Lecture 21: minimum norm solutions

Plan for rest of semester.

21 GD prefers low/minimum norm solutions
22 project
23 concentration
24 deep
25 network
26 generalization
27 double descent
28 ????? ~~kolmogorov~~-Arnold

Goal of course: we can get low test error.

$$R(\hat{f}) - R(g) = \underbrace{R(\hat{f}) - \hat{R}(\hat{f})}_{\text{generalization}} + \underbrace{\hat{R}(\hat{f}) - \hat{R}(f)}_{\text{opt}} + \underbrace{\hat{R}(f) - R(f)}_{\text{generalization}} + \underbrace{R(f) - R(g)}_{\text{apx}}$$

$\hat{f} \in \mathcal{F}$ alg

$f \in \mathcal{F}$ ref inside

g ref outside

"rethinking generalization: ..." Zhang et al.

" ... " Tomioka - Neyshabur - Srebro

One solution to issue:

* $\mathcal{F}$ is adapted to data.

(our approach: $\mathcal{F}$ is norm bounded, & the norm we need depends on data).

In optimisation:

① Near-initialization ("mostly convex" / "neural tangent kernel") 2009

Adaptivity to low norm: $\frac{1}{2t}\|w_t - z\|^2 + \hat{R}(w_t) \le \frac{1}{2}\|w_0 - z\|^2 + \hat{R}(z)$

② Far from initialization

* norm preservation
* next two lectures

Low-norm predictors

Seen it before: $\frac{1}{2}\|w_t - z\|^2 + \hat{R}(w_t) \le \frac{1}{2}\|w_0 - z\|^2 + \hat{R}(z)$ (convex + smooth, add slop for ReLU version near init)

Two topics we've missed:

- convergence to the minimum norm solution
- Far from initialization (& is that practical)

Recall appearance in linear regression:

$$\lim_{\lambda \downarrow 0} \operatorname*{argmin}_w \frac{1}{2}\|Xw - y\|^2 + \frac{\lambda}{2}\|w\|^2 = \lim_{\lambda \downarrow 0} (X^T X + \lambda I)^{-1} X^T y$$
$$= X^+ y$$
} ordinary least squares

Also GD on $\frac{1}{2}\|Xw - y\|_2^2 \xrightarrow[\#\text{iterations} \to \infty]{\text{step size to } 0} X^+ y$ $\qquad X^T(Xw - y)$

Good news: GD gets min norm solution.

Question: (a) how about classification losses, (b) how about nonlinear

Remark (how to prove min norm). Consider opt cond of
$\min \frac{1}{2}\|w\|^2$
s.t. $X^T X w = X^T y$

Realizability / separability assumptions

Suppose that our model perfectly labels the data

[Remark: why? without this assumption, norms seem to blow up.]

linear Regression version:

$\min \frac{1}{2}\|w\|_2^2$

s.t. $Xw = y$

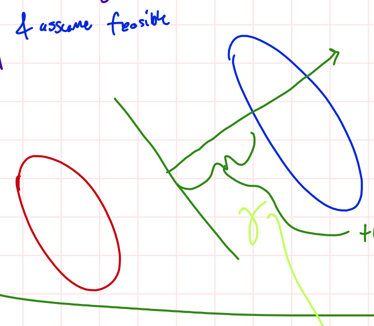
& assume feasible

classification version

$\min \frac{1}{2}\|w\|_2^2$

s.t. $y \odot (Xw) \ge 1$ $\quad$ [i.e., $\forall i,\ y x_i^T w \ge 1$]

& assume feasible



this "margin" is maximized

nonlinear version:
$\min_{w \in \mathbb{R}^p} \frac{1}{2}\|w\|^2$
s.t. $y_i f(x_i; w) \ge 1\ \forall i$.

We'll consider two cases: ① Linear case + exponential loss (similar to logistic / cross-entropy)

② L-homogeneous.

Linear case

Theorem (Δ). $\boxed{\text{Consider GF on } \mathcal{L}(w) = \sum_i \exp(-y_i x_i^T w) \quad [w_0 = 0,\ \dot{w}_t = -\nabla \mathcal{L}(w_t)].}$

& suppose $\exists u$, $\|u\|_2 = 1$, $\exists \gamma > 0$, s.t. $\forall i$ $y_i x_i^T u \ge \gamma$ (separability assumption).

$\boxed{\text{need } \|x\| \le 1}$

Then $\dfrac{\min_i y_i x_i^T w_t}{\|w_t\|} = \min_i y_i x_i^T \left(\dfrac{w_t}{\|w_t\|}\right) \ge \gamma - O\left(\dfrac{\ln n}{\ln t}\right)$.

Remark. $\frac{1}{\ln t}$ is slow; can be "accelerated", but in practice slow $1/\ln t$ seems real.

Lemma. Consider setting of (Δ). Then
$$\mathcal{L}(w_t) \le \frac{n}{t} + \frac{(\ln t)^2}{2t\gamma^2}, \qquad \|w_t\| \ge \ln(t\gamma^2) - \ln\left(1 + \frac{\ln t}{n}\right).$$

Remark. Note solutions unbounded & $\|w_t\| \to \infty$

Proof. Recall $\forall z \in \mathbb{R}^d$ $\frac{1}{2t}\|w_t - z\|^2 + \mathcal{L}(w_t) \le \frac{1}{2t}\|z\|^2 + \mathcal{L}(z)$.

Now clairvoyantly guess $z = u \frac{\ln t}{\gamma}$ $\qquad y_i x_i^T u \frac{\ln t}{\gamma} \ge \gamma \cdot \frac{\ln t}{\gamma} = \ln t$

$$\Rightarrow \mathcal{L}(w_t) \le \frac{1}{2t}\frac{(\ln t)^2}{\gamma^2} + \sum \exp(-y_i x_i^T z) \le \frac{(\ln t)^2}{2t\gamma^2} + \frac{n}{t}.$$

(Remark: if we included dropped $\frac{1}{2}\|w_t - z\|^2$, then can show $\|w_t\| = O\left(\frac{\ln t}{\gamma}\right)$.)

for norm lower bound:
$$\|w_t\| = -\ln \exp(-\|w_t\|)$$
$$\ge -\ln \sum \exp(-\|w_t\|) \overset{\text{uses } \|x_i\| \le 1}{\ge} -\ln \underbrace{\sum_i \exp(-y_i x_i^T w_t)}_{\mathcal{L}(w_t)}$$
$$\ge -\ln\left[\frac{(\ln t)^2}{2t\gamma^2} + \frac{n}{t}\right].$$

//

Proof of (Δ).

$$\min_i y_i x_i^T w_t = -\ln \min_i \exp\ y_i x_i^T w_t \ge -\ln \sum_i \exp(-y_i x_i^T w_t)$$
$$= \underbrace{-\ln \sum \exp(-y_i x_i^T w_t) + \ln \sum \exp(-y_i x_i^T w_0)}\ \underbrace{-\ln \sum \exp(-y_i x_i^T w_0)}_{\ln n}$$

FTC
$$= \int_0^t \frac{d}{ds} - \ln \sum \exp(-y_i x_i^T w_s)\, ds = \int_0^t \left\langle \frac{-\nabla \mathcal{L}(w_s)}{\sum \exp(-y_i x_i^T w_s)}, \dot{w}_s \right\rangle ds$$
$$= \int_0^t \frac{\|\nabla \mathcal{L}(w_s)\|^2}{\sum \exp(-y_i x_i^T w_s)}\, ds \ge \int_0^t \|\nabla \mathcal{L}(w_s)\| \cdot \frac{\sum_i \exp(-y_i x_i^T w_s) \overbrace{\langle y_i x_i, u \rangle}^{\ge \gamma}}{\sum \exp(-y_i x_i^T w_s)}\, ds$$
$$\ge \gamma \int_0^t \|\nabla \mathcal{L}(w_s)\|\, ds \ge \gamma \left\| \int_0^t \nabla \mathcal{L}(w_s)\, ds \right\| = \gamma \|w_t - w_0\|$$
$$= \gamma \|w_t\|$$

$$\Rightarrow \frac{\min_i y_i x_i^T w_t}{\|w_t\|} \ge \gamma \left(\frac{\cancel{\|w_t\|}}{\cancel{\|w_t\|}}\right) - O\left(\frac{\ln n}{\ln t}\right)$$
[lemma.

//