Spectrally-normalized margin bounds for neural networks

Peter Bartlett, Dylan Foster, and Matus Telgarsky

peter@berkeley.edu, djfoster@cs.cornell.edu, mjt@illinois.edu

Cornell University · ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN · Berkeley UNIVERSITY OF CALIFORNIA

Problem

- **Modern neural nets can fit random noise perfectly.**
- **Yet, same nets generalize from train to test on real-world datasets.**
- **How to reconcile these observations?**

Cf. [Zhang et al. 2017]

Contributions

1. A generalization bound

   $$\text{test error} \le \text{train error} + \text{complexity term},$$

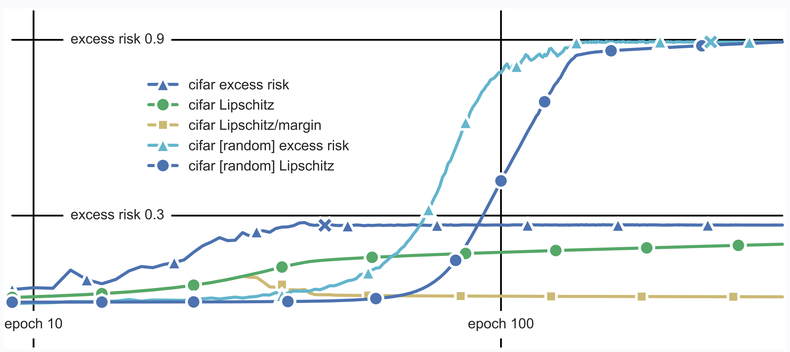
   where complexity term scales with lipschitz/margin.

2. An empirical study of neural nets chosen by SGD, showing
   - problem complexity correlates with lipschitz/margin;
   - observed test − train correlates with 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.



('x' marks the first epoch with perfect classification.)

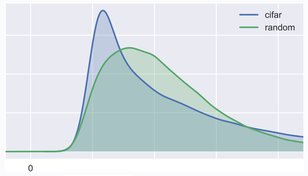
Margins?

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

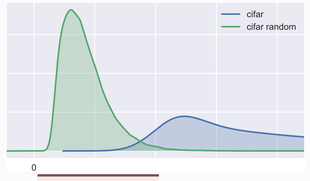
Margins give:

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

But margins require proper normalization!

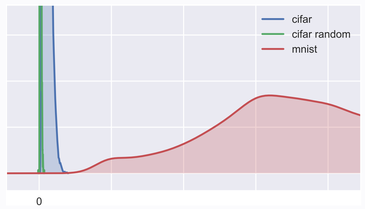


Margin distribution.

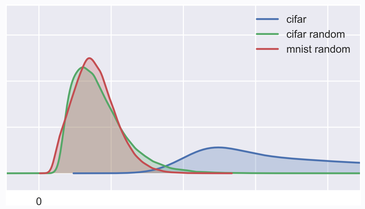


Margin/Lipschitz distribution.

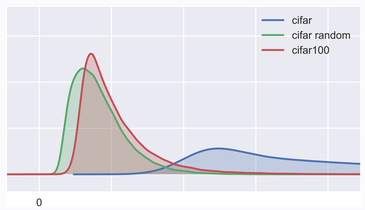
Lipschitz-normalized margin distributions



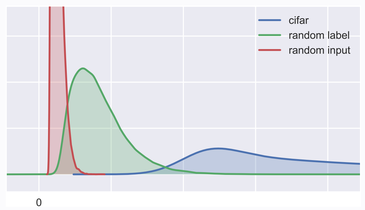
mnist is easier than cifar10.



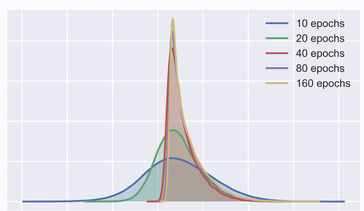
Random mnist is as hard as random cifar10!



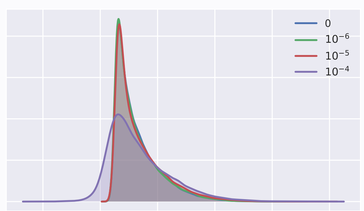
cifar100 is as hard as cifar10 with random labels!



Random inputs are harder than random labels.



Margins across epochs for cifar10.



Various levels of l_2 regularization for cifar10.

Generalization bound

With probability $1 - \delta$,
margin γ, data $X \in \mathbb{R}^{d \times n}$, weight matrices $\mathcal{A} = (A_1, \dots, A_L)$,
network $F_{\mathcal{A}}(x) = \sigma_L(A_L \sigma_{L-1} \cdots \sigma_1(A_1 x) \cdots)$
satisfy

$$\Pr[F_{\mathcal{A}}(x) \neq y] \le \widehat{\mathcal{R}}_\gamma(F_{\mathcal{A}}) + \widetilde{\mathcal{O}}\left(\left(\frac{\prod_{i=1}^n \rho_i \|A_i\|_\sigma}{\gamma} \right) \cdot \left(\frac{\|X\|_2}{n} \right) \cdot \left(\sum_{i=1}^L \frac{\|A_i\|_{2,1}^{2/3}}{\|A_i\|_\sigma^{2/3}} \right)^{3/2} \right)$$

where $\widehat{\mathcal{R}}_\gamma(F_{\mathcal{A}}) \le \Pr[F_{\mathcal{A}}(x) - \max_{i \neq y} F_{\mathcal{A}}(x)_i \le \gamma]$,
$\|A\|_\sigma = \max_i |\sigma_i(A)|$ (spectral norm) and $\|A\|_{2,1} = \sum_i \|A_{i\cdot}\|_2$ (group norm).

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.
- $2/3$ comes from optimizing per-layer covers.
- No combinatorial parameters!
  Prior work has *exponential dependence* in L
  (Bartlett & Mendelson; 2003) (Neyshabur, Tomioka and Srebro; 2015).
- $\frac{\|X\|_2}{n} = O(1/\sqrt{n})$ when data is bounded.

Proof

Step 1: Matrix covering (via Maurey Sparsification).
Given conjugate exponents (p, q) and (r, s) with $p \le 2$,
positive reals (a, b, ϵ), positive integer m,
matrix $X \in \mathbb{R}^{d \times n}$ with $\|X\|_p \le b$;

$$\ln \mathcal{N}\left(\{AX : A \in \mathbb{R}^{m \times d}, \|A\|_{q,s} \le a\}, \epsilon, \|\cdot\|_2 \right) \le \left\lceil \frac{a^2 b^2 m^{2/r}}{\epsilon^2} \right\rceil \ln(2dm).$$

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

$$\begin{aligned} \|X_{i+1} - \widehat{X}_{i+1}\|_2 &\le \rho_i \|A_i X_i - \widehat{A}_i \widehat{X}_i\|_2 \\ &\le \rho_i \left(\|A_i X_i - A_i \widehat{X}_i\|_2 + \|A_i \widehat{X}_i - \widehat{A}_i \widehat{X}_i\|_2 \right) \\ &\le \rho_i \|A_i\|_\sigma \|X_i - \widehat{X}_i\|_2 + \rho_i \epsilon_i, \end{aligned}$$

and continue by induction.

Step 3: Combining pieces with Dudley, Rademacher, and friends.
Let $\mathcal{F}$ denote networks $F_{\mathcal{A}}$ where matrices $\mathcal{A} = (A_1, \dots, A_L)$
satisfy $\|A_i\|_\sigma \le s_i$, $\|A_i\|_{2,1} \le b_i$; σ_i is ρ_i-Lipschitz with $\sigma_i(0) = 0$.
Combining above pieces gives

$$\ln \mathcal{N}(\mathcal{F}, \epsilon, \|\cdot\|_2) \le \left(\prod_{i=1}^L \rho_i^2 s_i^2 \right) \cdot \|X\|_2^2 \cdot \left(\sum_{i=1}^L \left(\frac{b_i}{s_i} \right)^{2/3} \right)^3 \cdot \left\lceil \frac{\ln(2W^2)}{\epsilon^2} \right\rceil;$$

final bound follows via Dudley and other standard Rademacher tools.