How To Make the Gradients Small Stochastically: Even Faster Convex and Nonconvex SGD*

Zeyuan Allen-Zhu
Microsoft Research AI
Redmond, WA 98052
zeyuan@csail.mit.edu

Abstract

Stochastic gradient descent (SGD) gives an optimal convergence rate when minimizing convex stochastic objectives $f(x)$. However, in terms of making the gradients small, the original SGD does not give an optimal rate, even when $f(x)$ is convex.

If $f(x)$ is convex, to find a point with gradient norm $\varepsilon$, we design an algorithm SGD3 with a near-optimal rate $\tilde{O}(\varepsilon^{-2})$, improving the best known rate $O(\varepsilon^{-8/3})$ of [17]. If $f(x)$ is nonconvex, to find its $\varepsilon$-approximate local minimum, we design an algorithm SGD5 with rate $\tilde{O}(\varepsilon^{-3.5})$, where previously SGD variants only achieve $O(\varepsilon^{-4})$ [6, 14, 30]. This is no slower than the best known stochastic version of Newton’s method in all parameter regimes [27].

1 Introduction

In convex optimization and machine learning, the classical goal is to design algorithms to decrease objective values, that is, to find points $x$ with $f(x) - f(x^*) \leq \varepsilon$. In contrast, the rate of convergence for the gradients, that is,

the number of iterations $T$ needed to find a point $x$ with $\|\nabla f(x)\| \leq \varepsilon$,

is a harder problem and sometimes needs new algorithmic ideas [25]. For instance, in the full-gradient setting, accelerated gradient descent alone is suboptimal for this new goal, and one needs additional tricks to get the fastest rate [25]. We review these tricks in Section 1.1.

In the convex (online) stochastic optimization, to the best of our knowledge, tight bounds are not yet known for finding points with small gradients. The best recorded rate was $T \propto \varepsilon^{-8/3}$ [17], and it was raised as an open question [1] regarding how to improve it.

In this paper, we design two new algorithms, SGD2 which gives rate $T \propto \varepsilon^{-5/2}$ using Nesterov’s tricks, and SGD3 which gives an even better rate $T \propto \varepsilon^{-2} \log^3 \frac{1}{\varepsilon}$ which is optimal up to log factors.

Motivation. Studying the rate of convergence for the minimizing gradients can be important at least for the following two reasons.

- In many situations, points with small gradients fit better our final goals.

---

*The full version of this paper can be found on https://arxiv.org/abs/1801.02982 When this paper was submitted to NIPS 2018, the “non-convex SGD” results were not included. We encourage the readers to go to our full version to find out these “non-convex SGD” results.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.
Nesterov [25] considers the dual approach for solving constrained minimization problems. He argued that “the gradient value \( \| \nabla f(x) \| \) serves as the measure of feasibility and optimality of the primal solution,” and thus is the better goal for minimization purpose.\(^2\)

In matrix scaling [7][10], given a non-negative matrix, one wants to re-scale its rows and columns to make it doubly stochastic. This problem has been applied in image reconstruction, operations research, decision and control, and other scientific disciplines (see survey [20]). The goal for matrix scaling is to find points with small gradients, but not small objectives.

- Designing algorithms to find points with small gradients can help us understand non-convex optimization better and design faster non-convex machine learning algorithms.

Without strong assumptions, non-convex optimization theory is always in terms of finding points with small gradients (i.e., approximate stationary points or local minima). Therefore, to understand non-convex stochastic optimization better, perhaps we should first figure out the best rate for convex stochastic optimization. In addition, if new algorithmic ideas are needed, can we also apply them to the non-convex world? We find positive answers to this question, and also obtain better rates for standard non-convex optimization tasks.

### 1.1 Review: Prior Work on Deterministic Convex Optimization

Suppose \( f(x) \) is a Lipschitz smooth convex function with smoothness parameter \( L \). Then, it is well-known that accelerated gradient descent (AGD) [23][24] finds a point \( x \) satisfying \( f(x) - f(x^*) \leq \delta \) using \( T = O(\sqrt{\frac{L}{\delta^2}}) \) gradient computations of \( \nabla f(x) \). To turn this into a gradient guarantee, we can apply the smoothness property of \( f(x) \) which gives \( \| \nabla f(x) \|^2 \leq L(f(x) - f(x^*)) \). This means

\[
\text{AGD converges in rate } T \propto \frac{1}{\delta}.
\]

Nesterov [25] proposed two different tricks to improve upon such rate.

**Nesterov’s First Trick: GD After AGD.** Recall that starting from a point \( x_0 \), if we perform \( T \) steps of gradient descent (GD) \( x_{t+1} = x_t - \frac{1}{L} \nabla f(x_t) \), then it satisfies \( \sum_{t=0}^{T-1} \| \nabla f(x_t) \|^2 \leq L(f(x_0) - f(x^*)) \). In addition, if this \( x_0 \) is already the output of AGD for another \( T \) iterations, then it satisfies \( f(x_0) - f(x^*) \leq O(\frac{L}{T^2}) \). Putting the two inequalities together, we have \( \min_{t=0}^{T-1} \{ \| \nabla f(x_t) \|^2 \} \leq O(L^2) \). We call this method “GD after AGD,” and

\[
\text{“GD after AGD” converges in rate } T \propto \frac{L^2}{\sqrt{\delta^3}}.
\]

**Nesterov’s Second Trick: AGD After Regularization.** Alternatively, we can also regularize \( f(x) \) by defining \( g(x) = f(x) + \frac{\sigma}{2} \| x - x_0 \|^2 \). This new function \( g(x) \) is \( \sigma \)-strongly convex, so AGD converges \textit{linearly}, meaning that using \( T \approx \frac{L}{\sqrt{\sigma}} \log \frac{1}{\delta} \) gradients we can find a point \( x \) satisfying \( \| \nabla g(x) \|^2 \leq L(g(x) - g(x^*)) \leq \epsilon^2 \). If we choose \( \sigma \approx \epsilon \), then this implies \( \| \nabla f(x) \| \leq \| \nabla g(x) \| + \epsilon \leq 2\epsilon \). We call this method “AGD after regularization,” and

\[
\text{“AGD after regularization” converges in rate } T \propto \frac{L^{1/2}}{\epsilon^{1/2}} \log \frac{L}{\delta}.
\]

This is optimal up to a log factor, because first-order methods need \( T = \Omega(\sqrt{L/\delta}) \) gradient computations to find \( f(x) - f(x^*) \leq \delta [23] \), but \( f(x) - f(x^*) \leq \| \nabla f(x) \| \cdot \| x - x^* \| \leq O(\| \nabla f(x) \|) \).

### 1.2 Our Results: Stochastic Convex Optimization

Consider the stochastic setting where the convex objective \( f(x) := \mathbb{E}_i[f_i(x)] \) and the algorithm can only compute stochastic gradients \( \nabla f_i(x) \) at any point \( x \) for a random \( i \). Let \( T \) be the number of stochastic gradient computations. It is well-known that stochastic gradient descent (SGD) finds a point \( x \) with \( f(x) - f(x^*) \leq \delta \) in (see for instance textbooks [8][18][26])

\[
T = O(\frac{L}{\delta^2}) \text{ iterations} \quad \text{or} \quad T = O(\frac{L}{\sigma \delta}) \text{ if } f(x) \text{ is } \sigma \text{-strongly convex}.
\]

Both rates are asymptotically optimal in terms of decreasing objective, and \( \mathcal{V} \) is an absolute bound on the variance of the stochastic gradients. Using the same argument \( \| \nabla f(x) \|^2 \leq L(f(x) - f(x^*)) \)

\[^2\text{Nesterov [25] studied } \min_{y \in Q} \{ g(y) : Ay = b \} \text{ with convex } Q \text{ and strongly convex } g(y). \text{ The dual problem is } \min_x \{ f(x) \} \text{ where } f(x) := \min_{y \in Q} \{ g(y) + \langle x, b - Ay \rangle \}. \text{ Let } y^*(x) \in Q \text{ be the (unique) minimizer of the internal problem, then } g(y^*(x)) - f(x) = \langle x, \nabla f(x) \rangle \leq \| x \| \cdot \| \nabla f(x) \|.
\]
Table 1: Comparison of first-order online stochastic methods for finding $\|\nabla f(x)\| \leq \varepsilon$. Following tradition, in these bounds, we hide variance and smoothness parameters in big-$O$ and only show the dependency on $\varepsilon$, the condition number $\kappa = \frac{L}{\sigma} \geq 1$ (if the objective is $\sigma$-strongly convex), or the nonconvexity parameter $\sigma$.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Gradient complexity $T$</th>
<th>2nd-order smooth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online convex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SGD (naive)</td>
<td>$O\left(\frac{1}{\varepsilon^4}\right)$</td>
<td></td>
</tr>
<tr>
<td>SGD1 (SGD after SGD)</td>
<td>$O\left(\frac{1}{\varepsilon^{8/3}}\right)$ (folklore, see Theorem 4.2)</td>
<td></td>
</tr>
<tr>
<td>SGD2 (SGD after regularization)</td>
<td>$O\left(\frac{1}{\varepsilon^{5/2}}\right)$ (see Theorem 2)</td>
<td></td>
</tr>
<tr>
<td>SGD3 (SGD + recursive regularization)</td>
<td>$O\left(\frac{1}{\varepsilon^{2} \cdot \log^{3/2} \frac{1}{\varepsilon}}\right)$ (see Theorem 3)</td>
<td></td>
</tr>
<tr>
<td>Online strongly convex</td>
<td></td>
<td>no</td>
</tr>
<tr>
<td>SGD$^{\text{IC}}$ (naive)</td>
<td>$O\left(\frac{1}{\varepsilon^{2}}\right)$ (see Theorem 4.2)</td>
<td></td>
</tr>
<tr>
<td>SGD1$^{\text{IC}}$ (SGD after SGD)</td>
<td>$O\left(\frac{1}{\varepsilon^{2} \cdot \kappa^{1/2}}\right)$ (see Theorem 1)</td>
<td></td>
</tr>
<tr>
<td>SGD3$^{\text{IC}}$ (SGD + recursive regularization)</td>
<td>$O\left(\frac{1}{\varepsilon^{2} \cdot \log^{3/2} \kappa}\right)$ (see Theorem 3)</td>
<td></td>
</tr>
<tr>
<td>Online nonconvex ($\sigma$-nonconvex)</td>
<td></td>
<td>needed</td>
</tr>
<tr>
<td>SGD (naive)</td>
<td>$O\left(\frac{1}{\varepsilon^4}\right)$</td>
<td></td>
</tr>
<tr>
<td>SCG</td>
<td>$O\left(\frac{1}{\varepsilon^{10/3}}\right)$ (see [16])</td>
<td></td>
</tr>
<tr>
<td>SGD4</td>
<td>$O\left(\frac{1}{\varepsilon^{2}} + \frac{\sigma}{\varepsilon^{4}}\right)$ (see Theorem 4)</td>
<td></td>
</tr>
<tr>
<td>Natasha1.5</td>
<td>$O\left(\frac{1}{\varepsilon^{3}} + \frac{\sigma^{1/4}}{\varepsilon^{10/3}}\right)$ (see [3])</td>
<td></td>
</tr>
<tr>
<td>SGD variants</td>
<td>$O\left(\frac{1}{\varepsilon^{4}}\right)$ (see [8][14][30])</td>
<td></td>
</tr>
<tr>
<td>SGD5</td>
<td>$O\left(\frac{1}{\varepsilon^{3.5}}\right)$ (see Theorem 5)</td>
<td></td>
</tr>
<tr>
<td>cubic Newton</td>
<td>$O\left(\frac{1}{\varepsilon^{3.5}}\right)$ (see [27])</td>
<td></td>
</tr>
<tr>
<td>Natasha2</td>
<td>$O\left(\frac{1}{\varepsilon^{3.25}}\right)$ (see [3])</td>
<td></td>
</tr>
</tbody>
</table>

As before, SGD finds a point $x$ with $\|\nabla f(x)\| \leq \varepsilon$ in

$$T = O\left(\frac{L^2 V}{\varepsilon^4}\right) \text{ iterations \quad or \quad } T = O\left(\frac{LV}{\sigma^2 \varepsilon^2}\right) \text{ if } f(x) \text{ is } \sigma \text{-strongly convex.} \quad \text{(SGD)}$$

These rates are not optimal. We investigate three approaches to improve such rates.

**New Approach 1: SGD after SGD.** Recall in Nesterov’s first trick, he replaced the use of the inequality $\|\nabla f(x)\|^2 \leq L(f(x) - f(x^*))$ by $T$ steps of gradient descent. In the stochastic setting, can we replace this inequality with $T$ steps of SGD? We call this algorithm SGD1 and prove that

**Theorem 1 (informal).** For convex stochastic optimization, SGD1 finds $x$ with $\|\nabla f(x)\| \leq \varepsilon$ in

$$T = O\left(\frac{L^{2/3} V}{\varepsilon^{8/3}}\right) \text{ iterations \quad or \quad } T = O\left(\frac{L^{1/2} V}{\sigma^{1/4} \varepsilon^2}\right) \text{ if } f(x) \text{ is } \sigma \text{-strongly convex.} \quad \text{(SGD1)}$$

The rate $T \propto \varepsilon^{-8/3}$, in the special case of unconstrained minimization, was first discovered by Ghadimi and Lan [17] using a more complicated algorithm. The rate $T \propto \frac{1}{\sigma^{1/2} \varepsilon^2}$ does not seem to be known before.

**New Approach 2: SGD after regularization.** Recall that in Nesterov’s second trick, he defined $g(x) = f(x) + \frac{\sigma}{2}\|x - x_0\|^2$ as a regularized version of $f(x)$, and applied the strongly-convex version of AGD to minimize $g(x)$. Can we apply this trick to the stochastic setting?

Note the parameter $\sigma$ has to be on the magnitude of $\varepsilon$ because $\nabla g(x) = \nabla f(x) + \sigma(x - x_0)$ and we wish to make sure $\|\nabla f(x)\| = \|\nabla g(x)\| \pm \varepsilon$. Therefore, if we apply SGD1 to minimize $g(x)$ to find a point $\|\nabla g(x)\| \leq \varepsilon$, the convergence rate is $T \propto \frac{1}{\sigma^{1/2} \varepsilon^2} = \frac{1}{\varepsilon^2}$. We call this algorithm SGD2.

**Theorem 2 (informal).** For convex stochastic optimization, SGD2 finds $x$ with $\|\nabla f(x)\| \leq \varepsilon$ in

$$T = O\left(\frac{L^{1/2} V}{\varepsilon^{3/2}}\right) \text{ iterations.} \quad \text{(SGD2)}$$

Again, this $T \propto \frac{1}{\varepsilon^{3/2}}$ rate does not seem to be known before.

**New Approach 3: SGD and recursive regularization.** In the second approach above, the $\varepsilon^{0.5}$ sub-optimality gap is due to the choice of $\sigma \propto \varepsilon$ which ensures $\|\sigma(x - x_0)\| \leq \varepsilon$. 


Intuitively, if \( x_0 \) were sufficiently close to \( x^* \) (and thus were also close to the approximate minimizer \( x \)), then we could choose \( \sigma \gg \varepsilon \) so that \( \|\sigma (x-x_0)\| \leq \varepsilon \) still holds. In other words, an appropriate warm start \( x_0 \) could help us break the \( \varepsilon^{-2.5} \) barrier and get a better convergence rate. However, how to find such \( x_0 \)? We find it by constructing a “less warm” starting point and so on. This process is summarized by the following algorithm which recursively finds the warm starts.

Starting from \( f^{(0)}(x) := f(x) \), we define \( f^{(s)}(x) := f^{(s-1)}(x) + \frac{\sigma_s}{2} \|x - \hat{x}_s\|^2 \) where \( \sigma_s = 2\sigma_{s-1} \) and \( \hat{x}_s \) is an approximate minimizer of \( f^{(s-1)}(x) \) that is simply calculated from the naive SGD. We call this method SGD3, and prove that

**Theorem 3** (informal). For convex stochastic optimization, SGD3 finds \( x \) with \( \|\nabla f(x)\| \leq \varepsilon \) in

\[
T = O\left(\frac{\log^3(L/\varepsilon)}{\varepsilon^2}\right) \quad \text{iterations} \quad \text{or} \quad T = O\left(\frac{\log^3(L/\varepsilon)}{\varepsilon^2}\right) \quad \text{if} \quad f(x) \text{ is } \sigma\text{-strongly convex.} \quad \text{(SGD3)}
\]

Our new rates in Theorem 3 not only improve the best known result of \( T \approx \varepsilon^{-8/3} \), but also are near optimal because \( \Omega((V/\varepsilon^2)) \) is clearly a lower bound: even to decide whether a point \( x \) has \( \|\nabla f(x)\| \leq \varepsilon \) or \( \|\nabla f(x)\| > 2\varepsilon \) requires \( \Omega(V/\varepsilon^2) \) samples of the stochastic gradient.Perhaps interestingly, our dependence on the smoothness parameter \( L \) (or the condition number \( \kappa := L/\sigma \) if strongly convex) is only polylogarithmic, as opposed to polynomial in all previous results.

### 1.3 Roadmap

We introduce notions in Section 2 and formalize the convex problem in Section 3. We review classical (convex) SGD theorems with objective decrease in Section 4. We give an auxiliary lemma in Section 5 that shows our SGD3 results in Section 6. We apply our techniques to non-convex optimization and give algorithms SGD4 and SGD5 in Section 7. We discuss more related work in Appendix A, and show our results on SGD1 and SGD2 respectively in Appendix B and Appendix C.

## 2 Preliminaries

Throughout this paper, we denote by \( \| \cdot \| \) the Euclidean norm. We use \( i \in \mathbb{R}[n] \) to denote that \( i \) is generated from \( |n| = \{1, 2, \ldots, n\} \) uniformly at random. We denote by \( \nabla f(x) \) the gradient of function \( f \) if it is differentiable, and \( \partial f(x) \) any subgradient if \( f \) is only Lipschitz continuous. We denote by \( \mathbb{I}[\text{event}] \) the indicator function of probabilistic events. We denote by \( \|A\|_2 \) the spectral norm of matrix \( A \). For symmetric matrices \( A \) and \( B \), we write \( A \succeq B \) to indicate that \( A - B \) is positive semidefinite (PSD). Therefore, \( A \succeq -\sigma I \) if and only if all eigenvalues of \( A \) are no less than \(-\sigma \). We denote by \( \lambda_{\min}(A) \) and \( \lambda_{\max}(A) \) the minimum and maximum eigenvalue of a symmetric matrix \( A \).

Recall some definitions on strong convexity and smoothness (and they have other equivalent definitions, see textbook [23]).

**Definition 2.1.** For a function \( f : \mathbb{R}^d \to \mathbb{R} \),

- \( f \) is \( \sigma\)-strongly convex if \( \forall x, y \in \mathbb{R}^d \), it satisfies \( f(y) \geq f(x) + \langle \partial f(x), y-x \rangle - \frac{\sigma}{2} \|x-y\|^2 \).
- \( f \) is of \( \sigma \)-bounded nonconvexity (or \( \sigma\)-nonconvex for short) if \( \forall x, y \in \mathbb{R}^d \), it satisfies \( f(y) \geq f(x) + \langle \partial f(x), y-x \rangle - \frac{\sigma}{2} \|x-y\|^2 \).
- \( f \) is \( L\)-Lipschitz smooth (or \( L\)-smooth for short) if \( \forall x, y \in \mathbb{R}^d \), \( \|\nabla f(x) - \nabla f(y)\| \leq L \|x-y\| \).
- \( f \) is \( L_2\)-second-order smooth if \( \forall x, y \in \mathbb{R}^d \), it satisfies \( \|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L_2 \|x-y\| \).

**Definition 2.2.** For composite function \( F(x) = \psi(x) + f(x) \) where \( \psi(x) \) is proper convex, given a parameter \( \eta > 0 \), the gradient mapping of \( F(\cdot) \) at point \( x \) is

\[
\mathcal{G}_{F,\eta}(x) := \frac{1}{\eta} \left( x - x^+ \right) \quad \text{where} \quad x^+ = \arg \min_y \left\{ \psi(y) + \langle \nabla f(x), y \rangle + \frac{1}{2\eta} \|y-x\|^2 \right\}
\]

In particular, if \( \psi(\cdot) = 0 \), then \( \mathcal{G}_{F,\eta}(x) = \nabla f(x) \).

Recall the following property about gradient mapping —see for instance [29, Lemma 3.7].

---

\(^3\)Previous authors also refer to this notion as “approximate convex”, “almost convex”, “hypo-convex”, “semi-convex”, or “weakly-convex.” We call it \( \sigma\)-nonconvex to stress the point that \( \sigma \) can be as large as \( L \) (any \( L\)-smooth function is automatically \( L\)-nonconvex).
Lemma 2.3. Let $F(x) = \psi(x) + f(x)$ where $\psi(x)$ is proper convex and $f(x)$ is $\sigma$-strongly convex and $L$-smooth. For every $x, y \in \{x \in \mathbb{R}^d : \psi(x) < +\infty\}$, letting $x^+ = x - \eta \cdot G_{F,\eta}(x)$, we have
\[
\forall \eta \in \left(0, \frac{1}{L}\right]: \quad F(y) \geq F(x^+) + \langle G_{F,\eta}(x), y - x \rangle + \frac{\eta}{2} \|G_{F,\eta}(x)\|^2 + \frac{\sigma}{2} \|y - x\|^2.
\]

The following definition and properties of Fenchel dual for convex functions is classical, and can be found for instance in the textbook [26].

Definition 2.4. Given proper convex function $h(y)$, its Fenchel dual $h^*(\beta) := \max_y \{y^\top \beta - h(y)\}$.

Proposition 2.5. $\nabla h^*(\beta) = \arg \max_y \{y^\top \beta - h(y)\}$.

Proposition 2.6. If $h(\cdot)$ is $\sigma$-strongly convex, then $h^*(\cdot)$ is $\frac{1}{\sigma}$-smooth.

3 Problem Formalization

Throughout this paper (except our nonconvex application Section 7), we minimize convex stochastic composite objective:
\[
\min_{x \in \mathbb{R}^d} \left\{ F(x) = \psi(x) + f(x) := \psi(x) + \frac{1}{n} \sum_{i \in [n]} f_i(x) \right\},
\]
where
1. $\psi(x)$ is proper convex (a.k.a. the proximal term),
2. $f_i(x)$ is differentiable for every $i \in [n]$,
3. $f(x)$ is $L$-smooth and $\sigma$-strongly convex for some $\sigma \in [0, L]$ that could be zero,
4. $n$ can be very large of even infinite (so $f(x) = E_i[f_i(x)]$) and
5. the stochastic gradients $\nabla f_i(x)$ have a bounded variance (over the domain of $\psi(\cdot)$), that is
\[
\forall x \in \{y \in \mathbb{R}^d : \psi(y) < +\infty\}: \quad \mathbb{E}_{i \in \mathbb{R}[n]} \|\nabla f(x) - \nabla f_i(x)\|^2 \leq \mathcal{V}.
\]

We emphasize that the above assumptions are all classical.

In the rest of the paper, we define $T$, the gradient complexity, as the number of computations of $\nabla f_i(x)$. We search for points $x$ so that the gradient mapping $\|G_{F,\eta}(x)\| \leq \varepsilon$ for any $\eta \approx \frac{1}{L}$. Recall from Definition 2.2 that if there is no proximal term (i.e., $\psi(x) \equiv 0$), then $G_{F,\eta}(x) = \nabla f(x)$ for any $\eta > 0$. We want to study the best tradeoff between the gradient complexity $T$ and the error $\varepsilon$.

We say an algorithm is online if its gradient complexity $T$ is independent of $n$. This tackles the big-data scenario when $n$ is extremely large or even infinite (i.e., $f(x) = E_i[f_i(x)]$ for some random variable $i$). The stochastic gradient descent (SGD) method and all of its variants studied in this paper are online. In contrast, GD, AGD [23, 24], and Katyusha [2] are offline methods because their gradient complexity depends on $n$ (see Table 2 in appendix).

4 Review: SGD with Objective Value Convergence

Recall that stochastic gradient descent (SGD) repeatedly performs proximal updates of the form
\[
x_{t+1} = \arg \min_{y \in \mathbb{R}^d} \left\{ \psi(y) + \frac{1}{2\alpha} \|y - x_t\|^2 + \langle \nabla f_i(x_t), y \rangle \right\},
\]
where $\alpha > 0$ is some learning rate, and $i$ is chosen in $1, 2, \ldots, n$ uniformly at random per iteration.

Note that if $\psi(y) = 0$ then $x_{t+1} = x_t - \alpha \nabla f_i(x_t)$. For completeness’ sake, we summarize it in Algorithm 1. If $f(x)$ is also known to be strongly convex, to get the tightest convergence rate, one can repeatedly apply SGD with decreasing learning rate $\alpha$ [19]. We summarize this algorithm as SGD$^\infty$ in Algorithm 2.

The following theorem describes the rates of convergence in objective values for SGD and SGD$^\infty$ respectively. Their proofs are classical (and included in Appendix D); however, for our exact statements, we cannot find them recorded anywhere.

\footnote{All of the results in this paper apply to the case when $n$ is infinite, because we focus on online methods. However, we still introduce $n$ to simplify notations.}

\footnote{In the special case $\psi(x) \equiv 0$, Theorem 4.1(a) and 4.1(b) are folklore (see for instance [26]). If $\psi(x) \neq 0$, Theorem 4.1(a) is recorded when $\psi(x)$ is Lipschitz or smooth [13], but we would not like to impose such assumptions. A variant of Theorem 4.1(b) is recorded for the accelerated version of SGD [15], but with a
Theorem 4.1. Let \( x^* \in \arg \min_x \{ F(x) \} \). To solve Problem (3.1) given a starting vector \( x_0 \in \mathbb{R}^d \),
(a) \( \text{SGD}(F, x_0, 0, T) \) outputs \( \overline{x} \) satisfying
\[
\mathbb{E}[F(\overline{x})] - F(x^*) \leq \frac{\alpha \sqrt{\mathbb{V}}}{\frac{T}{1-\alpha \mathbb{L}}} + \frac{\sqrt{\mathbb{V}}}{\frac{T}{1-\alpha \mathbb{L}}} \| x_0 - x^* \|^2.
\]
(b) If \( f(x) \) is \( \sigma \)-strongly convex and \( T \geq \frac{L}{2} \), then \( \text{SGD}(F, x_0, 0, T) \) outputs \( \overline{x} \) satisfying
\[
\mathbb{E}[F(\overline{x})] - F(x^*) \leq O\left( \frac{\sqrt{\mathbb{V}}}{\frac{T}{1-\alpha \mathbb{L}}} + \left(1 - \frac{\sigma}{T} \right)^{\Omega(T)} \sigma \| x_0 - x^* \|^2 \right).
\]
As a sanity check, if \( \mathbb{V} = 0 \), the convergence rate of SGD matches that of GD. (However, if \( \mathbb{V} = 0 \), one can apply accelerated gradient descent of Nesterov [22, 23] instead for a faster rate.)
To turn Theorem 4.1 into a rate of convergence for the gradients, we can simply apply Lemma 2.3 which implies
\[
\forall \eta \in (0, \frac{1}{L}] : \quad \frac{\eta}{2} \| \nabla F_{\eta}(\overline{x}) \|^2 \leq F(\overline{x}) - F(x^+) \leq F(\overline{x}) - F(x^*).
\]

Theorem 4.2. Let \( x^* \in \arg \min_x \{ F(x) \} \). To solve Problem (3.1) given a starting vector \( x_0 \in \mathbb{R}^d \) and any \( \eta = \frac{L}{T} \) where \( C \in (0, 1] \) is some absolute constant,
(a) \( \text{SGD} \) outputs \( \overline{x} \) satisfying
\[
\mathbb{E}[\| \nabla F(x^*) \|^2] \leq O\left( \frac{L^2 \| x_0 - x^* \|^2}{\frac{T}{1-\alpha \mathbb{L}}} + \frac{L \sqrt{\mathbb{V}} \| x_0 - x^* \|^2}{\sqrt{T}} \right).
\]
(b) if \( T \geq \frac{L}{2} \), then \( \text{SGD} \) outputs \( \overline{x} \) satisfying
\[
\mathbb{E}[\| \nabla F_{\eta}(\overline{x}) \|^2] \leq O\left( \frac{L^2 \sqrt{\mathbb{V}} \| x_0 - x^* \|^2}{\sqrt{T}} + \left(1 - \frac{\sigma}{T} \right)^{\Omega(T)} \sigma L \| x_0 - x^* \|^2 \right).
\]
Corollary 4.3. Hiding \( \mathbb{V}, L \), \( \| x_0 - x^* \| \) in the big-O notation, classical SGD finds \( x \) with
\[
F(x) - F(x^*) \leq O(T^{-1/2}) \quad \| \nabla F(x) \| \leq O(T^{-1/4}) \quad \text{for Problem (3.1)}
\]
or
\[
F(x) - F(x^*) \leq O((\sigma T)^{-1}) \quad \| \nabla F(x) \| \leq O((\sigma T)^{-1/2}) \quad \text{if } f(\cdot) \text{ is } \sigma \text{-strongly convex for } \sigma > 0.
\]

5 An Auxiliary Lemma on Regularization
Consider a regularized objective
\[
G(x) := \psi(x) + g(x) := \psi(x) + f(x) + \sum_{s=1}^{S} \frac{\sigma_s}{2} \| x - \tilde{x}_s \|^2,
\]
slightly worse rate \( T = O\left( \frac{\sqrt{\mathbb{V}}}{\frac{T}{1-\alpha \mathbb{L}}} + \frac{L^2 \| x_0 - x^* \|^2}{\frac{T}{1-\alpha \mathbb{L}}} \right) \). If the readers find either statement explicitly stated somewhere, please let us know and we would love to include appropriate citations.
where $\hat{x}_1, \ldots, \hat{x}_S$ are fixed vectors in $\mathbb{R}^d$. The following lemma says that, if we find an approximate stationary point $x$ of $G(x)$, then it is also an approximate stationary point of $F(x)$ up to some additive error.

**Lemma 5.1.** Suppose $\psi(x)$ is proper convex and $f(x)$ is convex and $L$-smooth. By definition, $g(x)$ is $\tilde{\sigma}$-strongly convex with $\tilde{\sigma} := \sum_{s=1}^{S} \sigma_s$. Let $x^*$ be the unique minimizer of $G(y)$ in (5.1), and $x$ be an arbitrary vector in the domain of $G$. Then, for every $\eta \in (0, \frac{1}{L\sigma}]$, we have

$$||G_{F,\eta}(x)|| \leq \sum_{s=1}^{S} \sigma_s ||x^* - \hat{x}_s|| + 3||G_{G,\eta}(x)||.$$ 

**Remark 5.2.** Lemma 5.1 should be easy to prove in the special case of $\psi(x) \equiv 0$. Indeed,

$$||\nabla f(x)|| = ||\nabla g(x) + \sum_{s} \sigma_s (x - \hat{x}_s)|| \leq ||\nabla g(x)|| + \sum_{s} \sigma_s ||x - \hat{x}_s||$$

$$\leq ||\nabla g(x)|| + \sum_{s} \sigma_s ||x^* - \hat{x}_s|| + \tilde{\sigma} ||x^* - x|| \leq 2||\nabla g(x)|| + \sum_{s} \sigma_s ||x^* - \hat{x}_s||.$$

Above, inequalities $\odot$ and $\oplus$ both use the triangle inequality; and inequality $\ominus$ is due to the $\tilde{\sigma}$-strong convexity of $g(x)$ (see for instance [23, Sec. 2.1.3]).

**Proof of Lemma 5.1.** See full version. \qed

### 6 Approach 3: SGD and Recursive Regularization

In this section, add a logarithmic number of regularizers to the objective, each centered at a different but carefully chosen point. Specifically, given parameters $\sigma_1, \ldots, \sigma_S > 0$, we define functions

$$F^{(0)}(x) := F(x) \quad \text{and} \quad F^{(s)}(x) := F^{(s-1)}(x) + \frac{\sigma_s}{2} ||x - \hat{x}_s||^2 \quad \text{for } s = 1, 2, \ldots, S$$

where each $\hat{x}_s$ (for $s \geq 1$) is an approximate minimizer of $F^{(s-1)}(x)$.

If $f(x)$ is $\sigma$-strongly convex, then we choose $S = \log_2 \frac{L}{\sigma}$ and let $\sigma_0 = \sigma$ and $\sigma_s = 2\sigma_{s-1}$. To calculate each $\hat{x}_s$, we apply SGD for $\frac{T}{\sigma}$ iterations. This totals to a gradient complexity of $T$. We summarize this method as $\text{SGD}^3_{\text{sc}}$ in Algorithm 3.

If $f(x)$ is not strongly convex, then we regularize it by $G(x) = F(x) + \frac{\sigma}{2} ||x - x_0||^2$ for some small parameter $\sigma > 0$, and then apply $\text{SGD}^3_{\text{sc}}$. We summarize this final method as $\text{SGD}^3$ in Algorithm 4.

We prove the following main theorem:

**Theorem 3 (SGD3).** Let $x^* \in \arg \min_x \{F(x)\}$. To solve Problem (3.1) given a starting vector $x_0 \in \mathbb{R}^d$ and any $\eta = \frac{C}{x}$ for some absolute constant $C \in (0, 1]$,

(a) If $f(x)$ is $\sigma$-strongly convex for $\sigma \in (0, L]$ and $T \geq \frac{L}{\sigma} \log \frac{L}{\sigma}$, then $\text{SGD}^3_{\text{sc}}(F, x_0, \sigma, L, T)$ outputs $\pi$ satisfying

$$\mathbb{E}[||G_{F,\eta}(\pi)||] \leq O\left(\frac{\sqrt{V} \cdot \log^{3/2} L}{\sqrt{T}} \frac{L}{\sigma} \right) + (1 - \frac{\sigma}{L})^{O(T/\log(L/\sigma))} ||x_0 - x^*||.$$

(b) If $\sigma \in (0, L]$ and $T \geq \frac{1}{\sigma} \log \frac{L}{\sigma}$, then $\text{SGD}(F, x_0, \sigma, L, T)$ outputs $\pi$ satisfying

$$\mathbb{E}[||G_{F,\eta}(\pi)||] \leq O\left(\sigma ||x_0 - x^*|| + \frac{\sqrt{V} \cdot \log^{3/2} L}{\sqrt{T}} \frac{L}{\sigma} \right) + (1 - \frac{\sigma}{L})^{\Theta(T/\log(L/\sigma))} ||x_0 - x^*||.$$

If $\sigma$ is appropriately chosen, then we find $\pi$ with $\mathbb{E}[||G_{F,\eta}(\pi)||] \leq \epsilon$ in gradient complexity

$$T \leq O\left(\frac{\sqrt{V} \cdot \log^{3} L ||x_0 - x^*||}{\epsilon^2} + \frac{L ||x_0 - x^*||}{\epsilon} \cdot \log \frac{L ||x_0 - x^*||}{\epsilon} \right).$$

**Remark 6.1.** All expected guarantees of the form $\mathbb{E}[||G_{F,\eta}(\pi)||] \leq \epsilon^2$ or $\mathbb{E}[||G_{F,\eta}(\pi)||] \leq \epsilon$ throughout this paper can be made into high-confidence bound by repeating the algorithm multiple times, each time estimating the value of $||G_{F,\eta}(\pi)||$ using roughly $O(\frac{1}{\epsilon^2})$ stochastic gradient computations, and finally outputting the point $\pi$ that leads to the smallest value $||G_{F,\eta}(\pi)||$. 

7
6.1 Proof of Theorem 3

Before proving Theorem 3, we state a few properties regarding the relationships between the objective-optimality of $\hat{x}_s$ and point distances.

Claim 6.2. Suppose for every $s = 1, \ldots, S$ the vector $\hat{x}_s$ satisfies

$$
\mathbb{E}[F^{(s-1)}(\hat{x}_s) - F^{(s-1)}(x^*_{s-1})] \leq \delta_s
$$

where $x^*_{s-1} \in \arg \min \{F^{(s-1)}(x)\}$, then,

(a) for every $s \geq 1$, $\mathbb{E}[\|\hat{x}_s - x^*_{s-1}\|^2] \leq \mathbb{E}[\|\hat{x}_s - x^*_{s-1}\|^2] \leq \frac{2\delta_s}{\sigma_{s-1}}$.

(b) for every $s \geq 1$, $\mathbb{E}[\|x^*_{s-1} - \hat{x}_s\|^2] \leq \mathbb{E}[\|x^*_{s-1} - \hat{x}_s\|^2] \leq \frac{\delta_s}{\sigma_s}$; and

(c) if $\sigma_s = \frac{2\sigma_{s-1}}{S}$ for all $s \geq 1$, then $\mathbb{E}[\sum_{s=1}^{S} \sigma_s \|x^*_{s} - \hat{x}_s\|^2] \leq 4 \sum_{s=1}^{S} \sqrt{\delta_s \sigma_s}.$

Proof of Claim 6.2

(a) $\mathbb{E}[\|\hat{x}_s - x^*_{s-1}\|^2] \leq \frac{\frac{\sigma_s}{2} \|x^*_{s-1} - \hat{x}_s\|^2 + \frac{\sigma_s}{2} \|x^*_{s-1} - \hat{x}_s\|^2}{\|x^*_{s-1} - \hat{x}_s\|^2} \leq \frac{\sigma_s}{2 \sigma_{s-1}} \mathbb{E}[F^{(s-1)}(\hat{x}_s) - F^{(s-1)}(x^*_{s-1})] \leq \frac{2\delta_s}{\sigma_{s-1}}.$ Here, inequality $\frac{\sigma_s}{2} \|x^*_{s-1} - \hat{x}_s\|^2$ is because $\mathbb{E}[x^2] \leq \mathbb{E}[X^2]$, and inequality $\frac{\sigma_s}{2}$ is due to the strong convexity of $F^{(s-1)}(x)$.

(b) We derive that

$$
\sigma_s \|x^*_{s} - \hat{x}_s\|^2 \leq \frac{\sigma_s}{2} \|x^*_{s} - \hat{x}_s\|^2 + F^{(s)}(\hat{x}_s) - F^{(s)}(x^*_{s}) = F^{(s)}(\hat{x}_s) - F^{(s-1)}(x^*_{s-1}) \leq \frac{2\delta_s}{\sigma_{s-1}}.
$$

Here, inequality $\frac{\sigma_s}{2} \|x^*_{s-1} - \hat{x}_s\|^2$ is due to the strong convexity of $F^{(s)}(x)$, and inequality $\frac{\sigma_s}{2}$ is because of the minimality of $x^*_{s-1}$. Taking expectation we have $\mathbb{E}[\|x^*_{s} - \hat{x}_s\|^2] \leq \mathbb{E}[\|x^*_{s-1} - \hat{x}_s\|^2] \leq \frac{\delta_s}{\sigma_s}$.

(c) Define $P_t := \sum_{s=1}^{t} \sigma_s \|x^*_{s-t} - \hat{x}_s\|$ for each $t \geq 0, 1, \ldots, S$. Then by triangle inequality we have

$$
P_s - P_{s-1} \leq \sigma_s \|x^*_{s} - \hat{x}_s\| + (\sum_{t=1}^{s-1} \sigma_t) \cdot \|x^*_{s} - x^*_{s-1}\|.
$$

Using the parameter choice of $\sigma_s = 2\sigma_{s-1}$, and plugging in Claim 6.2(a) and Claim 6.2(b) we have

$$
\mathbb{E}[P_s - P_{s-1}] \leq \sqrt{\delta_s \sigma_s} + \sigma_s \cdot \mathbb{E}[\|x^*_{s} - \hat{x}_s\| + \|x^*_{s-1} - \hat{x}_s\|] \leq 4 \sqrt{\delta_s \sigma_s}.
$$

Proof of Theorem 3(a) We first note that, when writing $f^{(s-1)}(x) = F^{(s-1)}(x) - \psi(x)$, each $f^{(s-1)}$ is at least $\sigma_{s-1}$-strongly convex and $L + \sum_{t=1}^{s-1} \sigma_t \leq 3L$ Lipschitz smooth. Therefore,
applying Theorem 4.1(b) we have
\[
\mathbb{E}[F(s^{-1})(\hat{x}_s) - F(s^{-1})(x_s^*)] \leq O \left( \frac{SV}{\sigma s^{-1} T} \right) + \left( 1 - \frac{\sigma s^{-1}}{3L} \right)^{\Omega(T/S)} \mathbb{E}[\sigma_s ||\hat{x}_s - x_s^*||^2].
\]
If \( s = 1 \), this means (recalling \( \hat{x}_0 = x_0 \) and \( x_0^* = x^* \))
\[
\mathbb{E}[F(0)(\hat{x}_0) - F(0)(x^*)] \leq O \left( \frac{SV}{\sigma_0 T} \right) + \left( 1 - \frac{\sigma_0}{L} \right)^{\Omega(T/S)} \sigma_0 ||x_0 - x^*||^2.
\]
If \( s > 1 \), this means
\[
\mathbb{E}[F(s^{-1})(\hat{x}_s) - F(s^{-1})(x_s^*)] \leq O \left( \frac{SV}{\sigma s^{-1} T} \right) + \left( 1 - \frac{\sigma s^{-1}}{L} \right)^{\Omega(T/S)} \mathbb{E}[F(s^{-2})(\hat{x}_{s-1}) - F(s^{-2})(x_{s-1}^*)].
\]
Together, this means to satisfy (6.1) it suffices to choose \( \delta_s \) so that
\[
\delta_s = O \left( \frac{SV}{\sigma s T} \right) + \left( 1 - \frac{\sigma}{L} \right)^{\Omega(sT/S)} \sigma_0 ||x_0 - x^*||^2.
\]
Using Lemma 2.3 with \( F(S^{-1}) \) and \( y = x = \hat{x}_s \), we have
\[
\frac{\sigma}{2} \mathbb{E}[\sigma_{F(\hat{x}_s),\eta}(\hat{x}_S)]^2 \leq F(S^{-1})(\hat{x}_S) - F(S^{-1})(\hat{x}_S) \leq F(S^{-1})(\hat{x}_S) - F(S^{-1})(x_{S-1}^*) \text{ and therefore}
\]
\[
\mathbb{E}[\mathbb{E}[\sigma_{F(\hat{x}_s),\eta}(\hat{x}_S)]^2] \leq \frac{2\delta_s}{\eta} = O(L\delta_s).
\]
Plugging this into Lemma 5.1 (with \( G(x) = F(S^{-1})(x) \)) and Claim 6.2(c), we have
\[
\mathbb{E}[\mathbb{E}[\sigma_{F,\eta}(\hat{x})]^2] \leq \mathbb{E}\left[ \sum_{s=1}^{S-1} \sigma_s ||\hat{x}_s - \hat{x}_s|| + 3\mathbb{E}[\sigma_{F(\hat{x}_s),\eta}(\hat{x}_S)] \right] \leq O\left( \sum_{s=1}^{S-1} \sqrt{\delta_s \sigma_s} + \sqrt{L\delta_s} \right)
\]
\[
= O\left( \sum_{s=1}^{S} \sqrt{\delta_s \sigma_s} \right) \leq O\left( \frac{S^{3/2}}{T/2} \right) + \left( 1 - \frac{\sigma}{L} \right)^{\Omega(T/S)} \sigma_0 ||x_0 - x^*||^2.
\]

**Proof of Theorem 3(b)** Define \( G(x) := F(x) + \frac{\sigma}{2} ||x-x_0||^2 \) and let \( x_G^* \) be the (unique) minimizer of \( G(\cdot) \). Note that \( x_G^* \) may be different from \( x^* \) which is a minimizer of \( F(\cdot) \). Applying Theorem 3(a) on \( G(x) \) and Lemma 5.1 with \( S = 1 \) and \( \hat{x}_1 = x_0 \), we have
\[
\mathbb{E}[\mathbb{E}[\sigma_{F,\eta}(\hat{x})]] \leq O\left( \frac{\sigma}{2} ||x_0 - x_0^*|| + \sqrt{\log \frac{3}{\delta_1}} \right) + \left( 1 - \frac{\sigma}{L} \right)^{\Omega(T/\log(1/\delta_1))} \sigma_0 ||x_0 - x_0^*||^2.
\]
Now, by definition \( \frac{\sigma}{2} ||x^* - x_0||^2 - \frac{\sigma}{2} ||x_G^* - x_0||^2 = (G(x^*) - F(x^*)) + (F(x_G^*) - G(x_G^*)) \geq 0 \) so we have \( ||x_G^* - x_0|| \leq ||x^* - x_0|| \). This completes the proof.

**Acknowledgements**

We would like to thank Lin Xiao for suggesting reference [29, Lemma 3.7], an anonymous researcher from the Simons Institute for suggesting reference [25], Yurii Nesterov for helpful discussions, Xinyu Weng for discussing the motivations, Sébastien Bubeck, Yuval Peres, and Lin Xiao for discussing notations, Chi Jin for discussing reference [27], and Dmitriy Drusvyatskiy for discussing the notion of Moreau envelope.

**References**


