Markov Chain Monte Carlo

Nadia Chirkova

Higher School of Economics, Samsung-HSE Laboratory
Moscow, Russia

NATIONAL RESEARCH UNIVERSITY

SAMSUNG Research

p(B|A)yesgroup.ru

Approximate inference

Probabilistic model: $p(Y, \theta | X) = p(Y | X, \theta) p(\theta)$ $\qquad p(\theta | X, Y) - ?$

Variational inference

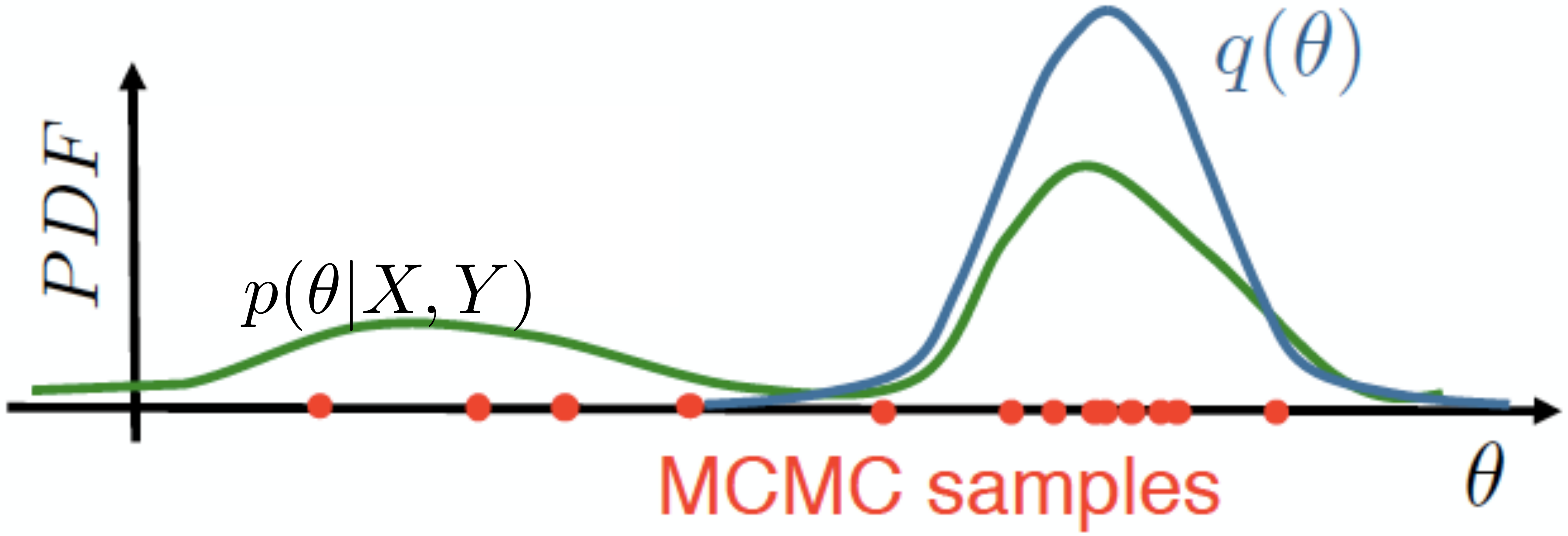
Approximate $p(\theta | X, Y) \approx q(\theta) \in \mathcal{Q}$

- Biased
- Faster and more scalable

MCMC

Sample from unnormalized $p(\theta | X, Y)$

- Unbiased
- A lot of samples are needed



Slide From Ekaterina Lobacheva's presentation, deepbayes.ru/2019

Prediction in Bayesian models

$$p(y_*|x_*, X, Y) = \mathbb{E}_{p(\theta|X,Y)} p(y_*|x_*, \theta)$$

Prediction in Bayesian models

$$p(y_*|x_*, X, Y) = \mathbb{E}_{p(\theta|X,Y)} p(y_*|x_*, \theta) \approx \frac{1}{K}\sum_{k=1}^{K} p(y_*|x_*, \theta^k)$$

$$\theta^k \sim p(\theta|X, Y)$$

We only need to know how to sample from from the posterior!

Prediction in Bayesian models

$$p(y_*|x_*, X, Y) = \mathbb{E}_{p(\theta|X,Y)} p(y_*|x_*, \theta) \approx \frac{1}{K}\sum_{k=1}^{K} p(y_*|x_*, \theta^k)$$

$$\theta^k \sim p(\theta|X, Y)$$

We only need to know how to sample from from the **unnormalized** posterior!

$$p(\theta|X,Y) = \frac{\overbrace{\boxed{p(Y|X,\theta)p(\theta)}}^{\text{unnormalized posterior}}}{\underbrace{\boxed{\int p(Y|X,\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}}}_{\text{intractable}}}$$

Approximating expectation

$$\mathbb{E}_{p(z)} f(z) \approx \frac{1}{K} \sum_{k=1}^{K} f(x^k), \quad z^k \sim p(z)$$

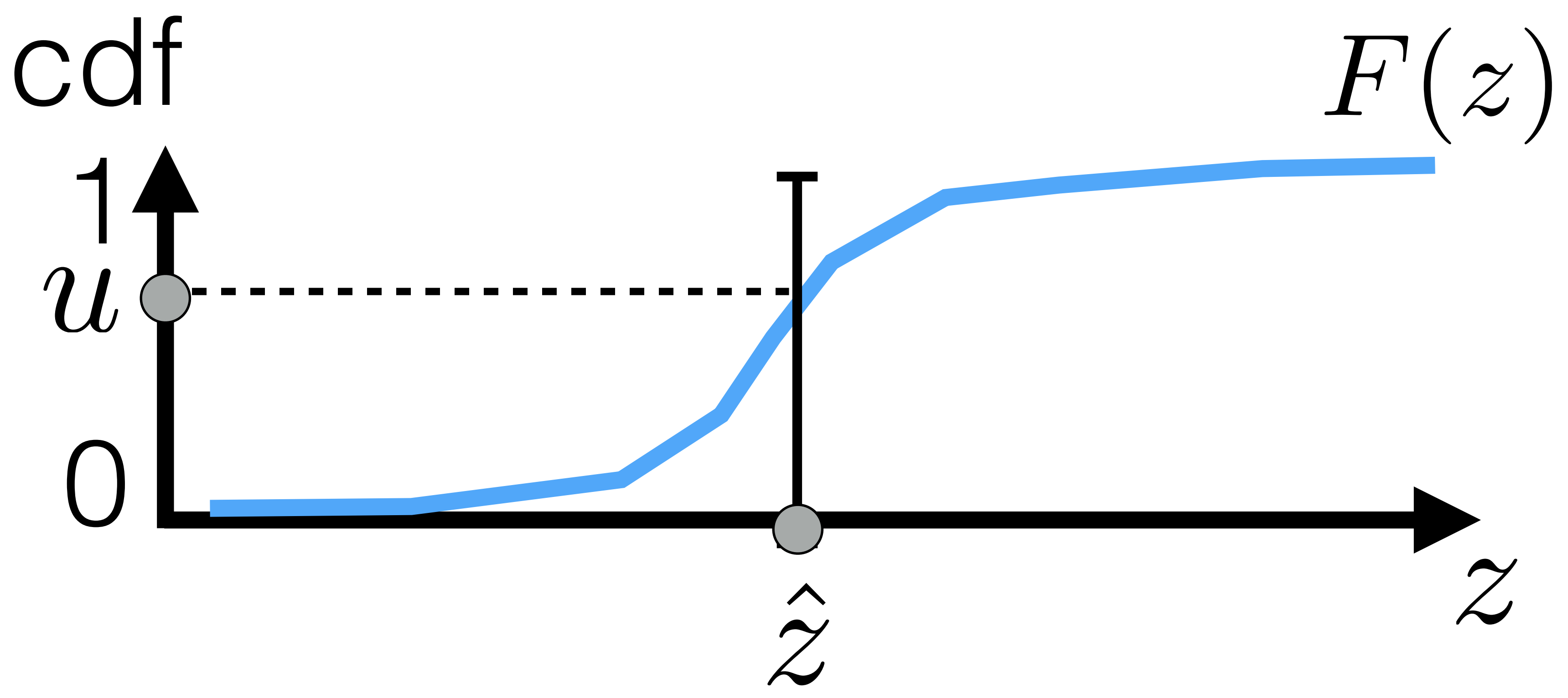
$$p(z) = \frac{1}{C} \hat{p}(z), \quad C = \int \hat{p}(z) dz$$

$$z^k \sim p(z) - ?$$

Inverse CDF-based sampling (1-dim r. v.)

Consider r. v. z with invertible cdf $F(z)$

1. Sample $u \sim U[0, 1]$
   $\hat{z} = F^{-1}(u)$

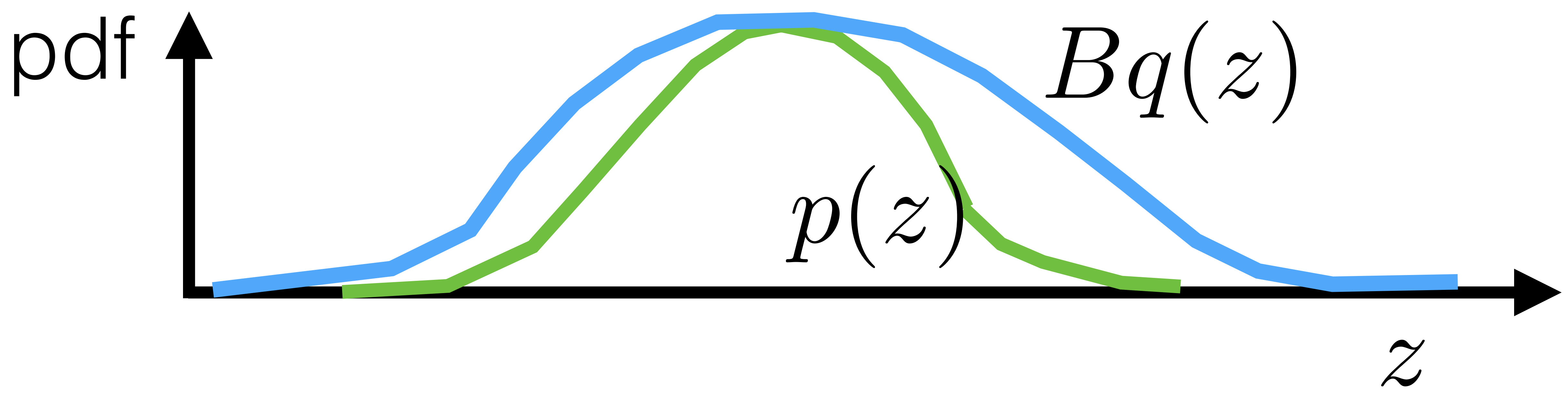


- Strict limitations of distribution
- Unnormalized pdf not supported

Rejection sampling

$$p(z) \leqslant Bq(z) \quad \forall z$$

and we know how
to sample from $q(z)$



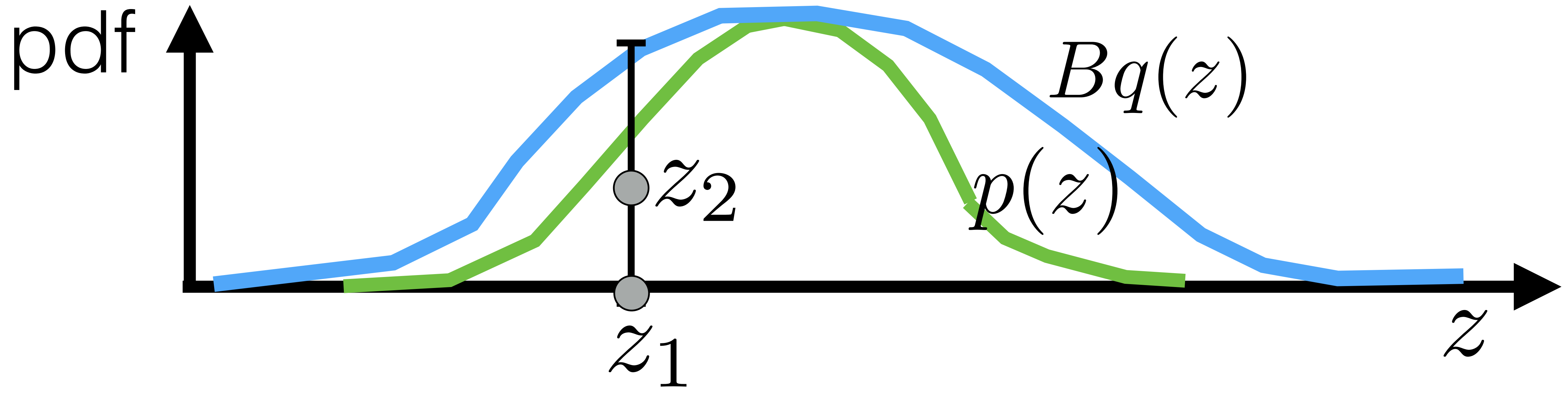
Rejection sampling

$p(z) \leqslant Bq(z) \quad \forall z$

and we know how
to sample from $q(z)$

1. Sample $z_1 \sim q(z)$
2. Sample $z_2 \sim U[0, Bq(z_1)]$
3. If $z_2 \leqslant p(z_1)$ save sample z_2

   Else skip sample



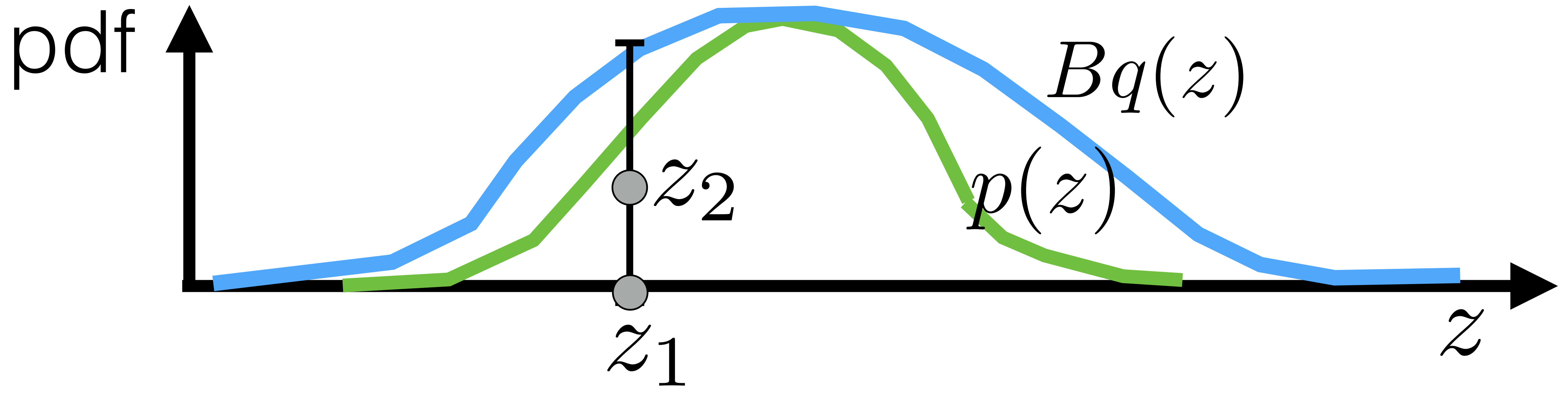
Rejection sampling

$p(z) \leqslant Bq(z) \quad \forall z$

and we know how
to sample from $q(z)$

1. Sample $z_1 \sim q(z)$
2. Sample $z_2 \sim U[0, Bq(z_1)]$
3. If $z_2 \leqslant p(z_1)$ save sample z_2

   Else skip sample



- If q and p are not similar to each other, we have a high rejection rate
- Unnormalized p is not supported

Importance sampling

Consider an arbitrary distribution $q(z)$
for which we know how to sample from

$$\mathbb{E}_{p(z)} f(z) = \int p(z) f(z) dz = \int q(z) \frac{p(z)}{q(z)} f(z) dz \approx \sum_{k=1}^{K} \underbrace{\frac{p(z^k)}{q(z^k)}}_{\text{weight}} f(z^k)$$

$$z^k \sim q(z)$$

- No rejects +
- Unnormalized p supported +
- A lot of “ineffective” samples with low weights —

Importance sampling

Consider an arbitrary distribution $q(z)$
for which we know how to sample from

$$\mathbb{E}_{p(z)} f(z) = \int p(z) f(z) dz = \int q(z) \frac{p(z)}{q(z)} f(z) dz \approx \sum_{k=1}^{K} \frac{p(z^k)}{q(z^k)} f(z^k)$$

$$z^k \sim q(z)$$

- **Unnormalized p supported**

$$p(z) = \frac{\hat{p}(z)}{C}, \quad C = \int \hat{p}(z) dz =$$

Importance sampling

Consider an arbitrary distribution $q(z)$
for which we know how to sample from

$$\mathbb{E}_{p(z)} f(z) = \int p(z) f(z) dz = \int q(z) \frac{p(z)}{q(z)} f(z) dz \approx \sum_{k=1}^{K} \frac{p(z^k)}{q(z^k)} f(z^k)$$

$$z^k \sim q(z)$$

- **Unnormalized p supported**

$$p(z) = \frac{\hat{p}(z)}{C}, \quad C = \int \hat{p}(z) dz = \int q(z) \frac{\hat{p}(z)}{q(z)} dz \approx \sum_{k=1} \frac{\hat{p}(z^k)}{q(z^k)}$$

Importance sampling

Consider an arbitrary distribution $q(z)$
for which we know how to sample from

$$\mathbb{E}_{p(z)} f(z) = \int p(z) f(z) dz = \int q(z) \frac{p(z)}{q(z)} f(z) dz \approx \sum_{k=1}^{K} \frac{p(z^k)}{q(z^k)} f(z^k)$$

$$z^k \sim q(z)$$

- **Unnormalized p supported**

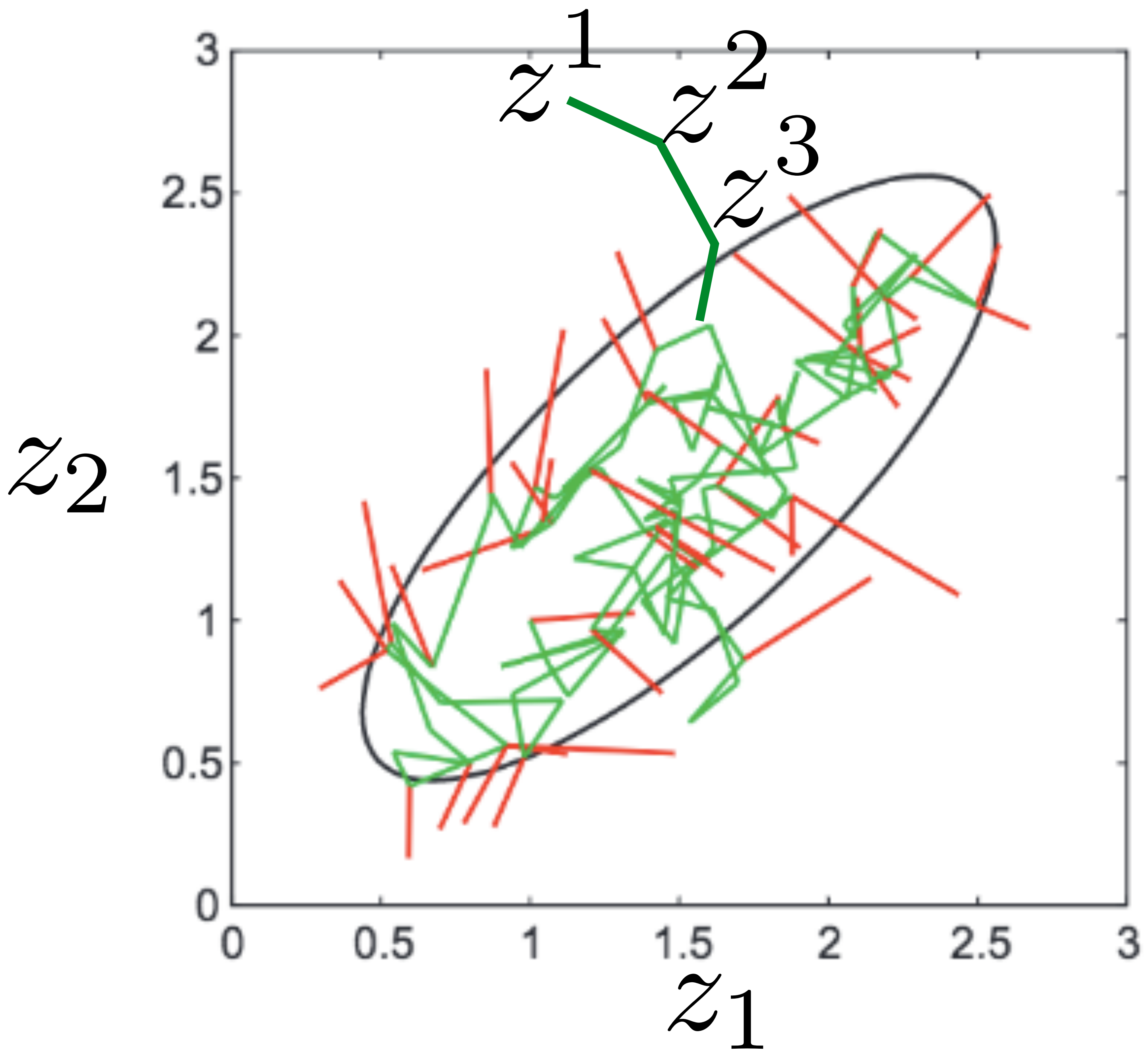
$$p(z) = \frac{\hat{p}(z)}{C}, \quad C = \int \hat{p}(z) dz = \int q(z) \frac{\hat{p}(z)}{q(z)} dz \approx \sum_{k=1} \frac{\hat{p}(z^k)}{q(z^k)}$$

$$\mathbb{E}_{p(z)} f(z) = \int \frac{1}{C} \hat{p}(z) f(z) = \frac{\int \hat{p}(z) f(z) dz}{\int \hat{p}(z) dz} = \frac{\int q(z) \frac{\hat{p}(z) f(z)}{q(z)} dz}{\int q(z) \frac{\hat{p}(z)}{q(z)} dz}$$

Markov Chain Monte Carlo

$$\mathbb{E}_{p(z)} f(z) \approx \frac{1}{K} \sum_{k=1}^{K} f(x^k), \quad z^k \sim p(z)$$

$$p(z) = \frac{1}{C}\hat{p}(z), \quad C = \int \hat{p}(z) dz \quad z^k \sim p(z) - ?$$



Consider we are given a Markov chain with some transition distribution $q(\cdot|\cdot)$:

$$z^0 \sim p_0(z^0)$$

$$z^1 \sim q(z|z^0)$$

$$z^2 \sim q(z|z^1) \ldots$$

Samples $z^0, z^1, z^2, z^3 \ldots$ are not independent but still can be used to estimate expectation

($p_0(\cdot)$ is an arbitrary initial distribution)

Slide From Dmitry Kropotov's presentation, deepbayes.ru/2019

Markov Chain Monte Carlo

$$\mathbb{E}_{p(z)} f(z) \approx \frac{1}{K} \sum_{k=1}^{K} f(x^k), \quad z^k \sim p(z)$$

$$p(z) = \frac{1}{C} \hat{p}(z), \quad C = \int \hat{p}(z) dz \qquad z^k \sim p(z) - ?$$

Consider we are given a Markov chain with some transition probability $q(\cdot|\cdot)$:

$$z^0 \sim p_0(z^0)$$
$$z^1 \sim q(z|z^0)$$
$$z^2 \sim q(z|z^1) \dots$$

How to choose the transition distribution so that the resulting samples come from p(z)?

Markov chain properties

Markov chain with transition distribution $q(\cdot|\cdot)$:

$$z^0 \sim p_0(z^0)$$

$$z^1 \sim q(z|z^0)$$

$$z^2 \sim q(z|z^1) \dots$$

$$p(z_0, z_1, \dots, z_n) = p_0(z_0) \prod_{i=1}^{n} q(z_i|z_{i-1})$$

$$p_1(z) = \int q(z|z_0) p_0(z_0) dz_0$$

$$p_2(z) = \int q(z|z_1) p_1(z_1) dz_1$$

Our goal: $p_n(z) \to p(z),\ n \to \infty$

Based on Dmitry Vetrov's lectures

Markov chain properties

Markov chain with transition distribution $q(\cdot|\cdot)$:

$$z^0 \sim p_0(z^0)$$
$$z^1 \sim q(z|z^0)$$
$$z^2 \sim q(z|z^1) \ldots$$

$$p(z_0, z_1, \ldots, z_n) = p_0(z_0) \prod_{i=1}^{n} q(z_i|z_{i-1})$$

$$p_1(z) = \int q(z|z_0) p_0(z_0) dz_0$$

$$\boxed{p_2(z)} = \int q(z|z_1) \boxed{p_1(z_1)} dz_1$$

Our goal: $p_n(z) \to p(z),\ n \to \infty$

Based on Dmitry Vetrov's lectures

Markov chain properties

Markov chain with transition distribution $q(\cdot|\cdot)$:

$$z^0 \sim p_0(z^0)$$
$$z^1 \sim q(z|z^0)$$
$$z^2 \sim q(z|z^1)\dots$$

Probability distribution $\pi(z)$ is called **invariant** under given Markov chain iff

$$\int q(z|y)\pi(y)dy = \pi(z)$$

<u>Sufficient condition for invariance:</u> $\pi(z)q(y|z) = \pi(y)q(z|y)$

Based on Dmitry Vetrov's lectures

Markov chain properties

Markov chain with transition distribution $q(\cdot|\cdot)$:

$$z^0 \sim p_0(z^0)$$
$$z^1 \sim q(z|z^0)$$
$$z^2 \sim q(z|z^1) \dots$$

Probability distribution $\pi(z)$ is called **invariant** under given Markov chain iff

$$\int q(z|y)\pi(y)dy = \pi(z)$$

<u>Sufficient condition for invariance:</u> $\pi(z)q(y|z) = \pi(y)q(z|y)$

Proof: $\int q(z|y)\pi(y)dy = \int q(y|z)\pi(z)dy = \pi(z)$

Based on Dmitry Vetrov's lectures

Markov chain properties

Markov chain with transition distribution $q(\cdot|\cdot)$:

$$z^0 \sim p_0(z^0)$$
$$z^1 \sim q(z|z^0)$$
$$z^2 \sim q(z|z^1) \dots$$

Probability distribution $\pi(z)$ is called **invariant** under given Markov chain iff

$$\int q(z|y)\pi(y)dy = \pi(z)$$

<u>Sufficient condition for invariance:</u> $\pi(z)q(y|z) = \pi(y)q(z|y)$

Markov chain may have more than one invariant distribution

Markov chain properties

Markov chain with transition distribution $q(\cdot|\cdot)$:

$$z^0 \sim p_0(z^0)$$
$$z^1 \sim q(z|z^0)$$
$$z^2 \sim q(z|z^1)\dots$$

Consider some initial distribution $p_0(z)$ and denote $p_n(z)$ — distribution of points after n steps of Markov chain. Then the Markov chain is called **ergodic** iff

$$p_n(z) \to \pi(z), \quad n \to \infty, \quad \forall p_0(z)$$

where $\pi(z)$ is invariant under given Markov chain.

Slide From Dmitry Kropotov's presentation, deepbayes.ru/2019

Markov chain properties

Markov chain with transition distribution $q(\cdot|\cdot)$:

$$z^0 \sim p_0(z^0)$$
$$z^1 \sim q(z|z^0)$$
$$z^2 \sim q(z|z^1)\dots$$

Consider some initial distribution $p_0(z)$ and denote $p_n(z)$ — distribution of points after n steps of Markov chain. Then the Markov chain is called **ergodic** iff

$$p_n(z) \to \pi(z), \quad n \to \infty, \quad \forall p_0(z)$$

where $\pi(z)$ is invariant under given Markov chain.

Sufficient condition for ergodic chain: $\forall z, y \quad \pi(z) \neq 0 \;\Rightarrow\; q(z|y) > 0$

Slide From Dmitry Kropotov's presentation, deepbayes.ru/2019

Markov chain properties

Markov chain with transition distribution $q(\cdot|\cdot)$:

$$z^0 \sim p_0(z^0)$$
$$z^1 \sim q(z|z^0)$$
$$z^2 \sim q(z|z^1) \dots$$

Probability distribution $\pi(z)$ is called **invariant** under given Markov chain iff

$$\int q(z|y)\pi(y)dy = \pi(z)$$

Consider some initial distribution $p_0(z)$ and denote $p_i(z)$ — distribution of points after i steps of Markov chain. Then the Markov chain is called **ergodic** iff

$$p_i(z) \to \pi(z), \quad i \to \infty, \quad \forall p_0(z)$$

where $\pi(z)$ is invariant under given Markov chain.

Markov chain properties

Markov chain with transition distribution $q(\cdot|\cdot)$:

$$z^0 \sim p_0(z^0)$$
$$z^1 \sim q(z|z^0)$$
$$z^2 \sim q(z|z^1) \dots$$

For generating samples from $p(z)$ using MCMC, the corresponding Markov chain should be **ergodic** with **invariant** distribution $p(z)$

Probability distribution $\pi(z)$ is called **invariant** under given Markov chain iff

$$\int q(z|y)\pi(y)dy = \pi(z)$$

Consider some initial distribution $p_0(z)$ and denote $p_i(z)$ — distribution of points after i steps of Markov chain. Then the Markov chain is called **ergodic** iff

$$p_i(z) \to \pi(z), \quad i \to \infty, \quad \forall p_0(z)$$

where $\pi(z)$ is invariant under given Markov chain.

Slide From Dmitry Kropotov’s presentation, deepbayes.ru/2019

Markov Chain Monte Carlo

Markov chain with transition distribution $q(\cdot|\cdot)$:

$$z^0 \sim p_0(z^0)$$
$$z^1 \sim q(z|z^0)$$
$$z^2 \sim q(z|z^1)\dots$$

Markov Chain Monte Carlo sampling from $p(z)$:

1. Select transition distribution $q(\cdot|\cdot)$ and starting distribution $p_0(\cdot)$
2. Is **Markov chain invariant** under $p(z)$?
3. Is **Markov chain ergodic** with invariant distribution $p(z)$?
4. If both conditions are satisfied, starting from some n, we can use $z^n, z^{n+1}, \dots$ as samples from $p(z)$

Metropolis-Hastings sampling

Algorithm of obtaining a new point while sampling from $p(z)$:

Input: previous sample z^n

Output: new sample z^{n+1}

Using proposal distribution $r(\cdot|z_n)$

1. Sample trial point $y \sim r(y|z^n)$
2. Calculate acceptance probability $A = \min\left(1, \frac{p(y)r(z^n|y)}{p(z^n)r(y|z^n)}\right)$
3. $z^{n+1} = y$ with probability A and z^n with probability $1 - A$

Transition distribution? Invariance? Ergodic chain?

Metropolis-Hastings sampling

Transition distribution:

$$q(z^{n+1}|z^n) = \begin{cases} A \cdot r(z^{n+1}|z^n),\ z^{n+1} \neq z^n \\ 1 - A \cdot r(z^{n+1}|z^n), z^{n+1} = z^n \end{cases}$$

Metropolis-Hastings sampling

Transition distribution:

$$q(z^{n+1}|z^n) = \begin{cases} A \cdot r(z^{n+1}|z^n),\ z^{n+1} \neq z^n \\ 1 - A \cdot r(z^{n+1}|z^n), z^{n+1} = z^n \end{cases}$$

Invariance? (using sufficient condition)

$$q(z^{n+1}|z^n)p(z^n) = q(z^n|z^{n+1})p(z^{n+1}) -?$$

=

Metropolis-Hastings sampling

Transition distribution:

$$q(z^{n+1}|z^n) = \begin{cases} A \cdot r(z^{n+1}|z^n), \; z^{n+1} \neq z^n \\ 1 - A \cdot r(z^{n+1}|z^n), z^{n+1} = z^n \end{cases}$$

Invariance? (using sufficient condition)

$$q(z^{n+1}|z^n)p(z^n) = q(z^n|z^{n+1})p(z^{n+1}) - ?$$

$$q(z^{n+1}|z^n)p(z^n) = A \cdot r(z^{n+1}|z^n)p(z^n) = \min\left(1, \frac{p(z^{n+1})r(z^n|z^{n+1})}{\underline{p(z^n)r(z^{n+1}|z^n)}}\right)\underline{r(z^{n+1}|z^n)p(z^n)} =$$

$$= \min\left(\underline{p(z^n)r(z^{n+1}|z^n)}, \; p(z^{n+1})r(z^n|z^{n+1})\right) \underset{\text{vice versa}}{=} q(z^n|z^{n+1})p(z^{n+1})$$

Metropolis-Hastings sampling

Transition distribution:

$$q(z^{n+1}|z^n) = \begin{cases} A \cdot r(z^{n+1}|z^n),\ z^{n+1} \neq z^n \\ 1 - A \cdot r(z^{n+1}|z^n), z^{n+1} = z^n \end{cases}$$

Invariance? (yes, according to sufficient condition)

$$q(z^{n+1}|z^n)p(z^n) = q(z^n|z^{n+1})p(z^{n+1})$$

Ergodic chain? (yes, if sufficient condition is satisfied)

Metropolis-Hastings sampling

Algorithm of obtaining a new point while sampling from $p(z)$:

Input: previous sample z^n

Output: new sample z^{n+1}

Using proposal distribution $r(\cdot|z_n)$

1. Sample trial point $y \sim r(y|z^n)$
2. Calculate acceptance probability $A = \min\left(1, \frac{\boxed{p(y)}r(z^n|y)}{\boxed{p(z^n)}r(y|z^n)}\right)$
3. $z^{n+1} = y$ with probability A and z^n with probability $1 - A$

Unnormalized p(z)?

Metropolis-Hastings sampling

Algorithm of obtaining a new point while sampling from $p(z)$:

Input: previous sample z^n

Using proposal distribution $r(\cdot|z_n)$

Output: new sample z^{n+1}

1. Sample trial point $y \sim r(y|z^n)$
2. Calculate acceptance probability $A = \min\left(1, \frac{\hat{p}(y)r(z^n|y)}{\hat{p}(z^n)r(y|z^n)}\right)$
3. $z^{n+1} = y$ with probability A and z^n with probability $1 - A$

Unnormalized p(z)? $p(z) = \frac{1}{C}\hat{p}(z)$

Metropolis sampling

Algorithm of obtaining a new point while sampling from $p(z)$:

Input: previous sample z^n

Output: new sample z^{n+1}

Using proposal distribution $r(\cdot|z_n)$

1. Sample trial point $y \sim r(y|z^n)$
2. Calculate acceptance probability $A = \min\left(1, \frac{\hat{p}(y)r(z^n|y)}{\hat{p}(z^n)r(y|z^n)}\right)$
3. $z^{n+1} = y$ with probability A and z^n with probability $1 - A$

Symmetric proposal? $r(z^n|y) = r(y|z^n) \;\; \forall z^n, y$

Metropolis sampling

Algorithm of obtaining a new point while sampling from $p(z)$:

Input: previous sample z^n

Output: new sample z^{n+1}

Using proposal distribution $r(\cdot|z_n)$

1. Sample trial point $y \sim r(y|z^n)$
2. Calculate acceptance probability $A = \min\left(1, \frac{\hat{p}(y)\cancel{r(z^n|y)}}{\hat{p}(z^n)\cancel{r(y|z^n)}}\right)$
3. $z^{n+1} = y$ with probability A and z^n with probability $1 - A$

Symmetric proposal?

Always accept the trial point if it increases pdf:

$\hat{p}(y) > \hat{p}(z^n)$

Simple example

Metropolis-Hastings sampling

1. Sample trial point $y \sim r(y|z^n)$
2. Calculate acceptance probability $A = \min\left(1, \frac{p(y)r(z^n|y)}{p(z^n)r(y|z^n)}\right)$
3. $z^{n+1} = y$ with probability A and z^n with probability $1 - A$

$$p(z) = \mathcal{N}(z|\mu, \Sigma), \; z \in \mathbb{R}^d$$

$$r(y|z^n) = \mathcal{N}(y|z^n, \sigma^2 I)$$

Tasks:

- Is proposal distribution symmetric?
- Write down the algorithm for MH sampling in the described setting

Simple example

- Is proposal distribution symmetric?

Yes: $\mathcal{N}(y|z^n, \sigma^2 I) = \mathcal{N}(z^n|y, \sigma^2 I)$

$$y = z^n + \epsilon \quad \Leftrightarrow \quad z^n = y + \epsilon, \quad \epsilon \sim \mathcal{N}(\epsilon|0, \sigma^2 I)$$

Simple example

- <u>Is proposal distribution symmetric?</u>

Yes: $\mathcal{N}(y|z^n, \sigma^2 I) = \mathcal{N}(z^n|y, \sigma^2 I)$

$$y = z^n + \epsilon \quad \Leftrightarrow \quad z^n = y + \epsilon, \quad \epsilon \sim \mathcal{N}(\epsilon|0, \sigma^2 I)$$

- <u>MH sampling algorithm:</u>

1. Sample trial point $y \sim \mathcal{N}(y|z^n, \sigma^2 I)$
2. Calculate $A = \min\left(1, \dfrac{\mathcal{N}(y|\mu, \Sigma)}{\mathcal{N}(z^n|\mu, \Sigma)}\right)$
3. $z^{n+1} = y$ with probability A and z^n with probability $1 - A$

Simple example

$$p(z) = \mathcal{N}(z|\mu, \Sigma), \;\; z \in \mathbb{R}^d$$

$$r(y|z^n) = \mathcal{N}(y|z^n, \sigma^2 I)$$

red lines: rejected trials
green lines: MH route

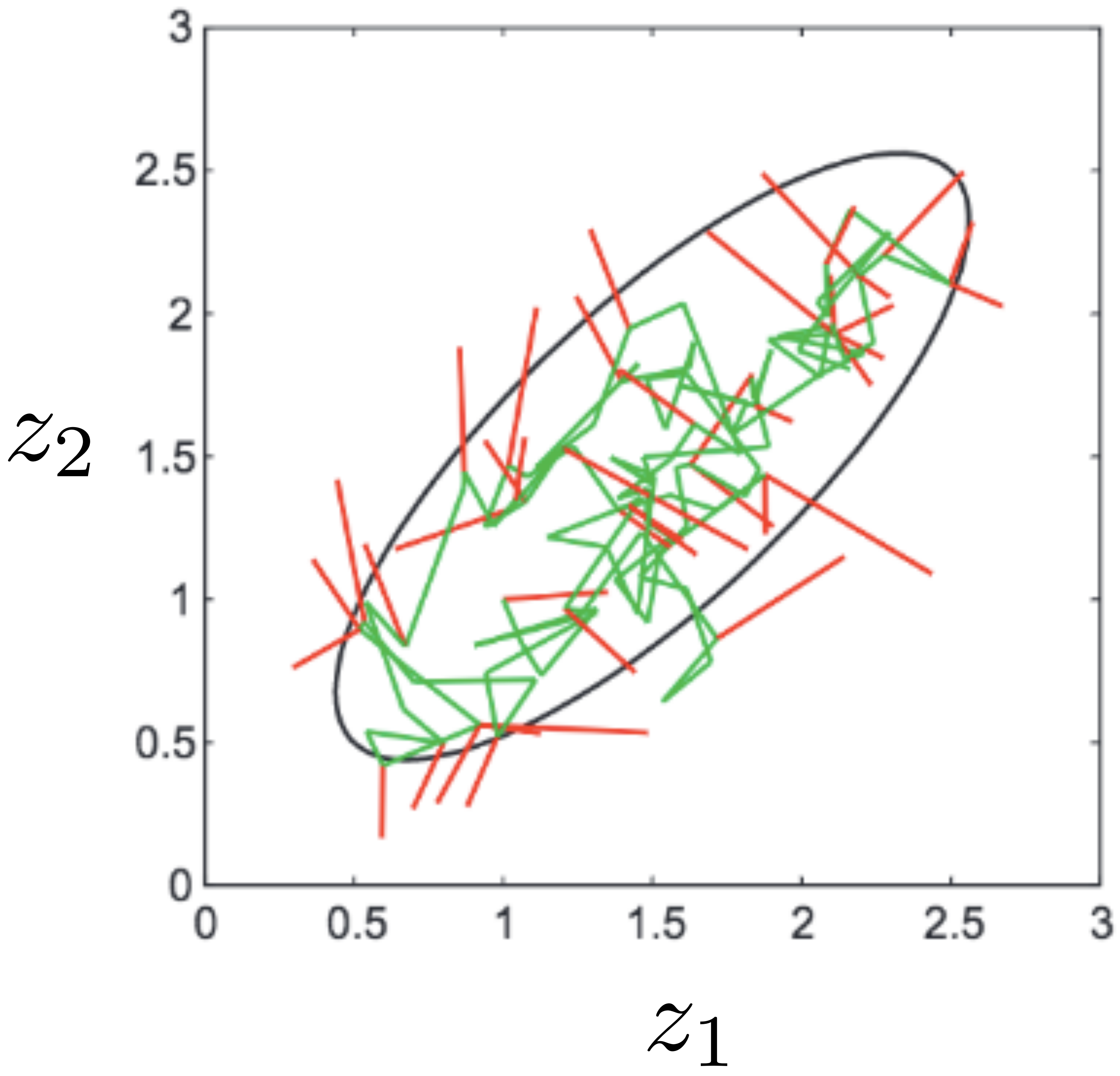


Image From Dmitry Kropotov's presentation, deepbayes.ru/2019

Langevin Monte-Carlo

Consider neural network with weights w, likelihood $p(y|x, w)$ and prior $p(w)$

Stochastic Gradient Langevin Dynamics updates the weights in the following way:

$$w_{n+1} = w_n + \frac{\epsilon}{2}\left(\nabla \log p(w_n) + \frac{N}{n}\sum_{i=1}^{n} \nabla \log p(y_{j_i}|x_{j_i}, w_n)\right) + \eta_n,$$

$$\eta_n \sim \mathcal{N}(\eta|0, \epsilon I),$$

$$j_i \sim Unif(1, \ldots, N)$$

(mini-batch objects)

<u>Two differences from SGD:</u>

- noise on gradients: η_n
- skipping samples according to Metropolis-Hastings test

<u>Pros:</u>

- generates weight samples around Maximum posterior (MAP) weights
- able to move from one MAP to another

Langevin Monte-Carlo

Consider neural network with weights w, likelihood $p(y|x,w)$ and prior $p(w)$

Stochastic Gradient Langevin Dynamics updates the weights in the following way:

$$w_{n+1} = w_n + \frac{\epsilon}{2}\left(\nabla \log p(w_n) + \frac{N}{n}\sum_{i=1}^{n} \nabla \log p(y_{j_i}|x_{j_i}, w_n)\right) + \eta_n,$$

$$\eta_n \sim \mathcal{N}(\eta|0, \epsilon I),$$

$$j_i \sim Unif(1, \dots, N)$$

(mini-batch objects)

Two differences from SGD:

- noise on gradients: η_n
- skipping samples according to Metropolis-Hastings test

Questions:

- Is proposal distribution symmetric?
- Write down the corresponding MH algorithm

Langevin Monte-Carlo

$$r(y|w^n) = \mathcal{N}(y|\mu^n, \epsilon I)$$

not symmetric (gradient in different point)

$$\mu^n = w_n + \frac{\epsilon}{2}\Big(\nabla \log p(w_n) + \frac{N}{n}\sum_{i=1}^{n} \nabla \log p(y_{j_i}|x_{j_i}, w_n)\Big)$$

<u>MH sampling algorithm:</u>

1. Sample trial point $\mathcal{N}(y|\mu^n, \epsilon I)$
2. Calculate $A = \min\left(1, \frac{p(y)r(w^n|y)}{p(w^n)r(y|w^n)}\right)$

$p(w_n)\prod_{i=1}^{n} p(y_{j_i}|x_{j_i}, w_n)$

$\mathcal{N}(y|\mu^n, \epsilon I)$

3. $w^{n+1} = y$ with probability A and w^n with probability $1 - A$

Prediction in Bayesian models

$$p(y_*|x_*, X, Y) = \mathbb{E}_{p(\theta|X,Y)} p(y_*|x_*, \theta) \approx \frac{1}{K}\sum_{k=1}^{K} p(y_*|x_*, \theta^k)$$

$$\theta^k \sim p(\theta|X, Y)$$

We only need to know how to sample from from the **unnormalized** posterior!

$$p(\theta|X,Y) = \frac{\boxed{p(Y|X,\theta)p(\theta)}}{\boxed{\int p(Y|X,\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}}}$$

unnormalized posterior

intractable

Summary

- To use Monte-Carlo sampling in Bayesian machine learning, we need techniques for sampling from unnormalized distribution.
  Examples: importance sampling, Metropolis-Hastings sampling

- For generating samples from p(z) using Markov Chain Monte Carlo (MCMC), the corresponding Markov chain should be ergodic with invariant distribution p(z)

- MCMC sampling is able to cover the whole posterior distribution, but is highly time-consuming