# Large Stepsize GD for Logistic Loss

## Non-Monotonicity of the Loss Improves Optimization Efficiency

Jingfeng Wu[1], Peter Bartlett[1,3], Matus Telgarsky[2], Bin Yu[1]

[1]UC Berkeley, [2]New York University, [3]Google DeepMind

### Background

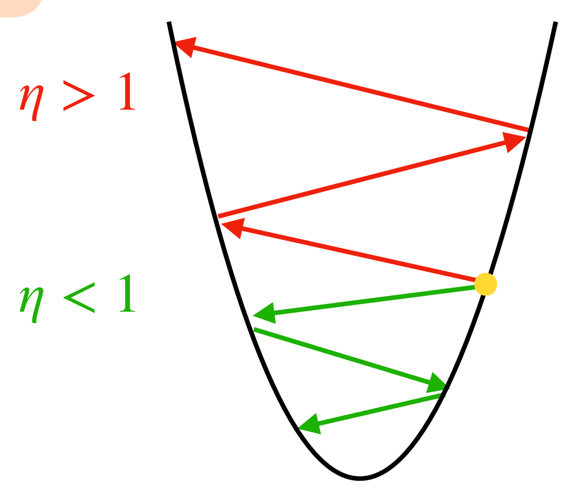
$$w_+ = w - \eta \nabla L(w)$$

How to choose stepsize?

#### Descent lemma

$L(w) = w^2$

$w_+ = (1-2\eta)w$




For **small** $\eta$, $L(w_t)$ decreases **monotonically**

For **large** $\eta$, $L(w_t)$ **diverges** for quadratics

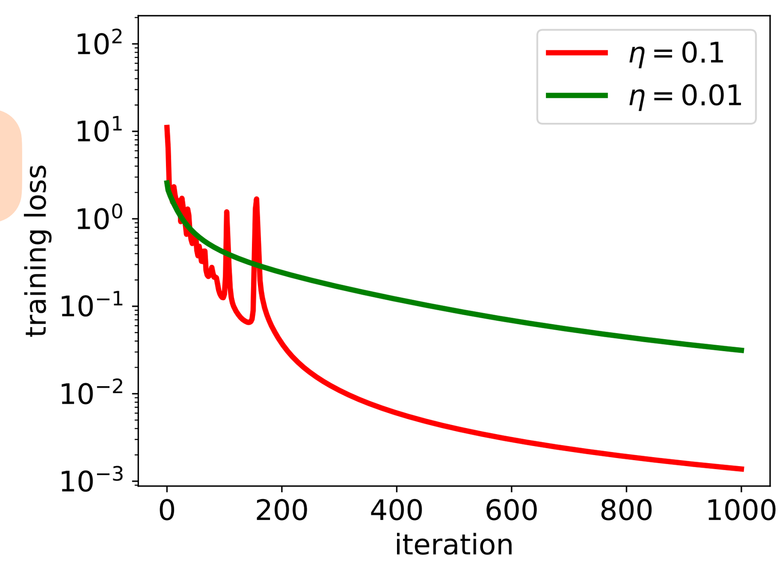
$$\begin{aligned} L(w_+) &= L(w - \eta \nabla L(w)) \\ &= L(w) - \eta \|\nabla L(w)\|^2 + \frac{\eta^2}{2} \nabla L(w)^\top \nabla^2 L(w) \nabla L(w) + O(\eta^3) \\ &\le L(w) - \eta \left(1 - \frac{\eta}{2}\|\nabla^2 L(w)\|_2\right) \|\nabla L(w)\|^2 + O(\eta^3) \end{aligned}$$

#### Edge of Stability (EoS)

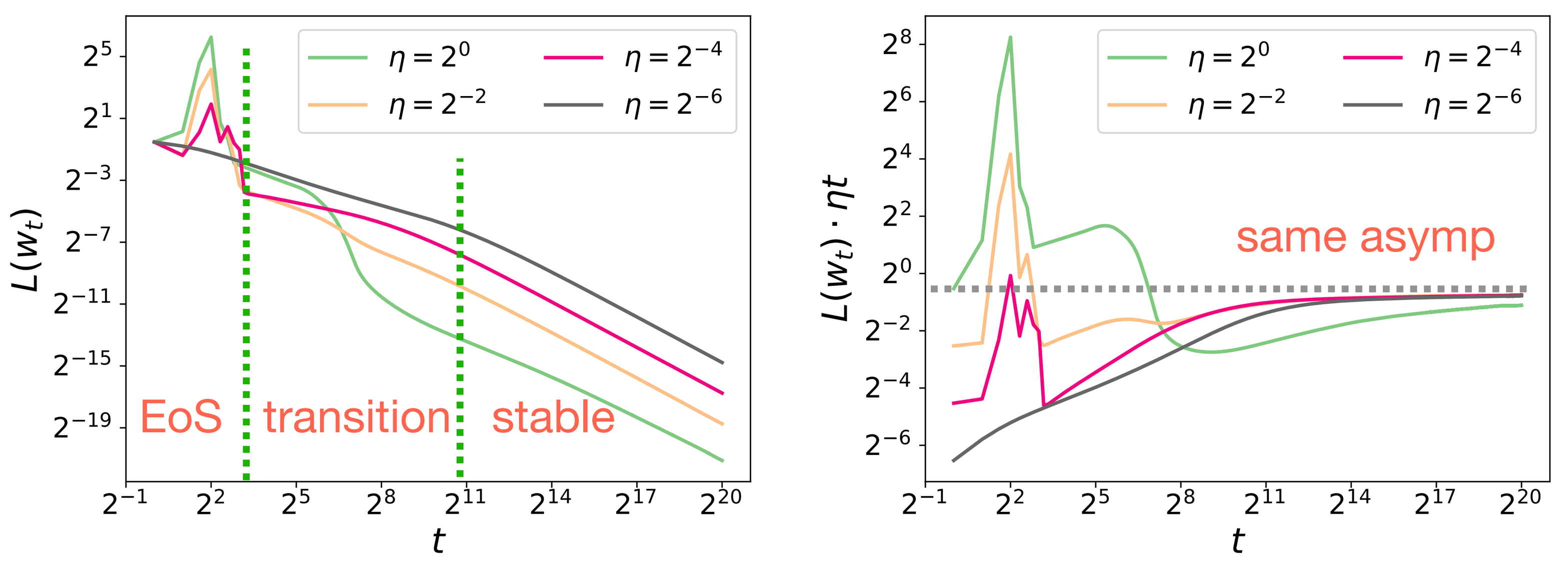
large stepsize works better;

"spikes" or "edge of stability"

unexplained by descent lemma




**3-layer net + 1,000 samples from MNIST**




**logistic regression + 1,000 samples from MNIST "0" or "8"**

### A EoS Theory in Logistic Regression

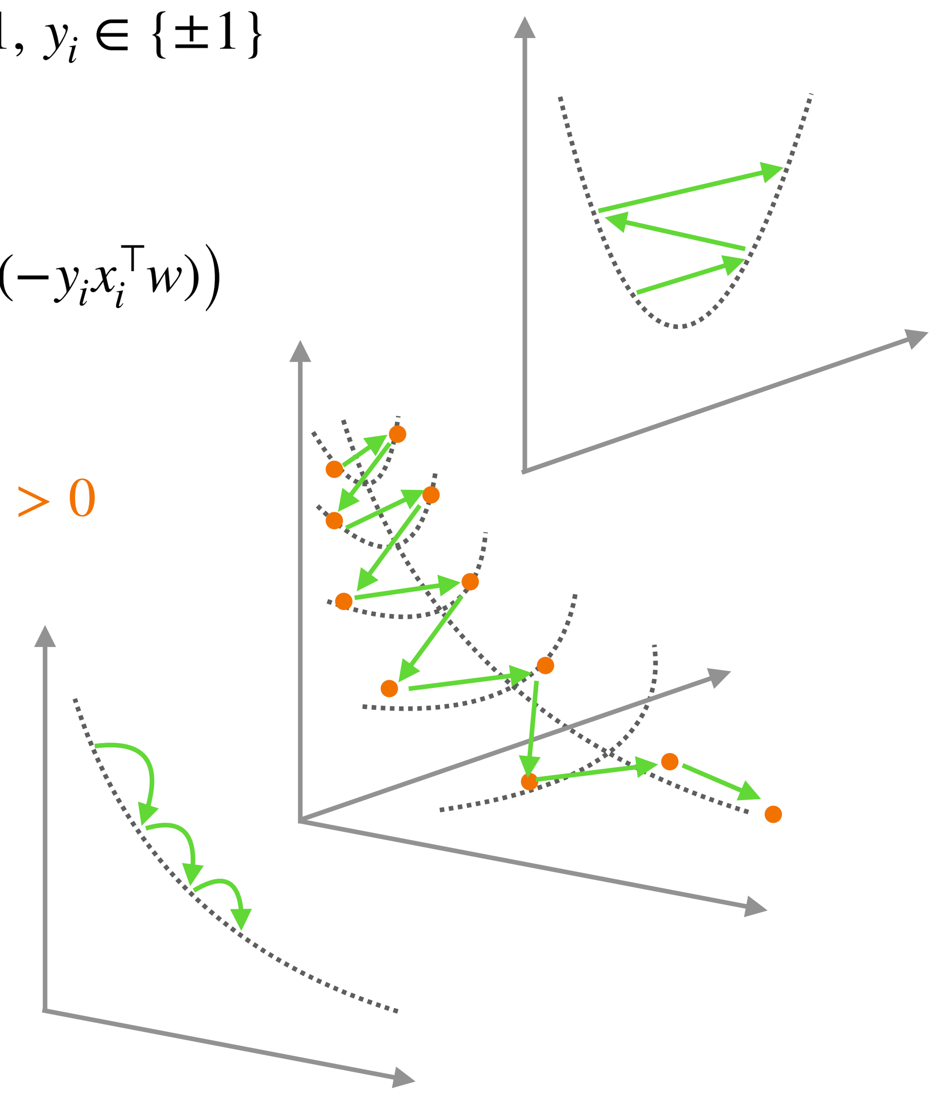
classification data $(x_i, y_i)_{i=1}^n$, $\|x_i\| \le 1$, $y_i \in \{\pm 1\}$

logistic loss + linear model

$$L(w) := \frac{1}{n}\sum_i \ln\left(1 + \exp(-y_i x_i^\top w)\right)$$

Assume [linear separability]:

$\exists$ vector $w_*$ such that $y x^\top w_* > \gamma > 0$



#### Theorem

- **EoS phase**. For every $t$
$$\frac{1}{t}\sum_{k=0}^{t-1} L(w_k) \le \tilde{O}\left(\frac{1+\eta^2}{\eta t}\right)$$
- **Stable phase**. If $L(w_s) \le 1/\eta$ for some $s$, then $L(w_{s+t}) \downarrow$ for $t \ge 0$ and
$$L(w_{s+t}) \le \tilde{O}\left(\frac{F(w_s)}{\eta t}\right), \quad F(w_s) := \hat{\mathbb{E}} \exp(-y x^\top w_s)$$
- **Phase transition**. We have $L(w_s) \le 1/\eta$ and $F(w_s) \le 1$ for
$$s \le \tau := \Theta\big(\max\{\eta,\ n,\ n/\eta \ln(n/\eta)\}\big)$$

#### Benefits of large stepsizes

1. Asymptotic $\tilde{O}(1/\eta t)$ for **every** $\eta$ (beyond 1/smoothness)
2. Larger $\eta$ => smaller const factor, but longer EoS
3. Given #steps $T \ge \Omega(n)$, if choose $\eta = \Theta(T)$, then
$$\tau \le T/2 \ \text{ and } \ L(w_T) \le \tilde{O}(1/T^2)$$

**"acceleration" by EoS**
**w/o momentum or varying stepsizes**

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

### Extensions — SGD — General Loss Functions — Neural Tangent Kernel

#### Theorem. SGD for logistic regression

Let $(w_k)_{k=1}^n$ be iterates of **const stepsize online SGD** for **logistic regression** on iid data from a **separable distribution**. Then for **every stepsize** $\eta$, w.p. $\ge 1-\delta$:

$$\frac{1}{n}\sum_{k=1}^n \mathbb{E}\ln\left(1+\exp(-y x^\top w_k)\right) \lesssim \frac{\ln^2(\gamma^2\eta n) + \eta^2}{\gamma^2 \eta n} + \frac{\left(\ln(\gamma^2\eta n)+\eta\right)\ln(1/\delta)}{\gamma n}$$

$$\frac{1}{n}\sum_{k=1}^n \Pr(y x^\top w_k \le 0) \lesssim \frac{\ln(\gamma^2 \eta n) + \eta}{\gamma^2 \eta n} + \frac{\ln(1/\delta)}{n}$$

**large stepsize works but no acceleration (upto log)**

#### A general loss function $\ell : \mathbb{R} \to \mathbb{R}_+$

A. **Regularity**. Assume $\ell$ is $\mathscr{C}^2$, convex, $\downarrow$, and $\ell(+\infty) = 0$,

define $\rho(\lambda) := \min_{z\in\mathbb{R}} \lambda \ell(z) + z^2, \quad \lambda \ge 1$ — minimizer away from init

B. **Lipschitzness**. Assume $g(\cdot) := |\ell'(\cdot)| \le C_g$ — prevents GD from diverging

C. **Self-boundedness**. Assume $g(\cdot) \le C_\beta \ell(\cdot)$ and — for entering stable phase

$\ell(z) \le \ell(x) + \ell'(z-x) + C_\beta g(x)(z-x)^2$, for $|z-x| \le 1$

D. **Exp-tail**. Assume $\ell(\cdot) \le C_e g(\cdot)$ — unnecessary but improves transition time

$$L(w) := \hat{\mathbb{E}}\ell\big(y f_x(w)\big), \quad f_x(w) := \frac{1}{\sqrt{m}}\sum_{s=1}^m a_s \max\{x^\top w^{(s)}, 0\}, \quad w \in \mathbb{R}^{md}$$

Assume NTK init: $w_0 \sim \mathcal{N}(0, I_{md})$, $(a_s)_{s=1}^m$ random from $\{\pm 1\}$ & fixed

Assume: "separable" in NTK RKHS — holds generically

#### Theorem. GD for NN/general losses

Assume $\ell$ satisfies A-B. Fix $T$, assume $m \ge \Omega(R^2)$ for $R := \Theta\big(\sqrt{\rho(\eta T)} + \eta\big)$.

- **Lazy training**. For $t \le T$, we have $\|w_t - w_0\| \le R$
- **EoS phase**. For $t \le T$, we have $\frac{1}{t}\sum_{k=0}^{t-1} L(w_k) \le O\left(\frac{\rho(\eta t) + \eta^2}{\eta t}\right)$
- **Stable phase**. Assume $\ell$ also satisfies C. If $L(w_s) \le \Theta(1/(\eta+n))$ for some $s$,
then $L(w_{s+t}) \downarrow$ and $L(w_{s+t}) \le O\left(\frac{\rho(\eta t)}{\eta t}\right), \ s+t \le T$
- **Phase transition**. We have $L(w_s) \le \Theta(1/(\eta+n))$ for some $s \le \tau$, where
$$\tau := \Theta\big(\max\{\psi^{-1}(\eta+n), \eta(\eta+n)\}\big), \quad \psi(\lambda) := \lambda/\rho(\lambda)$$
or $\tau := \Theta(\max\{\eta, n\ln(n)\})$ if $\ell$ also satisfies D

**large stepsize accelerates NTK & general losses**