CPSC 661: Sampling Algorithms in ML

Andre Wibisono

March 10, 2021

Yale University

Last time

I. Classical theory of sampling

- Reversible Markov chain
- Spectral gap $\Leftrightarrow$ Conductance
- Mixing time bound via s-conductance
- Metropolis-Hastings algorithm: MRW and MALA
- Mixing time: $\tilde{O}(n^2\kappa^2)$ for MRW, $\tilde{O}(n^2\kappa)$ for MALA

Today: Optimization and dynamics

Optimization

A universal language for describing and achieving goal

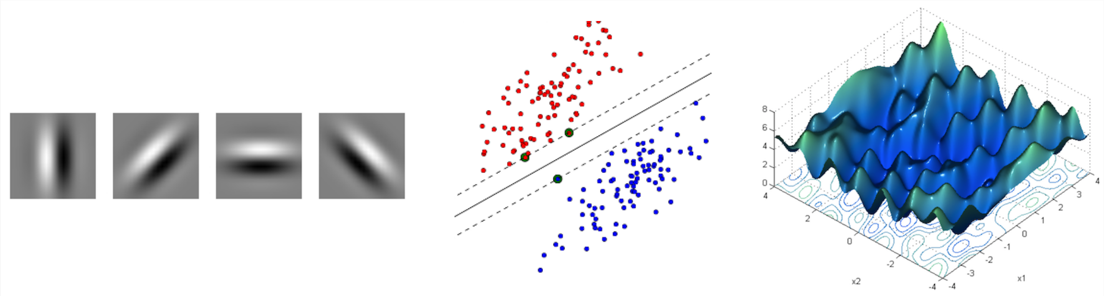
1. **Computer Science:** Greedy algorithms

   Modeling:

   - engineering (performance, cost)
   - economics (utility, reward)
   - biology (food, reproduction)
   - psychology (happiness?)
   - ...

Optimization

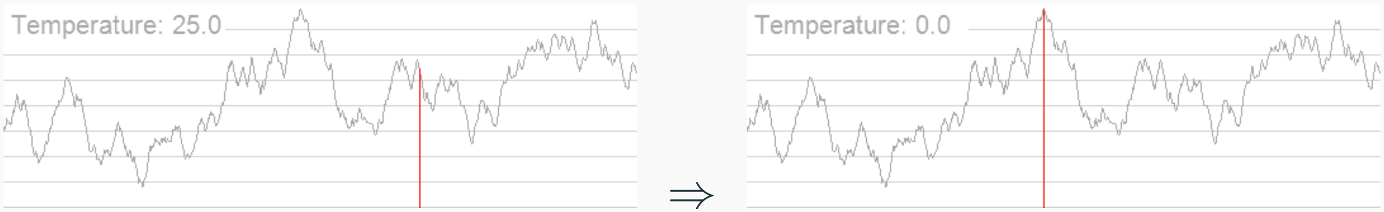
2. **Machine Learning:** Learning from data as optimization of objective function which encodes the goal



- Large-scale, high-dimensional, noisy data
- Classical models $\Rightarrow$ convex objectives
- Neural networks, variational inference $\Rightarrow$ non-convex
- Some *hidden convexity* via parameterization, manifold

Optimization

3. Can come from randomness (statistical physics)
    - As temperature $\to 0$, ensemble $\to$ ground state (lowest energy)
    - *Annealing*: Optimization via sampling from zero-noise distribution



Is everything optimization?

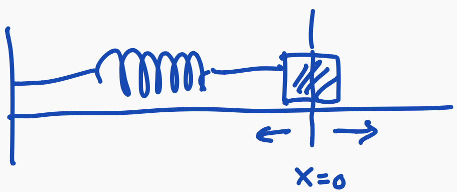
4. **Physics:** Newton's Law: Force = mass × acceleration

$$m\ddot{X}_t = -\nabla U(X_t)$$

- Conserves energy (*Hamiltonian*): $\mathcal{H} = \frac{m}{2}\|\dot{X}_t\|^2 + U(X_t)$

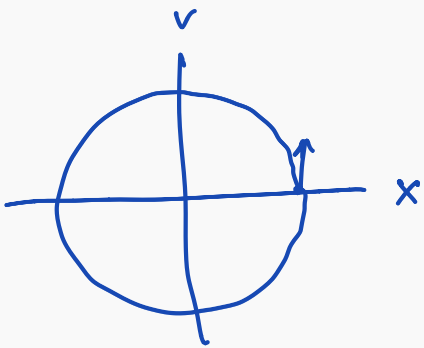
eg. $U(x) = \frac{1}{2}\|x\|^2$

Harmonic oscillator: $m\ddot{X}_t = -X_t$



Hamiltonian flow:

$$\begin{cases} \dot{X}_t = V_t \\ \dot{V}_t = -\frac{1}{m} X_t \end{cases}$$



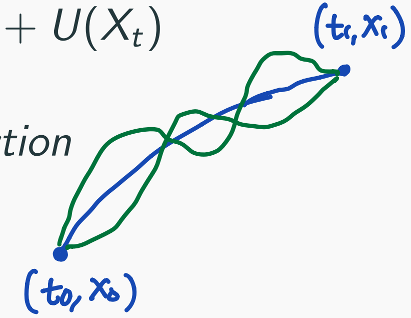
Is everything optimization?

4. **Physics:** Newton's Law: Force = mass × acceleration

$$\boxed{m\ddot{X}_t = -\nabla U(X_t)}$$

- Conserves energy (*Hamiltonian*): $\mathcal{H} = \frac{m}{2}\|\dot{X}_t\|^2 + U(X_t)$

- **Principle of least action:** Curve minimizes *action*



$$\mathcal{A} = \int_{t_0}^{t_1} \mathcal{L}(X_t, \dot{X}_t)\, dt$$

where $\mathcal{L}(X_t, \dot{X}_t) = \frac{m}{2}\|\dot{X}_t\|^2 - U(X_t)$ is the *Lagrangian* s.t. $X_{t_0} = x_0$, $X_{t_1} = x_1$

- Captures intrinsic geometry, covariant representation

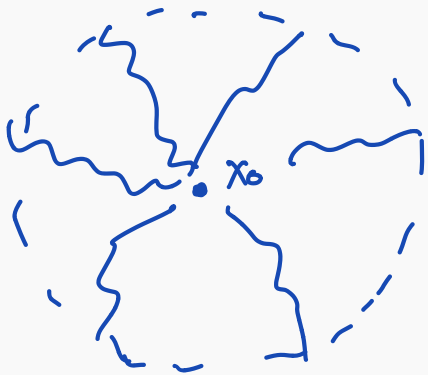
- Governs all physics: Electromagnetism, relativity, quantum, ...

Is everything optimization?

5. **Randomness:** Random walk, Brownian motion
    - Pure exploration, no objective

$$dX_t = dW_t$$

$$X_t \overset{d}{=} X_0 + \sqrt{t}\, Z, \quad Z \sim \mathcal{N}(0, I)$$



Is everything optimization?

5. **Randomness:** Random walk, Brownian motion

    - Pure exploration, no objective

    - This is maximizing *entropy* (a measure of randomness)

        - ⋆ Entropy increases along Brownian motion

        - ⋆ Brownian motion (heat flow) is gradient flow of $-$entropy

    - Also in discrete space: Random walk on graph

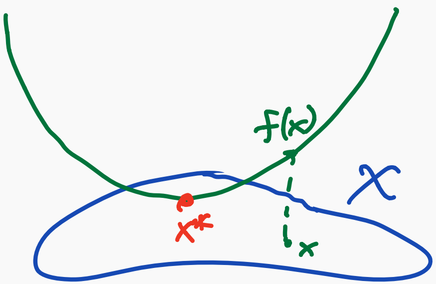
Exercise: What is *not* optimization?

Optimization

Given a space $\mathcal{X}$ and an objective function $f : \mathcal{X} \to \mathbb{R}$

Want to find minimizer

$$x^* = \arg\min_{x \in \mathcal{X}} f(x)$$



Optimization

Given a space $\mathcal{X}$ and an objective function $f : \mathcal{X} \to \mathbb{R}$

Want to find minimizer

$$x^* = \arg\min_{x \in \mathcal{X}} f(x)$$

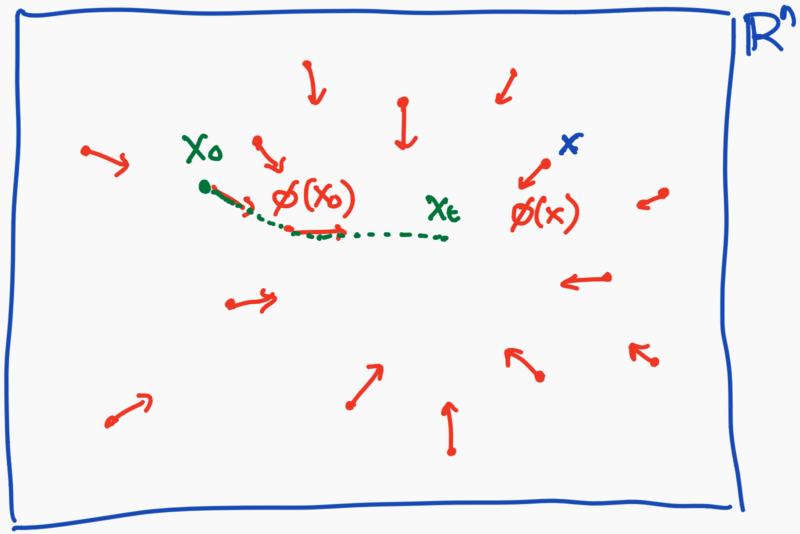
- Or find $\tilde{x}$ such that $f(\tilde{x}) - f(x^*) \leq \epsilon$ or $d(\tilde{x}, x^*) \leq \epsilon$
- In the worst case can be NP-hard (exponential time)
- With some structures (e.g. convexity) we can solve efficiently
- For now $\mathcal{X} = \mathbb{R}^n$, but also for manifold

Dynamics

A **dynamics** on $\mathbb{R}^n$ is determined by a vector field $\phi \colon \mathbb{R}^n \to \mathbb{R}^n$

From any $X_0 \in \mathbb{R}^n$, generate a *flow* $(X_t)_{t \geq 0}$ following:

$$\dot{X}_t = \phi(X_t)$$



Dynamics

A **dynamics** on $\mathbb{R}^n$ is determined by a vector field $\phi\colon \mathbb{R}^n \to \mathbb{R}^n$

From any $X_0 \in \mathbb{R}^n$, generate a *flow* $(X_t)_{t\geq 0}$ following:

$$\boxed{\dot{X}_t = \phi(X_t)} \quad (*)$$

- What does this mean? For small dt: $X_{t+dt} = X_t + \phi(X_t)\,dt + O(\cancelto{0}{dt^2})$
  (in discrete time many implementations, different performance)

- *Chain rule:*

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \dots, \frac{\partial f(x)}{\partial x_n} \right)$$

$$\frac{d}{dt} f(X_t) = \langle \nabla f(X_t), \dot{X}_t \rangle \overset{(*)}{=} \langle \nabla f(X_t), \phi(X_t) \rangle$$

$f(X_t) \in \mathbb{R}$

$X_t \in \mathbb{R}^n$

Dynamics for Optimization

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

1. **Gradient flow:**

$$\dot{X}_t = -\nabla f(X_t)$$

2. Heavy ball / accelerated gradient flow: (Polyak, Nesterov, ...)

$$\ddot{X}_t + \gamma \dot{X}_t + \nabla f(X_t) = 0$$

Gradient flow

$$\frac{d}{dt} X_t = \boxed{\dot{X}_t = -\nabla f(X_t)}$$

- First-order (in time) dynamics
- Greedy:

$$\cancel{-\nabla f(X_t) =}$$

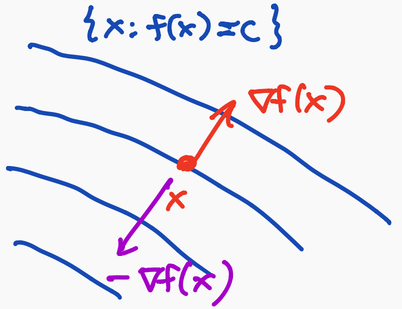
$$-\nabla f(x) = \operatorname*{argmin}_{v \in \mathbb{R}^n} \left\{ \langle \nabla f(x), v \rangle + \frac{1}{2} \|v\|^2 \right\}$$

- Descent method:

$$\frac{d}{dt} f(X_t) = \langle \nabla f(X_t), \dot{X}_t \rangle$$

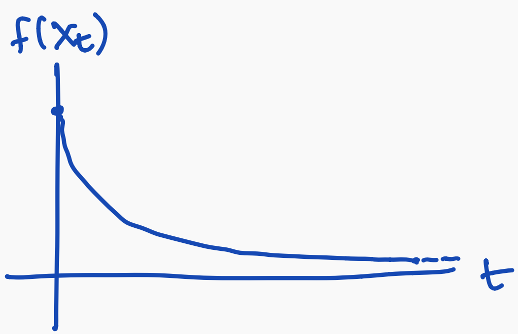
$$= -\|\nabla f(X_t)\|^2$$

$$\leq 0$$



$\nabla f(x)$ is direction of steepest ascent

$-\nabla f(x)$ is direction of steepest descent



Gradient flow

$$\dot{X}_t = -\nabla f(X_t)$$

- First-order dynamics
- Greedy:

$$-\nabla f(X_t) = \arg\min_{v \in \mathbb{R}^n} \left\{ \langle \nabla f(X_t), v \rangle + \frac{1}{2} \|v\|^2 \right\}$$

- Descent method:

$$\tfrac{d}{dt} f(X_t) = \langle \nabla f(X_t), \dot{X}_t \rangle = -\|\nabla f(X_t)\|^2 \leq 0$$

Example: Quadratic

$\mathcal{X} = \mathbb{R}^n$

Let $f(x) = \frac{1}{2}x^\top Ax$ for some $\overset{A^\top}{\overset{\shortparallel}{A}} \succeq 0$

$\nabla f(x) = Ax$

Gradient flow:

$$\dot{X}_t = -AX_t$$

$$\Rightarrow \quad X_t = \underbrace{e^{-At}}_{\text{matrix exponential}} X_0$$

$$e^{At} = I + tA + \frac{t^2}{2!}A^2 + \frac{t^3}{3!}A^3 + \dots$$

- for $n=1$: $X_t \in \mathbb{R}$, $A \geq 0$

$$\dot{X}_t = -AX_t \iff \frac{d}{dt}\log X_t = \frac{\dot{X}_t}{X_t} = -A \Rightarrow \log X_t = \log X_0 - At$$

$$\Rightarrow X_t = X_0 \cdot e^{-At}$$

note: in general, any $A \in \mathbb{R}^{n\times n}$ can be written

$$A = A_{sym} + A_{ant}$$

where $A_{sym} = A_{sym}^\top$

and $A_{ant} = -A_{ant}^\top$

and
$$f(x) = \frac{1}{2}x^\top Ax = \frac{1}{2}x^\top(A_{sym} + A_{ant})x = \frac{1}{2}x^\top A_{sym}x$$

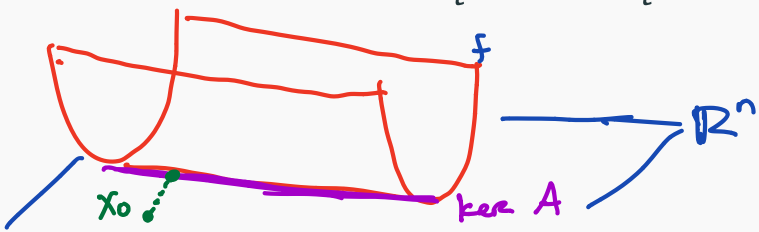
because $x^\top A_{ant} x = 0$

Example: Quadratic

Let $f(x) = \frac{1}{2}x^\top Ax$ for some $A \succeq 0$

Gradient flow:

$$\dot{X}_t = -AX_t \quad \Rightarrow \boxed{X_t = e^{-At}X_0}$$



If A has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_m > 0 = \lambda_{m+1} = \cdots = \lambda_n$, then

$$\|X_t - x^*\|^2 \leq e^{-2\lambda_m t}\|X_0 - x^*\|^2$$

where x^* is the projection of X_0 to the kernel of A

Hierarchy of Structures

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and $x^* = \arg\min_{x \in \mathbb{R}^n} f(x)$

1. f is α-strongly convex if

$\alpha > 0$

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \alpha \|x - y\|^2$$

$$\Leftrightarrow \quad \nabla^2 f(x) \succeq \alpha I$$

$$\Leftrightarrow \quad \forall\, v \in \mathbb{R}^n : \quad v^\top \nabla^2 f(x)\, v \geq v^\top (\alpha I) v = \alpha \|v\|^2$$

$\alpha = 0$: f is weakly convex

Hierarchy of Structures

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and $x^* = \arg\min_{x \in \mathbb{R}^n} f(x)$

1. f is α-strongly convex if

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \alpha \|x - y\|^2$$
$$\Leftrightarrow \quad \nabla^2 f(x) \succeq \alpha I$$

2. f is α-gradient dominated if

$$\|\nabla f(x)\|^2 \geq 2\alpha(f(x) - f(x^*))$$

(also known as Polyak-Łojaciewicz inequality)

Hierarchy of Structures

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and $x^* = \arg\min_{x \in \mathbb{R}^n} f(x)$

1. f is α-strongly convex if

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \alpha \|x - y\|^2$$
$$\Leftrightarrow \quad \nabla^2 f(x) \succeq \alpha I$$

2. f is α-gradient dominated if

$$\|\nabla f(x)\|^2 \geq 2\alpha(f(x) - f(x^*))$$

   (also known as Polyak-Łojaciewicz inequality)

3. f has α-sufficient growth if

$$f(x) - f(x^*) \geq \frac{\alpha}{2}\|x - x^*\|^2$$

Hierarchy of Structures

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and $x^* = \arg\min_{x \in \mathbb{R}^n} f(x)$

1. f is α-strongly convex if

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \alpha \|x - y\|^2$$
$$\Leftrightarrow \quad \nabla^2 f(x) \succeq \alpha I$$

2. f is α-gradient dominated if

$$\|\nabla f(x)\|^2 \geq 2\alpha(f(x) - f(x^*))$$

(also known as Polyak-Łojaciewicz inequality)

3. f has α-sufficient growth if

$$f(x) - f(x^*) \geq \frac{\alpha}{2} \|x - x^*\|^2$$

Theorem: (1) $\Rightarrow$ (2) $\Rightarrow$ (3)

Example: Quadratic

Let $f(x) = \frac{1}{2}x^\top Ax$ for some $A \succeq 0$

If A has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_m > 0 = \lambda_{m+1} = \cdots = \lambda_n$, then:

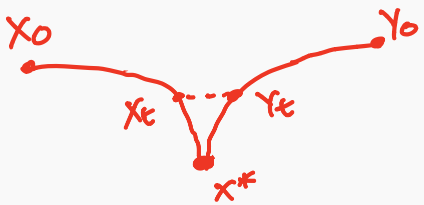
1. f is strongly convex with $\alpha = \lambda_n = 0$ $\qquad \alpha_{sc} = \lambda_{min}(\nabla^2 f(x)) = \lambda_{min}(A)$
2. f is gradient dominated with $\alpha = \lambda_m > 0$
3. f has sufficient growth with $\alpha = \lambda_m > 0$

Convergence Rates

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

Theorem

1. *If f is α-strongly convex, then gradient flow has exponential contraction: For $\dot{X}_t = -\nabla f(X_t)$, $\dot{Y}_t = -\nabla f(Y_t)$,*



$$\|X_t - Y_t\|^2 \leq e^{-2\alpha t}\|X_0 - Y_0\|^2$$

2. *If f is α-gradient dominated, then along gradient flow:*

$$\frac{\alpha}{2}\|X_t - x^*\|^2 \leq f(X_t) - f(x^*) \leq e^{-2\alpha t}(f(X_0) - f(x^*))$$

3. *If f is convex and has α-sufficient growth, along gradient flow:*

$$\|X_t - x^*\|^2 \leq e^{-\alpha t}\|X_0 - x^*\|^2$$

Proof

1. Consider $\left.\begin{aligned} \dot{X}_t &= -\nabla f(X_t) \\ \dot{Y}_t &= -\nabla f(Y_t) \end{aligned}\right\} (*)$

Compute : $\frac{d}{dt}\|X_t - Y_t\|^2 = 2\langle X_t - Y_t, \dot{X}_t - \dot{Y}_t\rangle$

$$\overset{(*)}{=} -2\langle X_t - Y_t, \nabla f(X_t) - \nabla f(Y_t)\rangle$$

$$\leq -2\alpha\|X_t - Y_t\|^2 \quad \text{by strong convexity}$$

Let $u_t = \|X_t - Y_t\|^2 \geq 0$

then $\dot{u}_t = \frac{d}{dt} u_t \leq -2\alpha u_t$

$\Leftrightarrow \frac{d}{dt}\log u_t = \frac{\dot{u}_t}{u_t} \leq -2\alpha$

$\Rightarrow \log u_t - \log u_0 \leq -2\alpha t$

Routine "Gronwall inequality"

$\Leftrightarrow u_t = \|X_t - Y_t\|^2 \leq u_0 \cdot e^{-2\alpha t} = \|X_0 - Y_0\|^2 \cdot e^{-2\alpha t}.$

2. Compute:

$$\frac{d}{dt}\Big(f(x_t) - f(x^*)\Big) = \langle \nabla f(x_t), \dot{x}_t \rangle$$

$$= -\|\nabla f(x_t)\|^2 \quad \text{since } \dot{x}_t = -\nabla f(x_t)$$

$$\leq -2\alpha\,(f(x_t) - f(x^*)) \quad \text{by grad-dominated}$$

then we are done (by Gronwall inequality):

$$f(x_t) - f(x^*) \leq e^{-2\alpha t}\Big(f(x_0) - f(x^*)\Big)$$

3. Compute

$$\frac{d}{dt}\|x_t - x^*\|^2 = 2\langle x_t - x^*, \dot{x}_t \rangle$$

$$= -2\langle x_t - x^*, \nabla f(x_t) \rangle$$

$$\leq -2\big(f(x_t) - f(x^*)\big) \quad \text{by convexity of } f$$

$$\leq -\alpha\,\|x_t - x^*\|^2$$

$\Rightarrow$ then $\|x_t - x^*\|^2 \leq e^{-\alpha t}\,\|x_0 - x^*\|^2$.

$\square$

Optimization references

- Boyd and Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004, available at `https://web.stanford.edu/~boyd/cvxbook`
- Nesterov, *Lectures on Convex Optimization*, Springer, 2004
- Nocedal and Wright, *Numerical Optimization*, Springer, 2006
- Vishnoi, *Algorithms for Convex Optimization*, available at `https://convex-optimization.github.io`

  CPSC 463: *Algorithms for Continuous Optimization*, Fall 2021
- Recht, *Optimization*, Big Data Bootcamp at Simons Institute, 2013, `https://simons.berkeley.edu/talks/ben-recht-2013-09-04`
- Boyd, *Convex Optimization*, Lecture videos for EE 364A at Stanford, 2008, `https://www.youtube.com/watch?v=McLq1hEq3UY`