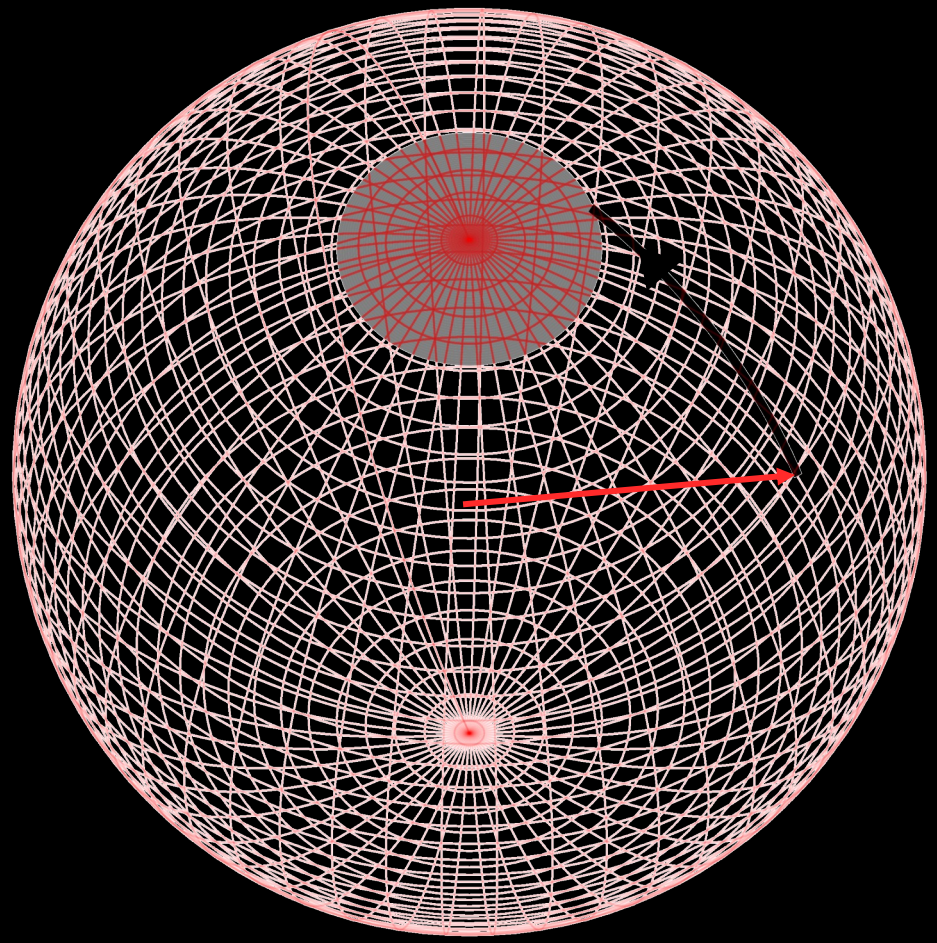
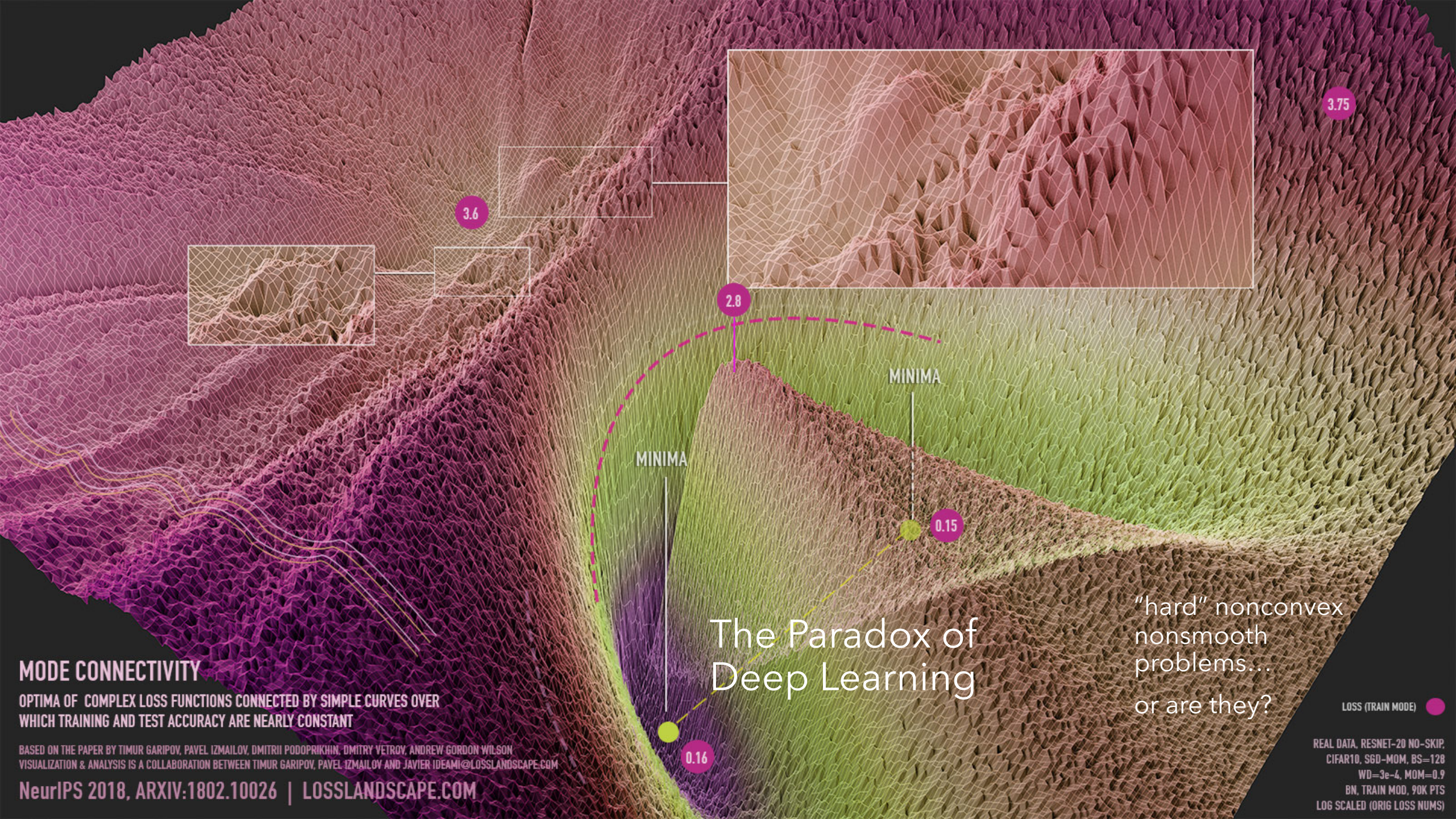


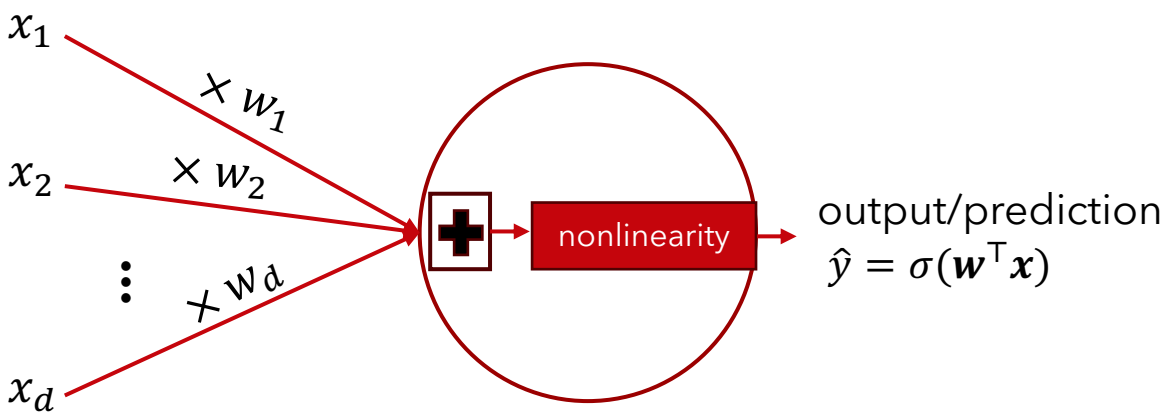
Gradient Alignment, Learning, and Optimization

Jelena Diakonikolas (UW-Madison)

ELLIIT Symposium: Optimization for Learning, May 2026


3.75
3.6
2.8
MINIMA
MINIMA
0.15
0.16
The Paradox of Deep Learning
"hard" nonconvex nonsmooth problems...
or are they?
LOSS (TRAIN MODE)
REAL DATA, RESNET-20 NO-SKIP,
CIFAR10, SGD-MOM, BS=128
WD=3e-4, MOM=0.9
BN, TRAIN MOD, 90K PTS
LOG SCALED (ORIG LOSS NUMS)
MODE CONNECTIVITY
OPTIMA OF COMPLEX LOSS FUNCTIONS CONNECTED BY SIMPLE CURVES OVER
WHICH TRAINING AND TEST ACCURACY ARE NEARLY CONSTANT
BASED ON THE PAPER BY TIMUR GARIPOV, PAVEL IZMAILOV, DMITRII PODOPRIKHIN, DMITRY VETROV, ANDREW GORDON WILSON
VISUALIZATION & ANALYSIS IS A COLLABORATION BETWEEN TIMUR GARIPOV, PAVEL IZMAILOV AND JAVIER IDEAMI@LOSSLANDSCAPE.COM
NeurIPS 2018, ARXIV:1802.10026 | LOSSLANDSCAPE.COM

Is learning one neuron easy or hard?



Joint Work With:

Puqian Wang

Nikos Zarifis

Ilias Diakonikolas

P. Wang*, N. Zarifis*, I. Diakonikolas, J. Diakonikolas, "*Robustly Learning a Single Neuron via Sharpness*," in Proc. ICML 2023. Oral Presentation.

N. Zarifis*, P. Wang*, I. Diakonikolas, J. Diakonikolas, "*Robustly Learning Single Index Models via Alignment Sharpness*," in Proc. ICML 2024.

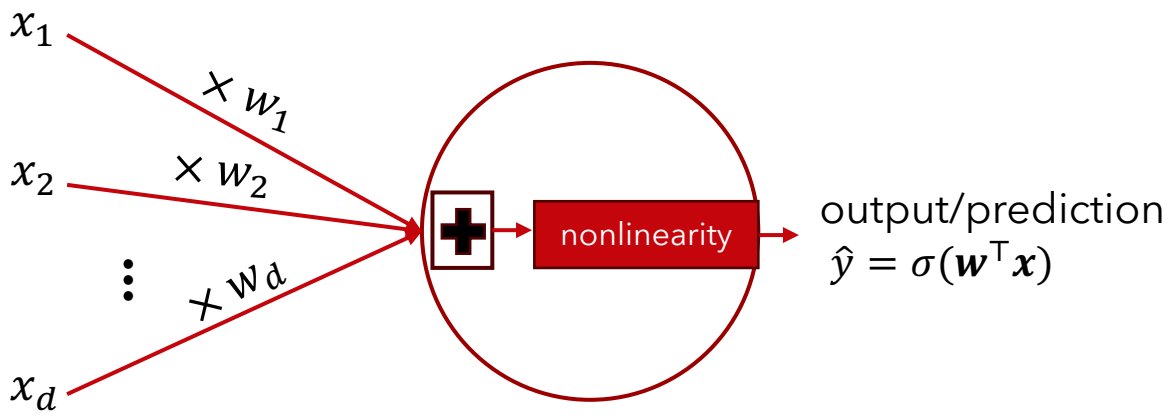
P. Wang*, N. Zarifis*, I. Diakonikolas, J. Diakonikolas, "*Sample and Computationally Efficient Robust Learning of Gaussian Single-Index Models*," in Proc. NeurIPS 2024.

N. Zarifis*, P. Wang*, I. Diakonikolas, J. Diakonikolas, "*Robustly Learning Monotone Generalized Linear Models via Data Augmentation*," in Proc. COLT 2025.

P. Wang*, N. Zarifis*, I. Diakonikolas, J. Diakonikolas, "*Robustly Learning Monotone Single-Index Models*," in Proc. NeurIPS 2025.

Generalized Linear Model (GLM) a.k.a. a Single Neuron

- Functions of the form $\sigma(w^\top x)$ for some activation σ: $\mathbb{R} \to \mathbb{R}$



$w \in \mathbb{R}^d$: (model) parameter vector

$x \in \mathbb{R}^d$: data vector

$y \in \mathbb{R}$: label

Data model:

(x, y) drawn from some distribution $\mathcal{D}$

$y = \sigma(w_*^\top x) +$ arbitrary noise

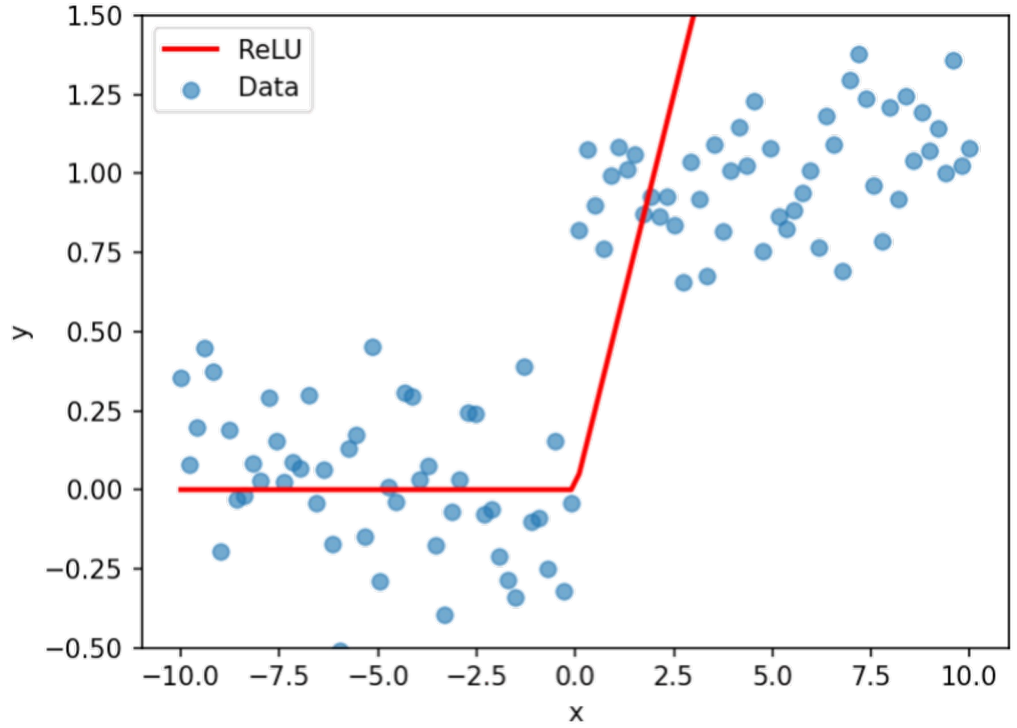
Goal: given σ, approximately* minimize

*we'll be more specific in a bit

$$F(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\sigma(w^\top x) - y)^2]$$

Generalized Linear Model (GLM) a.k.a. a Single Neuron

- Functions of the form $\sigma(w^\top x)$ for some activation σ: $\mathbb{R} \to \mathbb{R}$



$w \in \mathbb{R}^d$: (model) parameter vector

$x \in \mathbb{R}^d$: data vector

$y \in \mathbb{R}$: label

Data model:

(x, y) drawn from some distribution $\mathcal{D}$

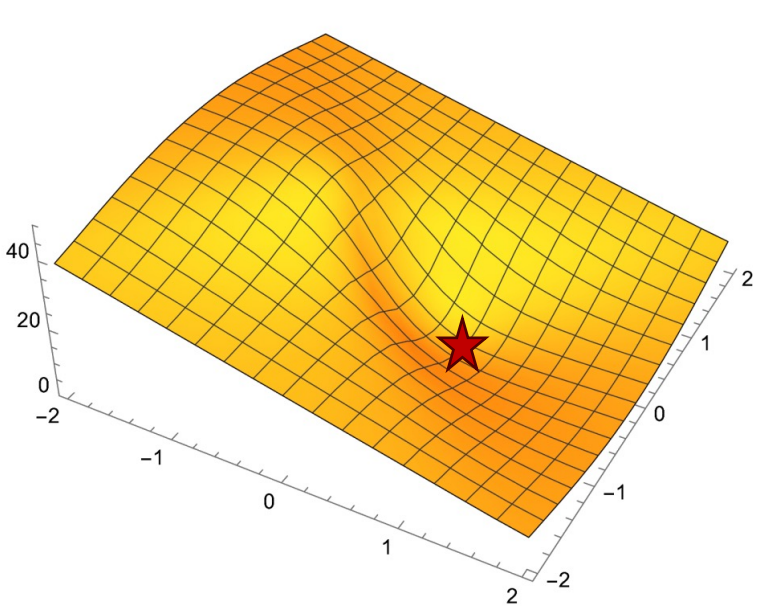
$y = \sigma(w_*^\top x) +$ arbitrary noise

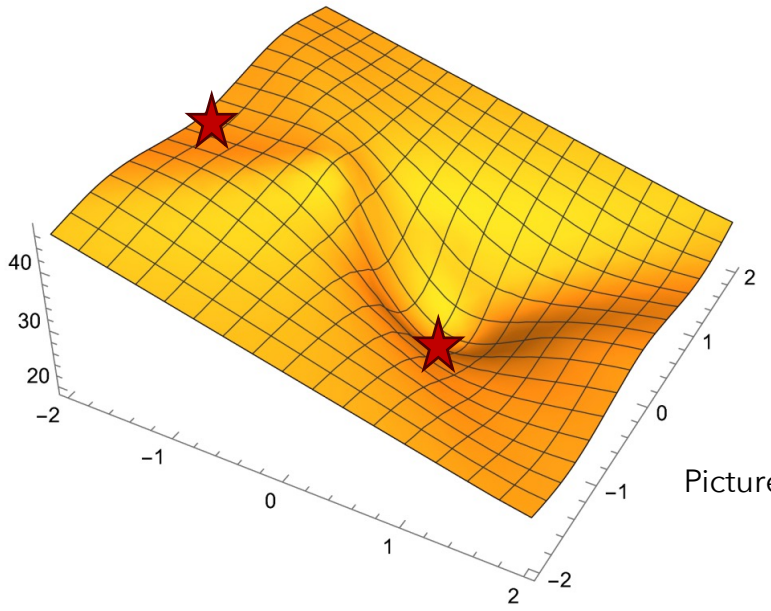
Goal: given σ, approximately* minimize

*we'll be more specific in a bit

$$F(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\sigma(w^\top x) - y)^2]$$

Label Noise Changes Already Nonconvex Loss Landscape





Picture credit to [DKTZ22], figure 1

Left: without label noise, Right: with label noise noise

- If label noise can be arbitrary, then:
  - it can add spurious local minima
  - it can change the gradient field to fool gradient-based algorithms

How Well Can We Approximate?

- Let $\mathrm{OPT} = \min_{w:\|w\|_2 \le W} F(w)$; $\quad F(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\sigma(w^\top x) - y)^2]$

- Without Distributional Assumptions, even for ReLU:
  - Finding w that gets error $O(\mathrm{OPT}) + \epsilon$ is **NP-Hard** [Manurangsi, Reichman (2018)]
  - Improperly learning with error $O(\mathrm{OPT}) + \epsilon$ requires super-polynomial time [I. Diakonikolas, Kane, Manurangsi, Ren (2022)]
- With Gaussian Covariates $x \sim N(0, I)$, even for ReLU:
  - Finding w that gets error $\mathrm{OPT} + \epsilon$ requires super-polynomial time [Goel, Karmalkar, Klivans (2019)], [I. Diakonikolas, Kane, Zarifis (2020)] , [Goel, Gollakota, Klivans (2021)], [I. Diakonikolas, Kane, Pittas, Zarifis (2021)], [I. Diakonikolas, Kane, Ren (2023)]
- With Gaussian Covariates $x \sim N(0, I)$, monotone activations, and "ideal" labels:
  - Information-theoretically impossible to learn an arbitrary monotone activation to error $\le 1/8$ [Zarifis, Wang, I. Diakonikolas, J. Diakonikolas (2025)]

How Well Can we Approximate?

- Let $\mathrm{OPT} = \min_{w:\|w\|_2 \le W} F(w)$; $F(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\sigma(w^\top x) - y)^2]$

Bottom Line:
Given $\epsilon > 0$, the best we can hope for is to find a vector w with $F(w) = C\ \mathrm{OPT} + \epsilon$, where $C > 1$, for structured classes of covariate distributions and activation functions.

Our Results

- Let $\mathrm{OPT} = \min_{w:\|w\|_2 \le W} F(w)$; $\quad F(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\sigma(w^\top x) - y)^2]$

Bottom Line:
Given $\epsilon > 0$, the best we can hope for is to find a vector w with $F(w) = C\ \mathrm{OPT} + \epsilon$, where $C > 1$, for structured classes of covariate distributions and activation functions.

What Our Results Look Like:
Given $\epsilon > 0$, we output w with $F(w) = C\ \mathrm{OPT} + \epsilon$ w.h.p., where $C > 1$ is an *absolute constant*, using $\mathrm{poly}(d, 1/\epsilon)$ samples and in $\mathrm{poly}(d, 1/\epsilon)$ time.

Our Results

- Let $\mathrm{OPT} = \min_{w:\|w\|_2 \le W} F(w)$; $\quad F(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\sigma(w^\top x) - y)^2]$

Bottom Line:
Given $\epsilon > 0$, the best we can hope for is to find a vector w with $F(w) = C\ \mathrm{OPT} + \epsilon$, where $C > 1$, for **structured classes of covariate distributions** and activation functions.

Fix a class of activations that contains ReLU, address distributions as general as possible

[WZD**D**'23], [ZWD**D**'24]

What Our Results Look Like:
Given $\epsilon > 0$, we output w with $F(w) = C\ \mathrm{OPT} + \epsilon$ w.h.p., where $C > 1$ is an *absolute constant*, using $\mathrm{poly}(d, 1/\epsilon)$ samples and in $\mathrm{poly}(d, 1/\epsilon)$ time.

Our Results

- Let $\mathrm{OPT} = \min_{w:\|w\|_2 \le W} F(w)$; $\quad F(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\sigma(w^\top x) - y)^2]$

Bottom Line:
Given $\epsilon > 0$, the best we can hope for is to find a vector w with $F(w) = C\ \mathrm{OPT} + \epsilon$, where $C > 1$, for structured classes of covariate distributions and **activation functions**.

Fix the standard Gaussian distribution, address activations as general as possible

[WZD**D**'24], [ZWD**D**'25, WZD**D**'25]

What Our Results Look Like:
Given $\epsilon > 0$, we output w with $F(w) = C\ \mathrm{OPT} + \epsilon$ w.h.p., where $C > 1$ is an *absolute constant*, using $\mathrm{poly}(d, 1/\epsilon)$ samples and in $\mathrm{poly}(d, 1/\epsilon)$ time.

Core Structural Result

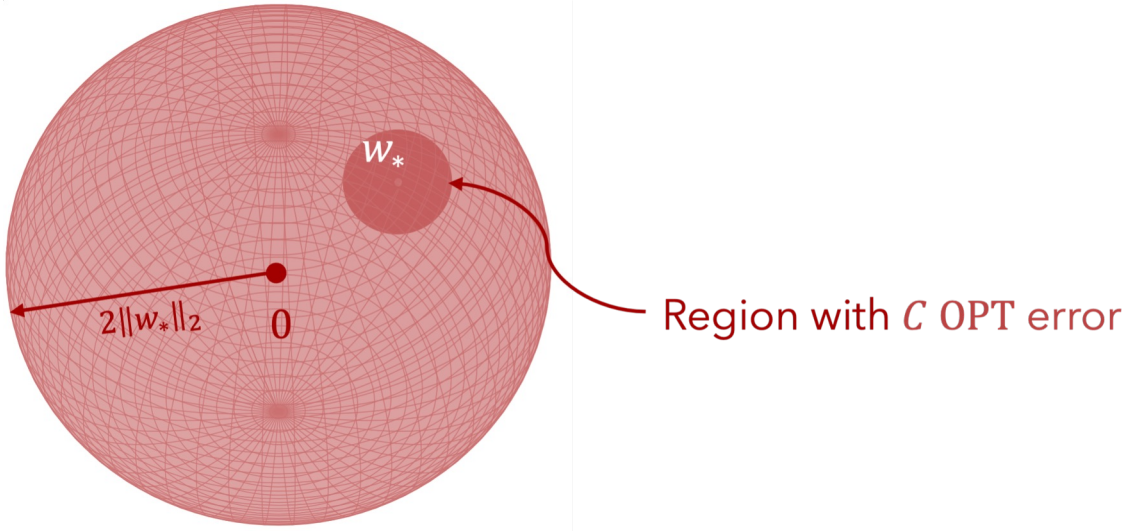
"Tame" Nonconvexity as a Local Error Bound

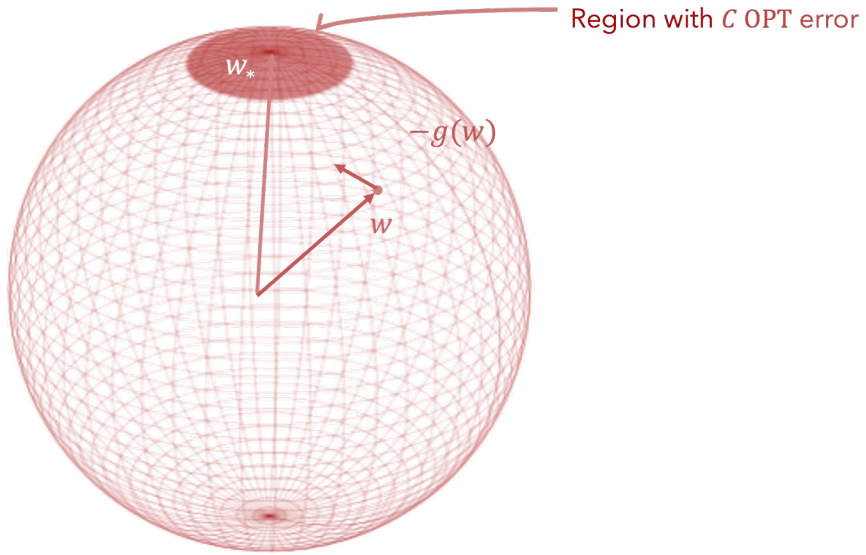
All our results begin from proving core structural results of the form:

"There is a 'signal' vector field $g(w)$ that, for any w in a region with loss $> C$ OPT satisfies

$$g(w)^{\mathsf{T}}(w - w_*) \geq \mu\, d(w, w_*)^2,$$

where $\mu > 0$ is an absolute constant, $d(\cdot,\cdot)$ a distance metric, and w_* the 'ground truth' parameter vector. Moreover, $g(w)$ can be efficiently estimated from labeled data examples (x_i, y_i)."





How we Choose $g(w)$

- Sometimes as the gradient of a classical surrogate function due to [Auer, Herbster, Warmuth (1995)]; in particular, in [WZD**D**, ICML'23], [ZWD**D**, ICML'24]
- Sometimes as the gradient of a surrogate we construct [WZD**D**, NeurIPS'24]
- Sometimes as the gradient of a smoothed loss, but with variable smoothing dependent on the distance to target [ZWD**D**, COLT'25]
- Sometimes there is no function of which $g(w)$ is the gradient [WZD**D**, NeurIPS'25]

In almost all these settings, we can argue that SGD on the original loss would not work in general.

Plan For the Rest of The Talk: Gaussian Covariates

- Part 1: Robust learning of GLMs (known activation)
  - Problem statement
  - Main result + related work
  - Main ideas
- Part 2: Robust learning SIMs (unknown activation)
  - Quick overview

Learning Monotone Gaussian GLMs

I.e., with a known activation/link function

Back to the Problem Statement

Given:

- sample access $(x, y) \sim \mathcal{D}$ for a Gaussian distribution of covariates $\mathcal{D}_x = \mathcal{N}(0, I)$,
- a monotone activation σ with bounded $2 + \zeta, \zeta > 0$, moment
- error parameter $\epsilon > 0$

Goal: find $w \in \mathbb{R}^d$: $\underbrace{\|w\|_2 = 1}_{\text{w.l.o.g.}}$ such that for an absolute constant $C > 1$,

$$F(w) \leq C\ \mathrm{OPT} + \epsilon,$$

where

$$\mathrm{OPT} = \min_{w:\|w\|_2=1} F(w); \qquad F(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\sigma(w^\top x) - y)^2]$$

Main Result

Theorem (Informal). Consider the problem of learning a GLM under Gaussian covariate distribution, where the activation σ is monotonically non-decreasing and has bounded $2+\zeta$, $\zeta > 0$ moment, meaning that there exists B_σ such that $\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\left(\sigma(z)\right)^{2+\zeta}\right] \leq B_\sigma$.

Given an additive error parameter $\epsilon \in (0,1)$, there exists an algorithm that draws $N = \tilde{O}\left(d(B_\sigma/\epsilon)^{2/\zeta} + d/\epsilon^2\right)$ samples, runs in time polynomial in N, d and outputs $\widehat{\boldsymbol{w}}$ such that with probability at least 2/3, $F(\widehat{w}) \leq C\,\mathrm{OPT} + \epsilon$, for C an absolute constant independent of any problem parameters (i.e., independent of $\epsilon, \zeta, d, B_\sigma, W$).

- The success probability 2/3 can be boosted to $1-\delta$ at $\log(1/\delta)$ cost, for any $\delta \in (0,1)$
- The class of monotone activations with bounded $2+\zeta$ moments contains all monotone Lipschitz activations (e.g., biased ReLU) and all monotone bounded activations (e.g., LTF)
- Some assumption on activation in addition to monotonicity is information-theoretically necessary, even without any label noise

Related Work: Noiseless or Structured-Noise Labels

Classical Results:

- Broad classes of covariate distributions (e.g., with just bounded support) and all monotone Lipschitz functions:

  [Kalai, Sastry (2009)], [Kakade, Kanade, Shamir, Kalai (2011)]

More Recent Results:

- Convergence of (stochastic) gradient descent for structured distributions and/or activations:

  [Soltanolkotabi (2017)], [Yehudai, Shamir (2020)], [Ben Arous, Gheissari, Jagannath (2021)],

  [Damian, Nichani, Ge, Lee (2023)]

Related Work: Adversarial Label Noise

Structured distributions:

[Frei, Cao, Gu (2020)], [I. Diakonikolas, Goel, Karmalkar, Klivans, Soltanolkotabi (2020)], [I. Diakonikolas, Kontonis, Tzamos, Zarifis (2022)], [Gollakota, Gopalan, Klivans, Stavropoulos (2023)], [Wang, Zarifis, I. Diakonikolas, JD (2023)], [Zarifis, Wang, I. Diakonikolas, JD (2024)]

Gaussian distribution:

[Guo, Vijayaraghavan (2024)], [Wang, Zarifis, I. Diakonikolas, JD (2024)]

Even when specialized to Gaussian covariates and 1-Lipschitz activations, **none of these results can get error $C\,\mathbf{OPT} + \epsilon$** with C being an (absolute!) constant.

Main Ideas

Data augmentation, staircase approximation, Hermite expansion

Original Loss Function vs Classical Convex Surrogate

Original Squared Loss	**Convex Surrogate** [Auer, Herbster, Warmuth (1995)]
$F(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\sigma(w^\top x) - y)^2]$	$\mathcal{L}_{\text{sur}}(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\int_0^{w^\top x} (\sigma(t) - y)dt\right]$
Nonconvex even for ReLU	**Convex** for monotone activations
Minimizer = solution	Minimizer may not be the solution
Stationary point ≠ minimizer, in general	Stationary point = minimizer, by convexity

$$\nabla F(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\sigma(w^\top x) - y)\sigma'(w^\top x)x]$$

$$\nabla \mathcal{L}_{\text{sur}}(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\sigma(w^\top x) - y)x]$$

"Useful" Vector Field to Guide the Algorithm Updates

- Both gradients (of the square and surrogate loss) can be written as

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[(\sigma(w^\top x) - y)h(w,x)x]$$

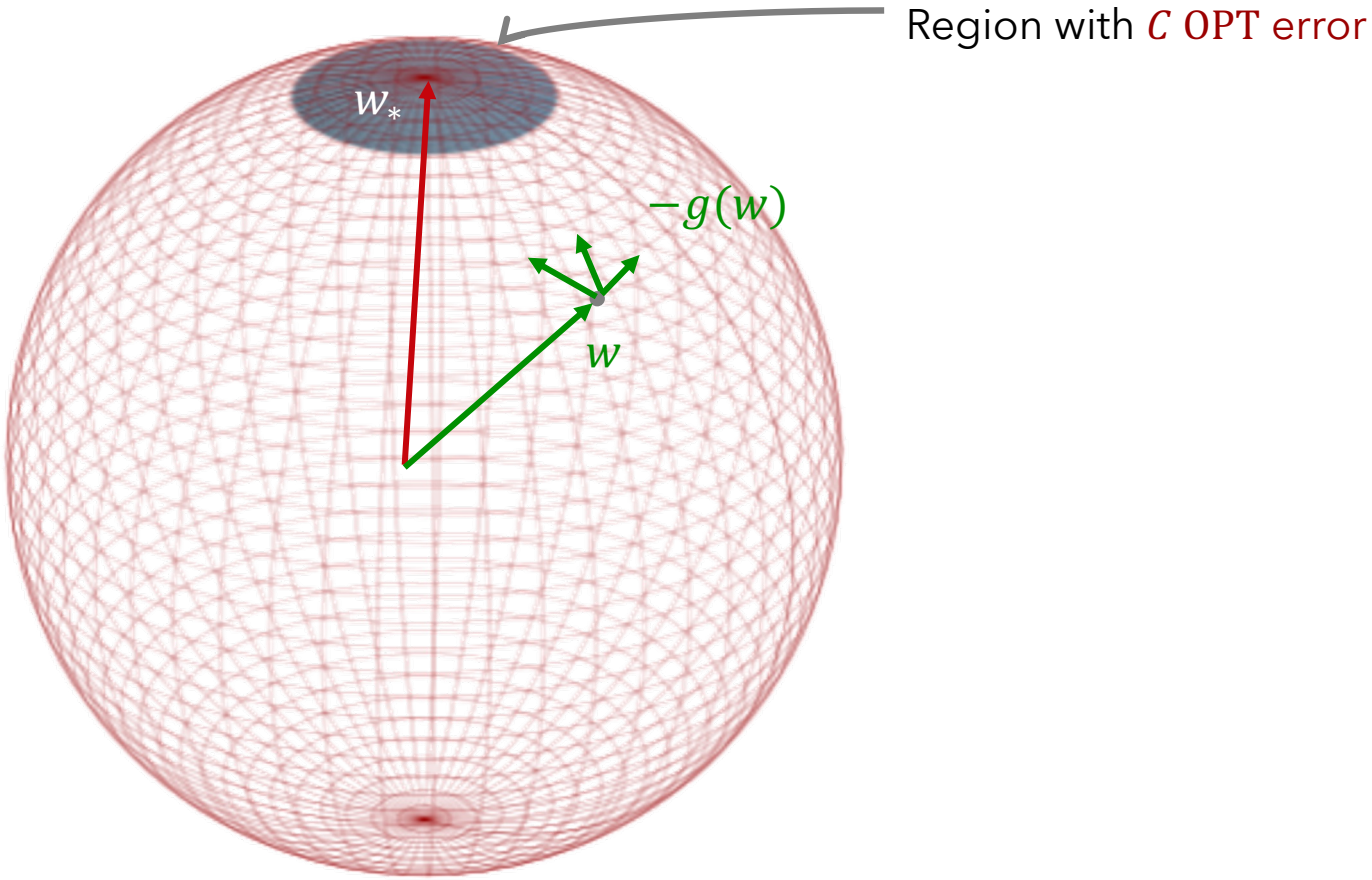
($h = \sigma'$ for the square loss; $h \equiv 1$ for the surrogate loss)

- Since specialized to the (unit) sphere ($\|w\| = 1$), can focus on the direction orthogonal to w:

$$g(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\sigma(w^\top x) - y)h(w,x)x^{\perp w}]$$

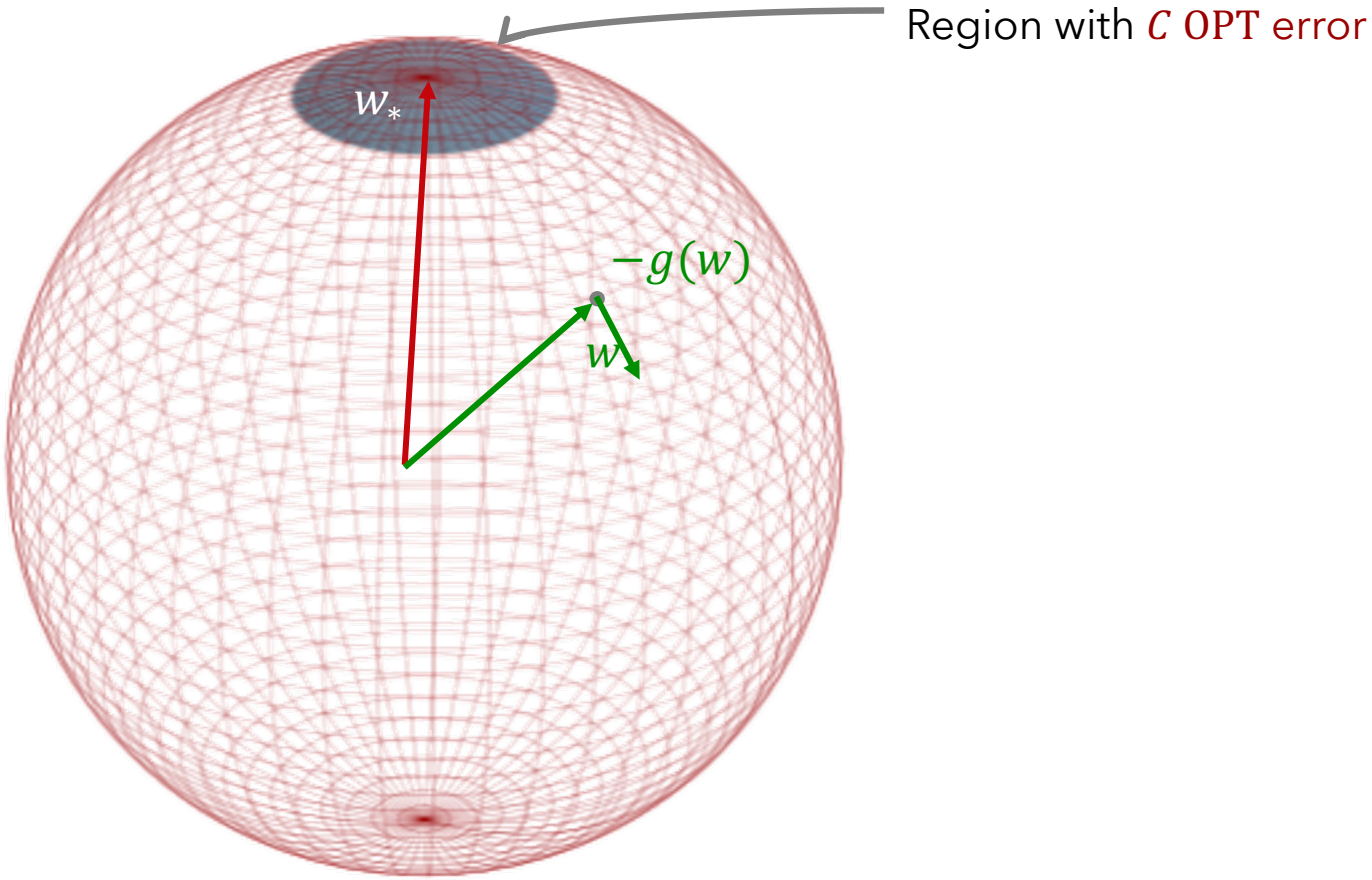
$$= -\mathbb{E}_{(x,y)\sim\mathcal{D}}[y\, x^{\perp w} h(w,x)]$$

A Geometric Perspective: Useful "Signal" in the Gradient



- Observation: if $-g(w)$ correlates (or "aligns") with w_*, then it can "pull" w towards w_*

Standard Choices of $g(w)$ Don't Work Here



- Can construct examples where $g(w)$ points in the wrong direction, even though the error is large

Designing the Update Vector Field

- What we want: if w is not already an $O(\mathrm{OPT})$ solution, then

$$-g(w)^\top w_* > 0$$

(we actually want something slightly stronger, but this is enough for now)

- Recall: $g(w) = -\mathbb{E}_{(x,y)\sim\mathcal{D}}[y\, x^{\perp w} h(w,x)]$. After a little bit of algebra (the definition of OPT, Cauchy-Schwarz, Stein's lemma), we can show:

$$-g(w)^\top w_* \geq \boxed{\mathbb{E}_{(x,y)\sim\mathcal{D}}[\sigma'(w_*^\top x) h(w,x)]} \sin^2\theta - \sqrt{OPT}\,\|h\|_{L_2} \ ,$$

where $\theta = \angle(w, w_*)$, $\|h\|_{L_2} = \left(\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\left(h(z)\right)^2\right]\right)^{1/2}$.

- Now we can consider maximizing $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\sigma'(w_*^\top x) h(w,x)]$.
- We can argue that the max is $h(w,x) \propto \underbrace{\mathbb{E}_{x,z\sim\mathcal{N}(0,I)}[\sigma'(\cos\theta\, w^\top x + \sin\theta\, w^\top z)]}_{\text{Gaussian-smoothed activation derivative}}$ for an independently drawn z

Other Analysis Ingredients

- Issue: we don't know $\theta = \angle(w, w_*)$, so we cannot use "the best" h
- Still, this motivates the use of randomized smoothing (Ornstein-Uhlenbeck semigroup) in the algorithm design as a means for controlling label noise
- The algorithm is Riemannian gradient descent (on the unit sphere), applied to the smoothed loss with variable smoothing
- Other ingredients:
  - initialization via solving a linear classification problem
  - approximation of monotone functions using "staircase" (piecewise-constant) functions
  - properties of Hermite polynomials of staircase functions

Plan: Focus on Gaussian Covariates

- Part 1: Robust learning of GLMs (known activation)
  - Problem statement
  - Main result + related work
  - Main ideas
- Part 2: Robust learning SIMs (unknown activation)
  - Quick overview

Single-Index Models

I.e., unknown activation/link function

Setup

Given:

- sample access $(x, y) \sim \mathcal{D}$ for a Gaussian distribution of covariates $\mathcal{D}_x = \mathcal{N}(0, I)$,
- class of activations $\mathcal{F}$
- error parameter $\epsilon > 0$

Goal: find $\sigma \in \mathcal{F}$ and $w \in \mathbb{R}^d$: $\|w\|_2 = 1$ such that for an absolute constant $C > 1$,

$$F(w; \sigma) \leq C\ \mathrm{OPT} + \epsilon,$$

unknown activation

where

$$\mathrm{OPT} = \min_{\substack{w:\|w\|_2 \leq W, \\ u \in \mathcal{F}}} F(w; u); \qquad F(w; u) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(u(w^\top x) - y)^2]$$

What We Already Knew

- If labels are realizable (OPT = 0) or with zero-mean noise (OPT = variance): has been known how to get OPT+ϵ for a while under mild assumptions; e.g., [Kalai-Sastry 2009], [Kakade-Kanade-Shamir-Kalai 2011]
- If labels can be arbitrary:
  - [Gollakota-Gopalan-Klivans-Stavropoulos 2023] applies to all 1-Lipschitz activations and $(x, y) \sim \mathcal{D}$ such that $\mathcal{D}_x$ has 2nd moment bounded by λ and $|y|$ is bounded by 1. Error guarantee:

$$O\left(W\sqrt{\lambda}\sqrt{\mathrm{OPT}}\right) + \epsilon$$

this may not be a constant; likely unavoidable without further distributional assumptions

not a linear scaling with OPT

  - Our own prior work [ZWD**D**, ICML 2024] gets the right error $O(\mathrm{OPT} + \epsilon)$ for a broader class than Gaussians, but a restricted class of activations (that contains ReLU but does not contain all Lipschitz activations or activations like sigmoid, LTFs, etc.)

Main Result

Theorem (Informal). Consider the problem of learning a SIM under Gaussian covariate distribution, where the activation class $\mathcal{F}$ contains all monotonically non-decreasing functions with bounded $2+\zeta,\ \zeta > 0$ moment, meaning that there exists B_σ such that $\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\left(\sigma(z)\right)^{2+\zeta}\right] \leq B_\sigma$.

Given an additive error parameter $\epsilon \in (0,1)$, there exists an algorithm that draws $N = d^2\mathrm{poly}((B_\sigma/\epsilon)^{1/\zeta}, 1/\epsilon)$ samples, runs in time polynomial in N and outputs $\widehat{\boldsymbol{w}}$ such that with probability at least 2/3, $F(\widehat{w}) \leq C\ \mathrm{OPT} + \epsilon$, for C an absolute constant independent of any problem parameters (i.e., independent of $\epsilon, \zeta, d, B_\sigma, W$).

Bonus Content: Learning a DRO Neuron

Robustly Learning a Neuron With Group Distributional Shifts

- The resulting problem is nonconvex-concave
- We leverage structural results about learning a neuron to develop a primal-dual algorithm, with a guarantee on the primal-dual gap; in [LKD**D**, NeurIPS 2024], [CLK**D**, AISTATS 2026]
- Algorithmic ideas seem useful more broadly; comparison to DoReMi [Xie et al., 2023], [Xia et al., 2024] for LLM pre-training:

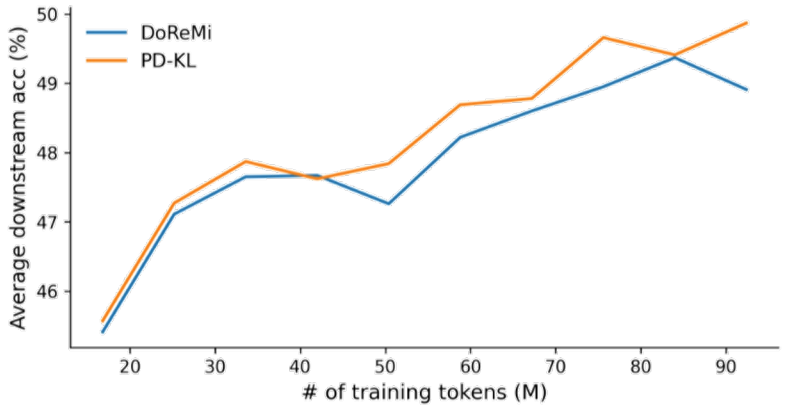


Figure 1: Compute-performance curve on *Sheared-LLaMA-1.3B*. Y-axis is the unweighted overall accuracy scores, X-axis is the number of tokens trained.

Table 1: Per-task results (%) at 92.4M tokens.

Bucket	Task	Metric	DoReMi	PD-KL
Commonsense & RC	**ARC-E**	acc_norm	**50.00**	49.83
Commonsense & RC	**ARC-C(25)**	acc_norm	29.95	**30.38**
Commonsense & RC	**HellaSwag(10)**	acc_norm	**54.78**	54.62
Commonsense & RC	**PICA**	acc	70.78	**71.87**
Commonsense & RC	**SciQ**	acc	85.00	**85.90**
Commonsense & RC	**WinoGrande**	acc	54.22	**55.25**
Commonsense & RC	**WSC**	acc	36.54	36.54
Continued & LM	**BoolQ(32)**	acc	56.39	**63.64**
Continued & LM	**LogicQA**	acc_norm	**28.11**	27.80
Continued & LM	**LAMBADA**	acc	48.61	**50.63**
World Knowledge	**TruthfulQA(5)**	acc	**23.62**	22.15
Unweighted Mean			48.91	**49.87**

Perspectives

- Main result: poly-time algorithms that learn any monotone Gaussian GLM or SIM that is information-theoretically possible to learn, with error guarantee as strong as possible
- Conceptual message: a case for looking broader than the function's (sub)gradient – there are other potentially useful update vector fields
- Structural local error bound results: intuitively, for structured distributions, there is a strong "signal" that can point the algorithm towards approximate solutions
- **Open Question**:
  - Can we go beyond Gaussians for monotone 1-Lipschitz or 1-bounded functions (in either setting)?
  - Address structured multi-index models with GD-like algorithms?

jelena@cs.wisc.edu

N. Zarifis*, P. Wang*, I. Diakonikolas, J. Diakonikolas, "*Robustly Learning Monotone Generalized Linear Models via Data Augmentation*," in Proc COLT 2025.

P. Wang*, N. Zarifis*, I. Diakonikolas, J. Diakonikolas, "*Robustly Learning Monotone Single-Index Models*," in Proc. NeurIPS 2025.