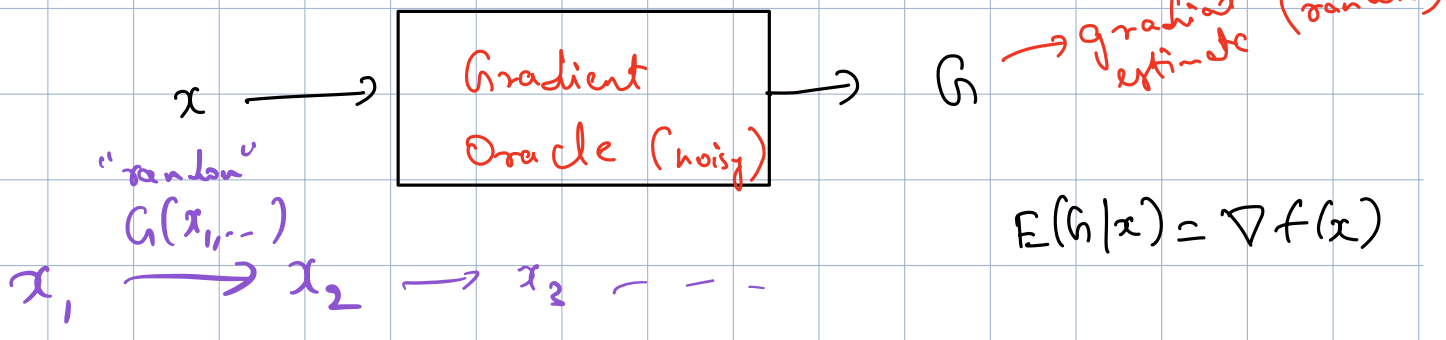
Lecture - 15 (contd)

Non-asymptotic analysis of stochastic gradient algorithms

Aim: Solve $\min_x f(x)$ → smooth

Setting: Unbiased gradient information



$$E(G|x) = \nabla f(x)$$

SG: algorithm

$$x_{k+1} = x_k - a(k)\, G(x_k, \xi_k) \qquad (1)$$

$a(k)$: step-size or learning rate

ξ_k: noise

Assumption:

(A1) $$E_{\xi_k}\left(G(x_k, \xi_k)\right) = \nabla f(x_k), \quad \forall k \geq 1$$

$$E_{\xi_k} \| G(x_k, \xi_k) - \nabla f(x_k) \|^2 \leq \sigma^2$$

for some $\sigma > 0$.

(A0) f is L-smooth

$$\|\nabla f(x_1) - \nabla f(x_2)\| \le L\|x_1 - x_2\|, \quad \forall x_1, x_2 \in \mathbb{R}^d$$

Asymptotic convergence:

(A2) $$\sum_k a(k) = \infty, \qquad \sum_k a(k)^2 < \infty$$

e.g. $a(k) = \frac{1}{k}$ → standard stochastic approximation conditions.

Under (A0) - (A2), the parameter x_k governed by ① converges almost surely to the set $\{x^* \mid \nabla f(x^*) = 0\}$

Asymptotic guarantee: $x_k \to \{x^* \mid \nabla f(x^*) = 0\}$ as $k \to \infty$ w.p. 1.

Tool used for proving this claim: Kushner-Clark Lemma

Eq ① (=)

$$x_{k+1} = x_k - a(k)\left(\nabla f(x_k) + M_{k+1}\right)$$

$$M_{k+1} \doteq G(x_k, \xi_k) - \nabla f(x_k)$$

("MDS")

We want to understand the non-asymptotic performance of the SG algorithm above.

Stationary point, say x^*, has $\nabla f(x^*) = 0$

ϵ-stationary point: $\bar{x}$ is an ϵ-stationary point if $E\|\nabla f(\bar{x})\| \leq \epsilon$, where the expectation is over the randomness in the SG algorithm.

[Ghadimi-Lan, 2014, SIAM. J. Opt, "Stochastic first/zeroth order]

"Randomized stochastic gradient" (RSG)

Suppose we run (1) for N iterations.

$\{x_1, x_2, x_3, \ldots x_N\}$ → (random) Set of iterates visited by SG algorithm.

RSG will pick a random iterate as follows:

R is picked uniformly at random from $\{x_1, \ldots x_N\}$

i.e., $\mathbb{P}(x_R = x_i) = \frac{1}{N} \quad \forall i$

$\mathbb{E}\, x_R = \frac{1}{N} \sum_{i=1}^{N} x_i$ } average iterate.

The non-asymptotic bound for RSG is of the form

$$\boxed{\mathbb{E}\, \|\nabla f(x_R)\|^2 \leq \frac{\text{Const}}{\sqrt{N}}}$$

under suitable choice of step-size.

RL connection: $x \to$ policy parameter

Objective $f \to$ $J_\pi(s^0)$ or $J(x)$

↓

Value function with start state s^0 in a discounted/SSP

"Policy gradient"

Proof of the non-asymptotic bound for RSG:

First → derive a bound for a general step-size

Second → Specialize the bound above with a particular stepsize.

First step: Recall $x_{k+1} = x_k - a(k)\, G(x_k, \xi_k)$

f is L-smooth:

$$f(x_{k+1}) \le f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2$$

$$= f(x_k) - a(k) \nabla f(x_k)^T G(x_k, \xi_k) + \frac{L}{2} a(k)^2 \|G(x_k, \xi_k)\|^2 \quad \text{— (1)}$$

Taking expectation wrt $\mathcal{F}_k$ (σ-field with info up to "k"), denote this by E_k

$$E_k f(x_{k+1}) \le f(x_k) - a(k) \|\nabla f(x_k)\|^2 + \frac{L}{2} a(k)^2 \left(\|\nabla f(x_k)\|^2 + \sigma^2 \right) \quad (2)$$

(why ↑ (A))

$$= f(x_k) - \left(a(k) - \frac{L}{2} a(k)^2 \right) \|\nabla f(x_k)\|^2 + \frac{L}{2} a(k)^2 \sigma^2$$

$$\left(a(k) - \frac{L}{2} a(k)^2 \right) \|\nabla f(x_k)\|^2 \le f(x_k) - E_k(f(x_{k+1})) + \frac{L}{2} a(k)^2 \sigma^2$$

(*) $$a(k) \|\nabla f(x_k)\|^2 \le \frac{2\left(f(x_k) - E_k[f(x_{k+1})] \right)}{2 - L a(k)} + \frac{L a(k)^2 \sigma^2}{(2 - L a(k))}$$

↳ assuming $a(k) \le 1/L \quad \forall k$

Summing ⊛ over $k = 1$ to N,

$$\sum_{k=1}^{N} a(k) \|\nabla f(x_k)\|^2 \le 2 \sum_{k=1}^{N} \frac{f(x_k) - E_k f(x_{k+1})}{2 - L\, a(k)} + L\sigma^2 \sum_{k=1}^{N} \frac{a(k)^2}{(2 - L\, a(k))} \quad \text{— (3)}$$

Take expectation wrt. $\{x_1, \ldots x_N\}$, i.e., $E_N(\cdot)$

$$\sum_{k=1}^{N} a(k) E_N \|\nabla f(x_k)\|^2 \le 2 \sum_{k=1}^{N} \frac{E_N f(x_k) - E_N f(x_{k+1})}{2 - L\, a(k)} + L\sigma^2 \sum_{k=1}^{N} \frac{a(k)^2}{(2 - L\, a(k))}$$

"End of lecture 15"

Set $a(k) = a \quad \forall\, k = 1 \ldots N$

$$a \sum_{k=1}^{N} E_N \|\nabla f(x_k)\|^2 \le \frac{2}{2 - La} \left(f(x_1) - f(x^*) \right) + \frac{L\sigma^2 N a^2}{2 - La}$$

(x^* → global optima of f)

Choose $a \le \frac{1}{L}$

used $E_N f(x_{N+1}) \ge f(x^*)$

$$E_N \|\nabla f(x_R)\|^2 = \frac{1}{N} \sum_{k=1}^{N} E_N \|\nabla f(x_k)\|^2$$

So,

$$E_N \|\nabla f(x_R)\|^2 \le \frac{1}{Na} \left[\frac{2}{2-La} \left(f(x_1) - f(x^*) \right) + \frac{L\sigma^2 N a^2}{2-La} \right]$$

Initial error (bias)

$$\le \frac{2}{Na} \left(f(x_1) - f(x^*) \right)$$

Sampling error (variance)

$$+ \frac{L\sigma^2 \cancel{N} a^{\cancel{2}}}{\cancel{N}a}$$

$$= \frac{2}{Na} \left(f(x_1) - f(x^*) \right) + L\sigma^2 a$$

Initial error

Variance due to gradient estimation

We have,

$$E \|\nabla f(x_R)\|^2 \le \frac{\text{Const}}{Na} + \text{Const}\, a$$

Setting $a = \min\left\{ \frac{1}{L}, \frac{1}{\sqrt{N}} \right\}$

$x_{k+1} = x_k - a_k g(x_k, \xi_k)$

$a_k = \frac{1}{k}$ diminishing

$a_k = a \to$ Constant

Nitration, $a = \frac{1}{\sqrt{N}}$

$$E \|\nabla f(x_R)\|^2 \le \frac{2\left(f(x_1) - f(x^*)\right)}{N} \max\left(L, \sqrt{N}\right)$$

$$+ \frac{L\sigma^2}{\sqrt{N}}$$

$$= \frac{2L\left(f(x_1) - f(x^*)\right)}{N} + \frac{1}{\sqrt{N}}\left(2\left(f(x_1) - f(x^*)\right) + L\sigma^2\right)$$

$$\boxed{\mathbb{E}\,\|\nabla f(x_R)\|^2 \leq \frac{\text{const}}{\sqrt{N}}}$$

→ Bound for smooth f + unbiased gradient oracle.

$\Rightarrow$ For RSh to converge to an ϵ-stationary point, i.e., $\mathbb{E}\,\|\nabla f(x_R)\|^2 \leq \epsilon$,

on order $\frac{1}{\epsilon^2}$ number of iterations

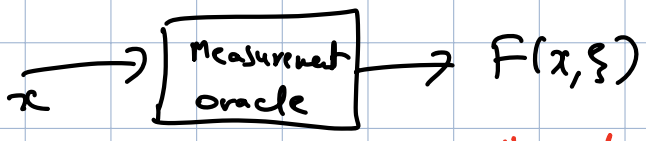
of RSh algorithm are sufficient.

Stochastic optimization with a biased gradient oracle

(Regular) stochastic optimization

$$\min_{x \in \mathcal{X}} \{ f(x) = E_\xi (F(x, \xi)) \}$$

Zeroth-order input



"gradients" not directly available

Stochastic gradient algorithm:

$$x_{k+1} = \Pi (x_k - \gamma_k g_k)$$

Π: projection operator; γ_k: stepsize; g_k: gradient estimate

With zeroth-order input, can form gradient estimates that satisfy

$$\| E_\xi [g(x, \xi)] - \nabla f(x) \| \leq c_1 \delta^2$$

$$E_\xi \| g(x, \xi) - E_\xi (g(x, \xi)) \|^2 \leq \frac{c_2}{\delta^2}$$

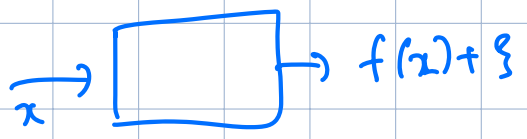
Bias

Variance

"δ" → bias-variance tradeoff.

How to form such an estimate! "Simultaneous perturbation" trick.

Here is a sample for $f \in \mathcal{C}^3$



Take two measurements : ① $f(x+\delta\Delta)+\xi^+$

② $f(x-\delta\Delta)-\xi^-$

$$\Delta = (\Delta_1, \dots, \Delta_d)$$

Δ_i = Rademacher $= \begin{cases} +1 & \text{w.p. } 1/2 \\ -1 & \text{w.p. } 1/2 \end{cases}$ iid

Gradient estimate $g^i = \dfrac{① - ②}{2\delta\Delta^i}$

Use Taylor expansion

$$f(x \pm \delta\Delta) = f(x) \pm \delta\Delta^T \nabla f(x) + \frac{\delta^2}{2}\Delta^T \nabla^2 f(x)\Delta + O(\delta^3)$$

$$\frac{f(x+\delta\Delta) - f(x-\delta\Delta)}{2\delta\Delta^i} = \nabla_i f(x) + \sum_{j \neq i} \frac{\Delta^j}{\Delta^i}\nabla_j f(x) + O(\delta^2)$$

↑ zero-mean

$$E(g^i) = \nabla_i f(x) + O(\delta^2)$$

Can show $\quad E(\|g - Eg\|^2) \le \dfrac{C\delta^2}{2}$

Δ = Rademacher vector $\Rightarrow$ SPSA "Spall 1992"

Can use other distributions for random perturbations.

Alternate framework: "Biased function measurements".

$$f(x) = \min_y \; \mathbb{E}_\xi\big(H_x(y,\xi)\big) \quad \text{——} (*)$$

↑ the objective at a certain input is the solution to an optimization problem

Suppose a batch-size parameter "m" decides how well (*) is solved, i.e.,

$$F(x,m) = \min_y \; \mathbb{E}_\xi\big[H_x(y,\xi)\big] + \underbrace{\epsilon(m)}$$

error term that has positive mean "increasing m lowers $\epsilon(m)$".

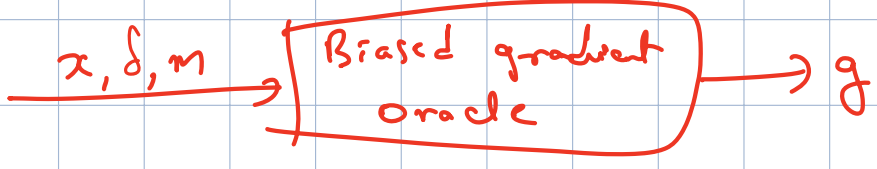
e.g. Risk-sensitive optimization ϱ: risk-measure $c_x = \varrho(Y_x)$

$\hat{\varrho}_{m,x}$ → estimate of $\varrho(Y_x)$ using m samples

$\mathbb{E}(\hat{\varrho}_{m,x}) \neq \varrho(Y_x)$, but

$$\mathbb{E}\,|\hat{\varrho}_{m,x} - \varrho(Y_x)|^2 \leq \frac{\text{const.}}{\sqrt{m}} \qquad \forall x$$

With biased function measurements, why the simultaneous perturbation trick leads to



Bias: $\| E_\xi [g(x,\xi,m)] - \nabla f(x) \| \leq c_1 \delta^2 + \frac{c_2}{\delta \sqrt{m}}$

Variance: $E_\xi \| g - E g \|^2 \leq \frac{c_3}{\delta^2}$

Algorithm: Randomized Stochastic gradient
[Ghadimi-Lan, SiOpt 2015]

Perform $x_{k+1} = \Pi\left(x_k - a_k g(x_k, \xi_k, m_k)\right)$
for N iterations

& return a random iterate R picked uniformly at random from $\{x_1, \dots, x_N\}$

Assumption (A1) $\| \nabla f(x) - \nabla f(y) \| \leq L \|x-y\| \quad \forall x, y$ fixed.

Assumption (A2) $\| \nabla f(x) \|_* \leq B < \infty, \forall x \in X$

Main claim: With $a_k = \min\left\{\frac{1}{L}, \frac{1}{N^{2/3}}\right\}$, $\delta_k = \frac{1}{N^{1/6}}$,

$m_k = N$

$\forall N \geq 1, \quad E \|\nabla f(x_R)\|^2 \leq \frac{2L(f(x_1) - f(x^*))}{N} + \frac{\text{const.}}{N^{1/3}}$

See [N. Bhavsar & P.C.A. "Zeroth-→

IEEE TAC, 2022]

A biased gradient oracle variant:

Bias: $\| E_\xi [g(x,\xi,m)] - \nabla f(x) \|_\infty \le C_1 \delta^2 + \frac{C_2}{\delta\sqrt{m}}$

Variance: $E_\xi \| g - Eg \|^2 \le C_3 \delta + \tilde{C}_3$

In this oracle model, we can obtain $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ rate

Motivation: Biased function measurements in a "noise-less" setting

Recall $g^i_2 \quad \dfrac{(f(x+\delta\Delta)+\xi^+) - (f(x-\delta\Delta)+\xi^-)}{2\Delta^i\delta}$

$E\|g - Eg\|^2 \le \dfrac{\text{Const}}{\delta^2}$ because of noise elements ξ^+, ξ^-

If no noise, the $E\|g - Eg\|^2 \le C_3\delta^2 + \tilde{C}_3$

How: Use a Gaussian-Smoothing estimator.

[Nesterov-Spokoiny, 2017]

$$g = \Delta\left[\frac{f(x+\delta\Delta) - f(x-\delta\Delta)}{2\delta}\right]$$

$\Delta = (\Delta^1, \dots, \Delta^d) \quad \Delta^i \sim N(0,1)$

$$x_{k+1} = x_k - \gamma_k g(x_k, \xi_k)$$

Main claim: With $\gamma_k = \min\left\{\frac{1}{L}, \frac{1}{\sqrt{N}}\right\}$, $\delta_k = \frac{1}{\sqrt{N}}$,

$$m_k = N$$

$$\forall N \geq 1, \quad \mathbb{E}\|\nabla f(x_R)\|^2 \leq \frac{2(f(x_1) - f(x^*))}{N} + \frac{\text{const.}}{\sqrt{N}}$$

For an oracle without bias-variance tradeoff

End of Lecture-16

Lecture-17

Non-asymptotic bounds for zeroth-order optimization

(0) Input: x Output: $g(x,\xi)$

(i) $\mathbb{E}_\xi(g(x,\xi)) = \nabla f(x) + c_1\delta^2 1_{d\times 1}$

(ii) $\mathbb{E}_\xi\left[\|g(x,\xi) - \mathbb{E}_\xi(g(x,\xi))\|^2\right] \le \frac{c_2}{\delta^2}$

} Oracle with bias variance tradeoff

Algorithm: Randomized Stochastic gradient

[Ghadimi-Lan, SiOpt 2015]

Perform $x_{k+1} = x_k - \gamma_k g(x_k,\xi_k)$

for N iterations

& return a random iterate R picked uniformly at random from $\{x_1, \dots, x_N\}$

Assumption (A1) $\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|$ $\forall x, y$ fixed.

Main claim: With $\gamma_k = \min\left\{\frac{1}{L}, \frac{1}{(d^2N)^{1/3}}\right\}$, $\delta_k = \frac{1}{(d^5N)^{1/6}}$,

$\forall N \ge 1$, $\mathbb{E}\|\nabla f(x_R)\|^2 \le \frac{2L(f(x_1) - f(x^*))}{N} + \frac{\text{const.}}{N^{1/3}}$

Pf: We prove a result for general γ_k & then specialize

With $P_R(k) = \text{Prob}(R=k) = \frac{\gamma_k}{\sum_{i=1}^N \gamma_i}$,

$$\mathbb{E}\|\nabla f(x_R)\|^2 \le \frac{1}{\sum_{i=1}^N \gamma_i}\left[\frac{2(f(x_1) - f(x^*))}{(2-L\gamma_1)} + 2B\sum_{k=1}^N c_1\delta_k^2\left(\frac{\gamma_k + L\gamma_k^2}{2-L\gamma_k}\right)\right.$$

$$+ L \sum_{k=1}^{N} \frac{\gamma_k^2}{(2-L\gamma_k)} \left(d c_1^2 \delta_k^4 + \frac{c_2}{\delta_k^2} \right) \Bigg]$$

Pf: Since f is L-smooth,

$$f(x_{k+1}) \le f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2$$

$$= f(x_k) - \gamma_k \langle \nabla f(x_k), g(x_k, \xi_k) \rangle + \frac{L}{2} \gamma_k^2 \|g(x_k, \xi_k)\|^2$$

$E_k(f(x_{k+1}))$ — conditional expectation, i.e., wrt info upto time k

$$\le f(x_k) - \gamma_k \langle \nabla f(x_k), \nabla f(x_k) + c_1 \delta_k^2 \mathbb{1}_{d\times 1} \rangle$$

$$+ \frac{L}{2} \gamma_k^2 \left[\| E_k(g(x_k, \xi_k)) \|^2 + \frac{c_2}{\delta_k^2} \right]$$

wrt $E_k \|g\|^2 \le c \|E_k g\|^2 + \frac{c_2}{\delta_k^2}$

wrt $-\|x\|_1 \le \sum_{i=1}^{d} x_i$

wrt $E_k g = \nabla f + \delta^2 \mathbb{1}_d$

$$\le f(x_k) - \gamma_k \|\nabla f(x_k)\|^2 + \gamma_k c_1 \delta_k^2 \|\nabla f(x_k)\|_1$$

$$+ \frac{L}{2} \gamma_k^2 \left[\|\nabla f(x_k)\|^2 + 2 c_1 \delta_k^2 \|\nabla f(x_k)\|_1 + d c_1^2 \delta_k^4 + \frac{c_2}{\delta_k^2} \right]$$

upper bound $\leftarrow \|\nabla f(x_k)\|_1$

$$\le f(x_k) - \left(\gamma_k - \frac{L}{2} \gamma_k^2 \right) \|\nabla f(x_k)\|^2 + c_1 \delta_k^2 (\gamma_k + L\gamma_k^2) B$$

$$+ \frac{L\gamma_k^2}{2} \left[d c_1^2 \delta_k^4 + \frac{c_2}{\delta_k^2} \right]$$

Rearranging,

$$\left(\gamma_k - \frac{L}{2} \gamma_k^2 \right) \|\nabla f(x_k)\|^2 \le f(x_k) - E_k f(x_{k+1}) + c_1 \delta_k^2 (\gamma_k + L\gamma_k^2) B$$

$$+ \frac{L\gamma_k^2}{2} \left[d c_1^2 \delta_k^4 + \frac{c_2}{\delta_k^2} \right]$$

7) Assuming $\gamma_k \le \frac{1}{L}$

$$\gamma_k \|\nabla f(x_k)\|^2 \le \frac{2}{(2-L\gamma_k)}\left[f(x_k) - E_k(f(x_{k+1})) + C_1\delta_k^2(\gamma_k + L\gamma_k^2)B\right] + \frac{L\gamma_k^2}{2-L\gamma_k}\left[\alpha C_1^2\delta_k^4 + \frac{C_2}{\delta_k^2}\right]$$

Summing & taking expectation,

$$\sum_{k=1}^{N} \gamma_k E_N \|\nabla f(x_k)\|^2 \le 2\sum_{k=1}^{N} \frac{E_N(f(x_k)) - E_N(f(x_{k+1}))}{(2-L\gamma_k)} + 2\sum_{k=1}^{N} C_1\delta_k^2\left(\frac{\gamma_k + L\gamma_k^2}{2-L\gamma_k}\right)B + L\sum_{k=1}^{N}\frac{\gamma_k^2}{(2-L\gamma_k)}\left(\alpha C_1^2\delta_k^4 + \frac{C_2}{\delta_k^2}\right)$$

$$\le 2\left[\frac{f(x_1)}{2-L\gamma_1} - \frac{E_N(f(x_{N+1}))}{2-L\gamma_1}\right] + 2\sum_{k=1}^{N} C_1\delta_k^2\left(\frac{\gamma_k + L\gamma_k^2}{2-L\gamma_k}\right)B + L\sum_{k=1}^{N}\frac{\gamma_k^2}{(2-L\gamma_k)}\left(\alpha C_1^2\delta_k^4 + \frac{C_2}{\delta_k^2}\right)$$

$$\le 2\left[\frac{f(x_1) - f(x^*)}{(2-L\gamma_1)}\right] + 2\sum_{k=1}^{N} C_1\delta_k^2\left(\frac{\gamma_k + L\gamma_k^2}{2-L\gamma_k}\right)B + L\sum_{k=1}^{N}\frac{\gamma_k^2}{(2-L\gamma_k)}\left(\alpha C_1\delta_k^4 + \frac{C_2}{\delta_k^2}\right)$$

$$E\|\nabla f(x_R)\|^2 \le \frac{1}{\sum_{k=1}^{N}\gamma_k}\left(\quad \right)$$

Specialize the result above for

$$\gamma_k = \left(\gamma = \min\left\{ \frac{1}{L}, \frac{1}{(d^2N)^{2/3}} \right\} \right), \quad \delta_k = \left(\delta = \frac{1}{(d^5N)^{1/6}} \right)$$

$$\mathbb{E}\|\nabla f(x_R)\|^2 = \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\|\nabla f(x_k)\|^2$$

$$\leq \frac{1}{N\gamma}\left[\frac{2(f(x_1)-f(x^*))}{2-L\gamma} + \frac{2BNC_1\delta^2(\gamma+L\gamma^2)}{(2-L\gamma)} + \frac{LN\gamma^2}{2-L\gamma}\left(dC_1^2\delta^4 + \frac{C_2}{\delta^2} \right) \right]$$

wog $\gamma \leq \frac{1}{L}$

$$\leq \boxed{\frac{2(f(x_1)-f(x^*))}{N\gamma} + 4BC_1\delta^2 + LN\gamma^2\left(dC_1^2\delta^4 + \frac{C_2}{\delta^2} \right)}$$

$$\leq \frac{2(f(x_1)-f(x^*))}{N} \max\left(L_1(d^2N)^{2/3} \right) + \frac{LBC_1}{(d^5N)^{1/3}} + L\left(\frac{dC_1^2}{(d^5N)^{2/3}} + \frac{2d^{5/3}}{N^{-1/3}} \right)\frac{1}{(d^2N)^{2/3}}$$

$$= 2L\frac{(f(x_1)-f(x^*))}{N} + \frac{2d^{4/3}}{N^{1/3}}(f(x_1)-f(x^*)) + \frac{4BC_1}{d^{5/3}N^{1/3}} + L\left(\frac{C_1^2}{d^{7/3}N^{2/3}} + \frac{C_2 d^{5/3}}{N^{-1/3}} \right]\frac{1}{(d^2N)^{2/3}}$$

$$= \frac{2L(f(x_1)-f(x^*))}{N} + \frac{1}{N^{1/3}}\left[2d^{4/3}(f(x_1)-f(x^*)) + \frac{4BC_1}{d^{5/3}} + \frac{LC_1^2}{d^{11/3}N} + C_2 d^{1/3} \right]$$

(first term: optimization error; bracketed term: sampling error)

$$\mathbb{E}\,\|\nabla f(x_k)\|^2 \le \frac{\text{Const}\cdot d^{4/3}}{N^{1/3}} \le \epsilon$$

$$\text{if} \quad N \ge \frac{d^4}{\epsilon^3}$$

Extend to biased gradient oracle without bias-variance tradeoff, i.e., a setting like

$$\|\mathbb{E}\,g - \nabla f\| \le C_1 \delta$$

$$\mathbb{E}\,\|g - \mathbb{E}g\|^2 \le C_2 \delta + C_3$$

End of Lecture-17

Lecture-18

chapter 7

Analysis of stochastic gradient algorithms: Strongly Convex Case

7.1 SG algorithm and a useful lemma

Stochastic gradient algo

$$x_{k+1} = x_k - \alpha_k g(x_k, \xi_k). \tag{7.1}$$

Assumption 7.1. $\mathbb{E}_{\xi_k}[g(x_k, \xi_k)] = \nabla f(x_k)$ *and* $\mathbb{E}_{\xi_k}[\|g(x_k, \xi_k)\|_2^2] - \|\mathbb{E}_{\xi_k}[g(x_k, \xi_k)]\|_2^2 \leq \sigma^2$.

$\mathbb{E}_\xi g^2 \leq \sigma^2 + \|\mathbb{E}g\|^2 = \sigma^2 + \|\nabla f\|^2$

Assumption 7.2. $f(x) \geq \bar{f}$ *for all* x

Lemma 7.1 (Improvement in gradient step)**.** *Under Assumption 7.1, we have*

$$\mathbb{E}_{\xi_k}[f(x_{k+1})] - f(x_k) \leq -\alpha_k(1 - \frac{1}{2}\alpha_k L)\|\nabla f(x_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L\sigma^2. \tag{7.2}$$

∇f is L-Lipschitz

Proof. Using L-smoothness of f and the update iteration (7.1), we obtain

$$f(x_{k+1}) - f(x_k) \leq \nabla f(x_k)^T(x_{k+1} - x_k) + \frac{1}{2}L\|x_{k+1} - x_k\|_2^2 \tag{7.3}$$

$$\leq -\alpha_k \nabla f(x_k)^T g(x_k, \xi_k) + \frac{1}{2}\alpha_k^2 L\|g(x_k, \xi_k)\|_2^2. \tag{7.4}$$

Taking expectation wrt ξ_k, we obtain

$$\mathbb{E}_{\xi_k}[f(x_{k+1})] - f(x_k) \leq -\alpha_k\|\nabla f(x_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L\mathbb{E}_{\xi_k}[\|g(x_k, \xi_k)\|_2^2] \tag{7.5}$$

$$\leq -\alpha_k(1 - \frac{1}{2}\alpha_k L)\|\nabla f(x_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L\sigma^2. \tag{7.6}$$

The claim follows. □

7.2 Strongly convex case

Theorem 7.2. *Let f be a m-strongly convex function. Then, the SG algorithm governed by* (7.1) *and with $\alpha_k = \alpha$ s.t. $0 < \alpha L < 1$, satisfies*

$$\mathbb{E}[f(x_n) - f(x^*)] \leq \frac{\alpha L \sigma^2}{2m} + (1 - \alpha m)^{n-1} \left(f(x_1) - f(x^*) - \frac{\alpha L \sigma^2}{2m} \right). \tag{7.7}$$

Proof. From Lemma 7.1, we have

$$\mathbb{E}_{\xi_k}[f(x_{k+1})] - f(x_k) \leq -\alpha_k (1 - \frac{1}{2}\alpha_k L) \|\nabla f(x_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L \sigma^2.$$

Since $\alpha_k = \alpha$ and $0 < \alpha L < 1$, we have

$$\mathbb{E}_{\xi_k}[f(x_{k+1})] - f(x_k) \leq -\frac{1}{2}\alpha \|\nabla f(x_k)\|_2^2 + \frac{1}{2}\alpha^2 L \sigma^2. \tag{7.8}$$

Since f is m-strongly convex, the PL-condition holds, i.e.,

$$f(x) - f(x^*) \leq \frac{1}{2m} \|\nabla f(x)\|_2^2, \ \forall x.$$

Using PL-condition in (7.8), we obtain

$$\mathbb{E}_{\xi_k}[f(x_{k+1})] - f(x_k) \leq -m\alpha (f(x_k) - f(x^*) + \frac{1}{2}\alpha^2 L \sigma^2. \tag{7.9}$$

Subtracting $f(x^*)$ on both sides and re-arranging, we obtain

$$\mathbb{E}_{\xi_k}[f(x_{k+1}) - f(x^*)] \leq (1 - \alpha m)[f(x_k) - f(x^*)] + \frac{1}{2}\alpha^2 L \sigma^2. \tag{7.10}$$

Taking expectations followed by straightforward simplifications, we obtain

$$\begin{aligned} \mathbb{E}[f(x_{k+1}) - f(x^*)] - \frac{\alpha L \sigma^2}{2m} &\leq (1 - \alpha m)\mathbb{E}[f(x_k) - f(x^*)] + \frac{\alpha^2 L \sigma^2}{2} - \frac{\alpha L \sigma^2}{2m} \\ &= (1 - \alpha m) \left(\mathbb{E}[f(x_k) - f(x^*)] - \frac{\alpha L \sigma^2}{2m} \right). \end{aligned} \tag{7.11}$$

Since $\alpha m < \frac{m}{L} < 1$, a repeated application of the above inequality leads to the following bound:

$$\mathbb{E}[f(x_n) - f(x^*)] \leq \frac{\alpha L \sigma^2}{2m} + (1 - \alpha m)^{n-1} \left(f(x_1) - f(x^*) - \frac{\alpha L \sigma^2}{2m} \right). \tag{7.12}$$

The claim follows. □

Remark 7.1. *Taking limits as $k \to \infty$ in* (7.7)*, we obtain*

$$\mathbb{E}[f(x_k) - f(x^*)] \to \frac{\alpha L \sigma^2}{2m} \text{ as } k \to \infty. \tag{7.13}$$

The result above implies that a constant stepsize SG algorithm does not converge to the optima, and instead gets to within a ball around the optima.

Diminishing Stepsize:

From improvement lemma,

$$E_k\left(f(x_{k+1})\right) - f(x_k) \le -\alpha_k\left(1 - \frac{\alpha_k L}{2}\right)\|\nabla f(x_k)\|^2 + \frac{\alpha_k^2 L \sigma^2}{2}$$

Suppose $\alpha_k L \le 1$

$$\le -\frac{\alpha_k}{2}\|\nabla f(x_k)\|^2 + \frac{\alpha_k^2 L \sigma^2}{2}$$

PL-condition $\Rightarrow$

$$\le -\alpha_k m\left(f(x_k) - f(x^*)\right) + \frac{\alpha_k^2 L \sigma^2}{2}$$

$$E_k f(x_{k+1}) - f(x^*) \le (1 - \alpha_k m)\left(f(x_k) - f(x^*)\right) + \frac{\alpha_k^2 L \sigma^2}{2}$$

Set $\alpha_k = \frac{\beta}{k+1}$, $\beta > \frac{1}{m}$ $\Rightarrow$ strong convexity parameter

Claim:

$$E\left(f(x_n) - f(x^*)\right) \le \frac{C}{n+1},$$

where $C = \max\left\{\frac{\beta^2 L \sigma^2}{2(\beta m - 1)},\ 2\left(f(x_1) - f(x^*)\right)\right\}$

Proof: by induction

Base case: holds trivially

Assume Claim holds for "n"

Ref:

Bottou, Nocedal, Curtis

"Opt for large-scale learning"

Survey, 2018

$$E\left(f(x_{n+1}) - f(x^*)\right) \leq (1-\alpha_n m) E\left(f(x_n) - f(x^*)\right) + \frac{\alpha_n^2 L \sigma^2}{2}$$

Induction hypothesis $\rightarrow$

$$\leq \left(1 - \frac{\beta m}{n+1}\right) \frac{c}{n+1} + \frac{\beta^2 L \sigma^2}{2(n+1)^2}$$

$$= \frac{(n+1-\beta m)c}{(n+1)^2} + \frac{\beta^2 L \sigma^2}{2(n+1)^2}$$

$$= \frac{n}{(n+1)^2} c - \frac{(\beta m - 1)}{(n+1)^2} c + \frac{\beta^2 L \sigma^2}{2(n+1)^2}$$

$$\frac{\beta^2 L \sigma^2}{2(n+1)^2} - c\,\frac{(\beta m - 1)}{(n+1)^2} \leq 0 \quad \text{Since}$$

$$c = \max\left(\frac{\beta^2 L \sigma^2}{2(\beta m - 1)}, 2\{f(x_1) - f(x^*)\}\right)$$

So,

$$E\left(f(x_{n+1}) - f(x^*)\right) \leq \frac{nc}{(n+1)^2} \leq \frac{c}{n+2}$$

SG for convex functions: unbiased gradients case

Problem: $\min_x f(x)$

SG: $x_{k+1} = x_k - \alpha_k g(x_k, \xi_k)$

Assume: (1) f is L-smooth

(2) **Assumption 7.1.** $\mathbb{E}_{\xi_k}[g(x_k,\xi_k)] = \nabla f(x_k)$ *and* $\mathbb{E}_{\xi_k}[\|g(x_k,\xi_k)\|_2^2] - \|\mathbb{E}_{\xi_k}[g(x_k,\xi_k)]\|_2^2 \le \sigma^2$.

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\alpha_k \langle g(x_k,\xi_k), x_k - x^* \rangle + \gamma_k^2 \|g(x_k,\xi_k)\|^2$$

$$\mathbb{E}\,\|x_{k+1} - x^*\|^2 \le \mathbb{E}\,\|x_k - x^*\|^2 - 2\alpha_k \langle \nabla f(x_k), x_k - x^* \rangle + \alpha_k^2 (\sigma^2 + \|\nabla f\|^2)$$

Convex f: $f(x^*) \ge f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle$

$\|\nabla f(x_k)\|^2 \le L \langle \nabla f(x_k), x_k - x^* \rangle$ ⟶ How? H.W.

$$(2\alpha_k - L\alpha_k^2)\left(\mathbb{E}\, f(x_k) - f(x^*)\right) \le \left(\mathbb{E}\,\|x_k - x^*\|^2 - \mathbb{E}\,\|x_{k+1} - x^*\|^2\right) + \alpha_k^2 \sigma^2$$

$$\alpha_k \left(\mathbb{E}(f(x_k)) - f(x^*)\right) \le \frac{1}{(2 - L\alpha_k)} \left(\mathbb{E}\,\|x_k - x^*\|^2 - \mathbb{E}\,\|x_{k+1} - x^*\|^2 + \alpha_k^2 \sigma^2\right)$$

$$\sum_{k=1}^{N} \alpha_k \left(E(f(x_k)) - f(x^*) \right) \leq$$

$$\sum_{k=1}^{N} \frac{1}{(2-L\alpha_k)} \left(E\|x_k - x^*\|^2 - E\|x_{k+1} - x^*\|^2 \right) + \sum_{k=1}^{N} \frac{\alpha_k^2 \sigma^2}{(2-L\alpha_k)}$$

Set $L\alpha_k \leq 1$

$$\sum_{k=1}^{N} \alpha_k \left(E(f(x_k) - f(x^*)) \right)$$

$$\leq \|x_1 - x^*\|^2 - E\|x_{N+1} - x^*\|^2$$

$$- \left(\sum_{k=2}^{N} \left(\frac{1}{2 - L\alpha_{k-1}} - \frac{1}{2 - L\alpha_k} \right) E\|x_k - x^*\|^2 \right)$$

Can be ignored

$$+ \sum_{k=1}^{N} \alpha_k^2 \sigma^2$$

$$E\, f(x_R) - f(x^*) \leq \frac{1}{\sum \alpha_k} \left(\|x_1 - x^*\|^2 + \sum \alpha_k^2 \sigma^2 \right)$$

$$\alpha_k = \alpha$$

$$E f(x_R) - f(x^*) \leq \frac{1}{N\alpha} \left(\|x_1 - x^*\|^2 + N\alpha^2\sigma^2 \right)$$

$$= \frac{\|x_1 - x^*\|^2}{N\alpha} + \alpha\sigma^2$$

$$\alpha = \frac{1}{\sqrt{N}} \Rightarrow \quad E\, f(x_R) - f(x^*) \leq \frac{\|x_1 - x^*\|^2}{\sqrt{N}} + \frac{\sigma^2}{\sqrt{N}}$$

Lecture - 19

SG for 200 case

↳ Convex

↳ Strongly convex

Algo: $x_{k+1} = \Pi\left(x_k - \alpha\, g(x_k, \xi_k)\right)$ → project onto a convex set C with diameter D ($\|x-y\| \le D \ \forall x, y \in C$)

Biased gradients: $\|E_k g - \nabla f\| \le C_1 \delta^2$

$E\|g - Eg\|^2 \le C_2/\delta^2$

Assume: (i) f is L-smooth & convex

(ii) $\|\nabla f(x)\| \le B$ ← Bounded gradients

$$\|x_{k+1} - x^*\|^2 = \|x_k - \alpha g(x_k, \xi_k) - x^*\|^2$$

$$= \|x_k - x^*\|^2 - 2\alpha\, g^T(x_k - x^*) + \alpha^2 \|g\|^2$$

Let $\Delta_k = g - \nabla f$

$$E_k\|x_{k+1} - x^*\|^2 \le \|x_k - x^*\|^2 - 2\alpha\, \nabla f^T(x_k - x^*) - 2\alpha\, \Delta_k^T(x_k - x^*) + \alpha^2\left(\|E_k g\|^2 + \frac{C_2}{\delta^2}\right)$$

$$\le \|x_k - x^*\|^2 - 2\alpha\, \nabla f^T(x_k - x^*) + 2\alpha C_1 \delta^2 \|x_k - x^*\| + \alpha^2\left(\|\nabla f\|^2 + 2\sqrt{2} C_1 \delta^2 \|\nabla f\| + \alpha C_1^2 \delta^4 + \frac{C_2}{\delta^2}\right)$$

(*)

Since $-\sum_{i=1}^{d} y_i \le \|y\|_1$

$$\|E_k g\|^2 = \|\nabla f + c_1 \delta^2 \mathbf{1}_d\|^2 \quad (\mathbf{1}_d \to \text{vector of ones})$$

$$\le \|\nabla f\|^2 + c_1^2 \delta^4 d + 2 c_1 \delta^2 \nabla f^T \mathbf{1}_d$$

$$\le \|\nabla f\|^2 + c_1^2 \delta^4 d + 2 c_1 \delta^2 \|\nabla f\| \sqrt{d}$$

Letting $z_k = x_k - x^*$ and using $\|\nabla f(x)\|^2 \le L \nabla f(x)^T (x - x^*)$

$(*) \Rightarrow$

$$E_k z_{k+1}^2 \le z_k^2 - 2\alpha \nabla f^T z_k + 2\alpha c_1 \delta^2 \|z_k\|_1 + \alpha^2 \left[L \nabla f^T z_k + 2\sqrt{d} c_1 \delta^2 L \|z_k\| + d c_1^2 \delta^4 + \frac{c_2}{\delta^2} \right]$$

Convex f: $f(x^*) \ge f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle$

Using this for

$$E_k z_{k+1}^2 \le z_k^2 - (2\alpha - L\alpha^2)(f(x_k) - f(x^*)) + 2\sqrt{d} \|z_k\|_2 c_1 \delta^2 (\alpha + L\alpha^2) + \alpha^2 \left[d c_1^2 \delta^4 + \frac{c_2}{\delta^2} \right]$$

(using $\|y\|_1 \le \sqrt{d} \|y\|_2$

Re-arranging,

$$\alpha (f(x_k) - f(x^*)) \le \frac{1}{2 - L\alpha} \left[z_k^2 - E_k z_{k+1}^2 + 2\sqrt{d} \|z_k\| c_1 \delta^2 (\alpha + L\alpha^2) + \alpha^2 \left(d c_1^2 \delta^4 + \frac{c_2}{\delta^2} \right) \right]$$

Setting $\alpha \le \frac{1}{L}$ & summing over $k = 1 \dots N$ & taking total expectations,

$$\sum_{k=1}^{N} \alpha \, E\left(f(x_k) - f(x^*)\right)$$

$$\le \sum_{k=1}^{N} \left(E z_k^2 - E z_{k+1}^2\right) + 2\sqrt{\alpha} \sum_{k=1}^{N} E\|z_k\| C_1 \delta^2 \left(\alpha + L\alpha^2\right)$$

$$+ \sum_{k=1}^{N} \alpha^2 \left(d C_1 \delta^2 + \frac{C_2}{\delta^2}\right)$$

$$\le z_1^2 + 2\sqrt{\alpha}\, D\, N\, C_1 \delta^2 \left(\alpha + L\alpha^2\right)$$

$$+ N\alpha^2 \left(d C_1^2 \delta^4 + \frac{C_2}{\delta^2}\right)$$

$$E\, f(x_R) - f(x^*) \qquad x_k \xrightarrow[\text{random}]{\text{unif. at}} \{x_1, \dots x_N\}$$

$$\le \frac{1}{N\alpha} \left[z_1^2 + 2\sqrt{\alpha}\, D\, N\, C_1 \delta^2 \left(\alpha + L\alpha^2\right) + \alpha^2 \left(d C_1^2 \delta^4 + \frac{C_2}{\delta^2}\right) \right]$$

Setting $\alpha = \min\left(\frac{1}{L}, \frac{1}{N^{2/3}}\right)$, $\delta = \frac{1}{N^{1/6}}$

$$E\, f(x_R) - f(x^*) \le \frac{\text{Const}}{N^{1/3}}$$

The claim follows. □

Remark 5.8. In contrast to the constant step size case handled previously, with a diminishing step size, we have a bound that vanishes as $m \to \infty$. However, the step size choice requires the knowledge of the strong convexity parameter μ, while the constant step size case in Theorem 5.4 did not assume such information. On a related note, it is possible to obtain a bound of $O\left(1/\sqrt{m}\right)$ with a step size choice that does not require the knowledge of μ, and more importantly, with a bound that does not scale inversely with μ. Such a bound may be preferable for ill-conditioned problems, where μ is very small. The reader is referred to (Nemirovski *et al.*, 2009) for the details.

5.3.2 SG with biased gradient information

As before, we consider the update iteration in (5.21). Unlike the previous section, where we assumed unbiased gradient estimates (i.e., the condition A5.1 holds), here the estimate $\widehat{\nabla} f(\theta_k)$ is a biased approximation to the gradient of the objective function f at θ_k.

As in the asymptotic analysis in Section 4.1.2, the biased gradient estimate $\widehat{\nabla} f(\theta_k)$ can be decomposed as follows:

$$\begin{aligned} \widehat{\nabla} f(\theta_k) &= \nabla f(\theta_k) + \beta_k + \eta_k, \text{ where} \\ \beta_k &= E\left[\widehat{\nabla} f(\theta_k) \mid \mathcal{F}_k\right] - \nabla f(\theta_k), \\ \eta_k &= \widehat{\nabla} f(\theta_k) - E\left[\widehat{\nabla} f(\theta_k) \mid \mathcal{F}_k\right], \end{aligned} \tag{5.29}$$

where $\mathcal{F}_k$ is a σ-field generated by $\{\theta_i, i \leq k\}$. In the above, β_k is the bias in the gradient estimate and η_k, $n \geq 0$, is a martingale difference sequence.

Using a simultaneous perturbation-based gradient estimate implies $\beta_k = O(\delta_k^2)$, where δ_k is the perturbation parameter used in forming the estimate (see Chapter 3 for several examples). While the bias goes down as δ_k^2, the variance of the gradient estimate scales inversely with δ_k^2. This has been formalized earlier in assumptions A5.2–A5.3.

We now present a non-asymptotic bound in expectation for the SG algorithm (5.21) with inputs from a biased gradient oracle that satisfies the aforementioned assumptions.

$\theta_{k+1} = \theta_k - a_k \hat{\nabla} f(\theta_k)$

f is μ-strongly convex

Proposition 5.1. Suppose the objective function f is L-smooth (see Definition 3.1), and assumptions A5.2–A5.3 hold. Then, we have

$$\mathbb{E}\,\|\theta_{m+1} - \theta^*\|^2 \leq \underbrace{2\exp(-2\mu\Gamma(m))\,\|\theta_0 - \theta^*\|^2}_{\textbf{initial error}} + \underbrace{2\sum_{k=1}^{n} a_k^2 \exp(-2\mu(\Gamma(m) - \Gamma_k))c_1^2\delta_k^4}_{\textbf{bias error}} + \underbrace{\sum_{k=1}^{n} a_k^2 \exp(-2\mu(\Gamma(m) - \Gamma_k))c_2\delta_k^{-2}}_{\textbf{sampling error}}, \tag{5.30}$$

where $\Gamma(k) := \sum_{i=1}^{k} a_i$.

$\int_0^1 f''(\theta^* + \lambda(\theta_m - \theta^*))(\theta_m - \theta^*)d\lambda = f'(\theta_m)$

Proof. Let $z_m = \theta_m - \theta^*$ denote the error at time instant n of the algorithm (5.21). Using $\nabla f(\theta^*) = 0$, we have

$$\left(\int_0^1 \nabla^2 f(\theta^* + \lambda(\theta_m - \theta^*))d\lambda\right) z_m = \nabla f(\theta_m).$$

Using the fact above, we arrive at a recursion for z_m from (5.29). Letting $J_m := \int_0^1 \nabla^2 f(\theta^* + \lambda(\theta_m - \theta^*)d\lambda$, we have

$$\begin{aligned} z_{m+1} &= (I - a(m)J_m)z_m - a(m)\,(\beta_m + \eta_m) \\ &= \Pi_m z_0 - \sum_{k=1}^{n} a(k)\Pi_m\Pi_k^{-1}(\beta_k + \eta_k), \end{aligned}$$

where $\Pi_m := \prod_{k=1}^{n} (I - a(k)J_k)$.

$\Pi_m\Pi_k^{-1}$ shorthand for $\prod_{j=k+1}^{n}(I - a(j)J_j)$

By the conditional Jensen's inequality, we obtain

$$\begin{aligned} (\mathbb{E}_m\,\|z_{m+1}\|)^2 &\leq \mathbb{E}_m(\langle z_m, z_m\rangle) \\ &= \mathbb{E}_m\left(\|\Pi_m z_0\|^2 + \left\|\sum_{k=1}^{n} a(k)\Pi_m\Pi_k^{-1}\beta_k\right\|^2 + \left\|\sum_{k=1}^{n} a(k)\Pi_m\Pi_k^{-1}\eta_k\right\|^2\right. \\ &\quad \left. -2\left\langle \Pi_m z_0, \sum_{k=1}^{n} a(k)\Pi_m\Pi_k^{-1}\beta_k\right\rangle - 2\left\langle \Pi_m z_0, \sum_{k=1}^{n} a(k)\Pi_m\Pi_k^{-1}\eta_k\right\rangle\right. \end{aligned}$$

$$+2\left\langle \sum_{k=1}^{n} a(k)\Pi_m\Pi_k^{-1}\beta_k, \sum_{k=1}^{n} a(k)\Pi_m\Pi_k^{-1}\eta_k \right\rangle\Bigg) \tag{5.31}$$

$$\leq 2\left\|\Pi_m z_0\right\|^2 + 2\sum_{k=1}^{n} a(k)^2 \left\|\Pi_m\Pi_k^{-1}\right\|^2 c_1^2\delta_k^4$$

$$+ \sum_{k=1}^{n} a(k)^2 \left\|\Pi_m\Pi_k^{-1}\right\|^2 \mathbb{E}\left\|\eta_k\right\|^2. \tag{5.32}$$

For the last inequality, we have used the following facts: (i) η_k is a martingale difference implying the last two cross terms in (5.31) are zero; (ii) $\beta_k \leq c_1\delta_k^2$ from A5.3; and (iii) Cauchy-Schwarz inequality for the first cross term in (5.31).

Now, we bound each of the square terms in (5.32) separately. Since the objective is strongly convex, we have that $\|I - a(m)J_m\| \leq \exp(-\mu a(m))$. Hence,

$$\begin{aligned}
\left\|\Pi_m\Pi_k^{-1}\right\|_2 &= \left\|\prod_{j=k+1}^{n}(I - a_j J_j)\right\|_2 \\
&\leq \prod_{j=k+1}^{n}\left\|(1-a_j\mu)I - a_j(J_j - \mu I)\right\|_2 \\
&\leq \prod_{j=k+1}^{n}\left\|(1-a_j\mu)I\right\|_2 \leq \prod_{j=k+1}^{n}(1-a_j\mu) \\
&\leq \exp\left(-\mu(\Gamma(m) - \Gamma(k))\right).
\end{aligned} \tag{5.33}$$

From A5.3, we can infer that the second moment of the martingale difference is bounded above by c_2/δ_k^2. The main claim now follows by plugging the bound on η_m and (5.33) into (5.32). □

By specializing the result in the proposition above, we derive a non-asymptotic bound of the order $O(1/\sqrt{m})$.

Theorem 5.6. (**Biased gradients and strongly convex objec-**

tive) Let $a(k) = c/k$ and $\delta_k = \delta_0/k^\delta$. Then,

$$\mathbb{E}\left\|\theta_m - \theta^*\right\| \leq \frac{\sqrt{2}\left\|\theta_0 - \theta^*\right\|}{m^{\mu c}} + \frac{\sqrt{2}cc_1\delta_0^2}{\sqrt{2\mu c - 4\delta - 1}} m^{-\frac{1}{2}-2\delta} + \frac{\sqrt{c_2}c}{\delta_0\sqrt{2\mu c + 2\delta - 1}} m^{\delta-\frac{1}{2}}.$$

Remark 5.9. Choosing $\delta = 0$, one can obtain a bound of the order $O\left(m^{-1/2}\right)$ for simultaneous perturbation schemes that lead to biased gradient estimates, and this bound matches the corresponding bound with unbiased gradient information up to constant factors. Contrast this with the difference in rates between biased and unbiased gradient information for the non-convex and convex cases in the previous sections.

Remark 5.10. Using L-smoothness of f and $\nabla f(\theta^*) = 0$, we have

$$\mathbb{E}\left[f\left(\theta_m\right)\right] - f(\theta^*) \leq \frac{L}{2}\mathbb{E}\left\|\theta_m - \theta^*\right\|^2 = O\left(\frac{1}{m}\right).$$

Proof. Bounding a sum by an integral, we obtain

$$\exp(-\mu\Gamma(m)) \leq \exp(-\mu c \ln m) = m^{-\mu c}.$$

Plugging $a(k) = c/k$ and $\delta_k = \delta_0/k^\delta$ into the bias error term in (5.30), we obtain

$$\begin{aligned}\sum_{k=1}^{m} a(k)^2 \exp(-2\mu(\Gamma(m) - \Gamma_k))c_1^2\delta_k^4 &\leq \sum_{k=1}^{m} \frac{c^2}{k^2} m^{-2\mu c} k^{2\mu c} c_1^2 \frac{\delta_0^4}{m^{4\delta}} \\ &\leq c^2 m^{-2\mu c} c_1^2\delta_0^4 \sum_{k=1}^{m} k^{2\mu c - 4\delta - 2} \\ &\leq \frac{c^2 c_1^2 \delta_0^4}{(2\mu c - 4\delta - 1)} m^{-1-4\delta}.\end{aligned}$$

Along similar lines, the sampling error term in (5.30) can be upper-bounded as follows:

$$\sum_{k=1}^{m} a(k)^2 \exp(-2\mu(\Gamma(m) - \Gamma_k))\frac{c_2}{\delta_k^2} \leq \frac{c^2 c_2}{\delta_0^2(2\mu c - 4\delta - 1)} m^{-1+2\delta}.$$

□