Advanced Machine Learning

Markov Chain Monte Carlo

(Based on slides by Ian Murray)

José Miguel Hernández–Lobato
Department of Engineering
University of Cambridge

Lent Term

Simple Monte Carlo

By the **law of large numbers**, integrals writen as an expected value can be approximated by the **empirical mean** of statistical samples.

General case:

$$\int f(x) p(x)\, dx \approx \frac{1}{N} \sum_{n=1}^{N} f(x_n)\,, \quad x_n \sim p(x)\,.$$

Predictions in Bayesian machine learning:

$$p(y|\mathcal{D}) = \int p(y|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta} \approx \frac{1}{N} \sum_{n=1}^{N} p(y|\boldsymbol{\theta}_n)\,, \quad \boldsymbol{\theta}_n \sim p(\boldsymbol{\theta}|\mathcal{D})\,.$$

More examples: EM algorithm, stochastic optimization, game tree search.

Properties of Monte Carlo estimate

Estimator:

$$\int f(x)p(x)\,dx \approx \hat{f} \equiv \frac{1}{N}\sum_{n=1}^{N} f(x_n)\,, \quad x_n \sim p(x)\,.$$

Unbiasedness:

$$\mathbf{E}_{x_1,\dots,x_N}\left[\hat{f}\right] = \frac{1}{N}\sum_{n=1}^{N}\mathbf{E}_{x_n}\left[f(x_n)\right] = \mathbf{E}_x\left[f(x)\right]\,.$$

Variance shrinkage:

$$\mathbf{Var}_{x_1,\dots,x_N}\left[\hat{f}\right] = \frac{1}{N^2}\sum_{n=1}^{N}\mathbf{Var}_{x_n}\left[f(x_n)\right] = \frac{\mathbf{Var}_x\left[f(x)\right]}{N}\,.$$

The error shrinks as $1/\sqrt{N}$, **independently of dimension of x**!

When to use Monte Carlo methods?

As numerical methods go, Monte Carlo is one of the least efficient; it should be used only on those intractable problems for which all other numerical methods are even less efficient.

— Alan D. Sokal

Sokal, A. Functional integration. Springer, 1997. 131-192.

The main advantage of Monte Carlo methods is their **unbiasedness**.

They are the best method when **computational cost** is not a key factor.

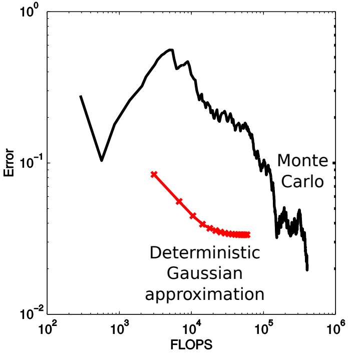
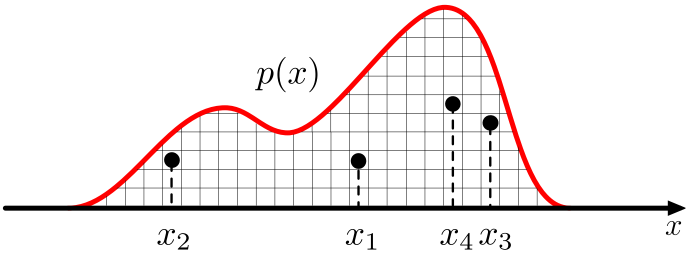


Figure: T. Minka. Phd thesis, MIT, 2001.

Exact sampling from arbitrary distributions

Select points uniformly at random from the area under the curve.



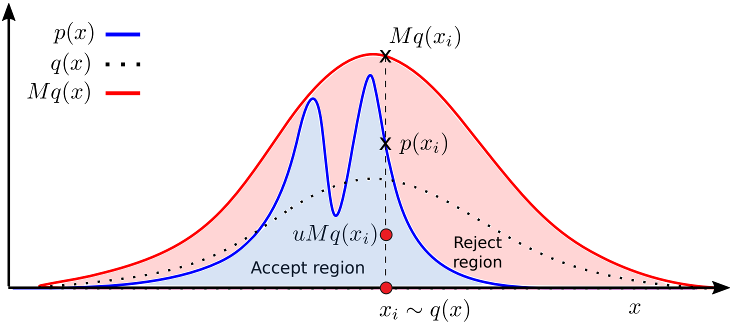
Area to the left of each sample x is uniformly distributed in $[0, 1]$. Why?

Rejection sampling

Simple alternative to sample from $p(x)$ when inverse CDF cannot be applied.

Based on sampling under a curve $Mq(x) \geq p(x)$ for all x.

1. Sample $x_i \sim q(x)$ and $u \sim \text{Uniform}[0, 1]$.
2. If $uMq(x_i) > p(x_i)$ then reject x_i and repeat.



No need for $p(x)$ to be normalized. What is the **acceptance** probability?

Importance sampling

Rejecting x_i seems **wasteful**. Could we avoid this?

Write instead the integral as an **expectation under** $q(x)$:

$$\int f(x)p(x)\,dx = \int f(x)\frac{p(x)}{q(x)}q(x)\,dx, \qquad q(x) > 0 \text{ if } p(x) > 0$$

$$\approx \frac{1}{N}\sum_{n=1}^{N} f(x_n)\underbrace{\frac{p(x_n)}{q(x_n)}}_{w_n} = \frac{1}{N}\sum_{n=1}^{N} f(x_n)w_n\,, \qquad x_i \sim q(x)\,.$$

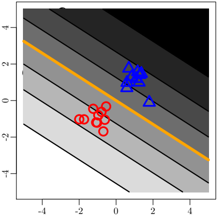
The w_n are known as **importance weights**.

Can be applied **even if the integral is not an expectation**.

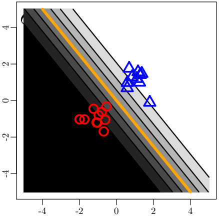
Given $p(x)$, what is the best sampling proposal q?

Importance sampling weights

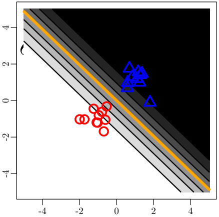
Weights obtained in probit regression when $q(\mathbf{x})$ is the prior.



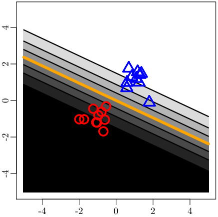
$w_1 = 2.24\text{e-}10$



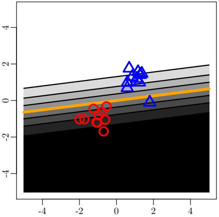
$w_2 = 0.097$



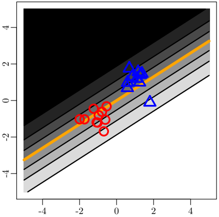
$w_3 = 3.19\text{e-}24$



$w_4 = 0.0516$



$w_5 = 0.00363$



$w_6 = 1.21\text{e-}08$

Many samples do not contribute to the expectation!

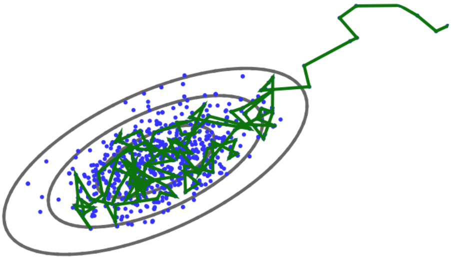
Markov Chain Monte Carlo (MCMC)

Main idea: construct a biased random walk that explores a target distribution $p_\star(\mathbf{x})$ (whose normalization constant may not be known).

The random walk **transition operator** follows the Markov assumption:

$$\mathbf{x}_t \sim T(\mathbf{x}_t|\mathbf{x}_{t-1})\,.$$

The stationary distribution of $\{\mathbf{x}_t\}$ will be $p_\star(\mathbf{x})$:



$\{\mathbf{x}_t\}$ are approximate, correlated samples from $p_\star(\mathbf{x})$.

Transition operator

Discrete example: $x \in 1, 2, 3$.

$$\mathbf{p}_\star = \begin{bmatrix} 3/5 \\ 1/5 \\ 1/5 \end{bmatrix}, \quad \mathbf{T} = \begin{bmatrix} 2/3 & 1/2 & 1/2 \\ 1/6 & 0 & 1/2 \\ 1/6 & 1/2 & 0 \end{bmatrix}, \quad [\mathbf{T}]_{a,b} \equiv T(x_t = a | x_{t-1} = b)\,.$$

$\mathbf{p}_\star$ is the **invariant distribution** of $\mathbf{T}$ because $\mathbf{p}_\star = \mathbf{T}\mathbf{p}_\star$:

$$\sum_b [\mathbf{T}]_{a,b} \, [\mathbf{p}_\star]_b = [\mathbf{p}_\star]_a\,.$$

$\mathbf{p}_\star$ is the **equilibrium distribution** of $\mathbf{T}$ because for any initial state distribution $\mathbf{p}_0$ we have that $\lim_{t \to \infty} [\mathbf{T}^t] \, \mathbf{p}_0 = \mathbf{p}_\star$ ($\mathbf{T}$ is **ergodic**).

Detailed balance

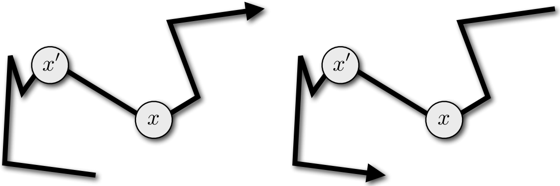
Means that transitions $a \to b$ and $b \to a$ are equally probable in the chain:

$$T(\mathbf{x}'|\mathbf{x})p_\star(\mathbf{x}) = T(\mathbf{x}|\mathbf{x}')p_\star(\mathbf{x}')\,. \tag{1}$$

Detailed balance implies that the invariant distribution is $p_\star(\mathbf{x}')$:

$$\sum_{\mathbf{x}} T(\mathbf{x}'|\mathbf{x})p_\star(\mathbf{x}) = p_\star(\mathbf{x}')\sum_{\mathbf{x}} T(\mathbf{x}|\mathbf{x}') = p_\star(\mathbf{x}')\,.$$

$\{\mathbf{x}\}$ satisfies detailed balanced $\Leftrightarrow$ $\{\mathbf{x}\}$ is reversible, that is, $x_1,\ldots,x_N$ and $x_N,\ldots,x_1$ have the same probability distribution:



To construct a chain that samples from $p_\star(\mathbf{x}')$, just find $T(\mathbf{x}'|\mathbf{x})$ satisfying (1).

Figure source: Ryan P. Adams

Metropolis-Hastings

One of the algorithms with highest influence in science and engineering!

Works by sampling from the **transition operator** given by

- Draw a *proposal* from an *easy* distribution $q(\mathbf{x}'|\mathbf{x})$, e.g., $\mathcal{N}(\mathbf{x}'|\mathbf{x}, \sigma\mathbf{I})$.
- Accept with probability $\min\left(1, \frac{p_\star(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')}{p_\star(\mathbf{x})q(\mathbf{x}'|\mathbf{x})}\right)$.
- Otherwise the next state $\mathbf{x}'$ in chain is a copy of current state $\mathbf{x}$.

Acceptance ratio does not change if $p_\star(\mathbf{x})$ **is not normalized**.

The MH transition operator can be shown to satisfy **detailed balance**.

Proposal $q(\mathbf{x}'|\mathbf{x})$ must have same or larger support than target $p_\star(\mathbf{x})$.

Example

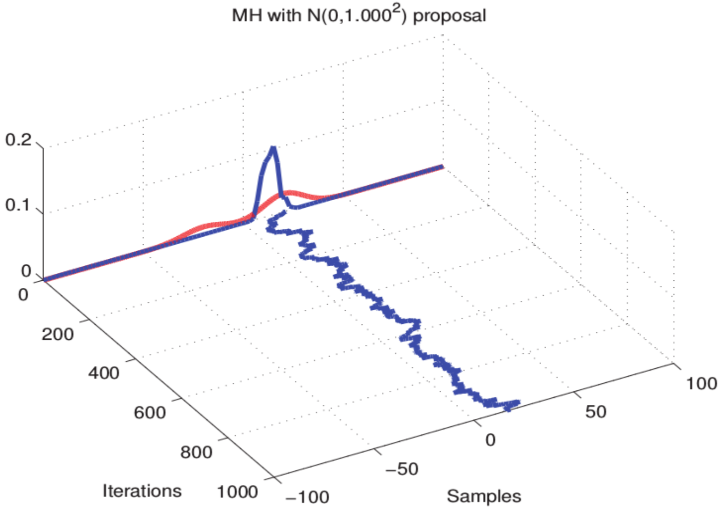


Figure source: K. Murphy

Example

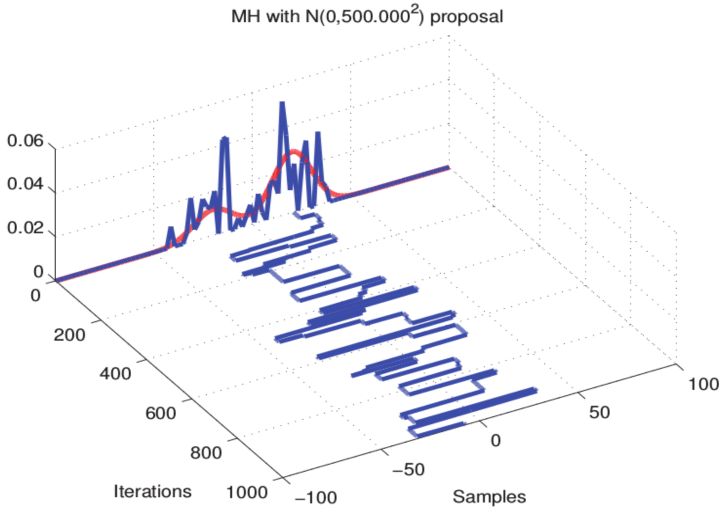


Figure source: K. Murphy

Example

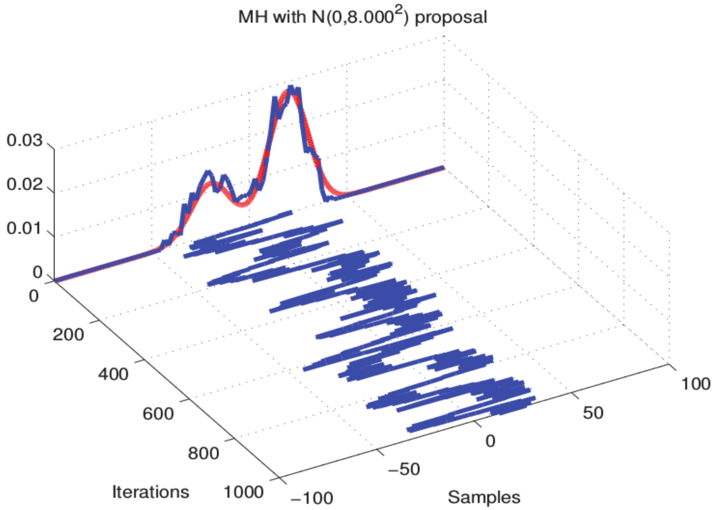
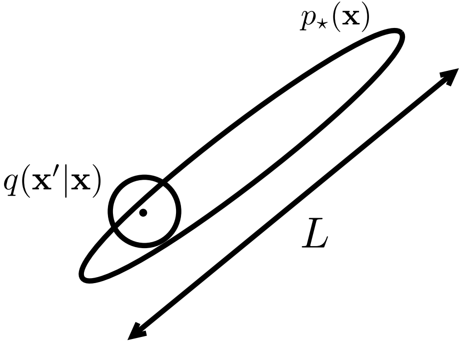


Figure source: K. Murphy

Limitations of Metropolis-Hastings (MH)



- Typically, $q(\mathbf{x}'|\mathbf{x}) = \mathcal{N}(\mathbf{x}'|\mathbf{x}, \sigma\mathbf{I})$ and proposals follow a **random walk**.
- If σ is large, we reject a lot!
- If σ is small, the chain diffuses very slowly: $\approx L^2/\sigma^2$ steps required to obtain independent samples.

Figure source: Ian Murray.