Minimax Optimal Convergence of Gradient Descent in Logistic Regression via Large and Adaptive Stepsizes

Ruiqi Zhang[1] Jingfeng Wu[1]
Licong Lin[1] Peter Bartlett[12]
[1]UC Berkeley [2]Google DeepMind

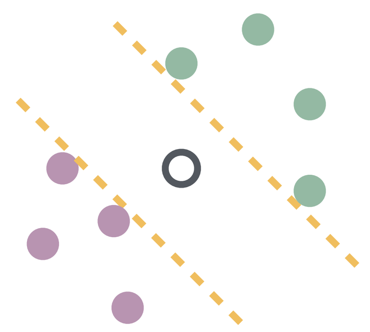
Berkeley UNIVERSITY OF CALIFORNIA

Background

$$L(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i x_i^\top w) \quad \ell(t) = \ln\big(1+\exp(-t)\big)$$

[Assumption (bounded + separable)]

- $\|x_i\| \le 1,\ y_i \in \{\pm 1\},\ i = 1,\dots,n$
- $\exists$ unit vector w^*, $\min_i y_i x_i^\top w^* \ge \gamma > 0$



Tasks

With a small number of GD steps,

1. minimize $L(w)$ up to ϵ error
2. find a linear separator, $\min_i y_i x_i^\top w > 0$

solving Task #1 with $\epsilon = \ln(2)/n$ solves Task #2

GD with a constant stepsize

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

[Ji & Telgarsky, 2018]

For $\eta = \Theta(1)$, we have $L(w_t) \downarrow$ and $L(w_t) \le \tilde{O}(1/t)$

[Wu et al, 2024] acceleration via unstable convergence

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

GD with (small) adaptive stepsizes

observe that $\|\nabla^2 L\| \le L$

$$w_{t+1} = w_t - \eta\big((-\ell^{-1})' \circ L(w_t)\big)\nabla L(w_t)$$

$$\approx w_t - \frac{\eta}{L(w_t)}\nabla L(w_t)$$

$$w_{t+1} = w_t - \eta \nabla \phi(w_t) \qquad \phi(w) = -\ell^{-1}(L(w))$$

$$\approx \ln \sum \exp(-y_i x_i^\top w)$$

[Ji & Telgarsky, 2021]

For $\eta = \Theta(1)$, we have $L(w_t) \downarrow$ and $L(w_t) \le \exp(-\Theta(t))$

Main results

Large adaptive stepsizes

[Theorem]

For $t > 1/\gamma^2$, we have

$$L(\bar{w}_t) \le \exp\big(-\Theta(\gamma^2 \eta t)\big), \quad \text{where } \bar{w}_t = \frac{1}{t}\sum_{k=1}^{t} w_k$$

after $1/\gamma^2$ burn-in steps, adaptive GD is arbitrarily fast as $\eta \to \infty$

averaging is needed, b/c $L(w_t)$ oscillates for large η

not always true if $L(w_t)$ is monotone => small η

[Theorem]

Fix $w_0 = 0$ and $0 < \gamma < 0.1$. Consider dataset

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

If the hyperparameter η for adaptive GD is such that $L(w_t) \downarrow$, then there is c that only depends on γ, such that

$$L(\bar{w}_t), L(w_t) \ge \exp(-ct)$$

A minimax lower bound

[Definition]

First-order batch method:

$$w_t \in w_0 + \mathrm{span}\{\nabla L(w_0), \dots, \nabla L(w_{t-1})\}$$

where $L(w) = \hat{\mathbb{E}}\ell(yx^\top w)$ for any ℓ

[Theorem]

$\forall w_0$, $\exists (x_i, y_i)_{i=1}^n$ with margin γ such that: for any first-order batch method, we have

$$\min_i y_i x_i^\top w_t > 0 \ \Rightarrow\ t \ge \Omega\big(\min\{1/\gamma^2, \ln n\}\big)$$

$$\Rightarrow\ t \ge \Omega(1/\gamma^2) \text{ when } n \text{ is large}$$

Other results

A step complexity comparison

\# steps needed by batch methods to find a linear separator (by achieving $L(w) < \ln(2)/n$)

(batch) methods	#steps
const-stepsize GD [J & T 2018]	$\tilde{O}(n/\gamma^2)$
small-stepsize adaptive GD [J & T, 2021]	$O(\ln(n)/\gamma^2)$
dual momentum [Ji et al, 2021]	$O\big(\sqrt{\ln(n)\ln\ln(n)}/\gamma\big)$
large-stepsize adaptive GD	$1/\gamma^2$
minimax lower bound	$\Omega(\min\{1/\gamma^2, \ln n\})$

For $n = \exp\big(\Omega(1/\gamma^2)\big)$,

- GD with large, adaptive stepsizes is minimax optimal
- other methods are strictly suboptimal
- *Perceptron*, an online method, also takes $1/\gamma^2$ steps

For $n = \exp\big(O(1/\gamma^2)\big)$, what's the correct trade-off between γ and n?

Extensions

Similar results hold for

- Two-layer networks w/ leaky ReLU, fixed outer layer, separable data
- Liner predictors w/ other loss functions

Key: transformed objective $\phi(\cdot)$ needs to be convex and Lipschitz

References

Ji & Telgarsky. "Risk and parameter convergence of logistic regression." COLT 2018

Ji, Srebro, Telgarsky. "Fast margin maximization via dual acceleration." ICML 2021

Ji & Telgarsky. "Characterizing the implicit bias via a primal-dual analysis." ALT 2021

Wu, Bartlett, Telgarsky, and Yu. "Large stepsize gradient descent for logistic loss: non-monotonicity of the loss improves optimization efficiency." COLT 2024