# Stein variational gradient descent

Thomas Bonis

# Optimal transport distance

Wasserstein distance between two measures $\mu$ and $\nu$

$$W_2(\mu, \nu)^2 = \inf_\pi \int \|y - x\|^2 d\pi(x, y),$$

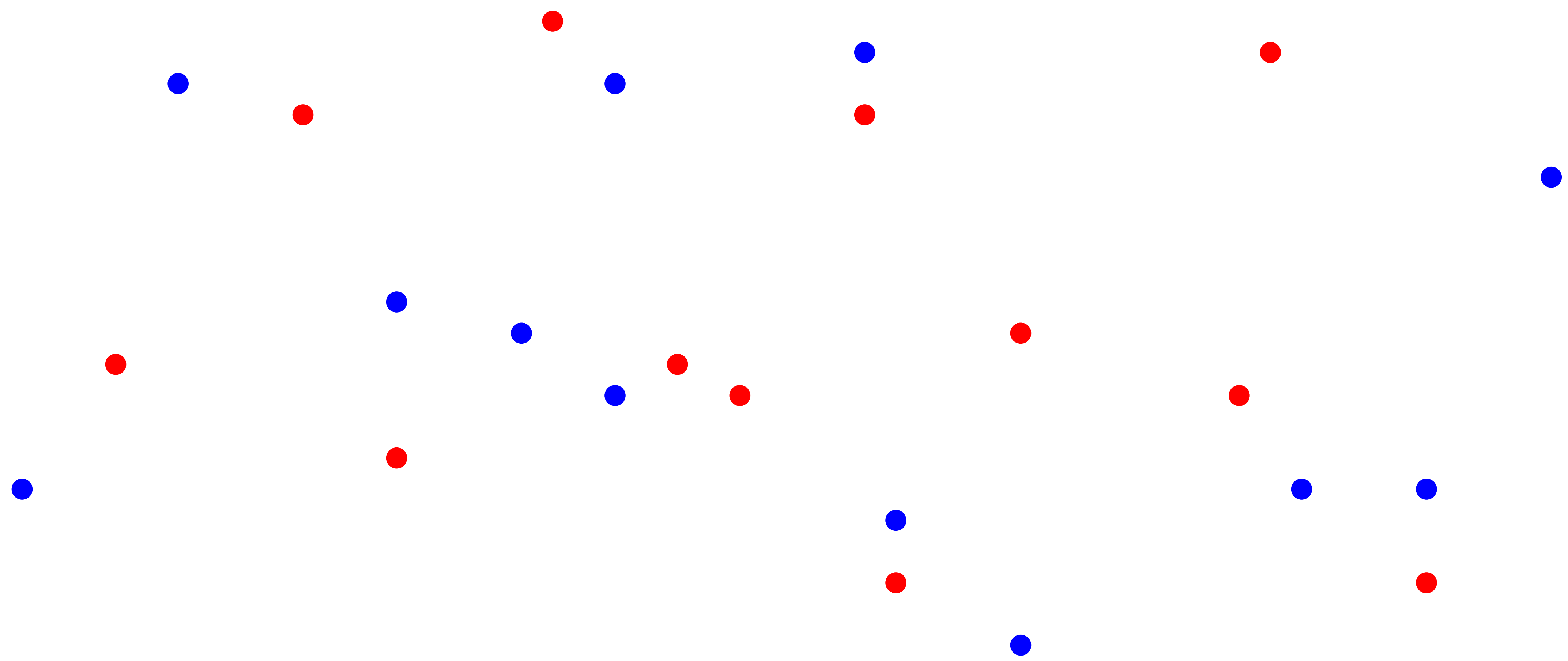
where the infimum is taken over all measures with marginals $\mu$ and $\nu$.

# Optimal transport distance

Wasserstein distance between two measures $\mu$ and $\nu$

$$W_2(\mu,\nu)^2 = \inf_\pi \int \|y-x\|^2 d\pi(x,y),$$

where the infimum is taken over all measures with marginals $\mu$ and $\nu$.

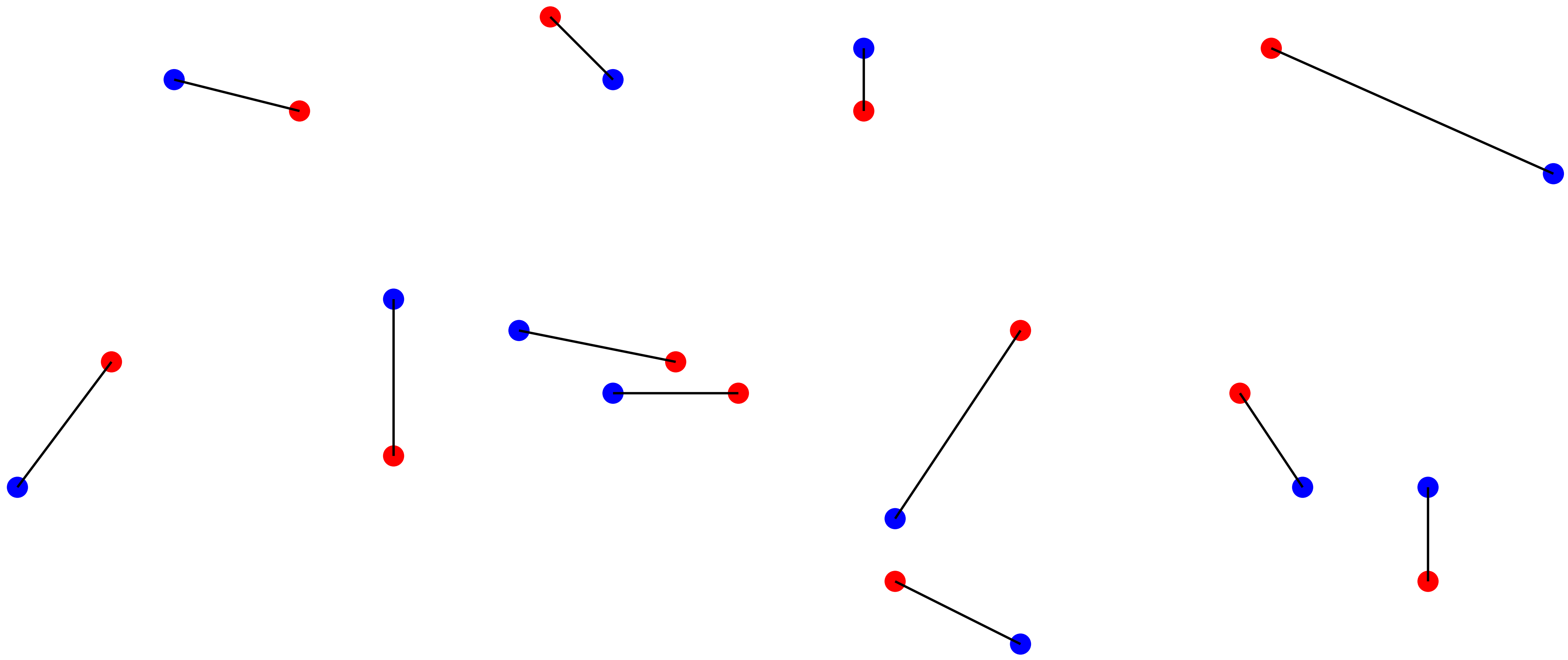


# Optimal transport distance

Wasserstein distance between two measures $\mu$ and $\nu$

$$W_2(\mu, \nu)^2 = \inf_\pi \int \|y - x\|^2 d\pi(x, y),$$

where the infimum is taken over all measures with marginals $\mu$ and $\nu$.

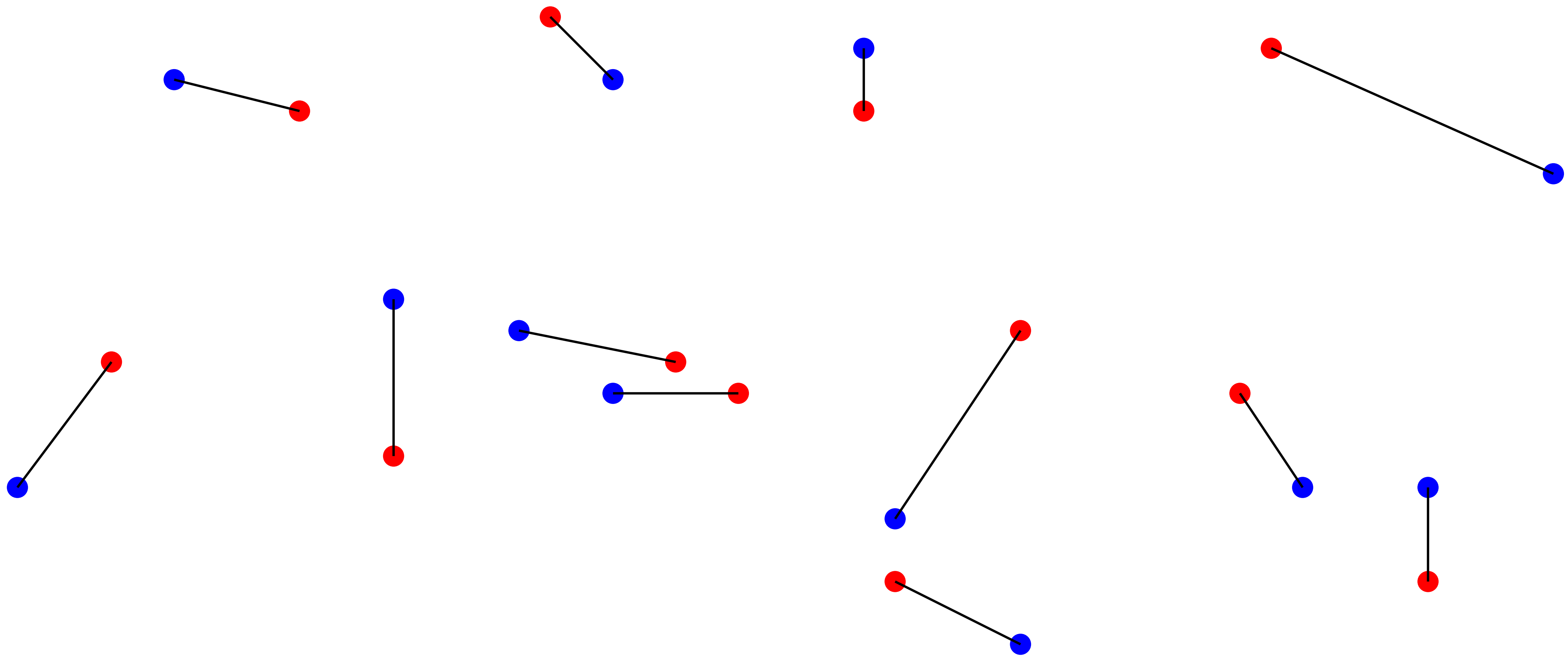


# Optimal transport distance

Wasserstein distance between two measures $\mu$ and $\nu$

$$W_2(\mu,\nu)^2 = \inf_\pi \int \|y-x\|^2 d\pi(x,y),$$

where the infimum is taken over all measures with marginals $\mu$ and $\nu$.



$W_2$ = mean length of matching.

# Optimal transport distance

Wasserstein distance between two measures $\mu$ and $\nu$

$$W_2(\mu, \nu)^2 = \inf_\pi \int \|y - x\|^2 d\pi(x, y),$$

where the infimum is taken over all measures with marginals $\mu$ and $\nu$.

Pros:

Well-suited for multivariate analysis

Can deal with discrete measures

Con: hard to compute

# Optimal transport distance

Wasserstein distance between two measures $\mu$ and $\nu$

$$W_2(\mu,\nu)^2 = \inf_\pi \int \|y-x\|^2 d\pi(x,y),$$

where the infimum is taken over all measures with marginals $\mu$ and $\nu$.

Pros:

Well-suited for multivariate analysis

Can deal with discrete measures

Good for statistical applications

Con: hard to compute

# Optimal transport distance

Wasserstein distance between two measures $\mu$ and $\nu$

$$W_2(\mu,\nu)^2 = \inf_\pi \int \|y-x\|^2 d\pi(x,y),$$

where the infimum is taken over all measures with marginals $\mu$ and $\nu$.

Pros:

Well-suited for multivariate analysis

Can deal with discrete measures

Good for statistical applications

Con: hard to compute

Need a more practical proxy

# Optimal transport distance

Wasserstein distance between two measures $\mu$ and $\nu$

$$W_2(\mu,\nu)^2 = \inf_\pi \int \|y-x\|^2 d\pi(x,y),$$

where the infimum is taken over all measures with marginals $\mu$ and $\nu$.

Pros:

Well-suited for multivariate analysis

Can deal with discrete measures

Good for statistical applications

Con: hard to compute

Need a more practical proxy

**Not** an integral probability metric

# Bayesian inference

$\mu$ is a measure on $\mathbb{R}^d$ such that

- $d >> 1$
- $d\mu(x) = \frac{e^{-V(x)}}{Z} dx$ with unknown $Z$

# Bayesian inference

$\mu$ is a measure on $\mathbb{R}^d$ such that

- $d >> 1$
- $d\mu(x) = \frac{e^{-V(x)}}{Z} dx$ with unknown $Z$

Objective: provide estimates for $\int f d\mu$, for various functions $f$.

# Bayesian inference

$\mu$ is a measure on $\mathbb{R}^d$ such that

- $d >> 1$
- $d\mu(x) = \frac{e^{-V(x)}}{Z} dx$ with unknown $Z$

Objective: provide a function $\nu$ close to $\mu$ s.t. $W_2(\nu, \mu)$ is small.

# Bayesian inference

$\mu$ is a measure on $\mathbb{R}^d$ such that

- $d >> 1$
- $d\mu(x) = \frac{e^{-V(x)}}{Z} dx$ with unknown $Z$

Objective: provide a function $\nu$ close to $\mu$ s.t. $W_2(\nu, \mu)$ is small.

Option 1: Langevin Monte-Carlo

Sample $n$ points from $X^h_{K+1}$ where $X^h_0 = 0$ and

$$X^h_{k+1} = -h \nabla V(X^h_k) + \sqrt{2h} Z, Z \sim \mathcal{N}(0,1)$$

# Bayesian inference

$\mu$ is a measure on $\mathbb{R}^d$ such that

- $d >> 1$
- $d\mu(x) = \frac{e^{-V(x)}}{Z} dx$ with unknown $Z$

Objective: provide a function $\nu$ close to $\mu$ s.t. $W_2(\nu, \mu)$ is small.

Option 1: Langevin Monte-Carlo

Sample $n$ points from $X^h_{K+1}$ where $X^h_0 = 0$ and

$$X^h_{k+1} = -h\nabla V(X^h_k) + \sqrt{2h}Z, Z \sim \mathcal{N}(0,1)$$

Can achieve any given acuracy with correct $h, K, n$ but slow in practice

# Bayesian inference

$\mu$ is a measure on $\mathbb{R}^d$ such that

- $d >> 1$
- $d\mu(x) = \frac{e^{-V(x)}}{Z}dx$ with unknown $Z$

Objective: provide a function $\nu$ close to $\mu$ s.t. $W_2(\nu, \mu)$ is small.

Option 1: Langevin Monte-Carlo

Sample $n$ points from $X^h_{K+1}$ where $X^h_0 = 0$ and

$$X^h_{k+1} = -h\nabla V(X^h_k) + \sqrt{2h}Z, Z \sim \mathcal{N}(0,1)$$

Can achieve any given acuracy with correct $h, K, n$ but slow in practice

Option 2: Variational inference

Find $\nu$ in a simple family of measure minimizing $KL(\nu|\mu)$. (Simple optimization problem)

# Bayesian inference

$\mu$ is a measure on $\mathbb{R}^d$ such that

- $d >> 1$
- $d\mu(x) = \frac{e^{-V(x)}}{Z} dx$ with unknown $Z$

Objective: provide a function $\nu$ close to $\mu$ s.t. $W_2(\nu, \mu)$ is small.

Option 1: Langevin Monte-Carlo

Sample $n$ points from $X_{K+1}^h$ where $X_0^h = 0$ and

$$X_{k+1}^h = -h\nabla V(X_k^h) + \sqrt{2h}Z, Z \sim \mathcal{N}(0,1)$$

Can achieve any given acuracy with correct $h, K, n$ but slow in practice

Option 2: Variational inference

Find $\nu$ in a simple family of measure minimizing $KL(\nu|\mu)$. (Simple optimization problem)

Fast but limited accuracy

# On Langevin Monte-Carlo

Why does Langevin Monte-Carlo work?

$\mu$ stationary measure of the diffusion process solution of

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t.$$

# On Langevin Monte-Carlo

Why does Langevin Monte-Carlo work?

$\mu$ stationary measure of the diffusion process solution of

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t.$$

Why is it slow?

All $n$ particles evolve independently.

# On Langevin Monte-Carlo

Why does Langevin Monte-Carlo work?

$\mu$ stationary measure of the diffusion process solution of

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t.$$

Why is it slow?

All $n$ particles evolve independently.

What would it look like if all particles moved together?

If $X_0 \sim \nu$, $X_t \sim \nu_t = h_t d\mu$ evolves given the SDE above.

# On Langevin Monte-Carlo

Why does Langevin Monte-Carlo work?

$\mu$ stationary measure of the diffusion process solution of

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t.$$

Why is it slow?

All $n$ particles evolve independently.

What would it look like if all particles moved together?

If $X_0 \sim \nu$, $X_t \sim \nu_t = h_t d\mu$ evolves given the SDE above.

$\nu_t$ defines a path between $\nu$ and $\mu$.

# On Langevin Monte-Carlo

Why does Langevin Monte-Carlo work?

$\mu$ stationary measure of the diffusion process solution of

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t.$$

Why is it slow?

All $n$ particles evolve independently.

What would it look like if all particles moved together?

If $X_0 \sim \nu$, $X_t \sim \nu_t = h_t d\mu$ evolves given the SDE above.

$\nu_t$ defines a path between $\nu$ and $\mu$.

At time $t$, the mass of $\nu_t$ at $x$ moves as $-\nabla \log h_t$.

$$\Rightarrow W_2(\nu, \mu) \leq \int_0^\infty \left( \int \|\nabla \log h_t\|^2 d\nu_t \right)^{1/2} dt$$

# On Langevin Monte-Carlo

Why does Langevin Monte-Carlo work?

$\mu$ stationary measure of the diffusion process solution of

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t.$$

Why is it slow?

All $n$ particles evolve independently.

What would it look like if all particles moved together?

If $X_0 \sim \nu$, $X_t \sim \nu_t = h_t d\mu$ evolves given the SDE above.

$\nu_t$ defines a path between $\nu$ and $\mu$.

At time $t$, the mass of $\nu_t$ at $x$ moves as $-\nabla \log h_t$.

Fisher information of $\nu_t$ w.r.t. $\mu$

$$\Rightarrow W_2(\nu, \mu) \leq \int_0^\infty \left( \int \|\nabla \log h_t\|^2 d\nu_t \right)^{1/2} dt$$

# Stein variational gradient descent

# Stein variational gradient descent

Idea: find an approximation $(\nu_k^h)_{k\in\mathbb{N}}$ of $(\nu_t)_{t\geq 0}$ and approximate $\nu_T$ for $T$ large enough.

# Stein variational gradient descent

Idea: find an approximation $(\nu_k^h)_{k\in\mathbb{N}}$ of $(\nu_t)_{t\geq 0}$ and approximate $\nu_T$ for $T$ large enough.

Problem: $(\nu_t)_{t\geq 0}$ tends to "diffuse" so it is hard to approximate through a discrete measure.

# Stein variational gradient descent

Idea: find an approximation $(\nu_k^h)_{k\in\mathbb{N}}$ of $(\nu_t)_{t\geq 0}$ and approximate $\nu_T$ for $T$ large enough.

Problem: $(\nu_t)_{t\geq 0}$ tends to "diffuse" so it is hard to approximate through a discrete measure.

Solution: smooth the mass movement using a kernel $K$.

$\Rightarrow$ if $d\nu_k^h = f d\mu$, move the mass at point $x$ by

$$-h\int K(x,y)\nabla \log f(y) d\nu(y) = h\int -\nabla V(y)K(x,y) + \nabla K(x,y) d\nu($$

# Stein variational gradient descent

Idea: find an approximation $(\nu_k^h)_{k\in\mathbb{N}}$ of $(\nu_t)_{t\geq 0}$ and approximate $\nu_T$ for $T$ large enough.

Problem: $(\nu_t)_{t\geq 0}$ tends to "diffuse" so it is hard to approximate through a discrete measure.

Solution: smooth the mass movement using a kernel $K$.

$\Rightarrow$ if $d\nu_k^h = f d\mu$, move the mass at point $x$ by

$$-h\int K(x,y)\nabla\log f(y)d\nu(y) = h\int -\nabla V(y)K(x,y)+\nabla K(x,y)d\nu($$

Mean displacement of $\nu_t$ if $\nu_t = \nu_k^h$

# Stein variational gradient descent

Idea: find an approximation $(\nu_k^h)_{k\in\mathbb{N}}$ of $(\nu_t)_{t\geq 0}$ and approximate $\nu_T$ for $T$ large enough.

Problem: $(\nu_t)_{t\geq 0}$ tends to "diffuse" so it is hard to approximate through a discrete measure.

Solution: smooth the mass movement using a kernel $K$.

$\Rightarrow$ if $d\nu_k^h = f d\mu$, move the mass at point $x$ by

$$-h\int K(x,y)\nabla \log f(y) d\nu(y) = h\int -\nabla V(y)K(x,y) + \nabla K(x,y) d\nu($$

Integration by parts

# Stein variational gradient descent

Idea: find an approximation $(\nu_k^h)_{k\in\mathbb{N}}$ of $(\nu_t)_{t\geq 0}$ and approximate $\nu_T$ for $T$ large enough.

Problem: $(\nu_t)_{t\geq 0}$ tends to "diffuse" so it is hard to approximate through a discrete measure.

Solution: smooth the mass movement using a kernel $K$.

$\Rightarrow$ if $d\nu_k^h = f d\mu$, move the mass at point $x$ by

$$-h\int K(x,y)\nabla \log f(y) d\nu(y) = h\int -\nabla V(y)K(x,y)+\nabla K(x,y)d\nu($$

Can be computed even for discrete $\nu$

# Stein variational gradient descent

Idea: find an approximation $(\nu_k^h)_{k\in\mathbb{N}}$ of $(\nu_t)_{t\geq 0}$ and approximate $\nu_T$ for $T$ large enough.

Problem: $(\nu_t)_{t\geq 0}$ tends to "diffuse" so it is hard to approximate through a discrete measure.

Solution: smooth the mass movement using a kernel $K$.

$\Rightarrow$ if $d\nu_k^h = f d\mu$, move the mass at point $x$ by

$$-h\int K(x,y)\nabla \log f(y) d\nu(y) = h\int -\nabla V(y)K(x,y) + \nabla K(x,y) d\nu($$

Algorithm:

- start with some discrete measure $\nu_0^h$ with $n$ particles $X_1^0, \dots X_n^0$.
  update the position of the particle $X_i^k$ with
- $$X_i^{k+1} = X_i^k + \frac{h}{n}\sum_j -\nabla V(X_j^k)K(X_i^k, X_j^k) + \nabla K(X_i^k, X_j^k)$$

# Guarantees for SVGD

# Guarantees for SVGD

Exisiting results:

# Guarantees for SVGD

Exisiting results:

- for $\nu_0$ continuous, SVGD is an approximation of a gradient flow in the space of measures.

# Guarantees for SVGD

Exisiting results:

- for $\nu_0$ continuous, SVGD is an approximation of a gradient flow in the space of measures.
- there are guarantees for the speed of convergence of this process. (Depends on $K$).

# Guarantees for SVGD

Exisiting results:

- for $\nu_0$ continuous, SVGD is an approximation of a gradient flow in the space of measures.
- there are guarantees for the speed of convergence of this process. (Depends on $K$).
- for $t > 0$, a discrete-time approximation can be used to approximate $\nu_t$ but with a continuous $\nu_0$.

# Guarantees for SVGD

Exisiting results:

- for $\nu_0$ continuous, SVGD is an approximation of a gradient flow in the space of measures.
- there are guarantees for the speed of convergence of this process. (Depends on $K$).
- for $t > 0$, a discrete-time approximation can be used to approximate $\nu_t$ but with a continuous $\nu_0$.
- if $\nu_0$ is discrete with $N$ particles, then the previous result is true as $N \to \infty$.

# Guarantees for SVGD

Exisiting results:

- for $\nu_0$ continuous, SVGD is an approximation of a gradient flow in the space of measures.
- there are guarantees for the speed of convergence of this process. (Depends on $K$).
- for $t > 0$, a discrete-time approximation can be used to approximate $\nu_t$ but with a continuous $\nu_0$.
- if $\nu_0$ is discrete with $N$ particles, then the previous result is true as $N \to \infty$.

Desirable results:

# Guarantees for SVGD

Exisiting results:

- for $\nu_0$ continuous, SVGD is an approximation of a gradient flow in the space of measures.
- there are guarantees for the speed of convergence of this process. (Depends on $K$).
- for $t > 0$, a discrete-time approximation can be used to approximate $\nu_t$ but with a continuous $\nu_0$.
- if $\nu_0$ is discrete with $N$ particles, then the previous result is true as $N \to \infty$.

Desirable results:

- Non-asymptotic result w.r.t. $N$ for the measure obtained after convergence of $SVGD$ for a standard metric $W_2$.

# Guarantees for SVGD

Exisiting results:

- for $\nu_0$ continuous, SVGD is an approximation of a gradient flow in the space of measures.
- there are guarantees for the speed of convergence of this process. (Depends on $K$).
- for $t > 0$, a discrete-time approximation can be used to approximate $\nu_t$ but with a continuous $\nu_0$.
- if $\nu_0$ is discrete with $N$ particles, then the previous result is true as $N \to \infty$.

Desirable results:

- Non-asymptotic result w.r.t. $N$ for the measure obtained after convergence of $SVGD$ for a standard metric $W_2$.

  $\Rightarrow$ Choice of $K$.

Thanks for your attention