Statistical Data Analysis 2, Sampling Methods

Ewa Szczurek

Faculty of Mathematics, Informatics and Mechanics, University of Warsaw

Nov 2018

Numerical sampling techniques

- In practice, for most probabilistic models, exact inference is intractable
- We are often interested in
  - posterior probabilities
  - evaluating expectations of the form

$$E[f] = \int f(z)p(z)dz,$$

    where $p(z)$ probability distribution, $f(z)$ some function.
- General idea:
  - obtain a set of samples $z^{(l)}$, for $l \in 1, \ldots, L$, drawn independently from the distribution $p(z)$.
  - approximate the expectation by

$$\frac{1}{L}\sum_{l=1}^{L} f(z^{(l)}).$$

Importance sampling

- Assume it is impractical to sample directly from $p(z)$ but that we can evaluate $p(z)$ easily for any given value of z.
- A simplistic strategy for evaluating expectations:
  - Discretize z-space into a uniform grid
  - Evaluate the integrand as

$$E[f] \simeq \sum_{l=1}^{L} p(z^{(l)}) f(z^{(l)}).$$

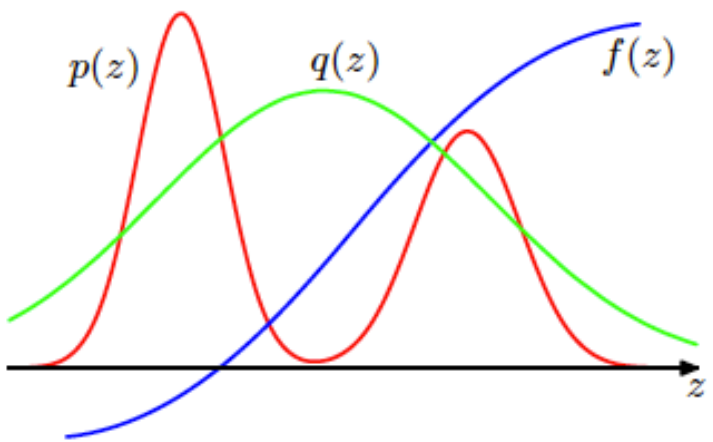
- Problems with this approach:
  - L grows exponentially with the dimensionality of z
  - $p(z)$ may have much of its mass confined to relatively small regions of z space-uniform sampling very inefficient
  - We would really like to choose the sample points to fall in regions where $p(z)$ is large, or ideally where the product $p(z)f(z)$ is large.

Importance sampling

- Instead, take a *proposal distribution*, $q(z)$, from which it is easy to draw samples.
- Take samples $\{z^{(l)}$ from $q(z)$ and estimate the expectation by:

$$E[f] = \int f(z)p(z)dz = \int f(z)\frac{p(z)}{q(z)}q(z)dz \simeq \frac{1}{L}\sum_{l=1}^{L}\frac{p(z)}{q(z)}f(z^{(l)})$$

where $r_l = \frac{p(z)}{q(z)}$ are the so called *importance weights*, correcting for sampling from the wrong distribution.



Importance sampling

- Often, $p(z) = \tilde{p(z)}/Z_p$ can only be evaluated up to the normalization constant Z_p.
- Use proposal distribution $q(z) = \tilde{p}(z)/Z_q$
- We have:

$$E[f] = \int f(z)p(z)dz = \frac{Z_q}{Z_p}\int f(z)\frac{\tilde{p}(z)}{\tilde{q}(z)}q(z)dz \simeq \frac{1}{L}\sum_{l=1}^{L}\tilde{r}_l f(z^{(l)},$$

where $\tilde{r}_l = \frac{\tilde{p}(z)}{\tilde{q}(z)}$.

Importance sampling

- The same sample set can be used to evaluate the ratio

$$\frac{Z_p}{Z_q} = \frac{1}{Z_q}\int \tilde{p}(z)dz = \int \frac{\tilde{p}(z)}{\tilde{q}(z)}q(z) = \frac{1}{L}\sum_{l=1}^{L}\tilde{r}_l,$$

  where the secon equality holds since $\frac{1}{Z_q} = \frac{q(z)}{\tilde{q}(z)}$.

- Thus,

$$E[f] \simeq \sum_{l=1}^{L} w_l f(z^{(l)}),$$

  where

$$w_l = \frac{\tilde{r}_l}{\sum_m \tilde{r}_m} = \frac{\tilde{p}(z^{(l)})/q(z^{(l)})}{\sum_m \tilde{p}(z^{(m)})/q(z^{(m)})}.$$

Problems with importance sampling

- The proposal distribution $q(z)$ cannot be small or zero where $p(z)$ is large
- Otherwise
  - the importance weights $\{r_l\}$ are small or have only few significant values
  - effective sample size is small.

Importance sampling example

- Assume $p(z)$ is given by some graphical model, specifying set of parents pa_i for each variable z_i
- If z_i observed, we say it belongs to evidence set $\mathbf{e}$
- For each i,
  - if $z_i \in \mathbf{e}$ it is set to its observed value
  - otherwise, its value is sampled from the LPD $p(z_i|\mathrm{pa}_i)$

  For the resulting sample z, its weight is

$$r(z) = \prod_{z_i \notin \mathbf{e}} \frac{p(z_i|\mathrm{pa}_i)}{p(z_i|\mathrm{pa}_i)} \prod_{z_i \in \mathbf{e}} \frac{p(z_i|\mathrm{pa}_i)}{1} = \prod_{z_i \in \mathbf{e}} p(z_i|\mathrm{pa}_i).$$

Markov Chain Monte Carlo

- Let $z(\tau)$ record of the current state of the sampling algorithm
- The proposal distribution $q(z|z(\tau))$ depends on the current state and is easy to sample from
- Thus, the sequence of samples, $z(1), z(2), \ldots$ forms a Markov chain.
- At each cycle of the algorithm, we generate a candidate sample z from the proposal distribution and then accept the sample according to an appropriate criterion.
- Notation $p(z) = \tilde{p}(z)/Z_p$ indicates that $\tilde{p}(z)$ can be evaluated for any given z, while Z_p may be unknown.

Metropolis algorithm

- Assumes symmetric proposal distribution, i.e., $q(z_A|z_B) = q(z_B|z_A)$ for all z_A, z_B.
- Algorithm iterates over steps $\tau \in 1, 2, \ldots$
  - Gnerate a candidate sample z^* from $q(z|z^{(\tau)})$
  - The candidate sample z^* is accepted with probability

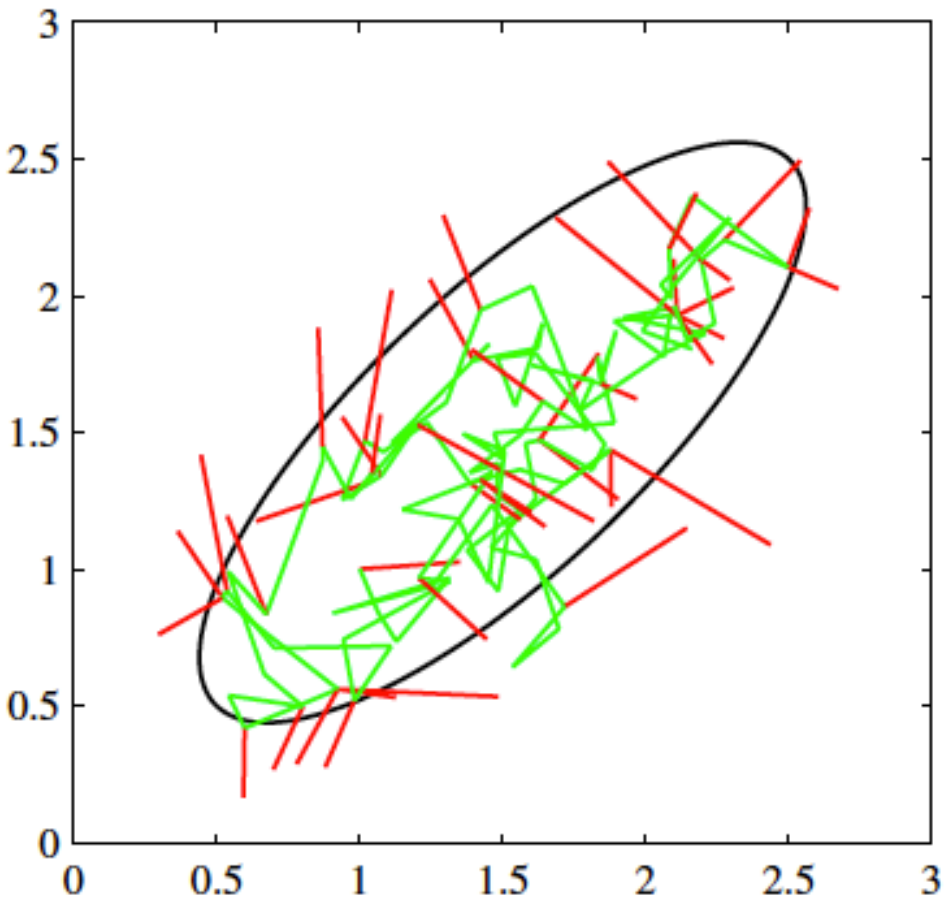
$$A(z^*, z^{(\tau)}) = \min\left(1, \frac{\tilde{p}(z^*)}{\tilde{p}(z^{(\tau)})}\right),$$

    (i.e., accept when $A(z^*, z^{(\tau)}) > u$, for u sampled uniformly from $(0, 1)$)
  - If z^* is accepted, set $z^{(\tau+1)}$ to z^*, otherwise set $z^{(\tau+1)}$ to $z^{(\tau)}$
- If $q(z_A|z_B) > 0$, for $\tau \to \infty$, the distribution of $z^{(\tau)}$ converges to $p(z)$
- The samples $z^{(1)}, z^{(2)}, \ldots z^{(T)}$ are not independent.
- To obtain independent samples, keep only every M^{th} sample.

Metropolis et al., 1953

Example run of the Metropolis algorithm

A simple illustration using Metropolis algorithm to sample from a Gaussian distribution whose one standard-deviation contour is shown by the ellipse. The proposal distribution is an isotropic Gaussian distribution whose standard deviation is 0.2. Steps that are accepted are shown as green lines, and rejected steps are shown in red. A total of 150 candidate samples are generated, of which 43 are rejected.



Homogeneous Markov Chains

- A Markov Chain $\{X_n\}$ is *homogeneous* if

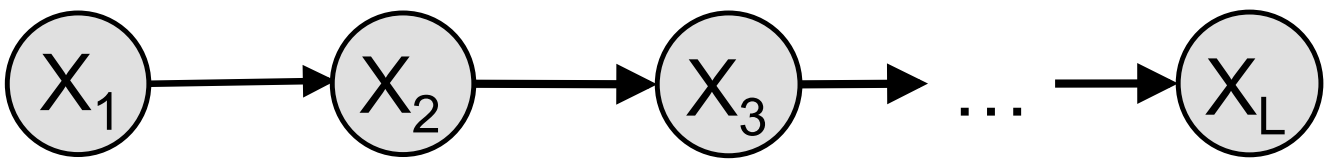
$$P(X_n|X_{n-1}) = P(X_2|X_1), \quad \text{for all } n \geq 2$$

- A homogeneous Markov Chain is defined by
  - the initial state distribution $\Pi \in \Delta_{K-1}$, where K is the number of variable states

$$\Pi_k = P(X_1 = k)$$

  - and the $K \times K$ transition matrix $T = (T_{kl})$ given by

$$T_{kl} = P(X_{n+1} = l|X_n = k)$$



Stationary/invariant distributions

- Marginal distribution of a particular variable can be expressed as

$$p(X_{n+1}) \sum_{X_n} P(X_{n+1}|X_n)p(X_n)$$

- For a homogeneous Markov chain

$$p(X_{n+1}) \sum_{X_n} T_{X_{n+1}X_n} p(X_n)$$

- Distribution $p^*(x)$ over states of a homogeneous Markov chain is stationary/invariant if

$$p^*(l) = \sum_k T_{kl} p^*(k)$$

Detailed balance

- Detailed balance condition is a sufficient (but not necessary) condition for a invariant distribution

$$p^*(l)\,T_{lk} = p^*(k)\,T_{kl}$$

- Markov chain that respects detailed balance is said to be reversible

Ergodicity

A Markov chain is ergodic if it is

1. aperiodic - return to any state is always possible
2. irreducible - any state is accessible from any other
3. positive recurrent - any state will evenually be reached with probability 1 and the mean recurrence time is finite

Stationary/equilibrium distribution in ergodic Markov chains

Theorem

An ergodic Markov chain has a unique equilibrium/stationary distribution $\pi = (\pi_l)_{l \in [K]}$ *such that*

$$\lim_{n \to \infty} T_{kl}^n = \pi_l = \sum_{k \in [K]} \pi_k T_{kl}, l \in [K], \sum_{l \in [K]} \pi_l = 1,$$

independent of the initial distribution Π.

In matrix notation, π is the solution of $\pi^T = \pi T$.

How to set up MCMC

- Goal: use sampling scheme that
  1. generates samples from a given distribution $p^*(z)$ of interest
  2. converges to $p^*(z)$ irrespective of choice of the initial distribution
- Solution: set up a Markov chain
  1. for which $p^*(z)$ is invariant
  2. which is ergodic
     - Then, $p^*(z)$ is its equilibrium distribution.

Constructing transition probabilities

- "Base" transition matrices $B_1, B_2, \ldots B_K$,

$$T = \sum_{b=1}^{K} \alpha_b B_b$$

- If a distribution is invariant with respect to each of the base distributions it will also be with respect to the mixture T_{kl}
- If each of the B_k satisfies detail balance, then also the mixture T_{kl} does.

Metropolis-Hastings algorithm

- The proposal distribution does not need to be symmetric
- Let $p(z) = \tilde{p}(z)/Z_p$
- Algorithm iterates over steps $\tau \in 1, 2, \ldots$
  - Generate a candidate sample z^* from $q_b(z|z^{(\tau)})$, where $z^{(\tau)}$ denotes current state
  - The candidate sample z^* is accepted with probability

$$A_b(z^*, z^{(\tau)}) = \min\left(1, \frac{\tilde{p}(z^*)q_b(z^{(\tau)}|z)}{\tilde{p}(z^{(\tau)})q_b(z^*|z^{(\tau)})}\right),$$

    where b labels the possible transitions being considered.
  - If z^* is accepted, set $z^{(\tau+1)}$ to z^*, otherwise set $z^{(\tau+1)}$ to $z^{(\tau)}$
- If the proposal distribution is symmetric, this procedure reduces to the Metropolis algorithm.

Hastings, 1970

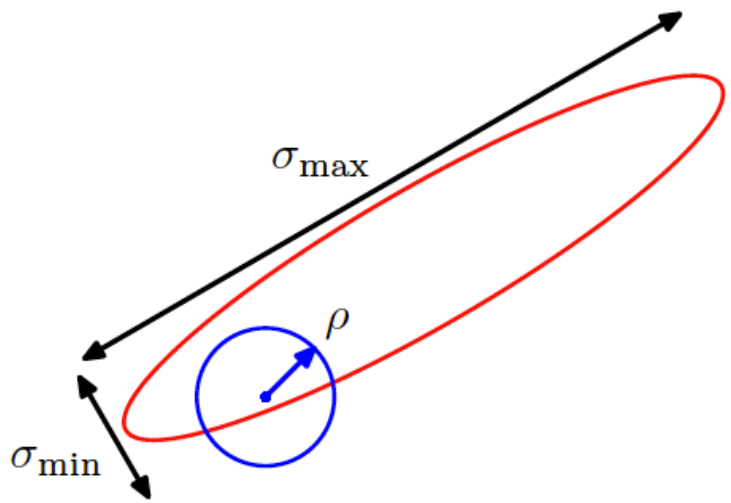
Metropolis-Hastings converges to $p(z)$

- Detailed balance is satisfied

$$
\begin{aligned}
p(l)\,T_{lk} &= p(l)q_b(k|l)A_b(l,k) \\
&= \frac{1}{Z_p}\min\left(\tilde{p}(l)q_b(k|l), \tilde{p}(k)q_b(l|k)\right) \\
&= \frac{1}{Z_p}\min\left(\tilde{p}(k)q_b(l|k), \tilde{p}(l)q_b(k|l)\right) \\
&= p(k)q_b(l|k)A_b(k,l) \\
&= p(k)\,T_{kl}.
\end{aligned}
$$

The choice of the proposal distribution impacts convergence rate

- For continuous state spaces, common choice is a Gaussian centered around the current state
- Trade-off in determining variance of the Gaussian
  - small: the proportion of accepted transitions high, but progress through the state space slow
  - large: the rejection rate high because many of the proposed steps to states for which the probability $p(z)$ is low



Gibbs sampling

- We wish to sample from $p(z) = p(z_1, \ldots, z_M)$
- Choose some initial state
- In each step, replace the value of one variable z_b by a value drawn from its distribution conditioned on the remaining variables $p(z_b|z_{\setminus b})$

Gibbs sampling

1. Initialize $\{z_i : i = 1, \ldots M\}$
2. For $\tau = 1, \ldots, T$
   - Sample $z_1^{(\tau+1)} \sim p(z_1 | z_2^{(\tau)}, z_3^{(\tau)} \ldots, z_M^{(\tau)})$
   - Sample $z_2^{(\tau+1)} \sim p(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)} \ldots, z_M^{(\tau)})$

     $\ldots$
   - Sample $z_i^{(\tau+1)} \sim p(z_i || z_1^{(\tau+1)}, \ldots, z_{i-1}^{(\tau+1)}, z_{i+1}^{(\tau)} \ldots, z_M^{(\tau)})$

     $\ldots$
   - Sample $z_M^{(\tau+1)} \sim p(z_M | z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$

Gibbs sampling as an instance of Metropolis-Hastings

- Regard the conditionals as the proposal distributions: $q_b(z^*|z) = p(z_b|z_{\setminus b})$
- Note $z^*_{\setminus b} = z_{\setminus b}$ because these components remain unchanged
- We have $p(z) = p(z_b|z_{\setminus b})p(z_{\setminus b})$
- The acceptance probability in Metropolis-Hastings scheme would be

$$\begin{aligned} A(z^*, z) &= \min\left(1, \frac{p(z^*)q_b(z|z^*)}{p(z)q_b(z^*|z)}\right) \\ &= \min\left(1, \frac{p(z^*_b|z_{\setminus b})p(z^*_{\setminus b})q_b(z|z^*)}{p(z_b|z_{\setminus b})p(z_{\setminus b})q_b(z^*|z)}\right) \\ &= 1. \end{aligned}$$

Gibbs sampling for graphical models

- Gibbs sampling is particularly useful, if it is much easier to sample from the conditionals $p(z_b|z_{\setminus b})$ than from the joint distribution $p(z_1, \ldots, z_M)$
- For graphical models, $p(z_b|z_{\setminus b}) = p(z_b|MB_b)$