Advanced Machine Learning

Bert Kappen
SNN Radboud University, Nijmegen
Gatsby Unit, UCL London

November 24, 2020
Course setup

- 6 ec course

- examination based on computer exercises and presentation of research topic

- Students work in groups of three and hand in a single result

- weekly exercises tutorials

- All course materials (slides, exercises) and schedule via [http://www.snn.ru.nl/~bertk/machinelearning/adv_ml.html](http://www.snn.ru.nl/~bertk/machinelearning/adv_ml.html)
Lecture 1. Monte Carlo sampling

Based on MacKay chapter 29 and 30.

- Uniform sampling
- Importance and Rejection sampling
- Metropolis-Hasting method
- Detour on Markov processes, stationary distribution, ergodicity
- Gibbs sampling
- Hybrid Monte Carlo
Machine learning is about inference

In many machine learning problems one needs to estimate high dimensional integrals.

1. Estimate expected parameter values in the posterior of a Bayesian learning problem:

\[
p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \quad \mathbb{E}_{\theta} = \int d\theta p(\theta|D)
\]

2. Estimate the ‘evidence’ for model selection

\[
p(D|H_i) = \int d\theta p(D|\theta, H_i)p(\theta|H_i)
\]

3. Computing statistics in graphical models \( p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_i(x_i|x_{Pa(i)}) \)

\[
p(x_1, x_2) = \sum_{x_3, \ldots, x_n} p(x_1, \ldots, x_n)
\]
Denote \( p(x) = \frac{p^*(x)}{Z} \) the probability distribution of interest with \( Z = \int dx p^*(x) \). We can assume that \( p^*(x) \) is easily evaluated for any \( x \), but \( Z \) is hard to evaluate.

Problems 1,3 are of the form: Estimate

\[
\Phi = \int dx p(x) \phi(x)
\]

with \( \phi(x) \) some function of \( x \). Problem 2 is of the form estimate \( Z = \int dx p^*(x) \).

All these problems are hard, but problem 2 (computing the partition sum) is actually harder than problems 1,3 (computing statistics).\(^1\)

We will consider estimating \( \Phi \).

\(^1\)For problem 2, more advanced methods, such as thermodynamic integration are needed.
An analogy

Imagine a lake with coordinates \( x = (x_1, x_2) \). You wish to estimate the average plankton concentration\(^2\)

\[
\Phi = \frac{1}{Z} \sum_x p^*(x)\phi(x) \quad Z = \sum_x p^*(x)
\]

\( p^*(x) \) is the depth of the lake at \( x \). \( \phi(x) \) is the plankton concentration at \( x \) (independent of the depth). So \( p^*(x)\phi(x) \) is the total amount of plankton at location \( x \).

\( Z \) is the total volume of the lake.

\(^2\)We assume for the simplicity of the argument that the lake has discrete locations labeled by \( x \).
An analogy

Go out with your boat to locations $x$, measure the depth with a plumbline and the local plankton concentration. If you sample all locations in the lake,

$$\hat{\Phi} = \frac{\sum x p^*(x) \phi(x)}{\sum x p^*(x)} = \Phi$$

you obtain the exact result.

But this is very expensive. If the lake has $a$ locations per dimension then this requires $a^2$ measurements. For an $n$ dimensional problem, the number becomes $a^n$ exponential in $n$. This is called the curse of dimensionality.
### An analogy

Suppose instead of visiting all locations, we visit \( N \) locations. A very good (actually the best) way to choose the locations is to choose them with a probability proportional to the depth \( p(x) = p^*(x)/Z \). Of course, this is impractical, because you need to measure first the depth at all lake locations before you can start rolling your dice!

But suppose you can, then,

\[
\hat{\Phi} = \frac{1}{N} \sum_{r=1}^{N} \phi(x^r) \quad x^r \sim p(x)
\]

\( \hat{\Phi} \) is a random variable: each time we repeat the experiment (collecting \( N \) samples) we get a different outcome. But it is a good estimator of \( \Phi \) because the average value over many such experiments is \( \mathbb{E}\hat{\Phi} = \Phi \). The estimator \( \hat{\Phi} \) is called unbiased.

The proof is simple. Since the samples are drawn independent, the probability to obtain outcomes \( x^1, \ldots, x^N \) is \( p(x^1, \ldots, x^N) = \prod_{r=1}^{N} p(x^r) \). The expected value of \( \hat{\Phi} \)
is

\[ \mathbb{E} \hat{\Phi} = \sum_{x^1, \ldots, x^N} \prod_{r=1}^{N} p(x^r) \hat{\Phi}(x^1, \ldots, x^N) = \frac{1}{N} \sum_{r=1}^{N} \sum_{x^r} p(x^r) \phi(x^r) = \Phi \]

The variance in \( \hat{\Phi} \) decreases with the number of samples \( N \):

\[ \nabla \hat{\Phi} = N \nabla \frac{\phi}{N} = \frac{1}{N} \nabla \phi \quad \nabla \phi = \sum_{x} p(x)(\phi(x) - \Phi)^2 \]

Thus,

\[ \hat{\Phi} = \Phi + O\left(\frac{1}{\sqrt{R}}\right) \]

Note, that the accuracy does not depend on the dimension of the problem. So this is great. But sampling from \( p \) is very hard!

\(^3\)Here we make use of the useful property that if \( X = \sum_i X_i \) is a sum of independent random variables, the variance in \( X \) is the sum of the variances of each \( X_i \).
Typical set

Consider the Ising model

$$p(x) = \frac{\exp(\beta \sum_{i>j} x_i x_j w_{ij})}{Z} = \frac{p^*(x)}{Z} \quad Z = \sum_x p^*(x)$$

with $x = (x_1, \ldots, x_n)$ and $x_i = \pm 1, i = 1, \ldots, n$.

Total number of states is $2^n$, but most probability is concentrated in the so-called typical set $T$. 
**Typical set**

Draw long sequence of $N$ states from the Ising model:

$$x^1, x^2, \ldots, x^N$$

The probability of the sequence is

$$P(x^1, x^2 \ldots, x^N) = P(x^1)P(x^2) \ldots P(x^N) = \prod_x p(x)^{N(x)}$$

The number of times the value $x$ occurs in the string is $N(x) \approx p(x)N$.

Thus,

$$P(x^1, x^2 \ldots, x^N) \approx \prod_x p(x)^{p(x)N} = \left(2^{-H}\right)^N$$

with $H = - \sum_x p_x \log p_x$ the entropy of the distribution.

Thus, in the large $N$ limit all typical strings have the same probability $2^{-NH}$ and all other (non-typical) strings have probability zero.
The formula suggests that 'on average' for a single sample, there are typical samples with probability $2^{-H}$ and non-typical samples with probability zero. Denote $T$ the set of typical samples. The number of typical samples is thus $|T| \approx 2^H$.

The typical set should be compared to the total number of states $2^n$ and the volume fraction is $2^{H-n}$. If we sample states $x$ uniformly at random, the probability to hit an element of the typical set is $2^{H-n}$. Thus, one needs on the order of $N_{\text{min}} = 2^{n-H}$ samples to hit the typical set once and therefore the number of samples

$$N \gg N_{\text{min}} = 2^{n-H}$$

For $n$ binary spins $0 \leq H \leq n$.

Left. Entropy of a 64 spin 2d Ising model with couplings $w_{ij} = 1$. Right. Sample from a 1024 spin Ising model near the critical temperature.
• For (very) high temperature (low \( \beta \)) \( H \approx n \) and \( N_{\text{min}} \approx 1 \) and uniform sampling feasible.

• Around the critical temperature, \( H \approx n/2 \) and \( N_{\text{min}} \approx 2^{n/2} \). For \( n = 1000 \) spins this is of order \( 10^{150} \) which is about the square of the number of particles in the universe.

In a nut shell: Uniform sampling only works for uniform distributions (in which case it is optimal!).
Sampling from a multi-variate Gaussian

How to sample from a multi-variate Gaussian distribution $\mathcal{N}(x|m, S)$?

Define $C$ such that $CC^T = S$ and define $y = C^{-1}(x - m)$. Since this is a linear transformation, $y$ is also Gaussian distributed with

$$
\mathbb{E}y = C^{-1}(\langle x \rangle - m) = 0
$$

$$
\mathbb{V}(y) = \mathbb{E}yy^T = C^{-1}\mathbb{E}(x - m)(x - m)^T (C^{-1})^T = C^{-1}S(C^{-1})^T = 1
$$

Thus sample from $y$ can be obtained by sampling each component independently:

$$
p(y) = \prod_i \mathcal{N}(y_i|0, 1)
$$

Sampling from a one dimensional Gaussian can be done by sampling from a uniform distribution using the Box-Muller transformation (Exercise).

Sample from $x$ is obtained by sampling from $y$ and $x = Cy + m$. 
Importance sampling

The task is to estimate

$$\Phi = \int dx p(x) \phi(x)$$

Sample from another distribution \(q(x)\).

Often one can propose a sample density that is
1) better than uniform and
2) easy to sample from.

For instance, a (spherical) Gaussian:

$$Q^*(x) \propto \exp(- \sum_i x_i^2 / 2)$$
Consider simple 1-d sampling problem. Given $p(x)$, compute

$$\Phi = \text{Prob}(x < 0) = \int_{-\infty}^{\infty} \phi(x)p(x)dx$$

with $\phi(x) = 1$ if $x \leq 0$ and zero otherwise.

Naive method: generate $N$ samples $X_i \sim p$

$$\hat{\Phi} = \frac{1}{N} \sum_{i=1}^{N} \phi(X_i)$$
Importance sampling

Consider another distribution \( q(x) \). Then

\[
\Phi = \text{Prob}(x < 0) = \int_{-\infty}^{\infty} \phi(x) \frac{p(x)}{q(x)} q(x) dx
\]

Importance sampling: generate \( N \) samples \( X_i \sim q \)

\[
\hat{\Phi} = \frac{1}{N} \sum_{i=1}^{N} \phi(X_i) \frac{p(X_i)}{q(X_i)}
\]

Unbiased (\( E\hat{\Phi} = \Phi \)) for any \( q \)!
The distribution

\[ q^*(x) = \frac{p(x)\phi(x)}{\Phi} \]

is the optimal importance sampler.

One sample \( X \sim q^* \) is sufficient to estimate \( a \):

\[ \hat{\Phi} = \phi(X) \frac{p(X)}{q^*(X)} = \Phi \]

The estimator has zero variance.
Normalization

So any importance sampler is unbiased \( E\hat{\Phi} = \Phi \), and the closer that \( q \) is to \( p \) the smaller the variance.

The importance weights \( p(x)/q(x) \) assume that we can evaluate \( p(x) \) (\( q \) we choose so that we can). However, most often \( p(x) \) can only be easily computed, up to a constant:

\[
p(x) = \frac{p^*(x)}{Z} \quad Z = \int dx p^*(x)
\]

Estimate both numerator and denominator by sampling.

\[
\Phi = \int dx p(x) \phi(x) = \frac{\int dx p^*(x) \phi(x)}{\int dx p^*(x)} = \frac{\int dx q(x) \frac{p^*(x)}{q(x)} \phi(x)}{\int dx q(x) \frac{p^*(x)}{q(x)}}
\]

Sample \( \{x^r\} \) from \( q(x) \) and compute

\[
w_r = \frac{p^*(x^r)}{q(x^r)} \quad \hat{\Phi} = \frac{\sum_r w_r \phi(x^r)}{\sum_r w_r}
\]
Importance sampling

The estimate is biased.

$$\mathbb{E} \hat{\Phi} = \mathbb{E} \left( \frac{\sum_{r=1}^{N} w_r \phi(x^r)}{\sum_{r=1}^{N} w_r} \right) \neq \frac{\mathbb{E} \sum_{r=1}^{N} w_r \phi(x^r)}{\mathbb{E} \sum_{r=1}^{N} w_r} = \frac{N \int dx p^*(x) \phi(x)}{N \int dx p^*(x)} = \Phi$$

However, for large $N$:

$$\sum_{r=1}^{N} w_r = N \mathbb{E} w_r + O(\sqrt{N}) \approx N \mathbb{E} w_r$$

$$\mathbb{E} \hat{\Phi} \approx \mathbb{E} \left( \frac{\sum_{r=1}^{N} w_r \phi(x^r)}{N \mathbb{E} w_r} \right) = \frac{\mathbb{E} \sum_{r=1}^{N} w_r \phi(x^r)}{N \int dx p^*(x)} = \frac{N \int dx p^*(x) \phi(x)}{N \int dx p^*(x)} = \Phi$$

The estimator $\hat{\Phi}$ is asymptotically unbiased.
Choose sufficiently broad importance sampler

Warning: rare events give large contributions. Compare

\[ q(x) \propto e^{-x^2/2} \quad q(x) \propto \frac{1}{x^2 + a^2} \]

The weights are \( \propto q(x)^{-1} \): Rare events have large weights.

Toy 1-d amino acid sampling problem, showing effect of poor importance sampler. Left: Gaussian suggests convergence after 500 iterations. Right: Cauchy is broader distribution and shows more robust behaviour.
**Rejection sampling**

Choose, \( c \) such that for all \( x : cQ^*(x) > P^*(x) \)

- generate \( x \) from \( Q^*(x) \)
- generate \( u \) uniform from \([0, cQ^*(x)]\)
- if \( u > P^*(x) \) reject \( x \), otherwise accept \( x \)

Probability of a sample \( x \) is \( Q^*(x) \frac{P^*(x)}{cQ^*(x)} \propto P^*(x) \).

\[
\hat{\Phi} = \sum_r \phi(x'^r) \to \int dx \phi(x)p(x)
\]
Let $p(x)$ and $q(x)$ be spherical Gaussians in $n$ dimensions with mean 0 and $\sigma_q = 1.01\sigma_p$. 
Since

\[
q(x) = \left( \frac{1}{\sqrt{2\pi\sigma^2_q}} \right)^n e^{-\frac{1}{2\sigma^2_q} \sum_i x_i^2} \quad \quad p(x) = \left( \frac{1}{\sqrt{2\pi\sigma^2_p}} \right)^n e^{-\frac{1}{2\sigma^2_p} \sum_i x_i^2}
\]

then

\[
c = \frac{p(0)}{q(0)} = \left( \frac{\sigma_q}{\sigma_p} \right)^n = 1.01^n
\]

With \( n = 1000 \) we find \( c = 20.000 \). Thus volume under \( cq \) is 20.000 times the volume under \( p \). Therefore, the acceptance rate = \( \frac{\text{volume } p}{\text{volume } cq} = \frac{1}{c} \)

Thus rejection sampling is inefficient in high dimensions.

A similar argument holds for importance sampling.
Lecture 2. Monte Carlo sampling

Based on MacKay chapter 29 and 30.

- Uniform sampling
- Importance and Rejection sampling
- Metropolis-Hasting method
- Detour on Markov processes, stationary distribution, ergodicity
- Gibbs sampling
- Hybrid Monte Carlo
- Sampling the Bayesian posterior for the perceptron
Metropolis algorithm

The Metropolis algorithm (1956) considers a sampling density which depends on the current sample value: \( q(x|x') \).

Initialize in some random state \( x^1 \)
Metropolis algorithm

At iteration $r$, sample $x'$ from $q(x'|x^r)$ and compute

$$a = \frac{p^*(x')q(x^r|x')}{p^*(x^r)q(x'|x^r)}$$

If $a \geq 1$, accept $x'$ as the new state: $x^{r+1} = x'$

Else, accept $x'$ as the new state with probability $a$

If accept: $x^{r+1} = x'$, else $x^{r+1} = x^r$
Markov processes

Let $S$ denote the set of all state vectors $x$. $x \in S$ is a binary vector of length $n$ and thus $x$ can take on $2^n$ different values.

Denote $p(x)$ a probability distribution over states. $p(x)$ is a vector of length $2^n$ with $\sum_x p(x) = 1$.

A Markov process is a discrete time stochastic dynamical process that is defined by a transition matrix $T(x'|x)$ that specifies the transition probability from an initial state $x$ to a final state $x'$ in a single time step. Thus, $T(x'|x)$ is a $2^n \times 2^n$ matrix. For any $x$, $T(x'|x)$ is a probability vector in $x'$:

$$\sum_{x'} T(x'|x) = 1$$

Matrices with this property are called stochastic matrices.

The Markov dynamics with initial state distribution $p_0(x)$ is

$$p_{t+1}(x') = \sum_x T(x'|x)p_t(x)$$
Markov processes

After long times, the Markov process reaches a stationary distribution.

\[ p_0 \rightarrow p_1 = T p_0 \rightarrow p_2 = T p_1 \rightarrow \ldots \]

The stationary distribution \( p_\infty \) is a vector such that it is invariant under the dynamics: \( T p_\infty = p_\infty \). Thus \( p_\infty \) is an eigenvector of \( T \) with eigenvalue 1. \( T \) always has at least one eigenvalue 1.

In the case that \( T \) has only one eigenvalue 1 the stationary distribution is unique. In this case the Markov process is called ergodic. A consequence is that the stationary distribution does not depend on the initial state.

If the eigenvalue 1 is degenerate the stationary distribution is not unique and the stationary distribution depends on the initial state.

The characteristic time it takes to reach stationarity is determined by the other eigenvalues of \( T \).
Let us denote the eigenvalues and left and right eigenvectors of $T$ by $\lambda_\alpha, l_\alpha, r_\alpha, \alpha = 1, \ldots, 2^n$, respectively.\footnote{In general, the number of eigenvalues of $T$ can be less than $2^n$. However, for our purposes we can ignore this case} In matrix notation we have

$$
Tr_\alpha = \lambda_\alpha r_\alpha \\
l_\alpha^\dagger T = \lambda_\alpha l_\alpha^\dagger
$$

Since $T$ is a non-symmetric matrix, the left and right eigenvectors are different, non-orthogonal and complex valued. $\dagger$ denotes complex conjugation and transpose. The eigenvalues are complex valued. Under rather general conditions each set of eigenvectors spans a non-orthogonal basis of $C^{2n}$. These two bases are dual in the sense that:

$$l_\alpha^\dagger r_\beta = \delta_{\alpha\beta}.$$ 

We see that the vector $a = (1, \ldots, 1)$ is a left eigenvector of $T$ with eigenvalue 1:

$$
(aT)(x) = \sum_{x'} a(x')T(x'|x) = \sum_{x'} T(x'|x) = 1 = a(x)
$$

The corresponding right eigenvector is $p_\infty$: $Tp_\infty = p_\infty$.\footnote{In general, the number of eigenvalues of $T$ can be less than $2^n$. However, for our purposes we can ignore this case}
Some properties

A Markov process is called\textit{ irreducible}, or \textit{ergodic}, on a subset of states $C \subset S$ if for any state $x \in C$ there is a finite probability to visit any other state $x' \in C$:

$$x = x^0, x^1, \ldots, x^k = x'$$

with $T(x^i|x^{i-1}) > 0$, $i = 1, \ldots, k$.

A subset of states $C \subset S$ is called \textit{closed} when the Markov process can never escape from $C$, once entered:

$$T(x'|x) = 0 \quad \text{for all } x \in C, x' \not\in C.$$  

In general, we can decompose the state space $S$ uniquely into closed irreducible subsets $C_i$

$$S = \mathcal{T} \cup C_1 \cup C_2 \ldots,$$

where $\mathcal{T}$ is a set of \textit{transient states}.
An irreducible Markov process $T$ of *periodicity* $d$ has $d$ eigenvalues given by

$$\lambda_m = \exp(2\pi im/d), \quad m = 0, \ldots, d - 1,$$

and all remaining eigenvalues of $T$ are inside the unit circle in the complex plane: $|\lambda_\alpha| < 1$.

In particular, $T$ has exactly one eigenvalue 1. Its corresponding right eigenvector is equal to the (unique) stationary distribution (ergodicity).
Characteristic times

We can expand $T$ on the basis of its eigenvectors:

$$T = \sum_{\alpha=1}^{2^n} \lambda_\alpha r_\alpha l_\alpha^\dagger$$

In the ergodic case with periodicity 1:

$$p_t = T p_{t-1} = \ldots = T^t p_0 = \sum_\alpha \lambda_\alpha r_\alpha l_\alpha^\dagger p_0 = p_\infty + \sum_{\alpha>1} \lambda_\alpha r_\alpha (l_\alpha^\dagger p_0)$$

It is easy to check that this satisfies the eigen equations:

$$Tr_\beta = \sum_\alpha \lambda_\alpha r_\alpha l_\alpha^\dagger r_\beta = \sum_\alpha \lambda_\alpha r_\alpha \delta_{\alpha \beta} = \lambda_\beta r_\beta$$

and similarly $l_\beta^\dagger T = \lambda_\beta l_\beta^\dagger$.  

---

5It is easy to check that this satisfies the eigen equations:
We can write

\[ \lambda^t_\alpha = |\lambda_\alpha|^t e^{i\phi_\alpha t} = e^{-t/\tau_\alpha} e^{i\phi_\alpha t} \quad \tau_\alpha = \frac{-1}{\log |\lambda_\alpha|} \]

The characteristic time to converge to the stationary distribution is determined by the largest $|\lambda_\alpha|$.
Non-ergodic behavior

A non-irreducible or non-ergodic Markov process has more than one eigenvalue 1 and therefore more than one left and right eigenvector with eigenvalue 1. Let us denote these eigenvectors by $l_1, \ldots, l_k$ and $r_1, \ldots, r_k$, respectively. Any linear combination of the right eigenvectors

$$p_\infty = \sum_{\alpha=1}^{k} \rho_\alpha r_\alpha$$

is therefore a stationary distribution, with parameters $\rho_\alpha$ such that $p_\infty(x) \geq 0$ for all $x$ and proper normalization: $\sum_x p_\infty(x) = 1$. Thus, there exists a manifold of dimension $k - 1$ of stationary distributions.
The $k$ left eigenvectors with eigenvalue 1 encode invariants of the dynamics

$$l_\alpha^\dagger p_{t+1} = l_\alpha^\dagger T p_t = l_\alpha^\dagger p_t \quad \alpha = 1, \ldots, k$$

Since $l_1 \propto (1, \ldots, 1)$ the first invariant simply ensures invariance of normalisation

$$\sum_x p_{t+1}(x) = \sum_x p_t(x).$$

Thus,

$$l_\alpha^\dagger p_0 = l_\alpha^\dagger p_\infty \quad \alpha = 1, \ldots, k$$

The invariants determine uniquely the stationary distribution in terms of the initial conditions.

$$p_\infty = \sum_{\alpha=1}^k \rho_\alpha r_\alpha = \sum_{\alpha=1}^k (l_\alpha^\dagger p_0)r_\alpha$$

because $\rho_\alpha = l_\alpha^\dagger p_\infty = l_\alpha^\dagger p_0$.

Note, that in the ergodic case ($k = 1$) the dependence on the initial state disappears, as it should, since $l_1^\dagger p_0 = 1$ for any initial distribution $p_0$
Summary

The Markov process can be analysed in terms of the eigenvalues and eigenvectors of the transition matrix $T$.

1. There exists always an eigenvalue $\lambda = 1$.
   - If this eigenvalue is non-degenerate, the Markov process is called ergodic. The stationary distribution is unique and independent on the initial state.
   - If this eigenvalue is degenerate, the Markov process is called non-ergodic. The stationary distribution is not unique and depends on the initial state.

2. Markov processes can multiple eigenvalues with $|\lambda| = 1$. In this case it is called periodic. In the most common case, the Markov process is non-periodic and has only $\lambda = 1$ on the rim.

3. Eigenvalues close to 1, imply long convergence times.
**Detailed balance**

Stationarity:

\[ \sum_{x'} T(x|x')p(x') = p(x) = \sum_{x'} T(x'|x)p(x) \]

There always exists a stationary solution \( p(x) \).

The Markov process \( T \) satisfies detailed balance if \( \exists p \) such that

\[ T(x|x')p(x') = T(x'|x)p(x) \text{ for all } x, x'. \]

Detailed balance is a property of the Markov dynamics \( T \). It is not true in general.

DB holds for thermodynamic systems and Metropolis Hastings, not for chaotic systems.

If DB \( p(x) \) is a stationary distribution of \( T \). The reverse is not true.
The Metropolis algorithm again

At iteration $r$, sample $x'$ from $q(x'|x^r)$ and compute

$$a = \frac{p^*(x')q(x^r|x')}{{p^*(x^r)}q(x'|x^r)}$$

If $a \geq 1$, accept $x'$ as the new state: $x^{r+1} = x'$

Else, accept $x'$ as the new state with probability $a$

If accept: $x^{r+1} = x'$, else $x^{r+1} = x^r$

The Metropolis algorithm is an ergodic (aperiodic) Markov process that satisfies detailed balance such that the stationary distribution is $p$. 
Convergence of Metropolis algorithm

Metropolis algorithm is example of Markov process. Given two states $x$ and $x'$. Define

$$a_{x'x} = \frac{p^*(x')q(x|x')}{p^*(x)q(x'|x')}, \quad a_{xx'} = \frac{1}{a_{x'x}}$$

Suppose $a_{x'x} \geq 1$. Then

**Given** $x$, the probability to accept $x'$ is

$$T(x'|x) = q(x'|x)$$

**Given** $x'$, the probability to accept $x$ is

$$T(x|x') = q(x|x')a_{xx'}$$

$$\frac{T(x'|x)}{T(x|x')} = a_{x'x} \frac{q(x'|x)}{q(x|x')} = \frac{p^*(x')}{p^*(x)} = \frac{p(x')}{p(x)},$$

i.e. detailed balance. This implies that the process $T(x'|x)$ converges to $p(x)$.
Metropolis Hasting for the Ising model

Use MH to sample from the Ising model

\[
p(x) = \frac{1}{Z} \exp \left( -E(x) \right) \quad E(x) = -\frac{1}{2} \sum_{i \neq j} w_{ij}x_ix_j = -\sum_{(ij)} w_{ij}x_ix_j
\]

Consider \( q(x'|x) \) to implement single spin flips. Define \( F_i x \) is the vector of spins obtained by flipping bit \( i \). Then

\[
q(F_i x|x) = \frac{1}{n} \quad q(x|x) = 0
\]

The MH ratio for the transition from state \( x \) to state \( F_i x \) is

\[
a_{x,F_i x} = \frac{p^*(F_i x)}{p^*(x)} = \exp \left( -\Delta E_{x,F_i x} \right)
\]

\[
\Delta E_{x,F_i x} = E(F_i x) - E(x) = 2x_i \sum_{j \neq i} w_{ij}x_j
\]
Example: 2 spins

\[ E = -wx_1x_2. \]

There are four states \( \alpha = 1 : 4 \) with energies:

\[ E_{1:4} = \{ E_{(++)}, E_{(+-)}, E_{(-+)}, E_{(--)} \} = \{-w, w, w, -w\} \]

We define \( \Delta E_{\alpha,\beta} = E_\beta - E_\alpha \):

\[
\Delta E = \begin{pmatrix}
0 & 2w & 2w & 0 \\
-2w & 0 & 0 & -2w \\
-2w & 0 & 0 & -2w \\
0 & 2w & 2w & 0
\end{pmatrix}
\]

We define \( q_{\alpha,\beta} = q(\beta|\alpha) \):

\[
q = \frac{1}{2} \begin{pmatrix}
0 & 1 & 1 & 0 \\
1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 \\
0 & 1 & 1 & 0
\end{pmatrix}
\]
Example: 2 spins

Assume $w > 0$. We have for $\alpha \neq \beta$ $a_{\alpha,\beta} = \min(\exp(-\Delta E_{\alpha,\beta}), 1)$:

$$a = \begin{pmatrix}
* & e^{-2w} & e^{-2w} & 1 \\
1 & * & 1 & 1 \\
1 & 1 & * & 1 \\
1 & e^{-2w} & e^{-2w} & *
\end{pmatrix}$$

For $\alpha \neq \beta$: $T_{\alpha,\beta} = T(\beta|\alpha) = a_{\alpha,\beta} q_{\alpha,\beta}$:

$$T = \begin{pmatrix}
1 - e^{-2w} & \frac{1}{2} e^{-2w} & \frac{1}{2} e^{-2w} & 0 \\
\frac{1}{2} & 0 & 0 & \frac{1}{2} \\
\frac{1}{2} & 0 & 0 & \frac{1}{2} \\
0 & \frac{1}{2} e^{-2w} & \frac{1}{2} e^{-2w} & 1 - e^{-2w}
\end{pmatrix}$$

Diagonal terms follow from $\sum_{\beta} T_{\alpha,\beta} = 1$.

Interpretation of $T$: from high energy states 2,3 ($(+, -)$ and $(-, +)$ spins transit with equal probability to one of the two low energy states 1,4 ($(+, +)$ and $(-, -)$). From low energy states 1,4 their is finite probability to transit to high energy states (decreasing with $w$).
Example: 2 spins

Eigenvalues of $T$ versus $w$. For low $w$ there is one eigenvalue 1 and is ergodic. For high $w$ there are two eigenvalues 1, system is non-ergodic.
Lecture 2. Monte Carlo sampling

Based on MacKay chapter 29 and 30.

- Uniform sampling
- Importance and Rejection sampling
- Metropolis-Hasting method
- Detour on Markov processes, stationary distribution, ergodicity
- Gibbs sampling
- Hybrid Monte Carlo
- Sampling the Bayesian posterior for the perceptron
Gibbs sampling

- Consider only change of one element \( i \) of \( (x_1, \ldots, x_n) \) at the time. Update \( x \rightarrow x' \) according to

\[
q(x'|x) = p(x'_i|x\setminus i)\delta_{x\setminus i, x'_i}
\]

with \( p(x'_i|x\setminus i) \) the conditional distribution from the target distribution \( p(x) \).

- Accept: \( x^{r+1} = x' \)

The one dimensional sampling can be done using for instance Rejection sampling.

\( x\setminus i \) is the vector of variables without \( x_i \). So if \( x = (x_1, x_2, x_3) \) then \( x\setminus 2 = (x_1, x_3) \).
Gibbs sampling

Each Gibbs sampling is a MH step with acceptance $a = 1$:

$$q(x'|x) = p(x'_i|x\backslash i)\delta_{x\backslash i,x'_i}$$

$$a = \frac{p(x')q(x|x')}{p(x)q(x'|x)} = \frac{p(x'_i|x\backslash i)p(x'_i)\ p(x_i|x'_i)}{p(x_i|x\backslash i)p(x_i)\ p(x'_i|x_i)} = \frac{p(x'_i|x\backslash i)\ p(x_i|x\backslash i)}{p(x_i|x\backslash i)\ p(x'_i|x\backslash i)} = 1$$

where we used $x'_i = x_i$. 
Comparison Gibbs sampling and MH sampling

\[ p(x) = \frac{1}{Z} \exp(-E(x)) \]

Given state \( x = (x_i, x_i) \), the probability to go to state \( x' = (x_i, -x_i) \) is

**Gibbs:** \( p(x'|x) = p(-x_i|x_i) = \frac{p(-x_i, x_i)}{p(x_i)} = \frac{e^{-E(-x_i, x_i)}}{e^{-E(x_i, x_i)} + e^{-E(-x_i, x_i)} = \frac{1}{1 + e^{\Delta E_{x,F_i}x}} \}

**MH:** \( p(x'|x) = \min(e^{-\Delta E_{x,F_i}x}, 1) \)

![Graph showing comparison between Gibbs and MH sampling]
The detailed balance is the same

Gibbs: \[
\frac{p(x'|x)}{p(x|x')} = \frac{p(-x_i|x_{\setminus i})}{p(x_i|x_{\setminus i})} = \frac{1 + e^{-\Delta E}}{1 + e^{\Delta E}} = e^{-\Delta E} = \frac{p(x')}{p(x)}
\]

MH: \[
\frac{p(x'|x)}{p(x|x')} = \frac{\min(e^{-\Delta E}, 1)}{\min(e^{\Delta E}, 1)} = e^{-\Delta E} = \frac{p(x')}{p(x)}
\]
Correlations slow down sampling

When $q(x'|x)$ is Gaussian centered on $x$, $\frac{q(x'|x)}{q(x|x')}$ independent of $x, x'$:

$$a_{x'x} = \frac{p^*(x')}{p^*(x)}$$

$\epsilon$ large:
Acceptance rate $a_{x'x}$ small.

$\epsilon$ small:
Strong dependence on starting value. Many samples needed to sample.
Correlations slow down sampling

With step $\epsilon$ random, the particle moves a distance $L \approx \sqrt{T\epsilon}$ in $T$ iterations.\(^7\)

If largest length scale is $L$ then

$$T \approx \left( \frac{L}{\epsilon} \right)^2$$

time steps are needed to sample the length $L$.

\(^7\)Denote the total distance travelled $x = \sum_{i=1}^{T} x_i$ with $x_i$ the distance in a single step with $\mathbb{E} x_i = 0$ and $\forall x_i = \epsilon^2$. Then $\mathbb{E} x = 0$ and $\forall x = \sum_{i=1}^{T} \forall x_i = T \epsilon^2$. Thus the typical distance travelled is $L = \sqrt{T\epsilon}$
Correlations slow down sampling

Convergence slow when variables correlated: When two variables have marginal width $L$ and conditional width $\epsilon$, the number of iterations scales as $T \approx L^2/\epsilon^2$. 
The Hybrid Monte Carlo Method

Let

\[ P(x) = \frac{e^{-E(x)}}{Z} \]

with \( E \) and its gradient \( \frac{\partial E}{\partial x_i} \) easy to compute.

Gradient information reduces random walk behaviour in Metropolis method.

Double the state space by introducing for each \( x_i \) a momentum \( p_i \) variable. Define the Hamiltonian and distribution

\[
H(p, x) = E(x) + \frac{\alpha}{2} \sum_i p_i^2
\]

\[
P_H(p, x) = \frac{1}{Z_H} \exp \left( -E(x) - \frac{\alpha}{2} \sum_i p_i^2 \right)
\]

The marginal is \( p(x) = \int dp P_H(p, x) \).
The Hybrid Monte Carlo Method

Generate sample \((x', p')\) starting from \((x, p)\) using the Hamiltonian dynamics and Metropolis-Hasting:

\[
\frac{dp_i}{dt} = -\frac{\partial H}{\partial x_i} \quad \frac{dx_i}{dt} = \frac{\partial H}{\partial p_i}
\]

leaves \(H\) invariant.

\[
\frac{dH}{dt} = \sum_i \frac{dH}{dp_i} \frac{dp_i}{dt} + \frac{dH}{dx_i} \frac{dx_i}{dt} = 0
\]
Example

Consider the double well cost \( E(x) = (x^2 - 1)^2 \).

The Hamiltonian \( H(x, p) = E(x) + \frac{1}{2} \alpha p^2 \) has the form.

α large yields large \( p \) values and dynamical trajectories to large distances.
Pseudo code

Choose initial $x^1$.

For $t = 1 : T$:

1. choose $p^t$ from $\mathcal{N}(0, \alpha^{-1})$, giving $(x^t, p^t)$

2. run Hamilton dynamics, giving $(x', p')$

3. Metropolis step: accept $(x^{t+1}, p^{t+1}) = (x', p')$ as new state with probability

   \[
   a = \min(1, a) = \frac{P_H(x', p')}{P_H(x, p)} = \frac{e^{-H(x', p')}}{e^{-H(x, p)}}
   \]

4. On rejection, $(x^{t+1}, p^{t+1}) = (x^t, p^t)$

NB: $a = 1$ by construction in theory, but deviates from 1 due to numerical integration errors.
```plaintext
\[
\begin{align*}
g &= \text{gradE}(x) \quad \text{# set gradient using initial } x \\
E &= \text{findE}(x) \quad \text{# set objective function too} \\
\text{for } l = 1:L \\
\quad p &= \text{randn(size(x))} \quad \text{# initial momentum is Normal(0,1)} \\
\quad H &= p' \ast p / 2 + E \quad \text{# evaluate } H(x,p) \\
x_{\text{new}} &= x; \quad g_{\text{new}} = g \\
\text{for } \tau = 1:\text{Tau} \\
\quad p &= p - \epsilon \ast g_{\text{new}} / 2 \quad \text{# make half-step in } p \\
\quad x_{\text{new}} &= x_{\text{new}} + \epsilon \ast p \quad \text{# make step in } x \\
\quad g_{\text{new}} &= \text{gradE}(x_{\text{new}}) \quad \text{# find new gradient} \\
\quad p &= p - \epsilon \ast g_{\text{new}} / 2 \quad \text{# make half-step in } p \\
\text{endfor} \\
E_{\text{new}} &= \text{findE}(x_{\text{new}}) \quad \text{# find new value of } H \\
H_{\text{new}} &= p' \ast p / 2 + E_{\text{new}} \\
\text{dH} &= H_{\text{new}} - H \text{ \quad # Decide whether to accept} \\
\text{if ( dH < 0 )} \\
\text{\quad accept} &= 1 \\
\text{elseif ( rand() < exp(-dH) )} \\
\text{\quad accept} &= 1 \\
\text{else} \\
\text{\quad accept} &= 0 \\
\text{endif} \\
\text{if ( accept )} \\
\quad g &= g_{\text{new}} \quad x = x_{\text{new}} \quad E = E_{\text{new}} \\
\text{endif} \\
\text{endfor}
\end{align*}
```
Example 2

\[ H(x, p) = \frac{1}{2} \beta_1 x_1^2 + \frac{1}{2} \beta_2 x_2^2 + \frac{1}{2} (p_1^2 + p_2^2) \]

The dynamics is

\[
\begin{align*}
\dot{x}_i &= \frac{\partial H(x, p)}{\partial p_i} = p_i \\
\dot{p}_i &= -\frac{\partial H(x, p)}{\partial x_i} = -\beta_i x_i \\
\ddot{x}_i &= -\beta_i x_i
\end{align*}
\]

The solution is of the form \( x_i(t) = \sin(\sqrt{\beta_i} t) \).

If \( \beta_1 \gg \beta_2 \), oscillations in \( x_1 \) much faster than in \( x_2 \).
Comparison of HMC and Metropolis

Two dimensional elongated Gaussian distribution. a-b) Hybrid Monte Carlo method c-d) Metropolis method.
Perceptron/logistic regression

\[ p(t = 1|x, w) = \sigma(\vec{w} \cdot \vec{x}) \]

\[ \sigma(x) = \frac{1}{1 + \exp(-2x)} \]

\[ \vec{w} \cdot \vec{x} = w_0 + w_1 x_1 + w_2 x_2 \]

\[ \mathbf{w} = (0, 2) \]
Perceptron/logistic regression

$w = (-2, 3)$

$w = (0, 2)$

$w = (2, -2)$

$w = (1, 0)$

$w = (3, 0)$

$w = (5, 1)$

$w = (1, 4)$

$w = (2, 2)$

$w = (5, 4)$
Learning as inference

Data set: \( \{x^\mu, t^\mu\}, \mu = 1, \ldots, P \) with \( t^\mu = \pm 1 \).

Probability of data point under the model: \( p(t^\mu | x^\mu, w) = \sigma(t^\mu w \cdot x^\mu) \)

Likelihood:
\[
p(D|w) = \prod_{\mu} p(t^\mu | x^\mu, w) = \exp(-G(w)) \quad G(w) = - \sum_{\mu} \log(p(t^\mu | x^\mu, w))
\]

Prior:
\[
p(w) = \frac{\exp(-\alpha E_w(w))}{Z_w(\alpha)} \quad E_w(w) = \sum_i w_i^2
\]
makes solutions with small weights more probable.

Posterior:
\[
p(w|D) = \frac{p(D|w)p(w)}{p(D)} \propto \exp(-M(w))
\]
\[
M(w) = G(w) + \alpha E_w(w)
\]
ML versus Bayesian

Standard in neural network learning is to compute the maximum likelihood or maximum posterior solution. For new test point $a$

$$D \rightarrow w_{ml}$$

$$p(t|a) = p(t|a, w_{ml})$$

Bayesian approach requires integration over multiple solutions:

$$D \rightarrow p(w|D)$$

$$p(t|a) = \int dw p(w|D)p(t|a, w) = \langle p(t|a, w) \rangle_{p(w|D)}$$
The maximum posterior solution

Minimizing the cost function $M(w) = G(w) + \alpha E_w(w)$ yields smoother solutions for larger $\alpha$. 
The full Bayesian solution
The full Bayesian solution

\[ \alpha = 0.01, \quad q(w'|w) = N(w'|w, \sigma), \quad \sigma = 0.1 \]

\[ p(t|x) = \int dw p(t|x, w)p(w|D) \approx \frac{1}{R} \sum_r p(t|x, w^r) \]
Lecture 3. The Ising model

- Illustration of Metropolis Hasting algorithm for the Ising model

- Phase transition: qualitative change when going from high temperature to low temperature. Ferro-magnetic Ising model.

- Critical slowing down

- Frustration in the anti-ferromagnetic Ising model
  - Transfer matrix method for computation of the partition sum

- Discrete optimization with iterative improvement and simulated annealing
The Ising model

Ising model is a probability distribution over vectors of binary variables. \( x = (x_1, \ldots, x_N) \) with \( x_i = \pm 1 \).

\[
p(x) = \frac{1}{Z} \exp (-\beta E(x)) \quad E(x) = -\frac{1}{2} \sum_{i \neq j} J_{ij} x_i x_j - \sum_i h_i x_i
\]

\[
Z = \sum_x \exp(-\beta E(x)) = \sum_{x_1} \ldots \sum_{x_N} \exp(-\beta E(x))
\]

\( \beta = 1/kT \) is the inverse temperature, can be absorbed in the coupling. 8

8It is sometimes convenient to write \( \frac{1}{2} \sum_{i \neq j} J_{ij} x_i x_j = \sum_{(ij)} J_{ij} x_i x_j \).
Some basic relations

The mean energy of the spin system \( \langle E \rangle = \sum_x E(x)p(x) \) can be computed from \( Z \):

\[
\frac{\partial \log Z}{\partial \beta} = \frac{1}{Z} \frac{\partial Z}{\partial \beta} = \frac{1}{Z} \sum_x -E(x) \exp(-\beta E(x)) = -\langle E \rangle
\]

The energy fluctuations are computed similarly:

\[
\frac{\partial^2 \log Z}{\partial \beta^2} = \frac{\partial}{\partial \beta} \left( \frac{1}{Z} \frac{\partial Z}{\partial \beta} \right) = -\frac{1}{Z^2} \left( \frac{\partial Z}{\partial \beta} \right)^2 + \frac{1}{Z} \frac{\partial^2 Z}{\partial \beta^2}
\]

\[
= -\langle E \rangle^2 + \frac{1}{Z} \sum_x E(x)^2 \exp(-\beta E(x)) = -\langle E \rangle^2 + \langle E^2 \rangle = \nabla (E)
\]

One calls \( F = -\frac{1}{\beta} \log Z \) the free energy, or a log partition sum.

\[
F = \langle E \rangle - \frac{1}{\beta} H \quad \text{ex.31.1}
\]


Metropolis Hasting for the Ising model

Use MH to sample from the Ising model

\[ p(x) = \frac{1}{Z} \exp(-E(x)) \quad E(x) = -\sum_{(ij)} w_{ij} x_i x_j \]

Consider \( q(x'|x) \) to implement single spin flips. Define \( F_i x \) is the vector of spins obtained by flipping bit \( i \). Then

\[ q(F_i x|x) = \frac{1}{n} \quad q(x|x) = 0 \]

The MH ratio for the transition from state \( x \) to state \( F_i x \) is

\[ a_{x,F_i x} = \frac{p^*(F_i x)}{p^*(x)} = \exp(-\Delta E_{x,F_i x}) \]

\[ \Delta E_{x,F_i x} = E(F_i x) - E(x) = 2x_i \sum_{j \neq i} w_{ij} x_j \]
Ferro-magnet

Rectangular grid with $J = 1$ and $h_i = 0$ Ferromagnet, periodic boundary.

Run system at different temperatures $\beta = 0.03 \rightarrow 10 \rightarrow 0.03$ in equilibrium (check for hysteresis) Run $T$ iterations at each temperature, discard first 1/3 for 'burn in'.

record $\langle E \rangle, \nabla E, \langle m^2 \rangle$ as time averages:

$$\langle E \rangle = \frac{1}{T} \sum_{t=1}^{T} E(x_t) \quad \nabla E = \frac{1}{T} \sum_{t=1}^{T} (E_t - \langle E \rangle)^2$$

$$\langle m^2 \rangle = \frac{1}{T} \sum_{t=1}^{T} m^2(x_t) \quad m(x_t) = \frac{1}{N} \sum_{i=1}^{N} x_i$$
Ferro-magnet: mean energy and magnetization

Left: Mean energy. Each spin has 4 links, thus $E$ has $2N$ terms.
- at low temperature $\langle E \rangle / N = -2 \langle x_i x_j \rangle = -2$ (for one of the two ground states).
- At high temperature $\langle E \rangle / N = -2 \langle x_i x_j \rangle \approx 0$ and $\nabla (E / N) \propto 1 / N$.

Right: Mean squared magnetisation
- is zero for high temperature
- and 'breaks' to $\pm 1$ for low temperature.
Ferro-magnet: energy fluctuations

Energy fluctuations increase around the critical temperature
Phase transitions

Phase transitions are values of $\beta$ for which the derivatives of

$$\log Z(\beta) = \log \sum_x e^{-\beta E(x)}$$

are discontinuous or diverge.

Phase transitions only occur when $N \to \infty$, because for finite $N$ $Z(\beta)$ is a smooth function.
Phase transitions

Consider the toy model: \( E(x) = -\epsilon N \) when \( x = (0, 0, \ldots, 0) \) and \( E(x) = 0 \) otherwise.

Then

\[
Z = e^{N\beta \epsilon} + 2^N - 1 \quad \text{lim log } Z = N\beta \epsilon \quad \text{lim log } Z = N \log 2
\]

\[
\frac{d \log Z}{d\beta} = N\epsilon \frac{e^{N\beta \epsilon}}{e^{N\beta \epsilon} + 2^N - 1}
\]

\[
\frac{d^2 \log Z}{d\beta^2} = N^2 \epsilon^2 \frac{(2^N - 1)e^{N\beta \epsilon}}{(e^{N\beta \epsilon} + 2^N - 1)^2}
\]

Fluctuations per spin are proportional to \( N \) and thus diverge when \( N \to \infty \).

\[
\langle E \rangle / N = -\epsilon O(1) \pm O(N)
\]
Phase transitions

Compare with the independent model $E(x) = \epsilon \sum_{i=1}^{N}(2x_i - 1)$. Then

$$Z = \left( e^{\beta \epsilon} + e^{-\beta \epsilon} \right)^N \quad \lim_{\beta \to \infty} \log Z = N\beta\epsilon \quad \lim_{\beta \to 0} \log Z = N \log 2$$

Since $\log Z = N f(\beta)$ we find that $\frac{d^2 \log Z}{d\beta^2} \propto N$ instead of $\propto N^2$. The fluctuations per spin do not diverge.

$\langle E \rangle / N = -\epsilon O(1) \pm O(1)$
Autocorrelation

During sampling $x_1, x_2, \ldots, x_N$ one obtains samples of the quantity of interest $Q_t = \phi(x_t)$. One can estimate the autocorrelation

$$A(\tau) = \frac{\langle Q_t Q_{t+\tau} \rangle - \langle Q_t \rangle^2}{\langle Q_t^2 \rangle - \langle Q_t \rangle^2} \propto e^{-\tau/\Theta}$$

Note that $A(0) = 1$. One expects for ergodic problems that for large $\tau$

$$\langle Q_t Q_{t+\tau} \rangle = \langle Q_t \rangle \langle Q_{t+\tau} \rangle = \langle Q_t \rangle^2 \quad \lim_{\tau \to \infty} A(\tau) = 0$$

$\Theta$ is the characteristic time on which the samples of the Markov process are correlated. Thus, the effective number of independent samples is $N/\Theta$. 

Bert Kappen, ML 77
Autocorrelation

Characteristic autocorrelation time $\Theta$ is finite independent of system size at $T \neq T_c$, but diverges with system size at $T = T_c$. This is known as critical slowing down. It is related to the divergence of the fluctuations at $T = T_c$.

Autocorrelation function for $\phi(x) = \left| \frac{1}{n} \sum_{i=1}^{n} x_i \right|$ in 2D Ising models of different sizes at $T = 3$ (left) and $T = T_c$ (right) with $T_c = 2 / \log(1 + 2 \sqrt{2}) \approx 2.269$.
Snapshots of the Markov process on a two dimensional $L = 32$ Ising lattice at three different times and three different temperatures. The middle row shows critical slowing down: the configuration hardly changes. The top and bottom rows show significant change.
Critical slowing down

The idea can be explained within the Landau theory of phase transitions. Model $V(x) = \alpha(T - T_c)x^2 + \frac{1}{4}\beta x^4$. When $T \ll T_c$ or $T \gg T_c$, the potential well can be approximated by a quadratic form. In both cases, the relaxation to the optimal value for $y = x - x_{opt}$ is described as a locally quadratic potential $V(y) = \frac{1}{2}\gamma y^2$. The relaxation dynamics is

$$\dot{y} = -\frac{dV}{dy} = -\gamma y \quad y(t) = y(0)e^{-\gamma t}$$

Thus, the relaxation is exponentially fast with characteristic time $\Theta = \gamma^{-1}$. 
For $T = T_c$, the quadratic term vanishes and the potential is $V(y) = \frac{1}{4} \beta y^4$ and

$$\dot{y} = -\beta y^3 \quad y(t) = \frac{1}{\sqrt{\beta t}}$$

The convergence is much slowed down.
Anti-ferromagnet

Anti-ferromagnet has $J = -1$ between neighbors on the grid.

Ground states of the anti-ferromagnet are one-to-one with ground states in ferromagnet.

The spins in the 2D lattice form a bi-partite graph $W, B$.

$$E_{\text{anti}}(x_B, x_W) = - \sum_{i \in B} \sum_{j \in W} (J = -1)x_ix_j = - \sum_{i \in B} \sum_{j \in W} (J = 1)x_i(-x_j)$$

$$= E_{\text{ferro}}(x_B, -x_W)$$

The ferro and anti-ferro magnetic system have the same energies $E(x)$ by relabeling $x$. And thus the same physics.
Anti-ferromagnet

When the number of spins per direction is odd, this is no longer true.

Rectangular grid of $5 \times 5$ spins with ferromagnetic (left) and anti-ferromagnetic (right) interaction.

Not all bonds $J x_i x_j$ can be minimized. This is a boundary effect that decreases with system size.
Frustration: (anti-)ferromagnetic triangular Ising model

When $J = -1$, $E(x)$ cannot be minimized by minimizing all link energies $-Jx_ix_j = x_ix_j$.

Of the eight possible configurations of three spins, six have energy $-1$ and two have energy $+3$.

As a result: no 'clear cut' unique minimal energy state(s) but many states with similar low energy (Spin glass).
Frustration: Anti-ferromagnetic triangular Ising model

Energy (top), fluctuations (middle), heat capacity (bottom) for ferromagnet $J = 1$ (left) and anti-ferromagnet ($J = -1$) on triangular lattice.

Note: different low energies due to frustration, absence of peak in $\nabla V(E)$ indicates no phase transition to state with long range order.
Transfer matrix method for computation of partition function

The summation of the partition sum $Z$ can be performed in the 2-dimensional Ising model on a lattice of size $W \times C$. The lattice is periodic in the $C$ direction.

$$Z = \sum_x \exp(-\beta E(x)) \quad E(x) = - \sum_{i>j} J x_i x_j$$

$$= \sum_{s_1} \sum_{s_2} \ldots \sum_{s_C} \exp\left( -\beta \sum_{c=1}^{C} E(s_c, s_{c+1}) \right)$$

$s_c$ is a binary vector of length $W$, $\sum s_c$ contains $2^W$ terms. The energy is

$$E(s_c, s_{c+1}) = - \sum_{i \in c, j \in c+1} J x_i x_j - \frac{1}{4} \sum_{i, j \in c} J x_i x_j - \frac{1}{4} \sum_{i, j \in c+1} J x_i x_j$$

$$Z = \sum_{s_1} \sum_{s_2} \ldots \sum_{s_C} M_{s_1,s_2} M_{s_2,s_3} \ldots M_{s_{C-1},s_C} M_{s_C,s_1} = \sum_{s_1} (M^C)_{s_1,s_1}$$

$$= \text{Trace}(M^C) \quad M_{s_c,s_{c+1}} = \exp(-\beta E(s_c, s_{c+1}))$$

$M$ is $2^W \times 2^W$ matrix.
Transfer matrix method for computation of partition function

\[
\lim_{C \to \infty} Z = \lim_{C \to \infty} \text{Trace}(M^C) = \lim_{C \to \infty} \sum_{\alpha} \mu_{\alpha}^C = \mu_{\text{max}}^C
\]

with \( \mu_{\text{max}} \) the largest eigenvalue of \( M \).

Free energy per spin

\[
f = -\frac{T}{WC} \log Z = -\frac{T}{W} \log \mu_{\text{max}}
\]

Free energy per spin \( f = \langle E \rangle - TH \). At high temperature \( \langle E \rangle \approx 0 \) and \( f \approx -TH \approx -T \log 2 \). Entropy \( H = -\frac{\partial F}{\partial T} \), (Ex. 31.1) at low temperature (slope), is zero for ferro and rect. anti-ferro, and non-zero for triang. anti-ferro.
Transfer matrix method for computation of partition function

Entropy and average energy is computed by finite differencing:

\[ H = -\frac{dF}{dT} \approx \frac{F(T + dT) - F(T)}{dT} \]

\[ \langle E \rangle = F - TH \]

Entropy of frustrated system (\(-\)) at low temperature is non-zero. Mean energy at low temperature is higher for frustrated (\(-\)) than for non-frustrated (\(+\)) system.
Direct computation of partition function

Energy fluctuation is computed by finite differencing of the mean energy

\[ \nabla E = \frac{d^2 \log Z}{d \beta^2} = -\frac{d \langle E \rangle}{d \beta} \]

\[ \nabla E = \frac{d^2 \log Z}{d \beta^2} \] versus temperature shows signs of phase transition for rectangular and triangular ferromagnet (+), but not for the frustrated triangular anti-ferromagnet (-).
Iterative improvement

Consider an discrete optimization problem to minimize a function $E(x)$:

$$\min_x E(x)$$

and denote the optimal value by $x^*$.

Iterative improvement. For any state, define a neighborhood $R(x)$ as the set of states $x'$ that can be reached from $x$. The neighborhood can be anything, for instance single spin flips or multiple spin flips.

1: Start with a random initial state $x_0$
2: for $t = 1, 2, \ldots$ do
3: Sample a state $x' \in R(x_t)$
4: if $E(x') < E(x_t)$ then
5: accept $x_{t+1} = x'$
6: else
7: reject $x_{t+1} = x_t$
8: end if
9: end for

Also referred to as a combinatoric optimization problem.
Iterative improvement

Iterative improvement is a so-called local search algorithm. These methods suffer from local minima, in the sense that the algorithm converges to a configuration $x_\infty \neq x^*$ that is locally optimal, but not globally $E(x_\infty) > E(x^*)$.

The depth $d(x)$ of state $x$ is defined as the lowest barrier to reach $x^*$. Define a path $\tau = (x_0, x_1, \ldots, x_p)$ with $x_0 = x$ and $x_p = x^*$ such that $x_i$ is in the neighborhood of $x_{i-1}$, $i = 1, \ldots, p$. The depth of states $x$ is

$$d(x) = \min_\tau \tilde{d}(\tau)$$

with $\tilde{d}(\tau) = \max_{x_i} E(x_i) - E(x)$ the depth of the path $\tau$. 
Simulated annealing

Simulated annealing approximates the optimization problem by a sequence of sampling problems:

$$\min_x E(x) \rightarrow p_k(x) = \frac{1}{Z} e^{-E(x)/T_k}$$

For each $k$ one can define a MH algorithm that samples from $p_k$.

For decreasing $T_k$, the distributions $p_k(x)$ become more and more centered on $x^*$ and

$$\lim_{T_k \to 0} p_k(x) = \delta_{x,x^*}$$

The analogy of a physical system freezing into a low temperature ordered state.
Simulated annealing

Instead of a sequence of homogeneous Markov processes, one can also define simulated annealing as a single inhomogeneous Markov process where $T_t$ decreases in each iteration $t$, as follows.

1. Start with a random initial state $x_0$
2. for $t = 1, 2, \ldots$ do
3. Sample a state $x' \in R(x_t)$
4. Compute $a = \exp \left( - \frac{E(x') - E(x)}{T_t} \right)$ \hspace{1cm} ▶ MH acceptance probability
5. if $a > 1$ then
6. accept $x_{t+1} = x'$
7. else
8. $x_{t+1} = x'$ w.p $a$ and $x_{t+1} = x_t$ w.p. $1 - a$
9. end if
10. end for

The sequence $T_t, t = 1, 2, \ldots$ is called the annealing schedule.
Simulated annealing [?]

The dynamics converges $x \rightarrow x^*$ [?] provided that the process is ergodic and

$$
T_t = \frac{T_1}{\log_2(t + 1)} \quad \beta_t = \beta_1 \log_2(t + 1) \quad T_1 \geq D \quad D = \max_x d(x)
$$

So the initial temperature must be larger than the deepest local minimum $D$.

The annealing is very slow. The number of iterations $t$ to reach a final temperature $T_f$ is

$$
t = e^{\frac{T_1}{T_f}}
$$

and can be (unacceptable) large.
Simulated annealing

In practice, one uses a sequence of homogeneous Markov processes (chains) of length $L$ and indexed by $k$.

Schedule
A simple choice is an exponential schedule $\beta_{k+1} = f \beta_k$.  \(^{10}\)

An alternative choice, proposed in [?], is $\beta_{k+1} = \beta_k + \frac{\Delta \beta}{\sqrt{\text{Var}_k E}}$ with $\text{Var}_k E$ the variance of the energy in chain $k$.

Estimate the initial temperature by doing an initial exploration of states $x$ and their energies $E(x)$ using the MH method with (too) high temperature. Set the initial temperature as $T_1 = \max(\Delta E)$, where $\Delta E = E(x') - E(x)$ when $x \rightarrow x'$. As a result,

$$a = \frac{p^*(x')}{p^*(x)} = e^{-\Delta E / T_1} \geq e^{-1}$$

at the start of the algorithm.

\(^{10}\beta_k = 1/T_k\) is the inverse temperature.
Simulated annealing

Estimate for each chain $k$ the mean energy $\langle E \rangle_k$ and its standard deviation $\nabla_k E$. The algorithm terminates when $\nabla_k E = 0$ and the estimate of the minimal cost is $\langle E \rangle_k$.

1: Run MH at high temperature to estimate $\beta_1$
2: while $\nabla E_k > 0$ do
3: $\beta_{k+1} = f \beta_k$ or $\beta_{k+1} = \beta_k + \frac{\Delta \beta}{\sqrt{\nabla_k E}}$
4: Run MH at temperature $\beta_k$. Compute $\langle E \rangle_k$ and $\nabla_k E$
5: end while

Choose $f = 1.01$ or similar.
Choose $L$ a multiple of the neighborhood size $|R|$.
Choose $\Delta \beta$ in the range $0.001, 0.01, 0.1$
Exercise Simulated annealing

Write your own SA algorithm to minimize \( E(x) = -\frac{1}{2} x' w x = - \sum_{(i,j)} x_i x_j w_{ij} \).

If \( w_{ij} > 0 \) random, optimization is easy (ferromagnetic). There are two global minima.

If \( w_{ij} \) has both signs, optimization is hard (spin glass). There are many minima of approx equal quality.

SA on frustrated \( w_{500} \) problem. AK annealing schedule with \( \Delta \beta = 0.001 \), chain length \( L = 1000 \). A solution \( E = -6600 \) is found in 20 sec. using 15758 chains. Left: mean \( E \) in chain \( k \) versus \( k \).

Middle: standard deviation of \( E \) in chain \( k \) versus \( k \). Right: \( \beta_k \) versus \( k \).
Lecture 4: Deterministic approximations of the Bayesian posterior

- Laplace approximation (the posterior for the perceptron)
- Variational approximation (the posterior for the Gaussian)
- Variational approximation for multi layer neural network
The Laplace approximation approximates a given a probability density $p(x)$ by a Gaussian distribution centered on the maximum $x^* = \arg\max_x p(x)$.

Without loss of generality we can write $p(x) = \frac{1}{Z} e^{-E(x)}$. Expand $E$ around $x^*$ to second order:

$$E(x) \approx E(x^*) + \sum_i (x_i - x_i^*) \frac{\partial E(x)}{\partial x_i} \bigg|_{x=x^*} + \frac{1}{2} \sum_{ij} (x_i - x_i^*)(x_j - x_j^*) \frac{\partial^2 E}{\partial x_i \partial x_j} \bigg|_{x=x^*}$$

Note, that $\frac{\partial E(x)}{\partial x_i} \bigg|_{x=x^*} = 0$ and define $H_{ij} = \frac{\partial^2 E}{\partial x_i \partial x_j} \bigg|_{x=x^*}$. Then

$$p(x) \approx \frac{1}{Z} e^{-E(x^*)} \left( e^{-\frac{1}{2}(x-x^*)^T H(x-x^*)} \right) = \frac{1}{\sqrt{\det(2\pi H^{-1})}} e^{-\frac{1}{2}(x-x^*)^T H(x-x^*)}$$

The Laplace approximation is not necessarily the 'best' approximation (for instance in terms of KL divergence), but is easy to compute.

$\int dx e^{-E(x)} \approx e^{-E(x^*)} \int e^{-\frac{1}{2}(x-x^*)^T H(x-x^*)} = e^{-E(x^*)} \sqrt{\det(2\pi H^{-1})}$
Laplace approximation for Bayesian posterior

Consider the Bayesian learning of the perceptron discussed in MK 39 and MK 41:

\[ p(D|w) = \exp(-G(w)) \quad G(w) = -\sum_{\mu=1}^{N} \log(p(t^{\mu}|x^{\mu}, w)) \]

\[ p(w|\alpha) \propto \frac{\exp\left(-\frac{1}{2}\alpha \sum_i w_i^2\right)}{Z_w(\alpha)} \]

\[ p(w|D, \alpha) = \frac{p(D|w)p(w)}{p(D)} = \frac{1}{Z} \exp(-M(w)) \quad M(w) = G(w) + \frac{1}{2}\alpha \sum_i w_i^2 \]

We wish to compute the Bayesian prediction the probability of class \( t = 0, 1 \) for a new data point

\[ p(t|x, D, \alpha) = \int dw p(t|x, w)p(w|D, \alpha) \]

\[ ^{12} \text{We use } t = 0, 1 \text{ and } p(t = 1|x, w) = \sigma(w \cdot x) \text{ with } \sigma(x) = 1/(1 + e^{-x}). \]
Laplace approximation for Bayesian posterior

We use the Laplace approximation to the posterior \( p(w|D, \alpha) \approx N(w|w^*, H^{-1}) \) as a Gaussian centered on the maximum posterior solution \( w^* = \text{argmin}_w M(w) \). \( w^* \) can be found by gradient descend.

The Hessian is

\[
H_{ij} = \frac{\partial^2 M}{\partial w_i w_j} \bigg|_{w=w^*} = \sum_{\mu=1}^{N} \sigma(h^{\mu})\sigma(-h^{\mu}) x_i^{\mu} x_j^{\mu} + \alpha \delta_{ij}
\]

\( H^{-1} \) is the covariance matrix of the posterior on \( w \) in the Gaussian approximation, which we can interpret as error bars on \( w \). With more data \( (N \to \infty) \) the variance in \( w \) goes to zero \( (H^{-1} \to 0) \).

Thus,

\[
p(t = 1|x, D, \alpha) \approx \int dw p(t = 1|x, w) N(w|w^*, H^{-1})
\]

The function \( p(t = 1|x, w) = \sigma(a) \) with \( a = w \cdot x \). Since \( w \) is Gaussian and \( x \) is fixed,
$a$ is also Gaussian and

$$
\mathbb{E}(a) = w^* \cdot x = a^* \quad \nabla(a) = x^T \nabla(w)x = x^T H^{-1} x = s^2
$$
Laplace approximation for Bayesian posterior

Thus, the $d$-dimensional integral is reduced to a 1-dimensional integral

$$p(t = 1|x, D, \alpha) \approx \int da \sigma(a) N(a|a^*, s^2)$$

The Gaussian integral of a sigmoid function can be approximated quite well as

$$\psi(a^*, s^2) = \int da \sigma(a) N(a|a^*, s^2) \approx \sigma(\kappa(s)a^*) \quad \kappa(s) = 1/ \sqrt{1 + \pi s^2/8}$$

Left: $\psi(a^*, s^2)$. Right: $\psi(a^*, s^2)$ and $\sigma(\kappa(2)a^*)$ versus $a^*$ for $s^2 = 4$. 
a) A projection of the Gaussian approximation onto the \((w_1, w_2)\) plane of weight space. The one- and two-standard-deviation contours are shown. Also shown are the trajectory of the optimizer, and the Monte Carlo method samples. (b) The predictive function obtained from the Gaussian approximation and the further approximation for the sigmoid integral (41.30).
Bayesian posterior for Gaussian

Hypothesis space for a Gaussian distribution in 1 dimension:

\[
p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)
\]

Give some data, we can evaluate \(p(x_1^n|\mu, \sigma)\)
Bayesian posterior for Gaussian: ML solution

The log likelihood is

\[ L = \log p(x_n^n | \mu, \sigma) = -N \log \left( \sqrt{2\pi\sigma} \right) - \frac{1}{2\sigma^2} \sum_n (x_n - \mu)^2 \]

The maximum likelihood solution is computed from \( \frac{\partial L}{\partial \mu} = \frac{\partial L}{\partial \sigma} = 0 \):

\[ \mu = \frac{1}{N} \sum_n x_n \quad \sigma^2 = \frac{1}{N} \sum_n (x_n - \mu)^2 \]

\( \mu \) independent of \( \sigma \), but \( \sigma \) depends on \( \mu \).
Bayesian posterior for Gaussian: Variational solution

Given the Gaussian model, flat prior on $\mu$ and the non-informative prior on the variance ($\beta = \frac{1}{\sigma^2}$):  

$$p(x|\mu, \beta) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{1}{2}\beta(x - \mu)^2\right) \quad p(\mu) \propto 1 \quad p(\beta) \propto \beta^{-1}$$

The posterior for $N$ data points $D = \{x_n\}_{n=1}^{N}$ is

$$p(\mu, \beta|D) \propto \beta^{N/2-1} \exp\left(-\frac{1}{2}\beta\left(N(\mu - \bar{x})^2 + S\right)\right)$$

with $\bar{x} = \frac{1}{N} \sum_n x_n$, $S = \sum_n (x_n - \bar{x})^2$.

---

13The non-informative prior is defined in the log domain. Define $x = \log \sigma$ then

$$p_x(x) = p_\sigma(\sigma) \left|\frac{d\sigma}{dx}\right|$$

The non-informative prior assumes $p_x(x)$ uniform. Then, since $\left|\frac{d\sigma}{d\sigma}\right| = \sigma^{-1}$ we obtain $p_\sigma(\sigma) \propto \sigma^{-1}$. In terms of $\beta$, we get $p_\beta(\beta) \propto \beta^{-1}$.
Bayesian posterior for Gaussian: Variational solution

We approximate the posterior \( p(\mu, \beta|D) \) by a factorized variational distribution \( q(\mu, \beta) = q_\mu(\mu)q_\beta(\beta) \).

The variational approximation is computed by minimizing

\[
KL(q|p) = \int d\mu d\beta q(\mu, \beta) \log \frac{q(\mu, \beta)}{p(\mu, \beta|D)} = \int d\mu q_\mu(\mu) \log q_\mu(\mu) + \int d\beta q_\beta(\beta) \log q_\beta(\beta) - \int d\mu d\sigma q_\mu(\mu)q_\beta(\beta) \log p(\mu, \beta|D)
\]
Bayesian posterior for Gaussian: Variational solution

We optimize with respect to $q_\mu(\mu)$ subject to normalization constraint:

$$\frac{\partial KL}{\partial q_\mu(\mu)} = \log q_\mu(\mu) + 1 - \int d\beta q_\beta(\beta) \log p(\mu, \beta | D)$$

$$= \log q_\mu(\mu) + \int d\beta q_\beta(\beta) \frac{1}{2} \beta N(\mu - \bar{x})^2 + \text{const.}$$

$$= \log q_\mu(\mu) + \frac{1}{2} \tilde{\beta} N(\mu - \bar{x})^2 + \text{const.} = 0$$

with $\tilde{\beta} = \int d\beta q_\beta(\beta) \beta$ and the constant term contains all terms independent of $\mu$. Thus, the solution $q_\mu$ is a Gaussian distribution with mean $\bar{x}$ and variance $(N\tilde{\beta})^{-1}$.

NB: $q_\mu$ depends on $q_\beta$ which we still have to determine!
Bayesian posterior for Gaussian: Variational solution

We optimize with respect to $q_\beta(\beta)$ subject to normalization constraint:

$$\frac{\partial KL}{\partial q_\beta(\beta)} = \log q_\beta(\beta) + 1 - \int dq_\mu(\mu) \log p(\mu, \beta|D)$$

$$= \log q_\beta(\beta) + 1 - \int dq_\mu(\mu) \left( \log \beta^{\frac{N}{2} - 1} - \frac{1}{2} \beta \left( N(\mu - \bar{x})^2 + S \right) \right)$$

$$= \log q_\beta(\beta) - \log \beta^{\frac{N}{2} - 1} + \frac{1}{2} \beta \left( \frac{1}{\beta} + S \right) + \text{const.}$$

$$= \log q_\beta(\beta) - \log \left( \beta^{c-1} e^{-\beta/s} \right) + \text{const.}$$

with the constant term contains all terms independent of $\beta$. In the second step we used $\int dq_\mu(\mu)(\mu - \bar{x})^2 = (N\bar{\beta})^{-1}$ and we defined $c = \frac{N}{2}, \frac{1}{s} = \frac{1}{2} \left( \frac{1}{\bar{\beta}} + S \right)$.

Thus, $q_\beta(\beta) = \Gamma(\beta|c, s)$ the Gamma distribution with parameters $c, s$.  

\footnote{The Gamma distribution is defined for non-negative $x$ as}

$$\Gamma(x|c, s) = \frac{1}{Z} \left( \frac{x}{s} \right)^{c-1} \exp \left( -\frac{x}{s} \right) \quad Z = \Gamma(c)s$$
Bayesian posterior for Gaussian: Variational solution

NB: \( q_\beta \) depends on \( \bar{\beta} \) which is given in terms of \( q_\beta \).

We can thus solve for \( \bar{\beta} \) self-consistently.

\[
\bar{\beta} = \int d\beta \Gamma(\beta|c, s)\bar{\beta} = cs = \frac{N}{\frac{1}{\beta} + S}
\]

which has solution \( \bar{\beta} = \frac{N-1}{s} \).

The final solution is

\[
q_\mu(\mu) = \mathcal{N}(\mu|\bar{x}, \frac{S}{N(N-1)})
\]

\[
q_\beta(\beta) = \Gamma(\beta|c, s) \quad c = \frac{N}{2} \quad \frac{1}{s} = \frac{NS}{2(N-1)}
\]

Note, that the variational approximation has mean \( \bar{\mu} = \bar{x} \) and \( \bar{\sigma}^2 = \frac{S}{N-1} \).
Bayesian posterior for Gaussian: Variational solution

Solid: $p(\mu, \sigma|D)$. Dashed: Iterative optimization of $q_\mu (b,d)$ and $q_\sigma (c,e)$ from initial (a) converges in 15 iterations (f)

The solution can also be obtained numerically by optimizing for $q_\mu$ for fixed $q_\beta$ and vise versa.
Variational multi-layered perceptron

[?] develops the variational approximation for the Bayesian posterior of multi-layered perceptron using a multi-variate Gaussian distribution.

They consider a two-layer neural network with $n$ inputs and one output

$$f(x, w) = \sum_{i=1}^{H} v_i \sigma(\sum_{j=1}^{n} u_{ij} x_j) \quad w = \{v_i, u_{ij}\}$$

The likelihood and prior:

$$p(D|w) = \exp \left( - \sum_{\mu} (f(x^\mu, w) - t^\mu)^2 \right) \quad p(w) \propto \exp \left( -\frac{1}{2} w' A w \right)$$

with $A$ a symmetric matrix.
**Variational multi-layered perceptron**

The posterior $p(w|D)$ is approximated using the variational approximation with a multivariate Gaussian distribution $q(w)$

$$KL(q|p) = \int dwq(w) \log q(w) - \int dwq(w) \log p(w|D)$$

$$= \int dwq(w) \log q(w) + \sum_{\mu} \int dwq(w) (f(x^\mu, w) - t^\mu)^2$$

$$+ \frac{1}{2} \int dwq(w) w'^A w$$

The entropy and quadratic term are easy to compute. The data term involve terms as

$$\int dwq(w) f(x, w)^2 = \int dwq(w) \sum_{ii'} v_i v_i' \sigma \left( \sum_j w_{ij} x_j \right) \sigma \left( \sum_{j'} w_{ij'} x_{j'} \right)$$

These can be reduced to 1-dimensional Gaussian integrals as we have seen in the Laplace approximation.
The $KL(q|p)$ is optimized using gradient descend in the parameters of $C, m$ of $q(w|C, m)$. Bayesian prediction for new data point $x$ is approximated as

$$
\langle f(x) \rangle = \int dw p(w|D) f(x, w) \approx \int dw \sum_{i=1}^{H} v_i \sigma \left( \sum_j u_{ij} x_j \right) q(w|C^*, m^*)
$$

Figure 2: Comparison of various approximations to the posterior distribution for a synthetic regression problem involving two adaptive parameters (details given in the text). (a) The true posterior distribution. (b) The local Gaussian approximation obtained by Laplace’s method, giving a Kullback-Leibler (KL) divergence value of 41. (c) The minimum KL fit obtained with a diagonal covariance Gaussian (KLD), giving a residual KL value of 4.6. (d) The minimum KL fit obtained using a full covariance Gaussian distribution, giving a residual KL value of 3.9.
Variational multi-layered perceptron

Application to Boston housing data. 13 inputs and one continuous output, 128 training samples, 250 test samples.

Covariance matrix of $q$ is diagonal + rank one: $C = \text{diag}(\sigma_1^2, \ldots, \sigma_k^2) + ss'$. This choice reduces the number of free parameters in $q$ from $O(k^2)$ to $O(k)$ with $k$ the number of weights in the neural network.

<table>
<thead>
<tr>
<th>Method</th>
<th>Test Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble ($s = 1$)</td>
<td>0.22</td>
</tr>
<tr>
<td>Ensemble (diagonal)</td>
<td>0.28</td>
</tr>
<tr>
<td>Laplace</td>
<td>0.33</td>
</tr>
</tbody>
</table>

References


