Lecture 24: Conclusion of margin maximization & opt.

Theorem. Suppose $L(w_0) < L(c_0)$. Then $t \mapsto \tilde{\gamma}(t) = \frac{e^{-t} L(w_t)}{\|w_t\|}$ is nondecreasing.

Remark: (Why?)
- Clean proof, holds under general conditions (Beyond NTK)
- Exercises \( L \)-homogeneity.
- \((\star)\) As \( \|w_t\| \to \infty \)
  \[ \tilde{\gamma}(t) \to \gamma(w_t) = \frac{\min_{j \in J} f(x_j; w_t)}{\|w_t\|} \]

\[ \Rightarrow \begin{cases} \text{A} & L(w_t) < 0 \\ \text{B} & \text{softmax}(f(x_j; w_t)) \to e_j \text{ for some } j. \end{cases} \]

NTK \implies $L(w_t) \leq \epsilon_j$.

Proof (soon).

Recall:

Theorem.\( \int \mathcal{F} \text{ convex & } \beta \text{-smooth,} \)
\[ \eta := \frac{1}{\kappa} \quad \forall \eta \geq 2/(1 - \beta) \]
\[ \leq 2\epsilon + \frac{1}{2} \beta \|w_0 - z\|^2 \]

Proof (soon).
Remark (omitted topics).

* "Mean-field perspective" (continuous time)
  Typically have infinite width, consider all times $t$ (not just "early phase")
* Typically $2$-homogeneous,
  \[ f(w_c) \leq c^2 f(w) \]
* [Chizat-Bach '18,] "on the global convergence...." converge to global min asymptotically
  (proof escapes saddle points)
* Acceleration / other step sizes / Adam / ...
  (I don't a clean theorem that beats GD in D.

* "Landscape analysis" (\$8)
  e.g. saying under some conditions, all local optima encountered by GD are global.
* Learning theoretic guarantees outside NTK
  e.g. better test error than NTK.
* "Escaping saddle points" (see also (41))
  \[ w' = w - \eta \nabla R(w) + \Delta(w, c^2 I) \]
  $\Rightarrow$ escapes saddles [Ge et al. '15]
  (can view this as time discretization of SGD $\Rightarrow$ global minimization)
Theorem. Suppose \( w \rightarrow f(x; w) \) locally Lipschitz & \( L \)-homogeneous.

Define \( m_i(w) := y_i f(x_i; w) \)

(which is \( L \)-homogeneous & locally Lipschitz)

Suppose \( L(w_0) = \sum_{i=1}^n l(m_i(w)) \)

\( \langle w, \dot{w} \rangle = \langle w, \nabla L(w) \rangle = \langle w, -\sum_i l'(m_i(w)) \dot{m}_i(w) \rangle \)

\( = \sum_i l'(m_i(w)) \langle w, \dot{m}_i(w) \rangle \)

\( = \sum_i l'(m_i(w)) \left( \nabla \cdot m_i(w) \right) \)

\( = L \left( \sum_i l'(m_i(w)) \left[ \ln \frac{1}{e^{-L w_i}} \right] \right) \)

\( = L \left( \sum_i l'(m_i(w)) \right) \ln \frac{1}{e^{-L w_i}} \)

Key property \( \langle w(t), \dot{w}(t) \rangle \geq L \cdot L(w) \ln \frac{1}{L(w)} \)

In linear case, used \( \langle \dot{w}, \ddot{w} \rangle \geq 0 \cdot L(w) \)

Proof. Plan: \( \dot{z} \geq 0 \)

\( \frac{d}{dt} L(w(t)) \)

\( = -\ln L(w(t)) \)

\( = -\frac{d}{dt} L(w(t)) \)

\( = \frac{d}{dt} \exp(-z) \)

\( = -\frac{d}{dt} \exp(-z) \)

\( = -L \left( \sum_i l'(m_i(w)) \right) \ln \frac{1}{e^{-L w_i}} \)

\( = L \left( \sum_i l'(m_i(w)) \right) \ln \frac{1}{e^{-L w_i}} \)

\( = 0 \Rightarrow \sum_i l'(m_i(w)) \left( \nabla \cdot m_i(w) \right) \)
\[ \dot{v}(t) = \frac{d}{dt} \|w_t\|^L = \frac{d}{dt} \langle w_t, w_t \rangle^{1/2} \]

\[ = \frac{1}{2} \langle w_t, w_t \rangle^{-1/2} \cdot 2 \langle w_t, \dot{w}_t \rangle \]

\[ = L \cdot \|w_t\|^{L-2} \cdot \langle w_t, \dot{w}_t \rangle \quad \text{(2)} \]

\[ \langle w_t, \dot{w}_t \rangle \leq L \cdot \|w_t\|^{L-1} \cdot \text{sup} \left( \|v\| / \|w_t\| \right) \]

\[ = L \cdot \|w_t\|^{L-1} \cdot \|\dot{w}_t\| \]

\[ \begin{align*}
\dot{v}(t) & = L \cdot \|w_t\|^{L-2} \cdot L(w_t) \frac{1}{L(w_t)} \\
& = L \cdot \|w_t\|^{L-2} \cdot \|w_t\| \langle \frac{w_t}{\|w_t\|}, \dot{w}_t \rangle \\
& \leq L \cdot \|w_t\|^{L-1} \cdot \text{sup} \left( \langle v, \dot{w}_t \rangle \right) \\
& = L \cdot \|w_t\|^{L-1} \cdot \|\dot{w}_t\|.
\end{align*} \]

\[ \dot{u}(t) = \frac{d}{dt} -\ln L(w_t) \]

\[ = \frac{1}{L(w_t)} \langle -\nabla L(w_t), \dot{w}(t) \rangle \]

\[ = \frac{\|\dot{w}_t\|^2}{2L(w_t)} \] \hspace{1cm} \text{Linear case} \hspace{1cm} \|\dot{w}_t\| \geq \langle \dot{w}_t, \dot{w}_t \rangle \geq L(w_t) \]

\[ \Rightarrow \frac{\|\dot{w}_t\|}{L(w_t)} \langle \dot{w}_t, \dot{w}_t \rangle \]

\[ \geq \frac{\|\dot{w}_t\|}{L(w_t) \cdot \|w_t\|} \left( L \cdot L(w_t) \ln \frac{1}{L(w_t)} \right) \]

Rest is in notes. (I tried to cover) (main points here.)