Stability and Convergence Tradeoff of Iterative Optimization Algorithms

Yuansi Chen[1], Chi Jin[2], Bin Yu[1,2]
[1]Department of Statistics, [2]Departmenet of EECS, University of California, Berkeley

Motivation

• Stability [1,2,3,4]: a "good" algorithm should not change its solution much if we modify training set slightly.
• Stability implies generalization : Unlike complexity based approaches, stability is an algorithmically feasible sanity check for generalization.
• [5] recently showed that fixed step-size stochastic gradient descent (SGD) has uniform stability, linearly dependent of iteration.
• We would like to provide a complete picture
 * stability can be estabilished for a wide range of iterative optimization algorithms
 * stability constrains the optimal convergence rate of optimizaiton algorithms

Generalization ⟷ Finite complexity

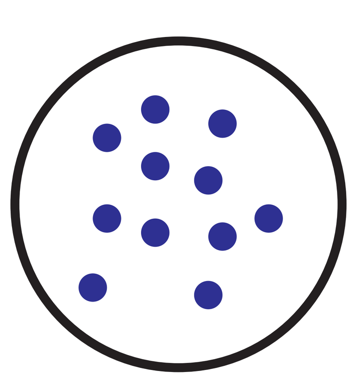
Stability

Algorithmic Stability

Uniform stability:

$$\epsilon_{\text{stab}} = \sup_{\substack{S,S' \\ \text{differ in one sample}}} \sup_{z} |l(\theta_S; z) - l(\theta_{S'}, z)|$$

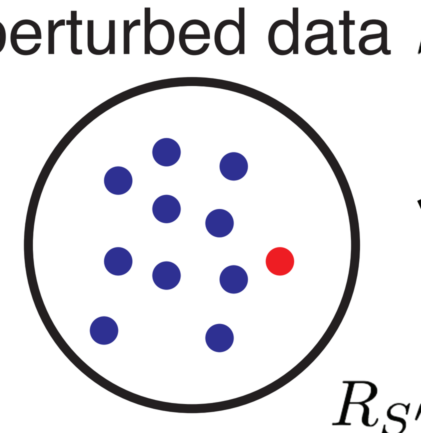
data S



$$R_S(\theta) = \frac{1}{n}\sum_{j=1}^{n} l(\theta; z_j)$$

Algorithm → θ_S, $\theta_{S'}$?

perturbed data S'



$$R_{S'}(\theta) = \left[\frac{1}{n}\sum_{i=1, i\neq k}^{n} l(\theta; z_j)\right] + \frac{1}{n} l(\theta; z'_k)$$

Stability and Convergence Tradeoff

$$\underbrace{R(\theta)}_{\epsilon_{\text{pop}}} = \underbrace{(R(\theta) - R_S(\theta))}_{\epsilon_{\text{gen}}} + \underbrace{R_S(\theta)}_{\epsilon_{\text{opt}}}$$

$$\mathbb{E}_S[R(\hat{\theta}_S)] = \mathbb{E}_S[\epsilon_{\text{gen}}] + \mathbb{E}_S[\epsilon_{\text{opt}}]$$
$$\leq \epsilon_{\text{stab}} + \mathbb{E}_S[\epsilon_{\text{opt}}]$$

A too stable algorithm can not converge too fast!

A lower bound on population risk combined with an good upper bound on stability, imples a lower bound on the optimal convergence rate.

Convex Smooth Loss

The loss function is convex, L-Lipschitz, β-smooth.
Le Cam's method for risk lower bound:

$$\mathbb{E}_S[R(\hat{\theta}_S)] \geq \frac{C}{\sqrt{n}}$$

Consequence:

$$\underbrace{\mathbb{E}_S[R(\hat{\theta}_S)]}_{O(\frac{1}{\sqrt{n}})} \leq \underbrace{\epsilon_{\text{stab}}}_{O(\frac{h(T)}{n})} + \underbrace{\mathbb{E}_S[\epsilon_{\text{opt}}]}_{?}$$

Gradient Descent's Stability

$$\theta_{t+1} = \theta_t - \eta \nabla R_S(\theta_t).$$

The stability of gradient descent for convex smooth objective mainly relies on its contracting property.

$$\|\theta - \eta\nabla R_S(\theta) - [\theta' - \eta \nabla R_S(\theta')]\| \leq \|\theta - \theta'\|$$

The error term caused by the data perturbation accumulates linearly as a function of iteration.

$$\frac{2\eta L^2 T}{n} \text{ stability} \Rightarrow O(\tfrac{1}{T}) \text{ convergence rate (optimal)}$$

Descreasing Step-size SGD's Stability

$$\theta_{t+1} = \theta_t - \eta \nabla l_{i_t}(\theta_t)$$
$$\eta = O(t^{-\alpha}), \alpha \in (2/3, 1)$$

$$O(\tfrac{T^{1-\alpha}}{n}) \text{ stability} \Rightarrow O(T^{-\alpha}) \text{ convergence rate (optimal)}$$

Nesterov Accelerated Gradient's Stability

$$w_{t+1} = \theta_t - \eta \nabla R_S(\theta_t)$$
$$\theta_{t+1} = (1 - \gamma_t) w_{t+1} + \gamma_t w_t$$

Contracting property of gradient descent is not enough to guarantee the stability!

• 1D or quadratic objective: (modified lower bound)

$$\frac{12\eta L^2 T^2}{n} \text{ stability} \Rightarrow O(\tfrac{1}{T^2}) \text{ convergence rate (optimal)}$$

• General case: Not clear yet.

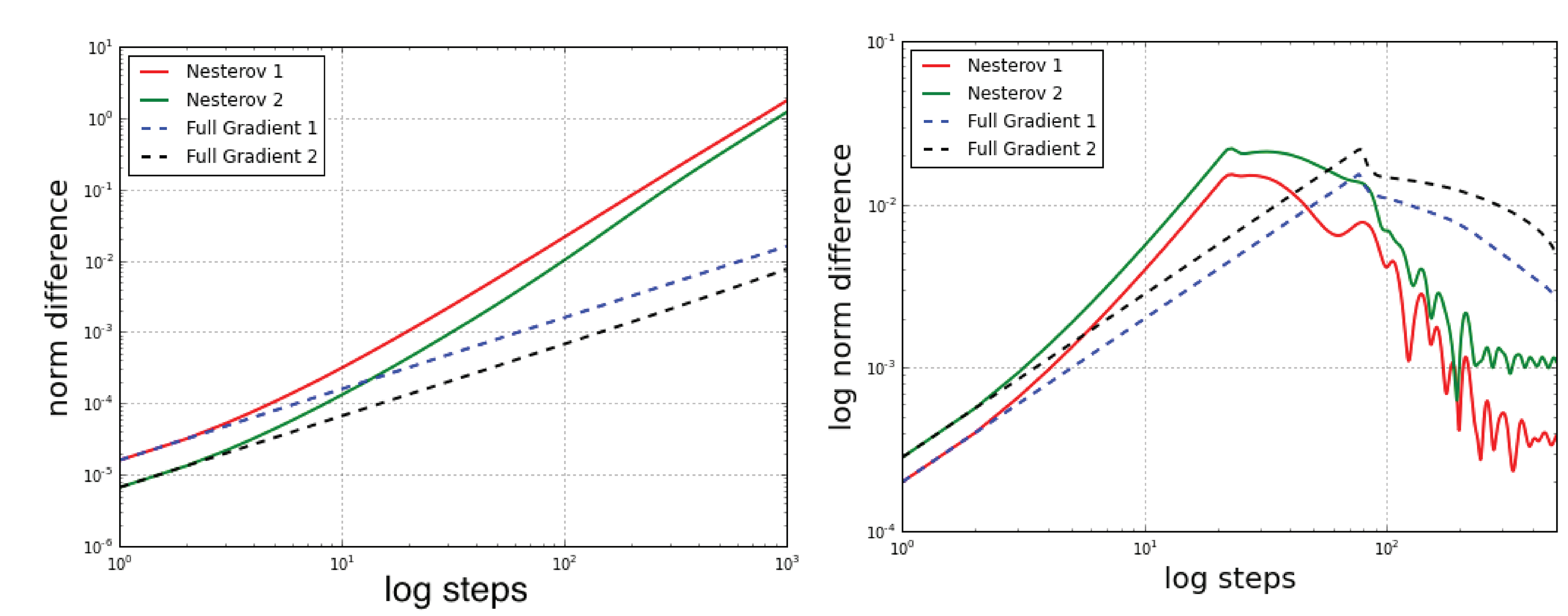
Strongly Convex Smooth Loss

All previous stability bounds hold. However, these bound might not be good enough, as an algorithm independent stability upper bound exists.

$$\epsilon_{\text{stab}} = \frac{2L^2}{\gamma n}$$

In the strongly convex setting, the population lower bound is of the same order. The tradeoff is not as notable as in convex smooth case.

Simulations



References

[1] Rogers, William H., and Terry J. Wagner. "A finite sample distribution-free performance bound for local discrimination rules." The Annals of Statistics (1978).
[2] Devroye, Luc, and T. Wagner. "Distribution-free performance bounds for potential function rules." IEEE Transactions on Information Theory 25.5 (1979): 601-604.
[3] Bousquet, Olivier, and André Elisseeff. "Stability and generalization." Journal of Machine Learning Research 2.Mar (2002): 499-526.
[4] Yu, Bin. "Stability." Bernoulli 19.4 (2013): 1484-1500.
[5] Hardt, Moritz, Benjamin Recht, and Yoram Singer. "Train faster, generalize better: Stability of stochastic gradient descent." ICML (2016).