Large Stepsize GD for Logistic Loss

Non-Monotonicity of the Loss Improves Optimization Efficiency

Jingfeng Wu[1], Peter Bartlett[13], Matus Telgarsky[2], Bin Yu[1]

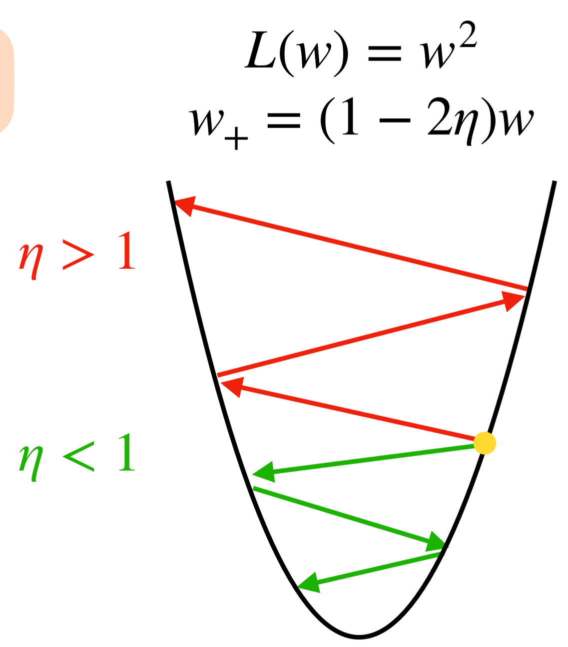
[1]UC Berkeley, [2]New York University, [3]Google DeepMind

Background

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

How to choose stepsize / learning rate?

Descent Lemma



For **small** η, $L(w_t)$ decreases **monotonically**

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

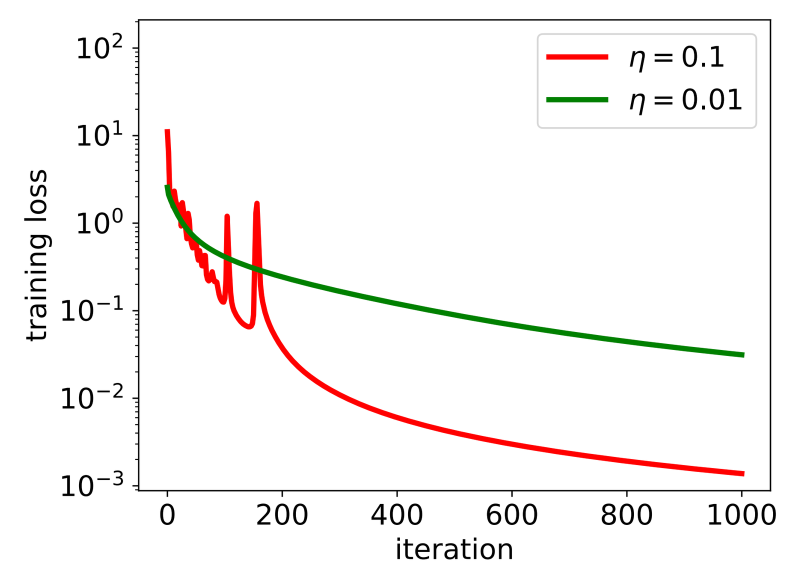
$$L(w_+) = L(w - \eta \nabla L(w))$$
$$= L(w) - \eta \|\nabla L(w)\|^2 + \frac{\eta^2}{2} \nabla L(w)^\top \nabla^2 L(w) \nabla L(w) - O(\eta^3)$$
$$\le L(w) - \eta\left(1 - \frac{\eta}{2}\|\nabla^2 L(w)\|_2\right)\|\nabla L(w)\|^2 - O(\eta^3)$$

Edge of Stability

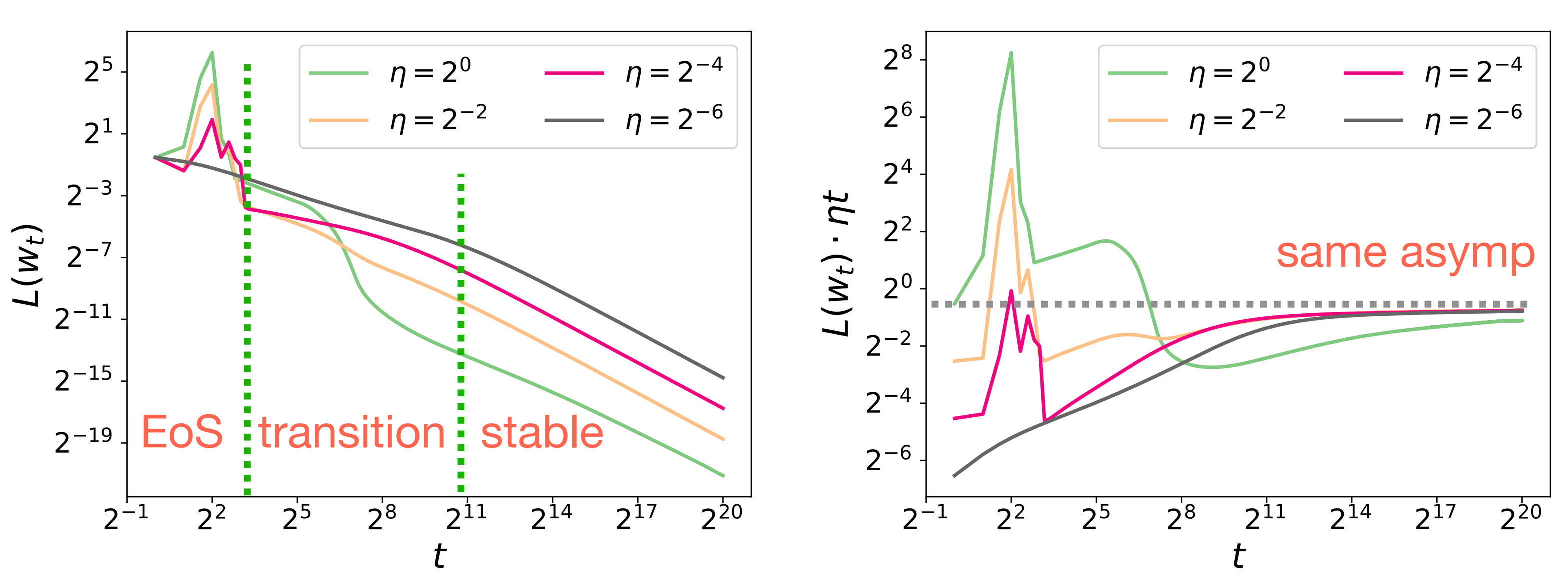
large stepsize works better

"spikes" or "edge of stability"

unexplained by descent lemma



3-layer net + 1,000 samples from MNIST



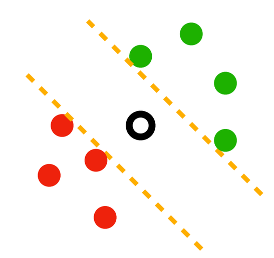
logistic regression + 1,000 samples from MNIST "0" or "8"

A Theory for EoS in Logistic Regression

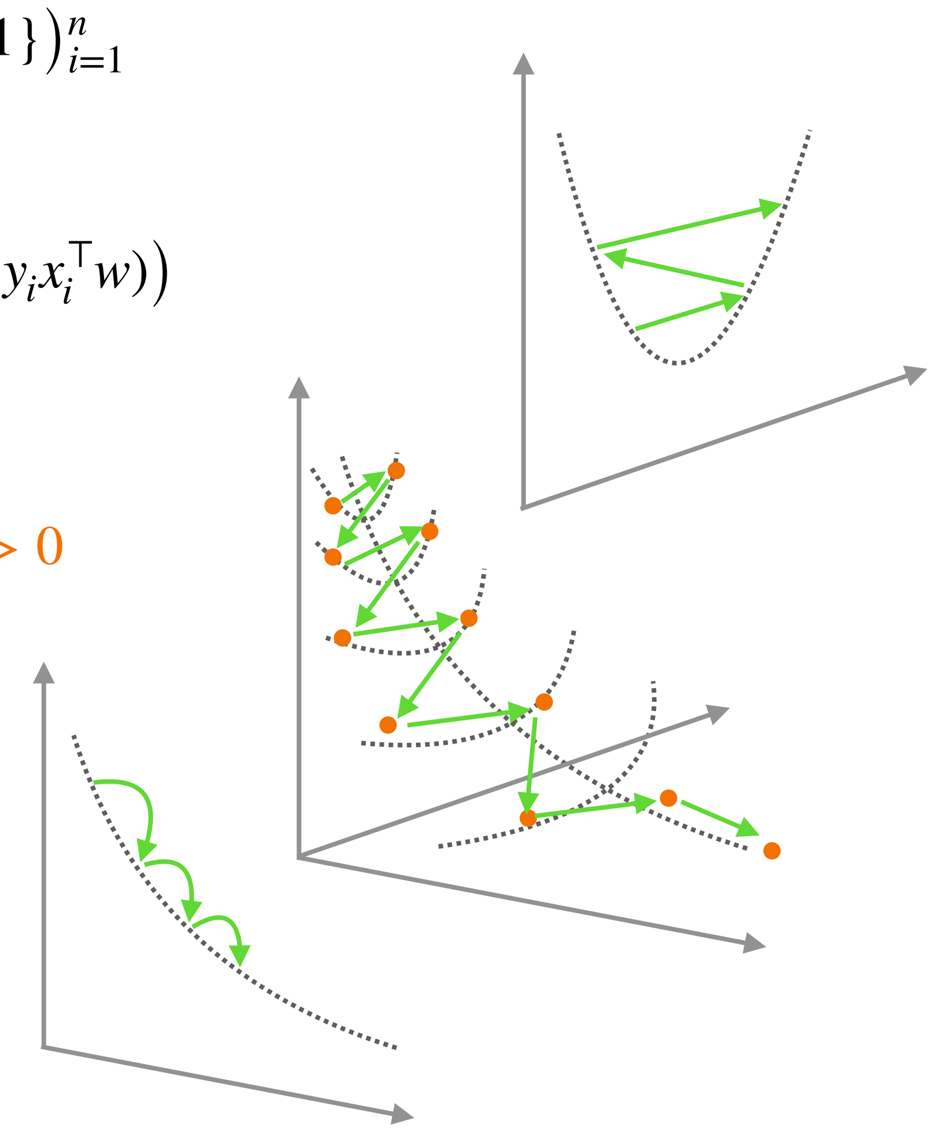
binary classification data $(x_i, y_i \in \{\pm 1\})_{i=1}^n$

logistic loss + linear model

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



Assume: $\exists$ 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
$$L(w_{s+t}) \le \tilde{O}\left(\frac{F(w_s)}{\eta t}\right), \quad F(w_s) := \hat{\mathbb{E}} \exp(-y x^\top w)$$
- **Phase transition**. We have $L(w_s) \le 1/\eta$ and $F(w_s) \le 1$ for
$$s \le \tau := \Theta\left(\max\{\eta, n, n/\eta \ln(n/\eta)\}\right)$$

Benefits of large stepsizes

1. Asymptotic $\tilde{O}(1/\eta t)$ for **every** η (beyond 1/smoothness)
2. Larger η => 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

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

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

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

B. **Lipschitzness**. Assume $g(\cdot) := |\ell'(\cdot)| \le C_g$

C. **Self-boundedness**. Assume $g(\cdot) \le C_\beta \ell(\cdot)$ and

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

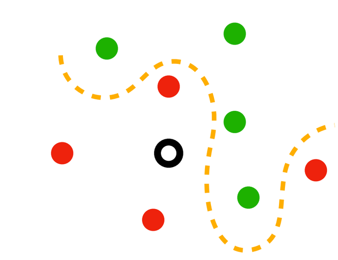
D. **Exp-tail**. Assume $\ell(\cdot) \le C_e g(\cdot)$

A two-layer network (kernel regime)

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

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



Theorem

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

- **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 ℓ also satisfies C. If $L(w_s) \le \Theta(1/(\eta + n))$ for some s, then
$$L(w_{s+t}) \downarrow \text{ and } L(w_{s+t}) \le O\left(\frac{\rho(\eta t)}{\eta t}\right), \quad s + t \le T$$
- **Phase transition**. We have $L(w_s) \le \Theta(1/(\eta + n))$ for $s \le \tau$, where
$$\tau := \Theta\left(\max\{\psi^{-1}(\eta + n), \eta(\eta + n)\}\right), \quad \psi(\lambda) := \lambda/\psi(\lambda)$$
or $\tau := \Theta(\max\{\eta, n \ln(n)\})$ if ℓ also satisfies D

Contributions: (1) EoS => faster optimization (2) open landscape (3) versatile techniques