

Esha Singh, Shoham Sabach, Yu-Xiang Wang

Introduction and Preliminaries

f : **non-convex**, possibly **non-smooth**
 g : computable **proximal** operator

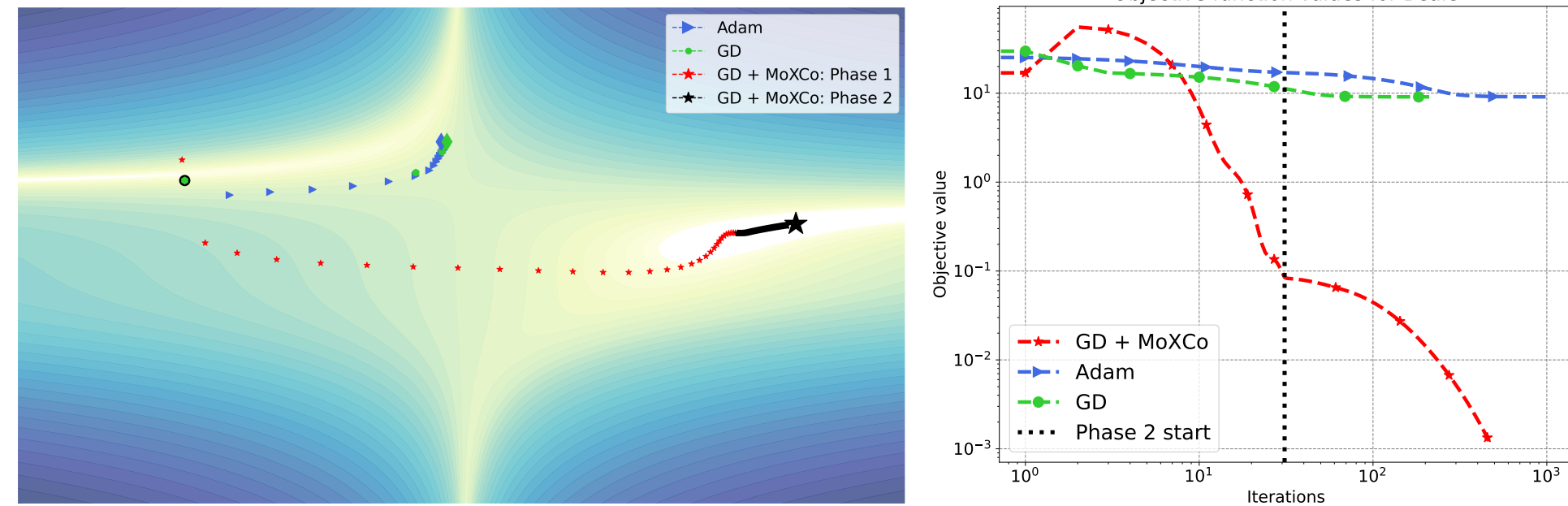
$$\arg \min_x f(x) + g(x)$$

Problems of interest

1. 2NN ReLU for Regression ($n > p$, $n \ll p$)
2. Quantized NNs: FP $\rightarrow$ binary weights

Goal: not only minimize but find a “good” solution that generalizes

Modern deep learning - high-dimensional loss landscapes pitfalls: plateaus & cliffs, sharp local minima, flat stationary points.



Toy loss landscapes such as Ackley/Beale mimic complex terrain of DNNs like ResNet-56 and quantization-inducing regularizers

Striving for **ERM global optimal** solution not always a right choice $\rightarrow$ No benign overfitting for ReLU NNs Noisy regression tasks [Joshi et al., 2023]

Many pitfalls for **optimizer (eg. SGD)** can get stuck into or explode from

Architectural designs $\rightarrow$ residual connections, batch norm; $\rightarrow$ **Not enough** to find optimal solutions

Optimization tricks $\rightarrow$ ADAM, gradient clipping, lr scheduling

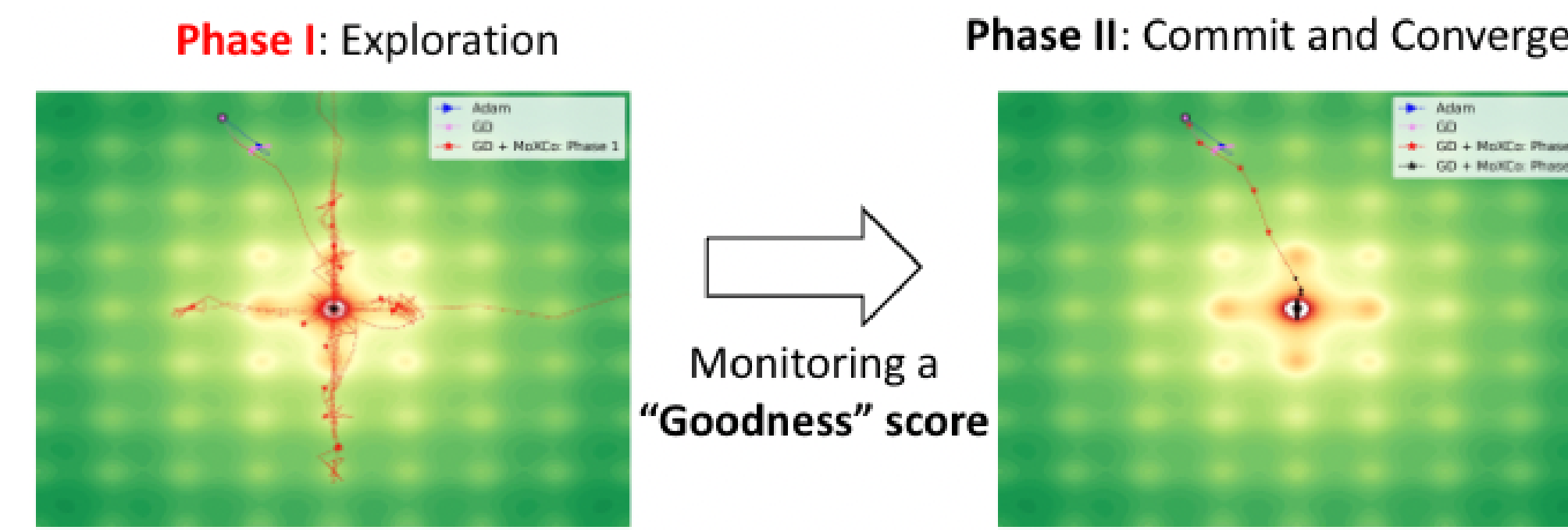
Observation: simultaneous optimization of loss values and loss sharpness for finding favorable local minima provides insufficient characterization of loss landscape’s local geometry.

Research Question: How to design optimization algorithms that converge to **good** local **stable** regions that generalizes better in HD loss landscapes?

Current answers in existing work

- Ochs et al. [2014]: Inertial force for use of a proximal gradient-type method handling structurally induced regularization.
- Cutkosky et al. [2023]: Estimation of stochastic gradients on a slightly perturbed point, for non-smooth & non-convex problems.
- Malviya et al. [2023]: Memory-augmented Adam variant that promotes exploration.

Two-phase optimization: **MoXCo**



Inertial Proximal Algorithm for exploration

Optimization trajectories on Ackley function with multiple local minima: MoXCo (red), ADAM (blue), and GD (pink). MoXCo achieves global minimum through 2 phases: initial aggressive exploration (red dots) followed by consistent convergence (black dots). ADAM and GD quickly converge to local minima.

Phase I: Inertial Proximal Algorithm for Promoting Exploration

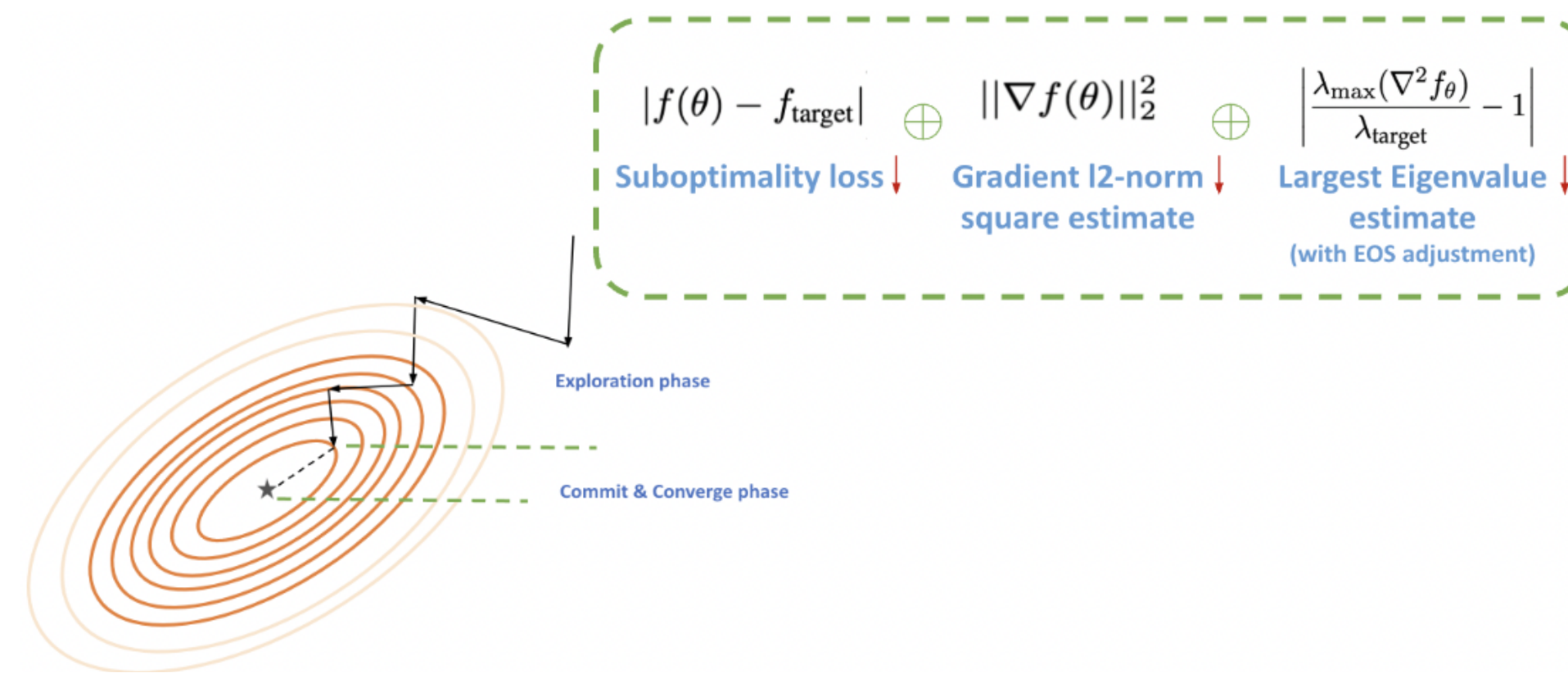
Momentum Exploration or “Inertial Adaptivity”: facilitating exploration & attaining faster convergence compared to non-momentum-accelerated methods. Can be used with any first-order optimizer!

$$\begin{aligned} \mathbf{u}_t &= \boldsymbol{\theta}_t + \alpha_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}) \\ \mathbf{v}_t &= \boldsymbol{\theta}_t + \beta_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}) \\ \boldsymbol{\theta}_{t+1} &= \text{Prox}_g(\mathbf{u}_t - \eta \hat{\nabla} f(\mathbf{v}_t)) \end{aligned}$$

Phase II: “Goodness” score and when to stop exploration

Monitor **goodness score** - a pre-defined criterion satisfied, in proximity of a *good* solution $\rightarrow$ commit to converge by reducing momentum.

$$\text{Goodness}(\theta) = \exp\left(-\tau \left[\|\nabla f(\theta)\|_2^2 + \left| \frac{\lambda_{\max}(\nabla^2 f_\theta)}{\lambda_{\text{target}}} - 1 \right| + |f(\theta) - f_{\text{target}}| \right]\right)$$

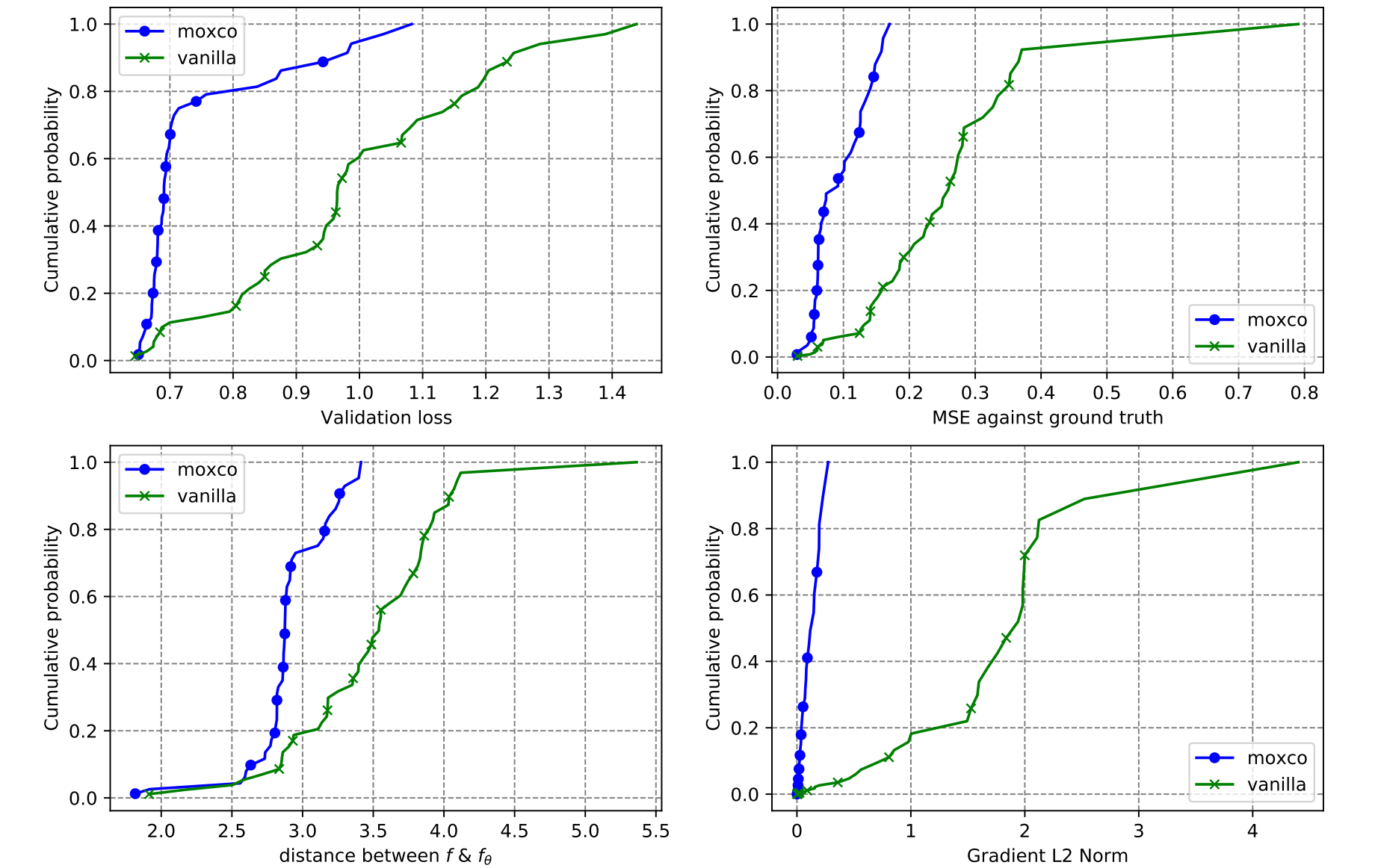


- τ : calibrates sensitivity of the score, Exponential transformation ensures score $0 \leq \text{Goodness} \leq 1$ as the exponent is nonzero.
- $\uparrow$ Goodness score, nearer to the solution that is approximately stationary, flatter than **Edge of Stability**, and close to an ideal target.

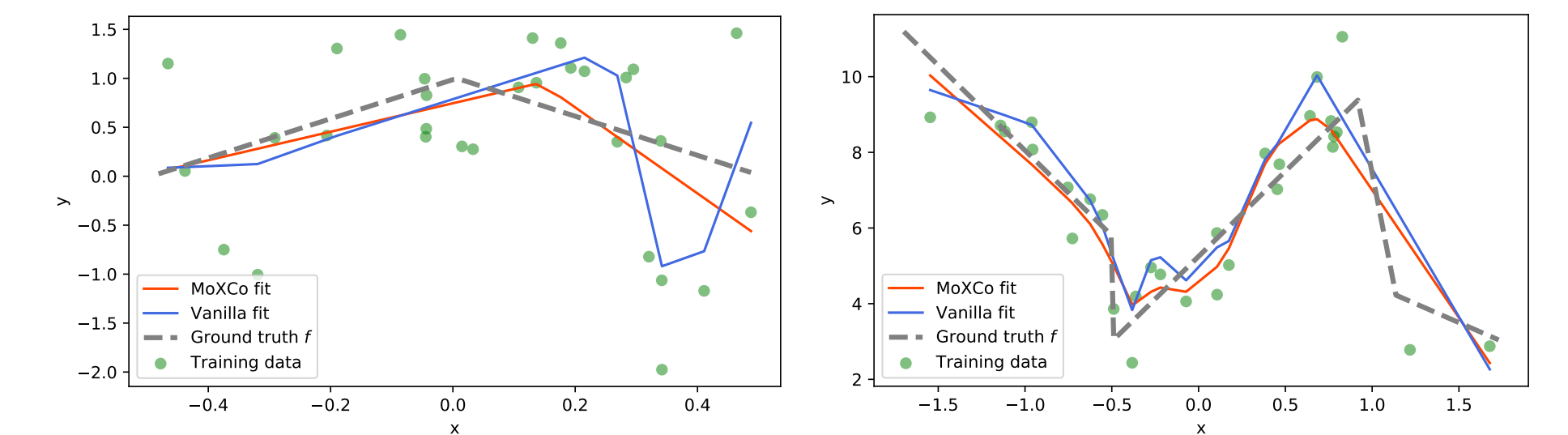
Main Theoretical and Experimental Results

Theorem 1: Let θ satisfies that (a) it is a local minimum of f ; (b) it is a stable fixed point of gradient descent with learning rate $\tilde{\eta}$, (c) $|f(\theta) - f_{\text{target}}| \leq \epsilon$.
Then $\text{Goodness}(\theta) \geq \exp(-\tau\epsilon)$.

Theorem 2: [Edge Of Stability (EOS) adjustment] If we consider running Inertial momentum to optimize a quadratic objective function of form $f(x) = \frac{1}{2}\mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + c$, then based on Cohen et al. [2021], if we consider running vanilla gradient descent on the $f(x)$ starting from any initialization and if $(\mathbf{q}, a)$ be an eigenvector/eigenvalue pair of $\mathbf{A}$, then if $a > \frac{2(1+\alpha)}{\eta(1+2\beta)}$, then the sequence $\{\mathbf{q}^T \mathbf{x}_t\}$ will diverge.



Hit rate analysis: 50 random initializations of under-parameterized 2NN ($m=5$) on 30 noisy-labeled data points. CDF (Y-axis) shows MoXCo+GD consistently outperforms vanilla GD across all “goodness” metrics measured at convergence.



Learned functions comparison: (L) Under-parameterized case ($m=5$): For same learning rate η , GD produces more complex solutions that fit noise, while MoXCo yields smoother fits. (R) $m=20000$.

Conclusions

A new take on developing optimizers that adapt to various pathological pitfalls in complex loss landscapes and converge to good solutions. Improvement (Top-1) in binarized ResNets on CIFAR10 by $> 2\%$!