Lecture 22: L-homogeneous KKT points & project info

Last time: minimum norm predictors (care because: generalize well).

local plan:

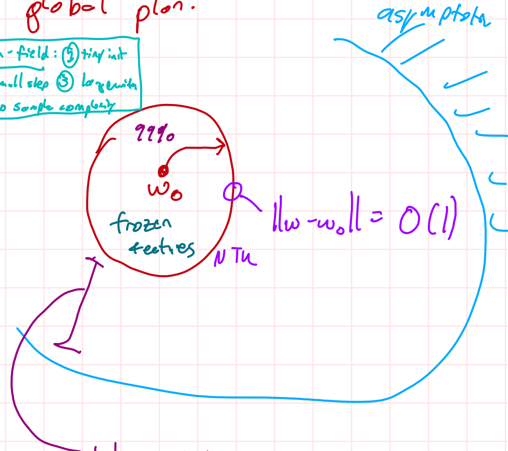
* GF on exponential or logistic loss (& linear pred) $\xrightarrow{t\to\infty}$ minimum norm (classification) soln.

$$\min \tfrac{1}{2}\|w\|_2^2 \quad \text{s.t.} \quad y_i x_i^T w \geq 1 \quad \forall i \qquad \Big| \qquad \max \min_i y_i x_i^T w \quad \text{s.t.} \quad \|w\|_2 \leq 1$$

* closest theorem L-homogeneous preds

global plan:

Mean-field: (1) tiny init (2) small step (3) large width (4) no sample complexity



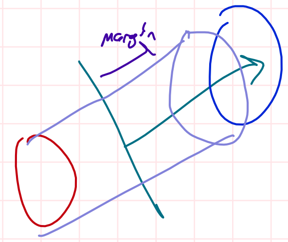
good news:
* includes feature learning

bad news
* slow convergence, hard to prove / unclear what's true

intermediate regime:
* still has some randomness
* little bit of feature learning
* matches practice

KKT points of L-homogeneous

Recall L-homogeneous predictor $F(x_i; c\,w) = c^L F(x_i; w)$, $c \geq 0$
(generalizes L-layer ReLU & leaky ReLU networks).



Linear case, margin $\max_{w \in \mathbb{R}^p} \frac{\min_i y_i F(x_i; w)}{\|w\|}$

$$= \max_{w \in \mathbb{R}^p} \frac{\min_i \|w\|^L y_i F(x_i; \frac{w}{\|w\|})}{\|w\|}$$

$$= \max_{w \in \mathbb{R}^p} \|w\|^{L-1} \min_i y_i F(x_i; w/\|w\|)$$

in linear case, $=1$; in $L \geq 2$ case, bogus defn.

Obvious fix: $\max_{w \in \mathbb{R}^p} \frac{\min_i y_i F(x_i; w)}{\|w\|^L} = \max_{\|w\| \leq 1} \min_i y_i F(x_i; w)$.

Remark: Modern networks are not L-homogeneous (e.g., softmax, biases).

Consider two opt problems

$$\min \tfrac{1}{2}\|w\|^2 \quad \text{s.t.} \quad y_i F(x_i; w) \geq 1 \ \forall i, \quad w \in \mathbb{R}^p \qquad \boxed{1}$$

$$\max_{w \in \mathbb{R}^p} \frac{-\ln \sum_i \exp(-y_i F(x_i; w))}{\|w\|^L} \qquad \boxed{2}$$

Mean-field

Theorem. Suppose GF on $\mathcal{L}(w) = \sum_i \exp(-y_i F(x_i; w))$ and $\exists \tau$ $\mathcal{L}(w_\tau) < 1$.

1. $t \mapsto -\ln \mathcal{L}(w_t)/\|w_t\|^L$ (i.e., objective in $\boxed{2}$) is nondecreasing over $[\tau, \infty)$. [Lyu-Li].
2. $\frac{w_t}{\|w_t\|} \to$ KKT point of $\boxed{1}$ ([Lyu-Li], [Ji-T.]) (under o-minimal definability).
   rules out oscillations; e.g., can't have $\sin(\cdot)$ activations

Proof remark: 1. note $v \mapsto \max_i v_i$ is 1-homogeneous, proof uses "approximate asymptotic 1-homogeneity" of $v \mapsto -\ln \sum \exp(-v_i)$.

2. KKT point proof shows "alignment": $\left\langle \frac{w_t}{\|w_t\|}, \frac{-\nabla \mathcal{L}(w_t)}{\|\nabla \mathcal{L}(w_t)\|} \right\rangle \to 1$.

Remark: goodness: regularity outside ntk.
bad news: no better than ntk.

Remark. People have shown convergence global max margin soln, but it requires infinitely many assumptions. [Chizat-Bach]