MFC CDT Probability and Statistics
Week 7

O. Deniz Akyildiz

Mathematics for our Future Climate: Theory, Data and Simulation (MFC CDT).

November 18, 2024

IMPERIAL

https://akyildiz.me/

𝕏: @odakyildiz

Recall our basic task:

Recall our basic task:

- We want to sample from a distribution $\pi(x) \propto \gamma(x)$ given only the knowledge of $\gamma(x)$.

Recall our basic task:

- We want to sample from a distribution $\pi(x) \propto \gamma(x)$ given only the knowledge of $\gamma(x)$.
- We want to use these samples to estimate an integral

$$(\varphi, \pi) = \int \varphi(x)\pi(x)\,\mathrm{d}x$$

Last week, we have covered the basic sampling techniques:

Last week, we have covered the basic sampling techniques:

- Uniform random number generation
  - Linear congruential generators

Last week, we have covered the basic sampling techniques:

- Uniform random number generation
  - Linear congruential generators
- Inversion (inverse transform) sampling
  - $U \sim \mathcal{U}(0, 1)$
  - $X = F^{-1}(U)$

Last week, we have covered the basic sampling techniques:

- Uniform random number generation
  - Linear congruential generators
- Inversion (inverse transform) sampling
  - $U \sim \mathcal{U}(0, 1)$
  - $X = F^{-1}(U)$
- Rejection sampling
  - $X' \sim q(x)$
  - Accept X' with probability $\gamma(X')/Mq(X')$

Last week, we have covered the basic sampling techniques:

- Uniform random number generation
  - Linear congruential generators
- Inversion (inverse transform) sampling
  - $U \sim \mathcal{U}(0, 1)$
  - $X = F^{-1}(U)$
- Rejection sampling
  - $X' \sim q(x)$
  - Accept X' with probability $\gamma(X')/Mq(X')$

The code is also available for these parts:

`https://akyildiz.me/mfc-probability-and-stats/Week-6/intro.html`

Now, we will first look at Monte Carlo integration and importance sampling.

Another popular approach to compute expectations (φ, π) is called *importance sampling*.

Another popular approach to compute expectations (φ, π) is called *importance sampling*.

Assume, as in the rejection sampling case, π is absolutely continuous w.r.t. q, denoted as $\pi \ll q$, meaning $q(x) = 0 \implies \pi(x) = 0$.

Another popular approach to compute expectations (φ, π) is called *importance sampling*.

Assume, as in the rejection sampling case, π is absolutely continuous w.r.t. q, denoted as $\pi \ll q$, meaning $q(x) = 0 \implies \pi(x) = 0$.

Then, we can write

$$(\varphi, \pi) = \int \varphi(x)\pi(\mathrm{d}x) = \int \varphi(x)\frac{\mathrm{d}\pi}{\mathrm{d}q}(x)q(x)\mathrm{d}x.$$

When π and q admit densities,

$$(\varphi, \pi) = \int \varphi(x)\pi(x)\mathrm{d}x = \int \varphi(x)\frac{\pi(x)}{q(x)}q(x)\mathrm{d}x.$$

Given

$$(\varphi, \pi) = \int \varphi(x) \frac{\pi(x)}{q(x)} q(x) \mathrm{d}x,$$

we can employ standard Monte Carlo by sampling $X_i \sim q$ and then constructing (by setting $w = \pi/q$)

$$\begin{aligned}(\varphi, \tilde{\pi}^N) &= \frac{1}{N} \sum_{i=1}^{N} \varphi(X_i) w(X_i), \\ &= \frac{1}{N} \sum_{i=1}^{N} \mathrm{w}_i \varphi(X_i).\end{aligned}$$

where $\mathrm{w}_i = w(X_i)$. We will call this estimator the importance sampling (IS) estimator.

Mini-quiz: Is this estimator unbiased?

Mini-quiz: Is this estimator unbiased?

Yes.

$$
\begin{aligned}
\mathbb{E}_q[(\varphi, \tilde{\pi}^N)] &= \mathbb{E}_q\left[\frac{1}{N}\sum_{i=1}^{N} \mathrm{w}_i \varphi(X_i)\right], \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_q\left[\frac{\pi(X_i)}{q(X_i)}\varphi(X_i)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N} \int \frac{\pi(x)}{q(x)}\varphi(x)q(x)\mathrm{d}x \\
&= \int \varphi(x)\pi(x)\mathrm{d}x = (\varphi, \pi).
\end{aligned}
$$

What is the variance?

$$\begin{aligned}
\mathrm{var}_q[(\varphi, \tilde{\pi}^N)] &= \mathrm{var}_q\left[\frac{1}{N}\sum_{i=1}^{N} \mathrm{w}_i \varphi(X_i)\right] \\
&= \frac{1}{N^2}\mathrm{var}_q\left[\sum_{i=1}^{N} w(X_i)\varphi(X_i)\right] \\
&= \frac{1}{N}\mathrm{var}_q\left[w(X)\varphi(X)\right] \qquad \text{where } X \sim q(x) \\
&= \frac{1}{N}\left(\mathbb{E}_q\left[w^2(X)\varphi^2(X)\right] - \mathbb{E}_q\left[w(X)\varphi(X)\right]^2\right) \\
&= \frac{1}{N}\left(\mathbb{E}_q\left[w^2(X)\varphi^2(X)\right] - \bar{\varphi}^2\right).
\end{aligned}$$

Finally, the basic IS estimator satisfies the following L_p bound just like the perfect Monte Carlo

$$\|(\varphi, \pi) - (\varphi, \tilde{\pi}^N)\|_p \leq \frac{\tilde{c}_p \|\varphi\|_\infty}{\sqrt{N}},$$

where $\tilde{c}_p$ is a constant depending on p and q.

What if we only have access to $\gamma(x) \propto \pi(x)$?

What if we only have access to $\gamma(x) \propto \pi(x)$?

Assume $\gamma \ll q$ and both abs. cont w.r.t. to the Lebesgue measure. Then we can write

$$\begin{aligned}(\varphi, \pi) &= \int \varphi(x)\pi(x)\mathrm{d}x \\ &= \frac{\int \varphi(x)\frac{\gamma(x)}{q(x)}q(x)\mathrm{d}x}{\int \frac{\gamma(x)}{q(x)}q(x)\mathrm{d}x}.\end{aligned}$$

We can then perform the same Monte Carlo integration idea but now both for the numerator and denominator.

We have

$$(\varphi, \pi) = \int \varphi(x)\pi(x)\mathrm{d}x$$
$$= \frac{\int \varphi(x)\frac{\gamma(x)}{q(x)}q(x)\mathrm{d}x}{\int \frac{\gamma(x)}{q(x)}q(x)\mathrm{d}x}.$$

Define $W(x) = \gamma(x)/q(x)$ and the SNIS approximation is given as

$$(\varphi, \pi) = \frac{\int \varphi(x)W(x)q(x)\mathrm{d}x}{\int W(x)q(x)\mathrm{d}x} \approx \frac{\frac{1}{N}\sum_{i=1}^{N}\varphi(X_i)W(X_i)}{\frac{1}{N}\sum_{i=j}^{N}W(X_j)}.$$

where $X_i \sim q(x)$. Let us write $\mathrm{W}_i = W(X_i)$ and $\mathrm{w}_i = \mathrm{W}_i/\sum_{j=1}^{N}\mathrm{W}_j$. Then the final estimator is

$$(\varphi, \tilde{\pi}^N) = \sum_{i=1}^{N}\mathrm{w}_i.\varphi(X_i)$$

Mini-quiz: Is this estimator unbiased?

Mini-quiz: Is this estimator unbiased?

No.

Mini-quiz: Is this estimator unbiased?

No.

The estimator is a ratio of two unbiased estimators. However, this ratio is *not* unbiased.

However, one can prove that

$$\|(\varphi, \pi) - (\varphi, \tilde{\pi}^N)\|_p \leq \frac{\tilde{c}_p \|\varphi\|_\infty}{\sqrt{N}},$$

where $\tilde{c}_p$ is a constant depending on p and q and φ is bounded.

Theorem 1

The MSE (i.e., set $p = 2$ and square both sides) is bounded by

$$\mathbb{E}\left[\left((\varphi, \pi) - (\varphi, \tilde{\pi}^N)\right)^2\right] \leq \frac{4\|\varphi\|_\infty \rho}{N},$$

where

$$\rho = \chi^2(\pi || q) + 1.$$

Suggests that the discrepancy between π and q controls the MSE.

Proof. We first note the following inequalities,

$$
\begin{aligned}
|(\varphi, \pi) - (\varphi, \tilde{\pi}^N)| &= \left| \frac{(\varphi W, q)}{(W, q)} - \frac{(\varphi W, q^N)}{(W, q^N)} \right| \\
&\leq \frac{\left|(\varphi W, q) - (\varphi W, q^N)\right|}{|(W, q)|} + |(\varphi W, q^N)| \left| \frac{1}{(W, q)} - \frac{1}{(W, q^N)} \right| \\
&= \frac{\left|(\varphi W, q) - (\varphi W, q^N)\right|}{|(W, q)|} + \|\varphi\|_\infty |(W, q^N)| \left| \frac{(W, q^N) - (W, q)}{(W, q)(W, q^N)} \right| \\
&= \frac{\left|(\varphi W, q) - (\varphi W, q^N)\right|}{(W, q)} + \frac{\|\varphi\|_\infty |(W, q^N) - (W, q)|}{(W, q)}.
\end{aligned}
$$

We take squares of both sides and apply the inequality $(a+b)^2 \leq 2(a^2+b^2)$ to further bound the rhs,

$$\cdots \leq 2\frac{\left|(\varphi W, q) - (\varphi W, q^N)\right|^2}{(W,q)^2} + 2\frac{\|\varphi\|_\infty^2 |(W,q^N) - (W,q)|^2}{(W,q)^2}$$

We can now take the expectation of both sides,

$$\mathbb{E}\left[\left((\varphi,\pi) - (\varphi,\tilde{\pi}^N)\right)^2\right] \leq \frac{2\mathbb{E}\left[\left((\varphi W, q) - (\varphi W, q^N)\right)^2\right]}{(W,q)^2} + \frac{2\|\varphi\|_\infty^2 \mathbb{E}\left[\left((W,q^N) - (W,q)\right)^2\right]}{(W,q)^2}.$$

Note that, both terms in the right hand side are perfect Monte Carlo estimates of the integrals.

Bounding the MSE of these integrals yields

$$\cdots \leq \frac{2}{N}\frac{(\varphi^2 W^2, q) - (\varphi W, q)^2}{(W, q)^2} + \frac{2\|\varphi\|_\infty^2}{N}\frac{(W^2, q) - (W, q)^2}{(W, q)^2},$$
$$\leq \frac{2\|\varphi\|_\infty^2}{N}\frac{(W^2, q)}{(W, q)^2} + \frac{2\|\varphi\|_\infty^2}{N}\frac{(W^2, q) - (W, q)^2}{(W, q)^2}.$$

Therefore, we can straightforwardly write,

$$\mathbb{E}\left[\left((\varphi, \pi) - (\varphi, \tilde{\pi}^N)\right)^2\right] \leq \frac{4\|\varphi\|_\infty^2}{(W, q)^2}\frac{(W^2, q)}{N}.$$

$$\mathbb{E}\left[\left((\varphi,\pi)-(\varphi,\tilde{\pi}^N)\right)^2\right] \leq \frac{4\|\varphi\|_\infty^2}{(W,q)^2}\frac{(W^2,q)}{N}.$$

Now it remains to show the relation of the bound to χ^2 divergence. Note that,

$$\begin{aligned}
\frac{(W^2,q)}{(W,q)^2} &= \frac{\int \frac{\Pi^2(x)}{q^2(x)}q(x)\mathrm{d}x}{\left(\int \frac{\Pi(x)}{q(x)}q(x)\mathrm{d}x\right)^2} \\
&= \frac{Z^2\int \frac{\pi^2(x)}{q^2(x)}q(x)\mathrm{d}x}{Z^2\left(\int \pi \mathrm{d}x\right)^2} \\
&= \mathbb{E}_q\left[\frac{\pi^2(X)}{q^2(X)}\right] := \rho.
\end{aligned}$$

Note that ρ is not exactly χ^2 divergence, which is defined as $\rho - 1$. Plugging everything into our bound, we have the result,

$$\mathbb{E}\left[\left((\varphi,\pi)-(\varphi,\pi^N)\right)^2\right] \leq \frac{4\|\varphi\|_\infty^2\rho}{N}.$$

■

Let us exemplify a few issues. Consider the following target distribution on $\mathbb{R}^d$:

$$\pi(x) = \frac{1}{\sigma_\pi^d (2\pi)^{d/2}} \exp\left(-\frac{1}{2\sigma_\pi^2}\|x\|^2\right)$$

and the following proposal distribution:

$$q(x) = \frac{1}{\sigma_q^d (2\pi)^{d/2}} \exp\left(-\frac{1}{2\sigma_q^2}\|x\|^2\right)$$

where $\sigma_q > \sigma_\pi$.

We know that the acceptance probability is

$$\alpha(x) = \frac{\pi(x)}{Mq(x)}.$$

Mini-quiz: How do we choose M?

The curse of dimensionality

Rejection sampling as $d \to \infty$

We know that the acceptance probability is

$$\alpha(x) = \frac{\pi(x)}{Mq(x)}.$$

Mini-quiz: How do we choose M?

$$M = \sup_{x \in \mathbb{R}^d} \frac{\pi(x)}{q(x)}.$$

We know that the acceptance probability is

$$\alpha(x) = \frac{\pi(x)}{Mq(x)}.$$

Mini-quiz: How do we choose M?

$$M = \sup_{x \in \mathbb{R}^d} \frac{\pi(x)}{q(x)}.$$

Then, we can write

$$M = \sup_{x \in \mathbb{R}^d} \frac{\sigma_q}{\sigma_\pi} \exp\left(-\frac{1}{2\sigma_\pi^2} \|x\|^2 + \frac{1}{2\sigma_q^2} \|x\|^2 \right)$$
$$= \frac{\sigma_q^d}{\sigma_\pi^d} \sup_{x \in \mathbb{R}^d} \exp\left(\frac{\sigma_\pi^2 - \sigma_q^2}{2\sigma_q^2 \sigma_\pi^2} \|x\|^2 \right) = \frac{\sigma_q^d}{\sigma_\pi^d}.$$

The curse of dimensionality

Rejection sampling as $d \to \infty$

Mini-quiz: Given M, what is the acceptance rate?

The curse of dimensionality

Rejection sampling as $d \to \infty$

Mini-quiz: Given M, what is the acceptance rate?

$$\hat{a} = \frac{1}{M} = \frac{\sigma_\pi^d}{\sigma_q^d}.$$

This means that as $d \to \infty$, given $\sigma_q > \sigma_\pi$, $\hat{a} \to 0$.

The curse of dimensionality for rejection samplers.

In standard Monte Carlo methods course, you would hear things like

The curse of dimensionality

Importance sampling as $d \to \infty$

In standard Monte Carlo methods course, you would hear things like

- Monte Carlo estimators are independent of the dimension of the problem.

The curse of dimensionality

Importance sampling as $d \to \infty$

In standard Monte Carlo methods course, you would hear things like

- Monte Carlo estimators are independent of the dimension of the problem.
- Importance sampling estimators are also independent of the dimension of the problem.

The curse of dimensionality

Importance sampling as $d \to \infty$

In standard Monte Carlo methods course, you would hear things like

- Monte Carlo estimators are independent of the dimension of the problem.
- Importance sampling estimators are also independent of the dimension of the problem.

These are false statements.

Importance sampling estimators also suffer badly as $d \to \infty$ (Li et al., 2005).

This motivates us to move on to our next topic: Markov chain Monte Carlo methods.

- In both high-dimensional sampling and more generally generative modelling, techniques based on MCMC and similar ideas are the state-of-the-art.

This motivates us to move on to our next topic: Markov chain Monte Carlo methods.

- In both high-dimensional sampling and more generally generative modelling, techniques based on MCMC and similar ideas are the state-of-the-art.
- Of course, there are many other techniques that are used in practice, but MCMC is the most popular one.

This motivates us to move on to our next topic: Markov chain Monte Carlo methods.

- In both high-dimensional sampling and more generally generative modelling, techniques based on MCMC and similar ideas are the state-of-the-art.
- Of course, there are many other techniques that are used in practice, but MCMC is the most popular one.

Next up: Introducing MCMC.

Why Markov chains?

Since we want i.i.d samples

Let K be a Markov kernel. Let $(X_1, X_2, \ldots)$ be a sequence of random variables such that $X_{n+1} \sim K(X_n, \cdot)$.

Why Markov chains?

Since we want i.i.d samples

Let K be a Markov kernel. Let $(X_1, X_2, \ldots)$ be a sequence of random variables such that $X_{n+1} \sim K(X_n, \cdot)$.

Theorem 1

If K is an irreducible, π-invariant kernel, then for any integrable function φ

$$\lim_{T\to\infty} \frac{1}{T}\sum_{i=1}^{T}\varphi(X_i) = \int \varphi(x)\pi(x)\mathrm{d}x = (\varphi, \pi),$$

almost surely, for almost all initial points x_0.

Let K be a Markov kernel. Let $(X_1, X_2, \ldots)$ be a sequence of random variables such that $X_{n+1} \sim K(X_n, \cdot)$.

Theorem 1

If K is an irreducible, π-invariant kernel, then for any integrable function φ

$$\lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{T} \varphi(X_i) = \int \varphi(x)\pi(x)\mathrm{d}x = (\varphi, \pi),$$

almost surely, for almost all initial points x_0.

Therefore, we can use these samples to estimate our integrals.

Why Markov chains?

Since we want i.i.d samples

Theorem 2

If K is irreducible, aperiodic, and π-invariant, then

$$\lim_{T\to\infty} \int_{\mathrm{X}} |K^T(y|x) - \pi(y)| \mathrm{d}y = 0,$$

for π-almost all starting values x.

We will first look at a surprisingly simple approach: the Metropolis-Hastings algorithm.

We will first look at a surprisingly simple approach: the Metropolis-Hastings algorithm.

This approach relies on the following idea:

We will first look at a surprisingly simple approach: the Metropolis-Hastings algorithm.

This approach relies on the following idea:

- We can sample from a proposal $q(x|x')$ (that is a Markov kernel)

We will first look at a surprisingly simple approach: the Metropolis-Hastings algorithm.

This approach relies on the following idea:

- We can sample from a proposal $q(x|x')$ (that is a Markov kernel)
- We can use accept/reject

We will first look at a surprisingly simple approach: the Metropolis-Hastings algorithm.

This approach relies on the following idea:

- We can sample from a proposal $q(x|x')$ (that is a Markov kernel)
- We can use accept/reject

We can design the process so that the stationary distribution of the chain is the target distribution.

We will first look at a surprisingly simple approach: the Metropolis-Hastings algorithm.

This approach relies on the following idea:

- We can sample from a proposal $q(x|x')$ (that is a Markov kernel)
- We can use accept/reject

We can design the process so that the stationary distribution of the chain is the target distribution.

This is however very different from the rejection sampling approach.

Metropolis-Hastings

A first look

Consider the following method:

- Sample $X' \sim q(x'|X_{n-1})$
- Set $X_n = X'$ with probability

$$\alpha(X'|X_{n-1}) = \min\left\{1, \frac{\pi(X')q(X_{n-1}|X')}{\pi(X_{n-1})q(X'|X_{n-1})}\right\}.$$

- Otherwise, set $X_n = X_{n-1}$.

Metropolis-Hastings

A first look

Consider the following method:

- Sample $X' \sim q(x'|X_{n-1})$
- Set $X_n = X'$ with probability

$$\alpha(X'|X_{n-1}) = \min\left\{1, \frac{\pi(X')q(X_{n-1}|X')}{\pi(X_{n-1})q(X'|X_{n-1})}\right\}.$$

- Otherwise, set $X_n = X_{n-1}$.

Note the last step: we discard the sample X' if rejected BUT set $X_n = X_{n-1}$.

The ratio

$$\mathrm{r}(x, x') = \frac{\pi(x')q(x|x')}{\pi(x)q(x'|x)},$$

is called acceptance ratio.

The MH algorithm automatically gives us a kernel.

The MH algorithm automatically gives us a kernel.

How to prove that the stationary distribution is the target distribution?

Let us figure out the kernel.

Let us figure out the kernel.

Let us say, we have the sample from the proposal x'. Fixing this sample, the acceptance step samples from the mixture (*intuitively*):

$$\alpha(x'|x)\delta_{x'}(y) + (1 - \alpha(x'|x))\delta_x(y).$$

To get the full kernel, we need to integrate over x':

$$\begin{aligned} K(y|x) &= \int q(x'|x) \left(\alpha(x'|x)\delta_{x'}(y) + (1 - \alpha(x'|x))\delta_x(y) \right) \mathrm{d}x', \\ &= \alpha(y|x)q(y|x) + (1 - a(x))\delta_x(y) \end{aligned}$$

where

$$a(x) = \int \alpha(x'|x)q(x'|x)\mathrm{d}x'.$$

More intuition in terms of x_n and x_{n-1}:

- What is the probability of being at x_{n-1} and getting accepted?

$$a(x_{n-1}) = \int_{\mathrm{X}} \alpha(x|x_{n-1}) q(x|x_{n-1}) \mathrm{d}x.$$

- Therefore, the probability of being at x_{n-1} and getting rejected is $1 - a(x_{n-1})$.

We can see that the kernel is

$$K(x_n|x_{n-1}) = \alpha(x_n|x_{n-1}) q(x_n|x_{n-1}) + (1 - a(x_{n-1}))\delta_{x_{n-1}}(x_n).$$

We can now prove that the kernel satisfies the detailed balance condition:

$$K(x'|x)\pi(x) = K(x|x')\pi(x').$$

$$
\begin{aligned}
\pi(x)K(x'|x) &= \pi(x)q(x'|x)\alpha(x',x) + \pi(x)(1-a(x))\delta_x(x') \\
&= \pi(x)q(x'|x)\min\left\{1, \frac{\pi(x')q(x|x')}{\pi(x)q(x'|x)}\right\} + \pi(x)(1-a(x))\delta_x(x') \\
&= \min\left\{\pi(x)q(x'|x), \pi(x')q(x|x')\right\} + \pi(x)(1-a(x))\delta_x(x') \\
&= \min\left\{\frac{\pi(x)q(x'|x)}{\pi(x')q(x|x')}, 1\right\}\pi(x')q(x|x') + \pi(x')(1-a(x'))\delta_{x'}(x) \\
&= K(x|x')\pi(x').
\end{aligned}
$$

Assume we are given an unnormalised density to sample γ where

$$\pi(x) = \frac{\gamma(x)}{Z},$$

where Z is the normalisation constant.

- Sample $X' \sim q(x'|X_{n-1})$
- Set $X_n = X'$ with probability

$$\alpha(X'|X_{n-1}) = \min\left\{1, \frac{\gamma(X')q(X_{n-1}|X')}{\gamma(X_{n-1})q(X'|X_{n-1})}\right\}.$$

- Otherwise, set $X_n = X_{n-1}$.

as the normalising constants of π would cancel out.

Metropolis-Hastings

How do we choose proposals?

- Independent proposals
- Symmetric (random walk) proposals
- Gradient-based proposals
- Adaptive proposals

Metropolis-Hastings

Independent proposals

Choose the proposal $q(x)$ independently of the current state X_{n-1}. Leads to

- $X' \sim q(x')$
- Accept with probability

$$\alpha(X'|X_{n-1}) = \min\left\{1, \frac{\pi(X')q(X_{n-1})}{\pi(X_{n-1})q(X')}\right\}.$$

- Otherwise, set $X_n = X_{n-1}$.

Metropolis-Hastings

Independent proposals

Let us say

$$\pi(x) = \mathcal{N}(x; \mu, \sigma^2)$$

For the example, assume we want to use MH to sample from it. Choose a proposal

$$q(x) = \mathcal{N}(x; \mu_q, \sigma_q^2).$$

How to compute the acceptance ratio?

$$
\begin{aligned}
\mathrm{r}(x, x') &= \frac{\pi(x')q(x)}{\pi(x)q(x')} \\
&= \frac{\mathcal{N}(x'; \mu, \sigma^2)\mathcal{N}(x; \mu_q, \sigma_q^2)}{\mathcal{N}(x; \mu, \sigma^2)\mathcal{N}(x'; \mu_q, \sigma_q^2)} \\
&= \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x'-\mu)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right)}{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(x'-\mu_q)^2}{2\sigma_q^2}\right)} \\
&= \frac{\exp\left(-\frac{(x'-\mu)^2}{2\sigma^2}\right) \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right)}{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \exp\left(-\frac{(x'-\mu_q)^2}{2\sigma_q^2}\right)} \\
&= e^{\left(-\frac{1}{2\sigma^2}\left[(x'-\mu)^2-(x-\mu)^2\right]\right)} e^{\left(-\frac{1}{2\sigma_q^2}\left[(x-\mu_q)^2-(x'-\mu_q)^2\right]\right)}
\end{aligned}
$$

We can choose:

$$q(x'|x) = \mathcal{N}(x'; x, \sigma_q^2)$$

The proposal looks at where we are and take a random step (random walk).

We can choose:

$$q(x'|x) = \mathcal{N}(x'; x, \sigma_q^2)$$

The proposal looks at where we are and take a random step (random walk).

Note that $q(x'|x)$ is symmetric, i.e. $q(x|x') = q(x'|x)$.

Acceptance ratio:

$$
\begin{aligned}
\mathrm{r}(x, x') &= \frac{\pi(x')q(x|x')}{\pi(x)q(x'|x)} \\
&= \frac{\pi(x')}{\pi(x)}, \\
&= \frac{\mathcal{N}(x'; \mu, \sigma^2)}{\mathcal{N}(x; \mu, \sigma^2)} \\
&= e^{\left(-\frac{1}{2\sigma^2}\left[(x'-\mu)^2-(x-\mu)^2\right]\right)}.
\end{aligned}
$$

Set a burnin period:

- Run the sampler for fixed number of iterations and discard the first n samples.
- This accounts for the convergence to the stationary measure.

We can *inform* the proposal by using the gradient of the target distribution.

$$q(x'|x) = \mathcal{N}(x'; x + \gamma \nabla \log \pi(x), 2\gamma I),$$

This tends to behave really well.

We can *inform* the proposal by using the gradient of the target distribution.

$$q(x'|x) = \mathcal{N}(x'; x + \gamma \nabla \log \pi(x), 2\gamma I),$$

This tends to behave really well.

This approach is called *Metropolis adjusted Langevin algorithm* (MALA). (more on these later)

Metropolis-Hastings

Caveats, design rules

- One has to be careful that $p/q < \infty$ (while no theoretical reason, the performance tends to be quite bad).

Metropolis-Hastings

Caveats, design rules

- One has to be careful that $p/q < \infty$ (while no theoretical reason, the performance tends to be quite bad).
- The proposal should attain a balance of acceptance rate and efficiency.

- One has to be careful that $p/q < \infty$ (while no theoretical reason, the performance tends to be quite bad).
- The proposal should attain a balance of acceptance rate and efficiency.
- Too high acceptance rate is **not** necessarily good: You might be taking too small steps and getting stuck in some regions

Consider the 2D density

$$p(x, y) \propto \exp\left(-\frac{x^2}{10} - \frac{y^4}{10} - 2(y - x^2)^2\right).$$

Assume we would like to sample from it.

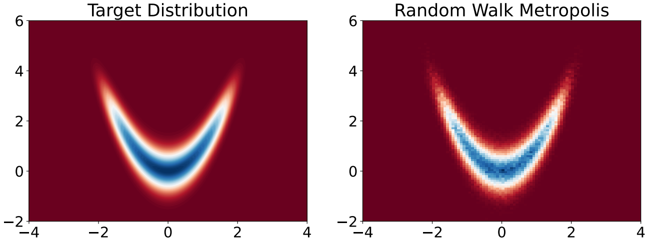

Figure: The banana density (unnormalised)

We have

$$\gamma(x,y) = \exp\left(-\frac{x^2}{10} - \frac{y^4}{10} - 2(y-x^2)^2\right).$$

and let us choose two alternative proposals

- The random walk proposal:

$$q(x',y'|x,y) = \mathcal{N}(x';x,\sigma_q^2)\mathcal{N}(y';y,\sigma_q^2).$$

- and the gradient-based proposal (MALA):

$$q(x',y'|x,y) = \mathcal{N}(z;z+\gamma\nabla\log\gamma(z),\sqrt{2\gamma}\mathbf{I}).$$

where $z=(x,y)$ and γ is a step size.

We have seen Metropolis-Hastings sampler.

We have seen Metropolis-Hastings sampler.

- Unfortunately, it may not be very efficient.

We have seen Metropolis-Hastings sampler.

- Unfortunately, it may not be very efficient.
- Acceptance ratios are very tricky to compute in a variety of settings:

We have seen Metropolis-Hastings sampler.

- Unfortunately, it may not be very efficient.
- Acceptance ratios are very tricky to compute in a variety of settings:
  - High-dimensional problems

We have seen Metropolis-Hastings sampler.

- Unfortunately, it may not be very efficient.
- Acceptance ratios are very tricky to compute in a variety of settings:
  - High-dimensional problems
  - Complex models

We have seen Metropolis-Hastings sampler.

- Unfortunately, it may not be very efficient.
- Acceptance ratios are very tricky to compute in a variety of settings:
  - High-dimensional problems
  - Complex models
  - Large datasets

We have seen Metropolis-Hastings sampler.

- Unfortunately, it may not be very efficient.
- Acceptance ratios are very tricky to compute in a variety of settings:
  - High-dimensional problems
  - Complex models
  - Large datasets

Next week, we will look at Langevin MCMC methods.

References I

Li, Bo, Thomas Bengtsson, and Peter Bickel (2005). "Curse-of-dimensionalit revisited: Collapse of importance sampling in very large scale systems". In: *Rapport technique* 85, p. 205.