Lecture 23: implicit bias continued

\[ \text{Want low test error, low training error is intermediate goal.} \]
\[ \text{DL gets nearly good test error. Why?} \]
\[ \text{Idea: GD prefers networks with low test error. (Noiseless.)} \]
\[ \text{Doing version: low norm.} \]

These lectures aim to prove GD prefers low norm solutions in some settings.

- Why low norm? E.g.,
  \[ \text{[test-train for linear } W_1 \text{] } \leq \frac{\|W_1\| + \|\nabla V(W)\|}{\sqrt{n}}. \]
  \[ \text{[test-train for deep } W_1 \ldots W_L \text{] } \leq \frac{\|W_1\| + \sum_{i=2}^{L} \|\nabla V(W)\|}{\sqrt{n}}. \]

\[ \text{Theorem. Suppose } \exists \omega \in \mathbb{R}^d, \min y, x \omega > 0. \]
\[ \text{Then GD } \omega(t) = -\nabla L(\omega(t)) \text{ satisfies} \]
\[ \frac{\min \| \omega \|}{\| \omega \|} = \frac{\partial^2 L(\omega(t))}{\| \omega \|} \geq \frac{1}{\sqrt{n} \log m} \implies \nabla^2 L(\omega(t)) \approx 0. \]

Remarks (intuition)

- For each unit of hull, \( \partial^2 L(\omega(t)) \) grows by \( \Delta \).
- L-homo genomic
  \[ \text{for each unit of hull, } \partial^2 L(\omega(t)) \text{ grows by current noise.} \]

Harvey & Bartlett & ?
"ReLU VC"

\[ \text{[test-train w/ param, liner, L] } \leq \frac{1}{\sqrt{n}} \sqrt{L}. \]
Proof write

\[-\ln L(w(t)) \leq \ln L(w(0)) + \int_0^t \frac{d}{ds} \ln L(w(s)) \, ds\]

where

\[\frac{d}{ds} \ln L(w(s)) = - \frac{\langle \nabla \ln L(w(s)), \dot{w}(s) \rangle}{L(w(s))}\]

Letting \( \tilde{u} := \arg\min_{L(w(s))} u^T x : y_i \) then

\[\| \dot{w}(s) \| \geq \langle \dot{w}(s), \tilde{u} \rangle = \langle - \sum_{i=1}^n \ell(y_i, x_i^T w(s)), \tilde{u} \rangle \]

\[= \sum_{i=1}^n \ell(y_i, x_i^T w(s)) \ell(y_i, x_i^T \tilde{u}) \geq \gamma\]

\[\Rightarrow L(w(s)) \geq \gamma.\]

\[L(w(t)) = \sum_{i=1}^n \ell(y_i, x_i^T w(s)) = \sum_{i=1}^n \exp(-y_i x_i^T w)\]

\[= \sum_{i=1}^n \exp(-y_i x_i^T w(t)) = \sum_{i=1}^n \exp(-y_i x_i^T w(0)) = \gamma \sum_{i=1}^n \exp(-y_i x_i^T w(0))\]

\[= \gamma \sum_{i=1}^n \exp(-y_i x_i^T w(0)) = \gamma.\]

Thus \( \Delta \geq \gamma \) and

\[-\ln L(w(t)) = -\ln n + \gamma.\]

"Done since \( \| w(t) \| = \| w_0 \| - \mathcal{O}(\ln t)\)."
Recall $L$-homogeneous means
\[ \forall c > 0, \quad m_c(c \cdot w) = c^L \, m_c(w) = c^L \, y_i \cdot f(x_i; w) \]

\[ \text{e.g.:} \quad m_x(c \cdot w) = c \, m(w) = c \, y_i \cdot x_i \cdot w \]

\[ \text{also} \quad \langle w, \partial m_c(w) \rangle = L \cdot m_c(w) \]
Theorem. Suppose \( L(w(0)) < 1 \) is \( L \)-homogeneous. Then \( t \mapsto \tilde{y}(w(t)) \) is non-decreasing.

Proof (due to Zhiwei Ji).

Based on Lyapunov.

Remark. \( L(w(0)) < 1 \) implies \( \min_i m_i(w) = \min_i \beta_i(x_i; w) > 0 \).

Proof plan: \[
\frac{d}{ds} \tilde{y}(w(t)) = \frac{d}{ds} \frac{-\ln L(w(t))}{\|w\|^2}.
\]

Many magic cancellations \& \( L \)-homogeneity.

Also: \[
\tilde{y}(w) = \frac{-\ln \sum_i l(m_i(w))}{\|w\|^2}.
\]

\( L(w) = \sum_i l(m_i(w)) \)

Remark (rates in linear case).

\( \dot{w}(t) = -\nabla L(w(t)) \)

Gave rate \[
\frac{\ln n}{\ln t}
\]

[Slow, but positive margin suffices.]

\( \dot{w}(t) = -\nabla \ln L(w(t)) \)

Rate becomes \[
\frac{\ln n}{t}
\]

[Proof becomes easier.]

[Can be interpreted as large step size.]