Course Description

Standard Supervised classification setup:
- Data distribution: \((x, y) \in X \times Y \in \mathbb{R}^d \times \mathbb{R}\)
  \[(x, y) \sim \nu, \nu: \text{probability measure in } X \times Y\]
- Loss function \(L(\hat{y}, y)\), convex with respect to \(\hat{y}\).
  \[\text{e.g. } L(\hat{y}, y) = (\hat{y} - y)^2, \quad L(\hat{y}, y) = \log(1 + e^{-\hat{y}y})\]
- Model \(\hat{y} = \phi(x; \theta), \theta \in \Theta\)
- Empirical Risk Minimization (ERM) / Structural Risk Minimization:
  \[\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} L(\phi(x_i; \theta), y_i)\]
- Empirical loss (Risk):
  \[\hat{R}(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} L(\phi(x_i; \hat{\theta}), y_i) + R(\hat{\theta})\]

- Optimization:
  \[\text{min}_{\theta} \hat{R}(\theta)\]

Many questions:
- Choice of model? (approximation)
- Choice of optimization? (deep learning)
- Generalization (Statistics specific)

Part I: Geometry of data

- Input \(x\) in a high-dimensional space \(x \in \mathbb{R}^d\)
- But in many applications, \(x\) may be itself modeled as a function defined over a low-dimensional domain \(\mathcal{X}\)
  \[\text{e.g. } x = \phi(x_{1:n}), \quad \mathcal{X} \in \mathbb{R}^n\]
- Impact of this extra structure on \(X\)?
  \[\text{Part II: Geometry of Optimization and Learning}\]
- Previous questions do not concern any learning/optimization.
  \[\text{Basics of convex optimization and Newton acceleration}\]
- Non-convex optimization (or how to escape saddle points off)
- Continuous-time analysis (stochastic and deterministic)
Lecture 1: The curse of dimensionality

This generally refers to an exponential scaling of some cost/budget with respect to the input dimensionality of the problem.

- RKHS: Hypothesis class or a function space $\mathcal{F}$: $\{ f: \mathcal{X} \rightarrow \mathbb{R} \mid \mathcal{X} = \mathbb{R}^d \}$
  - Observe $n$ points $((x_1, f(x_1)), \ldots, (x_n, f(x_n)))$.
  - $x_i \in \mathcal{X}$.
  - $f$ unknown, assumed in $\mathcal{F}$.
  - Q: How does the risk of our best estimator $\hat{f}$ grow with $n, d$?

1st case: $\mathcal{F}$ contains linear functions $f(x) = \langle x, \theta \rangle + w$, $w \sim \mathcal{N}(0, \Theta^{-1})$

- $f(x) = \langle \theta, x \rangle$
- $L(\theta) = \frac{1}{n} \sum_{i=1}^{n} (w_i + \langle x_i, \theta \rangle - \hat{y}(x_i))^2$
- $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} (w_i + \langle x_i, \theta \rangle - \hat{y}(x_i))$.
- $\nabla L(\theta) = \frac{1}{n} \sum_{i=1}^{n} x_i (w_i + \langle x_i, \theta - \hat{\theta} \rangle)$.
- $\Delta \theta = \frac{\Delta}{\Delta x} \sum x_i \Delta x = \frac{1}{n} \sum x_i (w_i + \langle x_i, \theta - \hat{\theta} \rangle)$.
Lecture 2: Geometric Stability in Euclidean Domain
(Scattering Transform)

\[ \| f(x) - f(x') \| \leq \beta \| x-x' \| \]

\[ \sup \| \hat{f}(x) - f^*(x) \| \leq \epsilon \? \]

As \( n \to \infty \)

\[ f(x), x \in X \rightarrow \text{high dim spec} \]

\[ f(x) = 0 \left( a^T \phi(x) + b \right) \text{ in a good approximation} \]

If we think about \( f(x) \) as encoding a two-class classifier, the change of variable \( x \rightarrow \phi(x) \) becomes

\[ \text{no boundary} \]

\[ \text{In particular, } a^T (\phi(x) - \phi(x')) = 0 \]

\[ \exists \ f(x) = f(x') \]

\[ \text{(level sets of } f \text{ mapped to hyperplanes by } \phi) \]

\[ \text{(intra class variability becomes flat)} \]

\[ \text{Invariant & Symmetry, } x \in \Omega \text{ (input domain)} \]

By a global symmetry \( ff \) is open \( f: \text{Aut}(\Omega) \rightarrow \Omega \)

\[ f(x) = f(\phi(x)) \forall x \]

\[ \text{They can be absorbed by } \phi \text{ as:} \]

\[ \text{Invariants: } \phi(\phi(x)) = \phi(x) \forall x \]

\[ \text{Equi-invariant: } \phi(\phi(x)) = \phi(\phi(x)) \forall x \]

(\( \phi \) symmetries in image recognition problem?)
Translations: $x \in L^2(\mathbb{R}^2)$, $y = (3 - 3)$

Dilations: $\phi_y: 5 \in \mathbb{R}^d, \phi_y(x) = s^\frac{1}{2} x / s^\frac{1}{2}$

Rotations: $\phi_\theta: \theta \in (0, 2\pi), \phi_\theta(x) = X(R_\theta(x))$

These are subgroups of the group of rigid motions $\text{Aff}(\mathbb{R}^d)$, we can assume if with the space of affine transforms

$(\psi_1, \psi_2) \rightarrow (\psi_1, \psi_2)$

L. Not commutative in general.

In particular, these are Abelian groups (why?).

By Rotation/Translation:

Theorem (Stone, 30's): $L^2(\mathbb{R}^d)$ Hilbert space.

Self-adjoint operators on $L^2(\mathbb{R}^d)$ form the unitary groups of $\text{Aut}(L^2(\mathbb{R}^d))$.

In particular, given $a \in \mathbb{R}^d$, $\exists A$ a self-adjoint operator such that $\forall t, A_t = e^{itA}$

Recall the Fourier Transform: for $x \in L^2(\mathbb{R}^d)$ we define

$\hat{x}(\zeta) = \int x(t) e^{-2\pi i \zeta \cdot x} dt$

Main properties:

- (Linearity): $x = a x + p y \Rightarrow \hat{x} = a \hat{x} + p \hat{y}$
- (Inversion): $\hat{x}(t) = x^\dagger x \hat{x}(\zeta)$
- $\langle x, y \rangle = \langle \hat{x}, \hat{y} \rangle$
- (IFT): $X(t) = \hat{x}(\zeta)$
- (Fourier): $\hat{f} = (\hat{x} \cdot x) ; \hat{f}(\zeta) = e^{i\zeta \cdot x} \hat{x}(\zeta)$
- (DOF): $f = x(t) \hat{x}(\zeta)$

Stone Theorem and Fourier (Translation are diagonal operators in the Fourier domain. Stone-Thom formulates. Possible to be simultaneously diagonalizable, all translations. Because $A = \int \frac{1}{2\pi} \text{tr} (\hat{x} \cdot x) dt$ because $\hat{x} \cdot x$)

One-parameter unitary groups

Def. A one-parameter unitary group $\{ U_t : t \in \mathbb{R} \} \in \text{U}(X)$

(1) $U_t \circ U_s = U_{t+s}$

In particular, these are Abelian groups (why?).

By Rotation/Translation:

Theorem (Stone, 30's): $L^2(\mathbb{R}^d)$ Hilbert space.

Self-adjoint operators on $L^2(\mathbb{R}^d)$ form the unitary groups of $\text{Aut}(L^2(\mathbb{R}^d))$.

In particular, given $a \in \mathbb{R}^d$, $\exists A$ a self-adjoint operator such that $\forall t, A_t = e^{itA}$

Recall the Fourier Transform: for $x \in L^2(\mathbb{R}^d)$ we define

$\hat{x}(\zeta) = \int x(t) e^{-2\pi i \zeta \cdot x} dt$

Main properties:

- (Linearity): $x = a x + p y \Rightarrow \hat{x} = a \hat{x} + p \hat{y}$
- (Inversion): $\hat{x}(t) = x^\dagger x \hat{x}(\zeta)$
- $\langle x, y \rangle = \langle \hat{x}, \hat{y} \rangle$
- (IFT): $X(t) = \hat{x}(\zeta)$
- (Fourier): $\hat{f} = (\hat{x} \cdot x) ; \hat{f}(\zeta) = e^{i\zeta \cdot x} \hat{x}(\zeta)$
- (DOF): $f = x(t) \hat{x}(\zeta)$

Stone Theorem and Fourier (Translation are diagonal operators in the Fourier domain. Stone-Thom formulates. Possible to be simultaneously diagonalizable, all translations. Because $A = \int \frac{1}{2\pi} \text{tr} (\hat{x} \cdot x) dt$ because $\hat{x} \cdot x$)

One-parameter unitary groups

Def. A one-parameter unitary group $\{ U_t : t \in \mathbb{R} \} \in \text{U}(X)$

(1) $U_t \circ U_s = U_{t+s}$
Given a group of operators 

\[ G = G_1 \times \ldots \times G_p \]

and a group of operators \( \Phi \),

we consider the following questions:

- How to obtain invariants in this case?

Let \( \lambda = V^* \text{diag}(e^{it_1}, \ldots, e^{it_k}) V \)

\[ \chi = \chi(x, t) \]

\[ \phi(x) = |Vx| \]

is thus a \( G \)-invariant:

\[ \forall x, t \quad \phi(\Phi_t(x)) = \phi(x) \]

Thus, in the commutative case, a single \( \phi \) is sufficient to obtain invariance.

But, even in the commutative case, is this good enough?

- Symmetry is a very strict criterion. How to relax?

A symmetry groups in images are low-dimensional. Not too helpful!

- Are the Fourier invariants stable to deformations?

- We know \( \phi(\Phi_t(x)) = \phi(x) \) for \( \Phi \) a transformation. But what happens for transformations \( G \neq G \)?

- In some applications, we are interested in local invariance instead.

- We consider the local deformation cost

\[ \|f - \hat{f}\| = \sup_{x, t} \|Df(x, t)\|_\infty \]

- Shallow (frame) invariant are unstable.

- Deformation of the form \( (e^{-i}x) \)?
\((\hat{\phi} \cdot x)(3) = (1+e)^{-1} \times (1+e)^{-1/3}\)

\(\hat{x}(3) = \hat{A}(3-\theta_0)\)

\((\hat{\phi} \cdot x)(15) = (1+e)^{-1} \phi_{15}(11+\theta^{-1} (15) X)\)

If \((1+e)^{-1} \theta_0 = \theta_0 \Rightarrow \theta_0 (2+\varepsilon)\) then supports are disjoint.

\[ ||\hat{x}|| = M \cdot \hat{x}, ||x|| \sim ||\hat{x}||\]

Moreover, \((\hat{x})\text{ loses x-function in phase correlations.}\)

Fix 7 Local Invariants:

\[ \sqrt{x_1} = \sqrt{x(7/2)} \]

\[ ||\phi(x) - \phi(\hat{\phi} \cdot x)|| \leq C 2^{-3} ||x||\]

\[ \forall \nu, ||\phi(x) - \phi(\hat{\phi} \cdot x)|| \leq C 2^{-3} ||x||\]

L) smooth along the orbits \(\rho\): i.e. measure on \(\rho\): 

\[ \Phi(x) = 2^{-1/2} \int \phi(2^{-3/2} x d\nu) = (\int \phi(2^{-3/2} x d\nu = 1)\]

= \(\int \phi_{15}(x) x(1+\varepsilon) d\nu = x \cdot \phi_{15}(11, \phi_{15} = 2^{-3} d\phi(2^{-3} x)\)

and in fact we have smoothness beyond the orbit, on general definitions:

\(\Phi(x) = x \cdot \phi \) satisfies \( \forall \|x\| = 1 \in C \|x\| \text{ for integral operators} \)

\[ ||\kappa \Phi(\nu) = \int \kappa(x, \nu) \Phi(x) d\nu - 1 \leq C \max \left( \sup \int |\kappa(x, \nu)| d\nu, \sup \int |\kappa(x, \nu)| d\nu \right) \]

\[ \phi(x) = \phi_{15}(\hat{\phi} \cdot x) \]

\[ (\nu) = \int \phi_{15}(x) x(2^{-3} x d\nu)\]

\[ \Phi(x) = \phi_{15}(x), \phi_{15}(x) = \int \phi_{15}(x) \Phi(x) d\nu\]

Q: Other stable linear operators?

\[ \forall \nu, \phi(x) = \phi(\hat{\phi} \cdot x) \Rightarrow \phi(x) = \frac{1}{\mu(2)} \int \phi(\hat{\phi} \cdot x) d\nu\]

\[ \Phi(x) = \phi \left( \frac{1}{\mu(2)} \int \phi(\hat{\phi} \cdot x) d\nu \right) = \Phi \left( \frac{1}{\mu(2)} \int \phi(\hat{\phi} \cdot x) d\nu \right)\]
We just saw that the only linear invariant is invariance of high frequency content. Now we want to consider translation invariance. Recall that we want a (1) Stable (2) Highly frequency content. 

**Translation invariance** 

- We expect the function \( f(x) \) to be almost the same for every point \( x \). 
- In order to extract all the frequencies, too much 
- Let's consider \( f(x) = \sum_{n} c_n e^{inx} \) and \( f(x) = \sum_{n} c_n e^{inx} \) with \( c_n \)s.

**Translation invariance** 

- In order to extract all the frequencies, too much 
- Let's consider \( f(x) = \sum_{n} c_n e^{inx} \) and \( f(x) = \sum_{n} c_n e^{inx} \) with \( c_n \)s.

**Translation invariance** 

- In order to extract all the frequencies, too much 
- Let's consider \( f(x) = \sum_{n} c_n e^{inx} \) and \( f(x) = \sum_{n} c_n e^{inx} \) with \( c_n \)s.
Theorem: For the purpose of local stability, we need local stability.

\[ \| \phi - \phi \|_{C^1} \leq C \| \Delta \|_{L^2} \]

with \( C \in \mathbb{R}^+ \).

Key tool behind this result: Cotler-Stein near orthogonality of a family of bounded linear operators between the Hilbert spaces.

Let \( \mathcal{S} \) be the family of bounded linear operators between two Hilbert spaces \( E \) and \( F \), and \( \mathcal{T} \) be a weakly compact operator.

If \( \mathcal{T} \) is almost orthogonal, then \( \mathcal{S} \) converges, and

\[ \| \mathcal{S} \| \leq \sqrt{AB} \]

Intuition:

\[ \psi_k \xrightarrow{\text{Local}} 2 \]

(i) Local scales, \( \psi_k \) has small support, and for \( \Delta \) small within support of \( \psi_k \), and because \( \Delta \) is smooth.

\[ \| \Delta ( \psi_k ) \|_{L^2} \leq \frac{1}{2} \| \psi_k \|_{L^2} \]

Thus

\[ \left\| \left( \Delta \psi \right)_k (x) \right\| \leq \left\| \Delta \psi \right\|_{L^2} \]

\[ \| \left( \Delta \psi \right)_k (x) \|_{L^2} \leq \frac{1}{2} \| \psi_k \|_{L^2} \]

(ii) at large scales, \( \psi_k \) is itself smooth, thus

\[ \| \left( \Delta \psi \right)_k (x) \|_{L^2} \leq \frac{1}{2} \| \psi_k \|_{L^2} \]

(similar to our result on local stability of \( \psi_k \)).

(iii) Finally, different scales interact weakly, in the sense of the Cotler-Stein lemma.

This result says that deformations in the input domain are approximately mapped to deformations in the wavelet domain.

\[ \psi_k (x) \psi_k (y) \approx \psi_k (x \Delta y) \]

\[ \psi_k (x \Delta y) \xrightarrow{w} \psi_k (x) \psi_k (y) \]

So we can hope to extract stable measurements as we did in the input with the average. But we need a non-linear operator \( \mathcal{S} \) between (why?)

Characterizing stable non-linearity: \( M : \mathcal{C}^1 (\Omega) \rightarrow \mathcal{C}^1 (\Omega) \)

- Preserve additive stability for Littlewood-Paley:

\[ \| M (x - \mathcal{F} \psi_k ) \| \leq \| x - \mathcal{F} \psi_k \| \]

- Preserve geometric stability for wavelet scales: it is sufficient...
Theorem (8.12): If $M$ is non-expansive in $L^2$ such that $\hat{M}(\xi,\eta) = \hat{M}(\xi,\eta)$ for $\xi, \eta$ in $\mathbb{R}$, then $M$ is point-wise:

$$\hat{M}(x,\eta) = \hat{M}(x,\eta)$$

for $x, \eta$ in $\mathbb{R}$. Set:

$$f : \mathbb{R} \to \mathbb{R}$$

Idea of the proof: Let $G$ be the isotopy group of $\mathbb{R}$.

$$G(x) = \{ \phi \in Diff(S^1) : x = \phi \}$$

(symmetries of the level sets).

Observation: $G(x) = G(\phi(x)) \forall x$.

For instance, we choose the complex modulus:

$$\phi \in \mathbb{R} \Rightarrow \phi \in \{ \phi \in \mathbb{R} : \phi \neq x \}$$

The modulus of the complex number:

$$\left| x + iy \right| = \sqrt{x^2 + y^2}$$

$$\left| x - iy \right| = \sqrt{x^2 + y^2}$$

Even then we extract an invariant, we loose information.

Q: How to systematically recover the info that is lost?

A new wavelet decomposition captures lost high-frequency:

$$\sum_{n \in \mathbb{Z}} \left| x + iy_n \right| \phi_n(x)$$

Def: For each "path" $p = (x_1 \ldots x_n)$ with $x_i \in \mathbb{R}$ and $x$,

the scattering transform of $x$ is

$$\tilde{S}_x(p) = \sum_{n \in \mathbb{Z}} \left| x + iy_n \right| \phi_n(x)$$

\text{Scattering Helix and Energy conservation.}

For a given set of paths $\Gamma$, the Euclidean norm defined by scattering coefficients is:

$$\| S_{\phi}(T) x \|^2 = \sum_{p \in \Gamma} \| S_{\phi}(p) x \|^2$$

is it well defined? What is $\beta$? denote all paths of length $N$ and denote by:

$$\bar{S}_x x = \sum_{p \in \Gamma} \left| x + iy_n \right| \phi_n(x)$$

the "propagator" operator.

It is non-expansive thanks to the Littlewood Paley property:

$$\| \bar{S}_x x - \bar{S}_x x' \|^2 = \sum_{n \in \mathbb{Z}} \left| x + iy_n \right|^2$$

But observe that scattering coefficients are contracted by cascading:

$$\tilde{S}_x : \tilde{S}(p(x) = \left| x + iy_n \right| \phi_n(x)$$

for $p = (x_1 \ldots x_n)$
\[ P_2 : \text{set of all paths} \quad P = V \cup \{v\} \]

\[ A = \text{paths of length } n \]  

(Inductive case: \[ ||S_x - S_{x'}|| \leq ||x - x'|| \].)  

Apply (1) to each \[ ||x \cdot y_k||, ||x \cdot y_k'\].

Lecture 4: Scattering Extensions, Graph Neural Networks:

1. Consider new unitary matrix decompositions:

\[ \tag{1} \sum ||x \cdot y_k||^2 = ||x||^2 ||y_k||^2 + \sum ||x \cdot y_k||^2. \]

Then for appropriate unitaries, \[ \sum ||x||^2 ||y_k||^2 \]

By applying (1) on each output \[ x_p = \sum ||x \cdot y_k|| \]

we have

\[ \sum ||x||^2 = \sum ||S(x)||^2 + \sum ||x \cdot y_k||^2, \]

the result amounts to showing that

\[ \lim_{n \to \infty} \sum \frac{||x \cdot y_k||}{||x||^2} = 0. \]

Energy moves progressively towards low frequencies.

Decay is exponential for low limited signals.

\[ \text{Challenges:} \quad \text{Expanding,} \quad \text{Approximations,} \quad \text{Structures} \]
\[ \left\| \text{AV}^{-1} - \text{AV}^{-1}(v_{c}) \right\| + \left\| \text{AV}^{-1}(v_{c}) \right\| \leq C \| \varepsilon_{c} \| + \varepsilon_{KL} \| \varepsilon_{c} \|. \]

Limitation of Separable Scattering

This is the specific convolutional architecture given by
\[ S \times \{ \frac{S}{G} \times |p_{i} \in \text{Map} \}. \]

1. No feature dimensionality reduction: each new layer increases the number of feature maps.
2. Each wavelet band is assumed to be deformed independently.
3. No adaptivity.

Joint vs Separable Invariance

\[ G = S \times S \]

Each copy of \( S \) acts on a different coordinate (but is vertical).

Ex: Roto-translation group. \( G = \mathbb{R}^{2} \times S^{1} \)

\[ g \in G, \quad g = (v, x) \quad \text{(V, \ x)} \]

\[ g \times (u, w) = (Rg_{x}T_{v}u, w) \]
Non-commutative:
\[ g \cdot g' = R_{g'} R(g) R_{g'} + R_{g'} R(g) R_{g'} \]
This group acts on 1st-layer features similarly as translation acts on input images:
\[ f((g \cdot x + y)_{g'}) = x \cdot (R_{g'} w; y \cdot x + y) \]

Let \( G \) be a compact group equipped with the
Hausdorff measure\( d\mu \), acting on \( N \), and let \( \mu(G) \)

\[ x \ast_{G} b \ast_{G} 1 = \int_{G} x((g \cdot u) \cdot w(g)) d\mu(g) \]

6. Scattering with \( \ast \)-translation wavelets:
[Simic & Mallet '13] [Oyallon & Mallet '15].

Proposals: Give feedback to the week.
1) Inverse Curriculum: Demand care of absence.
2) Applications of Scattering:
- In computer vision: [B & Mallet '13] MMST/Texture.

Lecture 5:
From Scattering to CNNS:
The essential stability properties of scattering stem from
the \( \ast \)-translation properties:

\[ \| W \ast_{G} x \|_{2} \leq \| x \|_{2} \] and
\[ \| W \ast_{G} x \|_{2} \leq \| x \|_{2} \]

A CNNS can be written as:
\[ x^{(2)} = \rho \left[ \sum_{k \in K} D_{k} x^{(1)}_{k}, \theta, x^{(1)} \right] \]
From Euclidean to Non-Euclidean Stability

So far, we measured geometric stability in terms of differentials. If $\mathcal{L} \cong \mathcal{L}'$, in the sense that $\mathcal{L} = \mathcal{L}'$, then a rigid translation should satisfy

$$\| \mathcal{L} - \mathcal{L}' \| = 0.$$

Q: If $\mathcal{L}$ is not Euclidean, can it be continued?

We can think of $\mathcal{L}$ as a change of metric:

$$\mathcal{L}, \mathcal{L}' \xrightarrow{\psi} \mathcal{L}, \psi(\mathcal{L}') \mathcal{L}'$$

$$\langle x, x' \rangle_{\mathcal{L}'} = \int x(\psi(\mathcal{L}')) x'(\mathcal{L}') \rho(\mathcal{L}') d\mathcal{L}'$$

$$\langle x, x' \rangle_{\mathcal{L}'} = \int x(\psi(\mathcal{L}')) x(\psi(\mathcal{L}')) \rho(\mathcal{L}') d\mathcal{L}'$$

$$\Delta \rho(\mathcal{L}) = \int x(\mathcal{L}) x'(\mathcal{L}) |I - \nabla \psi(\mathcal{L})| \rho(\mathcal{L}) d\mathcal{L}$$

If $\mathcal{L} = L^2(\mathcal{X})$ is our data input space: square-integrable functions defined over a metric space $\mathcal{X}$.

$\Phi: L^2(\mathcal{X}) \rightarrow \mathbb{R}^K$ rigid representation.

$\mathcal{X} = (\mathcal{L}, \rho)$ distance, e.g. $\text{skew}$ travel times depend on $\rho$.

$$\mathcal{X}(t_0), \mathcal{X}(t_1)$$

$\text{dist}(\mathcal{X}(t_0), \mathcal{X}(t_1))$ measures how close the metrics are, e.g. between $\mathcal{X}$.

Stability to metric changes:

$$\| \Phi \mathcal{E} - \Phi \mathcal{X} \| \leq \text{dist}(\mathcal{X}, \mathcal{X})$$

Q: How to define meaningful distances? Graph or flexible data structure to describe discrete metric domain.

→ Diffusion Wavelets on Graphs:

$G = (V, E)$ set of nodes.

$E$: set of edges $(i,j), i,j \in V$.

he forms mostly an undirected weighted graph.

Let $(V) = n$, and $W \in \mathbb{R}^{n \times n}$ be a symmetric, weighted adjacency matrix, $D = \text{diag}(W)$. degree matrix.

We define the diffusion operator $A = D^{-1/2} W D^{-1/2}$

$A^{t}$ or $D^{t} W$ is a Markov Chain.

$A$ and $A^{t}$ are similar → they share the

Q: Will $A$ and $A^{t}$ be similar?
Def: Diffusion distance at time $s$ between two modes in $d^s_G(x,x') = \| A^s x - A^s x' \|_{T^s}$.

Def: Diffusion distance between $G$ and $G'$ at time $s$ is defined as

$$d^s_G(G,G') = \inf_{T^{s+} T^{-s}} \| \tilde{A}^{2s} - \tilde{T}^{s+} (A')^{2s} \tilde{T}^{-s} \|_{T^{s+} T^{-s}}$$

1. As $s$ increases, distance is weakened (why?)
2. Our distance results in a stronger topology than another usual distance defined over metric space, the Gromov-Hausdorff distance:

$$d_G(G,G') = \inf_{\text{map } f} \| d^s_g(x,x') - d^s_{G'}(f(x),f(x')) \|_{T^{s+} T^{-s}}$$

3. Here we used $|V| = |V'| = n$, but this can be easily extended to varying size using transportation plans rather than permutations.

**G modal Scattering**

- We saw before that scattering transforms in the Euclidean domain are stable to deformations (= metric change).
- The above notion of metric stability extends to diffusion domains without any additional structure.
- We can extend that to non-Euclidean settings.

4. We have defined the equivalent of a deformation (= diffusion metric).

5. We need the equivalent of wavelets.

6. In fact, we can also use diffusions on the graph to define a wavelet decomposition.

Def: diffusion wavelets are obtained by using powers of the adjacency

$$A_0 = I - A, \quad A_j = A^{2^{-j}} (I - A^{2^{-j}}) \quad j \geq 0$$

7. $A$ is a low-pass filter, we are building multiresolution filter banks by combining diffusions at different time-scales.

- $\hat{A}^k$ represents $A^k$.

**Def:** Diffusion scattering transform $\Psi$:

$$L^2(G) \rightarrow (L^2(G))^\infty$$

$$x \mapsto \{ \Psi_j(x) \}$$

$$\Psi_j(x) = \frac{1}{\sqrt{\lambda_j}} A_j x, \quad A_j p(Wx), A_j p(W p(Wx)), \ldots$$
If \( f \) is differentiable, then
\[
\frac{f(tx + (1-t)y) - f(y)}{t-0} \leq f(x) - f(y) + \langle \nabla f(y), x - y \rangle.
\]
In general, we can extend this notion to non-differentiable functions, using the notion of sub-gradients.

**Definition**: If \( f : \mathbb{R}^m \to \mathbb{R} \) and \( g \in \mathbb{R}^d \) is a subgradient of \( f \) at \( x \) if, for all \( y \),
\[
f(x) - f(y) \leq \langle g, x - y \rangle.
\]
Fact: If \( f \) is convex, then \( \partial f(x) \) is non-empty. If \( f \) is differentiable and convex, then \( \partial f(x) = \text{range} D f(x) \).

**Gradient Descent for smooth functions**

- If \( f \) is continuously differentiable function is \( \beta \)-smooth if \( \forall x \), \( \| D^2 f(x) \| \leq \beta I \).

**Theorem**: If \( f \) is convex and \( \beta \)-smooth in \( \mathbb{R}^d \), then gradient descent

\[
x_{t+1} = x_t - \eta D f(x_t)
\]
with \( \eta \leq \beta^{-1} \) satisfies
\[
f(x_t) - \min_{x \in \mathbb{R}^d} f(x) \leq \frac{\eta \beta}{2} \| x_t - x^* \|^2.
\]
Lemma: For $f$ convex and $\beta$-smooth, we have

\[
\begin{align*}
    f(x) - f(y) - \langle \nabla f(y), x - y \rangle &\leq \frac{\beta}{2} ||x - y||^2 \\
    f(x) - f(y) = \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle \, dt \\
    f(x) - f(y) - \langle \nabla f(y), x - y \rangle &\leq \int_0^1 \langle \nabla f(y + t(x - y) - \nabla f(y)), x - y \rangle \, dt \\
    &\leq \int_0^1 \beta ||x - y|| \, ||\nabla f(y + t(x - y) - \nabla f(y))|| \, dt \\
    &\leq \int_0^1 \beta ||x - y||^2 \, dt = \frac{\beta}{2} ||x - y||^2.
\end{align*}
\]

Proof: By convexity, since $\nabla f(x) \in \partial f(x)$, we have

\[
0 \leq f(x) - f(y) - \langle \nabla f(y), x - y \rangle \leq \frac{\beta}{2} ||x - y||^2.
\]

Choosing $y = x - \frac{1}{\beta} \nabla f(x)$, we obtain

\[
0 \leq f(x) - f(y) - \langle \nabla f(x), y - x \rangle \leq \frac{\beta}{2} ||x - y||^2.
\]
using previous lemma, we have
\[ \frac{1}{\beta} \left\| Df(x) - Df(y) \right\| \leq \left\langle Df(x) - Df(y), x - y \right\rangle \]
Then
\[ \left\| x_{k+1} - x^* \right\| \leq \left\| x_k - x^* \right\| + \frac{1}{\beta} \left\| Df(x) - Df(x^*) \right\| \]
\[ \leq \left\| x_k - x^* \right\| + \frac{1}{\beta} \left\| Df(x) \right\| \]
\[ \leq \left\| x_k - x^* \right\| + \frac{1}{\beta} \left\| Df(x) \right\| \]
\[ \leq \left\| x_k - x^* \right\| \]

Rate 4 is not optimal in the class of $\beta$-smooth $\alpha$-strongly convex functions (we will see this later).
But first, let's push smoothness a bit further:

**def:** $f$ is $\alpha$-strongly convex if
\[ f(x) - f(y) \leq \left\langle Df(x), x - y \right\rangle - \frac{\alpha}{2} \left\| x - y \right\|^2 \]
If $f$ is twice differentiable, $D^2 f(x) \succ 0$.

Strong convexity and smoothness allows us to sandwich $f(y)$ between two quadratic forms:
\[ q^-_x(y) = f(x) + \left\langle Df(x), y - x \right\rangle + \frac{\alpha}{2} \left\| y - x \right\|^2 \]
\[ q^+_x(y) = f(x) + \left\langle Df(x), y - x \right\rangle + \frac{\alpha}{2} \left\| y - x \right\|^2 \]
\[ q^-_x(y) \leq f(y) \leq q^+_x(y) \]

Denote $k = \frac{\alpha}{\beta} (\beta + 1)$ the "condition" number of $f$.

**lemma:** $f$ is smooth, $\alpha$-strongly convex. Then
\[ \left\langle Df(x) - Df(y), x - y \right\rangle \geq \frac{\alpha^2}{\beta \alpha} \left\| x - y \right\|^2 + \frac{1}{\beta} \left\| Df(x) - Df(y) \right\|^2 \]
Recall that if $f$ is convex and $\beta$-smooth, then
\[ \left\langle Df(x) - Df(y), x - y \right\rangle \geq \frac{\beta}{\beta + 1} \left\| Df(x) - Df(y) \right\|^2 \]
If $f$ is $\beta$-smooth, $\alpha$-strongly convex, then
\[ f(x) = f(x^*) + \left( \frac{\alpha}{2} \right) \left\| x - x^* \right\|^2 + \frac{\beta}{\beta + 1} \left\| Df(x) - Df(x^*) \right\|^2 \]

**Theorem:** $f$ is smooth and $\alpha$-strongly convex, $\frac{\alpha}{\beta} \frac{x}{\gamma} \geq \frac{1}{\alpha}$ gives
\[ f(x_{k+1}) - f(x) \leq \frac{\beta}{2} \exp \left( - \frac{4\gamma t}{\beta + 1} \right) \left\| x_{k+1} - x^* \right\|^2 \]

By smoothness, $f(x_k) - f(x^*) \leq \frac{\beta}{2} \left\| x_k - x^* \right\|^2$.
Oracle lower bounds:

1) Smooth case: there exists $\beta$-smooth convex function $f$ st for any black-box procedure with $x_{i+1} = \text{opt}(g_i)$, have $f(x) - f(x^*) \geq \frac{\beta}{2} \| x - x^* \|^2$ ($\beta > 0$).

Remark: function is a Dirichlet Energy (quadratic).

2) Smooth + strongly convex case: $f$ is $\beta$-smooth and $\alpha$-strongly convex st for any $t$,

$$f(x_t) - f(x^*) \geq \frac{1}{2} \left( \sqrt{\frac{1}{1+t}} \right) \| x(t) - x^* \|^2$$

Gap: How to close it?

3) Quadratic case: $f(x) = \frac{1}{2} (x - x^*)^T H (x - x^*)$.

$$\nabla^2 f(x) = H,$$ so we assume $\alpha I \leq H \leq \beta I$.

$$x_{t+1} = x_t - \gamma \nabla f(x_t) = H(x_t - x^*)$$

$$= x_t - \gamma H (x_t - x^*) = \left( I - \gamma H \right)x_t + \gamma \frac{H x^*}{\|x^*\|}$$

Suppose $x = 0$;

$$x_{t+1} = \left( \sum_{k=0}^t \left[ I - \gamma H \right] \right) x^*$$

for $\alpha I \leq H \leq \beta I \Rightarrow (1 - \gamma H) \leq I - \gamma H \leq (1 - \gamma \alpha) I$.

$$\frac{\sum_{k=0}^t \left[ I - \gamma H \right] x^*}{\sqrt{t}} \leq \frac{1}{\sqrt{t}}$$

4) Discrete vs continuous time optimization.

Consider $x_{t+1} = x_t - \gamma \nabla f(x_t)$.

$\| x - x^* \| = \mathcal{O}\left( \frac{\|x^*\|}{\gamma \sqrt{t}} \right) = \mathcal{O}\left( \frac{1}{\gamma \sqrt{t}} \right)$.

5) Gradient Flow.

Gradient Descent is a forward Euler discretization.

$$x_{t+1} = H x_t + \nabla f(x_t)$$

$$\frac{d}{dt} x(t) = -\nabla f(x_t)$$

Then $x(t) \rightarrow x^*$. The path is well-defined for $t 

x(0) = x_0$.

$$x(t) = H t x_0$$

we can get a first understanding of optimization algorithms by first looking at their behavior for \( p = 0 \) continuous time.

Example:

\[
\begin{align*}
\mathbf{x}(t) &= \frac{1}{2} \mathbf{x}^T \mathbf{H} \mathbf{x} \\
\mathbf{H} &\in \text{pd} : \mathbf{H} > 0.5 \mathbf{I} \\
\mathbf{x}(t) &= -\mathbf{H} \mathbf{x}(t) \\
\|\mathbf{x}(t)\| &\leq \|\mathbf{x}(0)\| e^{-t} \\
\end{align*}
\]

Example:

\( \mathbf{f}(\mathbf{x}) \) convex, smooth.

\[
\mathbf{x}(t) = -\nabla \mathbf{f}(\mathbf{x}(t)).
\]

Recall the convergence rate we obtained:

\[
\mathbf{f}(\mathbf{x}) \leq \frac{\beta C}{\mathbf{K}} \| \mathbf{x} - \mathbf{x}^* \|^p = \frac{C}{\mathbf{K}^{1/p}} \| \mathbf{x} - \mathbf{x}^* \|^p.
\]

Can we use continuous-time to derive it much faster?

Consider the Lyapunov function:

\[
\begin{align*}
L(t) &= \frac{1}{2} \| \mathbf{f}(\mathbf{x}(t)) - \mathbf{f}(\mathbf{x}) \|^2 + \frac{1}{3} \| \mathbf{x}(t) - \mathbf{x}^* \|^3 \\
L(t) &= \mathbf{f}(\mathbf{x}(t)) - \mathbf{f}(\mathbf{x}) + \beta (\nabla \mathbf{f}(\mathbf{x}(t)), \mathbf{x}(t) - \mathbf{x}^*) \\
&\quad + \left( \nabla \mathbf{f}(\mathbf{x}(t)), \mathbf{x}(t) - \mathbf{x}^* \right) \\
&= -\| \nabla \mathbf{f}(\mathbf{x}(t)) \|^2 + \frac{\mathbf{f}(\mathbf{x}(t)) - \mathbf{f}(\mathbf{x}) + \left( \nabla \mathbf{f}(\mathbf{x}(t)), \mathbf{x}(t) - \mathbf{x}^* \right)}{\mathbf{K}} \\
&\quad + \frac{\beta C}{\mathbf{K}^{1/p}} \| \mathbf{x}(t) - \mathbf{x}^* \|^p.
\end{align*}
\]

\[
L(t) \leq L(0) = \frac{1}{2} \| \mathbf{x}(0) - \mathbf{x}^* \|^2
\]

\[
\mathcal{E}(\mathbf{f}(\mathbf{x}(t)) - \mathbf{f}(\mathbf{x})) = \mathbf{f}(\mathbf{x}(t)) - \mathbf{f}(\mathbf{x}) \leq \frac{\beta C}{\mathbf{K}^{1/p}} \| \mathbf{x}(0) - \mathbf{x}^* \|^p.
\]

Q: How good is gradient descent for the class of (smooth) convex functions?

A reasonable block-box procedure in a mapping from the past history to the next point:

\[
(x_0, g_1, g_2, \ldots, x_0, g_t) \rightarrow x_{t+1}
\]

We consider the "linear" case, where we ask

\[
x_{t+1} \in \text{span} \{ g_1, \ldots, g_t \}
\]

Q: How to close this gap?

A: Through the notion of acceleration.

Consider:

\[
\mathbf{x}_{t+1} = \mathbf{y}_t - \frac{1}{1+k} \nabla \mathbf{f}(\mathbf{y}_t)
\]

\[
x_{t+1} = \mathbf{x}_{t+1} + \frac{1}{k+1} (\mathbf{x}_{t+1} - \mathbf{x}_t)
\]

Theorem (Nesterov '03):

\[
\mathbf{f}(\mathbf{x}) - \mathbf{f}^* \leq \frac{2 \| \mathbf{x}_0 - \mathbf{x}_* \|^2}{(\mathbf{f}(\mathbf{x}) - \mathbf{f}^*))^{-1}}
\]
Accelerations for \( \alpha \)-strongly convex, \( \beta \)-smooth functions just need to slightly modify scheme:

\[
X_{k+1} = X_k - \frac{\gamma}{\beta} \nabla f(X_k)
\]

\[
X_{k+1} = X_k + \frac{\alpha}{\beta} (X_k - X_{k-1})
\]

**Theorem (Matrix):**

\[
f(x_k) - f(x) \leq \beta \left(1 - \sqrt{\frac{\alpha}{\beta}}\right)^k \| x_k - x \|^2
\]

\[
\sim \left(1 - \sqrt{\frac{\alpha}{\beta}}\right)^k \text{ vs } \left(1 - \kappa\right)^k \text{ from GP}
\]

1. Continuous-time interpretation?

2. What is the corresponding ODE?

\[
\dot{x} + \frac{3}{\epsilon} x + \nabla f = 0
\]

**Theorem:** [Su, Boyd, Candes, '15] The ODE has unique global solution \( x \in C^2([0,\infty), \mathbb{R}^n) \cap C^1([0,\infty), \mathbb{R}^n) \).

Moreover, as \( \gamma \to 0 \), Newton's scheme converges to the ODE for all fixed \( T > 0 \),

\[
\lim_{\gamma \to 0} \sup_{k \leq T} \| x_k - X_k(t) \| = 0
\]

**Simple Consequence:** We show easily that we get \( 1/t \)-convergence rate.

**Theorem:** \((\beta, \alpha, \epsilon)\): Consider now

\[
L(t) = t^2 \left( f(X) - f^* \right) + 2 \| x + t\dot{x} - x^* \|^2
\]

\[
\dot{L}(t) = 2t \left( f(X(t)) - f^* \right) + \epsilon t \left( \langle \nabla f(x), \dot{x} \rangle + \langle x, \dot{x} \rangle \right) - \frac{3}{\epsilon} \langle \dot{x} \rangle^2
\]

Since \( \dot{x} + \frac{3}{\epsilon} x = -t \nabla f \),

\[
\frac{3}{\epsilon} \dot{x} = -\frac{1}{\epsilon} \dot{x}
\]

\[
L(t) = 2t \left( f(X(t)) - f^* \right) + \epsilon t \langle \nabla f(x), \dot{x} \rangle - 2t \langle \nabla f(x), x - \frac{1}{2} \dot{x} \rangle
\]

\[
= 2t \left( f(X(t)) - f^* \right) - 2t \langle \nabla f(x), x - x^* \rangle \leq 0
\]

1. Interpretation in the strongly convex case: Clearly, show polynomials:

Recall we use approximately the inverse matrix \( (\gamma I)^{-1} \) with a finite power series \( \sum (I - \gamma I)^k \).

\[
\| \gamma I - \sum (I - \gamma I)^k \| \to 0 \text{ as } k \to \infty, \text{ with error } O((1 - \gamma I)^k)
\]

2. What is the best polynomial of degree \( k \) to approximate \( A^{-1} \)? Holy Grail? \( (A, (A - I)) \)?
Lemma: (Chebyshev polynomial): There exist a polynomial $q_c$ of degree $O\left(\sqrt{\log \frac{1}{\epsilon}}\right)$ such that

$$q_{c\ell}(0) = 1 \quad \text{and} \quad 1 - q_{c\ell}(x) \leq \epsilon \quad \forall x \in [a, b].$$

$q_{c\ell}$ is a Chebyshev polynomial, computed recursively, from $q_{c\ell-1}$, $q_{c\ell-1}$ \to Nesterov's Scheme.

To further explain Chebyshev:

$$\min || A' - q_{c\ell}(A) || \leq \min || I - A q_{c\ell}(A) ||$$

- $P_{c\ell}(A)$ commutes with $A$;

\[ \lambda \text{eigenvalue of } A \implies P_{c\ell}(A) = 1 - \lambda q_{c\ell}(A) \text{ eigenvalue of } P_{c\ell}(A). \]

- Eigenvalues of $A$ are in $\lambda \in [a, b]$.

$$\min \{ P_{c\ell}(x) \} \geq 1 \quad \implies \text{Chebyshev polynomials.}$$

Let polynomial $p_{c\ell}$ of degree $O\left(\sqrt{\log \frac{1}{\epsilon}}\right)$ such that

$$p_{c\ell}(0) = 1 \quad \text{and} \quad 1 - p_{c\ell}(x) \leq \epsilon \quad \forall x \in [a, b].$$

Today: Further on discrete vs continuous time.

- Stochastic Gradient Descent,

1) Previously: take a discrete scheme and derive a continuous limit.
2) $Q$: Can we take opposite route?
3) Given convex function $g$, we need two steps:

- $\text{constant } \epsilon > 0$ such that its solution $x(t)$ satisfies:

\[ g(x(t)) - g(x_\epsilon) \leq \epsilon^p \]

- Discretize the ODE such that convergence rate is preserved.

A note on discretization of ODEs:

\[ \frac{dy}{dt} = f(t, y) \quad \text{integrate from } t_n \text{ to } t_{n+1} \]

\[ y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} f(t, y(t)) \, dt \]

Two basic strategies:

(i) \( \int_{t_n}^{t_{n+1}} f(t, y(t)) \, dt \approx hf(t_n, y(t_n)) \), yields

\[ y_{n+1} = y_n + h f(t_n, y(t_n)) \quad \text{Explicit Euler Scheme.} \]

(ii) \( \int_{t_n}^{t_{n+1}} f(t, y(t)) \, dt \approx hf(t_{n+1}, y(t_{n+1})) \) thus

\[ y_{n+1} = y_n + h f(t_{n+1}, y(t_{n+1})) \quad \text{Implicit Euler Scheme.} \]
The Bregman Lagrangian \cite{Wisbou et al. 16}.

1. We assume $g$ minimizes $g(\theta)$ admits a unique minimum $\theta^*$.

2. Let $h$ be a convex function.

3. A Bregman divergence in $\Theta$ is
   \[ D_h(\theta, y) = h(\theta) - h(y) - \langle \nabla h(y), \theta - y \rangle \]

   non-negative since $h$ is convex, $D_h(\theta, 0) = 0$; but not symmetric.

   Locally like a Hessian metric:
   \[ D_h(\theta, y) = \frac{1}{2} (\theta - y) \nabla^2 h(y) (\theta - y) + o(\|\theta - y\|^2) \]  

4. Bregman Lagrangian is
   \[ \mathcal{L}(X, V, t) = e^{\alpha(t)} V^T (D_h(X + e^{-\alpha(t)} V, X) - e^{\beta(t)} g(X)) \]

   $X$ position

   $V$ velocity, with $\beta(t) \leq \alpha(t)$

   $e^{\beta(t)} \leq e^{\alpha(t)}$

   $X(t) \in \Theta$, $t \geq 0$.

   Action on the path is
   \[ J(X) = \int_{t_0}^{t_f} \mathcal{L}(X_t, \dot{X}_t, t) \, dt \]

   Lagrangian system:

   \[ \dot{X}_t + p(t) \frac{\alpha(t)}{t} X_t + C \alpha(t) p(t)^2 \left[ \nabla^2 h(X_t + e^{\alpha(t)} V) \right]^{1/3} = 0 \]

   Satisfy the Euler-Lagrange equation:
   \[ \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{X}} = \frac{\partial \mathcal{L}}{\partial X} \]

   by plugging the Bregman Lagrangian $\mathcal{L}$ becomes
   \[ \frac{d}{dt} D_h (X_t + e^{-\alpha(t)} \dot{X}_t) = -e^{\alpha(t)} + \beta(t) V(t) g(X_t) \]

   Q: Do these solutions minimize $g$? How fast?

   Theorem \cite{Wisbou et al. 16}: The solutions $X_t$ to the Euler-Lagrange satisfy $g(X_t) - g(\theta^*) \leq \mathcal{O}(e^{-\beta(t)})$

   5. Consider the Lyapunov function
   \[ \mathcal{E}_t = D_h (\theta^*, X_t + e^{-\alpha(t)} \dot{X}_t) + e^{\beta(t)} [g(X_t) - g(\theta^*)] \]

   Lyapunov Show as before that $\dot{\mathcal{E}}(t) \leq 0$.

   Optimal rate is achieved with the solution
   \[ \beta(t) = e^{\alpha(t)} \]

   Q: How to discretize these ODEs while preserving rate?

   \[ \alpha(t) = \log k - \log t, \quad \beta(t) = p \log t + \log C, \quad \gamma(t) = p \log t \]

   Resulting Euler-Lagrange becomes
   \[ \dot{X}_t + p \frac{\alpha(t)}{t} X_t + C \alpha(t) p(t)^2 \left[ \nabla^2 h(X_t + e^{\alpha(t)} V) \right]^{1/3} = 0 \]

   $C > 0$, $p > 0$. 
This is equivalent to a decoupled system of 1st order equations:

\[ Z_t = X_t + \frac{\epsilon}{\tau} X_t \]

\[ \frac{d}{dt} V(t) = -C P D^2 g(X_t) \]

A "naive" idea is to use a standard discretization scheme for (8), e.g., forward-backward Euler.

\[ \tilde{V}_k = \arg \min \left\{ C P K^{-1} \left\{ \tilde{V}(y_k) + \frac{K+1}{K} \delta \right\} \right\} \]

\[ \frac{\theta_{k+1}}{\theta_k} = \frac{\pi_k}{\pi_k + \frac{K}{K+1} \theta_k} \]

It is not stable, does not preserve the convergence rate.

As it turns out, in order to recover matching convergence, we need to recover some form of coercivity properly.

\[ \frac{\theta_{k+1}}{\theta_k} = \frac{\pi_k}{\pi_k + \frac{K}{K+1} \theta_k} \]

\[ \frac{\theta_{k+1}}{\theta_k} = \frac{\pi_k}{\pi_k + \frac{K}{K+1} \theta_k} \]

Stochastic Gradient Descent

- As before, consider the minimization of function \( f \) defined in \( \mathbb{R}^d \).

- However, we no longer have access to \( V \) directly, only given access to unbiased estimates \( \tilde{V}_k (\theta_k) \).

Ex: \( g_n \) is the loss for a single data point.

\[ g_n (\theta) = l ( \phi (x_n; \theta), y_n ) \]

\[ q (\theta) = \frac{1}{E(x_n) \delta P} l ( \phi (x; \theta), y ) \]

Stochastic Approximation (Robbins & Monro, '51)

General setting: find the zeros of a function \( h : \mathbb{R}^d \rightarrow \mathbb{R}^d \) from random observations at certain points.

- If \( h = 0 \) we have to stochastic optimization.

Robbins & Monro: \( \theta_n = \theta_{n-1} - \frac{\epsilon_n}{\delta_n} ( h ( \theta_{n-1} ) + \epsilon_n ) \)

Questions:

- Conditions for convergence?

- What is \( h \)?
Intuitively, the example: estimate the mean from samples.

Starting from \( \theta_0 = 0 \), we get data \( x_n \in \mathbb{R}^d \)

\[
\theta_n = (1 - x_n) \theta_{n-1} + x_n x_n = \theta_{n-1} - x_n (\theta_{n-1} - x_n)
\]

(Ex: \( x_n = \frac{x_n}{n} \to \theta_n = \frac{1}{n} \sum \limits_{i \in n} x_i \), \( \theta_n = \frac{1}{n(n+1)} \sum \limits_{i \in s_n} k x_i \).

\( \to \) If \( x_n \) are iid with \( \mathbb{E} x_n = x \), \( \mathbb{E} \| x_n - x \|^2 = \sigma^2 \)

\[
\theta_n - x = \Pi \left( 1 - x_n \right) (\theta_{n-1} - x) + \sum \limits_{i \in n} \Pi \left( 1 - x_n \right) x_i (x_n - x)
\]

\[
\mathbb{E} \| \theta_n - x \|^2 = \Pi \left( 1 - x_n \right)^2 \mathbb{E} \| x_n - x \|^2 + \sigma^2 \sum \limits_{i \in n} \mathbb{E} \| x_n - x \|^2 
\]

error has two contributions:

l) Error Initial conditions: \( \Pi \left( 1 - x_n \right) \to 0 \) as \( n \to \infty \).

\( \Pi \) we want to forget them.

l) Robustness to noise: \( \sum \limits_{i \in n} x_i \Pi \left( 1 - x_n \right) \to 0 \)

\( \sum x_i \to 0 \), \( \Pi \left( 1 - x_n \right) \to 0 \)

\( \to \) If \( x_n \to 0 \), \( \log \Pi \left( 1 - x_n \right) \to -\sum \limits_{i \in n} x_i \to 0 \)

\( \sum x_i \) should diverge \( \to \infty \) in order to \( \Pi \left( 1 - x_n \right) \to 0 \).

Ex: \( \theta_n = C / n \to 0 \)

\[
\sum \limits_{i \in n} x_i = \log n + C + O(n^{-\alpha})
\]

\( \alpha > 1 \) \( \sum \limits_{i \in n} x_i - C + O(n^{-\alpha}) \to \) init conditions not forgotten.

\( \alpha < 0 \) \( \sum \limits_{i \in n} x_i - C n^{1 - \alpha} + O(1) \).

Noise Term: assume \( x_n \) non increasing and \( \sigma < 1 \).

Then \( \forall x_n \), \( \sum \limits_{i \in n} \Pi \left( 1 - x_n \right)^2 = \sum \limits_{i \in n} \Pi \left( 1 - x_n \right) \)

Suppose \( \mu < 2 \). Then

\[
\sum \limits_{i \in n} \mathbb{E} \| x_n - x \|^2 \Pi \left( 1 - x_n \right) = \sum \limits_{i \in n} \mathbb{E} \| x_n - x \|^2 \Pi \left( 1 - x_n \right)
\]

\[
\sum \limits_{i \in n} \mathbb{E} \| x_n - x \|^2 \Pi \left( 1 - x_n \right) \to \Pi \left( 1 - x_n \right) \sum \limits_{i \in n} \mathbb{E} \| x_n - x \|^2 \Pi \left( 1 - x_n \right)
\]

\( \exp ( -\mu \sum \limits_{i \in n} x_i ) \sum \limits_{i \in n} x_i \to 0 \) as \( n \to \infty \).

So we need \( x_n \) to go to 0.

Ex: \( x_n = C / n \) noise term converges in \( O(n^{-\alpha}) \).
Convergence of stochastic optimization

- When $g$ is convex, we want to study how $\theta_n \to \theta^*$ (convergence in iterates) or $g(\theta_n) \to g(\theta^*)$ (convergence in value).

- Several convergence criteria at the stochastic case:
  - A.S. convergence: $\Pr \left( g(\theta_n) \to g(\theta^*) \right) = 1$
  - Convergence in proba: $\forall \varepsilon > 0$, $\Pr \left( |g(\theta_n) - g(\theta^*)| < \varepsilon \right) \to 1$
  - Convergence in moments: $\mathbb{E} |g(\theta_n) - g(\theta^*)|^p \to 0$.

- R.A.M asymptotic normality [Fabian, 98]

$$X_n := C_n^{-1}$$

$$\mathbb{E} (X_n - \theta)(X_n - \theta)^T \leq \kappa^{-2} C_n (X - \theta)(X - \theta)^T + n^{-1} C_n^2 (2C_n - I)^T \Sigma$$

$$A = D_n g(\theta_n) \Sigma = \mathbb{E} (X_n e_n)^T$$

$$e_n = D_n g(\theta_n) e_n$$

- $\kappa = \min(A)^{-1}$ for convergence.

- $C$ too small: no convergence (due to memory of not convex).

- For large $N$ cases, convergence.

Polyak-Ruppert Averaging

Promotes R.M also suffers from sensitivity of step-size, and dependence on unknown conditioning of the problem. We modify bias-variance tradeoff by considering averaging over the iterates:

$$\bar{\theta}_n = \frac{1}{n} \sum_{k \leq n} \theta_k \quad (\bar{\theta}_n = \frac{1}{n} \sum_{k \leq n} \theta_k)$$

Theorem (Cesaro's) Suppose $\theta_n \to \theta$ with rate $\| \theta_n - \theta \| \leq \frac{1}{\sqrt{n}}$.

Then $\bar{\theta}_n \to \theta$ with rate $\frac{\sigma}{\sqrt{n}}$.

One can show that if $\Sigma_n < \infty$, then the rate is always $\frac{\sigma}{\sqrt{n}}$.

- We lose convergence speed, but regain robustness. Several convergence rates for the different assumptions a good choice of step-size.

Global minimax rates E. Nemirovsky '83, A. Agarwal '12; Bach '17

- Strongly convex case: $O(\|\theta_n\|^{-1})$

- Non-strongly convex: $O(\|\theta_n\|^{-\frac{1}{2}})$

- Both attained with averaged stochastic gradient descent.

Take-home: In smooth problems, $\eta_n = n^{-\frac{1}{2}}$ + averaging gives adaptivity to strong convexity.

[Bach, Moulines, 14]
However, $\theta_n$ does not converge to $\theta$; it oscillates around it, with oscillations of order $1/\sqrt{n}$.

**Ergodic Theorem:** $\theta_n \to \theta_\infty$ at rate $O(1/n)$ (CLT):

$$\theta_n = \theta_{n-1} - \gamma Dg_n(\theta_{n-1})$$

**Stationary distribution** $P_\theta$

$$\int_\Theta T_\theta(d\theta) = \int_\Theta T_{\theta_\infty}(d\theta) - \gamma \int Dg(\theta) T_{\theta_\infty}(d\theta)$$

$$= \gamma \int Dg(\theta) P_{\theta}(d\theta) = 0$$

Hence, $\theta_{n} = \frac{1}{n} \sum\theta_\infty$ converges to $\theta_\infty$.

In the LS setting, it turns out that $\theta_{\infty}$ is independent of $\gamma$, and satisfies

$$\theta_{\infty} = \left[I - \gamma X^T X\right]^{-1} \gamma X^T y$$

For non-quadratic objectives, what happens?
\[ \Theta_n = \Theta_{n-1} - \gamma D\psi_n(\Theta_{n-1}) \]

is also a Homogeneous Markov chain.

Its stationary distribution \( P \) satisfies
\[ \int Dg(\theta) P(\theta d\theta) = 0 \]

However, its mean \( \bar{\Theta} = \int \theta P(\theta d\theta) \) does not minimize \( g \):

\[ Dg(\bar{\Theta}) = Dg(\int \theta P(\theta d\theta)) \neq \int Dg(\theta) P(\theta d\theta) = 0. \]

Thus, \( \Theta_n \) oscillates around the "wrong" stationary point. \( \bar{\Theta} \neq \Theta. \)

Moreover, \( \| \Theta - \bar{\Theta} \| = \mathcal{O}(\gamma) \)

\[ \bar{\Theta}_n - \bar{\Theta} = \bar{\Theta}_{n-1} - \bar{\Theta} + \bar{\Theta} - \Theta. \]

\[ \begin{align*}
\text{stochastic} & \quad \text{deterministic}
\end{align*} \]

[Batch, Demand, Gradient '17] If \( g \) is strongly convex, for small \( \gamma \) we have
\[ \delta_n = \delta_{n-1} + \lambda_n + \gamma \nabla \psi_n, \]

with \( \| \delta_n \| \leq C \gamma \), and
\[ \mathbb{E} \left[ \tilde{\Theta}_n^2 - \Theta^2 \right] = \frac{A(\Theta, \Theta)}{\lambda_n} + \delta_n + \gamma \delta_n. \]

Consider two chains \((\bar{\Theta}_n)\), \((\bar{\Theta}'_n)\) associated with \( \bar{\Theta} \) and \( \bar{\Theta}' \) respectively. Then we have that
\[ 2 \bar{\Theta}_n - \bar{\Theta}'_n \]

satisfies
\[ \mathbb{E} \left[ 2 \bar{\Theta}_n - \bar{\Theta}'_n \right] = \frac{2A(\bar{\Theta}, \bar{\Theta}) - A(\bar{\Theta}, \bar{\Theta}')}{\lambda_n} + \delta_n. \]

Non-convex and SCD (Batch, Curtis, Audet '16)

How about non-convex case? (Optim methods for Cags, Scales etc.)

We shall assume smoothness.

Assumption 1: \( F: \mathbb{R}^d \to \mathbb{R}^d \) is continuously differentiable, and \( \partial F \) is \( \beta \)-Lipschitz continuous.

\[ (F(x) \leq F(y) + \langle \nabla F(x), y-x \rangle + \frac{\beta}{2} \| x-y \| ^2) \]

Generic SCD algorithm

stochastic gradient:
\[ \Theta_{k+1} = \Theta_k - \gamma_k \nabla F(\Theta_k, z_k) \]

Algorithm
\[ \mathbb{E}_{\tilde{\Theta}_k} \left[ F(\Theta_{k+1}) \right] = F(\Theta_k) - \gamma_k \langle \nabla F(\Theta_k), \tilde{\Theta}_k \rangle + \frac{\gamma_k \beta}{2} \| \tilde{\Theta}_k \| ^2 \]

\[ F(\Theta_{k+1}) - F(\Theta_k) \leq \langle \nabla F(\Theta_k), \Theta_{k+1} - \Theta_k \rangle + \frac{\gamma_k \beta}{2} \| \tilde{\Theta}_k \| ^2 = -\gamma_k \langle \nabla F(\Theta_k), \tilde{\Theta}_k \rangle + \frac{\gamma_k \beta}{2} \| \tilde{\Theta}_k \| ^2 \]
Assumption: Denote by $\mathbb{V}_{3, k}(g(\theta, z_k)) = \mathbb{E} \|g\|^2 - (\mathbb{E} g)^2$ the variance of the stochastic gradient.

Assume $\mathbb{V}_{3, k}(g) \leq M^2 + \nu \|DF\|^2$

and $\mathbb{E} g = DF(\theta_c)$.

Then $\mathbb{E} \left[ (\mathbb{E} g(\theta_{k+1})) - g(\theta_c) \right] \leq (1 - \frac{1}{2} \gamma \beta (1 + \nu)) \mathbb{E} \|DF\|^2 + \frac{1}{2} \gamma \beta M$.

Theorem (Non-convex Objective, Fixed Step-size):

Suppose $F$ is bounded below, and $y$ fixed and $0 < \gamma < \frac{1}{\beta (1 + \nu)}$.

Then $\mathbb{E} \left[ \frac{1}{K} \sum_{k=1}^{K} \|DF(\theta_k)\|^2 \right] \leq \frac{\gamma \beta M + \mathbb{E} \|DF(\theta_c)\|^2}{1 - \gamma \beta (1 + \nu)}$.

Proof: $\mathbb{E} \left[ (\mathbb{E} g(\theta_{k+1})) - g(\theta_c) \right] \leq (1 - \frac{1}{2} \gamma \beta (1 + \nu)) \mathbb{E} \|DF(\theta_c)\|^2 + \frac{1}{2} \gamma \beta M \leq (1 - \frac{1}{2} \gamma \beta (1 + \nu)) \mathbb{E} \|DF(\theta_c)\|^2 + \frac{1}{2} \gamma \beta M$.

$\mathbb{E} \left[ \frac{1}{K} \sum_{k=1}^{K} \|DF(\theta_k)\|^2 \right] \leq \frac{2 (\mathbb{E} g - g(\theta_c))}{\gamma \beta} \frac{\gamma \beta M K}{1 - \gamma \beta (1 + \nu)}$.

Let $\delta > 0$. No work $\mathbb{E} \|DF(\theta_c)\|^2 \rightarrow 0$.

$\gamma$ is in a tradeoff; $\gamma$ small means we will be small, but it will take longer to converge.

Decreasing step-size?: Suppose $y_k$ satisfy $\sum_{k \in \mathbb{K}} y_k = +\infty$, $\sum_{k \in \mathbb{K}} y_k < +\infty$.

Theorem: Let $A_k = \sum_{k \in \mathbb{K}} y_k$. Then, under previous assumptions, $\mathbb{E} \left[ \frac{1}{A_k} \sum_{k \in \mathbb{K}} \|DF(\theta_k)\|^2 \right] \rightarrow 0$ ($1 < 1$).

Proof: We know by hypotheses that $y_k = 0$, so wlog $y_k \beta (1 + \nu) \leq 1$ and...
As before, we have
\[ \mathbb{E} \left[ F \left( \Theta_{k+1} \right) \right] - \mathbb{E} \left[ F \left( \Theta_k \right) \right] \leq -\frac{1}{2} \beta_k \mathbb{E} \left[ \| \nabla F (\Theta_k) \|^2 \right] + \frac{1}{2} \mu_k \sum_{k=1}^{\infty} \mathbb{E} [ F_k ].
\]

Summing again, we have
\[ \mathbb{E} \left[ F (\Theta_k) \right] \leq -\frac{1}{i} \sum_{k=1}^{\infty} \mathbb{E} \left[ \| \nabla F (\Theta_k) \|^2 \right] + \frac{1}{i} \beta_m \sum_{k=1}^{\infty} \mathbb{E} [ F_k ].
\]

\[ \Rightarrow \sum_{k=1}^{\infty} \mathbb{E} \left[ \| \nabla F (\Theta_k) \|^2 \right] \leq 2 \mathbb{E} [ F (\Theta_1) - F (\Theta_i) ] + \beta_m \sum_{k=1}^{\infty} \mathbb{E} [ F_k ].
\]

\[ \Rightarrow \sum_{k=1}^{\infty} \mathbb{E} \left[ \| \nabla F (\Theta_k) \|^2 \right] \leq \frac{C}{\Delta t} \rightarrow 0.
\]

**Corollary:** If we further assume that $F$ is twice differentiable and $\Theta \rightarrow \| \nabla F (\Theta) \|^2$ is Lipschitz, then
\[ \lim_{i \rightarrow \infty} \mathbb{E} \left[ \| \nabla F (\Theta_k) \|^2 \right] = 0.
\]

Let $\xi$-approximate first-order critical points
\[ \exists \delta; \| \nabla F (\Theta) \| < \epsilon \delta.
\]

Q: How fast can gradient descent reach first-order critical points?

**Theorem (Nesterov '98):** $\beta$-smooth. $F_0$ with
step size $\eta = \beta^2$ requires $g (\beta) - g^*$ iteration to reach a $\epsilon$-1st order stationary point of $g$.

**Proof (simplified for continuous time):**
\[ x (t) = -\nabla g (x (t)). \quad g \text{ bounded below.}
\]

\[ \Rightarrow g (x (T)) - g (x (0)) = \int_0^T \nabla g (x (t)) \cdot x (t) \, dt = -\int_0^T \| \nabla g (x (t)) \|^2 \, dt
\]

\[ \Rightarrow \int_0^T \| \nabla g (x (t)) \|^2 \, dt \leq K \forall T.
\]

\[ \Rightarrow h (t) = O (e^{-t}).
\]

Remarks: No curse of dimensionality in this bound.

Here we reach a 1st order stationary point. In the case of convex functions, 1st order stationarity is sufficient - no curvature.

**Classification of critical points.**

\[ d_f: \theta^* \text{ is a strict saddle point if } \nabla g (\theta^*) = 0 \text{ and } \nabla^2 g (\theta^*) \text{ is negative definite.}
\]
A local minimum satisfies \( \min (D^2 g(\theta)) \geq 0 \) (not necessarily strict).

Every critical point \( (Dg(\theta)) = 0 \) is an equilibrium point of gradient. Which are stable equilibria?

By intuition: Strict saddles are unstable.

Questions: A reasonable strategy to avoid unstable equilibria is to add noise (SGD). How is this necessary? Sufficient?

Worst-case analysis: Can gradient converge to saddle points with appropriately (bad) init. (Motivation 0?) But what about "generic" init?

Example: quadratic case: Consider the non-convex quadratic function \( g(\theta) = \frac{1}{2} \theta^T H \theta \), \( H = \text{diag}(h_1, h_2) \).

\( h_1 \cdot h_2 > 0 \), \( h_1 \cdot h_2 < 0 \)

A single critical point \( \theta^* = 0 \), which is a strict saddle.

Gradient Descent initialized at \( \theta_0 \) is
\[
\theta_{k+1} = \theta_k - \alpha Dg(\theta_k),
\]
\( \alpha \ll 1 \) is the step-size. Theorem: \( \lim_{k \to \infty} \theta_k = \theta^* \).

If \( \theta_k \) is sampled from a distribution which is absolutely continuous w.r.t. ambient dir, then \( \theta_k \)

avoids this strict saddle with probability \( 1 \).

How general is this? !

**Stable Manifold Theorem:** Assume \( g \) twice differentiable.

A discrete-time optim algorithm is a mapping \( f: \mathbb{R}^d \to \mathbb{R}^d \).

\( x_0 \in \mathcal{G}_0 \), \( f(x_k) = x_k - \nabla g(x_k) \).

Iterate \( k \) is obtained as \( \theta_k = (f^k)(\theta_0) \).

\[ X^* = \{ \text{strict saddle points of } g \} \]

**Definition (Global Stable Set):** The global stable set of strict saddles is
\[
S_{\eta} = \{ \theta_0 \mid \lim_{k \to \infty} (f^k)(\theta_0) \in X^* \}
\]

S_{\eta} thus contains the initial values that eventually will land in a strict saddle.

In a neighborhood of a critical point \( \theta^* \), how to capture the local attractive set, i.e. the points around \( \theta^* \) that will be pushed to \( \theta^* \)?

See e.g. \( \eta \in \mathcal{G}_{\eta} \).

\[ D^2 g(\theta) > 0 \]
If this local attractor set has zero measure within a small neighborhood of $0^*$, then with probability 1 an init close to $0^*$ will leave this neighborhood.

Q: How to go from local to global stability?
A: A proof: it is not immediate: escaping a saddle point does not guarantee no won't fall into another one later.

Def: Given a diffeomorphism $\phi(x)$, an unstable fixed point set is $A^* = \{ x \mid \phi(x) = x, \phi^n(x) \neq x \}$

$\|D\phi(x)\| > 2$ 

**Theorem (Stable Manifold, Smale’67)**: Let $\phi$ be $C^\infty$ mapping $X \rightarrow X$ and $\text{Det}(D\phi(x)) \neq 0 \forall x \in X$.
Then the set of initial points that converge to an unstable fixed point has measure zero:

$$\mu(\{ x_0 \mid \lim_{n \to \infty} \phi^n(x_0) \in A^* \}) = 0.$$ 

**Corollary**: If $x^* \in A^*$, then $\mu(S(x^*)) = 0$.

**Theorem (Gradient Descent Avoids Strict saddle, Lee 91)**
Assume $g$ is $p$-smooth, and $p > p^*$. Then if $0^*$ is a strict saddle, then $Pr(\lim \Theta = 0^*) = 0$ 

**Proof**: Every strict saddle of $g$ is an unstable fixed point of $\phi(x)$:

$$\phi(x) = x - \delta Dg(x).$$

$$D\phi(x) = I - \delta D^2 g(x)$$

$\Theta$ start saddle $\Rightarrow \chi_{min}(D^2 g(x)) < 0$ $\Rightarrow$

$$\chi_{max}(I - \delta D^2 g(x)) > 1$$

$\Rightarrow$ If $x < p^*$, then $\det(D\phi(x)) \neq 0 \forall \theta$

$$\chi_{min}(I - \delta D^2 g(x)) > 0 \forall x.$$

**Extraneous**:
Prove that $\phi(x)$ is a diffeo, it is invertible.
$\phi$ injective suppose $x, y s.t. \phi(x) = \phi(y)$

Integrate $x - y = \delta (Dg(x) - Dg(y))$

$$\|x - y\| = \delta \|Dg(x) - Dg(y)\| < \beta \|x - y\|$$

but $\beta < 1$.

$\phi$ surjective: Consider $x$ an implicit function

$$\phi(y) = \arg \min_{x} \frac{1}{2} \|x - y\|^2 - g(x) = \arg \min_{x} \|x - y\|^2$$

Since $x < p^*$, $F_y$ is strictly convex w.r.t $x$, $x$ unique and

Using KKT we have $\phi(y) = y - \delta Dg(\phi(y)) = 0$

$$y = \delta \chi_{max}(D^2 g(\phi(y))) = \phi(y)$$
Escaping from saddle points

- We are interested in finding $C^2$-second order stationary points: $D^2 g$ Lipschitz, they are defined as $\{ \nabla g(x) = 0 \}$ and $
abla^2 g(x) \succeq 0$ with $\ell = \text{Lip}(D^2 g)$.

- Stochastic GD finds $C^2$-second-order critical points: Thrift & Sattath. If $g$ satisfies the strict saddle, then noisy GD [13] converges to a local minimum in polynomial time.

- "Vanilla" SGD: $\theta_{k+1} = \theta_k - \eta (DG(x_k) + \epsilon_x)$.

  Time is dimension dependent $\ell \sim N/\epsilon_0 \ell$.

  (Several improvements in subsequent papers).

- We have seen that both GD and SGD escape strict saddles.$\infty$

  1. Depending on input dim? ensured?

  2. Does noise improve or decrease efficiency?

Theorem (Jin et al. '17) Noisy Gradient descent finds $C^2$-second-order stationary point w/p $O(\frac{\ell \| \theta_0 \|}{\eta})$ (up to polylog factors).

- Adding noise is sufficient to escape saddles.

- Up to log factors, this matches GD rate for first-order methods.

Extensions to proximal method, mirror descent, coordinate descent.

This result establishes that GD escapes strict saddles. Does this imply it converges to local minimizers?

Required additional properties

1. Set of strict saddles cannot be too "large" countably finite / isolated saddle, are ok.

2. $\lim \theta_k$ exists.

Two sufficient conditions are

1. Isolated critical points and compact sublevel sets.

2. Local Liapunov-like inequality: $\exists m, \alpha, \epsilon$ s.t.

   $$\| Dg(x) \|^2 m \| \nabla g(x) \|^2 \alpha, \alpha \leq 1$$

   for: $\theta \in \Theta \subset \mathbb{R}^d$; $\nabla g(x) \succeq 0$; $\| \nabla g(x) \|^2 \leq \epsilon$

   Ensures length traveled by iterate of GD is finite.

This result shows that the tiniest amount of randomness (at init) is sufficient in mild assumption to avoid saddles. What is the benefit of extra noise in the dynamics?
It turns out that noise is also necessary. Even with uniform input, one can construct smooth functions for which \( \mathbb{C}^O \) requires exponential time [Otto et al.'17].

High-dimensional Energy Landscapes

What are good models for non-linear, yet tractable high-dimensional energy landscapes?

Models for statical physics are very rich mathematically.

**E**g: Spherical Spin glass:

\[
E(\Theta) = \frac{1}{\Lambda (k-1)/2} \sum_{l=1}^{k-1} \sum_{i=1}^{k} \sum_{j=i+1}^{k} \Theta_{i,l} \Theta_{j,l} \Theta_{i,j} \Theta_{j,i}
\]

\( \Theta_{i,l} \sim N(0, \sigma^2), \quad \| \Theta \|_2 = 1 \)

Hamiltonian of the spherical \(1\)-spin spin glass.

Kac-Rice Formula

\[
\frac{1}{N} \log E(\Theta, \Theta_{i,i}) = \Theta(\rho, \mu)
\]

Reproducing Kernels and Kernel Dynamics

**Goal:** Infinite-dimensional least squares.

\( \mathbb{K}(x, y) \rightarrow \mathbb{H} \)

- Generic learning in a infinite-dimensional space

\[
\min_{f \in \mathbb{F}} \| f - \hat{f} \|^2.
\]

This problem is feasible.

- Since we measure error in an empirical norm, we need some form of regularization in the fitting phase. What is the most structure available?

**Hilbert Space \( \rightarrow \) Kernel.**

- **Def:** Given a set of objects \( X \), a psd kernel is a symmetric function \( K : X \times X \rightarrow \mathbb{R} \) s.t. for all finite sequences \( x_i \in X \) and \( \alpha \in \mathbb{R} \),

\[
\sum_{i,j=1}^{N} \alpha_i \alpha_j K(x_i, x_j) \geq 0.
\]

**Ex:** \( K(x, x') = \langle \phi(x), \phi(x') \rangle \) for arbitrary mappings \( \phi : X \rightarrow \mathbb{F} \), \( \mathbb{F} = \text{Hilbert space} \).
Theorem (Aronszajn’50): \( K \) is a PSD kernel if and only if there exists a Hilbert space \( \mathcal{F} \) and a mapping \( \phi: \mathcal{X} \to \mathcal{F} \) such that \( K(x, x') = \langle \phi(x), \phi(x') \rangle \).

\( \mathcal{F} \): Feature space; \( \phi(x) \): feature map of \( x \).

Example:

- Linear Kernel: \( K(x, y) = x^T y \) \( \phi(x) = x \)

- Polynomial Kernel: \( K(x, y) = (1 + x^T y)^d \) \( \phi(x) = \text{Homogeneous} \)

- If \( K_1 \) and \( K_2 \) are kernels, then
  \[ \alpha K_1 + \beta K_2 \text{ is a kernel.} \]
  \[ K_1, K_2 (x, x') = K_1(x, y) K_2(x, y) \text{ also kernels.} \]

- Gaussian Kernel: \( K(x, y) = \exp(-d(x, y)^2) \) \( \phi(x) = \text{finite-dimensional?} \)

Reproducing Kernel Hilbert Space

In Aronszajn Theorem, we have the representation
\[ K(x, x') = \langle \phi(x), \phi(x') \rangle \]

We instantiate \( \mathcal{F} \) as a function space over \( \mathcal{X} \). 
\[ \phi(x) = K(\cdot, x) \quad F : \mathcal{X} \to \mathcal{F} \]

\[ K(x', x) \]

\[ \text{Function evaluation: } f(x) = \langle f, \phi(x) \rangle \quad \text{for } f \in \mathcal{F} \]

Reproducing property: \( K(x, y) = \langle K(\cdot, x), K(\cdot, y) \rangle \)

Trick: Evaluating \( f \) at a point \( x \) is an inner product in the feature space.

Equivalently, the evaluation functional
\[ L_x : f \mapsto f(x) + \langle f, K \rangle \text{ is bounded:} \]
\[ \exists M > 0 \text{ such that } \| f \| \leq M \| f \|_2, \forall f. \]

Regulation and Representer Theorem

Suppose data \( x_i \in \mathcal{X} \), labels \( y_i \in \mathcal{Y} \), \( i = 1, \ldots, n \). Use a kernel \( K \) with RKHS \( \mathcal{F} \).

Learn \( f \) We consider
\[ \min_{f \in \mathcal{F}} \sum_{i=1}^n L(y_i, f(x_i)) + \frac{\lambda}{2} \| f \|_2^2 \]

Represent Theorem (Kimeldorf & Wahba, 70') The minimizer is of the form
\[ f = \sum_{j=1}^n \alpha_j \phi(x_j) = \sum_{j=1}^n \alpha_j K(x, x_j) \]

Proof: See.
Kernel Ridge Regression

Data \( x_1, \ldots, x_n, y_1, \ldots, y_n \in \mathbb{R}^d \), and kernel \( \kappa() \).

Infinite-dimensional Least Squares

\[
\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \| f \|_F^2
\]

Option 1: Use strong representable theorem.

\[ f = \mathbf{w} \cdot \Phi(x) \], obtained with

\[
\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (y_i - (\kappa \cdot \mathbf{w})(i))^2 + \lambda \| \mathbf{w} \|_2^2
\]

\[ \mathbf{w}^* = (\kappa + n \lambda I)^{-1} \mathbf{y} + \varepsilon, \quad \lambda \varepsilon = 0 \]

\( f^* \) is unique.

Option 2: Suppose finite-dimensional feature space can be kernel.

\[ F \subset \mathbb{R}^d, \quad \Phi \in \mathbb{R}^{d \times d} \]

\[
\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \| \mathbf{y} - \Phi \mathbf{w} \|_2^2 + \lambda \| \mathbf{w} \|_2^2
\]

\[ \mathbf{w} = (\Phi^T \Phi + n \lambda I)^{-1} \Phi^T \mathbf{y} \]

\[ \mathbf{y} = \Phi^T (\kappa + n \lambda I)^{-1} \mathbf{y} = \Phi^T \kappa \mathbf{x} \]

\[ \{ f(x) = \int_{\mathbb{R}^d} p(\nu) \Phi(x) d\nu \} \] is RKHS, with kernel

\[ k(x,y) = \int \Phi(x) \Phi(y) d\nu \]
As \( n \to \infty \), \( \hat{\theta} \to \theta^* \) (Rahimi & Recht '07)

Rate is essentially the Monte-Carlo Rate: \( \frac{1}{\sqrt{n}} \)

**Kernel Dynamics and the NTK**

1. Consider a generic parametric non-linear model
   \[ f(x; \theta) \quad (e.g. \ f(x; \theta) = \sum_{i=1}^{k} \alpha_i \phi(x, \beta_i)) \]

3. We evaluate this model with the cosine loss
   \[ R(f(x; \theta)) \quad (e.g. \ R(f(x; \theta)) = \| f - f(x; \theta) \|_2 \]

4. This update is the same as kernel regression, using the kernel \( K_{0, \theta} \)

Fact: If \( \theta(1) \) does not move too much,
\( K_{1} \approx K_{0} \), thus the general GD dynamics will be well approximated with kernel dynamics.
The assumption that $K_0 = K_0$ is equivalent to the assumption the model is linear: our parameterized model thus becomes

$$T_e(x; \theta) = f(x; \theta_0) + (\theta - \theta_0) \cdot \nabla f(x; \theta_0)$$

Up to an affine offset, this model is equivalent to a kernel method of the form

$$K(x, x') = \langle \nabla f(x; \theta_0), \nabla f(x'; \theta_0) \rangle$$

**Lazy Training Regime**

- Project the lazy regime using linear approximation in function space.
- What is the corresponding linear model?
  - It is defined by the kernel tangent kernel:
    $$K(x, x') = \nabla f(x; \theta_0)^T \nabla f(x'; \theta_0)$$
- Scaled objective function and companion with kernel dynamics.

Theorem by, Chizat & Bach.

- Analysis in terms of $f(x; \theta) = G(x) \sum_k \phi(x; \theta_k)$
  - The quantity that controls the weight of predictions.

Lazy approximation is roughly

$$\frac{\|D \hat{f}_0\|}{L D \tilde{f}}$$

If $f_n \to f$, this ratio converges to a constant.

If we scale by something large, it explodes.

$$\frac{\|D \hat{f}_0\|}{\|D f\|} = \frac{\|D \hat{f}_0\|}{\|D f\|}$$

$$L D \tilde{f} = \alpha(n) \|D f\|_n \|E \|\|D f\|_n$$

$$\|E \|_n \|f_0 - \tilde{f}_0\|_n \leq \text{max}(1, \alpha(n))$$

Q: What happens in the regime $\alpha(n) = 1$?

In this regime, we develop the particle interacting system:

$$f_n(x; \theta) = \frac{1}{n} \sum_{i=1}^n \phi(x; \theta_i) \quad \theta = (\theta_1, \ldots, \theta_n)$$

$$L(\theta) = \mathbb{E} \left[ \frac{1}{n} \| f_n(\theta) - f^* \|_n^2 \right]$$

$$= C f^* - \langle f_n(\theta), f^* \rangle + \frac{1}{n} \| f_n(\theta) \|_n^2$$

$$= C f^* - \frac{1}{n} \sum_{i=1}^n \langle \phi(\theta_i), f^* \rangle + \frac{1}{2n} \sum_{i,j} \langle \phi(\theta_i), \phi(\theta_j) \rangle$$

Let $\hat{\theta}_n = \frac{1}{n} \sum \delta_{\theta_i}$ empirical measure

$$f_n(\theta) = \left( \phi(\theta) \right) \hat{\theta}_n (\theta)$$
\[
L(\theta) = C_f - \int F(\theta) \mu_\theta (d\theta) + \sum K(\theta, \theta') \mu_\theta (d\theta) \mu_\theta (d\theta')
\]

\[
F(\theta) = \left\langle \Phi(\theta), \Phi^* \right\rangle, \quad K(\theta, \theta') = \left\langle \Phi(\theta), \Phi(\theta') \right\rangle.
\]

\[
\theta_i = -\nabla_{\theta_i} L(\theta) = -\frac{1}{n} \left\langle \nabla \Phi(\theta), \phi_i(\theta) \right\rangle + \frac{1}{n \lambda} \sum_{i, j} \left\langle \nabla \phi_i(\theta), \phi_j(\theta) \right\rangle.
\]

\[
\nabla \Phi(\theta) = \nabla V(\theta; \mu) \text{ with } V(\theta) = -F(\theta) + \int K(\theta, \theta') \mu_\theta (d\theta').
\]

1. The gradient of each neuron corresponds to feeling a velocity field evaluated at each location.

\[
\theta_i = \Phi_i. \text{ Describe the evolution of a measure as subject to a velocity field? }
\]

\[
\text{Continuity equation: } \begin{align*}
\frac{\partial}{\partial t} \mu_t &= \text{div} (DV \cdot \mu_t) \\
\frac{\partial}{\partial t} \left\langle \Phi, \mu_t \right\rangle &= -\left\langle \frac{\partial}{\partial t} \Phi, DV \right\rangle \mu_t & \forall \Phi \in C_c.
\end{align*}
\]