Large Stepsize Gradient Descent for Non-Homogeneous Two-Layer Networks

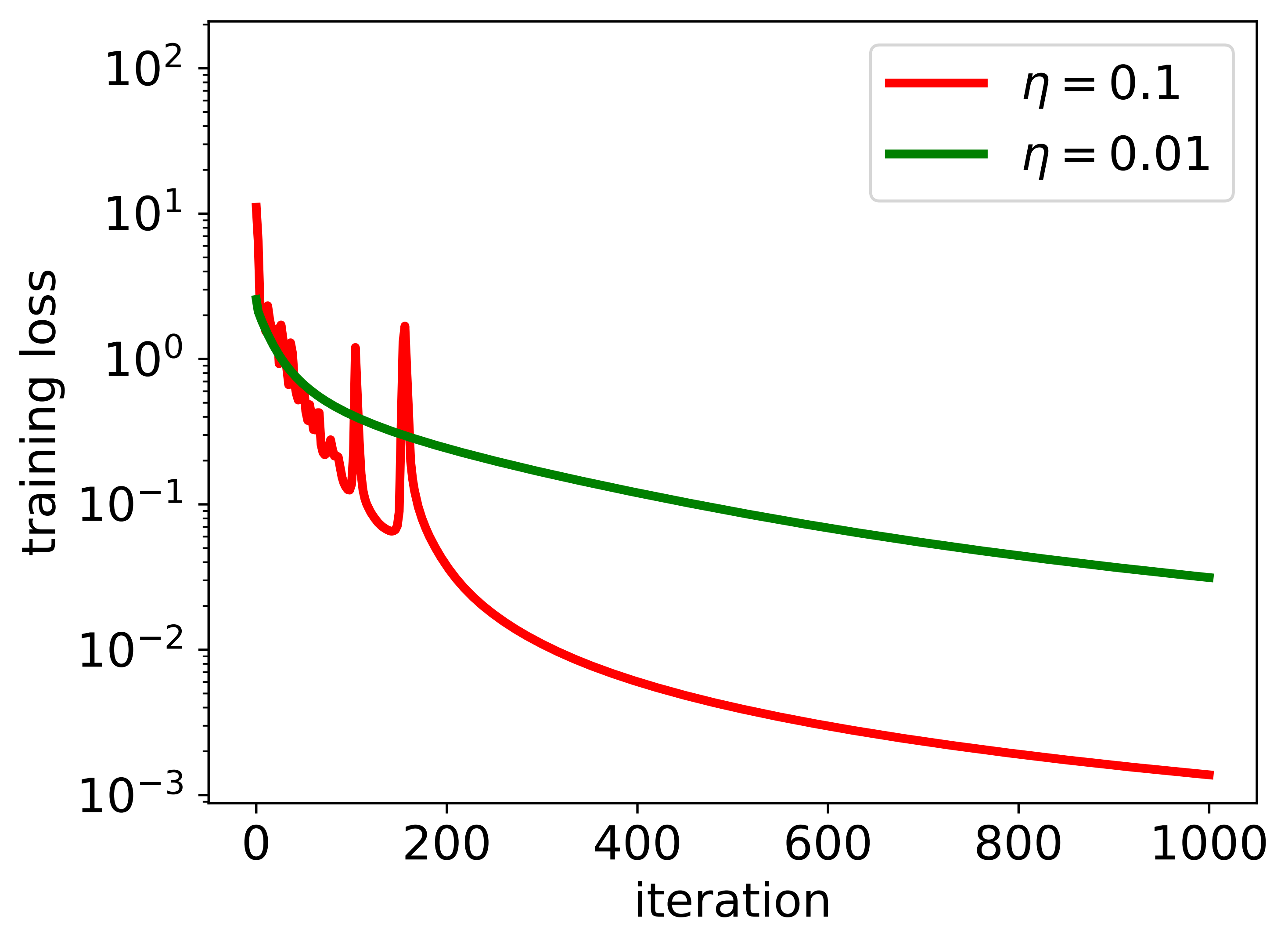
Margin Improvement and Fast Optimization

Yuhang Cai[1], Jingfeng Wu[1], Song Mei[1], Michael Lindsey[12], Peter Bartlett[13]

[1]UC Berkeley, [1]Lawrence Berkeley National Laboratory, [3]Google DeepMind

Background

- When training neural networks, **large stepsize** works better!
- "**Spikes**" or "**Edge of Stability**" unexplained by descent lemma.
- Implicit bias exists for **non-linear non-homogeneous models**!



3-layer net + 1,000 samples from MNIST

Setting

1. Binary classification data $\left(x_i,\ y_i \in \{\pm 1\}\right)_{i=1}^{n}$.
2. Logistic loss: $L(w) := \dfrac{1}{n} \sum_i \ln\left(1 + \exp(-y_i f(w; x_i))\right)$.
3. Gradient descent: $w_{t+1} = w_t - \eta \nabla_w L(w_t)$.

Stable phase and EoS phase

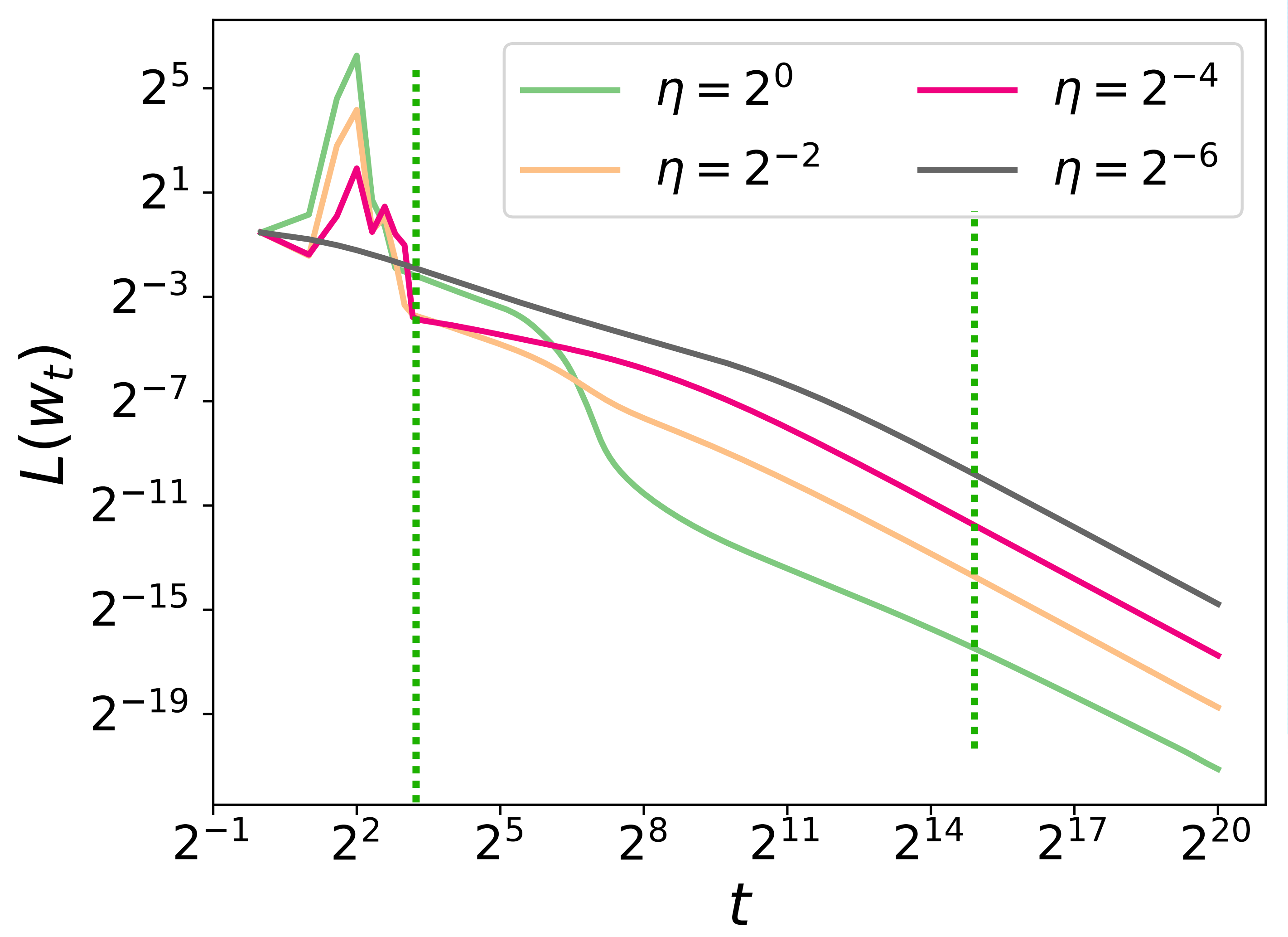
1. **EoS Phase.**

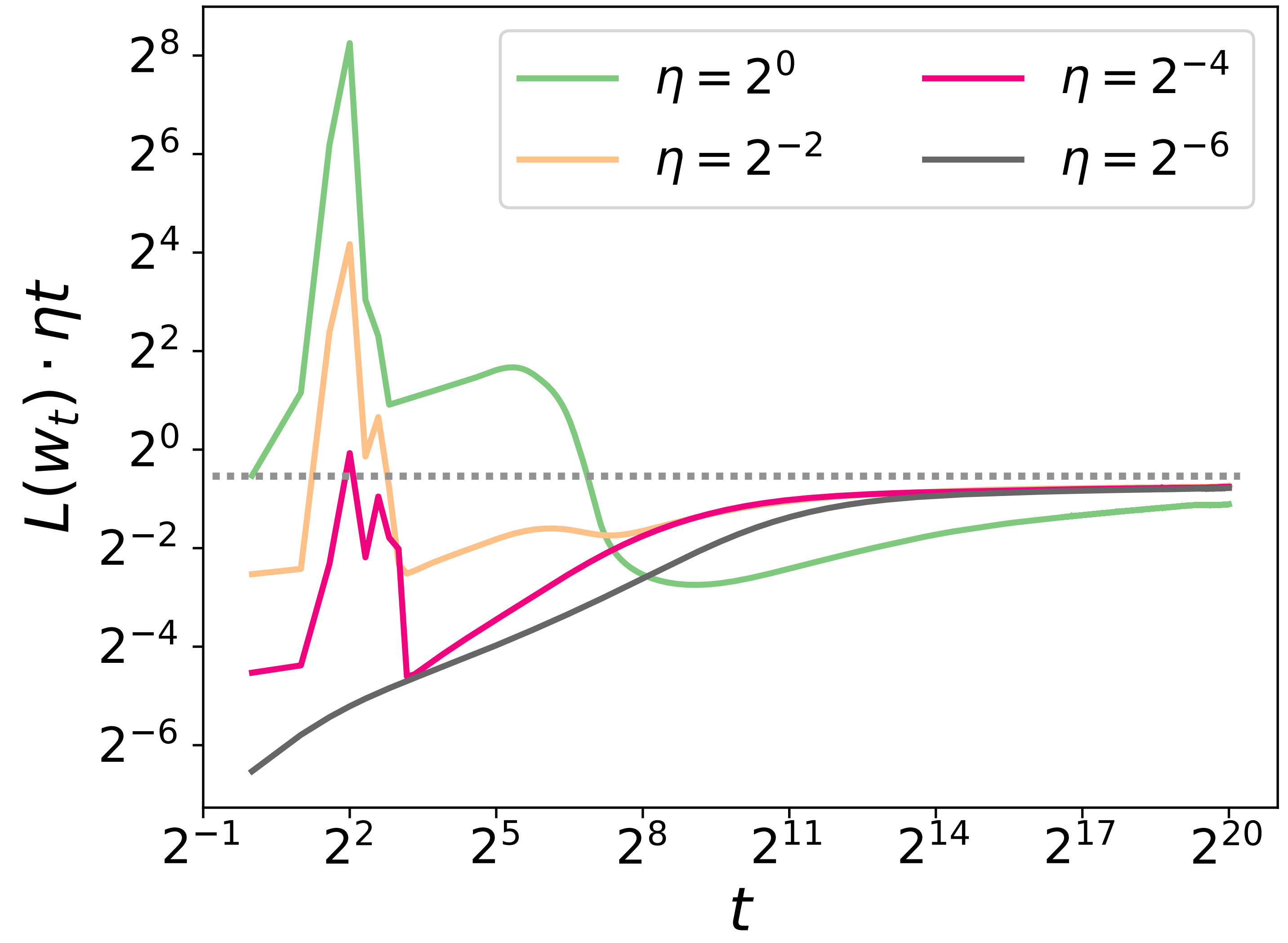
   - Loss oscillates but has a decreasing trend.

2. **Stable Phase.**
   - Loss monotonically decreases.
   - The parameter norm increases.
   - The parameter direction converges.
   - The normalized margin,

$$\bar{\gamma}(w) = \frac{\min_{i\in[n]} y_i f(w; x_i)}{\|w\|^M},$$

increases and stays positive.





A Theory for Non-homogeneous models

Stable phase

Near-homogeneous Models

- **Lipschitzness.** $\|\nabla_w f(w;x)\| \leq \rho$.
- **Smoothness**. $\|\nabla_w^2 f(w;x)\|_2 \leq \beta$.
- **Near-homogeneity**. $|\langle \nabla_w f(w;x), w\rangle - f(w;x)| \leq \kappa$.

Theorem 2.2 (Stable phase)

If $L(w_s) \leq \min\{1/2e^{\kappa+2}n, 1/(4\rho^2 + 2\beta)\eta\}$ for some s, then for $t \geq s$

- $L(w_t) = \Theta(1/t)$ decreases;
- $\|w_t\| = \Theta(\log t)$ increases;
- $\bar{\gamma}(w_t)$ stays positive and converges with a nearly increasing trend.

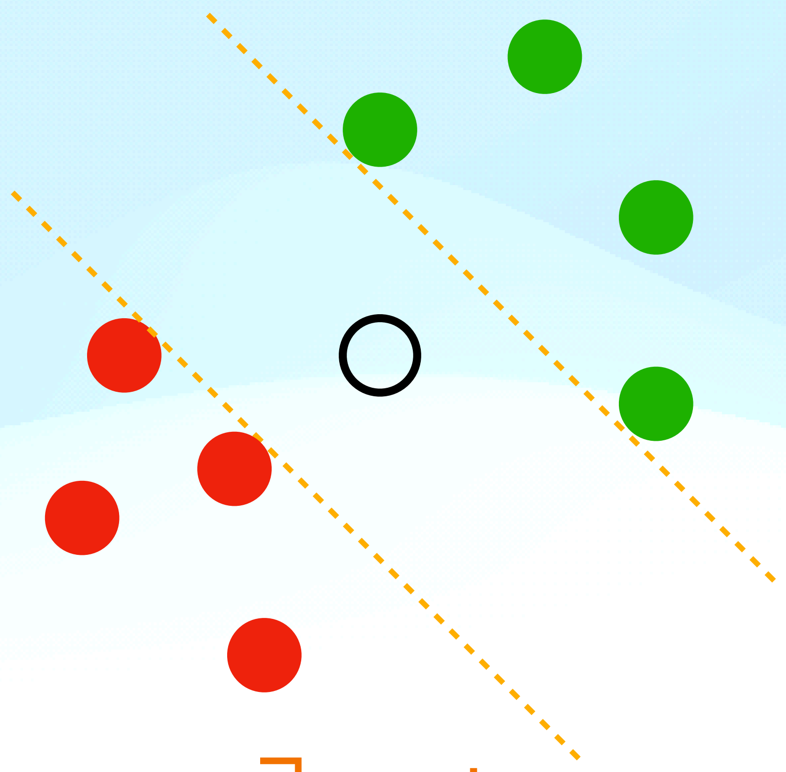
A Theory for Non-homogeneous models

EoS Phase

Two-layer Networks

$$f(w;x) = \frac{1}{m}\sum_{j=1}^{m} a_j \phi(x^T w^{(j)})$$

- **Lipschitzness.** $\alpha \leq \phi'(x) \leq 1.$
- **Near-homogeneity.** $|\phi'(x)x - \phi(x)| \leq \kappa.$



Assume: ∃ vector w_*
such that $yx^\top w_* > \gamma > 0$

Theorem 3.2 (EoS phase)

Given a two-layer NN. For every t, $\frac{1}{t}\sum_{k=0}^{t-1} L(w_k) \leq \tilde{O}\left(\frac{1+\eta^2}{\eta t}\right).$

A Theory for Non-homogeneous models

Phase transition

Theorem 4.1 (Phase transition)

For two-layer NNs, $L(w_s) \leq 1/\eta$ for
$s \leq \tau := \Theta\big(\max\{c_1\eta,\ c_2 n,\ c_2 n/\eta \ln(c_2 n/\eta)\}\big)$
Where $c_1 = 2e^{\kappa+2}, c_2 = (4\rho^2 + 2\beta)$.

Corollary 4.2 (Fast optimization)

For two-layer NNs, if $\eta = \Theta(T),$ then $L(T) = O(1/T^2)$.

Conclusion

Implicit bias with Near-homogeneity

We generalize the results for homogeneous models in [Lyu & Li 2020] to non-homogeneous models.

1. This includes **a broad class of activation functions!** Smooth Leaky ReLU, GELU, SiLU, Huberized ReLU etc..
2. Even with the non-homogeneous model, we show the **weak convergence of implicit bias.**
   The normalized margin converges!

Large stepsizes for non-linear model

We generalize the results for linear models in [Wu et al. 2024] to non-linear two-layer networks.

1. Asymptotic $\tilde{O}(1/\eta t)$ for **every** η (beyond 1/smoothness)
2. Given #steps $T \geq \Omega(n)$, if choose $\eta = \Theta(T)$, then

$$\tau \leq T/2 \ \text{ and } \ L(w_T) \leq \tilde{O}(1/T^2)$$

3. Theorem. In general, if not enter EoS, then $L(w_T) \geq \Omega(1/T)$

References

- Wu J, Bartlett P L, Telgarsky M, et al. Large Stepsize Gradient Descent for Logistic Loss: Non-Monotonicity of the Loss Improves Optimization Efficiency[J]. arXiv preprint arXiv:2402.15926, 2024.

- Lyu K, Li J. Gradient descent maximizes the margin of homogeneous neural networks[J]. arXiv preprint arXiv:1906.05890, 2019.

SCAN ME