Lecture 12: Implicit Regularization & Noisy Dynamics

Recap from last lecture:
* Role of symmetries & geometric priors: while helpful to improve sample complexity, not sufficient in itself to break the CoD.
* Scale separation: → Breaks the curse, at the expense of reducing approximation properties.

↳ Inductive biases coming from architecture.

Today: Another form of "inductive bias", coming from optimization algorithm.

↳ Implicit Regularization of GD

↳ " " " (S)GD.

↳ Noisy.

Next week: last lecture (Diffusion)

in two weeks: (May 5th): project presentations.

Learning in the overparametrised Regime

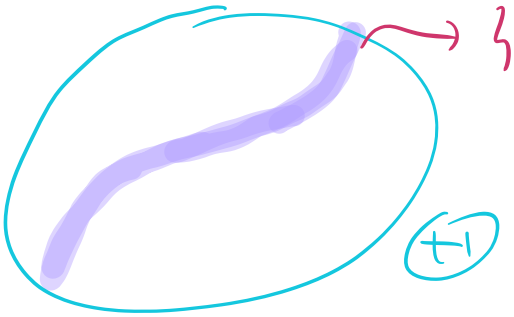
→ Let $\phi(\theta): X \to \mathbb{R}$ be a parametrized (+ differentiable) hypothesis class, $\theta \in \Theta \subseteq \mathbb{R}^p$.

→ Supervised learning: $L(\theta) = \hat{\mathbb{E}}_n [\ell(\phi(\theta)(x), y)] =$

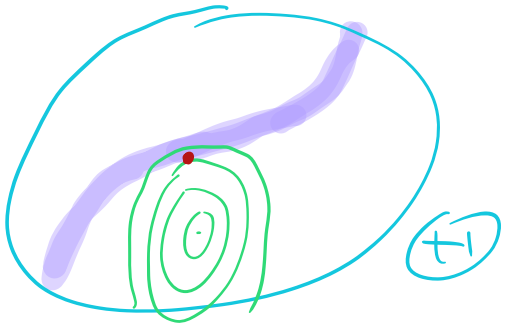
$$= \frac{1}{n} \sum_{i=1}^{n} \ell(\phi(\theta)(x_i), y_i)$$

→ Overparametrized regime: $p \gg n$. For models that are non-degenerate ($D\phi(\theta)(x_i)$ $i=1..n$ Jacobian matrix of size

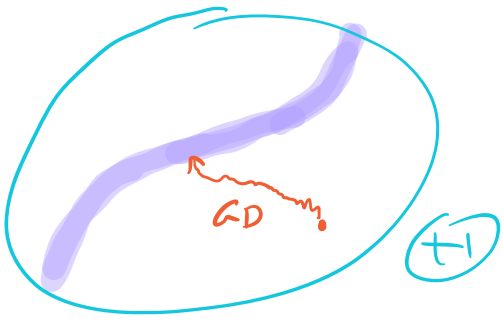
$p \times n$, has rank n generically), finding the parameters θ that fit the data becomes an underdetermined problem.



→ We saw in earlier lectures that adding regularisation enables non-vacuous generalisation bounds:



→ Yet, in many experimental setups, we seem to rely on another form of capacity control / regularisation, given by the algorithm itself.



Natural Questions:
- → Understand which solutions are picked by specific choices of algorithm.
- → Can we use this to establish generalisation bounds?

Part I: Implicit Bias of Gradient Descent

- (Simplest) setup: linear regression:

$$(*) \quad \min_{x \in \mathbb{R}^p} E(x) = \frac{1}{2}\|Ax - y\|^2, \qquad A \in \mathbb{R}^{n \times p} \text{ with } p > n.$$

- let's consider solving $(*)$ with GD:

$$x^{(k+1)} = x^{(k)} - \eta \cdot \nabla E(x^{(k)}) \quad ; \quad x^{(0)} = 0.$$

$$\nabla E(x) = A^T(Ax - y)$$

$$x^{(k+1)} = x^{(k)} - \eta \cdot A^T(Ax^{(k)} - y)$$

$$= [I - \eta A^T A] \cdot x^{(k)} + \eta A^T y.$$

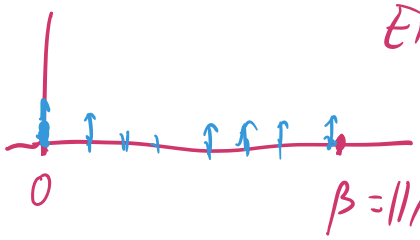
$$= [I - \eta A^T A]\left([I - \eta A^T A]x^{(k-1)} + \eta A^T y\right) + \eta A^T y$$

$$= \cdots$$

$$= \sum_{j=0}^{k} [I - \eta A^T A]^j \eta A^T y = \eta\left(\sum_{j=0}^{k} [I - \eta A^T A]^j\right) A^T y.$$

→ $A^T A$ is symmetric + psd.

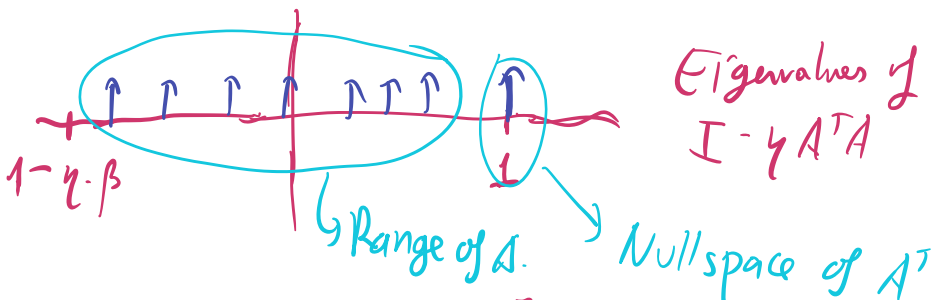
$\| \nabla^2 E$



→ If $1 - \eta\beta > -1$

$\Leftrightarrow$

$\eta < \frac{2}{\beta}$, then



$I - \eta A^T A$ has eigenvalues in $(-1, 1]$.

$\Rightarrow$ In the range of A, $\sum_{j=0}^{k} [I - \eta A^T A]^j \to (\eta A^T A)^{-1} = \eta^{-1}(A^TA)^{-1}$ as $k \to \infty$.

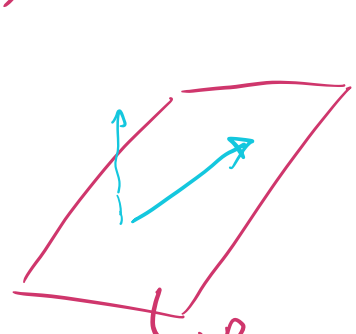
→ so $x^{(k)} \to (A^TA)^{-1}A^T$ ...

$x^{(k)} \xrightarrow[k\to\infty]{} (AA^T)^{-1} A y = A^\dagger y$

↳ pseudo-inverse of A.

→ variational characterisation: $A^\dagger y$ is the minimum L_2 norm solution of $\min_x \|Ax-y\|^2$; in other words,

$\min \|x\|^2$ st $Ax = y$.



↳ So, for this particular setting, GD finds the minimum L_2 solution that fits the data.

→ Note: GD in this overparametrised regime does not forget initial conditions ↳ GD initialised at $x^{(0)} = z$ converges to

$\min_x \|x - z\|^2$ st $Ax = y$.

Q: How general is this phenomena?

- → Different loss function
- → Different parametrisation Model.
- → Different flavor of GD?

→ Extension to logistic regression (binary classification) [Soudry et al '17].

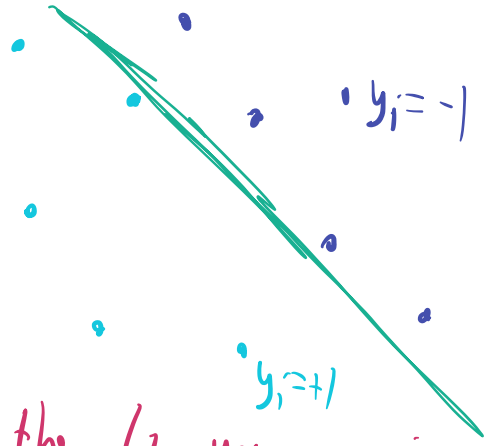
Using linear models. Consider a dataset $\{(x_i, y_i)\}_{i=1..n}$ ($x_i \in \mathbb{R}^d$) with $y_i = \pm 1$, and a loss function

$$\mathcal{L}(\theta) = \sum_{i=1}^n \ell(y_i \langle x_i, \theta \rangle) = \sum_{i=1}^n \ell(\langle \theta, x_i y_i \rangle)$$

$\ell(u) = \log(1 + e^{-u})$ logistic loss.

↳ Assume linearly separable data: $\exists\, \theta^* \in \mathbb{R}^d$ such that

$\langle \theta^*, x_i y_i \rangle > 0 \quad \forall i$



Theorem [Soudry et al.] GD on any initial condition satisfies

$$\lim_{k \to \infty} \frac{\theta^{(k)}}{\|\theta^{(k)}\|} = \frac{\hat{\theta}}{\|\hat{\theta}\|},$$ where $\hat{\theta}$ is the L_2 max-margin solution:

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \|\theta\|^2 \text{ s.t. } \langle \theta, y_i x_i \rangle \geq 1 \quad \forall i.$$

Idea of the proof: (·) Exploit the exponential tails of the logistic loss (+ the Laplace method)

(·) Implicit bias of direction $\frac{\theta^{(k)}}{\|\theta^{(k)}\|}$ rather than vector itself is inherent to logistic losses.

Q: Extension to non-linear models?

A: Yes! In [Chizat, Bach '20], extension to shallow NN.

· Margin of a predictor $f : \mathbb{R}^d \to \mathbb{R}$:

$$\min_{i \in \{1..n\}} y_i f(x_i) \quad \left(= \min_i \langle \theta, x_i y_i \rangle \text{ for linear } f(x) = \langle x, \theta \rangle \right)$$

· Max-Margin predictor in a function class $\mathcal{F}$:

$$\max_{\|f\|_{\mathcal{F}} \leq 1} \min_{i \in \{1..n\}} y_i f(x_i)$$

Theorem [Chizat, Bach]: Consider a two-layer ReLu network in the mean-field scaling regime: $f(x; \theta_1 .. \theta_m) = \frac{1}{m} \sum^m \phi(\theta_i, x)$

Then, under mild assumptions (non-quantitative on m!), gradient flow over logistic loss converges (in direction) to the F_1-max-margin solution, when F_1 is the variation-norm (a.k.a. Barron) space.

Other examples:

[Ji & Telgarsky '19]: Extend to more general losses, with rates.

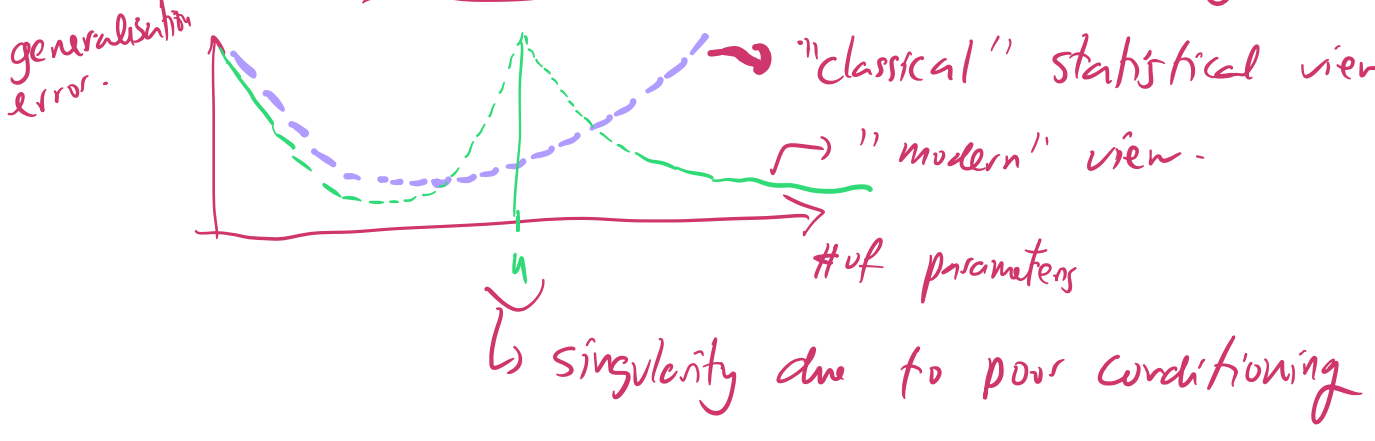
[Gunasekar et al.] [Li, Ma et al.] [Arora et al.] } Matrix / Tensor factorization
↳ GD finds low-rank solutions.

Takeaways so far:

↳ GD has some "built-in" capacity-control, sometimes encoded in its initialization + underlying choice of metric.

↳ While in some cases this implicit bias leads to strong statistical guarantees, it is not always clear what are the advantages over explicit regularization.

↳ Example: Double-descent [Belkin, Hsu, et al. + many others..]



→ Analysis is tedious: it needs to be done in a case-by-case

basis.; currently we don't have a good description for implicit bias on complex NN architectures.

→ Negative results: for certain non-linear models, implicit bias of GD provably cannot be characterised by any norm. [Vardi, Shamir] [Ratzin & Cohen]

→ We know that vanilla GD is sometimes brittle (trapped in bad local minima).

↳ Q: other forms of algorithmic regularization?

Part II: Noisy Gradient Dynamics

→ Vanilla GD step $\theta^{(k+1)} = \theta^{(k)} - \eta \underbrace{\nabla L(\theta^{(k)})}_{\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ell(\phi(\theta^{(k)})(x_i), y_i)}$

↳ Each update requires all training samples!

→ Since $\nabla L(\theta^{(k)}) = \hat{\mathbb{E}}_{\nu}[\nabla \ell(\phi(\theta^{(k)})(x), y)]$, a natural (stochastic) approximation is to do a MC estimate:

$$\widehat{\nabla L}(\theta^{(k)}) = \nabla \ell(\phi(\theta^{(k)})(\tilde{x}), \tilde{y}) \quad \text{with } (\tilde{x}, \tilde{y}) \sim \nu$$

↳ Stochastic Approximation / Optimization [Robbins & Munro, 51]:

$$\theta^{(k+1)} = \theta^{(k)} - \eta_k \cdot \underbrace{g(\theta^{(k)}, \xi_k)}_{\text{↳ stochastic gradient, } \mathbb{E}\, g(\theta^{(k)}, \xi) = \nabla L}$$

with

Folklore: Under mild assumptions on the variance $\mathbb{E}\|g(\theta^{(k)}, \xi) - \nabla L\|^2$ and for appropriate learning rate

schedules $(\eta_k)_k$, then SGD "recovers" same convergence guarantees as GD (global convergence for convex losses, local convergence for non-convex losses).

Q: What about implicit bias, in overparametrised settings? Is it the same as GD?

↳ This SGD is viewed as adding noise to the gradients. Next week: we are going to study precisely the convergence of a (simplified) version of this algorithm.

TLDR: (next week): solutions of SGD "concentrate" around local minimizers of the loss.

Q: Other forms of noisy gradient descent?

Label Noise as Implicit Regularization

Idea: Rather than adding noise into the gradients, we instead add noise to the targets:

→ For $k=0,1,\dots$ do:

(i) sample a datapoint $x^{(k)}$ uniformly from $\{(x_i, y_i)\}_{i=1\dots n}$

(ii) sample $\varepsilon \sim \{-\sigma, \sigma\}$.

(iii) $\theta^{(k+1)} = \theta^{(k)} - \eta \nabla_\theta \left[\left(f(\theta^{(k)})(x^{(k)}) - y^{(k)} - \varepsilon \right)^2 \right]$

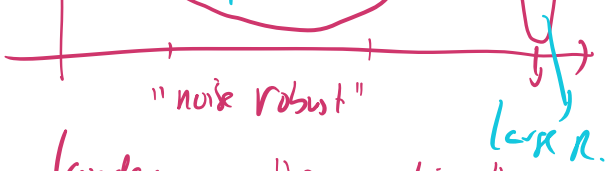
noisy label.

→ Consider the term

$$R(\theta) = \frac{-1}{2\eta} \operatorname{tr} \log\left(1 - \frac{\eta}{2} \nabla^2 L(\theta)\right)$$

(notice for $\eta \to 0$, $R(\theta) \approx \frac{1}{\eta} \operatorname{tr} \nabla^2 L(\theta)$)

↳ measure of "total" curvature of the loss.



landscape "perspective"

Theorem [Damian, Lee & Ma, follow up from Blanc et al]. Under smoothness and Polyak-Łojasiewicz sharpness assumptions, then GD with label noise finds an approximate stationary point of $\tilde{L}(\theta) = L(\theta) + \lambda \cdot R(\theta)$ (λ is explicit)

↳ In other words, label noise favors "flat" solutions over non-flat ones.

Summarise: Algorithmic regularization: powerful method to

⊕ obtain capacity-control and achieve statistical learning guarantees.

☹ sometimes it is the "wrong" way to achieve a desirable effect (least-squares), whereas sometimes leads to efficient algorithms (logistic regression in $\mathbb{F}_1$, label noise).