Markov Chain Monte Carlo

Kevin Duh

Mokuyokai 12/10/2009

Problems addressed

- Integration/Expectation

$$E_p[f(x)] = \int_X f(x)p(x)dx$$

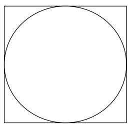
- Optimization

$$\arg\max_x p(x)$$

- Situation suitable for MCMC:
  - Direct integration/optimization is hard:
    - X is a large space or f(x) is complex
  - P(x) is easily computed (up to normalization constant)

Example: Integration

- Problem: Compute area of circle
  - Suppose we don't know the formula pi*r^2
- Monte Carlo approach:
  1. Bound the circle with a box
  2. Randomly (uniformly) spread seeds in the box
  3. Count the number of seeds inside the circle



Example: Optimization

- Problem: ASR decoding without Viterbi
  - argmax p(x) where p(x) is complex acoustic + language model
- Markov Chain Monte Carlo approach:
  1. Start with random hypothesis sentence x_0
  2. Randomly transform x_t into x_{t+1}
  3. If $p(x_t) > p(x_{t+1})$, let $x_{t+1} = x_t$
  4. Repeat steps 2 & 3 until convergence

my hi is name…
hi my is name…
hi is my name…
hi my name is…

What we'll cover

- Monte Carlo methods:
  - Rejection sampling
  - Importance sampling
- Markov Chain Monte Carlo (MCMC):
  - Markov Chain review
  - Metropolis-Hastings algorithm
  - Gibbs sampling
- Others: Monte Carlo EM, Slice sampling

Monte Carlo Principle

- Approximate density by samples from p(x):

$$p_N(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x^{(i)}}(x),$$

- Estimate is unbiased and converges to the truth

$$I_N(f) = \frac{1}{N}\sum_{i=1}^{N} f\left(x^{(i)}\right) \xrightarrow[N\to\infty]{a.s.} I(f) = \int_{\mathcal{X}} f(x)p(x)\,dx.$$

- Advantage over deterministic integration:
  - Samples from high-probability area, so more efficient
  - Question: how to sample from complex p(x)?

Rejection sampling

- Problem setup:
  - Want to sample p(x), but too hard
  - Instead sample from proposal distribution q(x)
    - Requirement: Mq(x) ≥ p(x) forall x

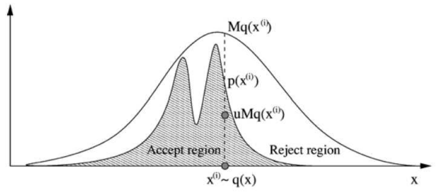


Figure 2. Rejection sampling: Sample a candidate $x^{(i)}$ and a uniform variable u. Accept the candidate sample if $uMq(x^{(i)}) < p(x^{(i)})$, otherwise reject it.

Issues with Rejection sampling

- Difficult to bound p(x) over the whole space with a small M
- If M is too large, most samples will be rejected → not efficient

$$\Pr(x \text{ accepted}) = \Pr\left(u < \frac{p(x)}{Mq(x)}\right) = \frac{1}{M}$$

Importance Sampling

- Use arbitrary proposal distribution q(x) whose support includes p(x)

$$I(f) = \int f(x)p(x)dx = \int f(x)\frac{p(x)}{q(x)}q(x)dx = \int f(x)w(x)q(x)dx$$

Importance weight

- Directly compute expectation (using all samples)

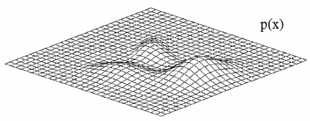
$$\hat{I}_N(f) = \sum_{i=1}^{N} f(x^{(i)})w(x^{(i)})$$

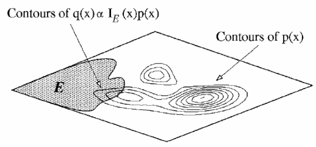
What is a good proposal distribution?

- Answer: one that is proportional to |f(x)|p(x)

$$q^{\star}(x) = \frac{|f(x)|p(x)}{\int |f(x)|p(x)\,dx}$$

- Sampling from q*(x) may be more efficient than p(x)
- But this is not always possible



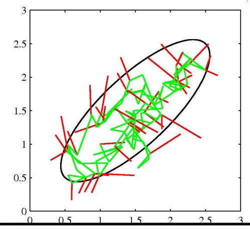


What we'll cover

- Monte Carlo methods:
  - Rejection sampling
  - Importance sampling
- Markov Chain Monte Carlo (MCMC):
  - Markov Chain review
  - Metropolis-Hastings algorithm
  - Gibbs sampling
- Others: Monte Carlo EM, Slice sampling

MCMC Motivation

- Monte Carlo methods may not be efficient in high dimensional spaces
- In MCMC, successive samples are correlated via a Markov chain
  - Explores high-density regions of the state-space



Green=accepted
Red=rejected

Markov Chain review

- An invariant p(x) = p(x)T exists if transition matrix T satisfies:
  - Irreducibility: fully-connected
  - Aperiodic: chain not trapped in cycles
- A sufficient (but not necessary) condition is *reversibility*:
  - $p(x_{t+1})\ T(x_t|x_{t+1}) = p(x_t)\ T(x_{t+1}|x_t)$
- In MCMC, we define proposal distribution $T=q(x_{t+1} \mid x_t)$:
  - After a period of “burn-in” / “mixing”, we’ll be sampling from the invariant distribution p(x)

Metropolis-Hastings (MH) algorithm

1. Initialise $x^{(0)}$.
2. For $i = 0$ to $N-1$
   - Sample $u \sim \mathcal{U}_{[0,1]}$.
   - Sample $x^\star \sim q(x^\star|x^{(i)})$.
   - If $u < \mathcal{A}(x^{(i)}, x^\star) = \min\left\{1, \frac{p(x^\star)q(x^{(i)}|x^\star)}{p(x^{(i)})q(x^\star|x^{(i)})}\right\}$

     $x^{(i+1)} = x^\star$

     else

     $x^{(i+1)} = x^{(i)}$

Reject new state?

MH Examples:
True p(x) is bimodal
Proposal $q(x_{t+1}|x_t) = N(x_t, \text{sigma})$

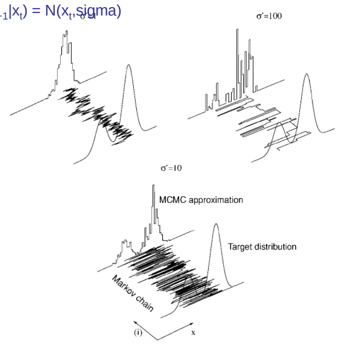


Figure 7. Approximations obtained using the MH algorithm with three Gaussian proposal distributions of different variances.

Why does MH work?

The transition kernel for the MH algorithm is

$$K_{\mathrm{MH}}(x^{(i+1)} \mid x^{(i)}) = q(x^{(i+1)} \mid x^{(i)})\mathcal{A}(x^{(i)}, x^{(i+1)}) + \delta_{x^{(i)}}(x^{(i+1)})r(x^{(i)}),$$

where $r(x^{(i)})$ is the term associated with rejection

$$r(x^{(i)}) = \int_{\mathcal{X}} q(x^\star \mid x^{(i)})(1 - \mathcal{A}(x^{(i)}, x^\star))\, dx^\star.$$

By construction, the reversibility condition is satisfied by K_{MH}, with p(x) emerging as the invariant distribution:

$$p(x^{(i)})K_{\mathrm{MH}}(x^{(i+1)} \mid x^{(i)}) = p(x^{(i+1)})K_{\mathrm{MH}}(x^{(i)} \mid x^{(i+1)})$$

The main question is choosing q() so that convergence is fast
- Gibbs sampling, Metropolis method, etc. are instances of MH with different q()

Gibbs sampler

- Assume a multivariate p(x)
- Gibbs sampling uses the conditional probabilities as q()
  - Very natural for graphical models
- Acceptance probability A() turns out to be 1

$$\begin{aligned}\mathcal{A}(x^{(i)}, x^\star) &= \min\left\{1, \frac{p(x^\star)q(x^{(i)} \mid x^\star)}{p(x^{(i)})q(x^\star|x^{(i)})}\right\} \\ &= \min\left\{1, \frac{p(x^\star)p(x_j^{(i)} \mid x_{-j}^{(i)})}{p(x^{(i)})p(x_j^\star|x_{-j}^\star)}\right\} \\ &= \min\left\{1, \frac{p(x_{-j}^\star)}{p(x_{-j}^{(i)})}\right\} = 1.\end{aligned}$$

Gibbs sampling pseudocode

1. Initialise $x_{0,1:n}$.
2. For $i = 0$ to $N-1$
   - Sample $x_1^{(i+1)} \sim p(x_1|x_2^{(i)}, x_3^{(i)}, \ldots, x_n^{(i)})$.
   - Sample $x_2^{(i+1)} \sim p(x_2|x_1^{(i+1)}, x_3^{(i)}, \ldots, x_n^{(i)})$.

     $\vdots$
   - Sample $x_j^{(i+1)} \sim p(x_j|x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)})$.

     $\vdots$
   - Sample $x_n^{(i+1)} \sim p(x_n|x_1^{(i+1)}, x_2^{(i+1)}, \ldots x_{n-1}^{(i+1)})$.

What we'll cover

- Monte Carlo methods:
  - Rejection sampling
  - Importance sampling
- Markov Chain Monte Carlo (MCMC):
  - Markov Chain review
  - Metropolis-Hastings algorithm
  - Gibbs sampling
- Others: Monte Carlo EM, Slice sampling

19

Monte Carlo EM

1. *E step*. Compute the expected value of the complete log-likelihood function with respect to the distribution of the hidden variables

$$Q(\theta) = \int_{\mathcal{X}_h} \log(p(x_h, x_v \mid \theta)) p(x_h \mid x_v, \theta^{(\mathrm{old})})\, dx_h,$$

where $\theta^{(\mathrm{old})}$ refers to the value of the parameters at the previous time step.

2. *M step*. Perform the following maximisation $\theta^{(\mathrm{new})} = \arg\max_\theta Q(\theta)$.

Use Monte Carlo sampling here to:
(1) approximate difficult integrals
(2) get out of local optima

20

Slice sampling

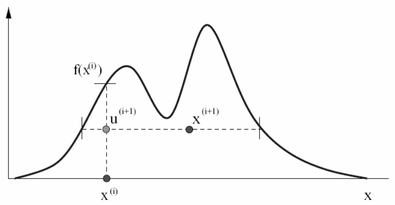
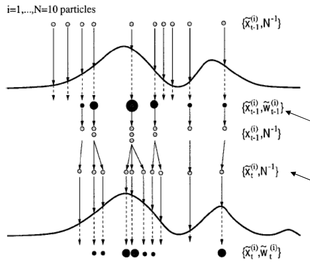


Figure 15. Slice sampling: given a previous sample, we sample a uniform variable $u^{(i+1)}$ between 0 and $f(x^{(i)})$. One then samples $x^{(i+1)}$ in the interval where $f(x) \geq u^{(i+1)}$.

21

Sequential Monte Carlo (Particle Filter)



Setup:
Dynamic model $p(x_t|x_{t-1})$
Observation model $p(y_t|x_t)$

Goal:
Estimate posterior $p(x_t|y_1,y_2,..y_t)$

$$w_t^{(i)} \propto \frac{p\left(y_t|\widetilde{x}_t^{(i)}\right) p\left(\widetilde{x}_t^{(i)}|x_{0:t-1}^{(i)}, y_{1:t-1}\right)}{q_t\left(\widetilde{x}_t^{(i)}|x_{0:t-1}^{(i)}, y_{1:t}\right)}.$$

Increase states with high w
Decrease states with low w

22

Summary

1. Problems addressed:
   - Difficult integration/optimization where p(x) is easily evaluated for a given x

$$I_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}) \xrightarrow[N\to\infty]{a.s.} I(f) = \int_{\mathcal{X}} f(x)p(x)\,dx.$$

2. All methods use a proposal q() to help sample from p(x)
   - Rejection sampling: $Mq(x) \geq p(x)$

3. MCMC differs from Monte Carlo in that successive samples are correlated by $q(x_{t+1} \mid x_t)$
   - Metropolis-Hastings: general $q(x_{t+1} \mid x_t)$

$$\mathcal{A}(x^{(i)}, x^\star) = \min\left\{1, \frac{p(x^\star)q(x^{(i)}|x^\star)}{p(x^{(i)})q(x^\star|x^{(i)})}\right\}$$

   - Gibbs: $q(x_{t+1} \mid x_t)$ is conditional probability of multivariate distribution

23

References

- Figures/Equations for these slides come from:
  - Andrieu et. al. "An Intro to MCMC for Machine Learning", Machine Learning 2003
  - Bishop, *Pattern Recognition and Machine Learning* (chapter 11)
- Other useful references:
  - Diaconis, "The MCMC Revolution"
  - Resnik, "Gibbs sampling for the uninitiated"
  - Robert/Casella, *Monte Carlo statistical methods,* 1999
  - Neal, "Probabilistic inference using MCMC methods", 1993

24