Introduction to Machine Learning

Sampling Methods

SHENG-JYH WANG

NATIONAL YANG MING CHIAO TUNG UNIVERSITY, TAIWAN

SPRING, 2024

Introduction (1/3)

- For most probabilistic models of practical interest, exact inference is intractable!
- We consider approximate inference methods based on numerical sampling, also known as *Monte Carlo* techniques.
- The fundamental problem involves finding the expectation of some function *f*(**z**) with respect to a probability distribution *p*(**z**).

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

Draw independently a set of samples $\mathbf{z}^{(l)}$, where $l = 1, \ldots, L$, from the distribution *p*(**z**).

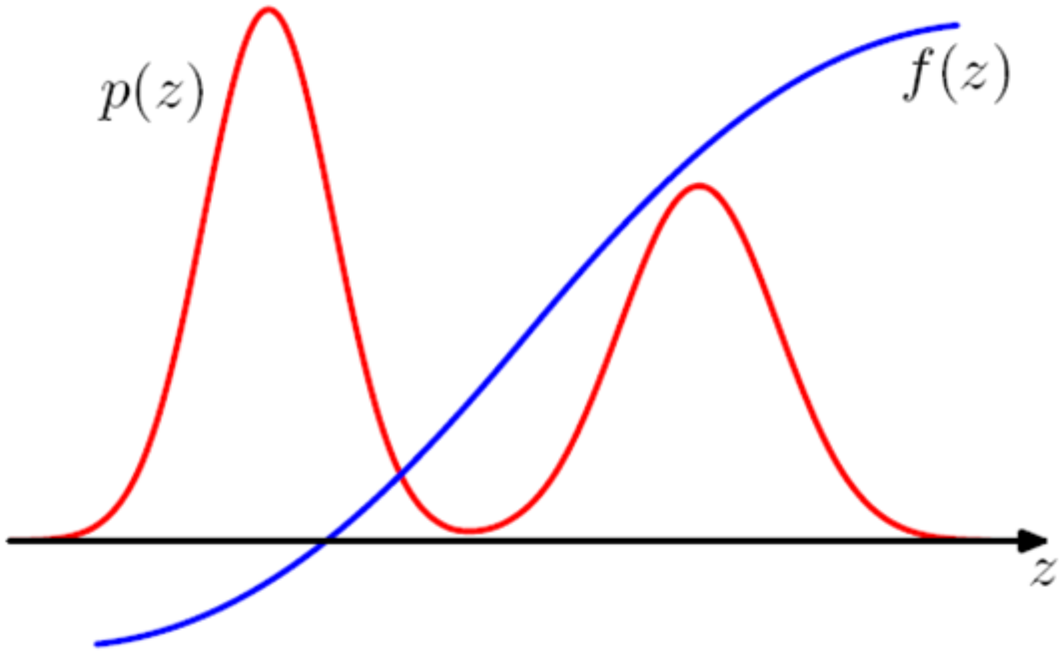
$$\widehat{f} = \frac{1}{L} \sum_{l=1}^{L} f(\mathbf{z}^{(l)}). \qquad \mathbb{E}[\widehat{f}] = \mathbb{E}[f] \qquad \mathrm{var}[\widehat{f}] = \frac{1}{L} \mathbb{E}\left[(f - \mathbb{E}[f])^2\right]$$

Introduction (2/3)

1. The accuracy of the estimator does not depend on the dimensionality of **z.**
2. In practice, ten or twenty independent samples may suffice to estimate an expectation to sufficient accuracy. However, the samples $\{\mathbf{z}^{(l)}\}$ might not be independent, and the effective sample size might be much smaller than the apparent sample size.
3. If $f(\mathbf{z})$ is small in regions where $p(\mathbf{z})$ is large, and vice versa, then the expectation may be dominated by regions of small probability.

   $\Rightarrow$ Relatively large sample sizes will be required to achieve sufficient accuracy.

Introduction (3/3)



Standard Distributions (1/6)

z is uniformly distributed over the interval (0, 1), and y = f(z).
The distribution of y will be

$$p(y) = p(z)\left|\frac{dz}{dy}\right|$$ where $p(z) = 1$.

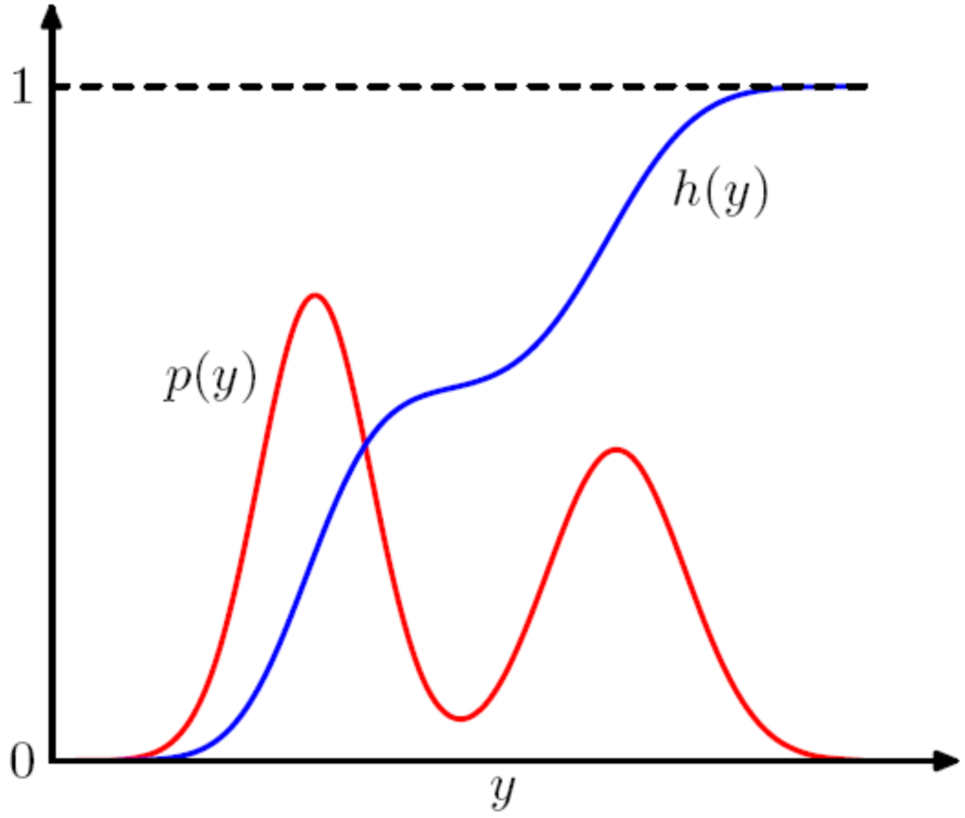
If we want to choose the function $f(z)$ such that the resulting values of y have some specific desired distribution $p(y)$,

$$z = h(y) \equiv \int_{-\infty}^{y} p(\widehat{y})\,\mathrm{d}\widehat{y}$$ CDF of y

$$y = h^{-1}(z)$$

$\Rightarrow$ We transform the uniformly distributed random numbers using the inverse of the indefinite integral of the desired distribution.

Standard Distributions (2/6)



Standard Distributions (3/6)

Example:

$$p(y) = \lambda \exp(-\lambda y)$$

where $0 \leq y < \infty$.

$$\Rightarrow \quad h(y) = 1 - \exp(-\lambda y)$$

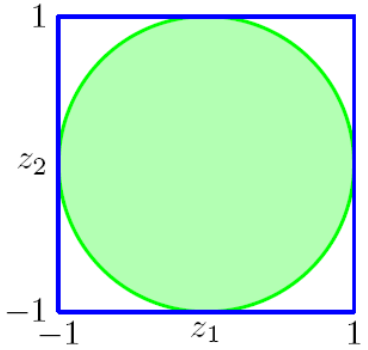
$$\Rightarrow \quad y = -\lambda^{-1} \ln(1 - z)$$

Standard Distributions (4/6)

The ***Box-Muller method*** for generating Gaussian samples.

(1) Generate pairs of uniformly distributed random numbers z_1, z_2 $\in$ (-1,1).

(2) Discard each pair unless it satisfies $z_1^2 + z_2^2 \leq 1$. This leads to a uniform distribution of points inside the unit circle with $p(z_1,z_2) = 1/\pi$.



Standard Distributions (5/6)

(3) For each pair (z_1,z_2), we evaluate the quantities

$$
\begin{aligned}
y_1 &= z_1 \left(\frac{-2 \ln z_1}{r^2} \right)^{1/2} \\
y_2 &= z_2 \left(\frac{-2 \ln z_2}{r^2} \right)^{1/2}
\end{aligned}
$$

where $r^2 = z_1^2 + z_2^2$.

$$
\Rightarrow \qquad
\begin{aligned}
p(y_1, y_2) &= p(z_1, z_2) \left| \frac{\partial(z_1, z_2)}{\partial(y_1, y_2)} \right| \\
&= \left[\frac{1}{\sqrt{2\pi}} \exp(-y_1^2/2) \right] \left[\frac{1}{\sqrt{2\pi}} \exp(-y_2^2/2) \right]
\end{aligned}
$$

Standard Distributions (6/6)

1. If y has a Gaussian distribution with zero mean and unit variance, then $\sigma y + \mu$ will have a Gaussian distribution with mean μ and variance σ^2.
2. To generate vector-valued variables having a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, we can make use of the *Cholesky decomposition* $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$.

   If $\mathbf{z}$ is a vector valued random variable whose components are independent and Gaussian distributed with zero mean and unit variance, then $\mathbf{y} = \boldsymbol{\mu} + \mathbf{L}\mathbf{z}$ will have mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

Rejection Sampling (1/7)

Suppose we wish to sample from a distribution p(z).

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

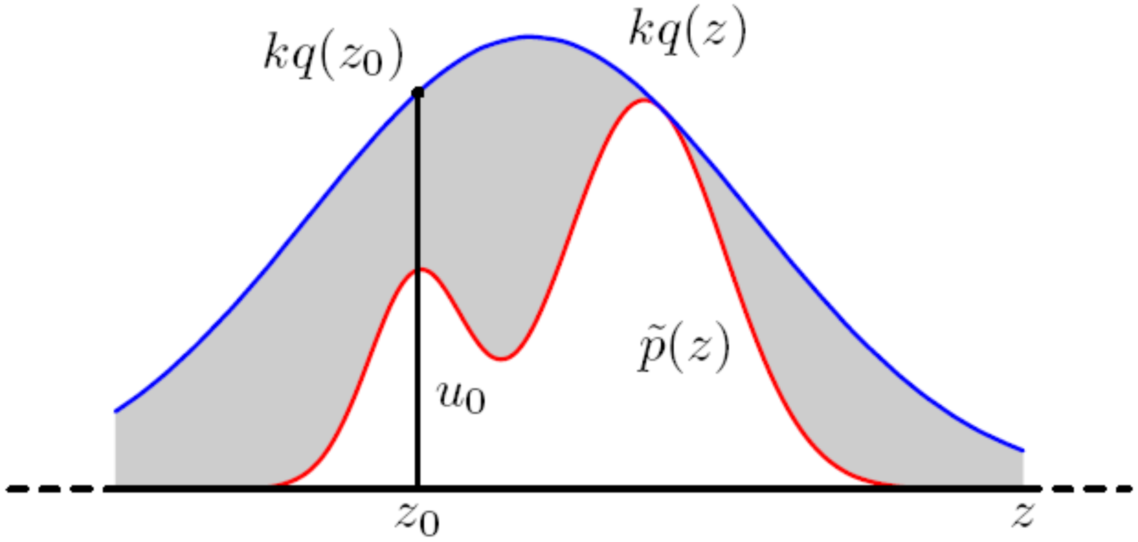
where $\widetilde{p}(z)$ can readily be evaluated, but Z_p is unknown.

proposal distribution: some simpler distribution $q(z)$, from which we draw samples.

k: a constant, whose value is chosen such that $kq(z) \geq \widetilde{p}(z)$ for all z.

$kq(z)$ is called the comparison function.

Rejection Sampling (2/7)



(1) Generate a number z_0 from $q(z)$.

(2) Generate a number u_0 from uniform distribution over $[0, kq(z_0)]$.

(3) If $u_0 > \tilde{p}(z_0)$, the sample is rejected; otherwise, z_0 is retained.

Rejection Sampling (3/7)

1. The pair of random numbers has uniform distribution under the curve of the function *kq*(*z*).
2. The probability that a sample will be accepted is

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

The fraction of points that are rejected by this method depends on the ratio of the area under the unnormalized distribution $\widetilde{p}(z)$ to the area under the curve *kq*(*z*).

$\Rightarrow$ The constant k should be as small as possible subject to the limitation that $kq(z) \geq \widetilde{p}(z)$.

Rejection Sampling (4/7)

Example: Gamma distribution

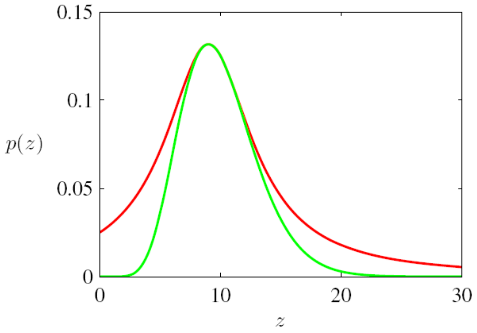
$$\mathrm{Gam}(z|a,b) = \frac{b^a z^{a-1}\exp(-bz)}{\Gamma(a)}$$

Proposal distribution:

Cauchy distribution, which can be generate by transforming a uniform random variable *y* using *z* = *b* tan *y* + *c.*

$$q(z) = \frac{k}{1+(z-c)^2/b^2}.$$

Plot showing the gamma distribution given by (11.15) as the green curve, with a scaled Cauchy proposal distribution shown by the red curve. Samples from the gamma distribution can be obtained by sampling from the Cauchy and then applying the rejection sampling criterion.



Rejection Sampling (5/7)

Adaptive Rejection Sampling

In many instances, it is difficult to determine a suitable analytic form for the envelope distribution *q*(*z*). An alternative approach is to construct the envelope function on the fly based on measured values of the distribution *p*(*z*).

Construction of an envelope function for a log concave *p*(*z*) is simple.

(1) The function ln *p*(*z*) and its gradient are evaluated at some initial set of grid points.

(2) The intersections of the resulting tangent lines are used to construct the envelope function.

$$q(z) = k_i \lambda_i \exp\{-\lambda_i(z - z_{i-1})\} \qquad z_{i-1} < z \leqslant z_i.$$

Importance Sampling (1/7)

Provide a framework for approximating expectations directly but does not itself provide a mechanism for drawing samples from distribution *p*(**z**).

Suppose it is impractical to sample directly from *p*(**z**) but we can evaluate *p*(**z**) easily for any given value of **z**.

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

Importance Sampling (2/7)

The probability distributions of interest often have much of their mass confined to relatively small regions of **z** space.

$\Rightarrow$ Uniform sampling will be very inefficient.

$\Rightarrow$ We would really like to choose the sample points to fall in regions where $p(\mathbf{z})$ is large, or ideally where the product $p(\mathbf{z})f(\mathbf{z})$ is large.

Choose a proposal distribution $q(\mathbf{z})$ from which it is easy to draw samples.

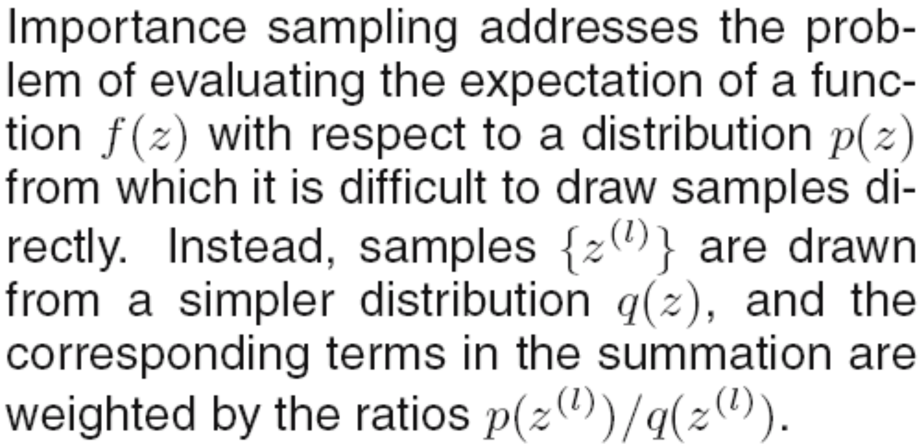
Importance Sampling (3/7)

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

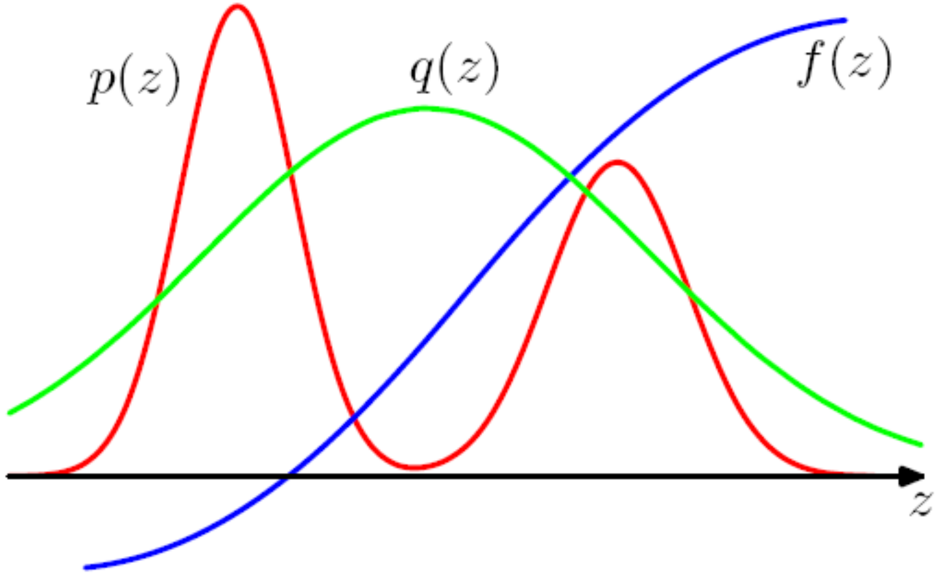
$r_l = p(\mathbf{z}^{(l)})/q(\mathbf{z}^{(l)})$: *importance weights*

Remark: Unlike rejection sampling, all of the samples are retained.

Importance Sampling (4/7)



Importance sampling addresses the problem of evaluating the expectation of a function $f(z)$ with respect to a distribution $p(z)$ from which it is difficult to draw samples directly. Instead, samples $\{z^{(l)}\}$ are drawn from a simpler distribution $q(z)$, and the corresponding terms in the summation are weighted by the ratios $p(z^{(l)})/q(z^{(l)})$.



Importance Sampling (5/7)

Usually, we have $p(\mathbf{z}) = \widetilde{p}(z)/Z_{p,}$ where $\widetilde{p}(z)$ can be evaluated easily, whereas Z_p is unknown.

Similarly, we have $q(\mathbf{z}) = \widetilde{q}(z)/Z_q$.

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

where

$$\widetilde{r}_l = \widetilde{p}(\mathbf{z}^{(l)})/\widetilde{q}(\mathbf{z}^{(l)})$$

Importance Sampling (6/7)

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

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

where

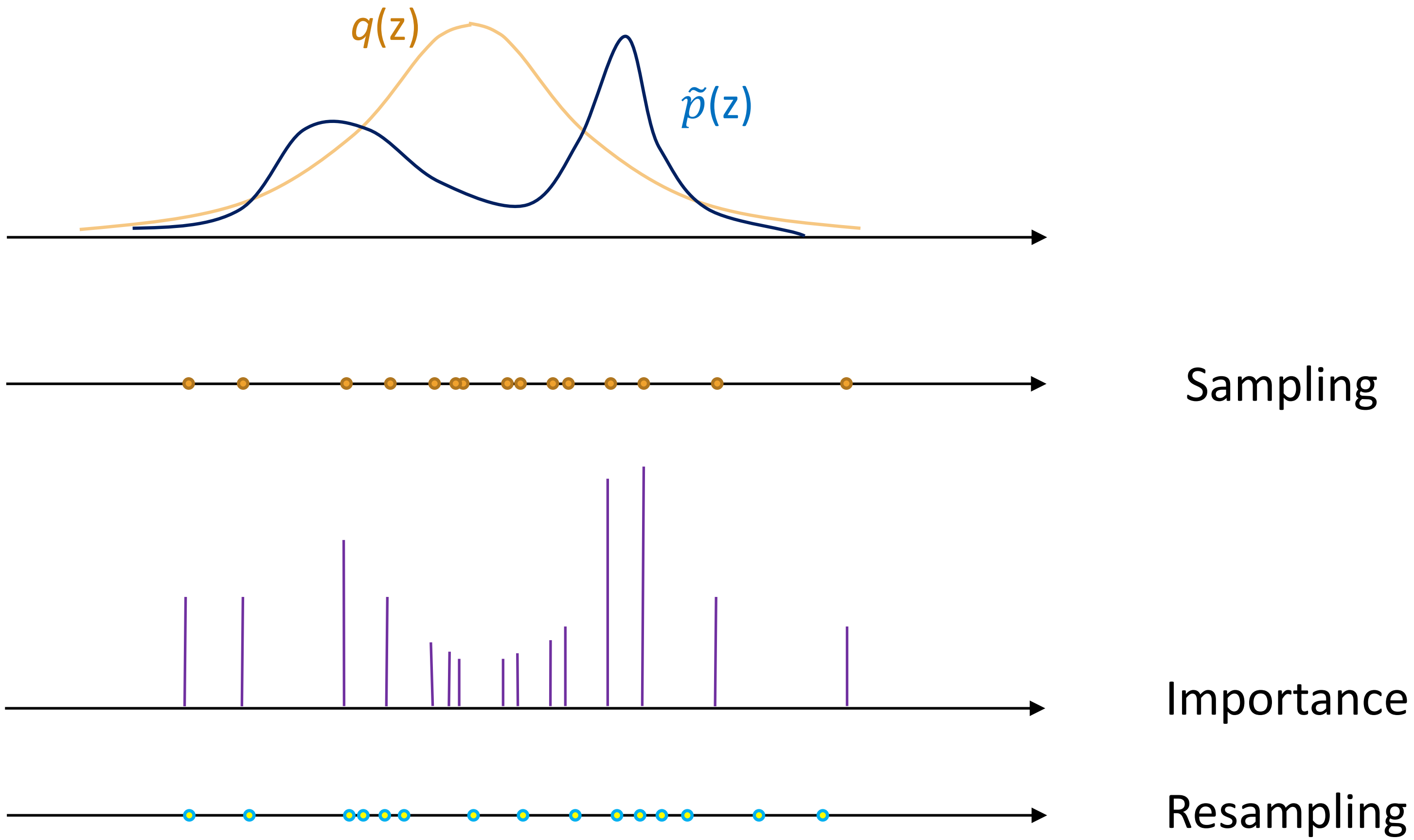
$$w_l = \frac{\widetilde{r}_l}{\sum_m \widetilde{r}_m} = \frac{\widetilde{p}(\mathbf{z}^{(l)})/q(\mathbf{z}^{(l)})}{\sum_m \widetilde{p}(\mathbf{z}^{(m)})/q(\mathbf{z}^{(m)})}.$$

Importance Sampling (7/7)

Remark: If $p(\mathbf{z})f(\mathbf{z})$ is strongly varying and has a significant proportion of its mass concentrated over relatively small regions of **z** space, the set of importance weights $\{r_l\}$ may be dominated by a few weights having large values, with the remaining weights being relatively insignificant. The problem is even more severe if none of the samples falls in the regions where $p(\mathbf{z})f(\mathbf{z})$ is large.

$\Rightarrow$ The sampling distribution $q(\mathbf{z})$ should not be small or zero in regions where $p(\mathbf{z})$ may be significant.

Sampling-Importance-Resampling (1/5)



Sampling-Importance-Resampling (2/5)

Make use of a sampling distribution $q(\mathbf{z})$ but avoid having to determine the constant k.

Two-stage scheme:

(1) L samples $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(L)}$ are drawn from $q(\mathbf{z})$.

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

(2) A second set of L samples is drawn from the discrete distribution $(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(L)})$ with probabilities given by $(w_1, \ldots, w_L)$.

Sampling-Importance-Resampling (3/5)

Remarks:

1. Sampling-Importance-Resampling: an approximation.
   Rejection sampling: draws samples from the true distribution.
2. The resulting L samples are only approximately distributed according to p(z), but the distribution becomes correct in the limit L $\rightarrow \infty$.

$$\begin{aligned} p(z \leqslant a) &= \sum_{l:z^{(l)} \leqslant a} w_l \\ &= \frac{\sum_l I(z^{(l)} \leqslant a)\widetilde{p}(z^{(l)})/q(z^{(l)})}{\sum_l \widetilde{p}(z^{(l)})/q(z^{(l)})} \end{aligned}$$

where *I*(.) is the indicator function.

Sampling-Importance-Resampling (4/5)

As L → ∞

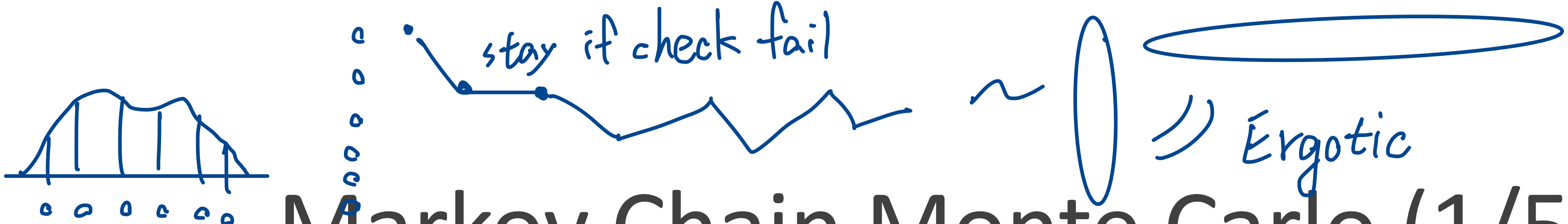
$$
\begin{aligned}
p(z \leqslant a) &= \frac{\displaystyle\int I(z \leqslant a)\,\{\widetilde{p}(z)/q(z)\}\,q(z)\,\mathrm{d}z}{\displaystyle\int \{\widetilde{p}(z)/q(z)\}\,q(z)\,\mathrm{d}z} \\
&= \frac{\displaystyle\int I(z \leqslant a)\widetilde{p}(z)\,\mathrm{d}z}{\displaystyle\int \widetilde{p}(z)\,\mathrm{d}z} \\
&= \int I(z \leqslant a)p(z)\,\mathrm{d}z
\end{aligned}
$$

The approximation improves as the sampling distribution q(z) gets closer to the desired distribution p(z). When q(z) = p(z), the initial samples $(z^{(1)}, \ldots, z^{(L)})$ have the desired distribution, and the weights $w_n = 1/L$.

Sampling-Importance-Resampling (5/5)

If moments with respect to the distribution *p*(**z**) are required, they can be evaluated similar to importance sampling.

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



Markov Chain Monte Carlo (1/5)

--- Have their origins in physics (Metropolis and Ulam, 1949).

--- Allow sampling from a large class of distributions.

--- Scale well with the dimensionality of the sample space.

- ✓ $p(\mathbf{z}) = \widetilde{p}(\mathbf{z})/Z_p$, where $\widetilde{p}(\mathbf{z})$ can be evaluated for any given value of **z**, but the value of Z_p may be unknown.
- ✓ We maintain a record of the current state $\mathbf{z}^{(\tau)}$, and use the proposal distribution $q(\mathbf{z}|\mathbf{z}^{(\tau)})$.
- ✓ The sequence of samples $\mathbf{z}^{(1)}$, $\mathbf{z}^{(2)}$, . . . forms a Markov chain.
- ✓ At each cycle of the algorithm, we generate a candidate sample **z*** from the proposal distribution and accept the sample according to an appropriate criterion.

Markov Chain Monte Carlo (2/5)

- **Markov Chains**

  A first-order Markov chain is a series of random variables $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(M)}$ such that $p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}) = p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)})$ for $m \in \{1, \ldots, M-1\}$.

  Example: Random Walk

$$\begin{aligned} p(z^{(\tau+1)} = z^{(\tau)}) &= 0.5 \\ p(z^{(\tau+1)} = z^{(\tau)} + 1) &= 0.25 \\ p(z^{(\tau+1)} = z^{(\tau)} - 1) &= 0.25 \end{aligned}$$

If the initial state is $z^{(1)} = 0$, then $E[z^{(\tau)}] = 0$, and $E[(z^{(\tau)})^2] = \tau/2$.
After τ steps, the random walk has only travelled a distance that on average is proportional to the square root of τ.
$\Rightarrow$ Very inefficient in exploring the state space.

Markov Chain Monte Carlo (3/5)

A central goal in designing Markov chain Monte Carlo methods is to avoid random walk behaviour.

A Markov chain is called *homogeneous* if the *transition probabilities* $T_m(\mathbf{z}^{(m)}, \mathbf{z}^{(m+1)}) \equiv p(\mathbf{z}^{(m+1)} | \mathbf{z}^{(m)})$ are the same for all m.

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

For a homogeneous Markov chain with transition probabilities $T(\mathbf{z}', \mathbf{z})$, the distribution $p^*(\mathbf{z})$ is invariant if

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

Markov Chain Monte Carlo (4/5)

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

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

for the particular distribution *p**(**z**).

(Pf)

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

Remark: A Markov chain that satisfies detailed balance is said to be *reversible*.

Markov Chain Monte Carlo (5/5)

Our goal is to use Markov chains to sample from a given distribution. We can achieve this if

- we set up a Markov chain such that the desired distribution is invariant; and
- for $m \rightarrow \infty$, the distribution $p(\mathbf{z}^{(m)})$ converges to the required invariant distribution $p^*(\mathbf{z})$, irrespective of the choice of initial distribution $p(\mathbf{z}^{(0)})$. (The *ergodicity* property)

Remarks:

1. The invariant distribution is called the *equilibrium* distribution.
2. It can be shown that a homogeneous Markov chain will be ergodic, subject only to weak restrictions on the invariant distribution and the transition probabilities.

The Metropolis-Hastings Algorithm (1/7)

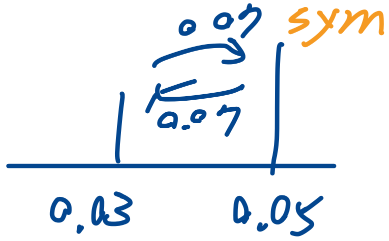
The Metropolis algorithm (Metropolis, 1953)

- Choose a symmetric proposal distribution:

  $q(\mathbf{z}_A|\mathbf{z}_B) = q(\mathbf{z}_B|\mathbf{z}_A)$ for all values of $\mathbf{z}_A$ and $\mathbf{z}_B$.

- The candidate sample is accepted with probability

$$A(\mathbf{z}^\star, \mathbf{z}^{(\tau)}) = \min\left(1, \frac{\widetilde{p}(\mathbf{z}^\star)}{\widetilde{p}(\mathbf{z}^{(\tau)})}\right) \quad \tau \to *$$



symmetric

$0.03 \times \underline{0.07} \times 1$

$0.05 \times \underline{0.07} \times \frac{0.03}{0.05}$

This can be achieved by choosing a random number u with uniform distribution over the period (0, 1) and then accepting the sample if $A(\mathbf{z}^*, \mathbf{z}^{(\tau)}) > u$.

- If the candidate sample is accepted, then $z^{(\tau+1)} = z^*$, otherwise the candidate point z^* is discarded and $z^{(\tau+1)}$ is set to $z^{(\tau)}$.

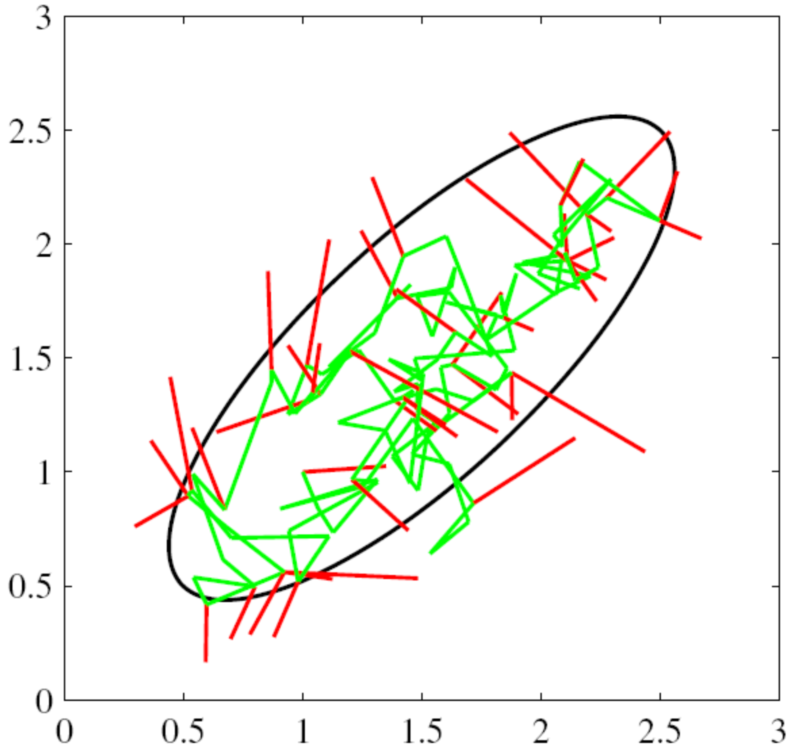
The Metropolis-Hastings Algorithm (2/7)

Remarks:

1. If the step from $\mathbf{z}^{(\tau)}$ to $\mathbf{z}^*$ causes an increase in the value of $p(\mathbf{z})$, then the candidate point is certain to be kept.
2. As long as $q(\mathbf{z}_A|\mathbf{z}_B) > 0$ for any values of $\mathbf{z}_A$ and $\mathbf{z}_B$ (this is a sufficient but not necessary condition), the distribution of $\mathbf{z}^{(\tau)}$ tends to $p(\mathbf{z})$ as $\tau \to \infty$.
3. The sequence $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots$ is not a set of independent samples from $p(\mathbf{z})$. To obtain "independent" samples, we can discard most of the sequence and just retain every M^{th} sample.

The Metropolis-Hastings Algorithm (3/7)

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.



The Metropolis-Hastings Algorithm (4/7)

The Metropolis-Hastings algorithm (Hastings, 1970)

The proposal distribution is no longer a symmetric function of its arguments.

At step τ of the algorithm, in which the current state is $\mathbf{z}^{(\tau)}$, we draw a sample $\mathbf{z}^*$ from the distribution $q_k(\mathbf{z}|\mathbf{z}^{(\tau)})$ and then accept it with probability $A_k(\mathbf{z}^*,\mathbf{z}^{(\tau)})$ where

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

k labels the members of the set of possible transitions being considered.

Remark: For a symmetric proposal distribution, the criterion reduces to the standard Metropolis criterion.

The Metropolis-Hastings Algorithm (5/7)

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

$\Rightarrow$ The detailed balance property is satisfied.

$\Rightarrow$ $p(\mathbf{z})$ is an invariant distribution of the Markov chain.

The Metropolis-Hastings Algorithm (6/7)

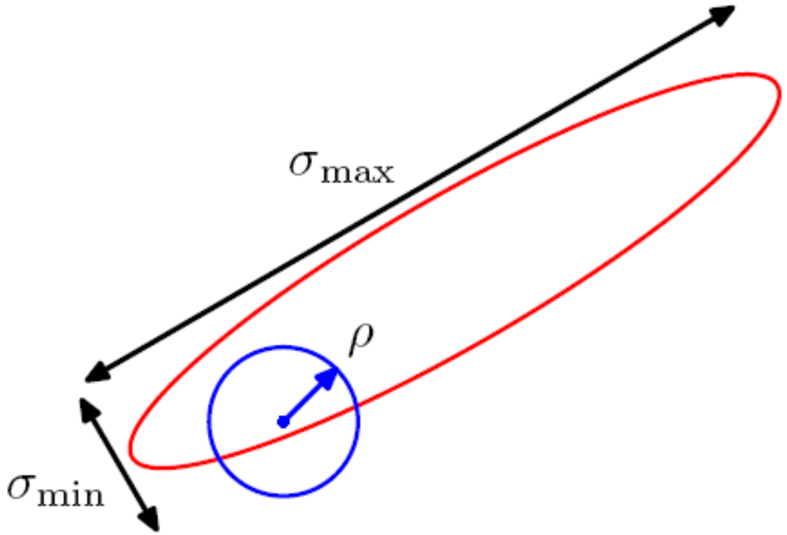
Remarks:

For continuous state spaces, a common choice is a Gaussian centered on the current state.

- If the variance is small, the proportion of accepted transitions will be high, but progress through the state space takes the form of a slow random walk leading to long correlation times.
- If the variance parameter is large, the rejection rate is high because many of the proposed steps will be to states for which the probability $p(\mathbf{z})$ is low.

The Metropolis-Hastings Algorithm (7/7)

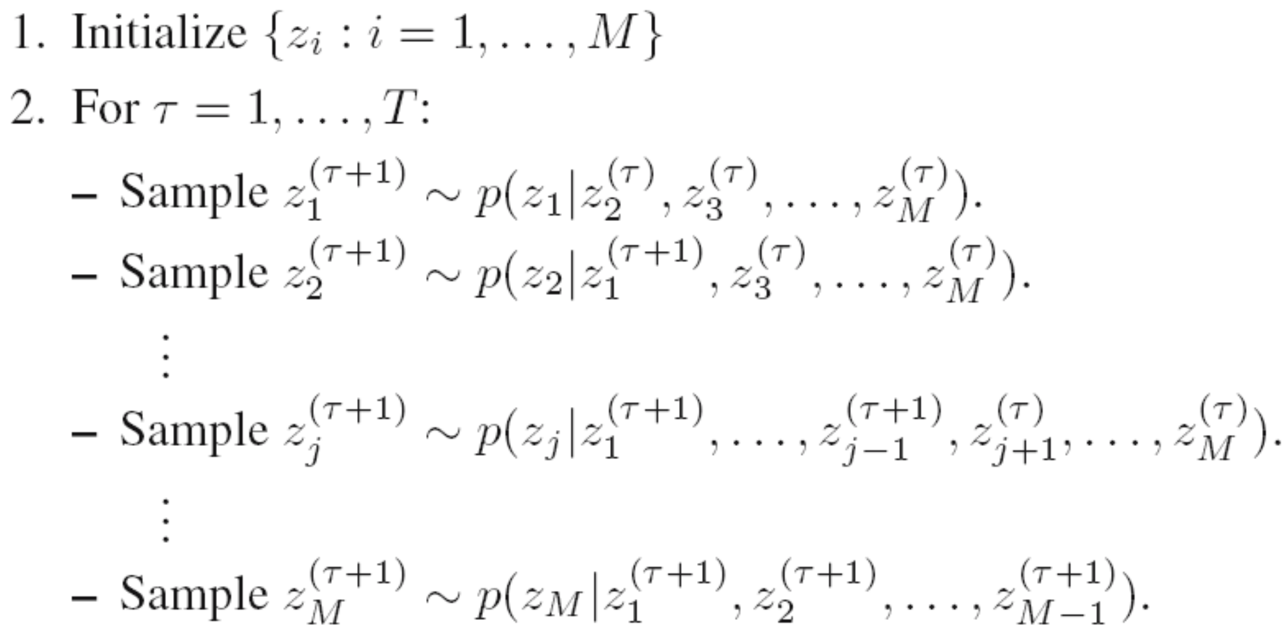
Schematic illustration of the use of an isotropic Gaussian proposal distribution (blue circle) to sample from a correlated multivariate Gaussian distribution (red ellipse) having very different standard deviations in different directions, using the Metropolis-Hastings algorithm. In order to keep the 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.



Gibbs Sampling (1/3)

-- A simple and widely applicable Markov chain Monte Carlo algorithm.

Consider the distribution $p(\mathbf{z}) = p(z_1, \ldots, z_M)$ from which we wish to sample, and suppose that we have chosen some initial state for the Markov chain.



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

一次在一個dimension 上找conditional pdf
更新該dimension 位置

Gibbs Sampling (2/3)

The Gibbs sampling procedure is a particular instance of the Metropolis-Hastings algorithm.

With $q_k(\mathbf{z}^* | \mathbf{z}) = p(z_k^* | \mathbf{z}_{\backslash k})$ and $\mathbf{z}^*_{\backslash k} = \mathbf{z}_{\backslash k}$

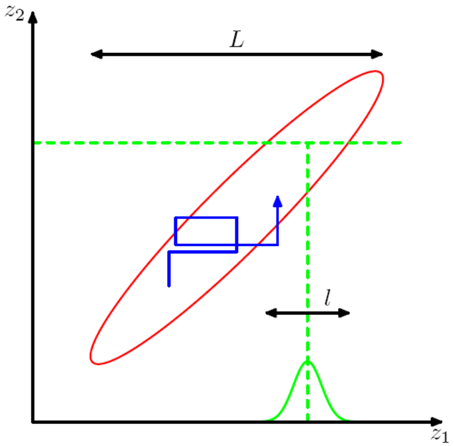
$$A(\mathbf{z}^\star, \mathbf{z}) = \frac{p(\mathbf{z}^\star) q_k(\mathbf{z}|\mathbf{z}^\star)}{p(\mathbf{z}) q_k(\mathbf{z}^\star|\mathbf{z})} = \frac{p(z_k^\star|\mathbf{z}^\star_{\backslash k}) p(\mathbf{z}^\star_{\backslash k}) p(z_k|\mathbf{z}^\star_{\backslash k})}{p(z_k|\mathbf{z}_{\backslash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

一定會跳過去

Gibbs Sampling (3/3)

Example: Consider a correlated Gaussian in two variables, having conditional distributions of width *l* and marginal distributions of width *L*.

Illustration of Gibbs sampling by alternate updates of two variables whose distribution is a correlated Gaussian. The step size is governed by the standard deviation of the conditional distribution (green curve), and is $O(l)$, leading to slow progress in the direction of elongation of the joint distribution (red ellipse). The number of steps needed to obtain an independent sample from the distribution is $O((L/l)^2)$.



For this simple case, we could rotate the coordinate system to decorrelate the variables. However, in practical applications it will generally be infeasible to find such transformations.