A Study of First-Order Methods with a Deterministic Relative-Error Gradient Oracle

Nadav Hallak [*1] **Kfir Y. Levy** [*2]

Abstract

This paper studies the theoretical guarantees of the classical projected gradient and conditional gradient methods applied to constrained optimization problems with biased relative-error gradient oracles. These oracles are used in various settings, such as distributed optimization systems or derivative-free optimization, and are particularly common when gradients are compressed, quantized, or estimated via finite differences computations. Several settings are investigated: Optimization over the box with a coordinate-wise erroneous gradient oracle, optimization over a general compact convex set, and three more specific scenarios. Convergence guarantees are established with respect to the relative-error magnitude, and in particular, we show that the conditional gradient is invariant to relative-error when applied over the box with a coordinate-wise erroneous gradient oracle, and the projected gradient maintains its convergence guarantees when optimizing a nonconvex objective function.

1. Introduction

1.1. Problem formulation

This paper studies the optimization process in which the goal is to minimize a smooth function over a closed and convex set using a first-order relative-error erroneous oracle. Formally, we seek to solve the problem

$$\begin{aligned} &\min f(x) \\ &\text{s.t. } x \in \mathcal{C}, \end{aligned} \tag{P}$$

where

- $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable function;
- $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous, i.e., $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$ for any $x, y \in \mathcal{C}$;
- $\mathcal{C} \subseteq \mathbb{R}^n$ is closed, convex, and bounded, inducing the bounds: $R := \max_{x,y\in C} \|x - y\|_2$, $f_{\text{opt}} := \min_{x\in\mathcal{C}} f(x)$, and $M := \max_{x\in C} \|\nabla f(x)\|_2$.

In this study, we investigate the performance of the two classical methodologies, the Conditional Gradient (CG) and the Projected Gradient (PG), when utilized to solve (P) using a biased erroneous gradient. We assume that the gradient is obtained via an *Erroneous Oracle (EO)*, denoted by $\mathcal{O}(\cdot, \cdot)$, such that for any $x \in \mathcal{C}$ and $\hat{g} \leftarrow O(x, \varepsilon)$

$$\|\hat{g} - \nabla f(x)\| \leq \varepsilon\|\nabla f(x)\|, \quad \textbf{(EO)}, \tag{1}$$

where $\varepsilon \in [0, 1)$ is the relative-error parameter.

Condition (1), sometimes referred to as the *norm condition* (Bollapragada et al., 2018; Conn et al., 2000; Berahas et al., 2022), was first introduced by (Polyak, 1987; Carter, 1991) in the context of Trust-Region methods, and in recent years acts as a central assumption in many settings involving gradient approximation schemes such as gradient compression in distributed optimization (Ajalloeian & Stich, 2020; Beznosikov et al., 2020; Richtárik et al., 2021; Condat et al., 2022), derivative-free methods (Cartis & Scheinberg, 2018; Berahas et al., 2019; 2022), gradient quantization (Chmiel et al., 2021), generalized finite difference gradient approximations (Cartis & Scheinberg, 2018; Paquette & Scheinberg, 2020), adaptive sampling optimization methods (Byrd et al., 2012; Bollapragada et al., 2018), and is also used in (Hintermüller & Vicente, 2005) which studies optimal control for nonlinear partial differential equations.

The norm condition (1) defining the EO can sometimes be refined coordinate-wise to obtain a *Coordinate-Wise Erroneous Oracle (CWEO)*. The CWEO will also be denoted by $\mathcal{O}(\cdot, \cdot)$, and we will indicate when we assume that the EO is a CWEO. The CWEO satisfies that for any $x \in \mathcal{C}$ and $\hat{g} \leftarrow O(x, \varepsilon)$, it holds that

$$|\hat{g}_i - \nabla f(x)_i| \leq \varepsilon|\nabla f(x)_i|, \quad \forall i \in [n] \quad \textbf{(CWEO)}. \tag{2}$$

The EO and CWEO are formally defined in Section 2.

[*] Equal contribution [1] The Faculty of Data and Decision Sciences, The Technion, Haifa, Israel. [2] Viterby Faculty of Electrical and Computer Engineering, The Technion, Haifa, Israel.. Correspondence to: Nadav Hallak <ndvhllk@technion.ac.il>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

A straightforward intuitive example for CWEO is the use of a coordinate-wise gradient floating-point based quantization, as demonstrated by Example 1.1 – see e.g., Section 5 in (Chmiel et al., 2021).

Example 1.1 (gradient quantization using floating-point)**.** *Floating-point representations exploit formulaic forms to approximately capture a wide range of numbers in the purpose of facilitating fast processing times in systems with very large or small numbers; see for example (Chmiel et al., 2021). A number $u \in \mathbb{R}_+$ is decomposed as $u = 2^{\ln u} = 2^{\ln u - \lfloor \ln u \rfloor} \cdot 2^{\lfloor \ln u \rfloor} = r \cdot 2^E$, where $r \in [1,2)$ and $E \in \mathbb{Z}$. The quantized floating-point number u_q of u is built by allocating bits for r and E, and the error between u_q and u is measured relatively. The optimizer may determine ε by tuning the number of bits used for r and E to guarantee that (2) holds true when quantizing the gradient.*

Another example is the gradient estimation via standard finite differences studied in the derivative-free methodology, see for example (Berahas et al., 2022) and references therein.

Example 1.2 (gradient approximation using standard finite differences)**.** *Gradient approximations via central finite differences (cf. Section 2.1 in (Berahas et al., 2022)) are based on the sample set $\mathcal{S} = \bigcup_{i\in[n]}\{x+\sigma e_i\} \cup \bigcup_{i\in[n]}\{x-\sigma e_i\}$ so that $\hat{g}_i = \frac{f(x+\sigma e_i)-f(x-\sigma e_i)}{2\sigma}$ for any $i \in [n]$. These approximations are guaranteed to satisfy the norm condition (1), as established by (Berahas et al., 2022).*

Many first-order methods rely on a gradient oracle to provide the exact, or an approximate, gradient for $f(\cdot)$ as a part of their optimization process, so that upon querying this oracle with $x \in \mathcal{C}$, it outputs $\hat{g} = \mathcal{O}(x;\varepsilon)$ such that $\hat{g} \approx \nabla f(x)$ having some error bound guarantee. Among the class of gradient oracles, the family of unbiased oracles dominates the literature, see for example the seminal paper (Nemirovski et al., 2009) and related.

In comparison, the literature on biased oracles is considerably more limited, and in particular, the theoretical performance of methods tackling constrained optimization with relative-error gradient oracles satisfying (1) are still not well-understood.

In the gradient compression distributed optimization settings, meaningful guarantees are restricted to unconstrained problems (Ajalloeian & Stich, 2020; Beznosikov et al., 2020; Richtárik et al., 2021), or require the Polyak-Lojasiewicz (PL) or the Kurdyka-Lojasiewicz (KL) conditions (Condat et al., 2022). The derivative-free methodology that explicitly assumes (1) is mainly devoted to unconstrained optimization (Conn et al., 2009; Berahas et al., 2022; Cartis & Scheinberg, 2018; Byrd et al., 2012), and as such, does not provide meaningful results for our constrained model. In general, the derivative-free approach tackles constraints in manners that are substantially different from the techniques in our work, and to the best of our knowledge, these do not (at least explicitly) rely on (1).

Indeed, this work focuses on the PG and CG utilizing oracles satisfying (1) applied to **constrained** problems **without** any special structural assumptions such as the aforementioned PL, KL, or strong convexity conditions. To demonstrate the difficulty in employing an optimization method with an EO to solve (P), let us consider the following trivial convex problem.

Example 1.3. *Let $f(x) = -x_1 - x_2$, $\mathcal{C} = \{x \in \mathbb{R}^2 : \|x\|_2 \leq 1\}$, and set $\varepsilon \in (0,1)$. The optimal solution is $x_1^* = x_2^* = 1/\sqrt{2}$, but by setting a persistent relative error of $(1+\varepsilon)$ for the first component and $(1-\varepsilon)$ for the second, due to the EO, the optimizer may not be able to determine that the point $x_1^{\text{error}} = \frac{1+\varepsilon}{\sqrt{2+2\varepsilon^2}}$, $x_2^{\text{error}} = \frac{1-\varepsilon}{\sqrt{2+2\varepsilon^2}}$ is not optimal. Moreover, if not given the function itself, it is impossible for any optimizer to determine based on the oracle whether the true function is either $f(x)$ or $\tilde{f}(x) = -(1+\varepsilon)x_1 - (1-\varepsilon)x_1$, or something in-between. Thus, in general, it is impossible to ensure convergence without a vanishing error. Specifically, if the PG or CG methods are initiated at $x^0 = x^*$, then they will diverge from the optimal solution of $f(\cdot)$, and will rather converge to x^{error}.*

To shortly summarize the contributions of this work:

- We show that for nonconvex optimization over box constraints with a **CWEO**, the CG and PG maintain their theoretical convergence guarantees. Moreover, the CG method maintains its guarantees when the objective function is convex.

- For general nonconvex constrained optimization with a **general EO** we derive convergence guarantees that are dependent on the relative-error parameter for both the PG and CG methods. We do the same for general convex constrained optimization with a **general EO** for the CG method; theoretical convergence guarantees of the PG remain unestablished.

- We develop two special instants for the CG framework where the first one may only access the sign of the elements of the gradient and optimizes over the box set; and the second is tailored for handling convex optimization when the feasible set is the Euclidean ball with a general EO, or when it belongs to a general class of sets, that contains the ℓ_p-norm balls ($p \geq 1$), with a CWEO.

Our results in particular produce an interesting insight that the theoretical performance of the CG method is much less sensitive to the EO compared to that of the PG, motivating further research of this phenomenon.

Outline. The general purpose *Erroneous Conditional Gradient (ECG)* and *Erroneous Projected Gradient (EPG)* methods are described in Section 3.1 and Section 3.2 respectively. The more specific *Sign Conditional gradient (SCG)* and *Rescaled Erroneous Conditional Gradient (RECG)* methods are described in Section 6. Table 1 summarizes the complexity part of our results and their respective locations (a larger version appears in Appendix A).

Table 1. Complexity results and their respective settings and locations; subsequence convergence guarantees do not appear in the table. When the relative error does not affect the theoretical guarantees, we say that these guarantees are 'preserved'.

Oracle	Set	Alg.	Function Type	Guarantee	Section
CWEO	Box	ECG	Convex/Nonconvex	Preserved	Sec. 4.1
		EPG	Nonconvex	Preserved (neglecting constants)	Sec. 4.2
		SCG	Convex/Nonconvex	Preserved	Sec. 6.1
	Sign-Preserving	RECG	Convex	$f(\bar{w}^T) - \theta f_{\text{opt}} \leq \frac{8L(R^2/\theta)}{T+1}$	Sec. 6.2.2
EO	General	ECG	Nonconvex	$\min_{d \in \mathcal{C}} \langle \nabla f(w^*), d - w^* \rangle \geq -\frac{2\varepsilon}{1-\varepsilon} MR$	Sec. 5.1
			Convex	$f(w^t) - f_{\text{opt}} \leq 2\varepsilon MR + \frac{4LR^2}{t+2}$	Sec. 5.1
		EPG	Nonconvex	$\min_{t \in [T]} \|q^{t+1} + \nabla f(w^{t+1})\|^2 \leq \frac{8L(f^0 - f_{\text{opt}})}{T} + \epsilon$	Sec. 5.2
	ℓ_2-Ball	RECG	Convex	$f(\bar{w}^T) - \theta f_{\text{opt}} \leq \frac{8L(R^2/\theta)}{T+1}$	Sec. 6.2

Literature. Mathematical optimization methods with approximate value/gradient estimates are ubiquitous in large-scale models, and in particular, in machine learning applications. This work focuses on **constrained** optimization with inexact gradient information deterministically satisfying the so-called norm condition in (1). The condition (1) was first introduced by (Polyak, 1987; Carter, 1991) in the context of Trust-Region methods, and since then, has played a significant role in the field of derivative-free optimization and optimization involving inexact gradient approximations – see for example the aforementioned papers and books and references therein. In recent years, it has also gained increased interest among the distributed optimization community, see e.g. (Ajalloeian & Stich, 2020; Richtárik et al., 2021; Condat et al., 2022), as it can capture the elemental bias that arises due the use s gradient compression schemes.

Unlike the vast and extensive literature on unbiased stochastic gradient methods, the volume of works on biased gradient estimators is quite limited. The seminal book (Bertsekas, 2003) discusses steepest descent methods under various gradient error scenarios in *unconstrained* problems with classical convergence guarantees depending on the scenario, see also references therein. Biased stochastic gradient estimators are further studied for *unconstrained* problems in the contemporary works (Ajalloeian & Stich, 2020; Richtárik et al., 2021; Condat et al., 2022). Previous works on biased oracles roughly divide into two: **(i)** assuming a gradient/value oracles with a fixed *additive bias* $\delta > 0$, i.e., $\mathcal{O}(x_t) = \nabla f(x_t) + b_t$ where $\|b_t\| \leq \delta$ (alternatively it is assumed that the gradient oracle fulfills the decent lemma, cf. Lemma 1.1, up to a fixed bias $\delta > 0$); and, **(ii)** assuming a gradient oracle with a multiplicative bias, i.e., $\mathcal{O}(x_t) = \nabla f(x_t) + b_t$ where $\|b_t\| \leq \rho \|\nabla f(x_t)\|$ for some $\rho \in [0, 1)$. Our work falls under the latter category.

In works assuming fixed additive bias (d'Aspremont, 2008; Devolder et al., 2014; Dvurechensky & Gasnikov, 2016; Dvurechensky, 2017; Stonyakin et al., 2019), it is shown both for convex and nonconvex functions that one cannot obtain an arbitrarily small error (or gradient norm) but rather that we can only reduce the error (or gradient norm) to a fixed value that depends on the bias.

Conversely, under the assumption of a multiplicative error, it was demonstrated in (Ajalloeian & Stich, 2020) that one can obtain meaningful guarantees in two cases: **(i)** For objectives $f(\cdot)$ that satisfy the Polyak-Łojasiewicz (PL) condition, it was shown that one can obtain an arbitrarily small error at a linear rate. **(ii)** For nonconvex smooth objectives, it was shown that one can obtain an arbitrarily small gradient norm, at the standard rate of $O(1/T)$ where T is the total number of gradient computations. Nevertheless, there do not exist guarantees for the general convex smooth case (either constrained or unconstrained), nor for the nonconvex constrained case. The works of Richtárik et al. (2021); Condat et al. (2022) have also studied such oracles with multiplicative error, but have mainly focused on utilizing them in the context of unconstrained federated learning.

To the best of our knowledge, the only work studying a notion of relative error in gradient approximation when minimizing **constrained** problems is the very recent (Condat et al., 2022). However, the model studied in (Condat et al., 2022) is that of a convex function minimization satisfying the PL property, and the optimization framework utilizes a proximal gradient -based approach. We consider both convex and nonconvex models, and do not assume any error-bound type property, especially we do not assume the PL property.

Notation. We use standard notation throughout, in particular, $\|\cdot\|$ stands for the Euclidean norm and $[n] := \{1, 2, \ldots, n\}$. Whenever the affiliation of $\nabla f(x)$ and $\mathcal{O}(x, \varepsilon)$ are clear from context, we will just write $\hat{g} \in \mathcal{O}(x, \varepsilon)$ and $g = \nabla f(x)$ respectively.

1.2. Mathematical preliminaries

The smoothness of $f(\cdot)$ facilitates the descent lemma.

Lemma 1.1 (descent lemma (Bertsekas, 2003)). *For any $x, y \in \mathcal{C}$ it holds that $f(x) \leq f(y) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2$.*

Box constraints are ubiquitous in mathematical modeling of problems in science and engineering.

Definition 1.2 (box set)**.** A set $\mathcal{C} \subseteq \mathbb{R}^n$ is called a box if $\mathcal{C} = \{x \in \mathbb{R}^n : x_i \in [l_i, u_i]\}$ where $l, u \in \mathbb{R}^n$ and $l_i \leq u_i$ for any $i \in [n]$.

In our analysis we use standard measures of (sub)-optimality. When analyzing the ECG or the EPG for convex problems, we simply use standard sub-optimality criteria which compares against the optimal value f_{opt}.

When analyzing the ECG for nonconvex problems, we use the optimality measure associated with the CG method sometimes called the *optimality gap* (Jaggi, 2013) or the *Conditional Gradient Norm* (Beck, 2017). It is defined at a point $w^* \in \mathcal{C}$ that satisfies $\min_{d \in \mathcal{C}} \langle \nabla f(w^*), d - w^* \rangle \geq 0$.

For the EPG approach in the nonconvex case we also use the above standard stationarity condition, but do so via an equivalent normal cone formulation in the purpose of simplifying the analysis. This is possible since the subdifferential of the indicator function of a closed and convex set is equal to a respective normal cone; see. e.g., Example 3.74 in (Beck, 2017) establishing that $w^* \in \mathcal{C}$ is stationary if and only if $-\nabla f(w^*) \in \partial \delta_{\mathcal{C}}(w^*) = \mathcal{N}_C(w^*)$, where $\mathcal{N}_{\mathcal{C}}(w^*)$ is the normal cone w.r.t $\mathcal{C}$ at w^*.

The relationship between stationarity and the normal cone formulation also implies that the bound on $\|q + \nabla F(w^*)\|$, where $q \in \partial \delta_{\mathcal{C}}(w^*)$, can be translated to approximate stationarity.

Lemma 1.3. *Let $w^* \in \mathcal{C}$ such that there exist $q \in \partial \delta_{\mathcal{C}}(w^*)$ and $\rho > 0$ satisfying that $\|\nabla f(w^*) + q\| \leq \rho$. Then for any $d \in \mathcal{C}$ it holds that $\langle \nabla f(w^*), d - w^*, \rangle \geq -\rho R$.*

Proof. Note that $\|\nabla f(w^*) + q\| \leq \rho$ implies that for any $u \in \mathbb{R}^d$, it holds that $\langle u, \nabla f(w^*) + q \rangle \geq -\rho \|u\|$. Now, pick $u := d - w^*$ for some $d \in \mathcal{C}$. Since $q \in \delta_{\mathcal{C}}(w^*) = \mathcal{N}_{\mathcal{C}}(w^*)$, $-\langle d - w^*, q \rangle \geq 0$, plugging this back into the above equation and using the boundedness of $\mathcal{C}$ yields,

$$\langle d - w^*, \nabla f(w^*) \rangle \geq -\langle d - w^*, q \rangle - \rho \|d - w^*\| \geq -\rho R.$$

The theorem holds since the above applies for any $d \in \mathcal{C}$. □

2. The Erroneous Oracle

The Erroneous Oracle (EO) is formally defined as follows.

Erroneous Oracles. Let $\varepsilon \in [0, 1)$ be the *relative-error parameter*, and let $x \in \mathcal{C}$. The oracle $\mathcal{O}(\cdot; \varepsilon)$ is called an *erroneous oracle (**EO**)* with respect to a function $f : \mathcal{C} \mapsto \mathbb{R}$, if for any vector $x \in \mathcal{C}$ it returns $\hat{g} = \mathcal{O}(x; \varepsilon) \in \mathbb{R}^n$ which satisfies

$$\|\hat{g} - \nabla f(x)\| \leq \varepsilon \|\nabla f(x)\|. \tag{3}$$

If for any $x \in \mathcal{C}$ it returns $\hat{g} = \mathcal{O}(x; \varepsilon) \in \mathbb{R}^n$ which satisfies

$$|\hat{g}_i - \nabla f(x)_i| \leq \varepsilon |\nabla f(x)_i|, \qquad \forall i \in [n]. \tag{4}$$

Then it is called a *coordinate-wise erroneous oracle (**CWEO**)*.

Remark 2.1. *We implicitly allow for the relative error to vary for any new access to the oracle, i.e., for $\hat{g}^1, \hat{g}^2 \leftarrow \mathcal{O}(x, \varepsilon)$ we may have that $\hat{g}^1 \neq \hat{g}^2$. Consequently, the oracle can capture a wide range of oracle error regimes such as adversarial, biased or unbiased stochastic, etc. We emphasize that no additional assumptions or limitations on the oracle will be imposed in the sequel.*

The definition of the EO implies the next useful bounds (see Appendix B).

Lemma 2.2 (EO bounds)**.** *Let $x \in \mathcal{C}$, $\varepsilon \in [0, 1)$, and $\hat{g} \leftarrow \mathcal{O}(x; \varepsilon)$. Then*

1. $(1 - \varepsilon)\|\nabla f(x)\| \leq \|\hat{g}\| \leq (1 + \varepsilon)\|\nabla f(x)\|$.
2. $\frac{\|\hat{g}\|}{1+\varepsilon} \leq \|\nabla f(x)\| \leq \frac{\|\hat{g}\|}{1-\varepsilon}$.
3. $\langle \nabla f(x), \hat{g} \rangle \geq (1 - \varepsilon)\|\nabla f(x)\|^2$.

When a CWEO is applied, we have the coordinate-wise properties detailed in Lemma 2.3; due to the technicality of the result, the proof is deferred to Appendix B.

Lemma 2.3 (CWEO properties)**.** *Let $x, d \in \mathbb{R}^n$, $\varepsilon \in [0, 1)$, and $\hat{g} \leftarrow \mathcal{O}(x; \varepsilon)$, where the EO is a CWEO. Then*

1. *Sign preservation property:* $\text{sign}(\hat{g}_i) = \text{sign}(\nabla f(x)_i)$ *for any $i \in [n]$;*
2. *Relative coordinate-wise error: for any $i \in [n]$ it holds that $(1 - \varepsilon)|\nabla f(x)_i| \leq |\hat{g}_i| \leq (1 + \varepsilon)|\nabla f(x)_i|$ and $\frac{1}{1+\varepsilon}|\hat{g}_i| \leq |\nabla f(x)_i| \leq \frac{1}{1-\varepsilon}|\hat{g}_i|$.*
3. *Direction error bound: Suppose that $x, d \in \mathcal{C}$. It holds that*

$$\begin{aligned} \frac{1}{1-\varepsilon^2}\left(\langle \hat{g}, d \rangle - \varepsilon(1+\varepsilon)MR\right) &\leq \langle g, d \rangle \\ \frac{1}{1-\varepsilon^2}\left(\langle \hat{g}, d \rangle + \varepsilon(1+\varepsilon)MR\right) &\geq \langle g, d \rangle. \end{aligned} \tag{5}$$

3. Methods

3.1. The Erroneous Conditional Gradient

The Erroneous Conditional Gradient (ECG) method described by Algorithm 1 is a variation of the classical CG method (Frank & Wolfe, 1956) obtained by replacing the gradient with an output of the EO. Accordingly, it determines the update direction by invoking the so-called *linear minimization oracle (LMO)* which, given a vector $d \in \mathbb{R}^n$, returns a solution to the optimization problem:

$$\mathrm{LMO}(d) := \underset{z}{\operatorname{argmin}}\{\langle d, z\rangle : z \in \mathcal{C}\}. \tag{6}$$

Since $\mathcal{C}$ is closed convex and bounded, the LMO (6) is a well-defined single valued operation.

Algorithm 1: Erroneous Conditional gradient (ECG)

```
Input: $w^0 \in \mathcal{C}, \varepsilon \geq 0$ .
for any $t \geq 0$ do
    retrieve $\hat{g}^t \leftarrow \mathcal{O}(\nabla f(w^t), \varepsilon)$ ;
    compute $p^{t+1} \leftarrow \mathrm{LMO}(\hat{g}^t) - w^t$;
    choose $\eta_t \in [0, 1]$ and set $w^{t+1} \leftarrow w^t + \eta_t p^{t+1}$;
end
```

In the next section we study the effect of relative error gradient feedback on the convergence guarantees of the CG by analyzing the ECG, and in particular, prove that CG is fully robust to the relative error. This outstanding property of the CG allows for example to use low precision gradients throughout the entire optimization process without any influence on the theoretical guarantees. It also suggests that the CG algorithmic framework may be more suited for models with intrinsic multiplicative gradient error, such as distributed optimization, than the PG algorithmic framework, motivating further research in this direction.

3.2. The Erroneous Projected Gradient

The Projected Gradient (PG) method essentially amounts to repeatedly executing the projection of the gradient step, $w^{t+1} = P_{\mathcal{C}}(w^t - \eta \nabla f(w^t))$ where $\eta > 0$ is the step-size and $P_{\mathcal{C}}(x) := \underset{z \in \mathcal{C}}{\operatorname{argmin}} \|z - x\|_2$ is the orthogonal projection operator onto the set $\mathcal{C}$. Given an erroneous gradient, we obtain the erroneous version of the PG described by Algorithm 2. To control the change in the function value, two update conditions are considered: (i) Descent: $h(w^t, w^{t+1}) = f(w^t) - f(w^{t+1})$; (ii) Sufficient descent: $h(w^t, w^{t+1}) = \|\eta^{-1}(w^{t+1} - w^t)\|^2 - \frac{2\varepsilon}{1-L\eta} MR$. The first condition only requires access to the function. In the context of the EO, this is suited for gradient approximations via standard finite differences or linear interpolation for example (cf. (Berahas et al., 2022)). The second condition requires some knowledge on the parameters of the problem at hand – the diameter of the feasible set and bounds on the Lipschitz constant and the norm of the gradient over the feasible set.

Algorithm 2: Erroneous Projected gradient (EPG)

```
Input: $w^0 \in C, \varepsilon \in [0, 1), \eta \in \left(0, \frac{1}{L(1+\varepsilon)}\right)$.
for any $t \geq 0$ do
    retrieve $\hat{g}^t \leftarrow \mathcal{O}(\nabla f(w^t), \varepsilon)$ ;
    set $w^{t+1} = P_{\mathcal{C}}(w^t - \eta \hat{g}^t)$;
    if $h(w^t, w^{t+1}) > 0$ then
        set $w^{t+1} \leftarrow w^t$;
    end
end
```

Remark 3.1 (update formula)**.** *Note that the update* $w^{t+1} = P_{\mathcal{C}}(w^t - \eta \hat{g}^t)$ *can be equivalently written as* $w^{t+1} = \underset{w \in \mathcal{C}}{\operatorname{argmin}} \langle \hat{g}^t, w - w^t\rangle + \dfrac{1}{2\eta}\|w - w^t\|^2$*, which immediately implies that* $\langle \hat{g}^t, w^{t+1} - w^t\rangle + \frac{1}{2\eta}\|w^{t+1} - w^t\|^2 \leq 0$.

4. Robustness with CWEO on the box set

In this section we analyze the ECG and EPG when applied to the box set using a **CWEO**.

Assumption 4.1. The set $\mathcal{C}$ is the box and the EO used by Algorithm 1 and Algorithm 2 is a CWEO.

4.1. ECG

When $\mathcal{C}$ is the box, the LMO in (6) is a separable problem whose solution $z^* \in \mathrm{LMO}(d)$ can be expressed equivalently component-wise by

$$z_i^* = \underset{z}{\operatorname{argmin}}\{d_i \cdot z_i : z_i \in \mathcal{C}_i\}, \qquad \forall i \in [n]. \tag{7}$$

Using this fact, we now establish that the trajectory of points generated by the ECG method is independent of the value of ε. This is correct if the LMO, which determines the trajectory, returns the same vector for both the gradient and its erroneous counterpart.

Theorem 4.2 (CG is robust under separability)**.** *Suppose that Assumption 4.1 holds true. Then for any* $\varepsilon \geq 0$*, Algorithm 1 generates that same sequence of points.*

Proof. We establish that any element in the sequence $\{p^{t+1}\}_{t\geq 0}$ generated by Algorithm 1 satisfies that $\mathrm{LMO}(\hat{g}^t) = \mathrm{LMO}(\nabla f(w^t))$, which readily implies the required. Indeed, denote $g^t = \nabla f(w^t)$, and let $\alpha_i \in [0, 1+\varepsilon]$ be the parameter satisfying that $\hat{g}_i^t = \alpha_i^t g_i^t$ for each $i \in [n]$. Note that: (i) the sign preservation property of the EO (cf. Lemma 2.3) guarantees that $\alpha_i^t > 0$ if $\hat{g}_i^t \neq 0$; (ii) the CWEO definition guarantees that $\alpha_i^t \leq 1 + \varepsilon$. Thus, due to the separability of $\mathcal{C}$ we have for any $i \in [n]$ that

$\underset{z}{\operatorname{argmin}}\{\hat{g}_i \cdot z : z \in C_i\} \equiv \underset{z}{\operatorname{argmin}}\{\alpha_i^t \hat{g}_i \cdot z : z \in C_i\} \equiv \underset{z}{\operatorname{argmin}}\{g_i \cdot z : z \in C_i\}$, as claimed. □

Obviously, Theorem 4.2 implies that *any result and guarantee* that holds true for the CG method with no relative-error oracle ($\varepsilon = 0$), holds true for the ECG with a relative-error oracle ($\varepsilon \in (0, 1)$).

Corollary 4.3 (informal)**.** *Suppose that Assumption 4.1 holds true. Then any theoretical guarantee satisfied for the CG holds true for the ECG with any $\varepsilon \in [0, 1]$.*

Remark 4.4 (on the CG with a relative-error gradient)**.** *Although simple, the proof of Theorem 4.2 provides a significant insight, that under the common settings of Assumption 4.1, the magnitude of the error is irrelevant and can actually have any value (even larger than one) without implications to the theoretical guarantees of the CG optimization scheme.*

4.2. EPG

Unlike the CG method, we only establish that the guarantees of the PG method are robust to relative error gradients in nonconvex minimization over the box set; the proof is deferred to Appendix C.

Theorem 4.5 (EPG robustness in nonconvex over the box)**.** *Suppose that Assumption 4.1 holds true. Let $\{w^t\}_{t\geq 0}$ be a sequence generated by Algorithm 2. Then for any $\eta \in (0, \frac{1}{(1+\varepsilon)L})$, it holds that:*

1. *For any $T \geq 0$ it holds that*

$$\min_{t\in[T]} \|q^{t+1} + \nabla f(w^{t+1})\|^2 \leq \frac{f(w^0) - f_{\text{opt}}}{Tc(\eta, \varepsilon, L)},$$

where $q^{t+1} \in \partial\delta_{\mathcal{C}}(w^{t+1})$ and $c(\eta, \varepsilon, L) = \frac{\eta(1-\varepsilon)^2}{2(1+\varepsilon)} \frac{1-L\eta(1+\varepsilon)}{(1+L\eta(1-\varepsilon))^2} > 0$. Consequently, for $\eta = \frac{1}{2(1+\varepsilon)L}$ the EPG achieves

$$\min_{t\in[T]} \|q^{t+1}+\nabla f(w^{t+1})\|^2 \leq \frac{2L(3+\varepsilon)^2(f(w^0) - f_{\text{opt}})}{(1-\varepsilon)^2 T}. \tag{8}$$

2. *Any accumulation point of $\{w^t\}_{t\geq 0}$ is a stationary point of* (P)*, and is in particular the optimal solution of* (P) *if f is convex.*

Remark 4.6 (convergence guarantees for convex problems)**.** *Improved rates of the classical PG method applied to convex problems under our general settings, i.e. without any error-bound type assumption, remains an open question. However, under the PL property, improved rates for biased gradient compression were established for unconstrained convex problems in (Ajalloeian & Stich, 2020; Richtárik et al., 2021), and for constrained convex problems (with the appropriate version of the PL) by the very recent (Condat et al., 2022).*

5. General Analysis

This section studies the ECG and EPG with a **general EO**.

5.1. ECG with General EO

To establish the theoretical guarantees of the ECG in the nonconvex setting, we require a descending property; all the proofs of this subsection are given in Appendix D.

Lemma 5.1 (ECG descending property)**.** *Let $\{(w^t, \hat{g}^t, p^t)\}$ be a sequence generated by Algorithm 1 with $\varepsilon \in [0, 1/2)$ and step-size η_t defined by: $\eta_t = 0$ when $\gamma_t \leq 0$, $\eta_t = \gamma_t$ when $\gamma_t \in [0, 1]$, and $\eta_t = 1$ when $\gamma_t \geq 1$, where $\gamma_t := -\frac{\langle \hat{g}^t, p^{t+1}\rangle + \frac{\varepsilon}{1-\varepsilon}\|\hat{g}^t\|\|p^{t+1}\|}{L\|p^{t+1}\|^2}$ for $p^{t+1} \neq 0$ and $\gamma_t = 0$ otherwise. Then*

$$f^t - f^{t+1} \geq \frac{1}{2}L\|p^{t+1}\|^2\eta_t^2. \tag{9}$$

Consequently, for any $t \geq 0$ satisfying that $-\langle \hat{g}^t, p^{t+1}\rangle > \frac{\varepsilon}{1-\varepsilon}\|\hat{g}^t\|\|p^{t+1}\|$ we have that $f^t > f^{t+1}$.

The convergence guarantees of ECG are stated in Theorem 5.2, where in its third part, we establish a bound on the CG optimality measure known as the *optimality gap* (Jaggi, 2013) or the *Conditional Gradient Norm* (Beck, 2017).

Theorem 5.2 (ECG nonconvex convergence properties)**.** *Let $\{(w^t, \hat{g}^t, p^t)\}$ be a sequence generated by Algorithm 1 with $\varepsilon \in [0, 1/2)$ and step-size as defined in Lemma 5.1. Then*

1. *The sequence $\{f^t\}_{t\geq 0}$ is monotonic non-ascending, and thus converges to some $f^* \leq f^t$ for any $t \geq 0$.*
2. $\lim_{t\to\infty} \left(\langle \hat{g}^t, p^{t+1}\rangle + \frac{\varepsilon}{1-\varepsilon}\|\hat{g}^t\|\|p^{t+1}\|\right) \geq 0.$
3. *Any accumulation point w^* of $\{w^t\}_{t\geq 0}$ satisfies that*

$$\min_{d\in\mathcal{C}} \langle \nabla f(w^*), d - w^*\rangle \geq -\frac{2\varepsilon}{1-\varepsilon}MR. \tag{10}$$

To establish the guarantees of the ECG in the convex scenario, we first derive a relation between the optimal function value, the error, and the erroneous descent term; the proof is given in Appendix D.

Lemma 5.3. *Suppose that f is convex. Let $w^* \in \mathcal{C}$ be the optimal solution of* (P) *and $\{(w^t, \hat{g}^t, p^t)\}$ be a sequence generated by Algorithm 1. Then $\langle \hat{g}^t, p^{t+1}\rangle \leq f_{\text{opt}} - f(w^t) + \varepsilon RM$.*

We can now establish the rate result.

Theorem 5.4 (ECG rate for convex objectives)**.** *Suppose that f is convex and let $w^* \in \mathcal{C}$ be the optimal solution of* (P). *Let $\{(w^t, \hat{g}^t, p^t)\}$ be the sequence generated by Algorithm 1 with step-size $\eta_t = \min\left\{1, \frac{2}{t+2}\right\}$ for any $t \geq 0$. Then*

$$f(w^t) - f_{\text{opt}} \leq 2\varepsilon MR + \frac{4LR^2}{t+2}. \tag{11}$$

5.2. EPG with General EO

To establish the EPG convergence properties for a nonconvex objective with a **general EO**, we first establish a descent property. Due to space limitations, the proofs of both Lemma 5.5 and Theorem 5.6 are detailed in Appendix E.

Lemma 5.5 (EPG descent property)**.** *Let $\{(w^t, \hat{g}^t, p^t)\}$ be a sequence generated by Algorithm 2. Then*

$$f^t - f^{t+1} \geq \frac{1-L\eta}{2}\|\eta^{-1}(w^{t+1}-w^t)\|^2 - \varepsilon MR. \tag{12}$$

Consequently, for any $t \geq 0$ in which

$$\|\eta^{-1}(w^{t+1}-w^t)\|^2 > \frac{2\varepsilon}{1-L\eta}MR, \tag{13}$$

we have a descent in the function value, i.e., $f^t > f^{t+1}$.

The EPG convergence guarantees below comprise a rate result bound to ϵ-stationarity and a subsequence convergence guarantee to a ϵ-stationary point where $\epsilon = 2\varepsilon M\left(\frac{8R}{\eta(1-L\eta)} + \varepsilon M\right)$.

Theorem 5.6 (EPG convergence properties)**.** *Let $\{(w^t, \hat{g}^t, p^t)\}$ be a sequence generated by Algorithm 2. Then*

1. *For any $T \geq 0$ it holds that*

   $$\min_{t\in[T]}\|q^{t+1}+\nabla f(w^{t+1})\|^2 \leq \frac{4(1+L\eta)^2(f^0 - f_{\text{opt}})}{T\eta(1-L\eta)} + \epsilon,$$

   where $q^{t+1} \in \partial\delta_{\mathcal{C}}(w^{t+1})$ and $\epsilon = 2\varepsilon M\left(\frac{8R}{\eta(1-L\eta)} + \varepsilon M\right)$. In particular for $\eta = 1/3L$ we obtain that

   $$\min_{t\in[T]}\|q^{t+1}+\nabla f(w^{t+1})\|^2 \leq \frac{8L(f^0 - f_{\text{opt}})}{T} + \epsilon.$$

2. *Any accumulation point w^* of $\{w^t\}_{t\geq 0}$ is an ϵ-stationary point of* (P)*. That is, there exists $q^* \in \partial\delta_{\mathcal{C}}(w^*)$ such that*

   $$\|q^* + \nabla f(w^*)\|^2 \leq \epsilon = 2\varepsilon M\left(\frac{8R}{\eta(1-L\eta)} + \varepsilon M\right).$$

6. Special scenarios

This section investigates methods tailored for specific settings. In Section 6.1 we show that for the CG scheme the magnitude of the relative error is in fact irrelevant when optimizing over the box set – only the sign matters; We do so via the *Sign-CG* method.

Section 6.2 develops the *Rescaled Erroneous Conditional Gradient* method specially devised for optimization of a convex function over sets satisfying a sign-preservation property which include any ℓ_p-norm ball ($p \geq 1$). Results for a general sign-preserving set requires a CWEO, while the guarantees specifically for ℓ_2-norm ball only require a general EO.

6.1. The Sign Conditional Gradient

In this section we utilize the proof of Theorem 4.2 to develop the Sign Conditional Gradient (SCG) scheme described in Algorithm 3 and to establish that its theoretical guarantees are the same as that of the ECG when the feasible set is the box; the proof is deferred to Appendix F.1.

Algorithm 3: Sign Conditional gradient (SCG)

```
Input: $w^0 \in \mathcal{C}, \varepsilon \geq 0$ .
for any $t \geq 0$ do
    retrieve $\hat{g}^t \leftarrow \text{sign}(\nabla f(x))$ ;
    compute $p^{t+1} \leftarrow \text{LMO}(\hat{g}^t) - w^t$;
    choose $\eta_t \in [0,1]$ and set $w^{t+1} \leftarrow w^t + \eta_t p^{t+1}$;
end
```

Theorem 6.1 (CG is robust under separability)**.** *Suppose that $\mathcal{C}$ is the box set. Then Algorithm 3 generates that same sequence of points as that of ECG using CWEO with $\varepsilon = 0$.*

Corollary 6.2 (informal)**.** *Suppose that $\mathcal{C}$ is the box set. Then any theoretical guarantee satisfied for the CG holds true for the SCG.*

6.2. The Rescaled Erroneous Conditional Gradient

So-far we have explored two settings for the CG approach that, to some extent, are on two opposing sides of the spectrum of structural assumptions. The first is the RCG applied over the box with a CWEO, and the other is the RCG applied over general (compact) convex sets with a general EO. For the former we showed that the relative-error does not affect the theoretical guarantees, while for the latter we obtained an additive error dependency.

This section introduces a rescaling CG approach that can be positioned between these two ends that exploits structural properties of a class of sets to derive a *relative guarantee* with respect to the optimal value independent of M. The method implementing this is called the *Rescaled Erroneous*

Conditional Gradient (RECG), and is described by Algorithm 4.

Here as-well we investigate two cases: (i) Minimization over the Euclidean ball, $\mathbb{B}[0,R] := \{x \in \mathbb{R}^n : \|x\| \leq R\}$, with a **general EO**; and, (ii) Minimization over sign-preserving sets (cf. Definition 6.5 below) with a **CWEO**. Note that any ℓ_p-norm ball ($p \geq 1$) is a sign-preserving set, which implies that the second case contains the first when the EO is a CWEO.

The relative guarantee is defined with respect to a multiplicative baseline θf_{opt}, where $\theta := (1-\varepsilon)/(1+\varepsilon) \in [0,1]$, and, as usual, ε is the error parameter of the gradient oracle. Note that θf_{opt} is a reasonable baseline only when $f_{\text{opt}} \leq 0$. Indeed, in this case we have $f_{\text{opt}} \leq \theta f_{\text{opt}}$, implying that we relax the baseline compared to the standard case where $\varepsilon = 0$.

Algorithm 4: Rescaled Erroneous Conditional Gradient (RECG)

```
Input: $w^0 \in \mathcal{C}, \mathcal{C}\ \varepsilon, \theta \in [0,1)$, Oracle $\mathcal{O}(\cdot, \varepsilon)$
for any $t \geq 0$ do
    retrieve $\hat{g}^t \leftarrow \mathcal{O}(w^t, \varepsilon)$ ;
    set $\hat{p}^t \leftarrow \text{LMO}(\hat{g}^t)$;
    choose $\eta^t \in [0,1]$ and set
      $w^{t+1} \leftarrow (1-\eta^t)w^t + \eta^t \hat{p}^t/\theta$;
end
Output: $\bar{w}^T = \theta w^T$
```

Remark. Note that as we prove below, the iterates w^t , $t \in [T]$ that are computed by Algorithm 4 belong to the ball or radius R/θ around the origin. Therefore, in the last step of the algorithm we re-scale $\bar{w}^T = \theta w^T$ to make sure that the output $\bar{w}^T$ resides in $\mathbb{B}[0,R]$.

6.2.1. Euclidean ball with a general EO

Before proceeding, we state concisely the setting of this subsection: (i) f is convex, and for simplicity of the analysis, $\max_{x \in \mathcal{C}} f(x) \leq 0$; (ii) $\mathcal{C} = \mathbb{B}[0,R]$; (iii) a **general EO** is used with $\varepsilon \in [0,1)$; and establish the following technical result detailed in Appendix F.2.

Lemma 6.3. *Let $r > 0$, and let $\{\delta^t\}_{t=1}^T$ be a sequence of non-negative real numbers such that $\delta^1 \leq r$ and*

$$\delta_{t+1} \leq (1-\eta^t)\delta^t + r(\eta^t)^2 \quad \forall t \geq 1,$$

where $\eta^t = 2/(t+1)$. Then $\delta^t \leq \frac{2r}{t+1}$ for any $t \geq 1$.

Theorem 6.4 shows that the RECG achieves a $\theta = (1-\varepsilon)/(1+\varepsilon)$ relative-error guarantee on the sequence of function values with respect to the optimal value; when $\varepsilon = 0$, we indeed recover the standard convergence result. The proof is given in Appendix F.2.

Theorem 6.4. *Assume that f is convex, $\mathcal{C} = \mathbb{B}[0,R]$, and $\max_{x \in \mathcal{C}} f(x) \leq 0$, also assume that $\mathcal{O}(\cdot, \varepsilon)$ is a general (EO) with parameter $\varepsilon \in [0,1)$. Set $\theta = (1-\varepsilon)/(1+\varepsilon)$ and $\eta^t = 2/(t+1)$. Then after $T \geq 1$ iterations the RECG generates a point $\bar{w}^T = \theta w^T$ satisfying,*

$$\bar{w}^T \in \mathbb{B}[0,R] \quad \text{and} \quad f(\bar{w}^T) - \theta f_{\text{opt}} \leq \frac{8L(R^2/\theta)}{T+1}.$$

6.2.2. Sign-preserving sets with a CWEO

In this section we show that by using a **CWEO** (cf. (2)), instead of the general EO, we can establish similar results as in Theorem 6.4 for a larger class of sets we call *Sign-Preserving Sets*.

Definition 6.5 (sign-preserving set)**.** A set $\mathcal{B} \subset \mathbb{R}^n$ is a sign-preserving set if it is closed, convex, contains the origin (i.e. $\mathbf{0} \in \mathcal{B}$), and satisfies that for any $g \in \mathbb{R}^n$ it holds that

$$-g_i \cdot p_i \geq 0 \ , \qquad \forall i \in [n],$$

where $p = \text{LMO}(g) \equiv \operatorname*{argmin}_{d \in \mathcal{B}} g^\top d$.

It is not hard to show that any closed and convex set $\mathcal{B}$ satisfying that zeroing a component in $x \in \mathcal{B}$ keeps it in the feasible set, i.e., $x \in \mathcal{B} \Rightarrow \forall i \in [n], x - x_i e_i \in \mathcal{B}$, is a sign-preserving set; this property is fundamental in sparse optimization for example, as the ability to zero components is essential. In particular, any ℓ_p-norm ball with $p \geq 1$ is a sign-preserving set.

Before proceeding, we state concisely the setting of this section: (i) f is convex, and for simplicity of the analysis, $\max_{x \in \mathcal{C}} f(x) \leq 0$; (ii) $\mathcal{C}$ is a sign-preserving set; (iii) and we assume the availability of **CWEO** with parameter $\varepsilon \in [0,1)$. We also note that in-spite of the similarity in techniques of the following results and those in Section 6.2.1, we prove them separately for simplicity and ease of reading. The proofs of this subsection are located in Appendix F.3.

We are now ready to state and establish our main theorem.

Theorem 6.6. *Assume that f is convex, $\mathcal{C}$ is a sign-preserving set, and $\max_{x \in \mathcal{C}} f(x) \leq 0$, also assume that $\mathcal{O}(\cdot, \varepsilon)$ is a CWEO with parameter $\varepsilon \in [0,1)$. Set $\theta = (1-\varepsilon)/(1+\varepsilon)$ and $\eta^t = 2/(t+1)$. Then after $T \geq 1$ iterations the RECG generates a point $\bar{w}^T = \theta w^T$ satisfying,*

$$\bar{w}^T \in \mathcal{C} \quad \text{and} \qquad f(\bar{w}^T) - \theta f_{\text{opt}} \leq \frac{8L(R^2/\theta)}{T+1} \ .$$

7. Conclusions and Future Work

This work investigates optimization procedures with a deterministic EO in a variety of constrained settings, establishing convergence guarantees with or without dependency on the relative-error under different structural assumptions. We showed that it is possible to obtain meaningful guarantees that go beyond what is known for the setting of a general and coordinate-wise erroneous gradient oracles, laying foundations for new strategies to handle erroneous gradients.

Two main veins of research stem from our work: (i) Studying random EOs appearing in both the derivative-free literature and in the distributed optimization literature, see (Berahas et al., 2022) and (Condat et al., 2022) and references therein, respectively; (ii) Extending the results to weaker types of EO such as the one associated with the inner-product test in (Bollapragada et al., 2018). Another interesting research direction is to incorporate acceleration mechanisms (à la Nesterov) with such an oracle at hand. Finally, the guarantees of the projected gradient scheme for convex constrained problems without the PL assumption remains an open question that calls for further study.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgments

Nadav Hallak's work was supported by the ISRAEL SCIENCE FOUNDATION (grant No. 637/21). Kfir Y. Levy has received funding from the Israeli Science Foundation (grant No. 447/20).

References

Ajalloeian, A. and Stich, S. U. On the convergence of sgd with biased gradients. *arXiv preprint arXiv:2008.00051*, 2020.

Beck, A. *Introduction to nonlinear optimization*, volume 19 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2014. ISBN 978-1-611973-64-8.

Beck, A. *First-Order Methods in Optimization*, volume 25. SIAM, 2017.

Berahas, A. S., Cao, L., Choromanski, K., and Scheinberg, K. Linear interpolation gives better gradients than gaussian smoothing in derivative-free optimization. *arXiv preprint arXiv:1905.13043*, 2019.

Berahas, A. S., Cao, L., Choromanski, K., and Scheinberg, K. A theoretical and empirical comparison of gradient approximations in derivative-free optimization. *Foundations of Computational Mathematics*, 22(2):507–560, 2022.

Bertsekas, D. P. *Nonlinear programming*. Athena scientific Belmont, 2003.

Beznosikov, A., Horváth, S., Richtárik, P., and Safaryan, M. On biased compression for distributed learning. *arXiv preprint arXiv:2002.12410*, 2020.

Bollapragada, R., Byrd, R., and Nocedal, J. Adaptive sampling strategies for stochastic optimization. *SIAM Journal on Optimization*, 28(4):3312–3343, 2018.

Byrd, R. H., Chin, G. M., Nocedal, J., and Wu, Y. Sample size selection in optimization methods for machine learning. *Mathematical programming*, 134(1):127–155, 2012.

Carter, R. G. On the global convergence of trust region algorithms using inexact gradient information. *SIAM Journal on Numerical Analysis*, 28(1):251–265, 1991.

Cartis, C. and Scheinberg, K. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming*, 169: 337–375, 2018.

Chmiel, B., Ben-Uri, L., Shkolnik, M., Hoffer, E., Banner, R., and Soudry, D. Neural gradients are near-lognormal: improved quantized and sparse training. *arXiv: Computer Vision and Pattern Recognition*, 2021.

Condat, L., Yi, K., and Richtárik, P. Ef-bv: A unified theory of error feedback and variance reduction mechanisms for biased and unbiased compression in distributed optimization. *arXiv preprint arXiv:2205.04180*, 2022.

Conn, A. R., Gould, N. I., and Toint, P. L. *Trust region methods*. SIAM, 2000.

Conn, A. R., Scheinberg, K., and Vicente, L. N. *Introduction to derivative-free optimization*. SIAM, 2009.

d'Aspremont, A. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, 2008.

Devolder, O., Glineur, F., and Nesterov, Y. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1):37–75, 2014.

Dvurechensky, P. Gradient method with inexact oracle for composite non-convex optimization. *arXiv preprint arXiv:1703.09180*, 2017.

Dvurechensky, P. and Gasnikov, A. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *Journal of Optimization Theory and Applications*, 171(1):121–145, 2016.

Frank, M. and Wolfe, P. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2): 95–110, 1956.

Hintermüller, M. and Vicente, L. N. Space mapping for optimal control of partial differential equations. *SIAM Journal on Optimization*, 15(4):1002–1025, 2005.

Jaggi, M. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pp. 427–435. PMLR, 2013.

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

Paquette, C. and Scheinberg, K. A stochastic line search method with expected complexity analysis. *SIAM Journal on Optimization*, 30(1):349–376, 2020.

Polyak, B. T. Introduction to optimization. 1987.

Richtárik, P., Sokolov, I., and Fatkhullin, I. Ef21: A new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34, 2021.

Stonyakin, F. S., Dvinskikh, D., Dvurechensky, P., Kroshnin, A., Kuznetsova, O., Agafonov, A., Gasnikov, A., Tyurin, A., Uribe, C. A., Pasechnyuk, D., et al. Gradient methods for problems with inexact model of the objective. In *International Conference on Mathematical Optimization Theory and Operations Research*, pp. 97–114. Springer, 2019.

A. Supplementals

Table 2 below serves as a more convenient quick guide to the content of this paper than Table 1 in the body of the text.

Table 2. Complexity results and their respective settings and locations; subsequence convergence guarantees do not appear in the table. When the relative error does not affect the theoretical guarantees, we say that these guarantees are 'preserved'.

Oracle	**Set**	**Alg.**	**fun.**	**Guarantee**	**Section**
CWEO	box	ECG	convex/nonconvex	preserved	Section 4.1
		EPG	nonconvex	preserved	Section 4.2
		SCG	convex/nonconvex	preserved	Section 6.1
	sign-preserving	RECG	convex	$f(\bar{w}^T) - \theta f_{\text{opt}} \leq \frac{8L(R^2/\theta)}{T+1}$	Section 6.2.2
EO	general	ECG	nonconvex	$\min_{d \in \mathcal{C}} \langle \nabla f(w^*), d - w^* \rangle \geq -\frac{2\varepsilon}{1-\varepsilon} MR$	Section 5.1
			convex	$f(w^t) - f_{\text{opt}} \leq 2\varepsilon MR + \frac{4LR^2}{t+2}$	Section 5.1
		EPG	nonconvex	$\min_{t \in [T]} \lVert q^{t+1} + \nabla f(w^{t+1}) \rVert^2 \leq \frac{8L(f^0 - f_{\text{opt}})}{T} + \epsilon$	Section 5.2
	ℓ_2-ball	RECG	convex	$f(\bar{w}^T) - \theta f_{\text{opt}} \leq \frac{8L(R^2/\theta)}{T+1}$	Section 6.2

B. Proofs for Claims in Section 2

Proof of Lemma 2.2. The first and second relations follow from the triangle inequality together with the error bound of the EO. The third relation follows from combining the Cauchy-Schwartz inequality with the first relation. □

Proof of Lemma 2.3. The first and second claims follow straightly from the coordinate-wise bound, $|\hat{g}_i - \nabla f(x)_i| \leq \varepsilon |\nabla f(x)_i|$. For the first claim, note that the CWEO condition implies that $\forall i \in [n]$,

$$\nabla f(x)_i - \varepsilon |\nabla f(x)_i| \leq \hat{g}_i \leq \nabla f(x)_i + \varepsilon |\nabla f(x)_i| \,.$$

Thus, if $\nabla f(x)_i \leq 0$ then $\hat{g}_i \leq \nabla f(x)_i + \varepsilon |\nabla f(x)_i| = (1 - \varepsilon) \nabla f(x)_i \leq 0$. Similarly, if $\nabla f(x)_i \geq 0$ then $\hat{g}_i \geq \nabla f(x)_i - \varepsilon |\nabla f(x)_i| = (1 - \varepsilon) \nabla f(x)_i \geq 0$. Which establishes the first part.

The second claim is derived by applying the triangle inequality and simple manipulations,

$$\max\{|\hat{g}_i| - |\nabla f(x)_i|, |\nabla f(x)_i| - |\hat{g}_i|\} \leq |\hat{g}_i - \nabla f(x)_i| \leq \varepsilon |\nabla f(x)_i|.$$

We will now prove the third claim. Set $g := \nabla f(x)$. From the first claim we have that $\text{sign}(g_i \cdot d_i) = \text{sign}(\hat{g}_i \cdot d_i)$ for any $i \in [n]$. Thus, from the second claim we have that

$$g_i \cdot d_i = \text{sign}(g_i \cdot d_i) |g_i \cdot d_i| = \text{sign}(\hat{g}_i \cdot d_i) |g_i \cdot d_i| \geq \begin{cases} -\frac{|\hat{g}_i||d_i|}{1-\varepsilon}, & g_i \cdot d_i \leq 0, \\ \frac{|\hat{g}_i||d_i|}{1+\varepsilon}, & g_i \cdot d_i > 0, \end{cases}$$

and

$$g_i \cdot d_i = \text{sign}(g_i \cdot d_i) |g_i \cdot d_i| = \text{sign}(\hat{g}_i \cdot d_i) |g_i \cdot d_i| \leq \begin{cases} -\frac{|\hat{g}_i||d_i|}{1+\varepsilon}, & g_i \cdot d_i \leq 0, \\ \frac{|\hat{g}_i||d_i|}{1-\varepsilon}, & g_i \cdot d_i > 0. \end{cases}$$

Utilizing the above we deduce that

$$\begin{aligned} \langle g, d \rangle = \sum_{i : g_i d_i > 0} |g_i||d_i| - \sum_{i : g_i d_i < 0} |g_i||d_i| &\geq \frac{1}{1+\varepsilon} \sum_{i : g_i d_i > 0} |\hat{g}_i||d_i| - \frac{1}{1-\varepsilon} \sum_{i : g_i d_i < 0} |\hat{g}_i||d_i| \\ &= \frac{1}{1-\varepsilon^2} \left(\langle \hat{g}, d \rangle - \varepsilon \langle |\hat{g}|, |d| \rangle \right) \end{aligned}$$

and

$$\begin{aligned} \langle g, d \rangle = \sum_{i : g_i d_i > 0} |g_i||d_i| - \sum_{i : g_i d_i < 0} |g_i||d_i| &\leq \frac{1}{1-\varepsilon} \sum_{i : g_i d_i > 0} |\hat{g}_i||d_i| - \frac{1}{1+\varepsilon} \sum_{i : g_i d_i < 0} |\hat{g}_i||d_i| \\ &= \frac{1}{1-\varepsilon^2} \left(\langle \hat{g}, d \rangle + \varepsilon \langle |\hat{g}|, |d| \rangle \right). \end{aligned}$$

The boundedness of $\mathcal{C}$ together with the second claim then yield that

$$\begin{aligned}\langle|\hat{g}|,|d|\rangle &\leq \|\hat{g}\|\|d\| \leq (1+\varepsilon)MR,\\ -\langle|\hat{g}|,|d|\rangle &\geq -\|\hat{g}\|\|d\| \geq -(1+\varepsilon)MR.\end{aligned}$$

Plugging these bounds to the former ones results with the required relations. □

C. Proofs for Claims in Section 4.2

Proof of Theorem 4.5. Since $\mathcal{C}$ is separable, updating w^{t+1} is equivalent to applying the coordinate-wise update for any $i \in [n]$

$$w_i^{t+1} = \operatorname{argmin}_{z\in\mathcal{C}_i}\left\{\hat{g}_i^t(z-w_i^t)+\frac{1}{2\eta}(z-w_i^t)^2\right\}. \tag{14}$$

Due to the sign preservation property and the relative coordinate-wise error bounds of the CWEO established in Lemma 2.3, there exists $\alpha_i^t \in [1-\varepsilon, 1+\varepsilon]$ for any $i \in [n]$ and $t \geq 0$ such that

$$w_i^{t+1} = \operatorname{argmin}_{z\in\mathcal{C}_i}\left\{\alpha_i^t g_i^t\cdot(z_i-w_i^t)+\frac{1}{2\eta}(z_i-w_i^t)^2\right\}.$$

That is, (14) can equivalently be written as

$$w_i^{t+1} = \operatorname{argmin}_{z\in\mathcal{C}_i}\left\{g_i^t\cdot(z_i-w_i^t)+\frac{1}{2\eta\alpha_i^t}(z_i-w_i^t)^2\right\}. \tag{15}$$

Equation (15) implies two results:

$$0 \geq g_i^t\cdot(w_i^{t+1}-w_i^t)+\frac{1}{2\eta\alpha_i^t}(w_i^{t+1}-w_i^t)^2, \tag{16}$$

$$0 \in \partial\delta_{C_i}(w_i^{t+1})+g_i^t+\frac{1}{\eta\alpha_i^t}(w_i^{t+1}-w_i^t), \tag{17}$$

where the former follows from the optimality of w_i^{t+1} and the fact that $w_i^t \in \mathcal{C}_i$, and the latter from the first-order optimality conditions associated with the problem in (15).

From (16) and the fact that $\alpha_i^t \in [1-\varepsilon, 1+\varepsilon]$, we obtain

$$\begin{aligned}-\frac{1}{2\eta(1+\varepsilon)}(w_i^{t+1}-w_i^t)^2 &\geq -\frac{1}{2\eta\alpha_i^t}(w_i^{t+1}-w_i^t)^2\\ &\geq g_i^t\cdot(w_i^{t+1}-w_i^t).\end{aligned} \tag{18}$$

Plugging (18) to the descent lemma (cf. Lemma 1.1) then yields $f(w^{t+1})-f(w^t) \leq \langle g^t, w^{t+1}-w^t\rangle+\frac{L}{2}\|w^{t+1}-w^t\|^2 \leq \left(\frac{L}{2}-\frac{1}{2\eta(1+\varepsilon)}\right)\|w^{t+1}-w^t\|^2$, which implies the sufficient decrease property

$$f(w^t)-f(w^{t+1}) \geq \frac{1-L\eta(1+\varepsilon)}{2\eta(1+\varepsilon)}\|w^{t+1}-w^t\|^2. \tag{19}$$

From (17) we obtain that there exists $q^{t+1} \in \partial\delta_{\mathcal{C}}(w^{t+1})$ such that for any $i \in [n]$

$$q_i^{t+1}+g_i^{t+1}+(g_i^t-g_i^{t+1})+\frac{1}{\eta\alpha_i^t}(w_i^{t+1}-w_i^t) = 0. \tag{20}$$

Thus, using the triangle inequality and the Lipschitz continuity of the gradient, we obtain that

$$\begin{aligned}\frac{1}{\eta(1-\varepsilon)}\|w^{t+1}-w^t\| &\geq \max_{i\in[n]}\left\{\frac{1}{\eta\alpha_i^t}\right\}\cdot\|w^{t+1}-w^t\|\\ &\geq \|q^{t+1}+g^{t+1}\|-L\|w^t-w^{t+1}\|,\end{aligned}$$

and subsequently,

$$\frac{1+L\eta(1-\varepsilon)}{\eta(1-\varepsilon)}\|w^{t+1}-w^t\| \geq \|q^{t+1}+g^{t+1}\|. \tag{21}$$

Plugging (21) to (19) yields

$$\begin{aligned} f(w^t)-f(w^{t+1}) &\geq \frac{1-L\eta(1+\varepsilon)}{2\eta(1+\varepsilon)}\|w^{t+1}-w^t\|^2 \\ &\geq \frac{\eta(1-\varepsilon)^2}{2(1+\varepsilon)}\frac{1-L\eta(1+\varepsilon)}{(1+L\eta(1-\varepsilon))^2}\|q^{t+1}+g^{t+1}\|^2 \\ &= c(\eta,\varepsilon,L)\|q^{t+1}+g^{t+1}\|^2. \end{aligned} \tag{22}$$

Summing (22) from $t=0$ to $t=T-1$ and the minimal element in the summation results with

$$f(w^0)-f_{\text{opt}} \quad \geq c(\eta,\varepsilon,L)T\min_{t=1,\dots,T-1}\|q^{t+1}+g^{t+1}\|^2. \tag{23}$$

Thus, $\|q^{t+1}+g^{t+1}\| \to 0$ as $t\to\infty$ and

$$\min_{t=1,\dots,T-1}\|q^{t+1}+g^{t+1}\|^2 \leq \frac{f(w^0)-f_{\text{opt}}}{c(\eta,\varepsilon,L)T}.$$

Plugging $\eta=\frac{1}{2(1+\varepsilon)L}$ yields the required (8).

Let $\{w^{t_j}\}_{j\geq 0}$ be a subsequence of $\{w^t\}_{t\geq 0}$ converging to w^*. Since f is lower bounded and monotonic decreasing in the sequence $\{w^t\}_{t\geq 0}$ due to the boundedness of $\mathcal{C}$ and (19), the sequence $\{f(w^t)\}_{t\geq 0}$ converges to a limit point f^*. Moreover, the continuity of f guarantees that $f^*=f(w^*)$. Then by taking a limit for the elements in the converging subsequence we obtain using the fact that $\|q^{t_j+1}+g^{t_j+1}\|\to 0$ as $j\to\infty$, and by utilizing the closeness of the graph of the subdifferential and the continuity of the gradient, that

$$0\in\partial\delta_{\mathcal{C}}(w^*)+\nabla f(w^*), \qquad \forall i\in[n], \tag{24}$$

meaning that w^* is a stationary point of (P). If the objective function is in addition convex, then this guarantees that w^* is an optimal solution due to the sufficiency of the stationarity conditions in convex problems. □

D. Proofs for Claims in Section 5.1

Proof of Lemma 5.1. By Lemma 1.1, the properties of the EO together with Lemma 2.2, and the triangle inequality,

$$\begin{aligned} f^{t+1}-f^t \leq \eta_t\langle g^t,p^{t+1}\rangle+\frac{L\eta_t^2}{2}\|p^{t+1}\|^2 &= \eta_t\left(\langle\hat{g}^t,p^{t+1}\rangle+\langle g^t-\hat{g}^t,p^{t+1}\rangle\right)+\frac{L\eta_t^2}{2}\|p^{t+1}\|^2 \\ &\leq \eta_t\left(\langle\hat{g}^t,p^{t+1}\rangle+\|g^t-\hat{g}^t\|\|p^{t+1}\|\right)+\frac{L\eta_t^2}{2}\|p^{t+1}\|^2 \\ &\leq \eta_t\left(\langle\hat{g}^t,p^{t+1}\rangle+\frac{\varepsilon}{1-\varepsilon}\|\hat{g}^t\|\|p^{t+1}\|\right)+\frac{L\eta_t^2}{2}\|p^{t+1}\|^2. \end{aligned}$$

If $\eta_t=\gamma_t=-\frac{\langle\hat{g}^t,p^{t+1}\rangle+\frac{\varepsilon}{1-\varepsilon}\|\hat{g}^t\|\|p^{t+1}\|}{L\|p^{t+1}\|^2}$. Then $f^{t+1}-f^t\leq-\eta_t^2L\|p^{t+1}\|^2+\frac{L\eta_t^2}{2}\|p^{t+1}\|^2=-\frac{1}{2}L\|p^{t+1}\|^2\eta_t^2$. Otherwise, assume that $\eta_t=1$, which requires that

$$\gamma_t=-\frac{\langle\hat{g}^t,p^{t+1}\rangle+\frac{\varepsilon}{1-\varepsilon}\|\hat{g}^t\|\|p^{t+1}\|}{L\|p^{t+1}\|^2}\geq 1.$$

Then

$$f^{t+1}-f^t\leq\eta_t\left(\langle\hat{g}^t,p^{t+1}\rangle+\frac{\varepsilon}{1-\varepsilon}\|\hat{g}^t\|\|p^{t+1}\|\right)+\frac{L\eta_t^2}{2}\|p^{t+1}\|^2\leq-\frac{L\|p^{t+1}\|^2}{2}.$$

Hence,

$$f^t - f^{t+1} \geq \frac{1}{2}L\|p^{t+1}\|^2 \min\{1, \eta_t^2\} \geq \frac{1}{2}L\|p^{t+1}\|^2\eta_t^2.$$

which concludes the correctness of (9).

The descending property then follows from (9) together with the definition of the step-size. □

Proof of Theorem 5.2. 1. Lemma 5.1 implies that $\{f^t\}_{t\geq 0}$ is monotonic non-ascending. Since f is bounded below over $\mathcal{C}$, this implies that $\{f^t\}_{t\geq 0}$ converges to a limit point, say f^*.

2. Consider the three possible range of values for γ_t for any $t \geq 0$: (i) $\gamma_t < 0$, (ii) $\gamma_t = 0$, (iii) $\gamma_t > 0$.

   For any $t \geq 0$, $\gamma_t < 0$ implies that $\langle \hat{g}^t, p^{t+1}\rangle + \frac{\varepsilon}{1-\varepsilon}\|\hat{g}^t\|\|p^{t+1}\| > 0$, and $\gamma_t = 0$ implies that $\langle \hat{g}^t, p^{t+1}\rangle + \frac{\varepsilon}{1-\varepsilon}\|\hat{g}^t\|\|p^{t+1}\| = 0$. Hence, for any $t \geq 0$ for which $\gamma_t \leq 0$ we have that $\langle \hat{g}^t, p^{t+1}\rangle + \frac{\varepsilon}{1-\varepsilon}\|\hat{g}^t\|\|p^{t+1}\| \geq 0$.

   Now consider $t \geq 0$ in which $\gamma_t > 0$ and set $J := \{t \geq 0 : \gamma_t > 0\}$. By the definition of γ_t, its positiveness implies that $\|p^{t+1}\|, \eta_t > 0$ for any $t \in J$. Due to the first part of this theorem, taking the limit $t \to \infty$ in relation (9) yields that $\|p^{t+1}\|\eta_t \to 0$. This is true in particular for any $t \in J$ for which $\|p^{t+1}\|\eta_t > 0$. Consequently, either J is finite (implying the correctness of the claim) or at least one of these terms must converge to zero (recall that both are positive). Obviously, if $\|p^{t+1}\| \to 0$ then $\lim_{t\to\infty}\left(\langle \hat{g}^t, p^{t+1}\rangle + \frac{\varepsilon}{1-\varepsilon}\|\hat{g}^t\|\|p^{t+1}\|\right) = 0$.

   Suppose that $\eta_t \to 0$. Then there exists $K_0 > 0$ such that $\eta_t < 1$ for any $t > K_0$. Without loss of generality, assume that J is a subsequence for which $t > K_0$ and $\gamma_t > 0$; note that we already established that for any $t > K_0$, $t \notin J$, it holds that $\langle \hat{g}^t, p^{t+1}\rangle + \frac{\varepsilon}{1-\varepsilon}\|\hat{g}^t\|\|p^{t+1}\| \geq 0$. Subsequently, by the definition of η_t, for any $t \in J$ we have that $\eta_t = \gamma_t$. Therefore, from the assumption that $\eta_t \to 0$, we conclude that $\lim_{t\in J, t\to\infty}\left(\langle \hat{g}^t, p^{t+1}\rangle + \frac{\varepsilon}{1-\varepsilon}\|\hat{g}^t\|\|p^{t+1}\|\right) = 0$.

   Finally, from combining the two deductions above: $\langle \hat{g}^t, p^{t+1}\rangle + \frac{\varepsilon}{1-\varepsilon}\|\hat{g}^t\|\|p^{t+1}\| \geq 0$ for any $t \geq K_0$ such that $\gamma_t \leq 0$, and $\lim_{t\to\infty}\left(\langle \hat{g}^t, p^{t+1}\rangle + \frac{\varepsilon}{1-\varepsilon}\|\hat{g}^t\|\|p^{t+1}\|\right) = 0$ for any $t \geq K_0$ such that $\gamma_t > 0$, we obtain the required relation.

3. From the definition of p^{t+1} and the property of the EO we have that for any $d \in \mathcal{C}$

$$\begin{aligned}\langle g^t, d - w^t\rangle &= \langle \hat{g}^t, d - w^t\rangle + \langle g^t - \hat{g}^t, d - w^t\rangle \\ &\geq \langle \hat{g}^t, p^{t+1}\rangle - \varepsilon\|g^t\|\|d - w^t\| \\ &\geq \langle \hat{g}^t, p^{t+1}\rangle - \varepsilon MR.\end{aligned}$$

   Thus, from Lemma 2.2, for any $d \in \mathcal{C}$ it holds that

$$\begin{aligned}&\langle g^t, d - w^t\rangle + \frac{\varepsilon(1+\varepsilon)}{1-\varepsilon}MR \\ &\quad\geq \langle g^t, d - w^t\rangle + \frac{\varepsilon}{1-\varepsilon}\|\hat{g}^t\|\|p^{t+1}\| \\ &\quad\geq \langle \hat{g}^t, p^{t+1}\rangle - \varepsilon MR + \frac{\varepsilon}{1-\varepsilon}\|\hat{g}^t\|\|p^{t+1}\|.\end{aligned}$$

   Taking a limit over the converging subsequence while invoking the previous part and rearranging terms finally yields (10).

□

Proof of Lemma 5.3. Set $u^{t+1} = \arg\min_{d\in\mathcal{C}}\langle g^t, d - w^t\rangle - w^t$. Then from the convexity of f, the choice of p^{t+1}, and the definition of the EO,

$$\begin{aligned}f(w^t) - f(w^*) \leq \langle g^t, w^t - w^*\rangle \leq -\min_{d\in\mathcal{C}}\langle g^t, d - w^t\rangle &= -\langle \hat{g}^t, p^{t+1}\rangle + \langle \hat{g}^t, p^{t+1}\rangle - \langle g^t, u^{t+1}\rangle \\ &\leq -\langle \hat{g}^t, p^{t+1}\rangle + \langle \hat{g}^t - g^t, u^{t+1}\rangle \\ &\leq -\langle \hat{g}^t, p^{t+1}\rangle + \|\hat{g}^t - g^t\|\|u^{t+1}\| \\ &\leq -\langle \hat{g}^t, p^{t+1}\rangle + \varepsilon\|g^t\|\|u^{t+1}\|.\end{aligned}$$

Finally, the bound follows from the boundedness of the feasible set $\|g^t\|\|u^{t+1}\| \leq RM$. □

Proof of Theorem 5.4. For convenience we denote $f^t := f(w^t)$ and $c = 4LR^2$; note that $f_{\text{opt}} = f(w^*)$. By the descent lemma (cf. Lemma 1.1) together with the boundedness of $\mathcal{C}$ and the step-size regime,

$$\begin{aligned} f^{t+1} - f_{\text{opt}} &\le f^t - f_{\text{opt}} + \eta_t \langle g^t, p^{t+1} \rangle + \frac{L\eta_t^2}{2}\|p^{t+1}\|^2 \\ &\le f^t - f_{\text{opt}} + \frac{2}{t+2}\langle g^t, p^{t+1}\rangle + \frac{4LR^2}{(t+2)^2} \\ &= f^t - f_{\text{opt}} + \frac{2}{t+2}\langle g^t, p^{t+1}\rangle + \frac{c}{(t+2)^2}. \end{aligned} \tag{25}$$

Using the properties of the EO together with Lemma 5.3, we obtain $\langle g^t, p^{t+1}\rangle = \langle \hat{g}^t, p^{t+1}\rangle + \langle g^t - \hat{g}^t, p^{t+1}\rangle \le f_{\text{opt}} - f^t + \varepsilon MR + \|g^t - \hat{g}^t\|\|p^{t+1}\| \le f_{\text{opt}} - f^t + \varepsilon MR + \varepsilon\|g^t\|\|p^{t+1}\| \le f_{\text{opt}} - f^t + 2\varepsilon MR$. Plugging this to (25) yields

$$\begin{aligned} f^{t+1} - f_{\text{opt}} &\le f^t - f_{\text{opt}} + \frac{2}{t+2}\langle g^t, p^{t+1}\rangle + \frac{c}{(t+2)^2} \\ &\le f^t - f_{\text{opt}} + \frac{2}{t+2}\left(f_{\text{opt}} - f^t + 2\varepsilon MR\right) + \frac{c}{(t+2)^2} \\ &= \frac{t}{t+2}(f^t - f_{\text{opt}}) + \frac{4\varepsilon MR}{t+2} + \frac{c}{(t+2)^2}. \end{aligned} \tag{26}$$

We will now use induction to conclude the claimed. For $t = 0$ we indeed have from the above that $f^1 - f_{\text{opt}} \le \frac{4\varepsilon MR}{2} + \frac{c}{4} \le 2\varepsilon MR + \frac{c}{2}$. Assume that

$$f^t - f_{\text{opt}} \le 2\varepsilon MR + \frac{c}{t+2}. \tag{27}$$

We will show that (11) holds true for $t+1$. Using the relation (26) and the assumption of the induction in (27),

$$\begin{aligned} f^{t+1} - f_{\text{opt}} &\le \frac{t}{t+2}(f^t - f_{\text{opt}}) + \frac{4\varepsilon MR}{t+2} + \frac{c}{(t+2)^2} \\ &\le \frac{t}{t+2}\left(2\varepsilon MR + \frac{c}{t+2}\right) + \frac{4\varepsilon MR}{t+2} + \frac{c}{(t+2)^2} \\ &= 2\varepsilon MR + \frac{t+1}{(t+2)^2}c. \end{aligned}$$

The required follows from the fact that $\frac{t+1}{(t+2)^2}c \le \frac{c}{t+3}$. □

E. Technical Proofs for Claims in Section 5.2

Proof of Lemma 5.5. By the descent lemma (Lemma 1.1), Cauchy-Schwartz inequality, properties of the EO, and the update procedure of Algorithm 2 (cf. Remark 3.1),

$$\begin{aligned} f^{t+1} - f^t &\le \langle g^t, w^{t+1} - w^t\rangle + \frac{L}{2}\|w^{t+1} - w^t\|^2 \\ &= \langle \hat{g}^t, w^{t+1} - w^t\rangle + \langle g^t - \hat{g}^t, w^{t+1} - w^t\rangle + \frac{L}{2}\|w^{t+1} - w^t\|^2 \\ &\le \|g^t - \hat{g}^t\|\|w^{t+1} - w^t\| + \frac{L\eta - 1}{2\eta}\|w^{t+1} - w^t\|^2 \\ &\le \varepsilon\|g^t\|R + \frac{L\eta - 1}{2\eta}\|w^{t+1} - w^t\|^2 \\ &\le \varepsilon MR + \frac{L\eta - 1}{2\eta}\|w^{t+1} - w^t\|^2. \end{aligned}$$

Thus,

$$f^t - f^{t+1} \ge \frac{(1 - L\eta)\eta}{2}\|\eta^{-1}(w^{t+1} - w^t)\|^2 - \varepsilon MR,$$

and the required follows. □

Proof of Theorem 5.6. By the update procedure of w^{t+1},

$$w^{t+1} = \mathrm{argmin}_{z\in\mathcal{C}} \left\{ \langle \hat{g}^t, z - w^t \rangle + \frac{1}{2\eta}\|z - w^t\|^2 \right\},$$

we have from the first-order optimality conditions that there exists $q^{t+1} \in \partial\delta_{\mathcal{C}}(w^{t+1})$ such that

$$q^{t+1} + \hat{g}^t + \frac{1}{\eta}(w^{t+1} - w^t) = 0.$$

Thus, using the triangle inequality, the EO definition, and the Lipschitz continuity of the gradient, we have that

$$\begin{aligned}\frac{1}{\eta}\|w^{t+1} - w^t\| = \|q^{t+1} + \hat{g}^t\| &= \|q^{t+1} + g^{t+1} + g^t - g^{t+1} + \hat{g}^t - g^t\| \\ &\geq \|q^{t+1} + g^{t+1}\| - \|g^t - g^{t+1}\| - \|\hat{g}^t - g^t\| \\ &\geq \|q^{t+1} + g^{t+1}\| - L\|w^{t+1} - w^t\| - \varepsilon\|g^t\| \\ &\geq \|q^{t+1} + g^{t+1}\| - L\|w^{t+1} - w^t\| - \varepsilon M.\end{aligned}$$

That is,

$$(1 + \eta L)\left(\|\eta^{-1}(w^{t+1} - w^t)\| + \frac{1}{1+\eta L}\varepsilon M\right) \geq \|q^{t+1} + g^{t+1}\|.$$

Subsequently, using Young's inequality

$$\begin{aligned}\frac{2(\varepsilon M)^2}{(1+L\eta)^2} + 2\|\eta^{-1}(w^{t+1} - w^t)\|^2 &\geq \left(\frac{\varepsilon M}{1+L\eta} + \|\eta^{-1}(w^{t+1} - w^t)\|\right)^2 \\ &\geq \frac{1}{(1+L\eta)^2}\|q^{t+1} + g^{t+1}\|^2.\end{aligned} \tag{28}$$

Plugging (28) to Lemma 5.5 yields that

$$\begin{aligned}f^t - f^{t+1} &\geq \frac{(1-L\eta)\eta}{2}\|\eta^{-1}(w^{t+1} - w^t)\|^2 - \varepsilon MR \\ &\geq \frac{(1-L\eta)\eta}{4(1+L\eta)^2}\left(\|q^{t+1} + g^{t+1}\|^2 - 2(\varepsilon M)^2\right) - \varepsilon MR \\ &= \frac{(1-L\eta)\eta}{4(1+L\eta)^2}\|q^{t+1} + g^{t+1}\|^2 - \varepsilon M\left(R + \frac{\varepsilon M\eta(1-L\eta)}{2(1+L\eta)^2}\right).\end{aligned} \tag{29}$$

Summing (29) from $t = 0$ to $t = T-1$ yields

$$\begin{aligned}f^0 - f_{\text{opt}} \geq f^0 - f^T &\geq \frac{(1-L\eta)\eta}{4(1+L\eta)^2}\sum_{t=0}^{T-1}\|q^{t+1} + g^{t+1}\|^2 - T\varepsilon M\left(R + \frac{\varepsilon M\eta(1-L\eta)}{2(1+L\eta)^2}\right) \\ &\geq T\left(\frac{(1-L\eta)\eta}{4(1+L\eta)^2}\min_{t=0,\ldots,T-1}\|q^{t+1} + g^{t+1}\|^2 - \varepsilon M\left(R + \frac{\varepsilon M\eta(1-L\eta)}{2(1+L\eta)^2}\right)\right).\end{aligned}$$

Thus, using the fact that $\eta L < 1$,

$$\begin{aligned}\frac{4(1+L\eta)^2(f^0 - f_{\text{opt}})}{T\eta(1-L\eta)} &\geq \min_{t=0,\ldots,T-1}\|q^{t+1} + g^{t+1}\|^2 - 2\varepsilon M\left(\frac{2R(1+L\eta)^2}{\eta(1-L\eta)} + \varepsilon M\right) \\ &\geq \min_{t=0,\ldots,T-1}\|q^{t+1} + g^{t+1}\|^2 - 2\varepsilon M\left(\frac{8R}{\eta(1-L\eta)} + \varepsilon M\right)\end{aligned}$$

which establishes the rate result bound.

To establish the subsequence convergence result, note that from Lemma 5.5 and the update criteria of EPG (the condition on h in Algorithm 2) we have that the sequence $\{f^t\}_{t\geq 0}$ is monotonic non-ascending, and thus converges to some value f_{opt}. Moreover, this implies that

$$\lim_{t\to\infty}\left\{\|\eta^{-1}(w^{t+1}-w^t)\|^2-\frac{2\varepsilon MR}{\eta(1-L\eta)}\right\}=0.$$

Let w^* be an accumulation point of the generated sequence where $w^{t_j+1}\to w^*$ as $j\to\infty$. By taking a limit over the converging subsequence in (28) we obtain that

$$\begin{aligned}
&\lim_{j\to\infty}\|q^{t_j+1}+g^{t_j+1}\|^2\\
&\quad\leq 2(\varepsilon M)^2+2(1+L\eta)^2\lim_{j\to\infty}\|\eta^{-1}(w^{t_j+1}-w^{t_j})\|^2\\
&\quad= 2(\varepsilon M)^2+(1+L\eta)^2\frac{4\varepsilon MR}{\eta(1-L\eta)}\\
&\quad\leq 2(\varepsilon M)^2+\frac{16\varepsilon MR}{\eta(1-L\eta)},
\end{aligned}$$

where the last inequality follows from the step-size regime $\eta L<1$. Thus, by the closeness of the graph of the subdifferential $\partial\delta_{\mathcal{C}}(\cdot)$ and the continuity of the gradient $\nabla f(\cdot)$, we have that there exists $q^*\in\partial\delta_{\mathcal{C}}(w^*)$ such that

$$\|q^*+\nabla f(w^*)\|^2\leq\epsilon=2\varepsilon M\left(\varepsilon M+\frac{8R}{\eta(1-L\eta)}\right)$$

as required. □

F. Proofs of Section 6

F.1. Proofs for claims in Section 6.1

Proof of Theorem 6.1. We establish that any element in the sequence $\{p^{t+1}\}_{t\geq 0}$ generated by Algorithm 1 satisfies that $\mathrm{LMO}(\hat{g}^t)=\mathrm{LMO}(\nabla f(w^t))$, which readily implies the required. Denote $g^t=\nabla f(w^t)$, and note that due to the separability of $\mathcal{C}$ we have that the LMO is a separable problem whose solution can be expressed equivalently component-wise as we did in (7).

Thus, due to the separability of $\mathcal{C}$ and the positive homogeneity of the optimization problem we have for any $i\in[n]$ that $\underset{z}{\mathrm{argmin}}\{\hat{g}_i\cdot z: z\in C_i\}\equiv\underset{z}{\mathrm{argmin}}\{\hat{g}_i\cdot|g_i|\cdot z: z\in C_i\}\equiv\underset{z}{\mathrm{argmin}}\{g_i\cdot z: z\in C_i\}$, as claimed. □

F.2. Proofs for claims in Section 6.2.1

Proof of Lemma 6.3. We will prove by induction. The base case holds since $\delta^1\leq r=\frac{2r}{1+1}$. Now, for the induction step, assume that the claim holds true for δ^{t-1} (where $t>1$), and we will show that this implies that it holds true for δ^t. Indeed, using the relation between δ^t,δ^{t-1} we have,

$$\delta^t\leq\left(1-\frac{2}{t}\right)\delta^{t-1}+\frac{r}{2}\frac{4}{t^2}\leq\frac{t-2}{t}\frac{2r}{t}+\frac{2r}{t^2}\leq\frac{2r}{t+1}\;.$$

where we have used the induction hypothesis as well as $\frac{t-1}{t^2}\leq\frac{1}{t+1}$ which holds for $t>1$. This concludes the proof. □

Proof of Theorem 6.4. We prove the assertion in three parts.

Part 1: Note that since $\mathcal{C}=\mathbb{B}[0,R]$ then for any $v\in\mathbb{R}^d$ we have $\mathrm{LMO}(v)=-R\cdot v/\|v\|$. Using this, we can bound the inner product between g^t and $\hat{p}^t$. Indeed,

$$\begin{aligned}
\langle g^t,\hat{p}^t\rangle&=-\frac{R}{\|\hat{g}^t\|}\langle g^t,\hat{g}^t\rangle\leq-(1-\varepsilon)\frac{R}{\|\hat{g}^t\|}\|g^t\|^2\\
&\leq-\frac{1-\varepsilon}{1+\varepsilon}\frac{R}{\|g^t\|}\|g^t\|^2=\theta\cdot\min_{v\in\mathbb{B}[0,R]}\langle g^t,v\rangle,
\end{aligned}\tag{30}$$

where we used the definition of $\hat{p}^t$, as well as Lemma 2.2, and the fact that $\min_{v\in\mathbb{B}[0,R]}\langle g^t, v\rangle = -R\|g^t\|$; we also used $\underset{v\in\mathbb{B}[0,R]}{\operatorname{argmin}}\langle g^t, v\rangle = -Rg^t/\|g^t\|$, and $\theta := (1-\varepsilon)/(1+\varepsilon)$.

Part 2: We now prove that the iterates w^t are bounded. Using induction, we show that $\|w^t\| \le R/\theta$, $\forall t \ge 0$. For the base case, note that $w_0 \in \mathbb{B}[0,R]$ and therefore $\|w_0\| \le R \le R/\theta$ (recall that $\theta \in (0,1]$). Now, for the induction step, assume that $\|w^t\| \le R/\theta$, and let us show that this implies $\|w^{t+1}\| \le R/\theta$. Indeed, by definition, w^{t+1} is a convex combination of two vectors w^t and $\hat{p}^t/\theta$, since these two vectors belong the the ball of radius R/θ, so is their convex combination w^{t+1}. This establishes the induction step.

Part 3: Let x^* be an optimal solution of (P), i.e., $f(x^*) = f_{\text{opt}} = \min_{x\in\mathbb{B}[0,R]} f(x)$. From the update rule of Algorithm 4 we have $w^{t+1} - w^t = \eta^t(\hat{p}^t/\theta - w^t)$. Using the above together with the smoothness of f implies,

$$\begin{aligned} f(w^{t+1}) - f(w^t) \;&\le\; \langle g^t, (w^{t+1} - w^t)\rangle + \frac{L}{2}\|w^{t+1} - w^t\|^2 \\ &= \eta^t\langle g^t, (\hat{p}^t/\theta - w^t)\rangle + \frac{L(\eta^t)^2}{2}\|\hat{p}^t/\theta - w^t\|^2 \\ &\le \eta^t \min_{v\in\mathbb{B}[0,R]}\langle g^t, (v - w^t)\rangle + \frac{L(\eta^t)^2}{2}\|\hat{p}^t/\theta - w^t\|^2 \\ &\le \eta^t\langle g^t, (x^* - w^t)\rangle + \frac{L(\eta^t)^2}{2}\|\hat{p}^t/\theta - w^t\|^2 \\ &\le \eta^t(f(x^*) - f(w^t)) + 2L(\eta^t)^2(R/\theta)^2, \end{aligned}$$

where the second inequality uses (30), and the third inequality uses the fact that $x^* \in \mathbb{B}[0,R]$. The last line uses the gradient inequality, and the fact that $\hat{p}^t/\theta$ and w^t belong to the ball of radius R/θ (see Part 2 of this proof).
Now, denoting $\delta^t = f(w^t) - f(x^*) = f(w^t) - f_{\text{opt}}$ and rearranging the above equation yields,

$$\delta^{t+1} \le (1-\eta^t)\delta^t + 2L(R/\theta)^2(\eta^t)^2.$$

We will now use the Lemma 6.3 to bound δ^t. Recalling that $w^0 \in \mathbb{B}[0,R]$, the smoothness of f implies that,

$$\delta^1 := f(w_0) - f(x^*) \le \frac{L\|x - x^*\|^2}{2} \le LR^2/2 \le 4L(R/\theta)^2,$$

where we used the fact that $\theta \in (0,1]$. Thus, applying Lemma 6.3 and taking $r := 4L(R/\theta)^2$ implies that for any $t \ge 0$,

$$\delta^t := f(w^t) - f_{\text{opt}} \le \frac{8L(R/\theta)^2}{t+1}. \tag{31}$$

Finally,

$$\begin{aligned} f(\bar{w}^T) - \theta f_{\text{opt}} \;&=\; f(\theta w^T + (1-\theta)\cdot 0) - \theta f_{\text{opt}} \\ &\le\; \theta f(w^T) + (1-\theta) f(0) - \theta f_{\text{opt}} \\ &\le\; \theta\left(f(w^T) - f_{\text{opt}}\right) \le\; \frac{8L(R^2/\theta)}{T+1}, \end{aligned}$$

where the first inequality uses the Jensen inequality together with the fact that $\theta \in (0,1)$, the second inequality uses $f(0) \le \max_{x\in\mathbb{B}[0,R]} f(x) \le 0$, and the last inequality uses (31). □

F.3. Proofs for claims in Section 6.2.2

The next lemma is a key element in establishing Theorem 6.6.

Lemma F.1. *Assume that f is convex, $\mathcal{C}$ is a sign-preserving set, and $\max_{x\in\mathcal{C}} f(x) \le 0$, also assume that $\mathcal{O}(\cdot,\varepsilon)$ is a CWEO with parameter $\varepsilon \in [0,1)$. Set $\theta = (1-\varepsilon)/(1+\varepsilon)$. Then for any $x \in \mathcal{C}$, $\hat{g} = \mathcal{O}(x,\varepsilon)$, and $\hat{p} = LMO(\hat{g})$ it holds that*

$$\langle \nabla f(x), \hat{p}\rangle \le \theta \cdot \min_{v\in\mathcal{C}}\langle g, v\rangle\ .$$

Proof of Lemma F.1. Denote $g = \nabla f(x)$ and $p = \mathrm{LMO}(g)$. From Lemma 2.3 for the CWEO, we have that for any $i \in [n]$ there exists $\varepsilon_i \in (-\varepsilon, \varepsilon)$ such that $\hat{g}_i = (1+\varepsilon_i)g_i$, and therefore $g_i = \frac{1}{1+\varepsilon_i}\hat{g}_i$. Consequently,

$$\begin{aligned}\langle g, \hat{p}\rangle = \sum_{i=1}^n g_i \cdot \hat{p}_i = \sum_{i=1}^n \frac{1}{1+\varepsilon_i}\hat{g}_i \cdot \hat{p}_i \leq \frac{1}{1+\varepsilon}\sum_{i=1}^n \hat{g}_i \cdot \hat{p}_i \leq \frac{1}{1+\varepsilon}\sum_{i=1}^n \hat{g}_i \cdot p_i = \frac{1}{1+\varepsilon}\sum_{i=1}^n (1+\varepsilon_i)\hat{g}_i \cdot p_i \\ \leq \frac{1-\varepsilon}{1+\varepsilon}\sum_{i=1}^n \hat{g}_i \cdot p_i = \theta \cdot \min_{v\in\mathcal{C}}\langle g, v\rangle, \end{aligned} \tag{32}$$

where: (i) The first inequality follows from the sign-preservation of $\mathcal{C}$, and due to the fact that $\varepsilon_i \in (-\varepsilon, \varepsilon) \subseteq (-1, 1)$; (ii) The second inequality follows since $\hat{p} = \mathrm{LMO}(\hat{g})$ implies that $\langle \hat{g}, \hat{p}\rangle \leq \langle g, p\rangle$; (iii) The last inequality uses $(1+\varepsilon_i) \geq (1-\varepsilon)$ together with the fact that $g_i \cdot p_i \leq 0$ for all $i \in [n]$ (since $\mathcal{C}$ is sign-preserving); (iv) Lastly, we use the definition of θ and the fact that $p = \mathrm{LMO}(g)$. □

The proof goes along the same lines as the proof of Theorem 6.4, for completeness we provide the full proof.

Proof of Theorem 6.6. We prove the assertion in three parts.

Part 1: Recall that due to Lemma F.1, the following property holds for any $t \in [T]$,

$$\langle g^t, \hat{p}^t\rangle \leq \theta \cdot \min_{v\in\mathcal{C}}\langle g^t, v\rangle, \tag{33}$$

Part 2: We now prove that the iterates w^t are bounded. Using induction, we show that $\|w^t\| \leq \theta^{-1} \cdot \mathcal{C}$, $\forall t \geq 0$. For the base case, note that $w_0 \in \mathcal{C}$ and therefore $w_0 \in \mathcal{C} \subseteq \theta^{-1} \cdot \mathcal{C}$, this holds since $\theta \in (0, 1]$ and since $\mathbf{0} \in \mathcal{C}$ (see Def. 6.5). Now, for the induction step, assume that $w^t \in \theta^{-1} \cdot \mathcal{C}$, and let us show that this implies $w^{t+1} \in \theta^{-1} \cdot \mathcal{C}$. Indeed, by definition, w^{t+1} is a convex combination of two vectors w^t and $\hat{p}^t/\theta$, since these two vectors belong the set $\theta^{-1} \cdot \mathcal{C}$, so is their convex combination w^{t+1}. This establishes the induction step.

Part 3: Let x^* be an optimal solution of (P), i.e., $f(x^*) = f_{\mathrm{opt}} = \min_{x\in\mathcal{C}} f(x)$. From the update rule of Algorithm 4 we have $w^{t+1} - w^t = \eta^t(\hat{p}^t/\theta - w^t)$. Using the above together with the smoothness of f implies,

$$\begin{aligned} f(w^{t+1}) - f(w^t) \ &\leq \ \langle g^t, (w^{t+1} - w^t)\rangle + \frac{L}{2}\|w^{t+1} - w^t\|^2 \\ &= \ \eta^t\langle g^t, (\hat{p}^t/\theta - w^t)\rangle + \frac{L(\eta^t)^2}{2}\|\hat{p}^t/\theta - w^t\|^2 \\ &\leq \ \eta^t \min_{v\in\mathcal{C}}\langle g^t, (v - w^t)\rangle + \frac{L(\eta^t)^2}{2}\|\hat{p}^t/\theta - w^t\|^2 \\ &\leq \ \eta^t\langle g^t, (x^* - w^t)\rangle + \frac{L(\eta^t)^2}{2}\|\hat{p}^t/\theta - w^t\|^2 \\ &\leq \ \eta^t(f(x^*) - f(w^t)) + 2L(\eta^t)^2(R/\theta)^2, \end{aligned}$$

where the second inequality uses (33), and the third inequality uses the fact that $x^* \in \mathcal{C}$. The last line uses the gradient inequality, and the fact that $\hat{p}^t/\theta$ and w^t belong the set $\theta^{-1} \cdot \mathcal{C}$ (see Part 2 of this proof) together with the diameter bound R. Now, denoting $\delta^t = f(w^t) - f(x^*) = f(w^t) - f_{\mathrm{opt}}$ and rearranging the above equation yields,

$$\delta^{t+1} \leq (1 - \eta^t)\delta^t + 2L(R/\theta)^2(\eta^t)^2.$$

We will now use the Lemma 6.3 to bound δ^t. Recalling that $w^0 \in \mathbb{B}[0, R]$, the smoothness of f implies that,

$$\delta^1 := f(w_0) - f(x^*) \leq \frac{L\|x - x^*\|^2}{2} \leq LR^2/2 \leq 4L(R/\theta)^2,$$

where we used the fact that $\theta \in (0, 1]$. Thus, applying Lemma 6.3 and taking $r := 4L(R/\theta)^2$ implies that for any $t \geq 0$,

$$\delta^t := f(w^t) - f_{\mathrm{opt}} \leq \frac{8L(R/\theta)^2}{t+1}. \tag{34}$$

Finally using the fact that $\mathbf{0} \in \mathcal{C}$,

$$\begin{aligned} f(\bar{w}^T) - \theta f_{\mathrm{opt}} =& \; f(\theta w^T + (1-\theta)\cdot 0) - \theta f_{\mathrm{opt}} \\ \leq& \; \theta f(w^T) + (1-\theta) f(0) - \theta f_{\mathrm{opt}} \\ \leq& \; \theta \left(f(w^T) - f_{\mathrm{opt}} \right) \leq \frac{8L(R^2/\theta)}{T+1}, \end{aligned}$$

where the first inequality uses the Jensen inequality together with the fact that $\theta \in (0,1)$, the second inequality uses $f(0) \leq \max_{x \in \mathbb{B}[0,R]} f(x) \leq 0$, and the last inequality uses (34). □