On The Convergence of First Order Methods for Quasar-Convex Optimization

Jikai Jin JKJIN@PKU.EDU.CN
Peking University, Beijing, China

Abstract

In recent years, the success of deep learning has inspired many researchers to study the optimization of general smooth non-convex functions. However, recent works have established pessimistic worst-case complexities for this class functions, which is in stark contrast with their superior performance in real-world applications (e.g. training deep neural networks). On the other hand, it is found that many popular non-convex optimization problems enjoy certain structured properties which bear some similarities to convexity. In this paper, we study the class of *quasar-convex functions* to close the gap between theory and practice. We study the convergence of first order methods in a variety of different settings and under different optimality criterions. We prove complexity upper bounds that are similar to standard results established for convex functions and much better that state-of-the-art convergence rates of non-convex functions. Overall, this paper suggests that *quasar-convexity* allows efficient optimization procedures, and we are looking forward to seeing more problems that demonstrate similar properties in practice.

1. Introduction

In this paper we consider the problem of minimizing a given objective function $f : \mathbb{R}^n \to \mathbb{R}$.

The study of this optimization problem has a long history. Early works mainly focus on the special case when f is convex, and we have access to the exact gradient at each point [16]. However, the situation begins to change with the advent of the big data era, and, in particular, the rise of machine learning. In many machine learning applications (e.g. deep neural networks [9]), the objective function is highly complicated and non-convex. Therefore, the classical theory of convex optimization can no longer produce meaningful implications in many real-world scenarios.

Motivated by the empirical success of optimization algorithms, in recent years there has been a flurry of research works that study algorithms in non-convex optimization [4, 7, 8, 12]. Specifically ,these works study efficient algorithms for finding an approximate stationary points for general smooth non-convex function. A standard result is that the simple Stochastic Gradient Descent(SGD) algorithm can find an ϵ-stationary point with a complexity of $\mathcal{O}\left(\epsilon^{-4}\right)$. However, it has been established recently that this complexity is already optimal among *first order methods* (i.e. methods that only use first-order information of the objective function) [2]. This gives a convergence rate which is considerably slower than the actual convergence rate we observe in practice, thereby suggesting that there is still a gap between theory and practice.

On the other hand, the study of specific optimization problems suggests that sometimes the objective function exhibits certain desirable properties. For instance it has been proved that there is no spurious local minima in a variety of low-rank matrix problems [6], policy optimization in reinforcement learning satisfies some Polyak-Łojasiewicz-type conditions [5, 14], the landscape

of neural network exhibits some convex-like properties [13], etc. These observations inspire us to consider the possibility of more efficient optimization when imposing structural assumptions to the objective function.

In this paper we study the optimization using first order methods under *quasar-convexity*, which is a generalization of the notion of *convexity*. While several prior works [3, 10, 11] have provided theoretical treatments of this class of functions in some specific settings, we extend the analysis of quasar-convex optimization to include a variety of setups that are of practical interest. We also provide sharper results compared with [10] in some special cases. The main results of this paper are summarized in Section A.

2. Preliminaries

We first introduce some definitions that will be useful in this paper.

Definition 1 *A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be L-smooth if its gradient is L-Lipschitz, i.e. $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^n$.*

Definition 2 *(Quasar-convexity) Suppose that the function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and has a global minimizer x^*, then we say that f is γ-quasar-convex w.r.t. x^* if $0 \leq \gamma \leq 1$ and the following holds for all $x \in \mathbb{R}^n$:*

$$f(x^*) \geq f(x) + \frac{1}{\gamma}\nabla f(x)^T(x^* - x) \tag{1}$$

Further we say that f is (γ, μ)-strongly-quasar-convex if additionally $\mu \geq 0$ and the following holds for all $x \in \mathbb{R}^n$:

$$f(x^*) \geq f(x) + \frac{1}{\gamma}\nabla f(x)^T(x^* - x) + \frac{\mu}{2}\|x - x^*\|^2 \tag{2}$$

Throughout this paper we describe the performance of optimization algorithms via providing their *oracle complexities*. Roughly speaking, the *oracle complexity* of an algorithm is the minimum time it needs to query a certain oracle in order to meet some optimality criterion.

In this paper we consider two classes of oracles:

Deterministic Oracle $\mathcal{D}(f)$. The algorithm sends a point x to the oracle and the oracle responds with a pair $(f(x), \nabla f(x))$.

Stochastic Oracle $\mathcal{S}(f, \sigma)$. The algorithm sends a point x to the oracle and the oracle responds with random vector $g(x)$ such that $\mathbb{E}g(x) = f(x)$ and $\mathbb{E}\|g(x) - \nabla f(x)\|^2 \leq \sigma^2$. However, note that our results in this paper can be easily extended to the more general setting with $\mathbb{E}\|g(x) - \nabla f(x)\|^2 \leq M\|\nabla f(x)\|^2 + \sigma^2$.

For briefness we will refer to the optimization problem equipped with these two oracles as *deterministic setting* and *stochastic setting*, respectively.

We will also consider two types of optimality criterion. Specifically, we consider finding an ϵ-optimal point (i.e. a point $\widetilde{x}$ with $f(\widetilde{x}) - \inf_x f(x) \leq \epsilon$) and finding an ϵ-stationary point (i.e. a point with $\|\nabla f(x)\| \leq \epsilon$).

Formal definitions of related concepts are given in Appendix B due to space limits.

2.1. Notations

Throughout this paper we let x_0 be the starting point of all the algorithms we consider, x^* be the global minima of function f, and let R, Δ be upper bounds of the quantities $\|x_0 - x^*\|$ and $f(x_0) - f(x^*)$ respectively. We use $\mathcal{O}$ to hide numerical constants and $\widetilde{\mathcal{O}}$ to hide $\log$ terms.

3. Convergence Results for Smooth Quasar-Convex Functions

In this section we study the convergence of first order methods for smooth non-strongly-quasar-convex functions. We consider the deterministic and stochastic setting separately.

3.1. Deterministic Setting

In Hinder et al. [11] the authors propose a near optimal method for finding ϵ-optimal points in this setting. We recall their result below.

Theorem 3 *([11, Theorem 2]) There exists an algorithm $\mathcal{A}_c$ such that its complexity of finding an ϵ-optimal point is $\mathcal{O}\left(\sqrt{\frac{LR^2}{\gamma\epsilon}}\log\left(\sqrt{\frac{LR^2}{\gamma\epsilon}}\right)\right)$.*

A direct application of the inequality $\|\nabla f(x)\|^2 \leq L\left(f(x) - f(x^*)\right)$ would give a $\widetilde{\mathcal{O}}\left(\epsilon^{-1}\right)$ complexity upper bound for finding approximate stationary point. However, we can improve this bound by using the *GD after AGD* trick proposed by Nesterov [15]. The result is summarized in the following theorem.

Theorem 4 *There exists an algorithm such that for L-smooth and γ-quasi convex functions its complexity of finding ϵ-stationary point is $\widetilde{\mathcal{O}}\left(LR^{\frac{2}{3}}\gamma^{-\frac{1}{3}}\epsilon^{-\frac{2}{3}}\right)$.*

The idea is to run the algorithm $\mathcal{A}_{\mathsf{c}}$ in Theorem 3 to reach an ϵ_1-optimal point x_1, and then run the standard Gradient Descent(GD) to reach an ϵ-stationary point. The complexity is then upper bounded by combining Theorem 3 and standard results of GD. Finally, choosing ϵ_1 optimally gives the desired result.

The formal proofs of all results in the main paper are deferred to the Appendix.

3.2. Stochastic Setting

We first recall the vanilla Stochastic Gradient Descent(SGD) algorithm:

Algorithm 1: $\mathtt{SGD}\left(f, x_0, \{\alpha_k\}_{k\geq 1}, T\right)$

Input : Objective function f, initial point x_0, parameters γ, L, σ, total iterations T
for $t \leftarrow 1$ *to* T **do**
  $x_t \leftarrow x_{t-1} - \alpha_t \nabla f(x_{t-1}, \xi_{t-1})$;
end
Output : $\widetilde{x} \in \{x_1, x_2, \cdots, x_T\}$ uniformly at random

In the following theorem we establish the convergence rate of SGD for smooth quasar-convex functions.

Theorem 5 *Suppose that f is an L-smooth, γ-quasar-convex function, and we run SGD for T iterations with some fixed step size $\alpha_t = \alpha = \min\left\{\frac{R}{2\sigma\sqrt{T}}, \frac{1}{2L}\right\}$, where $R = \|x_0 - x^*\|$. Then, we have*

$$\frac{1}{T}\mathbb{E}\sum_{t=1}^{T}\left(f(x_t) - f(x^*)\right) \leq 4\left(\frac{R\sigma}{\gamma\sqrt{T}} + \frac{1}{\gamma}\frac{R^2 L}{T}\right) \tag{3}$$

The only existing convergence guarantee in our setting that we are aware of is established in Gower et al. [10]. We observe that their bound in neither uniformly stronger nor weaker than ours; see Section E for a detailed discussion.

Corollary 6 *For the class of L-smooth, γ-quasar-convex functions the complexity os SGD for finding ϵ-optimal point is $\mathcal{O}\left(\frac{R^2\sigma^2}{\gamma^2\epsilon^2} + \frac{R^2 L}{\gamma\epsilon}\right)$*

Equipped with the above results, we can now establish a complexity upper bound for making gradient small.

Theorem 7 *There exists an algorithm such that for L-smooth and γ-quasi convex functions its complexity of finding ϵ-stationary point is $\mathcal{O}\left(\sigma^2\left(\frac{LR}{\gamma\epsilon^4}\right)^{\frac{2}{3}}\right)$ (here we omit the lower order terms for convenience).*

4. Convergence Results for Smooth Strongly-Quasar-Convex Functions

In this section we turn to the optimization of smooth strongly-quasar-convex functions.

4.1. Deterministic Setting

In the deterministic setting, the following result was established in Hinder et al. [11].

Theorem 8 *([11, Theorem 1])There exists an algorithm $\mathcal{A}_{sc}$ such that for L-smooth and (γ, μ)-strongly-quasar-convex functions, then the complexity of $\mathcal{A}_{sc}$ for finding an ϵ-suboptimal point is $\mathcal{O}\left(\sqrt{\frac{\kappa}{\gamma^2}}\log\left(\frac{\kappa\Delta}{\epsilon}\right)\right)$, where $\kappa = L/\mu$.*

An immediate consequence of the above result is the following:

Corollary 9 *The complexity of $\mathcal{A}_{sc}$ for finding ϵ-stationary point is $\mathcal{O}\left(\sqrt{\frac{\kappa}{\gamma^2}}\log\left(\frac{\kappa L\Delta}{\epsilon^2}\right)\right)$. This also implies a complexity of $\mathcal{O}\left(\sqrt{\frac{\kappa}{\gamma^2}}\log\left(\frac{\sqrt{\kappa}LR}{\epsilon}\right)\right)$*

4.2. Stochastic Setting

We have the following convergence result for vanilla SGD:

Theorem 10 *The vanilla SGD finds an ϵ-optimal point with complexity $\mathcal{O}\left(\frac{L\sigma^2}{\mu^2\gamma^2 T}\left(1 + \log\left(\frac{\mu\gamma R\sqrt{T}}{\sigma}\right)\right)\right)$.*

Next we refine the analysis and proves a new convergence guarantee which has better dependence on μ and L.

Theorem 11 *Suppose that we run vanilla SGD for* $T > \max\left\{\frac{3\sigma^2}{\gamma^2\mu^2R^2}, \frac{6L}{\gamma^2\mu}\left(\log\left(\frac{2L\mu R^2}{\sigma^2}\right)+1\right)\right\}$ *iterations with some fixed step size* $\alpha = \frac{1}{\gamma\mu T}\log\left(\frac{\gamma^2\mu^2TR^2}{\sigma^2}\right)$*, then it can output a random point* X *such that*

$$\mathbb{E}\left[f(X) - f(x^*)\right] \leq \widetilde{\mathcal{O}}\left(\frac{\sigma^2}{\gamma^2\mu T}\right) \tag{4}$$

Corollary 12 *The complexity of Algorithm 1 for finding* ϵ*-optimal point is* $\mathcal{O}\left(\frac{\sigma^2}{\gamma^2\mu\epsilon} + \frac{L}{\gamma\mu}\log\left(\frac{\mu R^2}{\epsilon}\right)\right)$ *with appropriate choice of step size.*

Finally we establish the following complexity upper bound of finding ϵ-stationary points.

Theorem 13 *There exists an algorithm that achieves a complexity of* $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\frac{\sigma^2}{\gamma\epsilon^2} + \frac{L}{\mu}\log\left(\frac{\gamma\mu\sqrt{L}R^2}{\epsilon^2}\right)\right)$ *for finding* ϵ*-stationary points.*

The idea is to use the *SGD after SGD* approach proposed by Allen-Zhu [1]. Specifically, we first run SGD to find an ϵ_1-optimal point, then run SGD starting from this point to find a point with small gradient. This approach is an extension of Nesterov's *GD after AGD* approach in the deterministic setting.

5. Conclusion & Future directions

In this paper we study smooth quasar- and strongly-quasar-convex functions with two different optimality criterions and in two different settings. However, there are still some interesting questions that remain unanswered. Firstly, it is unclear whether the dependency of our bounds on γ is optimal. Indeed the discussion in Section E suggests that they might be improved. Secondly we note that there exists another trick for finding ϵ-stationary points for *convex functions* in existing literature [1, 15]. The idea is to add a small perturbation to make the function strongly convex, which can be optimized very efficiently. This approach can yield complexities that match corresponding lower bounds. Unfortunately it cannot be applied to quasar-convex functions, since we cannot guarantee that x^* is still the global minima after perturbation. It is thus unknown whether there exists other efficient approaches, or whether our approach is already optimal. Finally, we are looking forward to exploring convergence guarantees for other types of structured non-convex functions in the future.

References

[1] Zeyuan Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex sgd. In *Advances in Neural Information Processing Systems*, pages 1157–1167, 2018.

[2] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

[3] Jingjing Bu and Mehran Mesbahi. A note on nesterov's accelerated method in nonconvex optimization: a weak estimate sequence approach. *arXiv preprint arXiv:2006.08548*, 2020.

[4] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

[5] Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *arXiv preprint arXiv:2007.06558*, 2020.

[6] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: a unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1233–1242, 2017.

[7] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[8] Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

[9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[10] Robert M Gower, Othmane Sebbouh, and Nicolas Loizou. Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation. *arXiv preprint arXiv:2006.10311*, 2020.

[11] Oliver Hinder, Aaron Sidford, and Nimit Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In *Conference on Learning Theory*, pages 1894–1938. PMLR, 2020.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in neural information processing systems*, pages 597–607, 2017.

[14] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. *arXiv preprint arXiv:2005.06392*, 2020.

[15] Yurii Nesterov. How to make the gradients small. *Optima. Mathematical Optimization Society Newsletter*, (88):10–11, 2012.

[16] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Appendix A. Summary of Main Results

We summarize all our complexity results in the table below.

	Smooth Quasar-Convex Function	
	Deterministic	Stochastic
Finding Approximate Minima	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{LR^2}{\gamma\epsilon}}\right)$ ([11, Theorem 2])	$\mathcal{O}\left(\frac{R^2\sigma^2}{\gamma^2\epsilon^2}+\frac{R^2L}{\gamma\epsilon}\right)$ (Corollary 6)
Making Gradient Small	$\mathcal{O}\left(LR^{\frac{2}{3}}\gamma^{-\frac{1}{3}}\epsilon^{-\frac{2}{3}}\right)$ (Theorem 4)	$\mathcal{O}\left(\sigma^2\left(\frac{LR}{\gamma\epsilon^4}\right)^{\frac{2}{3}}\right)$ (Theorem 7)
	Smooth Strongly-Quasar-Convex Function	
	Deterministic	Stochastic
Finding Approximate Minima	$\mathcal{O}\left(\sqrt{\frac{\kappa}{\gamma^2}}\log\left(\frac{\kappa\Delta}{\epsilon}\right)\right)$ ([11, Theorem 1])	$\widetilde{\mathcal{O}}\left(\frac{\sigma^2}{\gamma^2\mu\epsilon}+\frac{L}{\gamma\mu}\log\left(\frac{\mu R^2}{\epsilon}\right)\right)$ (Corollary 12)
Making Gradient Small	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{\kappa}{\gamma^2}}\log\left(\frac{\sqrt{\kappa}LR}{\epsilon}\right)\right)$ (Corollary 9)	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{L}{\mu}}\frac{\sigma^2}{\gamma\epsilon^2}\right)$ (Theorem 20)

Appendix B. Formal descriptions of the setup

In this section we introduce some useful concepts that allow us to rigorously describe the performance of a specific optimization algorithm.

B.1. Optimization Oracle

In order to describe the optimization process more conveniently, we assume that the optimization algorithm (only) has access to an *oracle* which answers successive queries of the algorithm.

In this paper we only consider two most commonly used oracle in the optimization literature.

Deterministic Oracle $\mathcal{D}(f)$. The algorithm sends a point x to the oracle and the oracle responds with a pair $(f(x), \nabla f(x))$.

Stochastic Oracle $\mathcal{S}(f, \sigma)$. The algorithm sends a point x to the oracle and the oracle responds with random vector $g(x)$ such that $\mathbb{E}g(x) = f(x)$ and $\mathbb{E}\|g(x) - \nabla f(x)\|^2 \leq \sigma^2$.

For briefness we will refer to the optimization problem equipped with these two oracles as deterministic setting and stochastic setting, respectively.

B.2. Optimization Algorithms

For a given oracle $\mathbb{O}$, we consider the set $\mathcal{A}(\mathbb{O})$ consisting of all algorithms that works as follows: starting from a point x_0, it produces a (random) sequence $\{x_t\}$ according the following recursive relation:

$$x_t = \mathsf{A}_t\left(r, \mathsf{O}_0, \cdots, \mathsf{O}_{t-1}\right) \tag{5}$$

where O_i is the oracle feedback at x_i, r is a random seed and A_t is a deterministic mapping.

B.3. Complexity Measures

Consider a function class $\mathcal{F}$ and oracle class $\mathbb{O}$, let $\mathcal{P}[\mathcal{F}]$ be the set of all distributions over $\mathcal{F}$,then for all $\epsilon > 0$ we define the complexity *for finding approximate stationary point* as

$$\sup_{\mathsf{O}\in\mathbb{O}} \sup_{P\in\mathcal{P}[\mathcal{F}]} \inf_{\mathsf{A}\in\mathcal{A}(\mathsf{O})} \inf\left\{T \in \mathbb{N} \middle| \mathbb{E}\|\nabla f(x_T)\| \leq \epsilon\right\}$$

and the complexity *for finding approximate global minima* as

$$\sup_{\mathsf{O}\in\mathbb{O}} \sup_{P\in\mathcal{P}[\mathcal{F}]} \inf_{\mathsf{A}\in\mathcal{A}(\mathsf{O})} \inf\left\{T \in \mathbb{N} \middle| \mathbb{E}\left[f(x_T) - \inf_{x\in\mathbb{R}^n} f(x)\right] \leq \epsilon\right\}$$

where we omit the dependence of x_T on P and $\mathcal{A}$ in the above expressions.

B.4. Notations

Throughout this paper we let x_0 be the starting point of all the algorithms we consider, x^* be the global minima of function f, and let R, Δ be upper bounds of the quantities $\|x_0 - x^*\|$ and $f(x_0) - f(x^*)$ respectively. We use $\mathcal{O}$ to hide numerical constants and $\widetilde{\mathcal{O}}$ to hide $\log$ terms.

The following two function classes appear regularly in the main paper:

$$\mathcal{F}_{\mathsf{c}}(\gamma, R) = \{f : \mathbb{R}^n \to \mathbb{R} : f \text{ is } L\text{-smooth and } \gamma\text{-quasar convex, and } \|x_0 - x^*\| \leq R\}$$

$$\mathcal{F}_{\mathsf{sc}}(\gamma, \mu, R) = \{f : \mathbb{R}^n \to \mathbb{R} : f \text{ is } L\text{-smooth and } (\gamma, \mu)\text{-quasar strongly-convex, and } \|x_0 - x^*\| \leq R\}$$

Appendix C. Proof of Theorems in Section 3

Theorem 14 *(Restatement of Theorem 4) There exists an algorithm such that for any L-smooth and γ-quasi convex function f such that its complexity of finding ϵ-stationary point is $\widetilde{\mathcal{O}}\left(LR^{\frac{2}{3}}\gamma^{-\frac{1}{3}}\epsilon^{-\frac{2}{3}}\right)$ function and gradient evaluations.*

Proof The idea is to use Nesterov's *'GD after AGD'* trick [1], where we replace Nesterov's AGD with the algorithm $\mathcal{A}_c$.

Specifically, for a fixed $\epsilon_1 > 0$, we first run $\mathcal{A}_c$ for $\widetilde{\mathcal{O}}\left(\sqrt{\frac{LR^2}{\gamma\epsilon_1}}\right)$ iterations and arrive at a point $\widetilde{x_0}$ such that $f\left(\widetilde{x_0}\right) - f(x^*) \leq \epsilon_1$. Then, starting from $\widetilde{x_0}$ we run gradient descent for $\mathcal{O}\left(L\epsilon_1\epsilon^{-2}\right)$ iterations. It is well known that we can then find a point x such that $\|\nabla f(x)\| \leq \epsilon$.

The complexity of the above procedure is then

$$\widetilde{\mathcal{O}}\left(\sqrt{\frac{LR^2}{\gamma\epsilon_1}} + L\epsilon_1\epsilon^{-2}\right)$$

The result follows by choosing $\epsilon_1 = \left(\gamma^{-1}R^2\epsilon^4\right)^{\frac{1}{3}}$. ■

Theorem 15 *(Restatement of Theorem 5) Suppose that f is an L-smooth, γ-quasar-convex function, and we run SGD for T iterations with some fixed step size $\alpha_t = \alpha$, where $R = \|x_0 - x^*\|$. Then, we have*

$$\frac{1}{T}\mathbb{E}\sum_{t=1}^{T}\left(f(x_t) - f(x^*)\right) \leq 4\left(\frac{R\sigma}{\gamma\sqrt{T}} + \frac{1}{\gamma}\frac{R^2L}{T}\right) \tag{6}$$

Proof First note that

$$\begin{aligned}\|x_t - x^*\|^2 &= \|x_{t+1} - x^*\|^2 + 2\left\langle x_t - x_{t+1}, x_{t+1} - x^*\right\rangle + \|x_{t+1} - x_t\|^2 \\ &= \|x_{t+1} - x^*\|^2 + \|x_{t+1} - x_t\|^2 + 2\alpha\left\langle \nabla f(x_t, \xi_t), x_{t+1} - x^*\right\rangle\end{aligned} \tag{7}$$

Denote $\Delta_t = \mathbb{E}\left[f(x_t) - f^*\right]$, then we have

$$\begin{aligned}\Delta_{t+1} &\leq \mathbb{E}\left[f(x_t) + \langle\nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 - f(x^*)\right] \\ &\leq \mathbb{E}\left[\langle\nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 + \frac{1}{\gamma}\left\langle\nabla f(x_t), x_t - x^*\right\rangle\right] \\ &= \mathbb{E}\left[\left\langle\nabla f(x_t), x_{t+1} + \left(\frac{1}{\gamma} - 1\right)x_t - \frac{1}{\gamma}x^*\right\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2\right] \\ &= \mathbb{E}\left[\frac{1}{\gamma}\left\langle\nabla f(x_t), x_{t+1} - x^*\right\rangle + \left(\frac{1}{\gamma} - 1\right)\left\langle\nabla f(x_t), x_t - x_{t+1}\right\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2\right]\end{aligned} \tag{8}$$

Now we handle the first and second term in the above expression respectively. First by (7), for some fixed $\lambda > 0$ we have

$$\begin{aligned}
&\mathbb{E}\left[\langle \nabla f(x_t), x_{t+1} - x^* \rangle\right] \\
&= \mathbb{E}\left[\langle \nabla f(x_t) - \nabla f(x_t, \xi_t), x_{t+1} - x^* \rangle\right] + \mathbb{E}\left[\langle \nabla f(x_t, \xi_t), x_{t+1} - x^* \rangle\right] \\
&= \mathbb{E}\left[\langle \nabla f(x_t) - \nabla f(x_t, \xi_t), x_{t+1} - x_t \rangle\right] + \mathbb{E}\left[\langle \nabla f(x_t, \xi_t), x_{t+1} - x^* \rangle\right] \\
&\leq \frac{1}{2\lambda}\mathbb{E}\|\nabla f(x_t) - \nabla f(x_t, \xi_t)\|^2 + \mathbb{E}\left[\frac{\lambda}{2}\|x_{t+1} - x_t\|^2 + \frac{1}{2\alpha}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 - \|x_{t+1} - x_t\|^2\right)\right] \\
&\leq \frac{\sigma^2}{2\lambda} + \mathbb{E}\left[\frac{1}{2\alpha}\|x_t - x^*\|^2 - \frac{1}{2\alpha}\|x_{t+1} - x^*\|^2 - \left(\frac{1}{2\alpha} - \frac{\lambda}{2}\right)\|x_{t+1} - x_t\|^2\right]
\end{aligned} \tag{9}$$

Next, as long as $\alpha < \frac{1}{2L}$, by the L-smoothness of f we have

$$\Delta_{t+1} - \Delta_t \leq -\alpha\mathbb{E}\|\nabla f(x_t)\|^2 + \frac{L}{2}\alpha^2\left(\mathbb{E}\|\nabla f(x_t)\|^2 + \sigma^2\right) \leq -\frac{\alpha}{2}\mathbb{E}\|\nabla f(x_t)\|^2 + \frac{L}{2}\alpha^2\sigma^2 \tag{10}$$

Therefore

$$\begin{aligned}
\mathbb{E}\left\langle \nabla f(x_t), x_t - x_{t+1} \right\rangle &= \alpha\mathbb{E}\|\nabla f(x_t)\|^2 \\
&\leq 2\left(\Delta_t - \Delta_{t+1}\right) + L\alpha^2\sigma^2
\end{aligned} \tag{11}$$

Now, by plugging (9) and (11) into (8) we have

$$\begin{aligned}
\Delta_{t+1} \leq \frac{\sigma^2}{2\gamma\lambda} + \frac{1}{2\gamma\alpha}\mathbb{E}\left[\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right] - \frac{1}{\gamma}\left(\frac{1}{2\alpha} - \frac{\lambda}{2} - \frac{L\gamma}{2}\right)\mathbb{E}\|x_{t+1} - x_t\|^2 \\
+ 2\left(\frac{1}{\gamma} - 1\right)\left(\Delta_t - \Delta_{t+1}\right) + \left(\frac{1}{\gamma} - 1\right)L\alpha^2\sigma^2
\end{aligned} \tag{12}$$

Choosing $\lambda = \frac{1}{\alpha} - L\gamma$ in the RHS and rearranging, we obtain

$$\frac{1}{2\gamma\alpha}\mathbb{E}\left[\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right] + \frac{\sigma^2\alpha}{2\gamma(1 - \alpha L\gamma)} + \left(\frac{1}{\gamma} - 1\right)L\alpha^2\sigma^2 \geq \left(\frac{2}{\gamma} - 1\right)\Delta_{t+1} - \left(\frac{2}{\gamma} - 2\right)\Delta_t \tag{13}$$

Perform a telescope sum for $t = 0, 1, \cdots, T-1$ gives

$$\left(\frac{2}{\gamma} - 1\right)\Delta_T + \sum_{t=1}^{T-1}\Delta_t \leq \frac{R^2}{2\gamma\alpha} + \frac{\sigma^2\alpha}{2\gamma(1 - \alpha L\gamma)}T + \left(\frac{1}{\gamma} - 1\right)L\alpha^2\sigma^2 T + \left(\frac{2}{\gamma} - 2\right)\Delta_0 \tag{14}$$

Again by L-smoothness we can see that $\Delta_0 \leq \frac{L}{2}R^2$, thus

$$\sum_{t=1}^{T}\Delta_t \leq \frac{R^2}{2\gamma\alpha} + \frac{\sigma^2\alpha}{2\gamma(1 - \alpha L\gamma)}T + \left(\frac{1}{\gamma} - 1\right)L\alpha^2\sigma^2 T + \left(\frac{1}{\gamma} - 1\right)LR^2 \tag{15}$$

Choose $\alpha = \frac{R}{2\sigma\sqrt{T}}$, which is smaller than $\frac{1}{2L}$ when $T > \frac{R^2L^2}{\sigma^2}$, then

$$\frac{1}{T}\sum_{t=1}^{T}\Delta_t \leq \frac{2R\sigma}{\gamma\sqrt{T}} + 2\left(\frac{1}{\gamma} - 1\right)\frac{R^2L}{T} < \frac{4R\sigma}{\gamma\sqrt{T}} \tag{16}$$

Otherwise if $T \leq \frac{R^2L^2}{\sigma^2}$, it follows from (15) that

$$\frac{1}{T}\sum_{t=1}^{T}\Delta_t \leq \frac{R^2}{2\gamma\alpha T} + \frac{\sigma^2\alpha}{2\gamma(1-\alpha L\gamma)} + \left(\frac{1}{\gamma}-1\right)L\alpha^2\sigma^2 + \left(\frac{1}{\gamma}-1\right)\frac{LR^2}{T} \tag{17}$$

We choose $\alpha = \frac{1}{2L}$ in the above inequality, so that by some calculation we have

$$\frac{1}{T}\sum_{t=1}^{T}\Delta_t \leq 4\left(\frac{R\sigma}{\gamma\sqrt{T}} + \frac{1}{\gamma}\frac{R^2L}{T}\right) \tag{18}$$

■

Theorem 16 *(Restatement of Theorem 7) There exists an algorithm that can output a point $\overline{x}$ such that $\mathbb{E}\left[f\left(\overline{x}\right) - f(x^*)\right] \leq \epsilon$ with at most $\mathcal{O}\left(\sigma^2\left(\frac{LR}{\gamma\epsilon^4}\right)^{\frac{2}{3}}\right)$ queries to the stochastic oracle.*

Proof The idea is to use the *SGD after SGD* approach proposed by Allen-Zhu [1]. Fixed $\epsilon_1 > 0$, we first run SGD for $\mathcal{O}\left(\frac{R^2\sigma^2}{\gamma^2\epsilon_1^2} + \frac{R^2L}{\gamma\epsilon_1}\right)$ iterations, then 6 ensures an output X_1 (a random variable) such that $\mathbb{E}\left[f(X_1) - f(x^*)\right] \leq \epsilon_1$. Let $\mathcal{F}_1$ denote the σ-algebra generated by all the randomness in this stage.

Next, we run SGD for another $\mathcal{O}\left(L\epsilon_1\sigma^2\epsilon^{-4}\right)$ iterations, starting from X_1, then, according to a standard result in non-convex optimization Ghadimi and Lan [7], we can find a point (random variable) X_2 such that $\mathbb{E}\left[\|\nabla f(X_2)\|\middle|\mathcal{F}_1\right] \leq \epsilon\left(\frac{f(X_1)-f(x^*)}{\epsilon_1}\right)^{\frac{1}{4}}$. This implies that

$$\mathbb{E}\|\nabla f(X_2)\| \leq \epsilon\mathbb{E}\left[\left(\frac{f(X_1)-f(x^*)}{\epsilon_1}\right)^{\frac{1}{4}}\right] \leq \epsilon\left(\frac{\mathbb{E}\left[f(X_1)-f^*\right]}{\epsilon_1}\right)^{\frac{1}{4}} \leq \epsilon \tag{19}$$

The total iterations is then given by $\mathcal{O}\left(\frac{R^2\sigma^2}{\gamma^2\epsilon_1^2} + \frac{R^2L}{\gamma\epsilon_1} + L\epsilon_1\sigma^2\epsilon^{-4}\right)$. Choose ϵ_1 optimally and omitting lower order terms, we obtain the desired result. ■

Appendix D. Proof of Theorems in Section 4

Corollary 17 *(Restatement of Corollary 9) The complexity of $\mathcal{A}_{sc}$ for finding ϵ-stationary point is* $\mathcal{O}\left(\sqrt{\frac{\kappa}{\gamma^2}}\log\left(\frac{\kappa L\Delta}{\epsilon^2}\right)\right)$

Proof Note that L-smoothness of f implies that $\|\nabla f(x)\|^2 \leq L\left(f(x)-f(x^*)\right)$. The first statement follows from Theorem 8. The second statement follows from the inequality $f(x_0)-f^* \leq \frac{L}{2}R^2$. ■

Theorem 18 *(Restatement of Theorem 10) The vanilla SGD finds an ϵ-optimal point with complexity* $\mathcal{O}\left(\frac{L\sigma^2}{\mu^2\gamma^2 T}\left(1+\log\left(\frac{\mu\gamma R\sqrt{T}}{\sigma}\right)\right)\right)$.

Proof According to Gower et al. [10] Theorem D.2, we have

$$\|x_T - x^*\|^2 \leq (1-\alpha\mu\gamma)^T + \frac{2\alpha\sigma^2}{\mu\gamma} \leq \exp\left(-\alpha\mu\gamma T\right)R^2 + \frac{2\alpha\sigma^2}{\mu\gamma} \tag{20}$$

Choosing $\alpha = \frac{1}{\mu\gamma T}\log\left(\frac{\mu^2\gamma^2R^2T}{2\sigma^2}\right)$ minimize the RHS of the above inequality. Thus we have

$$\|x_T - x^*\|^2 \leq \frac{L\sigma^2}{\mu^2\gamma^2 T}\left(1+\log\left(\frac{\mu\gamma R\sqrt{T}}{\sigma}\right)\right) \tag{21}$$

Finally by L-smoothness we have

$$f(x_T) - f(x^*) \leq \frac{L}{2}\|x_T - x^*\|^2 \tag{22}$$

and the conclusion follows. ■

Theorem 19 *(Restatement of Theorem 12) Suppose that we run vanilla SGD for*

$$T > \max\left\{\frac{3\sigma^2}{\gamma^2\mu^2R^2}, \frac{6L}{\gamma^2\mu}\left(\log\left(\frac{2L\mu R^2}{\sigma^2}\right)+1\right)\right\}$$

iterations with some fixed step size $\alpha = \frac{1}{\gamma\mu T}\log\left(\frac{\gamma^2\mu^2TR^2}{\sigma^2}\right)$, then it can output a random point X such that

$$\mathbb{E}\left[f(X) - f(x^*)\right] \leq \widetilde{\mathcal{O}}\left(\frac{\sigma^2}{\gamma^2\mu T}\right) \tag{23}$$

Proof Note that for any x we have

$$f(x^*) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|^2$$

Thus,

$$\begin{aligned}\mathbb{E}\|x_{t+1}-x^*\|^2 &\le \mathbb{E}\left[\|x_t-x^*\|^2 - 2\alpha(x_t-x^*)^T\nabla f(x) + \alpha^2\left(\|\nabla f(x)\|^2+\sigma^2\right)\right]\\ &\le \mathbb{E}\left[(1-\gamma\mu\alpha)\|x_t-x^*\|^2 - 2\gamma\alpha(f(x_t)-f(x^*)) + \alpha^2\left(\|\nabla f(x)\|^2+\sigma^2\right)\right]\\ &\le \mathbb{E}\left[(1-\gamma\mu\alpha)\|x_t-x^*\|^2 - 2\gamma\alpha(f(x_t)-f(x^*)) + 2\alpha^2L(f(x_t)-f(x^*)) + \alpha^2\sigma^2\right]\\ &= \mathbb{E}\left[(1-\gamma\mu\alpha)\|x_t-x^*\|^2 - 2\alpha(\gamma-\alpha L)(f(x_t)-f(x^*)) + \alpha^2\sigma^2\right]\end{aligned} \tag{24}$$

Recursively apply the above inequality we have that

$$2\alpha(\gamma-\alpha L)\sum_{t<T}(1-\gamma\mu\alpha)^{T-t-1}(f(x_t)-f(x^*)) \le \frac{\alpha\sigma^2}{\gamma\mu} + (1-\gamma\mu\alpha)^T R^2 \tag{25}$$

Suppose X is a random variable such that $X = x_t, t = 0, 1, \cdots, T-1$ with probability $(1-\gamma\mu\alpha)^{T-t-1}/Z$ where Z is a normalizing constant, then we have

$$\frac{2(\gamma-\alpha L)}{\gamma\mu}\left(1-(1-\gamma\mu\alpha)^T\right)\mathbb{E}\left[f(X)-f(x^*)\right] \le \frac{\alpha\sigma^2}{\gamma\mu} + (1-\gamma\mu\alpha)^T R^2 \tag{26}$$

$$\left(1-(1-\gamma\mu\alpha)^T\right)\mathbb{E}\left[f(X)-f(x^*)\right] \le \frac{\alpha\sigma^2}{2(\gamma-\alpha L)} + \frac{\gamma\mu(1-\gamma\mu\alpha)^T R^2}{2(\gamma-\alpha L)} \tag{27}$$

We choose $\alpha = \frac{1}{\gamma\mu T}\log\left(\frac{\gamma^2\mu^2 T R^2}{\sigma^2}\right)$. When $T > \frac{3\sigma^2}{\gamma^2\mu^2R^2}$ we have $\alpha > \frac{1}{\gamma\mu T}$, so $1-(1-\gamma\mu\alpha)^T > \frac{1}{2}$. On the other hand, if $T > \frac{6L}{\gamma^2\mu}\max\left\{\log\left(\frac{2L\mu R^2}{\sigma^2}\right), 1\right\}$, we have $\gamma - \alpha L > \frac{1}{2}\gamma$. Thus for large T, $\mathbb{E}\left[f(x)-f(x^*)\right] = \widetilde{\mathcal{O}}\left(\frac{\sigma^2}{\gamma^2\mu T}\right)$. ■

Theorem 20 *(Restatement of Theorem) There exists an algorithm that achieves a complexity of* $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\frac{\sigma^2}{\gamma\epsilon^2} + \frac{L}{\mu}\log\left(\frac{\gamma\mu\sqrt{L}R^2}{\epsilon^2}\right)\right)$ *for finding* ϵ*-stationary points.*

Proof We first run SGD for $\mathcal{O}\left(\frac{\sigma^2}{\gamma^2\mu\epsilon_1} + \frac{L}{\gamma\mu}\log\left(\frac{\mu R^2}{\epsilon_1}\right)\right)$ to arrive at a (random) point X_1, then starting from X_1 we run SGD for another $\mathcal{O}\left(L\sigma^2\epsilon_1\epsilon^{-4}\right)$ iterations. Then we can output a point with expected gradient norm smaller than ϵ. The details are the same as the proof of Theorem 7. ■

Appendix E. Discussion of Theorem 5

In Gower et al. [10], the authors prove the following result:

Theorem 21 *([10, Theorem 4.1]) With appropriately chosen step sizes, SGD for finding ϵ-optimal point of L-smooth, γ-quasar-convex functions has complexity $\mathcal{O}\left(\frac{R^2+c^2\sigma^2}{c\sqrt{T}}\right)$, where $c \in \left(0, \frac{\gamma}{L}\right)$.*

When $\gamma\sigma > LR$, we can choose $c = \Theta\left(\frac{R}{\sigma}\right)$ in the above result. In this case the complexity becomes $\mathcal{O}\left(\frac{R\sigma}{\sqrt{T}}\right)$, which is better than our bound in Theorem 5 when $\gamma \ll 1$. On the other hand, if $\gamma\sigma \ll LR$, then the quantity in Theorem 21 attains its minimum at $c = \frac{\gamma}{L}$, which gives a complexity of $\mathcal{O}\left(\frac{LR^2}{\gamma\sqrt{T}} + \frac{\gamma\sigma^2}{L\sqrt{T}}\right)$, which has worse dependence on parameters L, R, σ compared with the dominating term $\mathcal{O}\left(\frac{R\sigma}{\gamma\sqrt{T}}\right)$ in Theorem 5.

An interesting special case is $\gamma = 1$, in which quasi-convexity is easily seen to be a weaker condition than convexity. Ghadimi and Lan [7] established that the convergence rate of SGD for convex functions is $\mathcal{O}\left(\frac{R\sigma}{\sqrt{T}} + \frac{R^2L}{T}\right)$, which is the same as our bound with $\gamma = 1$. Therefore our result can be seen as a generalization of the result of Ghadimi and Lan [7].

Based on the observations above, it is natural to ask whether it is possible to derive an upper bound such that the dominating $\mathcal{O}\left(1/\sqrt{T}\right)$ term does not depend on γ. We hope to study this interesting question in the future.