Advanced Deep Learning

Monte Carlo Methods

U Kang
Seoul National University

In This Lecture

- Monte Carlo Sampling
- Importance Sampling
- Makov Chain Monte Carlo Methods
- Gibbs Sampling
- Challenge of Mixing between Separated Modes

Las Vegas vs Monte Carlo methods

- Las Vegas algorithms
  - Return precisely the correct answer
  - Random amount of resources
  - E.g., randomized quicksort
- Monte Carlo algorithms
  - Fixed computational budget
  - Approximate answer
    - Reducing error by expending more resources
  - E.g., sample n data points to get the average
- Cannot expect precise answers for many machine learning problems => MC

Outline

- **Sampling and Monte Carlo Methods**
- Importance Sampling
- Markov Chain Monte Carlo Methods
- Gibbs Sampling
- Challenge of Mixing between Separated Modes

Sampling and Monte Carlo Methods

- Many ML technologies are based on
  - Drawing samples from probability distribution,
  - Use the samples to form MC estimate

- Why Sampling?
  - Provides a flexible way to approximate sums and integral at reduced costs
  - Sampling may be the goal of an algorithm

e.g. model that can sample from training distribution

Sampling and Monte Carlo Methods

- Basics of MC sampling
  - Sum can not be computed exactly

  Idea: Sum = expectation under some distribution

$$s = \sum_{\mathbf{x}} p(\mathbf{x}) f(\mathbf{x}) = E_p[f(\mathrm{x})]$$

$$s = \int p(\mathbf{x}) f(\mathbf{x}) dx = E_p[f(\mathrm{x})]$$

  - Approximate the expectation by a corresponding average

Sampling and Monte Carlo Methods

- Drawing n samples from p and forming the empirical average

$$\hat{s}_n = \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}^{(i)})$$

- Expected value? Variance ?

Sampling and Monte Carlo Methods

$$\hat{s}_n = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}^{(i)})$$

- Expected Value

$$E[\hat{s}_n] = \frac{1}{n}\sum_{i=1}^{n} E[f(\mathbf{x}^{(i)})] = \frac{1}{n}\sum_{i=1}^{n} s = s$$

- Law of large numbers: if $x^{(i)}$ are i.i.d, then

$$\lim_{n\to\infty} \hat{s}_n = s$$

Sampling and Monte Carlo Methods

$$\hat{s}_n = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}^{(i)})$$

- Variance decreases and converges to 0

$$Var[\hat{s}_n] = \frac{1}{n^2}\sum_{i=1}^{n} Var[f(\mathrm{x})] = \frac{Var[f(\mathrm{x})]}{n}$$

  = estimate of the uncertainty in a MC average

  = expected error of the MC approximation

- By the Central Limit Theorem, the distribution of $\hat{s}_n$ converges to a normal distribution

Sampling and Monte Carlo Methods

- But, what if it is not feasible to sample from $p(x)$?
  - Importance Sampling
  - Monte Carlo Markov Chains

Outline

- ☑ Sampling and Monte Carlo Methods
- ☐ **Importance Sampling**
- ☐ Markov Chain Monte Carlo Methods
- ☐ Gibbs Sampling
- ☐ Challenge of Mixing between Separated Modes

Importance Sampling

- Main idea
  - Rather than sampling from p, we specify another probability density function q (called proposal distribution)

$$s = \int p(\mathbf{x}) f(\mathbf{x}) dx = E_p[f(\mathrm{x})]$$

$$p(\mathrm{x}) f(\mathrm{x}) = q(\mathrm{x}) \frac{p(\mathrm{x}) f(\mathrm{x})}{q(\mathrm{x})}$$

$$s = \int \frac{p(x) f(x)}{q(x)} q(x)\, dx$$

Importance Sampling

- Estimating s

$$p(\mathrm{x})f(\mathrm{x}) = q(\mathrm{x})\frac{p(\mathrm{x})f(\mathrm{x})}{q(\mathrm{x})}$$

  - We can transform

$$\hat{s}_p = \frac{1}{n}\sum_{i=1,\mathrm{x}^{(i)}\sim p}^{n} f(\mathrm{x}^{(i)})$$

into

$$\hat{s}_q = \frac{1}{n}\sum_{i=1,\mathbf{x}^{(i)}\sim q}^{n} \frac{p(\boldsymbol{x}^{(i)})f(\boldsymbol{x}^{(i)})}{q(\boldsymbol{x}^{(i)})}$$

where

$$E_q[\hat{s}_q] = E_p[\hat{s}_p] = s \qquad Var[\hat{s}_q] = Var[\frac{p(\mathrm{x})f(\mathrm{x})}{q(\mathrm{x})}]/n$$

Importance Sampling

- Mean and variance of $\hat{s}_q$
  - The expected value does not depend on q
  - The variance can be greatly sensitive to the choice of q
  - The minimum variance occurs when $\mathrm{q}(\mathrm{x}) = |f(x)|p(x)/c$
    - c = normalization constant to make $q^*(x)$ a probability distribution

Importance Sampling

- Proof of optimal q(x) to minimize variance of $\hat{s}_q$
  - Claim: Let σ_q^2 be the variance of $\frac{p(x)f(x)}{q(x)}$ wrt q(x). Let $\bar{q}(x) = |f(x)|p(x)/c$ where c is a normalization constant. Then for any probability density q, $\sigma_{\bar{q}}^2 \leq \sigma_q^2$.
    - (L1) $\sigma_q^2 = \int \frac{f(x)^2 p(x)^2}{q(x)} dx - \mu^2$ where $\mu = \int f(x)p(x)dx$
    - (L2) for any g, $var\big(g(x)\big) = E[g(x)^2] - [E\big(g(x)\big)]^2 \geq 0$
    - Note that $c = \int |f(x)|p(x)dx$. Then,
    - $\sigma_{\bar{q}}^2 + \mu^2 = \int \frac{f(x)^2 p(x)^2}{\bar{q}(x)} dx = c \int |f(x)|p(x)dx =$ $(\int |f(x)|p(x)dx)^2 = (\int \frac{|f(x)|p(x)}{q(x)} q(x)dx)^2 \leq$ $\int \frac{f(x)^2 p(x)^2}{q(x)^2} q(x)dx = \int \frac{f(x)^2 p(x)^2}{q(x)} dx = \sigma_q^2 + \mu^2$

use (L2)

Importance Sampling

- Discussion
  - Good q can greatly improve efficiency but poor choice makes it much worse
  - Danger
    - If q small for some x and not compensated by small p(x) or f(x), variance can become very large

$$Var[\hat{s}_q] = Var[\frac{p(x)f(x)}{q(x)}]/n$$

    - On the other hand, if q(x) >> p(x) and f(x), collects useless samples

Outline

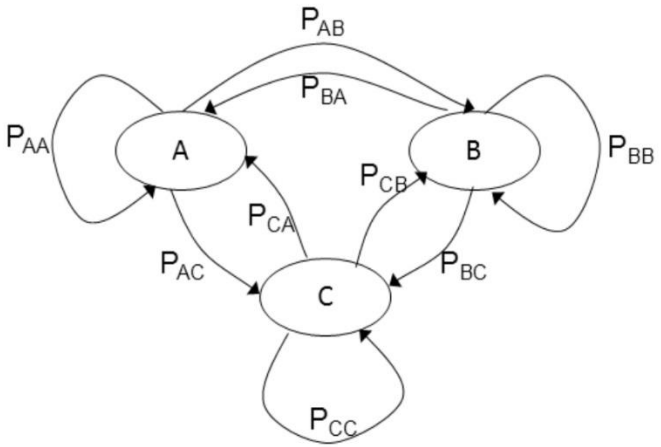
- ☑ Sampling and Monte Carlo Methods
- ☑ Importance Sampling
- ☐ **Markov Chain Monte Carlo Methods**
- ☐ Gibbs Sampling
- ☐ Challenge of Mixing between Separated Modes

Markov Chain Monte Carlo Methods

- In many cases, there is no tractable method for drawing exact samples from p(x) or from a low variance importance sampling distribution q(x)

- MCMC solves the problem
  - (Informal) Have a state x that begins as an arbitrary value. Over time, we randomly update x repeatedly. Eventually x becomes a fair sample from p(x)
  - (Formal) A Markov chain is defined by a random state x and a transition distribution T(x'|x) specifying transition probability. Running the Markov chain means repeatedly updating the state x to a value x' sampled from T(x'|x)

Markov Chain Monte Carlo Methods



$$\mathbf{P}=\begin{bmatrix} p_{AA} & p_{AB} & p_{AC} \\ p_{BA} & p_{BB} & p_{BC} \\ p_{CA} & p_{CB} & p_{CC} \end{bmatrix}$$

- Markov Chain
  - States x, x'
  - Probability T(x'|x)
  - Can run several various chains in parallel
  - States taken from distribution $q^{(t)}(x)$ where t = steps
  - Initialize $q^{(0)}$, goal is for $q^{(t)}(x)$ to converge to p(x)

Markov Chain Monte Carlo Methods

- Markov Chain with states = positive integers

$$q(x = i) = v_i$$

- Update of a single Markov chain's state to a new state

$$q^{(t+1)}(x') = \sum_x q^{(t)}(x) T(x'|x)$$

- Matrix formulation

$$A_{i,j} = T(\mathrm{x}' = i|\mathrm{x} = j)$$

Markov Chain Monte Carlo Methods

- We can describe how the entire distribution over all the different Markov chains (running in parallel) shifts as we apply an update

$$v^{(t)} = Av^{(t-1)}$$

- The matrix A does not change over the iterations, we can write

$$v^{(t)} = A^t v^{(0)}$$

  - A is a *stochastic matrix* = each of its columns represents a probability distribution

Markov Chain Monte Carlo Methods

- Perron-Frobenius theorem guarantees the following:
  - If there is a nonzero probability of transitioning from any state x to any other state x' for some power t, then the sequence converges to the largest eigenvector with eigenvalue 1

- Stationary distribution
  - A has only one eigenvector with $\lambda = 1$

$$Av = v$$

  - This condition holds for any additional step

Markov Chain Monte Carlo Methods

- Stationary distribution
  - The transition operator does change each individual state
  - But, once we have reached the stationary distribution, repeated applications of the transition sampling procedure do not change the distribution over the states of all the various Markov chains
  - If T (transition distribution) is chosen correctly, the stationary distribution q would be equal to p which is the distribution to sample from
- Convergence to the fixed point described by

$$q'(\mathrm{x}') = E_{x \sim q} T(\mathrm{x}' | \mathrm{x})$$

Markov Chain Monte Carlo Methods

- What do we do once the algorithm has converged?
  - A sequence of infinite samples can be drawn from the equilibrium distribution
  - They are identically distributed
  - However, there is a correlation between successive MCMC samples
    - Solution 1) return only every n successive samples; this takes time
    - Solution 2) run multiple Markov chains in parallel to eliminate latency; this takes extra computations

Markov Chain Monte Carlo Methods

- Another difficulty: Mixing time
  - We do not know how many steps are needed to reach equilibrium ("burning in" the Markov chain)
  - Testing if equilibrium reached is difficult
  - In practice, we cannot actually represent our Markov chain in terms of a matrix; the number of states may be exponentially large
  - Instead, use heuristic methods
    - Manually inspecting samples
    - Measuring correlations between successive samples

Outline

- ☑ Sampling and Monte Carlo Methods
- ☑ Importance Sampling
- ☑ Markov Chain Monte Carlo Methods
- ☐ **Gibbs Sampling**
- ☐ Challenge of Mixing between Separated Modes

Motivation

- How can we ensure that a distribution $q(x)$ is a useful distribution?
- A computationally simple and effective approach to building a Markov chain that samples from p_{model} is Gibbs sampling

Gibbs sampling

- A technique to draw **samples** from a joint distribution based on the full conditional distributions of all the associated random variables.
- Sampling from $T(\mathrm{x}'|\mathrm{x})$ is accomplished by
  - Selecting one variable x_i
  - Sampling it from p_{model} conditioned on its neighbors in the undirected graph G defining the structure of the model

Process

- Process: Repeatedly obtain one sample from the probability distribution $p(x_1, x_2, x_3)$ of three random variables

1. Initialize x_1, x_2, x_3 => $X^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)})$
2. Fix $x_2^{(0)}$, and $x_3^{(0)}$, sample $x_1^{(1)} \sim p(x_1^{(1)} | x_2^{(0)}, x_3^{(0)})$
3. Fix $x_1^{(1)}$, and $x_3^{(0)}$, sample $x_2^{(1)} \sim p(x_2^{(1)} | x_1^{(1)}, x_3^{(0)})$
4. Fix $x_1^{(1)}$, and $x_2^{(1)}$, sample $x_3^{(1)} \sim p(x_3^{(1)} | x_1^{(1)}, x_2^{(1)})$
5. We get $X^{(1)} = (x_1^{(1)}, x_2^{(1)}, x_3^{(1)})$
6. Repeat the above k steps

Process

- The initial values of the variables can be determined randomly or by some other algorithm
- It is common to ignore some number of samples at the beginning
  - Depends on initialization $X^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)})$

Advantages & Disadvantages

- Advantages
  - The algorithm is easier to implement
  - Less dependent on initial parameters

- Disadvantages
  - Mixing time may be long
    - Mixing time: The time until the Markov chain is "close" to its steady state distribution
  - Need to be able to compute conditional probability distributions

Block Gibbs Sampling

- Variations of the Gibbs sampling
- Task: Given $p(a, b, c)$ draw samples
- A block Gibbs sampling
  - Groups two or more variables together and samples from their joint distribution conditioned on all other variables
  - E.g., draw (a, b) given c, draw c given (a, b)

Outline

- ☑ Sampling and Monte Carlo Methods
- ☑ Importance Sampling
- ☑ Markov Chain Monte Carlo Methods
- ☑ Gibbs Sampling
- ➡ ☐ **Challenge of Mixing between Separated Modes**

The challenge of Mixing between Separated Modes

- Mixing
  - Means MCMC reaches the steady state distribution; in the state, successive samples from MCMC would be independent
  - Poor mixing: MCMC samples become very correlated

- Mode
  - A region of low energy (= high probability)
    - E.g., if the variables are pixels in an image, a region of low energy might be a connected manifold of images of the same object

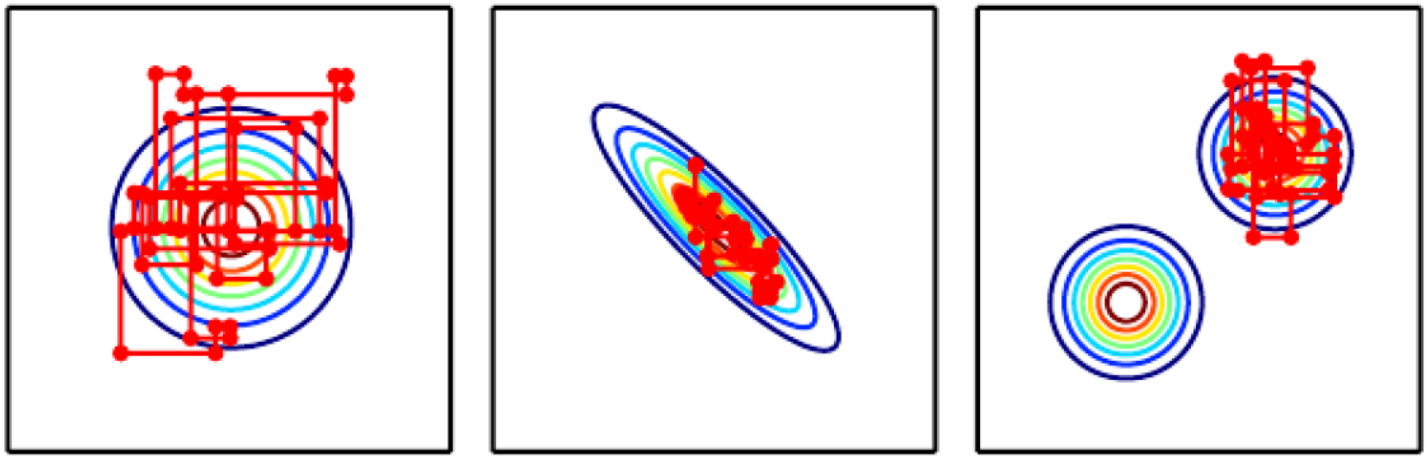
The challenge of Mixing between Separated Modes

- The primary difficulty involved with MCMC methods is that they have a tendency to mix poorly
- Problem 1) In high-dimensional cases, MCMC samples become very correlated
- Problem 2)
  - Successful escape routes are rare for many interesting distributions
  - The Markov chain will continue to sample the same mode longer than it should

The challenge of Mixing between Separated Modes

- Consider the Gibbs sampling algorithm
- Consider the probability of going from one mode to a nearby mode within a given number of steps
- Transitions between two modes that are separated by a region of low probability are less likely.

The challenge of Mixing between Separated Modes



- Paths followed by Gibbs sampling for three distributions, with the Markov chain initialized at the mode in both cases.

The challenge of Mixing between Separated Modes

- Case 1) *Center* figure
  - The correlation between variables makes it difficult for the Markov chain to mix
  - The update of each variable must be conditioned on the other variable, the correlation reduces the rate at which the Markov chain can move away from the starting point

The challenge of Mixing between Separated Modes

- Case 2) *Right* figure
  - Gibbs sampling mixes very slowly
  - Because it is difficult to change modes while altering only one variable at a time.

The challenge of Mixing between Separated Modes

- Difficult to mix between the different modes of the distribution
  - A distribution has sharp peaks of high probability surrounded by regions of low probability
- Several techniques for faster mixing
  - Constructing alternative versions of the target distribution in which the peaks are not as high and the surrounding valleys are not as low

Tempering to Mix between Modes

- An energy-based model as defining a probability distribution $p(x) \propto \exp(-E(x))$
  - Energy-based model with an extra parameter β
  - Controlling how sharply peaked the distribution is $p_\beta(x) \propto \exp(-\beta E(x))$
  - β is described as the reciprocal of the temperature in EBM; small β means high temperature, vice versa
- Tempering
  - A general strategy of mixing between modes by drawing samples with $\beta < 1$
  - Note that very high β implies deterministic p(x), while very small β implies uniform p(x)

Tempering to Mix between Modes

- Two approaches of tempering
- Approach 1) tempered transitions
  - Draw temporarily samples from higher-temperature distributions to mix to different modes, then resume sampling from the unit temperature distribution
    - As temperature rises high, β becomes close to zero, and the distribution becomes more uniform

Tempering to Mix between Modes

- Approach 2) parallel tempering
  - The Markov chain simulates many different states in parallel, at different temperatures
  - Stochastically swap states between two different temperature levels

What you need to know

- Monte Carlo Sampling
- Importance Sampling
- Makov Chain Monte Carlo Methods
- Gibbs Sampling
- Challenge of Mixing between Separated Modes

Questions?