Course Description

Standard supervised classification setup:

- Data distribution: $(x, y) \in \mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R}$
  $(x, y) \sim \nu$, ν: probability measure in $\mathcal{X} \times \mathcal{Y}$.
- Loss function $\ell(\hat{y}, y)$; convex with respect to $\hat{y}$.
  e.g. $\ell(\hat{y}, y) = \|\hat{y} - y\|^2$; $\ell(\hat{y}, y) = \log(1 + e^{-\hat{y} \cdot y})$
- Model $\hat{y} = \phi(x; \theta)$; $\theta \in \mathcal{D}$
- Empirical Risk Minimization (ERM) / Structural Risk Minim.

  $\{(x_1, y_1) \dots (x_n, y_n)\} \overset{iid}{\sim} \nu$ $\qquad \hat{\nu} = \frac{1}{n} \sum_{l=1}^{L} \delta_{(x_l, y_l)}$
  empirical measure.

  Empirical loss/Risk:

  $$\hat{R}(\theta) = \mathbb{E}_{\hat{\nu}} [\ell(\phi(x, \theta), y)] + \underbrace{\mathcal{R}(\theta)}_{\text{regularization}}$$

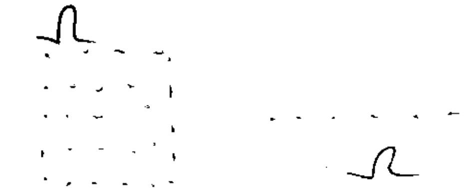
  $$= \frac{1}{n} \sum_{l} \ell(\phi(x_l, \theta), y_l) + \mathcal{R}(\theta).$$

- Optimization:
  $$\min_{\theta} \hat{R}(\theta)$$

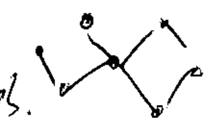
<u>Many questions</u>:
- choice of model? (approximation)
- choice of optimization? $\quad$ when is deep learning specific?
- Generalization? (statistics)

<u>Part I</u>: Geometry of data

- input x in a high-dimensional space. $x \in \mathbb{R}^d$ $d \gg 1$.
- But in many applications, x may be itself modeled as a function defined over a low-dimensional domain.
  $\hookrightarrow$ Eg $x = x(u)$, $u \in \Omega$



- Impact of this extra structure in Ω?
- In particular, assuming no noise in targets, $y = f(x)$, how to build representations $\Phi(x; \Omega)$ that are stable to known transformations T ~~that~~ st $|f(Tx) - f(x)| \simeq 0$?
- Domain Ω can be a group (lots of global symmetries) (torus, $\mathbb{R}^d$)
  a manifold
  a graph



- Q: general framework? • Applications to livesepl.
- Density estimation: choice of metric? [from MLE to OT].

<u>Part II</u>: Geometry of Optimization and Learning

- Previous questions do not concern any learning/optimization.
- Basics of convex optimization and Nesterov acceleration.
- Non-convex optimization (or how to escape saddle points eff).
- Continuous-time analysis (stochastic and deterministic).

~~"over-parametrised" regime~~,
role of batch normalization
positive / negative results.

- Asymptotic behavior of "over-parametrised" neural networks learning dynamics.
  - ↳ Reproducing Kernel Hilbert Spaces (RKHS).
  - ↳ Mean field and optimal transport.
- Implicit bias of gradient descent.

~~Lecture 1: The curse of dimensionality.~~

Parallel curricula:

DFL 1: The Neural ODE [Chen et al. '18]

$x \to \boxed{\Phi} \to z = \phi(x)$

$z = X(T)$, $X(0) = x$

$\dot{X} = A_\theta(X)$?

DFL 2: Exploration in RL.

- Unifying count-based Exploration and Intrinsic Motivation [Bellemare et al '16]

Lecture 1: The curse of dimensionality

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

- Hypothesis class : a function space

$\mathcal{F} : \{ f : \mathcal{X} \to \mathbb{R} \}$ $\quad \mathcal{X} = \mathbb{R}^d$

observe n points $(x_1, f^*(x_1)) \dots (x_n, f^*(x_n))$.

$x_i \sim \nu$ iid.

f^* unknown, assumed in $\mathcal{F}$.

Q: How does the risk of our best estimator f grow with n, d ?

1st case: $\mathcal{F}$ contains linear functions

$f^*(x) = \langle x, \theta^* \rangle + w, \quad w \sim N(0, \sigma^2 I)$

$f_\theta(x) = \langle \theta, x \rangle \quad L(\theta) = \frac{1}{2n} \sum_{i=1} \big(\underbrace{w_i + \langle x_i, \theta^* \rangle}_{y_i} - \langle x_i, \theta \rangle \big)^2$

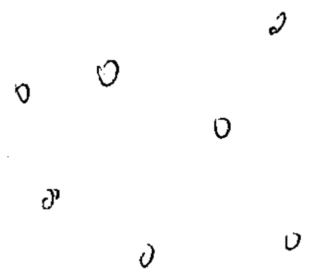
$\hat{\theta} \sim N(\theta, \sigma^2 (X^T X)^{-1}) \quad L(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (w_i + \langle x_i, \theta^* - \theta \rangle)^2$

$\nabla L(\theta) = \frac{1}{n} \sum x_i (w_i + \langle x_i, \theta^* - \theta \rangle)$

$\hat{\theta} = \hat{\Sigma}_X^{-1} \hat{\Sigma}_{XY} \qquad \hat{\Sigma}_X = \frac{1}{n} \sum x_i x_i^T$

$\|\hat{\theta} - \theta^*\| \sim \sqrt{d}$

2nd case: $\mathcal{F}$ contains Lipschitz functions.



$$\| f(x) - f(x') \| \leq \beta \| x - x' \|$$

$$\sup \| \hat{f}(x) - f^{\circ}(x) \| \leq \varepsilon \, ?$$

As $n \sim \varepsilon^{-d}$

Lecture 2: Geometric Stability in Euclidean Domain (Scattering Transform)

→ Linearization.

$f(x)$, $x \in \mathcal{X}$ → high-dim space.

$\hat{f}(x) = \sigma(a^T \phi(x) + b)$ is a good approximation.

If we think about $f(x)$ as encoding a two-class classif., the change of variables $x \mapsto \Phi(x)$ linearizes the boundary.

↳ in particular, $a^T(\phi(x) - \phi(x')) = 0$

$\Rightarrow \; f(x) = f(x')$

(level sets of f mapped to hyperplanes by Φ)

(intra-class variability becomes flat)

~~[illegible]~~ Q: How to construct such Φ?

→ Invariance & Symmetry. $x \in \Omega$ (input domain)

Def: a global symmetry of f is operator $\varphi \in \mathrm{Aut}(\Omega)$ st

$$f(x) = f(\varphi(x)) \quad \forall x.$$

$\Rightarrow$ They can be absorbed by ϕ as

- Invariants: $\phi(\varphi(x)) = \phi(x) \quad \forall x.$
- Equivariants: $\phi(\varphi(x)) = \tilde{\varphi}\, \phi(x) \quad \forall x$

Q: symmetries in image recognition problems?

Translations: $x \in L_2(\mathbb{R}^2)$, $\boxed{3} \to \boxed{3}$

$\{\varphi_v \,;\, v \in \mathbb{R}^2\}$, $\varphi_v(x)(u) = x(u-v)$

Dilations: $\{\varphi_s \,;\, s \in \mathbb{R}_+\}$, $\varphi_s(x)(u) = s^{-1} x(s^{-1}u)$

Rotations $\{\varphi_\theta \,;\, \theta \in [0, 2\pi)\}$, $\varphi_\theta(x)(u) = x(R_\theta u)$

- These are subgroups of the group of rigid motions Affine Group $Aff(\mathbb{R}^2)$, we can associate it with the space of affine transforms

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \mapsto \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} + \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

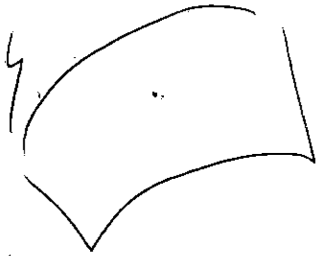
↳ Not-commutative in general.

↳ In audio, this group contains the Heisenberg group, with linear part of the form $\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}$

↳ Lie group; ~~symmetric space~~ (manifolds with symmetries)

- Invariant Representations;

the action of a transformation group G onto X is an orbit: $G \cdot x = \{\varphi_g(x) \,;\, g \in G\}$



~~What we would like what~~

Q: General approach to obtain Group Invariant Represent?

- One-parameter unitary Groups

Def. A one-parameter unitary group $\{\varphi_t \in Aut(\Omega)\}_{t \in \mathbb{R}_0}$

(i) $\forall t, s$ $\varphi_{s+t} = \varphi_s \varphi_t$

(ii) $\lim_{s \to t} \|\varphi_s - \varphi_t\| = 0$, (iii) $\|\varphi_t x\| = \|x\|$ $\forall t, x$.

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

(Eg Rotation / Translations).

Theorem [Stone, 30s'] Ω Hilbert Space.

$$\left\{ \begin{matrix} \text{self-adjoint operators on} \\ \Omega \end{matrix} \right\} \overset{1:1}{\longleftrightarrow} \left\{ \begin{matrix} \text{one-parameter} \\ \text{unitary groups of} \\ Aut(\Omega) \end{matrix} \right\}$$

In particular; given $\{\varphi_t\}_{t \in \mathbb{R}}$, $\exists$ A self-adjoint st $\forall t$, $\varphi_t = e^{itA}$ $\quad (\langle Au, w\rangle = \langle u, Aw\rangle)$

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

$$\hat{x}(\xi) = \int x(u) e^{-i\xi u} du$$

Main properties: [Linearity.] $z = \alpha x + \beta y \Rightarrow \hat{z} = \alpha \hat{x} + \beta \hat{y}$

[Parseval]: $\|\hat{x}\| = \|x\|$, $\langle x, y\rangle = \langle \hat{x}, \hat{y}\rangle$.

(IFT]: $x(u) = \int \hat{x}(\xi) e^{i\xi u} d\xi$

[Transl]: $y = \varphi_v x(u)$; $\hat{y}(\xi) = e^{iv\xi} \hat{x}(\xi)$.

[Dil]: $y = x(su)$; $\hat{y}(\xi) = s^{-1} \hat{x}(s^{-1}\xi)$.

- Stone thm and Fourier: (Translations are diagonal operators in the Fourier domain. Stone thm formalizes)

Fourier Simultaneously diagonalizes all translations. (possible because they commute)

Stone thm generalizes this notion. "Fourier transform"

$A = V^* \operatorname{diag}(\lambda_1 \ldots \lambda_d) V$

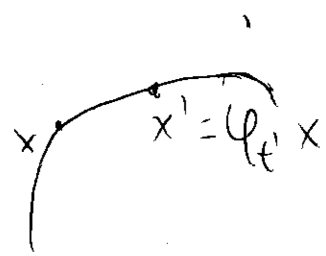
→ larger Abelian groups also ok

$(G = G_1 \times \dots G_\ell)$

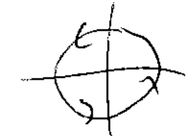
↳ one-parameter groups. ...

→ Generalized to non-abelian case by Peter-Weyl theorem using representation theory.

→ Q: How to obtain invariants in that case?



$A = V^* \mathrm{diag}(\lambda_1 - \lambda_d) V$

$y = Vx$  $y_k' = e^{i\lambda_k t'} y_k$

$y_k = \langle x, v_k \rangle$

· $\phi(x) = |Vx|$ is thus a G-invariant:

$\forall x, t \quad \Phi(\varphi_t(x)) = \phi(x)$

· $A = V^* \Lambda V \Rightarrow e^{itA} = V^* \mathrm{diag}(e^{it\lambda_1}, \dots e^{it\lambda_d}) V$

$V\varphi_t x = V e^{itA} x = VV^* \mathrm{diag}(e^{it\lambda_j}) Vx$

$\Rightarrow |V\varphi_t x| = |Vx| \quad \forall x, t.$

. Thus, in the commutative case, a single layer is sufficient to obtain invariance.

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

↳ symmetry is a very strict criteria. How to relax?

↳ Symmetry groups in Images are low-dimensional; not that helpful!

→ Deformations: $x \in L^2(\mathbb{R}^m)$ $\varphi: \mathbb{R}^m \to \mathbb{R}^m$ diffeomorphism.

$x_\tau(u) := x(u - \tau(u))$ $\quad \varphi(u) = u - \tau(u)$ $(\|\nabla \tau\|_\infty < 1)$

→ Geometric Stability: $\forall x, \tau$, ↳ deformation cost.

$$|f(x) - f(x_\tau)| \lesssim \|\tau\|$$

E.g. $\|\tau\| = \sup_u \|\nabla \tau(u)\|$: measures the stretching of the domain.

Stable representations: $\forall x, \tau, \ \|\Phi(x) - \Phi(x_\tau)\| \leq C\|\tau\|$

since $\hat{f}(x) = \langle \Phi(x), \beta \rangle$, it follows that

$|\hat{f}(x) - \hat{f}(x_\tau)| \leq \|\beta\| \, \|\Phi(x) - \Phi(x_\tau)\| \leq C\|\beta\|\|\tau\|.$

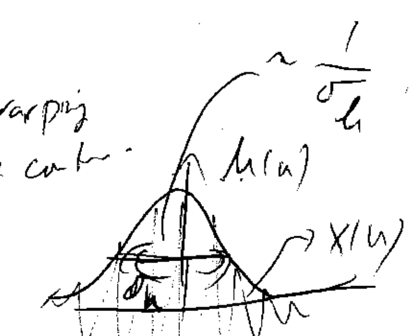
→ Are the Fourier Invariants stable to deformations?

↳ We know $\Phi(\varphi_t x) = \Phi(x)$ ($\varphi_t \in G$) $\forall t, x$, but what happens for transformations $\varphi \notin G$?

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

→ We consider the local deformation cost

$$\|\tau\| = \underbrace{2^{-J} \|\tau\|_\infty}_{\text{displacement} \leq 2^J} + \underbrace{\|\nabla \tau\|_\infty}_{\text{warping cost}}$$



→ Shallow (Fourier) invariant are unstable.

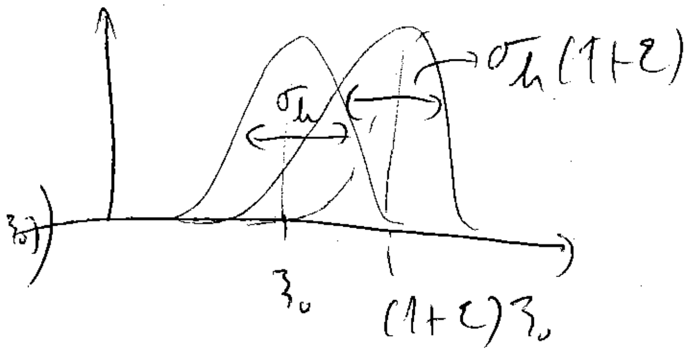
$h(u)$ lowpass window. $x(u) = h(u) e^{i\xi u}$

→ Deformation of the form $\varphi_\tau x(u) = x((1+\varepsilon)u)$, $\varepsilon << 1$

$(\widehat{\varphi_\tau x})(\xi) = (1+\varepsilon)^{-1}\,\hat{x}\left((1+\varepsilon)^{-1}\xi\right)$

$\hat{x}(\xi) = \hat{h}(\xi - \xi_0)$.

$(\widehat{\varphi_\tau x})(\xi) = (1+\varepsilon)^{-1}\,\hat{h}\left((1+\varepsilon)^{-1}(\xi) - \xi_0\right)$



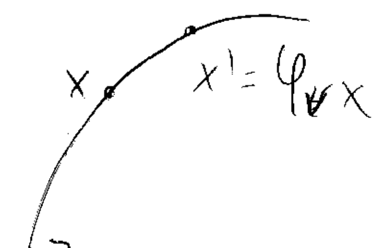
If $(1+\varepsilon)\xi_0 - \xi_0 = \varepsilon\xi_0 \gg \sigma_h(2+\varepsilon)$

then supports are disjoint,

$$\|\, |\hat{x}| - |\hat{x}_\tau| \,\| \sim \|x\|$$

→ Moreover, $|\hat{x}|$ loses information in phase correlations.

⇒ Fix? Local Invariants.



$$\|\Phi(x) - \Phi(\varphi_v x)\| \le C\, 2^{-J}\|v\|,$$

$\forall v, \|x\|=1$, $\Updownarrow$ $\dfrac{\|\Phi(x) - \Phi(\varphi_v x)\|}{\|v\|} \lesssim 2^{-J}$

↳ smooth along the orbits → Haar measure on G

$$\Phi(x) = 2^{-dJ}\int_v \phi(2^{-J}v)\,\varphi_v x\, dv, \quad \Big(\int \phi(v)\,dv = 1\Big) \quad \phi_{J,v}$$

$$= \int \phi_J(v)\, x(u-v)\, dv = x * \phi_J(u), \quad \phi_J = 2^{-Jd}\phi(2^{-J}v)$$

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

Prop: $\Phi(x) = x * \phi_J$ satisfies $\forall \|x\| = 1 \in L^2, \forall \tau$,

$$\|\Phi(x) - \Phi(\varphi_\tau x)\| \le C\|\tau\|.$$

↳ Proof: Schur lemma for integral operators

$$Kf(u) = \int k(u,u')\, f(u')\, du'$$

$$\|K\| \le \max\Big(\sup_u \int |k(u,u')|\, du', \ \sup_{u'} \int |k(u,u')|\, du\Big)$$

$\Phi(x) = x * \phi_J$; $\Phi(x)(u) = \int x(v)\,\phi_J(u-v)\, dv$

$\varphi_\tau(x)(u) = x(u - \tau(u))$

$$\Phi(x_\tau)(u) = \int x(v - \tau(v))\,\phi_J(u-v)\, dv = \int x(v)\,\phi_J(u - v + \rho(v))\cdot |I - \nabla\tau(v)|^{-1}\, dv$$

$$\Rightarrow \big(\Phi(x) - \Phi(\varphi_\tau x)\big)(u) = \int x(v)\left[\phi_J(u-v) - \phi_J(u-v+\rho(v))\cdot |I - \nabla\tau(v)|^{-1}\right] dv$$

Q: Other stable linear operators?

$\forall v$, $\Phi(x) = \Phi(\varphi_v x) \Rightarrow \Phi(x) = \dfrac{1}{\mu(G)}\displaystyle\int \Phi(\varphi_v x)\, dv$

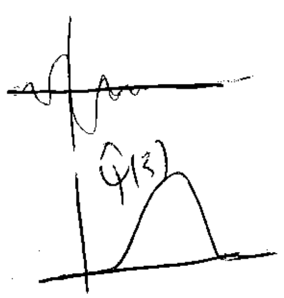
$$\Rightarrow \Phi(x) = \Phi\Big(\frac{1}{\mu(G)}\int \varphi_v x\, dv\Big) = \Phi\Big(\frac{1}{\mu(G)}\int x(u)\, du\Big)$$

Lecture 3 : The Scattering Transform

→ We just saw that the only linear invariant is essentially the average. This loses all the relevant information.

Corollary: High frequency representation must involve a non-linearity. → recall we want (i) Stability
(ii) "richness" ability to discr.

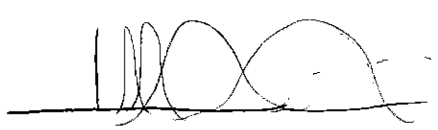
Wavelets ψ bandpass $L^1(\mathbb{R}^d)$ $\int \psi(u)\,du = 0$



$\hat{\psi}(0) = 0$ (We may ask for more vanishing moments) (a single vanishing moment)

equivalently $\perp$ $\hat{\psi}^{(l)} = 0 \quad \forall\, l \le l_0$.

In order to extract all the frequencies, we need wavelets at different scales: $\psi_j(u) = 2^{-j}\psi(2^{-j}u)$. (1D) $j \le J$

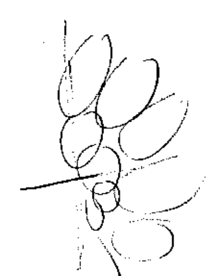


$\psi(u) = e^{i\omega_0 u} h(u)$

$\psi_{j,\theta}(u) = 2^{-2j}\psi(R_\theta 2^{-j}u)$ (2D).

Littlewood-Paley transform: let $x \in L^2(\Omega)$, then

$$Wx = \left\{ \begin{array}{l} x * \psi_{j,\theta} \\ x * \phi_J \end{array} \right\}_{\substack{j \le J \\ \lambda \in \Lambda}} \in L^2(\Omega \times \underbrace{G \times \mathbb{Z}}_{\Lambda})$$



Thm: (Littlewood-Paley): if $\exists\, \delta > 0$ such that

$$\forall \xi, \quad 1-\delta \le |\hat{\phi}(\xi)|^2 + \frac{1}{2}\sum_\lambda |\hat{\psi}(2^j R_\theta \xi)|^2 \le 1,$$

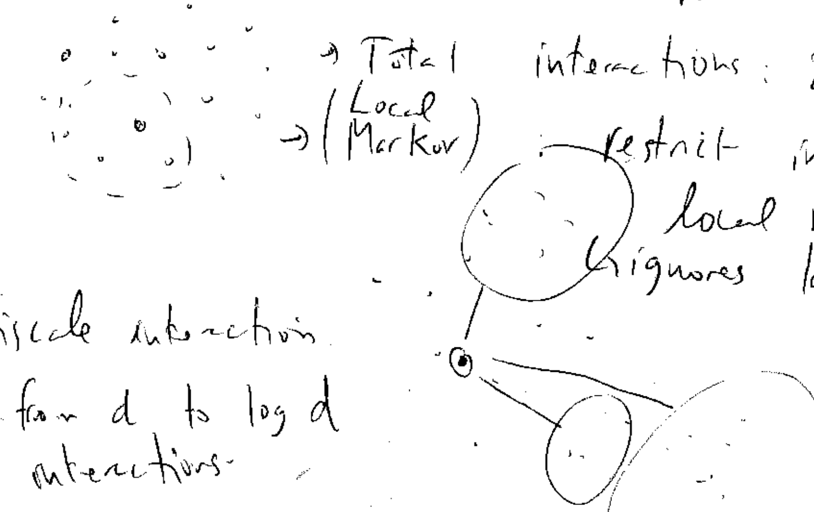
then $\forall x \in L^2$, $(1-\delta)\|x\|^2_{L_2(\Omega)} \le \|Wx\|^2_{L_2(\Omega\times\Lambda)} \le \|x\|^2_{L_2(\Omega)}$

→ Link this to additive stability.

→ wavelets used extensively in harmonic analysis to characterize local regularity (eg through Besov-spaces).

→ used successfully in non-parametric statistics to obtain optimal rates in denoising via th shrinkage estimators (Donoho).

→ wavelets and scale separation: suppose d particles interacting:



→ Total interactions: 2^d

→ (Local Markov): restrict interactions to local neighborhoods
↳ ignores large scales.

→ Multiscale interaction

each particle: from d to $\log d$ interactions.

↳ Related to Fast Multipole Method (Greengard & Rokhlin)

→ Q: Can we then simply concatenate W with an averaging? No, since they have 0 mean. We can create a new invariant out of each wavelet sub-band with a non-linearity that restores an informative average.

For example using a complex modulus: $\int |x * \psi_\lambda|(u)\,du$.

Wx is translation equivariant: $\tilde{x}(u) = x(u-u_0)$

$W_\lambda \tilde{x}(u) = W_\lambda x(u-u_0)$

Q: stable to deformations?

As before, let $\varphi_\tau x(u) := x(u-\tau(u))$ $\forall x$.

Theorem [Mallat '12]: There exists $C > 0$ such that $\forall J \in \mathbb{Z}$

$$\|[W_J, \varphi_\tau]\| \le C \left[\overbrace{\|\nabla\tau\|_\infty \max\left(1, \log \frac{\|\Delta\tau\|_\infty}{\|\nabla\tau\|_\infty}\right) + \|H\tau\|_\infty}^{\|\tau\|} \right]$$

min(J,)

↳ <u>Remark</u>: for the lowpass, we had local stability/<u>invariance</u> $\|\phi - \phi\circ\varphi_\tau\| \le C\cdot\|\tau\|$, here it's <u>equivariance</u>.

↳ Key tool behind this result: Cotlar-Stein near orthogonality lemma: $\{T_j\}_j$ a family of bounded linear operators between two Hilbert spaces E, F; $T_j : E \to F$.

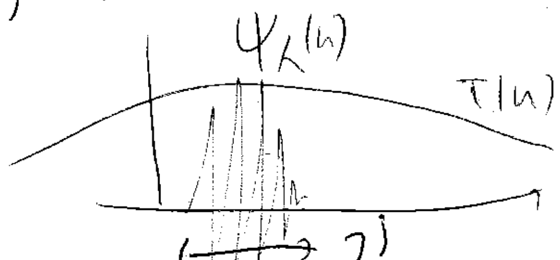
$\{T_j\}_j$ is <u>almost orthogonal</u> if

$$A = \sup_j \sum_k \sqrt{a_{jk}} < +\infty, \quad B = \sup_j \sum_k \sqrt{b_{jk}} < +\infty$$

$$a_{jk} := \|T_j T_k^*\|_E \qquad b_{jk} := \|T_j^* T_k\|_F .$$

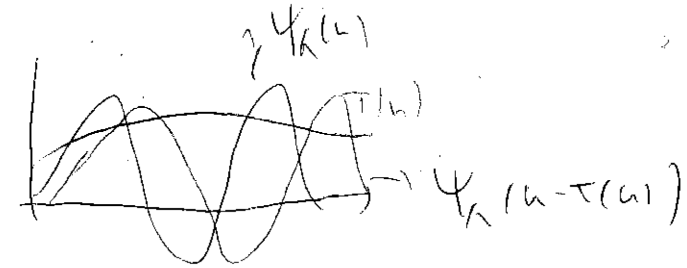
If $\{T_j\}_j$ are almost orthogonal, then $\sum_j T_j$ converges, and $\|\sum_j T_j\| \le \sqrt{AB}$.

<u>Intuition</u>:



(i) at small scales, ψ_λ has small support, and for u, v within support of ψ_λ, and because τ is smooth, $|\tau(u) - \tau(v)| \sim 2^j |\nabla\tau|_\infty$.

Thus $\left|((\varphi_\tau x) * \psi_\lambda)(u) - (x * \psi_\lambda)(u - \tau(u))\right| \sim |\nabla\tau|_\infty$.



(ii) at large scales, ψ_λ is itself smooth, thus $|\varphi_\tau(x * \psi_\lambda) - (\varphi_\tau x) * \psi_\lambda| \sim \|\nabla\tau\|_\infty$.

(similar as our result on local stability for ϕ)

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

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

$$\begin{tikzcd} \varphi_\tau x \arrow[r, "W"] & \varphi_\tau x * \psi_\lambda \approx \varphi_\tau(x * \psi_\lambda) \\ x \arrow[u, "\varphi_\tau"] \arrow[r, "W"'] & \{x * \psi_\lambda\}_\lambda \arrow[u, "\varphi_\tau"'] \end{tikzcd}$$

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

→ Characterizing stable non-linearity $M : L^2(\Omega) \to L^2(\Omega)$.

- Preserve additive stability from Littlewood-Paley.
  $\|Mx - Mx'\| \le \|x - x'\|$.
- Preserve geometric stability from wavelet commut: it is sufficient

Theorem [B'12]: If M is non-expansive in L^2 such that $M\varphi_\tau = \varphi_\tau M$ for φ_τ diffeo, then M is point-wise:

$$Mx(u) = \rho(x(u)) \qquad M: L^2 \to L^2 \text{ s.t. } \rho: \mathbb{R}\to\mathbb{R}$$

↳ Idea of the proof: Look at the isotropy group of $x \in L^2(\Omega)$ $\quad G(x) = \{\phi \in \mathrm{Diff}(\Omega);\ x = x\circ\phi\}$ (symmetries of the level sets).

observation: $G(x) \subseteq G(Mx) \quad \forall x$.

For instance we choose the complex modulus (Relu is ok for real wavelets too).

$$x \longrightarrow \{|x*\psi_\lambda|\}_\lambda \longrightarrow \{||x*\psi_\lambda|*\psi_{\lambda'}|\}_{\lambda,\lambda'} \to \cdots$$

$\downarrow$ $\int x(u)\,du$ $\qquad \downarrow$ $\int |x*\psi_\lambda|(u)\,du$ (Wavelet) $\qquad \downarrow$ $\int ||x*\psi_\lambda|*\psi_{\lambda'}|(u)\,du$

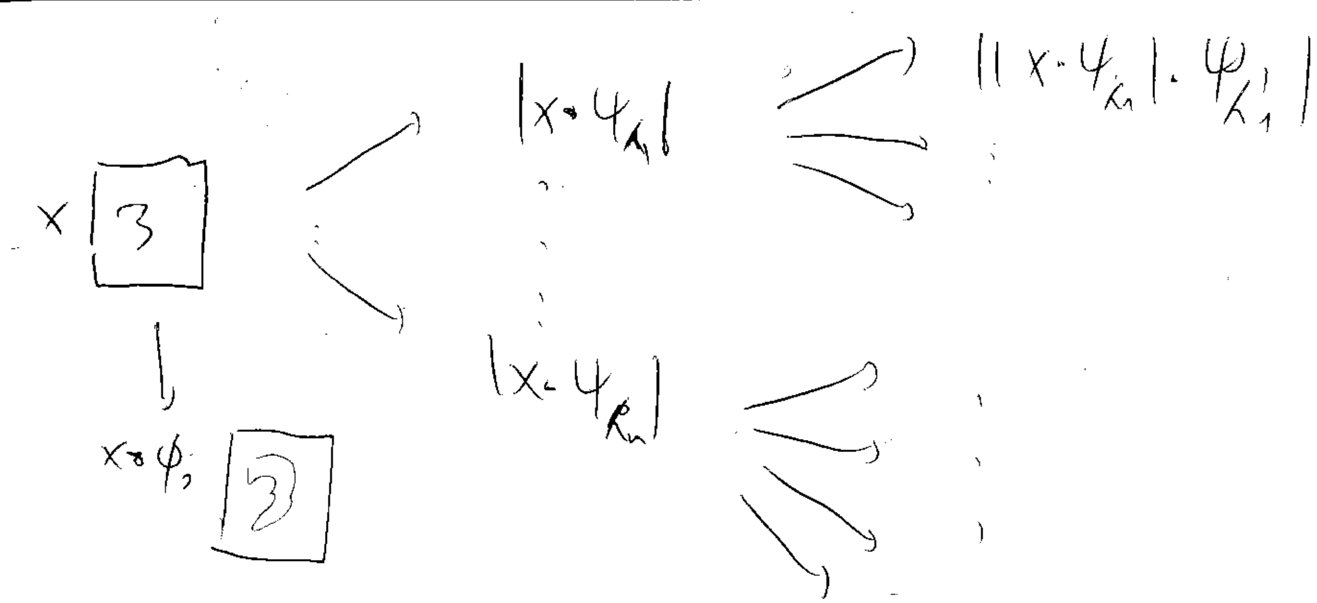
Every time we extract an invariant, we loose information

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

A new wavelet decomposition captures lost high-frequencies!

$$\int ||x*\psi_\lambda|*\psi_{\lambda'}|(u)\,du$$

Def: For each "path" $p = (\lambda_1 \cdots \lambda_m)$ with $\lambda_i \in \Lambda_J$ and x, the scattering transform of x is



→ Scattering Metric and Energy conservation.

For a given set of paths T, the Euclidean norm defined by scattering coefficients is

$$\|S_J[T]x\|^2 := \sum_{p\in T} \|S_J[p]x\|^2, \quad x \in L^2(\Omega).$$

is it well-defined? Let Λ_J^m denote all paths of length m, and denote by

$$\bar{U}_J x = \{x*\phi_J, |x*\psi_\lambda|\}_\lambda \quad \text{the "propagator" operator}$$

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

$$\|\bar{U}_J x - \bar{U}_J x'\|^2 = \|x*\phi_J - x'*\phi_J\|^2 + \sum_\lambda \||x*\psi_\lambda| - |x'*\psi_\lambda|\|^2 \quad (*)$$
$$\le \|x - x'\|^2.$$

But observe that scattering coeffs are constructed by cascading $\bar{U}_J$: if $U[p]x := ||\cdots x*\psi_{\lambda_1}|*\cdots \psi_{\lambda_m}|$, $p = (\lambda_1 \cdots \lambda_m)$

$\mathcal{P}_J$: set of all paths. $\mathcal{P}_J = \bigcup_m \Lambda_J^m$

Λ_J^m: paths of length m.

(Induction over m. $\|S_J x - S_J x'\| \le \|x - x'\|$.)

Apply (*) to each $|x * \psi_\lambda|$, $|x' * \psi_\lambda|$.

Lecture 4: Scattering Extensions, Graph Neural Networks.

↳ Consider now unitary wavelet decompositions:

(*) $\|x\|^2 = \|x * \phi_J\|^2 + \sum_\lambda \|x * \psi_\lambda\|^2$.

$\boxed{\text{Thm: For appropriate wavelets, } \|x\|^2 = \|Sx\|^2.}$

By applying (*) on each output $x_p := \| |x * \psi_{\lambda_1}| * \cdots * \psi_{\lambda_s}|$

we have

$$\forall m, \quad \|x\|^2 = \sum_{|p|<m} \|S_J[p]x\|^2 + \sum_{|p|=m} \| |x * \psi_{\lambda_1}| * \cdots * \psi_{\lambda_m}| \|^2$$

the result amounts to showing that

$$\lim_{m\to\infty} \sum_{|p|=m} \| |x * \psi_{\lambda_1}| \cdots |\psi_{\lambda_m}| \|^2 = 0.$$

- Energy moves progressively towards low frequencies
- Decay is Exponential for band-limited signals [Waldspurger '17, Czaja-Li '16].

[illegible]

· Geometric Stability of Euclidean Scattering

~~recall, $\|\nabla\tau\| \ldots$~~

recall $\|\tau\| := 2^{-J}\|\tau\|_\infty + \|\nabla\tau\|_\infty + \|H\tau\|_\infty$

Thm. [Mallat '12] There exists $C>0$ st for all $x \in L^2(\mathbb{R}^d)$ and all m, the m-th order scattering representation satisfies $\|S_J x - S_J x_\tau\| \le C \cdot m \|x\| \|\tau\|$.

Proof architecture:

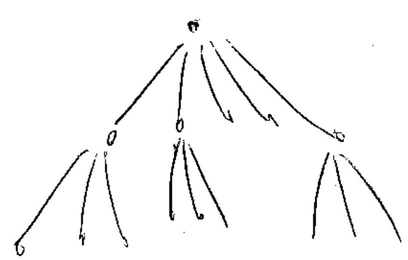
- Denote $Ax = x * \phi_J$, $Wx = \{x * \psi_\lambda\}$, $Mx = |x|$.
- We know that
$$\|A - A\varphi_\tau\| \le C\|\tau\|$$
$$\|W\varphi_\tau - \varphi_\tau W\| \le C\|\tau\|$$
$$M\varphi_\tau = \varphi_\tau M \qquad \left(\begin{array}{l} \|[W,\varphi_\tau]\| \le C\|\tau\| \\ [M,\varphi_\tau] = 0 \end{array}\right).$$
- By def, the operator S_J is
$\{A, AMW, AMWMW, \ldots\}$ (m times).
- $\|S_J - S_J\varphi_\tau\|^2 = \|A - A\varphi_\tau\|^2 + \|AMW - AMW\varphi_\tau\|^2 + \cdots$
$\cdots + \|A(MW)^{\otimes m} - A(\underbrace{MW}_{U})^m L_\varphi\|^2$
- $\|AU^k - AU^k\varphi_\tau\| \le \|AU^k - AU^{k-1}\varphi_\tau U\| +$
$\|AU^{k-1}\varphi_\tau U - AU^k\varphi_\tau\|$

$$\leq \| AU^{k-1} - AU^{k-1}\varphi_\tau \| + \| AU^{k-1}[U, \varphi_\tau] \|$$

$$\leq \| AU^{k-1} - AU^{k-1}\varphi_\tau \| + C\|\tau\|$$

$$\leq K C \|\tau\| + \| A - A\varphi_\tau \| \leq \tilde{C} K \|\tau\| .$$

<u>Limitations of separable scattering</u>



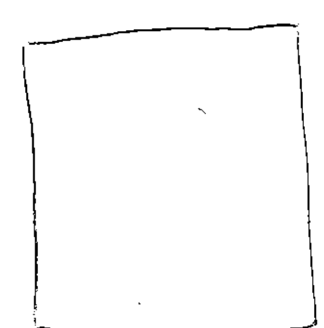
→ This is the specific convolutional architecture given by $Sx = \{ S_J[p]x \; ; \; |p| \leq m \}$.

↳ No feature dimensionality reduction: each new layer increases the number of feature maps.

↳ Each wavelet band is assumed to be deformed independently.

↳ No adaptivity.

- Joint vs Separable Invariance.

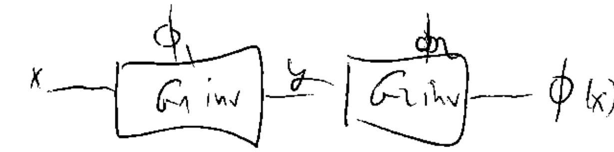


$$G = S^1 \times S^1$$

Each copy of S^1 acts on a different coordinate (horiz vs vertical).

$\varphi^1 x(u_1, u_2) = x(u_1 - a, u_2) \; ; \; \varphi^2 x(u_1, u_2) = x(u_1, u_2 - a)$

$$G = G_1 \times G_2$$



$y = \phi_1(x)$ such that $\phi_1(\varphi^1 x) = \phi_1(x) \; \forall \varphi^1 \in G_1$

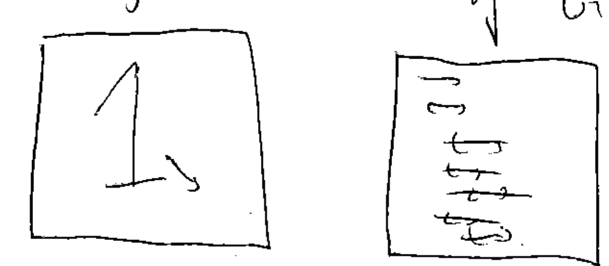
$\phi_2(y)$ such that $\phi_2(\varphi^2 x) = \phi_2(x) \; \forall \varphi^2 \in G_2$.

So, if ϕ_1 is $\left\{ \begin{array}{l} G_1 \text{ invariant} \\ G_2 \text{ equivariant} \end{array} \right.$ and ϕ_2 is G_2 invariant, then $\phi = \phi_2 \circ \phi_1$ is G-invariant.

$$\phi_2 \phi_1(\varphi^1 \varphi^2 x) = \phi_2 \phi_1(\varphi^2 x) = \phi_2 \varphi^2 \phi_1(x) = $$
$$= \phi_2 \phi_1(x) = \phi(x).$$

So we can compose invariances by working in separate group structures.

However, this factorization does not take into account the joint action of G along (u_1, u_2).



<u>Ex</u>: Roto-translation group. $G = \mathbb{R}^2 \times S^1$
(u, α)

$g \in G$, $g = (u, \alpha)$

$$(g \cdot x)(u) = x(R_\alpha T_v u)$$

Non-Commutative:

$$g \cdot g' = R_\alpha \cdot T_v R_{\alpha'} T_{v'} \neq R_{\alpha'} T_{v'} R_\alpha T_v$$

This group acts on 1st-layer features similarly as translation acts on input images:

$$\rho(\underset{(v,\alpha)}{\underset{\shortparallel}{g}} \cdot x * \psi_\lambda) = X_1(R_\alpha T_v u;\ j, \alpha + \theta).$$

↳ so we can replace separable spatial convolution convolution with roto-transl. convolution:

def Let G be a compact group equipped with the Haar measure $d\mu$, acting on Ω, and $h \in L^1(G)$,

$$x *_G h(u) = \int_G x(\varphi_g u) \cdot h(g) d\mu(g)$$

↳ If $x = x(u, j, \theta)$ and G are roto-translations, this convolution recombines the channels (not separable across orientation).

↳ Scattering with roto-translation wavelets:

[Sifre & Mallat '13] [Oyallon & Mallat '15].

→ Proposals: Give feedback this week.

↳ Inverse Curriculum; Aaron leave of absence.

→ Applications of Scattering.

- In Computer vision: [B & Mallat '11/12] MNIST/ Texture.
  - [Sifre & Mallat '13] Texture.
  - [Oyallon & Mallat '15] Cifar 10/100.
- Speech Recognition / Music / Audio Processing: [B. et al '16]: Super-Resolution. [Anden & Mallat '14] [Anden & Lostanlen '16].
- Multifractal Analysis: [B et al. '14, '15]
- Unsupervised Quantum Chemistry: [Eickenberg & Hirn, '16]
- Unsupervised Learning [Angles & Mallat '18] GAN
  - [B & Mallat '18]: Geometric Probability Theory.

Lecture 5:

From Scattering to CNNs.

- The essential stability properties of scattering stem from the commutation properties $\|[W, L_\tau]\| \leq \|\tau\|$ and $\|A - AL_\tau\| \leq \|\tau\|$.

$W = \{\psi_\lambda\}$ contains a family of differential operators $D_\lambda x = x * \psi_\lambda$ which nearly commute with deformation.

- A CNN can be written as

$$x^{(l+1)}_{\lambda'} = \rho\left[\sum_\lambda D_\lambda x^{(l)} \cdot \theta_{\lambda,\lambda'}\right]$$

We learn linear combinations of stable differential

From Euclidean to Non-Euclidean Stability

↳ so far, we measured geometric stability in terms of diffeomorphisms $\varphi : \Omega \to \Omega$, in the sense that $\varphi(u)$ "close" to a rigid translation should satisfy
$\| \phi(x) - \phi(\varphi x) \| \simeq 0$.

↳ Q: If Ω is not Euclidean / not even continuous?
We can think of a diffeo as a change of metric.

$$(\Omega, d\mu) \xrightarrow{\varphi} (\Omega, \varphi(d\mu))$$

$$\langle x, x' \rangle = \int x(u)\, x'(u)\, \mu(du)$$

$$\langle x, x' \rangle_\varphi = \int x(\varphi(u)) \cdot x'(\varphi(u))\, d\mu(du) =$$

$$= \int x(u) \cdot x'(u)\, | I - \nabla \varphi(u) |\, d\mu(du)$$

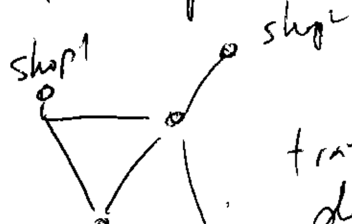
↳ $d\tilde{\mu}(u) = | I - \nabla \varphi(u) |\, d\mu(u)$ is a change of metric.

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

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

$\mathcal{X} = (\Omega, d)$ ↳ distance.

Eg:



travel times depend on

$\mathcal{X}^{(t=0)}, \mathcal{X}^{(t=1)} \ldots$

$\mathrm{dist}(\mathcal{X}^{(0)}, \mathcal{X}^{(1)})$ measures how close the metrics are.

Stability to metric changes:

$$\| \Phi_{\mathcal{X}} - \Phi_{\mathcal{X}'} \| \lesssim \mathrm{dist}(\mathcal{X}, \mathcal{X}')$$

Q: How to define meaningful distances? Graphs are flexible data structures to describe discrete metric domains.

→ Diffusion Wavelets on Graphs.

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

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

We focus mostly on undirected weighted graphs:

Let $|V| = n$, and $W \in \mathbb{R}^{n \times n}$ be a symmetric weighted adjacency matrix. $D = \mathrm{diag}(W\mathbb{1})$ degree matrix.

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

$\tilde{A} = D^{-1} W$ is a Markov Chain,

$\tilde{A}$ and A are similar ⇒ they have the same eigenvalues

→ $\|A\| \leq 1$

[Coifman, Lafon, Nadler, '06]

def: Diffusion distance at time s between two nodes is $d^s_G(x,x') = \| A^s \delta_x - A^s \delta_{x'} \|$

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

$$d^s(G,G') = \inf_{\Pi \in \Pi_n} \| A^{2s} - \Pi^T (A')^{2s} \Pi \|$$

↳ As s increases, distance is weaker (why?)

↳ Our distance results in a stronger topology than another usual distance defined over metric space, the Gromov-Hausdorff distance:

$$d_{GH}(G,G') = \inf_{\Pi \in \Pi_n} \sup_{x,x'} | d^s_G(x,x') - d^s_{G'}(\pi(x),\pi(x')) |$$

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

Graph Diffusion Scattering

Q: We saw before that scattering transforms in the Euclidean domain are stable to deformations (= metric changes).

Can we extend that to non-Euclidean setting?

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

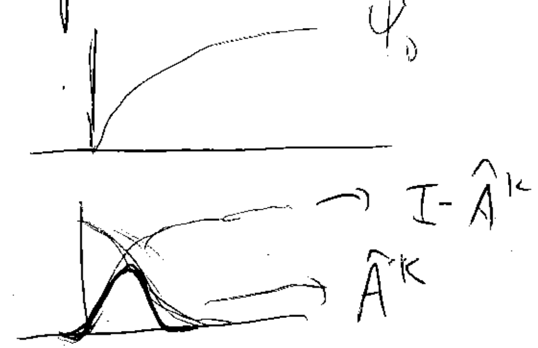
↳ We need the equivalent of wavelets.

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:

$$\psi_0 = I - A, \quad \psi_j := A^{2^{j-1}} (I - A^{2^{j-1}}) \quad j > 0.$$

↳ A is a low-pass filter: we are building multiscale filter bank by combining diffusions at different time-scales.



Def: Diffusion Scattering transform:

$$\Psi : L^2(G) \to (L^2(G))^J$$
$$x \longmapsto \{ \psi_j x \}_j$$

$$\Phi(x) = \{ A^J x,\ A^J \rho(\Psi x),\ A^J \rho(W \rho(Wx)), \dots \}.$$

Thm: [Gama, ~~B~~ Ribeiro, B. '19]: The graph scattering representation is stable to metric perturbations, as measured by graph diffusion metric:

$$\|\Phi_G(x) - \Phi_{G'}(x)\| \leq \sqrt{n}\, d(G, G') \;\|x\|$$

GNN: Instead of predefining how to combine normalize the low-pass diffusion $A = D^{-1}W$, we replace it by a data-driven diffusion.

Part II: Optimization and Geometry in DL

- Prototypical problem in ~~stat~~ ML is of the form $\min_{x \in X} f(x)$ (exceptions: saddle point problems, e.g. adversarial training)
- Characteristic: $X \subseteq \mathbb{R}^d$, d huge.
- Worst-case performance: searching for minima is hard. exponential in d. How to overcome this?

First important class of functions that overcome curse of dimensionality: convex functions.

Assume for now $X = \mathbb{R}^d$.

Def: f convex if $\forall x, y, t$, $f(tx + (1-t)y) \leq t f(x) + (1-t) f(y)$.

If f is differentiable, then

$$\frac{f(tx + (1-t)y) - f(y)}{t} \leq f(x) - f(y)$$

$t \to 0$ ↓

$$\langle \nabla f(y), x - y \rangle \leq f(x) - f(y).$$

In general, we can extend this notion to non-differentiable functions, using the notion of sub-gradients:

def Let $f: \mathbb{R}^d \to \mathbb{R}$. $g \in \mathbb{R}^d$ is a subgradient of f at x if $\forall y$, $f(x) - f(y) \leq \langle g, x - y \rangle$.

fact: If f is convex, then $\partial f(x)$ is non-empty. $\partial f(x) = \{ g ;\ g \text{ is a subgradient at } x \}$. If f is differentiable and convex, then $\nabla f(x) \in \partial f(x)$.

Gradient Descent for smooth functions

→ f continuously differentiable function is β-smooth if ∇f is β-Lipschitz: $\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$.

→ if f is twice differentiable, $\nabla^2 f \preceq \beta I$.

Thm: f convex and β-smooth in $\mathbb{R}^d$. Then gradient descent $x_{t+1} = x_t - \eta \nabla f(x_t)$ with $\eta = \beta^{-1}$ satisfies

$$f(x_t) - \underbrace{\min_x f(x)}^{f(x^*)} \leq \frac{2\beta \|x_1 - x^*\|^2}{t-1}$$

We need two structural lemma's first.

Lemma: f β-smooth. Then $\forall x, y$

$$|f(x)-f(y) - \langle \nabla f(y), x-y \rangle| \leq \frac{\beta}{2}\|x-y\|^2$$

$$f(x)-f(y) = \int_0^1 \langle \nabla f(y+t(x-y)), x-y \rangle \, dt$$

$$|f(x)-f(y) - \langle \nabla f(y), x-y \rangle| =$$

$$= \left| \int_0^1 \langle \nabla f(y+t(x-y)) - \nabla f(y), x-y \rangle \, dt \right|$$

$$\leq \int_0^1 \|x-y\| \, \|\nabla f(y+t(x-y)) - \nabla f(y)\| \, dt$$

$$\leq \int_0^1 \|x-y\|^2 t \beta \, dt = \frac{\beta}{2}\|x-y\|^2. \quad \square$$

↳ If f is convex, since $\nabla f(x) \in \partial f(x)$, we have

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

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

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

~~f(x-β⁻¹∇f(x)) ...~~

$$f(x-\beta^{-1}\nabla f(x)) - f(x) + \langle \nabla f(x), \beta^{-1}\nabla f(x) \rangle \leq \frac{\beta}{2}\|\beta^{-2}\nabla f(x)\|^2$$

$$f(x-\beta^{-1}\nabla f(x)) - f(x) + \beta^{-1}\|\nabla f(x)\|^2 \leq \frac{\beta^{-1}}{2}\|\nabla f(x)\|^2$$

$$f(x-\beta^{-1}\nabla f(x)) - f(x) \leq -\frac{\beta^{-1}}{2}\|\nabla f(x)\|^2.$$

Lemma: (co-coercivity). For f convex and β-smooth, we have

$$f(x) - f(y) \leq \langle \nabla f(x), x-y \rangle - \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2$$

$$z = y - \frac{1}{\beta}(\nabla f(y) - \nabla f(x)).$$

$$f(x) - f(y) = f(x) - f(z) + f(z) - f(y)$$

$$\leq \langle \nabla f(x), x-z \rangle + \langle \nabla f(y), z-y \rangle + \frac{\beta}{2}\|z-y\|^2$$

$$= \langle \nabla f(x), x-y \rangle + \langle \nabla f(x) - \nabla f(y), y-z \rangle + \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2$$

$$= \langle \nabla f(x), x-y \rangle - \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 \quad \square$$

Proof of thm: Using one-step improvement we have

$$f(x_{s+1}) - f(x_s) \leq -\frac{1}{2\beta}\|\nabla f(x_s)\|^2.$$

$$\delta_s = f(x_s) - f(x^*) \rightarrow \delta_{s+1} \leq \delta_s - \frac{1}{2\beta}\|\nabla f(x_s)\|^2$$

using convexity

$$\delta_s \leq \langle \nabla f(x_s), x_s - x^* \rangle \leq \|x_s - x^*\| \, \|\nabla f(x_s)\|$$

Assume for now that $\|x_s - x^*\|$ in decreasing s.

Then $\|\nabla f(x_s)\| \geq \frac{\delta_s}{\|x_s - x^*\|} \geq \frac{\delta_s}{\|x_1 - x^*\|}$

$$\Rightarrow \delta_{s+1} \leq \delta_s - \frac{\delta_s^2}{2\beta \, \|x_1 - x^*\|^2}, \quad (2\beta \|x_1 - x^*\|^2)^{-1} = w$$

$$\Rightarrow w\delta_s^2 + \delta_{s+1} \leq \delta_s \Rightarrow \frac{w\delta_s}{\delta_{s+1}} + \frac{1}{\delta_s} \leq \frac{1}{\delta_{s+1}} \Rightarrow \frac{1}{\delta_{s+1}} - \frac{1}{\delta_s} \geq w$$

$$\Rightarrow \frac{1}{\delta_t} \geq w(t-1)$$

using previous lemma, we have

$$\frac{1}{\beta}\|\nabla f(x)-\nabla f(y)\|^2 \le \langle \nabla f(x)-\nabla f(y),\ x-y\rangle$$

Then $\|x_{s+1}-x^*\|^2 = \|x_s - \beta^{-1}\nabla f(x_s) - x^*\|^2$

$$= \|x_s-x^*\|^2 - \frac{2}{\beta}\langle \underset{\nabla f(x^*)}{\nabla f(x_s)},\ x_s-x^*\rangle + \frac{1}{\beta^2}\|\nabla f(x_s)-\nabla f(x^*)\|^2$$

$$\le \|x_s-x^*\|^2 - \frac{2}{\beta^2}\|\nabla f(x_s)\|^2 + \frac{1}{\beta^2}\|\nabla f(x_s)\|^2$$

$$\le \|x_s-x^*\|^2 \quad \square$$

↳ Rate $1/t$ is <u>not</u> optimal in the class of β-smooth convex functions (we will see this later).

· But first, let's push smoothness a bit further:

<u>def</u>: f is α-strongly convex if

$$f(x)-f(y) \le \langle \nabla f(x),\ x-y\rangle - \frac{\alpha}{2}\|x-y\|^2$$

↳ if f is twice differentiable, $\nabla^2 f(x) \succeq \alpha I$.

↳ Strong convexity + smoothness means we can sandwich $f(y)$ between two quadratic forms.

$$\begin{cases} q_x^-(y) = f(x) + \langle \nabla f(x),\ y-x\rangle + \frac{\alpha}{2}\|x-y\|^2 \\ q_x^+(y) = f(x) + \langle \nabla f(x),\ y-x\rangle + \frac{\beta}{2}\|x-y\|^2 \end{cases}$$

$$q_x^-(y) \le f(y) \le q_x^+(y)$$

→ denote $K = \frac{\beta}{\alpha}$ (≥ 1) the "condition" number of f.

<u>Lemma</u>: f β-smooth, α-strongly convex. Then

$$\langle \nabla f(x)-\nabla f(y),\ x-y\rangle \ge \frac{\alpha\beta}{\alpha+\beta}\|x-y\|^2 + \frac{1}{\beta+\alpha}\|\nabla f(x)-\nabla f(y)\|^2$$

Recall that if f is convex and β-smooth, then

$$\langle \nabla f(x)-\nabla f(y),\ x-y\rangle \ge \frac{1}{\beta}\|\nabla f(x)-\nabla f(y)\|^2.$$

If f is β-smooth, α-strongly convex, then $\varphi(x) := f(x) - \frac{\alpha}{2}\|x\|^2$ is convex and $(\beta-\alpha)$ smooth.

$$\Rightarrow \underset{\langle \nabla f(x)-\nabla f(y) - \alpha(x-y),\ x-y\rangle}{\underset{\|}{\langle \nabla\varphi(x)-\nabla\varphi(y),\ x-y\rangle}} \ge \frac{1}{\beta-\alpha}\|\nabla\varphi(x)-\nabla\varphi(y)\|^2$$

$$\langle \nabla f(x)-\nabla f(y),\ x-y\rangle \ge \alpha\|x-y\|^2 + \frac{1}{\beta-\alpha}\Big[\|\nabla f(x)-\nabla f(y)\|^2 + \alpha\|x-y\|^2 - 2\alpha\langle \nabla f(x)-\nabla f(y),\ x-y\rangle\Big]$$

$\square$

<u>Theorem</u>: f β-smooth and α-strongly convex; $\eta = \frac{2}{\alpha+\beta}$ gives

$$f(x_{t+1}) - f(x^*) \le \frac{\beta}{2}\exp\left(-\frac{4t}{K+1}\right)\|x_1-x^*\|^2$$

Proof. By smoothness, $f(x_t) - f(x^*) \le \frac{\beta}{2}\|x_t-x^*\|^2$

$$\|x_{t+1}-x^*\|^2 = \|x_t - \eta\nabla f(x_t) - x^*\|^2$$

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

$$\le \left(1 - 2\frac{\eta\alpha\beta}{\alpha+\beta}\right)\|x_t-x^*\|^2 + \left(\eta^2 - 2\frac{\eta}{\beta+\alpha}\right)\|\nabla f(x_t)\|^2$$

$$\le \left(K-1\right)^2 \|x_t - x^*\|^2 \le \exp\left(-4t \ldots\right)$$

Oracle lower bounds:

↳ Smooth case: there exists β-smooth convex function f st for any black-box procedure with $x_{t+1} \in \text{span}\{g_1 \dots g_t\}$,

$$\min_{s \le t} f(x_s) - f(x^*) \geq \frac{3}{32}\beta \frac{\|x_1 - x^*\|^2}{(t+1)^2} \quad \left(t < \frac{n}{2}\right)$$

→ Remark: function is a Dirichlet Energy (quadratic).

↳ smooth + strongly convex case: $\exists$ β-smooth and α-strongly convex f st for any t,

$$f(x_t) - f(x^*) \geq \frac{\alpha}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2(t+1)} \|x_1 - x^*\|^2$$

Gap: How to close it?

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

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

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

$= x_t - \eta H(x_t - x_*) = [I - \eta H] x_t + \eta \underbrace{H x_*}_{b}$

Suppose $x_0 = 0$; $x_{t+1} = \left(\sum_{k \le t} [I - \eta H]^k\right) \eta b$

$\alpha I \leq H \leq \beta I \Rightarrow (1-\eta\beta) I \leq I - \eta H \leq (1-\eta\alpha) I$

$\|I - \eta H\| < 1$ if $\eta < \beta^{-1}$ $\Rightarrow (\eta H)^{-1} = [I - (I - \eta H)]^{-1} = \sum_{k \ge 0}^{\infty} [I - \eta H]^k$

$x_{t+1} = H_t \eta b = [(\eta H)^{-1} + H_t - (\eta H)^{-1}] \eta b$

$= x_* + (H_t - (\eta H)^{-1}) \eta b$

$\|H_t - (\eta H)^{-1}\| = O(\|(I - \eta H)^t\|) = O\left(\left(1 - \frac{\alpha}{\beta}\right)^t\right)$

$= O((1 - \kappa^{-1})^t)$. $\square$

→ Discrete vs continuous time optimization.

- Consider $x_{k+1} = x_k - \gamma \nabla f(x_k)$ $\quad x_1 \; x_2 \; x_3$

Q: What happens when $\gamma \to 0$? Do trajectories "pile-up"?

- Assume ∇f is bounded and Lipschitz (smooth).
- Introduce the Ansatz $x_k \approx X(k \cdot \gamma)$, where $X(t)$ is a smooth curve defined for $t \geq 0 \in \mathbb{R}_+$. Let $\frac{t}{\gamma} = k$

$X(t) \leftrightarrow x_{t/\gamma} = x_k$

$X(t+\gamma) \leftrightarrow x_{k+1}$

$$\frac{x_{k+1} - x_k}{\gamma} = -\nabla f(x_k)$$

$$\downarrow$$

$$\frac{X(t+\gamma) - X(t)}{\gamma} = -\nabla f(X_t)$$

$$\downarrow$$

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

↳ Gradient Flow.

Gradient Descent is a forward-Euler discretization we can get a first understanding of optimization algorithms by first looking at their behavior for $\eta \to 0$ continuous time.

Ex. $f(x) = \frac{1}{2} x^T H x$, $\quad H$ (~~symmetric~~) pd : $H \succ \alpha I$.

$$\dot{x}(t) = -Hx(t) \Rightarrow x(t) = e^{-tH} x(0)$$

$$\Rightarrow \|x(t)\| \le \|x(0)\| e^{-t\alpha}$$

Ex: $f(x)$ convex, smooth.

$\dot{x}(t) = -\nabla f(x(t))$. Recall the convergence rate we obtained:

$$f(x_k) \le \frac{\beta C \|x_1 - x^*\|^2}{k} = \frac{C\|x_1 - x^*\|^2}{k/\beta} \quad k\cdot\eta \quad t$$

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

Consider the ~~Lyapunov~~ function

$$L(t) = t \cdot [f(X(t)) - f(x^*)] + \frac{1}{2}\|X(t) - x^*\|^2$$

$$\dot{L}(t) = f(X(t)) - f(x^*) + t\langle \nabla f(X(t)), -\nabla f(X(t))\rangle + \langle \nabla f(X(t)), X(t) - x^* \rangle$$

$$= \underbrace{-t\|\nabla f(X_t)\|^2}_{\le 0} + \underbrace{f(X(t)) - f(x^*) + \langle \nabla f(X_t), x^* - X_t \rangle}_{\le 0 \text{ by convexity}}$$

$$L(t) \le L(0) = \frac{1}{2}\|X(0) - x^*\|^2$$

$$\vee\!|$$

$$t[f(X(t)) - f(x^*)] \Rightarrow f(X(t)) - f(x^*) \le \frac{\|X(0) - x^*\|^2}{2t}$$

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

↳ A reasonable black-box procedure is a mapping from the past history to the next point:

$$(x_1, g_1, x_2, g_2, \dots, x_t, g_t) \rightsquigarrow x_{t+1}$$

we consider the "linear" case, where we ask

$$x_{t+1} \in \mathrm{Span}(g_1 \dots g_t)$$

Q: How to close this gap?

A: Through the notion of acceleration.

Consider

$$x_{k+1} = y_k - \eta \nabla f(y_k)$$

$$y_{k+1} = x_{k+1} + \frac{k}{k+3}(x_{k+1} - x_k)$$

<u>Theorem</u> (Nesterov '83):

$$f(x_t) - f^* \le \frac{2\|x_0 - x_*\|^2}{(t+1)^2 \eta}$$

→ Acceleration for α-strongly convex, β-smooth functions: just need to slightly modify scheme:

$$x_{k+1} = y_k - \beta^{-1}\nabla f(y_k)$$

$$y_{k+1} = x_{k+1} + \frac{1-\sqrt{\alpha/\beta}}{1+\sqrt{\alpha/\beta}}\,(x_{k+1} - x_k)$$

Theorem (Nesterov): $f(x_k) - f(x^*) \le \beta\left(1-\sqrt{\frac{\alpha}{\beta}}\right)^k \|x_0 - x_*\|^2$

$\left(\sim (1-\sqrt{\kappa^{-1}})^k\right.$ vs $\left.(1-\kappa^{-1})^k \text{ from GD}\right)$.

↳ continuous-time interpretation?

↳ What is the corresponding ODE?

$$\ddot X + \frac{3}{t}\dot X + \nabla f = 0$$

↳ <u>Theorem:</u> [Su, Boyd, Candès, '15] The ODE has unique global solution in $X \in C^2((0,\infty),\mathbb{R}^d) \cap C^1([0,\infty)\,\mathbb{R}^d)$.

Moreover, as $\eta \to 0$, Nesterov's scheme converges to the ODE: for all fixed $T > 0$,

$$\lim_{\eta\to 0}\ \sup_{k \le T/\sqrt{\eta}} \|x_k - X(k\sqrt{\eta})\| = 0$$

Simple <u>Consequence</u>: we show easily that we get $1/t^2$ convergence rate.

<u>Theorem</u> [S,B,C]: Consider now

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

$$\dot L(t) = 2t\left(f(X(t)) - f^*\right) + t^2\langle \nabla f, \dot X\rangle + 4\left\langle X + \tfrac{t}{2}\dot X - x_*,\ \tfrac{3}{2}\dot X + \tfrac{t}{2}\ddot X\right\rangle$$

since $\ddot X + \frac{3}{t}\dot X = -\nabla f$,

$$\tfrac{3}{2}\dot X + \tfrac{t}{2}\ddot X = -\tfrac{t}{2}\nabla f$$

$$\dot L(t) = 2t\left[f(X(t)) - f^*\right] + t^2\langle\nabla f, \dot X\rangle - 2t\left\langle \nabla f, X - x_* + \tfrac{t}{2}\dot X\right\rangle$$

$$= 2t\left(f(X(t)) - f^*\right) - 2t\langle \nabla f, X - x_*\rangle \le 0. \quad \square$$

↳ Interpretation in the strongly convex case: Chebyschev polynomials.

Recall we were approximating the inverse matrix $(\eta H)^{-1}$ with a finite power series $\sum_j [I - \eta H]^j$

$\|\underbrace{(\eta H)^{-1}}_{A} - \sum_{j\le k}[I-\eta H]\| \to 0$ as $k \to \infty$, with error $O((1-\kappa)^k)$

Q: What is the best polynomial of degree k to approximate A? $\Leftrightarrow \min \|I - A\,p(A)\|$

Lemma: (Chebyshev polynom): There exist a polynomial q_k of degree $O\left(\sqrt{\left(\frac{\beta}{\alpha}\right)\log\left(\frac{1}{\varepsilon}\right)}\right)$ such that

$q_k(0) = 1$ and $|q_k(x)| \leq \varepsilon \ \forall x \in [\alpha, \beta]$.

q_k is a Chebyshev polynomial, computed recursively from q_{k-1}, q_{k-2}. $\Rightarrow$ Nesterov Scheme.

→ Further explain Chebyscher.

$$\min \| A^{-1} - q_k(A) \| \cong \min \| \underbrace{I - A q_k(A)}_{P_k(A)} \|$$

- $P_k(A)$ commutes with A;

  λ eigenvalue of A $\Leftrightarrow$ $P_k(\lambda) = 1 - \lambda q_k(\lambda)$ eigenvalue of $P_k(A)$.

- Eigenvalues of A are in $\lambda \in [\alpha, \beta]$.

- $\min\limits_{x \in (\alpha, \beta)} |P_k(x)|$ st $P_k(0) = 1$ → Chebyscher polynomials.

  ↳ ∃ polynomial P_k of degree $O\left(\sqrt{\left(\frac{\beta}{\alpha}\right)\log(1/\varepsilon)}\right)$ st $P_k(0) = 1$ and $|P_k(x)| \leq \varepsilon$ for $x \in (\alpha, \beta)$.

→ Today: Further on discrete vs continuous time.

- Stochastic Gradient Descent.

↳ Previously: take a discrete scheme and derive a continuous time limit.

↳ Q: Can we take opposite route?

- given convex function g, we need two steps:

1) Construct an ODE such that its solutions $X(t)$ satisfy $g(X(t)) - g(\theta_*) \lesssim t^{-p}$

2) Discretize the ODE such that convergence rate is preserved.

A note on discretization of ODEs.

$\frac{dy}{dt} = f(t, y)$ integrate from t_n to $t_{n+1} = t_n + h$

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

Two basic strategies:

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

$y(t_{n+1}) = y(t_n) + h f(t_n, y(t_n))$: Explicit Euler Scheme.

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

~~$y(t_{n+1})$~~ $y_{n+1} = y_n + h f(t_{n+1}, y_{k+1})$ Implicit

The Bregman Lagrangian [Wisibono et al. '16]

↳ we assume ~~of~~ $\min_{\theta\in\Theta} g(\theta)$ admits a unique minimiser θ^*.

↳ let h be another convex function.

df: A Bregman divergence in Θ is

$$D_h(\theta;y) = h(\theta) - h(y) - \langle \nabla h(y), \theta - y \rangle$$

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

Locally like a Hessian metric:

$$D_h(\theta,y) = \frac{1}{2}(\theta-y)^T \nabla^2 h(y)(\theta-y) + o(\|\theta-y\|).$$

df: Bregman Lagrangian is

$$\mathcal{L}(X,V,t) := e^{\alpha(t)+\gamma(t)}\left(D_h(X+e^{-\alpha(t)}V, X) - e^{+\beta(t)} g(X)\right)$$

X position
V velocity
t time,

with $\dot\beta(t) \le e^{\alpha(t)}$
$\dot\gamma(t) = e^{\alpha(t)}$

$X_t \in \Theta$, $t \geq 0$.

Action on the path is $J(X) = \int_{t\geq 0} \mathcal{L}(X_t, \dot X_t, t)\,dt$

Lagrangian system: ~~minimize act~~ curves minimizing

satisfy the Euler-Lagrange equation:

$$\frac{d}{dt}\frac{\partial \mathcal{L}}{\partial V} = \frac{\partial \mathcal{L}}{\partial X};$$ by plugging the Bregman Lagrangian it becomes

$$\frac{d}{dt}\nabla h\left(X_t + e^{-\alpha(t)}\dot X_t\right) = -e^{\alpha(t)+\beta(t)}\nabla g(X_t)$$

Q: Do these solutions minimize g? How fast?

Theorem: [Wisibono et al '17] The solutions ~~of~~ to the Euler-Lagrange satisfy $g(X_t) - g(\theta_*) \le O(e^{-\beta(t)})$

↳ Consider the Lyapunov function

$$\mathcal{E}_t = D_h\left(\theta^*, X_t + e^{-\alpha(t)}\dot X_t\right) + e^{\beta(t)}\left(g(X_t) - g(\theta_*)\right)$$

~~[illegible]~~ Show as before that $\dot{\mathcal{E}}(t) \le 0$.

Optimal rate is achieved with the relation $\dot\beta(t) = e^{\alpha(t)}$

Q: How to discretize these ODEs while preserving rate?

$\alpha(t) = \log p - \log t$, $\beta(t) = p\log t + \log C$, $\gamma(t) = p\log t$

Resulting Euler-Lagrange becomes $p>0$, $C>0$.

$$\ddot X_t + \frac{p+1}{t}\dot X_t + Cp^2 t^{p-2}\left[\nabla^2 h(X_t + tp^{-1}\dot X_t)\right]^{-1}\nabla g(X_t) = 0$$

This is equivalent to a decoupled system of 1st order eqs:

$$(*)\quad \begin{cases} Z_t = X_t + \frac{t}{p}\dot{X}_t \\ \frac{d}{dt}\nabla h(Z_t) = -C p t^{p-1} \nabla g(X_t) \end{cases}$$

A "naive" idea is to use a standard discretization scheme for (*); eg forward-backward Euler.

$$\begin{cases} y_k = \arg\min_z \{ C p k^{p-1} \langle \nabla g(\theta_k), z\rangle + \delta^{-p} D_h(z, y_{k-1}) \} \\ \theta_{k+1} = \frac{p}{k} y_k + \frac{k-p}{k}\theta_k \end{cases}$$

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

↳ As it turns out, in order to recover matching convergence, we need to ~~recover~~ have some form of co-coercivity property,

"Nesterov estimate sequence"

$$\begin{cases} \theta_{k+1} = \frac{p}{k+p}\eta_k + \frac{k}{k+p}\xi_k, \\ \eta_k = \arg\min_z \{ C p k^{p-1} \langle \nabla g(\xi_k), z\rangle + \delta^{-p} D_h(z, \eta_{k-1}) \} \end{cases}$$

ξ_k s.t. $\langle \nabla g(\xi_k), \theta_k - \xi_k \rangle \geq M \delta^{p/(p-1)} \|\nabla g(\xi_k)\|^{p/(p-1)}$ (explicit).

↳ This can be used to derive high-order ... methods.

→ Some other related works:
- Symplectic Optim. Betancourt, Wilson, Jordan
- Lyapunov analysis [Wilson, Recht, Jordan]

Stochastic Gradient Descent

→ As before, consider the minimisation of function g defined in $\mathbb{R}^d$

→ However, we no longer have access to $\nabla g(\theta)$ directly; only given access to unbiased estimates $\nabla g_n(\theta_n)$.

Ex: g_n is the loss for a single data-point:

$$g_n(\theta) = \ell(\phi(x_n;\theta), y_n)$$

$$g(\theta) = \mathbb{E}_{(x,y)\sim P}\, \ell(\phi(x;\theta), y).$$

Stochastic Approximation [Robbins & Munro '51]

General setting: find the zeros of a function $h: \mathbb{R}^d \to \mathbb{R}^d$ from random observations at certain points

→ If $h = \nabla g$ we lead to stochastic optimization.

Robbins & Munro: $\theta_n = \theta_{n-1} - \gamma_n [h(\theta_{n-1}) + \varepsilon_n]$

Questions: ① Conditions for convergence?
②

Motivating example: estimate the mean from sample.

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

$$\theta_n = (1-\gamma_n)\theta_{n-1} + \gamma_n x_n = \theta_{n-1} - \gamma_n(\theta_{n-1} - x_n)$$

(Ex: $\gamma_n = 1/n \to \theta_n = \frac{1}{n}\sum_{k\le n} x_k$

$\gamma_n = \frac{2}{n+1}$, $\theta_n = \frac{2}{n(n+1)}\sum_{k\le n} k x_k$)

→ If x_n are iid with $\mathbb{E}x_n = x$, $\mathbb{E}\|x_n - x\|^2 = \sigma^2$

$$\theta_n - x = \prod_{k\le n}(1-\gamma_k)(\theta_0 - x) + \sum_{i\le n}\prod_{k>i}(1-\gamma_k)\gamma_i(x_i - x)$$

$$\mathbb{E}\|\theta_n - x\|^2 = \prod_{k\le n}(1-\gamma_k)^2\|\theta_0 - x\|^2 + \sigma^2\sum_{i\le n}\gamma_i^2\prod_{k>i}(1-\gamma_k)^2$$

error has two contributions:

↳ ~~Appr~~ Initial conditions: $\prod_{k\le n}(1-\gamma_k) \to 0$ as $n\to\infty$.

we want to forget them.

↳ Robustness to noise: $\sum_{i\le n}\gamma_i^2\prod_{k>i}(1-\gamma_k)^2 \to 0$ as $n\to\infty$

→ If $\gamma_n \to 0$, $\log\prod(1-\gamma_k) \simeq -\sum_{k\le n}\gamma_k$, so

$\sum_{k\le n}\gamma_k$ should diverge to ∞. in order to

Ex: if $\gamma_n = Cn^{-\alpha}$,

$\alpha = 1$: $\sum_{i\le n} i^{-1} = \log n + C' + O(n^{-1})$

$\alpha > 1$: $\sum_{i\le n} i^{-\alpha} = C' + O(n^{1-\alpha})$ → init condition not forgotten.

$\alpha < 0$: $\sum i^{-\alpha} = C' n^{1-\alpha} + O(1)$.

Noise Term: assume γ_n non increasing and $\gamma_n \le \frac{1}{\mu}$.

Then $\forall m \le n$, $\sum_{i\le n}\gamma_i^2\prod_{k>i}(1-\gamma_k)^2 \cdot \sum_{i\le n}\gamma_i^2\prod_{k>i}(1-2\gamma_k)$

Suppose $\mu < 2$. Then

$$\sum_i \gamma_i^2\prod_{k>i}(1-2\gamma_k) \le \sum_i \gamma_i^2\prod_{k>i}(1-\mu\gamma_k) =$$

$$= \sum_{i=1}^{m}\prod_{k>i} + \sum_{i=m+1}^{n}\prod_{k>i}$$

$$\le \prod_{i=m+1}^{n}(1-\mu\gamma_i)\sum_{k=1}^{n}\gamma_k^2 + \gamma_m\sum_{i=m+1}^{n}\gamma_i\prod_{k>i}(1-\mu\gamma_k)$$

concavity

$$\le \exp\left(-\mu\sum_{i=m+1}^{n}\gamma_i\right)\sum_{k=1}^{m}\gamma_k^2 + \frac{\gamma_m}{\mu} \leftarrow \text{OK}.$$

↳ so we ~~need~~ require γ_n to go to 0.

Ex: $\gamma_n = C/n$ → noise term converges in $O(1/n)$

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 in the stochastic case:

- A.S. convergence: $\text{Prob}(g(\theta_n) \to g(\theta_*)) = 1$

$\Downarrow$

- Convergence in proba: $\forall \varepsilon > 0$, $\text{Prob}(|g(\theta_n) - g(\theta_*)| \geq \varepsilon) \to 0$.

$\Uparrow$

- Convergence in Moments: $\mathbb{E}|g(\theta_n) - g(\theta_*)|^r \to 0$.

→ R & M asymptotic Normality [Fabian '68]

$\gamma_n = Cn^{-1}$:

$$\mathbb{E}(\theta_n - \theta_*)(\theta_n - \theta_*)^T \simeq n^{-2CA}(\theta_0 - \theta_*)(\theta_0 - \theta_*)^T + n^{-1}C^2(2CA - I)^{-1}\Sigma$$

$A = D^2 g(\theta_*)$ $\Sigma = \mathbb{E}(\varepsilon_n \varepsilon_n^T)$ ~~sto~~ covariance
$(\nabla g_n = Dg + \varepsilon_n)$

~~Rem~~

$\Rightarrow$ $2C\lambda_{min}(A) > 1$ for convergence.

C too small: no convergence (due to memory of init conds)

C too large: large variance

Polyak-Ruppert Averaging

· Previous R&M algo ~~is~~ suffers from sensitivity of step-size, and dependence on unknown conditioning of the problem.

· We modify bias-variance tradeoff by considering an averaging over the iterates:

$$\bar{\theta}_n = \frac{1}{n}\sum_{k \leq n} \theta_k \quad \left(\bar{\theta}_n = (1 - \tfrac{1}{n})\bar{\theta}_{n-1} + n^{-1}\theta_n\right)$$

Theorem: (Cesaro's) Suppose $\theta_n \to \theta_*$ with rate $\|\theta_n - \theta_*\| \leq \alpha_n$

Then $\bar{\theta}_n \to \theta_*$ with rate $\bar{\alpha}_n \leq \frac{1}{n}\sum_{k \leq n} \alpha_k$

One can show that ~~consequently~~: if $\sum \alpha_n < +\infty$ then rate is always $\frac{1}{n}$

We lose convergence speed, but we gain robustness.

Several convergence rates for ~~th~~ different assumptions on g and choices of step-size:

Global minimax rates [Nemirovsky '83, Agarwal '12, Bach '17]

Strongly convex case: $O((\alpha n)^{-1})$

Non-strongly convex: $O(n^{-1/2})$

both attained with averaged stochastic gradient descent

Take-home: In smooth problems, $\gamma_n \sim n^{-1/2}$ + averaging gives adaptivity to strong convexity

[Bach, Moulines '11]

Q: constant step-size?

Setup with least squares optim:

$$g(\theta) = \mathbb{E}\left[(Y - \langle X, \theta \rangle)^2 \right], \quad \theta \in \mathbb{R}^d$$

- Generic Covariance $H = \mathbb{E}[XX^T]$ (No assumption p.d.!)
- constant step-size: $\gamma = 1/4R^2$, assuming $\|X\| \leq R$.

Theorem (Bach, Moulines): Averaged stochastic GD satisfies

$$\mathbb{E}\, g(\bar\theta_{n-1}) - g(\theta_*) \leq \frac{C}{n}$$

Matches statistical lower bound (Tsybakov '03).

→ The SGD step is

$$\theta_n = \theta_{n-1} - \gamma X_n (\langle X_n, \theta_{n-1} \rangle - y_n)$$

$$\theta_n | \theta_{n-1} = [I - \gamma x_n x_n^T]\, \theta_{n-1} + \gamma \underbrace{y_n x_n}_{z_n}$$

Homogeneous Markov chain

convergence to a stationary distribution Π_γ.

Expectation is $\bar\theta_\gamma = \int \theta\, \Pi_\gamma(d\theta)$

- In the LS setting, it turns out that $\bar\theta_\gamma$ is independent of γ, and satisfies

$\bar\theta_\gamma = \mathbb{E}[XX^T]^+ \mathbb{E}[X^T Y]$ pseudo-inverse

However, θ_n does NOT converge to θ_*; it oscillates around it, with oscillations of order $\sqrt{\gamma}$.

How to produce a convergent sequence? Average!

Ergodic Theorem: $\bar\theta_n \to \bar\theta_\gamma$ at rate $O(1/\sqrt{n})$ (CLT).

$$\theta_n = \theta_{n-1} - \gamma D g_n(\theta_{n-1}).$$

Stationary dis Π_γ

$$\int \theta\, \Pi_\gamma(d\theta) = \int \theta\, \Pi_\gamma(d\theta) - \gamma \int Dg(\theta)\, \Pi(d\theta)$$

$$\Rightarrow \int Dg(\theta)\, \Pi(d\theta) = 0.$$

- $\bar\theta_\gamma = \theta_*$ and $\bar\theta_n = \frac{1}{n} \sum \theta_k$ converges to $\bar\theta_\gamma$.

~~$\bar\theta_\gamma \neq \theta_*$~~

- ~~under slightly stronger assumptions~~

~~$\lim_{n\to\infty} n \mathbb{E}(\bar\theta_n - \theta_*)^2$~~

For non-quadratic objectives what happens?

→ $\theta_n = \theta_{n-1} - \gamma \nabla g_n(\theta_{n-1})$

is also a Homongeous Markov chain

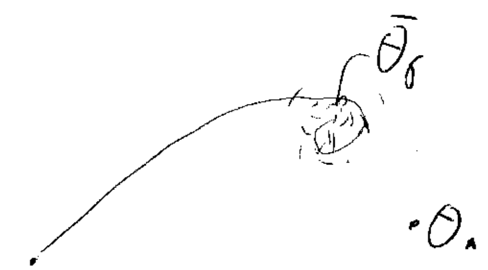
→ Its stationary distribution π_γ satisfies ← $(g = \mathbb{E} g_n)$.

$$\int \nabla g(\theta)\, \pi_\gamma(d\theta) = 0$$

→ However, its mean $\bar\theta_\gamma = \int \theta\, \pi_\gamma(d\theta)$ does not minimize g:

$$\nabla g(\bar\theta_\gamma) = \nabla g\left(\int \theta\, \pi_\gamma(d\theta)\right) \neq \int \nabla g(\theta)\, \pi_\gamma(d\theta) = 0.$$

→ Thus, θ_n oscillates around the "wrong" stationary point. $\bar\theta_\gamma \neq \theta_*$



Moreover, $\|\theta_* - \bar\theta_\gamma\| = O(\gamma)$

→ $$\bar\theta_n - \theta_* = \underbrace{\bar\theta_n - \bar\theta_\gamma}_{\text{stochastic error}} + \underbrace{\bar\theta_\gamma - \theta_*}_{\text{deterministic}}$$

→ [Bach, Durmus, Dieuleveut '17] If g is strongly convex, For small γ we have

$$\bar\theta_\gamma = \theta^* + \gamma\Delta + r_\gamma$$

with $\|r_\gamma\| \le C\gamma^2$, and

$$\mathbb{E}[\bar\theta_k^{(\gamma)} - \theta^*] = \frac{A(\theta_0, \gamma)}{k} + \gamma\Delta + r_\gamma^{(2)}$$

→ Consider now two chains $(\bar\theta_k^{(2\gamma)})_k$, $(\bar\theta_k^{\gamma})_k$ associated with γ and 2γ respectively. Then we have that $2\bar\theta_k^{(\gamma)} - \bar\theta_k^{2\gamma}$ satisfies

$$\mathbb{E}[2\bar\theta_k^{(\gamma)} - \bar\theta_k^{2\gamma} - \theta^*] = \frac{2A(\theta_0,\gamma) - A(\theta_0, 2\gamma)}{k} + O(\gamma^2)$$

<u>Non-convex and SGD</u> [Bottou, Curtis, Nocedal '16] (Optim Methods for Large-Scale ML)

↳ How about non-convex case? We still assume smoothness.

<u>Assumption 1</u>: $F : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, and ∇F is β-Lipschitz continuous

$$\left(F(x) \le F(y) + \langle \nabla F(y), x - y\rangle + \tfrac{1}{2}\beta\|x-y\|^2\right)$$

Generic SGD algorithm

$$\theta_{k+1} = \theta_k - \eta_k \overbrace{g(\theta_k, \xi_k)}^{\text{stochastic gradient}}$$

<u>Lemma</u> $\mathbb{E}_{\xi_k}[F(\theta_{k+1})] - F(\theta_k) \le -\eta_k \langle \nabla F(\theta_k), \mathbb{E}_{\xi_k} g(\theta_k, \xi_k)\rangle + \frac{1}{2}\eta_k^2 \beta\, \mathbb{E}_{\xi_k}[\|g(\theta_k, \xi_k)\|^2]$

$$F(\theta_{k+1}) - F(\theta_k) \le \langle \nabla F(\theta_k), \theta_{k+1} - \theta_k\rangle + \tfrac{1}{2}\beta\|\theta_k - \theta_{k+1}\|^2$$

$$\le -\eta_k \langle \nabla F(\theta_k), g(\theta_k, \xi_k)\rangle + \frac{\eta_k^2}{2}\beta\|g(\theta_k, \xi_k)\|^2$$

→ So, if $g(\theta_k, \xi_k)$ is an unbiased estimator of $\nabla F(\theta_k)$, we obtain

$$\mathbb{E}_{\xi_k}[F(\theta_{k+1})] - F(\theta_k) \leq -\eta_k \|\nabla F(\theta_k)\|^2 + \frac{1}{2}\eta_k^2 \beta \mathbb{E}\|g(\theta_k, \xi_k)\|^2]$$

→ ~~Assumption~~: Denote by $\mathbb{V}_{\xi_k}(g(\theta_k, \xi_k)) = \mathbb{E}\|g\|^2 - \|\mathbb{E}g\|^2$ the variance of the stochastic gradient.

→ Assume $\mathbb{V}_{\xi_k}[g] \leq M + M_V \|\nabla F\|_2^2$ $\forall k$ for some scalars M, M_V.

and $\mathbb{E}_{\xi_k} g = \nabla F(\theta_k)$.

→ Lemma: $\mathbb{E}_{\xi_k}[F(\theta_{k+1})] - F(\theta_k) \leq -(1 - \frac{1}{2}\eta_k \beta (M+1))\eta_k \|\nabla F(\theta_k)\|_2^2 + \frac{1}{2}\beta \eta_k^2 M.$

□

Theorem (Non-convex Objective, Fixed Stepsize).

Suppose F is bounded below, and η fixed and $0 < \eta < \frac{1}{\beta(1+M_V)}$

Then $\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \|\nabla F(\theta_k)\|^2\right] \leq \eta\beta M + \frac{2(F(\theta_1) - F_{inf})}{K\eta}$

$\to \eta\beta M.$

Proof: $\mathbb{E}[F(\theta_{k+1})] - \mathbb{E}[F(\theta_k)] \leq$

$\leq -(1 - \frac{1}{2}\eta\beta(1+M))\eta \mathbb{E}\|\nabla F(\theta_k)\|^2 + \frac{1}{2}\eta^2\beta M$

$\leq -\frac{1}{2}\eta \mathbb{E}\|\nabla F(\theta_k)\|^2 + \frac{1}{2}\eta^2 \beta M$

$F_{inf} - F(\theta_1) \leq \mathbb{E}[F(\theta_{K+1})] - F(\theta_1) \leq -\frac{1}{2}\eta \sum \mathbb{E}\|\nabla F(\theta_k)\|^2 + \frac{K}{2}\eta^2\beta M$

$\Rightarrow \mathbb{E}\left[\frac{1}{K}\sum \|\nabla F(\theta_k)\|^2\right] \leq \frac{2(F(\theta_1) - F_{inf})}{K\eta} + \eta\beta M$ □

↳ $M = 0$? No noise $\Rightarrow \|\nabla F(\theta_k)\|_2 \to 0$

→ η is in a tradeoff $\Big\{$ η small → average norm will be smaller, but it will take longer to converge.

→ Decreasing step-size? Suppose η_k satisfy

$\sum_k \eta_k = +\infty, \quad \sum_k \eta_k^2 < +\infty.$

Theorem: Let $A_K := \sum_{k \leq K} \eta_k$. Then, under previous assumptions, $\mathbb{E}\left[\frac{1}{A_K}\sum_{k\leq K} \eta_k \|\nabla F(\theta_k)\|^2\right] \to 0 \; (K \to \infty).$

Proof: We know by hypothesis that $\eta_k \to 0$, so wlog $\eta_k \beta (1+M) \leq 1 \quad \forall k.$

As before, we have

$$\mathbb{E}[F(\theta_{k+1})] - \mathbb{E}[F(\theta_k)] \leq -\frac{1}{2}\eta_k \mathbb{E}[\|\nabla F(\theta_k)\|^2] + \frac{1}{2}\eta_k^2 \beta M$$

Summing again, we have

$$F_{inf} - \mathbb{E}F(\theta_1) \leq -\frac{1}{2}\sum_{k=1}^{K} \eta_k \mathbb{E}\|\nabla F(\theta_k)\|^2 + \frac{1}{2}\beta M \sum_{k=1}^{K}\eta_k^2$$

$$\Rightarrow \sum_{k=1}^{K} \eta_k \mathbb{E}\|\nabla F(\theta_k)\|^2 \leq 2(\mathbb{E}F(\theta_1) - F_{inf}) + \beta M \sum_{k=1}^{K}\eta_k^2$$

$$\Rightarrow \frac{1}{A_K}\sum_{k\leq K} \eta_k \mathbb{E}\|\nabla F(\theta_k)\|^2 \leq \frac{C}{A_K} \to 0.$$

Corollary If we further assume that F is twice diff and $\theta \mapsto \|\nabla F(\theta)\|^2$ is Lipschitz, then

$$\lim_{k\to\infty} \mathbb{E}[\|\nabla F(\theta)\|^2] = 0.$$

↳ ε-approximate first-order critical points $\{\theta;\ \|\nabla F(\theta)\| \leq \varepsilon\}$

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

Theorem (Nesterov '98): g β-smooth. GD with step size $\gamma = \beta^{-1}$ requires $\beta \cdot \frac{g(\theta_1) - g^*}{\varepsilon^2}$ iterations to reach a ε-1st order stationary point of g.

Proof (simplified for continuous-time).

$\dot{X}(t) = -\nabla g(X(t))$. g bounded below.

$$g(X(T)) - g(X(0)) = \int_0^T \langle \nabla g(X(t)), \dot{X}(t)\rangle dt$$

$$= -\int_0^T \|\nabla g(X(t))\|^2 dt$$

$$\Rightarrow \int_0^T \underbrace{\|\nabla g(X(t))\|^2}_{h(t)} dt \leq K \quad \forall T.$$

so $h(t) = O(t^{-1})$.

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 — NOT anymore.

→ Classification of critical points.

def: θ^* is a strict saddle point if $\nabla g(\theta^*) = 0$ and

→ A local minima satisfies $\lambda_{min}(\nabla^2 g(\theta_0)) \geq 0$ (not necessarily strict).

→ Every critical point ($\nabla g(\theta^*) = 0$) is an equilibrium point of gradient. Which are <u>stable</u> equilibria?

↳ Intuition: strict saddles are unstable.

<u>Questions</u>: A reasonable strategy to avoid unstable equilibria is to add noise (SGD). ~~How~~ Is noise necessary? sufficient?

→ Worst case analysis: GD can provably converge to saddle points with appropriately (bad) init [Nesterov '04]. But what about "generic" init?

→ Example: quadratic case: Consider the non-convex quadratic function $g(\theta) = \frac{1}{2}\theta^T H \theta$, $H = \mathrm{diag}(\lambda_1 \cdots \lambda_d)$,
$\lambda_1 \cdots \lambda_s > 0$ $\lambda_{s+1} \cdots \lambda_d < 0$

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

↳ Gradient Descent initialised at θ_0 is

$$\theta_{k+1} = \theta_k - \gamma \nabla g(\theta_k) = \sum_{i=1}^{d} (1 - \gamma\lambda_i)^{k+1} \theta_{0,i}\, e_i$$

as $\gamma < \frac{1}{\ldots}$ Then $\|\theta_k\| \ldots$ iff $\theta_0 \in \mathrm{span}\{\ldots e_s\}$.

→ If θ_0 is sampled from a distribution which is absolutely continuous wrt ambient dim, then GD avoids this strict saddle with probability 1.

→ How general is that?

<u>Stable Manifold Theorem</u>: . assume g twice differentiable.

· A discrete-time optim algorithm is a mapping $\varphi: \mathbb{R}^d \to \mathbb{R}^d$

ex: GD $\varphi(x) = x - \gamma \nabla g(x)$

· Iterate k is obtained as $\theta_k = \varphi^k(\theta_0)$.

· $\mathcal{X}^* = \{$ strict saddle points of $g\}$.

· <u>Definition</u> (Global Stable Set): The Global Stable Set of strict saddles is

$$S_\varphi = \{\theta_0 ;\ \lim_{k \to \infty} \varphi^k(\theta_0) \in \mathcal{X}^*\}$$

S_φ thus contains the initial values that eventually will land in a strict saddle.

→ In a neighborhood of a critical point θ^*, how to capture the local attractive set, ie  the points around θ^* that will be pushed to θ^*?

Span$\{e_i ;\ e_i^T \nabla^2 g(\theta^*) e_i > 0\}$

- If this local attractor set has zero measure within a small neighborhood of θ^*, then with probability 1 an init close to θ^* will leave this neighborhood.
- Q: How to go from local to global stability? A priori, it is not immediate: escaping a saddle point does not guarantee we won't fall into another one later.

Def: Given a diffeomorphism φ, an <u>unstable fixed point set</u> is $A_\varphi^* = \{\theta;\ \varphi(\theta) = \theta;\ \|D\varphi(\theta)\| > 1\}$

Theorem [Stable Manifold, Smale][67]: Let φ be C^1 (diffeo) mapping $\mathcal{X} \to \mathcal{X}$ and $\mathrm{Det}(D\varphi(\theta)) \neq 0$ $\forall \theta \in \mathcal{X}$. Then the set of initial points that converge to an unstable fixed point has measure zero:

$$\mu\left(\{\theta_0;\ \lim_{k\to\infty} \varphi^k(\theta_0) \in A_\varphi^*\}\right) = 0.$$

Corollary: If $\mathcal{X}^* \subseteq A_\varphi^*$, then $\mu(S_\varphi) = 0$.

Theorem [Gradient Descent Avoids Strict saddles, Lee et al '16] Assume g is β-smooth and $\gamma < \beta^{-1}$. Then if θ^* is a strict saddle, then $\Pr(\lim_k \theta_k = \theta^*) = 0$

Proof: $\to$ Every strict saddle of g is an unstable fixed point of $\varphi(x) = x - \gamma\, Dg(x)$.

$$D\varphi(x) = I - \gamma D^2 g(x)$$

θ^* strict saddle $\Rightarrow$ $\lambda_{min}(D^2 g(\theta^*)) < 0$ $\Rightarrow$

$\Rightarrow$ $\lambda_{max}(I - \gamma D^2 g(\theta^*)) > 1$.

$\to$ If $\gamma < \beta^{-1}$, then $\det(D\varphi(\theta)) \neq 0$ $\forall \theta$, $\lambda_{min}(I - \gamma D^2 g(x)) > 0$ $\forall x$. $\square$

~~extensions~~

$\to$ Prove that $\varphi(x)$ is a diffeo. it is invertible.

- φ injective: suppose x, y s.t. $\varphi(x) = \varphi(y)$

$\Rightarrow$ $x - y = \gamma(Dg(x) - Dg(y))$

$\|x-y\| = \gamma\|Dg(x) - Dg(y)\| \leq \gamma\beta\|x - y\|$

but $\gamma\beta < 1$.

- φ surjective: Consider the implicit function

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

Since $\gamma < \beta^{-1}$, F_y is strongly convex wrt x, $\Rightarrow$ unique min.

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

$y = \phi(y) - \gamma Dg(\phi(y)) = \varphi\,\phi\, y$.

→ 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

(i) Set of strict saddles cannot bee too "large"
countably finite / isolated saddles are ok.

(ii) $\lim_k \theta_k$ exists.

Two sufficient conditions are
(i) Isolated critical points and compact sublevel sets.
(ii) Local Łojasiewicz inequality: $\exists\ m, a, \varepsilon$ st

$$\|\nabla g(\theta)\| \geq m\,|g(\theta) - g(\theta^*)|^a, \quad a < 1.$$

for $\theta \in \{\theta \in N(\theta^*)\ ;\ g(\theta^*) < g(\theta) < g(\theta^*) + \xi\}$

↳ ensures length traveled by iterates of GD is finite.

→ This result shows that the tiniest amount of randomness (at init) is sufficient in mild assumptions to avoid saddles. What is the benefit of extra noise in the dynamics?

Escaping from saddle points

→ We are interested in finding ε-second order stationary points: $\nabla^2 g$ Lipschitz, they are defined as

$$\|\nabla g(\theta^*)\| \leq \varepsilon \quad \text{and} \quad \lambda_{min}(\nabla^2 g(\theta^*)) \geq -\sqrt{\varepsilon \rho}$$
$$\text{with } \rho = Lip(\nabla^2 g).$$

→ Stochastic GD finds ε-second-order critical points.

Thm [Ge'15] If g satisfies the strict saddle, then Noisy GD converges to a local minimum in polynomial time.

"Vanilla" SGD: $\theta_{k+1} = \theta_k - \gamma(\nabla g(\theta_k) + \xi_k)$, $\xi \sim N(0, I)$.

Time is dimension-dependent
(Several improvements in subsequent papers).

→ We have seen that both GD and SGD escape strict saddles.
↳ Dependency on input dim? cursed?
↳ Does noise improve or decrease efficiency?

Theorem [Jin et al '17] Noisy Gradient descent finds ε-second-order stationary point whp with $\tilde{O}\left(\frac{\beta(g(\theta_0) - g^*)}{\varepsilon^2}\right)$ (up to polylog factors).

→ Adding noise is sufficient to escape saddles.
→ Up to log factors, this matches GD rate for 1st-order

→ It turns out that noise is also necessary. Even with uniform init, one can construct smooth functions for which GD requires exponential time ~~Theorem~~ [Du et al '17].

High-dimensional Energy Landscapes

- What are good models for non-convex, yet tractable high-dimensional energy landscapes?
- Models from statistical physics are very rich mathematical structure.

Eg Spherical Spin glass:

$$E(\theta) = \frac{1}{\Lambda^{(k-1)/2}} \sum_{l_1 \cdots l_p = 1}^{\Lambda} z_{l_1 \cdots l_p} \theta_{l_1} \cdots \theta_{l_p}$$

$z_p \sim N(0, \sigma^2)$, $\|\theta\|^2 = \Lambda$

Hamiltonian of the spherical k-spin spin glass.

Kac-Rice Formula

$$\lim_{\Lambda \to \infty} \frac{1}{N} \log \mathbb{E}\, Crt_{N,k}(u) = \theta_{k,p}(u)$$

Reproducing Kernels and Kernel Dynamics

Goal: Infinite-dimensional least squares.
~~[illegible]~~

→ Generic learning in a infinite-dimensional space:

$$\min_{f \in F} \|f^* - f\|^2.$$ ~~This problem~~

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

Hilbert space → Kernels.

→ Def: Given a set of objects $\mathcal{X}$, a psd kernel is a symmetric function $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ st for all finite sequences $x_i \in \mathcal{X}$ and $\alpha_i \in \mathbb{R}$, $i = 1 \ldots n$,

$$\sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) \geq 0.$$

Ex: $K(x, x') = \langle \Phi(x), \phi(x') \rangle$ for arbitrary mappings

$\phi: \mathcal{X} \to \mathcal{F}$. $\mathcal{F}$ Hilbert space.

~~Def~~

Theorem (Aronszajn '50): K is a psd kernel if and only if there exists $\mathcal{F}$ Hilbert space and a mapping $\phi: \mathcal{X} \to \mathcal{F}$ st $\forall (x, x')$ $K(x,x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{F}}$.

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

Examples:

→ Linear kernel: $K(x,y) = x^T y$. ($\phi(x) = x$)

→ Polynomial kernel: $K(x,y) = (1 + x^T y)^d$ ($\phi(x)$ = monomials)

↳ If K_1 and K_2 are kernels, then

- $\alpha K_1 + \beta K_2$ is a kernel. ($\alpha, \beta \geq 0$)
- $K_1 K_2 (x, x') := K_1(x,y) \cdot K_2(x,y)$

are also kernels.

→ Gaussian kernel $K(x,y) = \exp(-d\|x-y\|^2)$

$\phi(x) = ?$ finite-dimensional?

Reproducing Kernel Hilbert Space

In Aronszajn Theorem, we have the representation

$$K(x,x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{F}}.$$

↳ We instantiate $\mathcal{F}$ as a function space over $\mathcal{X}$.

$$\phi(x) = K(\cdot, x) \qquad \mathcal{F}: \mathcal{X} \to \mathbb{R}$$
$$x' \mapsto K(x', x).$$

· function evaluation: $f(x) = \langle f, \phi(x) \rangle$ $f \in \mathcal{F}$

· Reproducing property: $K(x,y) = \langle K(\cdot, x), K(\cdot, y) \rangle$ → "kernel Trick"

· Evaluating of f at a point x is an inner product in the feature space.

Equivalently, the evaluation functional

$$L_x: f \mapsto f(x) \quad \forall f \in H \quad \text{is bounded:}$$

$$\exists M > 0 \text{ st } |L_x(f)| = |f(x)| \leq M \|f\|_H \quad \forall f.$$

Regularization and representer theorem

· Suppose data $x_i \in \mathcal{X}$, labels $y_i \in \mathcal{Y}$ $i = 1 \dots n$ kernel K with RKHS $\mathcal{F}$.

Learning: We consider

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(y_i, f^T \phi(x_i)) + \frac{\lambda}{2} \|f\|^2$$

Representer Theorem (Kimeldorf & Wahba, '70) The minimiser is of the form $f = \sum_{j=1}^{n} \alpha_j \Phi(x_j) = \sum_{j=1}^{n} \alpha_j K(x_j, \cdot)$

Proof: ok.

Kernel Ridge Regression

- Data $x_1 \cdots x_n \in X$ $y_1 \cdots y_n \in \mathbb{R}$, psd kernel k.
- Infinite-dimensional Least squares

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{F}}^2$$

Option 1: Use previous representer theorem

$f = \sum_{i=1}^{n} \alpha_i K(\cdot, x_i)$, obtained with

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} (y_i - (K\alpha)_i)^2 + \lambda \alpha^T K \alpha$$

$\alpha^* = (K + n\lambda I)^{-1} y + \varepsilon, \quad K\varepsilon = 0.$

f^* is unique.

Option 2: Suppose finite-dimensional feature space

$\mathcal{F} \subset \mathbb{R}^d, \quad \Phi \in \mathbb{R}^{n \times d}$:

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

$\rightarrow w = (\Phi^T \Phi + n\lambda I)^{-1} \Phi^T y$

$w = \Phi^T (\Phi\Phi^T + n\lambda I)^{-1} y = \Phi^T (K + n\lambda I)^{-1} y = \Phi^T \alpha.$

$\Phi w = K\alpha$

→ Primal / Dual perspective: linear systems of different sizes. (n: # of data-points, d: # of features) → dual perspective is more natural for generic kernels, where $d = \infty$!).

Kernels and Neural Networks

→ we have seen that kernels are dot products in generic feature spaces and that adding kernels is also a kernel.

→ Let $\varphi(x, v)$ $x \in X$, $v \in \mathcal{V}$ be a "neuron"

- we consider now a fixed probability measure τ on $\mathcal{V}$, assumed to be compact. ($\varphi_v : X \to \mathbb{R}$ measurable.) Let

$$\mathcal{F} = \left\{ f ;\ f(x) = \int_{\mathcal{V}} p(v) \varphi_v(x) d\tau(v) \quad \forall x ;\ p : \mathcal{V} \to \mathbb{R} \text{ is square-}\tau\text{-integrable} \right\}.$$

Prop (Bach '16): $\mathcal{F}$ is an RKHS, with kernel

$$K(x, y) = \int_{\mathcal{V}} \varphi_v(x) \varphi_v(y) d\tau(v)$$

Proof: consider $T : L_2(d\tau) \to \mathcal{F}$ defined by $(Tp)(x) := \int_V p(v)\varphi_v(x)\,d\tau(v)$. Let K be the null space of T. $T_{|K^\perp}$ is a bijection from $K^\perp$ to $\mathcal{F}$. We define $\langle f, g\rangle_{\mathcal{F}} := \int_V (T^{-1}f)(v)(T^{-1}g)(v)\,d\tau(v)$.

- $K(\cdot, y) \in \mathcal{F}$ for all $y \in X$.
- $\langle f, K(\cdot, y)\rangle = \int_V (T^{-1}f)(v) \cdot p(v)\,d\tau(v) =$

$$= \int (T^{-1}f)(v) \cdot \varphi_v(y)\,d\tau(v) = T(U^{-1}f)(y) = f(y)$$

↳ Reproducing Property.

↳ Kernels as expectations are well-suited to approximations by Monte-Carlo sampling.

→ Consider $v_1 \dots v_m$ iid samples from τ, and let $\hat{K}(x, y) = \frac{1}{m}\sum_{i=1}^{m} \varphi_{v_i}(x)\varphi_{v_i}(y)$.

→ What is kernel regression in this setting? Our predictors have the form

$\frac{1}{m}\sum \alpha_i \varphi_{v_i}(x)$ adjusting α_i : ... no feature learning.

As $m \to \infty$, $\hat{K} \to K$ (Rahimi & Recht '07)

Rate is essentially the Monte-Carlo Rate: $\frac{1}{\sqrt{m}}$.

Kernel Dynamics and the NTK

↳ Consider a generic parametric non-linear model

$f(x;\theta)$ (eg $f(x;\theta) = \sum_{i=1}^{n} \alpha_i \varphi(x, \beta_i)$)

↳ We evaluate this model with the convex loss

$R(f(x;\theta))$; (eg $R(f(x;\theta)) = \|f^* - f(x;\theta)\|^2$)

$\dot{\theta}(t) = \nabla_\theta f(\cdot, \theta(t))^T R'(f(\cdot, \theta(t)))$, thus

$$\dot{f}(\cdot, \theta(t)) = \underbrace{\nabla_\theta f(\cdot, \theta(t))^T \nabla_\theta f(\cdot, \theta(t))}_{K(t)\,\cancel{(x, x')}} \cdot R'(f(\cdot, \theta(t)))$$

↳ This update is the same as kernel regression, using the kernel $K_t(x, x') = \nabla f(x;\theta(t))^T \nabla f(x';\theta(t))$.

Fact: If $\theta(t)$ does not move too much, $K_t \approx K_0$, thus the general GD dynamics will be well approximated with kernel dynamics.

↳ The assumption that $K_t \approx K_0$ is equivalent to the assumption the model is linear: our parametric model thus becomes

$$T_f(x;\theta) = f(x;\theta_0) + (\theta-\theta_0)\cdot\nabla_\theta f(x;\theta_0)$$

Up, to an affine offset, this model is equivalent to a kernel method of the form $K(x,x') = \nabla_\theta f(x;\theta_0)^T \nabla_\theta f(x';\theta_0)$.

→ Lazy Training ~~Regime~~.

→ Present the lazy regime using linear approximation in function space.

Q: What is the corresponding linear model?

It is defined by the neural Tangent Kernel:

$$K(x,x') = \nabla_\theta f(x;\theta_0)^T \nabla_\theta f(x';\theta_0)$$

→ Scaled objective function and comparison with kernel dynamics.

Theorem by Chizat & Bach.

→ Analysis in terms of $f(x;\theta) = \alpha(n)\sum_{i=1}^{n}\phi(x;\theta_i)$

The quantity that controls the validity of linear approximation is roughly $\dfrac{\|Df_{\theta_0}\|^2}{L_{Df}}$

If $f_n \to f$ then this ratio converges to a constant. If we scale by something larger, it explodes.

$$\mathbb{E}\|Df_{\theta_0}\|^2 = \cancel{n}\,\alpha(n)^2\,\mathbb{E}\|D\phi_\theta\|^2$$

$$L_{Df} = \alpha(n) L_{D\phi} \; ; \; \mathbb{E}\|f(\theta_0)\|^2 = n\alpha(n)^2\,\mathbb{E}\|\phi\|^2$$

$$\mathbb{E}\|f_i(\theta_0) - \bar f(\theta_0)\| \lesssim \max(1, \alpha(n)\sqrt{n})$$

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

In this regime, we develop the particle ~~system~~ interacting system:

$$f_n(\cancel{x};\theta) = \frac{1}{n}\sum_{i=1}^{n}\phi(\theta_i) \qquad \theta = (\theta_1 \dots \theta_n), \quad \theta_i \in D.$$

$$L(\theta) = \frac{1}{2}\|\bar f_n(\theta) - f^*\|^2$$

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

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

Let $\hat\mu_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{\theta_i}$ empirical measure

$$\bar f_n(\theta) = \int \phi(\theta)\,\mu_n(d\theta).$$

$$L(\theta) = C_f - \int F(\theta)\mu_n(d\theta) + \iint K(\theta,\theta')\mu_n(d\theta)\mu_n(d\theta')$$

$$F(\theta) = \langle \phi(\theta), f^* \rangle_V, \quad K(\theta,\theta') = \langle \phi(\theta), \phi(\theta') \rangle_V$$

$$\dot\theta_i = -\nabla_{\theta_i} L(\theta) = -\frac{1}{n}\langle f^*, \nabla\phi(\theta_i)\rangle + \frac{1}{n^2}\sum_{i'} \langle \nabla\phi(\theta_i), \phi(\theta_{i'})\rangle$$

→ we have $\frac{n}{2}\nabla_{\theta_i} L(\theta) = \nabla V(\theta_i; \mu_n)$ with

$$V(\theta;\mu_n) = -F(\theta) + \int K(\theta,\theta')\mu_n(d\theta').$$

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

Q: Describe the evolution of a measure ~~as~~ subject to a velocity field? <u>Continuity equation</u>:

$$\partial_t \mu_t = \mathrm{div}(\nabla V \cdot \mu_t)$$

$$\partial_t\left(\int f \mu_t\right) = -\int \langle \nabla f, \nabla V\rangle \mu_t \quad \forall f \in C_c^1$$

→ Important properties/questions:

(i) Convergence of this PDE?

(ii) Behavior of finite-particle dynamics around mean-field? Fluctuation analysis.

(iii) How is generalization affected under this regime?

↳ Now features move at nominal speed.

(iv) Extension to deeper architectures?