CPSC 661: Sampling Algorithms in ML

Andre Wibisono

February 1, 2021

Logistics

CLASS: CPSC 661: **Sampling Algorithms in Machine Learning**
Yale University, Spring 2021

TIME: Monday + Wednesday, 11:35 am – 12:50 pm

PLACE: `https://yale.zoom.us/j/95006937328` (see Canvas)

INSTRUCTOR: Andre Wibisono (`andre.wibisono@yale.edu`)

Please fill out survey: `https://forms.gle/1L8ohnyU2PW3onm27`

About Me

Andre Wibisono

andre.wibisono@yale.edu

Joined **Yale University** Spring 2021

(PhD Berkeley, BS MIT)

Research interests: Machine learning, optimization, sampling, statistics, information theory, game dynamics, ...

Learning from Data

How to Learn from Data?

- Learning as Optimization of objective encoding goal
- Machine Learning: Large-scale, high-dimensional, noisy data

Learning from Data

How to Learn from Data?

- Learning as Optimization of objective encoding goal
- Machine Learning: Large-scale, high-dimensional, noisy data
- Practical successes $\rightarrow$ societal impact
- Need better theory $\leftrightarrow$ practice
- Optimization and Sampling as fundamental primitives

This Class

A study of **Sampling** problems

and efficient **Algorithms**

for problems motivated by **Machine Learning**

This Class

A study of **Sampling** problems

and efficient **Algorithms**

for problems motivated by **Machine Learning**

(Beautiful mathematics: geometry, dynamics, probability)

Logistics

Grading

- Problem sets / Scribe notes: $3 \times 15\% = 45\%$
- Project:
  - ⋆ Presentation $= 25\%$
  - ⋆ Written report $= 30\%$

Problem Set

- 3 problem sets, every $\sim$4 weeks
- Choose some from a list of questions (theory / code)
- 1 problem set can be substituted for 1 scribe notes

Scribe Notes

Scribe: Write lecture in Latex notes

- Good to practice writing
- Helps clarify ideas
- Helps as reference for fellow students
- 1 scribe notes can be exchanged for 1 problem set

Please sign up at: `https://tinyurl.com/y3j246v3`

Project

Explore a question about sampling

- Will suggest some in lectures, or related to your research!
- Minimum: Literature review, known results (theory / practice)
- Ideal: Extend to other settings, new algorithms, experiments
- Presentation: 20–30 minutes in last week of class
- Report: $\sim$10 pages on last day of class

Why Sampling?

Why Sampling?

1. Statistics
2. Bayesian inference
3. Optimization (annealing)
4. Randomized approximation
5. Benefits of randomness

...

1. Statistics

- Estimate statistics of a distribution ν from samples

$$\mathbb{E}_\nu[f] = \int_X f(x)\, d\nu(x)$$

→ draw $X_1, \ldots, X_n \sim \nu$ iid

estimate $\mathbb{E}_{\hat{\nu}}[f] = \frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow{n\to\infty} \mathbb{E}_\nu[f]$

- Simulate ensemble behavior, statistical mechanics

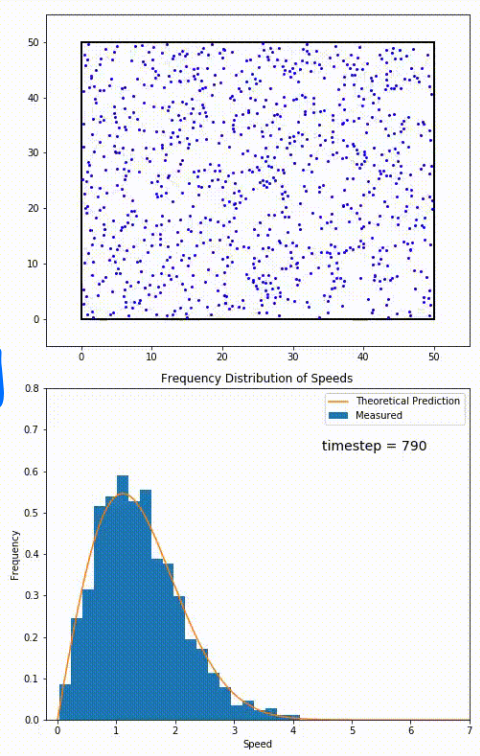


Figure from Wikipedia:

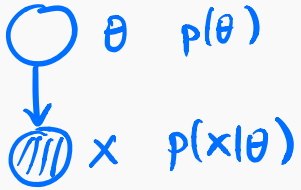
`https://en.wikipedia.org/wiki/Maxwell-Boltzmann_distribution`

2. Bayesian Inference

Bayes Rule:

$$p(\theta \mid x) = \frac{p(\theta) \cdot p(x \mid \theta)}{p(x)}$$

posterior, prior, likelihood



- Inference: Posterior reflects our belief after observing data
- Statistics of posterior, uncertainty quantification
  (MAP = mode, Bayes estimator = mean of posterior)

Bayes Rule:

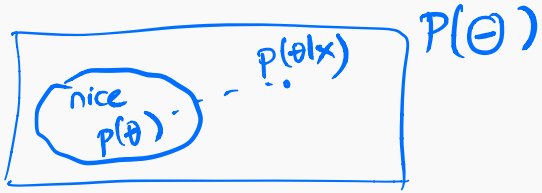
$$p(\theta \mid x) = \frac{p(\theta) \cdot p(x \mid \theta)}{p(x)}$$

- Issue: Can only compute posterior up to constant

$$p(x) = \int_{\Theta} p(\theta) \cdot p(x \mid \theta)\, d\theta$$

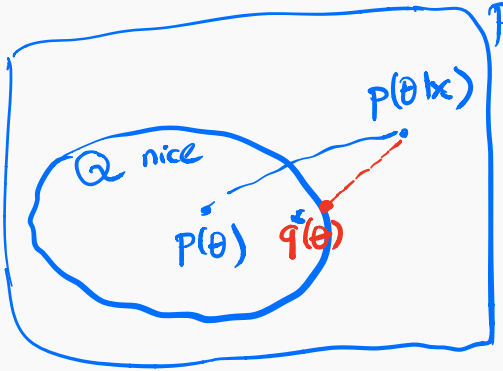
large $p(x, \theta)$

- Posterior usually not a nice distribution (cf. *conjugate prior*)



How to approximate posterior?

- **Optimization:** Variational Bayes (VB)
  - Find the best tractable approximation to posterior
  - Usually non-convex, local optima, no approximation guarantee



choose Q nice (exp. family, Gaussian, mean-field ...)

$$q^*(\theta) = \operatorname{argmin}_{q \in Q} KL(q \,||\, p(\theta|x))$$

↑ relative entropy

$\Updownarrow$

$\max$ ELBO (evidence lower bound)

How to approximate posterior?

- **Optimization:** Variational Bayes (VB)
  - Find the best tractable approximation to posterior
  - Usually non-convex, local optima, no approximation guarantee

- **Sampling:** Markov Chain Monte Carlo (MCMC)
  - Draw samples from posterior $\theta_1, \ldots, \theta_n \sim p(\theta|x)$
  - Only need to evaluate posterior up to a constant

(eg. Metropolis-Hastings)

- Can compute $\nabla_\theta \log p(\theta|x)$ $\longrightarrow$ Langevin algorithm
- Q: How fast converge?

3. Optimization

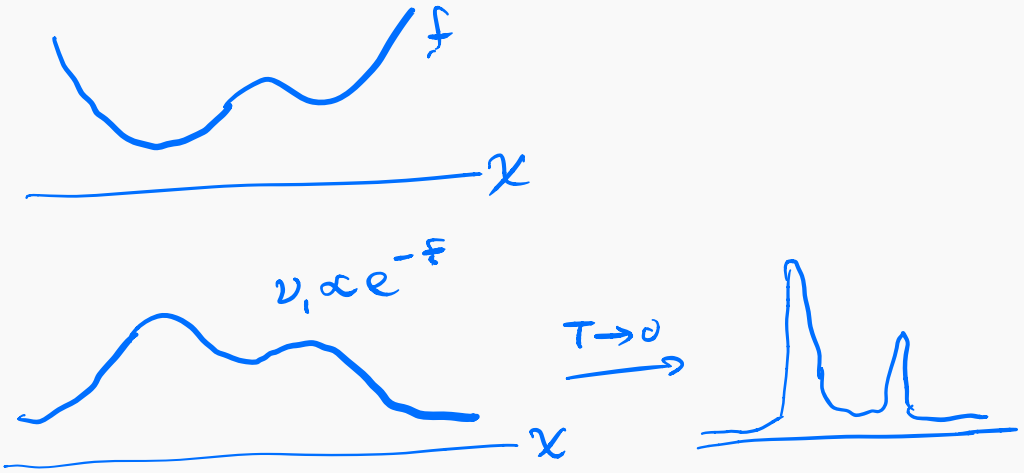
Annealing: Optimization via Sampling from zero-noise distribution

$$\nu_T(x) \propto e^{-f(x)/T}$$

T = temperature

$T \to 0$

$$\lim_{T \to 0} \nu_T = \delta_{\arg\min f}$$



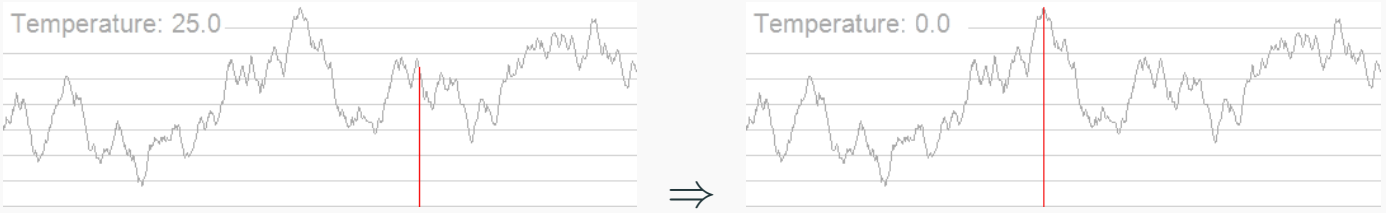


Figure from Wikipedia: `https://en.wikipedia.org/wiki/Simulated_annealing`

4. Randomized Approximation

Some hard deterministic problems have randomized approximations

- Counting perfect matchings $\Leftrightarrow$ computing permanent
- Computing volume of a convex body in $\mathbb{R}^n$

4. Randomized Approximation

Some hard deterministic problems have randomized approximations

- Counting perfect matchings $\Leftrightarrow$ computing permanent
- Computing volume of a convex body in $\mathbb{R}^n$
  - ⋆ [Dyer, Frieze '88]: #P-hard even for an explicit polytope
  - ⋆ Any deterministic algorithm must take $O(1/\epsilon^n)$ time to reach relative error $(1+\epsilon)$

4. Randomized Approximation

Some hard deterministic problems have randomized approximations

- Counting perfect matchings $\Leftrightarrow$ computing permanent
- Computing volume of a convex body in $\mathbb{R}^n$ ←
  - ⋆ [Dyer, Frieze '88]: #P-hard even for an explicit polytope
  - ⋆ Any deterministic algorithm must take $O(1/\epsilon^n)$ time to reach relative error $(1+\epsilon)$
  - ⋆ [Dyer, Frieze, Kannan '89]: $\exists$ randomized algorithm that achieves relative error $(1+\epsilon)$ with probability $\geq 1-\delta$ in time $\text{poly}(n, \frac{1}{\epsilon}, \log \frac{R}{\delta})$
  - ⋆ Key step: Sample from uniform distribution on convex body

4. Randomized Approximation

Volume Computation: an ongoing adventure

	Exponent	New aspects
Dyer-Frieze-Kannan 89	23	everything
Lovász-Simonovits 90	16	localization
Applegate-K 90	10	logconcave integration
L 90	10	ball walk
DF 91	8	error analysis
LS 93	7	multiple improvements
KLS 97	5	speedy walk, isotropy
LV 03,04	4	annealing, wt. isoper.
LV 06	4	integration, local analysis
Cousins-V. 15 (well-rounded)	3	Gaussian cooling

Slide from Vempala, *Reducing Isotropy to KLS: An Almost Cubic Volume*, 2021,

`https://www.youtube.com/watch?v=gmItyFygBfc`

5. Benefits of Randomness: Smoothness

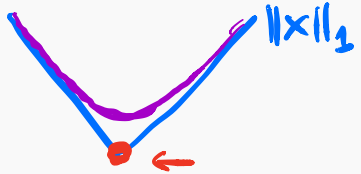
Opt: $\min_{x \in \mathbb{R}^n} f(x)$ ← ∃ nice theory

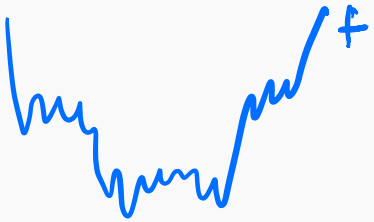
when f smooth: ∇f, $\nabla^2 f$

$\|\nabla f\| \le L_1$

$\|\nabla^2 f\| \le L_2$ ...

what if f non-smooth?





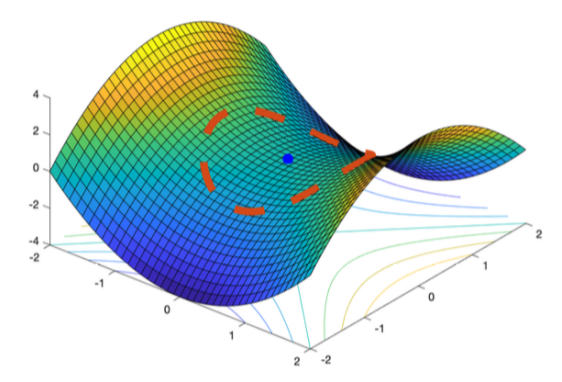
Smoothing: $f_\sigma(x) = \mathbb{E}_v[f(x+\sigma z)]$, $z \sim v$

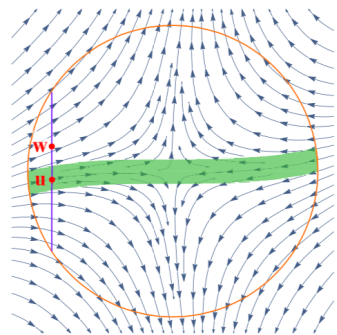
$\nabla_x f_\sigma(x) = \mathbb{E}_v[\nabla f(x+\sigma z)]$

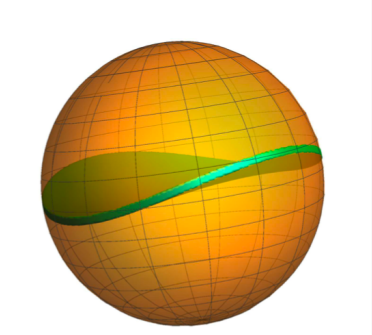
↑

$\nabla f(x+\sigma z)$

5. Benefits of Randomness: Escape from Saddle Points

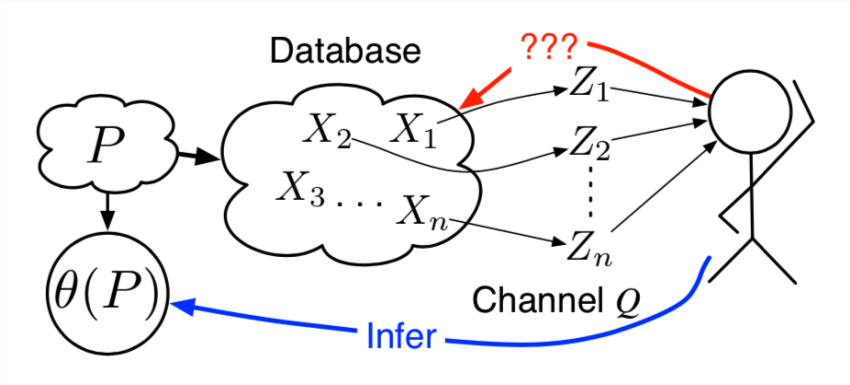


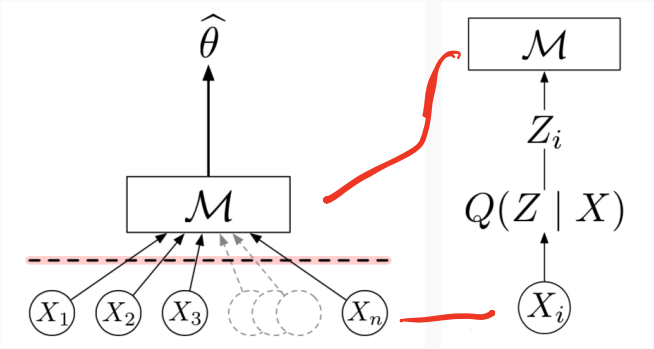


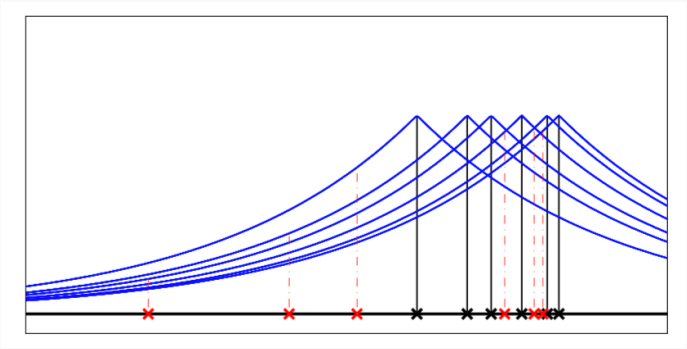


Figures from Jin et al., *How to Escape Saddle Points Efficiently*, ICML 2017

5. Benefits of Randomness: Privacy







Figures from Duchi, *Local Privacy and Statistical Minimax Rates*, 2013,

`https://simons.berkeley.edu/talks/john-duchi-2013-12-13`

How to Sample?

Sampling

Goal: Draw sample (random variable) $X \sim \nu$ on state space $\mathcal{X}$

$\mathcal{X}$ discrete:

- $\mathcal{X} = \{0, 1\}$, $\nu = (\frac{1}{2}, \frac{1}{2})$ or any $(p, 1-p)$
- $\mathcal{X} = \{1, \ldots, n\}$, $\nu = (p_1, \ldots, p_n)$

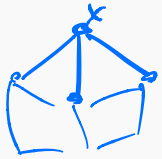
Sampling

Goal: Draw sample (random variable) $X \sim \nu$ on state space $\mathcal{X}$

$\mathcal{X}$ discrete:

- $\mathcal{X} = \{0, 1\}$, $\nu = (\frac{1}{2}, \frac{1}{2})$ or any $(p, 1-p)$
- $\mathcal{X} = \{1, \ldots, n\}$, $\nu = (p_1, \ldots, p_n)$
- $\mathcal{X} =$ graph, group, matroid, ...
- $\mathcal{X} =$ permutations, matchings, subsets, ...

$G = (V, E)$, $\mathcal{X} = V = \{1, \ldots, n\}$



$\mathcal{X} = S_n$ $\quad |\mathcal{X}| = n! \sim n^n$

$\mathcal{X} = \{0,1\}^n$ $\quad |\mathcal{X}| = 2^n$

Sampling

Goal: Draw sample (random variable) $X \sim \nu$ on state space $\mathcal{X}$

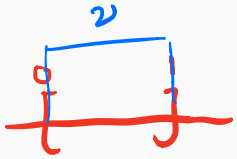
$\mathcal{X}$ discrete:

- $\mathcal{X} = \{0, 1\}$, $\nu = (\frac{1}{2}, \frac{1}{2})$ or any $(p, 1-p)$
- $\mathcal{X} = \{1, \ldots, n\}$, $\nu = (p_1, \ldots, p_n)$
- $\mathcal{X} =$ graph, group, matroid, ...
- $\mathcal{X} =$ permutations, matchings, subsets, ...

$\mathcal{X}$ continuous:

- $\mathcal{X} = \mathbb{R}^n$ or a subset $K \subset \mathbb{R}^n$
- $\mathcal{X} =$ manifold, sphere, hyperbolic space, faces, ...

How to Sample?



1. $\mathcal{X} = [0, 1]$, $\nu =$ Uniform
   - Can generate from fair coin flips (Uniform on $\{0, 1\}$)

How to Sample?

1. $\mathcal{X} = [0,1]$, $\nu =$ Uniform
   - Can generate from fair coin flips (Uniform on $\{0,1\}$)

2. $\mathcal{X} = \mathbb{R}$, $\nu =$ any probability distribution with CDF $F_\nu(t) = \Pr_\nu(X \le t)$
   - **Inverse transform**:

$$U \sim \text{Uniform}([0,1]) \quad \Rightarrow \quad X = F_\nu^{-1}(U) \sim \nu$$

$$\Pr(X \le t) = \Pr(F_\nu^{-1}(U) \le t)$$
$$= \Pr(U \le F_\nu(t))$$
$$= F_\nu(t)$$

How to Sample?

$x = (x_1, \ldots, x_n)$

3. $\mathcal{X} = \mathbb{R}^n$, $\nu =$ any probability distribution with marginals ν_i
   - **Gibbs Sampling:** Update coordinates $i \in \{1, \ldots, n\}$:

     $$X_i \sim \nu_i(\cdot \mid X_{\setminus i})$$

     (also for $\mathcal{X}$ discrete, e.g. *Glauber dynamics* for Ising model)

How to Sample?

3. $\mathcal{X} = \mathbb{R}^n$, $\nu =$ any probability distribution with marginals ν_i

   - **Gibbs Sampling:** Update coordinates $i \in \{1, \ldots, n\}$:

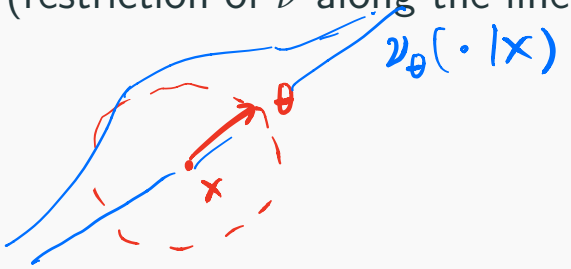
     $$X_i \sim \nu_i(\cdot \mid X_{\setminus i})$$

     (also for $\mathcal{X}$ discrete, e.g. *Glauber dynamics* for Ising model)

   - **Hit-and-Run:** Update along random direction $\theta \in \mathbb{S}^{n-1}$:

     $$X \sim \nu_\theta(\cdot \mid X)$$

     (restriction of ν along the line passing X at direction θ)



How to Sample?

4. $\mathcal{X} = \mathbb{R}^n$, $\nu =$ Gaussian $\mathcal{N}(0, I)$ with density

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

- Can sample by n independent samples of $\mathcal{N}(0,1)$ in $\mathbb{R}$:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \quad \Rightarrow \quad X = (X_1, \ldots, X_n) \sim \mathcal{N}(0, I)$$

How to Sample?

4. $\mathcal{X} = \mathbb{R}^n$, $\nu =$ Gaussian $\mathcal{N}(0, I)$ with density

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

- Can sample by n independent samples of $\mathcal{N}(0,1)$ in $\mathbb{R}$:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \quad \Rightarrow \quad X = (X_1, \ldots, X_n) \sim \mathcal{N}(0, I)$$

- Can sample from $\mathcal{N}(\mu, \Sigma)$ by linear transformation:

$$X \sim \mathcal{N}(0, I) \quad \Rightarrow \quad X' = \mu + \Sigma^{1/2} X \sim \mathcal{N}(\mu, \Sigma)$$

How to Sample?

4. $\mathcal{X} = \mathbb{R}^n$, $\nu =$ Gaussian $\mathcal{N}(0, I)$ with density

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

- Can sample by n independent samples of $\mathcal{N}(0,1)$ in $\mathbb{R}$:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \quad \Rightarrow \quad X = (X_1, \ldots, X_n) \sim \mathcal{N}(0, I)$$

- Can sample from $\mathcal{N}(\mu, \Sigma)$ by linear transformation:

$$X \sim \mathcal{N}(0, I) \quad \Rightarrow \quad X' = \mu + \Sigma^{1/2} X \sim \mathcal{N}(\mu, \Sigma)$$

- **Brownian motion** (Gaussian walk):

$$X' = X + 2\eta Z, \qquad Z \sim \mathcal{N}(0, I)$$

How to Sample?

5. $\mathcal{X} = \mathbb{R}^n$, $\nu =$ any probability distribution
   - Assume can evaluate ν up to normalizing constant
   - **Metropolis-Hastings** (with Gaussian walk):

$$X' = X + \sqrt{2\eta}\, Z, \qquad Z \sim \mathcal{N}(0, I)$$

accept X' with probability $\min\left\{1, \dfrac{\nu(X')}{\nu(X)}\right\}$

How to Sample?

5. $\mathcal{X} = \mathbb{R}^n$, $\nu =$ any probability distribution

   - Assume can evaluate ν up to normalizing constant
   - **Metropolis-Hastings** (with Gaussian walk):

$$X' = X + \sqrt{2\eta}\, Z, \qquad Z \sim \mathcal{N}(0, I)$$

$$\text{accept } X' \text{ with probability } \min\left\{1, \frac{\nu(X')}{\nu(X)}\right\}$$

   - Needs zero-order access (value of ν up to a constant)
   - Ball walk: $Z \sim$ Uniform on a ball

How to Sample?

6. $\mathcal{X} = \mathbb{R}^n$, $\nu =$ any distribution with density $\nu(x) \propto e^{-f(x)}$
   - Assume can evaluate gradient $\nabla f = -\nabla \log \nu$
   - **Langevin algorithm**:

$$X' = X - \eta \nabla f(X) + \sqrt{2\eta}\, Z, \qquad Z \sim \mathcal{N}(0, I)$$

How to Sample?

6. $\mathcal{X} = \mathbb{R}^n$, $\nu =$ any distribution with density $\nu(x) \propto e^{-f(x)}$
   - Assume can evaluate gradient $\nabla f = -\nabla \log \nu$
   - **Langevin algorithm**:

     $$X' = X - \eta \nabla f(X) + \sqrt{2\eta}\, Z, \qquad Z \sim \mathcal{N}(0, I)$$

     (+ Metropolis-Hastings $\Rightarrow$ MALA)
   - Needs first-order access (gradient of ν)
   - Where does this come from?

Optimization

$$\min_{x \in \mathbb{R}^n} f(x)$$

Gradient flow:

$$\dot{X}_t = -\nabla f(X_t) \quad \leftarrow \text{cts time}$$

Gradient descent:

$$x_{k+1} = x_k - \eta \nabla f(x_k) \quad \leftarrow \text{disc time}$$

Noise

Brownian motion:

$$dX_t = \sqrt{2} dW_t$$

Gaussian walk:

$$x_{k+1} = x_k + \sqrt{2\eta} Z_k, \qquad Z_k \sim \mathcal{N}(0, I)$$

Sampling $\nu \propto e^{-f}$

Langevin dynamics:

$$dX_t = -\nabla f(X_t)\, dt + \sqrt{2} dW_t$$

Langevin algorithm:

$$x_{k+1} = x_k - \eta \nabla f(x_k) + \sqrt{2\eta}\, Z_k, \qquad Z_k \sim \mathcal{N}(0, I)$$

Optimization + **Noise** $\rightarrow$ **Sampling**

What is Happening?

Optimization + **Noise** $\rightarrow$ **Sampling**

- e.g. Gradient flow + Gaussian noise = Langevin dynamics
- Can we generalize?

What is Happening?

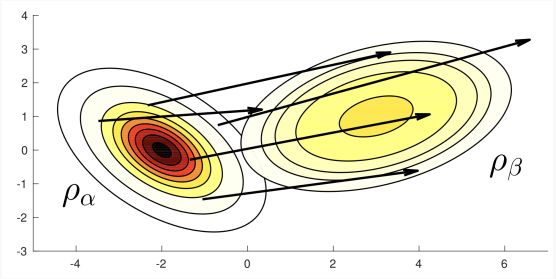
Optimization + **Noise** $\rightarrow$ **Sampling**

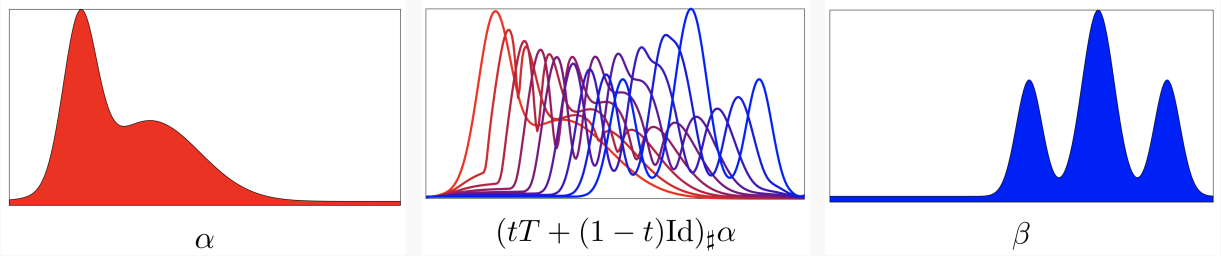
- e.g. Gradient flow + Gaussian noise = Langevin dynamics
- Can we generalize?
- Other methods? e.g. Acceleration/momentum?
- Other noise? Gaussian noise $\leftrightarrow$ Entropy
- Formalize via *optimal transport*, Wasserstein geometry

Geometry of Probability Distributions

Optimal transport $\Rightarrow$ Wasserstein geometry

- Earth-mover, mass transportation, Monge-Kantorovich (W_1)
- Applications in computer vision, ML, statistics, economics, ...





Figures from Peyre and Cuturi, *Computational Optimal Transport*, 2019, available at: `https://optimaltransport.github.io/book`

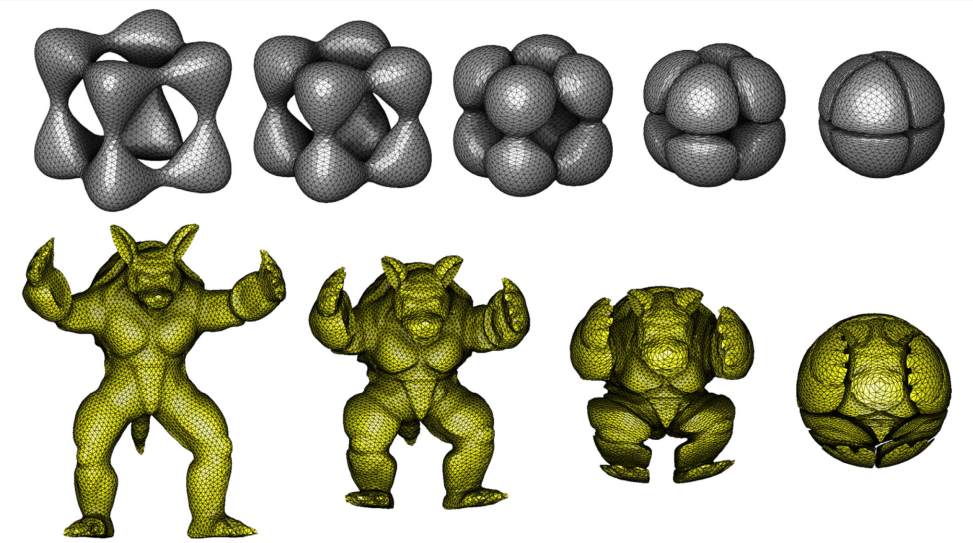


Fig. 13. Interpolation of 3D volumes using optimal transport.

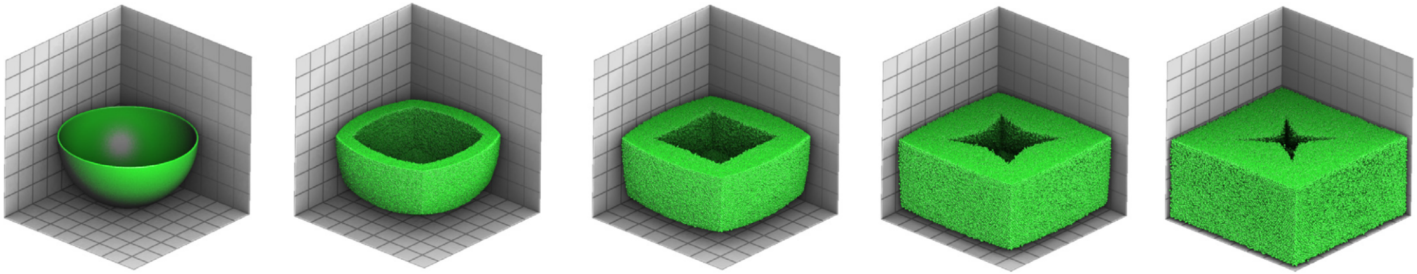


Fig. 14. Computing the transport between objects of different dimension, from a sphere to a cube. The sphere is sampled with 10 million points. The cross-section reveals the formation of a singularity that has some similarities with the medial axis of the cube.

Figures from Lévy and Schwindt, *Notions of optimal transport theory and how to implement them on a computer*, Computers & Graphics, 2018

Geometry of Probability Distributions

Wasserstein metric W_2 on $\mathcal{P}(\mathbb{R}^n)$

- Differentiable manifold, Otto calculus
- Deep connection between geometry of $\mathcal{X}$ and $\mathcal{P}(\mathcal{X})$
  - Convexity of Entropy $\Leftrightarrow$ curvature of $\mathcal{X}$
  - Also when $\mathcal{X}$ = manifold or $\mathcal{X}$ = graph / discrete space

Geometry of Probability Distributions

Wasserstein metric W_2 on $\mathcal{P}(\mathbb{R}^n)$

- Differentiable manifold, Otto calculus
- Deep connection between geometry of $\mathcal{X}$ and $\mathcal{P}(\mathcal{X})$
  - Convexity of Entropy $\Leftrightarrow$ curvature of $\mathcal{X}$
  - Also when $\mathcal{X}$ = manifold or $\mathcal{X}$ = graph / discrete space
- Geometric view of information-theoretic objects:
  Entropy, Fisher information, Brownian motion, ...
- Villani (Fields Medal 2010): **Boltzmann's *H*-Theorem**:
  Entropy is increasing (in Boltzmann gas model)

Good references: Villani, *Topics in Optimal Transportation*, AMS, 2003

Villani, *Optimal Transport: Old and New*, Springer, 2009

Sampling as Optimization

Wasserstein metric W_2 on $\mathcal{P}(\mathbb{R}^n)$

- Optimization perspective on stochastic dynamics for sampling
  - ⋆ Brownian motion = Gradient flow for maximizing Entropy
  - ⋆ Langevin dynamics = Gradient flow for KL divergence

Sampling as Optimization

Wasserstein metric W_2 on $\mathcal{P}(\mathbb{R}^n)$

- Optimization perspective on stochastic dynamics for sampling
  - ⋆ Brownian motion = Gradient flow for maximizing Entropy
  - ⋆ Langevin dynamics = Gradient flow for KL divergence
- Study Sampling as Optimization in the space of distributions
  - ⋆ Systematic translation Optimization ⇒ Sampling algorithms?

Sampling as Optimization

Wasserstein metric W_2 on $\mathcal{P}(\mathbb{R}^n)$

- Optimization perspective on stochastic dynamics for sampling
  - ⋆ Brownian motion = Gradient flow for maximizing Entropy
  - ⋆ Langevin dynamics = Gradient flow for KL divergence
- Study Sampling as Optimization in the space of distributions
  - ⋆ Systematic translation Optimization ⇒ Sampling algorithms?
- Complexity theory of Sampling (parallel to Optimization)
  - ⋆ Convex f $\Leftrightarrow$ Log-concave $\nu \propto e^{-f}$
  - ⋆ Beyond convexity / log-concavity: isoperimetry, ...
  - ⋆ Active area of research, many interesting questions!

Course Plan

Course Plan

I. Sampling via Random Walk / Markov Chain
   - ⋆ Reversibility: Spectral gap $\Leftrightarrow$ conductance
   - ⋆ Metropolis-Hastings scheme (accept-reject)
   - ⋆ MH as projection in space of reversible Markov chains

Course Plan

I. Sampling via Random Walk / Markov Chain
- ⋆ Reversibility: Spectral gap $\Leftrightarrow$ conductance
- ⋆ Metropolis-Hastings scheme (accept-reject)
- ⋆ MH as projection in space of reversible Markov chains

II. Sampling as Optimization
- ⋆ Review on optimization, optimal transport
- ⋆ Langevin dynamics for sampling, interpretation as gradient flow
- ⋆ Convergence and bias of Langevin algorithm in discrete time
- ⋆ Other discretization/variants: proximal, Hessian manifold, ...

Course Plan

I. Sampling via Random Walk / Markov Chain
- ⋆ Reversibility: Spectral gap $\Leftrightarrow$ conductance
- ⋆ Metropolis-Hastings scheme (accept-reject)
- ⋆ MH as projection in space of reversible Markov chains

II. Sampling as Optimization
- ⋆ Review on optimization, optimal transport
- ⋆ Langevin dynamics for sampling, interpretation as gradient flow
- ⋆ Convergence and bias of Langevin algorithm in discrete time
- ⋆ Other discretization/variants: proximal, Hessian manifold, ...

II. Other Sampling Algorithms
- ⋆ Underdamped Langevin dynamics (momentum)
- ⋆ Hamiltonian Monte Carlo
- ⋆ Normalizing flow, Stein variational gradient descent, ...

IV. Student presentations

Thank you!

Questions?

Please fill out survey: `https://forms.gle/1L8ohnyU2PW3onm27`

Scribe sign-up: `https://tinyurl.com/y3j246v3`