Implicit regularization

We saw last class a theorem (by Lyu-Li '19, Ji-Telgarsky '20) which shows implicit bias of G.D. towards $\ell^2$-margin maximization.
We will now prove a number of implicit bias results.

We'll first consider regression.

Let $X \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$, assume $n<d$ & $X$ full rank (over-parameterized/high-dimensional!)

Consider $\hat{L}(\beta) := \frac{1}{2}\|X\beta - y\|_2^2 = \frac{1}{2}\sum_1^n (\langle x_i, \beta\rangle - y_i)^2$.

Lemma Let $X^+$ denote pseudo-inverse of $X$. Then $\beta$ is a global min of $\hat{L} \iff \beta = X^+ y + \xi$ for some $\xi \perp \operatorname{span}\{x_1, \dots, x_n\}$.

Pf. Any $\beta \in \mathbb{R}^d$ can be represented as

$$\beta = X^+ y + \xi \text{ for some } \xi \in \mathbb{R}^d \quad (\xi = \beta - X^+ y).$$

Since

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

$$= X \cdot X^T (XX^T)^{-1} y + X\xi$$

$$= y + X\xi.$$

Since $X$ is full rank, $X^+ = X^T(XX^T)^{-1}$

Now, $\beta$ is a global min of $\hat{L} \iff \|X\beta - y\|_2^2 = 0$.

$$\iff \|y + X\xi - y\|_2^2 = \|X\xi\|^2 = 0.$$

$$\|X\xi\|^2 = 0 \iff \sum_{i=1}^n \langle x_i, \xi \rangle^2 = 0 \iff \langle x_i, \xi \rangle = 0 \ \forall i = 1, \dots, n.$$

$$\iff \xi \perp \operatorname{span}\{x_1, \dots, x_n\}.$$

$\boxed{\xi}$

So every global min of squared loss lies in subspace spanned by data, and in that subspace is given by $X^+ \beta$.

<u>Lemma</u> Let $\beta^* := \arg\min\{\|\beta\|_2^2 : \hat{L}(\beta) = 0\}$.

Then $\beta^* = X^+ y$.

<u>pf</u>. Let $\beta$ be st $\hat{L}(\beta) = 0$. By prev lemma, $\exists \xi \perp \{x_1, \dots, x_n\}$ s.t. $\beta = X^+ y + \xi$. Thus, $\|X\xi\|_2 = 0$, and we have

$$\|\beta\|_2^2 = \|X^+y\|^2 + \|\xi\|^2 + 2\langle X^+y, \xi\rangle$$
$$= \|X^+y\|^2 + \|\xi\|^2 + 2\langle X^T(XX^T)^{-1}y, \xi\rangle$$
$$= \|X^+y\|^2 + \|\xi\|^2 + 2\langle (XX^T)^{-1}y, X\xi\rangle$$
$$= \|X^+y\|^2 + \|\xi\|^2 \geq \|X^+y\|^2.$$

We get equality $\iff \xi = 0$. $\boxed{\xi}$

Now consider iterates of G.D. started from $\beta_0$:

$\beta_{t+1} = \beta_t - \eta \nabla \hat{L}(\beta_t)$. $\hat{L}$ is convex and 1-smooth, so we know $\beta_t \to$ global min of $\hat{L}$. We now show it converges to min. $l_2$ norm solution, if $\beta_0 = 0$.

<u>Theorem</u> Let $\beta_0 = 0$, $\{\beta_t\}$ G.D. iterates. Suppose $\beta_t \to \hat{\beta}$ with $\hat{L}(\hat{\beta}) = 0$. Then $\hat{\beta} = \beta^* = \arg\min\{\|\beta\|_2^2 : \hat{L}(\beta) = 0\}$.

<u>pf</u> First, note that $\beta_t \in \operatorname{span}\{x_1, \dots, x_n\}$ $\forall t$: clear @ $t=0$.

$-\nabla \hat{L}(\beta) = \frac{1}{n}\sum_i^n (y_i - \langle w, x_i\rangle)\cdot x_i \in \operatorname{span}\{x_1, \dots, x_n\}$, So an induction arg clearly shows $\beta_t \in \operatorname{span}\{x_1, \dots, x_n\}$ $\forall t$.

In particular, $\beta_t \to \hat{\beta} \in \operatorname{span}\{x_1, \dots, x_n\}$.

So $\hat{\beta} = X^T v$ for some $v \in \mathbb{R}^n$.

Since by assumption $\hat{L}(\hat{\beta}) = 0$, $0 = \|X\hat{\beta} - y\|_2^2$

implies $0 = X\hat{\beta} - y = XX^T v - y$, so

$$v = (XX^T)^{-1} y \implies \hat{\beta} = X^T v = X^T (XX^T)^{-1} y = X^+ y = \beta^*.$$

So, G.D. on squared loss has <u>implicit bias</u> towards minimum $\ell^2$-norm solution (an <u>implicit regularization</u> effect).

We'll now look at classification setting. Ref: [T], ch. 10.

<u>Def</u> Data $\{x_i, y_i\}$, $x_i \in \mathbb{R}^d$, $y_i \in \{\pm 1\}$, is <u>linearly separable</u> if $\exists w \in \mathbb{R}^d$ st $\min_i y_i \langle w, x_i \rangle > 0$. For linearly sep. data, The <u>$\ell_2$-max margin</u> predictor is $\bar{u} := \arg\min\{\|w\|_2^2 : y_i \langle w, x_i \rangle \geq 1 \ \forall i\}$.

Equivalently, $\bar{u} := \arg\max \{ \min_i y_i \langle w, x_i \rangle \; : \|w\|_2 = 1 \}$.

<u>Exercise</u> Show that if the $\ell_2$-max-margin predictor exists, it is unique.

$\hat{L}(\theta) = \sum_1^n$ not $\frac{1}{n}$.

<u>Prop</u> Suppose $f(x;\theta)$ is $L$-homog. in $\theta$, $\ell$ is exp-loss, and $\exists\, \hat{\theta}$ st $\hat{L}(\hat{\theta}) = \sum_1^n \ell(y_i f(x_i;\hat{\theta})) < \frac{\ell(0)}{n}$.
Then $\inf_\theta \hat{L}(\theta) = 0$, and the inf is <u>not</u> attained.

<u>Pf</u>. Let $m_i(\theta) := y_i f(x_i;\theta)$ (<u>margin</u> of ex. $i$).
Since $\ell$ is decreasing

$y = \exp(-x)$
$\log y = -x \to x = -\log y$

$$\ell(\min_i m_i(\theta)) = \max_i \ell(m_i(\theta)) \leq \sum_{i=1}^n \ell(m_i(\theta)) = \hat{L}(\theta) < \ell(0)/n \leq \ell(0)$$

Since $\ell^{-1}(t) = -\log t$ is strictly decreasing, $\ell^{-1}(\ell(\min_i m_i(\theta))) = \min_i m_i(\theta) > \underbrace{\ell^{-1}(\ell(0))}_{=0}$.

Thus: $0 \leq \inf_\theta \hat{L}(\theta) \leq \limsup_{c\to\infty} \hat{L}(c\hat{\theta}) = \limsup_{c\to\infty} \frac{1}{n}\sum_1^n \ell(c\, m_i(\theta))$

Since $m_i(\theta) > 0$, $\ell$ decreasing, $c>0$,

$$\ell(c \cdot m_i(\theta)) \to 0 \text{ as } c \to \infty.$$

$$0 \le \inf_\theta \hat{L}(\theta) \le \limsup_{c\to\infty} \frac{1}{n}\sum_i^n \ell(c \cdot m_i(\theta)) \le \frac{1}{n}\sum_i^n \limsup_{c\to\infty} \ell(c \cdot m_i(\theta))$$

$$= 0.$$

Thus $\inf_\theta \hat{L}(\theta) = 0$. Since $\ell(t) > 0 \ \forall t$, impossible to have $\hat{L}(\theta) = 0$.

$\boxed{\Sigma}$

So we cannot "find" an "optimum": solutions are off at $\infty$.

To compare predictors; first note

$$\min_i m_i(\theta) = \|\theta\|_2^L \cdot \min_i m_i\left(\frac{\theta}{\|\theta\|_2}\right) \text{ by homogeneity.}$$

$\rightarrow$ we can compare by normalized margin $\theta / \|\theta\|_2$.

Moreover, for exp loss, we have: $\ell^{-1}(b) = -\log t$ (decreasing)

$$\frac{\ell^{-1}(\hat{L}(\theta))}{\|\theta\|_2^L} = \frac{\ell^{-1}\left(\sum_i^n \ell(m_i(\theta))\right)}{\|\theta\|_2^L} \le \frac{\ell^{-1}\left(\max_i \ell(m_i(\theta))\right)}{\|\theta\|_2^L} = \frac{\min_i m_i(\theta)}{\|\theta\|_2^L}$$

$$\frac{\ell^{-1}(\hat{L}(\theta))}{\|\theta\|_2^L} + \frac{\log(n)}{\|\theta\|_2^L} = \frac{\ell^{-1}\left(\frac{1}{n}\sum_i^n \ell(m_i(\theta))\right)}{\|\theta\|^L}$$

$$\geq \frac{\ell^{-1}(\max_i \ell(m_i(\theta)))}{\|\theta\|^L} \quad \text{since } \ell^{-1} \text{ is decreasing \& avg} \leq \max$$

$$= \frac{\min_i m_i(\theta)}{\|\theta\|^L} \quad \text{since } \ell \text{ is decreasing}$$

$$\geq \frac{\ell^{-1}\left(\sum_{n=1}^n \ell(m_i(\theta))\right)}{\|\theta\|^L} . \quad \text{since } \ell^{-1} \text{ is decr. \& } \max_i \ell(m_i(\theta)) \leq \sum_i \ell(m_i(\theta))$$

$$\Rightarrow \quad \frac{\min_i m_i(\theta)}{\|\theta\|_2^L} \in \left[ \frac{\ell^{-1}(\hat{L}(\theta))}{\|\theta\|_2^L}, \frac{\ell^{-1}(\hat{L}(\theta))}{\|\theta\|_2^L} + \frac{\log n}{\|\theta\|_2^L} \right].$$

We thus can approximate the normalized margin $\frac{\min_i m_i(\theta)}{\|\theta\|^L}$ by $\ell^{-1}(\hat{L}(\theta)) / \|\theta\|_2^L$.

**Def.** Call data m-separable if $\exists \theta$ st $\min_i m_i(\theta) > 0$.

$$\underbrace{\gamma(\theta) := \min_i m_i(\theta/\|\theta\|_2)}_{\text{(normalized) margin}}, \quad \underbrace{\bar{\gamma} := \max_{\|\theta\|_2=1} \gamma(\theta)}_{\text{max margin}} ; \quad \underbrace{\tilde{\gamma}(\theta) := \frac{\ell^{-1}(\hat{L}(\theta))}{\|\theta\|_2^L}}_{\text{smooth (normalized) margin.}}$$

Prop Suppose data is m-separable. Then:

① $\bar{\gamma} := \max_{\|\theta\|_2=1} \gamma(\theta) > 0$ is well-defined

② For any $\theta \neq 0$, $\lim_{c\to\infty} \tilde{\gamma}(c\theta) = \gamma(\theta)$.

Pf ① We want to show that the maximum is attained.

Note that by assumption, $\exists \hat{\theta}$ st $\gamma(\hat{\theta}) > 0$. Since $m_i(\theta)$ is homogeneous, we know that $m_i(\alpha\theta) = \alpha^L m_i(\theta)$ for $\alpha > 0$. Thus $m_i(0) = 0$, hence $\hat{\theta} \neq 0$, so consider $\theta := \hat{\theta}/\|\hat{\theta}\|_2$. Then $\|\theta\|_2 = 1$. Thus,

$$\gamma(\theta) = \min_i m_i(\theta/\|\theta\|_2) = \|\theta\|_2^{-L} \min_i m_i(\theta) > 0.$$

Since $m_i(\theta)$ is continuous, & ptwise min of cts fcns in cts, $\gamma(\theta)$ is continuous in $\theta$, and is strictly positive on at least one point on the domain $\{\|\theta\|_2 = 1\} =: D$

Thus its maximum must be strictly positive, and attained on $D$ due to compactness.

② Recall that $\tilde{\gamma}(\theta) := \frac{\ell^{-1}(\hat{L}(\theta))}{\|\theta\|_2^L} \leq \gamma(\theta) = \frac{\min_i m_i(\theta)}{\|\theta\|_2^L} \leq \frac{\ell^{-1}(\hat{L}(\theta))}{\|\theta\|_2^L} + \frac{\log n}{\|\theta\|_2^L} = \tilde{\gamma}(\theta) + \frac{\log n}{\|\theta\|_2^L}$.

$$\Rightarrow \tilde{\gamma}(c\theta) \leq \gamma(c\theta) = \gamma(\theta) \leq \tilde{\gamma}(c\theta) + \frac{\log n}{c^L \|\theta\|_2^L}.$$

$\Rightarrow \limsup_{c\to\infty} \tilde{\gamma}(c\theta) \leq \gamma(\theta)$, and $\liminf_{c\to\infty} \left\{\tilde{\gamma}(c\theta) + \frac{\log n}{c^L\|\theta\|_2^L}\right\} = \liminf_{c\to\infty} \tilde{\gamma}(c\theta) \geq \gamma(\theta)$.

$\blacksquare$

## Gradient flow maximizes the margin of linear predictors.

Let $\hat{L}(\theta) = \sum_{i=1}^{n} \ell(y_i \langle \theta, x_i \rangle)$. Gradient flow:

$$\frac{d\theta}{dt} = -\nabla \hat{L}(\theta(t)), \quad \text{with (assume) } \theta(0) = 0.$$

First, note that G.F. is always decreasing (even nonconvex):

$$\hat{L}(\theta(t)) - \hat{L}(\theta(0)) = \int_0^t \langle \nabla \hat{L}(\theta(s)), \frac{d}{ds} w(s) \rangle \, ds$$

$$= -\int_0^t \|\nabla \hat{L}(\theta(s))\|^2 \, ds \le 0.$$

Thus, $\min_{s \in [0,t]} \hat{L}(\theta(s)) = \hat{L}(\theta(t))$ for $t > 0$.

**Theorem** For any $z \in \mathbb{R}^d$, $t \ge 0$, GF satisfies

$$t \hat{L}(\theta(t)) + \frac{1}{2}\|\theta(t) - z\|_2^2 \le t \hat{L}(z) + \frac{1}{2}\|\theta(0) - z\|^2.$$

Pf. for any $z$,

$$\frac{1}{2}\|\theta(t) - z\|^2 - \frac{1}{2}\|\theta(0) - z\|^2 = \frac{1}{2}\int_0^t \frac{d}{ds}\|\theta(s) - z\|^2 \, ds$$

$$= \int_0^t \langle \frac{d\theta}{ds}, \theta(s) - z \rangle \, ds$$

$$= \int_0^t \langle \nabla \hat{L}(\theta(s)), \theta(s) - z \rangle \, ds$$

$$\leq \int_0^t \left(\hat{L}(z) - \hat{L}(\theta(s))\right) ds$$
$$= t\,\hat{L}(z) - \int_0^t \hat{L}(\theta(s))\, ds$$

Now, $\displaystyle\int_0^t \hat{L}(\theta(s))\,ds \geq \int_0^t \min_{s\in[0,t]} \hat{L}(\theta(s))\,ds$
$$= t \cdot \min_{s\in[0,t]} \hat{L}(\theta(s))$$
$$= t\cdot \hat{L}(\theta(t)) \text{ by } 4.$$

$$\Rightarrow \tfrac{1}{2}\|\theta(t) - z\|^2 - \tfrac{1}{2}\|\theta(0) - z\|_2^2 \leq t\,\hat{L}(z) - t\,\hat{L}(\theta(t)). \qquad \blacksquare$$

**Lemma** For linearly sep. data w/ $y_i\langle \bar{u}, x_i\rangle \geq \gamma$, $\|\bar{u}\| = 1$, $\|x_i\| \leq 1$,

$$\hat{L}(\theta(t)) \leq \frac{1 + \log(2nt\gamma^2)}{2t\gamma^2} \quad \& \quad \|\theta(t)\| \geq \log(2tn\gamma^2) - \log(1 + \log(2tn\gamma^2))$$

**Pf** Take $z = \log(c)\,\bar{u}\,\gamma^{-1}$ (for some $c > 0$ to be determined) in the preceding theorem to get:

$$\hat{L}(\theta(t)) \le \hat{L}(z) + \frac{1}{2t}\left(\|z\|^2 - \|\theta(t) - z\|^2\right)$$

$$\le \sum_i \ell(m_i(z)) + \frac{\|z\|^2}{2t}$$

$$\le \sum_i \exp(-\log(c)) + \frac{\log^2(c)}{2t\gamma^2}$$

$m_i(z)$
$= \log(c)\,\gamma^{-1}\, y_i\langle \bar{u}, x_i\rangle$
$\ge \log c$
as

$$= \frac{n}{c} + \frac{\log^2(c)}{2t\gamma^2}.$$

Take $c := 2tn\gamma^2$.

For lower bound on $\|\theta_t\|$, note that

$\|\theta(t)\| \ge y_i\langle \theta, x_i\rangle = m_i(\theta)$ for any $i$ by C-S & $\|x_i\| \le 1$.

$$\Rightarrow \ell(\|\theta(t)\|) \le \ell(\max_i m_i(\theta(t)))$$

$$= \min_i \ell(m_i(\theta(t)))$$

$$\le \frac{1}{n}\hat{L}(\theta(t)) \le \frac{1 + \log^2(2tn\gamma^2)}{2tn\gamma^2} \quad \text{by .}$$

Now take $\ell^{-1}$ of both sides & use $\ell^{-1}$ decreasing.

Thus:

1. $\hat{L}(\theta(t)) \to 0$
2. $\|\theta(t)\| \to \infty$.

Now we show margin maximiz.

<u>Theorem</u> Consider linearly sep. data w/ exp loss & $\|x_i\| \le 1$. Then:

$$\gamma(\theta_t) \ge \tilde{\gamma}(\theta_t) \ge \bar{\gamma} - \frac{\log n}{\log t + \log(2n\gamma^2) - 2\log\log(2 t n e \gamma^2)}$$

Pf: Let $u(t) := \ell^{-1}(\hat{L}(\theta_t))$, $v(t) := \|\theta(t)\|$. Thus,

$$\tilde{\gamma}(\theta(t)) = \frac{u(t)}{v(t)} = \frac{u(0) + \int_0^t \frac{du(s)}{dt}\, ds}{v(t)}.$$

Want: $u(t)$ grows fast, $v(t)$ not too large.

Since $-\ell' = \ell$,

$$\frac{du}{dt} = \frac{d}{dt} -\ell(\hat{L}(\theta_t)) = \left\langle \frac{-\nabla \hat{L}(\theta_t)}{\hat{L}(\theta_t)}, \frac{d\theta}{dt} \right\rangle = \frac{\|d\theta/dt\|^2}{\hat{L}(\theta_t)}.$$

$$\left\|\frac{d\theta}{dt}\right\| \ge \left\langle \frac{d\theta}{dt}, \bar{u} \right\rangle = \left\langle \sum_i^n -\ell'(m_i(\theta)) y_i x_i, \bar{u} \right\rangle$$
$$= \left\langle \sum_i^n \ell(m_i(\theta)) y_i x_i, \bar{u} \right\rangle$$
$$\ge \gamma \sum_i^n \ell(m_i(\theta)) = \gamma \hat{L}(\theta).$$

OTOH, $\nu(t) = \|\theta(t)\| = \|\theta(t) - \theta(0)\| = \left\| \int_0^t \frac{d\theta(s)}{ds}\, ds \right\| \le \int_0^t \left\| \frac{d\theta(s)}{ds} \right\| ds$.

$$\Rightarrow \frac{\int_0^t \frac{du}{ds}\, ds}{\nu(t)} \ge \frac{\int_0^t \frac{\|d\theta/ds\|^2}{\hat{L}(\theta(s))}\, ds}{\int_0^t \|d\theta/ds\|\, ds} = \frac{\int_0^t \|d\theta/ds\| \cdot \frac{\|d\theta/ds\|}{\hat{L}(\theta(s))}\, ds}{\int_0^t \|d\theta/ds\|\, ds}$$

$$\ge \frac{\int_0^t \|d\theta/ds\| \cdot \gamma\, ds}{\int_0^t \|d\theta/ds\|\, ds} = \gamma.$$

prev. lemma

Now: $\frac{u(0)}{\nu(t)} = \frac{-\log(\hat{L}(\theta(0)))}{\|\theta(t)\|} = \frac{-\log(n)}{\|\theta(t)\|} \ge \frac{-\log n}{\log t + \log(2n\gamma^2) - \log\log(2tne\gamma^2)}$

Putting into we get claimed thm. $\boxed{\Sigma}$

This shows smooth normalized margin $\longrightarrow$ max margin as $t \to \infty$.

Since $\gamma(\theta) \ge \tilde{\gamma}(\theta)$, this implies $\theta(t)$ achieves max margin as $t \to \infty$.