Large Stepsizes Accelerate Gradient Descent for Regularized Logistic Regression

Jingfeng Wu*[1] Pierre Marion*[3] Peter Bartlett[12]

[1]UC Berkeley [2]Google DeepMind
[3]Inria, DI ENS, PSL University

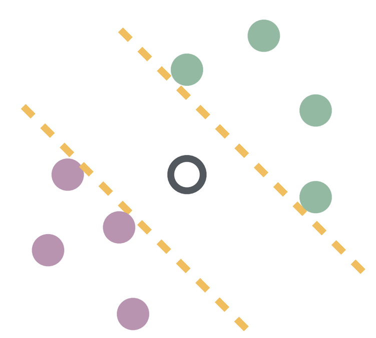
NEURAL INFORMATION PROCESSING SYSTEMS

Background

$$\tilde{L}(w) = L(w) + \frac{\lambda}{2}\|w\|^2 \quad L(w) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + e^{-y_i x_i^\top w}\right)$$

[Assumption (bounded + separable)]

- $\|x_i\| \le 1,\ y_i \in \{\pm 1\},\ i = 1,\ldots,n$
- $\exists$ unit vector w^*, $\min_i y_i x_i^\top w^* \ge \gamma = \Theta(1)$



typical case when overparameterized

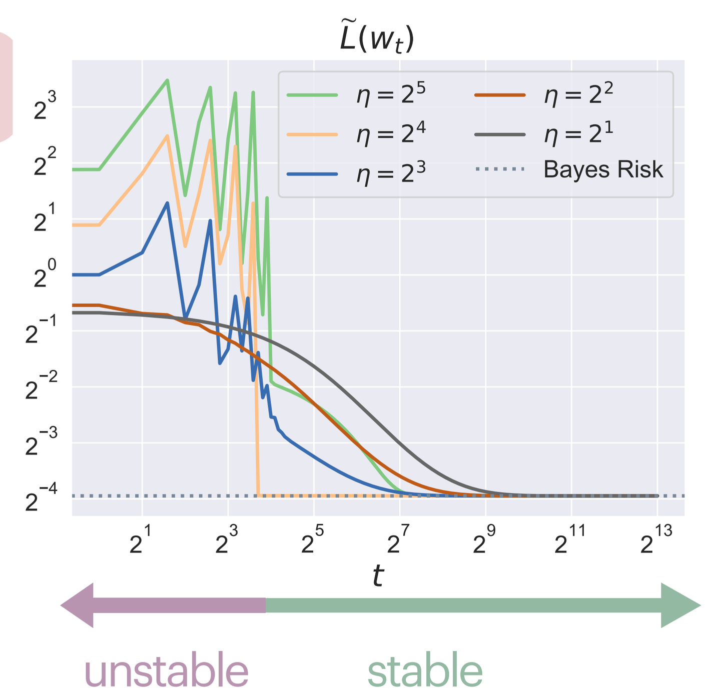
[Basic properties]

- objective $\tilde{L}$ is $\Theta(1)$-smooth and λ-strongly convex
- condition number is $\kappa = \Theta(1/\lambda)$
- minimizer $w_\lambda = \arg\min \tilde{L}(\cdot)$ is unique, $\|w_\lambda\| = \Theta(\ln(1/\lambda))$

Gradient descent

$$w_{t+1} = w_t - \eta \nabla \tilde{L}(w_t) \qquad w_0 = 0$$

large stepsize is
- unstable
- but faster



Prior results

[Without regularization, $\lambda = 0$; Wu, Bartlett, Telgarsky, Yu, 2024]

- For $\eta = \Theta(1)$, we have $L(w_t) \downarrow$ and $L(w_t) \le 1/t$
- For $T = \Omega(n)$ and $\eta = \Theta(T)$, we have $L(w_T) \le \tilde{O}(1/T^2)$

[With regularization; classical optimization theory]

For $\eta = \Theta(1)$, we have $\tilde{L}(w_t) \downarrow$ and $\tilde{L}(w_t) - \min \tilde{L} \le \epsilon$ for

$$t = O(\ln(1/\epsilon)/\lambda) = \tilde{O}(1/\lambda)$$

improved to $\tilde{O}(1/\lambda^{1/2})$ by Nesterov's momentum

Acceleration via large stepsizes

Small regularization

[Theorem] for small λ, large stepsize improves step complexity to $\tilde{O}(1/\lambda^{1/2})$

Assume separability and

$$\lambda \le \Theta\left(\frac{1}{n \ln n}\right) \qquad \eta \le \Theta\left(\min\left\{\frac{1}{\lambda^{1/2}}, \frac{1}{n\lambda}\right\}\right)$$

$\eta_{\max} = \Theta(1/\lambda^{1/2})$

Unstable phase. GD is *unstable* for at most τ steps for

$$\tau := \Theta\big(\max\{\eta, n, n/\eta \ln(n/\eta)\}\big)$$

$\tau = \Theta(\eta) \le \Theta(1/\lambda^{1/2})$

Stable phase. From τ and onward, $\tilde{L}(w_{\tau+t}) \downarrow$ and

$$\tilde{L}(w_{\tau+t}) - \min \tilde{L} \lesssim \exp(-\lambda \eta t)$$

General regularization

[Theorem] large stepsize improves step complexity to $\tilde{O}(1/\lambda^{2/3})$

Assume separability and

$$\lambda \le \Theta(1) \qquad \eta \le \Theta(1/\lambda^{1/3})$$

$\eta_{\max} = \Theta(1/\lambda^{1/3})$

Unstable phase. GD is *unstable* for at most τ steps for

$$\tau := \Theta(\eta^2)$$

$\tau \le \Theta(1/\lambda^{2/3})$

Stable phase. From τ and onward, $\tilde{L}(w_{\tau+t}) \downarrow$ and

$$\tilde{L}(w_{\tau+t}) - \min \tilde{L} \lesssim \exp(-\lambda \eta t)$$

A lower bound

[Theorem] small stepsize cannot accelerate

Fix $0 < \gamma < 0.1$ and consider a separable dataset

$$x_1 = (\gamma, 0.9) \quad x_2 = (\gamma, -0.5) \quad y_1 = y_2 = 1$$

For all $\lambda \lesssim 1$ and $\epsilon \lesssim \lambda \ln^2(1/\lambda)$, if η is such that $\tilde{L}(w_t) \downarrow$ for $t \ge 0$,

$$\text{then } \tilde{L}(w_t) - \min \tilde{L} \le \epsilon \ \Rightarrow\ t = \Omega\left(\frac{\ln(1/\epsilon)/\ln^2(1/\lambda)}{\lambda}\right)$$

Acceleration without overfitting

[Assumption] Let $(x_i, y_i)_{i=1}^n$ be iid copies of (x, y), where a.s.

- $\|x\| \le 1,\ y \in \{\pm 1\}$
- $\exists$ unit vector w^*, $yx^\top w^* \ge \gamma = \Theta(1)$

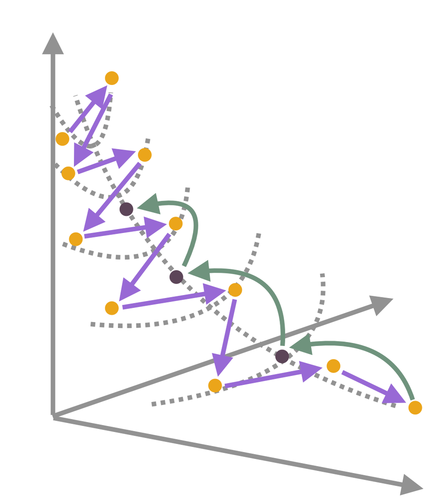
[Classical fast rate]

For the test error of any estimator $\hat{w}$, w.h.p.

data fitting vs estimator norm

$$L_{\text{test}}(\hat{w}) := \mathbb{E}\ln(1 + e^{-yx^\top \hat{w}}) \lesssim L(\hat{w}) + \tilde{O}(1)\frac{\max\{1, \|\hat{w}\|^2\}}{n}$$

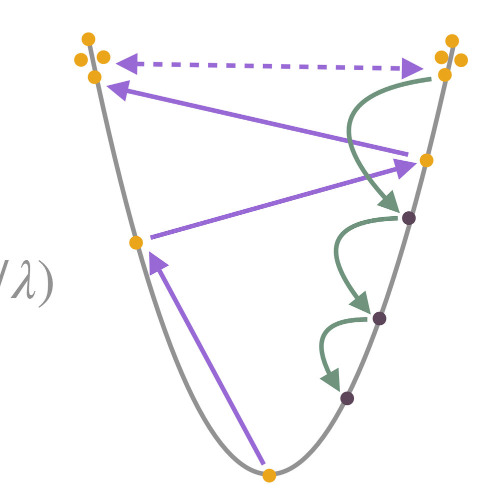
algorithms	λ	η	#steps to get 1/n test error
GD	0	$\Theta(1)$	$O(n)$
	$1/n$	1	$\tilde{O}(n)$
	$1/n$	$\Theta(n^{1/3})$	$\tilde{O}(n^{2/3})$
Nesterov	$1/n$	1	$\tilde{O}(n^{1/2})$



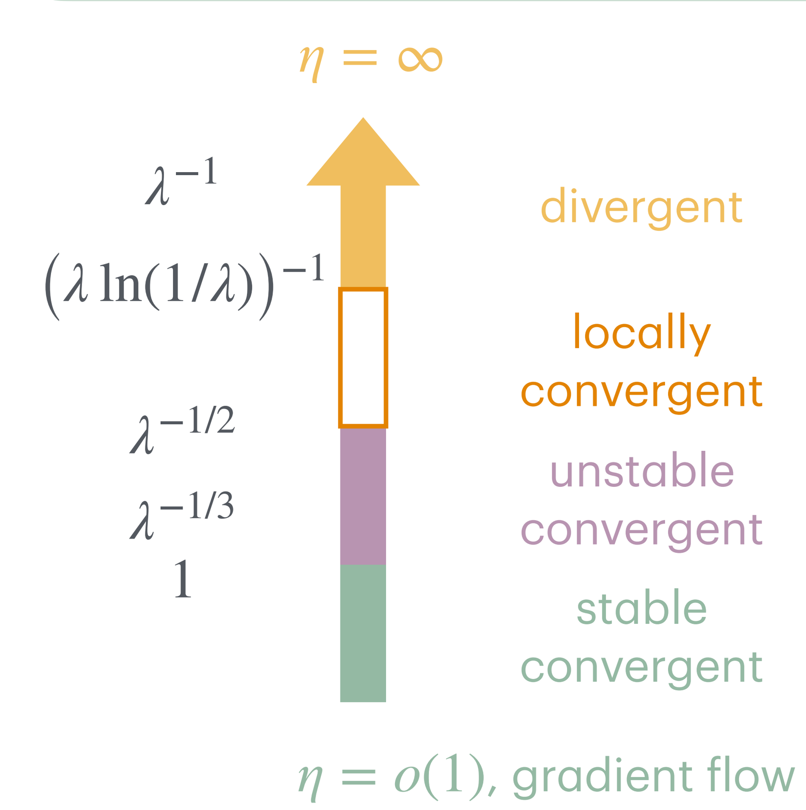
left: $L(w)$ right: $\frac{\lambda}{2}\|w\|^2$

unstable phase: overshoot
$\sup \|w_t\| = \Theta(\eta) = \mathrm{poly}(1/\lambda)$

stable phase: move back
$\|w_\lambda\| = \tilde{O}(1)$



Contribution & open problems



generic support vectors
unknown global behavior
match Nesterov
sample size independent