Lecture 17: Shallow smooth NTK

Pick up your papers:

* Crawl citation graph both directions from papers you already know & like.
* Look over conference acceptance lists.
  
  Crawl arXiv (cs.LG)

Context:

* 2018, multiple authors presented NTK as a way to understand $E^2$ on wide networks (*overparameterized*).

Hoffer

Jacob-Gabriel '18, Allen-Zhu/Li/Lin '18,
Du-Lee-Li-Wang-Zhai '18.

* Today’s version is due to Chizat & Bach ’19; they highlight

Shared whiteboard:

Caveats:

* Needs VPN
* No undo
* No palm/pen detection

IMO:

* Please write here or dit below, but not in my wiki
  except to mark some expression
  (To help stay organized/focused)
Scaling phenomenon: coefficient on predictor (variance of Gaussian) distorts much of the setting. Intuition: function becomes flat as we zoom in (rescale).

Notation & setting:

\[ f(\omega) = \begin{bmatrix} f(x_1; \omega) \\ f(x_2; \omega) \\ \vdots \\ f(x_n; \omega) \end{bmatrix} \in \mathbb{R}^n \]

\( \ell(\text{predictor notation}) \) overloads to \( \ell(\text{in training set}) \).

Squared loss regression:

\[ R(\alpha f(\omega)) = \frac{1}{2} \| y - \alpha f(\omega) \|^2, \quad R(t) = R(\alpha f(t^0)) \] for scale \( t > 0 \).

Gradient flow on \( R \):

\[ J_w := \begin{bmatrix} \nabla f(x_1; \omega)^T \\ \vdots \\ \nabla f(x_n; \omega)^T \end{bmatrix} \in \mathbb{R}^{n \times p} \]

\[ J_e := J_w e \]
\[ \dot{w}(t) = -V_w R(\alpha + \dot{w}(t)) = -a_j D R(\alpha F(w(t))) \]

Next comes \( u(t) \), the flow on the tangent model:

\[ f_0(u) := f(w(0)) + J_0(u - w(0)) \in \mathbb{R}^p \]

\[ u(t) = -\nabla_u R(\alpha f_0(u(t))) = -a_j D R(\alpha f_0(u(t))) \]

Initial conditions are:

\[ u(0) = w(0), \quad \text{thus} \quad f_0(u(0)) = f(w(0)) = f(w(0)) \]

Proof idea will be:

\* \( u(t) \) easy to analyze

\* We'll show \( u(t) \) stays close to inherits the behavior of \( u(0) \).

Assumptions / notation:

\[ \text{If } \text{rank}(J_0) = n \]

\[ \text{ Least squares always has solutions } \Rightarrow \text{ approximation + Hech } \]

\[ a \begin{cases} 
\sigma_{\min} := \sigma_{\min}(J_0) = \sqrt{\lambda_{\min}(J_0^T J_0)} > 0 \\
\sigma_{\max} := \sigma_{\max}(J_0) > 0 \\
\|J_w - J_v\| \leq \beta \|w - v\|.
\end{cases} \]
Theorem. Suppose \( a \geq \frac{\beta N^2 88^2 \sigma_{\text{max}} R_0}{\sigma_{\text{min}}^2}, \) for:

\[
\begin{align*}
(1) & \quad \max \{ R(o f(u(t)) \}, R(o f_0(u(t))) \} \leq R(0) \exp \left( -\frac{t e^2}{2} \right), \\
(2) & \quad \max \{ \| u(t) - w(0) \|, \| u(t) - w(d) \| \} \leq \frac{3 \sqrt{8} \sigma_{\text{max}} R(0)}{\sigma_{\text{min}}^2}, \\
\end{align*}
\]

\[\| u(t) - w(0) \| = \| u(t) - w(d) \|.\]

Remark. \( K := \frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}, \quad \alpha_0 := \frac{K}{\sigma_{\text{min}}}, \quad \sigma_{\text{max}} := \frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}.\)

\[
(1) \leq R(0) \exp \left( -\frac{t e^2}{2} \right), \\
(2) \leq 3 \sqrt{8} R(0),
\]
and \( \alpha_0 \) large enough when \( \sigma_{\text{min}} := \sqrt{\beta \sqrt{R(0)}}.\)

Proof idea.

Consider \( w \) s.t.

\[\| w - w(0) \| \leq \beta := \frac{\sigma_{\text{min}}}{2 \beta},\]

define

\[
T := \inf \{ t > 0 : \| w(t) - w(0) \| > \beta^2 \}.
\]

Proof structure:

1. Prove \( w(T) \) for \( \| \epsilon \| < \delta \)
Consequence: $T = 00$. 