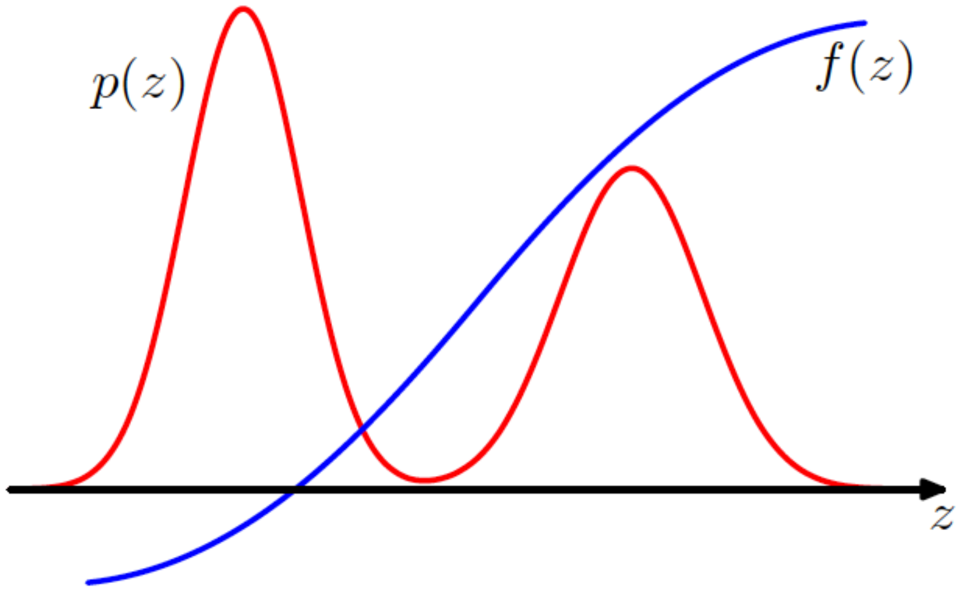
Sampling Methods

Approximate Inference

$$\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}$$



- Variational inference
  - Use a simple (tractable) distribution to approximate the original distribution

- **Monte Carlo sampling**

Unbiased estimation

$$\mathbb{E}[\widehat{f}] = \mathbb{E}[f]$$

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

i.i.d

$$\operatorname{var}[\widehat{f}] = \frac{1}{L}\mathbb{E}\left[(f - \mathbb{E}[f])^2\right]$$

Basic Sampling Alg.

- Change of variable rule for generating samples from simple non-uniform distr.

$$p(y) = p(z)\left|\frac{dz}{dy}\right| \qquad y = f(z)$$

$$p(y_1, \ldots, y_M) = p(z_1, \ldots, z_M)\left|\frac{\partial(z_1, \ldots, z_M)}{\partial(y_1, \ldots, y_M)}\right|$$

Determinant of Jacobian

- Question
  - How to obtain samples from multi-variate Gaussian distribution with specific mean and covariance matrix? (Exercise)

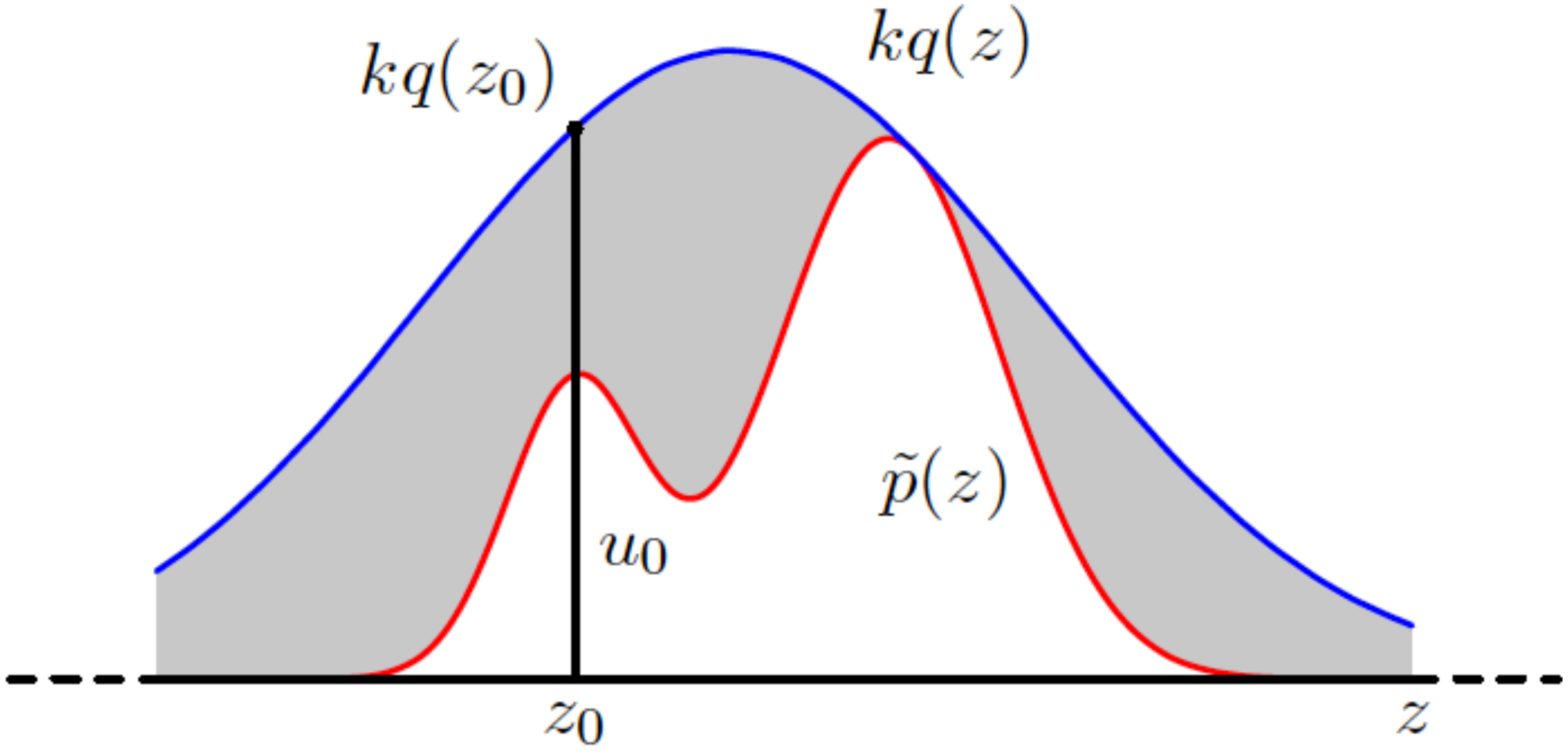
Rejection Sampling

$$p(z) = \frac{1}{Z_p}\widetilde{p}(z)$$

easy to evaluate, unnormalized density

- Relying on certain proposal distribution that is easy to sample from.

In the rejection sampling method, samples are drawn from a simple distribution $q(z)$ and rejected if they fall in the grey area between the unnormalized distribution $\widetilde{p}(z)$ and the scaled distribution $kq(z)$. The resulting samples are distributed according to $p(z)$, which is the normalized version of $\widetilde{p}(z)$.



$$kq(z) \geqslant \widetilde{p}(z)$$

Comparison function

z0 sampled from q(z)

u0 uniformly sampled from $[0, kq(z_0)]$

if $u_0 > \widetilde{p}(z_0)$ then the sample is rejected, otherwise

- Choosing k is important

$$
\begin{aligned}
p(\mathrm{accept}) &= \int \{\widetilde{p}(z)/kq(z)\}\, q(z)\, \mathrm{d}z \\
&= \frac{1}{k}\int \widetilde{p}(z)\, \mathrm{d}z.
\end{aligned}
$$

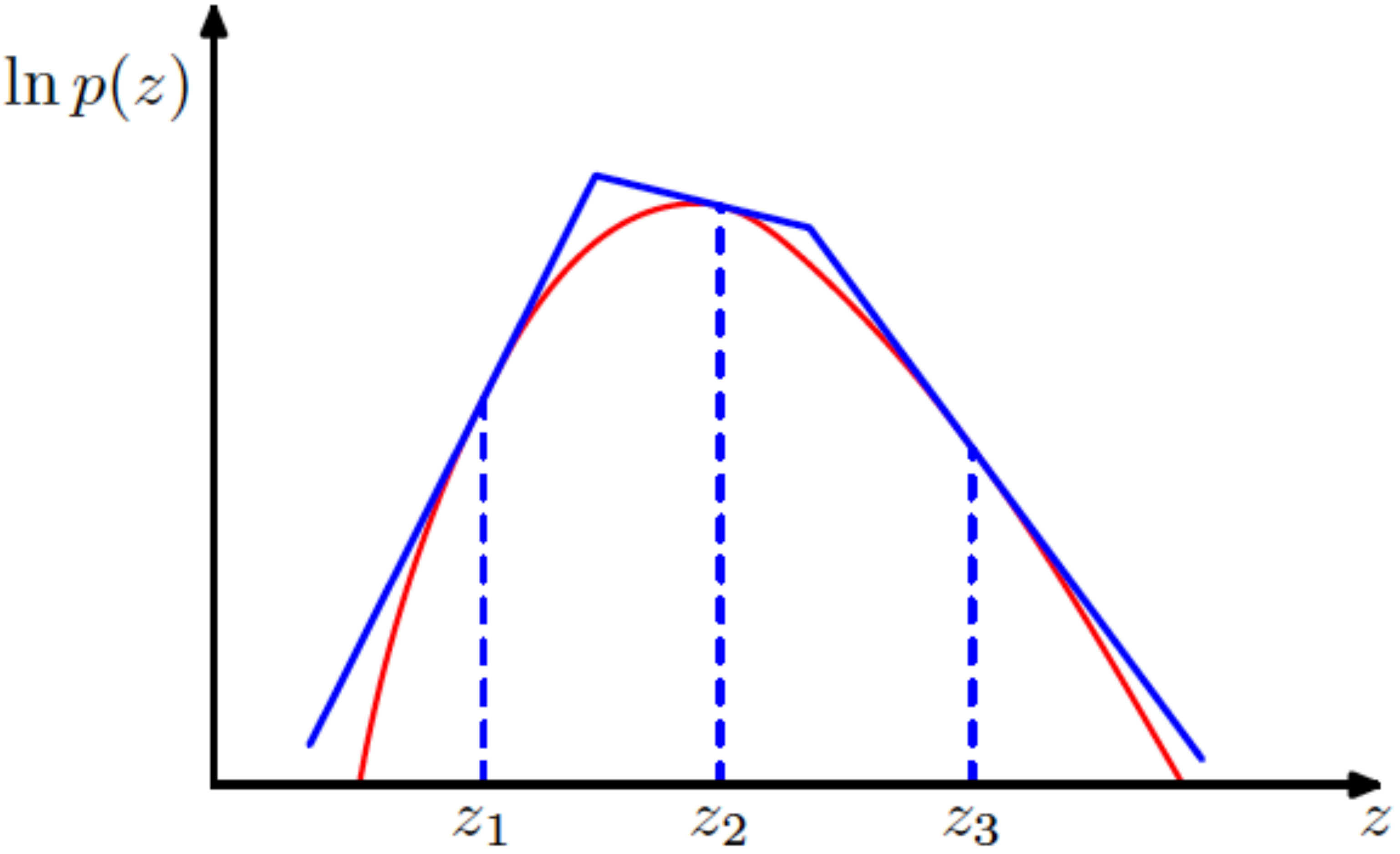
  - Exercise

    - How to sample from Gamma distribution?

$$\mathrm{Gam}(z|a,b) = \frac{b^a z^{a-1}\exp(-bz)}{\Gamma(a)}$$

- Any better rejection sampling strategy? (Adaptive rejection sampling)

In the case of distributions that are log concave, an envelope function for use in rejection sampling can be constructed using the tangent lines computed at a set of grid points. If a sample point is rejected, it is added to the set of grid points and used to refine the envelope distribution.



Importance Sampling

- A framework for approximating expectations, not directly drawing samples from the target distribution.

- $$\begin{aligned}\mathbb{E}[f] &= \int f(\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}\\ &= \int f(\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})}q(\mathbf{z})\,\mathrm{d}\mathbf{z}\\ &\simeq \frac{1}{L}\sum_{l=1}^{L}\frac{p(\mathbf{z}^{(l)})}{q(\mathbf{z}^{(l)})}f(\mathbf{z}^{(l)}).\end{aligned}$$

$$\begin{aligned}\mathbb{E}[f] &= \int f(\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}\\ &= \frac{Z_q}{Z_p}\int f(\mathbf{z})\frac{\widetilde{p}(\mathbf{z})}{\widetilde{q}(\mathbf{z})}q(\mathbf{z})\,\mathrm{d}\mathbf{z}\\ &\simeq \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L}\widetilde{r}_l f(\mathbf{z}^{(l)}).\end{aligned}$$

$$\begin{aligned}\frac{Z_p}{Z_q} &= \frac{1}{Z_q}\int \widetilde{p}(\mathbf{z})\,\mathrm{d}\mathbf{z} = \int \frac{\widetilde{p}(\mathbf{z})}{\widetilde{q}(\mathbf{z})}q(\mathbf{z})\,\mathrm{d}\mathbf{z}\\ &\simeq \frac{1}{L}\sum_{l=1}^{L}\widetilde{r}_l\end{aligned}$$

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

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

Both of rejection and importance sampling suffer from selection of proper proposal distributions and high dimensionality.

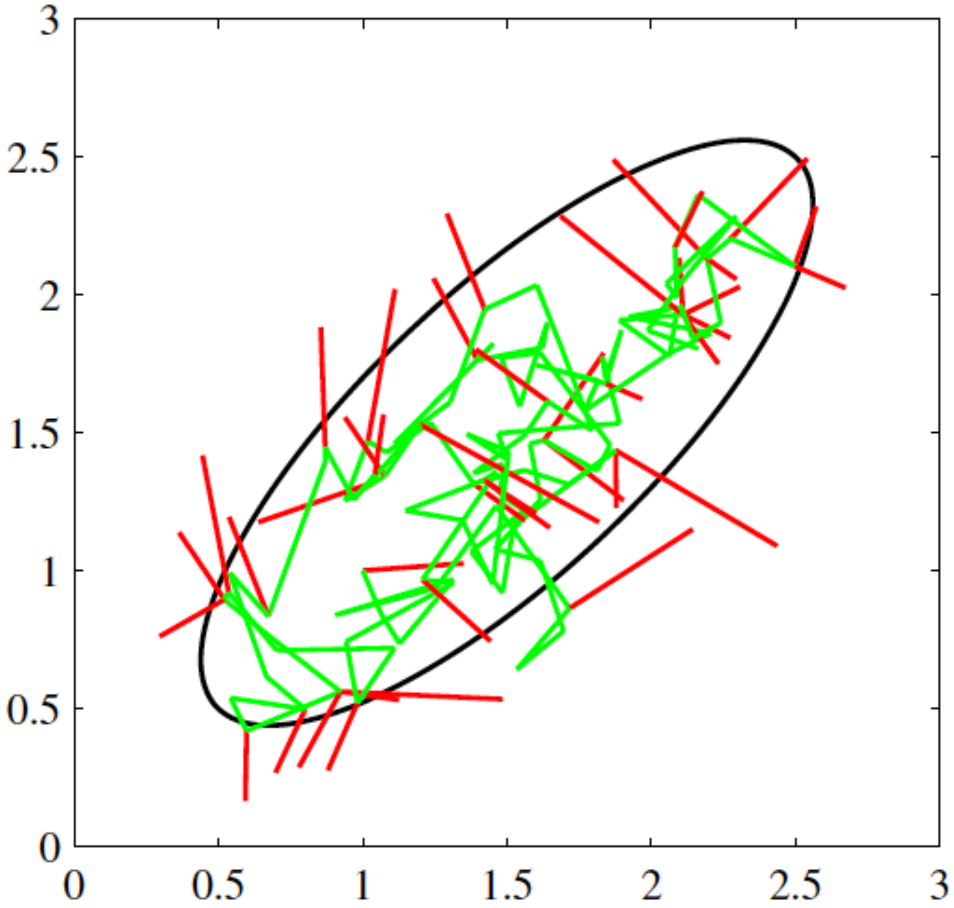
How to efficiently sample from high-dimensional distributions?

Markov Chain Monte Carlo

- MCMC originates from physics (Metropolis and Ulam, 1949), and started to have significant impact in statistics in 1980s.

- General MCMC procedure: construct a Markov chain of samples from a simple proposal distribution, and accept the each sample according to certain criterion such that the invariant distribution the chain is the target distribution.

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.



Markov Chain

$$p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(m)}) = p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)})$$

Transition prob.: $$T_m(\mathbf{z}^{(m)},\mathbf{z}^{(m+1)}) \equiv p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)})$$

A distribution is said to be invariant, or stationary, with respect to a Markov chain if each step in the chain leaves that distribution invariant. Thus, for a homogeneous Markov chain with transition probabilities $T(\mathbf{z}',\mathbf{z})$, the distribution $p^\star(\mathbf{z})$ is invariant if

$$p^\star(\mathbf{z}) = \sum_{\mathbf{z}'} T(\mathbf{z}',\mathbf{z})p^\star(\mathbf{z}').$$

A sufficient (but not necessary) condition for ensuring that the required distribution $p(\mathbf{z})$ is invariant is to choose the transition probabilities to satisfy the property of *detailed balance*, defined by

$$p^\star(\mathbf{z})T(\mathbf{z},\mathbf{z}') = p^\star(\mathbf{z}')T(\mathbf{z}',\mathbf{z})$$

$$\downarrow$$

$$\sum_{\mathbf{z}'} p^\star(\mathbf{z}')T(\mathbf{z}',\mathbf{z}) = \sum_{\mathbf{z}'} p^\star(\mathbf{z})T(\mathbf{z},\mathbf{z}') = p^\star(\mathbf{z})\sum_{\mathbf{z}'} p(\mathbf{z}'|\mathbf{z}) = p^\star(\mathbf{z})$$

The Metropolis-Hastings Alg.

In each step:

Proposal distribution

$$\text{draw a sample } \mathbf{z}^\star \text{ from the distribution } q_k(\mathbf{z}|\mathbf{z}^{(\tau)})$$

Accept it with following prob.:

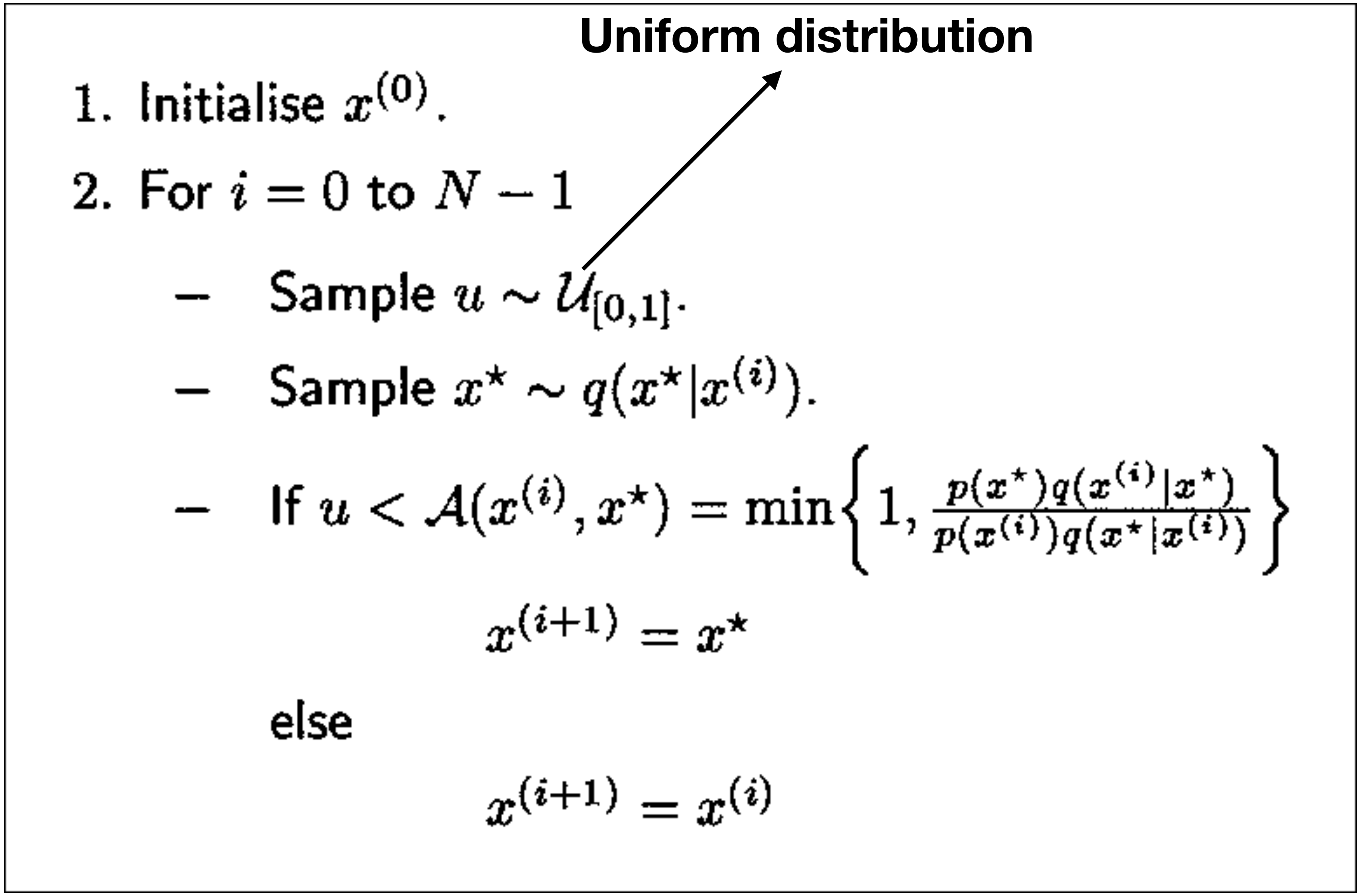
$$A_k(\mathbf{z}^\star, \mathbf{z}^{(\tau)}) = \min\left(1, \frac{\widetilde{p}(\mathbf{z}^\star)q_k(\mathbf{z}^{(\tau)}|\mathbf{z}^\star)}{\widetilde{p}(\mathbf{z}^{(\tau)})q_k(\mathbf{z}^\star|\mathbf{z}^{(\tau)})}\right)$$

if symmetric, vanished. E.g. Gaussian distr. centered on current state

Detailed balanced satisfied:

$$\begin{aligned} p(\mathbf{z})q_k(\mathbf{z}|\mathbf{z}')A_k(\mathbf{z}',\mathbf{z}) &= \min\left(p(\mathbf{z})q_k(\mathbf{z}|\mathbf{z}'), p(\mathbf{z}')q_k(\mathbf{z}'|\mathbf{z})\right) \\ &= \min\left(p(\mathbf{z}')q_k(\mathbf{z}'|\mathbf{z}), p(\mathbf{z})q_k(\mathbf{z}|\mathbf{z}')\right) \\ &= p(\mathbf{z}')q_k(\mathbf{z}'|\mathbf{z})A_k(\mathbf{z},\mathbf{z}') \end{aligned}$$

Ergodicity: from p(z0) to p(zm), when number of steps m goes to infinity, converge to the target distribution.



Uniform distribution

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

MH Alg.

The Metroplis-Hasting Alg.

- Choice of proposal distribution is extremely important for the performance of M.H. Alg., typically Gaussian.

Trade-off of variance with Gaussian proposals.
Small variance: high accepting, slow exploration; vice versa.

Schematic illustration of the use of an isotropic Gaussian proposal distribution (blue circle) to sample from a correlated multivariate Gaussian distribution (red ellipse) having very different standard deviations in different directions, using the Metropolis-Hastings algorithm. In order to keep the rejection rate low, the scale ρ of the proposal distribution should be on the order of the smallest standard deviation $\sigma_{\min}$, which leads to random walk behaviour in which the number of steps separating states that are approximately independent is of order $(\sigma_{\max}/\sigma_{\min})^2$ where $\sigma_{\max}$ is the largest standard deviation.

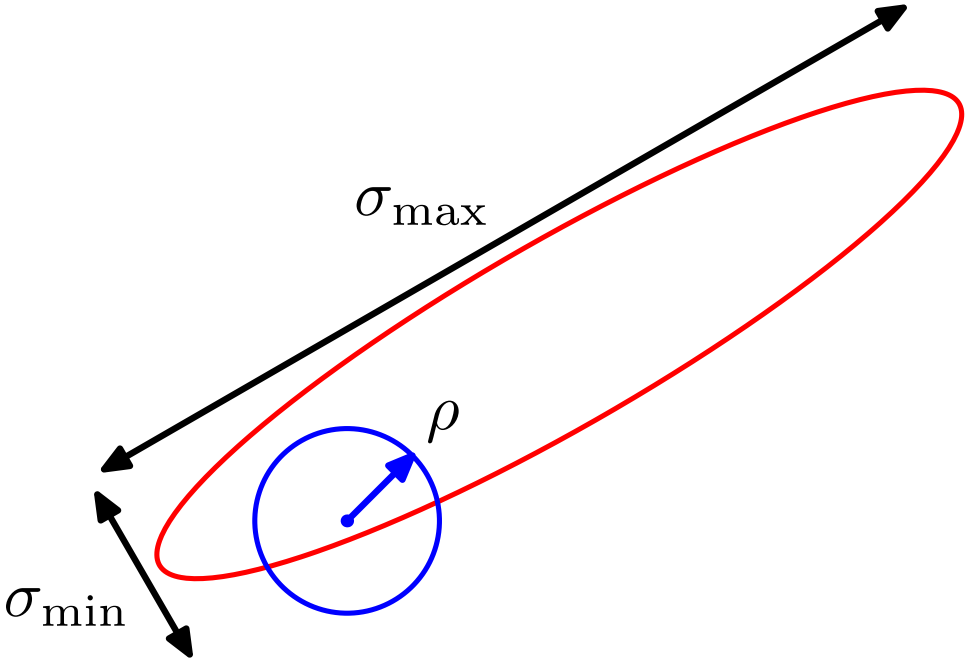


Fig. from Bishop (2016)

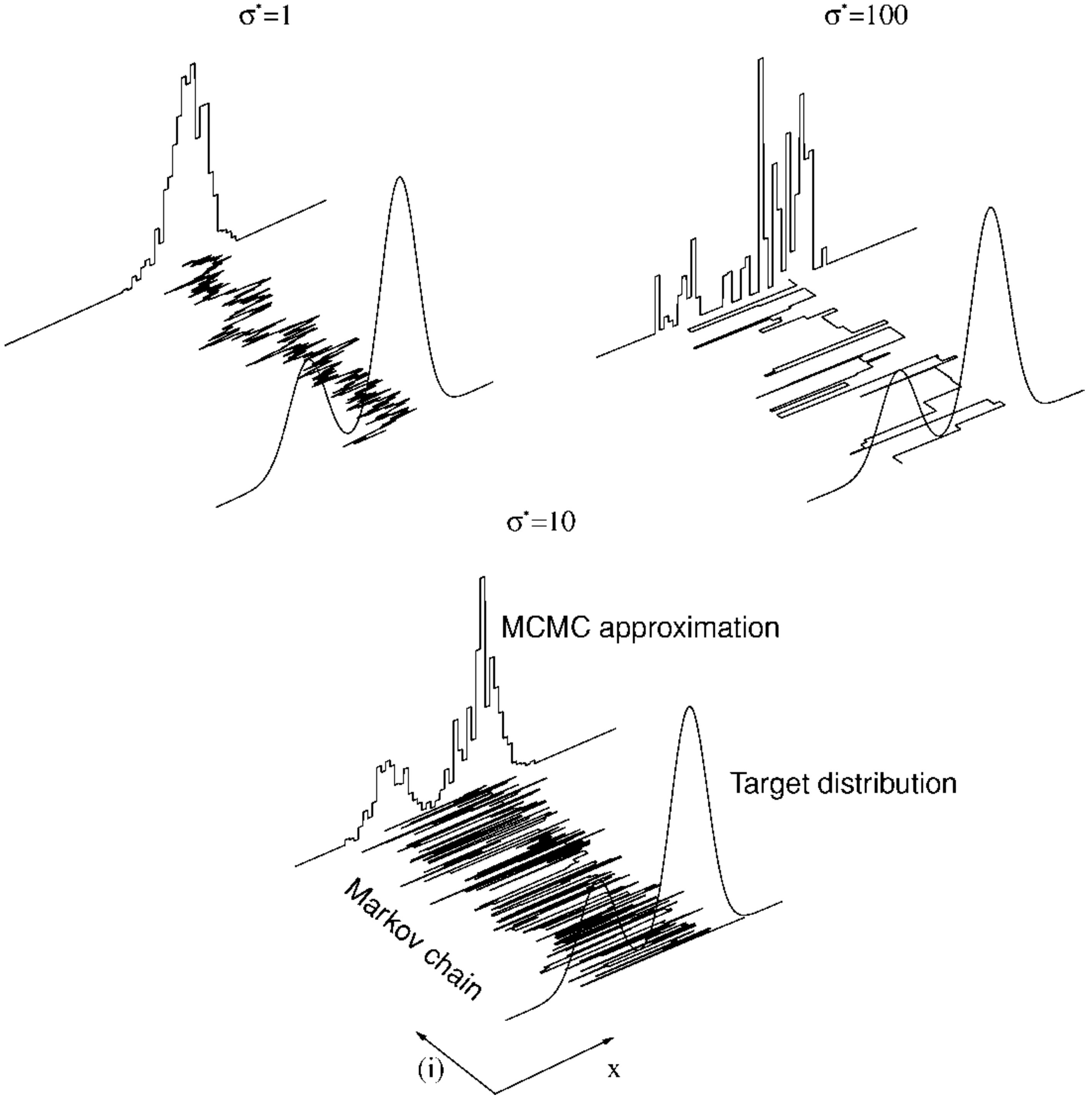
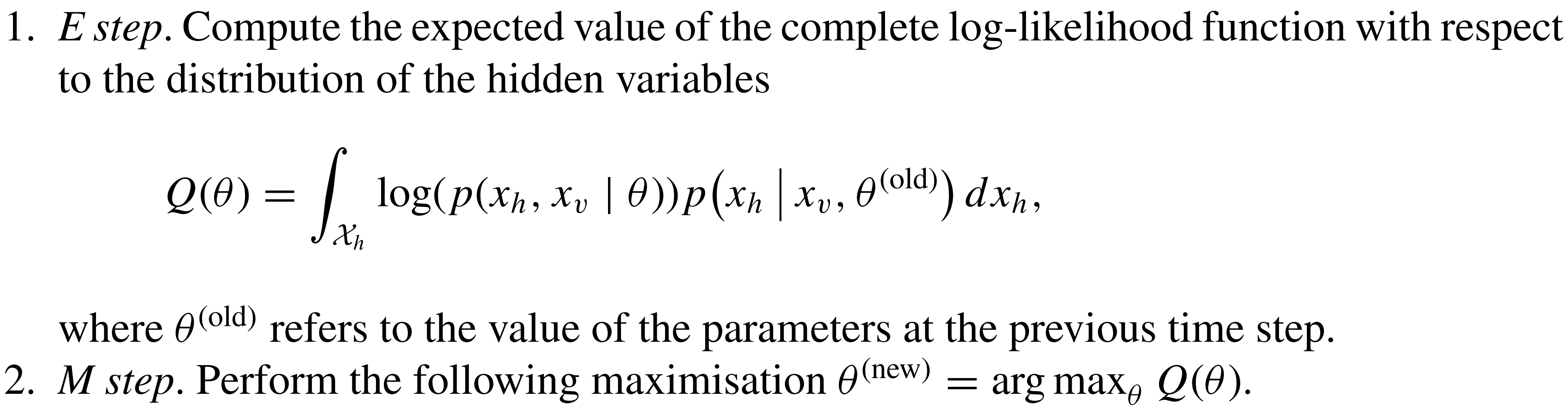


Figure 7. Approximations obtained using the MH algorithm with three Gaussian proposal distributions of different variances.

Monte Carlo EM

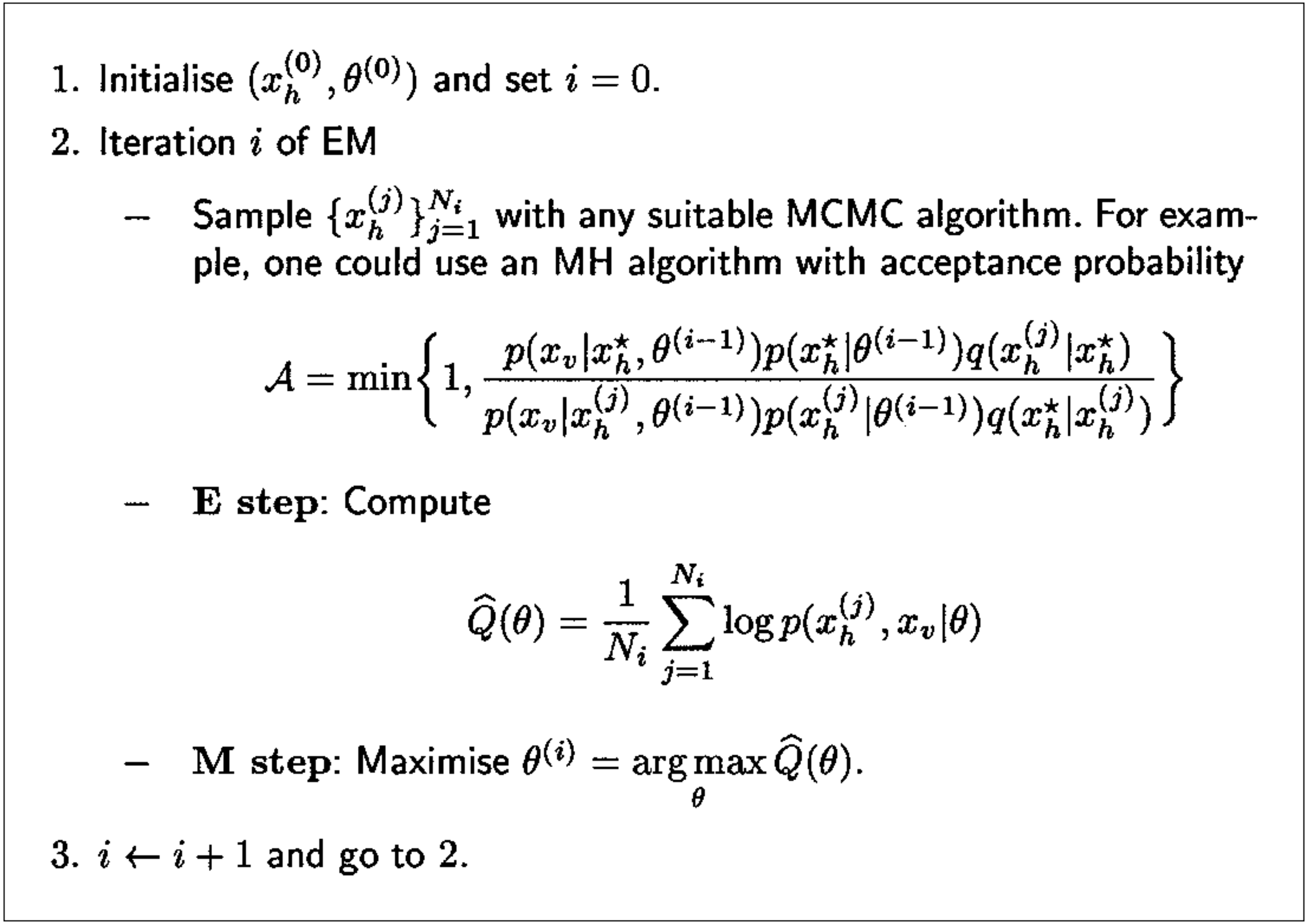
- EM alg.
    - Sometimes, in E step, the integral of Q function is hard to evaluate.
    - Monte Carlo approximation could be used.



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

$$Q(\theta) = \int_{\mathcal{X}_h} \log(p(x_h, x_v \mid \theta)) p\left(x_h \,\middle|\, x_v, \theta^{(\text{old})}\right) dx_h,$$

   where $\theta^{(\text{old})}$ refers to the value of the parameters at the previous time step.
2. *M step*. Perform the following maximisation $\theta^{(\text{new})} = \arg\max_\theta Q(\theta)$.



1. Initialise $(x_h^{(0)}, \theta^{(0)})$ and set $i = 0$.
2. Iteration i of EM
   - Sample $\{x_h^{(j)}\}_{j=1}^{N_i}$ with any suitable MCMC algorithm. For example, one could use an MH algorithm with acceptance probability

     $$\mathcal{A} = \min\left\{1, \frac{p(x_v|x_h^{\star}, \theta^{(i-1)})p(x_h^{\star}|\theta^{(i-1)})q(x_h^{(j)}|x_h^{\star})}{p(x_v|x_h^{(j)}, \theta^{(i-1)})p(x_h^{(j)}|\theta^{(i-1)})q(x_h^{\star}|x_h^{(j)})}\right\}$$

   - **E step**: Compute

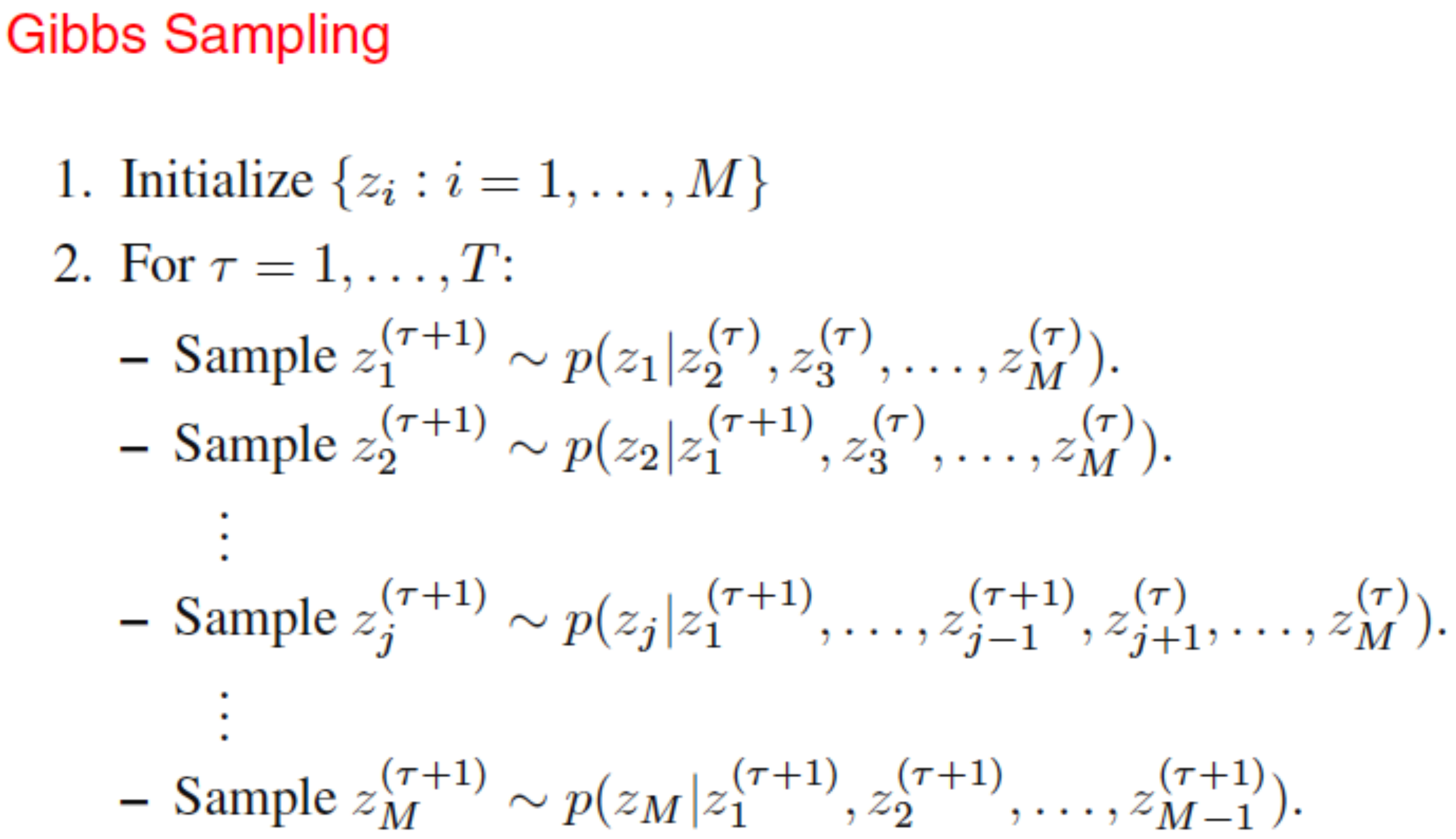
     $$\widehat{Q}(\theta) = \frac{1}{N_i}\sum_{j=1}^{N_i} \log p(x_h^{(j)}, x_v|\theta)$$

   - **M step**: Maximise $\theta^{(i)} = \arg\max_{\theta} \widehat{Q}(\theta)$.
3. $i \leftarrow i + 1$ and go to 2.

Monte Carlo EM. If only one sample is used, resulting stochastic EM.

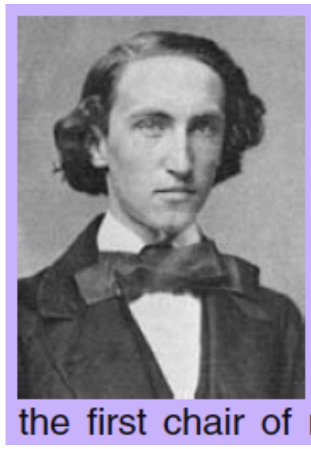
Gibbs Sampling

- A simple and widely used MCMC alg.: a special case of MH alg.
- Per variable conditional distr. as proposal distr. $p(z_i|\mathbf{z}_{\backslash i})$



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 $z_2^{(\tau+1)} \sim p(z_2|z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
      $\vdots$
    – Sample $z_j^{(\tau+1)} \sim p(z_j|z_1^{(\tau+1)}, \ldots, z_{j-1}^{(\tau+1)}, z_{j+1}^{(\tau)}, \ldots, z_M^{(\tau)})$.
      $\vdots$
    – Sample $z_M^{(\tau+1)} \sim p(z_M|z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$.
```



Josiah Willard Gibbs

1839–1903

Gibbs spent almost his entire life living in a house built by his father in New Haven, Connecticut. In 1863, Gibbs was granted the first PhD in engineering in the United States, and in 1871 he was appointed to the first chair of mathematical physics in the United States at Yale, a post for which he received no salary because at the time he had no publications. He developed the field of vector analysis and made contributions to crystallography and planetary orbits. His most famous work, entitled *On the Equilibrium of Heterogeneous Substances*, laid the foundations for the science of physical chemistry.

- Why always accepting?

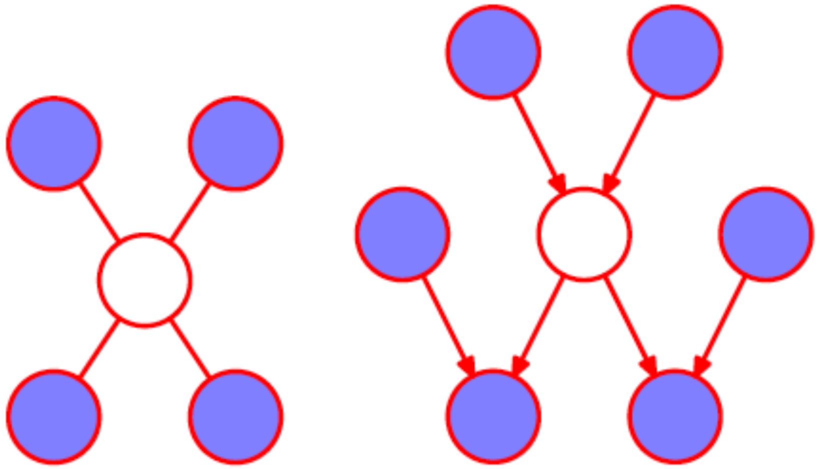
$$\mathbf{z}^\star_{\setminus k} = \mathbf{z}_{\setminus k}$$

$$A(\mathbf{z}^\star, \mathbf{z}) = \frac{p(\mathbf{z}^\star)q_k(\mathbf{z}|\mathbf{z}^\star)}{p(\mathbf{z})q_k(\mathbf{z}^\star|\mathbf{z})} = \frac{p(z_k^\star|\mathbf{z}^\star_{\setminus k})p(\mathbf{z}^\star_{\setminus k})p(z_k|\mathbf{z}^\star_{\setminus k})}{p(z_k|\mathbf{z}_{\setminus k})p(\mathbf{z}_{\setminus k})p(z_k^\star|\mathbf{z}_{\setminus k})} = 1$$

- Scenarios where conditional distribution is easy

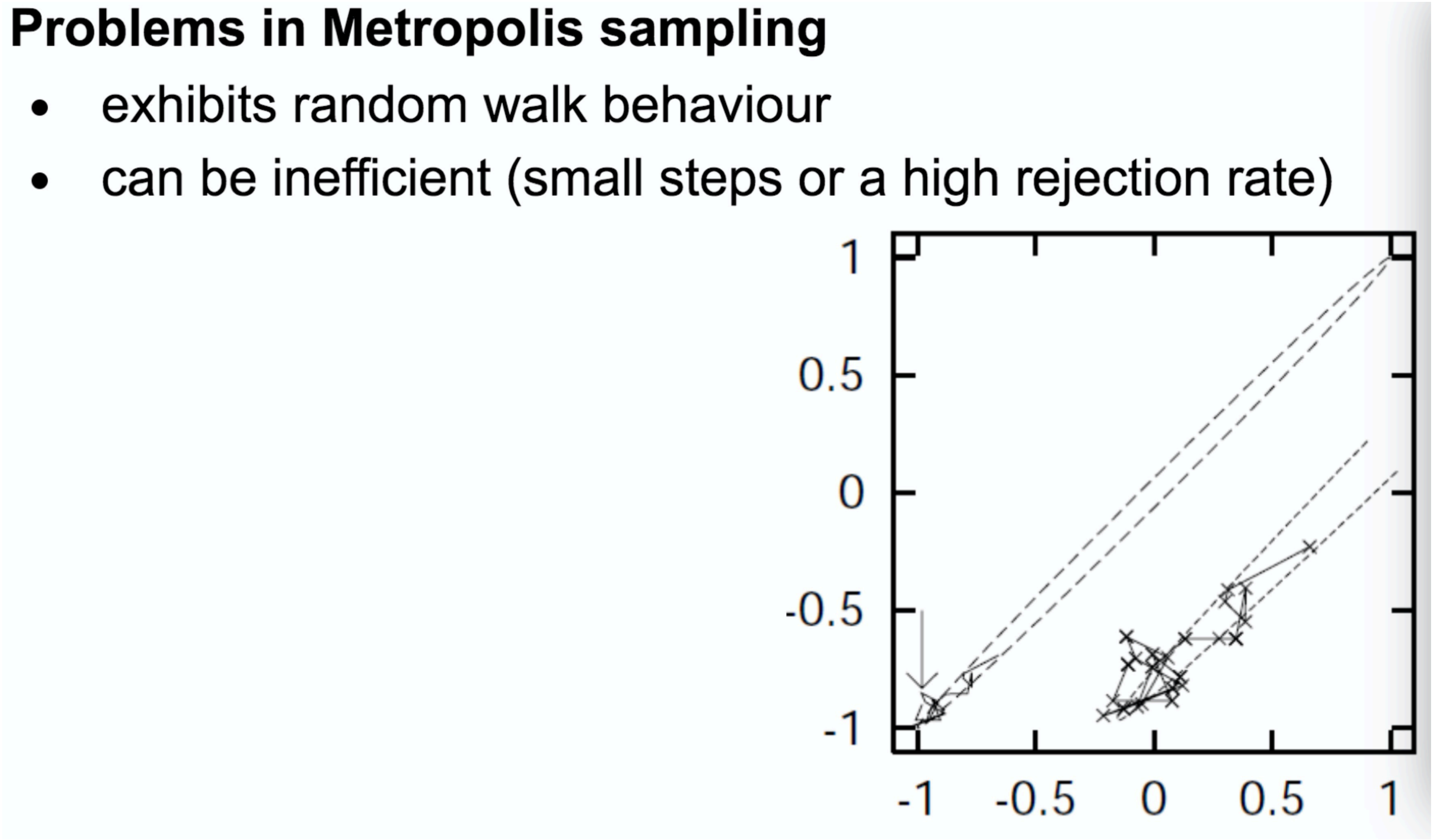
- Variant of Gibbs: block Gibbs

The Gibbs sampling method requires samples to be drawn from the conditional distribution of a variable conditioned on the remaining variables. For graphical models, this conditional distribution is a function only of the states of the nodes in the Markov blanket. For an undirected graph this comprises the set of neighbours, as shown on the left, while for a directed graph the Markov blanket comprises the parents, the children, and the co-parents, as shown on the right.



Hamiltonian Monte Carlo

- A very efficient MCMC algorithm based on Hamiltonian dynamics, using gradient information.

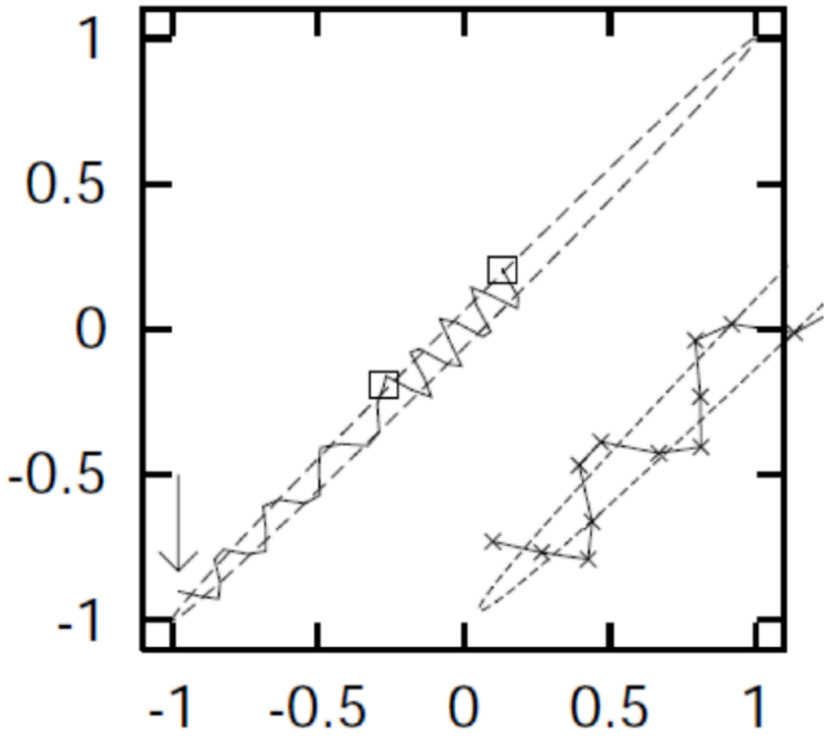


Problems in Metropolis sampling

- exhibits random walk behaviour
- can be inefficient (small steps or a high rejection rate)

HMC

- uses gradient information to make large steps with low rejection rate
- is applicable to distributions over continuous variables, for which we can evaluate the gradient of the log probability



Hamiltonian Dynamics

$$p(\mathbf{z}) = \frac{1}{Z_p} \exp\left(-E(\mathbf{z})\right)$$

Potential energy

$$\frac{\mathrm{d}r_i}{\mathrm{d}\tau} = -\frac{\partial E(\mathbf{z})}{\partial z_i}$$

Kinetic energy: $$K(\mathbf{r}) = \frac{1}{2}\|\mathbf{r}\|^2 = \frac{1}{2}\sum_i r_i^2$$

Hamiltonian function: $$H(\mathbf{z}, \mathbf{r}) = E(\mathbf{z}) + K(\mathbf{r})$$

Dynamical systems:

$$\begin{aligned} \frac{\mathrm{d}z_i}{\mathrm{d}\tau} &= \frac{\partial H}{\partial r_i} \\ \frac{\mathrm{d}r_i}{\mathrm{d}\tau} &= -\frac{\partial H}{\partial z_i} \end{aligned}$$

$$\begin{aligned}\frac{\mathrm{d}H}{\mathrm{d}\tau} &= \sum_i \left\{ \frac{\partial H}{\partial z_i}\frac{\mathrm{d}z_i}{\mathrm{d}\tau} + \frac{\partial H}{\partial r_i}\frac{\mathrm{d}r_i}{\mathrm{d}\tau} \right\} \\ &= \sum_i \left\{ \frac{\partial H}{\partial z_i}\frac{\partial H}{\partial r_i} - \frac{\partial H}{\partial r_i}\frac{\partial H}{\partial z_i} \right\} = 0\end{aligned}$$

$$\mathbf{V} = \left(\frac{\mathrm{d}\mathbf{z}}{\mathrm{d}\tau}, \frac{\mathrm{d}\mathbf{r}}{\mathrm{d}\tau} \right) \qquad \begin{aligned}\operatorname{div} \mathbf{V} &= \sum_i \left\{ \frac{\partial}{\partial z_i}\frac{\mathrm{d}z_i}{\mathrm{d}\tau} + \frac{\partial}{\partial r_i}\frac{\mathrm{d}r_i}{\mathrm{d}\tau} \right\} \\ &= \sum_i \left\{ -\frac{\partial}{\partial z_i}\frac{\partial H}{\partial r_i} + \frac{\partial}{\partial r_i}\frac{\partial H}{\partial z_i} \right\} = 0\end{aligned}$$

$$p(\mathbf{z}, \mathbf{r}) = \frac{1}{Z_H} \exp(-H(\mathbf{z}, \mathbf{r}))$$

Using the two results of conservation of volume and conservation of H, it follows that the Hamiltonian dynamics will leave $p(\mathbf{z}, \mathbf{r})$ invariant.

Although H is invariant, the values of $\mathbf{z}$ and $\mathbf{r}$ will vary, and so by integrating the Hamiltonian dynamics over a finite time duration it becomes possible to make large changes to $\mathbf{z}$ in a systematic way that avoids random walk behaviour.

$$p(\mathbf{z},\mathbf{r}) = \frac{1}{Z_H}\exp(-H(\mathbf{z},\mathbf{r}))$$

$$\frac{\mathrm{d}z_i}{\mathrm{d}\tau} = \frac{\partial H}{\partial r_i}$$

$$\frac{\mathrm{d}r_i}{\mathrm{d}\tau} = -\frac{\partial H}{\partial z_i}$$

Leapfrog integrator

$$r(t+\epsilon/2) = r(t) - \frac{\epsilon}{2}\frac{\partial E(x(t))}{\partial x}$$

$$x(t+\epsilon) = x(t) + \epsilon r(t+\epsilon/2)$$

$$r(t+\epsilon) = r(t+\epsilon/2) - \frac{\epsilon}{2}\frac{\partial E(x(t+\epsilon))}{\partial x}$$

HMC Procedures

1. Draw a momentum from Gaussian: $r \sim \exp(-K(r))/Z_K$
2. Simulate Hamiltonian dynamics by L leapfrog steps

$$r(t+\epsilon/2) = r(t) - \frac{\epsilon}{2}\frac{\partial E(x(t))}{\partial x}$$

$$x(t+\epsilon) = x(t) + \epsilon r(t+\epsilon/2)$$

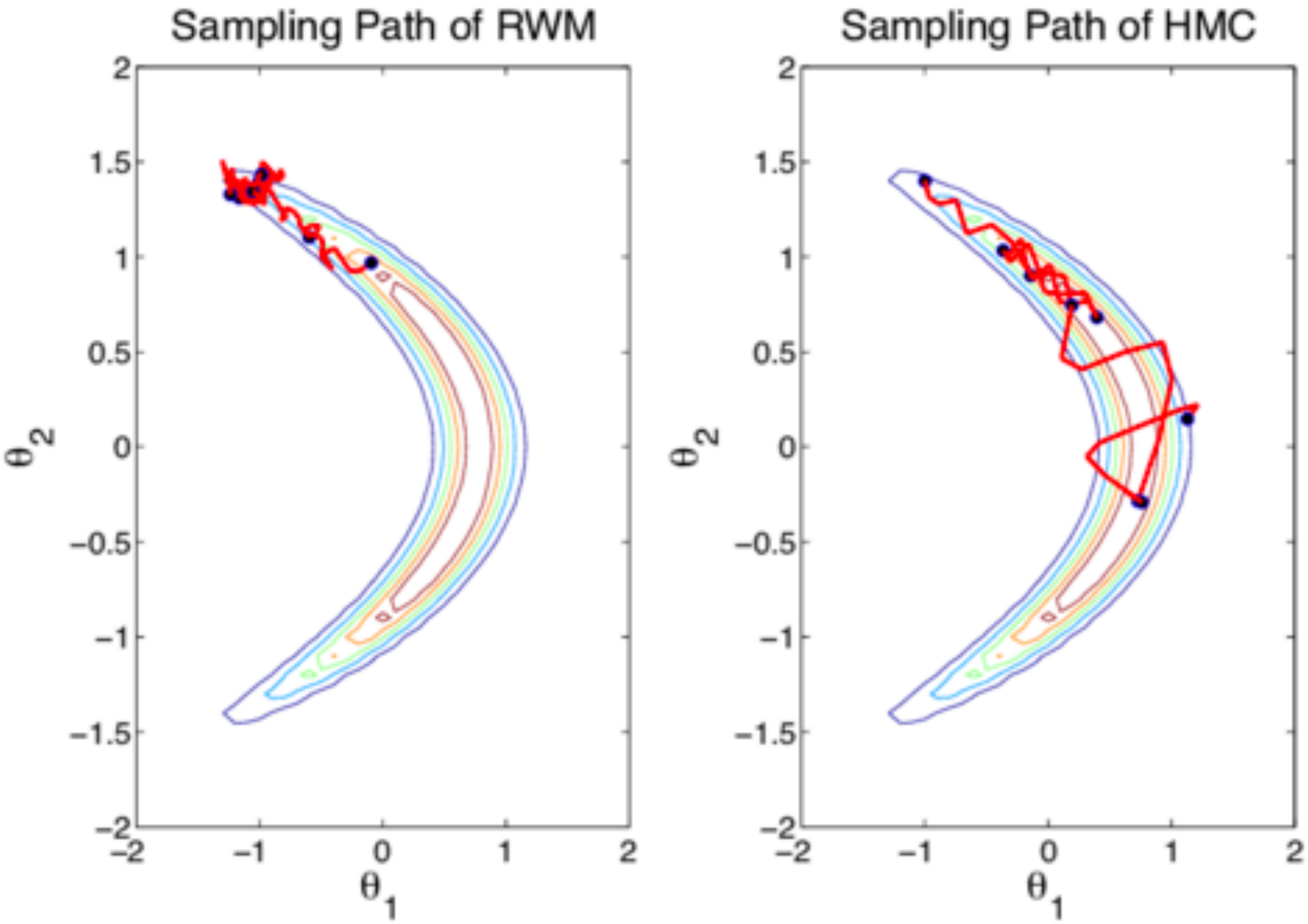
$$r(t+\epsilon) = r(t+\epsilon/2) - \frac{\epsilon}{2}\frac{\partial E(x(t+\epsilon))}{\partial x}$$

3. Decide whether to accept the state (x', r') after the leapfrog steps

$$\min(1, \exp[H(x,r) - H(x',r')])$$

✔ can take bigger steps to high probability states with low rejection rate
◎ applicable when we can evaluate the gradient
✘ need to tune the parameters (step size ϵ and #iterations L)

- The optimal acceptance rate is 65% ☞ Neal (2011)



Random walk MH v.s. HMC

Estimating Partition Function

- Energy based distribution

$$p_E(\mathbf{z}) = \frac{1}{Z_E}\exp(-E(\mathbf{z}))$$

- Partition function estimation is useful for model comparison, where the ratio of partition function is required.

$$\begin{aligned}\frac{Z_E}{Z_G} &= \frac{\sum_{\mathbf{z}}\exp(-E(\mathbf{z}))}{\sum_{\mathbf{z}}\exp(-G(\mathbf{z}))}\\ &= \frac{\sum_{\mathbf{z}}\exp(-E(\mathbf{z})+G(\mathbf{z}))\exp(-G(\mathbf{z}))}{\sum_{\mathbf{z}}\exp(-G(\mathbf{z}))}\\ &= \mathbb{E}_{G(\mathbf{z})}[\exp(-E+G)]\\ &\simeq \sum_l \exp(-E(\mathbf{z}^{(l)})+G(\mathbf{z}^{(l)}))\end{aligned}$$

- The issue: large variance of Monte Carlo estimation.

- Solution: annealed important sampling (AIS)

$$\frac{Z_M}{Z_1} = \frac{Z_2}{Z_1}\frac{Z_3}{Z_2}\cdots\frac{Z_M}{Z_{M-1}}$$

$$E_\alpha(\mathbf{z}) = (1-\alpha)E_1(\mathbf{z}) + \alpha E_M(\mathbf{z})$$