ANITA: An Optimal Loopless Accelerated Variance-Reduced Gradient Method

Zhize Li ZHIZE.LI@KAUST.EDU.SA
KAUST, Saudi Arabia

Abstract

In this paper, we propose a novel accelerated gradient method called ANITA for solving the fundamental finite-sum optimization problems. Concretely, we consider both general convex and strongly convex settings: i) For general convex finite-sum problems, ANITA improves previous state-of-the-art result given by Varag [17]. In particular, for large-scale problems or the target error is not very small, i.e., $n \geq \frac{1}{\epsilon^2}$, ANITA obtains the *first* optimal result $O(n)$, matching the lower bound $\Omega(n)$ provided by Woodworth and Srebro [46], while previous results are $O(n \log \frac{1}{\epsilon})$ of Varag [17] and $O(\frac{n}{\sqrt{\epsilon}})$ of Katyusha [1]. ii) For strongly convex finite-sum problems, we also show that ANITA can achieve the optimal convergence rate $O\big((n+\sqrt{\frac{nL}{\mu}}) \log \frac{1}{\epsilon}\big)$ matching the lower bound $\Omega\big((n+\sqrt{\frac{nL}{\mu}}) \log \frac{1}{\epsilon}\big)$ provided by Lan and Zhou [15]. Besides, ANITA enjoys a simpler loopless algorithmic structure unlike previous accelerated algorithms such as Varag [17] and Katyusha [1] where they use an inconvenient double-loop structure. Moreover, we provide a new *dynamic multi-stage convergence analysis*, which is the key technical part for improving previous results to the optimal rates. Finally, the numerical experiments show that ANITA converges faster than the previous state-of-the-art Varag [17], validating our theoretical results and confirming the practical superiority of ANITA. We believe that our new theoretical rates and convergence analysis for this fundamental finite-sum problem will directly lead to key improvements for many other related problems, such as distributed/federated/decentralized optimization problems. For instance, Li and Richtárik [26] obtain the first compressed and accelerated result, substantially improving previous state-of-the-art results, by applying ANITA to the distributed optimization problems with compressed communication.

1. Introduction

In this paper, we consider the fundamental finite-sum problems of the form

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x), \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a smooth and convex function. We consider two settings in this paper, i) general convex setting ($\mu = 0$); ii) strongly convex setting ($\mu > 0$), where μ is the strongly convex parameter for $f(x)$, i.e., $f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq \frac{\mu}{2}\|x - y\|^2$. Note that the case $\mu = 0$ reduces to the standard convexity. Also note that the strong convexity is only corresponding to the average function f, is not needed for these component functions f_is.

Finite-sum problem (1) captures the standard empirical risk minimization (ERM) problems in machine learning [42]. There are n data samples and f_i denotes the loss associated with i-th data sample, and the goal is to minimize the loss over all data samples. This optimization problem has found a wide range of applications in machine learning, statistical inference, and image processing. In recent years, there has been extensive research in designing gradient-type methods for solving this problem (1). To measure the efficiency of algorithms for solving (1), it is standard to bound the number of stochastic gradient computations for finding a suitable solution. In particular, our goal is to find a point $\widehat{x} \in \mathbb{R}^d$ such that $\mathbb{E}[f(\widehat{x}) - f(x^*)] \le \epsilon$, where the expectation is with respect to the randomness inherent in the algorithm. We use the term *ϵ-approximate solution* to refer to such a point $\widehat{x}$, and use the term *stochastic gradient complexity* to describe the convergence result (convergence rate) of algorithms.

Two of the most classical gradient-type algorithms are gradient descent (GD) and stochastic gradient descent (SGD) (e.g., 7, 10, 11, 14, 33, 34, 36). However, GD requires to compute the full gradient over all n data samples for each iteration ($x_{t+1} = x_t - \eta\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_t)$) which is inefficient especially for large-scale machine learning problems where n is very large. Although SGD only needs to compute a single stochastic gradient (e.g., $\nabla f_i(x)$) for each iteration ($x_{t+1} = x_t - \eta\nabla f_i(x_t)$), it requires an additional bounded variance assumption for the stochastic gradients (i.e.,$\exists\sigma > 0$, $\mathbb{E}_i[\|\nabla f_i(x) - \nabla f(x)\|^2] \le \sigma^2$) since it does not compute the full gradients ($\nabla f(x)$, i.e., $\frac{1}{n}\sum_{i=1}^n \nabla f_i(x)$). More importantly, for strongly convex problems, SGD only obtains a sublinear convergence rate $O(\frac{\sigma^2}{\mu\epsilon})$ rather than a linear rate $O(\cdot \log\frac{1}{\epsilon})$ achieved by GD.

To remedy the variance term $\mathbb{E}[\|\nabla f_i(x) - \nabla f(x)\|^2]$ in SGD, the variance reduction technique has been proposed and it has been widely-used in a lot of algorithms in recent years. In particular, Le Roux et al. [18] (later version [41]) propose the first variance-reduced algorithm called SAG and show that by incorporating new gradient estimators into SGD one can possibly achieve the linear convergence rate for strongly convex problems. Then this variance reduction direction is followed by many works such as [6, 12, 31, 32, 37, 43]. Particularly, SAG [18] uses a biased gradient estimator while SAGA [6] modifies it to an unbiased estimator and provides better convergence results. Johnson and Zhang [12] propose a novel unbiased stochastic variance reduced gradient (SVRG) method which directly incorporates the full gradient term $\nabla f(x)$ into SGD. More specifically, each epoch of SVRG starts with the computation of the full gradient $\nabla f(\tilde{x})$ at a snapshot point $\tilde{x} \in \mathbb{R}^n$ and then runs SGD for a fixed number of steps using the modified stochastic gradient estimator

$$\widetilde{\nabla}_t = \nabla f_i(x_t) - \nabla f_i(\tilde{x}) + \nabla f(\tilde{x}), \tag{2}$$

i.e., $x_{t+1} = x_t - \eta\widetilde{\nabla}_t$, where i is randomly picked from $\{1, 2, \ldots, n\}$. In particular, if each full gradient $\nabla f(\tilde{x})$ (which requires n stochastic gradient computations) at the snapshot point $\tilde{x}$ is reused for n iterations (i.e., $\tilde{x}$ is changed after every n iterations), then the amortized stochastic gradient computations for each iteration is the same as SGD. Note that $\mathbb{E}[\widetilde{\nabla}_t] = \nabla f(x_t)$ is an unbiased estimator, and its variance $\mathbb{E}[\|\widetilde{\nabla}_t - \nabla f(x_t)\|^2] \le 4L\big(f(x_t) - f(x^*) + f(\tilde{x}) - f(x^*)\big)$ is reduced as the algorithm converges $x_t, \tilde{x} \to x^*$, while the variance term is uncontrollable for plain SGD where $\widetilde{\nabla}_t = \nabla f_i(x_t)$. Johnson and Zhang [12] also show that SVRG obtains the linear convergence $O((n + \frac{L}{\mu})\log\frac{1}{\epsilon})$ which can be better than the sublinear convergence rate $O(\frac{\sigma^2}{\mu\epsilon})$ of plain SGD, for strongly convex problems. The SVRG gradient estimator (2) is adopted in many algorithms (e.g., 2, 4, 9, 13, 19, 20, 23, 25, 39, 40, 47, 48) and also is used in our ANITA.

The aforementioned variance-reduced methods are not accelerated and hence they do not achieve the optimal convergence rates for convex finite-sum problem (1). See the non-accelerated variance-reduced algorithms listed in the first part of Table 1, i.e., SAG, SVRG, SAGA and SVRG^{++}, they do not achieve the accelerated rates, i.e., $\frac{L}{\mu}$ vs. $\sqrt{\frac{L}{\mu}}$ (strongly convex case) and $\frac{L}{\epsilon}$ vs. $\sqrt{\frac{L}{\epsilon}}$ (general convex case). Note that we do not list the SCSG [19] and SARAH [37] in Table 1 since SCSG requires an additional bounded variance assumption (without this assumption, its result is the same as SVRG and SAGA) and SARAH uses $\mathbb{E}[\|\nabla f(\widehat{x})\|^2] \le \epsilon$ as the convergence criterion which can not be directly converted to $\mathbb{E}[f(\widehat{x}) - f(x^*)] \le \epsilon$. SARAH is usually used for solving nonconvex problems where the convergence criterion is typically the norm of gradient (e.g., 8, 21, 22, 27, 29, 38, 45). Also both SCSG and SARAH are non-accelerated methods and thus do not achieve the optimal convergence results. Therefore, much recent research effort has been devoted to the design of accelerated gradient methods (e.g., 1, 3, 14, 16, 17, 24, 28, 30, 36, 44). As can be seen from Table 1, for strongly convex finite-sum problems, existing accelerated methods such as RPDG [15], Katyusha [1], Varag [17] and our ANITA are optimal since their convergence results are $O\big((n+\sqrt{\frac{nL}{\mu}})\log\frac{1}{\epsilon}\big)$ matching the lower bound $\Omega\big((n+\sqrt{\frac{nL}{\mu}})\log\frac{1}{\epsilon}\big)$ given by Lan and Zhou [15].

However, for general (non-strongly) convex finite-sum problems, all previous accelerated methods do not achieve the optimal convergence result. In particular, Varag [17] obtains the current best result $O\big(n\min\{\log\frac{1}{\epsilon},\log n\}+\sqrt{\frac{nL}{\epsilon}}\big)$, while the lower bound in this general convex case is $\Omega\big(n+\sqrt{\frac{nL}{\epsilon}}\big)$ provided by Woodworth and Srebro [46]. More importantly, for large-scale problems where the number of data samples n is very large, or the target error ϵ is not very small, then the convergence result of Varag is $O(n\log\frac{1}{\epsilon})$ which is not optimal since the lower bound is $\Omega(n)$ (see Table 2). Note that the case of large-scale problems or the case of moderate target error often exists in machine learning applications. We show that our ANITA takes an important step towards the ultimate limit of accelerated methods and it is the first algorithm to achieve the optimal convergence rate $O(n)$ in this case matching the lower bound $\Omega(n)$. See Table 1 and Table 2 for more details.

2. Our Contributions

In this paper, we propose a novel simple accelerated variance-reduced gradient method, called ANITA (Algorithm 1), for solving both general convex and strongly convex finite-sum problems given in the form of (1). Table 1 and Table 2 summarize the convergence results of ANITA and previous algorithms. The proposed ANITA takes an important step towards the ultimate limit of accelerated methods and can achieve the optimal convergence rates.

As we mentioned before (see the last paragraph of Section 1), although accelerated methods have been widely studied in the optimization and machine learning literature, the limit of accelerated methods is still not be achieved for general convex finite-sum problems. Especially for the case of large-scale finite-sum problems or moderate target error, they do not achieve the optimal result $O(n)$. Motivated by this, in this paper we mainly focus on further improving the convergence result in order to close the gap between the upper bound and lower bound. Now, we highlight the following results achieved by ANITA:

- For general convex problems, ANITA obtains the rate $O\big(n\min\big\{1+\log\frac{1}{\epsilon\sqrt{n}},\ \log\sqrt{n}\big\}+\sqrt{\frac{nL}{\epsilon}}\big)$ for finding an ϵ-approximate solution of problem (1), which improves previous best result $O\big(n\min\{\log\frac{1}{\epsilon},\ \log n\}+\sqrt{\frac{nL}{\epsilon}}\big)$ given by Varag [17] (see the 'general convex' column of Table 1). Moreover, for a very wide range of ϵ, i.e., $\epsilon\in(0,\frac{L}{n\log^2\sqrt{n}}]\cup[\frac{1}{\sqrt{n}},+\infty)$, or the number of data

Table 1: Convergence rates for finding an ϵ-approximate solution $\mathbb{E}[f(\widehat{x}) - f(x^*)] \le \epsilon$ of (1)

Algorithms	μ-strongly convex	General convex	Loopless (Simple)
GD	$O\left(\frac{nL}{\mu}\log\frac{1}{\epsilon}\right)$	$O\left(\frac{nL}{\epsilon}\right)$	Yes
Nesterov's accelerated GD [35, 36]	$O\left(n\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	$O\left(n\sqrt{\frac{L}{\epsilon}}\right)$	Yes
SAG [18]	$O\Big(\big(n + n^2\lfloor\frac{L}{n\mu}\rfloor\big)\log\frac{1}{\epsilon}\Big)$	—	Yes
SVRG [12]	$O\left(\left(n+\frac{L}{\mu}\right)\log\frac{1}{\epsilon}\right)$	—	No
SAGA [6]	$O\left(\left(n+\frac{L}{\mu}\right)\log\frac{1}{\epsilon}\right)$	$O\left(\frac{n+L}{\epsilon}\right)$	Yes
SVRG++ [4]	—	$O\left(n\log\frac{1}{\epsilon}+\frac{L}{\epsilon}\right)$	No
RPDG [15]	$O\left(\left(n+\sqrt{\frac{nL}{\mu}}\right)\log\frac{1}{\epsilon}\right)$	$O\left(\left(n+\sqrt{\frac{nL}{\epsilon}}\right)\log\frac{1}{\epsilon}\right)$ [1]	Yes
Catalyst [30]	$O\left(\left(n+\sqrt{\frac{nL}{\mu}}\right)\log\frac{1}{\epsilon}\right)$ [1]	$O\left(\left(n+\sqrt{\frac{nL}{\epsilon}}\right)\log^2\frac{1}{\epsilon}\right)$ [1]	No
Katyusha [1]	$O\left(\left(n+\sqrt{\frac{nL}{\mu}}\right)\log\frac{1}{\epsilon}\right)$	$O\left(n\log\frac{1}{\epsilon}+\sqrt{\frac{nL}{\epsilon}}\right)$ [1]	No
Katyusha$^{\text{ns}}$ [1]	—	$O\left(\frac{n}{\sqrt{\epsilon}}+\sqrt{\frac{nL}{\epsilon}}\right)$	No
Varag [17]	$O\left(\left(n+\sqrt{\frac{nL}{\mu}}\right)\log\frac{1}{\epsilon}\right)$	$O\Big(n\min\big\{\log\frac{1}{\epsilon},\ \log n\big\}+\sqrt{\frac{nL}{\epsilon}}\Big)$	No
ANITA (this paper)	$O\left(\left(n+\sqrt{\frac{nL}{\mu}}\right)\log\frac{1}{\epsilon}\right)$	$O\Big(n\min\big\{1+\log\frac{1}{\epsilon\sqrt{n}},\ \log\sqrt{n}\big\}+\sqrt{\frac{nL}{\epsilon}}\Big)$	Yes
	$O\left(\left(n+\sqrt{\frac{nL}{\mu}}\right)\log\frac{1}{\epsilon}\right)$	$O\left(n+\sqrt{\frac{nL}{\epsilon}}\right)$ [2]	Yes
Lower bound	$\Omega\left(\left(n+\sqrt{\frac{nL}{\mu}}\right)\log\frac{1}{\epsilon}\right)$ [15]	$\Omega\left(n+\sqrt{\frac{nL}{\epsilon}}\right)$ [46]	—

[1] These gradient complexity bounds are obtained via indirect approaches, i.e., by adding strongly convex perturbation.
[2] ANITA can achieve this optimal result for a very wide range of ϵ, i.e., $\epsilon \in (0, \frac{L}{n\log^2\sqrt{n}}] \cup [\frac{1}{\sqrt{n}}, +\infty)$ or the number of data samples $n \in (0, \frac{L}{\epsilon\log^2\sqrt{n}}] \cup [\frac{1}{\epsilon^2}, +\infty)$ (see Table 2 for more details). Note that the term $\min\{\log\frac{1}{\epsilon},\ \log n\}$ in Varag [17] cannot be removed regardless of the value of ϵ or n. Thus ANITA is the first accelerated algorithm that can exactly achieve the optimal convergence result.

Table 2: Direct accelerated stochastic algorithms for *general convex setting* wrt. ϵ

Algorithms	The target error ($\mathbb{E}[f(\widehat{x}) - f(x^*)] \le \epsilon$): large $\epsilon \longrightarrow$ small ϵ (or the number of data samples: large $n \longrightarrow$ small n)			
	$\epsilon \ge \frac{1}{\sqrt{n}}$ (or $n \ge \frac{1}{\epsilon^2}$)	$\frac{1}{\sqrt{n}} > \epsilon \ge \frac{1}{n}$ (or $\frac{1}{\epsilon^2} > n \ge \frac{1}{\epsilon}$)	$\frac{1}{n} > \epsilon \ge \frac{L}{n\log^2\sqrt{n}}$ (or $\frac{1}{\epsilon} > n \ge \frac{L}{\epsilon\log^2\sqrt{n}}$)	$\frac{L}{n\log^2\sqrt{n}} > \epsilon$ (or $\frac{L}{\epsilon\log^2\sqrt{n}} > n$)
Katyusha$^{\text{ns}}$ [1]	$O\left(\frac{n}{\sqrt{\epsilon}}\right)$	$O\left(\frac{n}{\sqrt{\epsilon}}\right)$	$O\left(\frac{n}{\sqrt{\epsilon}}\right)$	$O\left(\frac{n}{\sqrt{\epsilon}}+\sqrt{\frac{nL}{\epsilon}}\right)$
Varag [17]	$O\left(n\log\frac{1}{\epsilon}\right)$	$O\left(n\log\frac{1}{\epsilon}\right)$	$O\left(n\log n\right)$	$O\left(n\log n+\sqrt{\frac{nL}{\epsilon}}\right)$
ANITA (this paper)	$O\left(n\right)$	$O\left(n\left(1+\log\frac{1}{\epsilon\sqrt{n}}\right)\right)$	$O\left(n\log\sqrt{n}\right)$	$O\left(\sqrt{\frac{nL}{\epsilon}}\right)$
Lower bound [46]	$\Omega\left(n\right)$	$\Omega\left(n\right)$	$\Omega\left(n\sqrt{\frac{L}{\epsilon n}}\right)$	$\Omega\left(\sqrt{\frac{nL}{\epsilon}}\right)$

Remark: ANITA achieves the optimal result $O(n)$ for large-scale problems (large n) or moderate target error (not too small ϵ). It should be pointed out that all parameter settings of ANITA (i.e., $\{p_t\}$, $\{\theta_t\}$, $\{\eta_t\}$, and $\{\alpha_t\}$ in Algorithm 1) do not require the value of ϵ in advance. The convergence rate of ANITA will automatically switch to different results listed in Table 2.

samples $n \in (0, \frac{L}{\epsilon \log^2 \sqrt{n}}] \cup [\frac{1}{\epsilon^2}, +\infty)$, ANITA can exactly achieve the optimal convergence result $O\big(n + \sqrt{\frac{nL}{\epsilon}}\big)$ matching the lower bound $\Omega\big(n + \sqrt{\frac{nL}{\epsilon}}\big)$ provided by Woodworth and Srebro [46] (see Table 1 and its Footnote 2).

• In particular, we would like to point out that none of previous algorithms with/without acceleration can obtain the optimal result $O(n)$ for finite-sum problems (1) where the number of data samples is very large or the target error is not very small, ANITA is the first algorithm that achieves the optimal result $O(n)$ for these typical machine learning problems (see the second column of Table 2 and its Remark).

• We also note that ANITA is the first loopless direct accelerated stochastic algorithm for solving general convex finite-sum problems, while previous accelerated stochastic algorithms use indirect approaches (RPDG, Catalyst, Katyusha) and/or use inconvenient double-loop algorithmic structures (Katyusha$^{\text{ns}}$, Varag) (see Table 1). Moreover, by exploiting the loopless structure of ANITA, we provide a new *dynamic multi-stage convergence analysis* which is the key technical part for improving previous results to the optimal rates.

• For strongly convex finite-sum problems (i.e., under strong convexity Assumption 2), we also prove that ANITA achieves the optimal convergence rate $O\big((n + \sqrt{\frac{nL}{\mu}}) \log \frac{1}{\epsilon}\big)$ matching the lower bound $\Omega\big((n + \sqrt{\frac{nL}{\mu}}) \log \frac{1}{\epsilon}\big)$ provided by Lan and Zhou [15] (see Table 1).

• Finally, the experiments show that ANITA converges faster than the previous state-of-the-art Varag [17], validating our theoretical results and confirming the practical superiority of ANITA.

2.1. ANITA algorithm

In this section, we describe the simple novel ANITA method in Algorithm 1. Note that μ is the strongly convex parameter (see Assumption 2). We point out that Algorithm 1 can deal with *both* general convex ($\mu = 0$) and strongly convex ($\mu > 0$) problems.

In each iteration t, the stochastic gradient estimator $\widetilde{\nabla}_t$ of ANITA (Line 5 of Algorithm 1) uses the gradient information of only one randomly sampled function f_i. Note that for the last term $\nabla f(w_t)$, it reuses previous $\nabla f(w_{t-1})$ with probability $1-p_{t-1}$ or needs to compute the full gradient

Algorithm 1 ANITA

Input: initial point x_0, parameters $\{p_t\}$, $\{\theta_t\}$, $\{\eta_t\}$, $\{\alpha_t\}$

1: $w_0 = \bar{x}_0 = \underline{x}_0 = x_0$
2: **for** $t = 0, 1, 2, \ldots, T-1$ **do**
3: $\quad \underline{x}_t = \theta_t x_t + (1-\theta_t) w_t$
4: $\quad$ Randomly pick $i \in \{1, 2, \ldots, n\}$
5: $\quad \widetilde{\nabla}_t = \nabla f_i(\underline{x}_t) - \nabla f_i(w_t) + \nabla f(w_t)$
6: $\quad x_{t+1} = \frac{1}{1+\mu\eta_t}(x_t + \mu\eta_t \underline{x}_t) - \frac{\eta_t}{\alpha_t} \widetilde{\nabla}_t$
7: $\quad \bar{x}_{t+1} = \theta_t x_{t+1} + (1-\theta_t) w_t$
8: $\quad w_{t+1} = \begin{cases} \bar{x}_{t+1} & \text{with probability } p_t \\ w_t & \text{with probability } 1 - p_t \end{cases}$
9: **end for**

Output: w_T

$\nabla f(\bar{x}_t)$ with probability p_{t-1} (see Line 8). Thus we know that ANITA uses $(n+2)p_{t-1}+2(1-p_{t-1})$ stochastic gradients in expectation for iteration t. In particular, if $p_t \equiv \frac{1}{n}$, then ANITA only uses constant stochastic gradients for each iteration which maintains the same computational cost as SGD. The snapshot point w_t is updated in the last Line 8, it is a probabilistic step which is the key part for removing double-loop structures to obtain a simple loopless algorithm, similar to [13, 29].

However, beyond the two interpolation steps (momentum) (see Line 3 and 7), we propose a new dynamic multi-stage convergence analysis which uses a dynamic control of the probability $\{p_t\}$ in Line 8, unlike directly fixing it to a constant $p_t \equiv p$ as in [13, 29]. This is also the first time that a loopless algorithm uses a dynamic control of $\{p_t\}$. More importantly, our new convergence analysis exploiting this dynamic multiple stages can lead to better convergence rates.

3. Convergence Results for ANITA

Here we state two Corollaries 1 and 2 (from Theorems 3 and 4) for solving finite-sum problems (1) in the general convex and strongly convex settings, respectively. The main Theorems 3 and 4 and all detailed proofs are deferred to the appendix.

Corollary 1 (General convex case) *Suppose that Assumption 1 holds. Choose the parameters $\{p_t\}$, $\{\theta_t\}$, $\{\eta_t\}$, $\{\alpha_t\}$ as stated in Theorem 3. Then ANITA (Algorithm 1) can find an ϵ-approximate solution for problem (1) such that*

$$\mathbb{E}[f(w_T) - f(x^*)] \le \epsilon$$

within T iterations, where

$$T \le \begin{cases} 2n & \text{if } \epsilon \ge O(\frac{1}{n}) \\ n + \sqrt{\frac{24(n+3)L\|x_0 - x^*\|^2}{\epsilon}} & \text{if } \epsilon < O(\frac{1}{n}) \end{cases},$$

and the number of stochastic gradient computations can be bounded by

$$\#\text{grad} = O\left(n \min\left\{1 + \log\frac{1}{\epsilon\sqrt{n}},\ \log\sqrt{n}\right\} + \sqrt{\frac{nL}{\epsilon}}\right).$$

Remark: Note that all parameter settings $\{p_t\}$, $\{\theta_t\}$, $\{\eta_t\}$, $\{\alpha_t\}$ of ANITA in Corollary 1 (Theorem 3) do not require the value of ϵ in advance. The convergence rate of ANITA will automatically switch to different results as stated in Table 2.

Corollary 2 (Strongly convex case) *Suppose that Assumptions 1 and 2 hold. Choose the parameters $\{p_t\}$, $\{\theta_t\}$, $\{\eta_t\}$, $\{\alpha_t\}$ as stated in Theorem 4. Then ANITA (Algorithm 1) can find an ϵ-approximate solution for problem (1) such that $\mathbb{E}[f(w_T) - f(x^*)] \le \epsilon$ within T iterations, where*

$$T \le \frac{5}{4p\theta}\log\frac{\Phi_0}{\epsilon}.$$

Moreover, by choosing $p = \frac{1}{n}$, the number of stochastic gradient computations can be bounded by

$$\#\text{grad} = O\left(\max\left\{n, \sqrt{\frac{nL}{\mu}}\right\}\log\frac{1}{\epsilon}\right).$$

4. Experiments

In the experiments, we consider the following logistic regression problem:

$$\min_{x\in\mathbb{R}^d} f(x) := \frac{1}{n}\sum_{i=1}^{n} \log\left(1+\exp(-b_i a_i^T x)\right), \tag{3}$$

where $\{a_i, b_i\}_{i=1}^n \in \mathbb{R}^d \times \{\pm 1\}$ are data samples. All datasets used in our experiments are downloaded from LIBSVM [5].

We present the numerical experiments of ANITA (Algorithm 1) compared with previous state-of-the-art Varag [17]. We also present the standard gradient descent (GD) as a benchmark. We directly use the parameter settings according to the theoretical convergence theorems or corollaries of these algorithms, i.e., we do not tune any hyperparameters. Note that for the logistic function in (3), one can precompute the smoothness parameter L satisfying Assumption 1, i.e., $L \le 1/4$ if the data samples are normalized. Given the parameter L, we are ready to set all other hyperparameters for GD (Corollary 2.1.2 in [36]), for Varag (Theorem 1 in [17]) and for ANITA (our Theorem 3).

In the following Figure 1, the x-axis and y-axis represent the number of data passes (i.e., we compute n stochastic gradients for each data pass) and the training loss, respectively. The numerical results presented in Figure 1 are conducted on different datasets. Each plot corresponds to one dataset (six datasets in total). The experimental results show that ANITA indeed converges faster than Varag [17] in the earlier stage (moderate target error), validating our theoretical results (see the second column of Table 2 and its Remark). More importantly, ANITA is the first accelerated algorithm which can obtain the optimal convergence result $O(n)$ in this range. Besides, ANITA also enjoys a simpler loopless algorithmic structure while Varag uses an inconvenient double-loop structure.

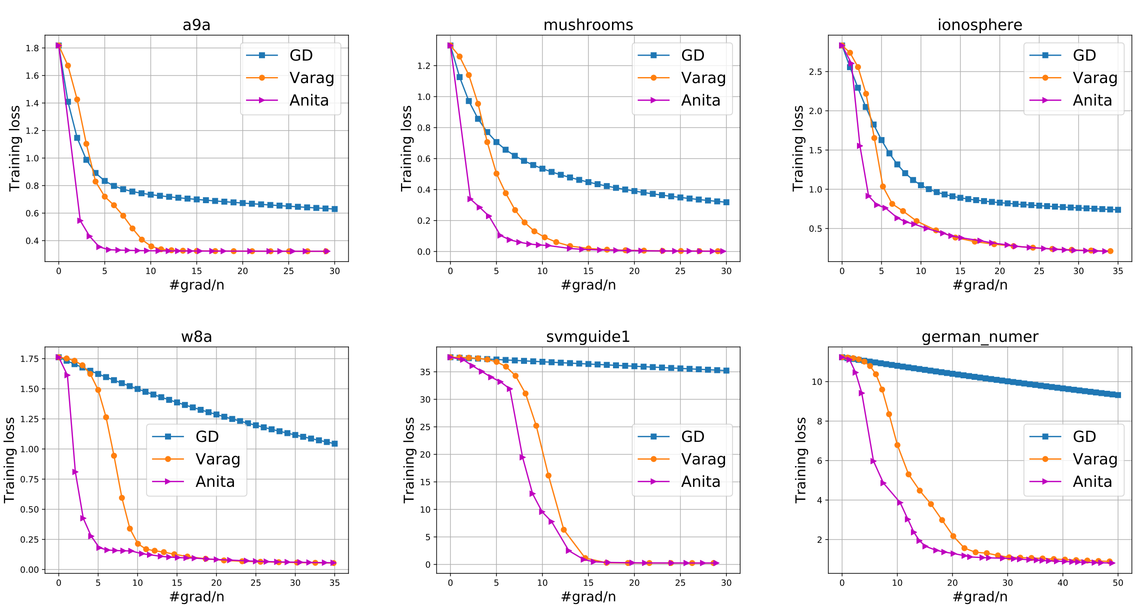

Figure 1: The convergence performance of GD, Varag and ANITA under different datasets.

References

[1] Zeyuan Allen-Zhu. Katyusha: the first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1200–1205. ACM, 2017.

[2] Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, pages 699–707, 2016.

[3] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

[4] Zeyuan Allen-Zhu and Yang Yuan. Improved SVRG for non-strongly-convex or sum-of-non-convex objectives. *arXiv preprint arXiv:1506.01972*, 2015.

[5] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

[6] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[7] John C Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *Conference on Learning Theory*, pages 14–26, 2010.

[8] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 687–697, 2018.

[9] Rong Ge, Zhize Li, Weiyao Wang, and Xiang Wang. Stabilized SVRG: Simple variance reduction for nonconvex optimization. In *Conference on Learning Theory*, pages 1394–1448, 2019.

[10] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[11] Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

[12] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[13] Dmitry Kovalev, Samuel Horváth, and Peter Richtárik. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, 2020.

[14] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

[15] Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *arXiv preprint arXiv:1507.02000*, 2015.

[16] Guanghui Lan and Yi Zhou. Random gradient extrapolation for distributed and stochastic optimization. *SIAM Journal on Optimization*, 28(4):2753–2782, 2018.

[17] Guanghui Lan, Zhize Li, and Yi Zhou. A unified variance-reduced accelerated gradient method for convex optimization. In *Advances in Neural Information Processing Systems*, pages 10462–10472, 2019.

[18] Nicolas Le Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems*, pages 2663–2671, 2012.

[19] Lihua Lei and Michael I Jordan. Less than a single pass: Stochastically controlled stochastic gradient method. *arXiv preprint arXiv:1609.03261*, 2016.

[20] Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via SCSG methods. In *Advances in Neural Information Processing Systems*, pages 2345–2355, 2017.

[21] Zhize Li. SSRGD: Simple stochastic recursive gradient descent for escaping saddle points. In *Advances in Neural Information Processing Systems*, pages 1521–1531, *arXiv:1904.09265*, 2019.

[22] Zhize Li. A short note of PAGE: Optimal convergence rates for nonconvex optimization. *arXiv preprint arXiv:2106.09663*, 2021.

[23] Zhize Li and Jian Li. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 5569–5579, *arXiv:1802.04477*, 2018.

[24] Zhize Li and Jian Li. A fast Anderson-Chebyshev acceleration for nonlinear optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1047–1057. PMLR, *arXiv:1809.02341*, 2020.

[25] Zhize Li and Peter Richtárik. A unified analysis of stochastic gradient methods for nonconvex federated optimization. *arXiv preprint arXiv:2006.07013*, 2020.

[26] Zhize Li and Peter Richtárik. CANITA: Faster rates for distributed convex optimization with communication compression. In *Advances in Neural Information Processing Systems*, 2021. arXiv:2107.09461.

[27] Zhize Li and Peter Richtárik. ZeroSARAH: Efficient nonconvex finite-sum optimization with zero full gradient computation. *arXiv preprint arXiv:2103.01447*, 2021.

[28] Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtárik. Acceleration for compressed gradient descent in distributed and federated optimization. In *International Conference on Machine Learning*, pages 5895–5904. PMLR, *arXiv:2002.11364*, 2020.

[29] Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pages 6286–6295. PMLR, *arXiv:2008.10898*, 2021.

[30] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pages 3384–3392, 2015.

[31] Julien Mairal. Optimization with first-order surrogate functions. In *International Conference on Machine Learning*, pages 783–791. PMLR, 2013.

[32] Julien Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

[33] Arkadi Nemirovski and David Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley, XV, 1983.

[34] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4): 1574–1609, 2009.

[35] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. In *Doklady AN USSR*, volume 269, pages 543–547, 1983.

[36] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer, 2004.

[37] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621, 2017.

[38] Nhan H Pham, Lam M Nguyen, Dzung T Phan, and Quoc Tran-Dinh. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *arXiv preprint arXiv:1902.05679*, 2019.

[39] Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning*, pages 314–323, 2016.

[40] Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alexander J Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems*, pages 1145–1153, 2016.

[41] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

[42] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: from theory to algorithms*. Cambridge University Press, 2014.

[43] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(2), 2013.

[44] Weijie Su, Stephen P Boyd, and Emmanuel J Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

[45] Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. Spiderboost: A class of faster variance-reduced algorithms for nonconvex optimization. *arXiv preprint arXiv:1810.10690*, 2018.

[46] Blake E Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems*, pages 3639–3647, 2016.

[47] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

[48] Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3925–3936, 2018.

Contents

Appendix A. Main Convergence Theorems for ANITA

In this appendix, we present two main convergence theorems of ANITA (Algorithm 1) for solving finite-sum problems (1), i.e., Theorem 3 under Assumption 1 (general convex setting in Appendix A.1) and Theorem 4 under Assumptions 1–2 (strongly convex setting in Appendix A.2). We will provide the proof sketches for Theorems 3 and 4 in the next Appendix B. The detailed proofs for Theorems 3–4 and Corollaries 1–2 are deferred to Appendix C.

Assumption 1 (L-smoothness) *Functions $f_i : \mathbb{R}^d \to \mathbb{R}$ are convex and L-smooth such that*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \le L\|x - y\| \tag{4}$$

for some $L \ge 0$ and all $i \in [n]$.

Assumption 2 (μ-strong convexity) *A function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex such that*

$$f(x) - f(y) - \langle \nabla f(y), x - y\rangle \ge \frac{\mu}{2}\|x - y\|^2, \tag{5}$$

for some $\mu \ge 0$.

Note that the case $\mu = 0$ reduces to the standard convexity. We will denote $\mu = 0$ as the general convex setting and $\mu > 0$ as the strongly convex setting in this paper. Also note that the strong convexity is only corresponding to the average function f in (1), is not needed for the component functions f_is.

A.1. General convex setting

Now, we provide the main convergence theorem of ANITA for general convex problems and then obtain a corollary for providing the detailed convergence result. Before presenting the theorem, we first recall some basics for the geometric distribution. For a geometric distribution with parameter $p > 0$, denoted as $N \sim \mathrm{Geom}(p)$, i.e., $N = k$ with probability $(1-p)^k p$ for $k = 0, 1, 2, \ldots$ (after k failures until the first success). We know that $\mathbb{E}[N] = \frac{1-p}{p}$. It is not hard to see that if we fix the probability p_t in Line 8 of Algorithm 1 to a constant p, then the update of w_t follows from a geometric distribution $\mathrm{Geom}(p)$. For instance, if we choose $p_t \equiv p = \frac{1}{n+1}$ in Algorithm 1, then we know that $\mathbb{E}[N] = \frac{1-p}{p} = n$ and w_{t+1} will maintain the same as previous values and only change to $\bar{x}_{t+1}$ after n iterations in expectation. In the *first stage* of ANITA, we indeed uses constant probability $p_t \equiv p = \frac{1}{n+1}$. Let t_1 be the first time such that w changes to $\bar{x}$, i.e., $w_{t_1+1} = \bar{x}_{t_1+1}$ and $w_{t_1} = w_{t_1-1} = \cdots = w_0$. Thus $t_1 \sim \mathrm{Geom}(p)$ and $\mathbb{E}[t_1] = n$, where $p = \frac{1}{n+1}$. Note that this first stage where we fix $p_t \equiv p$ is similar to loopless SVRG [13], SCSG [19] and PAGE [29]. One can also derandomize the special case of constant probability p in this first stage to a deterministic double-loop with loop length $\frac{1-p}{p}$ algorithms like the original SVRG [12] and SARAH [37]. The difference is that our ANITA will use a dynamic change of p_t after this first stage, while previous algorithms always keep fixing the probability $p_t \equiv p$.

Theorem 3 (General convex case) *Suppose that Assumption 1 holds. For $0 \le t \le t_1$, let $p_t \equiv \frac{1}{n+1}$, $\theta_t \equiv 1 - \frac{1}{2\sqrt{n}}$, $\eta_t \le \frac{1}{L(1+1/(1-\theta_t))}$ and $\alpha_t = \theta_t$. For $t > t_1$, let $p_t = \max\{\frac{4}{t-t_1+3\sqrt{n}}, \frac{4}{n+3}\}$,*

$\theta_t = \frac{2}{p_t(t-t_1+3\sqrt{n})}$, $\eta_t \le \frac{1}{3L}$ *and* $\alpha_t = \theta_t$. *Then the following equation holds for* ANITA *(Algorithm 1) for any iteration* $t > t_1 + 1$*:*

$$\mathbb{E}[f(w_t) - f(x^*)] \le \frac{32\|x_0 - x^*\|^2}{\eta_{t-1}p_{t-1}(t - t_1 + 3\sqrt{n})^2}.$$

Remark: From the choice of probability $\{p_t\}$ in Theorem 3, we know that there are three stages of ANITA: i) the first stage $p_t \equiv \frac{1}{n+1}$ for $0 \le t \le t_1$; ii) the second stage $p_t = \frac{4}{t-t_1+3\sqrt{n}}$ for $t_1 < t \le t_1 + n + 3 - 3\sqrt{n}$; iii) the third stage $p_t \equiv \frac{4}{n+3}$ for $t > t_1 + n + 3 - 3\sqrt{n}$. This multi-stage convergence analysis is key part for the improvement of ANITA. Roughly speaking, the number of stochastic gradient computations in the first stage is $\#\text{grad} = O(n)$, in the second stage is $\#\text{grad} = O\big(n \min\big\{\log \frac{1}{\epsilon\sqrt{n}}, \log\sqrt{n}\big\}\big)$, and in the third stage is $\#\text{grad} = O\big(\sqrt{\frac{nL}{\epsilon}}\big)$. We will provide a proof sketch of Theorem 3 in the next Appendix B.1. The detailed proofs of Theorem 3 and its Corollary 1 (in Section 3) are deferred to Appendix C.1. Also note that all parameter settings $\{p_t\}$, $\{\theta_t\}$, $\{\eta_t\}$, $\{\alpha_t\}$ of ANITA in Theorem 3 do not require the value of ϵ in advance. The convergence rate of ANITA will automatically switch to different results as stated in Table 2.

A.2. Strongly convex setting

In this section, we provide the main convergence theorem of ANITA for strongly convex problems ($\mu > 0$ in Assumption 2) and then obtain a corollary for providing the detailed convergence result.

Theorem 4 (Strongly convex case) *Suppose that Assumptions 1 and 2 hold. For any* $t \ge 0$*, let* $p_t \equiv p$, $\theta_t \equiv \theta = \frac{1}{2}\min\{1, \sqrt{\frac{\mu}{pL}}\}$, $\eta_t \le \frac{1}{L\theta_t(1+1/(1-\theta_t))}$ *and* $\alpha_t = 1 + \mu\eta_t$. *Then the following equation holds for* ANITA *(Algorithm 1) for any iteration* $t \ge 0$*:*

$$\mathbb{E}[\Phi_t] \le \Big(1 - \frac{4p\theta}{5}\Big)^t \Phi_0, \tag{6}$$

where $\Phi_t := f(w_t) - f(x^*) + \frac{(1+\mu\eta)p\theta}{2\eta}\|x_t - x^*\|^2$.

Remark: In this strongly convex case, the parameter setting of ANITA in Theorem 4 is simpler than the general convex case in Theorem 3. Here, the choice of probability $\{p_t\}$ can be fixed to a constant p and $\{\theta_t\}$ also can be chosen as a constant θ. Then according to Theorem 4, we know that $\{\eta_t\}$ and $\{\alpha_t\}$ also reduce to constant values. Thus there is only one stage in this strongly convex case rather than three stages in previous general convex case. Also here the function value decreases in an exponential rate, i.e., $\mathbb{E}[\Phi_t] \le \big(1 - \frac{4p\theta}{5}\big)^t\Phi_0$ (see (6) in Theorem 4). It is easy to see that the number of iterations T can be bounded by $O(\cdot \log\frac{1}{\epsilon})$ for finding an ϵ-approximate solution $\mathbb{E}[f(w_T) - f(x^*)] \le \epsilon$. Then, by choosing $p = \frac{1}{n}$ (thus each iteration only computes constant stochastic gradients in expectation), the number of total stochastic gradient computations can be bounded by $\#\text{grad} = O\big(\max\big\{n, \sqrt{\frac{nL}{\mu}}\big\}\log\frac{1}{\epsilon}\big)$. This convergence result of ANITA is optimal which matches the lower bound $\Omega\Big(\big(n + \sqrt{\frac{nL}{\mu}}\big)\log\frac{1}{\epsilon}\Big)$ given by Lan and Zhou [15] (see Table 1). Similarly, we will provide a proof sketch of Theorem 4 in the next Appendix B. The detailed proofs of Theorem 4 and its Corollary 2 (in Section 3) are deferred to Appendix C.2. Note that all parameter settings $\{p_t\}$, $\{\theta_t\}$, $\{\eta_t\}$, $\{\alpha_t\}$ of ANITA in Theorem 4 also do not require the value of ϵ in advance.

Appendix B. Proof Sketches for Main Theorems of ANITA

In this appendix, we provide the proof sketches for the two main convergence theorems of ANITA for general convex and strongly convex cases, i.e., for Theorem 3 (in Section B.1) and Theorem 4 (in Section B.2).

B.1. Proof sketch for general convex case (Theorem 3)

Now, we provide the proof sketch of Theorem 3. As we discussed at the end Remark of Section A.1, we know that there are three stages of ANITA. First, we provide a key lemma for the first stage.

Lemma 5 *Suppose Assumption 1 holds. For* $0 \le t \le t_1$*, let* $p_t \equiv p$*,* $\theta_t \equiv \theta$*,* $\eta_t \le \frac{1}{L(1+1/(1-\theta_t))}$ *and* $\alpha_t = \theta_t$*. Then the following equation holds for* ANITA *(Algorithm 1):*

$$\mathbb{E}[f(w_{t_1+1}) - f(x^*)] \le \mathbb{E}\Big[(1-\theta)\big(f(x_0) - f(x^*)\big) + \Big(\frac{\theta^2 p}{2\eta} + (1-p)L(1-\theta)\theta^2\Big)\|x_0 - x^*\|^2 - \Big(\frac{\theta^2 p}{2\eta} - (1-p)L(1-\theta)\theta^2\Big)\|x_{t_1+1} - x^*\|^2\Big]. \tag{7}$$

Particularly, we choose $p_t \equiv p = \frac{1}{n+1}$ in the first stage of ANITA in Theorem 3. As we discussed before Theorem 3, we know that $\mathbb{E}[t_1] = \frac{1-p}{p} = n$. One can also derandomize this first stage by running Line 3–7 of Algorithm 1 for n iterations and then letting $w_{n+1} = \bar{x}_{n+1}$.

After the first stage, for iterations $t > t_1$, we will use a dynamic change of p_t. We first provide the following technical lemma which describes the change of function value between two adjacent iterations.

Lemma 6 *Suppose Assumption 1 holds. Choose stepsize* $\eta_t \le \frac{1}{L(1+1/(1-\theta_t))}$ *and* $\alpha_t = \theta_t$ *for any* $t \ge 0$*. Then the following equation holds for* ANITA *(Algorithm 1) for any iteration* $t \ge 0$*:*

$$\mathbb{E}\Big[\frac{\eta_t}{p_t\theta_t^2}\big(f(w_{t+1}) - f(x^*)\big)\Big] \le \mathbb{E}\Big[\frac{(1-p_t\theta_t)\eta_t}{p_t\theta_t^2}\big(f(w_t) - f(x^*)\big) + \frac{1}{2}\Big(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\Big)\Big]. \tag{8}$$

According to (8), in order to get a recursion formula, we need to show that

$$\frac{(1-p_t\theta_t)\eta_t}{p_t\theta_t^2} \le \frac{\eta_{t-1}}{p_{t-1}\theta_{t-1}^2} \tag{9}$$

by further choosing appropriate parameters $\{p_t\}$, $\{\theta_t\}$ and $\{\eta_t\}$. In particular, choosing $p_t = \max\{\frac{4}{t-t_1+3\sqrt{n}}, \frac{4}{n+3}\}$, $\theta_t = \frac{2}{p_t(t-t_1+3\sqrt{n})}$ and $\eta_t \equiv \eta \le \frac{1}{3L}$ for $t > t_1$ (as chosen in Theorem 3) can satisfy (9) for any $t > t_1 + 1$. Combining this choice of $\{p_t\}$, $\{\theta_t\}$ and $\{\eta_t\}$ with Lemma 6 and summing up from iteration $t_1 + 1$ to t, we obtain the following Lemma 7.

Lemma 7 *Suppose Assumption 1 holds. For* $t > t_1$*, let* $p_t = \max\{\frac{4}{t-t_1+3\sqrt{n}}, \frac{4}{n+3}\}$*,* $\theta_t = \frac{2}{p_t(t-t_1+3\sqrt{n})}$*,* $\eta_t \le \frac{1}{3L}$ *and* $\alpha_t = \theta_t$*. Then the following equation holds for* ANITA *(Algorithm 1) for any iteration* $t > t_1 + 1$*:*

$$\mathbb{E}\Big[\frac{\eta_{t-1}}{p_{t-1}\theta_{t-1}^2}\big(f(w_t) - f(x^*)\big)\Big] \le \mathbb{E}\Big[\frac{(1-p_{t_1+1}\theta_{t_1+1})\eta_{t_1+1}}{p_{t_1+1}\theta_{t_1+1}^2}\big(f(w_{t_1+1}) - f(x^*)\big) + \frac{1}{2}\Big(\|x_{t_1+1} - x^*\|^2 - \|x_t - x^*\|^2\Big)\Big]. \tag{10}$$

Also note that we can bound the term $f(x_0) - f(x^*)$ in (7) as $f(x_0) - f(x^*) \leq \frac{L}{2}\|x_0 - x^*\|^2$ according to the L-smoothness of f (Assumption 1). Now, we combine Lemma 5 and Lemma 7 to prove the main Theorem 3, i.e., by plugging (7) into (10) and plugging in the value of parameters, we can obtain, for any iteration $t > t_1 + 1$,

$$\mathbb{E}[f(w_t) - f(x^*)] \leq \frac{32\|x_0 - x^*\|^2}{\eta_{t-1}p_{t-1}(t - t_1 + 3\sqrt{n})^2}.$$

The proof sketch of Theorem 3 is finished.

B.2. Proof sketch for strongly convex case (Theorem 4)

Now, we provide the proof sketch of Theorem 4. As we discussed at the end Remark of Section A.2, the parameter setting of ANITA in this strongly convex case is simpler than the general convex case in Theorem 3. As a result, we only need one technical Lemma 8 in this proof sketch of Theorem 4 rather than three Lemmas 5–7 in previous general convex case.

Lemma 8 *Suppose that Assumptions 1 and 2 hold. Choose stepsize* $\eta_t \leq \frac{1}{L\theta_t(1+1/(1-\theta_t))}$ *and* $\alpha_t = 1 + \mu\eta_t$ *for any* $t \geq 0$*. Then the following equation holds for* ANITA *(Algorithm 1) for any iteration* $t \geq 0$*:*

$$\mathbb{E}\left[f(w_{t+1}) - f(x^*) + \frac{(1+\mu\eta_t)p_t\theta_t}{2\eta_t}\|x_{t+1} - x^*\|^2\right] \leq \mathbb{E}\left[(1 - p_t\theta_t)\big(f(w_t) - f(x^*)\big) + \frac{p_t\theta_t}{2\eta_t}\|x_t - x^*\|^2\right].$$

Then, if we further choosing the probability $\{p_t\}$ as a constant p and $\{\theta_t\}$ as a constant θ, we know that the parameters η_t and α_t will also be fixed to the constant η and α (see Lemma 8). Now, if we further define

$$\Phi_t := f(w_t) - f(x^*) + \frac{(1+\mu\eta)p\theta}{2\eta}\|x_t - x^*\|^2,$$

then Lemma 8 can be changed to, for any iteration $t \geq 0$,

$$\mathbb{E}[\Phi_{t+1}] \leq \mathbb{E}\left[\max\left\{1 - p\theta,\ \frac{1}{1+\mu\eta}\right\}\Phi_t\right]. \tag{11}$$

Now if we further let $\theta_t \equiv \theta = \frac{1}{2}\min\{1, \sqrt{\frac{\mu}{pL}}\}$, we have

$$\frac{1}{1+\mu\eta} \leq 1 - \frac{4p\theta}{5}. \tag{12}$$

By plugging (12) into (11), we finish the proof sketch of Theorem 4:

$$\mathbb{E}[\Phi_{t+1}] \leq \mathbb{E}\left[\Big(1 - \frac{4p\theta}{5}\Big)\Phi_t\right] \leq \Big(1 - \frac{4p\theta}{5}\Big)^{t+1}\Phi_0.$$

Appendix C. Missing Detailed Proofs

Now, we provide the detailed proofs of main convergence theorems and corollaries of ANITA for both general convex case (Theorem 3 and Corollary 1) and strongly convex case (Theorem 4 and Corollary 2).

Before proving these theorems and corollaries, we first recall some basic properties for smooth convex functions and some basic facts for the geometric distribution, then we provide some important technical lemmas.

Lemma 9 (Lemma 1 in [17]) *If $f : X \to \mathbb{R}$ has L-Lipschitz continuous gradients (L-smooth), then we have*

$$\frac{1}{2L}\|\nabla f(x) - \nabla f(z)\|^2 \le f(x) - f(z) - \langle \nabla f(z), x - z\rangle, \qquad \forall x, z \in X. \tag{13}$$

We also recall the proof in Lan et al. [17] for completeness.

Proof of Lemma 9. Denote $\phi(x) = f(x) - f(z) - \langle \nabla f(z), x - z\rangle$. Clearly ϕ also has L-Lipschitz continuous gradients. It is easy to check that $\nabla\phi(z) = 0$, and hence that $\min_x \phi(x) = \phi(z) = 0$, which implies

$$\begin{aligned}
\phi(z) &\le \phi\Big(x - \frac{1}{L}\nabla\phi(x)\Big) \\
&= \phi(x) + \int_0^1 \Big\langle \nabla\phi\Big(x - \frac{\tau}{L}\nabla\phi(x)\Big), -\frac{1}{L}\nabla\phi(x)\Big\rangle d\tau \\
&= \phi(x) + \Big\langle \nabla\phi(x), -\frac{1}{L}\nabla\phi(x)\Big\rangle + \int_0^1 \Big\langle \nabla\phi\Big(x - \frac{\tau}{L}\nabla\phi(x)\Big) - \nabla\phi(x), -\frac{1}{L}\nabla\phi(x)\Big\rangle d\tau \\
&\le \phi(x) - \frac{1}{L}\|\nabla\phi(x)\|^2 + \int_0^1 L\Big\|\frac{\tau}{L}\nabla\phi(x)\Big\| \, \Big\|\frac{1}{L}\nabla\phi(x)\Big\| d\tau \\
&= \phi(x) - \frac{1}{2L}\|\nabla\phi(x)\|^2.
\end{aligned}$$

Therefore, we have $\frac{1}{2L}\|\nabla\phi(x)\|^2 \le \phi(x) - \phi(z) = \phi(x)$, and the result follows immediately from this relation. □

For a geometric distribution with parameter $p > 0$, denoted as $N \sim \mathrm{Geom}(p)$, i.e., $N = k$ with probability $(1-p)^k p$ for $k = 0, 1, 2, \ldots$ (after k failures until the first success). We know the following facts hold (see, e.g., [20]).

Fact 1 *Let $N \sim \mathrm{Geom}(p)$. Then for any sequence $D_0, D_1, \ldots$ with $\mathbb{E}|D_N| < \infty$, we have*

$$\mathbb{E}[N] = \frac{1-p}{p}, \tag{14}$$

$$\mathbb{E}[D_N - D_{N+1}] = \frac{p}{1-p}\big(D_0 - \mathbb{E}[D_N]\big), \tag{15}$$

$$\mathbb{E}[D_N] = pD_0 + (1-p)\mathbb{E}[D_{N+1}]. \tag{16}$$

Now, we provide some important technical lemmas which are useful for proving the main convergence theorems of ANITA. Concretely, Lemma 10 provides some ways to upper bound the variance of the gradient estimator in ANITA. Lemma 11 describes the change of function value after a gradient update step in ANITA.

Lemma 10 *Suppose that Assumption 1 holds. The gradient estimator*

$$\widetilde{\nabla}_t = \nabla f_i(\underline{x}_t) - \nabla f_i(w_t) + \nabla f(w_t) \tag{17}$$

is defined in Line 5 of Algorithm 1, then conditional on the past, we have

$$\mathbb{E}[\widetilde{\nabla}_t] = \nabla f(\underline{x}_t), \tag{18}$$
$$\mathbb{E}[\|\widetilde{\nabla}_t - \nabla f(\underline{x}_t)\|^2] \leq L^2 \|\underline{x}_t - w_t\|^2, \tag{19}$$
$$\mathbb{E}[\|\widetilde{\nabla}_t - \nabla f(\underline{x}_t)\|^2] \leq 2L\big(f(w_t) - f(\underline{x}_t) - \langle \nabla f(\underline{x}_t), w_t - \underline{x}_t\rangle\big). \tag{20}$$

Proof of Lemma 10. For (18), it is easy to see that (note that the expectation is taken over the random choice of i in iteration t (see Line 4 of Algorithm 1))

$$\begin{aligned}
\mathbb{E}[\widetilde{\nabla}_t] &\overset{(17)}{=} \mathbb{E}[\nabla f_i(\underline{x}_t) - \nabla f_i(w_t) + \nabla f(w_t)] \\
&= \nabla f(\underline{x}_t) - \nabla f(w_t) + \nabla f(w_t) = \nabla f(\underline{x}_t).
\end{aligned}$$

Then, for (19), we obtain it from Assumption 1 as follows:

$$\begin{aligned}
\mathbb{E}[\|\widetilde{\nabla}_t - \nabla f(\underline{x}_t)\|^2] &\overset{(17)}{=} \mathbb{E}[\|\nabla f_i(\underline{x}_t) - \nabla f_i(w_t) + \nabla f(w_t) - \nabla f(\underline{x}_t)\|^2] \\
&\leq \mathbb{E}[\|\nabla f_i(\underline{x}_t) - \nabla f_i(w_t)\|^2] && (21) \\
&\leq L^2 \|\underline{x}_t - w_t\|^2, && (22)
\end{aligned}$$

where (21) follows from the fact that $\mathbb{E}[\|x - \mathbb{E}x\|^2] \leq \mathbb{E}[\|x\|^2]$ for any random variable x, and (22) follows from Assumption 1, i.e., the L-Lipschitz continuous gradients $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|$.

Now, for the last one (20), we obtain it from (21) and Assumption 1 as follows:

$$\begin{aligned}
\mathbb{E}[\|\widetilde{\nabla}_t - \nabla f(\underline{x}_t)\|^2] &\overset{(21)}{\leq} \mathbb{E}[\|\nabla f_i(\underline{x}_t) - \nabla f_i(w_t)\|^2] \\
&\leq \mathbb{E}\big[2L\big(f_i(w_t) - f_i(\underline{x}_t) - \langle \nabla f_i(\underline{x}_t), w_t - \underline{x}_t\rangle\big)\big] && (23) \\
&= 2L\big(f(w_t) - f(\underline{x}_t) - \langle \nabla f(\underline{x}_t), w_t - \underline{x}_t\rangle\big),
\end{aligned}$$

where (23) uses Lemma 9 with x and z replaced by w_t and $\underline{x}_t$, and f replaced by f_i since f_i has L-Lipschitz continuous gradients according to Assumption 1. □

Lemma 11 *Suppose that Assumptions 1 and 2 hold. Let stepsize $\eta_t \leq \frac{\alpha_t}{L(1+\mu\eta_t)\theta_t(1+1/(1-\theta_t))}$, then the following equation holds for* ANITA *(Algorithm 1) for any iteration $t \geq 0$:*

$$\begin{aligned}
\mathbb{E}[f(w_{t+1}) - f(x^*)] \leq \mathbb{E}\Bigg[&(1 - p_t\theta_t)\big(f(w_t) - f(x^*)\big) \\
&+ \frac{p_t\alpha_t\theta_t}{(1+\mu\eta_t)\eta_t}\Big(\frac{1}{2}\|x_t - x^*\|^2 - \frac{1+\mu\eta_t}{2}\|x_{t+1} - x^*\|^2\Big) \\
&- \frac{\mu(1+\mu\eta_t - \alpha_t)p_t\theta_t}{2(1+\mu\eta_t)}\|\underline{x}_t - x^*\|^2\Bigg].
\end{aligned} \tag{24}$$

Note that for the case of $\mu = 0$ (general (non-strongly) convex setting), only the smoothness Assumption 1 is required, i.e., the strong convexity Assumption 2 is not needed for obtaining (24) *with $\mu = 0$.*

Proof of Lemma 11. First, in view of L-smoothness of f (Assumption 1), we have

$$
\begin{aligned}
&\mathbb{E}[f(\bar{x}_{t+1})] \\
&\le \mathbb{E}\bigg[f(\underline{x}_t) + \langle \nabla f(\underline{x}_t), \bar{x}_{t+1} - \underline{x}_t\rangle + \frac{L}{2}\|\bar{x}_{t+1} - \underline{x}_t\|^2\bigg] \\
&= \mathbb{E}\bigg[f(\underline{x}_t) + \langle \nabla f(\underline{x}_t), \theta_t(x_{t+1} - x_t)\rangle + \frac{L\theta_t^2}{2}\|x_{t+1} - x_t\|^2\bigg] \qquad (25)\\
&= \mathbb{E}\bigg[f(\underline{x}_t) + \langle \nabla f(\underline{x}_t) - \widetilde{\nabla}_t, \theta_t(x_{t+1} - x_t)\rangle + \langle \widetilde{\nabla}_t, \theta_t(x_{t+1} - x_t)\rangle + \frac{L\theta_t^2}{2}\|x_{t+1} - x_t\|^2\bigg] \\
&\le \mathbb{E}\bigg[f(\underline{x}_t) + \frac{\beta_t}{2L}\|\nabla f(\underline{x}_t) - \widetilde{\nabla}_t\|^2 + \frac{L\theta_t^2}{2\beta_t}\|x_{t+1} - x_t\|^2 + \langle \widetilde{\nabla}_t, \theta_t(x_{t+1} - x_t)\rangle + \frac{L\theta_t^2}{2}\|x_{t+1} - x_t\|^2\bigg] \\
&\qquad (26)\\
&= \mathbb{E}\bigg[f(\underline{x}_t) + \frac{\beta_t}{2L}\|\nabla f(\underline{x}_t) - \widetilde{\nabla}_t\|^2 + \frac{L(1+1/\beta_t)\theta_t^2}{2}\|x_{t+1} - x_t\|^2 + \langle \widetilde{\nabla}_t, \theta_t(x^* - x_t)\rangle \\
&\qquad + \langle \widetilde{\nabla}_t, \theta_t(x_{t+1} - x^*)\rangle\bigg] \\
&\overset{(18)}{=} \mathbb{E}\bigg[f(\underline{x}_t) + \frac{\beta_t}{2L}\|\nabla f(\underline{x}_t) - \widetilde{\nabla}_t\|^2 + \frac{L(1+1/\beta_t)\theta_t^2}{2}\|x_{t+1} - x_t\|^2 + \langle \nabla f(\underline{x}_t), \theta_t(x^* - x_t)\rangle \\
&\qquad + \langle \widetilde{\nabla}_t, \theta_t(x_{t+1} - x^*)\rangle\bigg] \qquad (27)\\
&\overset{(20)}{\le} \mathbb{E}\bigg[f(\underline{x}_t) + \beta_t\big(f(w_t) - f(\underline{x}_t) - \langle \nabla f(\underline{x}_t), w_t - \underline{x}_t\rangle\big) + \frac{L(1+1/\beta_t)\theta_t^2}{2}\|x_{t+1} - x_t\|^2 \\
&\qquad + \langle \nabla f(\underline{x}_t), \theta_t(x^* - x_t)\rangle + \langle \widetilde{\nabla}_t, \theta_t(x_{t+1} - x^*)\rangle\bigg] \\
&= \mathbb{E}\bigg[(1-\theta_t)f(w_t) + \theta_t f(\underline{x}_t) - \langle \nabla f(\underline{x}_t), (1-\theta_t)(w_t - \underline{x}_t)\rangle + \langle \nabla f(\underline{x}_t), \theta_t(x^* - x_t)\rangle \\
&\qquad + \frac{L(1+1/(1-\theta_t))\theta_t^2}{2}\|x_{t+1} - x_t\|^2 + \langle \widetilde{\nabla}_t, \theta_t(x_{t+1} - x^*)\rangle\bigg] \qquad (28)\\
&= \mathbb{E}\bigg[(1-\theta_t)f(w_t) + \theta_t\Big(f(\underline{x}_t) + \langle \nabla f(\underline{x}_t), x^* - \underline{x}_t\rangle\Big) \\
&\qquad + \frac{L(1+1/(1-\theta_t))\theta_t^2}{2}\|x_{t+1} - x_t\|^2 + \langle \widetilde{\nabla}_t, \theta_t(x_{t+1} - x^*)\rangle\bigg], \qquad (29)
\end{aligned}
$$

where (25) holds since $\bar{x}_{t+1} - \underline{x}_t = \theta_t(x_{t+1} - x_t)$ according to the two interpolation steps of ANITA (see Line 3 and Line 7 of Algorithm 1), (26) uses Young's inequality with $\beta_t > 0$, (28) holds by further choosing $\beta_t = 1 - \theta_t$, (29) removes w_t and x_t via the interpolation step $\underline{x}_t = \theta_t x_t + (1-\theta_t)w_t$ (see Line 3 of Algorithm 1).

Now, we use the (strong) convexity of f (see Assumption 2) in (29) to obtain

$$
\begin{aligned}
\mathbb{E}[f(\bar{x}_{t+1})] \le \mathbb{E}\bigg[&(1-\theta_t)f(w_t) + \theta_t\Big(f(x^*) - \frac{\mu}{2}\|\underline{x}_t - x^*\|^2\Big) \\
&+ \frac{L(1+1/(1-\theta_t))\theta_t^2}{2}\|x_{t+1} - x_t\|^2 + \langle \widetilde{\nabla}_t, \theta_t(x_{t+1} - x^*)\rangle\bigg]. \qquad (30)
\end{aligned}
$$

Then, we deduce the last inner product term in (30) as follows:

$$
\begin{aligned}
&\mathbb{E}\big[\langle\widetilde{\nabla}_t, \theta_t(x_{t+1}-x^*)\rangle\big] \\
&= \mathbb{E}\bigg[\frac{\alpha_t\theta_t}{(1+\mu\eta_t)\eta_t}\big\langle x_t+\mu\eta_t-(1+\mu\eta_t)x_{t+1}, x_{t+1}-x^*\big\rangle\bigg] \qquad (31)\\
&= \mathbb{E}\bigg[\frac{\alpha_t\theta_t}{(1+\mu\eta_t)\eta_t}\Big(\big\langle x_t-x_{t+1}, x_{t+1}-x^*\big\rangle+\mu\eta_t\big\langle \underline{x}_t-x_{t+1}, x_{t+1}-x^*\big\rangle\Big)\bigg]\\
&= \mathbb{E}\bigg[\frac{\alpha_t\theta_t}{(1+\mu\eta_t)\eta_t}\Big(\frac{1}{2}\big(\|x_t-x^*\|^2-\|x_t-x_{t+1}\|^2-\|x_{t+1}-x^*\|^2\big)\\
&\qquad\qquad +\frac{\mu\eta_t}{2}\big(\|\underline{x}_t-x^*\|^2-\|\underline{x}_t-x_{t+1}\|^2-\|x_{t+1}-x^*\|^2\big)\Big)\bigg]\\
&\le \mathbb{E}\bigg[\frac{\alpha_t\theta_t}{(1+\mu\eta_t)\eta_t}\Big(\frac{1}{2}\|x_t-x^*\|^2-\frac{1+\mu\eta_t}{2}\|x_{t+1}-x^*\|^2+\frac{\mu\eta_t}{2}\|\underline{x}_t-x^*\|^2-\frac{1}{2}\|x_t-x_{t+1}\|^2\Big)\bigg], \qquad (32)
\end{aligned}
$$

where (31) follows from the gradient update step of ANITA (see Line 6 of Algorithm 1).

Now we plug (32) into (30) to get

$$
\begin{aligned}
&\mathbb{E}[f(\bar{x}_{t+1})]\\
&\le \mathbb{E}\bigg[(1-\theta_t)f(w_t)+\theta_t f(x^*)-\frac{\mu(1+\mu\eta_t-\alpha_t)\theta_t}{2(1+\mu\eta_t)}\|\underline{x}_t-x^*\|^2+\frac{L(1+1/(1-\theta_t))\theta_t^2}{2}\|x_{t+1}-x_t\|^2\\
&\qquad +\frac{\alpha_t\theta_t}{(1+\mu\eta_t)\eta_t}\Big(\frac{1}{2}\|x_t-x^*\|^2-\frac{1+\mu\eta_t}{2}\|x_{t+1}-x^*\|^2\Big)-\frac{\alpha_t\theta_t}{2(1+\mu\eta_t)\eta_t}\|x_{t+1}-x_t\|^2\bigg]\\
&\le \mathbb{E}\bigg[(1-\theta_t)f(w_t)+\theta_t f(x^*)-\frac{\mu(1+\mu\eta_t-\alpha_t)\theta_t}{2(1+\mu\eta_t)}\|\underline{x}_t-x^*\|^2\\
&\qquad +\frac{\alpha_t\theta_t}{(1+\mu\eta_t)\eta_t}\Big(\frac{1}{2}\|x_t-x^*\|^2-\frac{1+\mu\eta_t}{2}\|x_{t+1}-x^*\|^2\Big)\bigg], \qquad (33)
\end{aligned}
$$

where the last inequality (33) holds by letting $\eta_t \le \frac{\alpha_t}{L(1+\mu\eta_t)\theta_t(1+1/(1-\theta_t))}$.

Finally, according to the probabilistic update of w_{t+1} in Line 8 of Algorithm 1, we have

$$\mathbb{E}[f(w_{t+1})] = \mathbb{E}\big[p_t f(\bar{x}_{t+1})+(1-p_t)f(w_t)\big] \qquad (34)$$

The proof is finished by combining (33) with (34), i.e., (24) is obtained by adding $p_t \times$ (33) and (34). □

C.1. Proofs for general convex case

In Appendix C.1.1, we provide the proof for the main convergence Theorem 3 in the general convex case (i.e., $\mu = 0$). Note that the strong convexity Assumption 2 is not needed in this case. Then we provide the proof for a following Corollary 1 with detailed convergence result in Appendix C.1.2.

C.1.1. PROOF OF THEOREM 3

First, according to the probabilistic update of w_{t+1} in Line 8 of Algorithm 1, i.e.,

$$w_{t+1} = \begin{cases} \bar{x}_{t+1} & \text{with probability } p_t \\ w_t & \text{with probability } 1-p_t \end{cases} \qquad (35)$$

Let $p_t \equiv p$ for $0 \le t \le t_1$, where t_1 denotes the first time such that $w_{t_1+1} = \bar{x}_{t_1+1}$, i.e., $w_{t_1} = w_{t_1-1} = \cdots = w_0$. It is easy to see that $t_1 \sim \mathrm{Geom}(p)$, i.e., $t_1 = k$ with probability $(1-p)^k p$ for $k = 0, 1, 2, \ldots t_1$ (after k failures until the first success). Also we know that $\mathbb{E}[t_1] = \frac{1-p}{p}$ according to Fact 1 (see (14)). Now, we restate the technical Lemma 5 which shows the decrease of function value in iterations $0 \le t \le t_1$, and then provide its proof.

Lemma 5 *Suppose Assumption 1 holds. For $0 \le t \le t_1$, let $p_t \equiv p$, $\theta_t \equiv \theta$, $\eta_t \le \frac{1}{L(1+1/(1-\theta_t))}$ and $\alpha_t = \theta_t$. Then the following equation holds for* ANITA *(Algorithm 1):*

$$\begin{aligned}\mathbb{E}[f(w_{t_1+1}) - f(x^*)] \le \mathbb{E}\Big[(1-\theta)\big(f(x_0) - f(x^*)\big) + \Big(\frac{\theta^2 p}{2\eta} + (1-p)L(1-\theta)\theta^2\Big)\|x_0 - x^*\|^2 \\ - \Big(\frac{\theta^2 p}{2\eta} - (1-p)L(1-\theta)\theta^2\Big)\|x_{t_1+1} - x^*\|^2\Big]. \qquad (36)\end{aligned}$$

Proof of Lemma 5. First, in view of L-smoothness of f (Assumption 1), we recall (27) (where $\forall \beta_t > 0$):

$$\begin{aligned}
&\mathbb{E}[f(\bar{x}_{t+1})] \\
&\le \mathbb{E}\Big[f(\underline{x}_t) + \frac{\beta_t}{2L}\|\nabla f(\underline{x}_t) - \widetilde{\nabla}_t\|^2 + \frac{L(1+1/\beta_t)\theta_t^2}{2}\|x_{t+1} - x_t\|^2 + \langle \nabla f(\underline{x}_t), \theta_t(x^* - x_t)\rangle \\
&\qquad + \langle \widetilde{\nabla}_t, \theta_t(x_{t+1} - x^*)\rangle\Big] \\
&= \mathbb{E}\Big[f(\underline{x}_t) + \frac{\beta_t}{2L}\|\nabla f(\underline{x}_t) - \widetilde{\nabla}_t\|^2 + \frac{L(1+1/\beta_t)\theta_t^2}{2}\|x_{t+1} - x_t\|^2 \\
&\qquad + \Big\langle \nabla f(\underline{x}_t), \theta_t(x^* - \underline{x}_t) + (1-\theta_t)(w_t - \underline{x}_t)\Big\rangle + \langle \widetilde{\nabla}_t, \theta_t(x_{t+1} - x^*)\rangle\Big] \qquad (37) \\
&\le \mathbb{E}\Big[(1-\theta_t)f(w_t) + \theta_t f(x^*) + \frac{\beta_t}{2L}\|\nabla f(\underline{x}_t) - \widetilde{\nabla}_t\|^2 + \frac{L(1+1/\beta_t)\theta_t^2}{2}\|x_{t+1} - x_t\|^2 \\
&\qquad + \langle \widetilde{\nabla}_t, \theta_t(x_{t+1} - x^*)\rangle\Big] \qquad (38) \\
&\overset{(19)}{\le} \mathbb{E}\Big[(1-\theta_t)f(w_t) + \theta_t f(x^*) + \frac{L\beta_t}{2}\|\underline{x}_t - w_t\|^2 + \frac{L(1+1/\beta_t)\theta_t^2}{2}\|x_{t+1} - x_t\|^2 \\
&\qquad + \langle \widetilde{\nabla}_t, \theta_t(x_{t+1} - x^*)\rangle\Big] \\
&= \mathbb{E}\Big[(1-\theta_t)f(w_t) + \theta_t f(x^*) + \frac{L\beta_t\theta_t^2}{2}\|x_t - w_t\|^2 + \frac{L(1+1/\beta_t)\theta_t^2}{2}\|x_{t+1} - x_t\|^2 \\
&\qquad + \langle \widetilde{\nabla}_t, \theta_t(x_{t+1} - x^*)\rangle\Big] \qquad (39) \\
&= \mathbb{E}\Big[(1-\theta_t)f(w_t) + \theta_t f(x^*) + \frac{L(1-\theta_t)\theta_t^2}{2}\|x_t - w_t\|^2 + \frac{L(1+1/(1-\theta_t))\theta_t^2}{2}\|x_{t+1} - x_t\|^2 \\
&\qquad + \langle \widetilde{\nabla}_t, \theta_t(x_{t+1} - x^*)\rangle\Big], \qquad (40)
\end{aligned}$$

where (37) and (39) use the interpolation step $\underline{x}_t = \theta_t x_t + (1-\theta_t)w_t$ (see Line 3 of Algorithm 1), (38) uses the convexity of f, and (40) holds by choosing $\beta_t = 1-\theta_t$.

For the last inner product term in (40), we recall (32) here:

$$\begin{aligned}
&\mathbb{E}\big[\langle \widetilde{\nabla}_t, \theta_t(x_{t+1}-x^*)\rangle\big] \\
&\le \mathbb{E}\bigg[\frac{\alpha_t\theta_t}{(1+\mu\eta_t)\eta_t}\Big(\frac{1}{2}\|x_t-x^*\|^2 - \frac{1+\mu\eta_t}{2}\|x_{t+1}-x^*\|^2 + \frac{\mu\eta_t}{2}\|\underline{x}_t-x^*\|^2 - \frac{1}{2}\|x_t-x_{t+1}\|^2\Big)\bigg] \\
&= \mathbb{E}\bigg[\frac{\alpha_t\theta_t}{\eta_t}\Big(\frac{1}{2}\|x_t-x^*\|^2 - \frac{1}{2}\|x_{t+1}-x^*\|^2 - \frac{1}{2}\|x_t-x_{t+1}\|^2\Big)\bigg], \qquad (41)
\end{aligned}$$

where the last equality (41) holds due to $\mu = 0$ in this general non-strongly convex case.

Now, we plug (41) into (40) to get

$$\begin{aligned}
\mathbb{E}[f(\bar{x}_{t+1})] \le \mathbb{E}\bigg[&(1-\theta_t)f(w_t) + \theta_t f(x^*) + \frac{\alpha_t\theta_t}{\eta_t}\Big(\frac{1}{2}\|x_t-x^*\|^2 - \frac{1}{2}\|x_{t+1}-x^*\|^2\Big) \\
&+ \frac{L(1-\theta_t)\theta_t^2}{2}\|x_t-w_t\|^2 - \frac{\alpha_t\theta_t - L(1+1/(1-\theta_t))\theta_t^2\eta_t}{2\eta_t}\|x_t-x_{t+1}\|^2\bigg]. \qquad (42)
\end{aligned}$$

According to the parameter setting in Lemma 5, we know that $p_t \equiv p$, $\theta_t \equiv \theta$, $\eta_t \le \frac{1}{L(1+1/(1-\theta_t))} \equiv \frac{1}{L(1+1/(1-\theta))}$ and $\alpha_t = \theta_t \equiv \theta$ for iterations $0 \le t \le t_1$ in the first stage. By plugging these parameters into (42), we obtain

$$\begin{aligned}
&\mathbb{E}[f(w_{t_1+1})] \\
&= \mathbb{E}[f(\bar{x}_{t_1+1})] \\
&\le \mathbb{E}\bigg[(1-\theta)f(w_0) + \theta f(x^*) + \frac{\theta^2}{2\eta}\Big(\|x_{t_1}-x^*\|^2 - \|x_{t_1+1}-x^*\|^2\Big) + \frac{L(1-\theta)\theta^2}{2}\|x_{t_1}-w_0\|^2\bigg] \\
&\overset{(15)}{=} \mathbb{E}\bigg[(1-\theta)f(w_0) + \theta f(x^*) + \frac{\theta^2 p}{2\eta(1-p)}\Big(\|x_0-x^*\|^2 - \|x_{t_1}-x^*\|^2\Big) + \frac{L(1-\theta)\theta^2}{2}\|x_{t_1}-w_0\|^2\bigg] \\
&\overset{(16)}{=} \mathbb{E}\bigg[(1-\theta)f(w_0) + \theta f(x^*) + \frac{\theta^2 p}{2\eta(1-p)}\Big(\|x_0-x^*\|^2 - (1-p)\|x_{t_1+1}-x^*\|^2 - p\|x_0-x^*\|^2\Big) \\
&\qquad + (1-p)\frac{L(1-\theta)\theta^2}{2}\|x_{t_1+1}-w_0\|^2 + p\frac{L(1-\theta)\theta^2}{2}\|x_0-w_0\|^2\bigg] \\
&\le \mathbb{E}\bigg[(1-\theta)f(x_0) + \theta f(x^*) + \frac{\theta^2 p}{2\eta(1-p)}\Big(\|x_0-x^*\|^2 - (1-p)\|x_{t_1+1}-x^*\|^2 - p\|x_0-x^*\|^2\Big) \\
&\qquad + (1-p)L(1-\theta)\theta^2\Big(\|x_{t_1+1}-x^*\|^2 + \|x_0-x^*\|^2\Big)\bigg], \qquad (43)
\end{aligned}$$

where (43) uses Cauchy-Schwarz inequality and $w_0 = x_0$. Now, the proof of Lemma 5 is finished since (36) directly follows from (43). □

Now, for iterations $t > t_1$. We restate the key Lemma 7 which shows the decrease of function value in iterations $t > t_1$, and then provide its proof.

Lemma 7 *Suppose Assumption 1 holds. For $t > t_1$, let $p_t = \max\{\frac{4}{t-t_1+3\sqrt{n}}, \frac{4}{n+3}\}$, $\theta_t = \frac{2}{p_t(t-t_1+3\sqrt{n})}$, $\eta_t \le \frac{1}{3L}$ and $\alpha_t = \theta_t$. Then the following equation holds for* ANITA *(Algorithm 1)* *for any iteration $t > t_1 + 1$:*

$$\mathbb{E}\left[\frac{\eta_{t-1}}{p_{t-1}\theta_{t-1}^2}\big(f(w_t) - f(x^*)\big)\right] \le \mathbb{E}\left[\frac{(1 - p_{t_1+1}\theta_{t_1+1})\eta_{t_1+1}}{p_{t_1+1}\theta_{t_1+1}^2}\big(f(w_{t_1+1}) - f(x^*)\big) + \frac{1}{2}\Big(\|x_{t_1+1} - x^*\|^2 - \|x_t - x^*\|^2\Big)\right]. \tag{44}$$

Proof of Lemma 7. For proving this lemma, we will use our technical Lemma 11 with $\mu = 0$ (general convex case). In particular, by choosing $\alpha_t = \theta_t$ and multiplying $\frac{\eta_t}{p_t\theta_t^2}$ for both sides in (24) with $\mu = 0$, we obtain the Lemma 6 and here we restate it:

Lemma 6 *Suppose Assumption 1 holds. Choose stepsize $\eta_t \le \frac{1}{L(1+1/(1-\theta_t))}$ and $\alpha_t = \theta_t$ for any $t \ge 0$. Then the following equation holds for* ANITA *(Algorithm 1) for any iteration $t \ge 0$:*

$$\mathbb{E}\left[\frac{\eta_t}{p_t\theta_t^2}\big(f(w_{t+1}) - f(x^*)\big)\right] \le \mathbb{E}\left[\frac{(1 - p_t\theta_t)\eta_t}{p_t\theta_t^2}\big(f(w_t) - f(x^*)\big) + \frac{1}{2}\Big(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\Big)\right]. \tag{45}$$

Then, we are going to sum up (45) from iteration $t_1 + 1$ to t for obtaining (44). In order to get a recursion formula for (45), we further choose appropriate parameters $\{p_t\}$, $\{\theta_t\}$ and $\{\eta_t\}$ to obtain

$$\frac{(1 - p_t\theta_t)\eta_t}{p_t\theta_t^2} \le \frac{\eta_{t-1}}{p_{t-1}\theta_{t-1}^2}. \tag{46}$$

It is not hard to verify that (46) can be satisfied for any $t > t_1 + 1$ by choosing

$$p_t = \max\left\{\frac{4}{t - t_1 + 3\sqrt{n}}, \frac{4}{n+3}\right\}, \tag{47}$$

$$\theta_t = \frac{2}{p_t(t - t_1 + 3\sqrt{n})}, \tag{48}$$

$$\eta_t \equiv \eta \le \frac{1}{3L}, \tag{49}$$

for any $t > t_1$. The proof of Lemma 7 is finished by summing up (45) from iteration $t_1 + 1$ to t and noting that (46) holds for any $t > t_1 + 1$. □

Now, we are ready to prove Theorem 3 by combining Lemma 5 (iterations $0 \le t \le t_1$) and Lemma 7 (iterations $t > t_1$).

Proof of Theorem 3. First, we note that $p_{t_1+1} = \frac{4}{1+3\sqrt{n}}$, $\theta_{t_1+1} = \frac{1}{2}$ and $\eta_{t_1+1} \le \frac{1}{L(1+1/(1-\theta_{t_1+1}))} = \frac{1}{3L}$. Then, by plugging (36) into (44) and noting that

$$\frac{(1 - p_{t_1+1}\theta_{t_1+1})\eta_{t_1+1}}{p_{t_1+1}\theta_{t_1+1}^2} = \frac{(1 - \frac{2}{1+3\sqrt{n}})\frac{1}{3L}}{\frac{1}{1+3\sqrt{n}}} = \frac{3\sqrt{n} - 1}{3L} \le \frac{4\sqrt{n}}{L}, \tag{50}$$

we have

$$\mathbb{E}\left[\frac{\eta_{t-1}}{p_{t-1}\theta_{t-1}^2}\big(f(w_t) - f(x^*)\big)\right]$$

$$\leq \mathbb{E}\Big[\frac{4\sqrt{n}}{L}\Big((1-\theta)\big(f(x_0)-f(x^*)\big)+\Big(\frac{\theta^2 p}{2\eta}+(1-p)L(1-\theta)\theta^2\Big)\|x_0-x^*\|^2$$
$$-\Big(\frac{\theta^2 p}{2\eta}-(1-p)L(1-\theta)\theta^2\Big)\|x_{t_1+1}-x^*\|^2\Big)+\frac{1}{2}\Big(\|x_{t_1+1}-x^*\|^2-\|x_t-x^*\|^2\Big)\Big]$$
$$\leq \mathbb{E}\Big[\frac{4\sqrt{n}}{L}\Big((1-\theta)\big(f(x_0)-f(x^*)\big)+\Big(\frac{\theta^2 p}{2\eta}+(1-p)L(1-\theta)\theta^2\Big)\|x_0-x^*\|^2\Big)$$
$$-\Big(\Big(\frac{\theta^2 p}{2\eta}-(1-p)L(1-\theta)\theta^2\Big)\frac{4\sqrt{n}}{L}-\frac{1}{2}\Big)\|x_{t_1+1}-x^*\|^2\Big]$$
$$\leq \mathbb{E}\Big[\frac{4\sqrt{n}}{L}\Big((1-\theta)\frac{L}{2}+\frac{\theta^2 p}{2\eta}+(1-p)L(1-\theta)\theta^2\Big)\|x_0-x^*\|^2$$
$$-\Big(\Big(\frac{\theta^2 p}{2\eta}-(1-p)L(1-\theta)\theta^2\Big)\frac{4\sqrt{n}}{L}-\frac{1}{2}\Big)\|x_{t_1+1}-x^*\|^2\Big] \tag{51}$$
$$\leq 8\|x_0-x^*\|^2, \tag{52}$$

where (51) holds due to the L-smoothness of f, (52) follows from the constant parameters i.e., $p_t \equiv p = \frac{1}{n+1}$, $\theta_t \equiv \theta = 1-\frac{1}{2\sqrt{n}}$ and $\eta_t \leq \frac{1}{L(1+1/(1-\theta_t))} \equiv \frac{1}{(1+2\sqrt{n})L}$ for $t \leq t_1$.

Finally, the proof of Theorem 3 is finished by multiplying $\frac{p_{t-1}\theta_{t-1}^2}{\eta_{t-1}}$ for both sides of (52), i.e., we have for any iteration $t > t_1+1$

$$\mathbb{E}[(f(w_t)-f(x^*)] \leq \frac{8p_{t-1}\theta_{t-1}^2\|x_0-x^*\|^2}{\eta_{t-1}} \overset{(48)}{=} \frac{32\|x_0-x^*\|^2}{\eta_{t-1}p_{t-1}(t-t_1+3\sqrt{n})^2}. \tag{53}$$

□

C.1.2. PROOF OF COROLLARY 1

Now, we provide the proof for Corollary 1 with detailed convergence result of ANITA in the general convex case (i.e., $\mu = 0$).

Proof of Corollary 1. Note that the output of ANITA (Algorithm 1) is w_T after T iterations. To show that w_T is an ϵ-approximate solution, we recall (53) with iteration $t = T > t_1+1$ here:

$$\mathbb{E}[(f(w_T)-f(x^*)] \leq \frac{32\|x_0-x^*\|^2}{\eta_{T-1}p_{T-1}(T-t_1+3\sqrt{n})^2}. \tag{54}$$

According to (47), we know $p_t = \max\left\{\frac{4}{t-t_1+3\sqrt{n}}, \frac{4}{n+3}\right\}$ for any $t > t_1$. Thus we divide (54) into two cases, i) $p_t = \frac{4}{t-t_1+3\sqrt{n}}$ for $t_1 < t \leq t_1+n+3-3\sqrt{n}$; ii) $p_t = \frac{4}{n+3}$ for $t > t_1+n+3-3\sqrt{n}$.

Now, we know that for Case i) $t \leq t_1+n+3-3\sqrt{n}$, then $p_t = \frac{4}{t-t_1+3\sqrt{n}}$, $\theta_t = \frac{2}{p_t(t-t_1+3\sqrt{n})} = \frac{1}{2}$, $\eta_t \leq \frac{1}{3L}$, and (54) turns to

$$\mathbb{E}[(f(w_T)-f(x^*)] \leq \frac{24L\|x_0-x^*\|^2}{T-t_1+3\sqrt{n}} \leq \epsilon. \tag{55}$$

The last inequality of (55) holds by choosing $T = t_1 - 3\sqrt{n} + \frac{24L\|x_0-x^*\|^2}{\epsilon}$. In particular, if $\epsilon \geq O(\frac{1}{n})$, then (recall that $\mathbb{E}[t_1] = n$ and also it can be derandomized to n iterations)

$$T = t_1 - 3\sqrt{n} + \frac{96L\|x_0-x^*\|^2}{\epsilon} \leq 2n. \tag{56}$$

For the other case $\epsilon < O(\frac{1}{n})$ (small target error), it corresponds to Case ii) $t > t_1 + n + 3 - 3\sqrt{n}$ (i.e., more iterations are needed), then $p_t = \frac{4}{n+3}$, $\theta_t = \frac{2}{p_t(t-t_1+3\sqrt{n})} = \frac{n+3}{2(t-t_1+3\sqrt{n})} \le \frac{1}{2}$, $\eta_t \le \frac{1}{3L}$ and (54) turns to

$$\mathbb{E}[(f(w_T) - f(x^*)] \le \frac{24(n+3)L\|x_0 - x^*\|^2}{(T - t_1 + 3\sqrt{n})^2} \le \epsilon. \tag{57}$$

The last inequality of (57) holds by choosing

$$T = t_1 - 3\sqrt{n} + \sqrt{\frac{24(n+3)L\|x_0 - x^*\|^2}{\epsilon}} \le n + \sqrt{\frac{24(n+3)L\|x_0 - x^*\|^2}{\epsilon}}. \tag{58}$$

Now, the remaining thing is to bound the number of stochastic gradient computations of ANITA for achieving the ϵ-approximate solution w_T. As we discussed in Section 2.1, we know that ANITA (Algorithm 1) uses $(n+2)p_t + 2(1-p_t) = np_t + 2$ stochastic gradients in expectation for iteration t. According to the choice of probability $\{p_t\}$ in Corollary 1 (Theorem 3), we know that there are three stages. 1) The first stage $p_t \equiv \frac{1}{n+1}$ for $0 \le t \le t_1$; 2) the second stage $p_t = \frac{4}{t-t_1+3\sqrt{n}}$ for $t_1 < t \le t_1 + n + 3 - 3\sqrt{n}$; 3) the third stage $p_t \equiv \frac{4}{n+3}$ for $t > t_1 + n + 3 - 3\sqrt{n}$.

First, let us consider the case of large ϵ (i.e., $\epsilon \ge O(\frac{1}{n})$). Then we know that only the first two stages of ANITA is enough for finding an ϵ-approximate solution in this case. According to (55), the total number of stochastic gradient computations is

$$\begin{aligned}
\#\text{grad} = \sum_{t=0}^{T-1}(np_t + 2) &= n\Big(\sum_{t=0}^{t_1} p_t + \sum_{t=t_1+1}^{T-1} p_t\Big) + 2T \\
&= n\Big(\sum_{t=0}^{t_1} \frac{1}{n+1} + \sum_{t=t_1+1}^{T-1} \frac{4}{t - t_1 + 3\sqrt{n}}\Big) + 2T \\
&\le O\bigg(n\Big(1 + \log\frac{1}{\epsilon\sqrt{n}}\Big)\bigg),
\end{aligned} \tag{59}$$

where the last inequality (59) follows from (56).

Then, for the other case $\epsilon < O(\frac{1}{n})$ (small target error), we know that more iterations are needed for finding an ϵ-approximate solution. According to (57), the total number of stochastic gradient computations is

$$\begin{aligned}
\#\text{grad} &= \sum_{t=0}^{T-1}(np_t + 2) \\
&= n\Big(\sum_{t=0}^{t_1} \frac{1}{n+1} + \sum_{t=t_1+1}^{t_1+n+3-3\sqrt{n}} \frac{4}{t - t_1 + 3\sqrt{n}} + \sum_{t=t_1+n+4-3\sqrt{n}}^{T-1} \frac{4}{n+3}\Big) + 2T \\
&\le O\left(n\log\sqrt{n} + \sqrt{\frac{nL}{\epsilon}}\right),
\end{aligned} \tag{60}$$

where the last inequality (60) follows from (58). □

C.2. Proofs for strongly convex case

Similar to Appendix C.1, we first provide the proof of the main convergence Theorem 4 for the strongly convex case (i.e., $\mu > 0$) in Appendix C.2.1. Then we provide the proof for its Corollary 2 with detailed convergence result in Appendix C.2.2.

C.2.1. PROOF OF THEOREM 4

In this strongly convex case, the parameter setting of ANITA in Theorem 4 is simpler than the general convex case in Theorem 3. Here, the choice of probability $\{p_t\}$ can be fixed to a constant p and $\{\theta_t\}$ also can be chosen as a constant θ. Then according to Theorem 4, we know that $\{\eta_t\}$ and $\{\alpha_t\}$ also reduce to constant values. Thus there is only one stage in this strongly convex case rather than three stages in previous general convex case. Also here the function value decreases in an exponential rate, i.e., ANITA obtains a linear convergence rate.

Proof of Theorem 4. First, we restate the key Lemma 8 for this strong convex setting, which describes the change of function value after a gradient update step in ANITA. Note that Lemma 8 directly follows from our technical Lemma 11 with $\alpha_t = 1 + \mu\eta_t$.

Lemma 8 *Suppose that Assumptions 1 and 2 hold. Choose stepsize* $\eta_t \le \frac{1}{L\theta_t(1+1/(1-\theta_t))}$ *and* $\alpha_t = 1 + \mu\eta_t$ *for any* $t \ge 0$. *Then the following equation holds for* ANITA *(Algorithm 1) for any iteration* $t \ge 0$*:*

$$\mathbb{E}\left[f(w_{t+1}) - f(x^*) + \frac{(1+\mu\eta_t)p_t\theta_t}{2\eta_t}\|x_{t+1} - x^*\|^2\right] \le \mathbb{E}\left[(1 - p_t\theta_t)\big(f(w_t) - f(x^*)\big) + \frac{p_t\theta_t}{2\eta_t}\|x_t - x^*\|^2\right]. \tag{61}$$

Then, according to the parameter settings chosen in Theorem 4, we know that $p_t \equiv p$ and $\theta_t \equiv \theta = \frac{1}{2}\min\{1, \sqrt{\frac{\mu}{pL}}\}$ for any $t \ge 0$, and the stepsize $\eta_t \le \frac{1}{L\theta_t(1+1/(1-\theta_t))} \equiv \eta = \frac{1}{L\theta(1+1/(1-\theta))}$. Now, we further define

$$\Phi_t := f(w_t) - f(x^*) + \frac{(1+\mu\eta)p\theta}{2\eta}\|x_t - x^*\|^2, \tag{62}$$

then (61) in Lemma 8 can be changed to, for any iteration $t \ge 0$,

$$\begin{aligned}
\mathbb{E}[\Phi_{t+1}] &\le \mathbb{E}\left[\max\left\{1 - p\theta,\ \frac{1}{1+\mu\eta}\right\}\Phi_t\right] \\
&\le \mathbb{E}\left[\left(1 - \frac{4p\theta}{5}\right)\Phi_t\right] \qquad (63) \\
&\le \left(1 - \frac{4p\theta}{5}\right)^{t+1}\Phi_0. \qquad (64)
\end{aligned}$$

where (63) uses $\frac{1}{1+\mu\eta} \le 1 - \frac{4p\theta}{5}$ since the choice of parameters $\theta = \frac{1}{2}\min\{1, \sqrt{\frac{\mu}{pL}}\}$ and $\eta = \frac{1}{L\theta(1+1/(1-\theta))}$, and the last inequality (64) holds by telescoping (63) from iteration t to 0. $\square$

C.2.2. PROOF OF COROLLARY 2

Now, we provide the proof for Corollary 2 with detailed convergence result of ANITA in the strongly convex case (i.e., $\mu > 0$).

Proof of Corollary 2. Note that the output of ANITA (Algorithm 1) is w_T after T iterations. To show that w_T is an ϵ-approximate solution, we recall (64) with iteration $t = T - 1$:

$$\mathbb{E}[(f(w_T) - f(x^*)] \le \mathbb{E}[\Phi_T] \le \Big(1 - \frac{4p\theta}{5}\Big)^T \Phi_0 \le \epsilon, \tag{65}$$

where the first inequality is due to the definition of Φ_T (see (62)), and the last inequality holds by letting the number of iterations

$$T = \frac{5}{4p\theta} \log \frac{\Phi_0}{\epsilon}.$$

Moreover, by choosing $p = \frac{1}{n}$ and recalling that $\theta = \frac{1}{2}\min\{1, \sqrt{\frac{\mu}{pL}}\}$, then the total number of stochastic gradient computations of ANITA for achieving the ϵ-approximate solution w_T is

$$\#\text{grad} = \sum_{t=0}^{T-1}(np_t + 2) = \Big(n\frac{1}{n} + 2\Big)T = \frac{15}{4p\theta}\log\frac{\Phi_0}{\epsilon} = O\left(\max\left\{n, \sqrt{\frac{nL}{\mu}}\right\}\log\frac{1}{\epsilon}\right).$$

$\square$