Proceedings of Machine Learning Research vol 99:1–20, 2019 32nd Annual Conference on Learning Theory

# Stochastic Approximation of Smooth and Strongly Convex Functions: Beyond the $O(1/T)$ Convergence Rate

**Lijun Zhang** ZHANGLJ@LAMDA.NJU.EDU.CN
**Zhi-Hua Zhou** ZHOUZH@LAMDA.NJU.EDU.CN
*National Key Laboratory for Novel Software Technology*
*Nanjing University, Nanjing 210023, China*

**Editors:** Alina Beygelzimer and Daniel Hsu

## Abstract

Stochastic approximation (SA) is a classical approach for stochastic convex optimization. Previous studies have demonstrated that the convergence rate of SA can be improved by introducing either smoothness or strong convexity condition. In this paper, we make use of smoothness and strong convexity *simultaneously* to boost the convergence rate. Let be the modulus of strong convexity, be the condition number, $F_*$ be the minimal risk, and $> 1$ be some *small* constant. First, we demonstrate that, in expectation, an $O(1/[\ T^{\alpha}] + \ F_*/T)$ risk bound is attainable when $T = \Omega(\ ^{\alpha})$. Thus, when $F_*$ is small, the convergence rate could be faster than $O(1/[\ T])$ and approaches $O(1/[\ T^{\alpha}])$ in the ideal case. Second, to further benefit from small risk, we show that, in expectation, an $O(1/2^{T/\kappa} + F_*)$ risk bound is achievable. Thus, the excess risk reduces exponentially until reaching $O(F_*)$, and if $F_* = 0$, we obtain a global linear convergence. Finally, we emphasize that our proof is constructive and each risk bound is equipped with an efficient stochastic algorithm attaining that bound.

**Keywords:** Stochastic Approximation, Stochastic Convex Optimization, Excess Risk, Smoothness, Strong Convexity

## 1. Introduction

Stochastic optimization (SO) is frequently encountered in a vast number of areas, including telecommunication, medicine, and finance, to name but a few (Shapiro et al., 2014). SO aims to minimize an objective function which is given in a form of the expectation. Formally, the problem can be formulated as

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{f\sim\mathbb{P}}[f(\mathbf{w})] \tag{1}$$

where $f(\ ): W \mapsto \mathbb{R}$ is a random function sampled from a distribution $\mathbb{P}$. A well-known special case is the risk minimization in machine learning, whose objective function is

$$F(\mathbf{w}) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell\big(y, \langle\mathbf{w}, \mathbf{x}\rangle\big)\right]$$

where $(\mathbf{x}, y)$ denotes a random instance-label pair sampled from certain distribution $\mathbb{D}$, $\mathbf{w}$ is the model for prediction, and $\ell(\ ,\ )$ is a loss that measures the prediction error (Vapnik, 1998).

In this paper, we focus on stochastic convex optimization (SCO), in which both the domain $W$ and the expected function $F(\ )$ are convex. A basic difficulty of solving stochastic optimization problem is that the distribution $\mathbb{P}$ is generally unknown, or even if known, it is hard to evaluate the expectation exactly (Nemirovski et al., 2009). To address this challenge, two different ways have been proposed: sample average approximation (SAA) (Kim et al., 2015) and stochastic approximation (SA) (Kushner and Yin, 2003). SAA collects a set of random functions $f_1; : : : ; f_T$ from $\mathbb{P}$, and constructs the empirical average $\sum_{i=1}^{T} f_i(\ )=T$ to approximate the expected function $F(\ )$. In contrast, SA tackles the stochastic optimization problem directly, at each iteration using a noisy observation of $F(\ )$ to improve the current iterate.

Compared with SAA, SA is more efficient due to the low computational cost per iteration, and has received significant research interests from optimization and machine learning communities (Zhang, 2004; Duchi et al., 2011; Ge et al., 2015; Wang et al., 2017). The performance of SA algorithms is typically measured by the excess risk:

$$F(\mathbf{w}_T) \quad \min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$$

where $\mathbf{w}_T$ is the solution returned after $T$ iterations. For Lipschitz continuous convex functions, stochastic gradient descent (SGD) achieves the unimprovable $O(1=\sqrt{T})$ rate of convergence. Alternatively, if the optimization problem has certain curvature properties, then faster rates are sometimes possible. Specifically, for smooth functions, SGD is equipped with an $O(1=T+\sqrt{F_*=T})$ risk bound, where $F_* = \min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$ is the minimal risk (Srebro et al., 2010). Thus, the convergence rate for smooth functions could be faster than $O(1=\sqrt{T})$ when the minimal risk is small. For strongly convex functions, the convergence rate can also be improved to $O(1=[\ T])$, where is the modulus of strong convexity (Hazan and Kale, 2011).

From the above discussions, we observe that either smoothness or strong convexity could be exploited to improve the convergence rate of SA. This observation motivates subsequent studies that boost the convergence rate by considering smoothness and strong convexity *simultaneously*. However, existing results are unsatisfactory because they either rely on strong assumptions (Mahdavi and Jin, 2013; Schmidt and Roux, 2013), are only applicable to unconstrained domains (Moulines and Bach, 2011; Needell et al., 2014), or limited to the special problems (Roux et al., 2012; Shalev-Shwartz and Zhang, 2013; Zhang et al., 2013a; Dieuleveut et al., 2017). This paper demonstrates that for the general SO problem, the convergence rate of SA could be faster than $O(1=T)$ when both smoothness and strong convexity are present and the minimal risk is small. Our work is similar in spirit to a recent study of SAA (Zhang et al., 2017a), which also establishes faster rates under similar conditions. The main contributions of our paper are summarized below.

- First, we propose a fast algorithm for stochastic approximation (FASA), which applies epoch gradient descent (Epoch-GD) (Hazan and Kale, 2011) with carefully designed initial solution and step size. Let be the condition number and $> 1$ be some small constant. Our theoretical analysis shows that, in expectation, FASA achieves an $O\left(1=[\ T^{\alpha}] + \ F_*=T\right)$ risk bound when the number of iterations $T = \Omega(\ ^{\alpha})$. As a result, the convergence rate could be faster than $O(1=[\ T])$ when $F_*$ is small, and approaches $O(1=[\ T^{\alpha}])$ when $F_* = O(1=T^{\alpha-1})$.
- Second, to further benefit from small risk, we propose to use a fixed step size in Epoch-GD, and establish an $O(1=2^{T/\kappa} + F_*)$ risk bound which holds in expectation. Thus, the excess risk reduces exponentially until reaching $O(F_*)$, and if $F_* = 0$, we obtain a global linear convergence.

## 2. Related Work

In this section, we review related work on SA and SAA.

### 2.1. Stochastic Approximation (SA)

concave, an $\widetilde{O}(d/[\lambda T])$ risk bound is attainable (Koren and Levy, 2015; Mehta, 2016). Lower bounds of ERM for stochastic optimization are investigated by Feldman (2016). In a recent work, Zhang et al. (2017a) establish an $\widetilde{O}(d/T + \sqrt{F_*/T})$ risk bound when $f(\cdot)$ is smooth and $F(\cdot)$ is Lipschitz continuous. The most surprising result is that when $f(\cdot)$ is smooth and $F(\cdot)$ is Lipschitz continuous and $\lambda$-strongly convex, Zhang et al. (2017a) prove an $O(1/[\lambda T^2] + F_*/T)$ risk bound, when $T = \widetilde{\Omega}(\kappa d)$. Thus, the convergence rate of ERM could be faster than $O(1/[\lambda T])$ when both smoothness and strong convexity are present and the number of training samples is large enough.

## 3. Our Results

We first introduce assumptions used in our analysis, then present our algorithms and theoretical guarantees.

### 3.1. Assumptions

**Assumption 1** *The random function $f(\cdot)$ is nonnegative.*

**Assumption 2** *The random function $f(\cdot)$ is (almost surely) $L$-smooth over $\mathcal{W}$, that is,*

$$\left\| \nabla f(\mathbf{w}) - \nabla f(\mathbf{w}') \right\| \leq L\|\mathbf{w} - \mathbf{w}'\|, \ \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}. \tag{2}$$

**Assumption 3** *The expected function $F(\cdot)$ is $\lambda$-strongly convex over $\mathcal{W}$, that is,*

$$F(\mathbf{w}) + \langle \nabla F(\mathbf{w}), \mathbf{w}' - \mathbf{w} \rangle + \frac{\lambda}{2}\|\mathbf{w}' - \mathbf{w}\|^2 \leq F(\mathbf{w}'), \ \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}. \tag{3}$$

**Assumption 4** *The gradient of the random function is (almost surely) upper bounded by $G$, that is,*

$$\|\nabla f(\mathbf{w})\| \leq G, \ \forall \mathbf{w} \in \mathcal{W}. \tag{4}$$

**Remark 1** We have the following comments regarding our assumptions.

- The above assumptions hold for many popular machine learning problems, such as (regularized) linear regression or logistic regression.
- Based on Assumptions 2 and 3, we define the condition number $\kappa = L/\lambda$, which will be used to characterize the performance of our methods. For simplicity, we assume $L$ is a constant, and thus $\kappa$ and $1/\lambda$ are on the same order.
- Let $\mathbf{w}_* = \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}} F(\mathbf{w})$ be the optimal solution to (1). Assumption 3 implies (Hazan and Kale, 2011)

$$\frac{\lambda}{2}\|\mathbf{w} - \mathbf{w}_*\|^2 \leq F(\mathbf{w}) - F(\mathbf{w}_*), \ \forall \mathbf{w} \in \mathcal{W} \tag{5}$$

  which is referred to as the quadratic growth condition (Necoara et al., 2019). Actually, in our analysis, we only make use of (5) instead of (3).
- The bounded gradient condition in Assumption 4 is not essential to our analysis. We introduce this assumption because our first algorithm uses Epoch-GD (Hazan and Kale, 2011) as a subroutine and Epoch-GD relies on Assumption 4. However, Epoch-GD can be replaced with any optimal algorithm for strongly convex stochastic optimization. In particular, if we choose the AC-SA algorithm of Ghadimi and Lan (2012), Assumption 4 can be dropped.

**Algorithm 1** Epoch Gradient Descent (Epoch-GD)

**Input**: parameters $_1$, $T_1$, $T$, and $\mathbf{w}_0$

Initialize $\mathbf{w}_1^1 = \mathbf{w}_0$, and set $k = 1$
**while** $\sum_{i=1}^{k} T_i \quad T$ **do**
  **for** $t = 1$ to $T_k$ **do**
    Sample a random function $f_t^k(\ )$ from $\mathbb{P}$
    Update
$$\mathbf{w}_{t+1}^k = \Pi_{\mathcal{W}} \left[ \mathbf{w}_t^k \quad {}_k\, r\, f_t^k(\mathbf{w}_t^k) \right]$$
  **end for**
  $\mathbf{w}_1^{k+1} = \frac{1}{T_k} \sum_{t=1}^{T_k} \mathbf{w}_t^k$
  $T_{k+1} = 2T_k$ and ${}_{k+1} = {}_k$=2
  $k = k + 1$
**end while**
**return** $\mathbf{w}_1^k$

**Algorithm 2** Fast Algorithm for Stochastic Approximation (FASA)

**Input**: parameters $L$, , $T$, and

Let $\bar{\mathbf{w}}$ be any point in $W$, and set $\quad = L$=
Invoke Epoch-GD($1$= ,$4$,$T$=$2$, $\bar{\mathbf{w}}$), and denote the solution by $\widehat{\mathbf{w}}$
Invoke Epoch-GD($1$=$4L$,$2^{\alpha+3}$ ,$T$=$2$, $\widehat{\mathbf{w}}$), and denote the solution by $\widetilde{\mathbf{w}}$
**return** $\widetilde{\mathbf{w}}$

### 3.2. A General Algorithm

We first introduce a general algorithm for SA, which always achieves an $O(1$= $T)$ rate, and becomes faster when $F_*$ is small.

#### 3.2.1. FAST ALGORITHM FOR STOCHASTIC APPROXIMATION (FASA)

Our fast algorithm for stochastic approximation (FASA) takes epoch gradient descent (Epoch-GD) as a subroutine. Although Hazan and Kale (2011) have established the convergence rate of Epoch-GD under the strong convexity condition, they did not utilize smoothness in their analysis. The procedures of Epoch-GD and FASA are described in Algorithm 1 and Algorithm 2, respectively.

Epoch-GD is an extension of stochastic gradient descent (SGD). It divides the optimization process into a sequence of epochs. In each epoch, Epoch-GD applies SGD multiple times, and the averaged iterate is passed to the next epoch. In the algorithm, we use $\Pi_{\mathcal{W}}[\ ]$ to denote the projection onto the nearest point in $W$. There are $4$ input parameters of Epoch-GD: (1) $_1$, the step size used in the first epoch; (2) $T_1$, the size of the first epoch; (3) $T$, the total number of stochastic gradients that can be consumed; and (4) $\mathbf{w}_0$, the initial solution. In each consecutive epoch, the step size decreases exponentially and the size of epoch increases exponentially.

In FASA, we first invoke Epoch-GD with an arbitrary initial solution, and the number of stochastic gradients is set to be $T$=$2$. The purpose of this step is to get a good solution $\widehat{\mathbf{w}}$ at the expense of $T/2$ stochastic gradients.[1] Then, Epoch-GD is invoked again with $\widehat{\mathbf{w}}$ as its initial solution and a budget of $T/2$ stochastic gradients. This time, we set a large epoch size to utilize the fact that the initial solution is of high quality. The convergence rate of FASA is given below.

**Theorem 1** *Suppose*

$$T \geq \kappa^\alpha \tag{6}$$

*where* $\alpha > 1$ *is some constant. Under Assumptions 1, 2, 3 and 4, the solution* $\widetilde{\mathbf{w}}$ *returned by Algorithm 2 satisfies*

$$\mathrm{E}\left[F(\widetilde{\mathbf{w}})\right] - F_* \leq \frac{2^{\alpha^2+5\alpha+5}G^2}{T^\alpha} + \frac{2^{2\alpha+5}\kappa F_*}{(2^{\alpha-1}-1)T}$$

*where* $F_* = F(\mathbf{w}_*) = \min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$ *is the minimal risk.*

**Remark 2** The above theorem implies that when $T$ is large enough, i.e., $T = \Omega(\kappa^\alpha)$, FASA achieves an

$$O\left(\frac{1}{T^\alpha} + \frac{F_*}{T}\right)$$

rate of convergence, which is faster than $O(1/[\lambda T])$ when the minimal risk is small. In particular, when $F_* = O(1/T^{\alpha-1})$, the convergence rate is improved to $O(1/[\lambda T^\alpha])$. Note that the upper bound has an exponential dependence on $\alpha$, so it is meaningful only when $\alpha$ is chosen as a *small* constant.

**Remark 3** Note that our algorithm is translation-invariant, i.e., it does not change if we translate the function by a constant. Since the upper bound in Theorem 1 depends on the minimal risk $F_*$, one may attempt to subtract a constant from the function to make the bound tighter. However, because of the nonnegative requirement in Assumption 1, the best we can do is to redefine

$$f(\mathbf{w}) \leftarrow f(\mathbf{w}) - \operatorname*{ess\,inf}_{f\sim\mathbb{P}} \inf_{\mathbf{w}\in\mathcal{W}} f(\mathbf{w})$$

and replace $F_*$ in Theorem 1 with $F_* - \operatorname{ess\,inf}_{f\sim\mathbb{P}} \inf_{\mathbf{w}\in\mathcal{W}} f(\mathbf{w})$.

To simplify Theorem 1, we provide the following corollary by setting $\alpha = 2$.

**Corollary 2** *Suppose* $T \geq \kappa^2$*. Under the same conditions as Theorem 1, we have*

$$\mathrm{E}\left[F(\widetilde{\mathbf{w}})\right] - F_* \leq \frac{2^{19}G^2}{T^2} + \frac{2^9\kappa F(\mathbf{w}_*)}{T} = O\left(\frac{1}{T^2} + \frac{F_*}{T}\right).$$

Finally, we present the excess risk when $\alpha$ is set to be the largest possible value, i.e., $\alpha = \log_\kappa T$, and then the first term in the upper bound of Theorem 1 decreases at a very fast rate.

**Corollary 3** *Suppose* $\alpha = \log_\kappa T > 1$*. Under the same conditions as Theorem 1, we have*

$$\begin{aligned}\mathrm{E}\left[F(\widetilde{\mathbf{w}})\right] - F_* \leq & \frac{2^5 G^2}{T^{(1-\log_\kappa 2 - 5\log_T 2)\log_\kappa T}} + \frac{2^5\kappa F_*}{(2^{\log_\kappa T - 1}-1)T^{1-2\log_\kappa 2}} \\ = & O\left(\frac{1}{T^{(1-\log_\kappa 2-5\log_T 2)\log_\kappa T}} + \frac{F_*}{T^{1-2\log_\kappa 2}}\right).\end{aligned}$$

1. In this step, Epoch-GD can be replaced with any algorithm that achieves the optimal $O(1/\lambda T)$ rate for strongly convex stochastic optimization, e.g., the AC-SA algorithm (Ghadimi and Lan, 2012) and SGD with $\alpha$-suffix averaging (Rakhlin et al., 2012).

**Remark 4** Note that any constant that is larger than $\kappa$ can also be used as the condition number. Thus, without loss of generality, we can assume $\log_\kappa 2$ is much smaller than $1$. Furthermore, as $T$ goes to infinity, $5\log_T 2$ converges to $0$. So, when $F_*$ is sufficiently small, the convergence rate will approach $O(1/[\lambda T^{c\log_\kappa T}])$ where $c = 1 - \log_\kappa 2$.

### 3.2.2. COMPARISONS WITH PREVIOUS RESULTS

In the following, we compare our Theorem 1 and Corollary 2 with related work in SA (Ghadimi and Lan, 2012; Dieuleveut et al., 2017; Jain et al., 2018; Moulines and Bach, 2011; Needell et al., 2014) and SAA (Zhang et al., 2017a).

For smooth and strongly convex functions, Ghadimi and Lan (2012, Proposition 9) have established an $O(1/T^2 + \sigma^2/[\lambda T])$ rate for the expected risk, where $\sigma^2$ is the variance of the stochastic gradient. Note that this rate is worse than that in Corollary 2 because $\sigma^2$ is a constant in general, even when $F_*$ is small. For example, consider the problem of least squares

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[(\mathbf{x}^\top\mathbf{w} - y)^2\right],$$

and assume $y = \mathbf{x}^\top\mathbf{w}_* + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$ is the Gaussian random noise and $\mathbf{w}_* \in \mathcal{W}$. Then, $F_* = \mathrm{E}[\epsilon^2] = \sigma^2$, which approaches zero as $\sigma \to 0$. On the other hand, the variance of the stochastic gradient at solution $\mathbf{w}_t$ can be decomposed as

$$\begin{aligned}\sigma^2 =& \mathrm{E}\left[\left\|2(\mathbf{x}^\top\mathbf{w}_t - y)\mathbf{x} - \mathrm{E}\left[2(\mathbf{x}^\top\mathbf{w}_t - y)\mathbf{x}\right]\right\|^2\right]\\ =& 4\mathrm{E}\left[\left\|(\mathbf{x}\mathbf{x}^\top - \mathrm{E}[\mathbf{x}\mathbf{x}^\top])(\mathbf{w}_t - \mathbf{w}_*)\right\|^2\right] + 4\mathrm{E}\left[\|\epsilon\mathbf{x}\|^2\right].\end{aligned}$$

As can be seen, even when there is no noise, i.e., $\sigma = 0$, the variance is nonzero due to the randomness of $\mathbf{x}$. We note that recent studies of least squares do not suffer this limitation, and in particular, Dieuleveut et al. (2017, Theorem 2) and Jain et al. (2018, Corollary 2) have proved $O(d/T^2 + d\sigma^2/T)$ and $O(\exp(-\mu T/\kappa) + d\sigma^2/T)$ rates, respectively. These results have a similar spirit with our Corollary 2, but they are limited to least squares.

For unconstrained problems, Moulines and Bach (2011) and Needell et al. (2014) have analyzed the distance between the SGD iterate and the optimal solution under the smoothness and strong convexity condition. In particular, Theorem 1 of Moulines and Bach (2011) (with $\alpha = 1$ and $C = 2$) implies the following convergence rate for the expected risk

$$O\left(\frac{\exp(\kappa^2)}{n^2} + \frac{F_*\log T}{\mu^2 T}\right)$$

which is worse than our Corollary 2 because of the additional $\log T/\mu$ factor in the second term. Theorem 2.1 of Needell et al. (2014) leads to the following rate

$$O\left(\left(1 - \frac{\mu}{T}\right)^T + \frac{F_*}{T}\right) \tag{7}$$

which is also worse than our Corollary 2 because $(1 - \mu/T)^T$ becomes a constant when $T \to \infty$. We note that it is possible to extend the analysis of Needell et al. (2014) to constrained problems, but

**Algorithm 3** Epoch Gradient Descent with Fixed Step Size (Epoch-GD-F)

**Input**: parameters , $T'$, $T$, and $\mathbf{w}_0$

Set $\mathbf{w}_1^1 = \mathbf{w}_0$ and $k = 1$
**while** $k \quad T = T'$ **do**
  **for** $t = 1$ to $T'$ **do**
    Sample a random function $f_t^k(\ )$ from $\mathbb{P}$
    Update
$$\mathbf{w}_{t+1}^k = \Pi_{\mathcal{W}} \left[ \mathbf{w}_t^k \qquad r\, f_t^k(\mathbf{w}_t^k) \right]$$
  **end for**
  $\mathbf{w}_1^{k+1} = \frac{1}{T'} \sum_{t=1}^{T'} \mathbf{w}_t^k$
  $k = k + 1$
**end while**
**return** $\widetilde{\mathbf{w}} = \mathbf{w}_1^k$

the convergence rate becomes slower, and thus is worse than our rate. Detailed discussions about how to simplify and extend the result of Needell et al. (2014) are provided in Appendix E.

The convergence rate in Corollary 2 matches the state-of-the-art convergence rate of SAA (Zhang et al., 2017a). Specifically, under similar conditions, Zhang et al. (2017a, Theorem 3) have proved an $O(1=[\ T^2] + \ F_*=T)$ risk bound for SAA, when $T = \widetilde{\Omega}(\ d)$. Compared with the results of Zhang et al. (2017a), our theoretical guarantees have the following advantages:

The lower bound of $T$ in our results is independent from the dimensionality, and thus our results can be applied to infinite dimensional problems, e.g., learning with kernels. In contrast, the lower bound of $T$ given by Zhang et al. (2017a, Theorem 3) depends on the dimensiona-
ults

**Theorem 4** *Set*

$$= \frac{1}{4 \ L}; \ T' = 16 \tag{8}$$

*where* $\ > 1$ *is some constant, and* $\mathbf{w}_0$ *be any point in* $W$. *Under Assumptions 1, 2 and 3, the solution* $\widetilde{\mathbf{w}}$ *returned by Algorithm 3 satisfies*

$$\mathrm{E}\left[F(\widetilde{\mathbf{w}})\right] \quad F_* \qquad \frac{F(\mathbf{w}_0) \quad F_*}{2^{k^\dagger}} + \frac{2F_*}{}$$

*where* $k^\dagger = bT{=}T'c$.

**Remark 5** From the above theorem, we observe that the excess risk is upper bounded by two terms: the first one decreases *exponentially* w.r.t. the number of epoches and the second one depends on $F_*$. When $\ = O(1)$, the excess risk is on the order of

$$O\left(\frac{1}{2^{T/\kappa}} + F_*\right.$$

Their sample complexity has a linear dependent on the dimensionality $d$, in contrast ours is dimensionality-independent. Thus, our results can be applied to the non-parametric setting where hypotheses lie in a functional space of infinite dimension.

The dependence of their sample complexity on  and  is much higher than ours.

Under a strong growth condition (Solodov, 1998), Schmidt and Roux (2013) have established the following linear convergence rate for SGD when applied to unconstrained problems:

$$O\left(\left(1 \quad \frac{1}{\ }\right)^{T}\right):$$

This strong growth condition requires that all stochastic gradients are $0$ at $\mathbf{w}_*$, which is itself a necessary condition for $F_* = 0$, because all the random functions are nonnegative. In this case, our Theorem 4 also achieves a linear rate at the same order. However, our results have the following advantages:

Our Theorem 4 is more general because it covers the cases that $F_*$ is nonzero.

Our results are applicable even when there is a domain constraint.

For unconstrained problems, Theorem 2.1 of Needell et al. (2014) with a suitable step size also implies the following rate

$$O\left(\left(1 \quad \frac{1}{\ }\right)^{T} + \quad F_*\right) \tag{9}$$

which is slower than our $O(2^{-T/\kappa} + F_*)$ rate in Theorem 4, because of the additional dependence on  in the second term. Besides, Needell et al. (2014, (2.4) and (2.2)) provided the iteration complexity of their algorithm, as well as that of Moulines and Bach (2011) when the minimal risk $F_*$ is *known*. Specifically, the iteration complexities of Moulines and Bach (2011) and Needell et al. (2014) for finding an  -optimal solution are

$$\Omega\left(\log\frac{1}{\ }\left(\ ^2 + \frac{^2F_*}{\ }\right)\right) \text{ and } \Omega\left(\log\frac{1}{\ }\left(\ \ + \frac{^2F_*}{\ }\right)\right); \tag{10}$$

respectively. In this case, our Theorem 4 with $\ = \max(1; 4F_*= \ )$ implies the following iteration complexity

$$\Omega\left(\log\frac{1}{\ }\left(\ \ + \frac{F_*}{\ }\right)\right): \tag{11}$$

Compared with the lower bounds in (10), our iteration complexity is better because (i) it has a smaller dependence on  , and (ii) it holds for constrained problems.

## 4. Analysis

Due to the limitation of space, we only prove Theorem 1 and Corollary 3. The omitted proofs are provided in the appendices. Our analysis follows from well-known and standard techniques, including the analysis of stochastic gradient descent (Zinkevich, 2003), self-bounding property of smooth functions (Srebro et al., 2010), and the quadratic growth condition of strong convexity (Hazan and Kale, 2011).

### 4.1. Proof of Theorem 1

We first state the excess risk of $\widehat{\mathbf{w}}$, the solution returned by the first call of Epoch-GD. From Theorem 5 of Hazan and Kale (2014), we have

$$\mathrm{E}\left[F(\widehat{\mathbf{w}})\right] - F(\mathbf{w}_*) \leq \frac{32G^2}{T} \overset{(6)}{\leq} \frac{32G^2}{\alpha}. \tag{12}$$

We proceed to analyze the solution returned by the second call of Epoch-GD. In each epoch, the standard stochastic gradient descent (SGD) (Zinkevich, 2003) is applied. The following lemma shows how the excess risk decreases in each epoch.

**Lemma 1** *Apply $T$ iterations of the update*

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}\left[\mathbf{w}_t - \eta \nabla f_t(\mathbf{w}_t)\right]$$

*where $f_t(\cdot)$ is a random function sampled from $\mathbb{P}$, and $\eta < 1/(2L)$. Assume $F(\cdot)$ is convex and Assumptions 1 and 2 hold, for any $\mathbf{w} \in \mathcal{W}$, we have*

$$\mathrm{E}\left[F(\bar{\mathbf{w}})\right] - F(\mathbf{w}) \leq \frac{1}{2\eta T(1-2\eta L)}\mathrm{E}\left[\|\mathbf{w}_1 - \mathbf{w}\|^2\right] + \frac{2\eta L}{(1-2\eta L)}F(\mathbf{w})$$

*where $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$.*

Compared with the traditional analysis of SGD, e.g., Lemma 7 of Hazan and Kale (2011), the main difference is that Lemma 1 exploits smoothness to control the squared norm of stochastic gradients, and in this way, the excess risk depends on the function value instead of the upper bound of stochastic gradients.

Based on the above lemma, we establish the following result for bounding the excess risk of the intermediate iterate.

**Lemma 2** *Consider the second call of Epoch-GD with parameters $(1/4L, 2^{\alpha+3}, T/2, \widehat{\mathbf{w}})$. For any $k$, we have*

$$\mathrm{E}[F(\mathbf{w}_1^{k+1})] - F(\mathbf{w}_*) \leq \frac{2^{\alpha^2+2\alpha+5}G^2}{(T_k)^{\alpha}} + \frac{2^{\alpha+3}\ F(\mathbf{w}_*)}{T_k}\left(\sum_{i=1}^{k}\frac{1}{2^{(i-1)(\alpha-1)}}\right). \tag{13}$$

The number of epochs made is given by the largest value of $k$ satisfying $\sum_{i=1}^{k} T_i \leq T/2$, i.e.,

$$\sum_{i=1}^{k} T_i = T_1 \sum_{i=1}^{k} 2^{i-1} = T_1(2^k - 1) \leq \frac{T}{2}.$$

This value is

$$k^{\dagger} = \left\lfloor \log_2\left(\frac{T}{2T_1} + 1\right)\right\rfloor,$$

and the final solution is $\widetilde{\mathbf{w}} = \mathbf{w}_1^{k^\dagger+1}$. From Lemma 2, we have

$$
\begin{aligned}
& F(\mathbf{w}_1^{k^\dagger+1}) \quad F(\mathbf{w}_*) \\
& \frac{2^{\alpha^2+2\alpha+5} G^2}{(T_{k^\dagger})^\alpha} + \frac{2^{\alpha+3} \quad F(\mathbf{w}_*)}{T_{k^\dagger}} \left( \sum_{i=1}^{k^\dagger} \frac{1}{2^{(i-1)(\alpha-1)}} \right) \\
& \frac{2^{\alpha^2+2\alpha+5} G^2}{}
\end{aligned}
$$

## References


Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 699–707, 2016.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems 26*, pages 773–781, 2013.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, 18(101):1–51, 2017.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Vitaly Feldman. Generalization of ERM in stochastic convex optimization: The dimension strikes back. In *Advances in Neural Information Processing Systems 29*, pages 3576–3584, 2016.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points — online stochastic gradient for tensor decomposition. In *Proceedings of the 28th Annual Conference on Learning Theory*, pages 797–842, 2015.

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.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 421–436, 2011.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15:2489–2512, 2014.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18(223):1–42, 2018.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26*, pages 315–323, 2013.

Sujin Kim, Raghu Pasupathy, and Shane G. Henderson. *A Guide to Sample Average Approximation*, pages 207–243. 2015.

Tomer Koren and Kfir Levy. Fast rates for exp-concave empirical risk minimization. In *Advances in Neural Information Processing Systems 28*, pages 1477–1485, 2015.

Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, second edition, 2003.

Gábor Lugosi. Concentration-of-measure inequalities. Technical report, Department of Economics, Pompeu Fabra University, 2009.

Mehrdad Mahdavi and Rong Jin. Passive learning with target risk. In *Proceedings of the 26th Annual Conference on Learning Theory*, pages 252–269, 2013.

Mehrdad Mahdavi, Lijun Zhang, and Rong Jin. Lower and upper bounds on the generalization of stochastic exponentially concave optimization. In *Proceedings of the 28th Annual Conference on Learning Theory*, 2015.

Nishant A. Mehta. Fast rates with high probability in exp-concave statistical learning. *ArXiv e-prints*, arXiv:1605.01288, 2016.

Eric Moulines and Francis R. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems 24*, pages 451–459, 2011.

I. Necoara, Yu. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1):69–107, 2019.

Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in Neural Information Processing Systems 27*, pages 1017–1025, 2014.

A. Nemirovski and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons Ltd, 1983.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Yurii Nesterov. Random gradient-free minimization of convex functions. Core discussion papers, 2011.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 449–456, 2012.

Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczós, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

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 25*, pages 2672–2680, 2012.

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *ArXiv e-prints*, arXiv:1308.6370, 2013.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14:567–599, 2013.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning*, pages 71–79, 2013.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, second edition, 2014.

genceM.V.Lijuorze07567(StrinkI)-Meh31Sdrze075Mahd(eragiirinkI25Andrze075RZhang.075Jin)]TJ/3(LLeaarrze075

Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient langevin dynamics. In *Proceedings of the 30th Annual Conference on Learning Theory*, pages 1980–2022, 2017b.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936, 2003.

## Appendix A. Proof of Lemma 1

We first introduce the self-bounding property of smooth functions (Srebro et al., 2010, Lemma 4.1).

**Lemma 3** *For an $H$-smooth and nonnegative function $f : \mathcal{W} \mapsto \mathbb{R}$,*

$$\|\nabla f(\mathbf{w})\| \leq \sqrt{4Hf(\mathbf{w})}, \ \forall \mathbf{w} \in \mathcal{W}.$$

Assumptions 1 and 2 imply $f_t(\cdot)$ is nonnegative and $L$-smooth. From Lemma 3, we have

$$\|\nabla f_i(\mathbf{w})\|^2 \leq 4Lf_i(\mathbf{w}), \ \forall \mathbf{w} \in \mathcal{W}. \tag{14}$$

Let $\mathbf{w}'_{t+1} = \mathbf{w}_t - \eta \nabla f_t(\mathbf{w}_t)$. Following the analysis of online gradient descent (Zinkevich, 2003), for any $\mathbf{w} \in \mathcal{W}$, we have

$$\begin{aligned}
& F(\mathbf{w}_t) - F(\mathbf{w}) \\
\leq & \langle \nabla F(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle \\
= & \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle + \langle \nabla F(\mathbf{w}_t) - \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle \\
= & \frac{1}{2\eta}\left(\|\mathbf{w}_t - \mathbf{w}\|^2 - \|\mathbf{w}'_{t+1} - \mathbf{w}\|^2\right) + \frac{\eta}{2}\|\nabla f_t(\mathbf{w}_t)\|^2 + \langle \nabla F(\mathbf{w}_t) - \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle
\end{aligned}$$

$$\frac{\eta}{2}\|\nabla f_t(\mathbf{w}_t)\|^2 + \langle \nabla F(\mathbf{w}_t) - \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle$$

Rearranging the above inequality, we obtain

$$\mathrm{E}\left[\sum_{t=1}^{T}\left(F(\mathbf{w}_t) - F(\mathbf{w})\right)\right] \leq \frac{1}{2\eta(1-2\eta L)}\mathrm{E}\left[\|\mathbf{w}_1 - \mathbf{w}\|^2\right] + \frac{2\eta LT}{(1-2\eta L)}F(\mathbf{w}).$$

Dividing both sides by $T$, we have

$$\begin{aligned}
&\frac{1}{2\eta T(1-2\eta L)}\mathrm{E}\left[\|\mathbf{w}_1 - \mathbf{w}\|^2\right] + \frac{2\eta L}{(1-2\eta L)}F(\mathbf{w}) \\
\geq &\frac{1}{T}\mathrm{E}\left[\sum_{t=1}^{T}\left(F(\mathbf{w}_t) - F(\mathbf{w})\right)\right] \geq \mathrm{E}\left[F(\bar{\mathbf{w}})\right] - F(\mathbf{w})
\end{aligned}$$

where the last step is due to Jensen's inequality.

## Appendix B. Proof of Lemma 2

Recall that the following parameters are used in the second call of Epoch-GD

$$\eta_1 = \frac{1}{4L}, \; T_1 = 2^{\alpha+3}\kappa, \; T_{k+1} = 2T_k, \; \eta_{k+1} = \frac{\eta_k}{2}, \; k \geq 1.$$

Then, we have

$$\eta_k L \leq \eta_1 L = \frac{1}{4}, \tag{15}$$

$$\eta_k T_k = 2^{\alpha+1}. \tag{16}$$

We prove this lemma by induction on $k$. When $k = 1$, from Lemma 1, we have

$$\begin{aligned}
&\mathrm{E}\left[F(\mathbf{w}_1^2)\right] - F(\mathbf{w}_*) \\
\leq &\frac{1}{2\eta_1 T_1(1-2\eta_1 L)}\mathrm{E}\left[\|\mathbf{w}_1^1 - \mathbf{w}_*\|^2\right] + \frac{2\eta_1 L}{(1-2\eta_1 L)}F(\mathbf{w}_*) \\
\overset{(15)}{=} &\frac{1}{\eta_1 T_1}\mathrm{E}\left[\|\mathbf{w}_1^1 - \mathbf{w}_*\|^2\right] + 4\eta_1 L F(\mathbf{w}_*) \\
\overset{(16)}{=} &\frac{1}{2^{\alpha+1}}\mathrm{E}\left[\|\mathbf{w}_1^1 - \mathbf{w}_*\|^2\right] + \frac{2^{\alpha+3}F(\mathbf{w}_*)}{T_1} \\
\overset{(5)}{\leq} &\frac{1}{2^{\alpha+1}}\frac{2}{\lambda}\mathrm{E}\left[F(\mathbf{w}_1^1) - F(\mathbf{w}_*)\right] + \frac{2^{\alpha+3}F(\mathbf{w}_*)}{T_1} \\
\overset{(12)}{\leq} &\frac{1}{2^{\alpha}}\left(\frac{32G^2}{\kappa^{\alpha}}\right) + \frac{2^{\alpha+3}F(\mathbf{w}_*)}{T_1} \\
\overset{(T_1=2^{\alpha+3}\kappa)}{=} &\frac{2^{\alpha^2+2\alpha+5}G^2}{(T_1)^{\alpha}} + \frac{2^{\alpha+3}F(\mathbf{w}_*)}{T_1}.
\end{aligned}$$

Assume that (13) is true for some $k \quad 1$, and we prove the inequality for $k+1$. According to Lemma 1, we have

$$
\begin{aligned}
&\mathrm{E}\left[F(\mathbf{w}_1^{k+2})\right] \quad F(\mathbf{w}_*) \\
&\frac{1}{2 \quad {}_{k+1}T_{k+1}(1 \quad 2 \quad {}_{k+1}L)}\mathrm{E}\left[k\mathbf{w}_1^{k+1} \quad \mathbf{w}_*k^2\right] + \frac{2 \quad {}_{k+1}L}{(1 \quad 2 \quad {}_{k+1}L)}F(\mathbf{w}_*) \\
\overset{(15)}{}\ &\frac{1}{{}_{k+1}T_{k+1}}\mathrm{E}\left[k\mathbf{w}_1^{k+1} \quad \mathbf{w}_*k^2\right] + 4 \quad {}_{k+1}LF(\mathbf{w}_*) \\
\overset{(16)}{=}&\frac{}{2^{\alpha+1}}\mathrm{E}\left[k\mathbf{w}_1^{k+1} \quad \mathbf{w}_*k^2\right] + \frac{2^{\alpha+3} \quad F(\mathbf{w}_*)}{T_{k+1}} \\
\overset{(5)}{}\ &\frac{}{2^{\alpha+1}}\frac{2}{}\mathrm{E}\left[F(\mathbf{w}_1^{k+1}) \quad F(\mathbf{w}_*)\right] + \frac{2^{\alpha+3} \quad F(\mathbf{w}_*)}{T_{k+1}} \\
\overset{(13)}{}\ &\frac{1}{2^\alpha}\left(\frac{2^{\alpha^2+2\alpha+5}G^2}{(T_k)^\alpha} + \frac{2^{\alpha+3} \quad F(\mathbf{w}_*)}{T_k}\left(\sum_{i=1}^{k}\frac{1}{2^{(i-1)(\alpha-1)}}\right)\right) + \frac{2^{\alpha+3} \quad F(\mathbf{w}_*)}{T_{k+1}} \\
=&\frac{2^{\alpha^2+2\alpha+5}G^2}{(T_{k+1})^\alpha} + \frac{2^{\alpha+3} \quad F(\mathbf{w}_*)}{T_{k+1}}\left(\sum_{i=1}^{k+1}\frac{1}{2^{(i-1)(\alpha-1)}}\right):
\end{aligned}
$$

## Appendix C. Proof of Theorem 4

We first establish the following lemma for bounding the excess risk of the intermediate iterate.

**Lemma 4** *For any $k$, we have*

$$
\mathrm{E}[F(\mathbf{w}_1^{k+1})] \quad F(\mathbf{w}_*) \quad \frac{F(\mathbf{w}_1^1) \quad F(\mathbf{w}_*)}{2^k} + \frac{F(\mathbf{w}_*)}{}\left(\sum_{i=1}^{k}\frac{1}{2^{i-1}}\right): \tag{17}
$$

The number of epochs made is given by $k^\dagger = bT{=}T'c$ and the final solution is $\widetilde{\mathbf{w}} = \mathbf{w}_1^{k^\dagger+1}$. From Lemma 4, we have

$$
\begin{aligned}
&F(\mathbf{w}_1^{k^\dagger+1}) \quad F(\mathbf{w}_*) \\
&\frac{F(\mathbf{w}_1^1) \quad F(\mathbf{w}_*)}{2^{k^\dagger}} + \frac{F(\mathbf{w}_*)}{}\left(\sum_{i=1}^{k^\dagger}\frac{1}{2^{i-1}}\right) \\
&\frac{F(\mathbf{w}_1^1) \quad F(\mathbf{w}_*)}{2^{k^\dagger}} + \frac{2F(\mathbf{w}_*)}{}:
\end{aligned}
$$

## Appendix D. Proof of Lemma 4

From (8), we know that

$$
L = \frac{1}{4} \quad \frac{1}{4}; \tag{18}
$$

$$
T' = 4: \tag{19}
$$

We prove this lemma by induction on $k$. When $k = 1$, from Lemma 1, we have

$$
\begin{aligned}
&\mathrm{E}\left[F(\mathbf{w}_1^2)\right] \quad F(\mathbf{w}_*) \\
&\frac{1}{2\ \ T'(1 \quad 2\ \ L)} k \mathbf{w}_1^1 \quad \mathbf{w}_* k^2 + \frac{2\ \ L}{(1 \quad 2\ \ L)} F(\mathbf{w}_*) \\
\overset{(18)}{}\ &\frac{1}{T'} k \mathbf{w}_1^1 \quad \mathbf{w}_* k^2 + \frac{F(\mathbf{w}_*)}{} \\
\overset{(19)}{=} &\frac{}{4} k \mathbf{w}_1^1 \quad \mathbf{w}_* k^2 + \frac{F(\mathbf{w}_*)}{} \overset{(5)}{} \frac{f(\mathbf{w}_1^1) \quad f(\mathbf{w}_*)}{2} + \frac{F(\mathbf{w}_*)}{} :
\end{aligned}
$$

Assume that (17) is true for some $k \quad 1$, and we prove the inequality for $k + 1$. According to Lemma 1, we have

$$
\begin{aligned}
&\mathrm{E}\left[F(\mathbf{w}_1^{k+2})\right] \quad F(\mathbf{w}_*) \\
&\frac{1}{2\ \ T'(1 \quad 2\ \ L)} \mathrm{E}\left[k \mathbf{w}_1^{k+1} \quad \mathbf{w}_* k^2\right] + \frac{2\ \ L}{(1 \quad 2\ \ L)} F(\mathbf{w}_*) \\
\overset{(18)}{}\ &\frac{1}{T'} \mathrm{E}\left[k \mathbf{w}_1^{k+1} \quad \mathbf{w}_* k^2\right] + \frac{F(\mathbf{w}_*)}{} \\
\overset{(19)}{=} &\frac{}{4} \mathrm{E}\left[k \mathbf{w}_1^{k+1} \quad \mathbf{w}_* k^2\right] + \frac{F(\mathbf{w}_*)}{} \\
\overset{(5)}{}\ &\frac{}{4} \frac{2}{} \mathrm{E}\left[F(\mathbf{w}_1^{k+1}) \quad F(\mathbf{w}_*)\right] + \frac{F(\mathbf{w}_*)}{} \\
\overset{(17)}{}\ &\frac{1}{2}\left(\frac{F(\mathbf{w}_1^1) \quad F(\mathbf{w}_*)}{2^k} + \frac{F(\mathbf{w}_*)}{}\left(\sum_{i=1}^{k} \frac{1}{2^{i-1}}\right)\right) + \frac{F(\mathbf{w}_*)}{} \\
= &\frac{F(\mathbf{w}_1^1) \quad F(\mathbf{w}_*)}{2^{k+1}} + \frac{F(\mathbf{w}_*)}{}\left(\sum_{i=1}^{k+1} \frac{1}{2^{i-1}}\right) :
\end{aligned}
$$

## Appendix E. Comparison with Needell et al. (2014)

First, we provide the following basic inequality that allows us to bound the excess risk by the distance. From Assumption 2, we have

$$
F(\mathbf{w}_t) \quad F(\mathbf{w}_*) \quad h\, r\, F(\mathbf{w}_*); \mathbf{w}_t \quad \mathbf{w}_* i + \frac{L}{2} k \mathbf{w}_t \quad \mathbf{w}_* k^2 : \tag{20}
$$

Using notations of our paper, Theorem 2.1 of Needell et al. (2014) establishes the following convergence rate for unconstrained problems:

$$
\mathrm{E}\left[k \mathbf{w}_t \quad \mathbf{w}_* k^2\right] \quad [1 \quad 2 \quad (1 \quad L)]^T k \mathbf{w}_0 \quad \mathbf{w}_* k^2 + \frac{4\ \ L F_*}{(1 \quad L)} \tag{21}
$$

where $\mathbf{w}_t$ is the SGD iterate in the $t$-th round and $< 1=$ is the step size. Note that $r F(\mathbf{w}_*) = 0$ in the unconstrained case. Combining (20) and (21