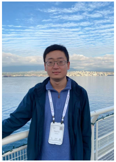
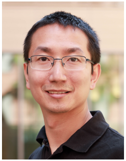
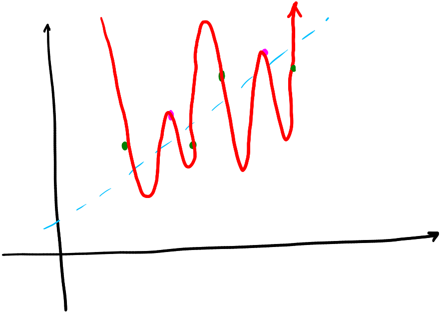
# Provable Generalization of SGD-trained Neural Networks of Any Width in the Presence of Adversarial Label Noise

Spencer Frei* Yuan Cao° Quanquan Gu°

*UCLA Department of Statistics
°UCLA Department of Computer Science

# Nonconvexity, Overparameterization, and Noise



- How does SGD-training succeed at minimizing training error when the problem is *nonconvex?*
- Why can *overparameterized* neural networks generalize well when trained on *noisy data*?

# Problem setup: adversarial label noise

- Underlying halfspace $y = \mathrm{sgn}(\langle v, x \rangle)$, but $(x, y) \sim \mathcal{D}$ has label corrupted $y \mapsto -y$ w.p. $p(x) \in [0, 1]$.
$$\mathsf{OPT}_{\mathsf{lin}} = \mathbb{E}_{x \sim \mathcal{D}} p(x).$$

- We will show SGD-trained NNs have classification error of at most $C\sqrt{\mathsf{OPT}_{\mathsf{lin}}}$.

- Consider neural networks with one hidden layer,

$$f_x(W) := \sum_{i=1}^{m} a_j \sigma(\langle w_j, x \rangle),$$

  $W \in \mathbb{R}^{m \times d}$ has rows $w_j^\top$; $\vec{a} \in \mathbb{R}^m$: second layer weights.
- $\sigma$: Leaky ReLU.
- Population-level cross entropy loss and classif. error:

$$L(W) := \mathbb{E}_{(x,y)} \ell(y f_x(W)), \quad \mathrm{err}(W) = \mathbb{P}_{(x,y)} \Big( y \neq \mathrm{sgn}\big(f_x(W)\big) \Big).$$

- Online SGD: $(x_t, y_t) \overset{\text{i.i.d.}}{\sim} \mathcal{D}$, with per-sample loss

$$\widehat{L}_t(W) := \ell(y_t f_{x_t}(W)) = \ell(y_t f_t(W)).$$

- Updates given by

$$W^{(t+1)} = W^{(t)} - \eta \nabla \widehat{L}_t(W^{(t)}).$$

# Learning noisy halfspaces with neural networks

**Theorem**

*If $\mathcal{D}_x$ satisfies anti-concentration (e.g. log-concave isotropic), then with small initialization, constant step size, and time/sample complexity $T = C \cdot \mathsf{OPT}_{\mathsf{lin}}^{-3}$ we have*

$$\exists t^* < T \text{ s.t. } \mathbb{P}_{(x,y)\sim\mathcal{D}}\Big(y \neq \operatorname{sgn}\big(f_x(W^{(t^*)})\big)\Big) \leq C \cdot \sqrt{\mathsf{OPT}_{\mathsf{lin}}}$$

- All bounds ($T$, error) independent of width $m$ of network
- Overparameterized NN will *not* overfit any more than a single neuron
- Optimization problem is significantly more nonconvex

# Proof Overview

- Standard Polyak-Łojasiewicz (PL) inequality:

$$\left\|\nabla \widehat{L}(W)\right\|^2 \geq \frac{\mu}{2}[\widehat{L}(W) - L^*]$$

leads to efficient guarantees of the form $L(W^{(t)}) \leq L^* + \varepsilon$.

We show a *proxy PL inequality* holds:

$$\left\|\nabla \widehat{L}(W)\right\| \geq \frac{\mu}{2}\left[\widehat{\mathcal{E}}(W) - C \cdot \sqrt{\mathsf{OPT}_{\mathsf{lin}}}\right],$$

where $\mathcal{E}(W)$ is a surrogate to the 0-1 loss. This leads to $\mathcal{E}(W^{(t)}) \leq C\sqrt{\mathsf{OPT}_{\mathsf{lin}}} + \varepsilon$.

# Summary

- First result to show that SGD-trained NNs can generalize under adversarial label noise.
- Holds for NNs of arbitrary width and initialization.
    - Cannot be explained using $\infty$-width approximations like neural tangent kernel or mean field approximation
- Implies that SGD-trained networks will always be *weak learners* if linear classifiers are weak learners.