CS281A/Stat241A Lecture 21

Monte Carlo Methods

Peter Bartlett

Announcements

- My office hours:
  Tuesday Nov 10 (today), 1-2pm, in 723 SD Hall.
  Thursday Nov 12, 1-2pm, in 723 SD Hall.
- Homework 5 due 5pm Monday, November 16.

Key ideas of this lecture

- Monte Carlo methods for approximate inference: Approximating expectations
- Applications:
  - E-step of EM.
  - Data augmentation in Bayesian analysis.
- Basic sampling methods
  - Multivariate Gaussians.
  - Directed graphical models.
- Rejection sampling
- Importance sampling
- Particle filters
- Markov Chain Monte Carlo

Approximate Inference

- When the cliques are large, exact inference is intractable.
- We resort to *approximate* inference methods.
  - Monte Carlo methods.
  - Variational methods.
- Today: Monte Carlo methods.

Approximating Expectations

- The inference problem:

  Given observations x_E

  of variables in an evidence set, $E \subset V$,

  and a set of variables $F \subset V$,

  $\ldots$ find $p(x_F | x_E = \bar{x}_E)$.

- We focus on approximating expectations:

$$\mathbb{E}\left[f(x) | x_E = \bar{x}_E\right].$$

Approximating Expectations

$$\mathbb{E}\left[f(x)|x_E = \bar{x}_E\right].$$

- If the functions f are indicators for events, these expectations are probabilities.
- These expectations are useful, for example, for the E-step of the EM algorithm:

$$\mathbb{E}\left[\ell_c(\theta)|x_E = \bar{x}_E\right].$$

Approximating Expectations

$$\mathbb{E}\left[f(x)|x_E = \bar{x}_E\right].$$

- If we can generate iid samples from the conditional distribution, we can approximate expectations.
- For $x^1, \ldots, x^m$ drawn i.i.d. from $p(x|x_E)$, we estimate $\mathbb{E}\left[f(x)|x_E = \bar{x}_E\right]$ with

$$\hat{\mathbb{E}}f = \frac{1}{m}\sum_{t=1}^{m} f(x^t).$$

- Estimate is unbiased: $\mathbb{E}\hat{\mathbb{E}}f = \mathbb{E}[f|x_E]$.
- Variance decreases: $\mathrm{V}ar(\hat{\mathbb{E}}f) = \mathrm{V}ar(f|x_E)/m$.

Bayesian Inference

- In a Bayesian setting, we have a joint distribution

$$p(x, \theta) = p(x|\theta)p(\theta).$$

- Given some observations $x_E = \bar{x}_E$, we wish to sample from the posterior, $p(\theta|x_E)$.
- The same inference problem (the names have changed).

Data Augmentation Algorithm 1

We want to approximate the posterior distribution:

$$p(\theta|x_E) = \int p(\theta|x)p(x_{E^C}|x_E)dx_{E^C}$$

$$\approx \frac{1}{m}\sum_{i=1}^{m} p(\theta|x^i_{E^C}, x_E),$$

where $x^1_{E^C}, x^2_{E^C}, \ldots x^m_{E^C}$ are chosen (approximately) from $p(x_{E^C}|x_E)$.

Data Augmentation Algorithm 2

$$p(x_{E^C}|x_E) = \int p(x_{E^C}|\theta, x_E)p(\theta|x_E)d\theta$$
$$\approx \frac{1}{m}\sum_{i=1}^{m} p(x_{E^C}|\theta^i, x_E),$$

where $\theta^1, \theta^2, \ldots \theta^m$ are chosen (approximately) from $p(\theta|x_E)$.

Data Augmentation Algorithm

I-step (Imputation): Use the sample $\theta^1, \ldots, \theta^m$ to approximately sample $x^1_{E^C}, \ldots, x^m_{E^C}$ from $p(x_{E^C} | x_E)$.

P-step (Posterior): Use the sample $x^1_{E^C}, \ldots, x^m_{E^C}$ to approximately sample $\theta^1, \ldots, \theta^m$ from $p(\theta | x_E)$.

Need to:

1. Sample from $p(\theta | x)$.
2. Sample from $p(x_{E^C} | \theta, x_E)$.

Key ideas of this lecture

- Monte Carlo methods for approximate inference: Approximating expectations
- Applications:
  - E-step of EM.
  - Data augmentation in Bayesian analysis.
- Basic sampling methods
  - Multivariate Gaussians.
  - Directed graphical models.
- Rejection sampling
- Importance sampling
- Particle filters
- Markov Chain Monte Carlo

Sampling Multivariate Gaussians

- Suppose we wish to sample $x \sim \mathcal{N}(\mu, \Sigma)$, and we have a source of (one-dimensional) Gaussians.
- If $Z \sim \mathcal{N}(0, I)$, then

$$x = \mu + LZ$$

  has distribution $\mathcal{N}(\mu, LL')$.
- Cholesky decomposition of a symmetric positive semidefinite matrix:

$$\Sigma = LL',$$

  where L is lower triangular.

Unconditional Sampling

- Consider a directed graphical model:

$$p(x) = \prod_i p(x_i | x_{\pi(i)}).$$

- Suppose that we wish to sample from p. unconditionally; no evidence.
- Algorithm:
  for each i (in a topological order):
  - Sample x_i from $p(x_i | x_{\pi(i)})$.

Key ideas of this lecture

- Monte Carlo methods for approximate inference: Approximating expectations
- Applications:
  - E-step of EM.
  - Data augmentation in Bayesian analysis.
- Basic sampling methods
  - Multivariate Gaussians.
  - Directed graphical models.
- Rejection sampling
- Importance sampling
- Particle filters
- Markov Chain Monte Carlo

Rejection Sampling

To generate m i.i.d. samples from $p(x|x_E)$:

- $S = \emptyset$.
- While $|S| < m$
  - Generate x from $p(x)$.
  - If $x_E = \bar{x}_E$, set $S := S \cup \{x\}$.

Each element x of the set S has distribution $p(x|x_E = \bar{x}_E)$.

Rejection Sampling

To generate m i.i.d. samples from $p(x)$:

- Fix a proposal distribution q satisfying

$$\exists C,\ \forall x,\ q(x) \geq Cp(x).$$

- $S = \emptyset$.
- While $|S| < m$
  - Generate x from $q(x)$.
  - Generate u uniformly from $[0, q(x)/C]$.
  - If $u \leq p(x)$, set $S := S \cup \{x\}$.

Rejection Sampling

- Why are the samples from $p(x)$?
  For any (x, u) for which x is accepted,

$$\begin{aligned}\Pr(x|u \leq p(x)) &= \frac{\Pr(x)\Pr(u \leq p(x)|x)}{\Pr(u \leq p(x))} \\ &= \frac{q(x)Cp(x)/q(x)}{\Pr(u \leq p(x))} \\ &= p(x)\frac{C}{\Pr(u \leq p(x))} \\ &= p(x),\end{aligned}$$

  from which we also see that $\Pr(u \leq p(x)) = C$.
- Thus, the expected time to sample m points from p is m/C.

Rejection Sampling

The same argument works when we do not know a normalizing constant for p:
To generate m i.i.d. samples from $p(x)$,

- Fix a proposal distribution q satisfying

$$\exists C, \forall x, \; q(x) \geq CZp(x).$$

- $S = \emptyset$.
- While $|S| < m$
  - Generate x from $q(x)$.
  - Generate u uniformly from $[0, q(x)/C]$.
  - If $u \leq Zp(x)$, set $S := S \cup \{x\}$.

Rejection Sampling

- Why are the samples from $p(x)$?
  For any (x, u) for which x is accepted,

$$\begin{aligned}
\Pr(x|u \leq p(x)) &= \frac{\Pr(x)\Pr(u \leq Zp(x)|x)}{\Pr(u \leq Zp(x))} \\
&= \frac{q(x)CZp(x)/q(x)}{\Pr(u \leq Zp(x))} \\
&= p(x)\frac{CZ}{\Pr(u \leq Zp(x))} \\
&= p(x),
\end{aligned}$$

  from which we also see that $\Pr(u \leq p(x)) = CZ$.

Rejection Sampling: $p(x|x_E)$

- Why is $p(x|x_E)$ a special case?
- Set $q(x) = p(x)$, the joint distribution.
- If $x_E = \bar{x}_E$,

$$
\begin{aligned}
q(x) &= p(x_E)p(x|x_E) \\
&= C\, p(x|x_E),
\end{aligned}
$$

  and since u is uniform on $[0, q(x)/C]$, we accept with probability 1.
- If $x_E \neq \bar{x}_E$, $q(x)/C = p(x|x_E) = 0$, so we reject with probability 1.

Rejection Sampling: Drawbacks

- Acceptance ratio can be small: it typically decreases exponentially with the dimension/number of variables.
- Thus, may need to do a lot of computation to gather a sample.

Key ideas of this lecture

- Monte Carlo methods for approximate inference: Approximating expectations
- Applications:
  - E-step of EM.
  - Data augmentation in Bayesian analysis.
- Basic sampling methods
  - Multivariate Gaussians.
  - Directed graphical models.
- Rejection sampling
- Importance sampling
- Particle filters
- Markov Chain Monte Carlo

Importance Sampling

- Key Idea: replace the random accept/reject decision in rejection sampling with a weighting, equal to the probability of acceptance.
- Again, choose a proposal distribution $q(x)$.

$$\mathbb{E}_p f(X) = \int f(x) p(x) dx$$

$$= \int f(x) \frac{p(x)}{q(x)} q(x) dx = \mathbb{E}_q \left[f(X) \underbrace{\frac{p(X)}{q(X)}}_{w(X)} \right].$$

We call $w(X)$ the *importance weights*.

Importance Sampling

$$\mathbb{E}_p f(X) = \mathbb{E}_q \left[f(X) \frac{p(X)}{q(X)} \right].$$

- c.f. accept with probability $Cp(X)/q(X)$.
- Again, we do not need to know normalization: suppose

$$p(x) = \frac{\tilde{p}(x)}{Z_p}, \qquad q(x) = \frac{\tilde{q}(x)}{Z_q}.$$

Then

$$\mathbb{E}_p f(X) = \frac{\mathbb{E}_q \left[f(X) \frac{\tilde{p}(X)}{\tilde{q}(X)} \right]}{\mathbb{E}_q \left[\frac{\tilde{p}(X)}{\tilde{q}(X)} \right]}$$

Importance Sampling

$$p(x) = \frac{\tilde{p}(x)}{Z_p}, \qquad q(x) = \frac{\tilde{q}(x)}{Z_q}.$$

$$\mathbb{E}_p f(X) = \frac{1}{Z_p} \int f(x)\tilde{p}(x)dx = \frac{Z_q}{Z_p} \mathbb{E}_q \left[f(X) \frac{\tilde{p}(X)}{\tilde{q}(X)} \right]$$

and $\displaystyle \frac{Z_p}{Z_q} = \int \frac{\tilde{p}(x)}{Z_q} dx = \int \frac{\tilde{p}(x)}{\tilde{q}(x)} q(x) dx = \mathbb{E}_q \left[\frac{\tilde{p}(X)}{\tilde{q}(X)} \right].$

So

$$\mathbb{E}_p f(X) = \frac{\mathbb{E}_q \left[f(X) \frac{\tilde{p}(X)}{\tilde{q}(X)} \right]}{\mathbb{E}_q \left[\frac{\tilde{p}(X)}{\tilde{q}(X)} \right]}$$

Importance Sampling

$$\mathbb{E}_p f(X) = \frac{\mathbb{E}_q \left[f(X) \frac{\tilde{p}(X)}{\tilde{q}(X)} \right]}{\mathbb{E}_q \left[\frac{\tilde{p}(X)}{\tilde{q}(X)} \right]}$$

We estimate this with

$$\frac{\sum_{i=1}^m w^i f(x^i)}{\sum_{i=1}^m w^i},$$

where

$$x^i \sim q \qquad \text{and} \qquad w^i = \frac{\tilde{p}(x^i)}{\tilde{q}(x^i)}.$$

Example: Likelihood Weighting

To calculate a single (x, w) pair from $p(x|x_E = \bar{x}_E)$ in a directed graphical model:

- Set $w := 1$
- For all i in a topological order

  if $i \in E$**:** set

$$x_i := \bar{x}_i$$
$$w := w\ p(\bar{x}_i|x_{\pi(i)})$$

  else: sample x_i from $p(x_i|x_{\pi(i)})$.

Example: Likelihood Weighting

Think of each (x, w) pair as a particle at x with weight w. We approximate the distribution by this set of weighted particles.

$$\hat{\mathbb{E}} f = \frac{\sum_{i=1}^m w^i f(x^i)}{\sum_{i=1}^m w^i}.$$

Here,

$$\tilde{p}(x) = p(x) = p(x|x_E)p(x_E)$$

$$\tilde{q}(x) = \prod_{i \notin E} p(x_i|x_{\pi(i)}),$$

so $w(x) = \frac{\tilde{p}(x)}{\tilde{q}(x)} = \frac{\prod_{i \in V} p(x_i|x_{\pi(i)})}{\prod_{i \notin E} p(x_i|x_{\pi(i)})} = \prod_{i \in E} p(x_i|x_{\pi(i)}).$

Importance Sampling

The variance of the estimate

$$\hat{\mathbb{E}}f = \frac{1}{m}\sum_{i=1}^{m} f(x^i)\frac{p(x^i)}{q(x^i)}$$

is

$$\frac{1}{m}\mathsf{V}ar\left(f(x^i)\frac{p(x^i)}{q(x^i)}\right).$$

This is minimized when

$$q(x) = \frac{f(x)p(x)}{\mathbb{E}f}.$$

Key ideas of this lecture

- Monte Carlo methods for approximate inference: Approximating expectations
- Applications:
  - E-step of EM.
  - Data augmentation in Bayesian analysis.
- Basic sampling methods
  - Multivariate Gaussians.
  - Directed graphical models.
- Rejection sampling
- Importance sampling
- Particle filters
- Markov Chain Monte Carlo

Particle Filters

- Consider a filtering problem, $p(x_t|y_1, \ldots, y_t)$:
  - HMM
  - Kalman filter
- Suppose $p(y_t|x_t)$ is complex.
  e.g., x_t is location of robot, y_t is (possibly multipath) sonar measurement of distance to a landmark.
- Then $p(x_t|y_1, \ldots, y_t)$ is complex.
- We can approximate these distributions with weighted particles $(x_t^1, w_t^1), \ldots, (x_t^m, w_t^m)$.

Particle Filters

- We have samples $x_t^1, \ldots, x_t^m$, approximately distributed as $p(x_t|y_1, \ldots, y_{t-1})$, and we use these to compute expectations under $p(x_t|y_1, \ldots, y_t)$:

$$\hat{\mathbb{E}} f(X_t) = \sum_{i=1}^{m} w_t^i f(x_t^i),$$

where

$$w_t^i = \frac{p(y_t|x_t^i)}{\sum_{j=1}^{m} p(y_t|x_t^j)}.$$

Particle Filters

- To see that this makes sense:

$$\begin{aligned}
\mathbb{E}f(X_t) &= \int f(x_t)p(x_t|y_1,\ldots,y_t)dx_t \\
&= \frac{\int f(x_t)p(x_t,y_t|y_1,\ldots,y_{t-1})dx_t}{\int p(x_t,y_t|y_1,\ldots,y_{t-1})dx_t} \\
&= \frac{\int f(x_t)p(y_t|x_t)p(x_t|y_1,\ldots,y_{t-1})dx_t}{\int p(y_t|x_t)p(x_t|y_1,\ldots,y_{t-1})dx_t} \\
&\approx \sum_{i=1}^{m} f(x_t^i)w_t^i.
\end{aligned}$$

Particle Filters

We update our weighted particles (x_t^i, w_t^i) by sampling x_{t+1}^i from

$$p(x_{t+1}|y_1,\ldots,y_t) = \int p(x_{t+1}|x_t,y_1,\ldots,y_t)p(x_t|y_1,\ldots,y_t)dx_t$$
$$\approx \sum_{j=1}^{m} p(x_{t+1}|x_t^j)w_t^j.$$

and by setting

$$w_{t+1}^i = \frac{p(y_{t+1}|x_{t+1}^i)}{\sum_{j=1}^m p(y_{t+1}|x_{t+1}^j)}.$$

Particle Filters

$$
\begin{aligned}
&p(x_{t+1}|y_1,\ldots,y_t) \\
&= \int p(x_{t+1}|x_t,y_1,\ldots,y_t)p(x_t|y_1,\ldots,y_t)dx_t \\
&= \int p(x_{t+1}|x_t)p(x_t|y_1,\ldots,y_t)dx_t \\
&= \frac{\int p(x_{t+1}|x_t)p(x_t|y_1,\ldots,y_{t-1})p(y_t|x_t,y_1,\ldots,y_{t-1})dx_t}{\int p(x_t|y_1,\ldots,y_{t-1})p(y_t|x_t,y_1,\ldots,y_{t-1})dx_t} \\
&= \frac{\int p(x_{t+1}|x_t)p(x_t|y_1,\ldots,y_{t-1})p(y_t|x_t)dx_t}{\int p(x_t|y_1,\ldots,y_{t-1})p(y_t|x_t)dx_t} \\
&\approx \sum_{i=1}^{m} p(x_{t+1}|x_t^i)w_t^i.
\end{aligned}
$$

Particle Filters

$$p(x_{t+1}|y_1,\ldots,y_t) \approx \sum_{i=1}^{m} p(x_{t+1}|x_t^i)w_t^i.$$

This distribution is a mixture of the m components $p(x_{t+1}|x_t^i)$.
We draw $x_{t+1}^1,\ldots,x_{t+1}^m$ from it.

Particle Filter Updates

1. Draw x_{t+1}^i from the mixture $\sum_{j=1}^m p(x_{t+1}|x_t^j)w_t^j$.
2. Weight each particle by $w_{t+1}^i \propto p(y_{t+1}|x_{t+1}^i)$.

Then expectations under $p(x_{t+1}|y_1, \ldots, y_{t+1})$ are approximated by

$$\hat{\mathbb{E}}f(X_{t+1}) = \sum_{i=1}^m w_{t+1}^i f(x_{t+1}^i).$$

Key ideas of this lecture

- Monte Carlo methods for approximate inference: Approximating expectations
- Applications:
  - E-step of EM.
  - Data augmentation in Bayesian analysis.
- Basic sampling methods
  - Multivariate Gaussians.
  - Directed graphical models.
- Rejection sampling
- Importance sampling
- Particle filters
- Markov Chain Monte Carlo

Markov Chain Monte Carlo

- To sample from $p(x)$ on a space $\mathcal{X}$:
  - Choose a Markov chain with state space $\mathcal{X}$.
  - Choose transition probabilities A so that the distribution over states converges (quickly) to p.
  - Simulate the Markov chain, and use the samples

$$x_t, x_{t+k}, x_{t+2k}, \ldots$$

MCMC: Terminology

- The transition probability matrix of a Markov chain determines the state evolution:

$$A_{ij} = \Pr(x_{t+1} = j | x_t = i).$$

- Recall that a distribution over states $p_t(x)' = (\Pr(x_t = 1), \ldots, \Pr(x_t = N))$ evolves as

$$p'_{t+1} = p'_t A.$$

- A *stationary distribution* p on $\mathcal{X}$ satisfies $p' A = p'$.

MCMC: Terminology

- An *ergodic* Markov chain is irreducible (no islands) and aperiodic. It always has a *unique* stationary distribution: for all p_0,

$$p_0' A^t \to p.$$

- An ergodic MC *mixes* exponentially: for some C, τ and stationary distribution p,

$$\|p_0' A^t - p\|_1 \le C e^{-t/\tau}.$$

MCMC: Terminology

- If p satisfies the detailed balance equations

$$p_i A_{ij} = p_j A_{ji},$$

then p is a stationary distribution, and the chain is called *reversible*:

$$\Pr(x_t = i, x_{t+1} = j) = \Pr(x_t = j, x_{t+1} = i).$$

Key ideas of this lecture

- Monte Carlo methods for approximate inference: Approximating expectations
- Applications:
  - E-step of EM.
  - Data augmentation in Bayesian analysis.
- Basic sampling methods
  - Multivariate Gaussians.
  - Directed graphical models.
- Rejection sampling
- Importance sampling
- Particle filters
- Markov Chain Monte Carlo