Today: Monte Carlo methods and sampling problems

Monte Carlo methods: computational algorithms that rely on repeated random sampling to obtain numerical results.

General strategy:

1. Define input domain
2. Generate samples from a probability distribution over the domain
3. Compute output for each input and aggregate the results

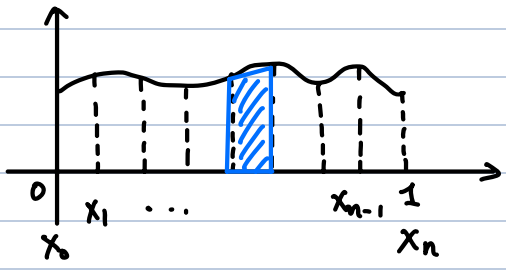
ex. Monte Carlo integration

$$I(f) = \int_0^1 f(x)\,dx$$

Numerical integration method:

Trapezoidal rule

$$I(f) \approx \sum_{i=1}^{n} \frac{1}{2}[f(x_i) + f(x_{i-1})]\,h$$



$$h = x_i - x_{i-1} = \frac{1}{n}$$

$$= \left[\frac{1}{2}f(x_0) + \sum_{i=1}^{n-1} f(x_i) + \frac{1}{2}f(x_n)\right] h$$

$$=: I_h(f)$$

$$|I(f) - I_h(f)| \le \frac{1}{12}h^2 \|f\|_\infty = O\left(\frac{1}{n^2}\right)$$

Monte Carlo integration view $I(f)$ as the mathematical expectation $I(f) = \mathbb{E} f(X)$

where X is uniformly distributed over $[0, 1]$.

According to the Law of Large Numbers, a natural estimator of $\mathbb{E}f(X)$ is

$$I(f) = \mathbb{E}f(X) \approx \frac{1}{n}\sum_{i=1}^{n} f(X_i) =: I_n(f)$$

where $X_1, \dots, X_n$ are identically independently distributed (i.i.d.) samples from $\mathrm{Unif}[0,1]$.

Actually we know $I_n(f)$ is unbiased:

$$\mathbb{E}\, I_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}f(X_i) = I(f)$$

and the error $e_n := |I_n(f) - I(f)|$ has variance

$$\mathbb{E}|e_n|^2 = \frac{1}{n^2}\,\mathbb{E}\left[\sum_{i=1}^{n}(f(X_i) - I(f))\right]^2$$

$$= \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbb{E}\left[(f(X_i) - I(f))(f(X_j) - I(f))\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left(f(X_i) - I(f)\right)^2$$

$$= \frac{1}{n}\,\mathbb{E}\left[(f(X) - I(f))^2\right] = \frac{1}{n}\,\mathrm{Var}(f)$$

$$\Rightarrow \mathbb{E}|e_n| \le \sqrt{\mathbb{E}|e_n|^2} \le \sqrt{\frac{\mathrm{Var}(f)}{n}} = O\left(\frac{1}{n^{1/2}}\right)$$

For general integral

$$\int f(x)\, p(x)\, dx$$

with $p(x)$ being a probability density function. We can still do

$$\int f(x)\, p(x)\, dx \approx \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

where $X_1, \dots, X_n$ are i.i.d. random variables sampled from $p(x)$

Remark: One can reduce the variance by importance sampling technique, i.e., take a probability density function $q(x)$, and $X_1, \dots, X_n \overset{iid}{\sim} q(x)$, and

$$\int f(x)\, p(x)\, dx = \int f(x) \frac{p(x)}{q(x)} q(x)\, dx \approx \frac{1}{n} \sum_{i=1}^{n} f(X_i) \frac{p(X_i)}{q(X_i)}$$

and $\mathbb{E}|e_n| \le \sqrt{\frac{\mathrm{Var}(f\frac{p}{q})}{n}}$

by appropriate choice of $q \approx f$, we should have $\mathrm{Var}(f\frac{p}{q}) < \mathrm{Var}(f)$

For high-dimensional integration, we can extend the trapezoidal rule to a grid with n^d grid points, and the approximation error is still $O(\frac{1}{n^2})$

But for Monte Carlo integration, with n samples, we still have $O(\frac{1}{n^{1/2}})$ error

So as soon as $d > 4$, the Monte Carlo method's computational cost is smaller than standard trapezoidal rule.

Key problem: given a probability density $p_*(x)$ over $\mathbb{R}^d$ how to sample $x_1, \dots, x_n \overset{iid}{\sim} p_*$

Method 1: Inverse transform

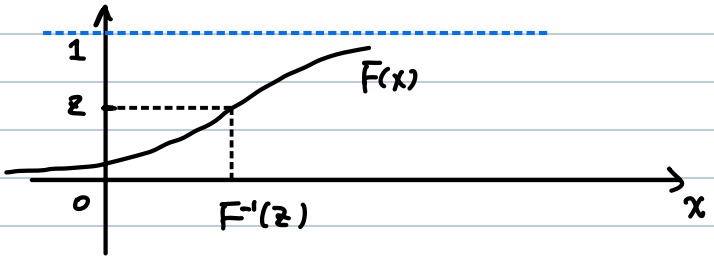
Let $Z_1, \dots, Z_n \overset{iid}{\sim} \mathrm{Unif}[0,1]$

$F(x) := \mathbb{P}(X \le x)$ cumulative distribution func.

$F^{-1}(z) := \inf\{x \in \mathbb{R} : F(x) \geq z\}$ pseudo-inverse of F

Then $X_i := F^{-1}(Z_i)$ and $X_1, \dots, X_n \overset{iid}{\sim} p(x)$

since $\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(Z) \leq x) = \mathbb{P}(Z \leq F(x)) = F(x)$



extension to high-dim suffers from curse of dimensionality

Method 2 : Markov chain Monte Carlo

$$\begin{array}{ccc} X_1, \dots, X_n & \xrightarrow{t \to \infty} & X_1^*, \dots, X_n^* \\ \text{iid} & & \text{iid} \\ p_0(x) & \xrightarrow{t \to \infty} & p_*(x) \propto \exp(-V(x)) \end{array}$$

Start from samples that are sampled from a simple distribution, evolve each sample independently, hopefully after large enough # steps, the samples become iid samples from target distribution $p_*(x)$

A standard way to evolve the samples is the following (over-damped) Langevin algorithm

$$X_i(t+1) = X_i(t) - h\nabla V(X_i(t)) + \sqrt{2h}\,\xi_i(t) \qquad (h>0)$$

where $\xi_i(t) \sim N(0, I_d)$ GD + noise

Since all $X_i(t)$'s are independent, we only focus on the distribution of a single sample, denoted by ρ_t, hopefully, $\rho_t \approx \rho_*$ as $t \to \infty$

Actually, this process is running GD for ρ_t w.r.t. a special functional

Optimization over $\mathbb{R}^d$	Optimization over probability measure space
Objective $f : \mathbb{R}^d \to \mathbb{R}$	Objective $E : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$
$\min_{x \in \mathbb{R}^d} f(x)$	$\min_{\rho \in \mathcal{P}(\mathbb{R}^d)} E[\rho]$
Gradient flow: $\frac{dx}{dt} = -\nabla f(x(t))$	Wasserstein gradient flow: $\partial_t \rho_t = -\nabla_{W_2} E[\rho] = \nabla \cdot (\rho_t \nabla \frac{\delta E}{\delta \rho}[\rho_t])$
	$\Updownarrow \quad X_t \sim \rho_t$
	$\frac{dX_t}{dt} = -\nabla \frac{\delta E}{\delta \rho}[\rho_t](X_t)$

Over-damped Langevin algorithm can be viewed as a discretization of the Wasserstein GF for Kullback-Leibler divergence $KL(\rho \| \rho_*) := \int_{\mathbb{R}^d} \rho(x) \log \frac{\rho(x)}{\rho_*(x)} dx$

By this equivalence, we can prove the following theorem

Thm: Suppose V is α-strongly convex, and β-smooth, then

$$W_2^2(\rho_t, \rho_*) \le (1-\alpha h)^t W_2^2(\rho_0, \rho_*) + \underbrace{2\frac{\beta}{\alpha} h d}_{\hookrightarrow \text{bias}}$$

Other MCMC methods:

- (Under-damped) Langevin algorithm is a generalization of the heavy-ball method in optimization

  Heavy-ball flow:

$$\begin{cases} \frac{dx}{dt} = v \\ \frac{dv}{dt} = -\gamma v - \nabla V(x) \end{cases}$$

  Under-damped Langevin:

$$\begin{cases} dx = v\,dt \\ dv = -\gamma v\,dt - \nabla V(x)\,dt + \sqrt{2\gamma}\,dW_t \end{cases}$$

  tight non-asymptotic convergence rate is still open

- Hamiltonian Monte Carlo

  Hamiltonian dynamic

$$\begin{cases} \frac{dx}{dt} = v \\ \frac{dv}{dt} = -\nabla V(x) \end{cases}$$

  Idea:
  At each step,
  1. Sample fresh momentum $v(0) \sim N(0, Id)$
  2. Simulate Hamiltonian dynamics up to time T (deterministic / random)
  3. Drop momentum, keep new position