Advanced Simulation - Lecture 3

George Deligiannidis

January 25th, 2016

Rejection Sampling

Recall

Algorithm (Rejection Sampling). Given two densities π, q with $\pi(x) \leq M\, q(x)$ for all x, we can generate a sample from π by

1. Draw $X \sim q$, draw $U \sim \mathcal{U}_{[0,1]}$.
2. Accept $X = x$ as a sample from π if

$$U \leq \frac{\pi(x)}{M\, q(x)},$$

otherwise go to step 1.

Proposition

The distribution of the samples accepted by rejection sampling is π.

Rejection Sampling

- Often we only know π and q up to some normalising constants; i.e.

$$\pi = \widetilde{\pi}/Z_\pi \quad \text{and} \quad q = \widetilde{q}/Z_q$$

where $\widetilde{\pi}, \widetilde{q}$ are known but Z_π, Z_q are unknown.
You still need to be able to sample from $q(\cdot)$.

- If you can upper bound:

$$\widetilde{\pi}(x)/\widetilde{q}(x) \leq \widetilde{M},$$

then using $\widetilde{\pi}$, $\widetilde{q}$ and $\widetilde{M}$ in the algorithm is correct.

- Indeed we have

$$\frac{\widetilde{\pi}(x)}{\widetilde{q}(x)} \leq \widetilde{M} \Leftrightarrow \frac{\pi(x)}{q(x)} \leq \widetilde{M}\frac{Z_q}{Z_\pi} = M.$$

Let T denote the number of pairs (X, U) that have to be generated until X is accepted for the first time.

Lemma

T is geometrically distributed with parameter $1/M$ and in particular $\mathbb{E}(T) = M$.

In the unnormalised case, this yields

$$\mathbb{P}(X \text{ accepted}) = \frac{1}{M} = \frac{Z_\pi}{\widetilde{M} Z_q},$$

$$\mathbb{E}(T) = M = \frac{Z_q \widetilde{M}}{Z_\pi},$$

and it can be used to provide unbiased estimates of Z_π / Z_q and Z_q / Z_π.

Examples:Uniform from bounded subset of $\mathbb{R}^p$

- Let $B \subset \mathbb{R}^p$, a bounded subset of $\mathbb{R}^p$:

$$\pi(x) \propto \mathbb{I}_B(x).$$

Let R be a rectangle containing $B \subset R$ and

$$q(x) \propto \mathbb{I}_R(x).$$

- Then we can use $\widetilde{M} = 1$ and

$$\widetilde{\pi}(x) / \left(\widetilde{M}' \widetilde{q}(x)\right) = \mathbb{I}_B(x).$$

- The probability of accepting a sample is then Z_π / Z_q.

Example: Normal density

- Let $\widetilde{\pi}(x) = \exp\left(-\frac{1}{2}x^2\right)$ and $\widetilde{q}(x) = 1/\left(1+x^2\right)$. We have

$$\frac{\widetilde{\pi}(x)}{\widetilde{q}(x)} = (1+x^2)\exp\left(-\frac{1}{2}x^2\right) \leq 2/\sqrt{e} = \widetilde{M}$$

  which is attained at ± 1.
- Let $X \sim \widetilde{q}$. The acceptance probability is

$$\mathbb{P}\left(U \leq \frac{\widetilde{\pi}(X)}{\widetilde{M}\widetilde{q}(X)}\right) = \frac{Z_\pi}{\widetilde{M}Z_q} = \frac{\sqrt{2\pi}}{\frac{2}{\sqrt{e}}\pi} = \sqrt{\frac{e}{2\pi}} \approx 0.66,$$

  and the mean number of trials to success is approximately $1/0.66 \approx 1.52$.

Examples: Genetic Linkage model

- We observe

$$(Y_1, Y_2, Y_3, Y_4) \sim \mathcal{M}\left(n; \frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right)$$

  where $\mathcal{M}$ is the multinomial distribution and $\theta \in (0,1)$.
- The likelihood of the observations is thus

$$\begin{aligned} & p(y_1, ..., y_4; \theta) \\ & = \frac{n!}{y_1! y_2! y_3! y_4!} \left(\frac{1}{2} + \frac{\theta}{4}\right)^{y_1} \left(\frac{1}{4}(1-\theta)\right)^{y_2+y_3} \left(\frac{\theta}{4}\right)^{y_4} \\ & \propto (2+\theta)^{y_1} (1-\theta)^{y_2+y_3} \theta^{y_4}. \end{aligned}$$

- Bayesian approach where we select $p(\theta) = \mathbb{I}_{[0,1]}(\theta)$ and are interested in

$$p(\theta | y_1, ..., y_4) \propto (2+\theta)^{y_1} (1-\theta)^{y_2+y_3} \theta^{y_4} \mathbb{I}_{[0,1]}(\theta).$$

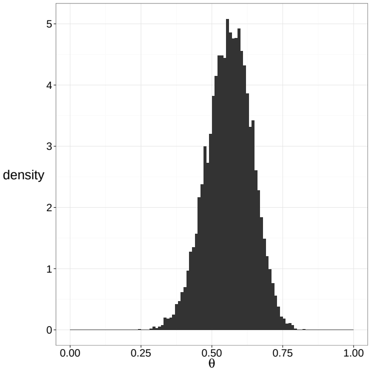
Examples: Genetic linkage model

- Rejection sampling using the prior as proposal $q(\theta) = \widetilde{q}(\theta) = p(\theta)$ to sample from $p(\theta | y_1, ..., y_4)$.
- To use accept-reject, we need to upper bound

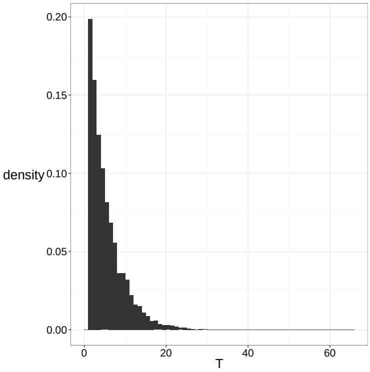
$$\frac{\widetilde{\pi}(\theta)}{\widetilde{q}(\theta)} = \widetilde{\pi}(\theta) = (2+\theta)^{y_1} (1-\theta)^{y_2+y_3} \theta^{y_4}$$

- Maximum of $\widetilde{\pi}$ can be found using standard optimization procedure to perform rejection sampling.
- For a realisation of (Y_1, Y_2, Y_3, Y_4) equal to $(69, 9, 11, 11)$ obtained with $n = 100$ and $\theta^\star = 0.6$, results shown in following figure.

Examples: Genetic linkage model



(a) Figure A



(b) Figure B

Figure: Histogram of 10,000 samples drawn from posterior obtained by rejection sampling (left); and histogram of waiting time distribution before acceptance (right).

Rejection Sampling Recap

Rejection sampling requires

1. Samples from some distribution q;

2. evaluation of $\pi(\cdot)$ point-wise, or unnormalized $\widetilde{\pi}$;

3. an upper bound M on $\pi(x)/q(x)$, or $\widetilde{\pi}/q$ and so on.

Sometimes the upper bound is not feasible.

Importance Sampling

- We want to compute

$$I = \mathbb{E}_\pi(\varphi(X)) = \int_{\mathbb{X}} \varphi(x)\,\pi(x)\,dx.$$

- We do not know how to sample from the target π but have access to a proposal distribution of density q.

- We only require that

$$\pi(x) > 0 \Rightarrow q(x) > 0;$$

  i.e. the support of q includes the support of π.

- q is called the proposal, or importance distribution.

Importance Sampling

- We have the following identity

$$I = \mathbb{E}_\pi(\varphi(X)) = \mathbb{E}_q(\varphi(X)w(X)),$$

where $w : \mathbb{X} \to \mathbb{R}^+$ is the importance weight function

$$w(x) = \frac{\pi(x)}{q(x)}.$$

- Hence for $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} q$,

$$\widehat{I}_n^{\text{IS}} = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i) w(X_i).$$

Importance Sampling Properties

Proposition

(a) ***Unbiased:*** $\mathbb{E}_q[\widehat{I}_n^{IS}] = I$;

(b) ***Strongly consistent:*** *If* $\mathbb{E}_q(|\varphi(X)|\, w(X)) < \infty$ *then*

$$\lim_{n\to\infty} \widehat{I}_n^{IS} = I, \quad a.s.$$

(c) ***CLT:*** $\mathbb{V}_q(\widehat{I}_n^{IS}) = \sigma_{IS}^2 / n$ *where*

$$\sigma_{IS}^2 := \mathbb{V}_q\left(\varphi(X) w(X)\right)$$

If $\sigma_{IS}^2 < \infty$ *then*

$$\lim_{n\to\infty} \sqrt{n}\left(\widehat{I}_n^{IS} - I\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_{IS}^2\right).$$

Importance Sampling: Practical Advice

- Consistency does not require $\sigma_{\text{IS}}^2 < \infty$ but highly recommended in practice (!).

- **Sufficient condition**: If $\mathbb{E}_\pi\left(\varphi^2(X)\right) < \infty$ and $w(x) \leq M$ for all x for some $M < \infty$, then $\sigma_{\text{IS}}^2 < \infty$.

- In practice ensure $w(x) \leq M$ although it is neither necessary nor sufficient, as seen in the following example.

Importance Sampling: Example

- $\pi(x) = \mathcal{N}(x; 0, 1)$, $q(x) = \mathcal{N}(x; 0, \sigma^2)$

$$w(x) = \frac{\pi(x)}{q(x)} \propto \exp\left[-x^2\left(1 - \frac{1}{\sigma^2}\right)\right].$$

- For $\sigma^2 \geq 1$, $w(x) \leq M$ for all x,
  and for $\sigma^2 < 1$, $w(x) \to \infty$ as $|x| \to \infty$.

- For $\varphi(x) = x^2$, we have $\sigma^2_{\text{IS}} < \infty$ for all $\sigma^2 > 1/2$.

- For $\varphi(x) = \exp\left(\frac{\beta}{2}x^2\right)$, we have $I < \infty$ for $\beta < 1$
  but $\sigma^2_{\text{IS}} = \infty$ for $\beta > 1 - \frac{1}{2\sigma^2}$.

Optimal Importance Distribution I

Question

Is there a best proposal that minimizes the variance σ_{IS}^2?

Proposition

The optimal proposal minimising $\mathbb{V}_q\left(\widehat{I}_n^{IS}\right)$ *is given by*

$$q_{opt}(x) = \frac{|\varphi(x)|\,\pi(x)}{\int_{\mathbb{X}} |\varphi(x)|\,\pi(x)\,dx}.$$

Optimal Importance Distribution II

Proof.

We have indeed

$$\sigma_{\text{IS}}^2 = \mathbb{V}_q\left(\varphi(X)w\left(X\right)\right) = \mathbb{E}_q\left(\varphi^2(X)w^2\left(X\right)\right) - I^2.$$

We also have by Jensen's inequality for any q

$$\mathbb{E}_q\left(\varphi^2(X)w^2\left(X\right)\right) \geq \left(\int_{\mathbb{X}} \left|\varphi(x)\right| \pi\left(x\right) dx\right)^2.$$

For $q = q_{\text{opt}}$, we have

$$\begin{aligned}\mathbb{E}_{q_{\text{opt}}}\left(\varphi^2(X)w^2\left(X\right)\right) &= \int_{\mathbb{X}} \frac{\varphi^2(x)\pi^2\left(x\right)}{\left|\varphi(x)\right| \pi\left(x\right)} dx \times \int_{\mathbb{X}} \left|\varphi(x)\right| \pi\left(x\right) dx \\ &= \left(\int_{\mathbb{X}} \left|\varphi(x)\right| \pi\left(x\right) dx\right)^2.\end{aligned}$$

□

Optimal Importance Distribution

- $q_{\text{opt}}(x)$ can never be used in practice!
- For $\varphi(x) > 0$ we have $q_{\text{opt}}(x) = \varphi(x)\pi(x)/I$ and $\mathbb{V}_{q_{\text{opt}}}\left(\widehat{I}_n^{\text{IS}}\right) = 0$ but this is because

$$\varphi(x)\, w(x) = \varphi(x) \frac{\pi(x)}{q_{\text{opt}}(x)} = I,$$

  it requires knowing I!
- This can be used as a guideline to select q; i.e. select $q(x)$ such that $q(x) \approx q_{\text{opt}}(x)$.
- Particularly interesting in rare event simulation, not quite in statistics.

Normalised Importance Sampling

- Standard IS has limited applications in statistics as it requires knowing $\pi(x)$ and $q(x)$ exactly.

- Assume $\pi(x) = \widetilde{\pi}(x)/Z_\pi$ and $q(x) = \widetilde{q}(x)/Z_q$, $\pi(x) > 0 \Rightarrow q(x) > 0$ and and define

$$\widetilde{w}(x) = \frac{\widetilde{\pi}(x)}{\widetilde{q}(x)}.$$

- An alternative identity is

$$I = \mathbb{E}_\pi(\varphi(X)) = \frac{\int_{\mathbb{X}} \varphi(x)\,\widetilde{w}(x)\,q(x)dx}{\int_{\mathbb{X}} \widetilde{w}(x)q(x)dx}.$$

SLLN for NIS

Proposition (SLLN for NIS)

Let $X_1, ..., X_n \overset{i.i.d.}{\sim} q$ *and assume that* $\mathbb{E}_q(|\varphi(X)|\, w(X)) < \infty$. *Then*

$$\widehat{I}_n^{NIS} = \frac{\sum_{i=1}^{n} \varphi(X_i)\widetilde{w}(X_i)}{\sum_{i=1}^{n} \widetilde{w}(X_i)}$$

is strongly consistent.

Proof.

Divide numerator and denominator by n. Both converge almost surely by the strong law of large numbers. □

CLT for NIS

Proposition

If $\mathbb{V}_q(\varphi(X)w(X)) < \infty$ *and* $\mathbb{V}_q(w(X)) < \infty$ *then*

$$\sqrt{n}(\widehat{I}_n^{NIS} - I) \Rightarrow \mathcal{N}(0, \sigma_{NIS}^2),$$

where

$$\begin{aligned}\sigma_{NIS}^2 &:= \mathbb{V}_q\Big(\big[\varphi(X)w(X)) - Iw(X)\big]\Big) \\ &= \int \frac{\pi(x)^2\left(\varphi(x) - I\right)^2}{q(x)} \mathrm{d}x.\end{aligned}$$

Proof

Proof.

First notice that with $X_1, \ldots, X_n$ i.i.d. $\sim q$

$$\sqrt{n}(\widehat{I}_n^{\text{NIS}} - I) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{w}(X_i) \left[\varphi(X_i) - I\right]}{\frac{1}{n} \sum_{i=1}^{n} \widetilde{w}(X_i)}$$

where since $\widetilde{w}(x) = \widetilde{\pi}/\widetilde{q}$

$$\mathbb{E}_q \left[\widetilde{w}(X_n)(\varphi(X_i) - I)\right] = 0.$$

Since $\mathbb{V}_q(\varphi(X)w(X)) < \infty$ by standard CLT

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{w}(X_i) \left[\varphi(X_i) - I\right] \Rightarrow \mathcal{N}\Big(0, \mathbb{V}_q\Big(\widetilde{w}(X_1)[\varphi(X_1) - I]\Big)\Big).$$

Proof ctd...

Proof.

The strong law of large numbers applied to the denominator

$$\frac{1}{n}\sum_{i=1}^{n}\widetilde{w}(X_i) \to \mathbb{E}_q[\widetilde{w}(X_1)] = Z_\pi/Z_q, \quad \text{a.s.}$$

By Slutsky's theorem, combining the two

$$\sqrt{n}(\widehat{I}_n^{\mathrm{NIS}} - I) \Rightarrow \mathcal{N}\left(0, \mathbb{V}_q\left(\widetilde{w}(X_1)[\varphi(X_1) - I]\right)\frac{Z_q^2}{Z_\pi^2}\right)$$
$$\sim \mathcal{N}\left(0, \sigma^2_{\mathrm{NIS}}\right).$$

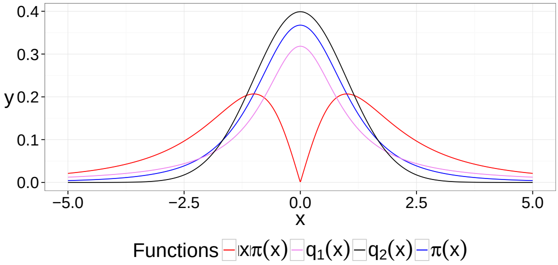
□

Alternatively, use Delta method.

Toy Example: t-distribution

- We want to compute $I = \mathbb{E}_{\pi}(|X|)$ where $\pi(x) \propto \left(1 + x^2/3\right)^{-2}$ (t_3-distribution).

1. Directly sample from π.
2. Use $q_1(x) = g_{t_1}(x) \propto \left(1 + x^2\right)^{-1}$ (t_1-distribution).
3. Use $q_2(x) \propto \exp\left(-x^2/2\right)$ (normal).



Toy Example: t-distribution

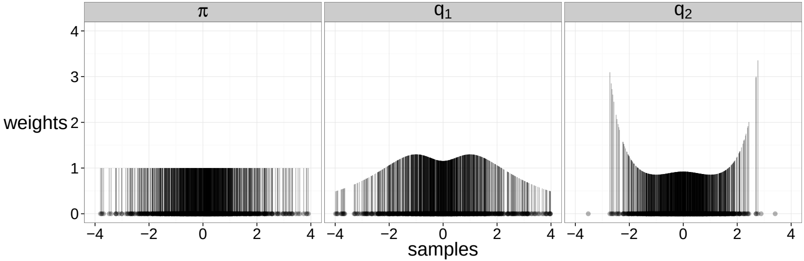


Figure: Sample weights obtained for 1000 realisations of X_i, from the different proposal distributions.

Toy Example: t-distribution

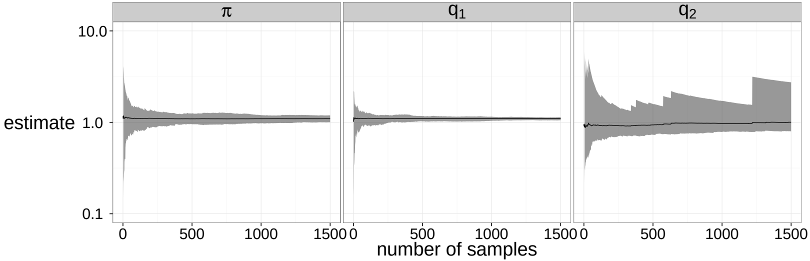


Figure: Estimates $\widehat{I_n}$ of I obtained after 1 to 1500 samples. The grey shaded areas correpond to the range of 100 independent replications.

Variance of importance sampling estimators

- Standard Importance Sampling: $X_1, \ldots, X_n \overset{iid}{\sim} q$,

$$\widehat{I}_n^{\mathrm{IS}} = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i) w(X_i).$$

- Asymptotic Variance:

$$\begin{aligned}\mathbb{V}_{as}\left(\widehat{I}_n^{\mathrm{IS}}\right) &= \mathbb{E}_q\left[\left(\varphi(X)w(X) - \mathbb{E}_q\left(\varphi(X)w(X)\right)\right)^2\right] \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\left(\varphi(X_i)w(X_i) - \widehat{I}_n^{\mathrm{IS}}\right)^2.\end{aligned}$$

- Thus the asymptotic variance can be estimated consistently with

$$\frac{1}{n}\sum_{i=1}^{n}\left(\varphi(X_i)w(X_i) - \widehat{I}_n^{\mathrm{IS}}\right)^2.$$

Variance of importance sampling estimators

- Normalised Importance Sampling: $X_1, \ldots, X_n \overset{iid}{\sim} q$,

$$\widehat{I}_n^{\text{NIS}} = \frac{\sum_{i=1}^n \varphi(X_i)\widetilde{w}(X_i)}{\sum_{i=1}^n \widetilde{w}(X_i)}.$$

- Asymptotic Variance:

$$\mathbb{V}_{as}\left(\widehat{I}_n^{\text{NIS}}\right) = \frac{\mathbb{E}_q\left[\left(\varphi(X)w(X) - I \times w(X)\right)^2\right]}{\mathbb{E}_q\left[w(X)\right]^2}.$$

- Thus the asymptotic variance can be estimated consistently with

$$\frac{\frac{1}{n}\sum_{i=1}^N \widetilde{w}(X_i)^2\left(\varphi(X_i) - \widehat{I}_n^{\text{NIS}}\right)^2}{\left(\frac{1}{n}\sum_{i=1}^N \widetilde{w}(X_i)\right)^2}.$$

Diagnostics

- Importance sampling works well when all weights roughly equal.
- If dominated by one $\widetilde{w}(X_j)$,

$$\widehat{I}_n^{\mathrm{NIS}} = \frac{\sum_{i=1}^n \varphi(X_i)\widetilde{w}(X_i)}{\sum_{i=1}^n \widetilde{w}(X_i)} \approx \widetilde{w}(X_j)\varphi(X_j).$$

  The "effective sample size" is one.
- To how many unweighted samples correspond our weighted samples of size n? Solve for n_e in

$$\frac{1}{n}\mathbb{V}_{as}\left(\widehat{I}_n^{\mathrm{NIS}}\right) = \frac{\sigma^2}{n_e},$$

  where σ^2/n_e corresponds to the variance of an unweighted sample of size n_e.

Diagnostics

- We solve by matching $\varphi(X_i) - \widehat{I}^{\text{NIS}}$ with $\varphi(X_i) - I \approx \sigma$ as if they were i.i.d samples:

$$\frac{1}{n}\frac{\frac{1}{n}\sum_{i=1}^{N}\widetilde{w}(X_i)^2\left(\varphi(X_i)-\widehat{I}_n^{\text{NIS}}\right)^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{w}(X_i)\right)^2} \approx \frac{\sigma^2}{n_e}$$

$$\text{i.e.} \qquad \frac{1}{n}\frac{\frac{1}{n}\sum_{i=1}^{N}\widetilde{w}(X_i)^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{w}(X_i)\right)^2} = \frac{1}{n_e}.$$

- The solution is

$$n_e = \frac{\left(\sum_{i=1}^{n}\widetilde{w}(X_i)\right)^2}{\sum_{i=1}^{n}\widetilde{w}(X_i)^2},$$

and is called the effective sample size.

Rejection and Importance Sampling in High Dimensions

- **Toy example:** Let $\mathbb{X} = \mathbb{R}^d$ and

$$\pi(x) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\sum_{i=1}^{d} x_i^2}{2}\right)$$

and

$$q(x) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\sum_{i=1}^{d} x_i^2}{2\sigma^2}\right).$$

- How do Rejection sampling and Importance sampling scale in this context?

Performance of Rejection Sampling

- We have

$$w(x) = \frac{\pi(x)}{q(x)} = \sigma^d \exp\left(-\frac{\sum_{i=1}^d x_i^2}{2}\left(1 - \frac{1}{\sigma^2}\right)\right) \leq \sigma^d$$

  for $\sigma > 1$.
- Acceptance probability is

$$\mathbb{P}(X \text{ accepted}) = \frac{1}{\sigma^d} \to 0 \text{ as } d \to \infty,$$

  i.e. exponential degradation of performance.
- For $d = 100$, $\sigma = 1.2$, we have

$$\mathbb{P}(X \text{ accepted}) \approx 1.2 \times 10^{-8}.$$

Performance of Importance Sampling

- We have

$$w(x) = \sigma^d \exp\left(-\frac{\sum_{i=1}^d x_i^2}{2}\left(1-\frac{1}{\sigma^2}\right)\right).$$

- Variance of the weights:

$$\mathbb{V}_q[w(X)] = \left(\frac{\sigma^4}{2\sigma^2-1}\right)^{d/2} - 1$$

where $\sigma^4/\left(2\sigma^2-1\right) > 1$ for any $\sigma^2 > 1/2$.

- For $d = 100$, $\sigma = 1.2$, we have

$$\mathbb{V}_q[w(X)] \approx 1.8 \times 10^4.$$

Wait a minute…

Lecture 1:

- Simpson's rule for approximating integrals: error in $\mathcal{O}(n^{-1/d})$.

Lecture 2:

- Monte Carlo for approximating integrals: error in $\mathcal{O}(n^{-1/2})$ with rate independent of d.

And now:

- Importance Sampling standard deviation in the Gaussian example in $\exp(d)n^{-1/2}$.

The rate is indeed independent of d but the "constant" (in n) explodes exponentially (in d).

Markov chain Monte Carlo

- Revolutionary idea introduced by Metropolis et al., J. Chemical Physics, 1953.
- **Key idea**: Given a target distribution π, build a Markov chain $(X_t)_{t\geq 1}$ such that, as $t \to \infty$, $X_t \sim \pi$ and

$$\frac{1}{n}\sum_{t=1}^{n} \varphi(X_t) \to \int \varphi(x)\, \pi(x)\, dx$$

  when $n \to \infty$ e.g. almost surely.
- Central limit theorems with a rate in $1/\sqrt{n}$.
- In some cases the constant (in n) does not explode exponentially with the dimension d, but polynomially.

Side Dish: Control Variates

- Variance reduction techniques, not always applicable but useful in some cases.
- Suppose that we want to compute

$$I = \int \varphi(x)\pi(x)dx$$

and that we know exactly

$$J = \int \psi(x)\pi(x)dx.$$

- Sample $X_1, \ldots, X_n$ from π and compute

$$\widehat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \left(\varphi(X_i) - \lambda(\psi(X_i) - J)\right).$$

- What is the benefit of $\widehat{I}_n$ over the standard Monte Carlo estimator?