ELE 539 / COS 512: Homework 3

due on Mar. 30, 2021 (11:59 PM Blackboard)

1 SGD for Smooth and Strongly Convex Functions

Consider SGD of update form

$$x_{t+1} = x_t - \eta g(x_t)$$

We assume that the stochastic gradient satisfies the following conditions: (a) $\forall x, \mathbb{E}g(x) = \nabla f(x)$; (b) $\forall x, \mathbb{E}\|g(x) - \nabla f(x)\| \leq \sigma^2$. In this question, we consider the unconstrained problem, and aim to prove the following theorem.

Theorem 1. *There exists an absolute constant* c*, for any* α*-strongly convex and* ℓ*-smooth function* f*, SGD with learning rate* $\eta = \min\{\frac{1}{\ell}, \frac{\iota}{\alpha t}\}$ *and* $\iota = \max\{1, 2\ln\frac{\alpha t\|x_1 - x^\star\|}{\sigma}\}$ *satisfies the following:*

$$\mathbb{E}f\left(\sum_{s=2}^{t+1}\lambda_s x_s\right) - f(x^\star) \leq \frac{\ell e^{-t/\kappa}}{2}\|x_1 - x^\star\|^2 + \frac{2\sigma^2\iota}{\alpha t}.$$

where $\lambda_s = (1-\eta\alpha)^{t+1-s}/\sum_{s=2}^{t+1}(1-\eta\alpha)^{t+1-s}$.

(a) [2 points] Prove that for any $s \in [t]$, we have

$$\mathbb{E}\|x_{s+1} - x^\star\|^2 \leq (1-\eta\alpha)\mathbb{E}\|x_s - x^\star\|^2 - 2\eta\mathbb{E}[f(x_{s+1}) - f(x^\star)] + 2\eta^2\sigma^2$$

(b) [2 point] Prove the following inequality

$$\mathbb{E}\sum_{s=2}^{t+1}\lambda_s(f(x_s) - f(x^\star)) \leq \frac{e^{-\eta\alpha t}}{2\eta}\|x_1 - x^\star\|^2 + \eta\sigma^2.$$

(c) [2 points] Use above results to prove Theorem 1.

2 Catalyst Acceleration for Finite-sum Problems

In this question, you are allowed to directly use the convergence result for the following algorithm: APPA.

Accelerated Proximal Point Algorithm (APPA). To solve $\min_x g(x)$, the update of APPA is as follows:

$$\begin{aligned} y_t =& x_t + \gamma(x_t - x_{t-1}) \\ x_{t+1} \approx& \operatorname*{argmin}_x \{g(x) + \ell\|x - y_t\|^2\} \text{ up to error tolerance } \zeta. \end{aligned} \tag{1}$$

The algorithm is very similar to the Nesterov's AGD introduced in the lecture except that in the second step—Nesterov's AGD performs a gradient descent step while APPA computes the proximal step (1), i.e., the minimal point under ℓ_2 regularization that penalizes points far from y_t. By "$\bar{x} \approx \operatorname{argmin}_x h(x)$ up to error tolerance ζ" in (1), we mean the solution $\bar{x}$ satisfies $h(\bar{x}) \le \min_x h(x) + \zeta$. Similar to Nesterov's AGD, we have the following guarantee for APPA:

Theorem 2. *For any $\epsilon > 0$, assume g is α-strongly convex, and $\ell \ge 0$, there exist choices of hyperparameters $\zeta = \Theta(\epsilon/\kappa^2)$, $\gamma = 1 - \Theta(1/\sqrt{\kappa})$ where $\kappa = (\ell + \alpha)/\alpha$, so that APPA satisfies $g(x_T) - g(x^\star) \le \epsilon$, after*

$$T = \mathcal{O}\left(\sqrt{\kappa} \cdot \log \frac{g(x_1) - g(x^\star)}{\epsilon}\right) \quad \textit{iterations}$$

We note here ℓ is not the smoothness of function g, but rather a arbitrary hyperparameter of APPA. Above Theorem provides the iteration complexity of APPA. It does not take into account the complexity of solving the proximal step (1).

Algorithm: Catalyst. To optimize $F(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$, the algorithm is described as follows

Algorithm 1 CATALYST

1: **for** $t = 1, \cdots, T$ **do**
2: $\quad y_t = x_t + \gamma(x_t - x_{t-1})$.
3: $\quad$ Use SVRG to solve $x_{t+1} \approx \operatorname{argmin}_x\{F(x) + \lambda\|x - y_t\|^2\}$ up to error tolerance ζ.
4: **Output:** $\mathbf{x}_{T+1}$.

In a high-level, Catalyst is APPA with SVRG for solving the inner proximal step. Let initialization $x_0 = x_1$. In this question, we aim to prove the following theorem:

Theorem 3. *For any $\epsilon, \delta > 0$, assume f_i is ℓ-smooth for any $i \in [n]$, and $F = \frac{1}{n}\sum_{i=1}^n f_i$ is α-strongly convex, then the output $\hat{x}$ of Catalyst will satisfies $F(\hat{x}) - F(x^\star) \le \epsilon$ with probability $1 - \delta$, after using*

$$\mathcal{O}\left(\left(n + \sqrt{\frac{n\ell}{\alpha}}\right) \log^2\left(\frac{[\max_{t\in[T]} F(y_t) - F(x^\star)]}{\epsilon} \cdot \frac{T}{\delta} \cdot \frac{\ell}{\alpha}\right)\right) \quad \textit{stochastic gradient queries}$$

We remark the complexity in Theorem 3 is always better than SVRG up to logarithmic factors. [1]

(a) [1 point] Case 1: $\ell \le (n+1)\alpha$, prove Theorem 3 by choosing appropriate λ and T.

(b) [2 points] Case 2: $\ell > (n+1)\alpha$, choose $\lambda = (\ell - \alpha)/n - \alpha$, compute the number of stochastic gradients required for SVRG to solve step 3 in Algorithm 1 with probability at least $1 - \delta$. [Hint: you can use the SVRG guarantee taught in the lecture.]

[1] The logarithmic factors in Theorem 3 can be further improved by warm-start for the inner SVRG subroutine.

(c) [2 points] Under case 2, and the choice of λ as in (b), compute the number of iterations T required for the outer loop (APPA) to guarantee $F(x_{T+1}) - F(x^\star) \le \epsilon$.

(d) [2 points] Combine above results to prove Theorem 3.