $w_1, \dots, w_n$ , show that  $1 \leq \text{ESS} = \frac{(\sum_i w_i)^2}{\sum_i w_i^2} \leq n$ , with  $\text{ESS} = n$  exactly when all weights are equal and  $\text{ESS} \rightarrow 1$  when one weight dominates. Show that with normalised weights  $W_i = w_i / \sum_j w_j$ ,  $\text{ESS} = 1 / \sum_i W_i^2$ , and that for large  $n$ ,  $\text{ESS} / n \approx 1 / \mathbb{E}_q[w^2]$  when  $w = p/q$  with  $p$  and  $q$  normalised.

$(\sum_i w_i)^2 = \sum_i w_i^2 + \sum_{i \neq j} w_i w_j \geq \sum_i w_i^2$ ,  
so  $\text{ESS} \geq 1$ .

Expand the square; the cross terms are positive.

$$(\sum_i w_i)^2 = (\sum_i 1 \cdot w_i)^2 \leq (\sum_i 1^2)(\sum_i w_i^2) = n \sum_i w_i^2, \text{ so ESS} \leq n.$$

Cauchy–Schwarz for the vectors  $\mathbf{1}$  and  $w$ ; equivalently the Jensen page's  $(\mathbb{E}[W])^2 \leq \mathbb{E}[W^2]$  for the uniform distribution on the  $w_i$ .

Equality in step 2 holds when  $w$  is proportional to  $\mathbf{1}$ , all weights equal; and if  $w_1 \gg w_i$  for  $i > 1$  then both  $(\sum_i w_i)^2$  and  $\sum_i w_i^2$  are  $\approx w_1^2$ , so  $\text{ESS} \approx 1$ .

The Cauchy–Schwarz equality case; dominant-term approximation.

$$\sum_i W_i^2 = \frac{\sum_i w_i^2}{(\sum_j w_j)^2} = \frac{1}{\text{ESS}}.$$

Substitute  $W_i = w_i / \sum_j w_j$ .

$$\frac{\text{ESS}}{n} = \frac{(\frac{1}{n} \sum_i w_i)^2}{\frac{1}{n} \sum_i w_i^2} \rightarrow \frac{(\mathbb{E}_q[w])^2}{\mathbb{E}_q[w^2]} = \frac{1}{\mathbb{E}_q[w^2]}.$$

Divide top and bottom by  $n^2$  and  $n$ ; each average converges to its mean by the law of large numbers, and  $\mathbb{E}_q[w] = 1$  for normalised  $p$  and  $q$  (Problem 6 with  $Z = 1$ ).

**$1 \leq \text{ESS} \leq n$ , equal to  $n$  for equal weights and near 1 for one dominant weight;**  $\text{ESS} = 1 / \sum_i W_i^2$ ;  
 $\text{ESS} / n \approx 1 / \mathbb{E}_q[w^2]$

The effective sample size is the number of plain Monte Carlo samples the weighted sample is worth, in the sense that the self-normalised estimator's variance is approximately  $\text{Var}_p(f) / \text{ESS}$  rather than  $\text{Var}_p(f) / n$ . It is the diagnostic for Problem 7's degeneracy: with  $\mathbb{E}_q[w^2] \approx 1329$  at  $d = 50$ , a million draws have an ESS near 750, and a histogram of the normalised weights shows a handful of samples carrying almost all the mass. Sequential Monte Carlo resamples whenever ESS falls below a threshold such as  $n/2$ . The ESS uses only the weights, not  $f$ , so it can report a healthy sample for an integrand that is badly estimated, which is the limitation Problem 4's  $|f|_p$  points at.

## Problem 9

Let  $g$  be a function with known mean  $\mathbb{E}_p[g] = \gamma$ . Show that the control-variate estimator  $\hat{\mu}_c = \frac{1}{n} \sum_i (f(x_i) - c(g(x_i) - \gamma))$  is unbiased for every  $c$ , that its variance is  $\frac{1}{n} (\text{Var}(f) - 2c \text{Cov}(f, g) + c^2 \text{Var}(g))$ , minimised at  $c^* = \text{Cov}(f, g) / \text{Var}(g)$  with value  $\frac{1}{n} \text{Var}(f)(1 - \rho^2)$ . Evaluate for  $f(U) = e^U$ ,  $g(U) = U$  with  $U \sim \mathcal{U}(0, 1)$ .

$$\mathbb{E}[f(x_i) - c(g(x_i) - \gamma)] = \mu - c(\gamma - \gamma) = \mu.$$

Linearity;  $\mathbb{E}_p[g] = \gamma$  by assumption, so the correction has mean zero for every  $c$ .

$$\text{Var}(f - c(g - \gamma)) = \text{Var}(f) - 2c \text{Cov}(f, g) + c^2 \text{Var}(g), \text{ and dividing by } n \text{ gives the variance of the average.}$$

The variance page's  $\text{Var}(X - cY) = \text{Var}(X) - 2c \text{Cov}(X, Y) + c^2 \text{Var}(Y)$ ; the constant  $\gamma$  changes nothing; Problem 1 for the  $1/n$ .

$$\frac{d}{dc}(\dots) = -2 \text{Cov}(f, g) + 2c \text{Var}(g) = 0 \iff c^* = \frac{\text{Cov}(f, g)}{\text{Var}(g)}, \text{ a minimum since the coefficient of } c^2 \text{ is positive.}$$

Differentiate the quadratic in  $c$ .

$$\text{At } c^*: \text{Var}(f) - \frac{\text{Cov}(f, g)^2}{\text{Var}(g)} = \text{Var}(f) \left( 1 - \frac{\text{Cov}(f, g)^2}{\text{Var}(f) \text{Var}(g)} \right) = \text{Var}(f)(1 - \rho^2).$$

Substitute  $c^*$ ; the bracket is  $1 - \rho^2$  by the definition of correlation.

$$U \sim \mathcal{U}(0, 1): \mathbb{E}[e^U] = e - 1, \mathbb{E}[e^{2U}] = \frac{1}{2}(e^2 - 1), \text{ so } \text{Var}(e^U) = \frac{1}{2}(e^2 - 1) - (e - 1)^2 \approx 0.2420; \mathbb{E}[Ue^U] = 1, \text{ so } \text{Cov}(e^U, U) = 1 - \frac{1}{2}(e - 1) \approx 0.1409; \text{Var}(U) = \frac{1}{12}.$$

$$\int_0^1 e^u du = e - 1, \int_0^1 e^{2u} du = \frac{1}{2}(e^2 - 1), \int_0^1 ue^u du = [ue^u - e^u]_0^1 = 1 \text{ by parts (the integration page).}$$

$$c^* = 0.1409 \cdot 12 \approx 1.690, \rho^2 = \frac{0.1409^2}{0.2420/12} \approx 0.984, \text{ and the variance falls from } 0.2420 \text{ to } 0.2420(1 - 0.984) \approx 0.0039, \text{ a factor of about } 61.$$

Steps 3 and 4 with the numbers of step 5.

$\hat{\mu}_c$  **is unbiased for every**  $c$ ;  $\text{Var}(\hat{\mu}_c) = \frac{1}{n} (\text{Var}(f) - 2c \text{Cov}(f, g) + c^2 \text{Var}(g))$ ,  
**minimised at**  $c^* = \text{Cov}(f, g) / \text{Var}(g)$   
**with value**  $\frac{1}{n} \text{Var}(f)(1 - \rho^2)$ ; **for**  $e^U$   
**with control**  $U$ :  $c^* \approx 1.690$ ,  $\rho^2 \approx 0.984$ ,  
**variance**  $0.2420 \rightarrow 0.0039$

A control variate subtracts the part of  $f$  that a known-mean function can explain, which is a regression of  $f$  on  $g$ :  $c^*$  is the least-squares slope and  $1 - \rho^2$  the unexplained fraction. Because  $c$  does not affect the mean, it can be estimated from the same samples at a cost of  $O(1/n)$  bias, which is how it is done in practice. This is the policy-gradient page's baseline in disguise: there  $g$  is the score function, whose mean is known to be zero, and  $c^*$  is the optimal baseline. Any  $c$  strictly between 0 and  $2c^*$  helps and any  $c$  outside that range hurts, so a default of  $c = 1$  is right only when  $c^* \approx 1$ .

### Problem 10

For  $U \sim \mathcal{U}(0, 1)$ , the antithetic pair estimator uses  $\frac{1}{2}(f(U) + f(1 - U))$ . Show that it is unbiased and that its variance is  $\frac{1}{2} \text{Var}(f(U))(1 + \rho_a)$  with  $\rho_a = \text{Corr}(f(U), f(1 - U))$ , so it beats two independent samples exactly when  $\rho_a < 0$ . Evaluate for  $f(u) = e^u$ .

$1 - U \sim \mathcal{U}(0, 1)$ , so  $\mathbb{E}[f(1 - U)] = \mathbb{E}[f(U)] = \mu$  and the pair average has mean  $\frac{1}{2}(\mu + \mu) = \mu$ .

Reflecting a uniform variable about  $\frac{1}{2}$  gives a uniform variable; linearity.

$\text{Var}\left(\frac{f(U) + f(1 - U)}{2}\right) = \frac{1}{4} (\text{Var}(f(U)) + \text{Var}(f(1 - U)) + 2 \text{Cov}(f(U), f(1 - U))) = \frac{1}{4} (2v + 2\rho_a v) = \frac{v}{2} (1 + \rho_a)$  with  $v = \text{Var}(f(U))$ .

The variance page's variance of a sum; both terms have variance  $v$  by step 1, and the covariance is  $\rho_a v$  by the definition of correlation.

Two independent samples have variance  $v/2$ , so the pair is better exactly when  $1 + \rho_a < 1$ , that is  $\rho_a < 0$ .

Compare  $\frac{v}{2}(1 + \rho_a)$  with  $\frac{v}{2}$ .

For  $f(u) = e^u$ :  $v = \frac{1}{2}(e^2 - 1) - (e - 1)^2 \approx 0.2420$  and  $\text{Cov}(e^U, e^{1-U}) = \mathbb{E}[e^U e^{1-U}] - \mu^2 = e - (e - 1)^2 \approx -0.2342$ , so  $\rho_a \approx -0.968$ .

$e^U e^{1-U} = e$  is constant; Problem 9's moments for  $v$  and  $\mu = e - 1$ .

The pair's variance is  $\frac{1}{2}(0.2420)(1 - 0.968) \approx 0.0039$ , against 0.1210 for two independent samples: a factor of about 31.

Step 2 with the numbers of step 4.

**The antithetic pair is unbiased with variance  $\frac{1}{2} \text{Var}(f(U))(1 + \rho_a)$ , better than two independent samples exactly when  $\rho_a < 0$ ; for  $e^u$ ,  $\rho_a \approx -0.968$  and the variance per pair is  $\approx 0.0039$  against 0.1210**

For a monotone  $f$ ,  $f(U)$  and  $f(1 - U)$  move in opposite directions, so  $\rho_a < 0$  and the pair's errors partly cancel; for  $e^u$  they cancel almost entirely because the product  $f(U)f(1 - U)$  is constant. For a symmetric  $f$  with  $f(u) = f(1 - u)$  the pair is one sample counted twice,  $\rho_a = 1$ , and the pair's variance is  $v$ , twice that of two independent samples (Mistake 4). The Gaussian version pairs  $\varepsilon$  with  $-\varepsilon$ , and it is why the ELBO page's reparameterised gradient estimates are sometimes computed on symmetric pairs; the mechanism, inducing negative correlation between samples that are averaged, is also what stratified and quasi-Monte Carlo sampling do more systematically.

## Where this goes wrong

### 1. Weights with the numerator and denominator swapped

Samples come from  $q$  and the weight corrects for  $q$ , so  $q$  goes on top.

$$\mathbb{E}_p[f] = \mathbb{E}_q[wf] \text{ with } x_i \sim q$$

Right so far: Problem 2.

“The weight is the proposal over the target: the samples are from  $q$ , so divide out  $q$ .”

The habit that causes the mistake: Problem 2's proof multiplied and divided by  $q$ , leaving  $p/q$  under  $\mathbb{E}_q$ ; the proposal's density goes in the denominator because it was put into the expectation, and the target's goes in the numerator because it is what is wanted.

$$\hat{\mu} = \frac{1}{n} \sum_i \frac{q(x_i)}{p(x_i)} f(x_i) \text{ with } x_i \sim q \quad \times$$

Its mean is  $\mathbb{E}_q[(q/p)f] = \sum_x q(x)^2 f(x)/p(x)$ , which is not  $\mu$  and need not be close to it. In Problem 5 the swapped weight is  $e^{3x-9/2}$ , which exceeds  $e^{4.5} \approx 90$  at  $x = 3$  and grows from there, so the estimate of a probability of 0.00135 comes out in the hundreds. The check for the orientation is Problem 6's  $\mathbb{E}_q[w] = 1$ : the weights of a correctly oriented normalised pair average to one over the samples, and the swapped ones average to  $\mathbb{E}_q[q/p] \geq 1$  with equality only when  $p = q$ , by the Jensen page's Problem 5.

## 2. Calling the self-normalised estimator unbiased

The weighted average with normalised weights looks exactly like Problem 2's estimator, which is unbiased.

$$\hat{\mu}_{\text{SN}} = \sum_i \tilde{w}_i f(x_i) / \sum_i \tilde{w}_i \rightarrow \mu$$

Right so far: Problem 6, step 3.

“The numerator is unbiased for  $Z\mu$  and the denominator for  $Z$ , so the ratio is unbiased for  $\mu$ .”

The shortcut that causes the mistake: the expectation of a ratio is not the ratio of expectations (the Jensen page's Mistake 3).

$$\mathbb{E}[\hat{\mu}_{\text{SN}}] = \mu \text{ for every } n \neq 1$$

For  $n = 1$  the estimator is  $f(x_1)$  with  $x_1 \sim q$ , whose mean is  $\mathbb{E}_q[f]$ , the integrand averaged under the wrong distribution (Problem 6); the check enumerates  $n = 2$  and finds the mean still off. The bias is  $O(1/n)$ , so the self-normalised estimator is consistent, and in practice its bias is swamped by its  $O(1/\sqrt{n})$  standard error, but it is not unbiased, and the difference matters when  $n$  is small or when many such estimates are averaged, because averaging reduces variance and not bias (the bias–variance page). When  $Z$  is known, Problem 2's estimator with the normalised weights is the unbiased one.

### 3. A proposal narrower than the target

A proposal concentrated on the region where the target is large seems like the efficient choice.

$$\text{Var}(\hat{\mu}_{\text{IS}}) = \frac{1}{n} (\mathbb{E}_p[w f^2] - \mu^2) \text{ with } w = p/q \quad \text{Right so far: Problem 3.}$$

“ $q = \mathcal{N}(0, \frac{1}{4})$  covers the mode of  $p = \mathcal{N}(0, 1)$  and wastes no samples on the tails.”

The habit that causes the mistake: the variance is an expectation under  $p$  of  $p/q$ , so what  $q$  does in the tails of  $p$ , where  $p/q$  can be huge, is what decides it.

$q = \mathcal{N}(0, \frac{1}{4})$  is a valid proposal with finite variance for  $f = 1$  ✗

Problem 7 with  $\sigma^2 = \frac{1}{4} < \frac{1}{2}$ :  $\mathbb{E}_q[w^2] = \int p^2/q$  diverges, because  $p^2/q \propto e^{-x^2+2x^2} = e^{x^2}$ . The estimator is still unbiased, so a run looks fine, with the weights small and the average stable, until a sample from the tail of  $p$  arrives with a weight of  $e^{x^2/2}$ -scale size and moves the average by more than everything before it; the empirical variance never settles, and the usual error bars are fiction. Any  $q$  whose tails are lighter than those of  $|f|p$  has this problem, which is why heavy-tailed proposals (a Student- $t$  around the mode, or a mixture with a wide component) are the standard safe choice, and why, in Problem 7's family,  $\sigma^2$  slightly above 1 is better than slightly below.

#### 4. Antithetic pairs on a symmetric integrand

Antithetic sampling is a free variance reduction, so it is switched on by default.

$$\text{Var} \left( \frac{f(U) + f(1-U)}{2} \right) = \frac{v}{2}(1 + \rho_a)$$

Right so far: Problem 10.

“ $f(U)$  and  $f(1-U)$  are negatively correlated because  $U$  and  $1-U$  are.”

The habit that causes the mistake:  $U$  and  $1-U$  have correlation  $-1$ , and  $f$  may not preserve that; Problem 10 needed  $\rho_a < 0$ , which holds for monotone  $f$  and not in general.

For  $f(u) = (u - \frac{1}{2})^2$ , the antithetic pair halves the variance ✗

$f(1 - u) = f(u)$ , so the pair is  $f(U)$  twice,  $\rho_a = 1$ , and the pair average has variance  $v$  where two independent samples would have  $v/2$ : the “free reduction” doubles the variance per sample. Any  $f$  symmetric about  $\frac{1}{2}$  does this, and any  $f$  with a symmetric component gains less than the monotone case suggests. The test is the sign of  $\rho_a$ , which can be estimated from a pilot run; for the Gaussian version with  $\varepsilon \rightarrow -\varepsilon$ , an even integrand such as  $\varepsilon^2$  or  $\|\varepsilon\|^2$  is the same failure, which is relevant to the ELBO page, where the KL term is even in  $\varepsilon$  and the antithetic pair does nothing for it.

## 5. Quoting the plain Monte Carlo variance for the weighted estimate

The importance-sampling estimator is unbiased for the same  $\mu$ , and its error is reported as if it were the plain one.

$$\mathbb{E}[\hat{\mu}_{\text{IS}}] = \mu = \mathbb{E}[\hat{\mu}_n]$$

Right so far: Problems 1 and 2.

“Same target, same mean, same error bars:  $\text{Var}_p(f)/n$ .”

The shortcut that causes the mistake: the two estimators average different random variables,  $f(X)$  under  $p$  and  $w(X)f(X)$  under  $q$ , and Problem 3's variance depends on  $q$  through  $\mathbb{E}_p[wf^2]$ .

$\text{Var}(\hat{\mu}_{\text{IS}}) = \text{Var}_p(f)/n$  whatever the proposal  $\times$

In Problem 5 the true variance is 218 times smaller than that, and with the proposal of Mistake 3 it is infinite; neither is visible from  $\text{Var}_p(f)$ . The right error bar is the empirical variance of the weighted terms  $w(x_i)f(x_i)$  divided by  $n$ , and the right diagnostic for whether that empirical variance can be trusted is the ESS of Problem 8. The same slip appears with the self-normalised estimator, whose variance uses  $f - \mu$  in place of  $f$  and ESS in place of  $n$ , and with the policy-gradient page's importance-weighted surrogate, whose variance grows with the ratios as the policy moves away from the one that generated the data.