Lecture 22: implicit bias & margin maximization

Closing remarks on Clarke differentials.
* Good to skeptical! (and out it.)

* Unclear if reflects practice (GD can easily bounce out).

* "pretend" chain rule holds everywhere.
  \[ \nabla w(i) = \min \text{non zero element} \]

* Norm preservation
  \[ \frac{1}{2} \| W_i(t) \|^2 - \frac{1}{2} \| W_i(b) \|^2 = \frac{1}{2} \| W_i(t) \|^2 - \frac{1}{2} \| W_i(b) \|^2 \]
  & priors at logit (i,j).

* Treatment of nonneutrinos is open.

Notation:

\[ m_i(w) = y_i f(x_i; w) \] "unnormalized margin"

\[ L(w) = \sum_i l(m_i(w)) \] "unnormalized risk"

Generally use explore \( \ell(z) = \exp(-z) \)

* "modified analysis" for "exp-tail losses", e.g. logistic \( z \to \text{logistic}(1 + \exp(-z)) \)

* Not sure why no explore in DL practice; "floating point issues"? [AdaBoost uses explore]
**Linear Separability**

Suppose \( \exists \omega \in \mathbb{R}^d \) s.t. \( \forall i : y_i \langle x_i, \omega \rangle > 0 \)

**Remark:** Why margins?

- Generalization theory for margin-based predictors, which improves on large margins.

  \[ \Rightarrow \text{perhaps a part of the DL generalization story.} \]

**Nonlinear \( \ell \)-Homogeneous Case.**

- \( m_i(c\omega) = c^\ell m_i(\omega) \) (\( c > 0 \))

**Proposition:** If \( \exists \hat{\omega} : L(\hat{\omega}) < L(\omega) \)

- \( \Rightarrow \inf_{\omega} L(\omega) = 0. \)

**Proof.**

- \[ 0 \leq \inf_{\omega} L(\omega) = \lim_{c \to \infty} \sup_{\omega} L(c\hat{\omega}) \]

- \[ = \lim_{c \to \infty} \sum_i \exp(-c m_i(c\hat{\omega})) \leq \sum_i \lim_{c \to \infty} \sup_{\omega} \exp(-c m_i(\omega)) = 0. \]

**Remark:** Before had a **cone** of solutions

Thanks to margins, have a unique one.
Corollary: If \( \exists \omega \text{ s.t. } L(\omega) < L(0) \)

\[ \Rightarrow \text{ any } \omega \text{ with } L(\omega) \to \inf L(\omega) \]

has \( \lim_{\omega \to \infty} \|w\|_{\omega} = \infty \).

Let's find reasonable definition of margin.

Note \( \min_i m_i(w) = m_i(\frac{\omega}{\|w\|}) = \|w\| L_i(\frac{\omega}{\|w\|}) \)

\[ \Rightarrow \text{ Margin } \max_{\|w\|} \min_i m_i(w) \]

\[ = \sup_{w \in \mathbb{R}^P} \min_i \frac{m_i(w)}{\|w\|^2} \]

\[ \leq -\ln \frac{\lambda}{n} \exp(-m_i(w)) = \frac{L^{-1}(L(\omega))}{\|w\|^2} + \frac{\ln n}{\|w\|^4} \]

\[ \min_i \frac{m_i(w)}{\|w\|^2} \leq \frac{L^{-1}(L(\omega))}{\|w\|^2} + \frac{\ln n}{\|w\|^4} \]

\[ \leq -\ln \frac{\lambda}{n} \exp(-m_i(w)) \]

\[ \Rightarrow \min \text{ minimizing } L \text{ could maximize margin.} \]
Remark. Multiclass margin.

Predictor $f : \mathbb{R}^d \to \mathbb{R}^d$,

multiclass prediction error $\sum \arg \max_i f(x)_i \neq y$

margin $f(x)_y - \max_{i \neq y} f(x)_i$

$$\frac{\|w\|^2}{\|w(t)\|^2} \geq \frac{\mathcal{L}(L(w))}{\|w\|} \geq y - \frac{\ln n}{\ln t + O(1)}.$$