

COSC 594/650

Stochastic gradient

Dongsheng Ding

dongshed@utk.edu

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

Where sample average approximation runs out

Practice: $\underset{\theta}{\text{minimize}} \quad \widehat{L}_m(\theta) = \frac{1}{m} \sum_{i=1}^m \ell(\xi_i, \theta)$

$$\theta_{k+1} = \theta_k - \alpha_k \frac{1}{m} \sum_{i=1}^m \nabla \ell(\xi_i, \theta_k)$$

■ EVERY STEP USES THE WHOLE SAMPLE

- ★ one gradient of $\widehat{L}_m$ costs m gradients of ℓ

■ THE SAMPLE IS FIXED IN ADVANCE

- ★ samples that arrive over time change the problem being solved

■ THE SAMPLE MUST BE STORED

- ★ m can exceed the memory available

QUESTION: Can each sample be used once, and never stored?

Stochastic approximation

Target: minimize $L(\theta) = \mathbb{E}_{\xi} [\ell(\xi, \theta)]$

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

■ ONE FRESH SAMPLE PER STEP

★ ξ_k is drawn at step k , independently of θ_k

■ THE STEP IS A GRADIENT ON AVERAGE

$$\mathbb{E} [\nabla \ell(\xi_k, \theta_k) \mid \theta_k] = \nabla L(\theta_k)$$

Noisy gradient descent: no sample is stored, and no sample is used twice

Convergence analysis: one-step descent

Strong convexity: L is σ -strongly convex

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

■ EXPAND THE DISTANCE TO $\theta^\star$

$$\|\theta_{k+1} - \theta^\star\|^2 = \|\theta_k - \theta^\star\|^2 - 2\alpha_k (\theta_k - \theta^\star)^\top \nabla\ell(\xi_k, \theta_k) + \alpha_k^2 \|\nabla\ell(\xi_k, \theta_k)\|^2$$

■ TAKE EXPECTATIONS, W/ $d_k = \mathbb{E}\|\theta_k - \theta^\star\|^2$

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

One inequality, one unknown term: the cross term

Convergence analysis: the cross term

■ CONDITION ON THE PAST

$$\mathbb{E} \left[(\theta_k - \theta^\star)^\top \nabla \ell(\xi_k, \theta_k) \right] = \mathbb{E} \left[(\theta_k - \theta^\star)^\top \nabla L(\theta_k) \right]$$

★ ξ_k is independent of θ_k – the inner expectation is $\nabla L(\theta_k)$

■ USE STRONG CONVEXITY TWICE

$$\begin{aligned} (\theta_k - \theta^\star)^\top \nabla L(\theta_k) &\geq L(\theta_k) - L(\theta^\star) + \frac{\sigma}{2} \|\theta_k - \theta^\star\|^2 \\ &\geq \frac{\sigma}{2} \|\theta_k - \theta^\star\|^2 + \frac{\sigma}{2} \|\theta_k - \theta^\star\|^2 = \sigma \|\theta_k - \theta^\star\|^2 \end{aligned}$$

$$\mathbb{E} [\cdot] \geq \sigma d_k \text{ – the cross term is a multiple of } d_k$$

Convergence analysis: the recursion

■ PUT THE TWO SLIDES TOGETHER

$$d_{k+1} \leq (1 - 2\sigma\alpha_k) d_k + \alpha_k^2 M^2$$

■ CHOOSE $\alpha_k = 1/(\sigma k)$ AND INDUCT ON k

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

★ the step size must shrink – a constant α leaves a floor of order $\alpha M^2/\sigma$

$\mathbb{E}\|\theta_k - \theta^*\|^2 = O(1/k)$ – k **samples of this method match** m **samples of SAA**

The step size is delicate

Example: minimize _{x} $\frac{1}{2}cx^2$, $c > 0$, $x_1 > 0$, no noise

■ THE WRONG CONSTANT

$$\alpha_k = \frac{1}{k} \quad \Rightarrow \quad x_{k+1} = x_1 \prod_{i=1}^k \left(1 - \frac{c}{i}\right) = \Theta(k^{-c})$$

★ $c = 1/5$ – after 10^{10} steps the error is still 10^{-2} of its start

■ THE RIGHT CONSTANT

$$\alpha_k = \frac{1}{ck} \quad \Rightarrow \quad x_2 = x_1 \left(1 - c \cdot \frac{1}{c}\right) = 0$$

The rate is set by σ , and σ is rarely known

Two ways to spend m samples

	Sample average approximation	Stochastic approximation
Sample	all m samples, repeatedly	one sample per step
Memory	the whole sample	one gradient
Output	$\hat{\theta}_m^*$, after solving a problem	θ_k , after k cheap steps
Error	$O_p(1/\sqrt{m})$	$\mathbb{E}\ \theta_k - \theta^*\ ^2 = O(1/k)$
Constant	$\text{tr}(H^{-1}\Sigma)$, nothing to tune	needs σ , and a bound M

Next: a step size that does not need σ – robust stochastic approximation

Questions?

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