Part 4-1 Non-Convex Optimization

Optimization Landscape

This chapter

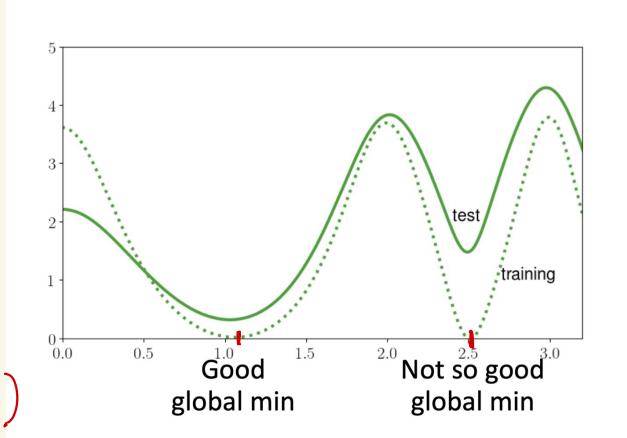
GD/SGD not only finds a global minimum,

but also finds the right global minimum.

Next Chapter (Implicit Bias/Regularization)

Overparameterization

$p \gg n$.



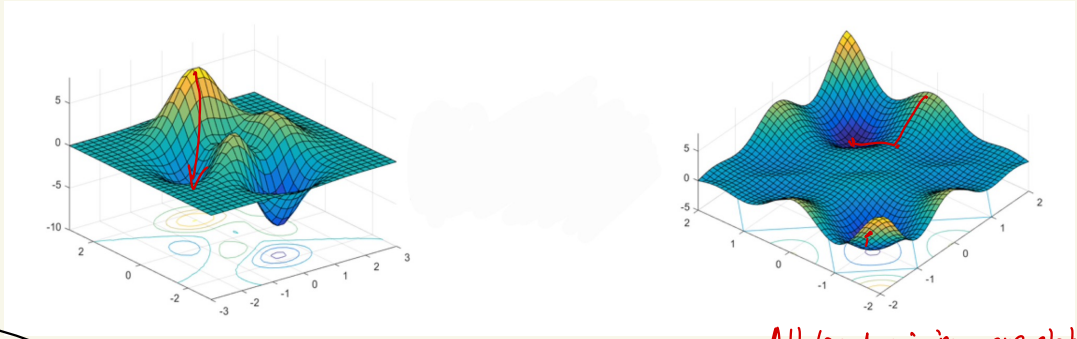
- Many existing optimizers are designed for optimizing convex functions.

GD ← steepest descent

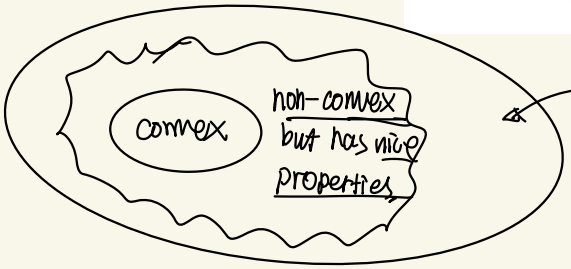
Why do they work well empirically for non-convex functions ?

① Overparameterization

② Non-convex objective has nice properties



All local minima are global



very hard functions

1) Geometry-Based: All local minima are global. All strict local minima are global

2) Local Dynamics (Lazy / Kernel Regime): Utilize local linear behavior of the network predictions,

3) Global Dynamics (Active Training, Mean Field): Characterized by large changes in the parameter.

GD: $\theta^0 =$ initialization

$$\theta^{t+1} = \theta^t - \eta \nabla g(\theta^t)$$

step-size

Observations:

① GD cannot always find the global minimum.

② Finding global minimum of non-convex function is NP-hard.

③ GD can solve convex functions

④ Objectives in Deep Learning are (almost) non-convex.

⑤ GD/SGD finds approximate global min. of loss functions in DL.

1. $\boxed{\text{Identify a large set of functions that GD/SGD can solve.}}$

2. Prove that some of the loss functions in ML problems belong to this set.

Basic Idea : GD can find local minimum

+ all local minimum of f are global minimum. ← problem-specific.

$\Rightarrow$ GD can find global minimum.

Let f be a twice-differentiable functions

def (local minimum) x is local minimum of f if $\exists$ open neighborhood B_x of x

s.t. $\forall x' \in B_x$. $f(x') \geq f(x)$

$$\Rightarrow \nabla f(x) = 0 \quad \nabla^2 f(x) \succeq 0$$

Note: not vice versa! (e.g. $f(x_1, x_2) = x_1^2 + x_2^3$. $(0,0) \rightarrow \nabla = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. $\nabla^2 = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$

$$\nabla f(x) = \begin{pmatrix} 2x_1 \\ 3x_2^2 \end{pmatrix} \qquad \nabla^2 f(x) = \begin{pmatrix} 2 & 0 \\ 0 & 6x_2 \end{pmatrix}$$

NOT a local minima

strict-saddle

Royer & Wright.

$$f(\theta) = f(\theta^0) + \underbrace{\nabla f(\theta^0)^T (\theta - \theta^0)}_{=0} + \frac{1}{2}(\theta - \theta^0)^T \nabla^2 f(\theta^0)(\theta - \theta^0)$$

until $\lambda_{min}(\nabla^2 f(\theta^t)) \geq -\varepsilon_s$

$$\theta^{t+1} = \theta^t - \eta \cdot \nabla f(\theta^t)$$

until $\|\nabla f(\theta^t)\| \leq \varepsilon_f$.

$v_t^T \nabla f(\theta^t) \leq 0$

$\rightarrow \lambda_t \leq -\varepsilon_s$.

$\rightarrow \theta^{t+1} = \theta^t + \eta \cdot v_t$.

i) $\nabla f(\theta) = 0$ (ε_f)

ii) $\nabla^2 f(\theta) \succeq 0$ (ε_s)

$\Rightarrow$ Second Order Stationary Point (SOSP)

but saddle point?

- Third-order saddles can be escaped.
- NP-hard to escape 4th order saddles ($\Rightarrow$ a 4th degree polynomial is NP-hard to optimize)
- Neural Networks of depth L will generally have saddles of order L. (x^L)

Thm It's NP-hard to check (find) whether a point is a local minimum or not.

Strict-saddle

Rong Ge et al. '15.
Jason Lee et al. '19

def. (strict-saddle)

We say f is (α, β, γ)-strict-saddle, if $\forall x \in \mathbb{R}^d$ satisfies one of the following,

① $\|\nabla f(x)\|_2 \geq \alpha$ $\qquad$ $\|\nabla f(x)\|_2 < \varepsilon_f$

② $\lambda_{\min}(\nabla^2 f(x)) \leq -\beta$ $\qquad$ $\lambda_{\min}(\nabla^2 f(x)) > -\varepsilon_s$

③ $\|x - x^*\|_2 \leq \gamma$: x is γ-close to a local minimum x^*.

local minimum

Thm (informally stated) Suppose f is (α, β, γ)-strict-saddle.

Then, many optimizers can converge to a local minimum with ε-error in Euclidean distance

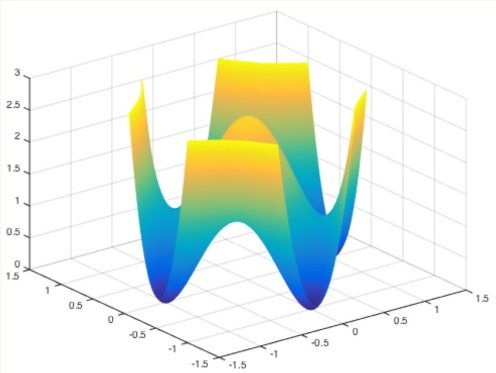
GD/SGD $\qquad$ in time $\text{poly}\left(d, \frac{1}{\alpha}, \frac{1}{\beta}, \frac{1}{\gamma}, \frac{1}{\varepsilon}\right)$

If all local minima are global
+ strict-saddle condition $\Rightarrow$ optimizers can converge to global minimum efficiently!

Thm Suppose $\exists\, \varepsilon_0, \tau_0, C > 0$, s.t.

if $x \in \mathbb{R}^d$ satisfies $\begin{cases} \|\nabla f(x)\|_2 \leq \varepsilon < \varepsilon_0 \\ \nabla^2 f(x) \succeq -\tau_0 I \end{cases}$ strict saddle condition

then x is ε^c-close to a global minimum of f.

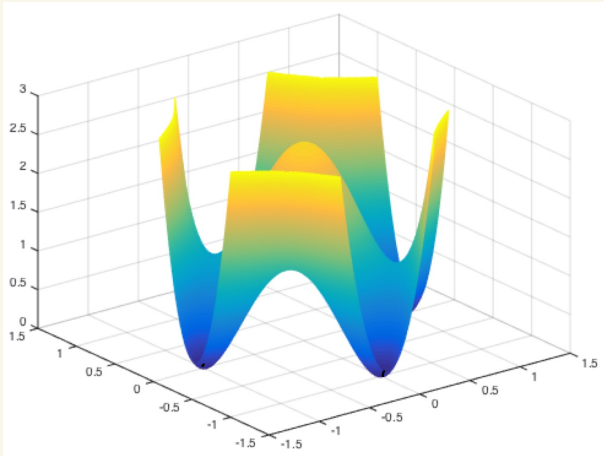


Then, many optimizers can converge to a global minimum of f up to δ-error in ℓ_2-distance in $\text{poly}\left(\frac{1}{\delta}, \frac{1}{\tau_0}, d\right)$ time

All Local Minima are Global Minima.

Recall: GD/SGD can find global minimum for non-convex functions

1. GD/SGD can find local minima (SOSP) efficiently in strict-saddle conditions.
2. All local minima of f are global minima.



e.g. 1) Matrix Completion

2) Rank-k Approximation (PCA)

3) Orthogonal Tensor Decomposition

4) Phase Retrieval

5) Dictionary Learning

6) Max-Cut via Burer-Monteiro

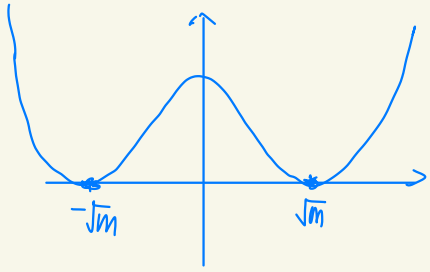
7) (Some) Neural Networks
- Overparameterized Networks with Quadratic Activation
- ReLU Network with Two Neurons
- ReLU Network via landscape design

△ Principle Component Analysis (PCA):

$M \in \mathbb{R}^{d\times d}$, symmetric PSD rank-1 approximation

$$\min_{x\in\mathbb{R}^d} g(x) = \frac{1}{2}\|M - xx^T\|_F^2 = \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}(M_{ij} - x_i x_j)^2$$

$d=1$: $g(x) = \frac{1}{2}(m - x^2)^2$



Thm: All local minima of g are global minima.

Proof: 1. find out all stationary points (gradient)

2. find out all local minima (Hessian)

$\|A\|_F = tr(A^TA)$

$$\nabla g(x) = -(M - xx^T)\cdot x = 0$$

$$\Rightarrow \boxed{Mx = xx^Tx = \|x\|_2^2 \cdot x}$$

eigenvalue eigenvector

△ global minimum:

eigenvector with largest eigenvalue.

(Eckart-Young-Mirsky theorem)

Suppose eigenvalues are distinct $\lambda_1 > \lambda_2 > \cdots > \lambda_d \geq 0$.

$v_1, v_2, \cdots, v_d$ (unit)

$$x = \pm\|x\| v_i = \pm\sqrt{\lambda_i}\cdot v_i$$

local minimum $\Rightarrow \nabla^2 g(x) \succeq 0$: $\forall v \in \mathbb{R}^d$, $\langle v, \nabla^2 g(x) v\rangle \geq 0$

Trick: Taylor Expansion $g(x+v) = g(x) +$ linear terms in v + quadratic terms in v + h.o. term

$\frac{1}{2}\langle v, \nabla^2 g(x) v\rangle$

$$g(x+v) = \frac{1}{2}\|M - (x+v)(x+v)^T\|_F^2$$

$$= \frac{1}{2}\|M - xx^T - (xv^T + vx^T) - vv^T\|_F^2$$

$$= \frac{1}{2}\|M - xx^T\|_F^2 - \langle M - xx^T, xv^T + vx^T\rangle + \frac{1}{2}\langle xv^T + vx^T, xv^T + vx^T\rangle - \langle M - xx^T, vv^T\rangle$$

2nd-order

$\langle A, B\rangle = tr(A^T B)$

+ h.o. terms

$$\Rightarrow \frac{1}{2}\langle v, \nabla^2 g(x) v\rangle$$

$$= \frac{1}{2}\langle xv^T + vx^T, xv^T + vx^T\rangle - \langle M - xx^T, vv^T\rangle$$

$$= \|x\|_2^2 \cdot \|v\|_2^2 + \langle x, v\rangle^2 - v^T M v + \langle x, v\rangle^2$$

$$= 2\langle x, v\rangle^2 + \|x\|_2^2 \cdot \|v\|_2^2 - v^T M v$$

$tr(AB) = tr(BA)$

$tr(ABC) = tr(CAB)$

Let $v = v_1$ (unit largest eigenvector) $\quad \frac{1}{2}\langle v, \nabla^2 g(x) v\rangle = 2\langle x, v_1\rangle^2 + \|x\|_2^2 - \underbrace{v_1^T M v_1}_{\lambda_1} \geq 0$

① $x = \pm\sqrt{\lambda_1}\cdot v_1$ $\quad x$ is global minimum.

② $x = \pm\sqrt{\lambda_i} v_i$ $(i \neq 1)$ $\quad \langle v_i, v_1\rangle = 0$

$$2\langle x_i, v_1\rangle^2 + \|x_i\|^2 - \lambda_1 = \lambda_i - \lambda_1 \geq 0$$

$\Rightarrow \lambda_i \geq \lambda_1$ contradiction!

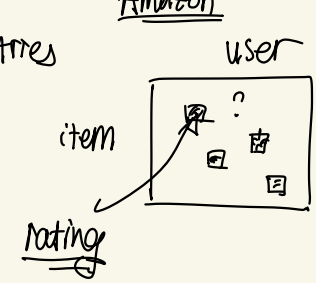
$\Downarrow$

If x is stationary point, and x is not a global minimum, then moving in v_1 direction would lead to a second-order improvement. $\square$

Matrix Completion [Ge et al. '16].

rank-1 version : M is rank-1 SPSD matrix $M = zz^T$, $z \in \mathbb{R}^d$

Given random entries of M, goal is to recover the rest of the entries



$\Omega \in [d] \times [d]$

Every entry is includeded in Ω independently w.p p

We observe $P_\Omega(M)$, the matrix obtained by zeroing out every entry of M outside Ω.

Goal: recover M

d parameters to describe a rank-1 matrix

$\#$ of observation $\approx p \cdot d^2 \gg d \Rightarrow p \gg \frac{1}{d}$

Non-convex objective:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{2} \cdot \sum_{(i,j)\in\Omega} (M_{ij} - x_i x_j)^2$$

$$= \frac{1}{2} \| P_\Omega(M - x x^T) \|_F^2$$

(e.g. convex transition)

e.g. $z = e_1$ $M = e_1 e_1^T = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$

def (Incoherence) w.l.o.g z satisfies $\|z\|_2 = 1$, $\|z\|_\infty \le \frac{\mu}{\sqrt{d}} \leftarrow O(\log d)$ or $O(1)$

Thm Suppose $p = \frac{\text{poly}(\mu, \log d)}{d\varepsilon}$, $\varepsilon > 0$. z is incoherent.

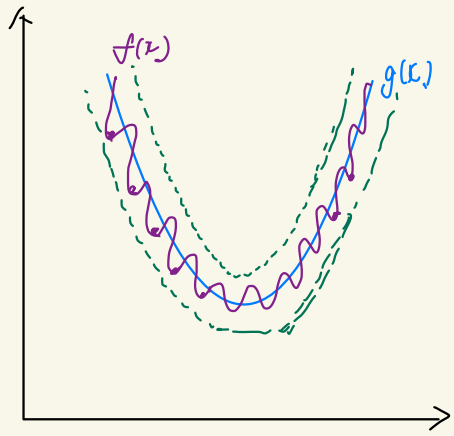
Then w.h.p. all local minima of f are $O(\sqrt{\varepsilon})$-close to z or $-z$

(global minimum of f)

Remark: $g(x) = \frac{1}{2} \|M - x x^T\|_F^2$

$f(x) = \frac{1}{2} \|P_\Omega(M - x x^T)\|_F^2$

$f(x) \approx g(x)$, $\nabla f(x) \approx \nabla g(x)$, $\nabla^2 f(x) \approx \nabla^2 g(x)$



$\boxed{\text{concentration}}$ of scalar: $\sum_{(i,j)\in\Omega} T_{ij} \approx p \cdot \sum_{(i,j)\in[d]\times[d]} T_{ij}$.

$$\left| \sum_{(i,j)\in\Omega} T_{ij} - p \cdot \sum_{(i,j)\in[d]\times[d]} T_{ij} \right| < \varepsilon, \text{ w.h.p}$$

$$\langle A, P_\Omega(B) \rangle = \sum_{(i,j)\in\Omega} A_{ij} B_{ij} \approx p \cdot \langle A, B \rangle$$

$f(x) = \frac{1}{2} \|P_\Omega(M - x x^T)\|_F^2$ then

- $\langle v, \nabla f(x) \rangle = -\langle v, \underset{(P_\Omega^2)}{P_\Omega}(M - x x^T) \cdot x \rangle = -\langle \underset{\text{rank-1 matrix}}{v x^T}, \underset{\text{rank-2 matrix}}{P_\Omega(M - x x^T)} \rangle$

- $\langle v, \nabla^2 f(x) v \rangle = \langle P_\Omega(v x^T + x v^T), v x^T + x v^T \rangle - 2 \langle P_\Omega(M - x x^T), v v^T \rangle$

Lemma Let $\varepsilon > 0$. $p = \dfrac{\mathrm{poly}(\mu, \log d)}{d\varepsilon^2}$. $A = uu^T$. $B = vv^T$. s.t. u, v incoherent.

$$\Rightarrow \boxed{\left| \frac{1}{p} \langle P_\Omega(A), B \rangle - \langle A, B \rangle \right| < \varepsilon \quad \text{w.h.p.}}$$

1. $\nabla g(x) = 0 \Rightarrow \langle x, \nabla g(x) \rangle = 0 \Rightarrow \langle xx^T - (zz^T - xx^T), \rangle = 0. \Rightarrow \langle x, z \rangle^2 = \|x\|_2^4$

2. $\nabla f(x) = 0 \Rightarrow \langle x, \nabla f(x) \rangle = 0 \Rightarrow \langle x, \nabla g(x) \rangle \approx \frac{1}{p} \langle x, \nabla f(x) \rangle \pm \varepsilon$ w.h.p.
(Lemma+1)

$\Rightarrow |\langle x, \nabla g(x) \rangle| = |\langle x, -(zz^T - xx^T)x \rangle| \le \varepsilon \Rightarrow \langle x, z \rangle^2 \ge \|x\|_2^4 - \varepsilon$ w.h.p.

3. $\nabla^2 g(x) \succeq 0 \Rightarrow \langle z, \nabla^2 g(x) z \rangle \ge 0$ ↳ by PCA, $\|z\|_2 = 1$.

$\Rightarrow 2\|x\|_2^2 + 4\langle x, z \rangle^2 - 2 \ge 0$

$\Rightarrow 3\|x\|_2^2 = \|x\|_2^2 + 2\|x\|_2^2 \ge \|x\|_2^2 + 2\langle x, z \rangle^2 \ge 1 \Rightarrow \|x\|_2^2 \ge \frac{1}{3}$

Cauchy-Schwartz.

4. $\nabla^2 f(x) \succeq 0 \Rightarrow \langle z, \nabla^2 f(x) z \rangle \ge 0$
(Lemma+3)

$\Rightarrow \langle z, \nabla^2 g(x) z \rangle \ge -\varepsilon$ w.h.p. $\Rightarrow \|x\|_2^2 \ge \frac{1}{3} - \frac{\varepsilon}{3}$. w.h.p

Lemma' (g has no bad local minima) All local minima of g are global.

Proof: $\nabla g(x) = 0 \Rightarrow \langle z, \nabla g(x) \rangle = 0$

$\Rightarrow \langle z, -(zz^T - xx^T)x \rangle = 0 \Rightarrow \langle x, z \rangle (1 - \|x\|_2^2) = 0$

$|\langle x, z \rangle| \ge \frac{1}{3}, \ne 0. \Rightarrow \|x\|_2^2 = 1 \Rightarrow \langle x, z \rangle^2 = 1 \Rightarrow x = \pm z$ □

global minima.

Thm (f has no bad local minima) $p = \dfrac{\mathrm{poly}(\mu, \log d)}{d\varepsilon^2}$.
(Lemma+Lemma')

Then w.h.p, all local minima of f are $O(\sqrt{\varepsilon})$-close to z or $-z$.

Proof: $\|x - z\|_2^2 = \|x\|_2^2 + \|z\|_2^2 - 2\langle x, z \rangle = \|x\|_2^2 + 1 - 2\langle x, z \rangle$

upper bound lower bound. w.h.p

$\nabla f(x) = 0 \rightarrow \langle x, z \rangle \ge \|x\|_2^4 - \varepsilon \ge \left(\frac{1}{3} - \frac{\varepsilon}{3}\right)^2 - \varepsilon \ge \frac{1}{32}$ .h.p $\quad \varepsilon \le \frac{1}{16}$

$\rightarrow \langle x, \nabla f(x) \rangle = 0 \Rightarrow |\langle z, \nabla g(x) \rangle| \le \varepsilon$

$$\Rightarrow |\langle x, z\rangle| \cdot \left|1 - \|x\|_2^2\right| \le \varepsilon$$

$$\Rightarrow \left|1 - \|x\|_2^2\right| \le 32\varepsilon = O(\varepsilon) \quad \Rightarrow \quad \|x\|_2^2 = 1 \pm O(\varepsilon) \text{ w.h.p}$$

$$\langle x, z\rangle^2 = \|x\|_2^4 - \varepsilon \ge (1 - O(\varepsilon))^2 - \varepsilon = 1 - O(\varepsilon)$$

$\to$ Case 1: $\langle x, z\rangle \ge 1 - O(\varepsilon)$: $\|x - z\|_2^2 \le 1 + O(\varepsilon) + 1 - 2(1 - O(\varepsilon)) \le O(\varepsilon)$

Case 2: $\langle x, z\rangle \le -(1 - O(\varepsilon))$: $\|x + z\|_2^2 \le 1 + O(\varepsilon) + 1 + 2(-(1 - O(\varepsilon))) \le O(\varepsilon)$

$\Rightarrow$ local minima x is $O(\varepsilon)$-close to global minima z or $-z$. why? □

Neural Tangent Kernel (NTK)

Recall: Geometric-Based Viewpoint

① strict-saddle condition
② no spurious local minima

$\Rightarrow$ GD/SGD efficiently finds global minima

- What neural networks are strict saddle?

1) Quadratic activation: $f(x) = \sum_{j=1}^{m} a_j (w_j^T x)^2$. $m \gtrsim \min(\sqrt{n}, d)$ overparameterization

2) ReLU activation: $f(x) = \sum_{i=1}^{m} \mathrm{ReLU}(w_j^{*T} x)$ with
   - Gaussian x
   - no negative output weight.
   - $m \le d$
   - loss function encodes tensor decomposition → smooth

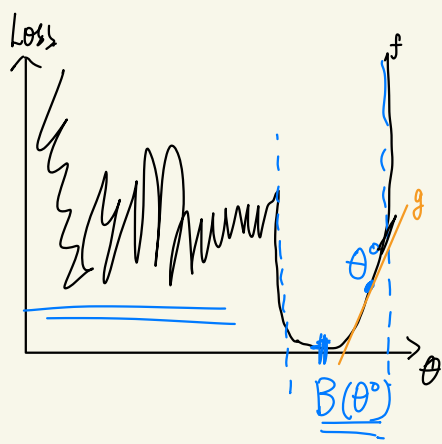
3) Two-neuron with orthogonal weights

4) One-convolutional filter with non-overlapping patches

.....

Local Dynamics: identify a particular region in which all local minima is global minima

↓ algorithmic-dependent



- nonlinear $f_\theta(x)$
- Taylor Expansion around initialization θ^0:

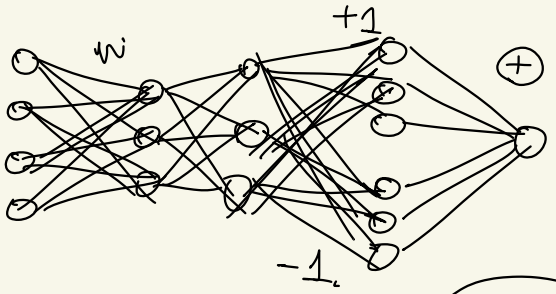
$$f_\theta(x) = \underbrace{f_{\theta^0}(x) + \langle \nabla_\theta f_{\theta^0}(x), \theta - \theta^0\rangle}_{\triangleq\, g_\theta(x)} + \text{h.o. term}$$

(linear approximation)

$g_\theta(x) = f_{\theta^0}(x) + \langle \nabla_\theta f_{\theta^0}(x), \theta - \theta^0 \rangle$ is an affine function in θ, or $\Delta\theta = \theta - \theta^0$

constant (for fixed x, θ^0)

sometimes for convenience, choose θ^0 s.t. $\boxed{f_{\theta^0}(x) = 0}$ $\forall x$.



$$\sum_{i=1}^{m} a_i \sigma(w_i^T x) + \sum_{i=1}^{m} -a_i \sigma(w_i^T x) \equiv 0 \quad \forall x$$

fixed x, θ^0

$$y' = g_\theta(x) = y - f_{\theta^0}(x) = \langle \nabla_\theta f_{\theta^0}(x), \Delta\theta \rangle$$

feature map $\phi(x) = \nabla_\theta f_{\theta^0}(x)$

parameter

$\phi(x)$ is fixed function of x, given the architecture and θ^0

For similicity, $f_{\theta^0}(x) = 0 \Rightarrow y' = y$

Kernel $\boxed{K(x, x') = \langle \phi(x), \phi(x') \rangle = \langle \nabla_\theta f_{\theta^0}(x), \nabla_\theta f_{\theta^0}(x') \rangle}$

Neural Tangent Kernel

gradient

We want to use $g_\theta(x)$ to approximate $f_\theta(x)$

$\Rightarrow$ kernel method

linear model on top of features

When $\theta \approx \theta^0$: $\ell(f_\theta(x), y) \approx \ell(g_\theta(x), y)$

not necessarily convex

convex

convex

linear.

- How valid is this approximation ?

$$\hat{L}(f_\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(x^{(i)}), y^{(i)}) \qquad \hat{L}(g_\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(g_\theta(x^{(i)}), y^{(i)})$$

Key idea: Taylor Expansion can work for certain cases in the following sense:

$\exists$ neighborhood of θ^0. s.t.

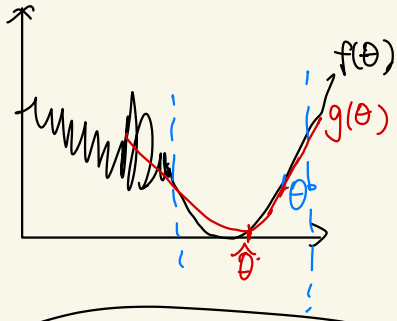
$\boxed{1}$ accurate approximation : $f_\theta(x) \approx g_\theta(x)$ $\Rightarrow$ $\hat{L}(f_\theta(x)) \approx \hat{L}(g_\theta(x))$ $\forall \theta \in B(\theta^0)$

$\boxed{2}$ it suffices to optimize in $B(\theta^0)$: $\exists$ an approximate global minimum $\hat{\theta} \in B(\theta^0)$

s.t. $\hat{L}(g_{\hat{\theta}}) \approx 0$ (thus $\hat{L}(f_{\hat{\theta}}) \approx 0$)

$\boxed{3.}$ optimizing $\hat{L}(f_\theta)$ is similiar to optimizing $\hat{L}(g_\theta)$

and does NOT leave $B(\theta^0)$



Note : $\boxed{3.}$ to some extent follows $\boxed{1.}$ and $\boxed{2.}$,

but still needs a proof.

$\boxed{1}$ $\boxed{2}$ $\boxed{3.}$ can be all true < Overparameterization and/or particular scaling of initialization θ^0

small (or zero) stochasicity, θ does not leave $B(\theta^0)$

↳ < small learning rate

full-batch GD.

Analysis of NTK Approach (I)

NTK: $K(x,x') = \langle \phi(x), \phi(x') \rangle = \langle \nabla_\theta f_{\theta^0}(x), \nabla_\theta f_{\theta^0}(x') \rangle$

gradient.

$g_\theta(x) = \nabla_\theta f_{\theta^0}(x)^T \cdot \Delta\theta$

Formalize $\boxed{1.}$ and $\boxed{2.}$: $\phi^{(i)} = \phi(x^{(i)}) = \nabla_\theta f_{\theta^0}(x^{(i)})$

$$\Phi = \begin{bmatrix} -\phi^{(1)}- \\ \vdots \\ -\phi^{(n)}- \end{bmatrix} \in \mathbb{R}^{n \times p}$$

< n : # of examples

p : # of parameters

$$\hat{L}(g_\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \phi(x^{(i)})^T \Delta\theta\right)^2 = \frac{1}{n}\|y - \Phi\Delta\theta\|_2^2. \quad y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} \in \mathbb{R}^n.$$

Assume $y^{(i)} = O(1)$ $\|y\|_2 = O(\sqrt{n})$

Lemma (for $\boxed{2}$) Suppose $p \geq n$. $\text{rank}(\Phi) = n$. $\boxed{\sigma_{\min}(\Phi) = \sigma} > 0$.

Then let $\Delta\hat{\theta}$ be the min-norm solution to $y = \Phi \cdot \Delta\theta$

nearest global minimum

then $\|\Delta\hat{\theta}\|_2 \leq O\left(\frac{\sqrt{n}}{\sigma}\right)$

Note: take the ball $B(\theta^0) = \{\theta = \theta^0 + \Delta\theta : \|\Delta\theta\|_2 \le O(\frac{\sqrt{n}}{\sigma})\}$

$B(\theta^0)$ contains a global minimum.

Proof: $\Delta\hat{\theta} = \Phi^{\dagger} \cdot y$ $\quad \|\Phi^{+}\|_{op} = \sup_x \frac{\|\Phi^{+}x\|}{\|x\|} = \sigma_{max}(\Phi^{+}) = \frac{1}{\sigma_{min}(\Phi)} = \frac{1}{\sigma}$

$\Rightarrow \|\Delta\hat{\theta}\|_2 \le \|\Phi^{+}\|_{op} \cdot \|y\|_2 = O(\sqrt{n}/\sigma)$ $\square$

Lemma (for ①) Suppose $\nabla_\theta f_\theta(x)$ is β-Lipschitz in θ

$\forall x: \forall \theta, \theta': \|\nabla_\theta f_\theta(x) - \nabla_\theta f_{\theta'}(x)\|_2 \le \beta \cdot \|\theta - \theta'\|_2$.

then: $|f_\theta(x) - g_\theta(x)| \le O(\beta \cdot \|\Delta\theta\|_2^2)$

bounded approximation error

Furthermore (restricted to $B(\theta^0)$). $\Rightarrow \forall \theta \in B(\theta^0): |f_\theta(x) - g_\theta(x)| \le O(\beta \cdot \frac{n}{\sigma^2})$

Proof: Suppose $h(\theta)$ is a function s.t. $\nabla h(\theta)$ is β-Lipschitz:

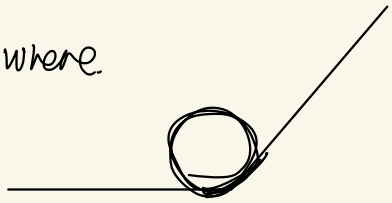
$\Rightarrow \|\nabla^2 h(\theta)\|_{op} \le \beta$ if h is twice-differentiable.

$$\Big|\underbrace{h(\theta)}_{f_\theta(x)} \underbrace{- h(\theta^0) - \langle \nabla h(\theta^0), \theta - \theta^0 \rangle}_{-g_\theta(x)}\Big| \le O(\beta \cdot \|\theta - \theta^0\|_2^2)$$

$\Rightarrow |f_\theta(x) - g_\theta(x)| \le O(\beta \cdot \|\Delta\theta\|_2^2)$ $\square$

Note. If f_θ involves ReLU. ∇f_θ is not continuous everywhere.

$\rightarrow$ require some fixes



In $B(\theta^0)$, $\exists$ global minimum $\hat{\theta}$.

and approximation error $|f_\theta(x) - g_\theta(x)| \le O(\beta \cdot \frac{n}{\sigma^2})$

Is it good?

Two cases where $\frac{\beta}{\sigma^2} \to 0$:

① Reparameterization with a scalar [Chizat and Bach '18]

& Lazy training

$$f_\theta(x) = \alpha \cdot \bar{f}_\theta(x)$$

standard NN. fixed width, depth (base)

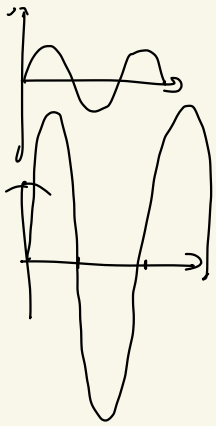
fix initialization θ^0. we only change α.

Let $\bar{\sigma} = \sigma_{min}\left(\begin{bmatrix} \nabla_\theta \bar{f}_{\theta^0}(x^{(1)})^T \\ \vdots \\ \nabla_\theta \bar{f}_{\theta^0}(x^{(n)})^T \end{bmatrix}\right)$ $\quad \bar{\beta}$: Lipschitzness of $\nabla_\theta \bar{f}_\theta(x)$ in θ

$\Rightarrow \sigma = \alpha \cdot \bar{\sigma}$. $\quad \beta = \alpha \cdot \bar{\beta}$

$\Rightarrow \frac{\beta}{\sigma^2} = \frac{\bar{\beta}}{\bar{\sigma}^2} \cdot \frac{1}{\alpha} \to 0$. $\alpha \to \infty$

$\alpha \uparrow$ f_θ: sharper. non-smoothness



e.g. $(x,y) = (1,1)$ $\quad \bar{f}_\theta(x) = x \cdot (\theta + \beta\theta^2) = \theta + \beta\theta^2$

1 datapoint. $\quad \hat{L}(\bar{f}_\theta) = (1 - (\theta + \beta\theta^2))^2$

Let $\theta^0 = 0$ $\quad \bar{g}_\theta(x) = \theta x = \theta$. $\Rightarrow \hat{L}(\bar{g}_\theta) = (1-\theta)^2$

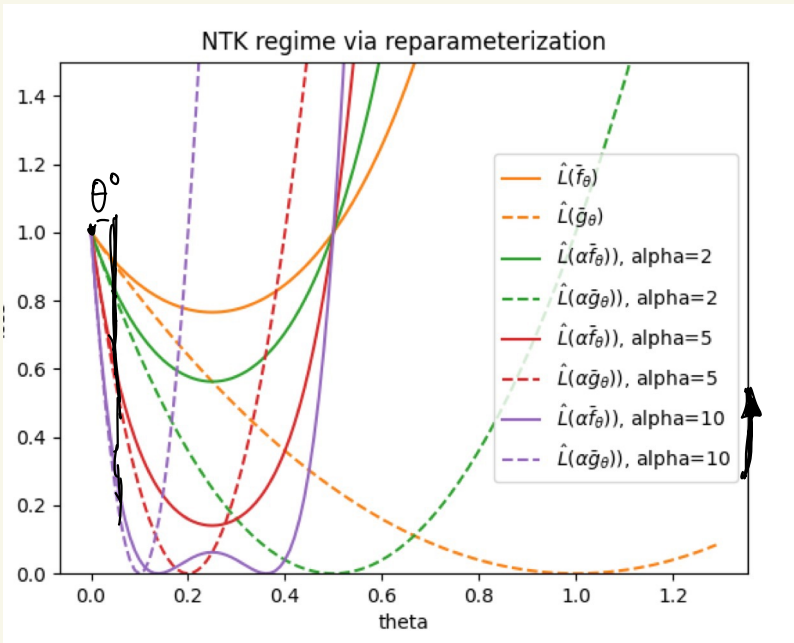
$\nabla f_{\theta^0}(x) = 2\beta\theta x = 2\beta\theta$ is 2β-Lipschitz. $\sigma = 1$.

$\hat{L}(\alpha \bar{f}_\theta) = (1 - \alpha(\theta + \beta\theta^2))^2$. $\quad \alpha\beta\theta^2$

$\hat{L}(\alpha \bar{g}_\theta) = (1 - \alpha\theta)^2$. minimizer $\frac{1}{\alpha}$

$\alpha \to \infty$. $\quad \frac{1}{\alpha} \to 0 = \theta^0$

$B(\theta^0) = [0, \frac{1}{\alpha}]$



However, reparameterization changes the loss function.

—— : $\bar{f}_\theta$

---- : $\bar{g}_\theta$

② Overparameterization (with specific initialization)

model (2-layer NN): $\hat{y} = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} a_i \cdot \sigma(w_i^T x)$

for convenience.

$$W = \begin{bmatrix} w_1^T \\ \vdots \\ w_m^T \end{bmatrix} \in \mathbb{R}^{m \times d}.$$

A: 1-Lipschitz. twice-differentiable

$a_i \overset{\text{uniform}}{\sim} \{\pm 1\}$ initialy not optimized at all

$w_i^0 \sim N(0, I_d)$.

$\|x\|_2 = \Theta(1)$

$\Theta = \text{Vector}(W) \in \mathbb{R}^{dm}$

$m \to \infty$ $(m = \text{poly}(n, d))$

overparameterization

Infinite-width Neural Network

signal of NTK regime.

$$A(w_i^T x) \approx 1.$$

$$\left| \sum_{i=1}^m a_i A(w_i^T x) \right| = \Theta(\sqrt{m})$$

$$f_{\theta^0}(x) = \Theta(1)$$

$$\sigma = \sigma_{min}(\Phi) = \sqrt{\sigma_{min}(\Phi \Phi^T)}$$

empirical kernel matrix

$$\boxed{(\Phi \Phi^T)_{ij} = \langle \nabla_\theta f_{\theta^0}(x^{(i)}), \nabla_\theta f_{\theta^0}(x^{(j)}) \rangle}$$

$$\frac{\partial f_\theta(x)}{\partial w_i} = \frac{1}{\sqrt{m}} \cdot A'(w_i^T x) \cdot x \in \mathbb{R}^d$$

$$\|\nabla f_\theta(x)\|_2^2 = \sum_{i=1}^m \left\| \frac{\partial f_\theta(x)}{\partial w_i} \right\|_2^2 = \frac{1}{m} \sum_{i=1}^m \| A'(w_i^T x) \cdot x \|_2^2$$

$$= \frac{1}{m} \|x\|_2^2 \cdot \sum_{i=1}^m \| A'(w_i^T x) \| \xrightarrow[m \to \infty]{\text{LLN}} \|x\|_2^2 \cdot \mathbb{E}_{w \sim N(0, I_d)} \left[A'(w^T x) \right] = O(1)$$

not depends on m.

Similiarly $\langle \nabla f_{\theta^0}(x), \nabla f_{\theta^0}(x') \rangle$

$$= \frac{1}{m} \langle x, x' \rangle \cdot \sum_{i=1}^m A'(w_i^T x) \cdot A'(w_i^T x') \xrightarrow{\text{LLN}} \langle x, x' \rangle \cdot \mathbb{E}_{w \sim N(0, I_d)} \left[A'(w^T x) \cdot A'(w^T x') \right]$$

$$\Rightarrow \boxed{\Phi \Phi^T \to K^\infty}$$: constant matrix as $m \to \infty$

NTK with $m \to \infty$ (infinite width)

$$K^\infty = \lim_{m \to \infty} \Phi \Phi^T$$

full-rank

$$\sigma_{min} \triangleq \sigma_{min}(K^\infty) > 0. \Rightarrow \sigma = \sqrt{\sigma_{min}(\Phi \Phi^T)} > \frac{1}{2} \sigma_{min}.$$

Intuition: $\Phi \Phi^T \to K^\infty$. then the spectrum of the matrix should also converge

$$\nabla f_\theta(x) - \nabla f_{\theta'}(x) = \left[\frac{1}{\sqrt{m}} \left(A'(w_i^T x) - A'(w_i'^T x) \right) \cdot x \right]_{i=1}^{m} \in \mathbb{R}^{m \times d}.$$

$$\|\nabla f_\theta(x) - \nabla f_{\theta'}(x)\|_2^2 = \frac{1}{m} \underbrace{\|x\|_2^2}_{\le 1} \cdot \sum_{i=1}^{m} \left(A'(w_i^T x) - A'(w_i'^T x) \right)^2.$$

↳ A' is $O(1)$-Lipschitz

$$\le O\left(\frac{1}{m} \cdot \sum_{i=1}^{m} (w_i^T x - w_i'^T x)^2 \right)$$

↳ Cauchy-Schwarz

$$= O\left(\frac{1}{m} \sum_{i=1}^{m} \|w_i - w_i'\|_2^2 \cdot \underbrace{\|x\|_2^2}_{\le 1} \right)$$

$$= O\left(\frac{1}{m} \|\theta - \theta'\|_2^2 \right)$$

$\Rightarrow \|\nabla f_\theta(x) - \nabla f_{\theta'}(x)\|_2 \le \frac{C}{\sqrt{m}} \cdot \|\theta - \theta'\|_2$ - C is a constant

$\Rightarrow$ Lipschitzness $\beta = \frac{C}{\sqrt{m}} \to 0 \quad (m \to \infty)$

$\frac{\beta}{\sigma^2} = \frac{1}{\sqrt{m}} \cdot \frac{1}{\sigma_{min}^2} \to 0$ as $m \to \infty$: We can add more neurons (overparameterization)

to make the function more smooth

Analysis of NTK Approach (II)

NTK: $K(x, x') = \langle \Phi(x), \Phi(x') \rangle = \langle \nabla_\theta f_{\theta^0}(x), \nabla_\theta f_{\theta^0}(x') \rangle$ θ^0

So far: $\boxed{1.}$ Good approximation: $f(\theta) \approx g(\theta)$, $\forall \theta \in B(\theta^0)$

$\boxed{2.}$ Nearby global minimum: $\exists$ global minimum $\hat{\theta} \in B(\theta^0)$ with $\hat{L}(g_{\hat{\theta}}) \approx 0$.

↳ bound: β/σ^2 ($\to 0$) $\begin{cases} \beta: \text{ Lipschitzness of } \nabla_\theta f_\theta(x) \text{ w.r.t } \theta \\ \sigma^2: \text{ minimum singular value of } \Phi \end{cases}$

To do: $\boxed{3.}$ Optimizing $\hat{L}(f_\theta)$ is similiar to optimizing $\hat{L}(g_\theta)$. ☆

linear function

Step 1: Analyze optimization of $\hat{L}(g_\theta)$ via GD

Step 2: Analyze optimization of $\hat{L}(f_\theta)$ via GD, by re-using the proofs in Step 1.

→ two approaches $\begin{cases} \text{strong convexity} \xrightarrow{GD} \text{exponential convergence rate} \\ \text{smoothness} \end{cases}$ ☆

Note. $g_\theta^t(x) = f_{\theta^t}(x) + \langle \nabla f_{\theta^t}(x), \theta - \theta^t \rangle$

$$\begin{cases} f(y) \geq f(x) + \nabla f(x)\cdot(y-x) + \frac{\alpha}{2}\cdot\|y-x\|^2 \\ f(y) \leq f(x) + \nabla f(x)\cdot(y-x) + \frac{\beta}{2}\cdot\|y-x\|^2 \end{cases}$$

$\nabla g_\theta^t(x) = \nabla f_{\theta^t}(x)$

$\Rightarrow \nabla \hat{L}(g_{\theta^t}^t) = \nabla \hat{L}(f_{\theta^t})$

$\Rightarrow$ f_θ and $\boxed{g_\theta^t}$ agree up to first-order at θ^t.

$\Rightarrow$ $\hat{L}(f_\theta)$ and $\hat{L}(g_\theta^t)$ agree up to first-order at θ^t.

GD on $\hat{L}(f_\theta)$ $\Leftrightarrow$ Online Gradient Descent on a sequence of functions $\hat{L}(g_\theta^0), \dots, \hat{L}(g_\theta^T)$

$\rightarrow$ use online learning approach.

Analyzing Optimization with g_θ

linear regression $\quad \min_{\Delta\theta} \|y - \Phi\cdot\Delta\theta\|_2^2$ with GD

$$\Phi = \begin{bmatrix} \nabla_\theta f_{\theta^0}(x^{(1)}) \\ \vdots \\ \nabla_\theta f_{\theta^0}(x^{(n)}) \end{bmatrix} \in \mathbb{R}^{n\times p}$$

$$\boxed{\Delta\theta^{t+1} = \Delta\theta^t - \eta\cdot\Phi^T(\Phi\cdot\Delta\theta^t - y)}$$

let $\hat{y}^t = \Phi\cdot\Delta\theta^t$ prediction

Convergence: error

$$\hat{\varepsilon}^{t+1} = \hat{y}^{t+1} - y = \Phi\cdot\Delta\theta^{t+1} - y$$
$$= \Phi\cdot\left(\Delta\theta^t - \eta\cdot\Phi^T(\Phi\Delta\theta^t - y)\right) - y$$
$$= (\Phi - \eta\cdot\Phi\Phi^T\Phi)\Delta\theta^t - (I - \eta\cdot\Phi\Phi^T)\cdot y$$
$$= (I - \eta\cdot\Phi\Phi^T)\Phi\cdot\Delta\theta^t - (I - \eta\cdot\Phi\Phi^T)\, y$$
$$= (I - \eta\cdot\Phi\Phi^T)\cdot\underbrace{(\hat{y}^t - y)}_{\hat{\varepsilon}^t} \quad \text{shrinking}$$

$\tau^2 = \sigma_{max}(\Phi\Phi^T)$

$\sigma = \sigma_{min}(\Phi) = \sqrt{\sigma_{min}(\Phi\Phi^T)}$ $\quad \eta \leq \frac{1}{\tau^2}$: $\|I - \eta\Phi\Phi^T\|_{op} \leq 1 - \eta\cdot\sigma^2$

$\Phi = U\Sigma V^T \Rightarrow I - \eta\cdot\Phi\Phi^T = I - \eta\cdot U\Sigma^2 U^T$

$= UU^T - \eta\cdot U\Sigma^2 U^T = U(I - \eta\cdot\Sigma^2)\cdot U^T$

$$\|I-\eta\Phi\Phi^T\|_{op} \le \max_j \underbrace{|1-\eta\cdot\tau_j^2|}_{\ge 0} = 1-\eta\cdot\tau_n^2 = 1-\eta\cdot\sigma^2.$$

$$\Rightarrow \|\hat{y}^{t+1}-y\|_2 = \|I-\eta\cdot\Phi\Phi^T\|_{op}\cdot\|\hat{y}^t-y\|_2$$

$$\le (1-\eta\sigma^2)\cdot\|\hat{y}^t-y\|_2 \le \cdots \le (1-\eta\cdot\sigma^2)^{t+1}\cdot\|\hat{y}^0-y\|_2$$

exponential rate

Up to $T=O\left(\frac{\log 1/\varepsilon}{\eta\sigma^2}\right)$: $\|\hat{y}^T-y\|_2 \le \varepsilon\cdot\|\hat{y}^0-y\|_2$

Analyzing Optimization with f_θ

reasonable, $\boxed{1.}$, $\boxed{2.}$

Thm $\exists$ constant $c_0\in(0,1)$ s.t. when $\boxed{\frac{\beta}{\sigma^2}\le\frac{c_0}{\sqrt{n}}}$ for sufficiently small η

(could depend on $\beta, \sigma, p, \cdots$)

$\hat{L}(f_{\theta^T})\le\varepsilon$ after $T=O\left(\frac{\log 1/\varepsilon}{\eta\sigma^2}\right)$ steps.

Proof: $\Phi^t = \begin{bmatrix}\nabla_\theta f_{\theta^t}(x^{(1)})^T\\ \vdots \\ \nabla_\theta f_{\theta^t}(x^{(n)})^T\end{bmatrix}\in\mathbb{R}^{n\times p}$.

$$\hat{L}(f_{\theta^t}) = \frac{1}{n}\sum_{i=1}^n\left(y^{(i)}-f_{\theta^t}(x^{(i)})\right)^2$$

$$\nabla_\theta\hat{L}(f_{\theta^t}) = -\sum_{i=1}^n\left(y^{(i)}-\boxed{f_{\theta^t}(x^{(i)})}\right)\cdot\nabla_\theta f_{\theta^t}(x^{(i)})$$

(scaled by n)

$$= -\sum_{i=1}^n\left(y^{(i)}-\hat{y}^{(i),t}\right)\cdot\Phi_i^t$$

$$= (\Phi^t)^T\cdot(\hat{y}^t-y) \triangleq b_t.$$

$$\boxed{\theta^{t+1}=\theta^t-\eta\cdot b_t.}$$

Because f is nonlinear, using Taylor Expansion:

$$f_{\theta^{t+1}}(x^{(i)}) = f_{\theta^t}(x^{(i)}) + \langle\nabla_\theta f_{\theta^t}(x^{(i)}), \theta^{t+1}-\theta^t\rangle + \text{h.o. terms}$$

$$= f_{\theta^t}(x^{(i)}) + \langle\nabla_\theta f_{\theta^t}(x^{(i)}), -\eta\cdot b_t\rangle + O\left(\|\theta^{t+1}-\theta^t\|_2^2\right)$$

ignore $O(\eta^2)$ $(\eta\to 0)$

Ignoring the high-order terms:

$\hat{y}^{t+1} = \hat{y}^t - \eta\cdot\Phi^t\cdot b_t.$ vector form

$$\Rightarrow \hat{y}^{t+1} = \hat{y}^t - \eta \cdot \Phi^t \cdot (\Phi^t)^T \cdot (\hat{y}^t - y)$$

$$\boxed{\hat{y}^{t+1} - y = \left(I - \eta \cdot \Phi^t (\Phi^t)^T\right) \cdot (\hat{y}^t - y)}$$ decreasing!

$$\hat{y}^{t+1} - y = (I - \eta \cdot \Phi \Phi^T) \cdot (\hat{y}^t - y)$$

Suppose $\boxed{\|\theta^t - \theta^0\| \le \frac{\sigma}{4\beta}}$ at time t $\quad (*)$

then $\|\Phi^t - \Phi\|_2 \le \frac{\sigma}{4}$ (by Lipschitzness: $\nabla_\theta f_\theta(x)$: β-Lipschitz in θ)

$$\sigma_{\min}(\Phi^t) = \min_{\|x\|_2 = 1} x^T \Phi^t x \qquad \Phi = \nabla_\theta f_{\theta^0}(x)$$

$$\ge \min_{\|x\|_2=1} x^T(\Phi^t - \Phi)x + \underbrace{\min_{\|x\|_2=1} x^T \Phi x}_{\sigma} \ge \sigma - \frac{\sigma}{4} = \frac{3}{4}\sigma$$

$$x^T(\Phi^t - \Phi)x \ge -|\langle x, (\Phi^t - \Phi)x\rangle|$$

$$\ge -\underbrace{\|x\|_2}_{(=1)} \cdot \|(\Phi^t - \Phi)x\|_2 \ge -\|\Phi^t - \Phi\|_2 \ge -\frac{\sigma}{4}$$

$$\Rightarrow \|I - \eta \cdot \Phi^t (\Phi^t)^T\|_{op} \le 1 - \eta \cdot \left(\frac{3}{4}\sigma\right)^2 \quad \star$$

$$\Rightarrow \|\hat{y}^{t+1} - y\|_2 \le \left(1 - \eta \cdot \left(\frac{3}{4}\sigma\right)^2\right) \cdot \|\hat{y}^t - y\|_2$$

$$\le \left(1 - \eta \cdot \left(\frac{3}{4}\sigma\right)^2\right)^{t+1} \cdot \|\hat{y}^0 - y\|_2$$

$$\|\hat{y}^{t+1} - y\|_2 \le (1 - \eta\sigma^2) \cdot \|\hat{y}^0 - y\|_2$$

Within $T = O\left(\frac{\log \frac{1}{\varepsilon}}{\eta \sigma^2}\right)$ steps, $\|\hat{y}^T - y\|_2 \le \varepsilon \cdot \|\hat{y}^0 - y\|_2$

We need now to show that $\|\theta^t - \theta^0\| \le \frac{\sigma}{4\beta}$ (induction) $\quad n \cdot \frac{\beta}{\sigma^2} \to 0$

expected!

Formally, $\frac{1}{\sqrt{n}} \|\hat{y}^0 - y\| \le O(1)$ $\qquad \|\Delta\hat{\theta}\|_2 \le \frac{\sqrt{n}}{\sigma} \ll \frac{\sigma}{4\beta}$

$$\Rightarrow \frac{1}{\sqrt{n}} \|\hat{y}^t - y\| \le O(1) \quad \text{(error decreasing)}$$

$$\Rightarrow \frac{1}{\sqrt{n}} \|\Phi \cdot (\theta^t - \hat{\theta})\| \le O(1) \qquad (\hat{\theta} = \Phi^\dagger y)$$

$$\Rightarrow \|\theta^t - \hat{\theta}\|_2 \le O\left(\frac{\sqrt{n}}{\sigma}\right)$$

Plus $\|\theta^0 - \hat{\theta}\|_2 \le O\left(\frac{\sqrt{n}}{\sigma}\right)$ by $\boxed{2.}$

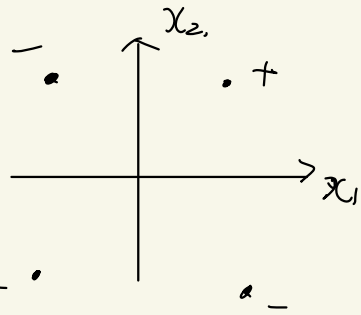
by triangular inequality, $\|\theta^t - \theta^0\|_2 \le \|\theta^t - \hat{\theta}\|_2 + \|\theta_0 - \hat{\theta}\|_2$

$$\le O\left(\frac{\sqrt{n}}{\sigma}\right) \le \frac{\sigma}{4\beta}. \quad \left(\frac{\beta}{\sigma^2} \le \frac{c_0}{\sqrt{n}} \text{ in assumption}\right) \quad \square$$

Limitations of NTK

Q: How well a kernel method can work?

Example. $x \in \mathbb{R}^d$, $y \in \{\pm 1\}$, $x_i \in \{\pm 1\}$, $y = x_1 x_2$



If using Neural Networks, regularizing ℓ_2-norm

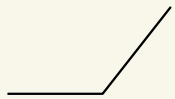
(equivalent to regularize $\boxed{C(\theta) = \sum_i |a_i| \cdot \|w_i\|}$)

"path norm"

$$\hat{y} = \sum_{i=1}^{m} a_i \cdot \sigma(w_i^T x)$$

the best solution (min-norm solution) uses a sparse combination of neurons.

$$y = \frac{1}{2}\left(\mathrm{ReLU}(x_1+x_2) + \mathrm{ReLU}(-x_1-x_2) - \mathrm{ReLU}(x_1-x_2) - \mathrm{ReLU}(x_2-x_1)\right)$$
$$= \frac{1}{2}\left(|x_1+x_2| - |x_1-x_2|\right) = x_1 x_2$$

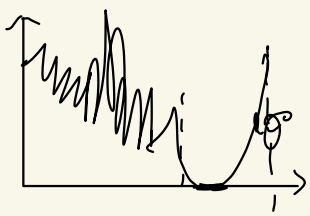


$\mathrm{ReLU}(t) + \mathrm{ReLU}(-t) = |t|$

However, if using NTK, doesn't learn any features

$$\hat{y} = \sum_{i=1}^{m} a_i \cdot \phi_i(x)$$

no sparsity (use all dimensions)



NN ($\Leftrightarrow \ell_1$-SVM): $\min \|a\|_1$ s.t. $\sum_{i=1}^{m} a_i \phi_i(x^{(j)}) = y^{(j)}$

NTK ($\Leftrightarrow \ell_2$-SVM): $\min \|a\|_2$ s.t. $\sum_{i=1}^{m} a_i \phi_i(x^{(j)}) = y^{(j)}$

Thm Kernel Method with NTK regime requires $n = \Omega(d^2)$ sample to learn the problem with error < 1. / Regularized NNs only need $n = O(d)$ samples.

Part 4-2 Implicit Regularization

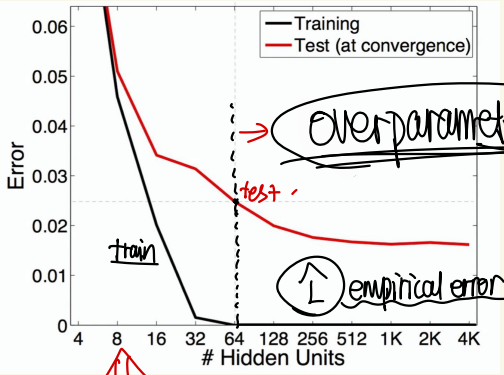
Implicit/Algorithmic Regularization Effect

Implicit Bias

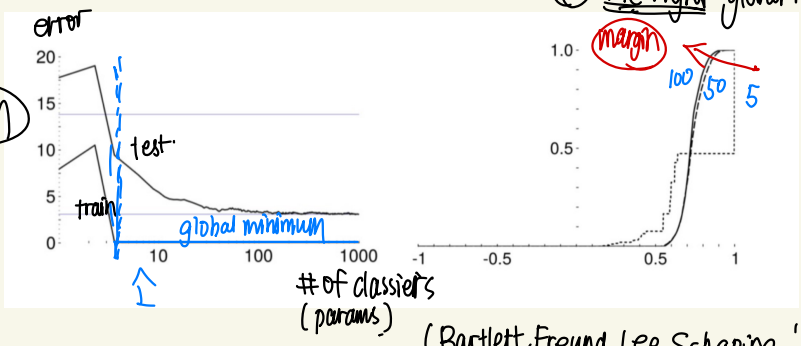
GD/SGD ① a global minimum

② the right global minimum

Neural Network:

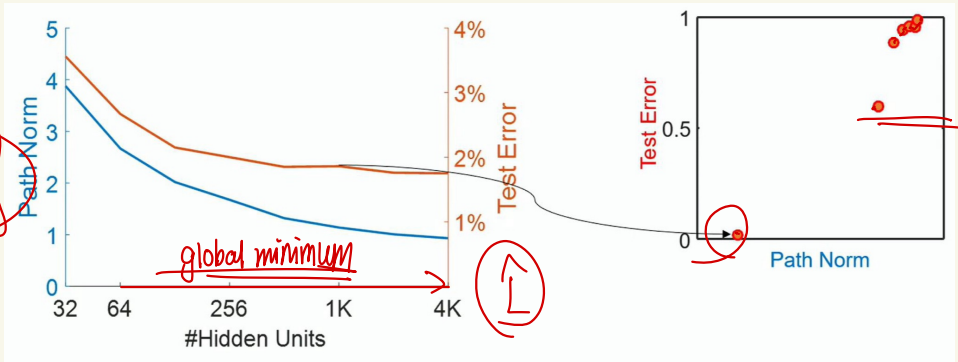


AdaBoost:



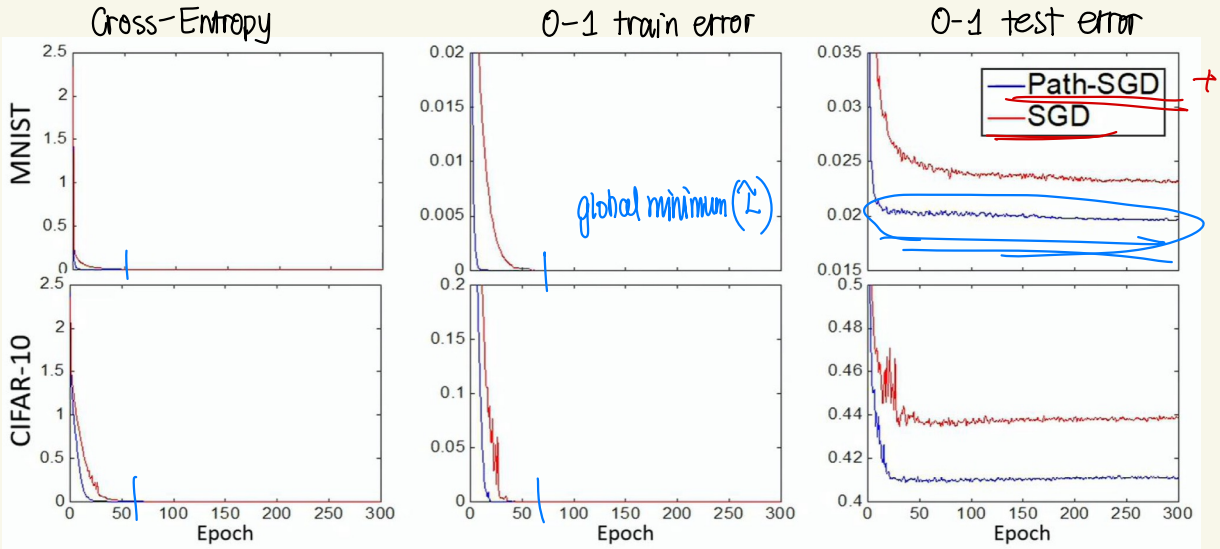
(Bartlett, Freund, Lee, Schapire '98)

Complexity measure.



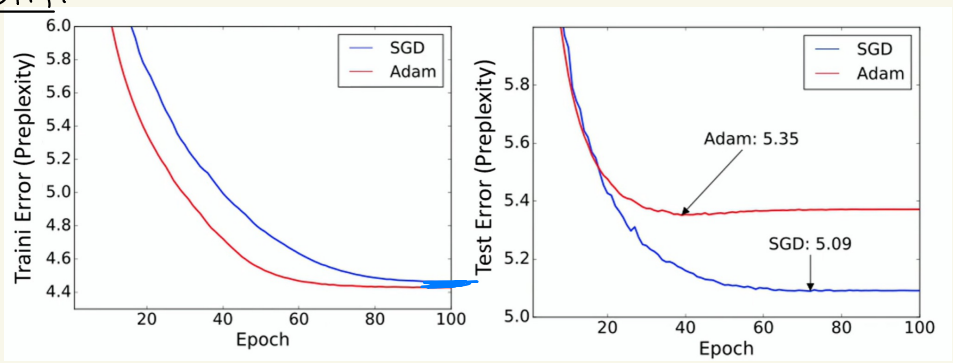
$\min \frac{1}{2}\|w\|_2^2$ s.t. $y^{(i)} f_w(x^{(i)}) \geq 1$

$\hat{L}(\theta) + \lambda \cdot R(\theta)$

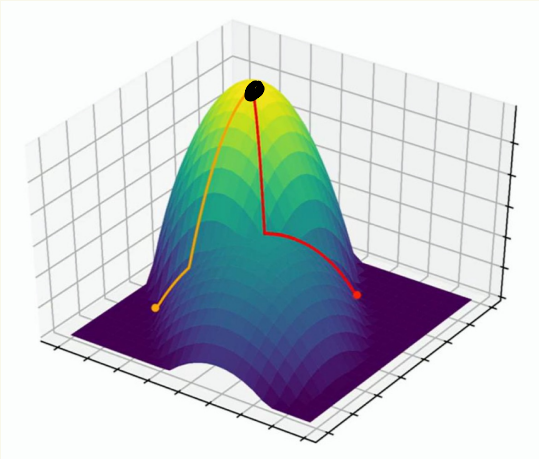


(Neyshabur, Salakhudtinov, Srebro '15)

3-layer LSTM:



(Wilson, Roelofs, Stein, Srebro, Recht '17)



Different Optimization Algorithm (e.g. GD. SGD)

→ Different bias in optimum reached

→ Different inductive bias

→ Different generalization properties

↳ population loss (test error)

Implicit Regularization

- multiple global minima of training loss exist. (overparameterization)
- Optimizer has some implicit preferences

Initialization

Stochasticity (SGD)

Overparameterized Linear Regression with Zero Initialization.

Takeaway: GD finds the min-norm solution.

Setup: $X = \begin{bmatrix} - x^{(1)T} - \\ \vdots \\ - x^{(n)T} - \end{bmatrix} \in \mathbb{R}^{n \times d}$. $\quad y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} \in \mathbb{R}^n$. $\quad \boxed{n \ll d}$ → overparameterized

full rank (rank: n)

$y \sim X\beta$

loss function: $\boxed{\hat{L}(\beta) = \frac{1}{2} \|y - X\beta\|_2^2}$

$\hat{L}(\beta)$ has infinite # of global minimum. (with loss 0)

Claim: β is a global minimizer $\Leftrightarrow$ $\beta = X^+ y + \xi$ s.t. $\xi \perp x^{(1)}, \dots, x^{(n)}$ $\quad (X^T)^\perp$

$\in \text{span}\{x^{(1)}, \dots, x^{(n)}\}$

Proof: $\forall \beta \in \mathbb{R}^d$. $\beta = X^+ y + \xi$ for some $\xi \in \mathbb{R}^d$.

$X^+ = X^T (XX^T)^{-1}$

$X\beta = X(X^+ y + \xi) = y + X\xi$

β is a global minimum $\Leftrightarrow$ $\hat{L}(\beta) = 0$

$\Leftrightarrow X\beta = y$

$\Leftrightarrow X\xi = 0$

$\Leftrightarrow \begin{bmatrix} x^{(1)T}\xi \\ \vdots \\ x^{(n)T}\xi \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$

$\Leftrightarrow \xi \perp x^{(1)}, \dots, x^{(n)}$

Moore-Penrose inverse.

A^+ s.t. $\begin{cases} AA^+A = A \\ A^+AA^+ = A^+ \\ (AA^+)^T = AA^+ \\ (A^+A)^T = A^+A. \end{cases}$

$A \in \mathbb{R}^{n \times d}$ (full-rank) $\begin{cases} n \geq d: & A^+ = (A^TA)^{-1}A^T \\ n \leq d: & A^+ = A^T(AA^T)^{-1} \end{cases}$

More generally, $A = U\Sigma V^T$ $\quad \begin{cases} U \in \mathbb{R}^{n \times r} \\ \Sigma \in \mathbb{R}^{r \times r} \\ V \in \mathbb{R}^{d \times r} \end{cases}$

$A^+ = V\Sigma^{-1}U^T$ $\quad \text{rank}(A) = r$.

Let $\boxed{\beta^* = \underset{\beta:\hat{L}(\beta)=0}{\operatorname{argmin}} \|\beta\|_2}$ be the min-norm solution.

Claim: $\beta^* = X^+ y$

Proof: global minimum $\beta = X^+ y + \xi$, $\xi \perp x^{(1)}, \dots, x^{(n)}$

$$\|\beta\|_2^2 = \|X^+ y\|_2^2 + \|\xi\|_2^2 + 2\underbrace{\langle X^+ y, \xi\rangle}_{=0}$$

$$= \|X^+ y\|_2^2 + \|\xi\|_2^2$$

$$\geq \|X^+ y\|_2^2 = \|\beta^*\|_2^2$$

$$\left(\begin{aligned} &\langle X^+ y, \xi\rangle \\ &= \langle X^T (XX^T)^{-1} y, \xi\rangle \\ &= \langle (XX^T)^{-1} y, \underbrace{X\xi}_{=0}\rangle = 0 \end{aligned} \right)$$

$\xi = 0 \Leftrightarrow \beta = \beta^*$ $\square$

GD: $\beta^t = \beta^{t-1} - \eta \cdot \nabla \hat{L}(\beta^{t-1})$ $\quad \beta^0 = 0$

Thm GD on $\hat{L}(\beta)$ with initialization $\beta^0 = 0$ converges to a solution $\hat{\beta} = \beta^*$ with $\hat{L}(\hat{\beta}) = 0$

min-norm global minimum $\qquad$ global minimum

Proof: 1. $\hat{L}$: convex. $\hat{L}(\beta^t) \to 0$ as $t \to \infty$ by standard convex optimization theory.

2. Key: $\beta^t \in \operatorname{span}\{x^{(1)}, \dots, x^{(n)}\}$

Induction: $\beta^0 = 0 \in \operatorname{span}\{x^{(1)}, \dots, x^{(n)}\}$

Assume $\beta^{t-1} \in \operatorname{span}\{x^{(1)}, \dots, x^{(n)}\}$

$$\text{GD: } \beta^t = \beta^{t-1} - \eta \cdot \underbrace{\nabla \hat{L}(\beta^{t-1})}_{\in \operatorname{span}\{x^{(1)}, \dots, x^{(n)}\}} = \beta^{t-1} - \eta \cdot X^T (X\beta^{t-1} - y)$$

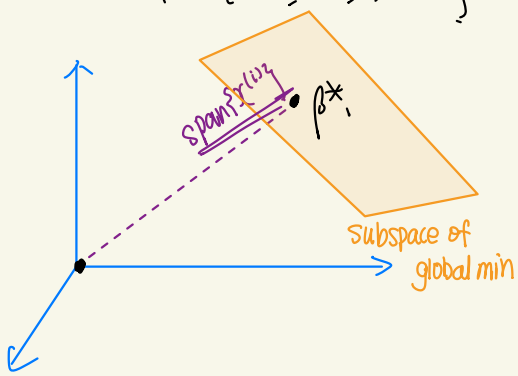
$$X^T c = \begin{bmatrix} | & & | \\ x^{(1)} & \cdots & x^{(n)} \\ | & & | \end{bmatrix} \cdot \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} | \\ c_1 x^{(1)} + \cdots + c_n x^{(n)} \\ | \end{bmatrix} \in \operatorname{span}\{x^{(1)}, \dots, x^{(n)}\}$$

① $\hat{L}(\hat{\beta}) = 0 \Rightarrow X\hat{\beta} = y$

② $\hat{\beta} \in \operatorname{span}\{x^{(1)}, \dots, x^{(n)}\} \Rightarrow \hat{\beta} = X^T v$ for some $v \in \mathbb{R}^n$,

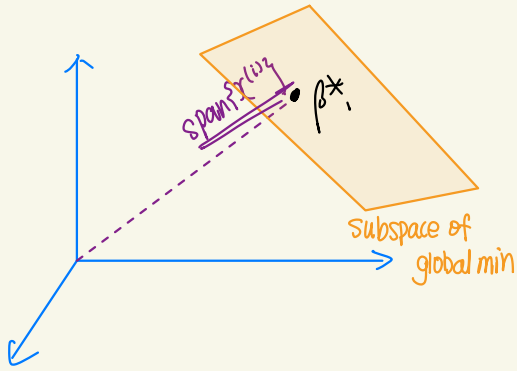
$$\Rightarrow \underbrace{XX^T}_{\in \mathbb{R}^{n\times n}} v = y \Rightarrow v = (XX^T)^{-1} y$$

$$\Rightarrow \hat{\beta} = X^T v = X(XX^T)^{-1} y = X^+ y = \beta^* \quad \square$$



$$\boxed{\text{GD converges to min-norm solution. } \hat{\beta} = \operatorname{argmin} \|\beta\|_2^2 \text{ s.t. } \hat{L}(\beta) = 0}$$

GD on Sparse Nonlinear Model



Recall: overparameterized linear model, gradient descent, zero initialization $\Rightarrow$ min-norm solution.

Non-linear Model (Li et al'17)

$$f_\beta(x) = \langle \beta \odot \beta, x \rangle$$

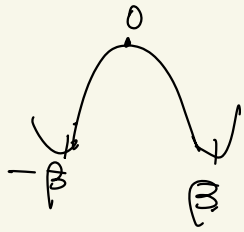
- linear in x
- non-linear in β.

$\odot$: Hadamard product (element-wise)

$$\begin{bmatrix}1\\2\\3\end{bmatrix} \odot \begin{bmatrix}4\\5\\6\end{bmatrix} = \begin{bmatrix}4\\10\\18\end{bmatrix}$$

$$\boxed{\hat{L}(\beta) = \frac{1}{4n}\sum_{i=1}^{n}\left(y^{(i)} - f_\beta(x^{(i)})\right)^2}$$

non-convex



$\hat{L}(0) > \hat{L}(\beta) = \hat{L}(-\beta)$

ground truth: $y = \langle \beta^* \odot \beta^*, x \rangle$

β^*: r-sparse. i.e. $\|\beta^*\|_0 \le r$ (r non-zero entries)

$$\begin{cases} n < d. & \Rightarrow n \ll d \\ n > \mathrm{poly}(r) & \Rightarrow n \gtrsim \tilde{\Omega}(r^2) \end{cases}$$

overparameterization

wlog. $\beta^* \succeq 0$. $\boxed{\beta^* = \mathbb{1}\{i \in S\}}$ for some $S \subseteq [d]$. $|S| = r$.

$x^{(1)}, \dots, x^{(n)} \overset{i.i.d}{\sim} \mathcal{N}(0, I_d)$

$y^{(i)} = \langle \beta^* \odot \beta^*, x^{(i)} \rangle$

Classical solution: ℓ_1-regularization (LASSO) to encourage sparsity

$$f_\theta(x) = \langle \theta, x \rangle$$

LASSO: $\boxed{\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \theta^T x^{(i)}\right)^2 + \lambda \cdot \|\theta\|_1 \quad ①}$ $\quad n \gtrsim \tilde{\Omega}(r)$

$\theta = \beta \odot \beta$ $\qquad \|\theta\|_1 = \sum_{i=1}^{d}|\theta_i| = \sum_{i=1}^{d}\beta_i^2 = \|\beta\|_2^2$.

$$① \Leftrightarrow \boxed{\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - f_\beta(x^{(i)})\right)^2 + \lambda \cdot \|\beta\|_2^2 \quad ②}$$

Implicit Regularization: use small initialization without explicit regularization $\Rightarrow$ ②

$$\hat{L}(\beta) = \frac{1}{4n}\sum_{i=1}^{n}\left(y^{(i)} - \langle \beta \odot \beta, x^{(i)} \rangle\right)^2$$

Optimizer: GD on $\hat{L}(\beta)$ with small initialization

$$\boxed{\beta^0 = \alpha \cdot \vec{1}} \quad \text{for some very small } \alpha > 0$$

$\notin S_r$

$$\beta^{t+1} = \beta^t - \eta \cdot \nabla \hat{L}(\beta^t)$$

$\boxed{n = \tilde{\Omega}(r^2) \text{ this optimization converges to } \beta^* \text{ with small } \alpha}$

Thm Let c be a sufficiently large constant, $n \geq c \cdot r^2 \cdot \log^2 d$ $\quad \alpha \leq \frac{1}{d^c}$

With $\frac{\log \frac{d}{\alpha}}{\eta} \lesssim T \lesssim \frac{1}{\eta \cdot \sqrt{d\alpha}}$ steps. $\|\beta^T \odot \beta^T - \beta^* \odot \beta^*\|_2^2 \leq O(\alpha \cdot \sqrt{d})$

Interpretation:

- $\hat{L}(\beta)$ has many global minima (b.c. overparameterization)
- Runtime lower bound depends only on $\log \frac{1}{\alpha}$

  α can be inverse poly, error depends on α. $\Rightarrow$ α small
- Runtime upper bound (pretty mild) $\rightarrow$ don't need early stopping
- α cannot be zero.: $\beta = 0$ is saddle point $\nabla \hat{L}(0) = 0$

GD prefers global minimum closest to initialization ?

In this case: $\beta^* = \arg\min_\beta \|\beta\|_2^2$ s.t. $y^{(i)} = \langle \beta \odot \beta, x^{(i)} \rangle$

$\Leftrightarrow \theta^* = \arg\min_\theta \|\theta\|_1$ s.t. $y^{(i)} = \langle \theta, x^{(i)} \rangle$

Some basic properties:

$L(\beta) \leftarrow \hat{L}(\beta)$ convergence

$$L(\beta) = \frac{1}{4} \mathbb{E}_x \left[\left(y - \langle \beta \odot \beta, x \rangle \right)^2 \right]$$

$$= \frac{1}{4} \mathbb{E}_x \left[\langle \beta^* \odot \beta^* - \beta \odot \beta, x \rangle \right]^2$$

$$= \frac{1}{4} \|\beta^* \odot \beta^* - \beta \odot \beta\|_2^2$$

$$\hat{L}(\beta) = \frac{1}{4n} \sum_{i=1}^{n} \langle \beta^* \odot \beta^* - \beta \odot \beta, x^{(i)} \rangle^2$$

$$\mathbb{E}_x \left[\langle x, v \rangle^2 \right] = \mathbb{E}_x \left[v^T x x^T v \right]$$

$$= v^T \mathbb{E}_x [x x^T] v$$

$$= v^T v \qquad \mathcal{N}(0, I)$$

$$= \|v\|_2^2$$

$\hat{L}(\beta) \approx L(\beta)$ Uniform Convergence for sparse β.

[but NOT for the entire space since overparameterization !!!]

Lemma $n \geq \tilde{\Omega}(r^2)$, W.h.p (over the randomness $x^{(1)}, \dots, x^{(n)}$)

$$(1-\delta)\|v\|_2^2 \leq \frac{1}{n}\sum_{i=1}^{n} \langle v, x^{(i)} \rangle^2 \leq (1+\delta)\|v\|_2^2 . \quad \forall \text{ 1-sparse vector } v \in \mathbb{R}^d \quad (\|v\|_0 \leq r) \quad ③$$

$$③ \Leftrightarrow v^T I v (1-\delta) \leq v^T \left(\frac{1}{n}\sum_{i=1}^{n} x^{(i)} x^{(i)T}\right) v \leq v^T I v (1+\delta)$$

Suppose require for every $v \in \mathbb{R}^d$: $(1-\delta) I \preceq \frac{1}{n}\sum_{i=1}^{n} x^{(i)} x^{(i)T} \preceq (1+\delta) I$

$\hookrightarrow$ covariance of X is close to I_d

But we only have ③ for sparse v!

$$\hat{L}(\beta) = \frac{1}{4}\cdot\frac{1}{n}\sum_{i=1}^{n} \langle \underbrace{\beta^* \odot \beta^* - \beta \odot \beta}_{v}, x \rangle^2 \approx \frac{1}{4}\|\beta^* \odot \beta^* - \beta \odot \beta\|_2^2 = L(\beta) \quad \boxed{\text{for sparse } \beta}$$

OTOH: $\exists$ dense β s.t. $\hat{L}(\beta) = 0$ but $L(\beta) \gg 0$!

no convergence

def (r, δ)-RIP condition: dataset X satisfying ③

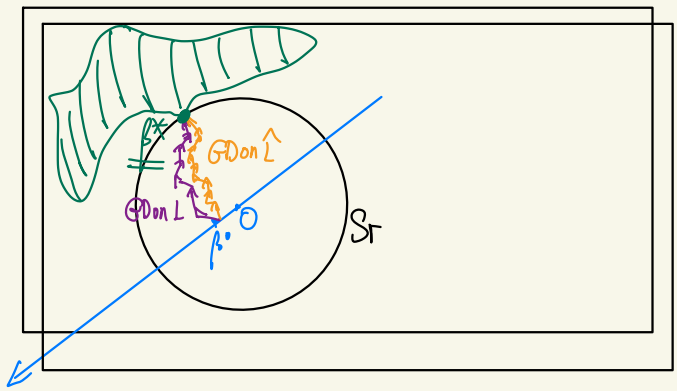
Restricted Isometry Property

Similiarly for sparse β:

$$\nabla \hat{L}(\beta) \approx \nabla L(\beta)$$

$$\boxed{S_r = \{\beta : \|\beta\|_0 \leq O(r)\}} \rightarrow \hat{L} \approx L, \ \nabla \hat{L} \approx \nabla L.$$

Main Intuition:



Initialization: $\alpha \cdot 1 \notin S_r$.

but $\alpha \cdot 1$ is close to $0 \in S_r$.

Step 1: GD on $L(\beta)$ converges to β^* in S_r.

Step 2: GD on $\hat{L}(\beta)$ is similiar to GD on $L(\beta)$ inside S_r.

Analysis of the population trajectory

$$L(\beta) = \frac{1}{4}\|\beta \odot \beta - \beta^* \odot \beta^*\|_2^2$$

$$\nabla L(\beta) = (\beta \odot \beta - \beta^* \odot \beta^*) \odot \beta .$$

$r_t \quad \eta \cdot (r_t^2 - 1) \quad \cdot r_t$

$$\rightarrow \text{GD: } \beta^{t+1} = \beta^t - \eta \cdot (\beta^t \odot \beta^t - \beta^* \odot \beta^*) \odot \beta^t \in \mathbb{R}^d \quad (*)$$

$\Downarrow$

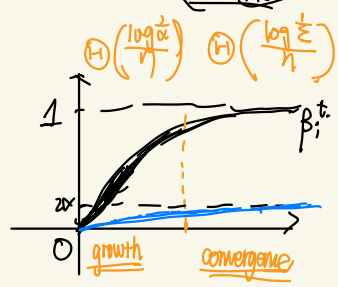
Thm GD on $L(\beta)$ converges to β^* in $O\left(\frac{\log\frac{1}{\varepsilon\alpha}}{\eta}\right)$ steps

within ε-error in ℓ_2-distance. small

d separate updates in d coordinates

Proof: $\beta^* = \mathbb{1}\{i\in S\} \Rightarrow \begin{cases} i\in S: & \beta_i^* = 1 \Rightarrow \beta_i^{t+1} = \beta_i^t - \eta\cdot\left((\beta_i^t)^2 - 1\right)\cdot\beta_i^t \\ i\notin S: & \beta_i^* = 0 \Rightarrow \beta_i^{t+1} = \beta_i^t - \eta\cdot(\beta_i^t)^3. \end{cases}$

$\beta_i^0 = \alpha$



Case (I): $i\in S$.

If $\beta_i^t \le \frac{1}{2}$: $\beta_i^{t+1} = \beta_i^t\left(1+\eta\cdot(1-(\beta_i^t)^2)\right)$

$\ge \beta_i^t\left(1+\frac{3}{4}\eta\right)$ expansion

β_i^t from α to $\frac{1}{2}$: $\left(1+\frac{3}{4}\eta\right)^{t_1} \ge \frac{\frac{1}{2}}{\alpha} \Rightarrow t_1 = \frac{\log\frac{1}{2\alpha}}{\log(1+\frac{3}{4}\eta)} = \Theta\left(\frac{\log\frac{1}{\alpha}}{\eta}\right)$

If $\beta_i^t \ge \frac{1}{2}$: $1-\beta_i^{t+1} = 1-\beta_i^t + \eta\cdot\left((\beta_i^t)^2-1\right)\beta_i^t$

$= 1-\beta_i^t - \eta(1-\beta_i^t)(1+\beta_i^t)\beta_i^t$

$\le 1-\beta_i^t - \eta(1-\beta_i^t)\beta_i^t$

$= (1-\beta_i^t)(1-\eta\cdot\beta_i^t) \le (1-\beta_i^t)\cdot\left(1-\frac{\eta}{2}\right)$ shrinkage.

$1-\beta_i^t$ from $\frac{1}{2}$ to ε: $\left(1-\frac{\eta}{2}\right)^{t_2} \le \frac{\varepsilon}{\frac{1}{2}} \Rightarrow t_2 = \frac{\log 2\varepsilon}{\log(1-\frac{\eta}{2})} = \Theta\left(\frac{\log\frac{1}{\varepsilon}}{\eta}\right)$

$\Rightarrow$ totally needs $\Theta\left(\frac{\log\frac{1}{\alpha\varepsilon}}{\eta}\right)$ steps.

Case (II): It's sufficient to show when $t \le \frac{1}{10\eta\alpha^2}$. $\beta_i^t \le 2\alpha < \varepsilon$. $\checkmark$

$\beta_i^{t+1} = \beta_i^t - \eta\cdot(\beta_i^t)^3$ $\boxed{\text{decreasing } \Delta}$

$t \le \frac{1}{10\eta\alpha^2}$ $\frac{1}{\alpha^2} > \log\frac{1}{\alpha}$

Suppose $\beta_i^s \le 2\alpha$ for all $s\le t$. $i\notin S$.

$\beta_i^{s+1} = \left(1-\eta\cdot(\beta_i^s)^2\right)\cdot\beta_i^s \le \left(1+\eta\cdot(\beta_i^s)^2\right)\cdot\beta_i^s \le (1+\eta\cdot 4\alpha^2)\cdot\beta_i^s$

$\beta_i^{t+1} \le \beta_i^0\cdot(1+4\eta\alpha^2)^t \le \alpha\cdot(1+4\eta\alpha^2)^{\frac{1}{10\eta\alpha^2}} \le \alpha\cdot\exp\left(\frac{4\eta\alpha^2}{10\eta\alpha^2}\right) = \alpha\cdot e^{\frac{2}{5}} \le 2\alpha$ $\square$

Analysis of empirical trajectory

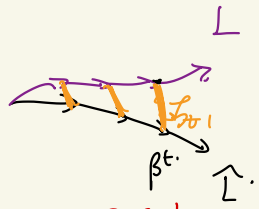
Simplification: β^* is 1-sparse ($r=1$) wlog. $\beta^* = e_1 = \begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}$

Thm Suppose $n \ge \tilde{\Omega}(1)$, $r=1$, GD on $\hat{L}(\beta)$ with $t = \Theta\left(\frac{\log\frac{1}{\alpha\delta}}{\eta}\right)$ steps:

$$\|\beta^t \odot \beta^t - \beta^* \odot \beta^*\|_2^2 \le \tilde{O}\left(\frac{1}{\sqrt{n}}\right)$$ (weaker b.c. error $\not\to 0$)

Proof: Idea: $\hat{L}(\beta) \approx L(\beta)$ $\forall \beta$ that is approximate sparse $\to$ never leave S_r significantly

$$\beta^t = \underbrace{r_t}_{\text{signal}} \cdot \beta^* + \underbrace{\xi_t}_{\text{noise}}$$



1) $r_t \to 1$. 2) $\|\xi_t\|_\infty \le O(\alpha)$ for $t \le \tilde{O}(\frac{1}{\eta})$

Step1: GD update

Step2, noise ξ_t dynamic

Step3: signal r_t dynamic

Step4, putting all together.

Step1: $\beta^* = e_1$. $r_t = \langle \beta^t, e_1 \rangle = \beta_1^t$

$\xi_t \perp e_1$ $\quad \xi_t = \begin{bmatrix} 0 \\ \beta_2^t \\ \vdots \\ \beta_d^t \end{bmatrix}$

$$\nabla \hat{L}(\beta^t) = \frac{1}{n}\sum_{i=1}^n \langle \beta^t \odot \beta^t - \beta^* \odot \beta^*, x^{(i)} \rangle x^{(i)} \odot \beta^t$$

$$= \frac{1}{n}\sum_{i=1}^n \langle r_t^2 \beta^* \odot \beta^* + \xi_t \odot \xi_t - \beta^* \odot \beta^*, x^{(i)} \rangle x^{(i)} \odot \beta^t$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^n \langle (r_t^2 - 1)\beta^* \odot \beta^* + \xi_t \odot \xi_t, x^{(i)} \rangle x^{(i)}}_{m_t} \odot \beta^t = m_t \odot \beta^t$$

Let $\tilde{X} = \frac{1}{n}\sum_{i=1}^n x^{(i)} x^{(i)T}$ be the sample covariance.

$$m_t = \frac{1}{n}\sum_{i=1}^n \langle (r_t^2 - 1)\beta^* \odot \beta^* + \xi_t \odot \xi_t, x^{(i)} \rangle x^{(i)}$$

$$= \tilde{X} \cdot \left((r_t^2 - 1)\beta^* \odot \beta^* + \xi_t \odot \xi_t\right)$$

$$= \tilde{X} \cdot (r_t^2 - 1)(\beta^* \odot \beta^*) + \tilde{X} \cdot (\xi_t \odot \xi_t)$$

Let $u_t = (r_t^2 - 1)(\beta^* \odot \beta^*) - \tilde{X} \cdot (r_t^2 - 1)(\beta^* \odot \beta^*)$. $\quad v_t = \tilde{X} \cdot (\xi_t \odot \xi_t)$

then $\nabla \hat{L}(\beta^t) = m_t \odot \beta^t = \left((r_t^2 - 1)(\beta^* \odot \beta^*) - u_t + v_t\right) \odot \beta^t$

Goal: $\nabla \hat{L}(\beta^t)$ close to $\nabla L(\beta^t)$ $\quad u_t \to 0$, $v_t \to 0$ (WTS)

by Lemma, $(1-\delta) \cdot \|v\|_2^2 \le \frac{1}{n}\sum_{i=1}^n \langle v, x^{(i)} \rangle^2 \le (1+\delta) \cdot \|v\|_2^2$. $\forall v \in S_r$. $\Rightarrow$ (r, δ)-RIP

$\forall v, w \in S_r$: $\left|\frac{1}{n}\sum_{i=1}^n \langle x^{(i)}, v \rangle \langle x^{(i)}, w \rangle - \langle v, w \rangle\right| = \left|v^T \tilde{X} w - v^T w\right| \le 4\delta \cdot \|v\|_2 \cdot \|w\|_2$

$\Downarrow$

$w = e_1, \dots, e_d$ $\quad \left\|\frac{1}{n}\sum_{i=1}^n \langle x^{(i)}, v \rangle x^{(i)} - v\right\|_\infty = \|\tilde{X} \cdot v - v\|_\infty \le 4\delta \cdot \|v\|_2$, $\quad$ ④

$\hookrightarrow (r_t^2 - 1)(\beta^* \odot \beta^*)$

$$\Rightarrow \|u_t\|_\infty \overset{④}{\le} 4\delta \cdot \|(r_t^2-1)\,\beta^*\odot\beta^*\|_2 \le 4\delta\cdot\|\underbrace{\beta^*\odot\beta^*}_{e_1}\|_2 = 4\delta. \quad \checkmark$$

$(|r_t|<1)$

$$\|v_t\|_\infty \le \|h_t\odot h_t\|_\infty + \|v_t - h_t\odot h_t\|_\infty \overset{④}{\le} \|h_t\odot h_t\|_\infty + 4\delta\cdot\|h_t\odot h_t\|_2$$

$$\le \|h_t\|_\infty^2 + 4\delta\cdot\|h_t\odot h_t\|_1 \quad (h_t \text{ small})$$

$$= \|h_t\|_\infty^2 + 4\delta\cdot\|h_t\|_2^2.$$

If we want to bound v_t, then $h_t \to 0$!

Step 2. Dynamics of h_t.

Lemma 1 $\forall\, t \le \frac{1}{c\eta\delta}$, for sufficiently large c, we have

$$\begin{cases} \|h_t\|_\infty \le 2\alpha, & \text{being controlled!} \\ \|h_t\|_2^2 \le \frac{1}{2}, \\ \|h_{t+1}\|_\infty \le (1+O(\eta\delta))\|h_t\|_\infty \end{cases}$$

Proof: $\nabla\hat{L}(\beta^t) = \left((r_t^2-1)\,\beta^*\odot\beta^* - u_t + v_t\right)\odot\beta^t.$

GD: $\beta^{t+1} = \underbrace{\left(\beta^t - \eta\cdot(r_t^2-1)\,\beta^*\odot\beta^*\odot\beta^t\right)}_{\text{GD update for population loss}} - \eta(v_t-u_t)\odot\beta^t,$

$$h_t = \begin{pmatrix} 0 \\ \beta_2^t \\ \vdots \\ \beta_d^t \end{pmatrix} = (I - e_1e_1^T)\beta_t.$$

$$h_{t+1} = (I-e_1e_1^T)\beta^{t+1}$$

$$= \boxed{(I-e_1e_1^T)\beta^t} - \eta\cdot(I-e_1e_1^T)\left((v_t-u_t)\odot\beta^t\right)$$

$$= h_t - \eta\left((I-e_1e_1^T)(v_t-u_t)\right)\odot h_t$$

$$= h_t - \eta\cdot\rho_t\odot h_t$$

$A\cdot(b\odot c) = (Ab)\odot(Ac)$

$A = I - e_1e_1^T$ diagnal $\checkmark$

$\Downarrow$

$\|h_{t+1}\|_\infty \le \|h_t\|_\infty(1+\eta\cdot\|\rho_t\|_\infty)$ where $\rho_t = (I-e_1e_1^T)(v_t-u_t)$

By induction $\|\rho_t\|_\infty \le \|v_t\|_\infty + \|u_t\|_\infty \le \|h_t\|_\infty^2 + 4\delta\cdot\|h_t\|_2^2 + 4\delta$

$$\le 4\alpha^2 + 6\delta \le 8\delta \quad (\alpha \text{ arbitrarily small})$$

$\alpha \le \text{poly}(\frac{1}{n}) \le O(\delta)$

$\delta = \tilde{O}(\frac{1}{\sqrt{n}})$

$\Downarrow$

$\|h_{t+1}\|_\infty \le (1+O(\eta\delta))\cdot\|h_t\|_\infty$

$$\|h_{t+1}\|_\infty \le (1+8\eta\delta)^{t+1}\cdot\|h_0\|_\infty \le (1+8\eta\delta)^{\frac{1}{c\eta\delta}}\cdot\alpha \le 2\alpha \quad \square$$

Step3. Dynamics of r_t

Lemma 2 $\forall t \le \frac{1}{c\eta\delta}$ with sufficiently large c, $\boxed{r_{t+1} = \underbrace{\left(1+\eta\cdot(1-r_t^2)\, r_t\right)}_{\text{GD on population loss}} + O(\eta\delta)\, r_t.}$

Proof: $\beta^{t+1} = \beta^t - \eta\cdot\left((r_t^2-1)\,\beta^*\odot\beta^* - u_t + v_t\right)\odot\beta_t.$

$$r_{t+1} = \langle \beta^{t+1}, e_1\rangle$$

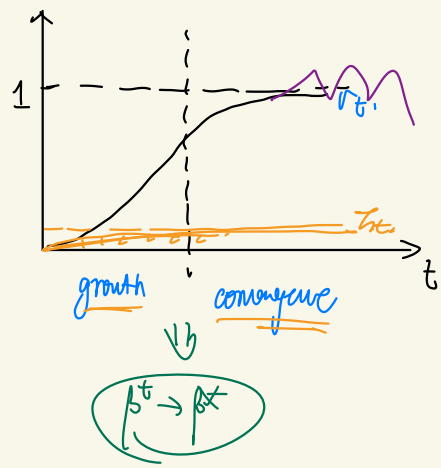
$$= \langle \beta^t, e_1\rangle - \eta\cdot(r_t^2-1)\cdot\langle \beta^t, e_1\rangle - \eta\cdot\langle v_t - u_t, e_1\rangle\cdot\langle \beta^t, e_1\rangle$$

$$= r_t - \eta\cdot(r_t^2-1)\, r_t + \eta\cdot\langle u_t - v_t, e_1\rangle\cdot r_t$$

$$= \left(1+\eta(1-r_t^2)\right) r_t + O(\eta\cdot\delta)\cdot r_t. \quad \square$$

$(u_t - v_t)_1 \le \|u_t - v_t\|_\infty \le \|u_t\|_\infty + \|v_t\|_\infty \le \beta\delta$

Last Step!!! To show that $r_t \to 1$ as $t\to\infty$



If $r_t \le \frac{1}{2}$, $r_{t+1} \ge \left(1+\frac{3}{4}\eta + O(\eta\delta)\right) r_t$

$$\ge \left(1+\frac{3}{4}\eta\right) r_t \Rightarrow t_1 = \Theta\left(\frac{\log\frac{1}{\alpha}}{\eta}\right)$$

If $\frac{1}{2} < r_t \le 1$, $1 - r_{t+1} \le 1 - r_t - \eta(1-r_t^2)\, r_t - O(\eta\delta)\cdot r_t.$

$$\le 1 - r_t - \eta(1-r_t)(1+r_t)\, r_t$$

$$\le \left(1-\frac{\eta}{2}\right)(1-r_t) \Rightarrow t_2 = \Theta\left(\frac{\log\frac{1}{\delta}}{\eta}\right)$$

$O(\delta)$-approximation

If $r_t > 1$ $\quad r_{t+1} - 1 = \left(1+\eta(1-r_t^2)\right) r_t + O(\eta\delta)\, r_t - 1$

$$= (1-\eta)(r_t - 1) + O(\eta\delta) \to t_3 = \Theta\left(\frac{\log\frac{1}{\delta}}{\eta}\right)$$

$\Rightarrow$ total $T = \Theta\left(\frac{\log\frac{1}{\alpha\delta}}{\eta}\right)$ to get $O(\delta)$-approximation $\quad \delta = \tilde{O}\left(\frac{1}{\sqrt{n}}\right)$ $\square$

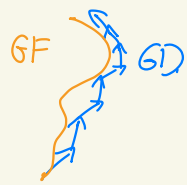
Initialization Effect

Recall: Certain initializations of GD converges to a minimum norm solution

↳ "constraint updating in some certain subspace"

Gradient Flow (GD with infinitestimal learning rate : continuous version of GD)

$L(W)$ loss function. GD: $W(t+1) = W(t) - \eta \cdot \nabla L(W(t))$



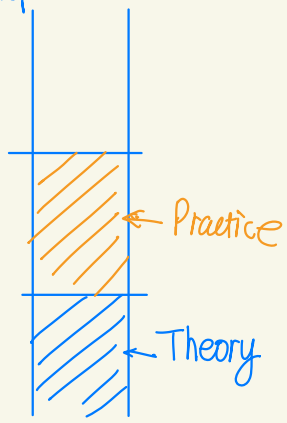
Scale the time by η: $W(t+\eta) = W(t) - \eta \cdot \nabla L(W(t))$

$\eta \to 0$: $W(t+dt) = W(t) - dt \cdot \nabla L(W(t))$.

$\Rightarrow$ DE: $\dot{W}(t) = -\nabla L(W(t))$. where $\dot{W}(t) = \frac{\partial W(t)}{\partial t}$.

allows us to ignore $\eta^2 (dt^2)$ term.

Step size:



Model: $f_W(x) = (W_+^{\odot 2} - W_-^{\odot 2})^T x = \theta_W^T x$.

$a^{\odot 2} = a \odot a$

$W_+, W_- \in \mathbb{R}^d$. $W = [W_+ \; W_-]$ is the parameter.

* $f_\beta(x) = (\beta \odot \beta)^T x$.

benefit: (i) $f_\beta(x)$ can only represent positive linear combinations of x.

$f_W(x)$ can represent any linear combinations of x.

(ii) initialize $W_+(0) = W_-(0)$ then $f_{W^0}(x) \equiv 0 \;\; \forall x$.

$\hat{L}(W) = \frac{1}{2} \sum_{i=1}^{n} (y^{(i)} - f_W(x^{(i)}))^2$.

for convenience.

initialization: $W_+(0) = W_-(0) = \alpha \cdot 1$. Let $\theta_W = W_+^{\odot 2} - W_-^{\odot 2}$. $\in \mathbb{R}^{n \times d}$. $(d \gg n)$

$W(\infty) = \lim_{t \to \infty} W(t)$. $\theta_\alpha(\infty) = \theta_{W(\infty)} \triangleq \theta_\alpha$. $X = \begin{bmatrix} - x^{(1)} - \\ \vdots \\ - x^{(n)} - \end{bmatrix}$. $y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}$

convergence solution

Thm For any $0 < \alpha < \infty$, Assume that we converge to a solution that fit the data.

$X \cdot \theta_\alpha = y$

can be proved

Then $\boxed{\theta_\alpha = \arg\min_\theta Q_\alpha(\theta)}$ such that $X\theta = y$

[Woodbury '20]
Rich & Kernel Regime in Overparameterized Models

complexity measure.

where $\boxed{Q_\alpha(\theta) = \alpha^2 \sum_{i=1}^{n} q\left(\frac{\theta_i}{\alpha^2}\right)}$

$q(\cdot): \mathbb{R} \to \mathbb{R}_{\geq 0}$. $\boxed{q(z) = 2 - \sqrt{4 + z^2} + z \cdot \operatorname{arcsinh}\left(\frac{z}{2}\right)}$

$\operatorname{arcsinh} x = \ln(x + \sqrt{x^2+1})$

"reverse engineering"

In particular: $\alpha \to \infty$. $q\left(\frac{\theta_i}{\alpha^2}\right) \asymp \frac{\theta_i^2}{\alpha^4}$.

$$Q_\alpha(\theta) \asymp \frac{1}{\alpha^2} \cdot \|\theta\|_2^2.$$

$\alpha \to \infty$: min ℓ_2-norm solution of θ.

(ℓ_4-norm solution of w)

$\alpha \to 0$ $\quad q\left(\frac{\theta_i}{\alpha^2}\right) \asymp \frac{|\theta_i|}{\alpha^2} \cdot \log \frac{1}{\alpha^2}$.

$$Q_\alpha(\theta) \asymp \|\theta\|_1 \cdot \log \frac{1}{\alpha^2}$$

$\alpha \to 0$: min ℓ_1-norm solution of θ

(ℓ_2-norm solution of w)

last video

$\alpha \in (0, \infty)$: intepolation between ℓ_1 and ℓ_2 regularization

---

More $\alpha \to \infty$: NTK regime

$\frac{\beta}{\sigma^2}$ β ← Lipschitzness of gradient, σ^2 ← condition number (both linearly depend on α)

$$W_-(0) = W_+(0) = \alpha \cdot 1$$

$$\nabla_w f_{w(0)}(x) = 2 \cdot \begin{bmatrix} W_+(0) \odot x \\ W_-(0) \odot x \end{bmatrix} = 2\alpha \begin{bmatrix} x \\ -x \end{bmatrix}$$

$\frac{\beta}{\sigma^2} \to 0$ as $\alpha \to \infty$

NTK $\langle \phi(x), \phi(x') \rangle$ → min ℓ_2-norm solution using feature $\begin{bmatrix} x \\ -x \end{bmatrix}$

Proof: (similiar to linear regression)

Step 1: find an invariance maintained by the optimizer (e.g. $\theta \in \text{span}\{x^{(i)}\}$)

Step 2: characterize the solution using the invariance.

$$\tilde{X} = [X \ -X] \in \mathbb{R}^{n \times 2d}. \quad W(t) = \begin{bmatrix} W_+(t) \\ W_-(t) \end{bmatrix} \in \mathbb{R}^{2d}.$$

$$\tilde{X} \cdot W(t)^{\odot 2} = [X \ -X] \cdot \begin{bmatrix} W_+(t)^{\odot 2} \\ W_-(t)^{\odot 2} \end{bmatrix} = \begin{bmatrix} (W_+(t)^{\odot 2} - W_-(t)^{\odot 2}) x^{(1)} \\ \vdots \\ (W_+(t)^{\odot 2} - W_-(t)^{\odot 2}) x^{(n)} \end{bmatrix} = \begin{bmatrix} f_{W(t)}(x^{(1)}) \\ \vdots \\ f_{W(t)}(x^{(n)}) \end{bmatrix} \in \mathbb{R}^n.$$

$$L(W(t)) = \|\tilde{X} \cdot W(t)^{\odot 2} - y\|_2^2.$$

$$\dot{W}(t) = -\nabla L(W(t)) = -\tilde{X}^T \left(\tilde{X} W(t)^{\odot 2} - y \right) \odot W(t)$$

$$= -\left(\tilde{X}^T \cdot \underbrace{r(t)}_{\text{residual term}} \right) \odot W(t) \quad (*)$$

as if it is a linear regression

Claim: $w(t) = w(0) \odot \exp\left(-2\tilde{X}^T \int_0^t r(s)\,ds\right)$

entrywise

a function of $w(t)$

$w_+(0) = \alpha \cdot 1$
$w_-(0) = \alpha \cdot 1$ $\Rightarrow w(0) = \alpha \cdot 1_{2d} \in \mathbb{R}^{2d}$.

$$\tilde{X}^T = \begin{bmatrix} X^T \\ -X^T \end{bmatrix} \in \mathbb{R}^{2d \times n}.$$

$$\exp(-2\tilde{X}^T v)^{\odot 2} = \begin{bmatrix} \exp(-2X^T v) \\ \exp(2X^T v) \end{bmatrix}^{\odot 2} = \begin{bmatrix} e^{-4X^T v} \\ e^{4X^T v} \end{bmatrix}$$

$\Downarrow$

$$\theta_\alpha(t) = w_+(t)^{\odot 2} - w_-(t)^{\odot 2}$$
$$= \alpha^2 \left(\exp\left(-4X^T \int_0^t r(s)\,ds\right) - \exp\left(4X^T \int_0^t r(s)\,ds\right) \right)$$
$$= 2\alpha^2 \cdot \sinh\left(-4X^T \int_0^t r(s)\,ds\right)$$

$$\dot{u}(t) = v(t) \cdot u(t) \quad \begin{cases} u: (w(t)); \\ v: (-\tilde{X}^T r(t)); \end{cases}$$

$$\Rightarrow \frac{\dot{u}(t)}{u(t)} = v(t)$$
$$\Rightarrow \frac{d}{dt} \log u(t) = v(t)$$
$$\Rightarrow \log u(t) - \log u(0) = \int_0^t v(s)\,ds$$
$$\Rightarrow u(t) = u(0) \cdot \exp \int_0^t v(s)\,ds$$

$$\sinh(x) = \frac{e^x - e^{-x}}{2}.$$

① $\theta_\alpha = \theta_\alpha(\infty) = 2\alpha^2 \cdot \sinh\left(-4X^T \int_0^\infty r(s)\,ds\right)$

② $X \cdot \theta_\alpha = y$ (assumption)

WTS: Any θ satisfying ①② is minimizer of $Q(\theta)$

It turned out to be the optimality condition of the program,

KKT

KKT: $\nabla Q(\theta) = X^T v$.

①: $\theta_\alpha = 2\alpha^2 \sinh\left(-4X^T \int_0^\infty r(s)\,ds\right)$

$$= 2\alpha^2 \sinh(-4X^T v')$$

$$\nabla Q(\theta) = \operatorname{arcsinh}\left(\frac{1}{2\alpha^2}\theta\right)$$

$\Downarrow$

$$\nabla Q(\theta_\alpha) = \operatorname{arcsinh}\left(\frac{1}{2\alpha^2} \cdot 2\alpha^2 \sinh(-4X^T v')\right)$$
$$= -4X^T v'$$
$$= X^T v \in \operatorname{span}\{x^{(1)}, \dots, x^{(n)}\}$$

$$\boxed{\begin{aligned} &\min_\theta Q(\theta) \\ &\text{s.t. } X\theta = y \end{aligned}}$$

optimality KKT $\begin{cases} \nabla Q(\theta) = X^T v & v \in \mathbb{R}^n \\ X\theta = y \end{cases}$

optimality: no 1st-order local improvement satisfying the constraint.

pertubation: $\Delta\theta \perp \operatorname{span}\{x^{(1)}, \dots, x^{(n)}\}$ $\quad X\Delta\theta = 0$

$\theta + \Delta\theta$ satisfies the constraint

$$Q(\theta + \Delta\theta) \geq \underset{\text{optimum}}{Q(\theta)} + \langle \Delta\theta, \nabla Q(\theta) \rangle \geq Q(\theta)$$

$\forall \Delta\theta \in \operatorname{span}\{x^{(1)}, \dots, x^{(n)}\}$

$\Downarrow$

$\nabla Q(\theta) \in \operatorname{span}\{x^{(1)}, \dots, x^{(n)}\}$ $\quad$ i.e. $\nabla Q(\theta) = X^T v$

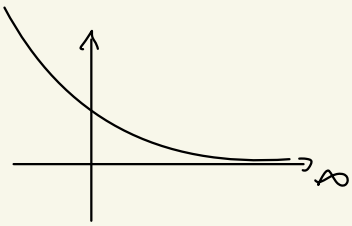
$\Rightarrow$ θ_α satisfies KKT condition of the program $\boxed{\begin{array}{l}\min\ Q(\theta) \\ \text{s.t. } X\theta = y\end{array}}$

$\Rightarrow$ global min !

Implicit Regularization of Classification Problem

Overparameterized Linear Models:

- regression problem: min-norm solution (certain initialization)
- classification problem: max-margin solution (any initialization)



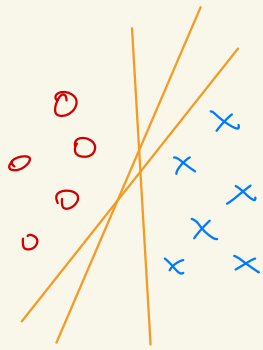
Set up: $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$, $x^{(i)} \in \mathbb{R}^d$, $y^{(i)} \in \{\pm 1\}$

linear model $h_w(x) = w^T x$

logistic loss function (cross-entropy): $\hat{L}(w) = \sum_{i=1}^{n} \ell(y^{(i)} \cdot h_w(x^{(i)})) = \sum_{i=1}^{n} \ell(y^{(i)} w^T x^{(i)})$

$$\ell(u) = \log(1 + \exp(-u))$$

- Multiple global minima if the data is separable.



$\exists$ infinite # of $\bar{w}$ s.t. $\bar{w}^T x^{(i)} \cdot y^{(i)} > 0$, $\forall i \in [n]$

where $\bar{w}$ is a unit vector.

$\hat{L}(\alpha \bar{w}) \to 0$ as $\alpha \to \infty$

"all of $\infty \cdot \bar{w}$ are global minima"

Scale

direction ✓

Question: Which direction will the model find via GD/GF ?

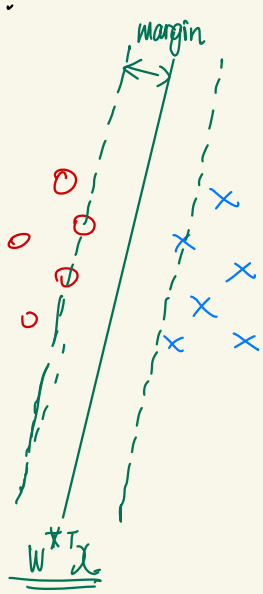
def margin: $\min_{i \in [n]} y^{(i)} w^T x^{(i)}$

normalized margin: $\gamma(w) = \dfrac{\min_{i \in [n]} y^{(i)} w^T x^{(i)}}{\|w\|_2}$

max-margin solution: $\gamma^* = \max_w \gamma(w) \triangleq \gamma(w^*)$

Implicit Bias !

maximizer (with unit norm)



Thm Gradient Flow converges to the direction of max-margin solution.

$$\boxed{\gamma(w(t)) \to \gamma^*} \text{ as } t \to \infty$$

Intuition: 1) ℓ is convex $\Rightarrow \hat{L}(w(t)) \to 0$ as $t \to \infty$

For large enough t, $\hat{L}(w(t)) \approx 0$,

2) $\ell(u) = \log(1+\exp(-u)) \approx \exp(-u)$ for large u

$\log(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \cdots$
$\geq x$

3) $\|w(t)\|_2 \to \infty$

(If $\|w(t)\|_2 \leq B$, then

$|y^{(i)} w(t)^T x^{(i)}| \leq B \cdot \|x^{(i)}\|_2$.

$\Rightarrow \hat{L}(w(t)) \geq \sum_{i=1}^n \exp(-|y^{(i)} w(t)^T x^{(i)}|)$

$\geq \sum_{i=1}^n \exp(-B \cdot \|x^{(i)}\|_2) > 0$ contradiction to 1))

"fixed"

4) Suppose $\|w\|_2 = q$ (very large)

$$\hat{L}(w) = \sum_{i=1}^n \ell(y^{(i)} w^T x^{(i)}) \approx \sum_{i=1}^n \exp(-y^{(i)} w^T x^{(i)})$$

logistic loss — exponential loss

$$\log \hat{L}(w) \approx \boxed{\log \sum_{i=1}^n \exp(-y^{(i)} w^T x^{(i)})} = \boxed{\log \sum_{i=1}^n \exp(-q \cdot y^{(i)} \bar{w}^T x^{(i)})}$$

☆ log-sum-exp

$$\approx \max_{i \in [n]} -q \cdot y^{(i)} \bar{w}^T x^{(i)}$$

softmax

$$\frac{e^{x_i}}{\sum_{i=1}^n e^{x_i}}$$

$$\log \sum_{i=1}^n \exp(q u_i) \geq \log \exp(q \cdot \max_i u_i) = q \cdot \max_i u_i$$

$$\log \sum_{i=1}^n \exp(q \cdot u_i) \leq \log(n \cdot \exp(q \cdot \max_i u_i)) = q \cdot \max_i u_i + \log n$$

$$= q \cdot \max_i u_i + o(q) \quad \text{as } q \to \infty$$

minimize the loss $\Leftrightarrow$ minimize $\max_{i \in [n]} -q \cdot y^{(i)} \bar{w}^T x^{(i)}$

$\Leftrightarrow$ maximize $\min_{i \in [n]} q\, y^{(i)} \bar{w}^T x^{(i)}$

margin

• $\ell(u) = \exp(-u)$

Proof: margin: $\gamma(w) = \dfrac{\min_{i\in[n]} y^{(i)} w^T x^{(i)}}{\|w\|_2}$

what GF finds

smooth-margin: $\tilde{\gamma}(w) \triangleq \boxed{\dfrac{-\log \hat{L}(w)}{\|w\|_2}} = \dfrac{-\log \sum_{i=1}^{n} \exp(-y^{(i)} w^T x^{(i)})}{\|w\|_2}$

$\forall i: y^{(i)} w^T x^{(i)} \geq \gamma(w) \cdot \|w\|_2$

$\tilde{\gamma}(w) = \gamma(w) - \dfrac{\log n}{\|w\|_2} \leq \gamma(w) \leq \gamma^*$ ①

max-margin

It's suffice to show: $\boxed{\lim_{t\to\infty} \tilde{\gamma}(w(t)) \geq \gamma^*}$

GF: $\dot{w}(t) = -\nabla \hat{L}(w(t))$

$$\frac{\partial}{\partial t}\left(-\log \hat{L}(w(t))\right) = -\frac{1}{\hat{L}(w(t))} \cdot \nabla_w \hat{L}(w(t)) \cdot \dot{w}(t) = +\frac{\|\dot{w}(t)\|_2^2}{\hat{L}(w(t))} \geq 0$$

Δ $\dfrac{\partial}{\partial t}(\log \text{loss}) \leq 0 \Rightarrow \dfrac{\partial}{\partial t}(\text{loss}) \leq 0$: loss decreasing!

$$-\log \hat{L}(w(T)) = -\log \hat{L}(w(0)) + \boxed{\int_0^T \frac{\|\dot{w}(t)\|_2^2}{\hat{L}(w(t))} dt.} \qquad (*)$$

Goal!

w^*: unit norm maximizer

$$\gamma^* = \max_w \gamma(w) = \max_w \frac{\min_{i\in[n]} y^{(i)} w^T x^{(i)}}{\|w\|_2} = \frac{\min_{i\in[n]} y^{(i)} w^{*T} x^{(i)}}{\|w^*\|_2} = \min_{i\in[n]} y^{(i)} w^{*T} x^{(i)}.$$

$$\boxed{\|\dot{w}(t)\|_2} \geq \langle \dot{w}(t), w^* \rangle \qquad \text{(Cauchy-Schwarz)}$$

$$= \langle -\nabla_w \hat{L}(w(t)), w^* \rangle$$

$$= \left\langle \sum_{i=1}^{n} y^{(i)} \exp(-y^{(i)} w^T x^{(i)}) \cdot x^{(i)}, w^* \right\rangle$$

$$= \sum_{i=1}^{n} y^{(i)} \exp(-y^{(i)} w^T x^{(i)}) \cdot \langle w^*, x^{(i)} \rangle$$

$y^{(i)} w^{*T} x^{(i)} \geq \min_{i\in[n]} y^{(i)} w^{*T} x^{(i)} = \gamma^*$

$$= \gamma^* \cdot \boxed{\sum_{i=1}^{n} \exp(-y^{(i)} w^T x^{(i)})} = \gamma^* \hat{L}(w(t))$$

$$\int_0^T \frac{\|\dot{w}(t)\|_2^2}{\hat{L}(w(t))} dt \geq \gamma^* \int_0^T \|\dot{w}(t)\|_2 \, dt \geq \gamma^* \cdot \left\| \int_0^T \dot{w}(t) \, dt \right\|_2 = \gamma^* \cdot \|w(T)\|_2.$$

Jenson.

$$\Rightarrow -\log \hat{L}(w(T)) \geq -\log \hat{L}(w(0)) + \gamma^* \cdot \|w(T)\|_2$$

$$\Rightarrow \frac{-\log \hat{L}(w(T))}{\|w(T)\|_2} \geq \frac{-\log \hat{L}(w(0))}{\|w(T)\|_2} + \gamma^*$$

$\to 0$

$w(T) \longrightarrow \infty$

$$\Rightarrow \lim_{T\to\infty} \tilde{\gamma}(w(T)) = \lim_{T\to\infty} \frac{-\log \hat{L}(w(T))}{\|w(T)\|_2} \geq \gamma^*. \quad ②$$

Combining ①-②: $\boxed{\lim_{T\to\infty} \tilde{\gamma}(w(T)) = \gamma^*}$ → GF converges to max-margin solution! □

Effect of Noise in SGD

TL;DR

1. one-dimensional quadratic objective: the noise makes the iterate bounce around the global minimum

2. multi-dimensional quadratic objective: similar to 1, while the fluctuation is closely related to the shape of the noise

Implicit Bias

3. non-quadratic objective: SGD with label noise on empirical loss $\hat{L}(\theta)$ has implicit regularization $tr(\nabla^2 \hat{L}(\theta))$, by third-order effect

loss function: $g(x)$.

SGD: $x_{t+1} = x_t - \eta \cdot (\nabla g(x_t) + \underbrace{\xi_t}_{\text{noise}})$. with $\mathbb{E}[\xi_t] = 0$.

---

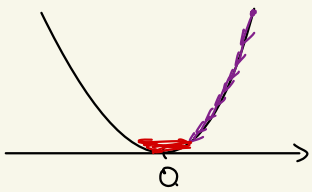
1. One-dimensional quadratic function

$g(x) = \frac{1}{2}x^2$. $\xi_t \sim N(0,1)$

SGD: $x_{t+1} = x_t - \eta \cdot (\nabla g(x_t) + \sigma \cdot \xi_t)$

$= x_t - \eta(x_t + \sigma \cdot \xi_t)$

$\Rightarrow x_{t+1} = \underbrace{(1-\eta)\cdot x_t}_{\text{contraction}} - \underbrace{\eta\sigma \cdot \xi_t}_{\text{stochastic}}$.



x_t large: contraction dominates

x_t small: noise dominates

→ eventually bouncing around global minimum $(x=0)$

$$x_{t+1} = (1-\eta)x_t - \eta\sigma \cdot \xi_t$$
$$= (1-\eta)\left((1-\eta)x_{t-1} - \eta\sigma\xi_{t-1}\right) - \eta\sigma\xi_t$$
$$= (1-\eta)^2 x_{t-1} - (1-\eta)\eta\sigma\xi_{t-1} - \eta\sigma\xi_t.$$
$$= \cdots$$
$$= \underbrace{(1-\eta)^{t+1}x_0}_{\text{contraction}} - \eta\sigma \underbrace{\sum_{k=0}^{t} (1-\eta)^k \xi_{t-k}}_{\text{accumulation of the noise}}$$

recent noise: more important

old noise: doesn't matter.

$$\text{Variance}[x_{t+1}] = \eta^2\sigma^2 \cdot \sum_{k=0}^{t} (1-\eta)^{2k}$$

As $t \to \infty$: $\text{Variance}[x_\infty] = \eta^2\sigma^2 \cdot \sum_{k=0}^{\infty}(1-\eta)^{2k} = \frac{\eta^2\sigma^2}{1-(1-\eta)^2} = \frac{\eta^2}{2\eta-\eta^2}\cdot\sigma^2 = \Theta(\eta\sigma^2)$

$\Rightarrow$ $x_\infty \sim \mathcal{N}(0, \Theta(\eta\sigma^2))$

$x_\infty \to 0$: $\begin{cases} \text{small learning rate } (\eta) \\ \text{small noise } (\sigma) \end{cases}$ $\Rightarrow$ small stochasticity

noise only makes it harder to converge to a global min.

Classical convex optimization:

noisy GD $\to$ less accurate solution
$\to$ faster to compute (less sample : mini-batch) } trade-off : gist.

Ornstein-Uhlenbeck (OU) Process

---

2. Multi-dimensional quadratic functions

$x \in \mathbb{R}^d$. PSD matrix $A \in \mathbb{R}^{d\times d}$. $g(x) = \frac{1}{2}x^T A x$. $\xi_t \sim \mathcal{N}(0, \Sigma)$

$$\begin{aligned} x_{t+1} &= x_t - \eta(\nabla g(x_t) + \xi_t) \\ &= x_t - \eta(Ax_t + \xi_t) \\ &= (I-\eta A)\cdot x_t - \eta\cdot\xi_t \qquad (*) \\ &= \cdots \\ &= \underbrace{(I-\eta A)^{t+1}\cdot x_0}_{\text{contraction}} - \underbrace{\eta\cdot\sum_{k=0}^{t}(I-\eta A)^k\cdot\xi_{t-k}}_{\text{accumulation of noise}} \end{aligned}$$

$\Rightarrow$ $x_\infty \sim \mathcal{N}\left(0, \eta^2\sum_{k=0}^{\infty}(I-\eta A)^k\Sigma(I-\eta A)^k\right)$

$$\begin{cases} (I-\eta A)^k = U\cdot\text{diag}((1-\eta d_i)^k)\,U^T \\ \Sigma = U\,\text{diag}(\sigma_i^2)\,U^T \end{cases}$$

$$= \eta^2 U\,\text{diag}\left(\sum_{k=0}^{\infty}\sigma_i^2\cdot(1-\eta d_i)^{2k}\right)U^T$$

$$\boxed{= \eta\cdot U\cdot\boxed{\text{diag}\left(\frac{\sigma_i^2}{d_i}\right)}\,U^T}$$

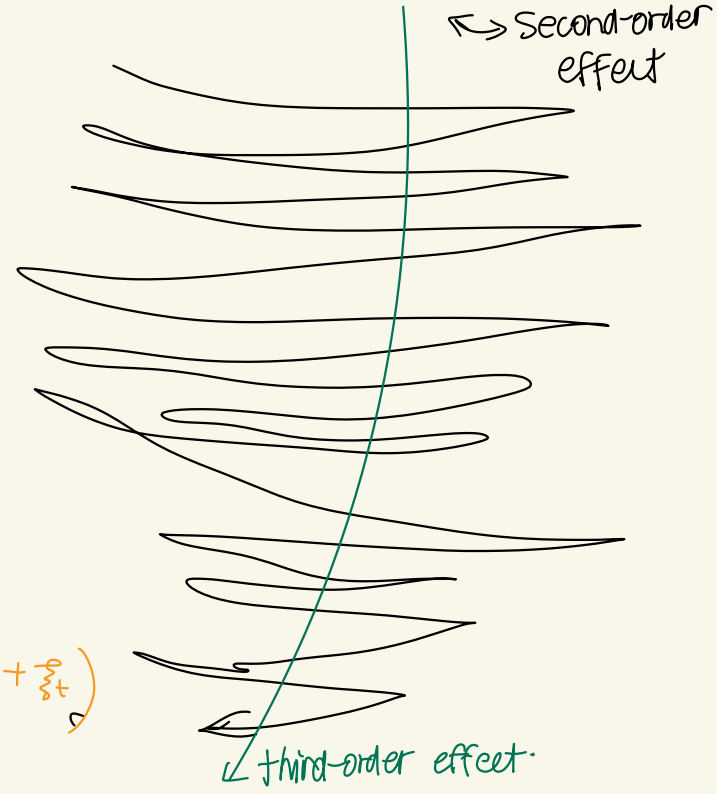
$\frac{\sigma_i^2}{d_i}$: iterate fluctuation level in the i-th eigenvector direction. $\to$ Shape of Σ and A.

Assume A, Σ simultaneously diagnonizable.

$$\begin{cases} A = U\,\text{diag}(d_i)\cdot U^T \\ \Sigma = U\cdot\text{diag}(\sigma_i^2)\,U^T \end{cases}$$

$\to$ viewed as d-separate OU processes in the eigen-coordinate system.

3. Non-quadratic loss functions



Wlog: a global minimizer of $g(x)$ is $x=0$

$\nabla g(0)=0$. $\nabla^2 g(0) \succeq 0$

Assume the iterate x_t are close to 0:

Taylor Expansion around 0:

$$x_{t+1} = x_t - \eta\cdot\left(\nabla g(x_t) + \xi_t\right) \qquad \xi_t \sim N(0, I_d)$$

$$g'(x_t) = g'(0) + g''(0)x_t + g'''(0)x_t^2 + \text{h.o.t.}$$

$$x_{t+1} = x_t - \eta\cdot\left(\underbrace{\nabla g(0)}_{=0} + \nabla^2 g(0)\cdot x_t + \nabla^3 g(0)[x_t, x_t] + \text{h.o.t.} + \xi_t\right)$$

$H = \nabla_x^2 g(0) \in \mathbb{R}^{d\times d}$

$T = \nabla_x^3 g(0) \in \mathbb{R}^{d\times d\times d}$ $\begin{cases} x\in\mathbb{R}^d,\ y\in\mathbb{R}^d,\ T[x,y]\in\mathbb{R}^d, & T[x,y]_i = \sum_{j,k\in[d]} T_{ijk}x_i y_k, \\ S\in\mathbb{R}^{d\times d},\ T[S]\in\mathbb{R}^d, & T[S]_i = \sum_{j,k\in[d]} T_{ijk}S_{jk} \end{cases}$

$$x_{t+1} = \underbrace{(I - \eta\cdot H)x_t - \eta\xi_t}_{\text{OU process. (similar as before)}} - \underbrace{\eta\cdot T[x_t, x_t]}_{\text{third-order term!}} + \text{h.o.term}$$

Δ

Assume we drop the third-order term: $x_{t+1} = (I - \eta H)\cdot x_t - \eta\xi_t$.

When convergence, $\|x_t\|_2 \approx \sqrt{\eta}$. $\Rightarrow \eta\cdot T[x_t, x_t] \approx \eta^2$. (3rd)

$\eta H x_t \approx \eta^{\frac{3}{2}}$ (2nd)

third-order term $\begin{cases} \ll \text{noise term } (\eta\xi_t) \\ \ll \text{second-order term} \end{cases}$

when $H \succ 0$, third-order term is negligible!

But typically, in overparameterization, H is low-dimensional.

In some directions where H is zero, $\eta H x_t = 0$ in that direction

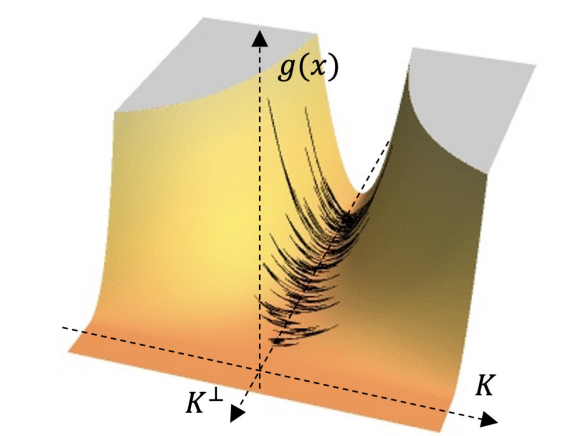
$\Rightarrow \eta^2$ becomes the largest update!

$\Rightarrow$ third-order term will introduce small bias!

(but very small: since noise dominates)

K: subspace spanned by those non-zero eigenvectors of H.

$$\begin{cases} 2^{\text{nd}} \text{ order term} \longrightarrow K \\ 3^{\text{rd}} \text{ order term} \longrightarrow K^{\perp} \text{ (no noise. no contraction)} \end{cases}$$



$$x_{t+1} = \underbrace{(1-\eta H)x_t - \eta \xi_t}_{\text{working in } K} \underbrace{- \eta \cdot T[x_t, x_t]}_{\text{working in } K^{\perp}} + \text{h.o.t.}$$

$$\downarrow \qquad\qquad \downarrow$$

$$u_t \ \checkmark \qquad\qquad r_t \ ?$$

$$r_{t+1} = x_{t+1} - u_{t+1} = (1-\eta H)(x_t - u_t) - \eta T[x_t, x_t] + \text{h.o.t.}$$

$$= \underbrace{(1-\eta H) r_t}_{\text{contraction}} - \underbrace{\eta T[x_t, x_t]}_{\text{bias} \in K^{\perp}} + \text{h.o.t.}$$

effect of 3rd order term

effect of 2nd order term and noise

$$\approx (1-\eta H) r_t - \eta T[u_t, u_t]$$

$(x_t \approx u_t)$

In the subspace of $K = \text{span}\{H\}$: contraction

In $K^{\perp}$: $\Pi_{K^{\perp}} r_{t+1} \approx \Pi_{K^{\perp}} r_t - \eta \cdot \Pi_{K^{\perp}} T[u_t, u_t] + \text{h.o.t.}$

$$= \cdots$$

$$= \Pi_{K^{\perp}} r_0 - \eta \cdot \sum_{k=0}^{t} \Pi_{K^{\perp}} T[u_k, u_k]$$

Assume $\mathbb{E}[u_k u_k^T] \to S$ as $k \to \infty$.

u_k mixes with a Markov Chain

$$\eta \cdot \sum_{k=0}^{t} T[u_k, u_k] \approx \eta \cdot t \cdot T[S].$$

Bias (informally): $-T[S]$ ☆

$$= -\nabla_x^3 g(0)[S]$$

$$= -\nabla_x \left(\langle \nabla^2 g(x), S \rangle \right)$$

$u \sim \mathcal{N}(0, S)$

$$\mathbb{E}[T[u,u]]_i$$

$$= \mathbb{E}\Big[\sum_{j,k \in [d]} T_{ijk} u_j u_k\Big]$$

$$= \sum_{j,k \in [d]} T_{ijk} \mathbb{E}[u_j u_k]$$

$$= \sum_{j,k \in [d]} T_{ijk} \left(\mathbb{E}[uu^T]\right)_{jk}$$

$$= \left(T[\mathbb{E}[uu^T]]\right)_i = \left(T[S]\right)_i$$

heuristically, $-T[S]$ is trying to make $R(x) = \langle \nabla^2 g(x), S \rangle$ small

With lots of assumption, SGD with noise converges to a stationary point of a regularized loss,

Implicit Regularization $\hat{L}(\theta) + \lambda \cdot R(\theta)$ with $R(\theta) \approx \text{tr}\left(\nabla^2 \hat{L}(\theta)\right)$ ☆☆