Natasha: Faster Non-Convex Stochastic Optimization Via Strongly Non-Convex Parameter

Zeyuan Allen-Zhu
zeyuan@csail.mit.edu
Princeton University / Institute for Advanced Study

February 2, 2017

Abstract

Given a non-convex function $f(x)$ that is an average of n smooth functions, we design stochastic first-order methods to find its approximate stationary points. The performance of our new methods depend on the smallest (negative) eigenvalue $-\sigma$ of the Hessian. This parameter σ captures how strongly non-convex $f(x)$ is, and is analogous to the strong convexity parameter for convex optimization.

Our methods outperform the best known results for a wide range of σ, and can also be used to find approximate local minima.

In particular, we find an interesting dichotomy: there exists a threshold σ_0 so that the fastest methods for $\sigma > \sigma_0$ and for $\sigma < \sigma_0$ have drastically different behaviors: the former scales with $n^{2/3}$ and the latter scales with $n^{3/4}$.

1 Introduction

We study the problem of composite *non-convex* minimization:

$$\min_{x\in\mathbb{R}^d} \Big\{ F(x) \stackrel{\text{def}}{=} \psi(x) + f(x) \stackrel{\text{def}}{=} \psi(x) + \frac{1}{n}\sum_{i=1}^{n} f_i(x) \Big\} \tag{1.1}$$

where each $f_i(x)$ is *nonconvex but smooth*, and $\psi(\cdot)$ is proper convex, possibly nonsmooth, but relatively simple. We are interested in finding a point x that is an *approximate local minimum* of $F(x)$.

- The finite-sum structure $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$ arises prominently in large-scale machine learning tasks. In particular, when minimizing loss over a training set, each example i corresponds to one loss function $f_i(\cdot)$ in the summation. This finite-sum structure allows one to perform stochastic gradient descent with respect to a random $\nabla f_i(x)$.
- The so-called *proximal* term $\psi(x)$ adds more generality to the model. For instance, if $\psi(x)$ is the indicator function of a convex set, then problem (1.1) becomes constraint minimization; if $\psi(x) = \|x\|_1$, then we can allow problem (1.1) to perform feature selection. In general, $\psi(x)$ has to be a simple function where the projection operation $\arg\min_x\{\psi(x) + \frac{1}{2\eta}\|x - x_0\|^2\}$ is efficiently computable. At a first reading of this paper, one can assume $\psi(x) \equiv 0$ for simplicity.

Many non-convex machine learning problems fall into problem (1.1). Most notably, training *deep neural networks* and classifications with *sigmoid loss* correspond to (1.1) where neither $f_i(x)$ or

$f(x)$ is convex [3]. However, our understanding to this challenging non-convex problem is very limited.

1.1 Strongly Non-Convex Optimization

Let L be the smoothness parameter for each $f_i(x)$, meaning all the eigenvalues of $\nabla^2 f_i(x)$ lie in $[-L, L]$.[1]

We denote by $\sigma \in [0, L]$ the *strong-nonconvexity* parameter of $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$, meaning that

$$\text{all the eigenvalues of } \nabla^2 f(x) \text{ lie in } [-\sigma, L].$$

We emphasize that parameter σ is analogous to the *strong-convexity* parameter μ for convex optimization, where all the eigenvalues of $\nabla^2 f(x)$ lie in $[\mu, L]$ for some $\mu > 0$.

We wish to find an ε-approximate stationary point (a.k.a. critical point) of $F(x)$, that is

$$\text{a point } x \text{ satisfying } \|\mathcal{G}(x)\| \le \varepsilon$$

where $\mathcal{G}(x)$ is the so-called gradient mapping of $F(x)$ (see Section 2 for a formal definition). In the special case of $\psi(\cdot) \equiv 0$, gradient mapping $\mathcal{G}(x)$ is the same as gradient $\nabla f(x)$, so x satisfies $\|\nabla f(x)\| \le \varepsilon$.

Since $f(\cdot)$ is σ-strongly nonconvex, any ε-approximate stationary point is automatically also an (ε, σ)-approximate local minimum — meaning that the Hessian of the output point $\nabla^2 f(x) \succeq -\sigma \mathbf{I}$ is approximately positive semidefinite (PSD).

Motivations and Remarks

- We focus on strongly non-convex optimization because introducing this parameter σ allows us to perform a *more refined study* of non-convex optimization. If σ equals L then L-strongly nonconvex optimization is equivalent to the general non-convex optimization.
- We focus only on finding stationary points as opposed to local minima, because in a recent study —see Appendix A— researchers have shown that finding (ε, δ)-approximate local minima *reduces* to finding ε-approximate stationary points in an $O(\delta)$-strongly nonconvex function.
- Parameter σ is often not constant and can be much smaller than L. For instance, second-order methods often find $(\varepsilon, \sqrt{\varepsilon})$-approximate local minima [18] and this corresponds to $\sigma = \sqrt{\varepsilon}$.

1.2 Known Results

Despite the widespread use of nonconvex models in machine learning and related fields, our understanding to non-convex optimization is still very limited. Until recently, nearly all research papers have been mostly focusing on either $\sigma = 0$ or $\sigma = L$:

- If $\sigma = 0$, the accelerated SVRG method [8, 21] finds x satisfying $F(x) - F(x^*) \le \varepsilon$, in gradient complexity $\widetilde{O}\big(n + n^{3/4}\sqrt{L/\varepsilon}\big)$.[2] This result is irrelevant to this paper because $f(x)$ is simply convex.
- If $\sigma = L$, the SVRG method [3] finds an ε-approximate stationary point of $F(x)$ in gradient complexity $O(n + n^{2/3} L/\varepsilon^2)$.

[1] This definition also applies to functions $f(x)$ that are not twice differentiable, see Section 2 for details.
[2] We use the $\widetilde{O}$ notation to hide poly-logarithmic factors in $n, L, 1/\varepsilon$.

- If $\sigma = L$, gradient descent finds an ε-approximate stationary point of $F(x)$ in gradient complexity $O(nL/\varepsilon^2)$.
- If $\sigma = L$, stochastic gradient descent finds an ε-approx. stationary point of $F(x)$ in gradient complexity $O(L^2/\varepsilon^4)$.

Throughout this paper, we refer to *gradient complexity* as the total number of stochastic gradient computations $\nabla f_i(x)$ and proximal computations $y \leftarrow \mathsf{Prox}_{\psi,\eta}(x) \stackrel{\text{def}}{=} \arg\min_y\{\psi(y) + \frac{1}{2\alpha}\|y-x\|^2\}$.[3]

Very recently, it was observed by two independent groups [1, 9] —although implicitly, see Section 2.1— that for solving the σ-strongly nonconvex problem, one can repeatedly regularize $F(x)$ to make it σ-strongly convex, and then apply the accelerated SVRG method to minimize this regularized function. Under mild assumption $\sigma \geq \varepsilon^2$, this approach

- finds an ε-approximate stationary point in gradient complexity $\widetilde{O}\big(\frac{n\sigma + n^{3/4}\sqrt{L\sigma}}{\varepsilon^2}\big)$.

We call this method repeatSVRG in this paper. Unfortunately, repeatSVRG is even slower than the vanilla SVRG for $\sigma = L$ by a factor $n^{1/3}$, see Figure 1.

Remark on SGD. Stochastic gradient descent (SGD) has a *slower* convergence rate (i.e., in terms of $1/\varepsilon^4$) than other cited first-order methods (i.e., in terms of $1/\varepsilon^2$). However, the complexity of SGD does not depend on n and thus is incomparable to gradient descent, SVRG, or repeatSVRG.[4] This is one of the main motivations to study how to reduce the complexity of non-SGD methods, especially in terms of n.

1.3 Our New Results

In this paper, we identify an interesting dichotomy with respect to the spectrum of the nonconvexity parameter $\sigma \in [0, L]$. In particular, we showed that if $\sigma \leq L/\sqrt{n}$, then our new method `Natasha` finds an ε-approximate stationary point of $F(x)$ in gradient complexity

$$O\Big(n \log\frac{1}{\varepsilon} + \frac{n^{2/3}(L^2\sigma)^{1/3}}{\varepsilon^2}\Big) .$$

In other words, together with repeatSVRG, we have improved the gradient complexity for σ-stringly nonconvex optimization to[5]

$$\widetilde{O}\Big(\min\Big\{\frac{n^{3/4}\sqrt{L\sigma}}{\varepsilon^2}, \frac{n^{2/3}(L^2\sigma)^{1/3}}{\varepsilon^2}\Big\}\Big)$$

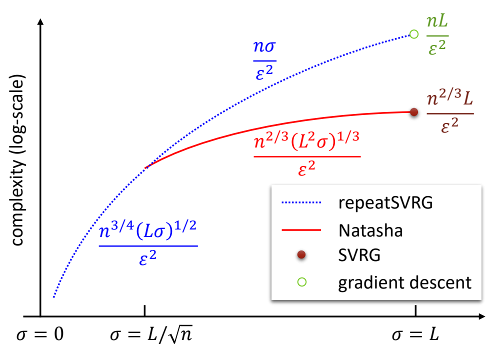

Figure 1: Comparison to prior works

and the first term in the min is smaller if $\sigma > L/\sqrt{n}$ and the second term is smaller if $\sigma < L/\sqrt{n}$. We illustrate our performance improvement in Figure 1. Our result matches that of SVRG for $\sigma = L$, and has a much simpler analysis.

Additional Results. One can take a step further and ask what if each function $f_i(x)$ is (ℓ_1, ℓ_2)-smooth for parameters $\ell_1, \ell_2 \geq \sigma$, meaning that all the eigenvalues of $\nabla^2 f_i(x)$ lie in $[-\ell_2, \ell_1]$.

[3]Some authors also refer to them as incremental first-order oracle (IFO) and proximal oracle (PO) calls. In most machine learning applications, each IFO and PO call can be implemented to run in time $O(d)$ where d is the dimension of the model, or even in time $O(s)$ if s is the average sparsity of the data vectors.

[4]In practice, there are examples in non-convex empirical risk minimization [3] and in training neural networks [3, 19] where SVRG alone can outperform SGD. Of course, for deep learning tasks, SGD remains to be the best practical method of choice.

[5]We remark here that this is under mild assumptions for ε being sufficiently small. For instance, the result of [1, 9] requires $\varepsilon^2 \leq \sigma$. In our result, the term $n \log\frac{1}{\varepsilon}$ disappears when $\varepsilon^6 \leq L^2\sigma/n$.

We show that a variant of our method, which we call $\texttt{Natasha}^{\mathsf{full}}$, solves this more refined problem of (1.1) with total gradient complexity $O\big(n\log\frac{1}{\varepsilon}+\frac{n^{2/3}(\ell_1\ell_2\sigma)^{1/3}}{\varepsilon^2}\big)$ as long as $\frac{\ell_1\ell_2}{\sigma^2}\le n^2$.

Remark 1.1. In applications, ℓ_1 and ℓ_2 can be of very different magnitudes. The most influential example is finding the leading eigenvector of a symmetric matrix. Using the so-called shift-and-invert reduction [12], computing the leading eigenvector reduces to the convex version of problem (1.1), where each $f_i(x)$ is $(\lambda,1)$-smooth for $\lambda \ll 1$. Other examples include all the applications that are built on shift-and-invert, including high rank SVD/PCA [5], canonical component analysis [4], online matrix learning [6], and approximate local minima algorithms [1, 9].

Mini-Batch. Our result generalizes trivially to the mini-batch stochastic setting, where in each iteration one computes $\nabla f_i(x)$ for b random choices of index $i\in[n]$ and average them. The stated gradient complexities of $\texttt{Natasha}$ and $\texttt{Natasha}^{\mathsf{full}}$ can be adjusted so that the factor $n^{2/3}$ is replaced with $n^{2/3}b^{1/3}$.

1.4 Our Techniques

Let us first recall the main idea behind stochastic variance-reduced methods, such as SVRG [14].

The SVRG method divides iterations into epochs, each of length n. It maintains a snapshot point $\widetilde{\mathbf{x}}$ for each epoch, and computes the full gradient $\nabla f(\widetilde{\mathbf{x}})$ only for snapshots. Then, in each iteration t at point x_t, SVRG defines gradient estimator $\widetilde{\nabla}=\nabla f_i(x_t)-\nabla f_i(\widetilde{\mathbf{x}})+\nabla f(\widetilde{\mathbf{x}})$ which satisfies $\mathbb{E}_i[\widetilde{\nabla}]=\nabla f(x_t)$, and performs proximal update $x_{t+1}\leftarrow \mathsf{Prox}_{\psi,\alpha}\big(x_t-\alpha\widetilde{\nabla}\big)$ for some learning rate α. (Recall that if $\psi(\cdot)\equiv 0$ then we would have $x_{t+1}\leftarrow x_t-\alpha\widetilde{\nabla}$.)

In nearly all the aforementioned results for nonconvex optimization, researchers have either directly applied SVRG [3] (for the case $\sigma=L$), or repeatedly applied SVRG [1, 9] (for general $\sigma\in[0,L]$). This puts some limitation in the algorithmic design, because SVRG requires each epoch to be of length exactly n.[6]

Our New Idea. In this paper, we propose $\texttt{Natasha}$ and $\texttt{Natasha}^{\mathsf{full}}$, two methods that are no longer black-box reductions to SVRG. Both of them still divide iterations into epochs of length n, and compute gradient estimators $\widetilde{\nabla}$ the same way as SVRG. However, we do not apply compute $x_t-\alpha\widetilde{\nabla}$ directly.

- In our base algorithm $\texttt{Natasha}$, we divide each epoch into p sub-epochs, each with a starting vector $\widehat{\mathsf{x}}$. Our theory suggests the choice $p\approx(\frac{\sigma^2}{L^2}n)^{1/3}$. Then, we replace the use of $\widetilde{\nabla}$ with $\widetilde{\nabla}+2\sigma(x_t-\widehat{\mathsf{x}})$. This is equivalent to replacing $f(x)$ with its regularized version $f(x)+\sigma\|x-\widehat{\mathsf{x}}\|^2$, where the center $\widehat{\mathsf{x}}$ varies across sub-epochs. We provide pseudocode in Algorithm 1 and illustrate it in Figure 2.

  We view this additional term $2\sigma(x_t-\widehat{\mathsf{x}})$ as a type of **retraction**, which stabilizes the algorithm by moving the vector a bit in the backward direction towards $\widehat{\mathsf{x}}$.

- In our full algorithm $\texttt{Natasha}^{\mathsf{full}}$, we add one more ingredient on top of $\texttt{Natasha}$. That is, we perform updates $z_{t+1}\leftarrow\mathsf{Prox}_{\psi,\alpha}(z_t-\alpha\widetilde{\nabla})$ with respect to a different sequence $\{z_t\}$, and then define $x_t=\frac{1}{2}z_t+\frac{1}{2}\widehat{\mathsf{x}}$ and compute gradient estimators $\widetilde{\nabla}$ at points x_t. We provide pseudocode in Algorithm 2.

  We view this averaging $x_t=\frac{1}{2}z_t+\frac{1}{2}\widehat{\mathsf{x}}$ as another type of **retraction**, which stabilizes the algorithm by moving towards $\widehat{\mathsf{x}}$. The technique of computing gradients at points x_t but moving

[6] The epoch length of SVRG is always n (or a constant multiple of n in practice), because this ensures the computation of $\widetilde{\nabla}$ is of amortized gradient complexity $O(1)$. The per-iteration complexity of SVRG is thus the same as the traditional stochastic gradient descent (SGD).

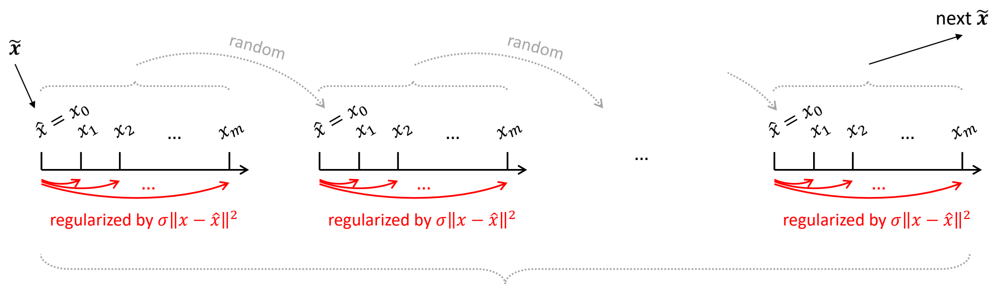

Figure 2: One full epoch of `Natasha`. The n iterations are divided into p sub-epochs, each consisting of $m = n/p$ steps.

a different sequence of points z_t is related to the *Katyusha momentum* recently developed for convex optimization [2].

1.5 Other Related Work

Methods based on variance-reduced stochastic gradients were first introduced for convex optimization. The first such method is SAG by Schmidt et al [20]. The two most popular choices for gradient estimators are the SVRG-like one we adopted in this paper (independently introduced by [14, 22], and the SAGA-like one introduced by [10]. In nearly all applications, the results proven for SVRG-like estimators and SAGA-like estimators are simply exchangeable (therefore, the results of this paper naturally generalize to SAGA-like estimators).

The first "non-convex use" of variance reduction is by Shalev-Shwartz [21] who assumes that each $f_i(x)$ is non-convex but their average $f(x)$ is still convex. This result has been slightly improved to several more refined settings [8]. The first truly non-convex use of variance reduction (i.e., for $f(x)$ being also non-convex) is independently by [3] and [19]. First-order methods only find stationary points (unless there is extra assumption on the randomness of the data), and converge no faster than $1/\varepsilon^2$.

When the second-order Hessian information is used, one can (1) find local minima instead of stationary points, and (2) improve the $1/\varepsilon^2$ rate to $1/\varepsilon^{1.5}$. The first such result is by cubic-regularized Newton's method [18]; however, its per-iteration complexity is very high. Very recently, two independent groups of authors tackled this problem from a somewhat similar viewpoint [1, 9]: if the computation of Hessian-vector multiplications (i.e., $\big(\nabla^2 f_i(x)\big)v$) is on the same order of the computation of gradients $\nabla f_i(x)$,[7] then one can obtain a $(\varepsilon, \sqrt{\varepsilon})$-approximate local minimum in gradient complexity $\widetilde{O}\big(\frac{n}{\varepsilon^{1.5}} + \frac{n^{3/4}}{\varepsilon^{1.75}}\big)$, if we use big-$O$ to also hide dependencies on the smoothness parameters.[8] Although Carmon et al. [9] only stated a complexity of $\widetilde{O}\big(\frac{n}{\varepsilon^{1.75}}\big)$ in the non-stochastic setting, their result generalizes to our stated complexity in the stochastic setting. As we argue in Appendix A, both these methods reduce the problem of finding $(\varepsilon, \sqrt{\varepsilon})$-approximate local minima to that of finding ε-approximate stationary points in $\sqrt{\varepsilon}$-strongly nonconvex functions.

Other related papers include Ge et al. [13] where the authors showed that a noise-injected version of SGD converges to local minima instead of critical points, as long as the underlying function is "strict-saddle." Their theoretical running time is a large polynomial in the dimension. Lee et al. [15] showed that gradient descent, starting from a random point, almost surely converges

[7] A lot of interesting problems satisfy this property, including training neural nets.

[8] More precisely, they obtain an $(\varepsilon, \sqrt{L_2\varepsilon})$-approximate local minimum using gradient complexity $\widetilde{O}\big(\frac{n\sqrt{L_2}}{\varepsilon^{1.5}} + \frac{n^{3/4}L_2^{1/4}L^{1/2}}{\varepsilon^{1.75}}\big)$ where L_2 is the second-order smoothness of $f(\cdot)$.

to a local minimum if the function is "strict-saddle". The rate of convergence required is somewhat unknown.

2 Preliminaries

Throughout this paper, we denote by $\|\cdot\|$ the Euclidean norm. We use $i \in_R [n]$ to denote that i is generated from $[n] = \{1, 2, \dots, n\}$ uniformly at random. We denote by $\nabla f(x)$ the full gradient of function f if it is differentiable, and $\partial f(x)$ any subgradient if f is only Lipschitz continuous at point x. We let x^* be any minimizer of $F(x)$.

Recall some definitions on strong convexity (SC), strongly nonconvexity, and smoothness.

Definition 2.1. *For a function $f\colon \mathbb{R}^d \to \mathbb{R}$,*

- *f is σ-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 σ-strongly nonconvex 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_1, ℓ_2)-smooth if $\forall x, y \in \mathbb{R}^d$, it satisfies*

$$f(x) + \langle \nabla f(x), y - x\rangle + \frac{\ell_1}{2}\|x - y\|^2 \geq f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle - \frac{\ell_2}{2}\|x - y\|^2 \ .$$

- *f is L-smooth if it is (L, L)-smooth.*

The (ℓ_1, ℓ_2)-smoothness parameters were introduced in [8] to tackle the convex setting of problem (1.1). The notion of strong nonconvexity is also known as "lower smoothness [8]" or "almost convexity [9]". We refrain from using the name "almost convexity" because it coincides with several other definitions in optimization literatures.

Definition 2.2. *Given a parameter $\eta > 0$, the* gradient mapping *of $F(\cdot)$ in (1.1) at point x is*

$$\mathcal{G}_\eta(x) \stackrel{\text{def}}{=} \frac{1}{\eta}\big(x - x'\big) \qquad \text{where} \qquad x' = \arg\min_y \big\{\psi(y) + \langle \nabla f(x), y\rangle + \frac{1}{2\eta}\|y - x\|^2\big\}$$

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

The following theorem for the SVRG method can be found for instance in [8], which is built on top of the results [11, 16, 21]:

Theorem 2.3 (SVRG)**.** *Let $G(y) \stackrel{\text{def}}{=} \psi(y) + \frac{1}{n}\sum_{i=1}^n g_i(y)$ be σ-strongly convex, then the SVRG method finds a point y satisfying $G(y) - G(y^*) \leq \varepsilon$*

- *with gradient complexity $O\big((n + \frac{L^2}{\sigma^2})\log\frac{1}{\varepsilon}\big)$, if each $g_i(\cdot)$ is L-smooth (for $L \geq \sigma$); or*
- *with gradient complexity $O\big((n + \frac{\ell_1\ell_2}{\sigma^2})\log\frac{1}{\varepsilon}\big)$, if each $g_i(\cdot)$ is (ℓ_1, ℓ_2)-smooth (for $\ell_1, \ell_2 \geq \sigma$).*

If one performs acceleration, the running times become $\widetilde{O}\big(n + n^{3/4}\sqrt{L/\sigma}\big)$ and $\widetilde{O}\big(n + n^{3/4}(\ell_1\ell_2\sigma^2)^{1/4}\big)$.

2.1 RepeatSVRG

We recall the idea behind a simple algorithm —that we call repeatSVRG— which finds the ε-approximate stationary points for problem (1.1) when $f(x)$ is σ-strongly nonconvex. The algorithm is divided into stages. In each stage t, consider a modified function $F_t(x) \stackrel{\text{def}}{=} F(x) + \sigma\|x - x_t\|^2$. It is easy to see that $F_t(x)$ is σ-strongly convex, so one can apply the accelerated SVRG method to minimize $F_t(x)$. Let x_{t+1} be any sufficiently accurate approximate minimizer of $F_t(x)$.[9]

[9] Since the accelerated SVRG method has a linear convergence rate for strongly convex functions, the complexity to find such x_{t+1} only depends logarithmically on this accuracy.

Now, one can prove (c.f. Section 4) that x_{t+1} is an $O(\sigma\|x_t - x_{t+1}\|)$-approximate stationary point for $F(x)$. Therefore, if $\sigma\|x_t - x_{t+1}\| \le \varepsilon$ we can stop the algorithm because we have already found an $O(\varepsilon)$-approximate stationary point. If $\sigma\|x_t - x_{t+1}\| > \varepsilon$, then it must satisfy that $F(x_t) - F(x_{t+1}) \ge \sigma\|x_t - x_{t+1}\|^2 \ge \Omega(\varepsilon^2/\sigma)$, but this cannot happen for more than $T = O\big(\frac{\sigma}{\varepsilon^2}(F(x_0) - F^*)$ stages. Therefore, the total gradient complexity is T multiplied with the complexity of accelerated SVRG in each stage (which is $\widetilde{O}(n + n^{3/4}\sqrt{L/\sigma})$ according to Theorem 2.3).

Remark 2.4. The complexity of repeatSVRG can be inferred from [1, 9], but is not explicitly stated. For instance, the paper [9] does not allow $F(x)$ to have a non-smooth proximal term $\psi(x)$, and applies accelerated gradient descent instead of accelerated SVRG.

3 Our Algorithms

We introduce two variants of our algorithms: (1) the base method `Natasha` targets on the simple regime when $f(x)$ and each $f_i(x)$ are both L-smooth, and (2) the full method `Natasha`$^{\mathsf{full}}$ targets on the more refined regime when $f(x)$ is L-smooth but each $f_i(x)$ is (ℓ_1, ℓ_2)-smooth.

Both methods follow the general idea of variance-reduced stochastic gradient descent: in each inner-most iteration, they compute a *gradient estimator* $\widetilde{\nabla}$ that is of the form $\widetilde{\nabla} = \nabla f(\widetilde{\mathbf{x}}) - \nabla f_i(\widetilde{\mathbf{x}}) + \nabla f_i(x)$ and satisfies $\mathbb{E}_{i \in_R [n]}[\widetilde{\nabla}] = \nabla f(x)$. Here, $\widetilde{\mathbf{x}}$ is a *snapshot* point that is changed once every n iterations (i.e., for each different $k = 1, 2, \dots, T'$ in the pseudocode), and we call it a *full epoch* for every distinct k. Notice that the amortized gradient complexity for computing $\widetilde{\nabla}$ is $O(1)$ per-iteration.

Base Method. In `Natasha` (see Algorithm 1), as illustrated by Figure 2, we divide each full epoch into p sub-epochs $s = 0, 1, \dots, p-1$, each of length $m = n/p$. In each sub-epoch s, we start with a point $x_0 = \widehat{\mathsf{x}}$, and replace $f(x)$ with its regularized version $f^s(x) \stackrel{\text{def}}{=} f(x) + \sigma\|x - \widehat{\mathsf{x}}\|^2$. Then, in each iteration t of the sub-epoch s, we

- compute gradient estimator $\widetilde{\nabla}$ with respect to $f^s(x_t)$, and
- perform update $x_{t+1} = \arg\min_y \big\{\psi(y) + \langle\widetilde{\nabla}, y\rangle + \frac{1}{2\alpha}\|y - x_t\|^2\big\}$ with learning rate α.

Effectively, the introduction of the regularizer $\sigma\|x - \widehat{\mathsf{x}}\|^2$ makes sure that when performing update $x_t \leftarrow x_{t+1}$, we also move a bit towards point $\widehat{\mathsf{x}}$ (i.e., retraction by regularization). Finally, when the sub-epoch is done, we define $\widehat{\mathsf{x}}$ to be a random one from $\{x_0, \dots, x_{m-1}\}$.

Full Method. In `Natasha`$^{\mathsf{full}}$, we also divide each full epoch into p sub-epochs. In each sub-epoch s, we start with a point $x_0 = z_0 = \widehat{\mathsf{x}}$ and define $f^s(x) \stackrel{\text{def}}{=} f(x) + \sigma\|x - \widehat{\mathsf{x}}\|^2$. However, this time in each iteration t, we

- compute gradient estimator $\widetilde{\nabla}$ with respect to $f^s(x_t)$,
- perform update $z_{t+1} = \arg\min_y \big\{\psi(y) + \langle\widetilde{\nabla}, y\rangle + \frac{1}{2\alpha}\|y - z_t\|^2\big\}$ with learning rate α, and
- choose $x_{t+1} = \frac{1}{2}z_{t+1} + \frac{1}{2}\widehat{\mathsf{x}}$.

Effectively, the regularizer $\sigma\|x - \widehat{\mathsf{x}}\|^2$ makes sure that when performing updates, we move a bit towards point $\widehat{\mathsf{x}}$ (i.e., retraction by regularization); at the same time, the choice $x_{t+1} = \frac{1}{2}z_{t+1} + \frac{1}{2}\widehat{\mathsf{x}}$ also helps us move towards point $\widehat{\mathsf{x}}$ (i.e., retraction by the so-called "Katyusha momentum"[10]). Finally, when the sub-epoch is over, we define $\widehat{\mathsf{x}}$ to be a random one from the set $\{x_0, \dots, x_{m-1}\}$, and move to the next sub-epoch.

[10]The idea for this second kind of retraction, and the idea of having the updates on a sequence z_t but computing gradients at points x_t, is largely motivated by our recent work on the Katyusha momentum and the Katyusha acceleration [2].

Algorithm 1 $\texttt{Natasha}(x^{\varnothing}, p, T', \alpha)$

Input: starting vector $x^{\varnothing}$, sub-epoch count $p \in [n]$, epoch count T', learning rate $\alpha > 0$.
Output: vector x^{out}.

$\widehat{\mathsf{x}} \leftarrow x^{\varnothing}$; $m \leftarrow n/p$; $X \leftarrow []$;
for $k \leftarrow 1$ **to** T' **do** ◇ *T' full epochs*
  $\widetilde{\mathbf{x}} \leftarrow \widehat{\mathsf{x}}$; $\mu \leftarrow \nabla f(\widetilde{\mathbf{x}})$;
  for $s \leftarrow 0$ **to** $p-1$ **do** ◇ *p sub-epochs in each epoch*
    $x_0 \leftarrow \widehat{\mathsf{x}}$; $X \leftarrow [X, \widehat{\mathsf{x}}]$;
    for $t \leftarrow 0$ **to** $m-1$ **do** ◇ *m iterations in each sub-epoch*
      $i \leftarrow$ a random choice from $\{1, \cdots, n\}$.
      $\widetilde{\nabla} \leftarrow \nabla f_i(x_t) - \nabla f_i(\widetilde{\mathbf{x}}) + \mu + 2\sigma(x_t - \widehat{\mathsf{x}})$ ◇ $\mathbb{E}_i[\widetilde{\nabla}] = \nabla\big(f(x) + \sigma\|x - \widehat{\mathsf{x}}\|^2\big)\big|_{x_t}$
      $x_{t+1} = \arg\min_{y \in \mathbb{R}^d} \big\{\psi(y) + \frac{1}{2\alpha}\|y - x_t\|^2 + \langle \widetilde{\nabla}, y \rangle\big\}$
    end for
    $\widehat{\mathsf{x}} \leftarrow$ a random choice from $\{x_0, x_1, \dots, x_{m-1}\}$; ◇ *for practitioners, choose the average*
  end for
end for
$\widehat{\mathsf{x}} \leftarrow$ a random vector in X; ◇ *for practitioners, choose the last*
$x^{\mathsf{out}} \leftarrow$ an approximate minimizer of $G(y) \stackrel{\text{def}}{=} F(y) + \sigma\|y - \widehat{\mathsf{x}}\|^2$ using SVRG.
return x^{out}. ◇ *it suffices to run SVRG for $O(n \log \frac{1}{\varepsilon})$ iterations.*

Algorithm 2 $\texttt{Natasha}^{\mathsf{full}}(x^{\varnothing}, p, T', \alpha)$

Input: starting vector $x^{\varnothing}$, sub-epoch count $p \in [n]$, epoch count T', learning rate $\alpha > 0$.
Output: vector x^{out}.

$\widehat{\mathsf{x}} \leftarrow x^{\varnothing}$; $m \leftarrow n/p$; $X \leftarrow []$;
for $k \leftarrow 1$ **to** T' **do** ◇ *T' full epochs*
  $\widetilde{\mathbf{x}} \leftarrow \widehat{\mathsf{x}}$; $\mu \leftarrow \nabla f(\widetilde{\mathbf{x}})$;
  for $s \leftarrow 0$ **to** $p-1$ **do** ◇ *p sub-epochs in each epoch*
    $z_0 \leftarrow \widehat{\mathsf{x}}$; $x_0 \leftarrow \widehat{\mathsf{x}}$; $X \leftarrow [X, \widehat{\mathsf{x}}]$;
    for $t \leftarrow 0$ **to** $m-1$ **do** ◇ *m iterations in each sub-epoch*
      $i \leftarrow$ a random choice from $\{1, \cdots, n\}$;
      $\widetilde{\nabla} \leftarrow \nabla f_i(x_t) - \nabla f_i(\widetilde{\mathbf{x}}) + \mu + 2\sigma(x_t - \widehat{\mathsf{x}})$; ◇ $\mathbb{E}_i[\widetilde{\nabla}] = \nabla\big(f(x) + \sigma\|x - \widehat{\mathsf{x}}\|^2\big)\big|_{x_t}$
      $z_{t+1} = \arg\min_{y \in \mathbb{R}^d} \big\{\psi(y) + \frac{1}{2\alpha}\|y - z_t\|^2 + \langle \widetilde{\nabla}, y \rangle\big\}$;
      $x_{t+1} = \frac{1}{2} z_{t+1} + \frac{1}{2}\widehat{\mathsf{x}}$; ◇ *Katyusha momentum $x_{t+1} = (1-\beta) z_{t+1} + \beta\widehat{\mathsf{x}}$*
      ◇ *theory predicts $\beta = \Theta\big(\frac{\sigma(\ell_1 + \ell_2)}{\ell_1 \ell_2}\big)$ gives the best performance*
      ◇ *$\beta = 1/2$ however leads to the simplest proof*
    end for
    $\widehat{\mathsf{x}} \leftarrow$ a random choice from $\{x_0, x_1, \dots, x_{m-1}\}$; ◇ *for practitioners, choose the average*
  end for
end for
$\widehat{\mathsf{x}} \leftarrow$ a random vector in X; ◇ *for practitioners, choose the last*
$x^{\mathsf{out}} \leftarrow$ an approximate minimizer of $G(y) \stackrel{\text{def}}{=} F(y) + \sigma\|y - \widehat{\mathsf{x}}\|^2$ using SVRG.
return x^{out}. ◇ *it suffices to run SVRG for $O(n \log \frac{1}{\varepsilon})$ iterations.*

4 A Sufficient Stopping Criterion

In this section, we present a sufficient condition for finding approximate stationary points in a σ-strongly nonconvex function. Lemma 4.1 below states that, if we regularize the original function and define $G(x) \stackrel{\mathrm{def}}{=} F(x) + \sigma\|x - \widehat{\mathsf{x}}\|^2$ for an arbitrary point $\widehat{\mathsf{x}}$, then the minimizer of $G(x)$ is an approximate saddle-point for $F(x)$.

Lemma 4.1. *Suppose $G(y) = F(y) + \sigma\|y - \widehat{\mathsf{x}}\|^2$ for some given point $\widehat{\mathsf{x}}$, and let x^* be the minimizer of $G(y)$. If we minimize $G(y)$ and obtain a point x satisfying*

$$G(x) - G(x^*) \le \delta^2\sigma \ ,$$

then for every $\eta \in \big(0, \frac{1}{\max\{L, 4\sigma\}}\big]$ we have the gradient mapping

$$\|\mathcal{G}_\eta(x)\|^2 \le 12\sigma^2\|x^* - \widehat{\mathsf{x}}\|^2 + O\big(\delta^2\big) \ .$$

Notice that when $\psi(x) \equiv 0$ this lemma is trivial, and can be found for instance in [9]. The main technical difficulty arises in order to deal with $\psi(x) \neq 0$.

Proof of Lemma 4.1. Let x^* be the (unique) minimizer of $G(y)$. Define auxiliary functions:

$$\Phi(y) \stackrel{\mathrm{def}}{=} \psi(y) + \frac{1}{2\eta}\|y - x\|^2 + \langle \nabla f(x), y - x\rangle - \psi(x) \quad \text{and} \quad \overline{\Phi}(y) \stackrel{\mathrm{def}}{=} \Phi(y) + \sigma\|y - \widehat{\mathsf{x}}\|^2 - \sigma\|x - \widehat{\mathsf{x}}\|^2$$

and letting $z = \arg\min_y \Phi(y)$ and $\overline{z} = \arg\min_y \overline{\Phi}(y)$. Observe that

- $\Phi(\cdot)$ is $\frac{1}{\eta}$-strongly convex so $-\Phi(z) = \Phi(x) - \Phi(z) \ge \frac{1}{2\eta}\|z - x\|^2$;
- $\overline{\Phi}(\cdot)$ is $\frac{1}{\eta}$-strongly convex so $\overline{\Phi}(z) \ge \overline{\Phi}(\overline{z}) + \frac{1}{2\eta}\|z - \overline{z}\|^2$;
- $\overline{\Phi}(\overline{z}) \ge G(\overline{z}) - G(x) \ge G(x^*) - G(x) \ge -\delta^2\sigma$ (because $\eta \le 1/L$ and $f(\cdot)$ is L-smooth).

Summing the three inequalities up we have

$$\sigma\|z - \widehat{\mathsf{x}}\|^2 - \sigma\|x - \widehat{\mathsf{x}}\|^2 \ge -\delta^2\sigma + \frac{1}{2\eta}\|z - x\|^2 + \frac{1}{2\eta}\|z - \overline{z}\|^2 \ .$$

Since we have inequality $\|z - \widehat{\mathsf{x}}\|^2 = \|(z - \overline{z}) + (\overline{z} - \widehat{\mathsf{x}})\|^2 \le (1 + 1/\beta)\|(z - \overline{z})\|^2 + (1 + \beta)\|(\overline{z} - \widehat{\mathsf{x}})\|^2$ for any $\beta > 0$, we can choose $\beta = 4\eta\sigma$ and obtain

$$(\sigma + \frac{1}{4\eta})\|(z - \overline{z})\|^2 + (\sigma + 4\eta\sigma^2)\|(\overline{z} - \widehat{\mathsf{x}})\|^2 - \sigma\|x - \widehat{\mathsf{x}}\|^2 \ge -\delta^2\sigma + \frac{1}{2\eta}\|z - x\|^2 + \frac{1}{2\eta}\|z - \overline{z}\|^2$$
$$\Longrightarrow \ (\sigma + 4\eta\sigma^2)\|\overline{z} - \widehat{\mathsf{x}}\|^2 - \sigma\|x - \widehat{\mathsf{x}}\|^2 \ge -\delta^2\sigma + \frac{1}{2\eta}\|z - x\|^2 \tag{4.1}$$

where the implication uses the fact that $\frac{1}{4\eta} \ge \sigma$. At this point, notice that:

- We have $\|x - x^*\|^2 \le \frac{2}{\sigma}(G(x) - G(x^*)) \le 2\delta^2$ by the strong convexity of $G(\cdot)$, and thus
$$-\sigma\|x - \widehat{\mathsf{x}}\|^2 \le -(\sigma - \eta\sigma^2)\|x^* - \widehat{\mathsf{x}}\|^2 + O(\delta^2/\eta) \ .$$
- We have $\|\overline{z} - x^*\|^2 \le \frac{2}{\sigma}(G(\overline{z}) - G(x^*)) \le 2\delta^2$ because $G(\overline{z}) \le G(x)$, and thus
$$(\sigma + 4\eta\sigma^2)\|\overline{z} - \widehat{\mathsf{x}}\|^2 \le (\sigma + 5\eta\sigma^2)\|x^* - \widehat{\mathsf{x}}\|^2 + O(\delta^2/\eta) \ .$$

Plugging them into (4.1), we have

$$(\sigma + 5\eta\sigma^2)\|x^* - \widehat{\mathsf{x}}\|^2 - (\sigma - \eta\sigma^2)\|x^* - \widehat{\mathsf{x}}\|^2 \ge \frac{1}{2\eta}\|z - x\|^2 - O\big(\delta^2/\eta\big)$$

and rearranging it we have

$$\|\mathcal{G}_\eta(x)\|^2 = \frac{1}{\eta^2}\|x - z\|^2 \le 12\sigma^2\|x^* - \widehat{\mathsf{x}}\|^2 + O\big(\delta^2\big) \ . \qquad \square$$

5 Base Method: Analysis for One Full Epoch

In this section, we consider problem (1.1) where each $f_i(x)$ is L-smooth and $F(x)$ is σ-approximate-convex. We use our base method `Natasha` to minimize $F(x)$, and analyze its behavior for one full epoch in this section. We assume $\sigma \le L$ without loss of generality, because any L-smooth function is also L-strongly nonconvex.

Notations. We introduce the following notations for analysis purpose only.

- Let $\widehat{\mathsf{x}}^s$ be the vector $\widehat{\mathsf{x}}$ at the beginning of sub-epoch s.
- Let x_t^s be the vector x_t in sub-epoch s.
- Let i_t^s be the index $i \in [n]$ in sub-epoch s at iteration t.
- Let $f^s(x) \stackrel{\text{def}}{=} f(x) + \sigma\|x - \widehat{\mathsf{x}}^s\|^2$, $F^s(x) \stackrel{\text{def}}{=} F(x) + \sigma\|x - \widehat{\mathsf{x}}^s\|^2$, and $x_*^s \stackrel{\text{def}}{=} \arg\min_x\{F^s(x)\}$.
- Let $\widetilde{\nabla} f^s(x_t^s) \stackrel{\text{def}}{=} \nabla f_i(x_t^s) - \nabla f_i(\widetilde{\mathbf{x}}) + \nabla f(\widetilde{\mathbf{x}}) + 2\sigma(x_t - \widehat{\mathsf{x}})$ where $i = i_t^s$.
- Let $\widetilde{\nabla} f(x_t^s) \stackrel{\text{def}}{=} \nabla f_i(x_t^s) - \nabla f_i(\widetilde{\mathbf{x}}) + \nabla f(\widetilde{\mathbf{x}})$ where $i = i_t^s$.

We obviously have that $f^s(x)$ and $F^s(x)$ are σ-strongly convex, and $f^s(x)$ is $(L + 2\sigma)$-smooth.

5.1 Variance Upper Bound

The following lemma gives an upper bound on the variance of the gradient estimator $\widetilde{\nabla} f^s(x_t^s)$:

Lemma 5.1. *We have* $\mathbb{E}_{i_t^s}\big[\|\widetilde{\nabla} f^s(x_t^s) - \nabla f^s(x_t^s)\|^2\big] \le pL^2\|x_t^s - \widehat{\mathsf{x}}^s\|^2 + pL^2\sum_{k=0}^{s-1}\|\widehat{\mathsf{x}}^k - \widehat{\mathsf{x}}^{k+1}\|^2$.

Proof. We have

$$\begin{aligned}
&\mathbb{E}_{i_t^s}\big[\|\widetilde{\nabla} f^s(x_t^s) - \nabla f^s(x_t^s)\|^2\big] = \mathbb{E}_{i_t^s}\big[\|\widetilde{\nabla} f(x_t^s) - \nabla f(x_t^s)\|^2\big] \\
&= \mathbb{E}_{i\in_R[n]}\Big[\big\|\big(\nabla f_i(x_t^s) - \nabla f_i(\widetilde{\mathbf{x}})\big) - \big(\nabla f(x_t^s) - \nabla f(\widetilde{\mathbf{x}})\big)\big\|^2\Big] \\
&\overset{①}{\le} \mathbb{E}_{i\in_R[n]}\Big[\big\|\nabla f_i(x_t^s) - \nabla f_i(\widetilde{\mathbf{x}})\big\|^2\Big] \\
&\overset{②}{\le} p\mathbb{E}_{i\in_R[n]}\Big[\big\|\nabla f_i(x_t^s) - \nabla f_i(\widehat{\mathsf{x}}^s)\big\|^2\Big] + p\sum_{k=0}^{s-1}\mathbb{E}_{i\in_R[n]}\Big[\big\|\nabla f_i(\widehat{\mathsf{x}}^k) - \nabla f_i(\widehat{\mathsf{x}}^{k+1})\big\|^2\Big] \\
&\overset{③}{\le} pL^2\|x_t^s - \widehat{\mathsf{x}}^s\|^2 + pL^2\sum_{k=0}^{s-1}\|\widehat{\mathsf{x}}^k - \widehat{\mathsf{x}}^{k+1}\|^2 \ .
\end{aligned}$$

Above, inequality ① is because for any random vector $\zeta \in \mathbb{R}^d$, it holds that $\mathbb{E}\|\zeta - \mathbb{E}\zeta\|^2 = \mathbb{E}\|\zeta\|^2 - \|\mathbb{E}\zeta\|^2$; inequality ② is because $\widehat{\mathsf{x}}^0 = \widetilde{\mathbf{x}}$ and for any p vectors $a_1, a_2, \dots, a_p \in \mathbb{R}^d$, it holds that $\|a_1 + \cdots + a_p\|^2 \le p\|a_1\|^2 + \cdots + p\|a_p\|^2$; and inequality ③ is because each $f_i(\cdot)$ is L-smooth. $\square$

5.2 Analysis for One Sub-Epoch

The following inequality is classically known as the "regret inequality" for mirror descent [7], and its proof is classical:

Fact 5.2. $\langle \widetilde{\nabla} f^s(x_t^s), x_{t+1}^s - u\rangle + \psi(x_{t+1}^s) - \psi(u) \le \frac{\|x_t^s - u\|^2}{2\alpha} - \frac{\|x_{t+1}^s - u\|^2}{2\alpha} - \frac{\|x_{t+1}^s - x_t^s\|^2}{2\alpha}$ *for every* $u \in \mathbb{R}^d$.

Proof. Recall that the minimality of $x_{t+1}^s = \arg\min_{y\in\mathbb{R}^d}\{\frac{1}{2\alpha}\|y-x_t^s\|^2+\psi(y)+\langle\widetilde{\nabla} f^s(x_t^s), y\rangle\}$ implies the existence of some subgradient $g \in \partial\psi(x_{t+1}^s)$ which satisfies $\frac{1}{\alpha}(x_{t+1}^s - x_t^s) + \widetilde{\nabla} f^s(x_t^s) + g = 0$. Combining this with $\psi(u) - \psi(x_{t+1}^s) \geq \langle g, u - x_{t+1}^s\rangle$, which is due to the convexity of $\psi(\cdot)$, we immediately have $\psi(u)-\psi(x_{t+1}^s)+\langle\frac{1}{\alpha}(x_{t+1}^s-x_t^s)+\widetilde{\nabla} f^s(x_t^s), u-x_{t+1}^s\rangle \geq \langle\frac{1}{\alpha}(x_{t+1}^s-x_t^s)+\widetilde{\nabla} f^s(x_t^s)+ g, u-x_{t+1}^s\rangle = 0$. Rearranging this inequality we have

$$\begin{aligned}\langle\widetilde{\nabla} f^s(x_t^s), x_{t+1}^s - u\rangle + \psi(x_{t+1}^s) - \psi(u) &\leq \langle -\frac{1}{\alpha}(x_{t+1}^s - x_t^s), x_{t+1}^s - u\rangle \\ &= \frac{\|x_t^s - u\|^2}{2\alpha} - \frac{\|x_{t+1}^s - u\|^2}{2\alpha} - \frac{\|x_{t+1}^s - x_t^s\|^2}{2\alpha} \ . \qquad \square\end{aligned}$$

The following lemma is our main contribution for the base method `Natasha`.

Lemma 5.3. *As long as* $\alpha \leq \frac{1}{2L+4\sigma}$*, we have*

$$\mathbb{E}\Big[\big(F^s(\widehat{\mathsf{x}}^{s+1}) - F^s(x_*^s)\big)\Big] \leq \mathbb{E}\Big[\frac{F^s(\widehat{\mathsf{x}}^s) - F^s(x_*^s)}{\sigma\alpha m/2} + \alpha p L^2\Big(\sum_{k=0}^{s}\|\widehat{\mathsf{x}}^k - \widehat{\mathsf{x}}^{k+1}\|^2\Big)\Big] \ .$$

Proof. We first compute that

$$\begin{aligned}&F^s(x_{t+1}^s) - F^s(u) = f^s(x_{t+1}^s) - f^s(u) + \psi(x_{t+1}^s) - \psi(u) \\ &\overset{①}{\leq} f^s(x_t^s) + \langle\nabla f^s(x_t^s), x_{t+1}^s - x_t^s\rangle + \frac{L+2\sigma}{2}\|x_t^s - x_{t+1}^s\|^2 - f^s(u) + \psi(x_{t+1}^s) - \psi(u) \\ &\overset{②}{\leq} \langle\nabla f^s(x_t^s), x_{t+1}^s - x_t^s\rangle + \frac{L+2\sigma}{2}\|x_t^s - x_{t+1}^s\|^2 + \langle\nabla f^s(x_t^s), x_t^s - u\rangle + \psi(x_{t+1}^s) - \psi(u) \ . \qquad (5.1)\end{aligned}$$

Above, inequality ① uses the fact that $f^s(\cdot)$ is $(L+2\sigma)$-smooth; and inequality ② uses the convexity of $f^s(\cdot)$. Now, we take expectation with respect to i_t^s on both sides of (5.1), and derive that:

$$\begin{aligned}&\mathbb{E}_{i_t^s}\big[F^s(x_{t+1}^s)\big] - F^s(u) \\ &\overset{①}{\leq} \mathbb{E}_{i_t^s}\Big[\langle\widetilde{\nabla} f^s(x_t^s) - \nabla f^s(x_t^s), x_t^s - x_{t+1}^s\rangle + \langle\widetilde{\nabla} f^s(x_t^s), x_{t+1}^s - u\rangle + \frac{L+2\sigma}{2}\|x_t^s - x_{t+1}^s\|^2 + \psi(x_{t+1}^s) - \psi(u)\Big] \\ &\overset{②}{\leq} \mathbb{E}_{i_t^s}\Big[\langle\widetilde{\nabla} f^s(x_t^s) - \nabla f^s(x_t^s), x_t^s - x_{t+1}^s\rangle + \frac{\|x_t^s - u\|^2}{2\alpha} - \frac{\|x_{t+1}^s - u\|^2}{2\alpha} - \big(\frac{1}{2\alpha} - \frac{L+2\sigma}{2}\big)\|x_{t+1}^s - x_t^s\|^2\Big] \\ &\overset{③}{\leq} \mathbb{E}_{i_t^s}\Big[\alpha\big\|\widetilde{\nabla} f^s(x_t^s) - \nabla f^s(x_t^s)\big\|^2 + \frac{\|x_t^s - u\|^2}{2\alpha} - \frac{\|x_{t+1}^s - u\|^2}{2\alpha}\Big] \\ &\overset{④}{\leq} \mathbb{E}_{i_t^s}\Big[\alpha p L^2\|x_t^s - \widehat{\mathsf{x}}^s\|^2 + \alpha p L^2\sum_{k=0}^{s-1}\|\widehat{\mathsf{x}}^k - \widehat{\mathsf{x}}^{k+1}\|^2 + \frac{\|x_t^s - u\|^2}{2\alpha} - \frac{\|x_{t+1}^s - u\|^2}{2\alpha}\Big] \ . \qquad (5.2)\end{aligned}$$

Above, inequality ① is follows from (5.1) together with the fact that $\mathbb{E}_{i_t^s}[\widetilde{\nabla} f^s(x_t^s)] = \nabla f^s(x_t^s)$ implies

$$\begin{aligned}&\mathbb{E}_{i_t^s}\big[\langle\nabla f^s(x_t^s), x_{t+1}^s - x_t^s\rangle + \langle\nabla f^s(x_t^s), x_t^s - u\rangle\big] \\ &\qquad = \mathbb{E}_{i_t^s}\big[\langle\widetilde{\nabla} f^s(x_t^s) - \nabla f^s(x_t^s), x_t^s - x_{t+1}^s\rangle + \langle\widetilde{\nabla} f^s(x_t^s), x_{t+1}^s - u\rangle\big] \ ;\end{aligned}$$

inequality ② uses Fact 5.2; inequality ③ uses $\alpha \leq \frac{1}{2L+4\sigma}$ together with Young's inequality $\langle a, b\rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$; and inequality ④ uses Lemma 5.1.

Finally, choosing $u = x_*^s$ to be the (unique) minimizer of $F^s(\cdot) = f^s(\cdot) + \psi(\cdot)$, and telescoping inequality (5.2) for $t = 0, 1, \dots, m-1$, we have

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{t=1}^{m-1}\big(F^s(x_t^s) - F^s(x_*^s)\big)\Big] \\
&\leq \mathbb{E}\Big[\frac{\|x_0^s - x_*^s\|^2}{2\alpha} + \sum_{t=0}^{m-1}\Big(\alpha p L^2 \|x_t^s - \widehat{\mathsf{x}}^s\|^2 + \alpha p L^2 \sum_{k=0}^{s-1}\|\widehat{\mathsf{x}}^k - \widehat{\mathsf{x}}^{k+1}\|^2\Big)\Big] \\
&\leq \mathbb{E}\Big[\frac{F^s(\widehat{\mathsf{x}}^s) - F^s(x_*^s)}{\sigma\alpha} + \alpha p m L^2 \Big(\sum_{k=0}^{s}\|\widehat{\mathsf{x}}^k - \widehat{\mathsf{x}}^{k+1}\|^2\Big)\Big] \enspace .
\end{aligned}
$$

Above, the second inequality uses the fact that $\widehat{\mathsf{x}}^{s+1}$ is chosen from $\{x_0^s, \dots, x_{m-1}^s\}$ uniformly at random, as well as the σ-strong convexity of $F^s(\cdot)$.

Dividing both sides by m and rearranging the terms (using $\frac{1}{2\sigma\alpha} \geq 1$), we have

$$
\mathbb{E}\Big[\big(F^s(\widehat{\mathsf{x}}^{s+1}) - F^s(x_*^s)\big)\Big] \leq \mathbb{E}\Big[\frac{F^s(\widehat{\mathsf{x}}^s) - F^s(x_*^s)}{\sigma\alpha m/2} + \alpha p L^2 \Big(\sum_{k=0}^{s}\|\widehat{\mathsf{x}}^k - \widehat{\mathsf{x}}^{k+1}\|^2\Big)\Big] \enspace . \qquad \square
$$

5.3 Analysis for One Full Epoch

One can telescope Lemma 5.3 for an entire epoch and arrive at the following lemma:

Lemma 5.4. *If $\alpha \leq \frac{1}{2L+4\sigma}$, $\alpha \geq \frac{4}{\sigma m}$ and $\alpha \leq \frac{\sigma}{p^2 L^2}$, we have*

$$
\sum_{s=0}^{p-1}\mathbb{E}\Big[\big(F^s(\widehat{\mathsf{x}}^s) - F^s(x_*^s)\big)\Big] \leq 2\mathbb{E}\Big[F(\widehat{\mathsf{x}}^0) - F(\widehat{\mathsf{x}}^p)\Big] \enspace .
$$

Proof. Telescoping Lemma 5.3 for all the subepochs $s = 0, 1, \dots, p-1$, we have

$$
\begin{aligned}
\sum_{s=0}^{p-1}\mathbb{E}\Big[\big(F^s(\widehat{\mathsf{x}}^{s+1}) - F^s(x_*^s)\big)\Big] &\leq \sum_{s=0}^{p-1}\mathbb{E}\Big[\frac{F^s(\widehat{\mathsf{x}}^s) - F^s(x_*^s)}{\sigma\alpha m/2} + \alpha p^2 L^2 \|\widehat{\mathsf{x}}^s - \widehat{\mathsf{x}}^{s+1}\|^2\Big] \\
&\overset{①}{\leq} \sum_{s=0}^{p-1}\mathbb{E}\Big[\frac{F^s(\widehat{\mathsf{x}}^s) - F^s(x_*^s)}{\sigma\alpha m/2} + \sigma \cdot \|\widehat{\mathsf{x}}^{s+1} - \widehat{\mathsf{x}}^s\|^2\Big] \\
&\overset{②}{=} \sum_{s=0}^{p-1}\mathbb{E}\Big[\frac{F^s(\widehat{\mathsf{x}}^s) - F^s(x_*^s)}{\sigma\alpha m/2} + \big(F^s(\widehat{\mathsf{x}}^{s+1}) - F^s(\widehat{\mathsf{x}}^s)\big) - \big(F(\widehat{\mathsf{x}}^{s+1}) - F(\widehat{\mathsf{x}}^s)\big)\Big]
\end{aligned}
$$

Above, ① uses $\alpha p^2 L^2 \leq \sigma$, and ② uses the definition $F^s(y) = F(y) + \sigma\|y - \widehat{\mathsf{x}}^s\|^2$. Finally, rearranging both sides, and using the fact that $\frac{1}{\sigma\alpha m} \leq \frac{1}{4}$, we have the desired inequality. $\square$

6 Base Method: Final Theorem

We are now ready to state and prove our main convergence theorem for `Natasha`:

Theorem 1. *Suppose in (1.1), each $f_i(x)$ is L-smooth and $F(x)$ is σ-approximate-convex for $\sigma \leq L$. Then, if $\frac{L^2}{\sigma^2} \leq n$, $p = \Theta\big((\frac{\sigma^2}{L^2}n)^{1/3}\big)$ and $\alpha = \Theta(\frac{\sigma}{p^2L^2})$, our base method* `Natasha` *outputs a point x^{out} satisfying*

$$\mathbb{E}[\|\mathcal{G}_\eta(x^{\mathsf{out}})\|^2] \leq O\Big(\tfrac{(L^2\sigma)^{1/3}n^{2/3}}{T'n}\Big) \cdot (F(x^{\varnothing}) - F^*) \ .$$

for every $\eta \in \big(0, \frac{1}{\max\{L,4\sigma\}}\big]$. In other words, to obtain $\mathbb{E}[\|\mathcal{G}_\eta(x^{\mathsf{out}})\|^2] \leq \varepsilon^2$, we need gradient complexity

$$O\Big(n \log\frac{1}{\varepsilon} + \frac{(L^2\sigma)^{1/3}n^{2/3}}{\varepsilon^2} \cdot (F(x^{\varnothing}) - F^*)\Big) \ .$$

In the above theorem, we have assumed $\sigma \leq L$ without loss of generality because any L-smooth function is also L-strongly nonconvex. Also, we have assumed $\frac{L^2}{\sigma^2} \leq n$ and if this inequality does not hold, then one should apply repeatSVRG for a faster running time (see Figure 1).

Proof of Theorem 1. We choose $p = \big(\frac{\sigma^2}{24L^2}n\big)^{1/3}$, $m = n/p$, and $\alpha = \frac{4}{\sigma m} = \frac{\sigma}{6p^2L^2} \leq \frac{1}{2L+4\sigma}$, so we can apply Lemma 5.4. If we telescope Lemma 5.4 for the entire algorithm (which has T' full epochs), and use the fact that $\widehat{\mathsf{x}}^p$ of the previous epoch equals $\widehat{\mathsf{x}}^0$ of the next epoch, we conclude that if we choose a random epoch and a random subepoch s, we will have

$$\mathbb{E}[F^s(\widehat{\mathsf{x}}^s) - F^s(x^s_*)] \leq \frac{2}{pT'}(F(x^{\varnothing}) - F^*) \ .$$

By the σ-strong convexity of $F^s(\cdot)$, we have $\mathbb{E}[\sigma\|\widehat{\mathsf{x}}^s - x^s_*\|^2] \leq \frac{4}{pT'}(F(x^{\varnothing}) - F^*)$.

Now, $F^s(x) = F(x) + \sigma\|x - \widehat{\mathsf{x}}^s\|^2$ satisfies the assumption of $G(x)$ in Lemma 4.1. If we use the SVRG method (see Theorem 2.3) to minimize the convex function $F^s(x)$, we get an output x^{out} satisfying $F^s(x^{\mathsf{out}}) - F^s(x^s_*) \leq \varepsilon^2\sigma$ in gradient complexity $O\big((n + \frac{L^2}{\sigma^2})\log\frac{1}{\varepsilon}\big) \leq O(n\log\frac{1}{\varepsilon})$.

We can therefore apply Lemma 4.1 and conclude that this output x^{out} satisfies

$$\mathbb{E}[\|\mathcal{G}_\eta(x^{\mathsf{out}})\|^2] \leq O\Big(\frac{\sigma}{pT'}\Big) \cdot (F(x^{\varnothing}) - F^*) = O\Big(\frac{(L^2\sigma)^{1/3}n^{2/3}}{T'n}\Big) \cdot (F(x^{\varnothing}) - F^*) \ .$$

In other words, we obtain $\mathbb{E}[\|\mathcal{G}_\eta(x^{\mathsf{out}})\|^2] \leq \varepsilon^2$ using

$$T'n = O\Big(n + \tfrac{(L^2\sigma)^{1/3}n^{2/3}}{\varepsilon^2} \cdot (F(x^{\varnothing}) - F^*)\Big)$$

computations of the stochastic gradients. Here, the additive term n is because $T' \geq 1$.

Finally, adding this with $O(n\log\frac{1}{\varepsilon})$, the gradient complexity for the application of SVRG in the last line of `Natasha`, we finish the proof of the total gradient complexity. □

7 Full Method: Analysis for One Full Epoch

In this section, we study a more refined version of problem (1.1), where $f(x)$ is L-smooth, each $f_i(x)$ is (ℓ_1, ℓ_2)-smooth, and $F(x)$ is σ-approximate-convex. As later argued in Remark 8.1, we can assume $\sigma \leq \min\{\ell_1, \ell_2, L\}$ almost without loss of generality.

We use our full method $\mathtt{Natasha}^{\mathsf{full}}$ to minimize $F(x)$, and analyze its behavior for one full epoch in this section. Note that parameter L is not needed in the specification of $\mathtt{Natasha}^{\mathsf{full}}$, but used only for analysis purpose.

Notations. We use the same notations as in Section 5, with an additional one highlighted here:

- Let $\widehat{\mathsf{x}}^s$ be the vector $\widehat{\mathsf{x}}$ at the beginning of sub-epoch s.

- Let x_t^s be the vector x_t in sub-epoch s.
- Let i_t^s be the index $i \in [n]$ in sub-epoch s at iteration t.
- Let $F^s(x) \stackrel{\text{def}}{=} F(x) + \sigma\|x - \widehat{\mathrm{x}}^s\|^2$ and $x_*^s \stackrel{\text{def}}{=} \arg\min_x\{F^s(x)\}$.
- Let $f^s(x) \stackrel{\text{def}}{=} f(x) + \sigma\|x - \widehat{\mathrm{x}}^s\|^2$ and $\boxed{f_i^s(x) \stackrel{\text{def}}{=} f_i(x) + \sigma\|x - \widehat{\mathrm{x}}^s\|^2}$.
- Let $\widetilde{\nabla} f^s(x_t^s) \stackrel{\text{def}}{=} \nabla f_i(x_t^s) - \nabla f_i(\widetilde{\mathbf{x}}) + \nabla f(\widetilde{\mathbf{x}}) + 2\sigma(x_t - \widehat{\mathrm{x}})$ where $i = i_t^s$.
- Let $\widetilde{\nabla} f(x_t^s) \stackrel{\text{def}}{=} \nabla f_i(x_t^s) - \nabla f_i(\widetilde{\mathbf{x}}) + \nabla f(\widetilde{\mathbf{x}})$ where $i = i_t^s$.

We obviously have that $f^s(x)$ and $F^s(x)$ are σ-strongly convex, and $f^s(x)$ is $(L + 2\sigma)$-smooth.

7.1 Variance Upper Bound

In this subsection we derive a new upper bound on the variance of the gradient estimator $\widetilde{\nabla}$. This bound will be tighter than Lemma 5.1, and will make use of the asymmetry between parameters ℓ_1 and ℓ_2. To achieve so, we first need to introduce the following lemma:

Lemma 7.1. *If $g(y) = \frac{1}{n}\sum_{i=1}^n g_i(y)$ is convex, and if each g_i is (ℓ_1, ℓ_2)-smooth, then we have*

$$\mathbb{E}_{i \in_R [n]}\big[\|\nabla g_i(y_1) - \nabla g_i(y_2)\|^2\big] \\ \leq 2(\ell_1 + \ell_2)(g(y_2) - g(y_1) - \langle \nabla g(y_1), y_2 - y_1 \rangle)\big] + 6\ell_1\ell_2\|y_2 - y_1\|^2 \enspace .$$

Proof. We consider two cases: $\ell_2 \leq \ell_1$ and $\ell_2 \geq \ell_1$.

- In the first case, we define $\phi_i(z) \stackrel{\text{def}}{=} g_i(z) - \langle \nabla g_i(y_1), z \rangle + \frac{\ell_2}{2}\|z - y_1\|^2$ for each $i \in [n]$. This function $\phi_i(z)$ is a convex, $(\ell_1 + \ell_2)$-smooth function that has a minimizer $z = y_1$ (which can be seen by taking the derivative). For this reason, we claim that

$$\forall z\colon \quad \phi_i(y_1) \leq \phi_i(z) - \frac{1}{\ell_1 + \ell_2}\|\nabla \phi_i(z)\|^2 \enspace , \tag{7.1}$$

  and this inequality is classical for smooth functions (see for instance Theorem 2.1.5 in textbook [17]). By expanding out the definition of $\phi_i(\cdot)$ in (7.1), we immediately have

$$g_i(y_1) - \langle \nabla g_i(y_1), y_1 \rangle \leq g_i(z) - \langle \nabla g_i(y_1), z \rangle + \frac{\ell_2}{2}\|z - y_1\|^2 \\ - \frac{1}{2(\ell_1 + \ell_2)}\|\nabla g_i(z) - \nabla g_i(y_1) + \ell_2(z - y_1)\|^2$$

  which then implies

$$\|\nabla g_i(z) - \nabla g_i(y_1)\|^2 \leq 2\|\nabla g_i(z) - \nabla g_i(y_1) + \ell_2(z - y_1)\|^2 + 2\|\ell_2(z - y_1)\|^2 \\ \leq 2(\ell_1 + \ell_2)(g_i(z) - g_i(y_1) - \langle \nabla g_i(y_1), z - y_1 \rangle) + (4\ell_2^2 + 2\ell_1\ell_2)\|z - y_1\|^2 \enspace . \tag{7.2}$$

  Now, by choosing $z = y_2$ and taking expectation with i in (7.2), we have

$$\mathbb{E}_i\big[\big\|\nabla g_i(y_2) - \nabla g_i(y_1)\big\|^2\big] \\ \leq 2(\ell_1 + \ell_2)\big(g(y_2) - g(y_1) - \langle \nabla g(y_1), y_2 - y_1 \rangle\big)\big) + (4\ell_2^2 + 2\ell_1\ell_2)\|y_2 - y_1\|^2 \tag{7.3}$$

- In the second case, we define $\phi_i(z) \stackrel{\text{def}}{=} -g_i(z) + \langle \nabla g_i(y_2), z \rangle + \frac{\ell_1}{2}\|z - y_2\|^2$ for each $i \in [n]$. It is clear that $\phi_i(z)$ is a convex, $(\ell_1 + \ell_2)$-smooth function that has a minimizer $z = y_2$ (which can be seen by taking the derivative). For this reason, we have

$$\forall z: \quad \phi_i(y_2) \le \phi_i(z) - \frac{1}{\ell_1 + \ell_2}\|\nabla \phi_i(z)\|^2 \ . \tag{7.4}$$

By expanding out the definition of $\phi_i(\cdot)$ in (7.4), we immediately have

$$\begin{aligned} - g_i(y_2) + \langle \nabla g_i(y_2), y_2 \rangle \le -g_i(z) + \langle \nabla g_i(y_2), z \rangle + \frac{\ell_1}{2}\|z - y_2\|^2 \\ - \frac{1}{2(\ell_1+\ell_2)}\|\nabla g_i(z) - \nabla g_i(y_2) - \ell_1(z - y_2)\|^2 \end{aligned}$$

which then implies that

$$\begin{aligned} \|\nabla g_i(z) - \nabla g_i(y_2)\|^2 &\le 2\|\nabla g_i(z) - \nabla g_i(y_2) - \ell_1(z-y_2)\|^2 + 2\|\ell_2(z-y_2)\|^2 \\ &\le 2(\ell_1+\ell_2)(g_i(y_2) - g_i(z) + \langle \nabla g_i(y_2), z - y_2\rangle) + (4\ell_1^2 + 2\ell_1\ell_2)\|z - y_2\|^2 \ . \end{aligned} \tag{7.5}$$

Now by choosing $z = y_1$ and taking expectation over i in (7.5), we have

$$\begin{aligned} &\mathbb{E}_i\big[\big\|\nabla g_i(y_1) - \nabla g_i(y_2)\big\|^2\big] \\ &\le 2(\ell_1+\ell_2)\big(g(y_2) - g(y_1) + \langle \nabla g(y_2), y_1 - y_2\rangle\big) + (4\ell_1^2 + 2\ell_1\ell_2)\|y_1 - y_2\|^2 \\ &\le (4\ell_1^2 + 2\ell_1\ell_2)\|y_1 - y_2\|^2 \\ &\le 2(\ell_1+\ell_2)\big(g(y_2) - g(y_1) - \langle \nabla g(y_1), y_2 - y_1\rangle)\big) + (4\ell_2^2 + 2\ell_1\ell_2)\|y_2 - y_1\|^2 \ . \end{aligned} \tag{7.6}$$

Above, the second and third inequalities use the convexity of $g(\cdot)$.

Combining (7.3) and (7.6) we finish the proof of the lemma. □

We are now ready to state our final variance upper bound:

Lemma 7.2 (variance bound)**.** *There exists constant $C \ge 1$ such that, if we define*

- $\Phi^s(y) \stackrel{\text{def}}{=} C(\ell_1+\ell_2)\cdot\big(f^s(\widehat{\mathsf{x}}^s) - f^s(y) - \langle \nabla f^s(y), \widehat{\mathsf{x}}^s - y\rangle\big)\big] + C(\ell_1\ell_2)\cdot\|y - \widehat{\mathsf{x}}^s\|^2 \ge 0;$
- $\Phi^s_t = \Phi^s(x^s_t)$ *and* $\Phi^s = \Phi^s(\widehat{\mathsf{x}}^{s+1})$,

then, we have $\mathbb{E}_i\big[\|\widetilde{\nabla} f^s(x^s_t) - \nabla f^s(x^s_t)\|^2\big] \le p\Phi^s_t + p\sum_{k=0}^{s-1}\Phi^k$ *where* $i = i^s_t$.

Before proceeding to the proof, we point out that if $\ell_1 = \ell_2 = L$ like in the base setting, then we shall have $\Phi^s(y) \le O(L^2)\|y - \widehat{\mathsf{x}}^s\|^2$ and Lemma 7.2 becomes identical to Lemma 5.1.

Proof. If we plug in $g = f^s$ and $g_i = f^s_i$ in Lemma 7.1, we have g_i is $(\ell_1 + 2\sigma, \ell_2 - 2\sigma)$-smooth and thus each g_i is also $(3\ell_1, \ell_2)$-smooth. Therefore, Lemma 7.1 implies there exists constant $C \ge 1$ such that

$$\begin{aligned} \mathbb{E}_i\big[\big\|\nabla f_i(y) - \nabla f_i(\widehat{\mathsf{x}}^s)\big\|^2\big] &\le 2\mathbb{E}_i\big[\big\|\nabla f^s_i(y) - \nabla f^s_i(\widehat{\mathsf{x}}^s)\big\|^2\big] + 2\big\|2\sigma(y - \widehat{\mathsf{x}}^s)\big\|^2 \\ &\le C(\ell_1+\ell_2)\cdot\big(f^s(\widehat{\mathsf{x}}^s) - f^s(y) - \langle \nabla f^s(y), \widehat{\mathsf{x}}^s - y\rangle\big)\big] + C(\ell_1\ell_2)\cdot\|y - \widehat{\mathsf{x}}^s\|^2 \\ &= \Phi^s(y) \ . \end{aligned} \tag{7.7}$$

Therefore, the variance term:

$$\begin{aligned}\mathbb{E}_i\big[\|\widetilde{\nabla} f^s(x_t^s)-\nabla f^s(x_t^s)\|^2\big] &= \mathbb{E}_i\big[\|\widetilde{\nabla} f(x_t^s)-\nabla f(x_t^s)\|^2\big] \\ &= \mathbb{E}_i\Big[\big\|\big(\nabla f_i(x_t^s)-\nabla f_i(\widetilde{\mathbf{x}})\big)-\big(\nabla f(x_t^s)-\nabla f(\widetilde{\mathbf{x}})\big)\big\|^2\Big] \\ &\overset{①}{\leq} \mathbb{E}_i\Big[\big\|\nabla f_i(x_t^s)-\nabla f_i(\widetilde{\mathbf{x}})\big\|^2\Big] \\ &\overset{②}{\leq} p\mathbb{E}_i\Big[\big\|\nabla f_i(x_t^s)-\nabla f_i(\widehat{\mathsf{x}}^s)\big\|^2\Big]+p\textstyle\sum_{k=0}^{s-1}\mathbb{E}_i\Big[\big\|\nabla f_i(\widehat{\mathsf{x}}^k)-\nabla f_i(\widehat{\mathsf{x}}^{k+1})\big\|^2\Big] \\ &\overset{③}{\leq} p\Phi_t^s+p\textstyle\sum_{k=0}^{s-1}\Phi^k \ . \end{aligned} \tag{7.8}$$

Above, inequality ① is because for any random vector $\zeta\in\mathbb{R}^d$, it holds that $\mathbb{E}\|\zeta-\mathbb{E}\zeta\|^2=\mathbb{E}\|\zeta\|^2-\|\mathbb{E}\zeta\|^2$; inequality ② is because $\widehat{\mathsf{x}}^0=\widetilde{\mathbf{x}}$ and for any p vectors $a_1,a_2,\dots,a_p\in\mathbb{R}^d$, it holds that $\|a_1+\cdots+a_p\|^2\leq p\|a_1\|^2+\cdots+p\|a_p\|^2$; and inequality ③ is from repeatedly applying (7.7). □

7.2 Analysis for One Sub-Epoch

The following fact is analogous to Fact 5.2, and the only difference is that in $\texttt{Natasha}^{\mathsf{full}}$ we are applying proximal updates on the $\{z_t^s\}_t$ sequence.

Fact 7.3. $\langle\widetilde{\nabla} f^s(x_t^s), z_{t+1}^s-u\rangle+\psi(z_{t+1}^s)-\psi(u)\leq\frac{\|z_t^s-u\|^2}{2\alpha}-\frac{\|z_{t+1}^s-u\|^2}{2\alpha}-\frac{\|z_{t+1}^s-z_t^s\|^2}{2\alpha}$ *for every* $u\in\mathbb{R}^d$.

Proof. Recall that the minimality of $z_{t+1}^s=\arg\min_{y\in\mathbb{R}^d}\{\frac{1}{2\alpha}\|y-z_t^s\|^2+\psi(y)+\langle\widetilde{\nabla} f^s(x_t^s),y\rangle\}$ implies the existence of some subgradient $g\in\partial\psi(z_{t+1}^s)$ which satisfies $\frac{1}{\alpha}(z_{t+1}^s-z_t^s)+\widetilde{\nabla} f^s(x_t^s)+g=0$. Combining this with $\psi(u)-\psi(z_{t+1}^s)\geq\langle g,u-z_{t+1}^s\rangle$, which is due to the convexity of $\psi(\cdot)$, we immediately have $\psi(u)-\psi(z_{t+1}^s)+\langle\frac{1}{\alpha}(z_{t+1}^s-z_t^s)+\widetilde{\nabla} f^s(x_t^s),u-z_{t+1}^s\rangle\geq\langle\frac{1}{\alpha}(z_{t+1}^s-z_t^s)+\widetilde{\nabla} f^s(x_t^s)+g,u-z_{t+1}^s\rangle=0$. Rearranging this inequality we have

$$\begin{aligned}\langle\widetilde{\nabla} f^s(x_t^s), z_{t+1}^s-u\rangle+\psi(z_{t+1}^s)-\psi(u) &\leq \langle-\frac{1}{\alpha}(z_{t+1}^s-z_t^s),z_{t+1}^s-u\rangle \\ &= \frac{\|z_t^s-u\|^2}{2\alpha}-\frac{\|z_{t+1}^s-u\|^2}{2\alpha}-\frac{\|z_{t+1}^s-z_t^s\|^2}{2\alpha} \ .\end{aligned}$$ □

The following lemma is our technical main contribution for the full method $\texttt{Natasha}^{\mathsf{full}}$.

Lemma 7.4. *If* $\alpha\leq\frac{1}{L+2\sigma}$, *then we have the following inequality for sub-epoch* s:

$$\begin{aligned}&\mathbb{E}\Big[\big(F^s(\widehat{\mathsf{x}}^{s+1})-F^s(x_*^s)\big)\Big] \\ &\leq\mathbb{E}\Big[\frac{F^s(\widehat{\mathsf{x}}^s)-F^s(x_*^s)}{\sigma\alpha m/2}+\alpha p\Big(\sum_{k=0}^{s}\Phi^k\Big)+\langle\nabla f^s(\widehat{\mathsf{x}}^{s+1}),\widehat{\mathsf{x}}^s-\widehat{\mathsf{x}}^{s+1}\rangle+\big(\psi(\widehat{\mathsf{x}}^s)-\psi(\widehat{\mathsf{x}}^{s+1})\big)\Big] \ .\end{aligned}$$

Proof. We first compute that

$$\begin{aligned}&2F^s(x_{t+1}^s)-F^s(x_t^s)-F^s(u)=2f^s(x_{t+1}^s)-f^s(x_t^s)-f^s(u)+2\psi(x_{t+1}^s)-\psi(x_t^s)-\psi(u) \\ &\overset{①}{\leq} f^s(x_t^s)+2\langle\nabla f^s(x_t^s),x_{t+1}^s-x_t^s\rangle+(L+2\sigma)\|x_t^s-x_{t+1}^s\|^2-f^s(u)+2\psi(x_{t+1}^s)-\psi(x_t^s)-\psi(u) \\ &\overset{②}{=} f^s(x_t^s)+\langle\nabla f^s(x_t^s),z_{t+1}^s-z_t^s\rangle+\frac{L+2\sigma}{4}\|z_t^s-z_{t+1}^s\|^2-f^s(u)+2\psi(x_{t+1}^s)-\psi(x_t^s)-\psi(u) \\ &\overset{③}{\leq} \langle\nabla f^s(x_t^s),z_{t+1}^s-z_t^s\rangle+\frac{L+2\sigma}{4}\|z_t^s-z_{t+1}^s\|^2+\langle\nabla f^s(x_t^s),x_t^s-u\rangle+\psi(z_{t+1}^s)+\psi(\widehat{\mathsf{x}}^s)-\psi(x_t^s)-\psi(u)\end{aligned} \tag{7.9}$$

Above, inequality ① uses the fact that $f^s(\cdot)$ is $(L+2\sigma)$-smooth; equality ② uses the fact that $z^s_{t+1} - z^s_t = 2(x^s_{t+1} - x^s_t)$; inequality ③ uses the convexity of $f^s(\cdot)$, the convexity of $\psi(\cdot)$, and the fact $x^s_{t+1} = \frac{1}{2}(z^s_{t+1} + \widehat{\mathsf{x}}^s)$.

Now, we take expectation with respect to i^s_t on both sides of (7.9), and derive that:

$$
\begin{aligned}
&2\mathbb{E}_{i^s_t}\big[F^s(x^s_{t+1})\big] - F^s(x^s_t) - F^s(u) \\
&\overset{①}{\leq} \mathbb{E}_{i^s_t}\Big[\langle \widetilde{\nabla} f^s(x^s_t) - \nabla f^s(x^s_t), z^s_t - z^s_{t+1}\rangle + \langle \widetilde{\nabla} f^s(x^s_t), z^s_{t+1} - u\rangle + \frac{L+2\sigma}{4}\|z^s_t - z^s_{t+1}\|^2 + \psi(z^s_{t+1}) - \psi(u)\Big] \\
&\qquad + \langle \nabla f^s(x^s_t), x^s_t - z^s_t\rangle + \psi(\widehat{\mathsf{x}}^s) - \psi(x^s_t) \\
&\overset{②}{\leq} \mathbb{E}_{i^s_t}\Big[\langle \widetilde{\nabla} f^s(x^s_t) - \nabla f^s(x^s_t), z^s_t - z^s_{t+1}\rangle + \frac{\|z^s_t - u\|^2}{2\alpha} - \frac{\|z^s_{t+1} - u\|^2}{2\alpha} - \big(\frac{1}{2\alpha} - \frac{L+2\sigma}{4}\big)\|z^s_{t+1} - z^s_t\|^2\Big] \\
&\qquad + \langle \nabla f^s(x^s_t), x^s_t - z^s_t\rangle + \psi(\widehat{\mathsf{x}}^s) - \psi(x^s_t) \\
&\overset{③}{\leq} \mathbb{E}_{i^s_t}\Big[\alpha\big\|\widetilde{\nabla} f^s(x^s_t) - \nabla f^s(x^s_t)\big\|^2 + \frac{\|z^s_t - u\|^2}{2\alpha} - \frac{\|z^s_{t+1} - u\|^2}{2\alpha}\Big] + \langle \nabla f^s(x^s_t), x^s_t - z^s_t\rangle + \psi(\widehat{\mathsf{x}}^s) - \psi(x^s_t) \\
&\overset{④}{\leq} \mathbb{E}_{i^s_t}\Big[\alpha p \Phi^s_t + \alpha p \sum_{k=0}^{s-1}\Phi^k + \frac{\|z^s_t - u\|^2}{2\alpha} - \frac{\|z^s_{t+1} - u\|^2}{2\alpha}\Big] + \langle \nabla f^s(x^s_t), \widehat{\mathsf{x}}^s - x^s_t\rangle + \psi(\widehat{\mathsf{x}}^s) - \psi(x^s_t) \enspace .
\end{aligned}
$$

Above, inequality ① is from (7.9) together with the fact that $\mathbb{E}_{i^s_t}[\widetilde{\nabla} f^s(x^s_t)] = \nabla f^s(x^s_t)$ implies

$$
\begin{aligned}
&\mathbb{E}_{i^s_t}\big[\langle \nabla f^s(x^s_t), z^s_{t+1} - z^s_t\rangle + \langle \nabla f^s(x^s_t), x^s_t - u\rangle\big] \\
&\qquad = \mathbb{E}_{i^s_t}\big[\langle \nabla f^s(x^s_t), x^s_t - z^s_t\rangle + \langle \widetilde{\nabla} f^s(x^s_t) - \nabla f^s(x^s_t), z^s_t - z^s_{t+1}\rangle + \langle \widetilde{\nabla} f^s(x^s_t), z^s_{t+1} - u\rangle\big] \enspace ;
\end{aligned}
$$

inequality ② uses Fact 7.3; inequality ③ uses $\alpha \leq \frac{1}{L+2\sigma}$ together with Young's inequality $\langle a, b\rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$; and inequality ④ uses Lemma 7.2 and the fact that $x^s_t = \frac{1}{2}z^s_t + \frac{1}{2}\widehat{\mathsf{x}}^s$.

Finally, choosing $u = x^s_*$ to be the (unique) minimizer of $F^s(\cdot) = f^s(\cdot) + \psi(\cdot)$, and telescoping the above inequality for $t = 0, 1, \dots, m-1$, we have

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{t=1}^{m-1}\big(F^s(x^s_t) - F^s(x^s_*)\big)\Big] \\
&\leq \mathbb{E}\Big[\frac{\|z^s_0 - x^s_*\|^2}{2\alpha} + \sum_{t=0}^{m-1}\Big(\alpha p \Phi^s_t + \alpha p \sum_{k=0}^{s-1}\Phi^k + \langle \nabla f^s(x^s_t), \widehat{\mathsf{x}}^s - x^s_t\rangle + \psi(\widehat{\mathsf{x}}^s) - \psi(x^s_t)\Big)\Big] \enspace .
\end{aligned}
$$

Using the fact $\widehat{\mathsf{x}}^{s+1}$ is chosen uniformly at random from $\{x^s_0, \dots, x^s_{m-1}\}$, and the fact that $x^s_0 = \widehat{\mathsf{x}}^s$, the above inequality implies

$$
\begin{aligned}
&\mathbb{E}\Big[m\big(F^s(\widehat{\mathsf{x}}^{s+1}) - F^s(x^s_*)\big) - \big(F^s(\widehat{\mathsf{x}}^s) - F^s(x^s_*)\big)\Big] \\
&\leq \mathbb{E}\Big[\frac{\|z^s_0 - x^s_*\|^2}{2\alpha} + \alpha p m\Big(\sum_{k=0}^{s}\Phi^k\Big) + m\langle \nabla f^s(\widehat{\mathsf{x}}^{s+1}), \widehat{\mathsf{x}}^s - \widehat{\mathsf{x}}^{s+1}\rangle + m\big(\psi(\widehat{\mathsf{x}}^s) - \psi(\widehat{\mathsf{x}}^{s+1})\big)\Big] \\
&\leq \mathbb{E}\Big[\frac{F^s(\widehat{\mathsf{x}}^s) - F^s(x^s_*)}{\sigma\alpha} + \alpha p m\Big(\sum_{k=0}^{s}\Phi^k\Big) + m\langle \nabla f^s(\widehat{\mathsf{x}}^{s+1}), \widehat{\mathsf{x}}^s - \widehat{\mathsf{x}}^{s+1}\rangle + m\big(\psi(\widehat{\mathsf{x}}^s) - \psi(\widehat{\mathsf{x}}^{s+1})\big)\Big] \enspace .
\end{aligned}
$$

Above, the second inequality uses the fact that $z^s_0 = \widehat{\mathsf{x}}^s$ and that $F^s(\cdot)$ is σ-strongly convex. Dividing both sides by m and rearranging the terms (using $\frac{1}{\sigma\alpha} \geq 1$), we have

$$\begin{aligned}
&\mathbb{E}\Big[\big(F^s(\widehat{\mathsf{x}}^{s+1}) - F^s(x^s_*)\big)\Big] \\
&\leq \mathbb{E}\Big[\frac{F^s(\widehat{\mathsf{x}}^s) - F^s(x^s_*)}{\sigma\alpha m/2} + \alpha p\Big(\sum_{k=0}^{s}\Phi^k\Big) + \langle \nabla f^s(\widehat{\mathsf{x}}^{s+1}), \widehat{\mathsf{x}}^s - \widehat{\mathsf{x}}^{s+1}\rangle + \big(\psi(\widehat{\mathsf{x}}^s) - \psi(\widehat{\mathsf{x}}^{s+1})\big)\Big] \ .
\end{aligned}$$

□

7.3 Analysis for One Full Epoch

We telescope Lemma 7.4 for an entire epoch and arrive at the following lemma:

Lemma 7.5. *If $\alpha \leq O(\frac{\sigma}{p^2\ell_1\ell_2})$ and $\alpha \geq \Omega(\frac{1}{\sigma m})$, we have*

$$\sum_{s=0}^{p-1}\mathbb{E}\Big[\big(F^s(\widehat{\mathsf{x}}^s) - F^s(x^s_*)\big)\Big] \leq 3\mathbb{E}\Big[F(\widehat{\mathsf{x}}^0) - F(\widehat{\mathsf{x}}^p)\Big] \ .$$

Proof. Telescoping Lemma 7.4 for all the subepochs $s = 0, 1, \dots, p-1$, we have

$$\begin{aligned}
&\sum_{s=0}^{p-1}\mathbb{E}\Big[\big(F^s(\widehat{\mathsf{x}}^{s+1}) - F^s(x^s_*)\big)\Big] \\
&\overset{①}{\leq} \sum_{s=0}^{p-1}\mathbb{E}\Big[\frac{F^s(\widehat{\mathsf{x}}^s) - F^s(x^s_*)}{\sigma\alpha m/2} + \alpha p^2\Phi^s + \langle \nabla f^s(\widehat{\mathsf{x}}^{s+1}), \widehat{\mathsf{x}}^s - \widehat{\mathsf{x}}^{s+1}\rangle + \big(\psi(\widehat{\mathsf{x}}^s) - \psi(\widehat{\mathsf{x}}^{s+1})\big)\Big] \\
&\overset{②}{\leq} \sum_{s=0}^{p-1}\mathbb{E}\Big[\frac{F^s(\widehat{\mathsf{x}}^s) - F^s(x^s_*)}{\sigma\alpha m/2} + \langle \nabla f^s(\widehat{\mathsf{x}}^{s+1}), \widehat{\mathsf{x}}^s - \widehat{\mathsf{x}}^{s+1}\rangle + \big(\psi(\widehat{\mathsf{x}}^s) - \psi(\widehat{\mathsf{x}}^{s+1})\big) \\
&\qquad + \alpha p^2 C(\ell_1+\ell_2)\cdot\big(f^s(\widehat{\mathsf{x}}^s) - f^s(\widehat{\mathsf{x}}^{s+1}) - \langle \nabla f^s(\widehat{\mathsf{x}}^{s+1}), \widehat{\mathsf{x}}^s - \widehat{\mathsf{x}}^{s+1}\rangle\big)\big] + \alpha p^2 C(\ell_1\ell_2)\cdot\|\widehat{\mathsf{x}}^{s+1} - \widehat{\mathsf{x}}^s\|^2\Big] \\
&\overset{③}{\leq} \sum_{s=0}^{p-1}\mathbb{E}\Big[\frac{F^s(\widehat{\mathsf{x}}^s) - F^s(x^s_*)}{\sigma\alpha m/2} + \langle \nabla f^s(\widehat{\mathsf{x}}^{s+1}), \widehat{\mathsf{x}}^s - \widehat{\mathsf{x}}^{s+1}\rangle + \big(\psi(\widehat{\mathsf{x}}^s) - \psi(\widehat{\mathsf{x}}^{s+1})\big) \\
&\qquad + \big(f^s(\widehat{\mathsf{x}}^s) - f^s(\widehat{\mathsf{x}}^{s+1}) - \langle \nabla f^s(\widehat{\mathsf{x}}^{s+1}), \widehat{\mathsf{x}}^s - \widehat{\mathsf{x}}^{s+1}\rangle\big)\big] + 2\sigma\cdot\|\widehat{\mathsf{x}}^{s+1} - \widehat{\mathsf{x}}^s\|^2\Big] \\
&= \sum_{s=0}^{p-1}\mathbb{E}\Big[\frac{F^s(\widehat{\mathsf{x}}^s) - F^s(x^s_*)}{\sigma\alpha m/2} + \big(F^s(\widehat{\mathsf{x}}^s) - F^s(\widehat{\mathsf{x}}^{s+1})\big) + 2\sigma\cdot\|\widehat{\mathsf{x}}^{s+1} - \widehat{\mathsf{x}}^s\|^2\Big] \\
&\overset{④}{=} \sum_{s=0}^{p-1}\mathbb{E}\Big[\frac{F^s(\widehat{\mathsf{x}}^s) - F^s(x^s_*)}{\sigma\alpha m/2} + \big(F^s(\widehat{\mathsf{x}}^{s+1}) - F^s(\widehat{\mathsf{x}}^s)\big) - 2\big(F(\widehat{\mathsf{x}}^{s+1}) - F(\widehat{\mathsf{x}}^s)\big)\Big]
\end{aligned}$$

Above, inequality ① uses Lemma 7.4 and $\Phi^s \geq 0$; inequality ② uses the definition of Φ^s from Lemma 7.2; inequality ③ uses $\alpha p^2 C(\ell_1+\ell_2) \leq 1$ and $\alpha p^2 C(\ell_1\ell_2) \leq 2\sigma$; and equality ④ uses the definition $F^s(y) = F(y) + \sigma\|y - \widehat{\mathsf{x}}^s\|^2$.

Finally, rearranging both sides, and using the fact that $\frac{1}{\sigma\alpha m} \leq \frac{1}{6}$, we have

$$\sum_{s=0}^{p-1}\mathbb{E}\Big[\big(F^s(\widehat{\mathsf{x}}^s) - F^s(x^s_*)\big)\Big] \leq 3\mathbb{E}\Big[F(\widehat{\mathsf{x}}^0) - F(\widehat{\mathsf{x}}^p)\Big] \ .$$

□

8 Full Method: Final Theorem

We are now ready to state and prove our main convergence theorem for $\mathtt{Natasha}^{\mathsf{full}}$:

Theorem 2. *Suppose $f(x)$ is L-smooth, each $f_i(x)$ is (ℓ_1, ℓ_2)-smooth, $F(x)$ is σ-strongly nonconvex, and $\sigma \le \min\{\ell_1, \ell_2, L\}$. If $\frac{\ell_1 \ell_2}{\sigma^2} \le n$, $p = \Theta\big((\frac{\sigma^2}{\ell_1\ell_2} n)^{1/3}\big)$ and $\Theta(\frac{\sigma}{p^2 \ell_1 \ell_2})$, $\texttt{Natasha}^{\mathsf{full}}$ outputs a point x^{out} satisfying*

$$\mathbb{E}[\|\mathcal{G}_\eta(x^{\mathsf{out}})\|^2] \le O\Big(\frac{(\ell_1\ell_2\sigma)^{1/3} n^{2/3}}{T'n}\Big) \cdot (F(x^{\varnothing}) - F^*)$$

for every $\eta \in \big(0, \frac{1}{\max\{L, 4\sigma\}}\big]$. In other words, to obtain $\mathbb{E}[\|\mathcal{G}_\eta(x^{\mathsf{out}})\|^2] \le \varepsilon^2$, we need gradient complexity

$$O\Big(n \log\frac{1}{\varepsilon} + \frac{(\ell_1\ell_2\sigma)^{1/3} n^{2/3}}{\varepsilon^2} \cdot (F(x^{\varnothing}) - F^*)\Big) \ .$$

Remark 8.1. One can assume $\sigma \le L$ without loss of generality because any L-smooth function is also L-strongly nonconvex. One can assume $\sigma \le \ell_2$ without loss of generality because $f(x)$ is ℓ_2-strongly nonconvex if each $f_i(x)$ is (ℓ_1, ℓ_2)-smooth. Only $\sigma \le \ell_1$ is a minor requirement for Theorem 2, but if this is not true, one can replace ℓ_1 with σ before applying Theorem 2.

Remark 8.2. In Theorem 2 we have assumed $\frac{\ell_1\ell_2}{\sigma^2} \le n^2$. If this inequality does not hold, one should apply repeatSVRG instead and it gives faster running time (see Figure 1). More specifically, repeatSVRG gives a complexity of $\widetilde{O}\big(\frac{n\sigma + n^{3/4}(\ell_1\ell_2\sigma^2)^{1/4}}{\varepsilon^2}\big)$ under a mild assumption of $\sigma \ge \varepsilon^2$ in this more refined (ℓ_1, ℓ_2)-smoothness setting.

Proof of Theorem 2. One can verify that our choices of p and α satisfy $p \in [n]$, $\alpha \le O(\frac{\sigma}{p^2\ell_1\ell_2})$ and $\alpha \ge \Omega(\frac{1}{\sigma m})$, so we can apply Lemma 7.5 and telescope it for the entire algorithm (which has T' full epochs). Use the fact that $\widehat{\mathsf{x}}^p$ of the previous epoch equals $\widehat{\mathsf{x}}^0$ of the next epoch, we conclude that if we choose a random epoch and a random subepoch s, we will have

$$\mathbb{E}[F^s(\widehat{\mathsf{x}}^s) - F^s(x^s_*)] \le \frac{3}{pT'}(F(x^{\varnothing}) - F^*) \ .$$

By the σ-strong convexity of $F^s(\cdot)$, we have $\mathbb{E}[\sigma\|\widehat{\mathsf{x}}^s - x^s_*\|^2] \le \frac{6}{pT'}(F(x^{\varnothing}) - F^*)$.

Now, $F^s(x) = F(x) + \sigma\|x - \widehat{\mathsf{x}}^s\|^2$ satisfies the assumption of $G(x)$ in Lemma 4.1. If we use the SVRG method (see Theorem 2.3) to minimize the convex function $F^s(x)$, we get an output x^{out} satisfying $F^s(x^{\mathsf{out}}) - F^s(x^s_*) \le \varepsilon^2\sigma$ in gradient complexity $O\big((n + \frac{\ell_1\ell_2}{\sigma^2})\log\frac{1}{\varepsilon}\big) \le O(n\log\frac{1}{\varepsilon})$.

We can therefore apply Lemma 4.1 and conclude that this output x^{out} satisfies

$$\mathbb{E}[\|\mathcal{G}_\eta(x^{\mathsf{out}})\|^2] \le O\Big(\frac{\sigma}{pT'}\Big) \cdot (F(x^{\varnothing}) - F^*) = O\Big(\frac{(\ell_1\ell_2\sigma)^{1/3} n^{2/3}}{T'n}\Big) \cdot (F(x^{\varnothing}) - F^*) \ .$$

In other words, we obtain $\mathbb{E}[\|\mathcal{G}_\eta(x^{\mathsf{out}})\|^2] \le \varepsilon^2$ using

$$T'n = O\Big(n + \frac{(\ell_1\ell_2\sigma)^{1/3} n^{2/3}}{\varepsilon^2} \cdot (F(x^{\varnothing}) - F^*)\Big)$$

computations of the stochastic gradients. Here, the additive term n is because $T' \ge 1$.

Finally, adding this with $O(n\log\frac{1}{\varepsilon})$, the gradient complexity for the application of SVRG in the last line of $\texttt{Natasha}^{\mathsf{full}}$, we finish the proof of the total gradient complexity. □

9 Conclusion

Stochastic gradient descent and gradient descent (including alternating minimization) have become the canonical methods for solving non-convex machine learning tasks. However, can we design new non-convex methods to run even faster than SGD or GD?

This present paper tries to tackle this general question, by providing a new `Natasha` method which is intrinsically different from GD or SGD. It runs faster than GD and SVRG-based methods at least in theory. We hope that this could be a non-negligible step towards our better understanding of non-convex optimization.

Finally, our results give rise to an interesting dichotomy in the complexity of first-order non-convex optimization: the complexity scales with $n^{3/4}$ for $\sigma < L/\sqrt{n}$ and with $n^{2/3}$ for $\sigma > L/\sqrt{n}$. It remains open to investigate whether this dichotomy is intrinsic, or we can design a more efficient algorithm that outperforms both.

Acknowledgements

This paper is partially supported by a Microsoft Research Award, no. 0518584, and an NSF grant, no. CCF-1412958. We thank Yuanzhi Li for enlightening conversations. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the National Science Foundation and Microsoft Corporation.

Appendix

A From Stationary Points to Local Minima

Recently, researchers have shown that the general problem of finding (ε,ρ)-approximate local minima, under mild conditions, reduces to (repeatedly) finding ε-approximate stationary points for an $O(\rho)$-strongly nonconvex function [1, 9]. We sketch the details here for the sake of completeness, in the special case of $\psi(x) \equiv 0$.[11]

We say that a point x is (ε,δ)-approximate local minimum, if $\|\nabla f(x)\| \leq \varepsilon$ and $\nabla^2 f(x) \succeq -\delta \mathbf{I}$.

Carmon et al. [9] showed that an (ε,δ)-approximate minimum for the general problem (1.1) can be solved via the following iterative procedure. In every iteration at point x_t, detect whether the smallest eigenvalue of $\nabla^2 f(x_t)$ is below $-\delta$:

- if yes, find the smallest eigenvector of $\nabla^2 f(x_t)$ approximately and move in this direction. (One can use for instance the shift-and-invert method [12].)
- if no, define $f_t(x) = f(x) + L\big(\max\big\{0, \|x-x_t\| - \frac{\delta}{L_2}\big\}\big)^2$ where L_2 is the second-order smoothness of $f(x)$ and $f_t(x)$ can be proven as $5L$-smooth and 3δ-strongly nonconvex; we then find an ε-approximate stationary point of $f'(x)$ and move there.

The Trade-Off on $\boldsymbol{\delta}$. The final running time of the above algorithm depends on the maximum between (1) the eigenvector computation and (2) the stationary-point computation. The larger δ is, the faster (1) becomes and the slower (2) becomes; the smaller δ is, the faster (2) becomes and the slower (1) becomes.

As argued in [1, 9], if the Hessian-vector multiplication $\big(\nabla^2 f_i(x)\big)v$ for an arbitrary vector runs in the same time as computing $\nabla f_i(x)$ —which is the case for training neural nets and other machine learning tasks— the optimum trade-off is $\delta = \sqrt{L_2 \varepsilon}$. This again confirms that in strongly non-convex optimization, the parameter δ can usually be much smaller than L.

[11] This reduction was only proved in the literature for $\psi(x) \equiv 0$; it is a simple exercise to generalize it to the proximal setting. Also, this reduction was explicitly given in [9] but only implicitly in [1].

References

[1] Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding Approximate Local Minima for Nonconvex Optimization in Linear Time. In *STOC*, 2017. Full version available at `http://arxiv.org/abs/1611.01146`.

[2] Zeyuan Allen-Zhu. Katyusha: The First Direct Acceleration of Stochastic Gradient Methods. In *STOC*, 2017. Full version available at `http://arxiv.org/abs/1603.05953`.

[3] Zeyuan Allen-Zhu and Elad Hazan. Variance Reduction for Faster Non-Convex Optimization. In *NIPS*, 2016.

[4] Zeyuan Allen-Zhu and Yuanzhi Li. Doubly Accelerated Methods for Faster CCA and Generalized Eigendecomposition. *ArXiv e-prints*, abs/1607.06017, July 2016.

[5] Zeyuan Allen-Zhu and Yuanzhi Li. LazySVD: Even Faster SVD Decomposition Yet Without Agonizing Pain. In *NIPS*, 2016.

[6] Zeyuan Allen-Zhu and Yuanzhi Li. Follow the Compressed Leader: Faster Algorithm for Matrix Multiplicative Weight Updates. *ArXiv e-prints*, abs/1701.01722, January 2017.

[7] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent. In *Proceedings of the 8th Innovations in Theoretical Computer Science*, ITCS '17, 2017. Full version available at `http://arxiv.org/abs/1407.1537`.

[8] Zeyuan Allen-Zhu and Yang Yuan. Improved SVRG for Non-Strongly-Convex or Sum-of-Non-Convex Objectives. In *ICML*, 2016.

[9] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Accelerated Methods for Non-Convex Optimization. *ArXiv e-prints*, abs/1611.00756, November 2016.

[10] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. In *NIPS*, 2014.

[11] Roy Frostig, Rong Ge, Sham M. Kakade, and Aaron Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *ICML*, volume 37, pages 1–28, 2015.

[12] Dan Garber, Elad Hazan, Chi Jin, Sham M. Kakade, Cameron Musco, Praneeth Netrapalli, and Aaron Sidford. Robust shift-and-invert preconditioning: Faster and more sample efficient algorithms for eigenvector computation. In *ICML*, 2016.

[13] 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*, COLT 2015, 2015.

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

[15] Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 1246–1257, 2016.

[16] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A Universal Catalyst for First-Order Optimization. In *NIPS*, 2015.

[17] Yurii Nesterov. *Introductory Lectures on Convex Programming Volume: A Basic course*, volume I. Kluwer Academic Publishers, 2004.

[18] Yurii Nesterov. Accelerating the cubic regularization of newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

[19] Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. *ArXiv e-prints*, abs/1603.06160, March 2016.

[20] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *arXiv preprint arXiv:1309.2388*, pages 1–45, 2013. Preliminary version appeared in NIPS

2012.

[21] Shai Shalev-Shwartz. SDCA without Duality, Regularization, and Individual Convexity. In *ICML*, 2016.

[22] Lijun Zhang, Mehrdad Mahdavi, and Rong Jin. Linear convergence with condition number independent access of full gradients. In *Advances in Neural Information Processing Systems*, pages 980–988, 2013.