Improved Complexities for Stochastic Conditional Gradient Methods under Interpolation-like Conditions

Tesi Xiao* Krishnakumar Balasubramanian† Saeed Ghadimi‡

January 28, 2022

Abstract

We analyze stochastic conditional gradient methods for constrained optimization problems arising in over-parametrized machine learning. We show that one could leverage the interpolation-like conditions satisfied by such models to obtain improved oracle complexities. Specifically, when the objective function is convex, we show that the conditional gradient method requires $\mathcal{O}(\epsilon^{-2})$ calls to the stochastic gradient oracle to find an ϵ-optimal solution. Furthermore, by including a gradient sliding step, we show that the number of calls reduces to $\mathcal{O}(\epsilon^{-1.5})$.

Keywords: Stochastic Conditional Gradient, Oracle Complexity, Overparametrization, Zeroth-order Optimization.

1 Introduction

Consider the following constrained stochastic optimization problem:

$$\min_{x\in\Omega} \quad \{f(x) := \mathbb{E}_\xi\left[F(x,\xi)\right]\}, \tag{1}$$

where $f:\mathbb{R}^d\to\mathbb{R}$ and $\Omega\subset\mathbb{R}^d$ is a closed and convex set and ξ is a random vector characterizing the stochasticity in the problem. In a machine learning setup, the function F could be interpreted as the loss function associated with a sample ξ and the function f could represent the risk, which is defined as the expected loss. Such constrained stochastic optimization problems arise frequently in statistical machine learning applications. The conditional gradient algorithm, also called as the Frank-Wolfe algorithm, is an efficient method for solving constrained optimization problems of the form in (1) due to their projection-free nature [25, 19, 14, 27, 5, 35]. In each step of the conditional gradient method, it is only required to minimize a linear objective over the set Ω. This operation could be implemented efficiently for a variety of sets arising in statistical machine learning, compared to the operation of projecting onto the set Ω, which is required for example by the projected gradient method. Hence, the conditional gradient method has regained popularity in the last decade in the optimization and machine learning community.

There has been extensive work in the past decade on analyzing the stochastic conditional gradient algorithm for optimization problems of the form in (1); see for example [15, 22, 28, 36, 16].

*Department of Statistics, University of California, Davis. texiao@ucdavis.edu.

†Corresponding author. One Shields Avenue, Department of Statistics, University of California, Davis, CA 95616 kbala@ucdavis.edu.

‡Department of Management Sciences, University of Waterloo. sghadimi@uwaterloo.ca.

However, existing works do not take into account certain favorable structures that are naturally available in modern over-parametrized machine learning problems. Specifically, it has been noted that modern machine learning models predict well on unseen data, despite fitting the training data perfectly [44, 24, 29, 30, 32, 21]. Examples include logistic regression or support vector machine with squared-hinge loss that are trained with linearly separable data [41, 42, 31] and deep neural networks [41, 3]. From an optimization point of view, for the problem in (1) with $\Omega \equiv \mathbb{R}^d$, the above interpolation condition means that at the optimal point, the gradient is not only zero (or close to zero) with respect to the risk function f but is also almost surely equal to zero for the random loss function F. Such a scenario helps to reduce the stochasticity in the gradient estimation process which in turn results in improved complexity results for several stochastic optimization procedures. Indeed in the recent past, several works have provided improved rates for algorithms like stochastic gradient descent [33, 30, 3, 18, 41, 42] and sub-sampled Newton's method [31]. In particular, for several settings, the above works demonstrate that the stochastic algorithm may perform as well as the corresponding deterministic counterpart. However, such works only study unconstrained optimization problems and do not have any consequences for constrained stochastic optimization problems of the form in (1).

Hence, in this work we consider the following question: *Can we obtain improvements in the oracle complexity of algorithms used for projection-free constrained stochastic optimization problems arising in the context of over-parametrized machine learning models, that are capable of perfectly interpolating the training data?* We give a positive answer to the above question by demonstrating that the stochastic conditional gradient method, a projection-free technique for solving constrained stochastic optimization problems, also enjoy improved oracle complexities when they are used to solve constrained stochastic optimization problems of the form in (1) under certain *interpolation-like* conditions. We elaborate on the specific form of improvement observed below. For stochastic conditional gradient algorithms, the oracle complexity is measured in terms of number of calls to the Stochastic First-order Oracle (SFO) and the Linear Minimization Oracle (LMO) used to the solve the subproblems (that are of the form of minimizing a linear function over the convex feasible set) arising in the algorithm. In this work, we make the following contribution to the literature on conditional gradient methods under interpolation-like assumptions (see Section 2 for the exact definitions) on the stochastic gradient:

1. For the case of convex f in (1), we show that the number of calls to the SFO for the *vanilla* stochastic conditional gradient method and stochastic conditional gradient sliding methods are given respectively by $\mathcal{O}(\epsilon^{-2})$ and $\mathcal{O}(\epsilon^{-1.5})$. For comparison, without such assumptions, the corresponding complexities are $\mathcal{O}(\epsilon^{-3})$ and $\mathcal{O}(\epsilon^{-2})$ respectively. The number of calls to the linear minimization oracle (LMO) is of the order $\mathcal{O}(\epsilon^{-1})$, in both cases.
2. We also demonstrate similar improvements in the context of zeroth-order conditional gradient methods, where one only observes noisy evaluations of the function being optimized. Specifically, the number of calls to the stochastic zeroth-order oracle for the *vanilla* stochastic conditional gradient method and stochastic conditional gradient sliding methods are given respectively by $\mathcal{O}(d\epsilon^{-2})$ and $\mathcal{O}(d\epsilon^{-1.5})$, with the same LMO complexity as the first-order setting.

We emphasize that, notably the above improvements are achieved without incorporating any double-loop based existing variance reduction techniques, for example SVRF [36] or SPIDER-FW [43]. It is also worth noting that [9, 39] argue that variance reduction techniques (at the least existing ones) are ineffective in the context of modern deep learning models which are invariably over-parametrized. We also remark that, in contrast to stochastic gradient methods for unconstrained optimization [41, 3], the above improved results still do not match the corresponding deterministic rates highlighting the subtlety with projection-free optimization.

2 Preliminaries and Assumptions

We now list and discuss the set of assumptions made in our work. We first list some regularity assumptions on the function f and the set Ω.

Assumption 1. *The function f has L-Lipschitz gradient ∇f, i.e., for any pair of points $x, y \in \Omega$, we have $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$, and the feasible set $\Omega \subset \mathbb{R}^d$ is bounded, i.e., $\max_{x,y\in\Omega} \|x - y\| \leq D$.*

The above set of assumptions are standard in the analysis of stochastic conditional gradient methods and has been used in prior works in the literature; see for example [17]. We make the above assumptions for both the first-order setting. We also require the following smoothness assumption in the zeroth-order setting.

Assumption 2. *The function F has Lipschitz continuous gradient with constant L, almost surely for any ξ, i.e., for any $x, y \in \mathbb{R}^d$, i.e., almost surely we have $\|\nabla F(x, \xi) - \nabla F(y, \xi)\| \leq L \|x - y\|$.*

Note that the above assumption is stronger than the first statement of Assumption 1 and implies it. However, we only use Assumption 2 for the analysis of zeroth-order algorithms.

2.1 Growth Conditions in the Convex Constrained Setting

We now state the main interpolation-like assumptions that we make in our work when f is convex and provide the main intuition behind such an assumption.

Assumption 3 (Moment-based Weak Growth Condition)**.** *Let x^* be the minimum point of f. We say that f satisfies the Moment-based Weak Growth Condition (WGC) with constant ρ, if for any point $x \in \Omega$, we have*

$$\mathbb{E}_\xi \|\nabla F(x, \xi)\|^2 \leq 2\rho L [f(x) - f(x^*)]. \tag{2}$$

Assumption 4 (Variance-based Weak Growth Condition)**.** *Let x^* be the minimum point of f. We say that the function f satisfies the Variance-based Weak Growth Condition (WGC) with constant ρ, if for any point $x \in \Omega$, we have*

$$\mathbb{E}_\xi \|\nabla F(x, \xi) - \nabla f(x)\|^2 \leq 2\rho L [f(x) - f(x^*)]. \tag{3}$$

The above conditions are motivated by the so-called strong growth condition: $\mathbb{E}\|\nabla F(x, \xi)\|^2 \leq \rho\|\nabla f(x)\|^2$, used in [41] for obtaining faster rates of convergence for stochastic gradient method in the unconstrained setting. Notice that in the interpolation setting, when $\nabla f(x^*) = 0$, we have $\nabla F(x^*, \xi) = 0$, almost surely. Thus, the strong growth condition is defined exactly to take advantage of this situation. Furthermore, in the smooth convex setting, [41] showed that the strong-growth condition is equivalent to the moment-based weak growth condition in Assumption 3. However, the moment-based weak growth condition as proposed in [41] is not directly suited for the constrained stochastic setting that we consider in this work. It is easy to construct examples for which there exists stationary point at the boundary of Ω with non-zero (stochastic) gradient, i.e., $\mathbb{E}\|\nabla F(x, \xi)\|^2$ could remain positive while the right hand side goes to 0 and hence the assumption is not satisfied. In order to resolve this issue, for the constrained setting, we relax the moment-based growth conditions to the variance-based versions. Note that we have

$$\mathbb{E}\|\nabla F(x, \xi) - \nabla f(x)\|^2 = \mathbb{E}\|\nabla F(x, \xi)\|^2 - \|\nabla f(x)\|^2 \leq \mathbb{E}\|\nabla F(x, \xi)\|^2.$$

Thus variance-based growth conditions naturally become the substitute for the moment-based version in constrained problems and could hold even the moment-based conditions do not hold. As they are

also motivated by the interpolation assumption, we refer to these conditions as interpolation-like conditions. Formally, under the variance-based growth conditions for a convex f, if we attain an optimal point $x^* \in \Omega$, the variance of the stochastic first-order oracle will be almost surely zero, i.e., $\nabla F(x^*, \xi) = \nabla f(x^*)$ almost surely. This property eventually leads to the improvements in the query complexity that we demonstrate. We emphasize that it is natural to construct counter-examples that violate Assumption 4. In those cases, the improved query complexities that we demonstrate are simply not applicable. Finally, we also have the following natural relationships between the two conditions.

Proposition 1. *The Weak Growth Conditions defined above have the following relations:*
(a) If f satisfies the Moment-based WGC (3) with ρ, then f satisfies the Variance-based WGC (4) with ρ and there exists $x^* \in \Omega$ such that $\nabla f(x^*) = 0$.
(b) If f satisfies the Variance-based WGC (4) with ρ and there exists $x^* \in \Omega$ such that $\nabla f(x^*) = 0$, then f satisfies the Moment-based WGC (3) with $\rho + 1$.

2.2 Growth Conditions in the Zeroth-Order Constrained Setting

In the zeroth-order setting, we only assume availability of the noisy function evaluations. This oracle setting is motivated by several applications where only noisy function queries of problem (1) is available, such as reinforcement learning [38, 7, 8], hyperparameter tuning [40], and black-box attacks to deep networks [6, 37]. Hence, we use the Gaussian Stein's identity based random gradient estimator, a standard gradient estimator in the zeroth-order optimization literature [17, 12, 34, 2]:

$$\bar{G}_\nu(x) = \frac{1}{b} \sum_{j=1}^{b} \frac{F(x + \nu u_j, \xi_j) - F(x, \xi_j)}{\nu} u_j,$$

where $u_1, \ldots, u_b$ are i.i.d. samples from $\mathsf{N}(0, \mathbf{I}_d)$. The above gradient estimator is a biased estimator of the true gradient $\nabla f(x)$, and was also used in [2], to develop zeroth-order conditional gradient descent algorithms.

While for the first-order setting, we use the relatively weaker variance-based conditions to obtain the improved bounds, in the zeroth-order setting, it turns out the stronger moment-based conditions are required. The reason is that the mean square error of the biased zeroth-order gradient estimator is bounded above by $\mathbb{E}\|\nabla F(x, \xi)\|^2$. Hence, to obtain improved rates, it makes it necessary to make assumptions on the moments of the stochastic gradient directly. We emphasize that this is required only for the constrained problems, since the moment-based conditions are equivalent to the variance-based conditions when there exists one zero-gradient point in the constraint set (see Proposition 1). In particular, we show in Appendix C that a zeroth-order version of Theorem 3 from [41], for stochastic gradient descent, to bound the gradient size in the nonconvex setting could be proved just under the variance-based growth conditions.

2.3 Motivating Examples

Before we present our main results in the next section, we briefly discuss some motivating examples of constrained stochastic optimization problems that arise in modern machine learning. In the convex setting, it is easy to see that kernel regression [29], squared-Hinge loss based linear SVM classifier or logistic regression on linearly separable data could be considered as operating in the over-parametrized regime and hence satisfy interpolation-like conditions [41, 31].

However, without any constraints, such predictors might be biased against certain sensitive features like race or gender. One way to build fair predictors is to explicitly encode fairness

constraints with respect to certain pre-defined sensitive features [11, 1]. Specifically, it was shown in [1] that several standard and well-accepted notions of fairness in classification setting, including equalized odds [20], demographic parity [13], balance for the negative class [26], treatment equality [4] could be formulated as empirical risk minimization problems subjected linear inequality constraints. In this case, the problem is exactly of the form in (1) with Ω being a polytope. Furthermore, [11] also proposed a general approach for fair empirical risk minimization. Similar to [1], the fundamental idea is to enforce constraints such that the conditional risk of a predictor is not varying much with respect to the sensitive features associated with the problem. Such formulations of fair empirical risk minimization in the interpolation regime also fall under the class of problems in (1).

Squared hinge loss with linearly separable data. As a concrete example, we extend the unconstrained examples presented in [41] to the constrained setting we consider. Assuming a finite support of features and the linearly separable data, it has been shown that the squared-hinge loss satisfies SGC with $\rho = c/\tau^2$ where c is the cardinality of the support and τ is the margin (Lemma 1 in [41]). In the above regime, the optimal classifier that minimizes the loss and achieves a stationary point with zero gradient is not always unique. In practice, to construct a fair classifier, enforcing constraints is a natural approach. Note that if there exists an $x^* \in \Omega$, by the convexity and the L-smoothness of f, we have

$$\|\nabla f(x)\|^2 \leq 2L(f(x) - f(x^*)). \tag{4}$$

That is to say, for linearly separable data with margin τ and a finite support of size c, if there exists one $x^* \in \Omega$, the squared-hinge loss satisfies Assumption 3 with $\rho = c/\tau^2$.

3 Improved Complexities for Stochastic Conditional Gradient Methods

We now provide improved complexities for stochastic conditional gradient methods under the interpolation-like assumption in Section 2. For convenience, we first introduce the following mini-batch stochastic gradients with first-order and zeroth-order oracle access: at t-th iteration, we uniformly pick i.i.d. samples $\{\xi_{t,1}, \ldots, \xi_{t,b_t}\}$ and estimate the gradient by

$$\tilde{\nabla}_t := \frac{1}{b_t}\sum_{i=1}^{b_t} \nabla F(x_{t-1}, \xi_{t,i}), \quad \bar{G}^t_\nu := \frac{1}{b_t}\sum_{j=1}^{b_t} \frac{F(x_{t-1} + \nu u_{t,j}, \xi_{t,j}) - F(x_{t-1}, \xi_{t,j})}{\nu} u_{t,j}$$

where $u_{t,1}, \ldots, u_{t.b_t}$ are i.i.d. samples from $\mathsf{N}(0, \mathbf{I}_d)$.

3.1 Stochastic Frank-Wolfe

In this section, we studied the oracle complexity of the vanilla stochastic Frank-Wolfe algorithm under the weak interpolation-like conditions in Assumption 4 and 3.

Theorem 2. *Consider solving problem* (1)*, by Algorithm 1, under Assumption 1 with f being convex.*

(a) Assuming access to stochastic first-order oracle, under Assumption 4, setting

$$\gamma_t = \frac{4}{t+3}, \quad b_t = \lceil (t+3)/2 \rceil,$$

we have the following convergence rate:

$$\mathbb{E}[f(x_t) - f(x^*)] \leq \frac{2(f(x_0) - f(x^*)) + 8(\rho+1)LD^2}{t+3}.$$

Algorithm 1 Stochastic Frank-Wolfe

Input: $x_0 \in \Omega$, number of iterations T, $\gamma_t \in [0,1]$, minibatch size b_t
for $t = 1, 2, \ldots, T$ **do**
  Compute the gradient g_t as follows:
    Set $g_t = \tilde{\nabla}_t$ (for the first-order setting).
    Set $g_t = \bar{G}^t_\nu$ (for the zeroth-order setting).
  Compute $d_t = \operatorname{argmin}_{d\in\Omega} \langle d, g_t \rangle$
  $x_t = x_{t-1} + \gamma_t(d_t - x_{t-1})$
end for
Output: x_T

Hence, the total number of calls to the stochastic first-order oracle and linear minimization oracle required to be solved to find an ϵ-optimal point of problem (1) *are, respectively, bounded by*

$$\mathcal{O}\left(\epsilon^{-2}\right), \quad \mathcal{O}\left(\epsilon^{-1}\right).$$

(b) Assuming access to stochastic zeroth-order oracle, under Assumptions 3 and 2, setting

$$\gamma_t = \frac{4}{t+3}, \quad b_t = (t+3)(d+4), \quad \nu = \frac{D}{(T+3)(d+6)^{3/2}}$$

we have

$$\mathbb{E}[f(x_t) - f(x^*)] \leq \frac{2(f(x_0) - f(x^*)) + 8(\rho + \rho^{-1} + 1)LD^2}{t+3}.$$

Hence, the total number of calls to the stochastic zeroth-order oracle and linear minimization oracle required to be solved to find an ϵ-optimal point of problem (1) *are, respectively, bounded by*

$$\mathcal{O}\left(d\epsilon^{-2}\right), \quad \mathcal{O}\left(\epsilon^{-1}\right).$$

The above oracle complexities in the first-order setting, match the results obtained by [43, 45]. However, the above works require double-loop based variance reduction techniques which in turn require the stronger mean-square gradient-Lipschitz assumption. Furthermore, the use of the variance reduction technique results in the increased wall-clock running time of the algorithm. Our result here is applicable to the vanilla version of the stochastic conditional gradient method, as long as the problem satisfies the interpolation-like conditions observed in modern machine learning problems.

3.2 Stochastic Conditional Gradient Sliding

In this section, we analyze the complexity of the stochastic gradient sliding (SCGS) algorithm under the weak growth condition. The SCGS was first proposed and thoroughly analyzed in [28]. It is a fundamental modification of the conditional gradient algorithm that achieved improved oracle complexities without relying on any variance reduction techniques. Below, we show that under the interpolation-like assumptions in Section 2, the oracle complexity of the SCGS could be further improved compared in both the first-order and zeroth-order methods.

Algorithm 2 SCGS:
Stochastic Conditional Gradient Sliding

Input: $x_0 \in \Omega$, T, $\beta_t \in \mathbb{R}_+$, $\gamma_t \in [0,1]$, b_t, $y_0 = x_0$
for $t = 1, 2, \ldots, T$ **do**
  Set $z_t = (1-\gamma_t)x_{t-1} + \gamma_t y_{t-1}$
  Compute the gradient g_t as follows:
    Set $g_t = \tilde{\nabla}_t$ (first-order).
    Set $g_t = \bar{G}^t_\nu$ (zeroth-order).
  Solve
    $y_t = \mathrm{ICG}(g_t, y_{t-1}, \beta_t, \eta_t)$
  by Algorithm 3
  Set $x_t = (1-\gamma_t)x_{t-1} + \gamma_t y_t$
end for
Output: x_T

Algorithm 3 ICG:
Inexact Conditional Gradient Method

Input: $g, u, \beta, \eta, u_1 = u, k = 1$
1. Let v_k be an optimal solution for the subproblem

$$\max_{v\in\Omega}\{h_k(v) = \langle g + \beta(u_k - u), u_k - v\rangle\}. \quad (5)$$

2. If $h_k(v_k) \le \eta$, terminate and output u_k.
3. $u_{k+1} = (1-\alpha_k)u_k + \alpha_k v_k$ with

$$\alpha_k = \min\left\{1, \frac{\langle \beta(u-u_k) - g, v_t - u_t\rangle}{\beta\|v_k - u_k\|^2}\right\}.$$

4. Set $k \leftarrow k+1$ and go to step 1.

Theorem 3. *Consider solving problem* (1)*, by Algorithm 2, under Assumption 1 with f being convex.*

(a) Assuming access to stochastic first-order oracle, under Assumption 4, setting

$$\beta_t = \frac{4L}{t+2}, \quad \gamma_t = \frac{3}{t+2}, \quad \eta_t = \frac{LD^2}{t(t+1)}, \quad b_t = \lceil 3\rho t(t+1)\rceil$$

we have

$$\mathbb{E}[f(x_t) - f(x^*)] \le \frac{6LD^2}{(t+2)^2} + \frac{15LD^2 + 3\|\nabla f(x^*)\|D}{(t+1)(t+2)}.$$

Hence, the total number of calls to the stochastic first-order oracle and linear minimization oracle required to be solved to find an ϵ-optimal point of problem (1) *are, respectively, bounded by*

$$\mathcal{O}\left(\epsilon^{-1.5}\right), \quad \mathcal{O}\left(\epsilon^{-1}\right).$$

(b) Assuming access to stochastic zeroth-order oracle, in addition, with Assumption 3, 2, setting

$$\beta_t = \frac{4L}{t+2}, \quad \gamma_t = \frac{3}{t+2}, \quad \eta_t = \frac{LD^2}{t(t+1)}, \quad b_t = \lceil 6\rho(d+4)t(t+1)\rceil, \quad \nu = \frac{D}{(T+2)^2(d+6)^{3/2}},$$

we have

$$\mathbb{E}[f(x_t) - f(x^*)] \le \frac{8LD^2}{(t+2)^2} + \frac{32LD^2}{(t+1)(t+2)}.$$

Hence, the total number of calls to the stochastic zeroth-order oracle and linear minimization oracle required to be solved to find an ϵ-optimal point of problem (1) *are, respectively, bounded by*

$$\mathcal{O}\left(d\epsilon^{-1.5}\right), \quad \mathcal{O}\left(\epsilon^{-1}\right).$$

To the best of our knowledge, the above complexity of $\mathcal{O}(\epsilon^{-1.5})$ is not achieved for any variance reduced versions of stochastic Frank-Wolfe methods. This improvement is solely obtained by the SCGS algorithm of [28] under the interpolation-like assumptions which are natural in modern machine learning problems, without any variance reduction methods. We also highlight that, in the unconstrained setting, the stochastic gradient method performs as well as its deterministic counterpart. However, the above result still falls short of the corresponding deterministic complexity of conditional gradient sliding, which is of the order $\mathcal{O}(\epsilon^{-0.5})$ [28]. This highlights the intrinsic difficulty associated with projection-free methods for constrained stochastic optimization problems.

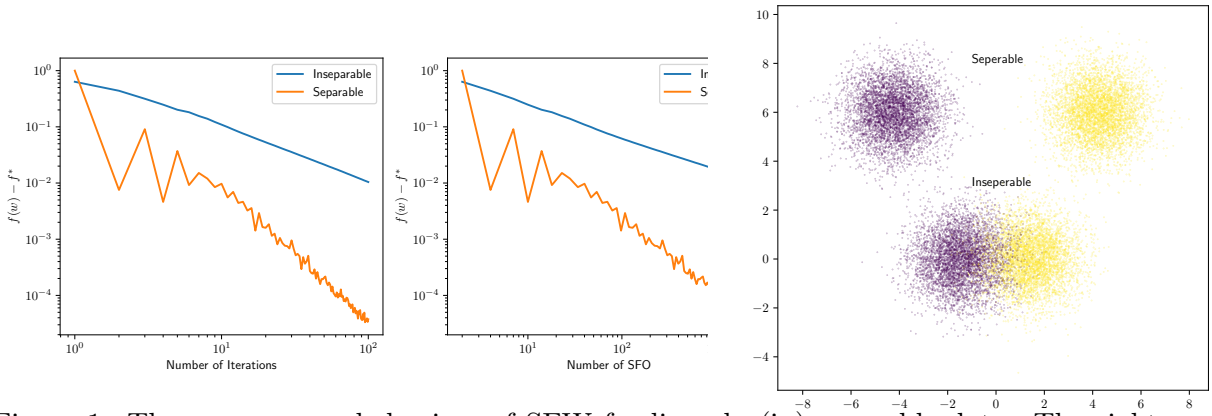

Figure 1: The convergence behaviors of SFW for linearly (in)-separable data. The right panel visualizes the first 2 dimensions of the synthetic data used for numerical analyses.

4 Experiments

We generate synthetic binary classification datasets with two isotropic Gaussian blobs symmetric with respect to the origin, with the sample size $n = 100,000$ and the dimension $d = 500$. We ensure that two blobs are linearly separable with a positive margin for one dataset while the other has an overlap. We seek to find a hyperplane $w^\top x$ that minimizes the squared-hinge loss $f(w) = \frac{1}{n}\sum_{i=1}^n f_i(w) = \frac{1}{n}\sum_{i=1}^n \max(0, 1 - y_i \cdot w^\top x_i)^2$ satisfying the constraint $\|w\|_1 \leq 1$. Note that $f(w)$ satisfies the weak growth condition for linearly separable data in view of sampling only a mini-batch of gradient (with replacement) in each iteration, and the parameter $\rho = L_{\max}/L$; see Proposition 2 in [41], and $L_{\max}$ is the largest Lipschitz constant for $\nabla f_i(w)$. In Figure 1, we plot the suboptimality $f(w) - f^*$ versus the number of iterations and the number of calls to the SFO. The results are obtained by averaging over 100 runs with random initialization w_0. We observe that SFW converges essentially faster for linearly separable data than the inseparable case.

5 Discussions

We briefly discuss extensions of our results to the nonconvex setting. Our proposed assumption is motivated by the notion of Frank-Wolfe gap [10, 23], which is defined as $\mathcal{G}_f(x) = \max_{y\in\Omega}\langle \nabla f(x), x - y\rangle$. With this, a nonconvex function f satisfies Constrained Growth Condition with constant ρ, if for any point $x \in \Omega$, $\mathbb{E}_\xi \|\nabla F(x,\xi) - \nabla f(x)\|^2 \leq 2\rho L \mathcal{G}_f(x)$. Note that if f is convex, then $\mathcal{G}_f(x) \geq f(x) - f(x^*)$. Hence, this generalizes Assumption 4 defined for the convex setting. Under this assumption in the nonconvex setting, it could be shown that the vanilla stochastic Frank-Wolfe algorithm can find an ϵ-stationary point of the problem within at most $\mathcal{O}\left(1/\epsilon^3\right)$ and $\mathcal{O}\left(1/\epsilon^2\right)$ number of calls to the SFO linear subproblem solver, respectively. However, although existence of functions satisfying the above asssumption could be shown, it is not clear if practical nonconvex functions appearing in machine learning context satisfy it. It would be extremely interesting to examine this as future work.

Acknowledgements

TX and KB were partially supported by a seed grant from the Center for Data Science and Artificial Intelligence Research, UC Davis and NSF TRIPODS Grant-1934568.

References

[1] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69, 2018.

[2] K. Balasubramanian and S. Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, pages 1–42, 2021.

[3] R. Bassily, M. Belkin, and S. Ma. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

[4] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 2018.

[5] L. Berrada, A. Zisserman, and M. P. Kumar. Deep frank-wolfe for neural network optimization. *arXiv preprint arXiv:1811.07591*, 2018.

[6] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26. ACM, 2017.

[7] K. Choromanski, M. Rowland, V. Sindhwani, R. Turner, and A. Weller. Structured evolution with compact architectures for scalable policy optimization. *arXiv preprint arXiv:1804.02395*, 2018.

[8] K. Choromanski, M. Rowland, V. Sindhwani, R. Turner, and A. Weller. Structured evolution with compact architectures for scalable policy optimization. In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018.

[9] A. Defazio and L. Bottou. On the ineffectiveness of variance reduced optimization for deep learning. In *Advances in Neural Information Processing Systems*, pages 1753–1763, 2019.

[10] V. Demyanov and A. Rubinov. *Approximate methods in optimization problems*. American Elsevier Publishing Co, 1970.

[11] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

[12] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[13] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133, 2018.

[14] R. M. Freund, P. Grigas, and R. Mazumder. An extended frank-wolfe method with in-face directions, and its application to low-rank matrix completion. *SIAM Journal on optimization*, 27(1):319–346, 2017.

[15] D. Garber and E. Hazan. A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization. *arXiv preprint arXiv:1301.4666*, 2013.

[16] S. Ghadimi. Conditional gradient type methods for composite nonlinear and stochastic optimization. *Mathematical Programming*, 173(1-2):431–464, 2019.

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

[18] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik. Sgd: General analysis and improved rates. *arXiv preprint arXiv:1901.09401*, 2019.

[19] Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1-2):75–112, 2015.

[20] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[21] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[22] E. Hazan and H. Luo. Variance-reduced and projection-free stochastic optimization. In *International Conference on Machine Learning*, pages 1263–1271, 2016.

[23] D. W. Hearn. The gap function of a convex program. *Operations Research Letters*, 1(2):67–71, 1982.

[24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[25] M. Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

[26] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[27] G. Lan, S. Pokutta, Y. Zhou, and D. Zink. Conditional accelerated lazy stochastic gradient descent. In *International Conference on Machine Learning*, pages 1965–1974, 2017.

[28] G. Lan and Y. Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization*, 26(2):1379–1409, 2016.

[29] T. Liang and A. Rakhlin. Just interpolate: Kernel" ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387*, 2018.

[30] S. Ma, R. Bassily, and M. Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334, 2018.

[31] S. Y. Meng, S. Vaswani, I. Laradji, M. Schmidt, and S. Lacoste-Julien. Fast and furious convergence: Stochastic second order methods under interpolation. *arXiv preprint arXiv:1910.04920*, 2020.

[32] A. Montanari, F. Ruan, Y. Sohn, and J. Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

[33] D. Needell, R. Ward, and N. Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in neural information processing systems*, pages 1017–1025, 2014.

[34] Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[35] S. N. Ravi, T. Dinh, V. S. R. Lokhande, and V. Singh. Constrained deep learning using conditional gradient and applications in computer vision. *arXiv preprint arXiv:1803.06453*, 2018.

[36] S. J. Reddi, S. Sra, B. Póczos, and A. Smola. Stochastic frank-wolfe methods for nonconvex optimization. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1244–1251. IEEE, 2016.

[37] A. K. Sahu, M. Zaheer, and S. Kar. Towards gradient free and projection free stochastic optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3468–3477, 2019.

[38] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

[39] M. Schmidt. Faster algorithms for deep learning? (presentation in vector institute: https://www.cs.ubc.ca/ schmidtm/documents/2020_vector_smallresidual.pdf), 2020.

[40] J. Snoek, H. Larochelle, and R. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[41] S. Vaswani, F. Bach, and M. Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204, 2019.

[42] S. Vaswani, A. Mishkin, I. Laradji, M. Schmidt, G. Gidel, and S. Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. In *Advances in Neural Information Processing Systems*, pages 3727–3740, 2019.

[43] A. Yurtsever, S. Sra, and V. Cevher. Conditional gradient methods via stochastic path-integrated differential estimator. In *Proceedings of the International Conference on Machine Learning-ICML 2019*, 2019.

[44] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[45] M. Zhang, Z. Shen, A. Mokhtari, H. Hassani, and A. Karbasi. One sample stochastic frank-wolfe. *arXiv preprint arXiv:1910.04322*, 2019.

SUPPLEMENTARY DOCUMENT

A Proof for Theorem 2

In order to prove Theorem 2, we require the following result from [34] for the zeroth-order case.

Lemma 4. [34] *Let the function f has lipschitz continuous gradient with constant L. Consider the smoothed function $f_\nu(x) = \mathbb{E}_u[f(x+\nu u)]$ where $u \sim \mathsf{N}(0, \mathbf{I}_d)$. Then for any $x \in \mathbb{R}^d$,*

$$\mathbb{E}_u\left[\frac{f(x+\nu u)-f(x)}{\nu}u\right] = \nabla f_\nu(x) \tag{6}$$

$$\|\nabla f_\nu(x) - \nabla f(x)\| \leq \frac{\nu}{2}L(d+3)^{\frac{3}{2}} \tag{7}$$

$$\frac{1}{\nu^2}\mathbb{E}_u[\{f(x+\nu u)-f(x)\}^2\|u\|^2] \leq \frac{\nu^2}{2}L^2(d+6)^3 + 2(d+4)\|\nabla f(x)\|^2. \tag{8}$$

We now present the lemma below to bound the mean squared error for the zeroth-order gradient estimator.

Lemma 5. *Under Assumption 1, 2, 3, we have*

$$\mathbb{E}\left\|\bar{G}_\nu^t - \nabla f_\nu(x_{t-1})\right\|^2 \leq \frac{4\rho L(d+4)(f(x_{t-1})-f(x^*))}{b_t} + \frac{\nu^2L^2(d+6)^3}{2b_t}, \tag{9}$$

$$\mathbb{E}\left\|\bar{G}_\nu^t - \nabla f(x_{t-1})\right\|^2 \leq \frac{4\rho L(d+4)(f(x_{t-1})-f(x^*))}{b_t} + \nu^2L^2(d+6)^3. \tag{10}$$

Proof. First note that by (6), we have

$$\mathbb{E}_{u,\xi}[\bar{G}_\nu^t] = \mathbb{E}_{u,\xi}[G_{t,j}] = \mathbb{E}_u\left[\frac{f(x_{t-1}+\nu u)-f(x_{t-1})}{\nu}u\right] = \nabla f_\nu(x_{t-1}),$$

Then by using (8) for F instead of f, under Assumption 2, 3, we can obtain

$$\begin{aligned}
\mathbb{E}_{u,\xi}\left\|\bar{G}_\nu^t - \nabla f_\nu(x_{t-1})\right\|^2 &= \frac{1}{b_t}\mathbb{E}_{u,\xi}\|G_{t,j} - \nabla f_\nu(x_{t-1})\|^2 \\
&\leq \frac{1}{b_t}\mathbb{E}_{u,\xi}\|G_{t,j}\|^2 \\
&\leq \frac{2(d+4)}{b_t}\mathbb{E}_\xi\|\nabla F(x_{t-1},\xi_{t,j})\|^2 + \frac{\nu^2L^2(d+6)^3}{2b_t} \\
&\leq \frac{4\rho L(d+4)(f(x_{t-1})-f(x^*))}{b_t} + \frac{\nu^2L^2(d+6)^3}{2b_t}
\end{aligned}$$

where the first inequality comes from the fact that the variance is less than the seocond moment.

To prove (10), we decompose the mean squared error into the bias and the variance by utilizing the results (9) and (7), i.e.,

$$\begin{aligned}
\mathbb{E}\left\|\bar{G}_\nu^t - \nabla f(x_{t-1})\right\|^2 &= \mathbb{E}\left\|\bar{G}_\nu^t - \nabla f_\nu(x_{t-1})\right\|^2 + \|\nabla f_\nu(x_{t-1}) - \nabla f(x_{t-1})\|^2 \\
&\leq \frac{4\rho L(d+4)(f(x_{t-1})-f(x^*))}{b_t} + \frac{\nu^2L^2(d+6)^3}{2b_t} + \frac{\nu^2L^2(d+3)^3}{4} \\
&\leq \frac{4\rho L(d+4)(f(x_{t-1})-f(x^*))}{b_t} + \nu^2L^2(d+6)^3.
\end{aligned}$$

□

We also need the following simple result in our proof.

Lemma 6. *Assume that sequences* $\{\phi_t\}_{t\geq 0} \geq 0$, $\{B_t\}_{t\geq 1}$, $\{\theta_t\}_{t\geq 1} \in [0,1]$ *are given such that*

$$\phi_t \leq (1-\theta_t)\phi_{t-1} + B_t. \tag{11}$$

Then, we have

$$\phi_T \leq \Theta_T \left[\phi_0 + \sum_{t=1}^{T} \frac{B_t}{\Theta_t}\right],$$

where, for any $t \geq 2$,

$$\Theta_t = \Theta_1 \prod_{k=2}^{t}(1-\theta_k), \quad \textit{where} \;\; \Theta_1 = 1-\theta_1 \;\; \textit{if} \;\; \theta_1 < 1, \quad \Theta_1 = 1 \;\; \textit{if} \;\; \theta_1 = 1. \tag{12}$$

Proof. Dividing both sides of (11) by Θ_t, summing them up from $t=1$ to $t=T$, noting non-negativity of ϕ_t and (12), we obtain the result. □

Proof. [Proof for Theorem 2] For convenience, let g_t be the gradient estimator at t step. Thus, $g_t = \tilde{\nabla}_t$ for the first order method while in the zeroth order setting $g_t = \bar{G}^t_\nu$.

$$\begin{aligned}
f(x_t) &\leq f(x_{t-1}) + \langle \nabla f(x_{t-1}), x_t - x_{t-1}\rangle + \frac{L}{2}\|x_t - x_{t-1}\|^2 \\
&= f(x_{t-1}) + \gamma_t \langle \nabla f(x_{t-1}), d_t - x_{t-1}\rangle + \frac{L\gamma_t^2}{2}\|d_t - x_{t-1}\|^2 \\
&\leq f(x_{t-1}) + \gamma_t \langle g_t, d_t - x_{t-1}\rangle + \gamma_t \langle \nabla f(x_{t-1}) - g_t, d_t - x_{t-1}\rangle + \frac{LD^2\gamma_t^2}{2} \\
&\leq f(x_{t-1}) + \gamma_t \langle g_t, x^* - x_{t-1}\rangle + \gamma_t \langle \nabla f(x_{t-1}) - g_t, d_t - x_{t-1}\rangle + \frac{LD^2\gamma_t^2}{2} \\
&= f(x_{t-1}) + \gamma_t \langle \nabla f(x_{t-1}), x^* - x_{t-1}\rangle + \gamma_t \langle \nabla f(x_{t-1}) - g_t, d_t - x^*\rangle + \frac{LD^2\gamma_t^2}{2} \\
&\leq f(x_{t-1}) + \gamma_t (f(x^*) - f(x_{t-1})) + \gamma_t \langle \nabla f(x_{t-1}) - g_t, d_t - x^*\rangle + \frac{LD^2\gamma_t^2}{2} \\
&\leq f(x_{t-1}) + \gamma_t (f(x^*) - f(x_{t-1})) + \frac{\gamma_t}{2\beta}\|\nabla f(x_{t-1}) - g_t\|^2 + \frac{D^2\gamma_t(L\gamma_t + \beta)}{2}.
\end{aligned}$$

The last inequality comes from the Young's inequality: for any $\beta > 0$,

$$\begin{aligned}
\langle \nabla f(x_{t-1}) - g_t, d_t - x^*\rangle &\leq \frac{1}{2\beta}\|\nabla f(x_{t-1}) - g_t\|^2 + \frac{\beta}{2}\|d_t - x^*\|^2 \\
&\leq \frac{1}{2\beta}\|\nabla f(x_{t-1}) - g_t\|^2 + \frac{D^2\beta}{2}.
\end{aligned}$$

Denote $\phi_t = f(x_t) - f(x^*)$. Substracting $f(x^*)$ from both sides of the inequality and taking the conditional expectation $\mathbb{E}[\cdot|\mathcal{F}_{t-1}]$, we have

$$\mathbb{E}[\phi_t|\mathcal{F}_{t-1}] \leq (1-\gamma_t)\phi_{t-1} + \frac{\gamma_t}{2\beta}\mathbb{E}[\|\nabla f(x_{t-1}) - g_t\|^2 |\mathcal{F}_{t-1}] + \frac{D^2\gamma_t(L\gamma_t+\beta)}{2}. \tag{13}$$

For the first-order gradient estimator $g_t = \tilde{\nabla}_t$, we have the following bound for its variance under Assumption 4:

$$\mathbb{E}[\left\|\nabla f(x_{t-1}) - \tilde{\nabla}_t\right\|^2 |\mathcal{F}_{t-1}] = \frac{1}{b_t}\mathbb{E}[\|\nabla f(x_{t-1}) - \nabla F(x_{t-1}, \xi_{t,j})\|^2 |\mathcal{F}_{t-1}] \leq \frac{2\rho L \phi_{t-1}}{b_t}.$$

Then by (13), we can obtain

$$\mathbb{E}[\phi_t|\mathcal{F}_{t-1}] \leq (1-\gamma_t)\phi_{t-1} + \frac{\gamma_t \rho L}{\beta b_t}\phi_{t-1} + \frac{D^2\gamma_t(L\gamma_t+\beta)}{2}.$$

Let $\gamma_t = \frac{4}{t+3}, \beta = \rho L \gamma_t = \frac{4\rho L}{t+3} > 0, b_t = \lceil (t+3)/2 \rceil$, then

$$\mathbb{E}[\phi_t|\mathcal{F}_{t-1}] \leq \left(1 - \frac{2}{t+3}\right)\phi_{t-1} + \frac{8(\rho+1)LD^2}{(t+3)^2}. \tag{14}$$

Now, letting $\theta_t = \frac{2}{t+3}$, it is easy to check that $\Theta_t = \frac{6}{(t+2)(t+3)}$ due to (12). Hence, in the view of Lemma 6, we have

$$\mathbb{E}[\phi_t] \leq \frac{6\phi_0}{(t+2)(t+3)} + \frac{8(\rho+1)LD^2}{t+3} \leq \frac{2[\phi_0 + 4(\rho+1)LD^2]}{t+3}.$$

The above inequality implies that to attain an ϵ-optimal point, the total number of interations T can be bounded by $\mathcal{O}(1/\epsilon)$. Hence, the number of the gradient calls $\sum_{t=1}^{T} b_t$ can be bounded by $\frac{T^2+7T}{4} = \mathcal{O}(T^2)$, and the number of calls to the linear minimization oracle immediately follows from this observation.

We now prove part (b). For the zeroth-order version, by (10) in Lemma 5 and (13), we can obtain

$$\begin{aligned}\mathbb{E}[\phi_t|\mathcal{F}_{t-1}] &\leq (1-\gamma_t)\phi_{t-1} + \frac{\gamma_t}{2\beta}\mathbb{E}[\left\|\nabla f(x_{t-1}) - \bar{G}_\nu^t\right\|^2 |\mathcal{F}_{t-1}] + \frac{D^2\gamma_t(L\gamma_t+\beta)}{2} \\ &\leq (1-\gamma_t)\phi_{t-1} + \frac{2\gamma_t\rho L(d+4)}{\beta b_t}\phi_{t-1} + \frac{\gamma_t\nu^2L^2(d+6)^3}{2\beta} + \frac{D^2\gamma_t(L\gamma_t+\beta)}{2}\end{aligned}$$

Let $\gamma_t = \frac{4}{t+3}, \beta = \gamma_t\rho L, b_t = (t+3)(d+4), \nu = D(T+3)^{-1}(d+6)^{-3/2} \leq D(t+3)^{-1}(d+6)^{-3/2}$, then we have

$$\mathbb{E}[\phi_t|\mathcal{F}_{t-1}] \leq \left(1 - \frac{2}{t+3}\right)\phi_{t-1} + \frac{8(\rho+\rho^{-1}+1)}{(t+3)^2LD^2}$$

Similarly, in the view of Lemma 6, we obtain

$$\mathbb{E}[f(x_t) - f(x^*)] \leq \frac{2[f(x_0) - f(x^*)] + 8(\rho+\rho^{-1}+1)LD^2}{t+3}.$$

The above inequality implies that to attain an ϵ-optimal point, the total number of interations T can be bounded by $\mathcal{O}(1/\epsilon)$. Hence, the number of calls to the zeroth-order oracles $2\sum_{t=1}^{T} b_t$ can be bounded by $(d+4)(T^2+7T) = \mathcal{O}(dT^2)$, and the number of calls to the linear minimization oracle immediately follows from this observation. □

B Proof of Theorem 3

Proof. [Proof of Theorem 3] For convenience, let g_t be the gradient estimator at t step. Thus, $g_t = \tilde{\nabla}_t$ for the first order method while in the zeroth order setting $g_t = \bar{G}^t_\nu$. First note that by the updates in Algorithm 2, the convexity and the smoothness of f, we have

$$\begin{aligned}
f(x_t) \leq & f(z_t) + \langle \nabla f(z_t), x_t - z_t \rangle + \frac{L}{2} \|x_t - z_t\|^2 \\
= & (1-\gamma_t)[f(z_t) + \langle \nabla f(z_t), x_{t-1} - z_t \rangle] + \gamma_t [f(z_t) + \langle \nabla f(z_t), y_t - z_t \rangle] + \frac{L\gamma_t^2}{2} \|y_t - y_{t-1}\|^2 \\
\leq & (1-\gamma_t) f(x_{t-1}) + \gamma_t [f(z_t) + \langle \nabla f(z_t), y_t - z_t \rangle] + \frac{L\gamma_t^2}{2} \|y_t - y_{t-1}\|^2 \\
= & (1-\gamma_t) f(x_{t-1}) + \gamma_t [f(z_t) + \langle \nabla f(z_t), y_t - z_t \rangle] + \frac{\beta_t \gamma_t}{2} \|y_t - y_{t-1}\|^2 \\
& - \frac{\gamma_t(\beta_t - L\gamma_t)}{2} \|y_t - y_{t-1}\|^2 . \qquad (15)
\end{aligned}$$

And by (5), we have

$$\langle g_t + \beta_t (y_t - y_{t-1}), y_t - x \rangle \leq \eta_t, \quad \forall x \in \Omega.$$

Let $x = x^*$ in the above inequality. Then we have

$$\begin{aligned}
\frac{1}{2} \|y_t - y_{t-1}\|^2 &= \frac{1}{2} \|y_{t-1} - x^*\|^2 - \langle y_{t-1} - y_t, y_t - x^* \rangle - \frac{1}{2} \|y_t - x^*\|^2 \\
&\leq \frac{1}{2} \|y_{t-1} - x^*\|^2 + \frac{1}{\beta_t} \langle g_t, x^* - y_t \rangle - \frac{1}{2} \|y_t - x^*\|^2 + \frac{\eta_t}{\beta_t}. \qquad (16)
\end{aligned}$$

Denoting $\delta_t = g_t - \nabla f(z_t)$ and combining (15) and (16), we obtain

$$\begin{aligned}
f(x_t) \leq & (1-\gamma_t) f(x_{t-1}) + \gamma_t f(x^*) + \gamma_t \langle \delta_t, x^* - y_t \rangle \\
& + \frac{\beta_t \gamma_t}{2} (\|y_{t-1} - x^*\|^2 - \|y_t - x^*\|^2) + \eta_t \gamma_t - \frac{\gamma_t}{2} (\beta_t - L\gamma_t) \|y_t - y_{t-1}\|^2 \\
= & (1-\gamma_t) f(x_{t-1}) + \gamma_t f(x^*) + \frac{\beta_t \gamma_t}{2} (\|y_{t-1} - x^*\|^2 - \|y_t - x^*\|^2) + \eta_t \gamma_t \\
& + \gamma_t \langle \delta_t, x^* - y_{t-1} \rangle + \gamma_t \langle \delta_t, y_{t-1} - y_t \rangle - \frac{\gamma_t}{2} (\beta_t - L\gamma_t) \|y_t - y_{t-1}\|^2 \\
\leq & (1-\gamma_t) f(x_{t-1}) + \gamma_t f(x^*) + \frac{\beta_t \gamma_t}{2} (\|y_{t-1} - x^*\|^2 - \|y_t - x^*\|^2) + \eta_t \gamma_t \\
& + \gamma_t \langle \delta_t, x^* - y_{t-1} \rangle + \frac{\gamma_t \|\delta_t\|^2}{2(\beta_t - L\gamma_t)},
\end{aligned}$$

where the last inequality comes from the fact that

$$\gamma_t \langle \delta_t, y_{t-1} - y_t \rangle \leq \frac{\gamma_t}{2(\beta_t - L\gamma_t)} \|\delta_t\|^2 + \frac{\gamma_t(\beta_t - L\gamma_t)}{2} \|y_t - y_{t-1}\|^2 .$$

Substracting $f(x^*)$ from both sides of the above inequality, denoting $\phi_t = f(x_t) - f(x^*)$, $\theta_t = \gamma_t$, and in the view of Lemma 6, we obtain

$$\phi_t \leq \Theta_t \left[\phi_0 + \sum_{k=1}^{t} \frac{B_k}{\Theta_k} \right], \qquad (17)$$

where

$$B_t = \frac{\beta_t\gamma_t}{2}(\|y_{t-1}-x^*\|^2 - \|y_t - x^*\|^2) + \eta_t\gamma_t + \gamma_t\langle\delta_t, x^* - y_{t-1}\rangle + \frac{\gamma_t\|\delta_t\|^2}{2(\beta_t - L\gamma_t)}.$$

Choosing $\gamma_t = \theta_t = \frac{3}{t+2}$, we can easily check that $\Theta_t = \frac{6}{t(t+1)(t+2)}$ due to (12). Moreover, letting $\beta_t = \frac{4L}{t+2}, \eta_t = \frac{LD^2}{t(t+1)}$, we have $\sum_{k=1}^t \frac{\eta_k\gamma_k}{\Theta_k} \le \frac{tLD^2}{2}$ and

$$\begin{aligned}
&\sum_{k=1}^t \frac{\beta_k\gamma_k}{\Theta_k}(\|y_{t-1}-x^*\|^2 - \|y_t-x^*\|^2)\\
&\le \frac{\beta_1\gamma_1}{\Theta_i}\|y_0 - x^*\|^2 + \sum_{k=2}^t\left(\frac{\beta_k\gamma_k}{\Theta_k} - \frac{\beta_{k-1}\gamma_{k-1}}{\Theta_{k-1}}\right)\|y_{t-1}-x^*\|^2\\
&\le \frac{\beta_1\gamma_1}{\Theta_i}D^2 + \sum_{k=2}^t\left(\frac{\beta_k\gamma_k}{\Theta_k} - \frac{\beta_{k-1}\gamma_{k-1}}{\Theta_{k-1}}\right)D^2 = \frac{\beta_t\gamma_t D^2}{\Theta_t} = \frac{2LD^2t(t+1)}{t+2},
\end{aligned}$$

where the last inequality comes from the fact $\frac{\beta_k\gamma_k}{\Theta_k} > \frac{\beta_{k-1}\gamma_{k-1}}{\Theta_{k-1}}$.

We now prove part (a). Let $g_t = \tilde{\nabla}_t$. Taking expectation for both sides of (17), and noting that $\mathbb{E}[\langle\delta_t, x^* - y_{t-1}\rangle] = 0$ and

$$\begin{aligned}
\mathbb{E}[\|\delta_t\|^2|\mathcal{F}_{t-1}] &\le \frac{2\rho L}{b_t}(f(z_t) - f(x^*)) \quad \triangleright \text{ by Assumption 4}\\
&\le \frac{2\rho L}{b_t}\Big((1-\gamma_t)\phi_{t-1} + \gamma_t(f(y_{t-1}) - f(x^*))\Big) \quad \triangleright z_t = (1-\gamma_t)x_{t-1} + \gamma_t y_{t-1}\\
&\le \frac{2\rho L}{b_t}\Big((1-\gamma_t)\phi_{t-1} + \gamma_t(\|\nabla f(x^*)\|D + \frac{LD^2}{2})\Big) \quad \triangleright \text{ by the smoothness}\\
&:= \frac{2\rho L}{b_t}\Big((1-\gamma_t)\phi_{t-1} + \gamma_t\frac{KLD^2}{2}\Big), \quad \triangleright K = \frac{\|\nabla f(x^*)\|}{LD} + 1
\end{aligned}$$

we can obtain

$$\mathbb{E}[\phi_t] \le \frac{6LD^2}{(t+2)^2} + \frac{3LD^2}{(t+1)(t+2)} + \frac{3}{t(t+1)(t+2)}\sum_{k=1}^t \frac{\rho k(k+1)\Big((k-1)\mathbb{E}[\phi_{k-1}] + \frac{3KLD^2}{2}\Big)}{b_k}$$

We now prove

$$\mathbb{E}[\phi_t] \le \frac{6LD^2}{(t+2)^2} + \frac{(12+3K)LD^2}{(t+1)(t+2)} \tag{18}$$

by induction. Set $b_k = \lceil 3\rho k(k+1)\rceil$. It is easy to check $\mathbb{E}[\phi_0] \le \frac{KLD^2}{2}$ by the smoothness of f which satisfies (18). If (18) holds for all $k \le t-1$, then with the above inequality we can obtain

$$\begin{aligned}
\mathbb{E}[\phi_t] &\le \frac{6LD^2}{(t+2)^2} + \frac{(3+\frac{3K}{2})LD^2}{(t+1)(t+2)} + \frac{1}{t(t+1)(t+2)}\sum_{k=1}^t(k-1)\mathbb{E}[\phi_{k-1}]\\
&\le \frac{6LD^2}{(t+2)^2} + \frac{(3+\frac{3K}{2})LD^2}{(t+1)(t+2)} + \frac{1}{t(t+1)(t+2)}\sum_{k=1}^t\left(\frac{6LD^2(k-1)}{(k+1)^2} + \frac{(12+3K)LD^2(k-1)}{k(k+1)}\right)
\end{aligned}$$

$$\leq \frac{6LD^2}{(t+2)^2} + \frac{(3+\frac{3K}{2})LD^2}{(t+1)(t+2)} + \frac{(18+3K)LD^2}{t(t+1)(t+2)} \sum_{k=1}^{t} \frac{1}{k+1}$$
$$\leq \frac{6LD^2}{(t+2)^2} + \frac{(12+3K)LD^2}{(t+1)(t+2)},$$

i.e., (18) holds for $k = t$. Therefore, to achieve an ϵ-optimal point, the number of outer iterations T can be bounded by $\mathcal{O}(1/\sqrt{\epsilon})$. Hence, the number of calls to the first order oracles can be bounded by

$$\sum_{t=1}^{T} b_t \leq 3\rho \sum_{t=1}^{T} t(t+1) = \rho T(T+1)(T+2) = \mathcal{O}(T^3).$$

Due to the fact that the inner iterations indeed solves a convex constrained optimization problem by the classical Frank-Wolfe method with the exact line search, one can show that the number of inner iterations N_t performed at the t-th out iteration can be bounded by

$$N_t \leq \left\lceil \frac{6\beta_t D^2}{\eta_t} \right\rceil = \mathcal{O}(t).$$

Thus, the number of calls to the linear minimization oracle can be bounded by

$$\sum_{t=1}^{T} N_t \leq \mathcal{O}(T^2).$$

We now prove part (b). Let $g_t = \bar{G}_\nu^t$. Notice that $\bar{G}_\nu^t$ is a biased estimator of $\nabla f(z_t)$. We can obtain the following results by (7):

$$\begin{aligned}\mathbb{E}[\langle \delta_t, x^* - y_{t-1}\rangle] &= \mathbb{E}[\langle \nabla f_\nu(z_t) - \nabla f(z_t), x^* - y_{t-1}\rangle] + \mathbb{E}[\langle \bar{G}_\nu^t - \nabla f_\nu(z_t), x^* - y_{t-1}\rangle] \\ &= \mathbb{E}[\langle \nabla f_\nu(z_t) - \nabla f(z_t), x^* - y_{t-1}\rangle] \leq \frac{\nu LD(d+3)^{3/2}}{2}.\end{aligned}$$

Besides, we can obtain a similar bound for $\mathbb{E}[\|\delta_t\|^2]$ by Lemma 5.

$$\begin{aligned}\mathbb{E}[\|\delta_t\|^2 \,|\mathcal{F}_{t-1}] &\leq \frac{4\rho L(d+4)(f(z_t) - f(x^*))}{b_t} + \nu^2 L^2 (d+6)^3 \\ &\leq \frac{4\rho L(d+4)\big((1-\gamma_t)\phi_{t-1} + \frac{LD^2\gamma_t}{2}\big)}{b_t} + \nu^2 L^2 (d+6)^3.\end{aligned}$$

where the last inequality is slightly different from the one for the first-order setting due to $\|\nabla f(x^*)\| = 0$ for convex cases under the moment-based WGC.

By (17), we have the following simplified inequality:

$$\begin{aligned}\mathbb{E}[\phi_t] \leq & \frac{6LD^2}{(t+2)^2} + \frac{3LD^2}{(t+1)(t+2)} + \frac{\nu LD(d+3)^{3/2}}{2} + \frac{3\nu^2 L(d+6)^3}{2} \\ & + \frac{6}{t(t+1)(t+2)} \sum_{k=1}^{t} \frac{\rho(d+4)k(k+1)\left((k-1)\mathbb{E}[\phi_{k-1}] + \frac{3LD^2}{2}\right)}{b_k}.\end{aligned}$$

Set $b_k = \lceil 6\rho k(k+1)(d+4) \rceil, \nu = \frac{D}{(T+2)^2(d+6)^{3/2}} \leq \frac{D}{(t+2)^2(d+6)^{3/2}}$. Then we have

$$\mathbb{E}[\phi_t] \leq \frac{8LD^2}{(t+2)^2} + \frac{12LD^2}{(t+1)(t+2)} + \frac{1}{t(t+1)(t+2)} \sum_{k=1}^{t} (k-1)\mathbb{E}[\phi_{k-1}].$$

Similar to the proof for part (a), we can finish the proof by induction and obtain the bounds for complexity. □

C Zeroth-order SGD under Growth Conditions

In this section, we highlight that one can extend the results in [41] only assuming access to stochastic zeroth-order oracle with corresponding variance-based growth conditions. Notice that both SGC and WGC are defined in the format of the relative shrinkage of $\mathbb{E}\|\nabla F(x,\xi)\|^2$. However, in the unconstrained setting, the corresponding variance-based versions are equivalent to the moment-based growth conditions (see Proposition 1 for WGC; for SGC, note that $\mathbb{E}\|\nabla F(x,\xi) - \nabla f(x)\|^2 = \mathbb{E}\|\nabla F(x,\xi)\|^2 - \|\nabla f(x)\|^2 = (\rho-1)\|\nabla f(x)\|^2$).

We present the following result for the zeroth-order setting which directly follows the proofs in [41]. We highlight that it is the zeroth-order version of Theorem 3 in [41]. Similar results for other setups considered in [41] can also be obtained for the zeroth-order setting.

Algorithm 4 Non-convex Zeroth-order SGD (ZO-SGD)

Input: $x_0 \in \Omega$, number of iterations T, η
for $t = 1, 2, \ldots, T$ **do**
Randomly pick ξ_t and compute

$$x_t = x_{t-1} - \eta \frac{F(x_{t-1} + \nu u_t, \xi_t) - F(x_{t-1}, \xi_t)}{\nu} u_t := x_{t-1} - \eta G_t.$$

where u_t is generated from $\mathsf{N}(0, \mathbf{I}_d)$.
end for
Output: x_R where R is uniformly distributed over $0, \ldots, T-1$

Theorem 7. *Consider solving the non-convex unconstrained L-smooth problem by Algorithm 4 with some appropriate constant step size η, if f satistifies SGC with constant ρ, then*

$$\mathbb{E}\|\nabla f(x_R)\|^2 \leq \mathcal{O}\left(\frac{1}{T}\right)$$

Proof Idea The zeroth-order SGD update is given by

$$x_t = x_{t-1} - \eta \frac{F(x_{t-1} + \nu u_t, \xi_t) - F(x_{t-1}, \xi_t)}{\nu} u_t := x_{t-1} - \eta G_t.$$

By the smoothness of f, we have

$$\begin{aligned} f(x_t) - f(x_{t-1}) &\leq \langle \nabla f(x_{t-1}, x_t - x_{t-1}) + \frac{L}{2}\|x_t - x_{t-1}\|^2 \\ &= -\eta \langle \nabla f(x_{t-1}), G_t \rangle + \frac{L\eta^2}{2}\|G_t\|^2 \end{aligned}$$

Consider the term $\langle \nabla f(x_{t-1}), G_t \rangle$. Taking expectation with respect to ξ_t, u_t, we have

$$\begin{aligned} \mathbb{E}[\langle \nabla f(x_{t-1}), G_t \rangle] &= \langle \nabla f(x_{t-1}), \nabla f_\nu(x_{t-1}) \rangle \\ &= \langle \nabla f(x_{t-1}), \nabla f(x_{t-1}) + \nabla f_\nu(x_{t-1}) - \nabla f(x_{t-1}) \rangle \end{aligned}$$

$$\geq \|\nabla f(x_{t-1})\|^2 - \frac{\nu L(d+3)^{3/2}}{2}\|\nabla f(x_{t-1})\|$$

Consider the term $\|G_t\|^2$. Taking expectation with respect to ξ_t, u_t, we have

$$\begin{aligned}\mathbb{E}\|G_t\|^2 &\leq \frac{\nu^2}{2}L^2(d+6)^3 + 2(d+4)\mathbb{E}\|\nabla F(x_{t-1},\xi_t)\|^2 \\ &\leq \frac{\nu^2}{2}L^2(d+6)^3 + 2\rho(d+4)\|\nabla f(x_{t-1})\|^2\end{aligned}$$

Then, by the above inequalities, we can obtain

$$\begin{aligned}&\mathbb{E}[f(x_t) - f(x_{t-1})] \\ &\leq -\eta\|\nabla f(x_{t-1})\|^2 + \eta\frac{\nu L(d+3)^{3/2}}{2}\|\nabla f(x_{t-1})\| + \eta^2 L\rho(d+4)\|\nabla f(x_{t-1})\|^2 + \eta^2\frac{\nu^2}{4}L^3(d+6)^2 \\ &\leq -\eta\|\nabla f(x_{t-1})\|^2 + \eta^2 L(d+3)\|\nabla f(x_{t-1})\|^2 + \frac{\nu^2 L(d+3)^2}{16} \\ &\quad + \eta^2 L\rho(d+4)\|\nabla f(x_{t-1})\|^2 + \eta^2\frac{\nu^2}{4}L^3(d+6)^2 \\ &\leq -\eta\|\nabla f(x_{t-1})\|^2 + \eta^2 L(\rho+1)(d+4)\|\nabla f(x_{t-1})\|^2 + \eta^2\frac{\nu^2}{4}L^3(d+6)^2 + \frac{\nu^2 L(d+3)^2}{16}.\end{aligned}$$

If $\eta = \frac{1}{2L(\rho+1)(d+4)}$, then we have

$$\begin{aligned}&\mathbb{E}[f(x_t) - f(x_{t-1})] \leq -\frac{\eta}{2}\|\nabla f(x_{t-1})\|^2 + \eta^2\frac{\nu^2}{4}L^3(d+6)^2 + \frac{\nu^2 L(d+3)^2}{16} \\ &\Rightarrow \|\nabla f(x_{t-1})\|^2 \leq \frac{2}{\eta}\mathbb{E}[f(x_{t-1}) - f(x_t)] + \eta\frac{\nu^2}{2}L^3(d+6)^2 + \frac{\nu^2 L(d+3)^2}{8\eta}\end{aligned}$$

Setting $\nu = \mathcal{O}(1/\sqrt{dT})$ and taking a telescoping sum of the above inequality, we can get the same $\mathcal{O}(1/T)$ rate for the non-convex setting. □

In the above proof, we did not pay careful attention to the exact constants of the tuning parameter, as our main point is to simply highlight it is possible to obtain a zeroth-order version of the results in [41] under variance-based growth conditions and the logic of the proof is the same as [41].