

Variance Minimisation and Exponential Tilting

Junqi Tang

Computational Statistics

The quantity we want to estimate

$$I = \mathbb{E}_p[f(X)] = \int f(x)p(x) dx.$$

- $p(x)$ is the original probability density.
- $f(x)$ is the function whose expectation we need.
- Standard Monte Carlo uses $X_i \stackrel{\text{iid}}{\sim} p$ and averages $f(X_i)$.

For a rare-event probability, $f(x) = \mathbf{1}_{\{x \in A\}}$ and

$$I = \mathbb{P}_p(X \in A).$$

A sample is useful when it tells us about the regions contributing to I .

Importance sampling and the density ratio

Draw $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} q$ and use

$$w(x) = \frac{p(x)}{q(x)}, \quad \hat{I}_n = \frac{1}{n} \sum_{i=1}^n f(Y_i) w(Y_i).$$

The correction works because

$$\mathbb{E}_q[f(Y)w(Y)] = \int f(x) \frac{p(x)}{q(x)} q(x) dx = I.$$

- Require $q(x) > 0$ wherever $f(x)p(x) \neq 0$, almost everywhere.

The proposal changes where we sample. The weight preserves the target.

A proposal can help or hurt

Teaching example: $X \sim \text{Bernoulli}(0.01)$ and $f(X) = X$. We estimate $I = 0.01$ using $Y \sim \text{Bernoulli}(r)$.

$$Z = f(Y)w(Y) = \begin{cases} 0.01/r, & Y = 1, \\ 0, & Y = 0, \end{cases} \quad \text{Var}(Z) = 0.01^2 \left(\frac{1}{r} - 1 \right).$$

Proposal	r	Variance of one contribution Z
Standard Monte Carlo	0.01	0.0099
More frequent successes	0.50	0.0001
Fewer successes	0.001	0.0999

For n independent draws, divide these variances by n .

Two approaches to proposal design

Problem-specific design

- Use knowledge of f and p to identify important regions.
- The ideal shape is $q^*(x) \propto |f(x)|p(x)$.

Systematic optimisation

- Choose a family $q(x; \theta)$.
- Select θ by minimising an estimation-error criterion.

For $f \geq 0$, the ideal density is $q^*(x) = f(x)p(x)/I$ when $I > 0$. Its normalising constant is the unknown target itself.

A parametric family of proposals

We restrict our search to

$$\{q(x; \theta) : \theta \in \Theta\}, \quad w(x; \theta) = \frac{p(x)}{q(x; \theta)}.$$

Examples:

- $q(\cdot; \theta) = \mathcal{N}(\theta, 1)$: optimise the mean.
- $q(\cdot; \mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2)$: optimise both parameters.

The parameter space must enforce valid distributions, such as $\sigma^2 > 0$.

An optimum within a family depends on the family we choose.

Why minimise variance

For a fixed valid proposal, $\hat{I}_n$ is unbiased:

$$\mathbb{E}_q[\hat{I}_n] = I.$$

The bias–variance decomposition gives

$$\text{MSE}(\hat{I}_n) = \text{Var}(\hat{I}_n) + (\mathbb{E}[\hat{I}_n] - I)^2 = \text{Var}(\hat{I}_n).$$

- At a fixed sample size, smaller variance means smaller MSE.

Variance of the importance sampling estimator

Let $Z_i = f(Y_i)w(Y_i; \theta)$ with $Y_i \stackrel{\text{iid}}{\sim} q(\cdot; \theta)$.

$$\begin{aligned}\text{Var}(\hat{I}_n) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n Z_i\right) \\ &= \frac{1}{n} \text{Var}(Z_1) \\ &= \frac{1}{n} \left(\mathbb{E}_{q(\cdot; \theta)}[f(Y)^2 w(Y; \theta)^2] - I^2 \right).\end{aligned}$$

The parameter changes the second moment. The target I stays fixed.

Writing the second moment under the original distribution

Substitute $w(x; \theta) = p(x)/q(x; \theta)$:

$$\begin{aligned}\mathbb{E}_{q(\cdot; \theta)}[f(Y)^2 w(Y; \theta)^2] &= \int f(x)^2 \frac{p(x)^2}{q(x; \theta)^2} q(x; \theta) dx \\ &= \int f(x)^2 \frac{p(x)}{q(x; \theta)} p(x) dx \\ &= \mathbb{E}_p[f(X)^2 w(X; \theta)].\end{aligned}$$

The variance minimisation objective

Define

$$F(\theta) = \mathbb{E}_p[f(X)^2 w(X; \theta)].$$

Then

$$\text{Var}(\hat{I}_n) = \frac{F(\theta) - I^2}{n}.$$

For a fixed n , neither I^2 nor $1/n$ changes the minimiser:

$$\theta^* \in \arg \min_{\theta \in \Theta} F(\theta).$$

We can optimise the proposal without knowing I .

An expectation can be hard to evaluate

We want to minimise

$$F(\theta) = \int f(x)^2 \frac{p(x)^2}{q(x; \theta)} dx.$$

- This integral may have no convenient closed form.
- An optimiser usually needs many objective evaluations.
- Repeated exact integration may be impractical.

Replace the expectation by a Monte Carlo average, then optimise it.

Sample average approximation

Generate a **pilot sample** and hold it fixed:

$$X_1, \dots, X_m \stackrel{\text{iid}}{\sim} p.$$

Approximate the population objective by

$$\hat{F}_m(\theta) = \frac{1}{m} \sum_{j=1}^m f(X_j)^2 w(X_j; \theta).$$

Then solve

$$\hat{\theta}_m \in \arg \min_{\theta \in \Theta} \hat{F}_m(\theta).$$

m is the fitting sample size. We reserve n for the final estimate.

What stays fixed during optimisation

Write $a_j = f(X_j)^2$. Then

$$\hat{F}_m(\theta) = \frac{1}{m} \sum_{j=1}^m a_j \frac{p(X_j)}{q(X_j; \theta)}.$$

- The sampled locations X_j and coefficients a_j stay fixed.
- Only the proposal parameter θ changes.
- Compute expensive values $f(X_j)$ once and reuse them.
- Use an appropriate constrained numerical optimisation routine.

A finite-sample minimiser $\hat{\theta}_m$ generally differs from θ^* .

Fitting and estimation use different samples

1. Draw $X_1, \dots, X_m \stackrel{\text{iid}}{\sim} p$ and fit $\hat{\theta}_m$.
2. Given the pilot sample, draw fresh observations

$$Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} q(\cdot; \hat{\theta}_m).$$

3. Estimate the target using

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^n f(Y_i) \frac{p(Y_i)}{q(Y_i; \hat{\theta}_m)}.$$

Under the support and integrability conditions,

$$\mathbb{E}[\hat{I}_n \mid X_1, \dots, X_m] = I.$$

A pilot sample can miss the important region

Suppose $f(x) = \mathbf{1}_{\{x \in A\}}$ and $\mathbb{P}_p(X \in A) = 0.01$. With $m = 50$ pilot samples,

$$\mathbb{P}(\text{no pilot observation in } A) = (1 - 0.01)^{50} \approx 0.605.$$

If all observations miss A , then

$$\hat{F}_m(\theta) = 0 \quad \text{for every valid } \theta.$$

An apparently small objective can reflect an uninformative pilot sample.

Proposal quality also depends on family choice, optimisation error and fitting cost.

Exponential tilting constructs a proposal family

Starting from p , define for $x, \theta \in \mathbb{R}^d$

$$q(x; \theta) = \frac{e^{\theta^\top x} p(x)}{C(\theta)}.$$

The normalising constant is

$$C(\theta) = \int e^{\theta^\top x} p(x) dx = \mathbb{E}_p[e^{\theta^\top X}].$$

- $\theta^\top x$ is an inner product.
- $C(\theta)$ depends on θ , and does not depend on x .
- Only parameters with $C(\theta) < \infty$ are admissible.

What the tilt changes

In one dimension,

$$\frac{q(x; \theta)}{p(x)} = \frac{e^{\theta x}}{C(\theta)}.$$

- $\theta > 0$ gives relatively more weight to larger values of x .
- $\theta < 0$ gives relatively more weight to smaller values of x .
- $\theta = 0$ gives $C(0) = 1$ and $q(x; 0) = p(x)$.

The exponential factor changes relative mass; normalisation restores total mass one.

The importance weight under exponential tilting

Taking the density ratio gives

$$\begin{aligned}w(x; \theta) &= \frac{p(x)}{e^{\theta^\top x} p(x) / C(\theta)} \\&= \boxed{C(\theta) e^{-\theta^\top x}}.\end{aligned}$$

- The exponent changes sign because we take a reciprocal.
- The factor $C(\theta)$ remains in the weight.

A normalising constant in x can still vary with the parameter being optimised.

The objective for choosing the tilt

Substitution into the second moment gives

$$F(\theta) = C(\theta)\mathbb{E}_p\left[f(X)^2 e^{-\theta^\top X}\right].$$

For fixed pilot observations,

$$\hat{F}_m(\theta) = \frac{C(\theta)}{m} \sum_{j=1}^m f(X_j)^2 e^{-\theta^\top X_j}.$$

- Changing f changes the coefficients $f(X_j)^2$.
- We can reuse the structure of the optimisation code.
- The best parameter still depends on the integrand.

Worked example: tilting a standard normal

Let $p(x) = (2\pi)^{-1/2}e^{-x^2/2}$. Complete the square:

$$-\frac{x^2}{2} + \theta x = -\frac{(x - \theta)^2}{2} + \frac{\theta^2}{2}.$$

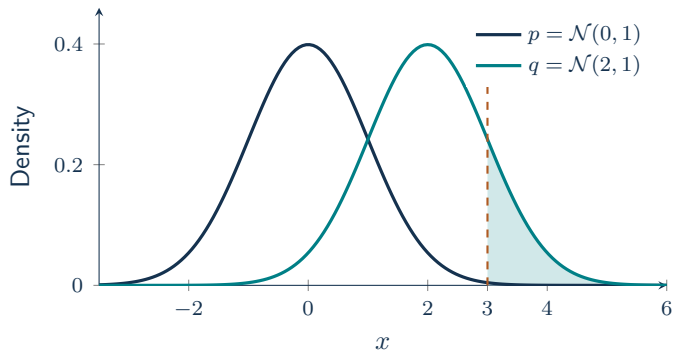
Therefore,

$$C(\theta) = e^{\theta^2/2}, \quad q(x; \theta) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}.$$

Exponential tilting changes $\mathcal{N}(0, 1)$ into $\mathcal{N}(\theta, 1)$.

$$w(x; \theta) = \exp\left(\frac{\theta^2}{2} - \theta x\right).$$

A Gaussian tilt puts more mass in the right tail



Target event: $X > 3$. The proposal makes observations above 3 more frequent. Each observation still receives its importance weight. Here $\theta = 2$ illustrates a tilt. It is not an asserted optimum.

Fitting a Gaussian tilt

For $p = \mathcal{N}(0, 1)$, the sample objective becomes

$$\hat{F}_m(\theta) = \frac{1}{m} \sum_{j=1}^m f(X_j)^2 \exp\left(\frac{\theta^2}{2} - \theta X_j\right).$$

Its derivative is available explicitly:

$$\hat{F}'_m(\theta) = \frac{1}{m} \sum_{j=1}^m f(X_j)^2 e^{\theta^2/2 - \theta X_j} (\theta - X_j).$$

- Compute the coefficients on the fixed pilot sample.
- Minimise this one-dimensional function numerically.
- Draw fresh $Y_i \sim \mathcal{N}(\hat{\theta}_m, 1)$ for estimation.

Practical conditions for exponential tilting

1. **Existence:** $C(\theta)$ must be finite.
2. **Weight evaluation:** we need a usable value of $C(\theta)$.
3. **Sampling:** we must be able to generate observations from q .
4. **Variance:** $\mathbb{E}_q[f(Y)^2 w(Y; \theta)^2]$ must be finite.

A familiar original distribution need not produce a familiar tilted distribution.

Exponential families help explain when the tilted form stays convenient.

The exponential family form

A parametric family is an exponential family if its density has the form

$$p(x; \lambda) = h(x) \exp\left(\eta(\lambda)^\top T(x) - A(\lambda)\right).$$

- λ parameterises the original distribution family.
- θ denotes a tilting parameter in this lecture.
- The representation uses a common support and base measure.
- For discrete distributions, use probability masses and sums.

The terms in an exponential family

$$p(x; \lambda) = h(x)e^{\eta(\lambda)^{\top}T(x) - A(\lambda)}.$$

Term	Depends on	Meaning
$h(x) \geq 0$	x	Base factor
$T(x)$	x	Vector of statistics
$\eta(\lambda)$	λ	Natural parameter vector
$A(\lambda)$	λ	Log-normalising term

$$A(\lambda) = \log \int h(x)e^{\eta(\lambda)^{\top}T(x)} dx.$$

$T(x)$ and $\eta(\lambda)$ must have the same dimension.

A normal family in exponential family form

For a normal distribution with mean μ and fixed variance 1,

$$p(x; \mu) = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2} = \underbrace{\frac{1}{\sqrt{2\pi}} e^{-x^2/2}}_{h(x)} e^{\mu x - \mu^2/2}.$$

We can identify

$$T(x) = x, \quad \eta(\mu) = \mu, \quad A(\mu) = \frac{\mu^2}{2}.$$

Multiplying by $e^{\theta x}$ shifts the natural parameter from μ to $\mu + \theta$.

After normalisation, the new distribution is $\mathcal{N}(\mu + \theta, 1)$.

Tilting the statistics shifts natural parameters

Write a family in natural parameter form as

$$p_{\eta}(x) = h(x)e^{\eta^{\top}T(x)-A(\eta)}.$$

Tilting by $e^{\theta^{\top}T(x)}$ gives

$$q_{\theta}(x) = h(x)e^{(\eta+\theta)^{\top}T(x)-A(\eta+\theta)},$$

provided $\eta + \theta$ remains an admissible parameter in the family. Our earlier linear tilt uses $T(x) = x$, or statistics that can represent the linear term.

The form of the tilt must match the statistics in the chosen family.

Recap and a short check

Proposal design by variance minimisation

$$F(\theta) = \mathbb{E}_p[f(X)^2 w(X; \theta)], \quad \hat{F}_m(\theta) = \frac{1}{m} \sum_j f(X_j)^2 w(X_j; \theta).$$

Exponential tilting

$$q(x; \theta) = \frac{e^{\theta^\top x} p(x)}{C(\theta)}, \quad w(x; \theta) = C(\theta) e^{-\theta^\top x}.$$

1. Why does the variance objective contain one weight under p ?
2. Why must $C(\theta)$ remain in the optimisation objective?
3. What happens if a rare-event pilot sample contains no hits?