CS395T: Foundations of Machine Learning for Systems Researchers

Fall 2025

Lecture 5: Monte Carlo Methods for Estimating Expectations and Gradients of Expectations in Reinforcement Learning

Monte Carlo method

Invented by John von Neumann and Stanislaw Ulam for solving neutron transport in materials

Key idea: use random sampling to estimate deterministic quantities that are hard to compute exactly, such as definite integrals

Key concern: variance reduction. How to reduce number of samples to reach given level of accuracy probabilistically

Terminology

Estimand: quantity whose value we want to know
- (e.g.) probability of head on coin toss (*P(h)*)

Estimate: approximation for estimand based on random observations (samples)
- (e.g.) *P(h) = 0.4*

Estimator: rule for computing estimate from samples (e.g., n tosses, h heads $\Rightarrow$ $P(h) \approx h/n$)
- Observations (samples) are random variables so estimates produced by an estimator are also random variables, which have a mean and variance

Sample variance: how different are estimates produced by an estimator for different sets of observations of same size
- Higher variance $\Rightarrow$ need more samples to obtain given level of accuracy in estimate

Unbiased estimator: mean of estimates from estimator = true value *at any given sample size*
Biased estimator: mean of estimates = true value + bias

Applications of sampling

$$E_{x\sim p(x;\underline{\theta})}[f(x)] = \int_x p(x;\underline{\theta}) f(x) dx$$
$$\nabla_{\underline{\theta}} E_{x\sim p(x;\underline{\theta})}[f(x)] = \nabla_{\underline{\theta}} \int_x p(x;\underline{\theta}) f(x) dx$$

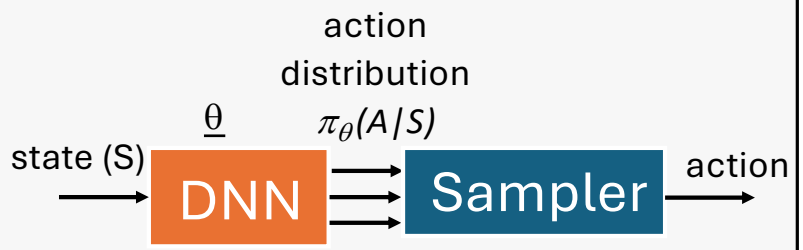
Estimate probabilities: frequency ≈ probability

Estimate expectations (definite integrals)

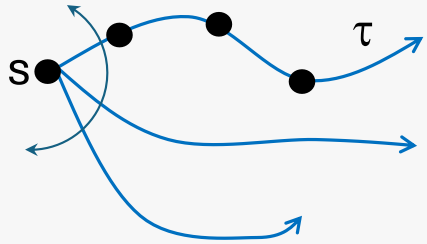
- $\underline{\theta}$ is a vector of parameters
- Evaluate definite integral for fixed parameter values
- Example: policy evaluation in RL

Estimate gradients of expectations

- How does expectation change when $\underline{\theta}$ is changed?
- Gradient of expectation is not an expectation
- Example: policy improvement in RL

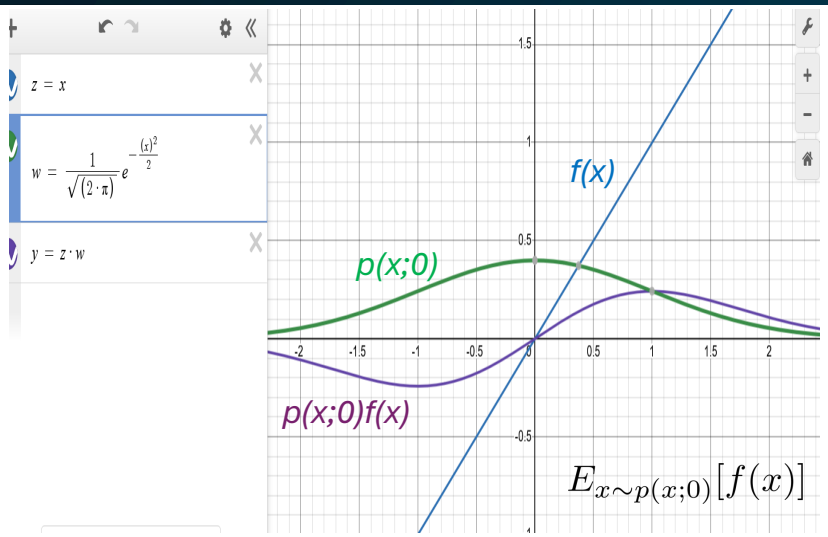


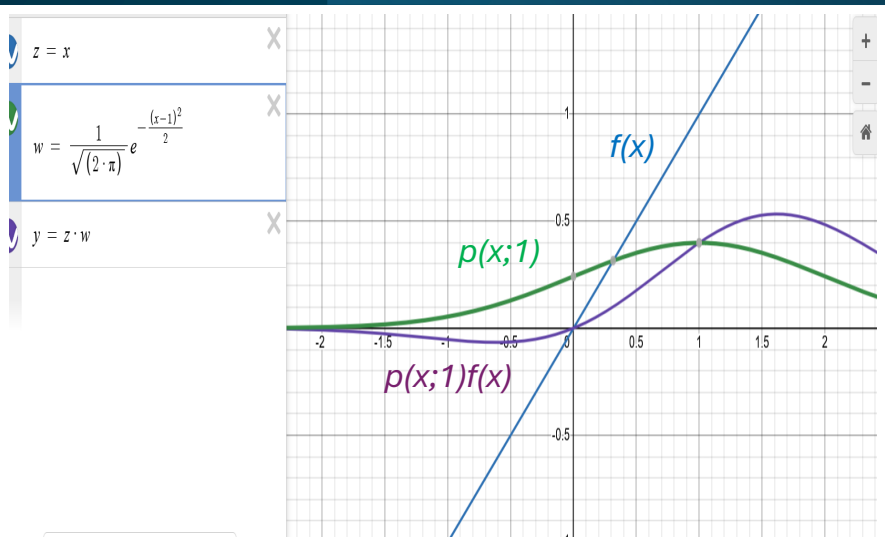
Deep RL Policy Network



$$J(s;\underline{\theta}) = E_{\tau\sim p(\tau;\underline{\theta})}[R(\tau)]$$
$$\nabla_{\underline{\theta}} J(s;\underline{\theta}) = \nabla_{\underline{\theta}} E_{\tau\sim p(\tau;\underline{\theta})}[R(\tau)]$$

Visualizing how expectation changes when parameter changes





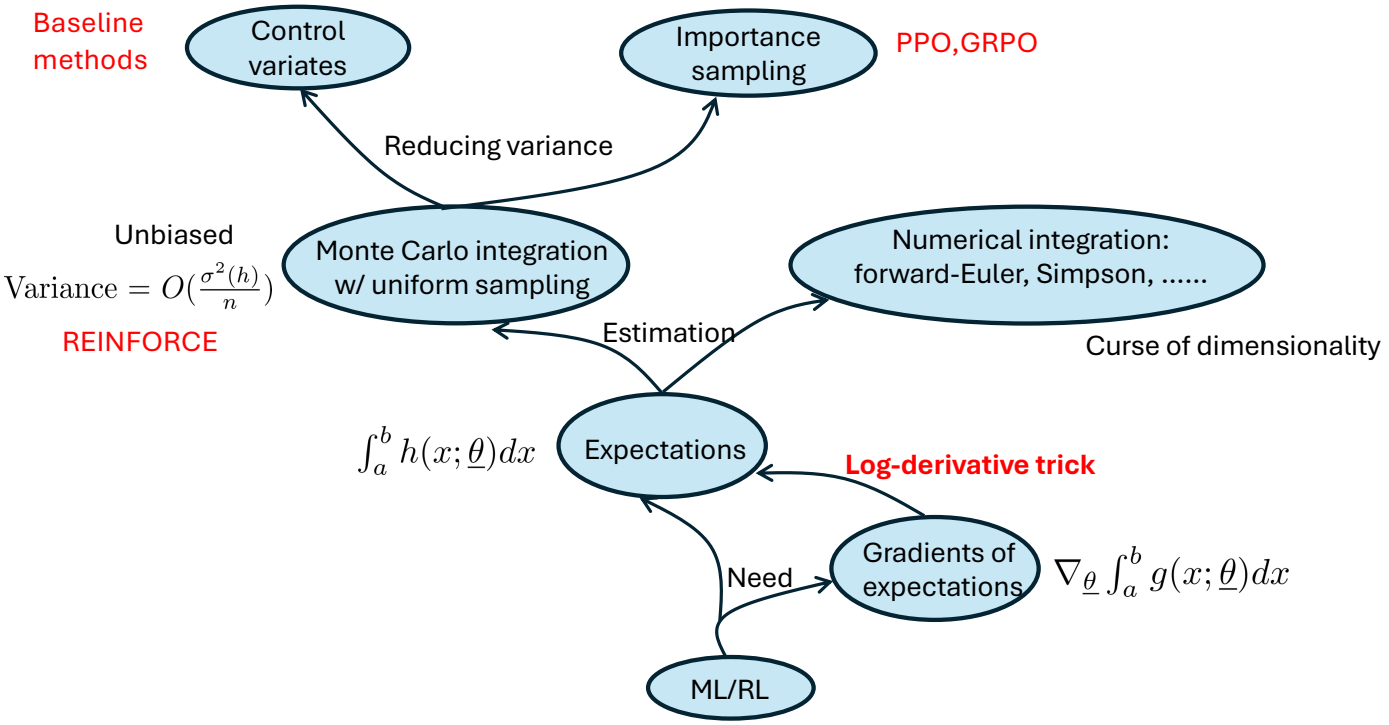
$f(x) = x$
$p(x;\mu) = \mathcal{N}(x;\mu,1)$
$E_{x\sim p(x;\mu)}[f(x)] = \int_{-\infty}^{\infty} x * p(x;\mu)dx$
Derivative wrt μ: How does $E_{x\sim p(x;\mu)}[f(x)]$ change as we change μ?
Visually:

Sign of expectation = sign of μ

Derivative of expectation: positive

$$\mathcal{N}(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

Organization



Log-derivative trick (I)

$$\boxed{\nabla_{\underline{\theta}} E_{x\sim p(x;\underline{\theta})}[f(x)] = E_{x\sim p(x;\underline{\theta})}[f(x)\nabla_{\underline{\theta}} log(p(x;\underline{\theta}))]}$$

Trick is based on two identities:

- **Leibniz's rule**: assuming bounds of integral do not depend on θ

$$\boxed{\frac{d}{d\theta}\int_x g(x;\theta)dx = \int_x \frac{d}{d\theta} g(x;\theta)dx}$$

$$\begin{aligned}
\frac{d}{d\theta}\int_x g(x;\theta)dx &= \lim_{\Delta\theta\to 0} \frac{\int_x g(x;\theta+\Delta\theta)dx \quad - \quad \int_x g(x;\theta)dx}{\Delta\theta} \\
&= \lim_{\Delta\theta\to 0} \frac{\int_x [g(x;\theta) + \Delta\theta * \frac{d}{d\theta} g(x;\theta) + O(\Delta\theta)^2]dx \quad - \quad \int_x g(x;\theta)dx}{\Delta\theta} \\
&= \int_x \frac{d}{d\theta} g(x;\theta)dx
\end{aligned}$$

- **Expectate rule**: for any function $h(x)$ and distribution $p(x)$ s.t. $h(x)\neq 0 \implies p(x)\neq 0$

$$\boxed{\int_x h(x)dx = \int_x p(x)\frac{h(x)}{p(x)}dx = E_{x\sim p(x)}\left[\frac{h(x)}{p(x)}\right]}$$

Log-derivative trick (II)

Single parameter θ

$$\frac{d}{d\theta}E_{x\sim p(x;\theta)}[f(x)] = \frac{d}{d\theta}\int_x p(x;\theta)f(x)dx \quad \text{(Definition of expectation)}$$

$$= \int_x f(x)\frac{d}{d\theta}p(x;\theta)dx \quad \text{(Leibniz rule)}$$

If we cannot evaluate integral analytically, rewrite as expectation and use Monte Carlo sampling

$$= E_{x\sim q(x;\theta)}\left[\frac{f(x)\frac{d}{d\theta}p(x;\theta)}{q(x;\theta)}\right] \quad \text{(Expectate rule)}$$

$$= E_{x\sim p(x;\theta)}[f(x)\frac{d}{d\theta}log(p(x;\theta))] \quad \text{(if } q{=}p\text{)}$$

Multiple parameters $\underline{\theta}$

$$\boxed{\nabla_{\underline{\theta}}E_{x\sim p(x;\underline{\theta})}[f(x)] = E_{x\sim p(x;\underline{\theta})}[f(x)\nabla_{\underline{\theta}}log(p(x;\underline{\theta}))]}$$

Check

$$\frac{d}{d\mu}(E_{x\sim p(x;\mu)}[x]) = \frac{d}{d\mu}(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-\frac{(x-\mu)^2}{2}}dx) \stackrel{?}{=} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x*\frac{d}{d\mu}(e^{-\frac{(x-\mu)^2}{2}})dx$$

$$(i) E_{x\sim p(x;\mu)}[x] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-\frac{(x-\mu)^2}{2}}dx = \mu \implies \frac{d}{d\mu}E_{x\sim p(x;\mu)}[x] = 1$$

$$(ii) \int_{-\infty}^{\infty} x * \frac{d}{d\mu}(\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2}})dx = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}} * (x-\mu)e^{-\frac{(x-\mu)^2}{2}}dx = 1$$

Definite Integrals Associated with Gaussian Distributions

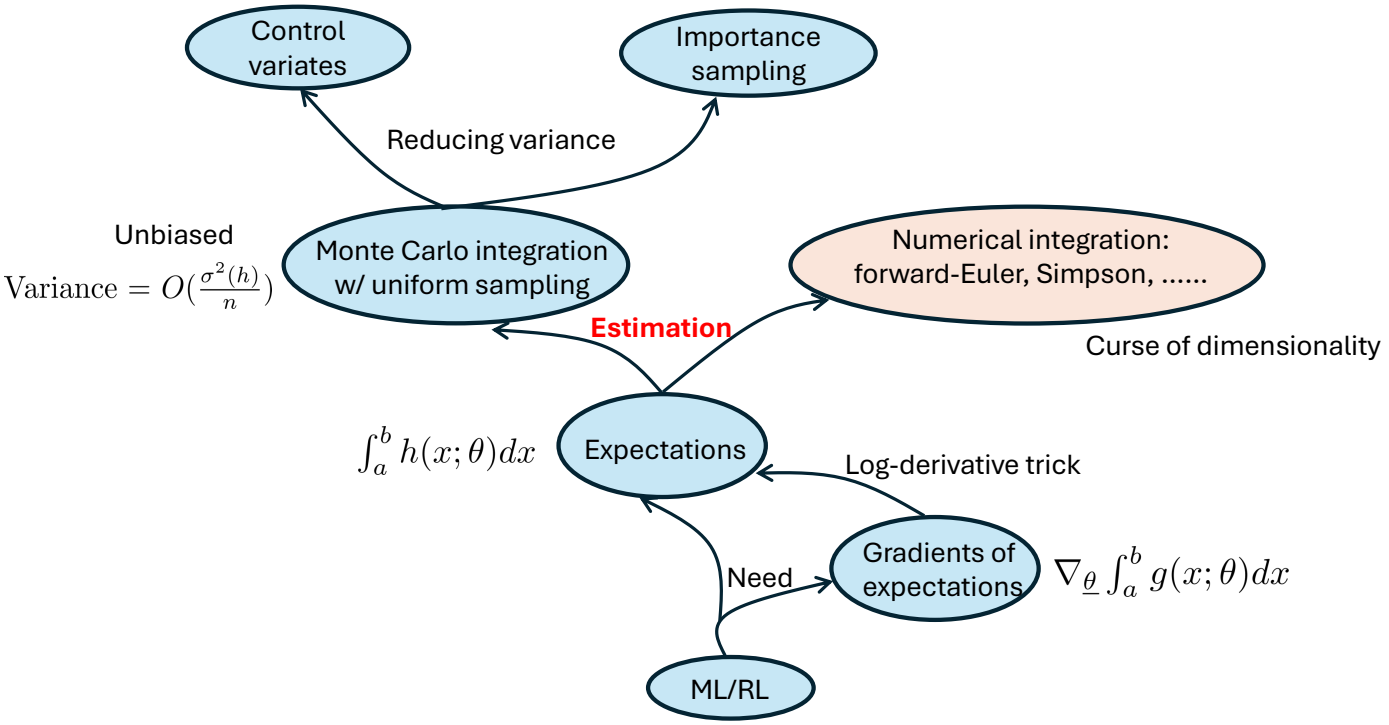
In physical systems which can be modeled by a Gaussian distribution, one sometimes needs to obtain the average or expectation value for physical quantities. If these properties depend on x, then they can be integrated to find the average value. For the first five powers of x, the integrals have the following forms:

$$\int_0^\infty e^{-x^2}dx = \frac{\sqrt{\pi}}{2} \qquad \int_0^\infty x^3e^{-x^2}dx = \frac{1}{2}$$

$$\int_0^\infty xe^{-x^2}dx = \frac{1}{2} \qquad \int_0^\infty x^4e^{-x^2}dx = \frac{3\sqrt{\pi}}{8}$$

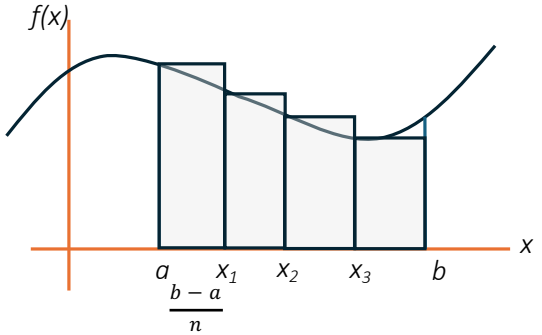
$$\int_0^\infty x^2e^{-x^2}dx = \frac{\sqrt{\pi}}{4} \qquad \int_0^\infty x^5e^{-x^2}dx = 1$$

Organization



Numerical integration (I)

$$A(f) = \int_a^b f(x)dx$$



- Divide interval [*a,b*] into n intervals (usually fixed size)
- Use a quadrature formula to estimate integral
  - Forward-Euler (FE), backward-Euler (BE), trapezoidal, Simpson's rule,....

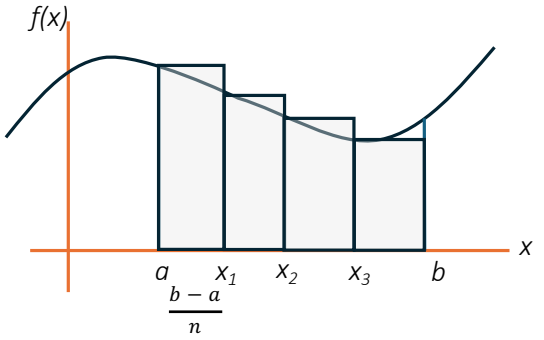
$$\hat{A}_{FE}(f;n) = \sum_{i=0}^{n-1} \frac{(b-a)}{n} * f(x_i) = \frac{(b-a)}{n} * \sum_{i=0}^{n-1} f(x_i) \text{ where } x_i = a + \frac{b-a}{n} * i$$

$$\hat{A}_{BE}(f;n) = \frac{(b-a)}{n} * \sum_{i=1}^{n} f(x_i) \text{ where } x_i = a + \frac{b-a}{n} * i$$

Convergence of forward-Euler

$$A(f) = \int_a^b f(x)dx$$

$$\hat{A}_{FE}(f;n) = \frac{(b-a)}{n} * \sum_{i=0}^{n-1} f(x_i) \text{ where } x_i = a + \frac{b-a}{n} * i$$



$$\boxed{\left|A(f) - \hat{A}_{FE}(f;n)\right| \leq \frac{m(b-a)^2}{2n} \text{ where } m \text{ is max value of } |f'(x)| \text{ in [a,b]}}$$

$$\boxed{\lim_{n\to\infty} \hat{A}_{FE}(f;n) = \int_a^b f(x)dx}$$

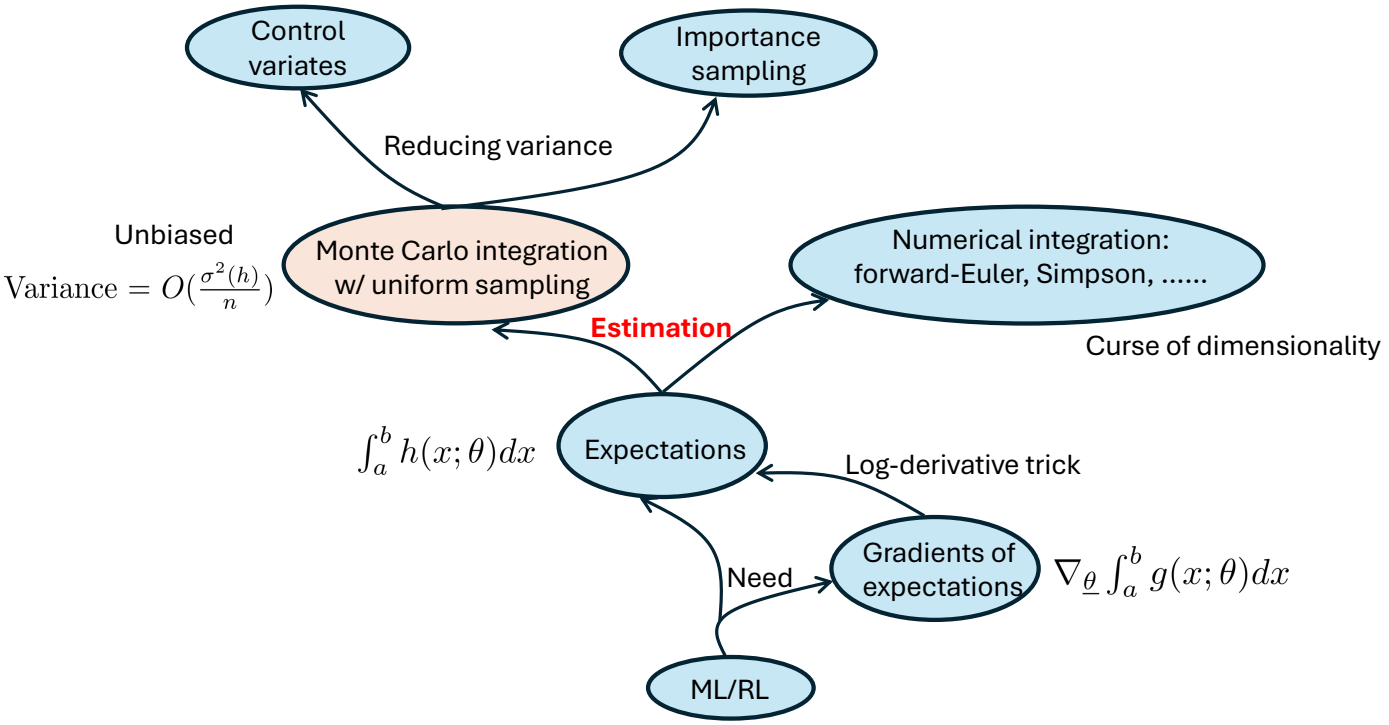
Other quadrature formulas converge faster (Simpson's rule error: $O(\frac{1}{n^4})$)

Bernhard Riemann
(1826-1866)

Drawback of numerical integration(II)

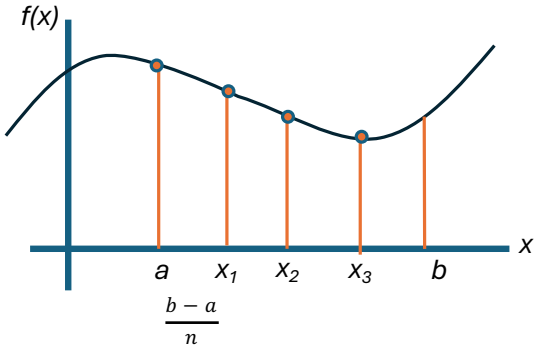
- Works well in 1D and 2D
- Curse of dimensionality: to achieve given accuracy, number of function evaluations grow as $O(2^d)$ for d dimensions
- Not used for high-dimensional integrals

Organization



Monte Carlo integration w/ uniform sampling

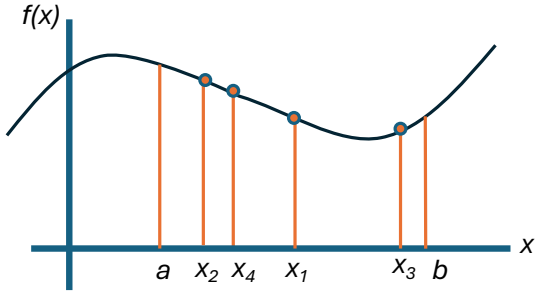
$$\hat{A}_{FE}(f;n) = (b-a) * \underbrace{\frac{1}{n}\sum_{i=0}^{n-1} f(x_i)}_{average\ of\ f(x_i)\ values} \quad \text{where } x_i = a + \frac{b-a}{n} * i$$



Reinterpret forward-Euler/backward-Euler

- Generate x_i values using arithmetic progression
- Take average of $f(x_i)$ values and multiply by $(b-a)$

$$\hat{A}_{MC}(f;n) = (b-a) * \underbrace{\frac{1}{n}\sum_{i=0}^{n-1} f(X_i)}_{average\ of\ f(X_i)\ values} \quad \text{where } X_i \sim U(a,b)$$

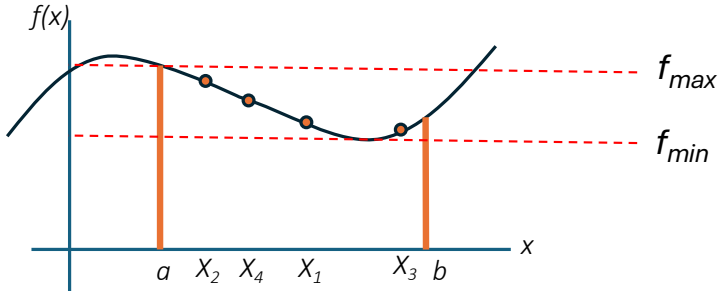


Monte Carlo integration (capital letters for random variables)

- Generate X_i values by sampling uniform distribution U(a,b)
- X_i's: *independent identically distributed (i.i.d.)* random variables
- Take average of f(X_i) values and multiply by (b-a)

Correctness of Monte Carlo integration (I)

$$\hat{A}_{MC}(f;n) = (b-a) * \underbrace{\frac{1}{n}\sum_{i=0}^{n-1} f(X_i)}_{average\ of\ f(X_i)\ values} \quad \text{where } X_i \sim U(a,b)$$



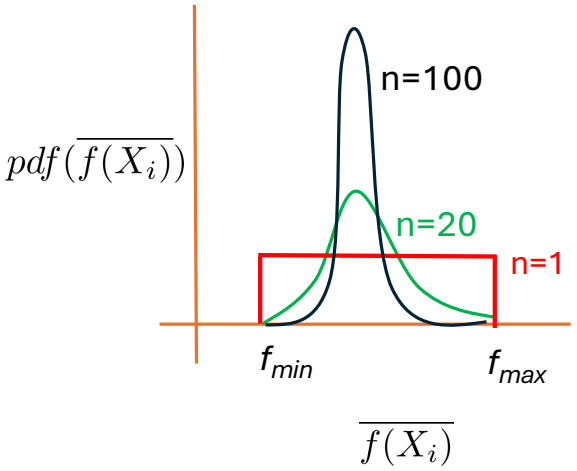
$\hat{A}_{MC}(f;n)$ is random variable

What does it mean for random variable to approximate a deterministic quantity? Look at its mean and variance.

Intuition

- Let f_{min} and f_{max} be minimum and maximum value of f in interval $[a,b]$
- Denote $\frac{1}{n}\sum_{i=0}^{n-1} f(X_i)$ by $\overline{f(X_i)}$

$\overline{f(X_i)} \in [f_{min}, f_{max}]$



https://www.scratchapixel.com/images/monte-carlo-methods/beanmachine.m4v

Correctness of Monte Carlo integration (II)

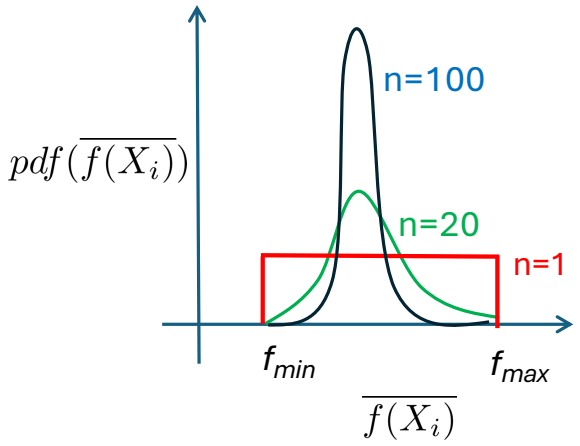
$$\hat{A}_{MC}(f;n) = (b-a) * \underbrace{\frac{1}{n}\sum_{i=0}^{n-1} f(X_i)}_{average\ of\ f(X_i)\ values} \quad \text{where } X_i \sim U(a,b)$$

Unbiased estimator:

$$E_{X_i \sim U(a,b)}\left[\hat{A}_{MC}(f;n)\right] = \frac{b-a}{n}\sum_{i=0}^{n-1}\int_a^b \frac{1}{b-a} f(X_i)dX_i = \int_a^b f(x)dx$$

$$\sigma_U^2(\hat{A}_{MC}(f;n)) = \frac{(b-a)^2}{n}\sigma_U^2(f) = O(\frac{\sigma_U^2(f)}{n})$$

$$\lim_{n\to\infty} \hat{A}_{MC}(f;n) = \int_a^b f(x)dx \quad \text{(Law of large numbers)}$$



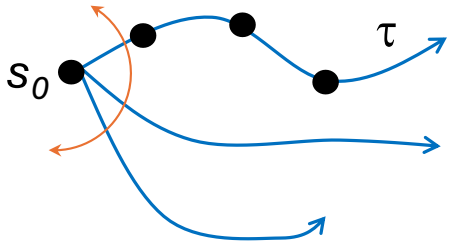
Convergence in probability

Monte Carlo estimation in deep RL: REINFORCE

REINFORCE algorithm

- Gradient ascent to optimize $\underline{\theta}$ using Monte Carlo estimates
- Sample trajectories out of s_0 to estimate $\nabla_{\underline{\theta}} J(s_0;\underline{\theta})$

Variance of estimator can grow *exponentially* with number of steps (T) in trajectories



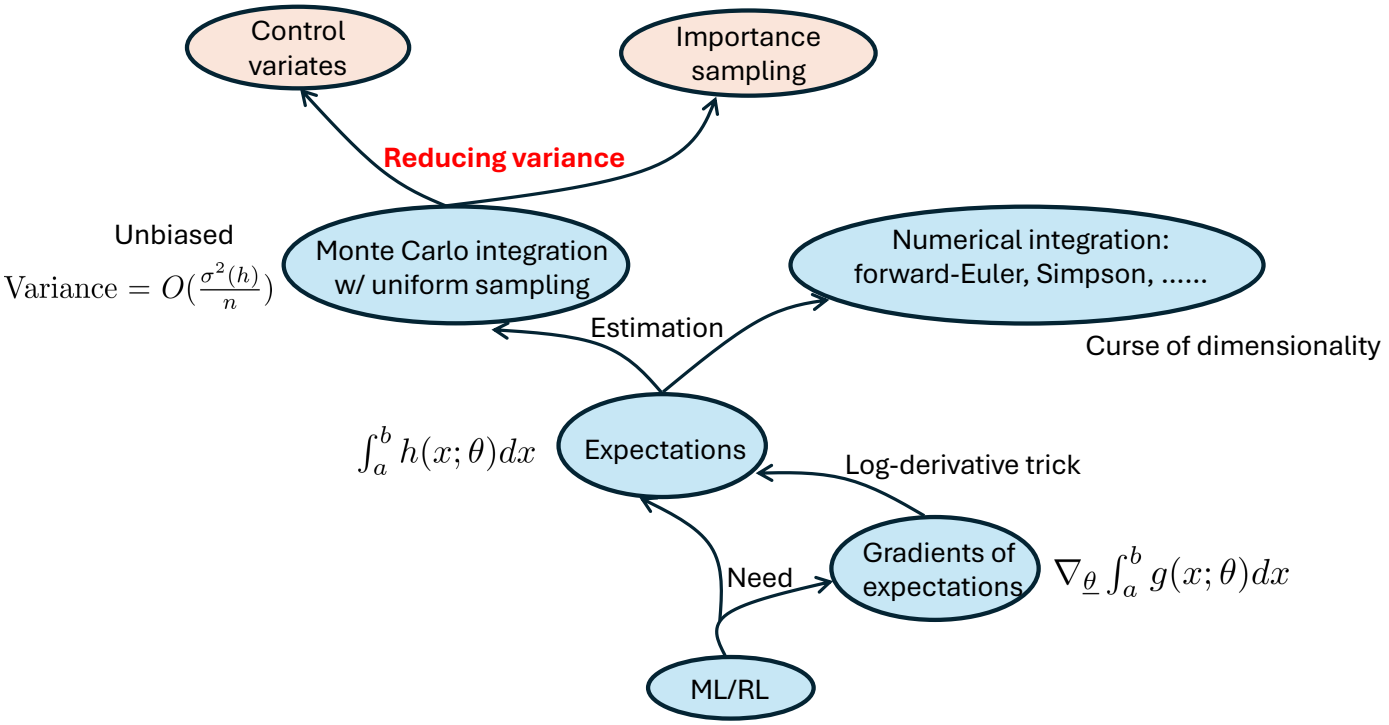
$$J(s;\underline{\theta}) = E_{\tau\sim p(\tau;\underline{\theta})}[R(\tau)] = \int_{\tau} p(\tau;\underline{\theta})R(\tau)d\tau$$

$$p(\tau;\underline{\theta}) = \prod_{i=0}^{T-1} \underbrace{\pi_{\underline{\theta}}(A_i|s_i)}_{policy\ network} * \underbrace{P(s_{i+1}|s_i,A_i)}_{environment}$$

n independent random variables X_i with means μ_i and variances σ_i

$$\sigma^2(X_1X_2...X_n) = \prod_{i=1}^{n}(\sigma_i^2+\mu_i^2) - \prod_{i=1}^{n}\mu_i^2 \quad \geq \prod_{i=1}^{n}\sigma_i^2$$

Organization



Reducing variance of Monte Carlo samples

$$\sigma_U^2(A_{MC}(f;n)) = O(\frac{\sigma_U^2(f)}{n})$$

Increase number of samples

- Expensive if obtaining samples is expensive

Stratified sampling: control distribution of samples

- Random may not be best because of *sample clumping* (birthday paradox)

Change function being integrated (*f*) to another function (*g*)

- If variance of *g* < variance of *f,* fewer samples needed
- Calculate estimate for integral of *f* from estimate for integral of *g*
- Many approaches
  - Importance sampling
  - Control variates
  - Antithetic variates
  - ……

Reducing sample clumping by stratified sampling

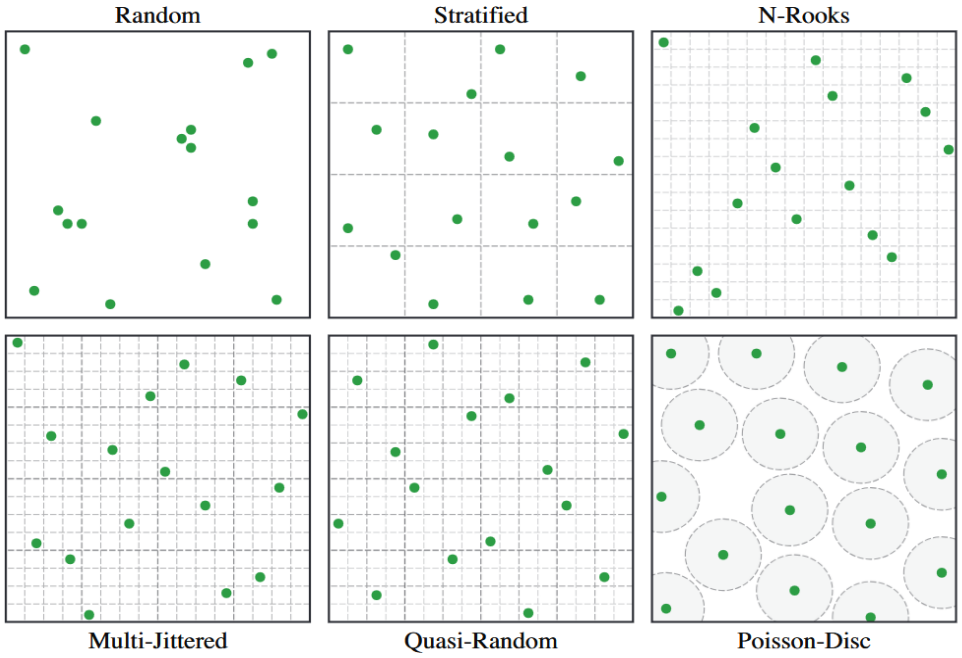


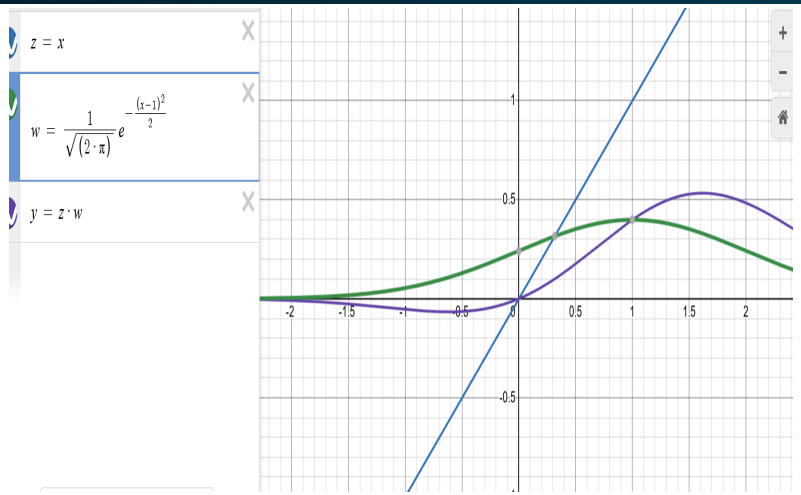
Figure A.4: An illustration comparing several 2D sampling approaches each using 16 samples. Purely random sampling (top left) can suffer from clumping, which increases variance by undersampling other regions of the integrand. All of the other approaches illustrated try to minimize this clumping to reduce variance.

Another approach: locality-sensitive hashing (LSH)

Stratified sampling useful for spatial domains but perhaps not for policy spaces

https://cs.dartmouth.edu/~wjarosz/publications/dissertation/appendixA.pdf

Importance sampling: intuitive idea



$$E_{x\sim\mathcal{N}(1,1)}\ [x]$$

Expectation = area under purple curve

One possibility: uniform sampling in some large interval like [-1000,1000]

- Most samples will contribute little to overall sum

Better idea: sample only in *important* regions where $|h(x)| = |p(x)f(x)| >> 0$

- One possibility: uniform sampling in some interval like [-2,10]
- Even better: skew samples to positive x values
- Even better: sample from distributions whose "shape" is close to that of $|h(x)|$

Importance sampling

$$
\begin{aligned}
E_{x\sim p(x)}[f(x)] &= \int_x p(x)f(x)dx = \int_x q(x)\frac{p(x)f(x)}{q(x)}dx \quad \text{(Expectate rule)} \\
&= E_{x\sim q(x)}\left[\frac{p(x)}{q(x)}f(x)\right] \\
&\approx \frac{1}{n}\sum_{i=0}^{n-1}\frac{p(X_i)}{q(X_i)}f(X_i) \quad \text{where } X_i \sim q(X_i)
\end{aligned}
$$

- Compute estimate for one expectation using a different distribution
  - *p(x)*: nominal/target distribution
  - *q(x)*: importance/proposal distribution
  - $\frac{p(x)}{q(x)}$: likelihood ratio

$$
\boxed{\text{If } \sigma_q^2\left(\frac{p(x)}{q(x)}f(x)\right) < \sigma_p^2(f(x)), \text{ sample efficiency is improved.}}
$$

Importance sampling in deep RL

Improve sample efficiency by reusing samples when $\underline{\theta}$ is updated to $\underline{\theta^1}$

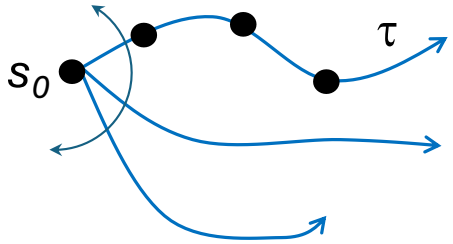
Trajectories collected under $\underline{\theta}$ are valid trajectories for $\underline{\theta^1}$ but transition probabilities will be different in general

If $\underline{\theta}$ and $\underline{\theta^1}$ are close to each other, samples collected for $\underline{\theta}$ will be representative for $\underline{\theta^1}$ well

- Ensuring this in PPO/GRPO: clipping, KL-divergence check

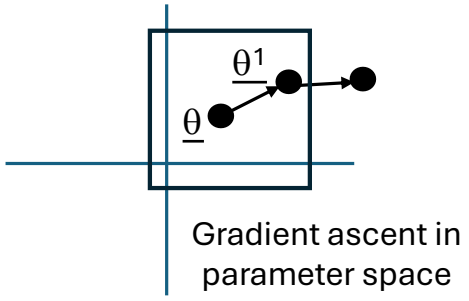
$$J(s;\underline{\theta^1}) = E_{\tau\sim p(\tau;\underline{\theta^1})}[R(\tau)] = \int_\tau p(\tau;\underline{\theta})\frac{p(\tau;\underline{\theta^1})}{p(\tau;\underline{\theta})}R(\tau)d\tau$$

$$= E_{\tau\sim p(\tau;\underline{\theta})}\left[\frac{p(\tau;\underline{\theta^1})}{p(\tau;\underline{\theta})}R(\tau)\right]$$



$$J(s;\underline{\theta}) = E_{\tau\sim p(\tau;\underline{\theta})}[R(\tau)] = \int_\tau p(\tau;\underline{\theta})R(\tau)d\tau$$

$$p(\tau;\underline{\theta}) = \prod_{i=0}^{T-1} \underbrace{\pi_{\underline{\theta}}(A_i|s_i)}_{policy\ network} * \underbrace{P(s_{i+1}|s_i,A_i)}_{environment}$$



Control variates in Deep RL: baseline methods, advantage

$$h(x) = \underbrace{f(x)}_{\text{function of interest}} - \underbrace{g(x)}_{\text{easy to compute}} + \underbrace{\mathbb{E}[g(x)]}_{\text{easy to compute}}$$

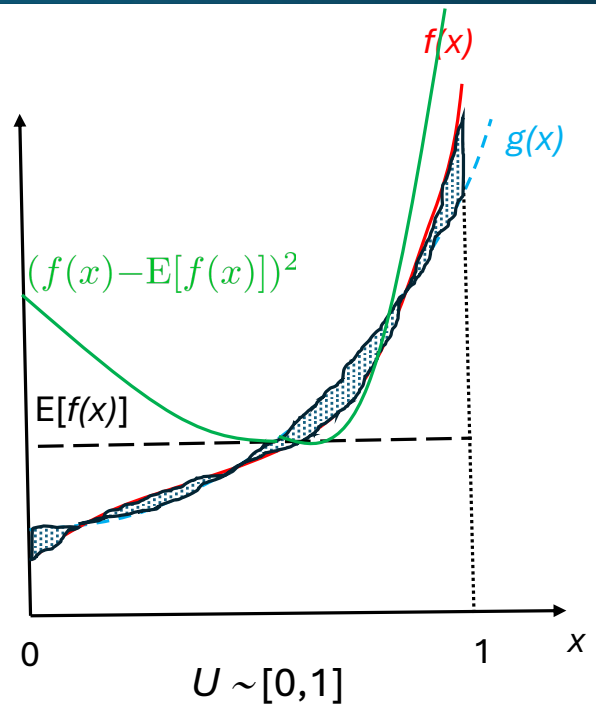
Distribution for expectations and variances: $U \sim [0,1]$
$\mathbb{E}[g(x)]$ should be a good approximation for $\mathbb{E}[f(x)]$
Special case of *control variates* in Monte Carlo literature

Intuition:
(i) $\mathbb{E}[h(x)] = \mathbb{E}[f(x)]$
(ii) $\mathbb{E}[g(x)]$ is computed analytically
(ii) $\mathbb{E}[f(x) - g(x)]$ provides correction to $\mathbb{E}[g(x)]$ and is estimated by sampling

$$\mathbb{E}[f(x)] = \mathbb{E}[h(x)] \approx \left(\frac{1}{n} \sum_{i=1}^{n} f(x_i) - g(x_i) \right) + \mathbb{E}[g(x)]$$

Variance: $\sigma^2(h(x)) = \sigma^2(f(x)) + \sigma^2(g(x)) - 2 * Cov(f(x), g(x))$
Win if $Cov(f(x), g(x)) > \frac{1}{2}\sigma^2(g(x))$ ($f(x)$, $g(x)$ are strongly correlated)

Note: variance can *increase* if *g(x)* is badly chosen!



g(x) is a baseline for *f(x)*

https://en.wikipedia.org/wiki/Control_variates

Control variates (general)

$$\underbrace{h(x)}_{\text{corrected function}} = \underbrace{f(x)}_{\text{function of interest}} + \underbrace{c}_{\text{any constant}} * (\underbrace{g(x)}_{\text{easy to compute}} - \underbrace{E[g(x)]}_{\text{known}})$$

$$\hat{E}[h(x)] = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) + c * (g(X_i) - E[g(x)])$$

Expectation: $E[h(x)] = E[f(x)] + c * (E[g(x) - E[g(x)]) = E[f(x)]$

So $\hat{E}[h(x)]$ is unbiased estimator for $E[f(x)]$

Variance: $\sigma^2(h(x)) = \sigma^2(f(x)) + c^2\sigma^2(g(x)) + 2c * Cov(f(x), g(x))$

Variance minimized when $c = -\frac{Cov(f(x),g(x))}{\sigma^2(g(x))} \implies$

$$\sigma^2(h(x)) = \sigma^2(f(x))\left[1 - \frac{Cov^2(f(x), g(x))}{\sigma^2(f(x))\sigma^2(g(x))}\right] = \sigma^2(f(x))(1 - \rho^2_{x,y})$$

Note that $\sigma^2(h(x) \leq \sigma^2(f(x))$. No win if zero correlation between f and g.

Pearson's correlation coefficient

$-1 \leq \rho_{x,y} \leq 1$

Control variates example (from Wikipedia)

We would like to estimate

$$I = \int_0^1 \frac{1}{1+x}\,\mathrm{d}x$$

using Monte Carlo integration. This integral is the expected value of $f(U)$, where

$$f(U) = \frac{1}{1+U}$$

and *U* follows a uniform distribution [0, 1]. Using a sample of size *n* denote the points in the sample as $u_1, \cdots, u_n$. Then the estimate is given by

$$I \approx \frac{1}{n}\sum_i f(u_i).$$

Now we introduce $g(U) = 1 + U$ as a control variate with a known expected value $\mathbb{E}\left[g\left(U\right)\right] = \int_0^1 (1+x)\,\mathrm{d}x = \frac{3}{2}$ and combine the two into a new estimate

$$I \approx \frac{1}{n}\sum_i f(u_i) + c\left(\frac{1}{n}\sum_i g(u_i) - 3/2\right).$$

Using $n = 1500$ realizations and an estimated optimal coefficient $c^\star \approx 0.4773$ we obtain the following results

	Estimate	Variance
Classical estimate	0.69475	0.01947
Control variates	0.69295	0.00060

https://en.wikipedia.org/wiki/Control_variates

Summary

Basic Monte Carlo estimation

- REINFORCE for policy optimization

Importance sampling

- Sample efficiency in PPO, TRPO, GRPO

Control variates for variance reduction

- Baseline methods, advantage

References

- Well-written introduction to Monte Carlo integration and variance reduction
  - https://cs.dartmouth.edu/~wjarosz/publications/dissertation/appendixA.pdf
- Using control variates for reducing variance in Monte Carlo rendering
  - https://cs.dartmouth.edu/~wjarosz/publications/rousselle16image.pdf