Sampling and Monte Carlo inference

Guillaume Obozinski

Ecole des Pontts - ParisTech

Master MVA

Outline

Monte Carlo estimation principle

Key idea: to approximate $\mathbb{E}[f(X)]$,

1. Draw $X^{(1)}, ..., X^{(n)} \overset{i.i.d.}{\sim} P$
2. Compute

$$\mu = \mathbb{E}[f(X)] \approx \hat{\mu} := \frac{1}{n} \sum_{i=1}^{n} f(X^{(i)})$$

In general $X = (X_1, \ldots, X_d)$ is the joint distribution of the variables of a graphical model

1. Draw $\big(X_1^{(1)}, \ldots, X_d^{(1)}\big), \ldots, \big(X_1^{(n)}, \ldots, X_d^{(n)}\big) \overset{i.i.d.}{\sim} P$
2. Compute $\mathbb{E}[f(X_1, \ldots, X_d)] \approx \frac{1}{n} \sum_{i=1}^{n} f\big(X_1^{(i)}, \ldots, X_d^{(i)}\big)$

Type of functions we would like to compute

$$\mathbb{E}[\phi(X)], \quad \mathbb{P}[X_k = 1] = \mathbb{E}[X_k], \quad \mathbb{P}[X_k X_l = 1] = \mathbb{E}[X_k X_l].$$

MC relies simply on the

Proposition (Law of Large Numbers (LLN))

$$\hat{\mu} \xrightarrow{a.s.} \mu \quad \textit{if} \quad ||\mu|| < \infty$$

Proposition (Central Limit Theorem (CLT))

For X a scalar random variable, if $\mathrm{Var}(f(X)) = \sigma^2 < \infty$, *then*

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$

and thus

$$\mathbb{E}(||\hat{\mu} - \mu||_2^2) = \frac{\sigma^2}{n}.$$

How to sample from a specific distribution?

- Uniform distribution on [0, 1] : use `rand`
- For Ber(p) : set $X = \mathbf{1}_{\{U<p\}}$ with $U \sim \mathcal{U}([0,1])$

Inverse transform sampling

$$\forall x \in \mathbb{R} \qquad F(x) = \int_{-\infty}^{x} p(t)dt = \mathbb{P}(X \in [-\infty, x])$$

$$X = F^{-1}(U) \text{ avec } U \sim \mathcal{U}([0,1])$$

proof: $\mathbb{P}(X \leq y) = \mathbb{P}(F^{-1}(U) \leq y) = \mathbb{P}(U \leq F(y)) = F(y)$

Example

Exponential distribution with $p(x) = \lambda e^{-\lambda x}\mathbf{1}_{\mathbb{R}_+}(x)$: use

$$X = -\frac{1}{\lambda}\ln(U).$$

Ancestral sampling

How do we sample from $$p(x_1, \dots, x_d) = \prod_{i=1}^{d} p(x_i \mid x_{\pi_i})\ ?$$

Algorithm 1 Ancestral sampling

1: **for** $i = 1$ to d **do**
2: $\quad z_i \leftarrow$ draw z_i from $\mathbb{P}(X_i = .|X_{\pi_i} = z_{\pi_i})$
3: **end for**
return $(z_1, \dots, z_d)$

Rejection sampling

Problem: We do not know of to sample directly from $p(x) = \dfrac{\tilde{p}(x)}{Z_p}$.

Assume:

- We can compute $\tilde{p}$ and Z_p is unknown
- We know
  - a distribution q that we can sample from
  - together with a constant $K \in \mathbb{R}$

  such that

$$\tilde{p}(x) < K\, q(x).$$

Algorithm 2 Rejection Sampling Algorithm

1: Draw X from q
2: Accept X with probability $\frac{\tilde{p}(x)}{K\, q(x)} \in [0, 1]$, otherwise reject X

For instance can be applied to UGMs with $\tilde{p}(x) = \prod_{c \in \mathcal{C}} \psi_c(x_c)$

Importance sampling

Assume $X \sim p$. Aim: compute the expectation of a function f .

$$
\begin{aligned}
\mathbb{E}_p(f(X)) &= \int f(x)p(x)\mathrm{d}x \\
&= \int \frac{f(x)p(x)}{q(x)} q(x)\mathrm{d}x \\
&= \mathbb{E}_q\left(f(Y)\frac{p(Y)}{q(Y)}\right) \qquad \text{with } Y \sim q \\
&= \mathbb{E}_q(g(Y)) \\
&\approx \frac{1}{n}\sum_{j=1}^{n} g(Y_j) \qquad \text{with } Y_j \overset{iid}{\sim} q \\
&= \frac{1}{n}\sum_{j=1}^{n} f(Y_j)\,\frac{p(Y_j)}{q(Y_j)}
\end{aligned}
$$

$w(Y_i) = \frac{p(Y_j)}{q(Y_j)}$ are called *importance weights*.

See lecture notes+ polycopie for more. Useful for particle filters.

Markov Chain Monte Carlo (MCMC)

Problem:

- Often too hard to just sample $(X_1, \dots, X_d)$ for an undirected graphical model.
- even with rejection sampling or importance sampling, because a good distribution q is too hard to find.

New key idea: Sample from an incorrect distribution and create a Markov chain that converges to the right distribution.

$$
\begin{aligned}
x^{(0)} &:= (x_1^{(0)}, \dots, x_d^{(0)}) \quad \text{drawn from} \quad P_0 \\
x^{(1)} &:= (x_1^{(1)}, \dots, x_d^{(1)}) \quad \text{drawn from} \quad Q(X^{(1)} = \cdot \mid X^{(0)} = x^{(0)}) \\
&\vdots \\
x^{(t+1)} &:= (x_1^{(t+1)}, \dots, x_d^{(t+1)}) \quad \text{drawn from} \quad Q(X^{(t+1)} = \cdot \mid X^{(t)} = x^{(t)})
\end{aligned}
$$

Idea: Design Q so that $x^{(t)} \sim P_t$ with $P_t \to P_\infty = P$.

Burn-in

Consider P_t is sufficiently close to $P_\infty = P$ for $t > T_0$.

- the first T_0 observations are discarded (burn-in)
- All observations after T_0 are used to approximate the expectation

$$\mathbb{E}_P[f(X)] \approx \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} f(X^{(t)})$$

Note that the observations are not i.i.d. ...

Review of Markov chains

Definition (Time Homogenous Markov chain)

$$
\begin{aligned}
\forall t \geq 0,\ \forall (x,y) \in \mathcal{X}, \qquad & p(X_{t+1} = y \mid X_t = x, X_{t-1}, \ldots, X_0) \\
= \quad & p(X_{t+1} = y \mid X_t = x) \\
= \quad & p(X_1 = y \mid X_0 = x) \\
= \quad & S(x,y)
\end{aligned}
$$

Definition (Transition matrix)

If $\mathcal{X}$ is a finite set, S is a matrix with

$$S_{x,y} = S(x,y) = \mathbb{P}(X_t = y | X_{t-1} = x).$$

Stationary distribution of a Markov chain

Definition (Stationary Distribution)

The distribution π on $\mathcal{X}$ is stationary if

$$S^T \pi = \pi \quad \text{with} \quad \pi = \big(\pi(x)\big)_{x \in \mathcal{X}}$$

or equivalently $\quad \forall y \in \mathcal{X},\ \pi(y) = \sum_{x \in \mathcal{X}} \pi(x) S(x, y).$

If $\mathbb{P}(X_n = x) = \pi(x)$ with π a stationary distribution of S, then

$$\begin{aligned} \mathbb{P}(X_{n+1} = y) &= \sum_x \mathbb{P}(X_{n+1} = y | X_n = x) \mathbb{P}(X_n = x) \\ &= \sum_x S(x, y)\, \pi(x) = \pi(y) \end{aligned}$$

Theorem (Perron-Frobenius)

Every stochastic matrix S has at least one stationary distribution.

Irreducibility and aperiodicity

$$S^m(x,y) := \mathbb{P}(X_{t+m} = y | X_t = x).$$

Definition (Irreducible Markov Chain)

An MC is *irreducible* if $\quad \forall x, y \in \mathcal{X}, \exists m \in \mathbb{N},\ S^m(x,y) > 0.$

Definition (Aperiodic Markov Chain)

Period of a state: $\quad \text{Period}(x) = \text{GCD}\big(\{m \mid S^m(x,x) > 0\}\big).$

$\text{Period}(x) = 1 \quad \Rightarrow \quad$ state x is called *aperiodic*.

A Markov chain is aperiodic if all states are aperiodic.

Definition (Regular Markov Chain)

A Markov chain is regular if $\quad \forall x, y \in \mathcal{X}, S(x,y) > 0.$

Convergence of irreducible aperodic Markov chains

Proposition

If a Markov chain on a ***finite*** *state space is*

- *irreducible and*
- *aperiodic*

then

- *its transition matrix has a* ***unique*** *stationary distribution* π*,*
- *for any initial distribution* P_0 *on* X_0*,*

$$\text{with} \quad P_t(\cdot) := \mathbb{P}(X_t = \cdot), \qquad \text{then} \qquad P_t \underset{t \to +\infty}{\longrightarrow} \pi.$$

Remark: If the state space is not finite, an additional assumption is needed on the Markov chain: that it is recurrent positive. We do not define this notion in this course.

Detailed balance

Definition (Detailed Balance)

A Markov chain is *reversible* if for the transition matrix S,

$$\exists \pi, \forall x, y \in \mathcal{X}, \qquad \pi(x)S(x,y) = \pi(y)S(y,x).$$

This equation is called the *detailed balance equation*. It can be reformulated as

$$\mathbb{P}(X_{t+1} = y, X_t = x) = \mathbb{P}(X_{t+1} = x, X_t = y)$$

Proposition

If π satisfies detailed balance, then π is a stationary distribution.

proof: $\sum_x S(x,y)\pi(x) = \sum_x \pi(y)S(y,x) = \pi(y)\sum_x S(y,x) = \pi(y).$

Metropolis-Hastings algorithm

Proposal transition

$$T(x, z) = \mathbb{P}(Z = z | X = x)$$

Acceptance probability

$$\alpha(x, t) = \mathbb{P}(\text{Accept } z | X = x, Z = z)$$

⚠ α is not a transition matrix.

Algorithm 3 Metropolis-Hastings

1: Initialize x_0 from $X_0 \sim q$
2: **for** $t = 1, \ldots, T$ **do**
3: Draw z_t from $\mathbb{P}(Z = \cdot | X_{t-1} = x_{t-1}) = T(x_{t-1}, \cdot)$
4: With proba $\alpha(z_t, x_{t-1})$, set $x_t = z_t$, otherwise, set $x_t = x_{t-1}$
5: **end for**
