Spectrally-normalized margin bounds for neural networks

Peter Bartlett
UC Berkeley and Qualcomm University of Technology

Dylan Foster
Cornell University

Matus Telgarsky
University of Illinois, Urbana-Champaign

Contributions.

1. A generalization bound

\[
\text{test error} \leq \text{train error} + \text{complexity term},
\]

where complexity term scales with \(\text{lipschitz/margin}\).

2. An empirical study of neural nets chosen by SGD, showing
   - problem complexity correlates with \(\text{lipschitz/margin}\);
   - observed test – train correlates with \(\text{lipschitz/margin}\).

SGD, lipschitz, and margins.

The Lipschitz constant of networks chosen by SGD correlates with problem complexity, and with test – train. Lipschitz is increasing, but lipschitz/margin is not.

Margins?

Define margin mapping \((x, y) \mapsto f(x)_y - \max_{i \neq y} f(x)_i\).

Margins give:

- Intuitive measures of confidence.
- Classification generalization via real-valued complexities.

But margins require proper normalization!

Generalization bound.

With probability \(1 - \delta\), margin \(\gamma\), data \(X \in \mathbb{R}^{d \times n}\), weight matrices \(A = (A_1, \ldots, A_l)\), network \(F_A(x) = \sigma_1(A_1; \sigma_1 \cdots \sigma_l(A_l)x) \cdots\)

satisfy

\[
\Pr[F_A(x) \neq y] \leq \hat{R}_n(F_A) + O \left( \frac{\prod_{i=1}^l \|A_i\|_\infty \|y\|}{\gamma} \cdot \left( \frac{\|X\|_2}{n} \right) \cdot \left( \sum_{i=1}^l \|A_i\|_2^2 \right)^{1/2} \right)
\]

where \(\hat{R}_n(F_A) \leq \Pr[F_A(x) - \max_{i \neq y} F_A(x)_i] \leq \gamma\).

Remarks.

- First term (purple) is the desired lipschitz/margin.
- Middle term (red) is standard.
- Last term (green) is worrisome; not present in lower bound; captures nonlinear/combinatorial structure.
- \(\gamma/3\) comes from optimizing per-layer covers.
- No combinatorial parameters!

Prior work has exponential dependence in \(L\) (Bartlett & Mendelson; 2003) (Neyshabur, Tomioka and Srebro; 2015).

Proof.

Step 1: Matrix covering (via Maurey Sparsification).
Given conjugate exponents \((p, q)\) and \((r, s)\) with \(p \leq 2\),

positive reals \((a, b, \epsilon)\), positive integer \(m\),

matrix \(X \in \mathbb{R}^{d \times n}\) with \(\|X\|_p \leq b\);

\[
\ln N \left( \{AX : A \in \mathbb{R}^{m \times d}, \|A\|_q \leq a, \epsilon, \|\cdot\|_2 \} \right) \leq \left( \frac{a^2 b^2 m^{2/s}}{\epsilon^2} \right) \ln(2dm).
\]

Step 2: Full-network cover via induction.
Suppose previous layer output \(X_i\) has cover \(\tilde{X}_i\).
Via matrix covering, apx \(X_{i+1} = \sigma(A_i X_i)\) with \(\tilde{X}_{i+1} := \sigma(\tilde{A}_i \tilde{X}_i)\):

\[
\|X_{i+1} - \tilde{X}_{i+1}\|_2 \leq \rho_i \|A_i X_i - \tilde{A}_i \tilde{X}_i\|_2 \\
\leq \rho_i \left( \|A_i X_i - A_i \tilde{X}_i\|_2 + \|A_i \tilde{X}_i - \tilde{A}_i \tilde{X}_i\|_2 \right) \\
\leq \rho_i \|A_i\|_2 \|X_i - \tilde{X}_i\|_2 + \rho_i \epsilon_i,
\]

and continue by induction.

Step 3: Combining pieces with Dudley, Rademacher, and friends.
Let \(F^\epsilon\) denote networks \(F_A\) where margins \(\mathcal{A} = (A_1, \ldots, A_l)\) satisfy \(\|y\| \leq \gamma, \|A_i\|_2 \leq b; \quad \sigma_i \leq \rho_i\)\-Lipschitz with \(\sigma_i(0) = 0\).

Combining above pieces gives

\[
\ln N(F^\epsilon, \epsilon, \|\cdot\|_2) \leq \left( \prod_{i=1}^l \rho_i^2 \|y\| \right) \cdot \left( \sum_{i=1}^l \|b_i\| \right)^{2/3} \left( \ln(2W^2) \right)^{2/3} / \epsilon^2;
\]

final bound follows via Dudley and other standard Rademacher tools.

(Lipschitz-normalized!) Margin distributions.

\[\text{margin} = \text{distribution.} \]

\[\text{margin/lipschitz} = \text{distribution.} \]