

COSC 594/650

Robust stochastic gradient

Dongsheng Ding

dongshed@utk.edu

Department of Electrical Engineering and Computer Science
University of Tennessee, Knoxville

Recap: stochastic gradient

Population problem: $\underset{\theta}{\text{minimize}} \quad L(\theta) = \mathbb{E}_{\xi} [\ell(\xi, \theta)]$

$$\theta_{k+1} = \theta_k - \alpha_k \nabla \ell(\xi_k, \theta_k), \quad \alpha_k = \frac{1}{\sigma k}$$

■ GUARANTEE

$$\mathbb{E} \|\theta_k - \theta^*\|^2 \leq \frac{D}{k}, \quad D = \max \left\{ \|\theta_1 - \theta^*\|^2, \frac{2M^2}{\sigma^2} \right\}$$

■ TWO DRAWBACKS

- ★ L must be σ -strongly convex
- ★ σ must be known – a wrong σ changes the rate, not just the constant

Fast with the right σ , and fragile without it

QUESTION: Can we drop strong convexity, and the modulus σ ?

Robust stochastic approximation

Convexity: L is convex – no σ

Bounded second moment: $\mathbb{E}\|\nabla\ell(\xi_k, \theta_k)\|^2 \leq M^2$ at every iterate

$$\theta_{k+1} = \theta_k - \alpha_k \nabla\ell(\xi_k, \theta_k), \quad \bar{\theta}_K = \frac{\sum_{k=1}^K \alpha_k \theta_k}{\sum_{k=1}^K \alpha_k}$$

■ TWO CHANGES

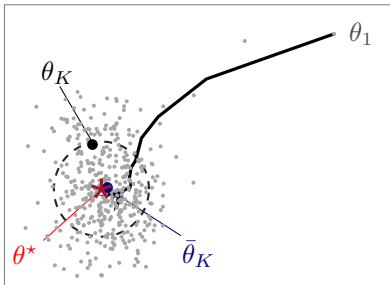
- ★ longer stepsize – $\alpha_k = \gamma/(M\sqrt{k})$ for any $\gamma > 0$, instead of $1/(\sigma k)$
- ★ report the weighted average $\bar{\theta}_K$, not the last iterate

Longer steps, and an **average** of the iterates

Why average the iterates

Example: $L(\theta) = \frac{1}{2} \theta^\top A \theta$, $A = \text{diag}(1, 0.4)$

Iterates: noisy gradients, constant step, $K = 800$



■ GRAY: THE ITERATES

★ 0.49 from θ^* (dashed circle)

■ BLUE: THEIR AVERAGE

★ $\bar{\theta}_K$ ends 0.07 from θ^*

The last iterate carries the noise, and the average cancels it

Convergence analysis: one step

■ FROM LECTURE 19, W/ $d_k = \mathbb{E} \|\theta_k - \theta^*\|^2$

$$d_{k+1} \leq d_k - 2\alpha_k \mathbb{E} \left[(\theta_k - \theta^*)^\top \nabla L(\theta_k) \right] + \alpha_k^2 M^2$$

★ the same expansion, after conditioning on the past

■ USE CONVEXITY ONCE

$$\begin{aligned} (\theta_k - \theta^*)^\top \nabla L(\theta_k) &\geq L(\theta_k) - L^* \\ \implies 2\alpha_k \mathbb{E} [L(\theta_k) - L^*] &\leq d_k - d_{k+1} + \alpha_k^2 M^2 \end{aligned}$$

Convexity alone turns the cross term into the optimality gap

Convergence analysis: sum over the iterations

■ ADD THE INEQUALITIES FOR $k = 1, \dots, K$

$$2 \sum_{k=1}^K \alpha_k \mathbb{E} [L(\theta_k) - L^*] \leq d_1 - d_{K+1} + M^2 \sum_{k=1}^K \alpha_k^2 \leq d_1 + M^2 \sum_{k=1}^K \alpha_k^2$$

★ the distances telescope, and $d_{K+1} \geq 0$ is dropped

■ USE CONVEXITY AGAIN

$$L(\bar{\theta}_K) \leq \frac{\sum_{k=1}^K \alpha_k L(\theta_k)}{\sum_{k=1}^K \alpha_k}$$

$$\mathbb{E} [L(\bar{\theta}_K)] - L^* \leq \frac{d_1 + M^2 \sum_{k=1}^K \alpha_k^2}{2 \sum_{k=1}^K \alpha_k} \quad \text{— one bound for every step size}$$

Rate with decreasing stepsize

Step size: $\alpha_k = \gamma/(M\sqrt{k})$, any $\gamma > 0$

■ TWO SUMS

$$\sum_{k=1}^K \frac{1}{k} \leq 1 + \log K, \quad \sum_{k=1}^K \frac{1}{\sqrt{k}} \geq 2(\sqrt{K+1} - 1)$$

■ PLUG INTO THE BOUND

$$\mathbb{E}[L(\bar{\theta}_K)] - L^* \leq \frac{M(d_1/\gamma + \gamma(1 + \log K))}{4(\sqrt{K+1} - 1)} = O\left(\frac{\log K}{\sqrt{K}}\right)$$

For every convex L , and with no σ

Robust to the stepsize

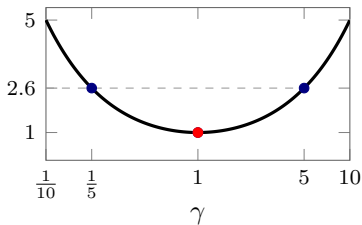
Fixed budget: K steps, constant $\alpha = \gamma R / (M\sqrt{K})$, $R \geq \|\theta_1 - \theta^*\|$

■ THE SAME BOUND

$$\mathbb{E}[L(\bar{\theta}_K)] - L^* \leq \frac{R^2 + KM^2\alpha^2}{2K\alpha} = \frac{1}{2} \left(\gamma + \frac{1}{\gamma} \right) \frac{RM}{\sqrt{K}}$$

■ A WRONG STEP

- ★ $\gamma = 1 - RM/\sqrt{K}$, no $\log K$
- ★ $\gamma = 5$ or $1/5$ – 2.6 times larger



A wrong step costs a constant factor, and not the rate

Two ways to choose the stepsize

	Stochastic gradient	Robust stochastic gradient
Assumes	L σ -strongly convex	L convex
Stepsize	$1/(\sigma k)$	$\gamma R/(M\sqrt{K})$
Output	last iterate θ_k	average $\bar{\theta}_K$
Guarantee	$\mathbb{E}\ \theta_k - \theta^*\ ^2 \leq D/k$	$\mathbb{E}[L(\bar{\theta}_K)] - L^* \leq \frac{1}{2} \left(\gamma + \frac{1}{\gamma} \right) \frac{RM}{\sqrt{K}}$
Wrong constant	changes the rate	changes the constant

Slower in the best case, and safe in every case

Empirical risk minimization is stochastic optimization

Empirical risk: $\hat{L}_m(\theta) = \frac{1}{m} \sum_{i=1}^m \ell(\xi_i, \theta)$

■ AN EXPECTATION OVER THE DATA

$$\hat{L}_m(\theta) = \mathbb{E}_i [\ell(\xi_i, \theta)], \quad i \text{ uniform on } \{1, \dots, m\}$$

★ the empirical distribution puts mass $1/m$ on each sample

■ SO STOCHASTIC GRADIENT APPLIES

$$\theta_{k+1} = \theta_k - \alpha_k \nabla \ell(\xi_{i_k}, \theta_k), \quad i_k \text{ uniform on } \{1, \dots, m\}$$

Stochastic gradient for ERM: sample a data point

Incremental gradient method

Finite sum: $\underset{\theta}{\text{minimize}} \quad f(\theta) = \frac{1}{m} \sum_{i=1}^m \ell(\xi_i, \theta)$

$$\theta_{k+1} = \theta_k - \alpha_k \nabla \ell(\xi_{i_k}, \theta_k)$$

■ HOW TO PICK i_k

- ★ cyclic – $i_k = 1, 2, \dots, m, 1, 2, \dots$
- ★ randomized – i_k uniform on $\{1, \dots, m\}$ (stochastic gradient)

One component per step, and the order is a design choice

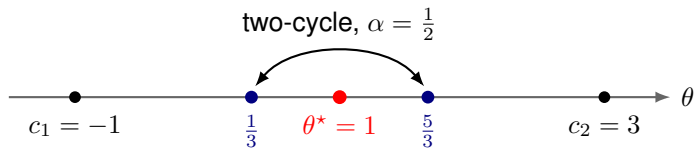
Example: two components, cyclic order

Example: $\underset{\theta}{\text{minimize}} \quad \frac{1}{2} (\theta - c_1)^2 + \frac{1}{2} (\theta - c_2)^2, \quad \theta^* = \frac{c_1 + c_2}{2}$

Cyclic, constant α : $\theta \leftarrow (1 - \alpha) \theta + \alpha c_1$, then $\theta \leftarrow (1 - \alpha) \theta + \alpha c_2$, repeat

■ THE ITERATES SETTLE ON A TWO-CYCLE

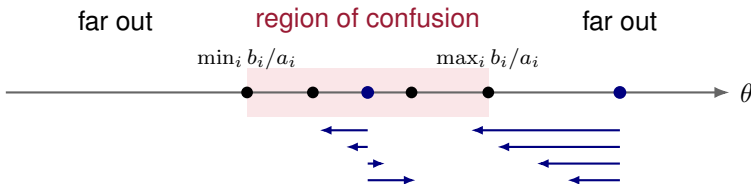
$$\frac{c_1 + (1 - \alpha)c_2}{2 - \alpha} \quad \text{and} \quad \frac{(1 - \alpha)c_1 + c_2}{2 - \alpha}, \quad \text{each} \quad \frac{\alpha |c_2 - c_1|}{2(2 - \alpha)} \quad \text{from} \quad \theta^*$$



A constant step leaves an $O(\alpha)$ cycle around θ^*

Far from θ^* , and near θ^*

Example: $f_i(\theta) = \frac{1}{2} (a_i\theta - b_i)^2$ with $a_i > 0$, component minimizers b_i/a_i



- FAR OUT: all $-\nabla f_i(\theta)$ point toward θ^* – m useful steps per pass
- INSIDE: the $-\nabla f_i(\theta)$ disagree – a constant step oscillates

Fast far from θ^* , and slow near it

Cyclic versus randomized order

Assumptions: f_i convex, $\|\nabla f_i(\theta)\| \leq M$ on Θ , $f^* = \min_{\theta \in \Theta} f(\theta)$

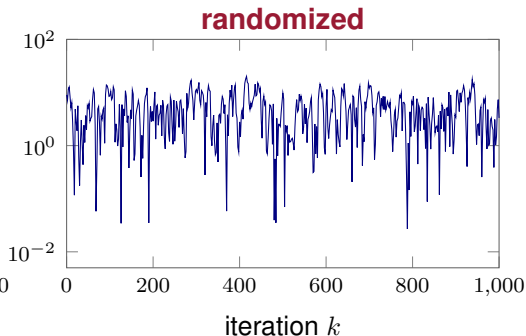
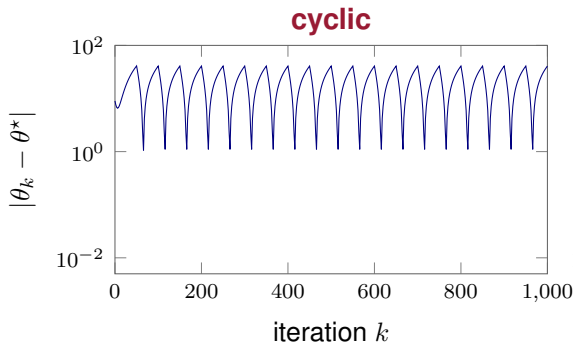
Order	Constant step α	$\alpha_k \rightarrow 0, \sum_k \alpha_k = \infty$
Cyclic	$\liminf_k f(\theta_k) \leq f^* + \alpha m M^2 / 2$	$\liminf_k f(\theta_k) = f^*$
Randomized, w.p. 1	$\liminf_k f(\theta_k) \leq f^* + \alpha M^2 / 2$	$\liminf_k f(\theta_k) = f^*$

A. Nedić and D. P. Bertsekas, *SIAM J. Optim.*, 12(1):109–138, 2001

Random order: an error floor m times lower

Example: cyclic versus randomized

Example: $f(\theta) = \frac{1}{100} \left(\sum_{i=1}^{50} (\theta - i)^2 + \sum_{i=1}^{50} (\theta + i)^2 \right), \quad \theta^* = 0, \quad \alpha = 0.05$



Cyclic: a sawtooth up to 40 — **randomized:** noise of size 7

Robust and incremental gradient methods

Method	Assumes	Stepsize	Guarantee
Stochastic gradient	σ -strongly convex	$1/(\sigma k)$	$\mathbb{E}\ \theta_k - \theta^*\ ^2 \leq D/k$
Robust stochastic gradient	convex	$\gamma R/(M\sqrt{K})$	$\frac{1}{2} \left(\gamma + \frac{1}{\gamma} \right) \frac{RM}{\sqrt{K}}$
Incremental, cyclic	convex f_i	constant α	floor $\alpha m M^2/2$
Incremental, randomized	convex f_i	constant α	floor $\alpha M^2/2$

Next: cheap steps and a linear rate – variance reduction

Questions?

(**Email: dongshed@utk.edu**)