

COSC 594/650

Stochastic variance reduced gradient

Dongsheng Ding

dongshed@utk.edu

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

Recap: batch versus stochastic

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

	Gradient descent	Stochastic gradient
Step	$\theta - \alpha \frac{1}{m} \sum_{i=1}^m \nabla f_i(\theta)$	$\theta - \alpha_k \nabla f_{i_k}(\theta)$
Cost per step	m gradients	1 gradient
Stepsize	constant	shrinking
Rate, strongly convex	linear	sublinear

Cheap steps, or a linear rate – so far, not both

QUESTION: Can cheap steps converge linearly?

Why the stepsize must shrink

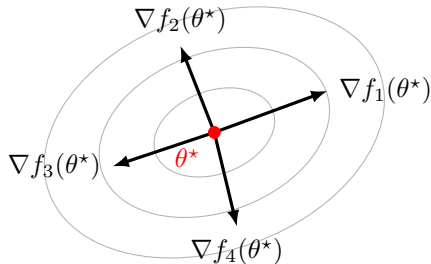
■ AT THE SOLUTION

$$\nabla f(\theta^*) = \frac{1}{m} \sum_{i=1}^m \nabla f_i(\theta^*) = 0$$

- ★ but each $\nabla f_i(\theta^*) \neq 0$ in general

■ THE NOISE STAYS

- ★ variance $\frac{1}{m} \sum_{i=1}^m \|\nabla f_i(\theta^*)\|^2 > 0$ at θ^*
- ★ a constant step stalls at $O(\alpha)$



The noise does not vanish at θ^* , so the step must

A control variate

Goal: estimate $\mathbb{E}[X]$ from one sample, with small variance

■ SUBTRACT A CORRELATED VARIABLE WITH A KNOWN MEAN

$$Z = X - Y + \mathbb{E}[Y] \implies \mathbb{E}[Z] = \mathbb{E}[X], \quad \text{Var}(Z) = \text{Var}(X - Y)$$

★ small when X and Y are close

■ FOR THE GRADIENT, WITH A SNAPSHOT $\tilde{\theta}$

$$X = \nabla f_i(\theta), \quad Y = \nabla f_i(\tilde{\theta}), \quad \mathbb{E}_i[Y] = \nabla f(\tilde{\theta})$$

$$\implies g = \nabla f_i(\theta) - \nabla f_i(\tilde{\theta}) + \nabla f(\tilde{\theta}), \quad \mathbb{E}_i[g] = \nabla f(\theta)$$

Unbiased, and its variance vanishes as θ and $\tilde{\theta}$ approach θ^*

Stochastic variance reduced gradient

Parameters: step size α , inner loop length T

for $s = 0, 1, 2, \dots$

full gradient at the snapshot: $\nabla f(\tilde{\theta}_s) = \frac{1}{m} \sum_{i=1}^m \nabla f_i(\tilde{\theta}_s)$, $\theta_0 = \tilde{\theta}_s$

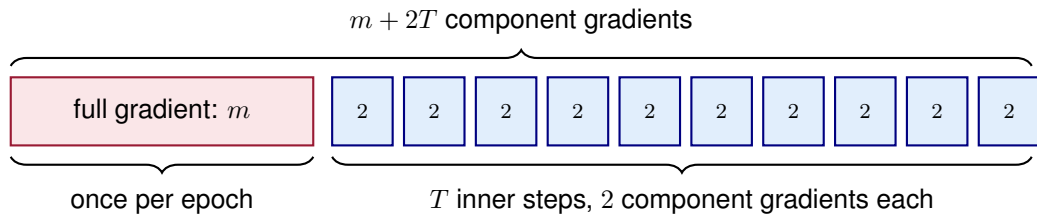
for $k = 0, 1, \dots, T - 1$: pick i_k uniformly at random

$$\theta_{k+1} = \theta_k - \alpha \left(\nabla f_{i_k}(\theta_k) - \nabla f_{i_k}(\tilde{\theta}_s) + \nabla f(\tilde{\theta}_s) \right)$$

new snapshot: $\tilde{\theta}_{s+1} = \theta_T$ (or θ_k with k uniform on $\{0, \dots, T - 1\}$)

One full gradient per epoch, then cheap steps with a constant α

The cost of one epoch



■ IN PRACTICE

- ★ $T = 2m$ – one epoch is $5m$ gradients, five passes over the data
- ★ memory – the snapshot $\tilde{\theta}_s$ and its full gradient, two vectors

An epoch costs a few passes over the data

Linear convergence

Components: each f_i convex and L -smooth

Average: f is σ -strongly convex, $\kappa = L/\sigma$

■ CONVERGENCE RATE

$$\rho = \frac{1}{\sigma\alpha(1-2L\alpha)T} + \frac{2L\alpha}{1-2L\alpha} < 1 \quad \implies \quad \mathbb{E} \left[f(\tilde{\theta}_s) \right] - f^* \leq \rho^s \left(f(\tilde{\theta}_0) - f^* \right)$$

★ with the random snapshot rule

The snapshots converge linearly, with a constant step

What the theorem buys

■ AN EXAMPLE

$$\alpha = \frac{0.1}{L}, \quad T = 50\kappa \quad \implies \quad \rho = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

■ WORK TO REACH ACCURACY ϵ , IN COMPONENT GRADIENTS

Gradient descent	Stochastic gradient	SVRG
$m\kappa \log(1/\epsilon)$	$O(1/\epsilon)$	$(m + \kappa) \log(1/\epsilon)$

★ $\kappa = m - m^2 \log(1/\epsilon)$ for gradient descent, and $m \log(1/\epsilon)$ for SVRG

$m + \kappa$ instead of $m\kappa$, and $\log(1/\epsilon)$ instead of $1/\epsilon$

Proof #1: the spread of the component gradients

Fix i : $h_i(\theta) = f_i(\theta) - f_i(\theta^*) - \nabla f_i(\theta^*)^\top (\theta - \theta^*)$

Properties: convex, L -smooth, $h_i \geq 0 = h_i(\theta^*) - \theta^*$ minimizes h_i

■ ONE GRADIENT STEP ON h_i , STEP $1/L$

$$0 \leq h_i\left(\theta - \frac{1}{L}\nabla h_i(\theta)\right) \leq h_i(\theta) - \frac{1}{2L}\|\nabla f_i(\theta) - \nabla f_i(\theta^*)\|^2$$

■ AVERAGE OVER i , USING $\nabla f(\theta^*) = 0$

$$\frac{1}{m} \sum_{i=1}^m \|\nabla f_i(\theta) - \nabla f_i(\theta^*)\|^2 \leq 2L(f(\theta) - f^*)$$

The spread is controlled by the optimality gap

Proof #2: the variance vanishes

Direction: $g_k = \nabla f_{i_k}(\theta_k) - \nabla f_{i_k}(\tilde{\theta}) + \nabla f(\tilde{\theta}), \quad \mathbb{E}[g_k] = \nabla f(\theta_k)$

■ SPLIT AT θ^*

$$\mathbb{E}\|g_k\|^2 \leq 2\mathbb{E}\|\nabla f_{i_k}(\theta_k) - \nabla f_{i_k}(\theta^*)\|^2 + 2\mathbb{E}\|\nabla f_{i_k}(\tilde{\theta}) - \nabla f_{i_k}(\theta^*) - \nabla f(\tilde{\theta})\|^2$$

$$\star \quad \|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2, \quad \text{and} \quad \mathbb{E}\|X - \mathbb{E}X\|^2 \leq \mathbb{E}\|X\|^2$$

■ APPLY PROOF #1 TWICE

$$\mathbb{E}\|g_k\|^2 \leq 4L \left(f(\theta_k) - f^* + f(\tilde{\theta}) - f^* \right)$$

The variance is bounded by the two optimality gaps

Proof #3: one inner step

■ EXPAND THE DISTANCE, AS IN LECTURE 19

$$\mathbb{E}\|\theta_{k+1} - \theta^*\|^2 = \|\theta_k - \theta^*\|^2 - 2\alpha (\theta_k - \theta^*)^\top \nabla f(\theta_k) + \alpha^2 \mathbb{E}\|g_k\|^2$$

★ expectation over i_k , given θ_k

■ USE CONVEXITY AND PROOF #2

$$\mathbb{E}\|\theta_{k+1} - \theta^*\|^2 \leq \|\theta_k - \theta^*\|^2 - 2\alpha (1 - 2L\alpha) (f(\theta_k) - f^*) + 4L\alpha^2 (f(\tilde{\theta}) - f^*)$$

Each inner step trades the gap at θ_k against the gap at the snapshot

Proof #4: one epoch

■ SUM OVER $k = 0, \dots, T - 1$, WITH $\theta_0 = \tilde{\theta}_s$

$$2\alpha(1 - 2L\alpha)T \mathbb{E} \left[f(\tilde{\theta}_{s+1}) - f^\star \right] \leq \mathbb{E} \|\tilde{\theta}_s - \theta^\star\|^2 + 4LT\alpha^2 \mathbb{E} \left[f(\tilde{\theta}_s) - f^\star \right]$$

★ random snapshot: $\mathbb{E} f(\tilde{\theta}_{s+1}) = \frac{1}{T} \sum_k \mathbb{E} f(\theta_k)$; drop $\mathbb{E} \|\theta_T - \theta^\star\|^2 \geq 0$

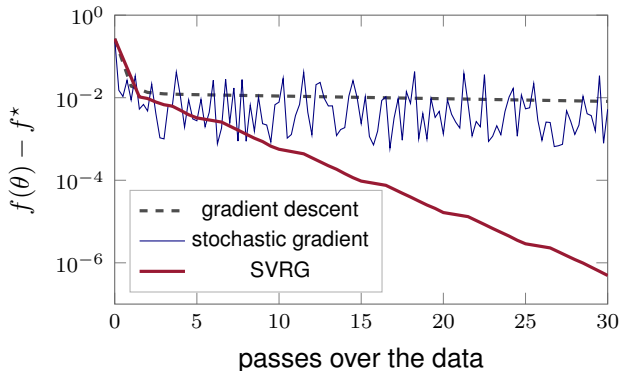
■ USE STRONG CONVEXITY

$$\|\tilde{\theta}_s - \theta^\star\|^2 \leq \frac{2}{\sigma} \left(f(\tilde{\theta}_s) - f^\star \right) \implies \mathbb{E} \left[f(\tilde{\theta}_{s+1}) - f^\star \right] \leq \rho \mathbb{E} \left[f(\tilde{\theta}_s) - f^\star \right]$$

Each epoch multiplies the gap by ρ

Experiment: logistic regression

Setup: ridge logistic regression, $m = 1000$, $d = 20$, $\lambda = 0.01$



■ SAME STEP $0.1/L$

- ★ SGD: a noise floor
- ★ SVRG: 10^{-6} in 30 passes

■ GRADIENT DESCENT

- ★ one step per pass

Variance reduction keeps the cheap steps, and adds a linear rate

Three ways to use the gradients

	Gradient descent	Stochastic gradient	SVRG
Per step	m gradients	1 gradient	2, plus m per epoch
Stepsize	constant	shrinking	constant
Rate	linear	$O(1/k)$	linear, per epoch
Total work	$m\kappa \log(1/\epsilon)$	$O(1/\epsilon)$	$(m + \kappa) \log(1/\epsilon)$

Cheap steps, and a linear rate – both at once

Questions?

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