Introduction to Monte Carlo

Introduction

- Long and rich history since computers were invented
- Contributed to the success of Bayesian Statistics
  - and is still very popular with practioners in many applications
- Important in many other topics:
  - optimisation (simulated annealing), computational physics, statistical mechanics, ...

Good historical account in Wikipedia:
`https://en.wikipedia.org/wiki/Monte_Carlo_method`

Introduction to Monte Carlo

- What is Monte Carlo?
  - Sampling from complex high dimensional distributions to compute integrals
  - use simulation to take advantage of computational power available.
- There are also other deterministic approximation methods:
  - difficult to tune or implement in higher dimensions
  - not very flexible in terms of setup and underlying approximations

Outline

- Perfect Monte Carlo
  - understanding basic principles and variance
- Other topics on variance reduction
  - control variates, Rao-Blackwellisation
- Discussion

Purpose of Monte Carlo

- Consider an arbitrary distribution on $\mathcal{X}$ with a density π w.r.t to dx, such that
$$\pi(dx) = \frac{\gamma(x)}{Z} dx$$
and is Z **unknown.**
- We want to compute
  - expectations:
$$\pi\left(\varphi\right) = \mathbb{E}_\pi[\varphi(X)] = \langle \pi, \varphi \rangle = \int_{\mathcal{X}} \varphi(x)\pi(dx)$$
here $\varphi : \mathcal{X} \longrightarrow \mathbb{R}^{n_x}$ is a function of interest - examples: $\varphi = x^n$, $\varphi = 1_A$,...
  - normalising const: $Z = \int \gamma(x)dx$
  - mode(s): $x^* = \arg\max \gamma$

Bayesian Inference

- Bayesian inference
  - Parameter X is a random variable and Y is some dataset
  - Bayes rule: **posterior**$\propto$ likelihood $\times$ prior

$$p(x|y) \propto \underbrace{p(y|x)p(x)}_{\gamma(x)}$$

Here *evidence*

$$Z = p(y) = \int p(y|x)p(x)dx$$

is very useful to compare models, but is **unknown**

Perfect Monte Carlo

- **IF** we can obtain i.i.d. samples $X^i \sim \pi, \quad i = 1, \ldots, N$
- One can use perfect Monte Carlo

$$\widehat{\pi}(\varphi) = \int_{\mathcal{X}} \varphi(x)\widehat{\pi}(dx) = \frac{1}{N}\sum_{i=1}^{N} \varphi(X^i). \tag{1}$$

with

$$\widehat{\pi}(dx) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X^i}(dx)$$

- In a way one can view samples forming an atomic approximation of π

$$\widehat{\pi} = \frac{1}{N}\sum_{i=1}^{N} \delta_{X^i}$$

Perfect Monte Carlo principles

- Perfect MC: Obtain i.i.d. samples $X^i \sim \pi$ and use sample average
$$\widehat{\pi}(\varphi) = \frac{1}{N}\sum_{i=1}^{N}\varphi(X^i)$$
- Principles:
  - Unbiasedness: $\mathbb{E}^N\left[\widehat{\pi}(\varphi)\right] = \pi(\varphi)$
  - SLLN: $\widehat{\pi}(\varphi) \to_{N\to\infty} \pi\left(\varphi\right)$
  - CLT: $\sqrt{N}(\widehat{\pi}(\varphi) - \pi\left(\varphi\right)) \to \mathcal{N}(0,\rho^2), \quad \rho^2 = \pi\left((\varphi - \pi(\varphi))^2\right)$

Unbiasedness

- Because we sample i.i.d., $X^i \sim \pi$, $\widehat{\pi}(\varphi)$ is an unbiased estimator:

$$\mathbb{E}_\pi \left[\sum_{i=1}^{N} \varphi\left(X^i\right) \right] = \sum_{i=1}^{N} \mathbb{E}_\pi \left[\varphi\left(X^i\right) \right] = N \mathbb{E}_\pi \left[\varphi\left(X\right) \right]$$

- Example 1:

$$\frac{1}{N} \mathbb{E}_\pi \left[\sum_{i} 1_{X^i < c} \right] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_\pi \left[1_{X^i < c} \right] = p(X^i < c)$$

Unbiasedness and variance

- Example 2: for $N > 1$ use $\frac{1}{N}\sum_{i=1}^{N} X^i$ to estimate $\mathbb{E}_\pi[X]$
- In fact $\underline{\text{a single sample from } \pi \text{ is an unbiased estimate for } \mathbb{E}_\pi[X]}$
- But we require many samples and high N
  - variance of estimator decreases with rate $1/N$

Perfect Monte Carlo variance

- Variance (non-asymptotic) is given by

$$Var\left[\widehat{\pi}(\varphi)\right] = \frac{1}{N} Var\left[\varphi(X^i)\right] = \frac{1}{N}\left(\int_{\mathcal{X}} \varphi^2(x)\pi(dx) - \pi(\varphi)^2\right)$$

- Note **rate** of decrease w.r.t N is not dependent on size of $\mathcal{X}$
- Dimensionality still important as integrals and π can depend implicitly on dimension and properties of φ

Issues with perfect Monte Carlo

- Very often direct sampling **not possible**
- Even if this is possible relative variance can still be very high:
  - when $\varphi = 1_A$ where A is a tail with very low probability ($p(X^i < c)$ is very low in Example 1 above)
- Curse of dimensionality:
  - for a required precision we might need exponential computational cost in the dimension
- Crucial question:
  - for a given problem how many samples do I need?

Tail estimation example using Perfect Monte Carlo

- Consider a continuous distribution P with density $p(x)$
- We are interested in computing $p^* = P(X \leq c) \approx 10^{-9}$
- Naive Monte Carlo setting:
  - For $i = 1 : N$ sample i.i.d. $x^i \sim p(\cdot)$, then compute

$$\widehat{p^*} = \frac{1}{N} \sum_{i=1}^{N} 1_{x \leq c}(x^i)$$

  - $\widehat{p^*}$ consistent, CLT $\sqrt{N}(\widehat{p^*} - p^*) \to \mathcal{N}(0, \mathbb{V}ar_P[1_{x \leq c}])$,

Tail estimation example using Perfect Monte Carlo

- Variance of estimator $\sigma^2_{\widehat{p^*}} = \frac{\mathbb{V}ar_p\left[1_{x\leq c}\right]}{N} = \frac{p^* - {p^*}^2}{N}$,
- Relative error:

$$RE = \sqrt{\mathbb{V}ar\left[\frac{\widehat{p^*}}{p^*}\right]} \approx \frac{1}{\sqrt{p^* N}}$$

- So would like at least $N \sim 10^{11}$ to get decent estimators - **Prohibitively long simulation times**

Control variates

- When estimating $\mathbb{E}_\pi[\varphi(X)]$ there are ways to reduce the variance
  - control variates or antithetic variables
  - conditioning or Rao Blackwellisation
  - Importance Sampling
  - ....

Control variates

- Let $\widehat{\varphi}$ be an unbiased estimate for $\mathbb{E}_\pi[\varphi(X)]$.
- For any Y such that $\mathbb{E}_\pi[Y] = 0$ and a constant β, then $\widehat{\varphi} + \beta Y$ is also an unbiased estimator

$$\mathbb{E}_\pi[\widehat{\varphi} + \beta Y] = \mathbb{E}_\pi[\widehat{\varphi}] + \beta \mathbb{E}_\pi[Y] = \mathbb{E}_\pi[\varphi(Y)]$$

and

$$Var_\pi[\widehat{\varphi} + \beta Y] = Var_\pi[\widehat{\varphi}] + \beta^2 \mathbb{V}ar_\pi[Y] + 2\beta \mathbb{C}ov_\pi[\widehat{\varphi}, Y]$$

Control variates

- In theory one can minimise variance w.r.t to β,

$$\beta = -\frac{\mathbb{C}ov_\pi\left[\widehat{\varphi}, Y\right]}{\mathbb{V}ar_\pi\left[Y\right]}$$

  to actually get a zero variance estimator!
- In practice it is difficult to achieve this, i.e. to find such β, Y
  - but can choose Y and tune β numerically and get good variance reduction
- Similar ideas appear in antithetic variates or Multi-level Monte Carlo

Rao Blackwell conditioning

- Consider a bivariate distribution $\tilde{\pi}(x, y) = \pi(x|y)p(y)$, i.e.

$$\int \tilde{\pi}(x, dy) = \pi(x),$$

  and assume **one can simulate** $\pi(x|y)$ and $p(y)$.
- Then $\mathbb{E}\left[\varphi(X)|\, Y\right]$ is an **unbiased** estimator for $\mathbb{E}_\pi\left[\varphi(X)\right]$

$$\mathbb{E}_\pi\left[\varphi(X)\right] = \mathbb{E}_p\left[\mathbb{E}\left[\varphi(X)|\, Y\right]\right]$$

- In addition, we have the variance conditioning identity

$$\mathbb{V}ar_\pi\left[\varphi(X)\right] \geq \mathbb{V}ar_p\left[\mathbb{E}\left[\varphi(X)|\, Y\right]\right]$$

Rao Blackwell conditioning

- Conclusion:
  - conditioning can improve on the variance.
- Procedure:
  - use perfect Monte Carlo from $p(y)$ and then sample from $\pi(x|y)$
  - Obtain i.i.d. samples: $Y^i \sim p \; X^i \sim \pi(\cdot|Y^i)$
  - Y^i acts as an auxiliary variable

Discussion

- Very often perfect Monte Carlo is possible only for simple cases
  - standard distributions for which random number generation is possible
- Even when it is possible to get direct samples from π, some test functions φ can result to estimates with very high Monte Carlo variance
  - e.g. Example 1 above for $\varphi = 1_A$
- Variance of estimators are a measure of efficiency
  - in some cases indirect sampling can be better

Other approaches for Monte Carlo sampling

- There are indirect ways for approximating $\pi(\varphi)$
- Basic approaches:
  - rejection sampling
  - importance sampling
- More advanced
  - Markov Chains - MCMC
  - particle systems & methods - Sequential Monte Carlo (SMC)
  - .... and various combinations of all the above