Probabilistic Inference and Learning
Lecture 21
Markov Chain Monte Carlo

Philipp Hennig

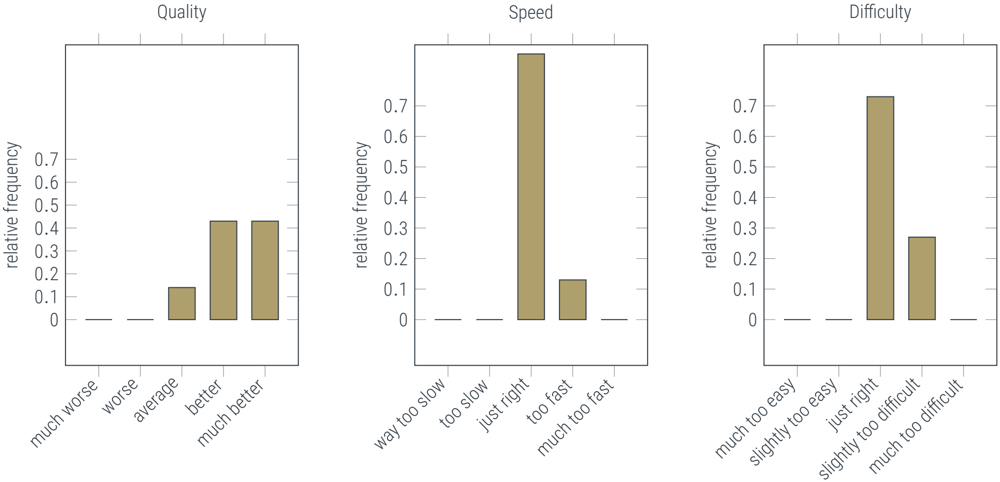
14 January 2019

Eberhard Karls Universität Tübingen

Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning

Last Lecture: Debrief

Feedback dashboard



Last Lecture: Debrief

Detailed Feedback

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Things you did not like:

-

Things you did not understand:

- Why would you want to use a **biased** estimator at all?
- slide 12: Why is it $\mathbb{E}[\frac{1}{S}\sum_{s=1}^{S}(f(x)-\phi)]^2$, shouldn't it be $\mathbb{E}[\frac{1}{S}(\sum_{s=1}^{S} f(x)) - \phi]^2$

Things you enjoyed:

- **discussion of randomness**
- **historical aside**
- discussion of biasedness
- figures
- π-example
- inefficiency of rejection sampling
- "erwartungstreu"
- repetitions from earlier lectures

Ein Hinweis in Eigener Sache …

http://sciencenotes.de

- 17.01.2019, Schlachthaus Tübingen
- Einlass: 19:30 Uhr, Beginn: 20:00 Uhr

1. Wie sehen Maschinen unsere Welt?,
   Dr. Wieland Brendel
2. tbd ,
   Dr. Caterina deBacco
3. Präzise Unsicherheit – Rechenalgorithmen für lernende Maschinen, Prof. Dr. Philipp Hennig
4. Understanding Self-Organization of the Brain,
   Dr. Anna Levina
5. Open Mic: Künstliche Intelligenz und Wir

LIVE-SET: STRÖME

Sampling (Monte Carlo) Methods

Recap from Monday

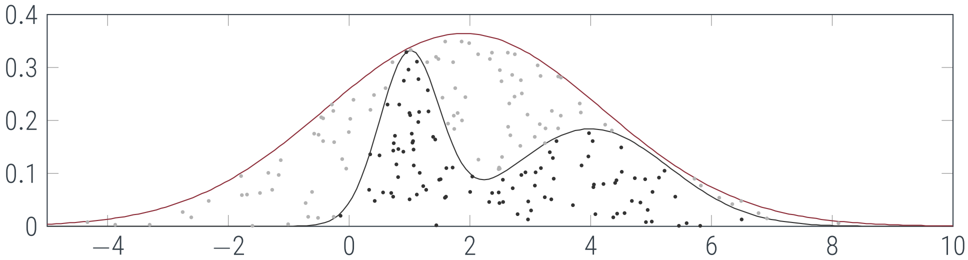
$$\mathbb{E}_p(f) = \int f(x)p(x)\,dx \approx \frac{1}{S}\sum_{s=1}^{S} f(x_s) := \hat{f} \quad \text{if} \quad x_s \sim p \quad \mathbb{E}[\hat{f}] = \mathbb{E}[f], \quad \operatorname{var}[\hat{f}] = \frac{\operatorname{var}(f)}{S}$$

Sampling is a way of performing rough probabilistic computations, in particular for **expectations** (including **marginalization**).

- samples from a probability distribution can be used to estimate expectations, roughly
- 'random numbers' don't need to be **unpredictable**, but rather **unstructured**
- uniformly distributed random numbers can be **transformed** into other distributions. This can be done numerically efficiently in some cases, and it is worth thinking about doing so
- Rejection sampling is a primitive but **exact** method that works with **intractable** models

Rejection Sampling

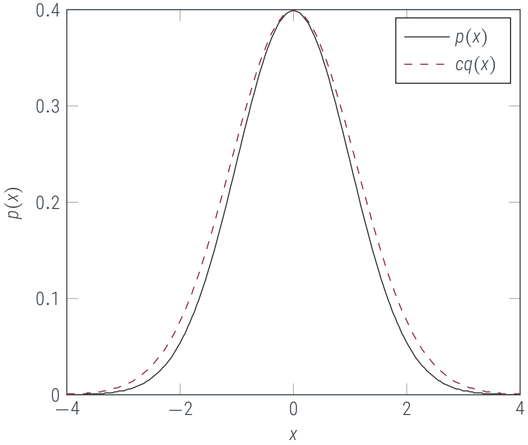
a simple method [Georges-Louis Leclerc, Comte de Buffon, 1707–1788]



- for any $p(x) = \tilde{p}(x)/Z$ (normalizer Z not required)
- choose $q(x)$ s.t. $cq(x) \geq \tilde{p}(x)$
- draw $s \sim q(x)$, $u \sim \text{Uniform}[0, cq(s)]$
- reject if $u > \tilde{p}(s)$

The Problem with Rejection Sampling

the curse of dimensionality [MacKay, §29.3]



Example:

- $p(x) = \mathcal{N}(x; 0, \sigma_p^2)$
- $q(x) = \mathcal{N}(x; 0, \sigma_q^2)$
- $\sigma_q > \sigma_p$
- optimal c is given by

$$c = \frac{(2\pi\sigma_q^2)^{D/2}}{(2\pi\sigma_p^2)^{D/2}} = \left(\frac{\sigma_q}{\sigma_p}\right)^D \exp\left(D \ln \frac{\sigma_q}{\sigma_p}\right)$$

- acceptance rate is ratio of volumes: $1/c$
- rejection rate rises **exponentially** in D
- for $\sigma_q/\sigma_p = 1.1$, $D = 100$, $1/c < 10^{-4}$

Importance Sampling

a slightly less simple method

- computing $\tilde{p}(x), q(x)$, then throwing them away seems wasteful
- instead, rewrite (assume $q(x) > 0$ if $p(x) > 0$)

$$\begin{aligned}\phi = \int f(x)p(x)\,\mathrm{d}x &= \int f(x)\frac{p(x)}{q(x)}q(x)\,\mathrm{d}x \\ &\approx \frac{1}{S}\sum_s f(x_s)\frac{p(x_s)}{q(x_s)} =: \frac{1}{S}\sum_s f(x_s)w_s \qquad \text{if } x_s \sim q(x)\end{aligned}$$

- this is just using a new function $g(x) = f(x)p(x)/q(x)$, so it is an unbiased estimator
- w_s is known as the **importance (weight)** of sample s
- if normalization unknown, can also use $\tilde{p}(x) = Zp(x)$

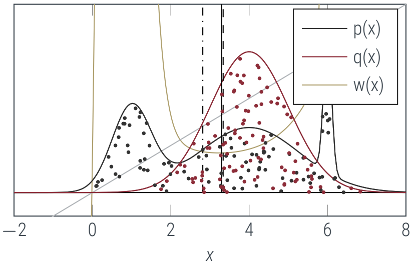
$$\begin{aligned}\int f(x)p(x) &= \frac{1}{Z}\frac{1}{S}\sum_s f(x_s)\frac{\tilde{p}(x_s)}{q(x_s)}\,\mathrm{d}x \\ &= \frac{1}{S}\sum_s f(x_s)\frac{\tilde{p}(x_s)/q(x_s)}{\frac{1}{S}\sum_{s'} 1\tilde{p}(x_s)/q(x_s)} =: \sum_s f(x_s)\tilde{w}_s\end{aligned}$$

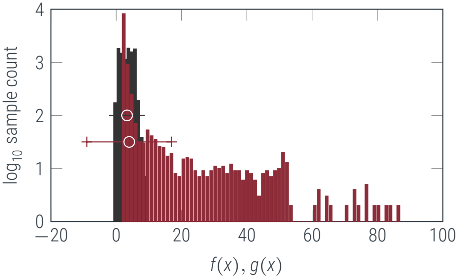
- this is consistent, but biased

What's wrong with Importance Sampling?

the curse of dimensionality, revisited

- recall that $\operatorname{var}\hat{\phi} = \operatorname{var}(f)/S$ – importance sampling replaces $\operatorname{var}(f)$ with $\operatorname{var}(g) = \operatorname{var}\left(f\frac{p}{q}\right)$
- $\operatorname{var}\left(f\frac{p}{q}\right)$ can be very large if $q \ll p$ somewhere. In many dimensions, usually all but everywhere!
- if p has "undiscovered islands", some samples have $p(x)/q(x) \to \infty$





Markov-Chain Monte Carlo

random walks drawing random numbers

- problem of importance sampling: samples generated independently, requires q good approximation to p everywhere.
- instead: generate samples **iteratively**, approximation q only needs to be good *locally*

Definition (Reminder: Markov Chains)

A joint distribution $p(X)$ over a sequence of random variabels $X := [x_0, \dots, x_N]$ is said to have **the Markov property** if

$$p(x_i \mid x_0, x_1, \dots, x_{i-1}) = p(x_t \mid x_{i-1}).$$

The sequence is then called a **Markov chain**.

An analogy

optimization

assume we wanted to find the maximum of $\tilde{p}(x)$

- given current estimate x_t
- draw proposal $x' \sim q(x' \mid x_t)$
- evaluate

$$a = \frac{\tilde{p}(x')}{\tilde{p}(x_t)}$$

- if $a \geq 1$, accept: $x_{t+1} \leftarrow x'$
- else stay: $x_{t+1} \leftarrow x_t$

Usually, throw away estimates at the end, only keep "best guess". But the estimates do contain information about the shape of $\tilde{p}$!

The Metropolis-Hastings* Method

* Authorship controversial. Likely inventors: M. Rosenbluth, A. Rosenbluth & E. Teller, 1953

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

we want to find representers (**samples**) of $\tilde{p}(x)$

- given current sample x_t
- **draw proposal** $x' \sim q(x' \mid x_t)$ (for example, $q(x' \mid x_t) = \mathcal{N}(x'; x_t, \sigma^2)$)
- **evaluate**

$$a = \frac{\tilde{p}(x')}{\tilde{p}(x_t)} \frac{q(x_t \mid x')}{q(x' \mid x_t)}$$

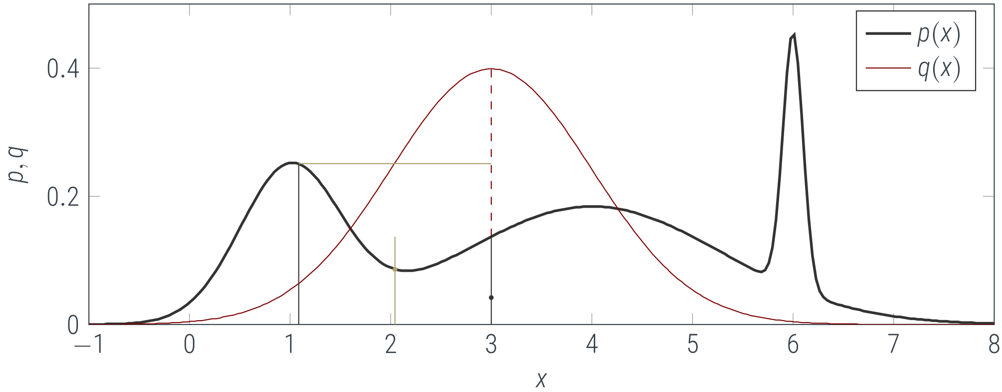
- if $a \geq 1$, **accept**: $x_{t+1} \leftarrow x'$
- else
  - **accept** with probability a: $x_{t+1} \leftarrow x'$
  - **stay** with probability $1 - a$: $x_{t+1} \leftarrow x_t$

Usually, assume symmetry $q(x_t \mid x') = q(x' \mid x_t)$ (the Metropolis method)

- no rejection. Every sample counts!
- like optimization, but with a chance to move downhill

Metropolis-Hastings in pictures

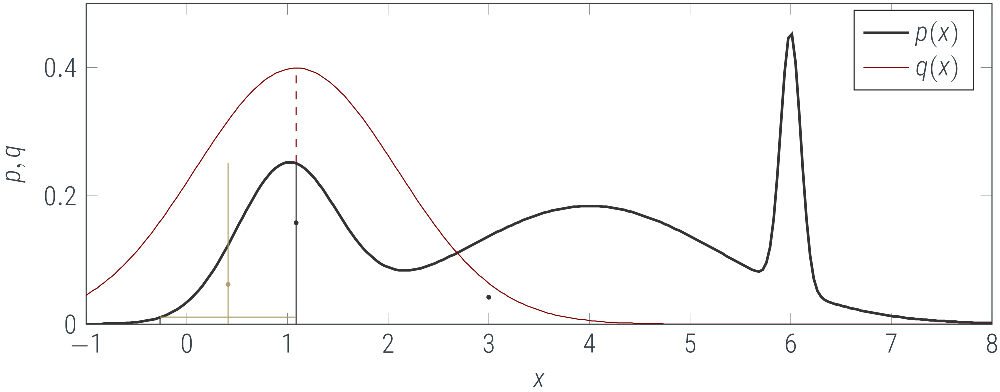
$t = 1$



$$a = \frac{\tilde{p}(x')}{\tilde{p}(x_t)} \frac{q(x_t \mid x')}{q(x' \mid x_t)} \qquad \text{accept with } p = \min(1, a)$$

Metropolis-Hastings in pictures

$t = 2$

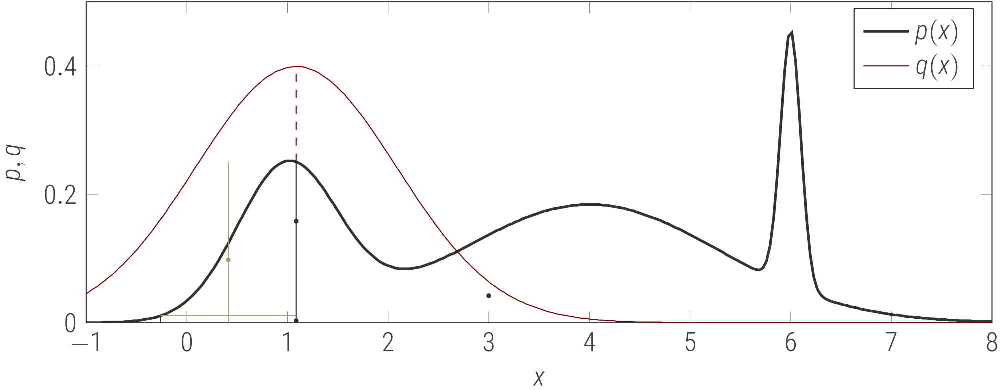


$$a = \frac{\tilde{p}(x')}{\tilde{p}(x_t)} \frac{q(x_t \mid x')}{q(x' \mid x_t)}$$

accept with $p = \min(1, a)$

Metropolis-Hastings in pictures

$t = 3$

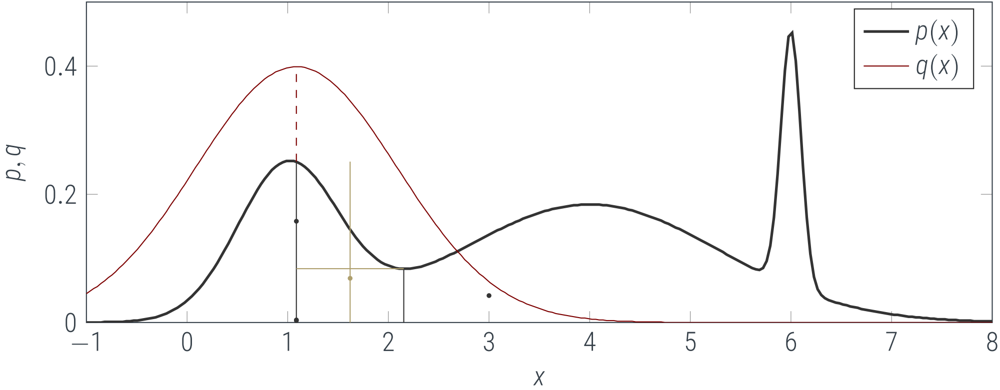


$$a = \frac{\tilde{p}(x')}{\tilde{p}(x_t)} \frac{q(x_t \mid x')}{q(x' \mid x_t)}$$

accept with $p = \min(1, a)$

Metropolis-Hastings in pictures

$t = 4$

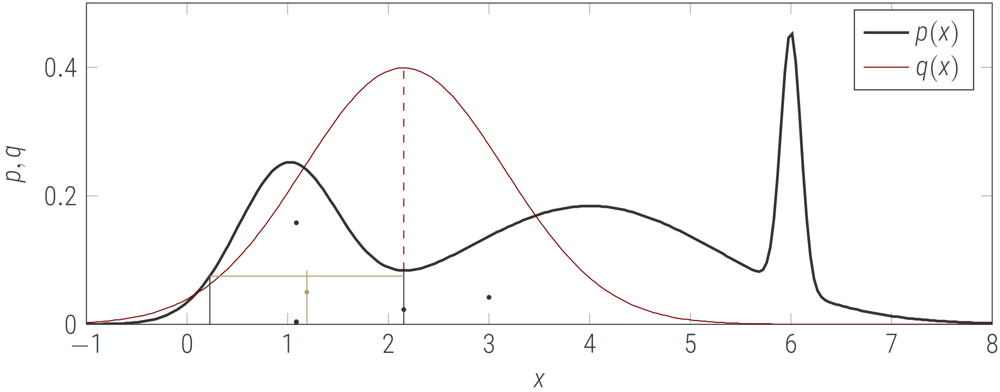


$$a = \frac{\tilde{p}(x')}{\tilde{p}(x_t)} \frac{q(x_t \mid x')}{q(x' \mid x_t)}$$

accept with $p = \min(1, a)$

Metropolis-Hastings in pictures

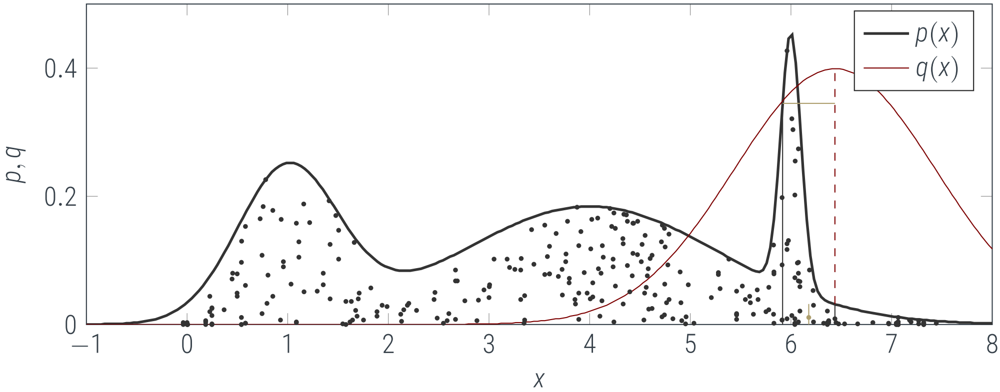
$t = 5$



$$a = \frac{\tilde{p}(x')}{\tilde{p}(x_t)} \frac{q(x_t \mid x')}{q(x' \mid x_t)} \qquad \text{accept with } p = \min(1, a)$$

Metropolis-Hastings in pictures

$t = 300$



$$a = \frac{\tilde{p}(x')}{\tilde{p}(x_t)} \frac{q(x_t \mid x')}{q(x' \mid x_t)} \qquad \text{accept with } p = \min(1, a)$$

Why is this a Monte Carlo Method?

MH draws from $p(x)$ in the limit of ∞ samples

proof (sketch) existence of stationary distribution: detailed balance

- MH satisfies detailed balance

$$
\begin{aligned}
p(x)T(x \to x') &= p(x) \cdot q(x' \mid x) \min\left[1, \frac{p(x')q(x \mid x')}{p(x)q(x' \mid x)}\right] \\
&= \min[p(x)q(x' \mid x), p(x')q(x \mid x')] \\
&= p(x') \cdot q(x \mid x') \min\left[\frac{p(x)q(x' \mid x)}{p(x')q(x \mid x')}, 1\right] \\
&= p(x')T(x' \to x)
\end{aligned}
$$

- Markov Chains satisfying detailed balance have **at least one *stationary distribution***

$$
\int p(x)T(x \to x')\,\mathrm{d}x = \int p(x')T(x' \to x)\,\mathrm{d}x = p(x') \int T(x' \to x)\,\mathrm{d}x = p(x')
$$

Why is this a Monte Carlo Method?

MH draws from $p(x)$ in the limit of ∞ samples

proof (sketch) uniqueness of stationary distribution:

Definition

Ergodicity A sequence $\{x_t\}_{t\in\mathbb{N}}$ is called **ergodic** if it

1. is *a-periodic* (contains no recurring sequence)
2. has *positive recurrence*: $x_t = x_*$ implies there is a $t' > t$ such that $p(x_{t'} = x_*) > 0$

→ $\{x_t\}_{t\in\mathbb{N}}$ is ergodic (by definition)

- ergodic Markov Chains have **at most one stationary distribution**

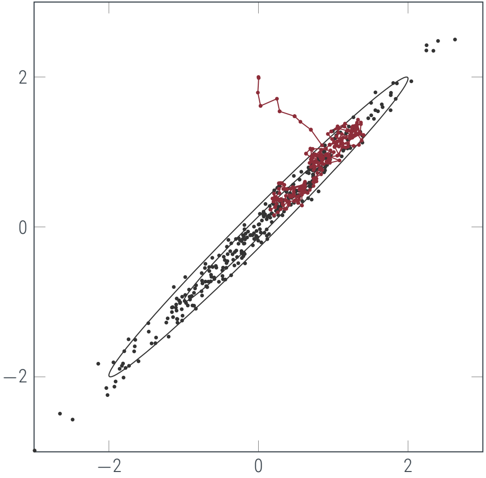
Theorem (convergence of Metropolis-Hastings, simplified)

If $q(x' \mid x_t) > 0 \ \forall (x', x_t)$, then, for any x_0, the density of $\{x_t\}_{t\in\mathbb{N}}$ approaches $p(x)$ as $t \to \infty$.

- this is not a statement about convergence rate!

Metropolis-Hastings performs a (biased) random walk

hence diffuses $\mathcal{O}(s^{1/2})$



Rule of Thumb: [MacKay, (29.32)]

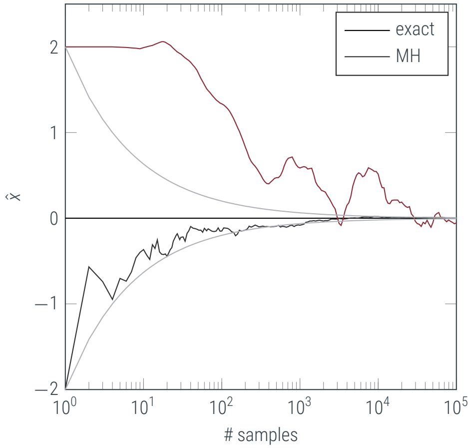
- typical use-case: high-dimensional D problem of largest length-scale L, smallest ε, isotropic proposal distribution
- have to set width of q to $\approx \varepsilon$, otherwise acceptance rate r will be very low.
- then Metropolis-Hastings does a **random walk** in D dimensions, moving a distance of $\sqrt{\mathbb{E}[\|x_t - x_0\|^2]} \sim \epsilon\sqrt{rt}$
- so, to create **one** independent draw at distance L, MCMC has to run for **at least**

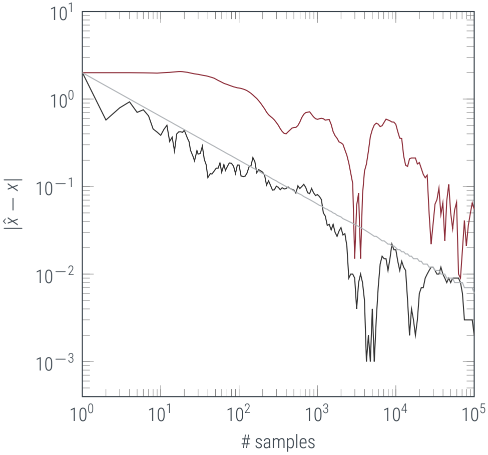
$$t \sim r\left(\frac{L}{\epsilon}\right)^2$$

iterations. In practice (e.g. if the distribution has *islands*), the situation can be **much** worse.

Metropolis-Hastings performs a (biased) random walk

estimating the mean of a correlated Gaussian





Gibbs Sampling

MCMC with optimal proposal distribution

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

[D. & S. Geman, 1984]

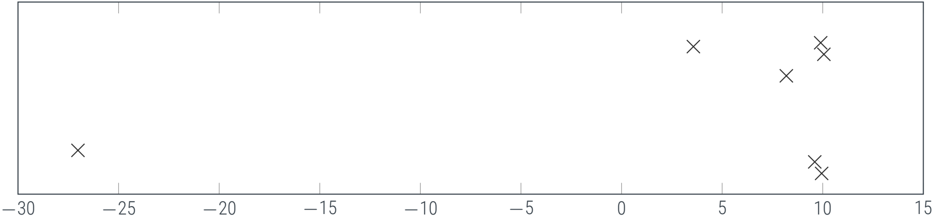
- $x_t \leftarrow x_{t-1}; x_{ti} \sim p(x_{ti} \mid x_{t1}, x_{t2}, \dots, x_{t(i-1)}, x_{t(i+1)}, \dots)$
- a special case of Metropolis-Hastings:
  - $q(x' \mid x_t) = \delta(x'_{\setminus i} - x_{t,\setminus i}) p(x'_i \mid x_{t,\setminus i})$
  - $p(x') = p(x'_i \mid x'_{\setminus i}) p(x'_{\setminus i}) = p(x'_i \mid x_{t,\setminus i}) p(x_{t,\setminus i})$
  - acceptance rate:

$$a = \frac{p(x')}{p(x_t)} \cdot \frac{q(x_t \mid x')}{q(x' \mid x_t)} \qquad = \frac{p(x'_i \mid x_{t,\setminus i}) p(x_{t,\setminus i})}{p(x_{ti} \mid x_{t,\setminus i}) p(x_{t,\setminus i})} \cdot \frac{q(x_t \mid x')}{\delta(x'_{\setminus i} - x_{t,\setminus i}) p(x'_i \mid x_{t,\setminus i})}$$

$$= \frac{q(x_t \mid x')}{p(x_{ti} \mid x_{t,\setminus i}) \delta(x'_{\setminus i} - x_{t,\setminus i})} \qquad = 1$$

The Seven Scientists

a simple example (DJC MacKay, Information Theory, Inference and Learning Algorithms, Ex. 22.15)



Scientist	A	B	C	D	E	F	G
x_n	-27.020	3.570	8.191	9.898	9.603	9.945	10.056

- likelihood: choose Gaussian, typical noise model

$$p(\boldsymbol{x} \mid \mu, \boldsymbol{\sigma}) = \prod_i^7 \mathcal{N}(x_i; \mu, \sigma_i^2) = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma_i^2}\right)$$

The Seven Scientists

setting up the model

- likelihood: choose Gaussian, typical noise model

$$p(\boldsymbol{x} \mid \mu, \boldsymbol{\sigma}) = \prod_i^7 \mathcal{N}(x_i; \mu, \sigma_i^2) = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma_i^2}\right)$$

- priors: **exponential families**
  - vague Gaussian prior on μ (e.g. $m = 0, s = 100$)

$$p(\mu) = \mathcal{N}(\mu; m, s^2) = \frac{1}{\sqrt{2\pi}s} \exp\left(-\frac{(\mu - m)^2}{2s^2}\right)$$

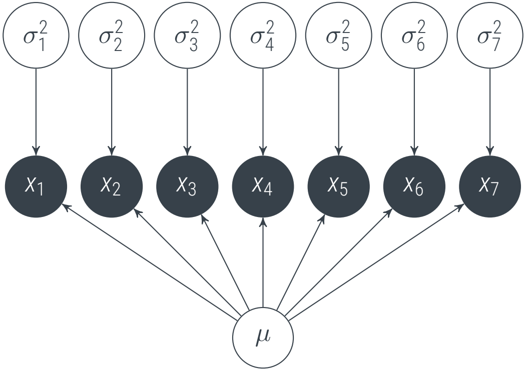
  - vague Gamma priors on σ_i^{-2} (e.g. $k = 1, \theta = 10$)

$$p(\sigma_i) = \mathcal{G}(\sigma_i^{-2}; k, \theta) = \frac{1}{\Gamma(k)\theta^k} (\sigma_i^{-2})^{k-1} \exp\left(-\frac{\sigma_i^{-2}}{\theta}\right)$$

(there are good algorithms for sampling from $\mathcal{G}(k, \theta)$).

The Seven Scientists

graphical model



$$p(\boldsymbol{x}, \boldsymbol{\sigma}, \mu) = \mathcal{N}(\mu; m, v^2) \prod_i \mathcal{N}(x_i; \mu, \sigma_i^2) \mathcal{G}(\sigma_i^{-2}; a, b)$$

The Seven Scientists

how about maximum likelihood?

- Can we get away without priors? Maximum likelihood inference

$$\log p(\boldsymbol{x} \mid \mu, \boldsymbol{\sigma}) = \sum_i -\frac{(x_i - \mu)^2}{2\sigma_i^2} - \frac{1}{2}\log \sigma_i^2 - \frac{1}{2}\log 2\pi$$

$$\frac{\partial p(\boldsymbol{x} \mid \mu, \boldsymbol{\sigma})}{\partial \mu} = \sum_i \frac{(x_i - \mu)}{\sigma_i^2} = 0 \quad \Rightarrow \quad \mu_* = \frac{\sum_i x_i/\sigma_i^2}{\sum_i 1/\sigma_i^2}$$

$$\frac{\partial p(\boldsymbol{x} \mid \mu, \boldsymbol{\sigma})}{\partial \sigma_i} = \frac{(x_i - \mu)^2}{\sigma_i^3} - \frac{1}{\sigma_i} = 0 \quad \Rightarrow \quad \sigma_i \rightarrow 0$$

- maximum likelihood solution: choose arbitrary i, set $\sigma_i \rightarrow 0$, $\mu \rightarrow x_i$

The Seven Scientists

conditional independences

- if $\boldsymbol{\sigma}$ were known, posterior on μ would be **analytical conjugate prior!** (1)

$$p(\mu \mid \mathbf{x}, \boldsymbol{\sigma}) \propto \mathcal{N}(\mu; m, s^2) \prod_i^7 \mathcal{N}(x_i; \mu, \sigma_i^2) \propto \mathcal{N}\left[\mu; \psi^2 \left(\frac{m}{s^2} + \sum_i \frac{x_i}{\sigma_i^2}\right), \psi^2 = \left(\frac{1}{s^2} + \sum_i \frac{1}{\sigma_i^2}\right)^{-1}\right]$$

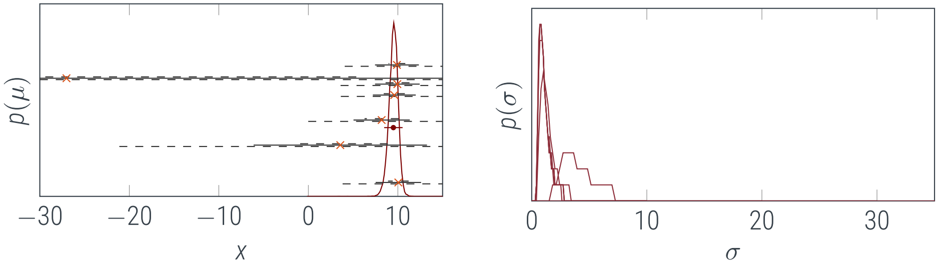
- if μ were known, posterior on $\boldsymbol{\sigma}$ would be **analytical conjugate prior!** (2)

$$p(\boldsymbol{\sigma} \mid \mathbf{x}, \mu) \propto \prod_i^7 \mathcal{N}(x_i; \mu, \sigma_i^2)\mathcal{G}(\sigma_i^{-2}; k, \theta) \propto \prod_i \mathcal{G}\left[\sigma_i^{-2}; k + \frac{1}{2}, \left(\frac{1}{\theta} + \frac{(x_i - \mu)^2}{2}\right)^{-1}\right]$$

→ Gibbs sampling: fix some $\boldsymbol{\sigma}_0$ (e.g. from prior). repeat:
  - draw $\mu_t \sim p(\mu \mid \mathbf{x}, \boldsymbol{\sigma}_{t-1})$ from (1)
  - draw $\boldsymbol{\sigma}_t \sim p(\boldsymbol{\sigma} \mid \mathbf{x}, \mu_t)$ from (2)

The Seven Scientists

sampling posterior from $3 \cdot 10^4$ samples



Further Reading

A field of its own. [MacKay, chapters 29 & 30]

Hamiltonian Monte Carlo (aka hybrid Monte Carlo)

- introduce momentum variables to reduce diffusion (requires gradient of p)
- various adaptations, e.g. for local shape of p ("Riemannian MCMC")
- NUTS (the No U-Turn Sampler) (M. Hoffman & A. Gelman, JMLR 15, 2014): a parameter free HMC method, currently arguably the gold standard for models allowing auto-diff

Slice Sampling

- very efficient (exponentially fast) exploration in one dimension
- almost no free parameters (but problems in many dimensions)
- elliptical slice sampling [Murray et al., 2010]: Very efficient for Gaussian priors, no free parameters

But:

- in nontrivial situations no sampling method except exact sampling gives exact finite-time bounds
- diagnostic tricks exist, but are not without flaws

Summary

Markov Chain Monte Carlo (MCMC) methods

- Methods drawing $\{x_t\}_{t\in\mathbb{N}}$ from $p(x_t \mid x_{t-1})$ are called **Markov Chains**
- if the density of $\{x_t\}$ converges to $p(x)$ as $t \to \infty$, the method is called **Markov Chain Monte Carlo (MCMC)**
- **Metropolis-Hastings** is a simple MCMC method requiring **few assumptions**
- **Gibbs** sampling is MH with optimal proposal distributions, requiring analytical samples from **conditional distributions**
- improved, more elaborate methods exist
- all MCMC methods suffer from **diffusion**, which can be very difficult to detect in practice
- MCMC approach **exact** answers after an **unknown** time