Efficient Projection-Free Online Convex Optimization with Membership Oracle

Zakaria Mhammedi MHAMMEDI@MIT.EDU
Massachusetts Institute of Technology

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

In constrained convex optimization, existing interior point methods do not scale well with the dimension of the ambient space. Alternative approaches such as Projected Gradient Descent only provide a computational benefit for simple convex sets where Euclidean projections can be performed efficiently, such as Euclidean balls. For other more complex sets, the cost of the projections can be too high. To circumvent these issues, alternative methods based on the famous Frank-Wolfe algorithm have been studied and widely used. Such methods use a Linear Optimization Oracle at each iteration instead of Euclidean projections; the former can often be performed efficiently. Such methods have also been extended to the online and stochastic optimization settings. However, the Frank-Wolfe algorithm and its variants do not achieve the optimal performance, in terms of regret or rate, for general convex sets. What is more, the Linear Optimization Oracle they use can still be computationally expensive in some cases. In this paper, we move away from Frank-Wolfe style algorithms and present a new reduction that turns any algorithm A over a Euclidean ball (where projections are cheap) to an algorithm over a general convex constraint set $\mathcal{C}$ contained within the ball, without sacrificing the performance of the original algorithm A by much. Our reduction requires $O(T \ln T)$ calls to a Membership Oracle on $\mathcal{C}$ after T rounds, and no linear optimization on $\mathcal{C}$ is needed. Using this reduction, we recover optimal regret bounds [resp. rates], in terms of the number of iterations, in online [resp. stochastic] convex optimization. Our guarantees are also useful in the offline convex optimization setting when the dimension of the ambient space is large.

1. Introduction

In this paper, we are interested in designing efficient algorithms for constrained convex optimization in the online, offline, and stochastic settings. Popular algorithms for optimizing a convex objective f defined on a bounded convex set $\mathcal{C} \subset \mathbb{R}^d$ include, for example, the ellipsoid or cutting plane methods (Grötschel et al., 1993, 2012; Bubeck, 2015). Though such algorithms enjoy linear convergence rates, where the optimality gap decreases exponentially fast with the number of iterations, their per-iteration computational complexity depends super-linearly in the dimension d. In particular, if $\text{Cost}(\mathcal{M}_{\mathcal{C}})$ is the computational cost of testing if some point $\boldsymbol{x} \in \mathbb{R}^d$ belongs to $\mathcal{C}$, then state-of-the-art cutting plane-based methods have a computational complexity of order $O(d^3 + d^2\text{Cost}(\mathcal{M}_{\mathcal{C}}))\log(1/\epsilon)$ to find an ϵ-suboptimal point (Lee et al., 2015, 2018). Thus, when the dimension d is large, such methods can become impractical.

For some convex sets like the Euclidean ball, alternative methods, such as the projected Gradient Descent (GD) algorithm, dispense of the "expensive" dimension dependence in their computational complexity at the cost of a worse dependence in $1/\epsilon$ (e.g. $1/\epsilon^2$ instead of $\log(1/\epsilon)$) in their convergence rate. Despite this cost, such methods are still favorable in practice when the dimension d is large. We refer the interested reader to Bubeck (2015) for a comprehensive comparison between the

convergence rates of different optimization algorithms. In some settings, evaluating the objective function f is itself expensive. For example, this is the case in many machine learning applications where the objective involves a sum over a large number of data points. In this case, a popular approach is to use (projected) Stochastic Gradient Descent (SGD), where it suffices to evaluate a stochastic version of the objective f at each iteration, which can often be done efficiently. SGD enjoys similar computational benefits as GD. However, when the domain $\mathcal{C}$ of interest is not a Euclidean ball, the main bottleneck of such methods is often the projections that need be performed each time the iterates of the algorithm step outside of $\mathcal{C}$. This has fueled a long line of research around the design of, so-called, projection-free algorithms that swap expensive projections for cheaper linear optimizations over the domain $\mathcal{C}$. The reader is referred to the vast literature on projection-free optimization that include, for instance, (Frank et al., 1956; Hazan, 2008; Jaggi, 2013; Garber, 2016; Kerdreux, 2020; Bomze et al., 2021).

The first projection-free algorithm (a.k.a. the Frank-Wolfe algorithm) was introduced by Frank et al. (1956) who applied it to smooth optimization on polyhedral sets. The Frank-Wolfe algorithm swaps expensive projections on the convex set $\mathcal{C}$ for linear optimization on $\mathcal{C}$, which for many sets of interest can be performed efficiently (see Table 2). A similar algorithm was later introduced for the Online Convex Optimization (OCO) setting with linear losses (Kalai and Vempala, 2005). This is a setting where instead of optimizing a fixed function, the goal is to minimize a quantity called *regret*. Formally, at each round t in OCO, a learner (algorithm) outputs an iterate $\boldsymbol{x}_t$ in some convex set $\mathcal{C}$, then observes a convex loss function $\ell_t : \mathcal{C} \to \mathbb{R}$ that may be chosen by an adversary based on $\boldsymbol{x}_t$ and the history up to round t. The goal is to minimize the regret $\text{Regret}_T(\boldsymbol{x}) := \sum_{t=1}^T \ell_t(\boldsymbol{x}_t) - \sum_{t=1}^T \ell_t(\boldsymbol{x})$, which represents the difference between the cumulative loss of the learner and that of any comparator $\boldsymbol{x} \in \mathcal{C}$.

Offline [resp. Stochastic] optimization are special cases of OCO, where ℓ_t is fixed [resp. stochastic with a fixed mean] for all t. Algorithms designed for OCO (e.g. Online Gradient Descent (Zinkevich, 2003)) often lead to optimal convergence rates when used in the offline and stochastic settings via online-to-batch conversion techniques (Cesa-Bianchi et al., 2004; Shalev-Shwartz et al., 2011; Cutkosky, 2019). For examples of problems that can be modeled via OCO we refer the reader to the OCO introduction by Hazan (2016). Unlike in the offline setting where there are algorithms such as those based on the cutting plane method that converge exponentially fast with the number of iterations, there are no equivalent options for OCO since the losses are different at each round and the goal is to minimize the regret. This means that designing efficient projection-free algorithms for OCO is even more crucial than in the offline and stochastic settings.

In this paper, we continue the long line of research in the design of efficient projection-free algorithms for online and stochastic optimization (see e.g. (Kalai and Vempala, 2005; Hazan and Kale, 2012; Hazan and Luo, 2016; Hazan and Minasyan, 2020)). We depart from the Frank-Wolfe-style approach by presenting a new algorithm that requires a Membership Oracle $\mathcal{M}_{\mathcal{C}}$ for $\mathcal{C}$ instead of a Linear Optimization Oracle (LOO). When given $\boldsymbol{x} \in \mathbb{R}^d$, the Membership Oracle returns $\mathbb{I}\{\boldsymbol{x} \in \mathcal{C}\}$. We present an algorithm that requires a total of $O(T \ln T)$ calls to a (possibly approximate) Membership Oracle and guarantees a $O(\sqrt{T})$ regret bound for general convex functions and convex sets. This regret bound is optimal in the number of rounds T for general OCO. This result improves upon previous results designed for more specific use cases, and closes the longstanding regret gap between projection-free (e.g. Frank-Wolfe-style algorithm) and projection-based methods by achieving optimal regret. We now discuss our contributions in more detail.

Contributions. For OCO with general convex losses, we provide an algorithm in the form of a (projection-free) reduction that achieves a $O(\kappa\sqrt{T})$ regret after T rounds using a total of $\widetilde{O}(T)$ [resp. $\widetilde{O}(dT)$] calls to an approximate Separation [resp. Membership] Oracle, for some $\kappa > 0$ that depends on $\mathcal{C}$. The constant κ, which may be assumed less than d without loss of generality (see Rem. 1), takes the role of a $\sqrt{d}$ factor that is present in the regret bounds of previous projection-free algorithms. In Table 2, we display bounds on κ as function of d for many sets of interest.

We also present a more efficient version of our first algorithm that makes only $\widetilde{O}(1)$ calls per round to a Membership Oracle (instead of $\widetilde{O}(d)$ calls) at the cost of a multiplicative $\sqrt{d}$ factor in the regret bound. Thus, our approach allows a trade-off between computation (i.e. Oracle calls) and regret. Our reduction also allows us to build an adaptive projection-free algorithm for OCO with strongly convex losses that achieves a logarithm regret without requiring the parameter of strong convexity as input, while ensuring a $\widetilde{O}(\sqrt{T})$ regret in the worst case.

Our guarantees for general OCO readily transfer to the stochastic setting via well-known online-to-batch conversion techniques (Cesa-Bianchi et al., 2004; Shalev-Shwartz et al., 2011), leading to a convergence rate of order $O(\kappa/\sqrt{T})$ which is optimal in T. We also present an algorithm for β-smooth stochastic optimization that achieves the rate $O(\beta\kappa^2/T^2 + \sigma\kappa/\sqrt{T})$, where σ^2 is the variance of the noise of the observed subgradients. Crucially, our algorithm does not require knowledge of either β or σ, and is thus fully adaptive. When instantiated in the offline setting (i.e. when $\sigma = 0$), theses rates imply that our algorithms represent viable alternatives to state-of-the-art cutting plane methods (Lee et al., 2018) whenever $d \geq \Omega(\kappa^2/\epsilon^2)$ for general convex functions or when $d \geq \Omega(\kappa/\sqrt{\epsilon})$ for smooth functions. We summarize our contributions in Table 1, where we compare our average regret/rates to the best known methods that apply to general convex sets.

We achieve these results by reducing an OCO problem on a general convex set $\mathcal{C}$ to an OCO problem on a ball, where projections can be performed efficiently. More specifically, we present an algorithm wrapper that takes in any base algorithm A defined on a Euclidean ball (where projections can be performed cheaply) containing $\mathcal{C}$ and turns it into an algorithm on $\mathcal{C}$ that does not incur any expensive projections and whose regret bound is at most a factor $O(\kappa)$ worse than that of A.

Related Works. This paper continues a long a line of work in projection-free online learning. Table 1 summarizes the guarantees of state-of-the-art Frank-Wolfe-style algorithms that are applicable to general convex sets. For OCO with general convex sets/functions, the best known regret is of order $O(T^{3/4})$ and was achieved by Hazan and Kale (2012). The corresponding Frank-Wolfe algorithm requires a single call to a Linear Optimization Oracle per round. Later, Garber and Hazan (2016) introduced an algorithm that achieves the optimal $O(\sqrt{T})$ [resp. $O(\log T)$] regret in OCO with general [resp. strongly convex] losses with one call to an LOO per round. However, their results only apply when the convex set is a polytope, and they use an Oracle that requires at least $O(\min(dT, d^2))$ arithmetic operations per round (translating into $O(\min(n^4, Tn^2))$ operations when $\mathcal{C}$ is a subset of $n \times n$ matrices). Closer to our approach is that of Mahdavi et al. (2012) and Levy and Krause (2019) who essentially also considered Membership and Separation Oracles for their algorithms instead of a Linear Optimization Oracle on $\mathcal{C}$. In their setting, the set $\mathcal{C}$ is of the form $\{\boldsymbol{x} \in \mathbb{R}^d \mid h(\boldsymbol{x}) \leq 0\}$, where $h : \mathbb{R}^d \to \mathbb{R}$ is a smooth convex function, which translates into the set $\mathcal{C}$ being smooth. The approach of Mahdavi et al. (2012) requires at least one orthogonal projection onto $\mathcal{C}$. Levy and Krause (2019) avoid this by providing a fast approximate projection and achieve optimal regret bounds for both convex and strongly convex losses while only requiring a single call to a Separating and Membership Oracle for $\mathcal{C}$ per round. However, their guarantees only applies for smooth sets.

Setting	Loss Function	Average Regret/Rate Previous Best		**This paper**	
Online	Non-Smooth	$T^{-1/4}$	(Hazan and Kale, 2012)	$T^{-1/2}$	(Thms. 8,15, 16)
Online	Smooth	$T^{-1/3}$	(Hazan and Minasyan, 2020)	$T^{-1/2}$	(Thm. 15)
Online	Strongly Convex	$T^{-1/3}$	(Kretzu and Garber, 2021)	$T^{-1} \cdot \ln T$	(Thm. 17)
Stochastic	Non-Smooth	$T^{-1/3}$	(Hazan and Kale, 2012)	$T^{-1/2}$	(Thms. 8,15,16)
Stochastic	β-Smooth (σ^2-var)	$T^{-1/2}$	(Lan et al., 2017)	$\beta T^{-2} + \sigma/T^{-1/2}$	(Thm. 18)
Stochastic	Strongly Convex	$T^{-1/3}$	(Hazan and Kale, 2012)	$T^{-1} \cdot \ln T$	(Thm. 17)
Offline	Non-Smooth	$T^{-1/3}$	(Hazan and Kale, 2012)	$T^{-1/2}$	(Thms. 8,15,16)
Offline	β-Smooth	βT^{-2}	(Lan and Zhou, 2016)	βT^{-2}	(Thm. 18)

Table 1: Only results that hold for general bounded convex sets are included. The algorithms that achieve the reported results for our paper require either $\widetilde{O}(d)$ or $\widetilde{O}(1)$ calls per round to a Membership Oracle for $\mathcal{C}$. In the latter case, the regret bound is a $\sqrt{d}$ factor worse.

In a way, their algorithm complements that of Garber and Hazan (2016) that works for polytopes (which are non-smooth sets). Our approach can be viewed as a generalization of that of Mahdavi et al. (2012); Levy and Krause (2019) to general sets. Finally, we mention the work of Abernethy et al. (2012) who used a self-concordant barrier for the set of interest to avoid projections. The efficient version of their algorithm, which is based on the damped Newton step, has a $O(d^3)$ per-round computational complexity (due to matrix inversion).

For OCO with smooth [resp. strongly] convex losses (Hazan and Minasyan, 2020) [resp. (Kretzu and Garber, 2021)] provide algorithms that achieve a $O(T^{2/3})$ regret (up to multiplicative factors in the dimension), requiring a single call to an LOO per round. For smooth projection-free stochastic optimization, the best rate we are aware of is of order $O(\beta/T^2 + \sigma/\sqrt{T})$, where β is the smoothness parameter and σ is the stochastic noise. The first algorithm to achieve this makes $O(T)$ LOO calls after T rounds and is due to Lan et al. (2017). This improves on the previous best $T^{-1/4}$ rate due to Hazan and Kale (2012) that uses the same number of Oracle calls. We achieve a similar rate as that of Lan et al. (2017), albeit we use a Membership Oracle instead of an LOO. The story is similar for offline smooth convex optimization where we achieve the optimal accelerated rate of $O(\beta/T^2)$ for β-smooth function as the algorithm of Lan and Zhou (2016), albeit without requiring knowledge of β.

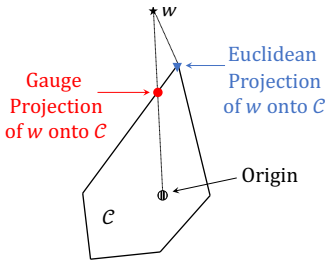

Figure 1: Gauge vs Euclidean projections.

Finally, the core idea behind our algorithm is inspired by recent techniques (Cutkosky and Orabona, 2018; Cutkosky, 2020) that leverage certain surrogate losses to reduce a constrained OCO problem to an unconstrained one. The key novelty in our work is a choice of surrogate losses that depend on the Gauge function (also known as the Minkowski function) of the set $\mathcal{C}$. In particular, we choose this dependence in a way that allows us to replace potentially expensive (Euclidean) projections by inexpensive Gauge projections (see Definition 5 and Figure 1), provided one has access to a Membership Oracle. We are able to perform efficient Gauge projections using recent techniques for estimating subgradients of convex functions that are not necessarily differentiable

(Lee et al., 2018). Our algorithm for strongly-convex OCO [resp. smooth stochastic optimization] builds on reductions due to Cutkosky and Orabona (2018) [resp. Cutkosky (2020)].

Outline. In §2, we describe our setting and provide some standard definitions from convex analysis. In §3, we describe tools we use to approximate the Gauge function of a set and its subgradients; these will be key to deriving our projection-free algorithms. In §4, we present our main method that reduces constrained OCO on an arbitrary compact convex set $\mathcal{C}$ to OCO on a Euclidean ball containing $\mathcal{C}$. There, we also discuss the implications of this result in the stochastic and offline optimization settings. We conclude with a discussion in §5. A table of contents is provided in the appendix.

Notation. For a real function $f\colon \mathcal{A} \to \mathbb{R}$, we denote by $\arg\max_{a\in\mathcal{A}} f(a)$ [resp. $\arg\min_{a\in\mathcal{A}} f(a)$] the subset $\mathcal{K}$ of $\mathcal{A}$ such that $f(b) = \sup_{a\in\mathcal{A}} f(a)$ [resp. $f(b) = \inf_{a\in\mathcal{A}} f(a)$], for all $b \in \mathcal{K}$. We denote by $\mathcal{B}(r)$ the Euclidean ball of radius r centered at the origin, and let $\|\cdot\|$ be the Euclidean norm. Finally, for any mathematical statement $\mathcal{E}$, we let $\mathbb{I}_{\mathcal{E}} = 1$ if $\mathcal{E}$ is true, and 0 otherwise. Additional notation needed in our proofs is included in Appendix A.

2. Setting and Definitions

We consider the standard OCO setting where at each *round* $t \geq 1$, a learner (the algorithm) outputs an iterate $\boldsymbol{x}_t$ in some convex set $\mathcal{C} \subset \mathbb{R}^d$, then the environment reveals a convex loss $\ell_t\colon \mathcal{C} \to \mathbb{R}$, and the learner suffers loss $\ell_t(\boldsymbol{x}_t)$. The goal of the learner is to minimize the *regret* $\mathrm{Regret}_T(\boldsymbol{x}) \coloneqq \sum_{t=1}^T \ell_t(\boldsymbol{x}_t) - \sum_{t=1}^T \ell_t(\boldsymbol{x})$ after $T \geq 1$ rounds against any fixed comparator $\boldsymbol{x} \in \mathcal{C}$.

We say that (ℓ_t) is an *adversarial loss sequence* when ℓ_t, $t \geq 1$, may depend on the learner's iterate $\boldsymbol{x}_t$ as well as the history $(\boldsymbol{x}_s, \ell_s)_{s\in[t-1]}$. This is the type of losses we will consider in our general OCO setting. We note that our bounds are often on the *linearized regret* $\sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle$, where, for all $t \geq 1$, $\boldsymbol{g}_t$ is in the subdifferential $\partial\ell_t(\boldsymbol{x}_t)$ of ℓ_t at $\boldsymbol{x}_t$ (see e.g. (Hiriart-Urruty and Lemaréchal, 2004) for a definition), and $\langle\cdot,\cdot\rangle$ denotes the standard inner product. Such bounds automatically transfer to bounds on the regret $\mathrm{Regret}_T(\boldsymbol{x})$ by convexity of the losses (ℓ_t); for any $\boldsymbol{x} \in \mathcal{C}$ and $t \geq 1$, we have $\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{x}) \leq \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle$, for all $\boldsymbol{g}_t \in \partial\ell_t(\boldsymbol{x}_t)$. Bounding the linearized regret is standard in OCO (see e.g. (Hazan, 2019)).

In this work, we will not assume knowledge of the horizon T, and so our regret bounds hold for all T's simultaneously; these are so-called *anytime* regret bound. We will allow the outputs $(\boldsymbol{x}_t)$ of the learner to be random. In this case, for any $t \geq 1$, we let $\mathcal{G}_t$ denote the σ-algebra generated by $(\boldsymbol{x}_s)_{s\in[t]}$, and we denote $\mathbb{E}_t[\cdot] \coloneqq \mathbb{E}[\cdot \mid \mathcal{G}_t]$. When, for any round t, the output $\boldsymbol{x}_t$ of an algorithm A is a deterministic function of the history $(\boldsymbol{x}_s, \ell_s)_{s\in[t-1]}$, we say that A is *deterministic*. Throughout, we assume that $\mathcal{C}$ is "sandwiched" between two Euclidean balls:

Assumption 1 *We assume that $\mathcal{C} \subseteq \mathbb{R}^d$ is a closed convex set s.t. for some $0 < r \leq R$, we have*

$$\mathcal{B}(r) \subseteq \mathcal{C} \subseteq \mathcal{B}(R). \tag{1}$$

Remark 1 Assumption 1 comes with no loss of generality as one can easily re-parametrize any OCO setting to satisfy (1) without any significant computational overhead. In Appendix F, we discuss the details of how this can be done in the general OCO setting as well as in various popular optimization settings. In Appendix F, we also derive corresponding upper bounds on the *asphericity* $\kappa \coloneqq R/r$ (Goffin, 1988), which appears as a multiplicative factor in our regret bounds. As can be

seen in Table 2, κ is less than $\sqrt{d}$ for many sets of interest. We also note that, in the worst case, κ can always be bounded by d [resp. $\sqrt{d}$] for general [resp. centrally symmetric] convex sets by applying an affine re-parametrization (Flaxman et al., 2005).

Our goal in this paper is to design algorithms with low regret in OCO without incurring expensive (Euclidean) projections onto the set $\mathcal{C}$. For this, we will use a Membership Oracle for the set $\mathcal{C}$ to perform what we call Gauge Projections, which can be done efficiently using a small number of calls to the Oracle. A Membership Oracle is a map $\mathcal{M}_{\mathcal{C}}\colon \mathbb{R}^d \to \{0,1\}$, where for any $\boldsymbol{x} \in \mathbb{R}^d$, $\mathcal{M}_{\mathcal{C}}(\boldsymbol{x}) = 1$ if and only if $\boldsymbol{x} \in \mathcal{C}$. Such Oracles are often the building blocks of many offline optimization algorithms (Grötschel et al., 1993; Bubeck, 2015; Lee et al., 2018). In many practical settings, one can only efficiently test membership approximately, and so for the sake of generality we will state our results when only such approximate Oracles are available. To formally define the notion of approximate Membership Oracle, we let $\mathcal{B}(\mathcal{C},\delta) \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d : \exists \boldsymbol{y} \in \mathcal{C} \text{ such that } \|\boldsymbol{x}-\boldsymbol{y}\| \leq \delta\}$. We also define $\mathcal{B}(\mathcal{C},-\delta) \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d : \forall \boldsymbol{y} \in \mathcal{B}(\delta),\ \boldsymbol{x}+\boldsymbol{y} \in \mathcal{C}\}$.

Definition 2 (Approximate Membership Oracle) *Let $\delta > 0$ and $\mathcal{C}$ be a convex set s.t. $\mathcal{B}(\delta) \subseteq \mathcal{C} \subset \mathbb{R}^d$. Then, the map $\mathsf{MEM}_{\mathcal{C}}(\cdot;\delta)\colon \mathbb{R}^d \to \{0,1\}$ is a δ-approximate Membership Oracle if*

$$\forall \boldsymbol{x} \in \mathbb{R}^d, \quad \mathbb{I}_{\{\boldsymbol{x} \in \mathcal{B}(\mathcal{C},-\delta)\}} \wedge \mathbb{I}_{\{\boldsymbol{x} \in \mathcal{B}(\mathcal{C},\delta)\}} \leq \mathsf{MEM}_{\mathcal{C}}(\boldsymbol{x};\delta) \leq \mathbb{I}_{\{\boldsymbol{x} \in \mathcal{B}(\mathcal{C},-\delta)\}} \vee \mathbb{I}_{\{\boldsymbol{x} \in \mathcal{B}(\mathcal{C},\delta)\}}. \tag{2}$$

Some definitions of (approximate) Membership Oracles only require (2) to hold with high probability (see e.g. Lee et al. (2018)). Our analysis can easily be extended to this case, but we make the deterministic assumption in (2) to simplify the presentation. We will use an approximate Membership Oracle to build a δ-approximate Linear Optimization Oracle $\mathcal{O}_{\mathcal{C}^\circ}(\cdot;\delta)$ for the polar set $\mathcal{C}^\circ$, where for any $\boldsymbol{w}$ in a ball containing $\mathcal{C}$, the random output $(\gamma, \boldsymbol{s})$ of $\mathcal{O}_{\mathcal{C}^\circ}(\boldsymbol{w};\delta)$ essentially satisfies $\sup_{\boldsymbol{u} \in \mathcal{C}} \langle \boldsymbol{w}, \boldsymbol{u}\rangle \leq \gamma \leq \sup_{\boldsymbol{u} \in \mathcal{C}} \langle \boldsymbol{w}, \boldsymbol{u}\rangle + \delta$ and $\langle \boldsymbol{s}, \boldsymbol{w}\rangle \geq \sup_{\boldsymbol{u} \in \mathcal{C}} \langle \boldsymbol{w}, \boldsymbol{u}\rangle - \delta$ with high probability[1].

We now present some standard definitions from convex analysis:

Definition 3 (Polar Set) *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a convex set. The* polar set *$\mathcal{C}^\circ$ of $\mathcal{C}$ is defined as $\mathcal{C}^\circ \coloneqq \{\boldsymbol{s} \in \mathbb{R}^d : \langle \boldsymbol{s}, \boldsymbol{u}\rangle \leq 1, \forall \boldsymbol{u} \in \mathcal{C}\}$.*

The notation of a polar set is key to our approach. We will essentially build an efficient Linear Optimization Oracle on $\mathcal{C}^\circ$, which will then allow us to perform, what we call, Gauge projections instead of Euclidean ones. Gauge projections (Def. 5 below) are defined with respect to a quasi-metric based on the *Gauge function* (also known as the Minkowski function) of the set $\mathcal{C}$:

Definition 4 (Gauge and Support Functions) *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a convex set. The* Gauge function *$\gamma_{\mathcal{C}}$ of $\mathcal{C}$ is defined as $\gamma_{\mathcal{C}}(\boldsymbol{w}) \coloneqq \inf\{\lambda > 0 : \boldsymbol{w} \in \lambda\mathcal{C}\}$, for $\boldsymbol{w} \in \mathbb{R}^d$. The* support function *$\sigma_{\mathcal{C}}$ of $\mathcal{C}$ is defined as $\sigma_{\mathcal{C}}(\boldsymbol{w}) \coloneqq \sup_{\boldsymbol{u} \in \mathcal{C}} \langle \boldsymbol{u}, \boldsymbol{w}\rangle$, for $\boldsymbol{w} \in \mathbb{R}^d$.*

The Gauge function satisfies all properties of a norm except for symmetry, which requires the set $\mathcal{C}$ to be centrally symmetric. Fortunately, we do not need this property to perform efficient projections with respect to the quasi-metric $\mathrm{d}_{\mathcal{C}}(\boldsymbol{u},\boldsymbol{w}) \coloneqq \gamma_{\mathcal{C}}(\boldsymbol{u}-\boldsymbol{w})$. The concept of support function in the previous definition will also be key in our analysis (it is the Gauge function of the polar set $\mathcal{C}^\circ$). We now introduce the concept of Gauge projection/distance to a set:

1. The guarantee provided by γ (i.e. $\langle \boldsymbol{w}, \boldsymbol{u}\rangle \leq \gamma \leq \sup_{\boldsymbol{u} \in \mathcal{C}} \langle \boldsymbol{w}, \boldsymbol{u}\rangle + \delta$) will be crucial in our analysis, which is why we make our LO Oracle $\mathcal{O}_{\mathcal{C}^\circ}$ return a scalar-vector pair instead of a single vector as is more common for such Oracles.

Definition 5 *Let $\mathcal{C}$ be as in Assump. 1, the Gauge projection on $\mathcal{C}$ is the set-valued map $\Pi_{\mathcal{C}}^{\mathrm{G}} : \mathbb{R}^d \rightrightarrows \mathcal{C}$ defined by $\Pi_{\mathcal{C}}^{\mathrm{G}}(\boldsymbol{w}) \coloneqq \arg\min_{\boldsymbol{u}\in\mathcal{C}} \gamma_{\mathcal{C}}(\boldsymbol{w}-\boldsymbol{u})$. We also define the Gauge distance to $\mathcal{C}$ as the map:*

$$S_{\mathcal{C}}(\boldsymbol{w}) = \inf_{\boldsymbol{u}\in\mathcal{C}} \gamma_{\mathcal{C}}(\boldsymbol{w}-\boldsymbol{u}) \geq 0, \quad \textit{for all } \boldsymbol{w} \in \mathbb{R}^d. \tag{3}$$

Gauge projections, which can be done efficiently using a Membership Oracle (see §3), will be all we need to ensure the iterates of our algorithms are in $\mathcal{C}$.

3. Preliminaries: Gauge Function Approximation

In this section, we briefly describe some of the technical tools we require for our method (additional details are provided in Appendix B). At the heart of our approach is the design of surrogate losses for online and stochastic optimization that depend on the Gauge distance function $S_{\mathcal{C}}$ in (3) in a way that leads to algorithms that perform Gauge projections instead of (potentially expensive) Euclidean projections. Naturally, this will require the ability to efficiently evaluate $S_{\mathcal{C}}$ and its subgradients. The next lemma (whose proof is in App. B) shows that $S_{\mathcal{C}}$ and $\partial S_{\mathcal{C}}$ can be expressed as concise functions of $\gamma_{\mathcal{C}}$ (the Gauge function):

Lemma 6 *For any set $\mathcal{C}$ satisfying Assump. 1, and any $\boldsymbol{u} \notin \mathcal{C}$ and $\boldsymbol{w} \in \mathbb{R}^d$, we have $\frac{\boldsymbol{u}}{\gamma_{\mathcal{C}}(\boldsymbol{u})} \in \Pi_{\mathcal{C}}^{\mathrm{G}}(\boldsymbol{u})$,*

$$S_{\mathcal{C}}(\boldsymbol{w}) = 0 \vee (\gamma_{\mathcal{C}}(\boldsymbol{w}) - 1), \quad \textit{and} \quad \partial S_{\mathcal{C}}(\boldsymbol{w}) = \begin{cases} \partial\gamma_{\mathcal{C}}(\boldsymbol{w}) = \arg\max_{\boldsymbol{s}\in\mathcal{C}^\circ} \langle \boldsymbol{s}, \boldsymbol{w}\rangle, & \textit{if } \boldsymbol{w} \notin \mathcal{C}; \\ \{\boldsymbol{0}\}, & \textit{otherwise}. \end{cases}$$

Note that for any $\boldsymbol{u} \notin \mathcal{C}$, Lemma 6 also provides the expression of a Gauge projection point in $\Pi_{\mathcal{C}}^{\mathrm{G}}(\boldsymbol{u})$ (i.e. $\boldsymbol{u}/\gamma_{\mathcal{C}}(\boldsymbol{u})$), which will be leveraged by our algorithm. Next, we describe how we build efficient algorithms for approximating $\gamma_{\mathcal{C}}$ and its subgradients using our approximate Membership Oracle $\mathsf{MEM}_{\mathcal{C}}$. As Lemma 6 shows, this is all we need to approximate $S_{\mathcal{C}}$, its subgradients, and Gauge projections onto $\mathcal{C}$.

Approximating $\gamma_{\mathcal{C}}$. By definition of the Gauge function, we have,

$$\forall \boldsymbol{w} \in \mathbb{R}^d, \quad \gamma_{\mathcal{C}}(\boldsymbol{w}) = \inf\{\lambda \in \mathbb{R}_{\geq 0} \mid \boldsymbol{w} \in \lambda\mathcal{C}\} = 1/\sup\{\nu \in \mathbb{R}_{\geq 0} \mid \nu\boldsymbol{w} \in \mathcal{C}\}. \tag{4}$$

Using the Membership Oracle $\mathsf{MEM}_{\mathcal{C}}$, we can approximate the largest $\nu \geq 0$ such that $\nu\boldsymbol{w} \in \mathcal{C}$ via bisection, which will lead to an approximation of $\gamma_{\mathcal{C}}(\boldsymbol{w})$ by (4). This is exactly what we do in Algorithm 2 in Appendix B. Algorithm 2 ($\mathsf{GAU}_{\mathcal{C}}$) requires at most $\lceil \log_2((4\kappa)^2/\delta)\rceil + 1$ calls to the Membership Oracle $\mathsf{MEM}_{\mathcal{C}}(\cdot; r\delta/(4\kappa)^2)$ to return a δ-approximate value of $\gamma_{\mathcal{C}}$, where $\kappa \coloneqq R/r$ and $r, R > 0$ are as in (1). In Lemma 10 of Appendix B, we state the full guarantee of Algorithm 2.

Approximating the subgradients of $\gamma_{\mathcal{C}}$. In addition to approximating $\gamma_{\mathcal{C}}$, we will also need to approximate its subgradients. By Lemma 6, the subdifferential of $S_{\mathcal{C}}$ coincides with that of $\gamma_{\mathcal{C}}$ at any point $\boldsymbol{w}$ outside $\mathcal{C}$. The lemma also shows that finding a subgradient of $\gamma_{\mathcal{C}}$ is equivalent to performing linear optimization on the polar set $\mathcal{C}^\circ$. In Appendix B, we present two algorithms, $\mathsf{OPT}_{\mathcal{C}^\circ}$ (Alg. 3) and $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$ (Alg. 4), based on (Lee et al., 2018, Algorithm 2), which use $\mathsf{GAU}_{\mathcal{C}}$ (Alg. 2) and a random partial difference along different coordinates to approximate the subgradients of $\gamma_{\mathcal{C}}$. The first algorithm $\mathsf{OPT}_{\mathcal{C}^\circ}$ requires at most $O(d\ln(d\kappa/\delta))$ calls to the approximate Membership Oracle $\mathsf{MEM}_{\mathcal{C}}$ to find a "δ-approximate" subgradient; the precise guarantee is stated in Proposition 11.

The algorithm $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$ is a stochastic version of $\mathsf{OPT}_{\mathcal{C}^\circ}$ that picks a single random coordinate I along which to estimate the subgradient of the Gauge function $\gamma_{\mathcal{C}}$. As a result, $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$ requires at most $O(\ln(d\kappa/\delta))$ calls to the approximate Membership Oracle $\mathsf{MEM}_{\mathcal{C}}$, which will provide a computational benefit over $\mathsf{OPT}_{\mathcal{C}^\circ}$ at the cost of a multiplicative $\sqrt{d}$ in the regret bound. Further, the output of $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$ is equal to that of $\mathsf{OPT}_{\mathcal{C}^\circ}$ in expectation for the same input (see Lemma 13).

We now move to the main section of the paper where we present our projection-free reduction that uses the tools we described in this section.

4. Efficient Projection-Free Online and Stochastic Convex Optimization

In this section, we will use our approximate Linear Optimization Oracles on $\mathcal{C}^\circ$ described in the previous section to build efficient algorithms for OCO with optimal regret bounds. Our final algorithms make at most $\widetilde{O}(T)$ calls to the Membership Oracle $\mathsf{MEM}_{\mathcal{C}}$ after $T \geq 1$ rounds, and are also time-uniform in the sense that they do not require a horizon T as input.

Algorithm 1 Projection-Free Algorithm-Wrapper for OCO on $\mathcal{C}$.

Require: **I)** An OCO algorithm A on $\mathcal{B}(R) \supseteq \mathcal{C}$; **II)** A tolerance sequence $(\delta_t) \subset (0, 1/3)$; and **III)** LO Oracle $\mathcal{O}_{\mathcal{C}^\circ}$ on $\mathcal{C}^\circ$ such that $\mathcal{O}_{\mathcal{C}^\circ}(\boldsymbol{w}; \delta) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^d$, $\forall \boldsymbol{w} \in \mathcal{B}(R), \delta \in (0, 1)$.
// In the ideal case, $(\gamma, \boldsymbol{s}) = \mathcal{O}_{\mathcal{C}^\circ}(\boldsymbol{w}; \delta) \implies \gamma = \langle \boldsymbol{s}, \boldsymbol{w} \rangle$ and $\boldsymbol{s} \in \arg\max_{\boldsymbol{u} \in \mathcal{C}^\circ} \langle \boldsymbol{u}, \boldsymbol{w} \rangle$

Initialize A and set $\boldsymbol{w}_1 \in \mathcal{B}(R)$ to A's first output.
for $t = 1, 2, \ldots$ **do**
  Set $(\gamma_t, \boldsymbol{s}_t) = \mathcal{O}_{\mathcal{C}^\circ}(\boldsymbol{w}_t; \delta_t)$
  Set $\boldsymbol{v}_t = \mathbb{I}_{\{\gamma_t \geq 1\}} \boldsymbol{s}_t$. // Subgradient of $S_{\mathcal{C}}$ at $\boldsymbol{w}_t$
  Play $\boldsymbol{x}_t = \mathbb{I}_{\{\gamma_t \geq 1\}} \boldsymbol{w}_t / \gamma_t + \mathbb{I}_{\{\gamma_t < 1\}} \boldsymbol{w}_t$. // Gauge projection of $\boldsymbol{w}_t$ onto $\mathcal{C}$
  Observe subgradient $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t)$.
  Set $\tilde{\boldsymbol{g}}_t = \boldsymbol{g}_t - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \boldsymbol{v}_t$ // $\tilde{\boldsymbol{g}}_t \in \partial \tilde{\ell}_t(\boldsymbol{w}_t)$; $\tilde{\ell}_t(\boldsymbol{w}) \coloneqq \langle \boldsymbol{g}_t, \boldsymbol{w} \rangle - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle S_{\mathcal{C}}(\boldsymbol{w})$
  Set A's tth loss function to $f_t : \boldsymbol{w} \mapsto \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w} \rangle$.
  Set $\boldsymbol{w}_{t+1} \in \mathcal{B}(R)$ to A's $(t+1)$th output given the history $((\boldsymbol{w}_i, \boldsymbol{x}_i, f_i)_{i \leq t})$.
end for

To avoid expensive projections, our algorithms make use of convex surrogate losses of the form $\tilde{\ell}(\boldsymbol{w}) \coloneqq \langle \boldsymbol{g}, \boldsymbol{w} \rangle + b S_{\mathcal{C}}(\boldsymbol{w})$, for $\boldsymbol{w} \in \mathbb{R}^d$ and $b \geq 0$, where $S_{\mathcal{C}}$ is the Gauge distance function in (3) and $\boldsymbol{g}$ is a subgradient of the loss of interest $\ell \colon \mathcal{C} \to \mathbb{R}$. Note that $\tilde{\ell}$ is not constrained to the set $\mathcal{C}$, unlike the actual loss of interest ℓ. The choice of such a surrogate loss function is inspired by existing constrained-to-unconstrained reductions in OCO due to Cutkosky and Orabona (2018); Cutkosky (2020). Similar to the latter, our projection-free reduction allows us to bound the regret of our Algorithm 1 by the regret of any OCO subroutine A that is fed subgradients of the surrogate losses. Here, the iterates of A need not be constrained to the set $\mathcal{C}$ since the domain of the surrogate losses is technically unconstrained. Thus, to build efficient projection-free OCO algorithms on $\mathcal{C}$ with optimal regret, it suffices to pick a sub-algorithm A that I) has an optimal regret bound on a set $\mathcal{W}$ containing $\mathcal{C}$, and II) does not incur any expensive projections. We will pick $\mathcal{W} = \mathcal{B}(R) \supseteq \mathcal{C}$ so that Euclidean projections onto $\mathcal{W}$, which may be required by A, can be performed efficiently.

We now state our main reduction result when the LO Oracle $\mathcal{O}_{\mathcal{C}^\circ}$ is set to the approximate LO Oracle $\mathsf{OPT}_{\mathcal{C}^\circ}$ (Alg. 3) we described in §3. The full proof of the lemma is in App. I.1.

Lemma 7 *Let $\kappa := R/r$ and A be any OCO algorithm on $\mathcal{B}(R)$. Further, for $t \geq 1$, let $\boldsymbol{w}_t, \tilde{\boldsymbol{g}}_t, \boldsymbol{x}_t$, and $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t)$ be as in Alg. 1 with $\mathcal{O}_{\mathcal{C}^\circ} \equiv \mathsf{OPT}_{\mathcal{C}^\circ}$ (Alg. 3) and any tolerance sequence $(\delta_s) \subset (0, 1/3)$. Then, for all $t \geq 1$, we have $\boldsymbol{x}_t \in \mathcal{C}$, and there exists a random variable $\Delta_t \in \left[0, 15^2 d^4 \kappa^3 \delta_t^{-2}\right]$ such that $\mathbb{E}_{t-1}[\Delta_t] \leq \delta_t$, $\|\tilde{\boldsymbol{g}}_t\| \leq (1 + \Delta_t + \kappa)\|\boldsymbol{g}_t\|$, and for all $\boldsymbol{x} \in \mathcal{C}$,*

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \leq \tilde{\ell}_t(\boldsymbol{w}_t) - \tilde{\ell}_t(\boldsymbol{x}) + \delta_t R\|\boldsymbol{g}_t\| \leq \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x} \rangle + (2\delta_t + \Delta_t)R\|\boldsymbol{g}_t\|, \tag{5}$$

where $\tilde{\ell}_t(\boldsymbol{w}) := \langle \boldsymbol{g}_t, \boldsymbol{w} \rangle - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle S_{\mathcal{C}}(\boldsymbol{w})$.

Proof Sketch. For simplicity, we provide a proof of the lemma when Alg. 1 has access to a perfect LO Oracle $\mathcal{O}_{\mathcal{C}^\circ}$ on $\mathcal{C}^\circ$, i.e. when $\forall \boldsymbol{w} \in \mathbb{R}^d, \delta > 0$, $(\gamma, \boldsymbol{s}) = \mathcal{O}_{\mathcal{C}^\circ}(\boldsymbol{w}; \delta)$ only if $\boldsymbol{s} \in \arg\max_{\boldsymbol{x} \in \mathcal{C}^\circ} \langle \boldsymbol{x}, \boldsymbol{w} \rangle$ and $\gamma = \langle \boldsymbol{s}, \boldsymbol{w} \rangle$. Thus, we will show the claims of the lemma with $\Delta_t = \delta_t = 0$.

Let γ_t, $\boldsymbol{\nu}_t$, $\boldsymbol{w}_t$, $\tilde{\boldsymbol{g}}_t$, and $\boldsymbol{x}_t$ be as in Algorithm 1. By the expression of the subdifferential of $S_{\mathcal{C}}$ in Lemma 6 and the definition of $\tilde{\ell}_t$, we have that $\boldsymbol{g}_t - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \boldsymbol{\nu}_t \in \partial \tilde{\ell}_t(\boldsymbol{w}_t)$. Thus, since $-\mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle > 0$ (see definition of $\boldsymbol{x}_t$ in Alg. 1), $\tilde{\ell}_t$ is convex and so

$$\forall \boldsymbol{x} \in \mathcal{C}, \quad \tilde{\ell}_t(\boldsymbol{w}_t) - \tilde{\ell}_t(\boldsymbol{x}) \leq \langle \boldsymbol{g}_t - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \boldsymbol{\nu}_t, \boldsymbol{w}_t - \boldsymbol{x} \rangle = \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x} \rangle.$$

It remains to show that $\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \leq \tilde{\ell}_t(\boldsymbol{w}_t) - \tilde{\ell}_t(\boldsymbol{x})$, for all $t \geq 1$ and $\boldsymbol{x} \in \mathcal{C}$. First note that for all $\boldsymbol{x} \in \mathcal{C}$, we have $S_{\mathcal{C}}(\boldsymbol{x}) = 0$, and so

$$\tilde{\ell}_t(\boldsymbol{x}) = \langle \boldsymbol{g}_t, \boldsymbol{x} \rangle, \quad \forall \boldsymbol{x} \in \mathcal{C}. \tag{6}$$

We will now compare $\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle$ to $\tilde{\ell}_t(\boldsymbol{w}_t)$ by considering cases. Suppose that $\gamma_t < 1$. In this case, we have $\boldsymbol{x}_t = \boldsymbol{w}_t$ and so $\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle = \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle = \tilde{\ell}_t(\boldsymbol{w}_t)$. Now suppose that $\gamma_t \geq 1$ and $\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle \geq 0$. By the fact that $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) = \gamma_t \geq 1$ and $\boldsymbol{x}_t = \boldsymbol{w}_t / \gamma_t$, we have

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \leq \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle = \tilde{\ell}_t(\boldsymbol{w}_t). \tag{7}$$

Now suppose that $\gamma_t \geq 1$ and $\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0$. Using the fact that $\boldsymbol{x}_t = \boldsymbol{w}_t / \gamma_t = \boldsymbol{w}_t / \gamma_{\mathcal{C}}(\boldsymbol{w}_t)$ and that $S_{\mathcal{C}}(\boldsymbol{w}_t) = \gamma_{\mathcal{C}}(\boldsymbol{w}_t) - 1$ (Lemma 6), we have

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle + \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \cdot S_{\mathcal{C}}(\boldsymbol{w}_t) = \langle \boldsymbol{g}_t, \boldsymbol{w}_t / \gamma_{\mathcal{C}}(\boldsymbol{w}_t) \rangle + \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle - \langle \boldsymbol{g}_t, \boldsymbol{w}_t / \gamma_{\mathcal{C}}(\boldsymbol{w}_t) \rangle = \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle.$$

Rearranging this, we get

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle = \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle - \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle S_{\mathcal{C}}(\boldsymbol{w}_t) = \tilde{\ell}_t(\boldsymbol{w}_t). \tag{8}$$

By combining (6), (7), and (8), we obtain $\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \leq \tilde{\ell}_t(\boldsymbol{w}_t) - \tilde{\ell}_t(\boldsymbol{x}) \leq \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x} \rangle$, for all $\boldsymbol{x} \in \mathcal{C}$, which shows (5) with $\Delta_t = \delta_t = 0$. It remains to bound $\|\tilde{\boldsymbol{g}}_t\|$ in terms of $\|\boldsymbol{g}_t\|$ and show that $\boldsymbol{x}_t \in \mathcal{C}$. When $\gamma_t < 1$, we have $\tilde{\boldsymbol{g}}_t = \boldsymbol{g}_t$ and $\boldsymbol{x}_t = \boldsymbol{w}_t$, and so $\|\tilde{\boldsymbol{g}}_t\| = \|\boldsymbol{g}_t\|$. Furthermore, since $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) = \gamma_t < 1$, the definition of the Gauge function implies that $\boldsymbol{w}_t \in \mathcal{C}$ and so $\boldsymbol{x}_t \in \mathcal{C}$ (since $\boldsymbol{x}_t = \boldsymbol{w}_t$). Now suppose that $\gamma_t \geq 1$. In this case, we have $\boldsymbol{x}_t = \boldsymbol{w}_t / \gamma_t = \boldsymbol{w}_t / \gamma_{\mathcal{C}}(\boldsymbol{w}_t)$ and $\tilde{\boldsymbol{g}}_t = \boldsymbol{g}_t - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \boldsymbol{\nu}_t$. Using this and that $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) = \gamma_t \geq 1$, we get that $\|\tilde{\boldsymbol{g}}_t\|$ is equal to

$$\|\boldsymbol{g}_t - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \boldsymbol{\nu}_t\| = \|\boldsymbol{g}_t\| + \|\boldsymbol{g}_t\| \frac{\|\boldsymbol{w}_t\|}{\gamma_{\mathcal{C}}(\boldsymbol{w}_t)} \|\boldsymbol{\nu}_t\| \overset{(*)}{\leq} \|\boldsymbol{g}_t\| \left(1 + \frac{\|\boldsymbol{w}_t\|}{r}\right) \leq (1 + \kappa)\|\boldsymbol{g}_t\|.$$

where $(*)$ follows by the fact that $\|\boldsymbol{\nu}_t\| \leq 1/r$ (since $\boldsymbol{\nu}_t \in \mathcal{C}^\circ \subseteq \mathcal{B}(1/r)$–see Lem. 42 for the set inclusion). Further, when $\gamma_t \geq 1$, we have $\gamma_{\mathcal{C}}(\boldsymbol{x}_t) = \gamma_{\mathcal{C}}(\boldsymbol{w}_t/\gamma_{\mathcal{C}}(\boldsymbol{w}_t)) = \gamma_{\mathcal{C}}(\boldsymbol{w}_t)/\gamma_{\mathcal{C}}(\boldsymbol{w}_t) = 1$, and so $\boldsymbol{x}_t \in \mathcal{C}$. ■

Lemma 7 allows us to bound the instantaneous (linearized) regret $\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle$ of Alg. 1 w.r.t. the instantaneous regret $\langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle$ of subroutine A. Thus by summing (5) in Lemma 7 for $t = 1, 2, \ldots,$ we can bound the regret of Alg. 1 with respect to that of A. Further, Lemma 7 shows that the scale of the subgradients $(\tilde{\boldsymbol{g}}_t)$ of the losses fed to A is at most a constant times the scale of the subgradients $(\boldsymbol{g}_t)$ at the iterates $(\boldsymbol{x}_t)$ of Alg. 1; in particular, $\|\tilde{\boldsymbol{g}}_t\| \leq (1 + \Delta_t + \kappa)\|\boldsymbol{g}_t\|, \forall t \geq 1$, where $\kappa := R/r$. This latter fact ensures that the regret bound of A is not too large as a function of the scale of $(\boldsymbol{g}_t)$.

4.1. Algorithm for General Online Convex Optimization

To streamline the discussions that follow, we will instantiate Alg. 1 and state its regret bound with a specific subroutine A (for a general A, see App. D). In particular, we will set A to Follow-The-Regularized-Leader Proximal (FTRL-prox) (McMahan, 2017). FTRL-prox is summarized in Alg. 8 in App. G. The algorithm takes input parameter $R > 0$ and performs at most one Euclidean projection onto the ball $\mathcal{B}(R)$ per iteration. An advantage of FTRL-prox is that it has a regret bound (see Prop. 25) of the form $O(\sqrt{V_T})$, where $V_T := \sum_{t=1}^T \|\boldsymbol{g}_t\|^2$ and $(\boldsymbol{g}_t)$ are the observed subgradients. This is desirable for reasons we outline in §4.2. We now state the regret bound of Alg. 1 when A is set to FTRL-prox:

Theorem 8 *Let $\delta \in (0, 1/3)$ and $\kappa := R/r$, with r and R as in (1). Let (ℓ_t) be any adversarial sequence of convex losses on $\mathcal{C}$ and $(\boldsymbol{x}_t)$ be the iterates of Alg. 1 in response to (ℓ_t). If Alg. 1 is run with subroutine A set to FTRL-prox with parameter R; $\mathcal{O}_{\mathcal{C}^\circ} \equiv \mathsf{OPT}_{\mathcal{C}^\circ}$ (Alg. 3); and $\delta_t = \delta/t^2$, $t \geq 1$, then for all $T \geq 1$ and $\rho \in (0, 1)$, we have $(\boldsymbol{x}_t)_{t\in[T]} \subset \mathcal{C}$ and with probability at least $1 - \rho$:*

$$\forall \boldsymbol{x} \in \mathcal{C}, \quad \sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \leq 4(1+\kappa)R\sqrt{V_T} + R \cdot (12 + 10\delta/\rho) \cdot \max_{t\leq T} \|\boldsymbol{g}_t\|, \tag{9}$$

where $V_T := \sum_{t=1}^T \|\boldsymbol{g}_t\|^2$ and $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t)$, for all $t \geq 1$.

Proof Sketch. By Lemma 7, we have $\boldsymbol{x}_t \in \mathcal{C}, \forall t \geq 1$. Lemma 7 also says that there exists a sequence $(\Delta_t) \subset \mathbb{R}_{\geq 0}$ of non-negative random variables satisfying $\mathbb{E}_{t-1}[\Delta_t] \leq \delta_t$, for all $t \in [T]$, such that for all $\boldsymbol{x}$ and $t \geq 1$, $\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \leq \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle + (2\delta_t + \Delta_t)R\|\boldsymbol{g}_t\|$. Letting $Q_t := \sum_{t=1}^T \|\tilde{\boldsymbol{g}}_t\|^2$, and summing the latter inequality for $t = 1, \ldots, T$, we obtain,

$$\sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle - \sum_{t=1}^T (2\delta_t + \Delta_t)R\|\boldsymbol{g}_t\| \leq \sum_{t=1}^T \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle \leq 2R\sqrt{2Q_T}, \quad \text{for all } \boldsymbol{x} \in \mathcal{C}, \tag{10}$$

where the last inequality follows by the fact that A is FTRL-prox and the regret bound of the latter in Prop. 25. Now, by the fact that $\|\tilde{\boldsymbol{g}}_t\| \leq (1 + \kappa + \Delta_t)\|\boldsymbol{g}_t\|$ (Lem. 7) and $(a+b)^2 \leq 2a^2 + 2b^2$ for all $a, b \in \mathbb{R}_{>0}$, the RHS of (10) can be further bounded by $4(1+\kappa)R\sqrt{V_T} + 4\sum_{t=1}^T \Delta_t R\|\boldsymbol{g}_t\|$. It remains to bound the sum $\sum_{t=1}^T (2\delta_t + 5\Delta_t)R\|\boldsymbol{g}_t\|$ in terms of $\delta R \max_{t\leq T}\|\boldsymbol{g}_t\|/\rho$, which we do using Doob's martingale inequality (see App. I.2 for the full proof). ■

The $O(\sqrt{V_T})$ regret bound in Theorem 8 is known as an *adaptive* regret bound, and can be much smaller than the standard $O(\sqrt{T})$ worst-case regret; e.g. when the losses are smooth (Srebro et al., 2010a; Cutkosky and Orabona, 2018). We say more on this in §4.2 below.

Oracle Complexity. Note that at each round t, the instance of Alg. 1 in Thm. 8 invokes $\mathsf{OPT}_{\mathcal{C}^\circ}(\cdot;\delta_t)$ once. Thus, in light of the discussion in §3 (see also Rem. 12 in App. B), the algorithm makes at most $O(dT\ln(dT\kappa/\delta)$ calls to the Membership Oracle $\mathsf{MEM}_{\mathcal{C}}$ after T rounds. We also note that if one has a Separation Oracle for $\mathcal{C}$, then Alg. 1 can be executed with only $\widetilde{O}(1)$ calls-per-round to the latter (instead of $\widetilde{O}(d)$ calls to $\mathsf{MEM}_{\mathcal{C}}$) without compromising the regret bound in Thm. 8.

We can reduce the number of $\mathsf{MEM}_{\mathcal{C}}$ Oracle calls to a total of $O(T\ln(dT\kappa/\delta))$ by using the "one-dimensional" stochastic version of $\mathsf{OPT}_{\mathcal{C}^\circ}$ we discussed in §3; i.e. $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$ (Alg. 4). This will come at the cost of a $O(\sqrt{d})$ multiplicative factor in the regret bound (see App. C.1).

The strongly convex case. In App. C.2, we design a subroutine A (Alg. 5) such that the instance of Alg. 1 with this subroutine achieves a logarithmic regret whenever the losses are strongly convex, while maintaining a $\widetilde{O}(\sqrt{T})$ regret in the worse case. Further, the algorithm does no require any curvature parameter as input. Finally, in Appendix E we present a general reduction that makes our algorithm for the strongly convex case as well as many previous online algorithms (such as those by (Ross et al., 2013; Wintenberger, 2017; Kotłowski, 2017; Kempka et al., 2019)) scale-invariant; in the sense that scaling the losses by some positive constant does not change the outputs of the algorithm (consequently the algorithm does not require any scale information as input). The reduction is a generalization of previous algorithms by Mhammedi et al. (2019); Mhammedi and Koolen (2020).

4.2. Algorithms for Stochastic and Offline Optimization

The stochastic setting. The guarantee of Alg. 1 in Thm. 8 readily transfers to the stochastic setting (formally described in App. C.3) via well-known online-to-batch conversion techniques (Cesa-Bianchi et al., 2004; Shalev-Shwartz et al., 2011), leading to a convergence rate of order $O(\kappa/\sqrt{T})$ which is optimal in T. What is more, it can be shown that the instance of Alg. 1 in Thm. 8 achieves a rate of order $\widetilde{O}(\beta\kappa^2/T+\sigma\kappa/\sqrt{T})$ when the objective function is β-smooth, without requiring β or σ (the noise in the subgradients) as input. This follows by the fact that the bound in Thm. 8 is of the form $O(\sqrt{V_T})$—enabled by the use of $\mathsf{FTRL\text{-}prox}$—and (Cutkosky, 2019, Cor. 6).

In App. C.3, we build a subroutine A (Alg. 6) based on an existing reduction due to Cutkosky (2019), such that the instance of Alg. 1 with this subroutine and $\mathcal{O}_{\mathcal{C}^\circ}\equiv\mathsf{OPT}_{\mathcal{C}^\circ}$ achieves the faster rate $\widetilde{O}(\beta\kappa^2/T^2+\sigma\kappa/\sqrt{T})$, without knowing β or σ (see Thm. 18).

The offline setting. Finally, in the offline setting (a special case of the stochastic setting with zero noise, i.e. $\sigma=0$), the rates we just mentioned for general [resp. β-smooth] convex functions imply that at most $\widetilde{O}(d\kappa^2/\epsilon^2)$ [resp. $\widetilde{O}(d\kappa/\sqrt{\epsilon})$] Membership Oracle calls are required to find an ϵ-suboptimal point. Since state-of-the-art algorithms based on the cutting plane method require $O(d^2\ln(1/\epsilon))$ calls to a Membership Oracle to achieve the same guarantee (Lee et al., 2018), our algorithm provides a viable alternative to the latter whenever $d\geq\Omega(\kappa^2/\epsilon^2)$ for general convex functions or when $d\geq\Omega(\kappa/\sqrt{\epsilon})$ for smooth functions. We note that the algorithm of Lan and Zhou (2016) has a similar guarantee to ours in the offline setting, though they require a Linear Optimization Oracle on $\mathcal{C}$ instead of a Membership Oracle.

5. Discussion

In this paper, we presented a novel projection-free reduction that allowed us to turn any OCO algorithm defined on a Euclidean ball $\mathcal{B}$, to an algorithm on any convex set $\mathcal{C}$ contained in $\mathcal{B}$, without sacrificing the performance of the original algorithm by much. Thanks to this, we were able to

build explicit algorithms that achieve optimal regret bounds in OCO without incurring expensive Euclidean projections. In particular, we swapped the latter for Gauge projections, which can be performed efficiently using a Membership Oracle; our final algorithms make at most $O(dT \ln T)$ [resp. $O(T \ln T)$] calls to such an Oracle after T rounds and guarantee a $O(\kappa\sqrt{T})$ [resp. $O(\kappa\sqrt{dT})$] regret bound, where $\kappa := R/r$ and $R, r > 0$ are such that $\mathcal{B}(r) \subseteq \mathcal{C} \subseteq \mathcal{B}(R)$.

The multiplicative factor κ in our regret bounds, which depends implicitly on the dimension, replaces the $\sqrt{d}$ factor in the regret bounds of previous projection-free algorithms (see e.g. (Hazan and Kale, 2012; Hazan and Minasyan, 2020)). From Table 2, we see that $\kappa \leq \sqrt{d}$ for many sets of interests (the bounds on κ are derived in App. F). Furthermore, κ can always be assumed to be less than d [resp. $\sqrt{d}$], without loss of generality, for general [resp. centrally symmetric] convex sets by applying a certain affine reparametrization if necessary (Flaxman et al., 2005). We note that κ may be thought of as a measure of how well the set $\mathcal{C}$ can be approximated by a Euclidean ball (a larger κ means that $\mathcal{C}$ is not very "round"). If the set $\mathcal{C}$ is more like a "box", then one way to reduce κ is to apply our reduction (Alg. 1) with a subroutine A defined on, for example, a hypercube containing $\mathcal{C}$ (where Euclidean projections are still cheap), rather than a Euclidean ball. We leave such investigations for future work.

For the same number of Oracle calls as in our reduction, no existing Frank-Wolfe-based algorithm guarantees a $O(\sqrt{T})$ regret for general convex sets (unlike with our approach). However, this would be comparing apples to oranges unless we consider the computational cost of the different Oracles involved: Membership versus Linear Optimization Oracle. In Table 2, we see that the computational complexities of these Oracles are comparable for various popular sets (see also Tab. 3). In many optimization settings, the set of interest is specified via a small number of inequalities. In such cases, a Membership Oracle, which simply checks all the inequalities, typically admits a more efficient implementation than an LOO (e.g. convex-hull of permutation matrices—see Tab. 2). On the other hand, there are sets with combinatorial structure such as Matroids (where the number of inequalities defining the set can be exponential), for which performing Linear Optimization can be more efficient than Membership evaluation. In this case, Frank-Wolfe-style algorithms may be more practical. More generally, since $(\mathcal{C}^\circ)^\circ = \mathcal{C}$ for a closed convex set $\mathcal{C}$, and LO on a convex set $\mathcal{K}$ is equivalent to performing separation on its polar $\mathcal{K}^\circ$, it is always possible to pick a set $\mathcal{C}$ on which LO is cheaper than performing separation/testing membership, and vice versa. For these reasons, our approach (which relies on separation/membership evaluation) may be viewed as complementary to ones that use an LOO.

Bonus results. Since our reduction works for any base algorithm, one can achieve different types of guarantees (e.g. a dynamic/anytime regret or one that scales with the norm of the comparator (Mhammedi and Koolen, 2020; Cutkosky, 2020)) by substituting A in Alg. 1 with an algorithm that is known to achieve the desired guarantee. We also note that our reduction can easily be extended to the setting where the losses are exp-concave instead of strongly convex using existing results due to Mhammedi et al. (2019). One can also easily extend our reduction to the Bandit setting (Chen et al., 2019; Garber and Kretzu, 2020).

Domain $\mathcal{C}$	Upper bound on κ post-reparam. (see App. F for param. details).	Computational Complexity of LOO $\mathcal{O}_{\mathcal{C}}(\boldsymbol{x};\delta)$	MO $\mathcal{M}_{\mathcal{C}}(\boldsymbol{x};\delta)$
ℓ_p-ball in $\mathbb{R}^d$	$O(d^{\lvert 1/p-1/2\rvert})$, where $\mathcal{C}\subset\mathbb{R}^d$	$O(d)$	$O(d)$
Simplex $\Delta_d\in\mathbb{R}^d$	$O(d)$, where $\mathcal{C}\subset\mathbb{R}^d$	$O(d)$	$O(d)$
Trace-norm-ball in $\mathbb{R}^{m\times n}$	$O(d^{1/4})$, where $\mathcal{C}\subset\mathbb{R}^d$	$O(\frac{\text{nnz}(\boldsymbol{x})}{\sqrt{\delta}})$	Cost(SVD)
Op-norm-ball in $\mathbb{R}^{m\times n}$	$O(d^{1/4})$, where $\mathcal{C}\subset\mathbb{R}^d$	Cost(SVD)	$O(\frac{\text{nnz}(\boldsymbol{x})}{\sqrt{\delta}})$
Conv-hull of Permutation Matrices in $\mathbb{R}^{n\times n}$	$O(d^{1/2})$, where $\mathcal{C}\subset\mathbb{R}^d$	$O(n^3)$	$O(n^2)$
Convex-hull of Rotation Matrices in $\mathbb{R}^{n\times n}$	$O(d^{1/4})$, where $\mathcal{C}\subset\mathbb{R}^d$	Cost(SVD)	Cost(SVD)
PSD matrices in $\mathbb{R}^{n\times n}$ with unit trace	$O(d)$, where $\mathcal{C}\subset\mathbb{R}^d$	$O(\frac{\text{nnz}(\boldsymbol{x})}{\sqrt{\delta}})$	$O(\frac{\text{nnx}(\boldsymbol{x})}{\sqrt{\delta}})$
PSD matrices in $\mathbb{R}^{n\times n}$ with diagonals ≤ 1	$O(d^{3/4})$, where $\mathcal{C}\subset\mathbb{R}^d$	$O(\frac{\text{nnz}(\boldsymbol{x})}{\sqrt{\delta^5}n^{-3}})$	$O(\frac{\text{nnx}(\boldsymbol{x})}{\sqrt{\delta}})$

Table 2: If $\mathcal{C}\subseteq\mathbb{R}^{n\times m}$, then $d\coloneqq mn$. $\text{nnz}(\cdot)\equiv$ # of non-zeros. Table entries derived in App. F.

Acknowledgments

We thank anonymous reviewers for detailed comments that helped improve the presentation of the paper. We acknowledge support from the ONR through awards N00014-20-1-2336 and N00014-20-1-2394.

References

Jacob Abernethy, Chansoo Lee, Abhinav Sinha, and Ambuj Tewari. Online linear optimization via smoothing. In *Conference on Learning Theory*, pages 807–823. PMLR, 2014.

Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Interior-point methods for full-information and bandit online learning. *IEEE Transactions on Information Theory*, 58(7):4164–4175, 2012.

Shalabh Bhatnagar. Adaptive newton-based multivariate smoothed functional algorithms for simulation optimization. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 18(1): 1–35, 2007.

Immanuel M. Bomze, Francesco Rinaldi, and Damiano Zeffiro. Frank–wolfe and friends: a journey into projection-free first-order optimization methods. *4OR*, 19(3):313–345, 2021.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Lin Chen, Mingrui Zhang, and Amin Karbasi. Projection-free bandit convex optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2047–2056. PMLR, 2019.

Ashok Cutkosky. Anytime online-to-batch, optimism and acceleration. In *International Conference on Machine Learning*, pages 1446–1454. PMLR, 2019.

Ashok Cutkosky. Parameter-free, dynamic, and strongly-adaptive online learning. In *International Conference on Machine Learning*, pages 2250–2259. PMLR, 2020.

Ashok Cutkosky and Róbert Busa-Fekete. Distributed stochastic optimization via adaptive sgd. *Advances in Neural Information Processing Systems*, 31:1910–1919, 2018.

Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in banach spaces. In *Conference On Learning Theory*, pages 1493–1529. PMLR, 2018.

John C Duchi, Peter L Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394, 2005.

Marguerite Frank, Philip Wolfe, et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Michael P Friedlander, Ives Macedo, and Ting Kei Pong. Gauge optimization and duality. *SIAM Journal on Optimization*, 24(4):1999–2022, 2014.

Dan Garber. Faster projection-free convex optimization over the spectrahedron. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 874–882, 2016.

Dan Garber and Elad Hazan. A linearly convergent variant of the conditional gradient algorithm under strong convexity, with applications to online and stochastic optimization. *SIAM Journal on Optimization*, 26(3):1493–1528, 2016.

Dan Garber and Ben Kretzu. Improved regret bounds for projection-free bandit convex optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2196–2206. PMLR, 2020.

JL Goffin. Affine and projective transformations in nondifferentiable optimization. *Trends in Mathematical Optimization*, pages 79–91, 1988.

M Grötschel, L Lovász, and A Schrijver. Geometric algorithms and combinatorial optimization. *Algorithms and Combinatorics*, 1993.

Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2. Springer Science & Business Media, 2012.

Elad Hazan. Sparse approximate solutions to semidefinite programs. In *Latin American symposium on theoretical informatics*, pages 306–316. Springer, 2008.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2 (3-4):157–325, 2016.

Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

Elad Hazan and Satyen Kale. Projection-free online learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1843–1850, 2012.

Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1263–1271, New York, New York, USA, 20–22 Jun 2016. PMLR.

Elad Hazan and Edgar Minasyan. Faster projection-free online learning. In *Conference on Learning Theory*, pages 1877–1893. PMLR, 2020.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

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

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Michal Kempka, Wojciech Kotłowski, and Manfred K. Warmuth. Adaptive scale-invariant online algorithms for learning linear models. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 3321–3330, 2019.

Thomas Kerdreux. *Accelerating conditional gradient methods*. PhD thesis, Université Paris sciences et lettres, 2020.

Wojciech Kotłowski. Scale-invariant unconstrained online learning. In *International Conference on Algorithmic Learning Theory, ALT 2017, 15-17 October 2017, Kyoto University, Kyoto, Japan*, pages 412–433, 2017.

Ben Kretzu and Dan Garber. Revisiting projection-free online learning: the strongly convex case. In *International Conference on Artificial Intelligence and Statistics*, pages 3592–3600. PMLR, 2021.

Jacek Kuczyński and Henryk Woźniakowski. Estimating the largest eigenvalue by the power and lanczos algorithms with a random start. *SIAM journal on matrix analysis and applications*, 13(4): 1094–1122, 1992.

Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization*, 26(2):1379–1409, 2016.

Guanghui Lan, Sebastian Pokutta, Yi Zhou, and Daniel Zink. Conditional accelerated lazy stochastic gradient descent. In *International Conference on Machine Learning*, pages 1965–1974. PMLR, 2017.

Yin Tat Lee, Aaron Sidford, and Sam Chiu-wai Wong. A faster cutting plane method and its implications for combinatorial and convex optimization. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1049–1065. IEEE, 2015.

Yin Tat Lee, Aaron Sidford, and Santosh S Vempala. Efficient convex optimization with membership oracles. In *Conference On Learning Theory*, pages 1292–1294. PMLR, 2018.

Kfir Levy and Andreas Krause. Projection free online learning over smooth sets. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1458–1466. PMLR, 2019.

Kfir Y Levy, Alp Yurtsever, and Volkan Cevher. Online adaptive methods, universality and acceleration. *Advances in Neural Information Processing Systems*, 31:6500–6509, 2018.

Mehrdad Mahdavi, Tianbao Yang, Rong Jin, Shenghuo Zhu, and Jinfeng Yi. Stochastic gradient descent with only one projection. *Advances in neural information processing systems*, 25:494–502, 2012.

H Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *The Journal of Machine Learning Research*, 18(1):3117–3166, 2017.

Karola Mészáros, Alejandro H Morales, and Jessica Striker. On flow polytopes, order polytopes, and certain faces of the alternating sign matrix polytope. *Discrete & Computational Geometry*, 62(1): 128–163, 2019.

Zakaria Mhammedi and Wouter M. Koolen. Lipschitz and comparator-norm adaptivity in online learning. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 2858–2887. PMLR, 2020.

Zakaria Mhammedi, Wouter M Koolen, and Tim Van Erven. Lipschitz adaptivity with multiple learning rates in online learning. In *Conference on Learning Theory*, pages 2490–2511. PMLR, 2019.

Marco Molinaro. Curvature of feasible sets in offline and online optimization. *arXiv preprint arXiv:2002.03213*, 2020.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Francesco Orabona and Dávid Pál. Open problem: Parameter-free and scale-free online algorithms. In *Conference on Learning Theory*, pages 1659–1664. PMLR, 2016.

Stephane Ross, Paul Mineiro, and John Langford. Normalized online learning. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 537–545, 2013.

James Saunderson, Pablo A Parrilo, and Alan S Willsky. Semidefinite descriptions of the convex hull of rotation matrices. *SIAM Journal on Optimization*, 25(3):1314–1343, 2015.

Alexander Schrijver. *Combinatorial optimization: polyhedra and efficiency*, volume 24. Springer Science & Business Media, 2003.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010a.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low-noise and fast rates. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems-Volume 2*, pages 2199–2207, 2010b.

Tim Van Erven and Wouter M. Koolen. Metagrad: Multiple learning rates in online learning. In *29 (NIPS)*, pages 3666–3674, 2016.

Tim Van Erven, Wouter M. Koolen, and Dirk Van der Hoeven. Metagrad: Adaptation using multiple learning rates in online learning. *arXiv preprint arXiv:2102.06622*, 2021.

Olivier Wintenberger. Optimal learning with Bernstein online aggregation. *Machine Learning*, 106 (1):119–141, 2017.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

Appendices

Contents

Appendix A. Additional Notation

In this section, we present some additional notation that we omitted from §2 of the main body due to space. We let $\ln_+(u) \coloneqq 0 \vee \ln(u)$, for $u > 0$. For $(\boldsymbol{x}_t) \subset \mathbb{R}^d$ and $t \in \mathbb{N}$, we write $\boldsymbol{x}_{1:t} \coloneqq (\boldsymbol{x}_1, \dots, \boldsymbol{x}_t) \in \mathbb{R}^{d \times t}$. For any real function $f \colon \mathcal{A} \to \mathbb{R}$, we denote by $\arg\max_{a \in \mathcal{A}} f(a)$ [resp. $\arg\min_{a \in \mathcal{A}} f(a)$] the subset $\mathcal{K}$ of $\mathcal{A}$ such that $f(b) = \sup_{a \in \mathcal{A}} f(a)$ [resp. $f(b) = \inf_{a \in \mathcal{A}} f(a)$], for all $b \in \mathcal{K}$. For any $p \in \mathbb{R} \cup \{\infty\}$, $\boldsymbol{u} \in \mathbb{R}^d$, and $r \geq 0$ we denote by $\mathcal{B}_p(\boldsymbol{u}, r)$, the $\|\cdot\|_p$-ball of radius r centered at $\boldsymbol{u}$, where $\|\cdot\|_p$ denotes the p-norm. When $\boldsymbol{u} = \boldsymbol{0}$, we simply write $\mathcal{B}_p(r)$ for $\mathcal{B}_p(\boldsymbol{0}, r)$. For the special case where $p = 2$, we let $\|\cdot\| \coloneqq \|\cdot\|_2$ and $\mathcal{B}(\cdot) \coloneqq \mathcal{B}_2(\cdot)$. We denote by $\Pi_{\mathcal{B}(r)} \colon \mathbb{R}^d \to \mathcal{B}(r)$ the Euclidean projection operator onto the ball $\mathcal{B}(r)$. For any set $\mathcal{K}$, we denote by $\operatorname{conv} \mathcal{K}$ [resp. $\operatorname{bd} \mathcal{C}$] the convex hull [resp. boundary] of $\mathcal{K}$. We let $\iota_{\mathcal{K}} \colon \mathbb{R}^d \to \{0, +\infty\}$ be the indicator function of the set $\mathcal{K}$, where $\iota_{\mathcal{K}}(\boldsymbol{w}) = 0$ if and only if $\boldsymbol{w} \in \mathcal{K}$. Finally, for any mathematical statement $\mathcal{E}$, we let $\mathbb{I}_{\mathcal{E}} = 1$ if $\mathcal{E}$ is true, and 0 otherwise.

Appendix B. Efficient Gauge Projections using a Membership Oracle $\mathcal{M}_{\mathcal{C}}$

In this section, we build explicit algorithms that use $\mathsf{MEM}_{\mathcal{C}}$ (see Definition 2) to efficiently approximate $\gamma_{\mathcal{C}}$ and its subgradients. As a result of this (and thanks to Lemma 6), we show that Gauge projections can be performed efficiently for any bounded convex set $\mathcal{C}$ satisfying Assumption 1 using a Membership Oracle—requiring only $\widetilde{O}(d)$ calls to the latter, where $\widetilde{O}$ hides log factors in the tolerated approximation error. Thanks to Lemma 6, Gauge projections have a very simple interpretation; the projection of a point $\boldsymbol{w} \notin \mathcal{C}$ onto $\mathcal{C}$ is the point of intersection of the ray $\{\lambda \boldsymbol{w} : \lambda \geq 0\}$ and the boundary of $\mathcal{C}$ (see Figure 1). Such a point always exists under Assumption 1. Gauge projections are all we need to ensure the iterates of our new algorithms are within $\mathcal{C}$ thanks to the carefully designed surrogate losses in Section 4.

Our starting point is Lemma 6. To prove it, we need the concept of a normal cone:

Definition 9 (Normal Cone) *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a convex set. The* normal cone *of $\mathcal{C}$ at $\boldsymbol{u} \in \mathcal{C}$ is the set* $\mathcal{N}_{\mathcal{C}}(\boldsymbol{u}) \coloneqq \{\boldsymbol{s} \in \mathcal{C} : \langle \boldsymbol{s}, \boldsymbol{w} - \boldsymbol{u} \rangle \leq 0, \forall \boldsymbol{w} \in \mathcal{C}\}$.

We also recall that $\iota_{\mathcal{C}}$ denotes the indicator function of a set $\mathcal{C}$ (see notation in App. A). With this, we now prove Lemma 6:

Proof of Lemma 6. Suppose that $\gamma_{\mathcal{C}}(\boldsymbol{w}) > 1$ (note that by Assumption 1, we have $\gamma_{\mathcal{C}}(\boldsymbol{w}) < +\infty$). Note that this is equivalent to $\boldsymbol{w} \notin \mathcal{C}$ by the definition of the Gauge function. We will show that $\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w}) \in \arg\min_{\boldsymbol{u} \in \mathcal{C}} \gamma_{\mathcal{C}}(\boldsymbol{w} - \boldsymbol{u})$, which is equivalent to showing that

$$\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w}) \in \operatorname*{arg\,min}_{\boldsymbol{u} \in \mathbb{R}^d} \{\gamma_{\mathcal{C}}(\boldsymbol{w} - \boldsymbol{u}) + \iota_{\mathcal{C}}(\boldsymbol{u})\} = \operatorname*{arg\,min}_{\boldsymbol{u} \in \mathbb{R}^d} \{\sigma_{\mathcal{C}^\circ}(\boldsymbol{w} - \boldsymbol{u}) + \iota_{\mathcal{C}}(\boldsymbol{u})\},$$

where the last equality follows by Lemma 42-(a). Thus, $\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w}) \in \arg\min_{\boldsymbol{u} \in \mathcal{C}} \gamma_{\mathcal{C}}(\boldsymbol{w} - \boldsymbol{u})$ if

$$\boldsymbol{0} \in \partial \phi(\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w})), \quad \text{where} \quad \phi(\boldsymbol{u}) \coloneqq \sigma_{\mathcal{C}^\circ}(\boldsymbol{w} - \boldsymbol{u}) + \iota_{\mathcal{C}}(\boldsymbol{u}). \tag{11}$$

Using the sub-differential chain-rule and the fact that $\partial \iota_{\mathcal{C}}(\boldsymbol{u}) = \mathcal{N}_{\mathcal{C}}(\boldsymbol{u})$ (see e.g. (Hiriart-Urruty and Lemaréchal, 2004)), where $\mathcal{N}_{\mathcal{C}}(\boldsymbol{u})$ is the normal cone at $\boldsymbol{u}$ (see Definition 3), we have

$$\partial \phi(\boldsymbol{u}) = -\partial \sigma_{\mathcal{C}^\circ}(\boldsymbol{w} - \boldsymbol{u}) + \mathcal{N}_{\mathcal{C}}(\boldsymbol{u}). \tag{12}$$

Let $\boldsymbol{s}_* \in \partial\sigma_{\mathcal{C}^\circ}(\boldsymbol{w} - \boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w}))$. We will show that $\boldsymbol{s}_* \in \mathcal{N}_{\mathcal{C}}(\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w}))$, where we recall that

$$\mathcal{N}_{\mathcal{C}}(\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w})) = \{\boldsymbol{s} \in \mathbb{R}^d : \langle \boldsymbol{s}, \boldsymbol{y}\rangle \leq \langle \boldsymbol{s}, \boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w})\rangle, \forall \boldsymbol{y} \in \mathcal{C}\}, \tag{13}$$

which will imply (11) thanks to (12). By Lemma 42-(b), we have $\boldsymbol{s}_* \in \partial\sigma_{\mathcal{C}^\circ}(\boldsymbol{w} - \boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w})) = \partial\sigma_{\mathcal{C}^\circ}(\boldsymbol{w}) = \arg\max_{\boldsymbol{s}\in\mathcal{C}^\circ}\langle \boldsymbol{s}, \boldsymbol{w}\rangle$, and so by Lemma 42-(a), we get $\langle \boldsymbol{s}_*, \boldsymbol{w}\rangle = \sigma_{\mathcal{C}^\circ}(\boldsymbol{w}) = \gamma_{\mathcal{C}}(\boldsymbol{w})$. On the other hand, by definition of the polar set $\mathcal{C}^\circ$, we have $\langle \boldsymbol{s}_*, \boldsymbol{y}\rangle \leq 1$, for all $\boldsymbol{y} \in \mathcal{C}$. Using this and the fact that $\langle \boldsymbol{s}_*, \boldsymbol{w}\rangle = \gamma_{\mathcal{C}}(\boldsymbol{w})$, we get

$$\forall \boldsymbol{y} \in \mathcal{C}, \quad \langle \boldsymbol{s}_*, \boldsymbol{y}\rangle \leq 1 = \langle \boldsymbol{s}_*, \boldsymbol{w}\rangle/\gamma_{\mathcal{C}}(\boldsymbol{w}) = \langle \boldsymbol{s}_*, \boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w})\rangle.$$

Combining this with the definition of the normal cone $\mathcal{N}_{\mathcal{C}}(\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w}))$ in (13), we get that $\boldsymbol{s}_* \in \mathcal{N}_{\mathcal{C}}(\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w}))$. This shows that $\boldsymbol{s}_* \in \partial\sigma_{\mathcal{C}^\circ}(\boldsymbol{w} - \boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w})) \cap \mathcal{N}_{\mathcal{C}}(\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w}))$, and so $\mathbf{0} \in \partial\phi(\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w}))$ by (12). This in turn implies that

$$\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w}) \in \arg\min_{\boldsymbol{u}\in\mathcal{C}} \gamma_{\mathcal{C}}(\boldsymbol{w} - \boldsymbol{u}).$$

From this result and the definition of the proximity function $S_{\mathcal{C}}(\boldsymbol{w})$, we get

$$S_{\mathcal{C}}(\boldsymbol{w}) = \gamma_{\mathcal{C}}(\boldsymbol{w} - \boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w})) = \gamma_{\mathcal{C}}(\boldsymbol{w} \cdot (1 - 1/\gamma_{\mathcal{C}}(\boldsymbol{w}))) = (1 - 1/\gamma_{\mathcal{C}}(\boldsymbol{w})) \cdot \gamma_{\mathcal{C}}(\boldsymbol{w}) = \gamma_{\mathcal{C}}(\boldsymbol{w}) - 1, \tag{14}$$

where the penultimate inequality follows by the current assumption that $\gamma_{\mathcal{C}}(\boldsymbol{w}) > 1$ and the positive homogeneity of Gauge functions (Lemma 42-(b)). It remains to consider the case where $\gamma_{\mathcal{C}}(\boldsymbol{w}) \leq 1$. In this case, we have $\boldsymbol{w} \in \mathcal{C}$, and so

$$S_{\mathcal{C}}(\boldsymbol{w}) = \inf_{\boldsymbol{u}\in\mathcal{C}} \gamma_{\mathcal{C}}(\boldsymbol{w} - \boldsymbol{u}) = \gamma_{\mathcal{C}}(\boldsymbol{w} - \boldsymbol{w}) = 0. \tag{15}$$

In combination with (14), (15) shows that $S_{\mathcal{C}}(\boldsymbol{w}) = 0 \vee (\gamma_{\mathcal{C}}(\boldsymbol{w}) - 1)$, which is a convex function (since it is the maximum of two convex functions). Finally, (14) [resp. (15)] shows that $\partial S_{\mathcal{C}}(\boldsymbol{w}) = \arg\max_{\boldsymbol{s}\in\mathcal{C}^\circ}\langle \boldsymbol{s}, \boldsymbol{w}\rangle$ when $\boldsymbol{w} \notin \mathcal{C} (\equiv \gamma_{\mathcal{C}}(\boldsymbol{w}) > 1)$ [resp. $\partial S_{\mathcal{C}}(\boldsymbol{w}) = \{\mathbf{0}\}$ when $\boldsymbol{w} \in \mathcal{C}$]. ■

As mentioned earlier, Lemma 6 shows the crucial fact that $\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w}) \in \Pi^{\mathrm{G}}_{\mathcal{C}}(\boldsymbol{w})$, for any $\boldsymbol{w} \notin \mathcal{C}$, and so to perform approximate Gauge projections (which we need to build our efficient algorithms), it suffices to approximate $\gamma_{\mathcal{C}}$. We technically also need to compute approximate subgradients of $S_{\mathcal{C}}$, which according to Lemma 6 reduces to linear optimization on $\mathcal{C}^\circ$. Since we do not assume access to a Linear Optimizer Oracle over $\mathcal{C}^\circ$, we will estimate subgradients of $S_{\mathcal{C}}$ using our Membership Oracle $\mathsf{MEM}_{\mathcal{C}}$.

B.1. Approximating the Gauge Function $\gamma_{\mathcal{C}}$ using $\mathcal{M}_{\mathcal{C}}$

By definition of the Gauge function, we have for any $\boldsymbol{w} \in \mathbb{R}^d$

$$\gamma_{\mathcal{C}}(\boldsymbol{w}) = \inf\{\lambda \in \mathbb{R}_{\geq 0} \mid \boldsymbol{w} \in \lambda\mathcal{C}\} = 1/\sup\{\nu \in \mathbb{R}_{\geq 0} \mid \nu\boldsymbol{w} \in \mathcal{C}\}. \tag{16}$$

Using the Membership Oracle $\mathsf{MEM}_{\mathcal{C}}$, we can approximate the largest $\nu \geq 0$ such that $\nu\boldsymbol{w} \in \mathcal{C}$ via bisection, which will lead to an approximation of $\gamma_{\mathcal{C}}(\boldsymbol{w})$ by (16). This is exactly what we do in Algorithm 2. We now state the guarantee of Algorithm 2 (the proof is Appendix H.1):

Algorithm 2 $\mathsf{GAU}_{\mathcal{C}}$: Approximate Gauge Function using Membership Oracle and the Bisection Method.

Require: Input $(\boldsymbol{w}, \delta) \in \mathcal{B}(6R/5) \times (0,1)$ with R as in (1). ϵ-Approximate Membership Oracle $\mathsf{MEM}_{\mathcal{C}}(\cdot;\epsilon)$ for $\mathcal{C}$ and $\epsilon > 0$ (see Definition 2).

Set $\epsilon = \delta r/(4\kappa)^2$. // $\kappa := R/r$, where r and R are as in (1)
if $\mathsf{MEM}_{\mathcal{C}}(2\boldsymbol{w};\epsilon) = 1$ or $\|\boldsymbol{w}\| \leq r/2$ **then**
  Return $\tilde{\gamma} = 0$.
end if
Set $\alpha = 0$, $\beta = 2$, and $\mu = (\alpha+\beta)/2$.
while $\beta - \alpha > \delta/(8\kappa^2)$ **do**
  Set $\alpha = \mu$ if $\mathsf{MEM}_{\mathcal{C}}(\mu\boldsymbol{w};\epsilon) = 1$; and $\beta = \mu$ otherwise.
  Set $\mu = (\alpha+\beta)/2$.
end while
Return $\tilde{\gamma} = (\alpha - \delta/(8\kappa^2))^{-1}$.

Lemma 10 ($\mathsf{GAU}_{\mathcal{C}}$: **Approximate Gauge Function**) *Let $r, R > 0$ be as in (1). For any $\delta \in (0,1)$ and $\boldsymbol{w} \in \mathcal{B}(6R/5)$, the output $\tilde{\gamma} = \mathsf{GAU}_{\mathcal{C}}(\boldsymbol{w};\delta)$ of Algorithm 2 satisfies*

$$\gamma_{\mathcal{C}}(\boldsymbol{w}) \leq \tilde{\gamma} \leq \gamma_{\mathcal{C}}(\boldsymbol{w}) + \delta, \quad \text{if } \gamma_{\mathcal{C}}(\boldsymbol{w}) \geq 9/16 \text{ or } \tilde{\gamma} \geq 1.$$

Furthermore, Alg. 2 calls the Membership Oracle $\mathsf{MEM}_{\mathcal{C}}(\cdot; r\delta/(4\kappa)^2)$ at most $\lceil \log_2((4\kappa)^2/\delta) \rceil + 1$ times.

Algorithm 3 $\mathsf{OPT}_{\mathcal{C}^\circ}$: Approximate Linear Optimization Algorithm on $\mathcal{C}^\circ$.

Require: Input point $\boldsymbol{w} \in \mathcal{B}(R)$ and $\delta \in (0, 1/3)$.

Set $\varepsilon = \frac{r^2\delta^3}{10^3 d^{7/2} R^2}$, $\nu_1 = \frac{r\delta}{10d}$, and $\nu_2 = \sqrt{\frac{\varepsilon\nu_1 r}{d^{1/2}}}$. // r and R are as in (1)
Sample $\boldsymbol{u} \in \mathcal{B}_\infty(\boldsymbol{w}, \nu_1)$ and $\boldsymbol{z} \in \mathcal{B}_\infty(\boldsymbol{u}, \nu_2)$ independently and uniformly at random.
for $i = 1, 2, \ldots, d$ **do**
  Let $\boldsymbol{w}_i'$ and $\boldsymbol{w}_i$ be the end point of the interval $\mathcal{B}_\infty(\boldsymbol{u}, \nu_2) \cap \{\boldsymbol{z} + \lambda \boldsymbol{e}_i : \lambda \in \mathbb{R}\}$.
  Set $\tilde{s}_i = \frac{1}{2\nu_2}(\mathsf{GAU}_{\mathcal{C}}(\boldsymbol{w}_i';\varepsilon) - \mathsf{GAU}_{\mathcal{C}}(\boldsymbol{w}_i;\varepsilon))$. // $\mathsf{GAU}_{\mathcal{C}}$ is as in Algorithm 2
end for
Set $\tilde{\boldsymbol{s}} = (\tilde{s}_1, \ldots, \tilde{s}_d)^\top$ and $\tilde{\gamma} = \mathsf{GAU}_{\mathcal{C}}(\boldsymbol{w};\delta)$.
Return $(\tilde{\gamma}, \tilde{\boldsymbol{s}})$.

B.2. Approximating the Subgradients of $\gamma_{\mathcal{C}}$ using $\mathcal{M}_{\mathcal{C}}$

In addition to approximating $\gamma_{\mathcal{C}}$, we will also need to approximate its subgradients, which would then lead to approximate subgradients of the Gauge distance function $S_{\mathcal{C}}$ by Lemma 6. The lemma also implies that approximating subgradients of $\gamma_{\mathcal{C}}$ essentially comes down to performing linear optimization on $\mathcal{C}^\circ$. Algorithm 3 ($\mathsf{OPT}_{\mathcal{C}^\circ}$), which is based on (Lee et al., 2018, Alg. 2), uses $\mathsf{GAU}_{\mathcal{C}}$ and a random partial difference in each coordinate to approximate the subgradients of $\gamma_{\mathcal{C}}$. In the next proposition, we state the precise guarantee of the algorithm. The proof of the proposition, which is in

Appendix H.2, is somewhat technical and relies heavily on existing results due to Lee et al. (2018) that we restate in Appendix H.

Proposition 11 ($\mathsf{OPT}_{\mathcal{C}^\circ}$: **Approximate LOO on** $\mathcal{C}^\circ$) *Let $\kappa \coloneqq R/r$ with $r, R > 0$ as in (1). For any $\boldsymbol{w} \in \mathcal{B}(R)$ and $\delta \in (0, 1/3)$, let $(\tilde{\gamma}, \tilde{\boldsymbol{s}})$ be the output of Alg. 3 with input $(\boldsymbol{w}, \delta)$. Then, $\|\tilde{\boldsymbol{s}}\|_\infty < +\infty$ almost surely, and there exists a positive random variable $\Delta \in [0, 15^2 d^4 \kappa^3 \delta^{-2}]$ satisfying $\mathbb{E}[\Delta] \le \delta$, and such that if $\tilde{\gamma} \ge 1$, then*

$$\begin{gathered}\gamma_{\mathcal{C}}(\boldsymbol{u}) \ge \gamma_{\mathcal{C}}(\boldsymbol{w}) + \langle \tilde{\boldsymbol{s}}, \boldsymbol{u} - \boldsymbol{w} \rangle - \Delta \cdot \max(1, \|\boldsymbol{u}\|/R), \quad \forall \boldsymbol{u} \in \mathbb{R}^d; \\ \|\tilde{\boldsymbol{s}}\|_\infty \le \frac{\delta}{R} + \frac{1}{r}; \quad \|\tilde{\boldsymbol{s}}\| \le \frac{\Delta}{R} + \frac{1}{r}; \text{ and } \|\tilde{\boldsymbol{s}}\|^2 \le \left(\frac{2}{r} + \frac{\delta}{R}\right)\frac{\Delta}{R} + \frac{1}{r^2}.\end{gathered} \tag{17}$$

Remark 12 (Complexity of $\mathsf{OPT}_{\mathcal{C}^\circ}$) *In the setting of Proposition 11, Algorithm 3 ($\mathsf{OPT}_{\mathcal{C}^\circ}$) requires $2d \cdot (\lceil \log_2((4\kappa)^2/\varepsilon) \rceil + 1)$ calls to the Membership Oracle $\mathsf{MEM}_{\mathcal{C}}(\cdot; r\varepsilon/(4\kappa)^2)$, where $\varepsilon = r^2\delta^3/(10^3 d^{7/2} R^2)$, and one call to $\mathsf{MEM}_{\mathcal{C}}(\cdot; r\varepsilon/(4\kappa)^2)$. This follows by Lemma 10 and the fact that $\mathsf{OPT}_{\mathcal{C}^\circ}$ makes $2 \cdot d$ calls to $\mathsf{GAU}_{\mathcal{C}}(\cdot; \varepsilon)$ and one call to $\mathsf{GAU}_{\mathcal{C}}(\cdot; \delta)$.*

In light of Remark 12, Proposition 11 implies that approximating the subgradient of $\gamma_{\mathcal{C}}$ at a point $\boldsymbol{w} \in \mathcal{B}(R)$ up to some random error $\Delta \ge 0$ satisfying $\mathbb{E}[\Delta] \le \delta$, requires only $O(d \ln(d\kappa/\delta))$ calls to the Membership Oracle $\mathsf{MEM}_{\mathcal{C}}$. This means that the approximation error decreases exponentially fast with the number of calls to $\mathsf{MEM}_{\mathcal{C}}$, which allows us to build our efficient projection-free algorithms in Section 4.

Algorithm 4 $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$: 'One-Dimensional' Stochastic Version of $\mathsf{OPT}_{\mathcal{C}^\circ}$.

Require: Input point $\boldsymbol{w} \in \mathcal{B}(R)$ and $\delta \in (0, 1/3)$.

1: Set $\varepsilon = \frac{r^2\delta^3}{10^3 d^{7/2} R^2}$, $\nu_1 = \frac{r\delta}{10d}$, and $\nu_2 = \sqrt{\frac{\varepsilon \nu_1 r}{d^{1/2}}}$. // r and R are as in (1)
2: Sample $I \in [d]$, $\boldsymbol{u} \in \mathcal{B}_\infty(\boldsymbol{w}, \nu_1)$, and $\boldsymbol{z} \in \mathcal{B}_\infty(\boldsymbol{u}, \nu_2)$ independently and uniformly at random.
3: Let $\boldsymbol{w}'_I$ and $\boldsymbol{w}_I$ be the end point of the interval $\mathcal{B}_\infty(\boldsymbol{u}, \nu_2) \cap \{\boldsymbol{z} + \lambda \boldsymbol{e}_I : \lambda \in \mathbb{R}\}$.
4: Set $\tilde{s}_I = \frac{1}{2\nu_2}(\mathsf{GAU}_{\mathcal{C}}(\boldsymbol{w}'_I; \varepsilon) - \mathsf{GAU}_{\mathcal{C}}(\boldsymbol{w}_I; \varepsilon))$. // $\mathsf{GAU}_{\mathcal{C}}$ is as in Algorithm 2
5: Set $\hat{\gamma} = \mathsf{GAU}_{\mathcal{C}}(\boldsymbol{w}; \delta)$ and $\hat{\boldsymbol{s}} = d\tilde{s}_I \cdot \boldsymbol{e}_I$.
6: Return $(\hat{\gamma}, \hat{\boldsymbol{s}})$.

In some settings, calling a Membership Oracle $\Omega(d)$ times per iteration might still be too expensive. A way around this is to use a stochastic version $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$ of $\mathsf{OPT}_{\mathcal{C}^\circ}$ that calls $\mathsf{MEM}_{\mathcal{C}}$ at most $\widetilde{O}(1)$ times and has the same output as $\mathsf{OPT}_{\mathcal{C}^\circ}$ in expectation. Algorithm 4, $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$, achieves this by randomly sampling a coordinate $I \in [d]$ for estimating the subgradient of $\gamma_{\mathcal{C}}$ (Line 4 of Alg. 4) and using importance weights. We now state the guarantee of Algorithm 4 (the proof is in Appendix H.3):

Lemma 13 *Let $\delta \in (0, 1/3)$, $\boldsymbol{w} \in \mathcal{B}(R)$, and $\kappa \coloneqq R/r$, where $r, R > 0$ are as in (1). Further, let $(\tilde{\gamma}, \tilde{\boldsymbol{s}}) = \mathsf{OPT}_{\mathcal{C}^\circ}(\boldsymbol{w}; \delta)$ and $(\hat{\gamma}, \hat{\boldsymbol{s}}) = \mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}(\boldsymbol{w}; \delta)$ (Alg. 4). Then, $\|\hat{\boldsymbol{s}}\| < +\infty$ a.s.; $\tilde{\gamma} = \hat{\gamma}$; and if $\hat{\gamma} \ge 1$, it follows that*

$$\mathbb{E}[\hat{\boldsymbol{s}}] = \mathbb{E}[\tilde{\boldsymbol{s}}], \quad \text{and} \quad \mathbb{E}[\|\hat{\boldsymbol{s}}\|^2] \le d \cdot (1/r + \delta/R)^2 .$$

Remark 14 (Complexity of $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$) *In the setting of Lemma 13, Algorithm 4 ($\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$) requires $2 \cdot (\lceil \log_2((4\kappa)^2/\varepsilon)\rceil + 1)$ calls to the Membership Oracle $\mathsf{MEM}_{\mathcal{C}}(\cdot; r\varepsilon/(4\kappa)^2)$, where $\varepsilon = r^2\delta^3/(10^3 d^{7/2}R^2)$, and one call to $\mathsf{MEM}_{\mathcal{C}}(\cdot;\delta)$. This follows by Lemma 10 and the fact that $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$ calls $\mathsf{GAU}_{\mathcal{C}}(\cdot;\varepsilon)$ twice and $\mathsf{GAU}_{\mathcal{C}}(\cdot;\delta)$ once.*

Note that $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$ randomly selects a single coordinate I along which to estimate the subgradient of the Gauge function $\gamma_{\mathcal{C}}$. Generalizing this idea to $k \leq d$ coordinates, one can build a version of $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$, call it $\mathsf{OPT}_{k\mathrm{d},\mathcal{C}^\circ}$, that samples $i \in [\lfloor d/k \rfloor]$ uniformly at random and selects coordinates $\{ik + j - 1 : j \in [k]\} \cap [d]$ along which to estimate the subgradients of $\gamma_{\mathcal{C}}$. In this case, $\mathsf{OPT}_{k\mathrm{d},\mathcal{C}^\circ}$ makes $\widetilde{O}(k)$ calls to the Membership Oracle and would lead to a natural trade-off between computation (i.e. Oracle calls) and regret (see discussion in Section C.1).

Appendix C. Projection-Free Online and Stochastic Optimization (Detailed)

In this section, we provide the details of the online and stochastic optimization algorithms that we omitted from §4 due to space. In §C.1, we consider the general OCO setting where the optimization Oracle $\mathcal{O}_{\mathcal{C}^\circ}$ in Algorithm 1 is set to $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$—the efficient stochastic version of $\mathsf{OPT}_{\mathcal{C}^\circ}$. In §C.2, we design a subroutine A (Alg. 5) such that Algorithm 1, instantiated with this subroutine, achieves a logarithmic regret whenever the losses are strongly convex, while maintain a $\widetilde{O}(\sqrt{T})$ regret in the worse case. In §C.3, we formally define the stochastic optimization setting and design a corresponding subroutine A which, together with Alg. 1, lead to a final algorithm that adapts to the noise and smoothness of the objective function.

C.1. A More Efficient Algorithm for General OCO

We now state the guarantee of Alg. 1 with $\mathcal{O}_{\mathcal{C}^\circ} \equiv \mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$ (the proof in in Appendix I.3):

Theorem 15 *Let $\delta \in (0, 1/3)$ and $\kappa := R/r$, with r and R as in (1). Let (ℓ_t) be any adversarial sequence of convex losses on $\mathcal{C}$ and $(\boldsymbol{x}_t)$ be the iterates of Alg. 1 in response to (ℓ_t). Further, let $\boldsymbol{g}_s \in \partial \ell_s(\boldsymbol{x}_s)$, $s \geq 1$. If Alg. 1 is run with subroutine A set to $\mathsf{FTRL\text{-}prox}$ (Alg. 8) with parameter R; $\mathcal{O}_{\mathcal{C}^\circ} \equiv \mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$; and $\delta_t = \delta/t^2$, $t \geq 1$, then $(\boldsymbol{x}_t) \subset \mathcal{C}$ and*

$$\forall \boldsymbol{x} \in \mathcal{C}, \quad \sum_{t=1}^{T} \mathbb{E}[\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle] \leq 4R\sqrt{(1 + d\cdot(\kappa+\delta)^2)\sum_{t=1}^{T}\mathbb{E}\left[\|\boldsymbol{g}_t\|^2\right]} + 6\delta R \cdot \mathbb{E}\left[\max_{t\in[T]}\|\boldsymbol{g}_t\|\right]. \tag{18}$$

The instance of Algorithm 1 in Theorem 15 invokes $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}(\cdot;\delta_t)$ at each iteration t. Thus, in light of Remark 14 in App. B, the algorithm makes at most $O(T\ln(dT\kappa/\delta)$ calls to the Membership Oracle $\mathsf{MEM}_{\mathcal{C}}$ after T rounds. We also remark that our choice of tolerance sequence (δ_t) in Theorems 8 and 15 is too conservative, requiring the Membership Oracle to be more accurate than necessary to achieve a $O(\sqrt{T})$ regret. In fact, our choice of (δ_t) in Theorems 8 and 15 ensures that the errors involved in the approximations of the subgradients of the surrogate losses add up to a lower order term in the regret bound; i.e. the right-most terms in (9) and (18). We can choose a larger sequence of tolerances as long as the sum of the approximation errors is of order $O(\sqrt{T})$. Next we derive a high probability regret bound for Algorithm 1 and show that one can pick δ_t as large as $O(t^{-1/2})$.

High Probability Regret Bound. We will now derive a high probability regret bound for Algorithm 1 where $\mathcal{O}_{\mathcal{C}^\circ} \equiv \mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$ (Alg. 4) and A is set to $\mathsf{FTRL\text{-}prox}$. For the sake of simplicity, we will assume that the losses are B-Lipschitz (see (20)). Technically, this condition is not needed to derive the result in our next theorem (up to log factors in the regret), but we make it to simplify the probabilistic argument we follow in the proof of the latter. We note that the algorithm does not require knowledge of B.

Theorem 16 *Let $\delta \in (0, 1/3)$, $B > 0$, and $\kappa := R/r$, with r and R as in (1). Let (ℓ_t) be any adversarial sequence of B-Lipschitz convex losses on $\mathcal{C}$ and $(\boldsymbol{x}_t)$ be the iterates of Alg. 1 in response to $(\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t))$. If Alg. 1 is run with subroutine A set to $\mathsf{FTRL\text{-}prox}$ with parameter R; $\mathcal{O}_{\mathcal{C}^\circ} \equiv \mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$ (Alg. 4); and $\delta_t = \delta/t^{-1/2}$, $t \geq 1$, then $(\boldsymbol{x}_t) \subset \mathcal{C}$ and for all $\rho \in (0, 1)$, $T \geq d \ln(1/\rho)$, and $\boldsymbol{x} \in \mathcal{C}$,*

$$\mathbb{P}\left[\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \leq 8RB(\kappa + \delta)\sqrt{dT(3 + 2\ln(1/\rho))} + 2dBR(2\kappa + 2\delta + 3\delta\sqrt{T}/\rho)\right] \geq 1 - \rho.$$

The proof of the theorem is in Appendix I.4. Once again, the instance of Algorithm 1 in Theorem 16 requires only $O(T \ln(Td\kappa/\delta))$ calls of the Membership Oracle $\mathsf{MEM}_{\mathcal{C}}$ (see the discussion proceeding Theorem 15).

Optimality of the bounds and computational trade-offs. All the regret bounds in Theorems 8-16 have an optimal dependence in T. As for the dependence in d, we see that the regret bounds of Algorithm 1 in the settings of Theorems 15 and 16 can be improved by a factor of $O(\sqrt{d})$ when making $\widetilde{O}(d)$ calls to $\mathsf{MEM}_{\mathcal{C}}$ per round (as in the setting of Theorem 8) or $\widetilde{O}(1)$ calls to a Separation Oracle for $\mathcal{C}$, if available. We also note that $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$ (Alg. 4), which is used in the settings of Theorems 15 and 16, randomly selects a coordinate in $[d]$ and estimates the subgradient of $\gamma_{\mathcal{C}}$ in that coordinate's direction. A version of this algorithm, call it $\mathsf{OPT}_{k\mathrm{d},\mathcal{C}^\circ}$, that samples $i \in [\lfloor d/k \rfloor]$ uniformly at random and selects coordinates $\{ik + j - 1 : j \in [k]\} \cap [d]$ to estimate the subgradients of $\gamma_{\mathcal{C}}$ would lead to a regret bound for Algorithm 1, with $\mathcal{O}_{\mathcal{C}^\circ} \equiv \mathsf{OPT}_{k\mathrm{d},\mathcal{C}^\circ}$, of order $O(\sqrt{dT/k})$, while making $\widetilde{O}(k)$ calls to $\mathsf{MEM}_{\mathcal{C}}$ per round. This leads to a natural trade-off between computation (Oracle calls) and regret.

Finally, we note that there is another potential dependence in d in the regret bounds through the asphericity κ. In Section F, we bound this quantity for the popular settings listed on Table 1. We are able to show that κ is often less than $d^{1/2}$ in many settings. Nevertheless, we note that κ is present in our bounds because of our pessimistic upper bounds on the norms of the subgradients of the surrogate losses; i.e. $\|\tilde{\boldsymbol{g}}_t\| \leq (1 + \Delta_t + \kappa)\|\boldsymbol{g}_t\|$, for all $t \geq 1$ (see Lemma 7). In fact, we do not expect the magnitude of $(\tilde{\boldsymbol{g}}_t)$ to be often much larger than that of $(\boldsymbol{g}_t)$ in practice; recall that $\tilde{\boldsymbol{g}}_t \neq \boldsymbol{g}_t$ only if the iterate $\boldsymbol{w}_t$ of the subroutine A is outside $\mathcal{C}$.

Implications for the stochastic and offline settings. The results we presented so far are also relevant in the stochastic and offline settings thanks to the online-to-batch conversion technique (Cesa-Bianchi et al., 2004; Shalev-Shwartz et al., 2011). In the latter settings (which we describe in more detail in Section C.3), the losses (ℓ_t) are i.i.d. and satisfy $\mathbb{E}[\ell_t] \equiv f$ for some fixed convex function $f : \mathcal{C} \to \mathbb{R}$. Thus, if we let $\mathrm{Regret}_T(\cdot)$ be the regret of Algorithm 1 in response to (ℓ_t) and

$x_* \in \arg\min_{\boldsymbol{x}\in\mathcal{C}} f(\boldsymbol{x})$, then the average iterate $\bar{\boldsymbol{x}}_T$ of Alg. 1 after T rounds satisfies

$$\mathbb{E}[f(\bar{\boldsymbol{x}}_T)] - \inf_{\boldsymbol{x}\in\mathcal{C}} f(\boldsymbol{x}) \overset{(*)}{\leq} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{x}_*)] \leq \frac{\text{Regret}_T(\boldsymbol{x}_*)}{T}, \tag{19}$$

where $(*)$ follows by Jensen's inequality and the fact that $\boldsymbol{x}_t$ is independent of ℓ_t. Plugging the regret bounds of Alg. 1 in the settings of Theorems 15-16 into (19) leads to a $O(\sqrt{d\kappa^2/T})$ rate in stochastic and offline optimization, which is optimal in T[2]. Therefore, Alg. 1 in the latter settings requires $\widetilde{O}(d\kappa^2/\epsilon^2)$ calls to the Membership Oracle to attain an ϵ-suboptimal point in offline optimization. Thus, whenever $d \geq \Omega(\kappa^2/\epsilon^2)$, our algorithm is a viable alternative to those based on the cutting plane method, which require at least $O(d^2\ln(1/\epsilon))$ calls to a Membership Oracle to achieve the same guarantee.

When the objective function f is β-smooth our adaptive bounds[3] in Theorems 8 and 15 together with an extension of the online-to-batch conversion analysis (Cutkosky, 2019, Corollary 6) automatically imply a rate of $O(\beta\kappa^2/T + \sigma\kappa/\sqrt{T})$, where σ^2 is the variance of the stochastic noise (see Section C.3 for a precise definition). Thus, in the offline setting (i.e. $\sigma = 0$), our algorithm achieves the fast $O(\beta\kappa^2/T)$ rate without knowing the smoothness parameter β. In Section C.3, we show how this rate can be improved further.

C.2. Algorithm for Strongly Convex Online Optimization

In this section, we will build a subroutine A for Algorithm 1 that will enable the latter to ensure a logarithmic regret when the losses are strongly convex, while maintaining the worst-case $O(\sqrt{T})$ regret up to log-factors. The subroutine in question is displayed in Algorithm 5. This subroutine is based on an exiting reduction due to Cutkosky and Orabona (2018) with the following key differences; I) we perform clipping of the subgradients $(\tilde{\boldsymbol{g}}_t)$ in Alg. 5; II) we slightly change the expression of the variables (Z_t) in Line 8 of Alg 5; and III) we use FreeGrad, a *parameter-free* OCO algorithm (Mhammedi and Koolen, 2020), as the underlying OCO subroutine. These differences will allow us to make our final algorithm scale-invariant in the sense that multiplying the losses by a positive constant does not change the outputs of the algorithm. Crucially, the algorithm does not require any prior scale information on the losses unlike many existing OCO algorithms (see e.g. (Mhammedi and Koolen, 2020)).

2. The rate $O(1/\sqrt{T})$ is optimal when no further assumptions on f are made (other than convexity).

3. Adaptive in the sense that they scale with $\sqrt{\sum_{t=1}^{T}\|\boldsymbol{g}_t\|^2}$ instead of $\sqrt{T}$.

Algorithm 5 Subroutine A for Algorithm 1 for Strongly Convex Online Optimization.

Require: Parameters $\epsilon, R > 0$, with R as in (1) and OCO Algorithm FreeGrad (Alg. 9). Input $(\boldsymbol{x}_t, f_t)$ from the Master Algorithm 1 at each round $t \geq 1$, where $\boldsymbol{x}_t \in \mathbb{R}^d$ and $f_t \colon \mathbb{R}^d \to \mathbb{R}$.

Initialize FreeGrad with parameters $\epsilon, R > 0$, and set $\boldsymbol{u}_0$ to FreeGrad's first output.
Set $\tilde{B}_0 = \epsilon$; $Z_0 = 2\epsilon^2$; and $\boldsymbol{v}_0 = \boldsymbol{0}$.
for $t = 1, 2, \ldots$ **do**
  Play $\boldsymbol{w}_t = \boldsymbol{u}_{t+1} + \boldsymbol{v}_t / Z_t$ and observe $\tilde{\boldsymbol{g}}_t \in \partial f_t(\boldsymbol{w}_t)$. `// ` f_t `is specified by Alg. 1`
  Set $\tilde{B}_t = \tilde{B}_{t-1} \vee \|\tilde{\boldsymbol{g}}_t\|$ and $\hat{\boldsymbol{g}}_t = \tilde{\boldsymbol{g}}_t \cdot \tilde{B}_{t-1} / \tilde{B}_t$.
  Send linear loss $\boldsymbol{w} \mapsto \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w} \rangle$ to FreeGrad as the tth loss function.
  Set $\boldsymbol{u}_{t+1} \in \mathcal{B}(R)$ to FreeGrad's $(t+1)$th output given the history $((\boldsymbol{u}_i, \boldsymbol{w} \to \langle \tilde{\boldsymbol{g}}_i, \boldsymbol{w} \rangle))_{i \leq t}$.
  Set $Z_t = Z_{t-1} + \|\hat{\boldsymbol{g}}_t\|^2 + \tilde{B}_t^2 - \tilde{B}_{t-1}^2$ and $\boldsymbol{v}_t = \boldsymbol{v}_{t-1} + \|\hat{\boldsymbol{g}}_t\|^2 \boldsymbol{x}_t$.
  `//` $\boldsymbol{x}_t$ `is specified by Alg. 1`
end for

The underlying subroutine FreeGrad is displayed in Algorithm 9 in Appendix G. The key feature of FreeGrad that allows us to adapt to strong convexity is that its regret is bounded from above by $O(\|\boldsymbol{w}\|\sqrt{V_T})$, up to log-factors in $\|\boldsymbol{w}\|$ and V_T, where $V_T = \sum_{t=1}^T \|\boldsymbol{g}_t\|^2$ and $(\boldsymbol{g}_t)$ are the observed subgradients at the iterates of the algorithm (the precise statement of the regret bound is differed to Appendix G). Thus, similar to FTRL-prox, FreeGrad also has an adaptive regret that can be much smaller than the worst-case $O(\sqrt{T})$; e.g. when the losses are smooth (Srebro et al., 2010a). Furthermore, FreeGrad's regret bound scales with the norm of the comparator $\|\boldsymbol{w}\|$, and thus becomes small for comparators close to the origin. This property will be crucial to prove a logarithmic regret for strongly convex losses (Cutkosky and Orabona, 2018).

To simplify the analysis, we will state our result for *B-Lipschitz* losses (ℓ_t) which are those that satisfy, for all $t \geq 1$ and $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x})$:

$$\|\boldsymbol{g}_t\| \leq B. \tag{20}$$

We also recall that (ℓ_t) are *μ-strongly convex*, for $\mu > 0$, if for all $t \geq 1$, $\boldsymbol{x} \in \mathcal{C}$ and $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x})$:

$$\forall \boldsymbol{y} \in \mathcal{C}, \quad \ell_t(\boldsymbol{y}) \geq \ell_t(\boldsymbol{x}) + \langle \boldsymbol{g}_t, \boldsymbol{y} - \boldsymbol{x} \rangle + \mu \|\boldsymbol{x} - \boldsymbol{y}\|^2 / 2,$$

where $\|\cdot\|$ is the Euclidean norm. With this, we now state our main result for this subsection (the proof is in Appendix I.5):

Theorem 17 *Let $\mu, B > 0$, $\delta \in (0, 1/3)$, and $\kappa := R/r$, where r and R as in (1). Suppose that Algorithm 1 is run with $\mathcal{O}_{\mathcal{C}^\circ} \equiv \mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$ (Alg. 4); $\delta_t = \delta / t^2$, $\forall t \geq 1$; and sub-routine* A *set to Alg. 5 with parameter $\epsilon > 0$. Then, for any adversarial sequence of convex [resp. μ-strongly convex] B-Lipschitz losses (ℓ_t) on $\mathcal{C}$ the iterates $(\boldsymbol{x}_t)$ of Algorithm 1 satisfy $(\boldsymbol{x}_t) \subset \mathcal{C}$, and for all $T \geq 1$ and $\boldsymbol{x} \in \mathcal{C}$, we have,*

$$\sum_{t=1}^T \mathbb{E}[\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{x})] \leq U_T R B \sqrt{T} + R B U_T^2 / \nu, \tag{21}$$

$$\left[\textit{resp.} \sum_{t=1}^T \mathbb{E}[\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{x})] \leq \left(\frac{R}{\nu} + \frac{B}{2\mu} \right) B U_T^2 \right],$$

where $U_T = O\left(\nu d^{1/2} \ln(e + \frac{\kappa d R T B}{\epsilon \delta})\right)$; $\nu := 1/(R \wedge 1) + \kappa + \delta$; and $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t)$, $\forall t \geq 1$.

We note that the instance of Algorithm 1 in the preceding theorem automatically adapts to the strong convexity constant $\mu > 0$ of the losses (ℓ_t), where it achieves a logarithmic regret. For general convex losses, the algorithm ensures the optimal worst-case $O(\sqrt{T})$ regret up to log-factors.

Computational Complexity. The instance of Algorithm 1 in Theorem 17 makes the same number of calls to the Oracle $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$ as in the setting of Theorems 15 and 16. Thus, this instance makes at most $O(T\ln(d\kappa T/\delta))$ calls to the Membership Oracle $\mathsf{MEM}_{\mathcal{C}}$. Further, we note that the sequence (δ_t) may be set to $\delta_t = \delta/t^{-1/2}$ [resp. $\delta_t = \delta_t/t$], for all $t \geq 1$, for general convex [resp. strongly-convex] functions, allowing the Membership Oracle $\mathsf{MEM}_{\mathcal{C}}$ to be less accurate while maintaining the same regret bound as Theorem 17 up to constant factors (see discussion after Theorem 15). Finally, we restricted the losses to be B-Lipschitz in Theorem 17 only to simplify the proofs—the algorithm need not know B.

Scale-invariance. Though the instance of Alg. 1 in Theorem 17 does not require a bound on the norm of the gradients (an information typically required by other OL algorithms (Orabona and Pál, 2016)), the algorithm is not scale-invariant; multiplying the sequence of losses (ℓ_t) by some fixed constant changes the iterates $(\boldsymbol{x}_t)$ of the algorithm. In general, this is an undesirable property for OL algorithms (Orabona and Pál, 2016).

Note also that the regret bound in Theorem 17 can technically be unbounded due to the fraction B/ϵ in the expression of U_T. This fraction can be arbitrarily large if ϵ (a parameter of the algorithm) is too small relative to the Lipschitz constant B. Such a problematic ratio has appeared in previous works such as (Ross et al., 2013; Wintenberger, 2017; Kotłowski, 2017; Mhammedi et al., 2019; Kempka et al., 2019). To tame such a ratio, Mhammedi et al. (2019) and Mhammedi and Koolen (2020) presented a technique for certain OCO algorithms, such as MetaGrad (Van Erven and Koolen, 2016; Van Erven et al., 2021) and FreeGrad, based on the idea of restarting these algorithms whenever the ratio between the maximum norm of the observed subgradients and the norm of the initial subgradient is too large. In Appendix E, we extend this technique and present a general reduction (Algorithm 7) that makes a large class of Online Learning algorithms (including the instance of Alg. 1 in the setting of Theorem 17) scale-invariant and gets rid of problematic ratios in their regret bounds.

Optimality and link to stochastic optimization. The regret bounds in Theorem 17 are optimal in T. For the dependence in d in the regret bounds, the computational trade-offs, and the link to stochastic and offline optimization, see the discussion at the end of Section C.1.

C.3. Efficient Projection-Free Smooth Stochastic Optimization

So far, we have mainly considered the setting where (ℓ_s) is a sequence convex losses that may be chosen in an adversarial fashion, and the goal was to choose a sequence of iterates $(\boldsymbol{x}_t) \subset \mathcal{C}$ such that the cumulative loss $\sum_{t=1}^T \ell_t(\boldsymbol{x}_t)$ is small. In this subsection, we are interested in minimizing a fixed convex differentiable function $f : \mathcal{C} \mapsto \mathbb{R}$, with only access to a Stochastic Gradient Oracle for the function f. Formally, we assume there exists a $\sigma > 0$ such that for any round $t \geq 1$ and some query point $\boldsymbol{x}_t \in \mathcal{C}$, we have access to a subgradient $\boldsymbol{g}_s \in \partial \ell_s(\boldsymbol{x}_t)$, where

$$\ell_s(\boldsymbol{x}) := f(\boldsymbol{x}) + \langle \boldsymbol{x}, \boldsymbol{\xi}_s \rangle, \tag{22}$$

and $(\boldsymbol{\xi}_s \in \mathbb{R}^d)$ are i.i.d. random vectors satisfying

$$\mathbb{E}[\boldsymbol{\xi}_s] = \mathbf{0} \quad \text{and} \quad \mathbb{E}[\|\boldsymbol{\xi}_s\|^2] \leq \sigma^2, \quad \forall s \geq 1. \tag{23}$$

We will also assume that the function f is β-smooth; that is, f is differentiable[4] on $\mathcal{C}$ and

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}, \quad f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\beta}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2. \tag{24}$$

It will be instrumental to use the following consequence of (24) (see e.g. (Srebro et al., 2010b; Levy et al., 2018; Cutkosky and Busa-Fekete, 2018)); if f is β-smooth, then $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}_*)\|^2 \leq 2\beta \cdot (f(\boldsymbol{x}) - f(\boldsymbol{x}_*))$, where $x_* \in \arg\min_{\boldsymbol{x} \in \mathcal{C}} f(\boldsymbol{x})$. Thus, when $\boldsymbol{x}_*$ is in the interior of $\mathcal{C}$, we have $\nabla f(\boldsymbol{x}_*) = \mathbf{0}$, and so it follows that

$$\|\nabla f(\boldsymbol{x})\|^2 \leq 2\beta \cdot (f(\boldsymbol{x}) - f(\boldsymbol{x}_*)). \tag{25}$$

For the sake of clarity, we now summarize the assumptions we make on the loss process in this section:

Assumption 2 *The sequence (ℓ_s) in Algorithm 6 satisfies (22) with I) $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots \in \mathbb{R}^d$ i.i.d. vectors as in (23); and II) f is β-smooth, for $\beta > 0$, and satisfies int$(\mathcal{C}) \cap \arg\min_{\boldsymbol{x} \in \mathcal{C}} f(\boldsymbol{x}) \neq \emptyset$, where int$(\mathcal{C})$ denotes the interior of $\mathcal{C}$.*

Without loss of generality (by making R larger if necessary), we assume that there exists $R' \leq R$ such that:

$$\mathcal{B}(r) \subseteq \mathcal{C} \subseteq \mathcal{B}(R') \subseteq \mathcal{B}(R)/(1+\nu), \quad \text{where} \quad \nu \coloneqq 4\sqrt{2}(1+\kappa), \quad \text{and} \quad \kappa \coloneqq R'/r. \tag{26}$$

We now state the main result of this section (the proof is in Appendix I.6):

Theorem 18 *Let $\delta \in (0, 1/3)$, and r, R, R', ν, and κ be as in (26). Further, suppose that Alg. 1 is run with $\mathcal{O}_{\mathcal{C}^\circ} \equiv \mathsf{OPT}_{\mathcal{C}^\circ}$ (Alg. 3); $\delta_t = \delta/t^3$, $\forall t \geq 1$; and sub-routine A set to Alg. 6 with input $\varepsilon > 0$. Then, under Assumption 2, the iterates $(\boldsymbol{x}_t)$ of Alg. 6, satisfy $(\boldsymbol{x}_t) \subset \mathcal{C}$, and for all $T \geq 1$,*

$$\mathbb{E}\left[f(\boldsymbol{x}_T) - f(\boldsymbol{x}_*)\right] \leq \frac{2\nu\varepsilon R' + \nu^2\beta(R')^2 U_T + 3\delta R(\ln T + 6)\sqrt{\sigma^2 + 2\beta R'}}{T^2} + \frac{2\nu R'\sigma\sqrt{U_T}}{\sqrt{T}},$$

where $\boldsymbol{x}_* \in \arg\min_{\boldsymbol{x} \in \mathcal{C}} f(\boldsymbol{x})$ and $U_T \coloneqq \ln\left(1 + 2\varepsilon^{-2}(\sigma^2 + 2R'\beta)T^3\right)$.

The proof of Theorem 18 is based on an extension of a result due to Cutkosky (2019). The instance of Algorithm 1 in Theorem 18 invokes $\mathsf{OPT}_{\mathcal{C}^\circ}(\cdot; \delta_t)$ at each iteration t. Thus, in light of Remark 12 in App. B, the algorithm makes at most $O(dT\ln(dT\kappa/\delta)$ calls to the Membership Oracle $\mathsf{MEM}_{\mathcal{C}}$ after T rounds.

Optimality of the rate and application to the offline setting. The rate in Theorem 18 is optimal in T and implies the fast $O(\beta\kappa^2/T^2)$ rate in the offline smooth setting (i.e. $\sigma = 0$), which is also optimal in T. Thus, Algorithm 1 in the setting of Theorem 18 reaches an ϵ-sub-optimal point in offline smooth optimization after $\widetilde{O}(d\kappa/\sqrt{\epsilon})$ calls to $\mathsf{MEM}_{\mathcal{C}}$. Since state-of-the-art algorithms based on the cutting plane method require $O(d^2\ln(1/\epsilon))$ calls to a Membership Oracle to reach an ϵ-sub-optimal point (Lee et al., 2018), our algorithm provides a viable alternative to the latter whenever $d \geq \Omega(\kappa/\sqrt{\epsilon})$ and the objective function is smooth. As we shall see in Section F, κ is less

4. To avoid boundary issues, we assume (similar to (Hiriart-Urruty and Lemaréchal, 2004, Section B.4.1)) that $\mathcal{C}$ is contained in a open set Ω on which f is differentiable.

than $\sqrt{d}$ in many settings of interest. We also recall that the presence of κ in our bounds is due to an over conservative upper bound on the norms of the subgradients of the surrogate losses, and so we expect the rates of our algorithms to scale better with d in practice.

Algorithm 6 Subroutine A for Algorithm 1 for Stochastic Convex Optimization.

Require: r, R, R', ν, and κ as in (26). Input $\varepsilon > 0$. Input $(\boldsymbol{x}_t, f_t)$ from the Master Algorithm 1 at each round $t \geq 1$, where $\boldsymbol{x}_t \in \mathbb{R}^d$ and $f_t \colon \mathbb{R}^d \to \mathbb{R}$.

Initialize FTRL-prox with parameter $R' > 0$ and set $\boldsymbol{u}_1 = \mathbf{0}$ (i.e. FTRL-prox's first output).
Set $\Lambda_1 = 0$; $Z_0 = \varepsilon^2$; and $\boldsymbol{w}_1 = \mathbf{0}$.
for $t = 1, 2, \ldots$ **do**
  Output $\boldsymbol{w}_t$ and observe $\tilde{\boldsymbol{g}}_t \in \partial f_t(\boldsymbol{w}_t)$. `// ` f_t `is provided by Alg. 1`
  Set $\boldsymbol{u}_{t+1} \in \mathcal{B}(R')$ to FTRL-prox's $(t+1)^{\text{th}}$ output given the history $(\boldsymbol{u}_i, \boldsymbol{w} \mapsto i\langle \tilde{\boldsymbol{g}}_i, \boldsymbol{w}\rangle)_{i \leq t}$.
  Set $Z_t = Z_{t-1} + \Lambda_t \|\boldsymbol{g}_t\|^2$ and $\eta_t = \nu R'/\sqrt{Z_t}$.
  Set $\Lambda_{t+1} = \Lambda_t + t + 1$ and $\mu_{t+1} = (t+1)/\Lambda_{t+1}$.
  Set $\boldsymbol{w}_{t+1} = (1 - \mu_{t+1})(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t) + \mu_{t+1}\boldsymbol{u}_{t+1}$. `// ` $\boldsymbol{x}_t$ `is provided by Alg. 1`
end for

Appendix D. General Regret Reduction

In this section, we state and prove a regret bound for Algorithm 1, when A is a general OCO algorithm.

Proposition 19 *Let $\delta \in (0,1)$, $B > 0$, $T \geq 1$, and A be the sub-routine of Algorithm 1. Suppose there exists a function $\mathcal{R}^{\mathsf{A}} : \mathcal{C} \times \cup_{t=1}^{\infty} \mathbb{R}^{d \times t} \to \mathbb{R}_{\geq 0}$ such that for any adversarial sequence of convex losses (f_t) on $\mathcal{C}$, the iterates $(\boldsymbol{w}_t)$ of A in response to (f_t), and the subgradients $\boldsymbol{\nabla}_t \in \partial f_t(\boldsymbol{w}_t)$, $t \in [T]$, satisfy*

$$\sum_{t=1}^{T} \langle \boldsymbol{\nabla}_t, \boldsymbol{w}_t - \boldsymbol{x} \rangle \leq \mathcal{R}^{\mathsf{A}}(\boldsymbol{x}, \boldsymbol{\nabla}_{1:T}), \forall \boldsymbol{x} \in \mathcal{C},$$

as long as $\|\boldsymbol{\nabla}_t\| \leq (1 + \frac{\delta}{t} + \kappa)B$, for all $t \in [T]$. Then, for any adversarial sequence of B-Lipschitz convex losses (ℓ_t), the iterates $(\boldsymbol{x}_t)$ of Alg. 1 with tolerance sequence $\delta_t := \frac{\delta}{t^3}$, and the subgradients $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t)$, satisfy

$$\mathbb{P}\left[\begin{array}{c} \forall \boldsymbol{x} \in \mathcal{C},\ \sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \leq \mathcal{R}^{\mathsf{A}}(\boldsymbol{x}, \tilde{\boldsymbol{g}}_{1:T}) + \left(4 + \frac{5\delta \ln T}{\rho}\right) RB \\ \text{and} \quad \forall t \in [T],\ \|\tilde{\boldsymbol{g}}_t\| \leq (1 + \delta/t + \kappa)B \end{array} \right] \geq 1 - 2\rho,$$

for all $\rho \in (0,1)$ and $(\tilde{\boldsymbol{g}}_t)$ as in Algorithm 1.

Proof By Lemma 7, we have, for all $\boldsymbol{w} \in \mathcal{C}$ and $t \geq 1$,

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \leq \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x} \rangle + (2\delta_t + \Delta_t) R \|\boldsymbol{g}_t\|, \tag{27}$$

where $\Delta_t \geq 0$ is a non-negative random variable satisfying $\mathbb{E}_{t-1}[\Delta_t] \leq \delta_t$. Thus, by the law of total expectation and Markov's inequality, we have

$$\mathbb{P}[\Delta_t \geq t^2 \delta_t / \rho] \leq \rho / t^2.$$

Let $\mathcal{E}_T$ be the event that $\{\Delta_t \le t^2\delta_t/\rho, \forall t \in [T]\}$. By a union bound and the fact that $\sum_{t=1}^{\infty} 1/t^2 \le 2$, we get that $\mathbb{P}[\mathcal{E}_T] \ge 1 - 2\rho$. For the rest of this proof, we will condition on the event $\mathcal{E}_T$. We have, by our choice of (δ_t),

$$\sum_{t=1}^{T} \Delta_t \le \sum_{t=1}^{T} t^2\delta_t/\rho = \sum_{t=1}^{T} \delta/(\rho t) \le \frac{\delta \ln T}{\rho}. \tag{28}$$

Also, note that by Lemma 7 (and the fact that $\mathcal{E}_T$ holds), we have,

$$\|\tilde{\boldsymbol{g}}_t\| \le (1 + \Delta_t + \kappa)\|\boldsymbol{g}_t\| \le (1 + \delta/t + \kappa)\|\boldsymbol{g}_t\| \le (1 + \delta/t + \kappa)B, \tag{29}$$

where the last inequality follows by the fact that ℓ_t is B-Lipschitz. Now, by summing (27) for $t = 1, \dots, T$, we obtain, for all $\boldsymbol{x} \in \mathcal{C}$

$$\begin{aligned}
\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle &\le \sum_{t=1}^{T} \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle + \sum_{t=1}^{T} (2\delta_t + \Delta_t)R\|\boldsymbol{g}_t\|, \\
&\le \mathcal{R}^{\mathsf{A}}(\boldsymbol{x}, \tilde{\boldsymbol{g}}_{1:t}) + \sum_{t=1}^{T} (2\delta_t + \Delta_t)R\|\boldsymbol{g}_t\|, \qquad (30) \\
&\le \mathcal{R}^{\mathsf{A}}(\boldsymbol{x}, \tilde{\boldsymbol{g}}_{1:t}) + (4 + 5\delta/\rho \ln(T))RB,
\end{aligned}$$

where (30) follows by (29) and the assumption made about the regret of Algorithm A, and the last inequality follows by (28) and the fact that ℓ_t is B-Lipschitz. ■

Appendix E. Algorithm Wrapper for Scale-Invariance

In this appendix, we extend the technique of Mhammedi et al. (2019); Mhammedi and Koolen (2020) by presenting a general reduction (Algorithm 7) that makes a large class of Online Learning algorithms scale-invariant and gets rid of problematic ratios in their regret bounds. In particular, we will consider all Online Learning algorithms whose regret bound can be expressed as a map $\mathcal{R} : \mathcal{C} \times \cup_{t\ge 1}\mathbb{R}^{d\times t} \times \mathbb{R} \to \mathbb{R}_{\ge 0}$ such that for any comparator $\boldsymbol{x}$ and sequence of observed subgradients $(\boldsymbol{\nabla}_t)$, the regret of the algorithm after $t \ge 1$ rounds is bounded from above by $\mathcal{R}(\boldsymbol{x}, \boldsymbol{\nabla}_{1:T}, L_t/\epsilon)$, where $\epsilon > 0$ is a parameter of the algorithm and $L_t := \epsilon \vee \max_{t\in[T]} \|\boldsymbol{\nabla}_t\|$. We note that so far we have not restricted the class of eligible algorithms by much. We will further require the following monotonicity property, which is satisfied by most popular Online Learning algorithms:

$$\mathcal{R}(\boldsymbol{u}, \boldsymbol{g}_{1:s}, p) \le \mathcal{R}(\boldsymbol{u}, \boldsymbol{g}_{1:t}, q), \quad \text{for all } p \le q, s \le t, \boldsymbol{u} \in \mathcal{C} \text{ and } (\boldsymbol{g}_t) \subset \mathbb{R}^d. \tag{31}$$

With this, we now state our reduction result for scale-invariance:

Proposition 20 *Let* A *be the sub-routine of Algorithm 7. Suppose there exists* $\mathcal{R}^{\mathsf{A}}$ *satisfying* (31) *and such that for any adversarial sequence of convex losses* (f_t) *on* $\mathcal{C}$*, the iterates* $(\boldsymbol{x}_t)$ *of* A *in response to* (f_t) *satisfy*

$$\boldsymbol{x}_1 = \boldsymbol{0} \quad \textit{and} \quad \sum_{s=1}^{t} \langle \boldsymbol{\nabla}_s, \boldsymbol{x}_s - \boldsymbol{x}\rangle \le \mathcal{R}^{\mathsf{A}}(\boldsymbol{x}, \boldsymbol{\nabla}_{1:t}, L_t/\epsilon), \quad \forall t \ge 1, \forall \boldsymbol{x} \in \mathcal{C}, \tag{32}$$

Algorithm 7 Scale-Invariant Wrapper via Restarts.

Require: OCO Algorithm A on $\mathcal{C}$ taking parameter $\epsilon > 0$.

Initialize $\tau = 1$, $\boldsymbol{x}_1^\tau = \mathbf{0}$, and $S_0 = B_0 = 0$.
for $t = 1, 2, \dots$ **do**
  Play $\boldsymbol{y}_t = \boldsymbol{x}_t^\tau$ and observe $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{y}_t)$.
  Set $B_t = B_{t-1} \vee \|\boldsymbol{g}_t\|$ and $S_t = S_{t-1} + \|\boldsymbol{g}_t\|/B_t$. `// With the convention that 0/0 = 0`
  if $B_t = 0$ **then**
    Set $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t$.
    Continue. `// play the zero vector until the first non-zero subgradients`
  else if $B_{t-1} = 0$ **then**
    Set $\tau = t$. `//` $(B_t/B_\tau \geq S_t) \equiv (B_t/B_\tau \geq \sum_{s=\tau}^t \|\boldsymbol{g}_s\|/B_s)$
    Initialize A with parameter $\epsilon = B_\tau$. `//` τ `is the new 'round 1'`
  end if
  Send linear loss $\boldsymbol{w} \mapsto \langle \boldsymbol{g}_t, \boldsymbol{w} \rangle$ to A.
  Set $\boldsymbol{x}_{t+1}^\tau$ to A's $(t+2-\tau)$th output given history $((\boldsymbol{x}_i^\tau, \boldsymbol{w} \mapsto \langle \boldsymbol{g}_i, \boldsymbol{w} \rangle))_{\tau \leq i \leq t}$.
end for

where $\epsilon > 0$ is a parameter of A; $\boldsymbol{\nabla}_s \in \partial f_s(\boldsymbol{x}_s)$, for all $s \in [t]$; and $L_t = \epsilon \vee \max_{s \in [t]} \|\boldsymbol{\nabla}_s\|$. Then, for any adversarial sequence of convex losses (ℓ_t), the iterates $(\boldsymbol{y}_t)$ of Algorithm 7 in response to (ℓ_t) satisfy

$$\sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{y}_t - \boldsymbol{x} \rangle \leq 2\mathcal{R}^{\mathsf{A}}(\boldsymbol{x}, \boldsymbol{g}_{1:T}, T) + 4RB_T, \quad \forall T \geq 1, \forall \boldsymbol{x} \in \mathcal{C},$$

where $\boldsymbol{g}_t$, $t \in [T]$, is any subgradients in $\partial \ell_t(\boldsymbol{y}_t)$ and $B_T = \epsilon \vee \max_{t \in [T]} \|\boldsymbol{g}_t\|$.

Proof of Proposition 20. We say that $\tau = t$ is the start of an epoch of Algorithm 7 if the condition in Line 8 is satisfied. We use the convention that the "last epoch" starts at $t = T + 1$. Let $\tau < t$ be the start of two consecutive epochs. Then, by the condition on Line 8 of Algorithm 7, we have

$$B_{t-1}/B_\tau \leq \sum_{s=1}^{t-1} \|\boldsymbol{g}_s\|/B_s \leq t - 1, \quad \text{and} \quad B_t/B_\tau > \sum_{s=1}^{t} \|\boldsymbol{g}_s\|/B_s. \tag{33}$$

Recall also that at round $s = \tau$, sub-routine A is initialized with $\epsilon = B_\tau$. Thus, by our assumption in (32), we have $\boldsymbol{x}_\tau^\tau = \mathbf{0}$ and

$$\begin{aligned} \forall \boldsymbol{x} \in \mathcal{C}, \quad \sum_{s=\tau}^{t-1} \langle \boldsymbol{g}_t, \boldsymbol{y}_t - \boldsymbol{x} \rangle = \langle \boldsymbol{y}_\tau, \boldsymbol{g}_\tau \rangle + \sum_{s=\tau}^{t-1} \langle \boldsymbol{g}_t, \boldsymbol{x}_t^\tau - \boldsymbol{x} \rangle &\overset{(32)}{\leq} \langle \boldsymbol{y}_\tau, \boldsymbol{g}_\tau \rangle + \mathcal{R}^{\mathsf{A}}(\boldsymbol{x}, \boldsymbol{g}_{\tau:t-1}, B_{t-1}/\epsilon), \\ &\leq RB_t + \mathcal{R}^{\mathsf{A}}(\boldsymbol{x}, \boldsymbol{g}_{\tau:t-1}, t-1), \\ &\leq RB_T + \mathcal{R}^{\mathsf{A}}(\boldsymbol{x}, \boldsymbol{g}_{1:T}, T), \end{aligned} \tag{34}$$

where the last inequality follows by the fact that $\mathcal{R}^{\mathsf{A}}$ satisfies (31). If there are two epochs or less, summing (34) across the epochs leads to the desired result. Now suppose that there are more than two epochs, and let τ [resp. t] be the start of the ante-penultimate [resp. penultimate] epoch (recall

that the last epoch starts at $T+1$ by convention). By (34), the regret across these two epochs is bounded as

$$\sum_{s=\tau}^{T}\langle \boldsymbol{g}_t, \boldsymbol{y}_t - \boldsymbol{x}\rangle \le 2RB_T + 2\mathcal{R}^{\mathsf{A}}(\boldsymbol{x}, \boldsymbol{g}_{1:T}, T). \tag{35}$$

We will now bound the regret across the earlier epochs. We have

$$\sum_{s=1}^{\tau-1}\langle \boldsymbol{g}_t, \boldsymbol{y}_t - \boldsymbol{x}\rangle \le 2R\sum_{s=1}^{\tau-1}\|\boldsymbol{g}_s\| \le 2RB_\tau \sum_{s=1}^{\tau-1}\frac{\|\boldsymbol{g}_s\|}{B_s} \le 2RB_\tau \sum_{s=1}^{t}\frac{\|\boldsymbol{g}_s\|}{B_s} \overset{(33)}{<} 2RB_t \le 2RB_T. \tag{36}$$

Combining (35) and (36) leads to the desired result. ■

Appendix F. Applying the Projection-Free Reduction in Practice

In this section, we consider various popular settings where our algorithm may be applied. We note that Assumption 1, i.e. the condition that $\mathcal{B}(r) \subseteq \mathcal{C} \subseteq \mathcal{B}(R)$ may not always be satisfied in these settings, but one can easily reparametrize the problem to satisfy the condition. We will derive explicit reparametrizations for the popular settings studied in (Hazan and Kale, 2012; Jaggi, 2013) (those in Table 3). But, first we will present a general reparametrization recipe.

Suppose that the available losses (f_t) are defined on a convex set $\mathcal{K} \subset \mathbb{R}^d$ that does not necessarily satisfy (1) (e.g. if $\mathcal{K}$ has an empty interior like the simplex). Further, suppose that we have a Membership Oracle $\mathcal{M}_{\mathcal{K}}$ for $\mathcal{K}$ and a subgradient Oracle for (f_t). We will show that one can easily reparametrize the problem on a set $\mathcal{C}$ satisfying (1) and whose Membership Oracle can easily be constructed from that of $\mathcal{K}$.

Let $\mathcal{H}$ be the subspace generated by the span of $\mathcal{K}$; that is, $\mathcal{H} \coloneqq \{\lambda \boldsymbol{x} : \lambda \in \mathbb{R} \text{ and } \boldsymbol{x} \in \operatorname{conv}\mathcal{K}\}$. Further, let $(\boldsymbol{u}_i)_{1\le i\le m}$, $m \le d$, be an orthogonal basis of $\mathcal{H}$ (this can be computed offline once for the given problem), and let $\boldsymbol{c}$ be a point in the relative interior of $\mathcal{K}$. Then, one can work with the surrogate losses $\ell_t : \mathbb{R}^m \to \mathbb{R}\cup\{+\infty\}$ defined by

$$\ell_t(\boldsymbol{x}) = \begin{cases} f_t\left(\boldsymbol{c} + \sum_{i=1}^m x_i \boldsymbol{u}_i\right), & \text{if } \boldsymbol{c} + \sum_{i=1}^m x_i \boldsymbol{u}_i \in \mathcal{K}; \\ +\infty, & \text{otherwise.} \end{cases}$$

The convexity of ℓ_t follows immediately from that of f_t. We also note that since $\boldsymbol{c}$ was chosen in the relative interior of $\mathcal{K}$, there exists $r, R > 0$ such that the domain $\mathcal{C} \subset \mathbb{R}^m$ of ℓ_t satisfies $\mathcal{B}(r) \subseteq \mathcal{C} \subseteq \mathcal{B}(R)$; that is, Assumption 1 is satisfied. We now show that a Membership Oracle for $\mathcal{C}$ [resp. subgradient Oracle for ℓ_t] can easily be constructed from a Membership Oracle for $\mathcal{K}$ [resp. subgradient Oracle for f_t].

Starting with the Membership Oracle: for any $\boldsymbol{x} \in \mathcal{C}$, a Membership Oracle $\mathcal{M}_{\mathcal{C}}$ for $\mathcal{C}$ can be implemented as $\mathcal{M}_{\mathcal{C}}(\boldsymbol{x}) = \mathcal{M}_{\mathcal{K}}(\boldsymbol{c} + \sum_{i=1}^m \boldsymbol{u}_i x_i)$. Now, given a subgradient Oracle for f_t, it is also easy to get a subgradient Oracle for ℓ_t; for any $\boldsymbol{x} \in \mathbb{R}^m$ such that $\boldsymbol{y} \coloneqq \sum_{i=1}^m \boldsymbol{u}_i x_i \in \mathcal{K}$, we have $\partial \ell_t(\boldsymbol{x}) = \{\sum_{i=1}^m \langle \boldsymbol{u}_i, \boldsymbol{\zeta}\rangle \boldsymbol{u}_i : \boldsymbol{\zeta} \in \partial f_t(\boldsymbol{y})\}$. Using this method, one needs to perform $O(md)$ arithmetic operations at each round to compute a subgradient of ℓ_t. It is possible to avoid this computational overhead by, instead, using a finite difference approach for estimating subgradients of

Domain $\mathcal{C}$	Operation Required by LOO $\mathcal{O}_{\mathcal{C}}(\boldsymbol{x};\cdot)$	MO $\mathcal{M}_{\mathcal{C}}(\boldsymbol{x};\cdot)$	Computational Complexity of LOO $\mathcal{O}_{\mathcal{C}}(\cdot;\delta)$	MO $\mathcal{M}_{\mathcal{C}}(\cdot;\delta)$
ℓ_p-ball in $\mathbb{R}^d$	$\|\boldsymbol{x}\|_q$	$\|\boldsymbol{x}\|_p$	$O(d)$	$O(d)$
Simplex $\Delta_d \in \mathbb{R}^d$	$\|\boldsymbol{x}\|_\infty$	$\langle \mathbf{1}, \boldsymbol{x}\rangle$	$O(d)$	$O(d)$.
Trace-norm-ball in $\mathbb{R}^{m\times n}$	$\|\boldsymbol{x}\|_{\text{op}}$	$\|\boldsymbol{x}\|_{\text{tr}}$	$O(\text{nnz}(\boldsymbol{x})/\sqrt{\delta})$	Cost(SVD)
Op-norm-ball in $\mathbb{R}^{m\times n}$	$\|\boldsymbol{x}\|_{\text{tr}}$	$\|\boldsymbol{x}\|_{\text{op}}$	Cost(SVD)	$O(\text{nnz}(\boldsymbol{x})/\sqrt{\delta})$
Conv-hull of Permutation Matrices in $\mathbb{R}^{n\times n}$		$\boldsymbol{e}_i^\top \boldsymbol{x}\mathbf{1} = 1;\ \forall i$ $\boldsymbol{e}_i^\top \boldsymbol{x}^\top \mathbf{1} = 1,$	$O(n^3)$	$O(n^2)$
Convex-hull of Rotation Matrices in $\mathbb{R}^{n\times n}$			Cost(SVD)	Cost(SVD)
PSD matrices in $\mathbb{R}^{n\times n}$ with unit trace	$\lambda_{\max}(\boldsymbol{x})$	$\lambda_{\min}(\boldsymbol{x})$ $\text{tr}(\boldsymbol{x})$	$O(\text{nnz}(\boldsymbol{x})/\sqrt{\delta})$	$O(\text{nnx}(\boldsymbol{x})/\sqrt{\delta})$
PSD matrices in $\mathbb{R}^{n\times n}$ with diagonals ≤ 1		$\lambda_{\min}(\boldsymbol{x})$ $\max_{i\in[n]}(x_{ii})$	$O(\text{nnz}(\boldsymbol{x})\sqrt{n^3/\delta^5})$	$O(\text{nnx}(\boldsymbol{x})/\sqrt{\delta})$
The flow polytope with (#nodes, #edges)=(d,m)			$\widetilde{O}(d+m)$	$O(d+m)$
The Matroid polytope for Matroid M; #elem.=d			$\widetilde{O}(d\text{Cost}(\mathcal{I}_M))$	$O(d^2\text{Cost}(\mathcal{I}_M)$ $+d^3)$

Table 3: Computational complexity of performing linear optimization [resp. testing membership] for different sets of interest. $\mathcal{O}_{\mathcal{C}}$ [resp. $\mathcal{M}_{\mathcal{C}}$] denotes a Linear Optimization Oracle (LOO) [resp. Membership Oracle (MO)]. $\delta > 0$ represents the allowed Oracle error (see Section 2). We hide any logarithmic dependence in $1/\delta$. Cost(SVD) [resp. Cost($\mathcal{I}_M$)] represents the computational cost of performing SVD [resp. testing if a set is independent in M (see Section F)]. For $\boldsymbol{x} \in \mathbb{R}^{n\times m}$, nnz($\boldsymbol{x}$) represents the number of non-zeros of $\boldsymbol{x}$. The details for the computational cost of the Linear Optimization Oracles listed can be found in (Hazan and Kale, 2012; Jaggi, 2013). The details for the Membership Oracle complexities can be found in Section F.

ℓ_t, requiring only $2m$ calls to a value Oracle for f_t per round. Lemma 29 provides the means for doing this[5].

We now show that in many popular settings there are natural parametrizations that do not require any expensive pre-processing step such as identifying a basis for the span of $\mathcal{H}$. In Table 4, we summarize the upper bounds we derive on the asphericity κ for different sets of interest after reparametrization. In Table 3, we summarize the computational complexity of a Membership Oracle for these sets.

F.1. ℓ_p-Norm Balls

Consider the setting where the losses are defined on $\mathcal{C} = \{\boldsymbol{x} : \|\boldsymbol{x}\|_p \leq 1\}$. In this case, the span of $\mathcal{C}$ is $\mathbb{R}^d$ and we can pick $\boldsymbol{c} = \mathbf{0}$. For $1 \leq p \leq 2$ Assumption 1 is satisfied with $r = d^{1/2-1/p}$ and $R = 1$; this follows by the fact that the ℓ_p norm satisfies $\|\boldsymbol{x}\| \leq \|\boldsymbol{x}\|_p \leq d^{1/p-1/2}\|\boldsymbol{x}\|, \forall \boldsymbol{x} \in \mathbb{R}^d$. When $2 \leq p$ Assumption 1 is satisfied with $r = 1$ and $R = d^{1/2-1/p}$. Thus, in either case, the asphericity is

$$\kappa = R/r \leq d^{|1/p-1/2|},$$

5. Technically, Lemma 29 provides a way of approximating the subgradients of a function whose domain is unconstrained. However, one can extend the result to the constrained case by a careful treatment of the region around the boundary of the set.

Domain	Upper bound on κ post-reparametrization	
ℓ_p-ball in $\mathbb{R}^d$	$d^{\lvert 1/p-1/2\rvert}$	$O(d^{\lvert 1/p-1/2\rvert})$, where $\mathcal{C} \subset \mathbb{R}^d$
Simplex $\Delta_d \in \mathbb{R}^d$	$2d$	$O(d)$, where $\mathcal{C} \subset \mathbb{R}^d$
Trace-norm-ball in $\mathbb{R}^{m\times n}$	$\sqrt{m \wedge n}$	$O(d^{1/4})$, where $\mathcal{C} \subset \mathbb{R}^d$
Op-norm-ball in $\mathbb{R}^{m\times n}$	$\sqrt{m \wedge n}$	$O(d^{1/4})$, where $\mathcal{C} \subset \mathbb{R}^d$
Conv-hull of Permutation Matrices in $\mathbb{R}^{n\times n}$	$\sqrt{5}n$	$O(d^{1/2})$, where $\mathcal{C} \subset \mathbb{R}^d$
Convex-hull of Rotation Matrices in $\mathbb{R}^{n\times n}$	$2n^{1/2}$	$O(d^{1/4})$, where $\mathcal{C} \subset \mathbb{R}^d$
PSD matrices in $\mathbb{R}^{n\times n}$ with unit trace	$8n^2$	$O(d)$, where $\mathcal{C} \subset \mathbb{R}^d$
PSD matrices in $\mathbb{R}^{n\times n}$ with diagonals ≤ 1	$4n^{3/2}$	$O(d^{3/4})$, where $\mathcal{C} \subset \mathbb{R}^d$
The flow polytope with (#nodes, #edges)=(d,m)		*Problem dependent*
The matroid polytope for matroid M; #elem.=d		*Problem dependent*

Table 4: Upper bounds on the asphericity κ for the different settings of Table 3 after reparametrization.

and there is no need to reparametrize. The operation needed for the Membership Oracle is simply computing $\|\boldsymbol{x}\|_p$, whereas linear optimization on $\mathcal{C}^\circ$ amounts to evaluation the dual norm $\|\boldsymbol{x}\|_q$, where $1/q + 1/p = 1$ (Jaggi, 2013).

F.2. Simplex Δ_d

Let $d > 1$ and consider the setting where the losses (f_t) are defined on the simplex $\Delta_d \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d_{\geq 0} : \mathbf{1}^\top \boldsymbol{x} = 1\}$. Since Δ_d has an empty interior, it does not satisfy Assumption 1. However, we can easily reparametrize the problem to ensure Assumption 1.

Let $\boldsymbol{c} \coloneqq (\frac{1}{2} + \frac{1}{2d})\boldsymbol{e}_d + \sum_{i=1}^{d-1} \frac{1}{2d}\boldsymbol{e}_i$ and consider the sequence of reparametrized losses (ℓ_t) given by

$$\ell_t(\boldsymbol{x}) \coloneqq f_t\left(\boldsymbol{c} + J^\top \boldsymbol{x}\right), \quad t \geq 1, \quad \text{where} \quad J \coloneqq \begin{bmatrix} I_{d-1} & -\mathbf{1} \end{bmatrix} \in \mathbb{R}^{d-1\times d}. \tag{37}$$

For any t, the function ℓ_t is convex and defined on the set

$$\mathcal{C} \coloneqq \left\{ \boldsymbol{x} \in \mathbb{R}^{d-1} : x_i \geq -\frac{1}{2d} \text{ and } \mathbf{1}^\top \boldsymbol{x} \leq \frac{1}{2} + \frac{1}{2d} \right\}. \tag{38}$$

We now show that for this reparametrized setting, Assumption 1 holds with $r = 1/(2d)$ and $R = 1$, which implies a asphericity of $\kappa = R/r \leq 2d$:

Proposition 21 *The set $\mathcal{C}$ in (38) satisfies $\mathcal{B}(1/(2d)) \subset \mathcal{C} \subset \mathcal{B}(1)$.*

Proof Note that the vertices of the set $\mathcal{C}$ are $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$, where

$$\forall i \in [d-1], \quad \boldsymbol{v}_i = \boldsymbol{e}_i - \sum_{j\in[d-1]} \frac{\boldsymbol{e}_j}{2d}, \quad \text{and} \quad \boldsymbol{v}_d = -\sum_{j\in[d-1]} \frac{\boldsymbol{e}_j}{2d}.$$

Since $\boldsymbol{v}_i \in \mathcal{B}(1)$ for all $i \in [d]$, we get that $\mathcal{C} \subseteq \mathcal{B}(1)$. We now show that $\mathcal{B}(1/(2d)) \subseteq \mathcal{C}$. For this, we need to find the point $\boldsymbol{u}$ in the boundary of $\mathcal{C}$ that is closest to the origin. This point must be the

orthogonal projection of the origin onto one of the $(d-2)$-dimensional faces of $\mathcal{C}$; there are d-many such faces corresponding to one of the inequalities defining $\mathcal{C}$ being satisfied with equality. Thus, $\boldsymbol{u}$ must satisfy one of the following:

- $\boldsymbol{u} = \left(\frac{1}{d-1} - \frac{1}{2d}\right) \sum_{i\in[d-1]} \boldsymbol{e}_i$.
- $\exists i \in [d-1]$, such that $\boldsymbol{u} = -\frac{\boldsymbol{e}_j}{2d}$.

In all cases, we have $\|\boldsymbol{u}\| \geq 1/(2d)$, and so this shows that $\mathcal{B}(1/(2d)) \subseteq \mathcal{C}$. ■

It is clear from the definition of the set $\mathcal{C}$ that the operations required to test the membership of a point $\boldsymbol{x} \in \mathbb{R}^{d-1}$ are I) computing $\langle \mathbf{1}, \boldsymbol{x}\rangle$; and II) evaluating x_i, for $i \in [d-1]$. Therefore, the corresponding computational complexity is $O(d)$. We now show how to build a subgradient Oracle for the reparametrized losses (ℓ_t). By the chain-rule, $\boldsymbol{g}$ is a subgradient of ℓ_t at $\boldsymbol{x}$ if and only if $\boldsymbol{g} = J\boldsymbol{\zeta}$, for $\boldsymbol{\zeta} \in \partial f_t(\boldsymbol{c} + J^\top \boldsymbol{x})$. Here, $J\boldsymbol{\zeta}$ can be evaluated in $O(d)$.

F.3. Trace and Operator Norm Balls

Trace norm ball. Let $m, n \geq 1$, $s \coloneqq m \wedge n$, and consider the setting where the losses are defined on the trace-norm ball $\mathcal{C} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^{m\times n} : \sum_{i=1}^s \sigma_i(\boldsymbol{x}) \leq 1\}$, where $\sigma_1(\boldsymbol{x}) \geq \cdots \geq \sigma_s(\boldsymbol{x})$ are the singular values of $\boldsymbol{x}$ in non-increasing order. Implementing a Membership Oracle $\mathcal{M}_{\mathcal{C}}$ for $\mathcal{C}$ requires computing the sum of singular values of a matrix $\boldsymbol{x}$, and so the computational complexity of $\mathcal{M}_{\mathcal{C}}$ is at most that of performing SVD. Since the trace-norm $\|\cdot\|_{\text{tr}}$ satisfies

$$\frac{1}{\sqrt{s}}\|\boldsymbol{x}\|_{\text{tr}} \leq \sqrt{\sum_{i=1}^s \sigma_i(\boldsymbol{x})^2} \leq \|\boldsymbol{x}\|_{\text{tr}},$$

and $\sqrt{\sum_{i=1}^s \sigma_i(\boldsymbol{x})^2} = \sqrt{\text{tr}(\boldsymbol{x}\boldsymbol{x}^\top)} = \|\boldsymbol{x}\|_{\text{F}}$ is just the Euclidean norm on $\mathbb{R}^{m\times n}$ ($\|\cdot\|_{\text{F}}$ denotes the Frobenius norm), we have that Assumption 1 is satisfied with $r = 1$ and $R = \sqrt{m \wedge n}$, and so the asphericity is $\kappa = \sqrt{m \wedge n}$.

Operator norm Ball. Let $m, n \geq 1$, $s \coloneqq m \wedge n$, and consider the setting where the losses are defined on the operator-norm ball $\mathcal{C} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^{m\times n} : \sigma_1(\boldsymbol{x}) \leq 1\}$, where $\sigma_1(\boldsymbol{x}) \geq \cdots \geq \sigma_s(\boldsymbol{x})$ are the singular values of $\boldsymbol{x}$. The operator norm $\|\cdot\|_{\text{op}}$ is the dual to the trace-norm $\|\cdot\|_{\text{tr}}$. Since $\|\boldsymbol{x}\|_{\text{op}}$ is the largest singular value of $\boldsymbol{x}$, we have $\|\boldsymbol{x}\|_{\text{op}} \leq \|\boldsymbol{x}\|_{\text{F}} \leq \sqrt{s}\|\boldsymbol{x}\|_{\text{op}}$, and so Assumption 1 is satisfied with $r = (m \wedge n)^{-1/2}$ and $R = 1$. Implementing a Membership Oracle for $\mathcal{C}$ requires computing the largest singular value of a given matrix $\boldsymbol{x}$. It is possible to approximate the largest singular value up to error δ using $O(\text{nnz}(\boldsymbol{x})/\sqrt{\delta})$ arithmetic operations, where $\text{nnz}(\boldsymbol{x})$ represents the number of non-zeros of $\boldsymbol{x}$ (Jaggi, 2013, Proposition 8). That is, the complexity of implementing a δ-approximate Membership Oracle $\mathcal{M}_{\mathcal{C}}(\cdot;\delta)$ (see Definition 2) is $O(\text{nnz}(\boldsymbol{x})/\sqrt{\delta})$.

F.4. Convex-hull of Permutation Matrices

We now consider the setting where the losses (f_t) are defined on the convex-hull of permutation matrices, also known as the Birkhoff polytope $\mathcal{K} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^{n\times n}_{\geq 0} : \boldsymbol{e}_i^\top \boldsymbol{x}\mathbf{1} = \boldsymbol{e}_i^\top \boldsymbol{x}^\top \mathbf{1} = 1\}$. Assumption 1 is not satisfies since the *Birkhoff polytope* has an empty interior. However, as we did in

the case of the simplex, we can easily reparametrize the problem to satisfy Assumption 1. We assume that $n > 1$. Before presenting the reparametrized losses, we first introduce some notation. Let

$$\boldsymbol{c} \coloneqq \sum_{1\le i\le n}\left(\frac{1}{2}+\frac{1}{2n}\right)\boldsymbol{e}_{in} + \sum_{1\le j\le n}\left(\frac{1}{2}+\frac{1}{2n}\right)\boldsymbol{e}_{nj} + \sum_{1\le i,j\le n}\frac{\boldsymbol{e}_{ij}}{2n},$$

and for any $\boldsymbol{x} \in \mathbb{R}^{n-1\times n-1}$ define the matrix $\bar{\boldsymbol{x}} \in \mathbb{R}^{n\times n}$ such that

$$\bar{x}_{ij} \coloneqq \begin{cases} x_{ij}, & \forall i,j \in [n-1]; \\ 0, & \text{otherwise.} \end{cases} \tag{39}$$

With this, consider the sequence of reparametrized losses (ℓ_t) given by

$$\begin{aligned} \ell_t(\boldsymbol{x}) &\coloneqq f_t(\boldsymbol{c} + M\bar{\boldsymbol{x}} + \bar{\boldsymbol{x}}M^\top), \ \ \forall \boldsymbol{x} \in \mathbb{R}^{n-1\times n-1}, \\ \text{where} \quad M &\coloneqq \begin{bmatrix} J^\top & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{n\times n} \ \text{ and } \ J \text{ as in (37) with } d = n. \end{aligned} \tag{40}$$

For any t, the function ℓ_t is convex and defined on the set

$$\mathcal{C} \coloneqq \left\{ \boldsymbol{x} \in \mathbb{R}^{n-1\times n-1} : \forall i,j \in [n-1], \ \ x_{ij} \ge -\frac{1}{2n}; \ \sum_{k=1}^{n-1} x_{ik} \le \frac{1}{2}+\frac{1}{2n}; \ \text{ and } \ \sum_{k=1}^{n-1} x_{kj} \le \frac{1}{2}+\frac{1}{2n} \right\}.$$

We now show that for this reparametrized setting, Assumption 1 holds with $r = 1/(2n)$ and $R = \sqrt{5}/2$. Note that the vertices of the set $\mathcal{C}$ are $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{(n-1)^2+1}$, where

$$\forall i \in [(n-1)^2], \quad \boldsymbol{v}_i = \tilde{\boldsymbol{e}}_i - \sum_{j\in[(n-1)^2]} \frac{\tilde{\boldsymbol{e}}_j}{2n}, \quad \text{and} \quad \boldsymbol{v}_{(n-1)^2+1} = -\sum_{j\in[(n-1)^2]} \frac{\tilde{\boldsymbol{e}}_j}{2n},$$

where $\tilde{\boldsymbol{e}}_i = \boldsymbol{e}_{pq}$ and p, q are the unique integers satisfying $i = p + (n-1)(q-1)$ and $p, q \in [n-1]$. Since $\boldsymbol{v}_i \in \mathcal{B}(\sqrt{5}/2)$, for all $i \in (n-1)^2 + 1$, we get that $\mathcal{C} \subseteq \mathcal{B}(\sqrt{5}/2)$. We now show that $\mathcal{B}(1/(2n)) \subseteq \mathcal{C}$. For this, we need to find the point $\boldsymbol{u}$ in the boundary of $\mathcal{C}$ that is closest to the origin. This point must be the orthogonal projection of the origin onto one of the $((n-1)^2-1)$-dimensional faces of $\mathcal{C}$; there are (n^2-1)-many such faces corresponding to one of the inequalities defining $\mathcal{C}$ being satisfied with equality. Thus, $\boldsymbol{u}$ must satisfy one of the following:

- $\exists i \in [n-1]$, such that $\boldsymbol{u} = \left(\frac{1}{n-1} - \frac{1}{2n}\right)\sum_{j\in[n-1]} \boldsymbol{e}_{ij}$.
- $\exists j \in [n-1]$, such that $\boldsymbol{u} = \left(\frac{1}{n-1} - \frac{1}{2n}\right)\sum_{i\in[n-1]} \boldsymbol{e}_{ij}$.
- $\exists i \in [(n-1)^2]$, such that $\boldsymbol{u} = \frac{-\tilde{\boldsymbol{e}}_i}{2n}$.

In all cases, we have $\|\boldsymbol{u}\| \ge 1/(2n)$, and so this shows that $\mathcal{B}(1/(2n)) \subseteq \mathcal{C}$. Thus, the asphericity for this reparametrized setting is

$$\kappa = R/r \le \sqrt{5}n.$$

It is clear from the definition of the set $\mathcal{C}$ that the operations required to test membership for a point $\boldsymbol{x} \in \mathbb{R}^{n\times n}$ are I) computing $\boldsymbol{e}_i^\top \boldsymbol{x}\mathbf{1}$ and $\boldsymbol{e}_i^\top \boldsymbol{x}^\top \mathbf{1}$ for $i, j \in [n]$; and II) evaluating x_{ij}, for $i, j \in [n]$. Thus the computational complexity of testing membership in $\mathcal{C}$ is $O(n^2)$. We now show how to build

a subgradient Oracle for the reparametrized losses (ℓ_t). By the chain-rule, $\boldsymbol{g}$ is a subgradient of ℓ_t at $\boldsymbol{x}$, if and only if, for all $i, j \in [n-1]$,

$$g_{ij} = 2\zeta_{ij} - \zeta_{nj} - \zeta_{in}, \quad \text{for} \quad \boldsymbol{\zeta} \in \partial f_t(\boldsymbol{c} + M\bar{\boldsymbol{x}} + \bar{\boldsymbol{x}}M^\top),$$

where $\bar{\boldsymbol{x}}$ and M are as in (39) and (40), respectively. Since M has $2(n-1)$ non-zero entries, $\boldsymbol{g}$ can be computed in $O(n^2)$ time (this is linear in the dimension of $\mathcal{C}$).

F.5. Convex-hull of Rotation Matrices

We now consider the setting where the losses (f_t) are defined on the convex-hull of rotation matrices; that is,

$$\mathcal{C} := \operatorname{conv} \mathrm{SO}(n), \quad \text{where} \quad \mathrm{SO}(n) := \{\boldsymbol{x} \in \mathbb{R}^{n\times n} : \boldsymbol{x}^\top \boldsymbol{x} = I, \ \det(\boldsymbol{x}) = 1\}.$$

This set satisfies Assumption 1 with $r = 1 - 2/n$ and $R = n$ (implying a asphericity of at most $\kappa = n^2/(n-2) = O(n)$), and so there is not need to reparametrize as we show next:

Proposition 22 *Let $n > 2$. The convex hull $\mathcal{C}$ of Orthogonal matrices in $\mathbb{R}^{n\times n}$ satisfies*

$$\mathcal{B}(1/2) \subseteq \mathcal{C} \subseteq \mathcal{B}(\sqrt{n}).$$

Proof Let $\mathrm{O}(n)$ the set of orthogonal matrices in $\mathbb{R}^{n\times n}$. By (Saunderson et al., 2015, Proposition 4.6), we have

$$\operatorname{conv} \mathrm{SO}(n) = (\operatorname{conv} \mathrm{O}(n)) \cap ((n-2)\mathrm{SO}^-(n)^\circ), \tag{41}$$

where $\mathrm{SO}^-(n) := \{\boldsymbol{x} \in \mathbb{R}^{n\times n} : \boldsymbol{x}^\top \boldsymbol{x} = I, \ \det(\boldsymbol{x}) = -1\}$. It is know that $\operatorname{conv} \mathrm{O}(n)$ coincides with the operator-norm ball (Saunderson et al., 2015), and so we have (see paragraph on the operator norm ball above)

$$\mathcal{B}(1) \subseteq \operatorname{conv} \mathrm{O}(n). \tag{42}$$

We will now show that $\mathcal{B}(1/\sqrt{n}) \subseteq \mathrm{SO}^-(n)^\circ$ from which we conclude that $\mathcal{B}(1/2) \subset \mathcal{B}(1 \vee (n^{1/2} - 2n^{-1/2})) \subset \operatorname{conv} \mathrm{SO}(n)$ using (41), (42), and the fact that $n > 2$. Let D be the diagonal matrix satisfying $D_{ii} = 1$ for all $i \in [n-1]$ and $D_{nn} = -1$. It is known that $\mathrm{SO}^-(n) = D \cdot \mathrm{SO}(n)$ (see e.g. (Saunderson et al., 2015)). Therefore, since $D \in \mathrm{O}(n)$, we have

$$\sup_{\boldsymbol{x}\in \mathrm{SO}^-(n)} \|\boldsymbol{x}\|_{\mathrm{op}} = \sup_{\boldsymbol{x}\in \mathrm{SO}(n)} \|\boldsymbol{x}\|_{\mathrm{op}} = 1.$$

Thus, by the fact that $\|\cdot\|_{\mathrm{F}} \leq \sqrt{n}\|\cdot\|_{\mathrm{op}}$, we have $\mathrm{SO}^-(n) \subseteq \mathcal{B}(\sqrt{n})$, which implies that $\mathcal{B}(1/\sqrt{n}) \subseteq \mathrm{SO}^-(n)^\circ$. In fact, since $\mathrm{SO}^-(n) \subseteq \mathcal{B}(\sqrt{n})$, we have $\langle \boldsymbol{u}, \boldsymbol{x}\rangle \leq 1$, for all $\boldsymbol{u} \in \mathcal{B}(1/\sqrt{n})$ and $\boldsymbol{x} \in \mathrm{SO}^-(n)$, and so $\mathcal{B}(1/\sqrt{n}) \subseteq \mathrm{SO}^-(n)^\circ$ by definition of a Polar set. Combining this with the fact that $\mathcal{B}(1) \subseteq \operatorname{conv} \mathrm{O}(n)$ and (41), we get that

$$\mathcal{B}(1/2) \overset{n\geq 2}{\subset} \mathcal{B}(1 \vee (n^{1/2} - 2n^{-1/2})) \subseteq \operatorname{conv} \mathrm{SO}(n) = \mathcal{C}.$$

We now show that $\mathcal{C} \subseteq \mathcal{B}(n)$. This follows by the fact that $\mathcal{C} = \operatorname{conv} \mathrm{SO}(n) \subset \operatorname{conv} \mathrm{O}(n)$ and that $\operatorname{conv} \mathrm{O}(n) \subseteq \mathcal{B}(\sqrt{n})$ since $\operatorname{conv} \mathrm{O}(n)$ is the operator norm ball. ■

To assess the complexity of a Membership Oracle for conv $\mathrm{SO}(n)$, we use the characterization of $\mathrm{SO}(n)$ in (41). Also, as argued in the proof of the previous proposition, $\operatorname{conv} \mathrm{O}(n)$ coincides with the operator-norm ball (Saunderson et al., 2015). Thus, in light of (41), to test if $\boldsymbol{x}$ is in $\operatorname{conv} \mathrm{SO}(n)$, it suffices to test if $\boldsymbol{x}$ is in the operator norm ball and in the set $\mathrm{SO}^-(n)^\circ$, simultaneously. The complexity of the former test is at most that of SVD (Saunderson et al., 2015). We now show that the complexity of testing for $\boldsymbol{x} \in \mathrm{SO}^-(n)^\circ$ is also at most that of SVD up to a constant factor. First, we note that testing membership for $\mathrm{SO}^-(n)^\circ$ can be performed using a single call to a Linear Optimization Oracle on $\mathrm{SO}^-(n)$ (by leveraging the definition of a polar set). Furthermore, LO on $\mathrm{SO}^-(n)$ can be done using one call to a LOO on $\mathrm{SO}(n)$. The latter follows by the fact that $\mathrm{SO}^-(n) = D \cdot \mathrm{SO}(n)$ (see e.g. (Saunderson et al., 2015)), where D is the diagonal matrix defined in the proof of Proposition 22, and so

$$\mathcal{O}_{\mathrm{SO}^-(n)}(\boldsymbol{x}) = \sup_{\boldsymbol{y} \in \mathrm{SO}^-(n)} \langle \boldsymbol{y}, \boldsymbol{x} \rangle = \sup_{\boldsymbol{y} \in \mathrm{SO}(n)} \langle \boldsymbol{y}, D\boldsymbol{x} \rangle = \mathcal{O}_{\mathrm{SO}(n)}(D\boldsymbol{x}).$$

Finally, since the complexity of linear optimization on $\mathrm{SO}^-(n)$ is at most the cost of SVD (Jaggi, 2013), we conclude, in light of (41), that the complexity of a Membership Oracle for $\operatorname{conv} \mathrm{SO}(n)$ is also at most that SVD up to a constant factor.

F.6. PSD Matrices with Unit Trace

We now consider the set PSD matrices with unit trace. This set does not satisfy Assumption 1 and so we need to reparametrize. It will be useful to introduce the operator $U \colon \mathbb{R}^{n(n-1)/2} \to \mathbb{R}^{n \times n}$, where for each $\boldsymbol{z} \in \mathbb{R}^{n(n-1)/2}$, $U(\boldsymbol{z})$ is the upper-triangular matrix whose ith column is equal to $(z_{i(i-1)/2+1}, \ldots, z_{i(i+1)/2}, 0, \ldots, 0)^\top \in \mathbb{R}^n$. Further, for any $\boldsymbol{x} \in \mathbb{R}^n$, we let $\operatorname{diag}(\boldsymbol{x})$ be the matrix whose diagonal constructed from the vector $\boldsymbol{x}$, and define

$$\Theta(\boldsymbol{y}, \boldsymbol{z}) \coloneqq \operatorname{diag}(J^\top \boldsymbol{y}) + U(\boldsymbol{z}) + U(\boldsymbol{z})^\top,$$

for all $\boldsymbol{y} \in \mathbb{R}^{n-1}$ and $\boldsymbol{z} \in \mathbb{R}^{n(n-1)/2}$, where J is as in (37) with $d = n$. With this, we consider the set of reparametrized losses (ℓ_t) given by

$$\ell_t(\boldsymbol{x}) \coloneqq f_t(\boldsymbol{c} + \Theta(\boldsymbol{y}, \boldsymbol{z})), \tag{43}$$

$$\text{where} \quad \boldsymbol{x} \coloneqq (\boldsymbol{y}, \boldsymbol{z}) \in \mathbb{R}^{n-1} \times \mathbb{R}^{n(n-1)/2} \quad \text{and} \quad \boldsymbol{c} \coloneqq \left(\frac{1}{2} + \frac{1}{2n}\right) \boldsymbol{e}_{nn} + \sum_{i=1}^{n-1} \frac{\boldsymbol{e}_{ii}}{2n}.$$

For any t, the function ℓ_t is convex and defined on the set

$$\mathcal{C} \coloneqq \left\{ (\boldsymbol{y}, \boldsymbol{z}) \in \mathbb{R}^{n-1} \times \mathbb{R}^{n(n-1)/2} : \boldsymbol{c} + \Theta(\boldsymbol{y}, \boldsymbol{z}) \succeq 0 \right\}. \tag{44}$$

Furthermore, this set satisfies Assumption 1 with $r = n^{-3/2}/4$ and $R = 2\sqrt{n}$, leading to an asphericity of at most $\kappa = 8n^2$ for the set (this is linear in the dimension of $\mathcal{C}$):

Proposition 23 *The set $\mathcal{C}$ in (44) satisfies $\mathcal{B}(n^{-3/2}/4) \subseteq \mathcal{C} \subseteq \mathcal{B}(2\sqrt{n})$.*

To implement a Membership Oracle for $\mathcal{C}$ one needs to be able to test if a matrix of the form $\boldsymbol{c} + \Theta(\boldsymbol{y}, \boldsymbol{z})$ is positive definite. Since this matrix is symmetric, it suffices to check if the smallest

eigenvalue of $\Theta(\boldsymbol{y}, \boldsymbol{z})$ is non-negative. We now present a way of approximating the smallest eigenvalue of a symmetric matrix, which will then lead to an approximate Membership Oracle for $\mathcal{C}$. Given a symmetric matrix M and $\delta > 0$, first approximate its largest singular value $\sigma_1(M)$ up to error $\delta/2$. This can be done using $\widetilde{O}(\text{nnz}(M)/\sqrt{\delta})$ arithmetic operations (see (Kuczyński and Woźniakowski, 1992) and (Jaggi, 2013, Proposition 8)). Next, approximate the largest singular value $\sigma_1(M')$ of $M' \coloneqq M - \sigma_1(M) \cdot I_n$ up to error $\delta/2$. This also requires $\widetilde{O}(\text{nnz}(M)/\sqrt{\delta})$ arithmetic operations. Now, since the smallest eigenvalue of M is given by $\lambda_{\min}(M) = \sigma_1(M) - \sigma_1(M')$, we can compute a δ-approximate value of $\lambda_{\min}(M)$, and thus implement a δ-approximate Membership Oracle, using $\widetilde{O}(\text{nnz}(M)/\sqrt{\delta})$ arithmetic operations.

We now show how to build a subgradient Oracle for the reparametrized losses (ℓ_t). By the chain-rule, $\boldsymbol{g} \coloneqq (\boldsymbol{g}_y, \boldsymbol{g}_z)$ is a subgradient of ℓ_t at $\boldsymbol{x} \coloneqq (\boldsymbol{y}, \boldsymbol{z})$ if and only if

$$\boldsymbol{g}_y = J\text{diag}^{-1}(\boldsymbol{\zeta}) \quad \text{and} \quad \boldsymbol{g}_z \coloneqq U^{-1}(\boldsymbol{\zeta}), \quad \text{for } \boldsymbol{\zeta} \in \partial f_t(\boldsymbol{c} + \Theta(\boldsymbol{y}, \boldsymbol{z})),$$

where J is as (37) and $U^{-1} \colon \mathbb{R}^{n \times n} \to \mathbb{R}^{n(n-1)/2}$ [resp. $\text{diag}^{-1} \colon \mathbb{R}^{n \times n} \to \mathbb{R}^n$] is any operator satisfying $U^{-1} \circ U(\boldsymbol{z}) = \boldsymbol{z}$, for all $\boldsymbol{z}$ [resp. $\text{diag}^{-1} \circ \text{diag}(\boldsymbol{x}) = \boldsymbol{x}$, for all $\boldsymbol{x}$]. Thus, the subgradient Oracle for ℓ_t requires only an additional $O(n^2)$ operations (this is linear in the dimension of $\mathcal{C}$). We now prove Proposition 23:

Proof of Proposition 23. By the fact that $\|\text{diag}(J^\top \boldsymbol{y})\|_{\text{F}} \leq 2\sqrt{n}\|\boldsymbol{y}\|$ and $\|U(\boldsymbol{z}) + U(\boldsymbol{z})^\top\|_{\text{F}} \leq 2\|U(\boldsymbol{z})\|_{\text{F}}$, for all $\boldsymbol{y} \in \mathbb{R}^{n-1}$ and $\boldsymbol{z} \in \mathbb{R}^{n(n-1)/2}$, we have

$$\|\Theta(\boldsymbol{y}, \boldsymbol{z})\|_{\text{F}}/(2\sqrt{n}) \leq \|(\boldsymbol{y}, \boldsymbol{z})\| = \|\boldsymbol{y}\| + \|\boldsymbol{z}\| \leq \|\Theta(\boldsymbol{y}, \boldsymbol{z})\|_{\text{F}}. \tag{45}$$

We will now bound the norm of $\Theta(\boldsymbol{y}, \boldsymbol{z})$. For any orthogonal matrix $H \in \mathcal{O}(n)$, $\boldsymbol{w} \in \mathbb{R}^{n \times n}$, and $\boldsymbol{\lambda} \in \mathbb{R}^{n-1}$, define

$$\Phi_H(\boldsymbol{w}) \coloneqq -\boldsymbol{e}_{nn}/(2n) + H^\top \boldsymbol{w} H; \qquad \Psi(\boldsymbol{\lambda}) \coloneqq \text{diag}(\boldsymbol{e}_n/2 + J^\top \boldsymbol{\lambda}),$$
$$\text{and} \quad \mathcal{C}' \coloneqq \left\{ \boldsymbol{\lambda} \in \mathbb{R}^{n-1} : \ \lambda_i \geq \frac{-1}{2n} \ \text{ and } \ \mathbf{1}^\top \boldsymbol{\lambda} \leq \frac{1}{2} + \frac{1}{2n} \right\},$$

where J is as in (37) with $d = n$. Since a similarity transformation does not change the trace, or the eigenvalues for that matter, we have $\boldsymbol{c} + \Phi_H \circ \Psi(\boldsymbol{\lambda}) \in \mathcal{C}, \forall H \in \mathcal{O}(n), \forall \boldsymbol{\lambda} \in \mathcal{C}'$. In particular, this implies that

$$\bigcup_{H \in \mathcal{O}(n)} \Phi_H \circ \Psi(\mathcal{C}') \subseteq \Theta(\mathcal{C}). \tag{46}$$

We will now show that for any $H \in \mathcal{O}(n)$ and $\boldsymbol{\lambda}$ on the boundary of $\mathcal{C}'$, we have $\|\Phi_H \circ \Psi(\boldsymbol{\lambda}) - \Phi_H \circ \Psi(\mathbf{0})\|_{\text{F}} \geq 1/(2n)$. This, combined with (46), would imply that $\mathcal{B}(1/(2n)) \subseteq \Theta(\mathcal{C})$. Let $H \in \mathcal{O}(n)$ and $\boldsymbol{\lambda} \in \text{bd}\, \mathcal{C}'$. Since Φ_H is an isometry with respect to the distance induced by the operator norm, we have

$$\begin{aligned} \|\Phi_H \circ \Psi(\boldsymbol{\lambda}) - \Phi_H \circ \Psi(\mathbf{0})\|_{\text{F}} &\geq \|\Phi_H \circ \Psi(\boldsymbol{\lambda}) - \Phi_H \circ \Psi(\mathbf{0})\|_{\text{op}} \\ &= \|\Psi(\boldsymbol{\lambda}) - \Psi(\mathbf{0})\|_{\text{op}} = \|J^\top \boldsymbol{\lambda}\| \geq \|\boldsymbol{\lambda}\| \geq 1/(2n), \end{aligned}$$

where the last inequality follows by Proposition 21. Thus, we have that $\mathcal{B}(1/(2n)) \subseteq \Theta(\mathcal{C})$. Combining this with (45) implies that $\mathcal{B}(n^{-3/2}/4) \subseteq \mathcal{C}$.

We will now show that $\mathcal{C} \subset \mathcal{B}(2\sqrt{n})$. For this, we will first show that (46) holds with equality. Let $\Theta'(\cdot,\cdot) \coloneqq \boldsymbol{e}_{nn}/2 + \Theta(\cdot,\cdot)$, and observe that the constraint defining the set $\mathcal{C}$ in (44) translates to $I_n/(2n) + \Theta'(\cdot,\cdot) \succeq 0$. Let $(\boldsymbol{y},\boldsymbol{z}) \in \mathcal{C}$. Since $\Theta'(\boldsymbol{y},\boldsymbol{z})$ is a real symmetric matrix, there exists an orthogonal matrix H such that $\Lambda \coloneqq H\Theta'(\boldsymbol{y},\boldsymbol{z})H^\top$ is a diagonal matrix. Let $\lambda_1,\dots,\lambda_n$ be the diagonal elements of Λ. Since these are the eigenvalues of $\Theta'(\boldsymbol{y},\boldsymbol{z})$, the fact that $I/(2n) + \Theta'(\boldsymbol{y},\boldsymbol{z}) \succeq 0$ implies $\lambda_i \geq -1/(2n)$, for $i \in [n]$. Furthermore, since $\operatorname{tr}(\Theta'(\boldsymbol{y},\boldsymbol{z})) = 1/2$, we have $\sum_{i=1}^n \lambda_i = 1/2$, and so since $\lambda_n \geq -1/(2n)$, it follows that $\sum_{i=1}^{n-1} \lambda_i \leq 1/2 + 1/(2n)$. Thus, we have $\boldsymbol{\lambda}' \coloneqq (\lambda_1,\dots,\lambda_{n-1}) \in \mathcal{C}'$ and

$$\boldsymbol{e}_{nn}/2 + \Theta(\boldsymbol{y},\boldsymbol{z}) = \Theta'(\boldsymbol{y},\boldsymbol{z}) = H^\top \operatorname{diag}(e_n/2 + J^\top\boldsymbol{\lambda}')H = \boldsymbol{e}_{nn}/2 + \Phi_H \circ \Psi(\boldsymbol{\lambda}').$$

Therefore, we have $\Theta(\mathcal{C}) \subseteq \bigcup_{H\in\mathcal{O}(n)} \Phi_H \circ \Psi(\mathcal{C}')$, and so by (46), we have that $\Theta(\mathcal{C}) = \bigcup_{H\in\mathcal{O}(n)} \Phi_H \circ \Psi(\mathcal{C}')$. Now, for any $H \in \mathcal{O}(n)$ and $\boldsymbol{\lambda} \in \mathcal{C}'$, we have

$$\begin{aligned}
\|\Phi_H \circ \Psi(\boldsymbol{\lambda}) - \Phi_H \circ \Psi(\boldsymbol{0})\|_{\mathrm{F}} &\leq \sqrt{n}\|\Phi_H \circ \Psi(\boldsymbol{\lambda}) - \Phi_H \circ \Psi(\boldsymbol{0})\|_{\mathrm{op}}, \\
&= \sqrt{n}\|\Psi(\boldsymbol{\lambda}) - \Psi(\boldsymbol{0})\|_{\mathrm{op}}, \\
&= \sqrt{n}\|J^\top\boldsymbol{\lambda}\| \leq \sqrt{n}\|\boldsymbol{\lambda}\| + \sqrt{n}|\langle \boldsymbol{1},\boldsymbol{\lambda}\rangle| \leq 2\sqrt{n},
\end{aligned}$$

where the last inequality follows by Proposition 21, and the fact that $\langle \boldsymbol{1},\boldsymbol{\lambda}\rangle \leq 1/2 + 1/(2n) \leq 1$ by definition of $\mathcal{C}'$. Therefore, by (45), we have $\mathcal{C} \subseteq \mathcal{B}(2\sqrt{n})$. ■

F.7. PSD Matrices with Bounded Diagonals

We now consider the set of PSD matrices with bounded diagonals; that is, $\mathcal{K} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^{n\times n} : \boldsymbol{x} \succeq 0, \text{and } 0 \leq x_{ii} \leq 1, \text{for all } i \in [n]\}$. This set does not satisfy Assumption 1 and so we need to reparametrize. As in the case of PSD matrices with unit trace, we let $U\colon \mathbb{R}^{n(n-1)/2} \to \mathbb{R}^{n\times n}$ be the operator such that for each $\boldsymbol{z} \in \mathbb{R}^{n(n-1)/2}$, $U(\boldsymbol{z})$ is the upper-triangular matrix whose ith column is equal to $(z_{i(i-1)/2+1},\dots,z_{i(i+1)/2},0,\dots,0)^\top \in \mathbb{R}^n$. Also, for any $\boldsymbol{y} \in \mathbb{R}$, we let $\operatorname{diag}(\boldsymbol{y})$ be the matrix whose diagonal constructed from the vector $\boldsymbol{y}$, and define

$$\Xi(\boldsymbol{y},\boldsymbol{z}) \coloneqq \operatorname{diag}(\boldsymbol{y}) + U(\boldsymbol{z}) + U(\boldsymbol{z})^\top.$$

With this, we consider the set of reparametrized losses (ℓ_t) given by

$$\ell_t(\boldsymbol{x}) \coloneqq f_t(\boldsymbol{c} + \Xi(\boldsymbol{y},\boldsymbol{z})), \quad \text{where} \quad \boldsymbol{x} \coloneqq (\boldsymbol{y},\boldsymbol{z}) \quad \text{and} \quad \boldsymbol{c} \coloneqq I_n/2.$$

For any t, the function ℓ_t is convex and defined on the set

$$\mathcal{C} \coloneqq \left\{(\boldsymbol{y},\boldsymbol{z}) \in \mathbb{R}^n \times \mathbb{R}^{n(n-1)/2} : \boldsymbol{c} + \Xi(\boldsymbol{y},\boldsymbol{z}) \succeq 0, \quad y_{ii} \leq 1/2, \forall i \in [n]\right\}, \tag{47}$$

Furthermore, this set satisfies Assumption 1 with $r = 1/4$ and $R = n^{3/2}/2$, and so the asphericity is $\kappa = 2n^{3/2}$:

Proposition 24 *The set $\mathcal{C}$ in (47) satisfies $\mathcal{B}(1/4) \subseteq \mathcal{C} \subseteq \mathcal{B}(n^{3/2})$.*

To implement a Membership Oracle for $\mathcal{C}$ one needs to be able to test if a matrix of the form $\boldsymbol{c} + \Xi(\boldsymbol{y}, \boldsymbol{z})$ is positive definite. Since this matrix is symmetric, it suffices to check that the smallest eigenvalue of $\Xi(\boldsymbol{y}, \boldsymbol{z})$ is non-negative. This can be done in the same way as in Subsection F.6 (PSD matrices with unit trace), and so a δ-approximate Membership Oracle for $\mathcal{C}$ can be implemented using $\widetilde{O}(\mathrm{nnz}(\boldsymbol{x})/\sqrt{\delta})$ arithmetic operations for any input $\boldsymbol{x} \in \mathbb{R}^n \times \mathbb{R}^{n(n-1)/2}$ and tolerance $\delta > 0$.

We now show how to build a subgradient Oracle for the reparametrized losses (ℓ_t). By the chain-rule, $\boldsymbol{g} \coloneqq (\boldsymbol{g}_y, \boldsymbol{g}_z)$ is a subgradient of ℓ_t at $\boldsymbol{x} \coloneqq (\boldsymbol{y}, \boldsymbol{z})$ if and only if

$$\boldsymbol{g}_y = \mathrm{diag}^{-1}(\boldsymbol{\zeta}) \quad \text{and} \quad \boldsymbol{g}_z \coloneqq U^{-1}(\boldsymbol{\zeta}), \quad \text{for } \boldsymbol{\zeta} \in \partial f_t(\boldsymbol{c} + \Xi(\boldsymbol{y}, \boldsymbol{z})),$$

where $U^{-1} \colon \mathbb{R}^{n \times n} \to \mathbb{R}^{n(n-1)/2}$ [resp. $\mathrm{diag}^{-1} \colon \mathbb{R}^{n \times n} \to \mathbb{R}^n$] is an operator satisfying $U^{-1} \circ U(\boldsymbol{z}) = \boldsymbol{z}$, for $\boldsymbol{z}$ [resp. $\mathrm{diag}^{-1} \circ \mathrm{diag}(\boldsymbol{x}) = \boldsymbol{x}$, for all $\boldsymbol{x}$]. Thus, the subgradient Oracle for ℓ_t only requires an additional $O(n^2)$ operations (this is linear in the dimension of $\mathcal{C}$).

Proof of Proposition 24. By the fact that $\|\mathrm{diag}(\boldsymbol{y})\|_{\mathrm{F}} = \|\boldsymbol{y}\|$ and $\|U(\boldsymbol{z}) + U(\boldsymbol{z})^\top\|_{\mathrm{F}} \leq 2\|U(\boldsymbol{z})\|_{\mathrm{F}}$, for all $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{z} \in \mathbb{R}^{n(n-1)/2}$, we have

$$\|\Xi(\boldsymbol{y}, \boldsymbol{z})\|_{\mathrm{F}}/2 \leq \|(\boldsymbol{y}, \boldsymbol{z})\| \coloneqq \|\boldsymbol{y}\| + \|\boldsymbol{z}\| \leq \|\Xi(\boldsymbol{y}, \boldsymbol{z})\|_{\mathrm{F}}. \tag{48}$$

We will now bound the norm of $\Xi(\boldsymbol{y}, \boldsymbol{z})$. Let $(\boldsymbol{y}, \boldsymbol{z}) \in \mathcal{C}$. Since $\Xi(\boldsymbol{y}, \boldsymbol{z})$ is a real symmetric matrix, there exists an orthogonal matrix H such that $\Lambda \coloneqq H\Xi(\boldsymbol{y}, \boldsymbol{z})H^\top$ is a diagonal matrix. Let $\lambda_1, \ldots, \lambda_n$ be the diagonal elements of Λ. Since these are also the eigenvalues of $\Xi(\boldsymbol{y}, \boldsymbol{z})$, the fact that $I_n/2 + \Xi(\boldsymbol{y}, \boldsymbol{z}) \succeq 0$ implies that $\lambda_i \geq -1/2$, for all $i \in [n]$. Furthermore, since $y_{ii} \leq 1/2$ for all $i \in [n]$, we have $\mathrm{tr}(\Xi(\boldsymbol{y}, \boldsymbol{z})) \leq n/2$ and so $\sum_{i=1}^n \lambda_i \leq n/2$. That is,

$$\boldsymbol{\lambda} \in \mathcal{C}' \coloneqq \left\{ \boldsymbol{x} \in \mathbb{R}^n : \; x_i \geq \frac{-1}{2} \;\text{ and }\; \mathbf{1}^\top \boldsymbol{x} \leq \frac{n}{2} \right\}.$$

The argument above implies that

$$\Xi(\mathcal{C}) \subseteq \bigcup_{H \in \mathrm{O}(n)} \Phi_H(\mathcal{C}'), \quad \text{where} \quad \Phi_H(\boldsymbol{\lambda}') = H^\top \mathrm{diag}\,(\boldsymbol{\lambda})\, H.$$

Using that $\|\boldsymbol{\lambda}\| \leq n$, for all $\boldsymbol{\lambda} \in \mathcal{C}'$, and the fact that multiplication by an orthogonal matrix does not change the operator norm, we have

$$n \geq \sup_{\boldsymbol{\lambda} \in \mathcal{C}'} \|\boldsymbol{\lambda}\| = \sup_{\boldsymbol{\lambda} \in \mathcal{C}', H \in \mathrm{O}(n)} \|\Phi_H(\boldsymbol{\lambda})\|_{\mathrm{op}} \geq \sup_{\boldsymbol{x} \in \mathcal{C}} \|\Xi(\boldsymbol{x})\|_{\mathrm{op}} \geq \sup_{\boldsymbol{x} \in \mathcal{C}} \|\Xi(\boldsymbol{x})\|_{\mathrm{F}}/\sqrt{n} \geq \sup_{\boldsymbol{x} \in \mathcal{C}} \|\boldsymbol{x}\|/\sqrt{n},$$

where the last inequality follows by (48). This implies that

$$\mathcal{C} \subseteq \mathcal{B}(n^{3/2}).$$

We now show that $\mathcal{B}(1/4) \subseteq \mathcal{C}$. For this, we need to evaluate the quantity $\inf_{\boldsymbol{x} \in \mathrm{bd}\mathcal{C}} \|\boldsymbol{x}\|$. Let $\boldsymbol{x} \in \mathrm{bd}\mathcal{C}$. The fact that $\boldsymbol{x}$ is on the boundary of $\mathcal{C}$ implies that at least one of the inequality constraints in the definition of $\mathcal{C}$ must be satisfied with equality; that is, one of the following must be true:

(a) $\exists i \in [n]$, such that $x_{ii} = 1/2$.

(b) $\exists \boldsymbol{u} \in \mathbb{R}^n$, such that $\|\boldsymbol{u}\| = 1$ and $\Xi(\boldsymbol{y}, \boldsymbol{z})\boldsymbol{u} = -\boldsymbol{c}\boldsymbol{u}$.

If (a) is true, then $\|\boldsymbol{x}\| \geq 1/2$. In case (b) holds, then by (48), we have

$$\|\boldsymbol{x}\| \geq \|\Xi(\boldsymbol{y},\boldsymbol{z})\|_{\mathrm{F}}/2 \geq \|\Xi(\boldsymbol{y},\boldsymbol{z})\|_{\mathrm{op}}/2 \geq \|\Xi(\boldsymbol{y},\boldsymbol{z})\boldsymbol{u}\|/2 = \|\boldsymbol{c}\boldsymbol{u}\|/2 = 1/4.$$

Since $\boldsymbol{x}$ was chosen arbitrarily on the boundary of $\mathcal{C}$, we have that $\mathcal{B}(1/4) \subseteq \mathcal{C}$, which completes the proof. ■

F.8. The Flow and Matroid Polytopes

Flow Polytope. For this polytopes, we do not present an explicit parametrization since it is highly dependent on the specific problem at hand. We only study the complexity of the Membership Oracles for this case. The flow polytope represents the convex hull of indicator vectors corresponding to paths in a directed acyclic graph with d nodes and m edges. This polytope can be described with $O(m+d)$ linear inequalities (Hazan and Kale, 2012; Mészáros et al., 2019). Therefore, a Membership Oracle for this polytope can be implemented using $O(d+m)$ arithmetic operations. Linear optimization on the flow polytope has the same complexity up to log-factor (Schrijver, 2003).

Matroid Polytope. Same as in the case of the flow polytope, we only comment on the computational complexity of a Membership Oracle. A Matroid Polytope is the convex hull of indicator vectors corresponding to the independent sets $A \in I$ of a matroid $M = (E, I)$. The polytope can be described using $O(2^d)$ linear inequalities where $d = |E|$. Thus, the naive implementation of the Membership Oracle that checks all these linear inequalities would be intractable. We will present an alternative approach to designing a δ-approximate Membership Oracle for M that requires only $O(d^3 + d^2 \ln(d)\mathrm{Cost}(\mathcal{I}_M)) \ln(1/\delta)$ arithmetic operations, where $\mathrm{Cost}(\mathcal{I}_M)$ is the computational cost (number of arithmetic operations) of testing if a subset of E is independent (i.e. an element of I).

Let $\mathcal{C}$ denote the matroid polytope corresponding to our matroid $M = (E, I)$. As we have shown in App. B, the complexity of a Membership Oracle on $\mathcal{C}$ is the same as linear optimization on $\mathcal{C}^\circ$. Furthermore, a δ-approximate Linear Optimization Oracle for $\mathcal{C}^\circ$ can be implemented using $O(d^3 + d\mathrm{Cost}(\mathcal{S}_{\mathcal{C}^\circ})) \ln(1/\delta)$ arithmetic operations, where $\mathrm{Cost}(\mathcal{S}_{\mathcal{C}^\circ})$ is the computational cost (number of arithmetic operations) of a Separation Oracle on $\mathcal{C}^\circ$ (Lee et al., 2018). The fact that $(\mathcal{C}^\circ)^\circ = \mathcal{C}$ for a closed convex set and our results from App. B (see also (Grötschel et al., 1993; Molinaro, 2020)) imply that the complexity of $\mathcal{S}_{\mathcal{C}^\circ}$ is the same as linear optimization on $\mathcal{C}$. The latter can be performed using $O(d\ln(d)\mathrm{Cost}(\mathcal{I}_M))$ arithmetic operations (Schrijver, 2003, Section 40.1). All in all, a Membership Oracle for $\mathcal{C}$ can be implemented using $O(d^3 + d^2\ln(d)\mathrm{Cost}(\mathcal{I}_M))$ arithmetic operations (omitting log-factors in $1/\delta$).

Appendix G. Adaptive OCO Algorithms

We now present two algorithms for Online Convex Optimization that we will build on to derive our results.

Algorithm 8 FTRL-proximal on $\mathcal{B}(R)$ (McMahan, 2017, Algorithm 2 & Section 3.3)

Require: Radius R.

Initialize $\boldsymbol{w}_1 = \mathbf{0}$, $\boldsymbol{G}_0 = \mathbf{0}$, and $V_0 = 0$.
Initialize $\eta_0 = \infty$ and $\Sigma_0 = 0$,
for $t = 1, 2, \dots$ **do**
  Play $\boldsymbol{w}_t$ and observe $\boldsymbol{\nabla}_t \in \partial \ell_t(\boldsymbol{w}_t)$.
  Set $\eta_t = \sqrt{2}R/\sqrt{V_t}$, $\sigma_t = 1/\eta_t - 1/\eta_{t-1}$, and $\Sigma_t = \Sigma_{t-1} + \sigma_t$.
  // With the convention $1/\infty := 0$
  Set $\boldsymbol{G}_t = \boldsymbol{G}_{t-1} + \boldsymbol{\nabla}_t$ and $V_t = V_{t-1} + \|\boldsymbol{\nabla}_t\|^2$.
  Set $\boldsymbol{w}_{t+1} = \Pi_{\mathcal{B}(R)}(\widetilde{\boldsymbol{w}}_{t+1})$, where $\widetilde{\boldsymbol{w}}_{t+1} := (-\boldsymbol{G}_t + \sum_{s=1}^t \sigma_s \boldsymbol{w}_s)/\Sigma_t$.
  // The vector $\boldsymbol{w}_{t+1}$ above satisfies $\boldsymbol{w}_{t+1} \in \arg\min_{\boldsymbol{w} \in \mathcal{B}(R)} \langle \boldsymbol{G}_t, \boldsymbol{w} \rangle + \sum_{s=1}^t \sigma_s^2 \|\boldsymbol{w} - \boldsymbol{w}_s\|^2/2$.
end for

The first algorithm, FTRL-prox(Alg. 8), requires R as input, but does not require an upper bound on the norm of the input loss vectors $(\boldsymbol{\nabla}_t)$ (i.e. the algorithm adapts to the norm of the loss vectors). We now state the guarantee of FTRL-prox which follows from (McMahan, 2017, Theorem 2 & Section 3.3) with the choice of learning rate $\eta_t := \sqrt{2}R/\sqrt{V_t}$, where $V_t := \sum_{s=1}^t \|\boldsymbol{\nabla}_s\|^2$:

Proposition 25 (FTRL-proximal's regret) *For any adversarial sequence of convex losses (ℓ_t) on $\mathcal{B}(R)$, the iterates $(\boldsymbol{w}_t)$ of* FTRL-prox *(Alg. 8) with parameter $R > 0$ in response to (ℓ_t), satisfy for all $T \geq 1$ and $\boldsymbol{w} \in \mathcal{B}(R)$,*

$$\sum_{t=1}^T (\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{w})) \leq \sum_{t=1}^T \langle \boldsymbol{\nabla}_t, \boldsymbol{w}_t - \boldsymbol{w} \rangle \leq 2R\sqrt{2V_T},$$

where $\boldsymbol{\nabla}_t$, $t \geq 1$, is any subgradient in $\partial \ell_t(\boldsymbol{w}_t)$ and $V_T := \sum_{t=1}^T \|\boldsymbol{\nabla}_t\|^2$.

Algorithm 9 FreeGrad (Mhammedi and Koolen, 2020) with the unconstrained-to-constrained reduction due to Cutkosky (2020).

Require: Parameters $\epsilon > 0$ and $R > 0$.

Initialize $\boldsymbol{w}_1 = \boldsymbol{s}_1 = \mathbf{0}$, $B_0 = \epsilon$, $\boldsymbol{G}_0 = \mathbf{0}$, and $Q_0 = \epsilon^2$.
for $t = 1, 2, \dots$ **do**
  Play $\boldsymbol{w}_t$ and observe $\boldsymbol{\nabla}_t \in \partial \ell_t(\boldsymbol{w}_t)$.
  Set $B_t = B_{t-1} \vee \|\boldsymbol{\nabla}_t\|$ and $\bar{\boldsymbol{\nabla}}_t := \boldsymbol{\nabla}_t \cdot B_{t-1}/B_t$.
  Set $\tilde{\boldsymbol{\nabla}}_t = \bar{\boldsymbol{\nabla}}_t - \mathbb{I}_{\langle \boldsymbol{\nabla}_t, \boldsymbol{w}_t \rangle < 0} \cdot \langle \bar{\boldsymbol{\nabla}}_t, \boldsymbol{s}_t \rangle \boldsymbol{s}_t$. // $\boldsymbol{s}_t$ is defined on Lines 1 and 8
  Set $\boldsymbol{G}_t = \boldsymbol{G}_{t-1} + \tilde{\boldsymbol{\nabla}}_t$ and $Q_t = Q_{t-1} + \|\tilde{\boldsymbol{\nabla}}_t\|^2$.
  Set $\boldsymbol{x}_{t+1} := -\boldsymbol{G}_t \cdot \frac{(2Q_t + B_t\|\boldsymbol{G}_t\|) \cdot \epsilon^2}{2(Q_t + B_t\|\boldsymbol{G}_t\|)^2 \sqrt{Q_t}} \cdot \exp\left(\frac{\|\boldsymbol{G}_t\|^2}{2Q_t + 2B_t\|\boldsymbol{G}_t\|}\right)$.
  Set $\boldsymbol{w}_{t+1} = \Pi_{\mathcal{B}(R)}(\boldsymbol{x}_{t+1})$ and $\boldsymbol{s}_{t+1} = \boldsymbol{x}_{t+1}/\|\boldsymbol{x}_{t+1}\| \cdot \mathbb{I}_{\|\boldsymbol{x}_{t+1}\| > R}$.
  // We use the convention that $0/0 = 0$.
end for

Another algorithm that will be instrumental to developing our projection-free (and scale-free) algorithms for strongly convex losses is FreeGrad (see Algorithm 9). An important property of

FreeGrad that will be useful to us when developing a projection-free algorithm for strongly convex losses (see Section C.2) is that its regret scales directly with the norm $\|\boldsymbol{w}\|$ of the comparator, as opposed to the worst-case R. We note that FreeGrad internally clips the sequence of observed sub-gradients. In our application of FreeGrad, it will be useful to state the guarantee of its iterates $(\boldsymbol{w}_t)$ in response to the sequence of clipped subgradients $(\bar{\boldsymbol{\nabla}}_t)$:

Proposition 26 (FreeGrad's clipped linearized regret) *For any adversarial sequence of convex losses (ℓ_t) on $\mathbb{R}^d$, the iterates $(\boldsymbol{w}_t)$ of* FreeGrad *with parameter $\epsilon, R > 0$, satisfy for all $T \geq 1$ and $\boldsymbol{w} \in \mathcal{B}(R)$,*

$$\sum_{t=1}^{T} \langle \bar{\boldsymbol{\nabla}}_t, \boldsymbol{w}_t - \boldsymbol{w} \rangle \leq 2\|\boldsymbol{w}\| \sqrt{\bar{V}_T \ln_+ \left(\frac{2\|\boldsymbol{w}\| \bar{V}_T}{\epsilon^2} \right)} + 4B_T \|\boldsymbol{w}\| \ln \left(\frac{4B_T \|\boldsymbol{w}\| \sqrt{\bar{V}_T}}{\epsilon^2} \right) + \epsilon,$$

where $\boldsymbol{\nabla}_t \in \partial \ell_t(\boldsymbol{w}_t)$ (any sub-grad.), $B_T \coloneqq \epsilon \vee \max_{t \in [T]} \|\boldsymbol{\nabla}_t\|$, $\bar{\boldsymbol{\nabla}}_t \coloneqq \boldsymbol{\nabla}_t \cdot B_{t-1}/B_t$, and $\bar{V}_T \coloneqq \epsilon^2 + \sum_{t=1}^{T} \|\bar{\boldsymbol{\nabla}}_t\|^2$.

Technically, the original version of FreeGrad applies to unbounded OCO, whereas the version of FreeGrad in Algorithm 9 generates outputs in $\mathcal{B}(R)$ using the same constrained-to-unconstrained reduction due to Cutkosky (2020). We constrain the outputs of FreeGrad to ensure that the iterates $(\boldsymbol{w}_t)$ of Algorithm 1 in the setting of Section C.2 are bounded, which in turn ensures that the approximation errors of $\mathsf{OPT}_{\mathcal{C}^\circ}$ in Algorithm 1 are not too large. The proof of Proposition 26 follows by (Mhammedi and Koolen, 2020, Theorem 6) and (Cutkosky, 2020, Theorem 2).

We close this section by mentioning that the unbounded version of FreeGrad, where $\tilde{\boldsymbol{\nabla}}_t = \boldsymbol{\nabla}_t$ and $\boldsymbol{w}_t = \boldsymbol{x}_t$, is an FTRL instance with the specific sequence of regularizers (ϕ_t^*), where ϕ_t^* is the Fenchel dual of

$$\phi_t(\boldsymbol{x}) \coloneqq \frac{1}{\sqrt{V_{t-1}}} \exp\left(\frac{\|\boldsymbol{x}\|^2}{2V_{t-1} + 2\|\boldsymbol{x}\|} \right), \quad V_t \coloneqq \epsilon^2 + \sum_{s=1}^{t} \|\boldsymbol{\nabla}_s\|^2.$$

In particular, the iterate $\boldsymbol{x}_{t+1}$ on Line 7 of Algorithm 9 is given by

$$\boldsymbol{x}_{t+1} \in \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathbb{R}^d} \{\langle \boldsymbol{G}_t, \boldsymbol{x} \rangle + \phi_t^*(\boldsymbol{x})\} = \{\nabla \phi_t(-\boldsymbol{G}_t)\}.$$

We recall that the Fenchel dual ϕ_t^* of ϕ_t is defined as $\phi_t^*(\boldsymbol{x}) = \sup_{\boldsymbol{u} \in \mathbb{R}^d} \{\langle \boldsymbol{u}, \boldsymbol{x} \rangle - \phi_t(\boldsymbol{u})\}$.

Appendix H. Linear Optimization on $\mathcal{C}^\circ$ using a Membership Oracle for $\mathcal{C}$

In this section, we restate and prove a slight extension of (Lee et al., 2018, Lemma 9 & 10), which we need in the proof of Proposition 11 at the end of this section. Our extension involves showing that the approximate subgradient in (Lee et al., 2018, Lemma 10) has bounded norm in high probability, which we need in the proof of Proposition 11. We also make the limiting argument used in the proof of (Lee et al., 2018, Lemma 9 & 10) more explicit; in particular, for any convex function $f : \mathbb{R}^d \to \mathbb{R}$ (not necessarily differentiable) and $\varepsilon > 0$, we explicitly construct a twice differentiable function $\tilde{f}_\varepsilon$ such that $\|f - \tilde{f}_\varepsilon\|_\infty \leq \varepsilon$. This will then allow us to make a limiting argument precise via Fatou's lemma to get the final result we want (see proof of Lemma 29).

Throughout this section, we let $U_\infty(\boldsymbol{u}, \nu)$ denote the uniform distribution over $\mathcal{B}_\infty(\boldsymbol{u}, \nu)$, for any $\boldsymbol{u} \in \mathbb{R}^d$ and $\nu > 0$. The next lemma is taken from (Lee et al., 2018, Lemma 9) with only minor notation adjustments.

Lemma 27 *For any $\boldsymbol{w} \in \mathbb{R}^d$, $0 < \nu_2 \leq \nu_1$, and twice differentiable convex function h defined on $B_\infty(\boldsymbol{w}, \nu_1 + \nu_2)$ with $\|\nabla h(\boldsymbol{z})\|_\infty \leq L$ for any $z \in \mathcal{B}_\infty(\boldsymbol{w}, \nu_1 + \nu_2)$ we have*

$$\mathbb{E}_{\boldsymbol{u} \sim U_\infty(\boldsymbol{w}, \nu_1)} \mathbb{E}_{\boldsymbol{z} \sim U_\infty(\boldsymbol{u}, \nu_2)} \|\nabla h(\boldsymbol{z}) - \boldsymbol{g}(\boldsymbol{u})\|_1 \leq \frac{\nu_2 d^{3/2} L}{\nu_1},$$

where $\boldsymbol{g}(\boldsymbol{u}) \coloneqq \mathbb{E}_{\boldsymbol{z} \sim U_\infty(\boldsymbol{u}, \nu_2)}[\nabla h(\boldsymbol{z})]$ and $U_\infty(\boldsymbol{u}, \nu)$ denotes the uniform distribution over $\mathcal{B}_\infty(\boldsymbol{u}, \nu)$.

Algorithm 10 Approximate Subgradient Oracle (Lee et al., 2018).

Require: Inputs $\nu_1 > 0$, $L > 0$, $\varepsilon > 0$, and $\boldsymbol{w} \in \mathbb{R}^d$.
Function $\tilde{f} : \mathbb{R}^d \to \mathbb{R}$.
Set $\nu_2 = \sqrt{\frac{\varepsilon \nu_1}{d^{1/2} L}}$.
Sample $\boldsymbol{u} \in \mathcal{B}_\infty(\boldsymbol{w}, \nu_1)$ and $\boldsymbol{z} \in \mathcal{B}_\infty(\boldsymbol{u}, \nu_2)$ independently and uniformly at random.
for $i = 1, 2, \ldots, d$ **do**
  Let $\boldsymbol{w}_i'$ and $\boldsymbol{w}_i$ be the end point of the interval $\mathcal{B}_\infty(\boldsymbol{u}, \nu_2) \cap \{\boldsymbol{z} + \lambda \boldsymbol{e}_i : \lambda \in \mathbb{R}\}$.
  Set $\tilde{s}_i = \frac{1}{2\nu_2}(\tilde{f}(\boldsymbol{w}_i') - \tilde{f}(\boldsymbol{w}_i))$.
end for
Set $\tilde{\boldsymbol{s}} = (\tilde{s}_1, \ldots, \tilde{s}_d)^\top$.
Return $\tilde{\boldsymbol{s}}$.

We now restate and slightly extend (Lee et al., 2018, Lemma 10) for twice differentiable functions. We then extend the result to non-differentiable functions using Fatou's lemma—see Lemma 29.

Lemma 28 *Let $L, \nu_1 > 0$, $\boldsymbol{w} \in \mathbb{R}^d$, and $h : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable convex function such that $\|\nabla h(\boldsymbol{x})\|_\infty \leq L$, for any $\boldsymbol{x} \in B_\infty(\boldsymbol{w}, 2\nu_1)$. Also, let $\varepsilon \in (0, \nu_1 \sqrt{d} L]$ and $\tilde{h} : \mathbb{R}^d \to \mathbb{R}$ be such that $\|\tilde{h} - h\|_\infty \leq \varepsilon'$ for some $\varepsilon' > 0$. Then, the variables ν_2, $\boldsymbol{u}$, $\boldsymbol{z}$, $\tilde{\boldsymbol{s}}$, and $(\boldsymbol{w}_i, \boldsymbol{w}_i')_{i \in [d]}$ generated during a run of Alg. 10 with input $(\tilde{h}, \nu_1, L, \varepsilon, \boldsymbol{w})$ satisfy*

$$\forall \boldsymbol{v} \in \mathbb{R}^d,\ h(\boldsymbol{v}) \geq h(\boldsymbol{w}) + \langle \tilde{\boldsymbol{s}}, \boldsymbol{v} - \boldsymbol{w} \rangle - \|\nabla h(\boldsymbol{z}) - \tilde{\boldsymbol{s}}\|_1 \cdot \|\boldsymbol{v} - \boldsymbol{w}\|_\infty - 4 d \nu_1 L,$$

Furthermore, for $\tilde{s}_i \coloneqq (h(\boldsymbol{w}_i') - h(\boldsymbol{w}_i))/(2\nu_2)$, for all $i \in [d]$, we have

$$\|\nabla h(\boldsymbol{z}) - \tilde{\boldsymbol{s}}\|_1 \leq \|\nabla h(\boldsymbol{z}) - \tilde{\boldsymbol{s}}\|_1 + d\varepsilon'/\nu_2 \quad \text{and} \quad \mathbb{E}[\|\nabla h(\boldsymbol{z}) - \tilde{\boldsymbol{s}}\|_1] \leq 2 d^{5/4} \sqrt{L \varepsilon / \nu_1}.$$

Proof Let $\boldsymbol{u}, \boldsymbol{w}_i$, and $\boldsymbol{w}_i'$ be the random vectors generated during the call to Algorithm 10 in the lemma's statement. Further, let $\tilde{\boldsymbol{s}} \in \mathbb{R}^d$ be such that $\tilde{s}_i \coloneqq (h(\boldsymbol{w}_i') - h(\boldsymbol{w}_i))/(2\nu_2)$, for $i \in [d]$. Applying the convexity of h yields, for any $\boldsymbol{v} \in \mathbb{R}^d$,

$$\begin{aligned} h(\boldsymbol{v}) &\geq h(\boldsymbol{z}) + \langle \nabla h(\boldsymbol{z}), \boldsymbol{v} - \boldsymbol{z} \rangle \\ &= h(\boldsymbol{z}) + \langle \tilde{\boldsymbol{s}}, \boldsymbol{v} - \boldsymbol{w} \rangle + \langle \nabla h(\boldsymbol{z}) - \tilde{\boldsymbol{s}}, \boldsymbol{v} - \boldsymbol{w} \rangle + \langle \nabla h(\boldsymbol{z}), \boldsymbol{w} - \boldsymbol{z} \rangle \\ &\geq h(\boldsymbol{w}) + \langle \tilde{\boldsymbol{s}}, \boldsymbol{v} - \boldsymbol{w} \rangle - \|\nabla h(\boldsymbol{z}) - \tilde{\boldsymbol{s}}\|_1 \|\boldsymbol{v} - \boldsymbol{w}\|_\infty - \|\nabla h(\boldsymbol{z})\|_\infty \|\boldsymbol{w} - \boldsymbol{z}\|_1 . \end{aligned}$$

Now, $\|\nabla h(\boldsymbol{z})\|_\infty \leq L$ and $\|\boldsymbol{w} - \boldsymbol{z}\|_1 \leq d \cdot \|\boldsymbol{w} - \boldsymbol{z}\|_\infty \leq 2d(\nu_1 + \nu_2)$ by definition of $\boldsymbol{z}$ in Algorithm 10. Furthermore, since $\varepsilon \leq \nu_1 \sqrt{d} L$, we have $\nu_2 = \sqrt{\frac{\varepsilon \nu_1}{d^{1/2} L}} \leq \nu_1$. Plugging these facts in the inequality of the previous display implies

$$h(\boldsymbol{v}) \geq h(\boldsymbol{w}) + \langle \tilde{\boldsymbol{s}}, \boldsymbol{v} - \boldsymbol{w} \rangle - \|\nabla h(\boldsymbol{z}) - \tilde{\boldsymbol{s}}\|_1 \|\boldsymbol{v} - \boldsymbol{w}\|_\infty - 4 d \nu_1 L.$$

This shows the first inequality we are after. Now, by the definition of $\tilde{\boldsymbol{s}}$ in Algorithm 10 and the fact that $\tilde{h}$ satisfies $\|\tilde{h} - h\|_\infty \le \varepsilon'$, we have

$$\|\nabla h(\boldsymbol{z}) - \tilde{\boldsymbol{s}}\|_1 \le \|\nabla h(\boldsymbol{z}) - \tilde{\boldsymbol{s}}\|_1 + d\varepsilon'/\nu_2.$$

It remains to bound $\mathbb{E}[\|\nabla h(\boldsymbol{z}) - \tilde{\boldsymbol{s}}\|]$. For this, let $\boldsymbol{g}(\boldsymbol{u}) \coloneqq \mathbb{E}_{\boldsymbol{z}\sim U_\infty(\boldsymbol{u},\nu_2)}[\nabla h(\boldsymbol{z})]$ and note that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{z}}\left[|\tilde{s}_i - [\boldsymbol{g}(\boldsymbol{u})]_i|\right] &= \mathbb{E}_{\boldsymbol{z}}\left[\left|\frac{h(\boldsymbol{w}_i') - h(\boldsymbol{w}_i)}{2\nu_2} - [\boldsymbol{g}(\boldsymbol{u})]_i\right|\right], \\
&\le \mathbb{E}_{\boldsymbol{z}}\left[\frac{1}{2\nu_2}\int\left|\frac{dh}{dw_i}(\boldsymbol{z} + \lambda \boldsymbol{e}_i) - [\boldsymbol{g}(\boldsymbol{u})]_i\right| d\lambda\right], \\
&= \mathbb{E}_{\boldsymbol{z}}\left[\left|\frac{dh}{dw_i}(\boldsymbol{z}) - [\boldsymbol{g}(\boldsymbol{u})]_i\right|\right],
\end{aligned}$$

where we used that both $\boldsymbol{z} + \lambda \boldsymbol{e}_i$ and $\boldsymbol{z}$ are uniform distribution on $\mathcal{B}_\infty(\boldsymbol{u}, \nu_2)$ in the last line. Hence, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{z}}\left[\|\tilde{\boldsymbol{s}} - \nabla h(\boldsymbol{z})\|_1\right] \le \mathbb{E}_z\left[\|\nabla h(\boldsymbol{z}) - \boldsymbol{g}(\boldsymbol{u})\|_1 + \mathbb{E}_{\boldsymbol{z}}\|\tilde{\boldsymbol{s}} - \boldsymbol{g}(\boldsymbol{y})\|_1\right] &\le 2\mathbb{E}_{\boldsymbol{z}}\left[\|\nabla h(\boldsymbol{z}) - \boldsymbol{g}(\boldsymbol{y})\|_1\right] \\
&\le 2d^{5/4}\sqrt{\frac{\varepsilon L}{\nu_1}},
\end{aligned}$$

where the last inequality follows by Lemma 27 and the fact that $\nu_2 = \sqrt{\frac{\varepsilon\nu_1}{d^{1/2}L}} \le \nu_1$. ■

Lemma 29 *Let $L, \nu_1 > 0$, $\boldsymbol{w} \in \mathbb{R}^d$, and $f : \mathbb{R}^d \to \mathbb{R}$ be convex (not necessarily differentiable) function such that $\sup_{\boldsymbol{g}\in\partial f(\boldsymbol{x})}\|\boldsymbol{g}\| \le L$, for any $\boldsymbol{x} \in B_\infty(\boldsymbol{w}, 2\nu_1)$. Also, let $\varepsilon \in (0, \nu_1\sqrt{d}L]$ and $\tilde{f} : \mathbb{R}^d \to \mathbb{R}$ be such that $\|\tilde{f} - f\|_\infty \le \varepsilon$. Then, the output $\tilde{\boldsymbol{s}}$ of Algorithm 10 with input $(\tilde{f}, \nu_1, L, \varepsilon, \boldsymbol{w})$ satisfies*

$$\begin{gathered}
\forall \boldsymbol{v} \in \mathbb{R}^d,\ f(\boldsymbol{v}) \ge f(\boldsymbol{w}) + \langle \tilde{\boldsymbol{s}}, \boldsymbol{v} - \boldsymbol{w}\rangle - X \cdot \|\boldsymbol{v} - \boldsymbol{w}\|_\infty - 4d\nu_1 L; \\
\|\tilde{\boldsymbol{s}}\|_\infty \le L + d^{1/4}\sqrt{L\varepsilon/\nu_1};\ \ \|\tilde{\boldsymbol{s}}\| \le X + L \le \|\tilde{\boldsymbol{s}}\|_1 + (d+1)L; \\
\text{and}\ \ \|\tilde{\boldsymbol{s}}\|^2 \le \left(4L + d^{1/4}\sqrt{L\varepsilon/\nu_1}\right)X + L^2,
\end{gathered}$$

where $X \ge 0$ is a non-negative random variable satisfying $\mathbb{E}[X] \le 3d^{5/4}\sqrt{L\varepsilon/\nu_1}$.

Proof Let ν_2, $\boldsymbol{u}$, $\boldsymbol{z}$, $\tilde{\boldsymbol{s}}$, and $(\boldsymbol{w}_i, \boldsymbol{w}_i')_{i\in[d]}$ be the variables generated during the call to Algorithm 10 in the lemma's statement. Further, let $\sigma > 0$ and $h_\sigma : \mathbb{R}^d \to \mathbb{R}$ be such that

$$h_\sigma(\boldsymbol{w}) \coloneqq \mathbb{E}[f(\boldsymbol{w} + \boldsymbol{x})], \quad \text{where} \quad \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 I_d).$$

It is known that h_σ is twice differentiable (Bhatnagar, 2007; Nesterov and Spokoiny, 2017; Abernethy et al., 2014), and satisfies (Duchi et al., 2012, Lemma 9),

$$\forall \boldsymbol{w} \in \mathbb{R}^d, \quad \|\nabla h_\sigma(\boldsymbol{w})\| \le L, \quad \text{and} \quad f(\boldsymbol{w}) \le h_\sigma(\boldsymbol{w}) \le f(\boldsymbol{w}) + L\sigma\sqrt{d}. \tag{49}$$

Thus, we have $\|\tilde{f} - h_\sigma\|_\infty \leq \varepsilon_\sigma \coloneqq \varepsilon + L\sigma\sqrt{d}$, and so by Lemma 28

$$h_\sigma(\boldsymbol{v}) \geq h_\sigma(\boldsymbol{w}) + \langle \tilde{\boldsymbol{s}}, \boldsymbol{v} - \boldsymbol{w}\rangle - \|\nabla h_\sigma(\boldsymbol{z}) - \tilde{\boldsymbol{s}}\|_1 \cdot \|\boldsymbol{v} - \boldsymbol{w}\|_\infty - 4d\nu_1 L, \quad \forall \boldsymbol{v} \in \mathbb{R}^d,$$

and for $\tilde{\boldsymbol{s}}_\sigma \coloneqq (h_\sigma(\boldsymbol{w}_1') - h_\sigma(\boldsymbol{w}_1), \ldots, h_\sigma(\boldsymbol{w}_d') - h_\sigma(\boldsymbol{w}_d))^\top/(2\nu_2)$,

$$\|\nabla h_\sigma(\boldsymbol{z}) - \tilde{\boldsymbol{s}}\|_1 \leq \|\nabla h_\sigma(\boldsymbol{z}) - \tilde{\boldsymbol{s}}_\sigma\|_1 + d\varepsilon_\sigma/\nu_2, \quad \text{and} \quad \mathbb{E}[\|\nabla h_\sigma(\boldsymbol{z}) - \tilde{\boldsymbol{s}}_\sigma\|_1] \leq 2d^{5/4}\sqrt{L\varepsilon/\nu_1}, \tag{50}$$

where the expectation is with respect to $\boldsymbol{z} \sim U_\infty(\boldsymbol{u}, \nu_2)$ and $\boldsymbol{u} \sim U_\infty(\boldsymbol{w}, \nu_1)$. Combining this with (49) leads to

$$f(\boldsymbol{v}) \geq f(\boldsymbol{w}) + \langle \tilde{\boldsymbol{s}}, \boldsymbol{v} - \boldsymbol{w}\rangle - \|\nabla h_\sigma(\boldsymbol{z}) - \tilde{\boldsymbol{s}}\|_1 \cdot \|\boldsymbol{v} - \boldsymbol{w}\|_\infty - 4d\nu_1 L - L\sigma\sqrt{d}, \quad \forall \boldsymbol{v} \in \mathbb{R}^d. \tag{51}$$

Now, define $X_n \coloneqq \|\nabla h_{1/n}(\boldsymbol{z}) - \tilde{\boldsymbol{s}}\|_1$, $n \geq 1$ and note that

$$0 \leq X_n \leq dL + \|\tilde{\boldsymbol{s}}\|_1 < +\infty, \quad \forall n.$$

Therefore, we have $X \coloneqq \liminf_{n\to\infty} X_n \leq \|\tilde{\boldsymbol{s}}\| + dL < +\infty$. Furthermore, by Fatou's lemma and (50), we have

$$\mathbb{E}[X] = \mathbb{E}[\liminf_{n\to\infty} X_n] \leq \liminf_{n\to\infty} \mathbb{E}[X_n] \leq \liminf_{n\to\infty} \mathbb{E}[\|\nabla h_{1/n}(\boldsymbol{z}) - \tilde{\boldsymbol{s}}_{1/n}\|_1] + d\varepsilon/\nu_2 \leq 3d^{5/4}\sqrt{L\varepsilon/\nu_1},$$

where the last inequality also follows by (50) and the fact that $\nu_2 = \sqrt{\frac{\varepsilon\nu_1}{d^{1/2}L}}$. Combining this with (51) leads to the first inequality of the lemma. Now, we have, for any $i \in [d]$,

$$\tilde{s}_i = \frac{\tilde{f}(\boldsymbol{w}_i') - \tilde{f}(\boldsymbol{w}_i)}{2\nu_2} \leq \frac{|f(\boldsymbol{w}_i') - f(\boldsymbol{w}_i)|}{2\nu_2} + \frac{\varepsilon}{\nu_2} \leq L + \frac{\varepsilon}{\nu_2} = L + d^{1/4}\sqrt{\frac{L\varepsilon}{\nu_1}},$$

where the last inequality follows by the definition of $\boldsymbol{w}_i'$ and $\boldsymbol{w}_i$, and the fact that $\sup_{\boldsymbol{g}\in\partial f(\boldsymbol{x})} \|\boldsymbol{g}\| \leq L$ (and so f is L-Lipschitz). In particular, this implies that

$$\forall i \in [d], \quad |\tilde{s}_i - [\nabla h_{1/n}(\boldsymbol{z})]_i| \leq 2L + d^{1/4}\sqrt{L\varepsilon/\nu_1}. \tag{52}$$

Using this, we get

$$\begin{aligned}
\|\tilde{\boldsymbol{s}}\|^2 &\leq \|\tilde{\boldsymbol{s}} - \nabla h_{1/n}(\boldsymbol{z})\|^2 + 2\|\tilde{\boldsymbol{s}} - \nabla h_{1/n}(\boldsymbol{z})\|\|\nabla h_{1/n}(\boldsymbol{z})\| + \|\nabla h_{1/n}(\boldsymbol{z})\|^2, \\
&\leq \left(2L + d^{1/4}\sqrt{L\varepsilon/\nu_1}\right)\|\tilde{\boldsymbol{s}} - \nabla h_{1/n}(\boldsymbol{z})\|_1 + 2\|\tilde{\boldsymbol{s}} - \nabla h_{1/n}(\boldsymbol{z})\|_1 L + L^2, \quad \text{(by (52))} \\
&\leq \left(4L + d^{1/4}\sqrt{L\varepsilon/\nu_1}\right)\|\tilde{\boldsymbol{s}} - \nabla h_{1/n}(\boldsymbol{z})\|_1 + L^2, \\
&\leq \left(4L + d^{1/4}\sqrt{L\varepsilon/\nu_1}\right)\liminf_{n\to\infty}\|\tilde{\boldsymbol{s}} - \nabla h_{1/n}(\boldsymbol{z})\|_1 + L^2, \\
&= \left(4L + d^{1/4}\sqrt{L\varepsilon/\nu_1}\right)X + L^2.
\end{aligned}$$

It remains to bound the norm $\|\tilde{\boldsymbol{s}}\|$. Similar to how we bounded $\|\tilde{\boldsymbol{s}}\|^2$, we have

$$\begin{aligned}
\|\tilde{\boldsymbol{s}}\| &\leq \|\tilde{\boldsymbol{s}} - \nabla h_{1/n}(\boldsymbol{z})\| + \|\nabla h_{1/n}(\boldsymbol{z})\|, \\
&\leq \|\tilde{\boldsymbol{s}} - \nabla h_{1/n}(\boldsymbol{z})\|_1 + L, \\
&\leq \liminf_{n\to\infty}\|\tilde{\boldsymbol{s}} - \nabla h_{1/n}(\boldsymbol{z})\|_1 + L = X + L.
\end{aligned}$$

This completes the proof. ■

H.1. Proof of Lemma 10 (Approximate Gauge Function)

Proof of Lemma 10. Let $\epsilon = \delta r/(4\kappa)^2$. We will first show that for all $\boldsymbol{w} \in \mathcal{B}(6R/5)$,

$$[\gamma_{\mathcal{C}}(\boldsymbol{w}) \geq 9/16 \;\text{ or }\; \tilde{\gamma} \geq 1] \implies [\mathsf{MEM}_{\mathcal{C}}(2\boldsymbol{w};\epsilon) = 0 \;\text{ and }\; \|\boldsymbol{w}\| \geq r/2]. \tag{53}$$

Suppose that $\tilde{\gamma} \geq 1$. By Lines 2 and 3 of Algorithm 2, this implies that $\mathsf{MEM}_{\mathcal{C}}(2\boldsymbol{w};\epsilon) = 0$. Now suppose that $\gamma_{\mathcal{C}}(\boldsymbol{w}) \geq 9/16$. Note that $\mathsf{MEM}_{\mathcal{C}}(2\boldsymbol{w};\epsilon) = 1$ only if $2\boldsymbol{w} \in \mathcal{B}(\mathcal{C},\epsilon)$. By the fact that $\mathcal{B}(r) \subseteq \mathcal{C}$ (Assumption 1) and $\delta < 1$, we have that $\mathcal{C} + \mathcal{B}(\epsilon) = \mathcal{B}(\mathcal{C},\epsilon)$. Thus, by a standard result in convex analysis, see e.g. (Hiriart-Urruty and Lemaréchal, 2004, Thm C.3.3.2), we have, for all $\boldsymbol{x} \in \mathcal{C}$,

$$\sigma_{\mathcal{B}(\mathcal{C},\epsilon)}(\boldsymbol{x}) = \sigma_{\mathcal{C}}(\boldsymbol{x}) + \sigma_{\mathcal{B}(\epsilon)}(\boldsymbol{x}). \tag{54}$$

Let $\boldsymbol{x}_* \in \partial\gamma_{\mathcal{C}}(\boldsymbol{w}) = \arg\max_{\boldsymbol{x}\in\mathcal{C}^\circ}\langle \boldsymbol{x},\boldsymbol{w}\rangle$. Since $\boldsymbol{x}_* \in \mathcal{C}^\circ$, we have $\langle \boldsymbol{x}_*,\boldsymbol{y}\rangle \leq 1$, for all $\boldsymbol{y} \in \mathcal{C}$, by definition of $\mathcal{C}^\circ$. Further, since $\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w}) \in \mathcal{C}$ (implied by Lemma 6) and $\langle \boldsymbol{x}_*,\boldsymbol{w}\rangle = \gamma_{\mathcal{C}}(\boldsymbol{w})$, we have

$$\sigma_{\mathcal{C}}(\boldsymbol{x}_*) = \sup_{\boldsymbol{y}\in\mathcal{C}}\langle \boldsymbol{x}_*,\boldsymbol{y}\rangle = \langle \boldsymbol{x}_*,\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w})\rangle = 1. \tag{55}$$

Plugging this into (54) and using the fact that $\sigma_{\mathcal{B}(\epsilon)}(\boldsymbol{x}_*) = \delta r\|\boldsymbol{x}_*\|/(8\kappa^2)$, we get

$$\sigma_{\mathcal{B}(\mathcal{C},\epsilon)}(\boldsymbol{x}_*) = 1 + \delta r\|\boldsymbol{x}_*\|/(8\kappa^2) \leq 1 + \delta/(8\kappa^2) < 9/8, \tag{56}$$

where the last inequality follows by the fact that $\delta \in (0,1)$ and $\|\boldsymbol{x}_*\| \leq 1/r$ (because $\boldsymbol{x}_* \in \mathcal{C}^\circ$, see Lemma 42-(c)). On the other hand, since $\gamma_{\mathcal{C}}(\boldsymbol{w}) \geq 9/16$, we have, by (55)

$$9/8 = 9/8\langle \boldsymbol{x}_*,\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w})\rangle \leq 9/8\langle \boldsymbol{x}_*,16\boldsymbol{w}/9\rangle = \langle \boldsymbol{x}_*,2\boldsymbol{w}\rangle.$$

This inequality together with (56) implies that the vector $\boldsymbol{x}_*$ separates $2\boldsymbol{w}$ from $\mathcal{B}(\mathcal{C},\epsilon)$ and so we have $\mathsf{MEM}_{\mathcal{C}}(2\boldsymbol{w};\epsilon) = 0$ by definition of the Membership Oracle $\mathsf{MEM}_{\mathcal{C}}(\cdot;\epsilon)$. It remains to show that $\|\boldsymbol{w}\| \geq r/2$. If $\gamma_{\mathcal{C}}(\boldsymbol{w}) \geq 9/16$, then $\|\boldsymbol{w}\| \geq 9r/16$ by the fact that $\gamma_{\mathcal{C}}(\boldsymbol{w}) \leq \|\boldsymbol{w}\|/r$ (Lemma 42-(c)). Also, if $\tilde{\gamma} \geq 1$, then Lines 2 and 3 of Algorithm 2 imply that $\|\boldsymbol{w}\| \geq r/2$. So far, we have shown that (53) holds. By definition of α and β in Algorithm 2, (53) implies that if $\gamma_{\mathcal{C}}(\boldsymbol{w}) \geq 9/16$ or $\tilde{\gamma} \geq 1$, then

$$\mathsf{MEM}_{\mathcal{C}}(\alpha\boldsymbol{w};\epsilon) = 1 \;\text{ and }\; \mathsf{MEM}_{\mathcal{C}}(\beta\boldsymbol{w};\epsilon) = 0. \tag{57}$$

Next, we will show that if either $\gamma_{\mathcal{C}}(\boldsymbol{w}) \geq 9/16$ or $\tilde{\gamma} \geq 1$, then for any $\mu > 0$,

$$\mathsf{MEM}_{\mathcal{C}}(\mu\boldsymbol{w};\epsilon) = 0 \implies \mu \geq \frac{1}{\gamma_{\mathcal{C}}(\boldsymbol{w})} - \frac{\delta}{8\kappa^2} \quad\text{and}\quad \mathsf{MEM}_{\mathcal{C}}(\mu\boldsymbol{w};\epsilon) = 1 \implies \mu \leq \frac{1}{\gamma_{\mathcal{C}}(\boldsymbol{w})} + \frac{\delta}{8\kappa^2}. \tag{58}$$

Suppose that $\mathsf{MEM}_{\mathcal{C}}(\mu\boldsymbol{w};\epsilon) = 0$. By (2), this implies that $\mu\boldsymbol{w} \notin \mathcal{B}(\mathcal{C},-\epsilon)$. On the other hand, since $\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w})$ is on the boundary of $\mathcal{C}$ (by definition of the Gauge function and Lemma 6), the vector $\boldsymbol{v} := \boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w}) - \epsilon\boldsymbol{w}/\|\boldsymbol{w}\|$ is in $\mathcal{B}(\mathcal{C},-\epsilon)$. Since $\boldsymbol{v}$ and $\mu\boldsymbol{w}$ are aligned, the fact that $\mu\boldsymbol{w} \notin \mathcal{B}(\mathcal{C},-\epsilon)$ and $\mathcal{C}$ is convex implies that

$$\|\mu\boldsymbol{w}\| \geq \|\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w}) - \epsilon\boldsymbol{w}/\|\boldsymbol{w}\|\| = \|\boldsymbol{w}\|/\gamma_{\mathcal{C}}(\boldsymbol{w}) - r\delta/(16\kappa^2) \geq \|\boldsymbol{w}\|/\gamma_{\mathcal{C}}(\boldsymbol{w}) - \|\boldsymbol{w}\|\delta/(8\kappa^2),$$

where the last inequality follows by the fact that $\|\boldsymbol{w}\| \geq r/2$ (see (53)). Dividing by $\|\boldsymbol{w}\|$ (this is non-zero by (53)) on both sides shows the first implication in (58). Similarly, if $\mathsf{MEM}_{\mathcal{C}}(\mu \boldsymbol{w}; \epsilon) = 1$. Again by (2), this implies that $\mu \boldsymbol{w} \in \mathcal{B}(\mathcal{C}, \epsilon)$. Combining this with the fact that $\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w}) + \epsilon \boldsymbol{w}/\|\boldsymbol{w}\|$ is on the boundary of $\mathcal{B}(\mathcal{C}, \epsilon)$ (since $\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w})$ is on the boundary of $\mathcal{C}$) implies that

$$\|\mu \boldsymbol{w}\| \leq \|\boldsymbol{w}/\gamma_{\mathcal{C}}(\boldsymbol{w}) + \epsilon \boldsymbol{w}/\|\boldsymbol{w}\|\| = \|\boldsymbol{w}\|/\gamma_{\mathcal{C}}(\boldsymbol{w}) + \delta r/(16\kappa^2) \leq \|\boldsymbol{w}\|/\gamma_{\mathcal{C}}(\boldsymbol{w}) + \delta \|\boldsymbol{w}\|/(8\kappa^2),$$

where the last inequality follows by the fact that $\|\boldsymbol{w}\| \geq r/2$ (see (53)). This shows the second implication in (58). Let $\tilde{\alpha}$ and $\tilde{\beta}$ be the values of α and β in Algorithm 2 after the while-loop is over. By combining (57), which we recall holds when $\gamma_{\mathcal{C}}(\boldsymbol{w}) \geq 9/16$ or $\tilde{\gamma} \geq 1$, and (58), we get

$$\tilde{\gamma} \geq \gamma_{\mathcal{C}}(\boldsymbol{w}); \;\; \text{and} \;\; \frac{1}{\gamma_{\mathcal{C}}(\boldsymbol{w})} - \frac{1}{\tilde{\gamma}} = \frac{1}{\gamma_{\mathcal{C}}(\boldsymbol{w})} - \left(\tilde{\alpha} - \frac{\delta}{8\kappa^2}\right) \leq \tilde{\beta} - \tilde{\alpha} + \frac{\delta}{8\kappa} + \frac{\delta}{8\kappa^2} \leq \frac{3\delta}{8\kappa^2}, \tag{59}$$

where the last inequality follows by the fact that the while-loop in Algorithm 2 terminates when $\beta - \alpha \leq \delta/(8\kappa^2)$. Multiplying both sides of (59) by $\tilde{\gamma} \cdot \gamma_{\mathcal{C}}(\boldsymbol{w})$, we get

$$\tilde{\gamma} - \gamma_{\mathcal{C}}(\boldsymbol{w}) \leq \tilde{\gamma} \cdot \gamma_{\mathcal{C}}(\boldsymbol{w}) \cdot 3\delta/(8\kappa^2). \tag{60}$$

Using (59), we get that $\tilde{\gamma} \leq (1/\gamma_{\mathcal{C}}(\boldsymbol{w}) - 3\delta/(8\kappa^2))^{-1}$. Plugging this into (60), we get

$$\tilde{\gamma} - \gamma_{\mathcal{C}}(\boldsymbol{w}) \leq \frac{\gamma_{\mathcal{C}}(\boldsymbol{w}) \cdot 3\delta/(8\kappa^2)}{1/\gamma_{\mathcal{C}}(\boldsymbol{w}) - 3\delta/(8\kappa^2)} \overset{(a)}{\leq} \frac{6\kappa/5 \cdot 3\delta/(8\kappa^2)}{5/(6\kappa) - 3\delta/(8\kappa^2)} \leq \delta,$$

where (a) follows by the fact that $\gamma_{\mathcal{C}}(\boldsymbol{w}) \leq \|\boldsymbol{w}\|/r \leq 6\kappa/5$, for all $\boldsymbol{w} \in \mathcal{B}(6R/5)$, and the last inequality follows by the fact that $\delta \in (0,1)$ and $\kappa \geq 1$.

We now consider the computational complexity. Note that at every iteration of the while-loop in Line 7 of Algorithm 2, the difference $\beta - \alpha$ halves. Thus, since $\beta - \alpha$ is initially equal to 2, the algorithm terminates (i.e. when $\beta - \alpha \leq \delta/(8\kappa^2)$) after at most $\lceil \log_2((4\kappa)^2/\delta) \rceil + 1$ steps. ■

H.2. Proof of Proposition 11 (Approximate LO Oracle on $\mathcal{C}^\circ$)

Proof of Proposition 11. Let ε and $\tilde{\gamma}$ be as in Algorithm 3. First, suppose that $\tilde{\gamma} \geq 1$. We will show (17) using the result of Lemma 29 with $(f, \tilde{f}) = (\gamma_{\mathcal{C}}, \mathsf{GAU}_{\mathcal{C}}(\cdot; \varepsilon))$ and $(\nu_1, \nu_2, \boldsymbol{w})$ as in Algorithm 3. First, note that by our choice of ν_1 and ε, the technical condition $\varepsilon \leq \nu_1 \sqrt{d} L$ under which Lemma 29 holds translates to $\delta \leq 10 d^{3/2} \kappa$. This holds since $\delta \in (0,1)$ and $\kappa, d \geq 1$. We also need to check the technical condition $\|f - \tilde{f}\|_\infty \leq \varepsilon$ of Lemma 29. We note that in Algorithm 3 we only ever evaluate $\tilde{f} = \mathsf{GAU}_{\mathcal{C}}(\cdot; \varepsilon)$ at the points $(\boldsymbol{w}_i, \boldsymbol{w}_i')_{i \in [d]}$, and so we only need to check the condition

$$|\gamma_{\mathcal{C}}(\boldsymbol{w}_i) - \mathsf{GAU}_{\mathcal{C}}(\boldsymbol{w}_i; \varepsilon)| \vee |\gamma_{\mathcal{C}}(\boldsymbol{w}_i') - \mathsf{GAU}_{\mathcal{C}}(\boldsymbol{w}_i'; \varepsilon)| \leq \varepsilon, \quad \forall i \in [d]. \tag{61}$$

The condition would follow from Lemma 10 if $\boldsymbol{w}_i$ and $\boldsymbol{w}_i'$ satisfy $\gamma_{\mathcal{C}}(\boldsymbol{w}_i) \wedge \gamma_{\mathcal{C}}(\boldsymbol{w}_i') \geq 9/16$ and $\boldsymbol{w}_i, \boldsymbol{w}_i' \in \mathcal{B}(6R/5)$, for all $i \in [d]$. Let $i \in [d]$ and $\boldsymbol{v} \in \{\boldsymbol{w}_i, \boldsymbol{w}_i'\}$. By definition of $\boldsymbol{w}_i$, $\boldsymbol{w}_i'$, and $\boldsymbol{u}$ in Algorithm 3, we have

$$\|\boldsymbol{w} - \boldsymbol{v}\| \leq \sqrt{d} \|\boldsymbol{w} - \boldsymbol{u}\|_\infty + \sqrt{d} \|\boldsymbol{u} - \boldsymbol{v}\|_\infty \leq \sqrt{d} \nu_1 + \sqrt{d} \nu_2 \leq \frac{11 r \delta}{100}, \tag{62}$$

where the last inequality follows by our choice of ν_1 and ν_2. Combining (62) with the fact that $\boldsymbol{w} \in \mathcal{B}(R)$ and triangular inequality, we get $\|\boldsymbol{v}\| \le \|\boldsymbol{w}\| + 11r\delta/100 \le 6R/5$, and so $\boldsymbol{v} \in \mathcal{B}(6R/5)$. On the other hand, by the fact that $\gamma_{\mathcal{C}} = \sigma_{\mathcal{C}^\circ}$ and the sub-additivity of the support function (Lemma 42-(e)), we have

$$\gamma_{\mathcal{C}}(\boldsymbol{v}) = \sigma_{\mathcal{C}^\circ}(\boldsymbol{v}) \ge \sigma_{\mathcal{C}^\circ}(\boldsymbol{w}) - \sigma_{\mathcal{C}^\circ}(\boldsymbol{w}-\boldsymbol{v}) = \gamma_{\mathcal{C}}(\boldsymbol{w}) - \gamma_{\mathcal{C}}(\boldsymbol{w}-\boldsymbol{v}) \overset{(a)}{\ge} \gamma_{\mathcal{C}}(\boldsymbol{w}) - \frac{11\delta}{100} \ge \frac{9}{16},$$

where (a) follows by the fact that $\gamma_{\mathcal{C}}(\boldsymbol{w}-\boldsymbol{v}) \le \|\boldsymbol{w}-\boldsymbol{v}\|/r$ (Lemma 42-(c)) and (62), and the last inequality follows by the fact that $\delta \le 1/3$ and $\gamma_{\mathcal{C}}(\boldsymbol{w}) \ge 2/3$; the latter is because $\tilde{\gamma} \ge 1$ and $\gamma_{\mathcal{C}}(\boldsymbol{w}) \ge \tilde{\gamma} - \delta$ (Lemma 10). Therefore, by Lemma 10, (61) holds, and so does the result of Lemma 29 with $(f, \tilde{f}) = (\gamma_{\mathcal{C}}, \mathsf{GAU}_{\mathcal{C}}(\cdot;\varepsilon))$ and $(\nu_1, \nu_2, \boldsymbol{w})$ as in Algorithm 3. We will now use this fact to show (17).

By Lemma 42 (points (a) and (b)), we have $\partial\gamma_{\mathcal{C}}(\boldsymbol{x}) \subseteq \mathcal{C}^\circ$ for any $\boldsymbol{x} \in \mathbb{R}^d$, and so we get

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d} \|\partial\gamma_{\mathcal{C}}(\boldsymbol{x})\| \le \sup_{\boldsymbol{y}\in\mathcal{C}^\circ} \|\boldsymbol{y}\| \le \frac{1}{r},$$

where the last inequality follows by Lemma 42-(c). This implies that the Lipschitz constant L in Lemma 29 can be set to $1/r$. Furthermore, by our choice of ε and ν_1 in Algorithm 3, we have $\varepsilon = \nu_1^3/(R^2 r\sqrt{d})$ and so Lemma 29 implies that there exists a non-negative random variable X satisfying $\mathbb{E}[X] \le 3\nu_1 d/(rR)$, and for all $\boldsymbol{u} \in \mathbb{R}^d$

$$\gamma_{\mathcal{C}}(\boldsymbol{u}) \ge \gamma_{\mathcal{C}}(\boldsymbol{w}) + \langle \tilde{\boldsymbol{s}}, \boldsymbol{u}-\boldsymbol{w}\rangle - X\|\boldsymbol{w}-\boldsymbol{u}\|_\infty - 4d\nu_1/r,$$

where $\tilde{\boldsymbol{s}}$ is as in Algorithm 3. Thus, with $\Delta \coloneqq 2RX + 4d\nu_1/r \ge 0$, we get

$$\gamma_{\mathcal{C}}(\boldsymbol{u}) \ge \gamma_{\mathcal{C}}(\boldsymbol{w}) + \langle \tilde{\boldsymbol{s}}, \boldsymbol{u}-\boldsymbol{w}\rangle - \Delta \cdot \max(1, \|\boldsymbol{u}\|/R), \quad \text{and} \quad \mathbb{E}[\Delta] \le 10d\nu_1/r, \tag{63}$$

where we used the fact that $\boldsymbol{w} \in \mathcal{B}(R)$. Further, Lemma 29 implies $\|\tilde{\boldsymbol{s}}\|_\infty \le 1/r + d^{1/4}\sqrt{L\varepsilon/\nu_1}$, $\|\tilde{\boldsymbol{s}}\| \le X + 1/r$, and $\|\tilde{\boldsymbol{s}}\|^2 \le \left(4/r + d^{1/4}\sqrt{L\varepsilon/\nu_1}\right) X + 1/r^2$, and so

$$\|\tilde{\boldsymbol{s}}\|_\infty \le d^{5/4}\sqrt{L\varepsilon/\nu_1} + 1/r; \quad \|\tilde{\boldsymbol{s}}\| \le \Delta/R + 1/r; \tag{64}$$

$$\text{and} \quad \|\tilde{\boldsymbol{s}}\|^2 \le \left(2/r + d^{1/4}\sqrt{L\varepsilon/\nu_1}\right)\Delta/R + 1/r^2. \tag{65}$$

Now, since $\varepsilon = \nu_1^3/(R^2 r\sqrt{d})$ and $\nu_1 = \delta r/(10d)$, we get that $10d\nu_1/r = \delta$ and $d^{1/4}\sqrt{L\varepsilon/\nu_1} = \delta/(10Rd)$. Plugging these into (63) and (65), we get (17) for the case when $\tilde{\gamma} \ge 1$.

It remain to show that $\|\tilde{\boldsymbol{s}}\|_\infty < +\infty$ almost surely and bound Δ. We do not assume that $\tilde{\gamma} \ge 1$ anymore. Let $i \in [n]$ and $(\boldsymbol{w}_i, \boldsymbol{w}_i')$ be as in Algorithm 3. Let $\boldsymbol{v} \in \{\boldsymbol{w}_i, \boldsymbol{w}_i'\}$ and suppose that $\mathsf{GAU}_{\mathcal{C}}(\boldsymbol{w};\varepsilon) \ge 1$, then by Lemma 10 and Lemma 42-(c), we have

$$1 \le \mathsf{GAU}_{\mathcal{C}}(\boldsymbol{v};\varepsilon) \le \gamma_{\mathcal{C}}(\boldsymbol{v}) + \varepsilon \le \|\boldsymbol{v}\|/r + \varepsilon \le \frac{R + \sqrt{d}\nu_1 + \sqrt{d}\nu_2}{r} + \varepsilon < +\infty.$$

Alternatively, we have $0 \le \mathsf{GAU}_{\mathcal{C}}(\boldsymbol{v};\varepsilon) < 1$. Therefore, for all $i \in [d]$, we have

$$\begin{aligned} |\tilde{s}_i| = \frac{|\mathsf{GAU}_{\mathcal{C}}(\boldsymbol{w}_i';\varepsilon) - \mathsf{GAU}_{\mathcal{C}}(\boldsymbol{w}_i;\varepsilon)|}{2\nu_2} &\le \frac{R + \sqrt{d}\nu_1 + \sqrt{d}\nu_2}{\nu_2 r} + \frac{\varepsilon}{\nu_2} \\ &= \frac{100d^{5/2}\kappa^2}{\delta^2 r} + \frac{10d^2\kappa}{\delta r} + \frac{\delta}{10dR} + \frac{\sqrt{d}}{r}, \\ &\le \frac{112d^{5/2}\kappa^2}{r\delta^2} < +\infty. \end{aligned} \tag{66}$$

Finally, by Lemma 29, we have $X \leq \|\tilde{\boldsymbol{s}}\|_1 + d/r$ and so combining this with (66), we get

$$\Delta = 2RX + 4d\nu_1/r \leq \frac{224d^4\kappa^3}{\delta^2} + 2d\kappa + \frac{4\delta}{10} \leq \frac{15^2 d^4\kappa^3}{\delta^2},$$

where the last inequality follows by the fact that $\delta \leq 1/3$. This completes the proof. ■

H.3. Proof of Lemma 13 (Efficient Stochastic LO Oracle on $\mathcal{C}^\circ$)

Proof of Lemma 13. Let $I \in [d]$ be the random variable in Algorithm 4 generated during the call to $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$ in the lemma's statement. Note that the approximate Gauge function $\mathsf{GAU}_{\mathcal{C}}$ (Alg. 2) is deterministic and so $\hat{\gamma} = \tilde{\gamma}$. On the other hand, by definition of $\tilde{\boldsymbol{s}}$ and $\hat{\boldsymbol{s}}$ in Algorithms 3 and 4, respectively, we have

$$\mathbb{E}[\hat{\boldsymbol{s}}] = \mathbb{E}[\hat{s}_I \cdot \boldsymbol{e}_I] = \sum_{i=1}^d d^{-1}\mathbb{E}[\hat{s}_i \mid I = i] \cdot \boldsymbol{e}_i \overset{(*)}{=} \sum_{i=1}^d d^{-1}\mathbb{E}[d\tilde{s}_i] \cdot \boldsymbol{e}_i = \tilde{\boldsymbol{s}},$$

where $(*)$ follows by the fact that, conditioned on $I = i$, $\hat{s}_i/d$ has the same distribution as $\tilde{s}_i$. Similarly,

$$\mathbb{E}[\|\hat{\boldsymbol{s}}\|^2] = \mathbb{E}[\|\hat{s}_I \boldsymbol{e}_I\|^2] = \sum_{i=1}^d d^{-1}\mathbb{E}[\hat{s}_i^2 \boldsymbol{e}_i \mid I = i] = \sum_{i=1}^d d^{-1}\mathbb{E}[d^2\tilde{s}_i^2 \boldsymbol{e}_i] = d\mathbb{E}[\|\tilde{\boldsymbol{s}}\|^2] \leq d \cdot \left(\frac{1}{r} + \frac{\delta}{R}\right)^2,$$

where the last inequality follows by the fact that $\|\tilde{\boldsymbol{s}}\|^2 \leq (2/r + \delta/R)\Delta/R + 1/r^2$ (Proposition 11), where $\Delta \geq 0$ is a random variable satisfying $\mathbb{E}[\Delta] \leq \delta$. Finally, the fact that $\|\hat{\boldsymbol{s}}\| < +\infty$ a.s. follows from the fact that $\|\tilde{\boldsymbol{s}}\|_\infty < +\infty$ a.s. (Prop. 11), and for any $i \in [d]$, conditioned on $I = i$, $\hat{s}_i/d$ has the same distribution as $\tilde{s}_i$. ■

Appendix I. Proofs of the Regret Bounds and Convergence Rates

I.1. Proof of Lemma 7 (Instantaneous Regret Bound)

To avoid expensive projections our main algorithm (Alg. 1) makes use of surrogate losses of the form $\tilde{\ell}(\boldsymbol{w}) := \langle \boldsymbol{g}, \boldsymbol{w}\rangle + bS_{\mathcal{C}}(\boldsymbol{w})$, $\boldsymbol{w} \in \mathbb{R}^d$ and $b \geq 0$. The choice of such a surrogate loss function is inspired by existing constrained-to-unconstrained reductions in OCO due to Cutkosky and Orabona (2018); Cutkosky (2020). We will use the approximate optimization Oracle $\mathsf{OPT}_{\mathcal{C}^\circ}$ from App. B to compute approximate subgradients of such surrogate losses. In particular, $\mathsf{OPT}_{\mathcal{C}^\circ}$ (Algorithm 3) guarantees the following:

Lemma 30 *Let $\boldsymbol{g} \in \mathbb{R}^d$, $b \geq 0$, and $\delta \in (0, 1/3)$. Let $S_{\mathcal{C}}$ be the Gauge distance function in (3) and define $\tilde{\ell}(\boldsymbol{w}) := \langle \boldsymbol{g}, \boldsymbol{w}\rangle + bS_{\mathcal{C}}(\boldsymbol{w})$. Then, for $\boldsymbol{w} \in \mathcal{B}(R)$, the output $(\tilde{\gamma}, \tilde{\boldsymbol{s}})$ of Alg. 3 with input $(\boldsymbol{w}, \delta)$ satisfies (17) and*

$$\forall \boldsymbol{u} \in \mathcal{B}(R),\ \ \tilde{\ell}(\boldsymbol{u}) \geq \tilde{\ell}(\boldsymbol{w}) + \langle \boldsymbol{g} + b\boldsymbol{\nu}, \boldsymbol{u} - \boldsymbol{w}\rangle - b \cdot (\Delta + \delta) \cdot \mathbb{I}_{\{\tilde{\gamma} \geq 1\}}, \quad \textit{where } \boldsymbol{\nu} := \mathbb{I}_{\{\tilde{\gamma}\geq 1\}}\tilde{\boldsymbol{s}},$$

and $\Delta \in [0, 15^2 d^4\kappa^3\delta^{-2}]$ is the same random variable satisfying (17); in particular, $\mathbb{E}[\Delta] \leq \delta$.

Proof First suppose that $\tilde{\gamma} < 1$. If $\gamma_{\mathcal{C}}(\boldsymbol{w}) \geq 9/16$, then by Lemma 10, we have $\gamma_{\mathcal{C}}(\boldsymbol{w}) \leq \tilde{\gamma} < 1$. Alternatively, $\gamma_{\mathcal{C}}(\boldsymbol{w}) < 9/16 < 1$. Therefore, by Lemma 6, $\partial S_{\mathcal{C}}(\boldsymbol{w}) = \{\mathbf{0}\}$, and so $\boldsymbol{g} \in \partial\tilde{\ell}(\boldsymbol{w})$. Since the function $\tilde{\ell}$ is convex, it follows that

$$\text{[case where } \tilde{\gamma} < 1] \quad \forall \boldsymbol{u} \in \mathcal{B}(R), \;\; \tilde{\ell}(\boldsymbol{u}) \geq \tilde{\ell}(\boldsymbol{w}) + \langle \boldsymbol{g}, \boldsymbol{u} - \boldsymbol{w} \rangle. \tag{67}$$

Now, suppose that $\tilde{\gamma} \geq 1$. In this case, by Lemma 10, we have $\gamma_{\mathcal{C}}(\boldsymbol{w}) \leq \tilde{\gamma} \leq \gamma_{\mathcal{C}}(\boldsymbol{w}) + \delta$. Combining this with the fact that $S_{\mathcal{C}}(\boldsymbol{w}) = 0 \vee (\gamma_{\mathcal{C}}(\boldsymbol{w}) - 1)$ (Lemma 6), we get

$$\gamma_{\mathcal{C}}(\boldsymbol{w}) - 1 \geq \tilde{\gamma} - 1 - \delta \overset{(a)}{\geq} 0 \vee (\gamma_{\mathcal{C}}(\boldsymbol{w}) - 1) - \delta = S_{\mathcal{C}}(\boldsymbol{w}) - \delta, \tag{68}$$

where (a) follows from the fact that $\tilde{\gamma} \geq 1$ (by assumption) and $\tilde{\gamma} \geq \gamma_{\mathcal{C}}(\boldsymbol{w})$. Using this, we get

$$\begin{aligned}
\text{[case where } \tilde{\gamma} \geq 1] \quad \forall \boldsymbol{u} \in \mathcal{B}(R), \;\; \tilde{\ell}(\boldsymbol{u}) &= \langle \boldsymbol{g}, \boldsymbol{u} \rangle + 0 \vee (b\gamma_{\mathcal{C}}(\boldsymbol{u}) - b), \\
&\geq \langle \boldsymbol{g}, \boldsymbol{u} \rangle + b\gamma_{\mathcal{C}}(\boldsymbol{u}) - b, \\
&\geq \langle \boldsymbol{g}, \boldsymbol{u} \rangle + b\gamma_{\mathcal{C}}(\boldsymbol{w}) - b + \langle b\tilde{\boldsymbol{s}}, \boldsymbol{u} - \boldsymbol{w} \rangle - b\Delta, && (69) \\
&\geq \langle \boldsymbol{g}, \boldsymbol{u} \rangle + bS_{\mathcal{C}}(\boldsymbol{w}) + \langle b\tilde{\boldsymbol{s}}, \boldsymbol{u} - \boldsymbol{w} \rangle - b\Delta - b\delta, && (70) \\
&= \tilde{\ell}(\boldsymbol{w}) + \langle \boldsymbol{g} + b\tilde{\boldsymbol{s}}, \boldsymbol{u} - \boldsymbol{w} \rangle - b\Delta - b\delta, \\
&= \tilde{\ell}(\boldsymbol{w}) + \langle \tilde{\boldsymbol{g}} + b\boldsymbol{\nu}, \boldsymbol{u} - \boldsymbol{w} \rangle - b\Delta - b\delta, && (71)
\end{aligned}$$

where (69) follows by the fact that $\tilde{\gamma} \geq 1$ and Proposition 11; (70) follows by (68); and (71) follows by the fact that $\tilde{\boldsymbol{s}} = \boldsymbol{\nu}$ since $\tilde{\gamma} \geq 1$. Combining (67) and (71) implies the desired result. ■

Lemma 30 shows that for any $t \geq 1$, the vector $\tilde{\boldsymbol{g}}_t$ in Algorithm 1 is an approximate subgradient of the surrogate loss $\tilde{\ell}_t(\cdot) \coloneqq \langle \boldsymbol{g}_t, \cdot \rangle - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle S_{\mathcal{C}}(\cdot)$, where $\boldsymbol{g}_t \in \partial\ell_t(\boldsymbol{x}_t)$ and $\boldsymbol{x}_t$ is the tth iterate of Alg. 1. We now use this to prove Lemma 7. We actually state and prove a slight generalization of Lemma 7, which will be useful when considering the stochastic optimization setting of Section C.3. In this generalization, we assume that for some $R' \in [r, R]$,

$$\mathcal{B}(r) \subseteq \mathcal{C} \subseteq \mathcal{B}(R') \subseteq \mathcal{B}(R). \tag{72}$$

Lemma 31 *Let $\kappa \coloneqq R'/r$ and A be any OCO algorithm on $\mathcal{B}(R)$, for $r, R, R' > 0$ as in (72). Further, for $t \geq 1$, let $\boldsymbol{w}_t, \tilde{\boldsymbol{g}}_t, \boldsymbol{x}_t$, and $\boldsymbol{g}_t \in \partial\ell_t(\boldsymbol{x}_t)$ be as in Alg. 1 with $\mathcal{O}_{\mathcal{C}^\circ} \equiv \mathsf{OPT}_{\mathcal{C}^\circ}$ and any tolerance sequence $(\delta_s) \subset (0, 1/3)$. Then, $(\boldsymbol{x}_t) \subset \mathcal{C}$, and for all $t \geq 1$, there exists a variable $\Delta_t \in \left[0, 15^2 d^4 (R/r)^3 \delta_t^{-2}\right]$ s.t. $\mathbb{E}_{t-1}[\Delta_t] \leq \delta_t$, $\|\tilde{\boldsymbol{g}}_t\| \leq (1 + \Delta_t + \kappa)\|\boldsymbol{g}_t\|$, and*

$$\forall \boldsymbol{x} \in \mathcal{C}, \quad \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \leq \tilde{\ell}_t(\boldsymbol{w}_t) - \tilde{\ell}_t(\boldsymbol{x}) + \delta_t R\|\boldsymbol{g}_t\| \leq \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x} \rangle + (2\delta_t + \Delta_t)R\|\boldsymbol{g}_t\|. \tag{73}$$

Proof of Lemma 7. Let γ_t, $\boldsymbol{\nu}_t$, $\boldsymbol{w}_t$, $\tilde{\boldsymbol{g}}_t$, and $\boldsymbol{x}_t$ be as in Algorithm 1. We will first show that

$$-R\|\boldsymbol{g}_t\| \leq \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \cdot \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \leq 0 \tag{74}$$

First suppose that $\gamma_t < 1$. In this case, $\boldsymbol{x}_t = \boldsymbol{w}_t$, and so

$$-R\|\boldsymbol{g}_t\| \leq \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \cdot \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle = \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \cdot \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle \leq 0.$$

Now suppose that $\gamma_t \geq 1$, then $\mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \cdot \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle = \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \cdot \langle \boldsymbol{g}_t, \boldsymbol{w}_t / \gamma_t \rangle$ and so

$$0 \geq \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \cdot \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \geq \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle \geq -R \|\boldsymbol{g}_t\|,$$

where the last inequality follows by the fact $\boldsymbol{w}_t \in \mathcal{B}(R)$. This shows (74). Now recall the definition of the surrogate function $\tilde{\ell}_t : \mathcal{B}(R) \to \mathbb{R}$:

$$\tilde{\ell}_t(\boldsymbol{w}) \coloneqq \langle \boldsymbol{g}_t, \boldsymbol{w} \rangle - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \cdot \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \cdot S_{\mathcal{C}}(\boldsymbol{w}),$$

where $S_{\mathcal{C}}$ is as in (3). By (74) and Lemma 30, there exists a r.v. $\Delta_t \in [0, \frac{15^2 d^4 \kappa^3}{\delta^2}]$ such that $\mathbb{E}_{t-1}[\Delta_t] \leq \delta_t$ and

$$\forall \boldsymbol{x} \in \mathcal{C}, \quad \tilde{\ell}_t(\boldsymbol{w}_t) - \tilde{\ell}_t(\boldsymbol{x}) \leq \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x} \rangle + R \|\boldsymbol{g}_t\| \cdot (\Delta_t + \delta_t). \tag{75}$$

It remains to show that $\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \leq \tilde{\ell}_t(\boldsymbol{w}_t) - \tilde{\ell}_t(\boldsymbol{x}) + \delta_t R \|\boldsymbol{g}_t\|$, for all $t \geq 1$ and $\boldsymbol{x} \in \mathcal{C}$. First, note that for all $\boldsymbol{x} \in \mathcal{C}$, we have $S_{\mathcal{C}}(\boldsymbol{x}) = 0$, and so

$$\tilde{\ell}_t(\boldsymbol{x}) = \langle \boldsymbol{g}_t, \boldsymbol{x} \rangle, \quad \forall \boldsymbol{x} \in \mathcal{C}. \tag{76}$$

We will now compare $\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle$ to $\tilde{\ell}_t(\boldsymbol{w}_t)$ by considering cases. Suppose that $\gamma_t < 1$. In this case, we have $\boldsymbol{x}_t = \boldsymbol{w}_t$ and so

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle = \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle = \tilde{\ell}_t(\boldsymbol{w}_t). \quad \text{[case where } \gamma_t < 1\text{]} \tag{77}$$

Now suppose that $\gamma_t \geq 1$ and $\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle \geq 0$. In this case, we have $\boldsymbol{x}_t = \boldsymbol{w}_t / \gamma_t$, and so

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle = \langle \boldsymbol{g}_t, \boldsymbol{w}_t / \gamma_t \rangle \leq \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle = \tilde{\ell}_t(\boldsymbol{w}_t). \quad \text{[case where } \gamma_t \geq 1, \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle \geq 0\text{]} \tag{78}$$

Now suppose that $\gamma_t \geq 1$, $\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0$, and $\boldsymbol{w}_t \in \mathcal{C}$. We note that this implies that $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \leq 1$ and so $S_{\mathcal{C}}(\boldsymbol{w}_t) = 0$ (Lemma 6). Thus, $\tilde{\ell}_t(\boldsymbol{w}_t) = \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle$. On the other hand, by Lemma 10 we have $\gamma_t \leq \gamma_{\mathcal{C}}(\boldsymbol{w}_t) + \delta_t \leq 1 + \delta_t$, and so since $\boldsymbol{x}_t = \boldsymbol{w}_t / \gamma_t$, we have

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \leq \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle / (1 + \delta_t) = \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle - \delta_t \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle / (1 + \delta_t).$$

Thus, since $\tilde{\ell}_t(\boldsymbol{w}_t) = \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle$, the previous display implies that

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \leq \tilde{\ell}_t(\boldsymbol{w}_t) + \delta_t R \|\boldsymbol{g}_t\|. \quad \text{[case where } \gamma_t \geq 1, \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0, \boldsymbol{w}_t \in \mathcal{C}\text{]} \tag{79}$$

We now consider the last case where $\gamma_t \geq 1$, $\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0$, and $\boldsymbol{w}_t \notin \mathcal{C}$. We note that this implies that $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \geq 1$. By the fact that $\gamma_t \leq \gamma_{\mathcal{C}}(\boldsymbol{w}_t) + \delta_t$ (Lemma 10), we have

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \leq \frac{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle}{\gamma_{\mathcal{C}}(\boldsymbol{w}_t) + \delta_t} = \frac{\langle \boldsymbol{g}_t, \boldsymbol{w}_t / \gamma_{\mathcal{C}}(\boldsymbol{w}_t) \rangle}{1 + \delta_t / \gamma_{\mathcal{C}}(\boldsymbol{w}_t)} = \langle \boldsymbol{g}_t, \boldsymbol{w}_t / \gamma_{\mathcal{C}}(\boldsymbol{w}_t) \rangle - \frac{\delta_t / \gamma_{\mathcal{C}}(\boldsymbol{w}_t)}{1 + \delta_t / \gamma_{\mathcal{C}}(\boldsymbol{w}_t)} \langle \boldsymbol{g}_t, \boldsymbol{w}_t / \gamma_{\mathcal{C}}(\boldsymbol{w}_t) \rangle, \tag{80}$$

Thus, since $S_{\mathcal{C}}(\boldsymbol{w}_t) = \gamma_{\mathcal{C}}(\boldsymbol{w}_t) - 1$ (by Lemma 6 and $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \geq 1$), we get

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \cdot S_{\mathcal{C}}(\boldsymbol{w}_t) \leq \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle - \frac{\delta_t \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle}{1 + \delta_t / \gamma_{\mathcal{C}}(\boldsymbol{w}_t)} - \langle \boldsymbol{g}_t, \boldsymbol{w}_t / \gamma_{\mathcal{C}}(\boldsymbol{w}_t) \rangle + \frac{\delta_t / \gamma_{\mathcal{C}}(\boldsymbol{w}_t)}{1 + \delta_t / \gamma_{\mathcal{C}}(\boldsymbol{w}_t)} \langle \boldsymbol{g}_t, \boldsymbol{w}_t / \gamma_{\mathcal{C}}(\boldsymbol{w}_t) \rangle.$$

Adding this together with (80) and using the fact that $\|\boldsymbol{w}_t\| \le R$, we get

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle + \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle S_{\mathcal{C}}(\boldsymbol{w}_t) \le \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle + \delta_t R \|\boldsymbol{g}_t\|,$$

and so after rearranging and using that $\tilde{\ell}_t(\boldsymbol{w}_t) = \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle - \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle S_{\mathcal{C}}(\boldsymbol{w}_t)$, we get

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \le \tilde{\ell}_t(\boldsymbol{w}_t) + \delta_t R \|\boldsymbol{g}_t\|. \quad \text{[case where } \gamma_t \ge 1, \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0, \boldsymbol{w}_t \notin \mathcal{C}] \tag{81}$$

By combining, (77), (78), (79), and (81), we obtain:

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \le \tilde{\ell}_t(\boldsymbol{w}_t) + \delta_t R \|\boldsymbol{g}_t\|.$$

Combining this with (75) and (76), we get

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \le \tilde{\ell}_t(\boldsymbol{w}_t) - \tilde{\ell}_t(\boldsymbol{x}) + \delta_t R \|\boldsymbol{g}_t\| \le \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x} \rangle + (2\delta_t + \Delta_t) R \|\boldsymbol{g}_t\|, \quad \forall \boldsymbol{x} \in \mathcal{C}.$$

This shows (73). It remains to bound $\|\tilde{\boldsymbol{g}}_t\|$ in terms of $\|\boldsymbol{g}_t\|$ and show that $\boldsymbol{x}_t \in \mathcal{C}$. When $\gamma_t < 1$, we have $\tilde{\boldsymbol{g}}_t = \boldsymbol{g}_t$ and $\boldsymbol{x}_t = \boldsymbol{w}_t$, and so $\|\tilde{\boldsymbol{g}}_t\| = \|\boldsymbol{g}_t\|$. Furthermore, we also have that $\boldsymbol{x}_t \in \mathcal{C}$. In fact, if $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \le 9/16$, then by definition of the Gauge function we have $\boldsymbol{w}_t \in \mathcal{C}$ and the same holds for $\boldsymbol{x}_t$ (since $\boldsymbol{x}_t = \boldsymbol{w}_t$). On the other hand, if $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \ge 9/16$, then by Lemma 10, we have $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \le \gamma_t < 1$ ($\gamma_t < 1$ is the case we are currently considering), and so $\boldsymbol{x}_t = \boldsymbol{w}_t \in \mathcal{C}$.

Now suppose that $\gamma_t \ge 1$. In this case, we have $\boldsymbol{x}_t = \boldsymbol{w}_t/\gamma_t$ and $\tilde{\boldsymbol{g}}_t = \boldsymbol{g}_t - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \boldsymbol{\nu}_t$. The latter fact together with the positive homogeneity of the Gauge function (Lemma 42-(a,b)) imply that $\gamma_{\mathcal{C}}(\boldsymbol{x}_t) = \gamma_{\mathcal{C}}(\boldsymbol{w}_t/\gamma_t) = \gamma_{\mathcal{C}}(\boldsymbol{w}_t)/\gamma_t \le \gamma_{\mathcal{C}}(\boldsymbol{w}_t)/\gamma_{\mathcal{C}}(\boldsymbol{w}_t) = 1$ (since $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \le \gamma_t$ by Lemma 10), and so $\boldsymbol{x}_t \in \mathcal{C}$. Using that $\boldsymbol{x}_t \in \mathcal{C}$ (which implies that $\|\boldsymbol{x}_t\| \le R'$) and that $\tilde{\boldsymbol{g}}_t = \boldsymbol{g}_t - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \boldsymbol{\nu}_t$, we get

$$\|\tilde{\boldsymbol{g}}_t\| = \|\boldsymbol{g}_t - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \boldsymbol{\nu}_t\| \le \|\boldsymbol{g}_t\|(1 + R'\|\boldsymbol{\nu}_t\|) \le (1 + \Delta_t + \kappa)\|\boldsymbol{g}_t\|,$$

where the last inequality follows the fact that $\|\boldsymbol{\nu}_t\| \le \Delta_t/R + 1/r$ (by (17) which is implied by Lemma 30). ■

Proof of Lemma 7. Follows from Lemma 31 with $R' = R$. ■

I.2. Proof of Theorem 8 (Regret Bound in High Probability using $\mathsf{OPT}_{\mathcal{C}^\circ}$)

Proof By Lemma 7, we have, for all $\boldsymbol{w}$ and $t \ge 1$, $\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle \le \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x} \rangle + (2\delta_t + \Delta_t) R \|\boldsymbol{g}_t\|$, where $(\Delta_t) \subset \mathbb{R}_{\ge 0}$ is a sequence of positive random variables satisfying $\mathbb{E}_{t-1}[\Delta_t] \le \delta_t$, for all $t \in [T]$. Summing this inequality for $t = 1, \dots, T$, we obtain, for all $\boldsymbol{x} \in \mathcal{C}$,

$$\begin{aligned}
\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle &\le \sum_{t=1}^{T} \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x} \rangle + \sum_{t=1}^{T} (2\delta_t + \Delta_t) R \|\boldsymbol{g}_t\|, \\
&\le 2R \sqrt{2 \sum_{t=1}^{T} \|\tilde{\boldsymbol{g}}_t\|^2} + \sum_{t=1}^{T} (2\delta_t + \Delta_t) R \|\boldsymbol{g}_t\|, \qquad (82) \\
&\le 4(1+\kappa) R \sqrt{\sum_{t=1}^{T} \|\boldsymbol{g}_t\|^2} + \sum_{t=1}^{T} (2\delta_t + 5\Delta_t) R \|\boldsymbol{g}_t\|, \qquad (83)
\end{aligned}$$

where (82) follows by our choice of the subroutine A and the regret bound of FTRL-prox in Proposition 25, and the last inequality follows by the fact that $\|\tilde{\boldsymbol{g}}_t\| \leq (1+\kappa+\Delta_t)\|\boldsymbol{g}_t\|$ (Lem. 7) and the fact that $(a+b)^2 \leq 2a^2+2b^2$ for all $a, b \in \mathbb{R}_{>0}$. By Lemma 7, we also have $\boldsymbol{x}_t \in \mathcal{C}$, $\forall t \geq 1$.

We now instantiate (83) with the specific choice of tolerance sequence (δ_t) in theorem's statement, where we explicitly bound the right-most sum in (82) involving (Δ_t). For this, we let $X_t := \sum_{i=1}^t (\Delta_i - \bar{\delta}_i)$, where $\bar{\delta}_i := \mathbb{E}_{i-1}[\Delta_i] \leq \delta_i$. The process (X_t) is a martingale; that is, for all $i \geq 1$, we have, $\mathbb{E}_i[X_t] = X_i$, for all $i < t$. Thus, by Doob's martingale inequality (Durrett, 2019, Theorem 4.4.2) we have, for any $\rho \in (0,1)$ and $T \geq 1$:

$$\mathbb{P}\left[\sum_{t=1}^T \Delta_t \geq (1+1/\rho)\sum_{t=1}^T \delta_t\right] \leq \mathbb{P}\left[X_T \geq \sum_{t=1}^T \delta_t/\rho\right] \leq \mathbb{P}\left[\max_{t\leq T} X_t \geq \sum_{t=1}^T \delta_t/\rho\right] \leq \frac{\rho \mathbb{E}\left[X_T \vee 0\right]}{\sum_{t=1}^T \delta_t} \leq \rho,$$

where the last inequality follows by the fact that $\mathbb{E}[X_T \vee 0] \leq \mathbb{E}[\sum_{t=1}^T \Delta_t] \leq \sum_{t=1}^T \delta_t$. Using this and the fact that $\sum_{t=1}^\infty 1/t^2 \leq 2$ in combination with (83), we obtain (9). ■

I.3. Proof of Theorem 15 (Regret Bound in Expectation using $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$)

Proof Let $(\tilde{\gamma}_t, \tilde{\boldsymbol{s}}_t) = \mathsf{OPT}_{\mathcal{C}^\circ}(\boldsymbol{w}_t; \delta_t)$ and $(\hat{\gamma}_t, \hat{\boldsymbol{s}}_t) = \mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}(\boldsymbol{w}_t; \delta_t)$. Further, let $\tilde{\boldsymbol{\nu}}_t := \mathbb{I}_{\{\tilde{\gamma}_t \geq 1\}}\tilde{\boldsymbol{s}}_t$, and γ_t and $\boldsymbol{\nu}_t$ be as in Algorithm 1. Note that by Proposition 11 and Lemma 13, we have $\|\tilde{\boldsymbol{s}}_t\| \vee \|\hat{\boldsymbol{s}}_t\| < +\infty$ almost surely and so the expectations $\mathbb{E}_{t-1}[\mathbb{I}_{\{\tilde{\gamma}_t \geq 1\}}\tilde{\boldsymbol{s}}_t]$ and $\mathbb{E}_{t-1}[\mathbb{I}_{\{\hat{\gamma}_t \geq 1\}}\hat{\boldsymbol{s}}_t]$ are well defined. We also note that $\gamma_t = \hat{\gamma}_t = \tilde{\gamma}_t$, and γ_t, $\boldsymbol{w}_t$, $\boldsymbol{g}_t$, and $\boldsymbol{x}_t$ are all deterministic functions of the past ($\boldsymbol{w}_t$ it the output of FTRL-prox, which is a deterministic function of the past).

By Lemma 7, there exists a random variable $\Delta_t \geq 0$ satisfying $\mathbb{E}_{t-1}[\Delta_t]$ and such that for all $t \geq 1$,

$$\forall \boldsymbol{x} \in \mathcal{C}, \quad \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \leq \langle \boldsymbol{g}_t - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle < 0}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle \tilde{\boldsymbol{\nu}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle + (2\delta_t + \Delta_t)R\|\boldsymbol{g}_t\|, \tag{84}$$

Now by Lemma 13, and the law of total expectation, we have, for all $\boldsymbol{x} \in \mathcal{C}$,

$$\begin{aligned}
\mathbb{E}[\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle] &= \mathbb{E}[\mathbb{E}_{t-1}[\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle]], \\
&\leq \mathbb{E}[\mathbb{E}_{t-1}[\langle \boldsymbol{g}_t - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle < 0}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle \tilde{\boldsymbol{\nu}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle + (2\delta_t + \Delta_t)R\|\boldsymbol{g}_t\|]], \quad \text{(by (84))} \\
&\leq \mathbb{E}[\mathbb{E}_{t-1}[\langle \boldsymbol{g}_t - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle < 0}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle \tilde{\boldsymbol{\nu}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle] + 3\delta_t R\|\boldsymbol{g}_t\|], \\
&= \mathbb{E}[\mathbb{I}_{\{\tilde{\gamma}_t < 1\}} \cdot \mathbb{E}_{t-1}[\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle]] \\
&\quad + \mathbb{E}[\mathbb{I}_{\{\tilde{\gamma}_t \geq 1\}} \cdot \mathbb{E}_{t-1}[\langle \boldsymbol{g}_t - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle < 0}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle \tilde{\boldsymbol{s}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle] + 3\delta_t R\|\boldsymbol{g}_t\|], \\
&= \mathbb{E}[\mathbb{I}_{\{\tilde{\gamma}_t < 1\}} \cdot \mathbb{E}_{t-1}[\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle]] \\
&\quad + \mathbb{E}[\mathbb{I}_{\{\tilde{\gamma}_t \geq 1\}} \cdot \mathbb{E}_{t-1}[\langle \boldsymbol{g}_t - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle < 0}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle \hat{\boldsymbol{s}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle] + 3\delta_t R\|\boldsymbol{g}_t\|], \qquad (85) \\
&= \mathbb{E}[\mathbb{E}_{t-1}[\langle \boldsymbol{g}_t - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle < 0}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle \boldsymbol{\nu}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle] + 3\delta_t R\|\boldsymbol{g}_t\|], \\
&= \mathbb{E}[\langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle + 3\delta_t R\|\boldsymbol{g}_t\|], \qquad (86)
\end{aligned}$$

where (85) follows by the facts that $\tilde{\gamma}_t = \hat{\gamma}_t$ and $\mathbb{E}_{t-1}[\hat{\boldsymbol{s}}_t] = \mathbb{E}_{t-1}[\tilde{\boldsymbol{s}}_t]$, when $\tilde{\gamma}_t \geq 1$ (Lemma 13). Summing (86) for $t = 1, \ldots, T$, we obtain,

$$\begin{aligned}
\forall \boldsymbol{x} \in \mathcal{C}, \quad \sum_{t=1}^{T} \mathbb{E}[\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle] &\leq \mathbb{E}\left[\sum_{t=1}^{T} \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x} \rangle + 3\sum_{t=1}^{T} \delta_t R \|\boldsymbol{g}_t\|\right], \\
&\leq \mathbb{E}\left[2R\sqrt{2\sum_{t=1}^{T} \|\tilde{\boldsymbol{g}}_t\|^2} + 3\sum_{t=1}^{T} \delta_t R \|\boldsymbol{g}_t\|\right], \qquad (87) \\
&\leq 2R\sqrt{\mathbb{E}\left[2\sum_{t=1}^{T} \mathbb{E}_{t-1}\left[\|\tilde{\boldsymbol{g}}_t\|^2\right]\right]} + \mathbb{E}\left[3\sum_{t=1}^{T} \delta_t R \|\boldsymbol{g}_t\|\right], \qquad (88)
\end{aligned}$$

where the last step follows by Jensen's inequality and (87) follows by our choice of the subroutine A ($\equiv$FTRL-prox) and the regret bound of FTRL-prox in Proposition 25.

It remains to bound $\|\tilde{\boldsymbol{g}}_t\|$ in terms of $\|\boldsymbol{g}_t\|$ and show that $\boldsymbol{x}_t \in \mathcal{C}$. First, if $\gamma_t < 1$, then $(\tilde{\boldsymbol{g}}_t, \boldsymbol{x}_t) = (\boldsymbol{g}_t, \boldsymbol{w}_t)$, and so $\|\tilde{\boldsymbol{g}}_t\| \leq \|\boldsymbol{g}_t\|(1 + R\|\boldsymbol{s}_t\|)$. Furthermore, we also have that $\boldsymbol{x}_t \in \mathcal{C}$. In fact, if $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \leq 9/16$, then by definition of the Gauge function we have $\boldsymbol{w}_t \in \mathcal{C}$ and the same holds for $\boldsymbol{x}_t$ (since $\boldsymbol{x}_t = \boldsymbol{w}_t$). On the other hand, if $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \geq 9/16$, then by Lemma 10, we have $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \leq \gamma_t < 1$, and so $\boldsymbol{x}_t = \boldsymbol{w}_t \in \mathcal{C}$.

Now suppose that $\gamma_t \geq 1$. In this case, we have $\boldsymbol{x}_t = \boldsymbol{w}_t/\gamma_t$ and $\boldsymbol{\nu}_t = \boldsymbol{s}_t$. Using this, the triangular inequality, and Cauchy Schwarz, we get

$$\|\tilde{\boldsymbol{g}}_t\| = \|\langle \boldsymbol{g}_t, \boldsymbol{w} \rangle - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \cdot \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \cdot \boldsymbol{s}_t\| = \|\boldsymbol{g}_t\| + \|\boldsymbol{g}_t\|\|\boldsymbol{w}_t\|/\gamma_t \cdot \|\boldsymbol{s}_t\| \leq \|\boldsymbol{g}_t\|(1 + R\|\boldsymbol{s}_t\|),$$

where the last inequality follows by the fact that $\gamma_t \geq 1$. Therefore, using Lemma 13, we get $\|\tilde{\boldsymbol{g}}_t\| \leq \|\boldsymbol{g}_t\|(1 + \Delta_t + \kappa)$ and

$$\mathbb{E}_{t-1}[\|\tilde{\boldsymbol{g}}_t\|^2] \leq 2\|\boldsymbol{g}_t\|^2(1 + R^2\mathbb{E}_{t-1}[\|\boldsymbol{s}_t\|^2]) = 2\|\boldsymbol{g}_t\|^2(1 + R^2\mathbb{E}_{t-1}[\|\hat{\boldsymbol{s}}_t\|^2]) \leq 2\|\boldsymbol{g}_t\|^2(1 + d \cdot (\kappa + \delta_t)^2).$$

Plugging this into (88) leads to, for all $\boldsymbol{x} \in \mathcal{C}$,

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle] &\leq 4R\sqrt{\sum_{t=1}^{T}(1 + d \cdot (\kappa + \delta_t)^2) \cdot \mathbb{E}\left[\|\boldsymbol{g}_t\|^2\right]} + 3\mathbb{E}\left[\sum_{t=1}^{T} \delta_t R \cdot \|\boldsymbol{g}_t\|\right], \\
&\leq 4R\sqrt{\sum_{t=1}^{T}(1 + d \cdot (\kappa + \delta)^2) \cdot \mathbb{E}\left[\|\boldsymbol{g}_t\|^2\right]} + 6\delta R \cdot \mathbb{E}\left[\max_{t \in [T]} \|\boldsymbol{g}_t\|\right],
\end{aligned}$$

where in the last inequality we used that $\sum_{t=1}^{+\infty} 1/t^2 \leq 2$. Furthermore, when $\gamma_t \geq 1$, we have $\gamma_{\mathcal{C}}(\boldsymbol{x}_t) = \gamma_{\mathcal{C}}(\boldsymbol{w}_t/\gamma_t) = \gamma_{\mathcal{C}}(\boldsymbol{w}_t)/\gamma_t \leq \gamma_{\mathcal{C}}(\boldsymbol{w}_t)/\gamma_{\mathcal{C}}(\boldsymbol{w}_t) = 1$ (since $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \leq \gamma_t$ by Lemma 10), and so $\boldsymbol{x}_t \in \mathcal{C}$. ■

I.4. Proof of Theorem 16 (Regret Bound in High Probability using $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$)

To prove Theorem 16, we need the following extension of Lemma 13:

Lemma 32 *Let $\delta \in (0,1)$, $\boldsymbol{w} \in \mathcal{B}(R)$, and $\kappa := R/r$, with $r, R > 0$ as in (1). Further, let $(\tilde{\gamma}, \tilde{\boldsymbol{s}}) = \mathsf{OPT}_{\mathcal{C}^\circ}(\boldsymbol{w}; \delta)$ and $(\hat{\gamma}, \hat{\boldsymbol{s}}) = \mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}(\boldsymbol{w}; \delta)$ (Alg. 4). Then, $\hat{\gamma} = \tilde{\gamma}$, $\|\hat{\boldsymbol{s}}\| < +\infty$ a.s., and if $\hat{\gamma} \geq 1$ it follows that*

$$\mathbb{E}[\hat{\boldsymbol{s}}] = \mathbb{E}[\tilde{\boldsymbol{s}}], \quad \mathbb{E}[\|\hat{\boldsymbol{s}}\|^2] \leq d \cdot (1/r + \delta/R)^2, \tag{89}$$

$$\|\hat{\boldsymbol{s}}\|_\infty \leq d \cdot (1/r + \delta/R), \quad \|\hat{\boldsymbol{s}}\| \leq d \cdot (1/r + \delta/R), \quad \textit{and} \quad \mathbb{E}[\|\hat{\boldsymbol{s}}\|^4] \leq d^3 \cdot (1/r + \delta/R)^4. \tag{90}$$

Proof Lemma 13 implies (89). We now show (90). Proposition 11 implies that $\tilde{s}_i \leq \delta/R + 1/r$, for all $i \in [d]$. Let $I \in [d]$ be the random variable in Algorithm 4 generated during the call to $\mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$ in the lemma's statement. Since, conditioned on $I = i$, $\hat{s}_i/d$ has the same distribution as $\tilde{s}_i$ and $\hat{s}_i = 0$ when $I \neq i$, we get that $\hat{s}_i \leq d \cdot (1/r + \delta/R)$. This implies the first inequality in (90). Using this, we get

$$\begin{aligned}\mathbb{E}[\|\hat{\boldsymbol{s}}\|^4] = \mathbb{E}[\|\hat{s}_I \boldsymbol{e}_I\|^4] = \mathbb{E}[\hat{s}_I^4] \leq d^2 \cdot \left(\frac{1}{r} + \frac{\delta}{R}\right)^2 \cdot \mathbb{E}[\hat{s}_I^2] &= d^2 \cdot \left(\frac{1}{r} + \frac{\delta}{R}\right)^2 \cdot \mathbb{E}[\|\hat{\boldsymbol{s}}\|^2] \\ &\leq d^3 \cdot \left(\frac{1}{r} + \frac{\delta}{R}\right)^4,\end{aligned}$$

where the last inequality follows the right-most inequality in (89). This shows the right-most inequality in (90). Finally, we have

$$\|\hat{\boldsymbol{s}}\| = |\hat{s}_I| \leq \max_{i \in [d]} |\hat{s}_i| \leq d \cdot (1/r + \delta/R),$$

where the last inequality follows by the first inequality in (90), which we already showed. Now we do not assume that $\hat{\gamma} \geq 1$ anymore. Finally, the fact that $\|\hat{\boldsymbol{s}}\| < +\infty$ follows by Lemma 13. ■

Proof of Theorem 16. The fact that $(\boldsymbol{x}_t) \subset \mathcal{C}$ follows from Lemma 15. Let $(\tilde{\gamma}_t, \tilde{\boldsymbol{s}}_t) = \mathsf{OPT}_{\mathcal{C}^\circ}(\boldsymbol{w}_t; \delta_t)$ and $(\hat{\gamma}_t, \hat{\boldsymbol{s}}_t) = \mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}(\boldsymbol{w}_t; \delta_t)$. Further, let $\tilde{\boldsymbol{\nu}}_t := \mathbb{I}_{\{\tilde{\gamma}_t \geq 1\}} \tilde{\boldsymbol{s}}_t$, and γ_t and $\boldsymbol{\nu}_t$ be as in Algorithm 1. Note that by Proposition 11 and Lemma 13, we have $\|\tilde{\boldsymbol{s}}_t\| \vee \|\hat{\boldsymbol{s}}_t\| < +\infty$ almost surely and so the expectations $\mathbb{E}_{t-1}[\mathbb{I}_{\{\tilde{\gamma}_t \geq 1\}} \tilde{\boldsymbol{s}}_t]$ and $\mathbb{E}_{t-1}[\mathbb{I}_{\{\hat{\gamma}_t \geq 1\}} \hat{\boldsymbol{s}}_t]$ are well defined. We also note that $\gamma_t = \hat{\gamma}_t = \tilde{\gamma}_t$, and γ_t, $\boldsymbol{w}_t$, and $\boldsymbol{x}_t$ are all deterministic functions of the past ($\boldsymbol{w}_t$ it the output of FTRL-prox, which is a deterministic function of the past).

By Lemma 7, there exists a random variable $\Delta_t \geq 0$ satisfying $\mathbb{E}_{t-1}[\Delta_t]$ and such that for all $t \geq 1$,

$$\begin{aligned}\forall \boldsymbol{x} \in \mathcal{C}, \quad \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x} \rangle &\leq \langle \boldsymbol{g}_t - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \tilde{\boldsymbol{\nu}}_t, \boldsymbol{w}_t - \boldsymbol{x} \rangle + (2\delta_t + \Delta_t) R \|\boldsymbol{g}_t\|, \\ &= \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x} \rangle + \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \langle \boldsymbol{\nu}_t - \tilde{\boldsymbol{\nu}}_t, \boldsymbol{w}_t - \boldsymbol{x} \rangle \\ &\quad + (2\delta_t + \Delta_t) R \|\boldsymbol{g}_t\|. \end{aligned} \tag{91}$$

Fix $\boldsymbol{x} \in \mathcal{C}$ and let $X_t := \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0} \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \cdot \langle \boldsymbol{\nu}_t - \tilde{\boldsymbol{\nu}}_t, \boldsymbol{w}_t - \boldsymbol{x} \rangle$. We start by bounding $|X_t|$ from above. By the fact that $\boldsymbol{x}_t \in \mathcal{C}$ and the B-Lipschitzness of ℓ_t, we have

$$|X_t| \leq 2\|\boldsymbol{g}_t\| R \|\boldsymbol{\nu}_t - \tilde{\boldsymbol{\nu}}_t\| \leq 2RB \cdot (\|\boldsymbol{\nu}_t\| + \|\tilde{\boldsymbol{\nu}}_t\|) \leq 4dRB \cdot (\kappa + \delta), \tag{92}$$

where the last inequality follows by the fact that $\boldsymbol{\nu}_t$ [resp. $\tilde{\boldsymbol{\nu}}_t$] has the same distribution as $\mathbb{I}_{\{\hat{\gamma}_t \geq 1\}} \hat{\boldsymbol{s}}_t$ [resp. $\mathbb{I}_{\{\tilde{\gamma}_t \geq 1\}} \tilde{\boldsymbol{s}}_t$], and $\|\hat{\boldsymbol{s}}_t \cdot \mathbb{I}_{\{\hat{\gamma}_t \geq 1\}}\| \leq d \cdot (\delta/R + 1/r)$ by Lemma 32 [resp. $\|\tilde{\boldsymbol{s}}_t \cdot \mathbb{I}_{\{\tilde{\gamma}_t \geq 1\}}\| \leq$

$d \cdot (\delta/R + 1/r)$ by Proposition 11]. We now show that $\mathbb{E}_{t-1}[X_t] = 0$. Using the definition of X_t, we have

$$\begin{aligned}\mathbb{E}_{t-1}[X_t] &= \mathbb{E}_{t-1}[\mathbb{I}_{\{\tilde{\gamma}_t<1\}} \cdot \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle<0}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle \cdot \langle \boldsymbol{\nu}_t - \tilde{\boldsymbol{\nu}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle] \\ &\quad + \mathbb{E}_{t-1}[\mathbb{I}_{\{\tilde{\gamma}_t\geq 1\}} \cdot \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle<0}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle \cdot \langle \boldsymbol{\nu}_t - \tilde{\boldsymbol{\nu}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle], \\ &= \mathbb{E}_{t-1}[\mathbb{I}_{\{\hat{\gamma}_t\geq 1\}} \cdot \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle<0}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle \cdot \langle \hat{\boldsymbol{s}}_t - \tilde{\boldsymbol{s}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle] = 0, \end{aligned} \tag{93}$$

where the equalities in (93) follow by the facts that $\tilde{\gamma}_t = \hat{\gamma}_t = \gamma_t$ (by Lemma 32); $\boldsymbol{\nu}_t = \hat{\boldsymbol{s}}_t = \mathbf{0}$ if $\gamma_t \geq 1$; and that $\mathbb{E}_{t-1}[\mathbb{I}_{\{\tilde{\gamma}_t\geq 1\}} \cdot (\hat{\boldsymbol{s}}_t - \tilde{\boldsymbol{s}}_t)] = 0$ (by Lemma 32). Using that $\boldsymbol{x}_t$ is a deterministic function of the past, we get

$$\begin{aligned}\mathbb{E}_{t-1}[X_t^2] &\leq \|\boldsymbol{x}_t\|^2\|\boldsymbol{g}_t\|^2\mathbb{E}_{t-1}[\|\boldsymbol{\nu}_t - \tilde{\boldsymbol{\nu}}_t\|^2 \cdot \|\boldsymbol{w}_t - \boldsymbol{x}\|^2], \\ &\leq 8R^4\|\boldsymbol{g}_t\|^2\mathbb{E}_{t-1}[(\|\boldsymbol{\nu}_t\|^2 + \|\tilde{\boldsymbol{\nu}}_t\|^2)], \quad (\boldsymbol{x}_t \in \mathcal{C}) \\ &\leq 8R^4\|\boldsymbol{g}_t\|^2\mathbb{E}_{t-1}[\mathbb{I}_{\{\hat{\gamma}_t\geq 1\}} \cdot (\|\hat{\boldsymbol{s}}_t\|^2 + \|\tilde{\boldsymbol{s}}_t\|^2)], \\ &\leq 8R^2B^2d(\kappa+\delta)^2. \end{aligned} \tag{94}$$

where the last inequality follows by Lemma 32 and Proposition 11. Thus, by (91), (92), (93), (94), and Freedman's inequality (Theorem 40), we have, for any $\rho \in (0,1)$, with probability at least $1-\rho$,

$$\begin{aligned}&\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \\ =&\sum_{t=1}^{T}\langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle + \sum_{t=1}^{T}(2\delta_t + \Delta_t)R\|\boldsymbol{g}_t\| + 4RB(\kappa+\delta)\sqrt{2dT\ln\rho^{-1}} + 4dRB\cdot(\kappa+\delta), \\ \leq& 2R\sqrt{2\sum_{t=1}^{T}\|\tilde{\boldsymbol{g}}_t\|^2} + \sum_{t=1}^{T}(2\delta_t + \Delta_t)R\|\boldsymbol{g}_t\| + 4RB(\kappa+\delta)\sqrt{2dT\ln\rho^{-1}} + 4dRB\cdot(\kappa+\delta), \end{aligned} \tag{95}$$

where the last inequality follows by the regret bound FTRL-prox in Proposition 25. We now bound the first two terms on the RHS of (95). We start with $\sum_{t=1}^{T}\|\tilde{\boldsymbol{g}}_t\|^2$. With $Y_t := \|\tilde{\boldsymbol{g}}_t\|^2$, we have

$$\begin{aligned}\mathbb{E}_{t-1}[Y_t] &\leq \mathbb{E}_{t-1}[2\|\boldsymbol{g}_t\|^2 + 2\|\boldsymbol{g}_t\|^2\|\boldsymbol{x}_t\|^2\|\boldsymbol{\nu}_t\|^2], \\ &\leq 2B^2R^2\mathbb{E}_{t-1}[\mathbb{I}_{\{\hat{\gamma}_t<1\}} \cdot \|\boldsymbol{\nu}_t\|^2] + 2B^2R^2\mathbb{E}_{t-1}[\mathbb{I}_{\{\hat{\gamma}_t\geq 1\}} \cdot \|\hat{\boldsymbol{s}}_t\|^2] + 2\|\boldsymbol{g}_t\|^2, \quad (\boldsymbol{x}_t \in \mathcal{C}) \\ &\leq 2dB^2(\kappa+\delta)^2 + 2B^2, \end{aligned} \tag{96}$$

where the last inequality follows by Lemma 32. Similarly, we also have

$$\begin{aligned}\mathbb{E}_{t-1}[Y_t^2] &\leq \mathbb{E}_{t-1}[8\|\boldsymbol{g}_t\|^4 + 8\|\boldsymbol{g}_t\|^4\|\boldsymbol{x}_t\|^4\|\boldsymbol{\nu}_t\|^4], \\ &= 8B^4R^4\mathbb{E}_{t-1}[\|\boldsymbol{\nu}_t\|^4] + 8B^4, \quad (\boldsymbol{x}_t \in \mathcal{C}) \\ &= 8B^4R^4\mathbb{E}_{t-1}[\mathbb{I}_{\{\hat{\gamma}_t<1\}} \cdot \|\boldsymbol{\nu}_t\|^4] + 8B^4R^4\mathbb{E}_{t-1}[\mathbb{I}_{\{\hat{\gamma}_t\geq 1\}} \cdot \|\hat{\boldsymbol{s}}_t\|^4] + 8B^4, \\ &\leq 8d^3B^4(\kappa+\delta)^4 + 8B^4, \end{aligned} \tag{97}$$

where the last inequality follows by Lemma 32. Also, we have

$$|Y_t| \leq \|\boldsymbol{g}_t\| + \|\boldsymbol{g}_t\|\|\boldsymbol{x}_t\|\|\boldsymbol{\nu}_t\| \leq B\cdot(1 + R\|\boldsymbol{\nu}_t\|) \leq B\cdot(1 + d\cdot(\kappa+\delta)), \tag{98}$$

where the last inequality follows by the fact that $\boldsymbol{\nu}_t$ has the same distribution as $\mathbb{I}_{\{\hat{\gamma}_t \geq 1\}} \hat{\boldsymbol{s}}_t$, and $\|\hat{\boldsymbol{s}}_t\| \leq d \cdot (\delta/R + 1/r)$ by Lemma 32. By combining (96), (97), and (98), and applying Freedman's inequality (Theorem 40), we get for all $\rho > 0$, with probability at least $1 - \rho$,

$$\begin{aligned}\sum_{t=1}^T \|\tilde{\boldsymbol{g}}_t\|^2 &\leq \sum_{t=1}^T \mathbb{E}_{t-1}[\|\tilde{\boldsymbol{g}}_t\|^2] + 4\sqrt{2B^4(1 + d^3(\kappa+\delta)^4)T\ln(1/\rho)} + B^2(1 + d(\kappa+\delta))^2 \ln(1/\delta), \\ &\leq 4B^2 d(\kappa+\delta)^2 T + 8d^{3/2}B^2(\kappa+\delta)^2\sqrt{T\ln(1/\rho)} + B^2(1 + d(\kappa+\delta))^2\ln(1/\delta), \\ &\leq 12B^2 d(\kappa+\delta)^2 T + 4B^2 d^2(\kappa+\delta)^2 \ln(1/\delta), \end{aligned} \tag{99}$$

where the last inequality follows by the fact that $d\ln(1/\rho) \leq T$. Finally, by Markov's inequality and the fact that ℓ_t is B-Lipschitz, we have, for all ρ,

$$\mathbb{P}\left[\sum_{t=1}^T \Delta_t \|\boldsymbol{g}_t\| \geq 2B\delta\sqrt{T}\rho^{-1}\right] \leq \frac{\rho\sum_{t=1}^T \mathbb{E}[\Delta_t\|\boldsymbol{g}_t\|]}{2\delta B\sqrt{T}} \leq \frac{\rho\sum_{t=1}^T \mathbb{E}[\Delta_t]}{2\delta\sqrt{T}} \leq \frac{\rho\sum_{t=1}^T \delta_t}{2\delta} \leq \rho, \tag{100}$$

where the last inequality follows the fact that $\delta_t = \delta/t^{-1/2}$ and $\sum_{t=1}^T 1/t^{-1/2} \leq 2\sqrt{T}$, for all $T \geq 1$. By combining (95), (99), and (100), and a union bound, we get, with probability at least $1 - 3\rho$,

$$\begin{aligned}&\sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \\ \leq& 4RB(\kappa+\delta)\sqrt{6dT + 2d\ln\rho^{-1}} + 6RB\delta\sqrt{T}/\rho + 4RB(\kappa+\delta)\sqrt{2dT\ln\rho^{-1}} + 4dRB\cdot(\kappa+\delta), \\ \leq& 8RB(\kappa+\delta)\sqrt{dT(3+\ln\rho^{-1}) + d\ln\rho^{-1}} + 2dBR\cdot(2\kappa + \delta\cdot(2 + 3\sqrt{T}/\rho)),\end{aligned}$$

where the last inequality follows by the fact that $\sqrt{a} + \sqrt{b} \leq \sqrt{2a+2b}$. This completes the proof. ■

I.5. Proof of Theorem 17 (The Strongly Convex Case)

Before proving Theorem 17, we present a result that may be viewed as an extension of Lemma 7:

Lemma 33 *Let $\boldsymbol{w}_t, \tilde{\boldsymbol{g}}_t$, and $\boldsymbol{x}_t$ be as in Algorithm 1 with $\mathcal{O}_{\mathcal{C}^\circ} \equiv \mathsf{OPT}_{\mathcal{C}^\circ}$, any tolerance sequence $(\delta_s) \subset (0, 1/3)$, and any OCO subroutine A defined on $\mathcal{B}(R)$ (R as in (1)). Then, for all $t \geq 1$ and all $\boldsymbol{x} \in \mathcal{C}$, we have*

$$\langle \tilde{\boldsymbol{g}}_t, \boldsymbol{x}_t\rangle \leq \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t\rangle + (2\delta_t + \Delta_t)R\|\boldsymbol{g}_t\|,$$

where $\Delta_t \in [0, 15^2 d^4\kappa^3\delta_t^{-2}]$ is the random variable in Lemma 7, which satisfies $\mathbb{E}_{t-1}[\Delta_t] \leq \delta_t$.

Proof Instantiating (5) in Lemma 7 with $\boldsymbol{x} = \boldsymbol{x}_t$, we get

$$\langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t\rangle \geq \tilde{\ell}_t(\boldsymbol{w}_t) - \tilde{\ell}_t(\boldsymbol{x}_t) + \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{x}_t\rangle - (\delta_t + \Delta_t)R\|\boldsymbol{g}_t\|. \tag{101}$$

Now, to get the desired result, we need to show that $\tilde{\ell}_t(\boldsymbol{w}_t) - \tilde{\ell}_t(\boldsymbol{x}_t) \geq -\delta_t R\|\boldsymbol{g}_t\|$. When $\gamma_t < 1$, we have $\boldsymbol{x}_t = \boldsymbol{w}_t$ and so

$$\tilde{\ell}_t(\boldsymbol{x}_t) = \tilde{\ell}_t(\boldsymbol{w}_t). \quad \text{[case where } \gamma_t < 1\text{]} \tag{102}$$

Now, suppose that $\gamma_t \geq 1$ and $\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle \geq 0$. By the fact that $\boldsymbol{x}_t = \boldsymbol{w}_t/\gamma_t$, we get

$$\tilde{\ell}_t(\boldsymbol{x}_t) = \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \leq \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle = \tilde{\ell}_t(\boldsymbol{w}_t). \quad \text{[case where } \gamma_t \geq 1, \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle \geq 0] \tag{103}$$

Now suppose that $\gamma_t \geq 1$, $\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0$, and $\boldsymbol{w}_t \in \mathcal{C}$. We note that this implies that $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \leq 1$ and so $S_{\mathcal{C}}(\boldsymbol{w}_t) = 0$ (Lemma 6). Thus, $\tilde{\ell}_t(\boldsymbol{w}_t) = \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle$. On the other hand, by Lemma 10 we have $\gamma_t \leq \gamma_{\mathcal{C}}(\boldsymbol{w}_t) + \delta_t \leq 1 + \delta_t$, and so since $\boldsymbol{x}_t = \boldsymbol{w}_t/\gamma_t$, we have

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \leq \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle/(1+\delta_t) = \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle - \delta_t \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle/(1+\delta_t).$$

Thus, since $\tilde{\ell}_t(\boldsymbol{w}_t) = \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle$ and $\tilde{\ell}_t(\boldsymbol{x}_t) = \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle$ (because $\boldsymbol{x}_t \in \mathcal{C}$), the previous display implies that

$$\tilde{\ell}_t(\boldsymbol{x}_t) = \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \leq \tilde{\ell}_t(\boldsymbol{w}_t) + \delta_t R \|\boldsymbol{g}_t\|. \quad \text{[case where } \gamma \geq 1, \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0, \boldsymbol{w}_t \in \mathcal{C}] \tag{104}$$

We now consider the last case where $\gamma_t \geq 1$, $\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0$, and $\boldsymbol{w}_t \notin \mathcal{C}$. We note that this implies that $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \geq 1$. By the fact that $\gamma_t \leq \gamma_{\mathcal{C}}(\boldsymbol{w}_t) + \delta_t$ (Lemma 10), we have

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \leq \frac{\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle}{\gamma_{\mathcal{C}}(\boldsymbol{w}_t) + \delta_t} = \frac{\langle \boldsymbol{g}_t, \boldsymbol{w}_t/\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \rangle}{1 + \delta_t/\gamma_{\mathcal{C}}(\boldsymbol{w}_t)} = \langle \boldsymbol{g}_t, \boldsymbol{w}_t/\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \rangle - \frac{\delta_t/\gamma_{\mathcal{C}}(\boldsymbol{w}_t)}{1 + \delta_t/\gamma_{\mathcal{C}}(\boldsymbol{w}_t)} \langle \boldsymbol{g}_t, \boldsymbol{w}_t/\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \rangle, \tag{105}$$

Thus, since $S_{\mathcal{C}}(\boldsymbol{w}_t) = \gamma_{\mathcal{C}}(\boldsymbol{w}_t) - 1$ (by Lemma 6 and $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \geq 1$), we get

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \cdot S_{\mathcal{C}}(\boldsymbol{w}_t) \leq \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle - \frac{\delta_t \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle}{1 + \delta_t/\gamma_{\mathcal{C}}(\boldsymbol{w}_t)} - \langle \boldsymbol{g}_t, \boldsymbol{w}_t/\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \rangle + \frac{\delta_t/\gamma_{\mathcal{C}}(\boldsymbol{w}_t)}{1 + \delta_t/\gamma_{\mathcal{C}}(\boldsymbol{w}_t)} \langle \boldsymbol{g}_t, \boldsymbol{w}_t/\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \rangle.$$

Adding this together with (105) and using the fact that $\|\boldsymbol{w}_t\| \leq R$, we get

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle + \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle S_{\mathcal{C}}(\boldsymbol{w}_t) \leq \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle + \delta_t R \|\boldsymbol{g}_t\|,$$

and so after rearranging and using that $\tilde{\ell}_t(\boldsymbol{x}_t) = \langle \boldsymbol{x}_t, \boldsymbol{g}_t \rangle$ (since $\boldsymbol{x}_t \in \mathcal{C}$) and $\tilde{\ell}_t(\boldsymbol{w}_t) = \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle - \langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle S_{\mathcal{C}}(\boldsymbol{w}_t)$, we get

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t \rangle \leq \tilde{\ell}_t(\boldsymbol{w}_t) + \delta_t R \|\boldsymbol{g}_t\|. \quad \text{[case where } \gamma \geq 1, \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle < 0, \boldsymbol{w}_t \notin \mathcal{C}] \tag{106}$$

By combining, (102), (103), (104), and (106), we obtain:

$$\tilde{\ell}_t(\boldsymbol{x}_t) \leq \tilde{\ell}_t(\boldsymbol{w}_t) + \delta_t R \|\boldsymbol{g}_t\|.$$

Combining this with (101), we get

$$\langle \tilde{\boldsymbol{g}}_t, \boldsymbol{x}_t \rangle \leq \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t \rangle + (2\delta_t + \Delta_t) R \|\boldsymbol{g}_t\|.$$

■

We now present the versions of Lemmas 7 and 33 when $\mathcal{O}_{\mathcal{C}^\circ} \equiv \mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$:

Lemma 34 *Let $\boldsymbol{w}_t, \tilde{\boldsymbol{g}}_t$, and $\boldsymbol{x}_t$ be as in Algorithm 1 with $\mathcal{O}_{\mathcal{C}^\circ} \equiv \mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$; any tolerance sequence $(\delta_s) \subset (0, 1/3)$; any tolerance sequence $(\delta_s) \subset (0, 1/3)$; and any OCO subroutine A defined on $\mathcal{B}(R)$ (R as in (1)). Then, for all $t \geq 1$, $\boldsymbol{x}_t \in \mathcal{C}$ and there exists a family of random variables $\{\xi_t(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{C}\}$ such that for all $\boldsymbol{x} \in \mathcal{C}$, $\mathbb{E}_{t-1}[\xi_t(\boldsymbol{x})] = 0$, $|\xi_t(\boldsymbol{x})| < 4d(\kappa + \delta_t)R\|\boldsymbol{g}_t\|$ almost surely, and*

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \leq \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle + (2\delta_t + \Delta_t)R\|\boldsymbol{g}_t\| + \xi_t(\boldsymbol{x}); \tag{107}$$

$$\langle \tilde{\boldsymbol{g}}_t, \boldsymbol{x}_t\rangle \leq \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t\rangle + (2\delta_t + \Delta_t)R\|\boldsymbol{g}_t\| + \xi_t(\boldsymbol{x}_t), \tag{108}$$

where $\Delta_t \in [0, 15^2 d^4 \kappa^3 \delta_t^{-2}]$ is a random variable satisfying $\mathbb{E}_{t-1}[\Delta_t] \leq \delta_t$. Furthermore, $\|\tilde{\boldsymbol{g}}_t\| \leq (1 + d\kappa + d\Delta_t)\|\boldsymbol{g}_t\|$ and $\mathbb{E}_{t-1}[\|\tilde{\boldsymbol{g}}_t\|^2] \leq 2(1 + d(\kappa + \delta_t)^2)\|\boldsymbol{g}_t\|^2$.

Proof Let $t \geq 1$; $\boldsymbol{x} \in \mathcal{C}$; $(\tilde{\gamma}_t, \tilde{\boldsymbol{s}}_t) = \mathsf{OPT}_{\mathcal{C}^\circ}(\boldsymbol{w}_t; \delta_t)$; and $(\hat{\gamma}_t, \hat{\boldsymbol{s}}_t) = \mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}(\boldsymbol{w}_t; \delta_t)$. Further, let $\tilde{\boldsymbol{\nu}}_t := \mathbb{I}_{\{\hat{\gamma}_t \geq 1\}}\tilde{\boldsymbol{s}}_t$ and $\xi_t(\boldsymbol{x}) := \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle < 0}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle\langle \boldsymbol{\nu}_t - \tilde{\boldsymbol{\nu}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle$. By Lemma 7, there exists a random variable $\Delta_t \geq 0$ satisfying $\mathbb{E}_{t-1}[\Delta_t] \leq \delta_t$ and such that for all $t \geq 1$ and $\boldsymbol{x} \in \mathcal{C}$,

$$\begin{aligned}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle &\leq \langle \boldsymbol{g}_t - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle < 0}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle\tilde{\boldsymbol{\nu}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle + (2\delta_t + \Delta_t)R\|\boldsymbol{g}_t\|,\\ &\leq \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle + (2\delta_t + \Delta_t)R\|\boldsymbol{g}_t\| + \xi_t(\boldsymbol{x}).\end{aligned}$$

By Proposition 11 and Lemma 32, we have $\|\tilde{\boldsymbol{s}}_t\| \vee \|\hat{\boldsymbol{s}}_t\| < d \cdot (1/R + \delta/r)$ whenever $\hat{\gamma}_t = \tilde{\gamma}_t \geq 1$. This implies that $\|\tilde{\boldsymbol{\nu}}_t\| \vee \|\hat{\boldsymbol{\nu}}_t\| < d \cdot (1/R + \delta/r)$, and so $|\xi_t(\boldsymbol{x})| < 4d(\kappa + \delta)R\|\boldsymbol{g}_t\|$. We now show that the conditional expectation of $\xi_t(\boldsymbol{x})$ is zero. By Lemma 13, and the law of total expectation, we have

$$\begin{aligned}\mathbb{E}_{t-1}[\xi_t(\boldsymbol{x})] &\leq \mathbb{E}_{t-1}[\mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle < 0}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle\langle \boldsymbol{\nu}_t - \tilde{\boldsymbol{\nu}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle],\\ &= \mathbb{I}_{\{\tilde{\gamma}_t \geq 1\}} \cdot \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle < 0}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle\langle \mathbb{E}_{t-1}[\boldsymbol{\nu}_t - \tilde{\boldsymbol{\nu}}_t], \boldsymbol{w}_t - \boldsymbol{x}\rangle, &(109)\\ &= \mathbb{I}_{\{\tilde{\gamma}_t \geq 1\}} \cdot \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle < 0}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle\langle \mathbb{E}_{t-1}[\boldsymbol{s}_t - \tilde{\boldsymbol{s}}_t], \boldsymbol{w}_t - \boldsymbol{x}\rangle,\\ &= 0. &(110)\end{aligned}$$

where (109) follows by the fact that $\boldsymbol{w}_t, \gamma_t, \tilde{\gamma}_t$, and $\boldsymbol{g}_t$ are deterministic function of the past, and (110) follows by the fact $\tilde{\gamma}_t = \hat{\gamma}_t$; and $\mathbb{E}_{t-1}[\hat{\boldsymbol{s}}_t] = \mathbb{E}_{t-1}[\tilde{\boldsymbol{s}}_t]$, when $\tilde{\gamma}_t \geq 1$ (Lemma 13). This shows (107). We follow similar steps to show (108), but we use the result of Lemma 33 instead of Lemma 7. By Lemma 33, we have

$$\langle \boldsymbol{g}_t - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle < 0}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle\tilde{\boldsymbol{\nu}}_t, \boldsymbol{x}_t\rangle \leq \langle \boldsymbol{g}_t - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle < 0}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle\tilde{\boldsymbol{\nu}}_t, \boldsymbol{w}_t\rangle + (2\delta_t + \Delta_t)R\|\boldsymbol{g}_t\|,$$

and so by rearranging, we get

$$\langle \tilde{\boldsymbol{g}}_t, \boldsymbol{x}_t\rangle \leq \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t\rangle + (2\delta_t + \Delta_t)R\|\boldsymbol{g}_t\| + \xi_t(\boldsymbol{x}_t).$$

This shows (108). We now bound $\|\tilde{\boldsymbol{g}}_t\|$. When $\tilde{\gamma}_t < 1$, we have $\boldsymbol{g}_t = \tilde{\boldsymbol{g}}_t$ and so $\|\tilde{\boldsymbol{g}}_t\| \leq (1 + d\kappa + d\Delta_t)\|\boldsymbol{g}_t\|$ holds trivially. Now suppose that $\tilde{\gamma}_t \geq 1$. In this case, $\boldsymbol{\nu}_t$ has the same distribution as $\mathbb{I}_{\{\hat{\gamma}_t \geq 1\}}\hat{\boldsymbol{s}}_t$, and so

$$\begin{aligned}\|\tilde{\boldsymbol{g}}_t\| = \|\langle \boldsymbol{g}_t, \boldsymbol{w}\rangle - \mathbb{I}_{\langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle < 0}\langle \boldsymbol{g}_t, \boldsymbol{x}_t\rangle\boldsymbol{\nu}_t\| = \|\boldsymbol{g}_t\| + \|\boldsymbol{g}_t\|\frac{\|\boldsymbol{w}_t\|}{\gamma_t}\|\boldsymbol{\nu}_t\| &\leq \|\boldsymbol{g}_t\|(1 + R\|\boldsymbol{\nu}_t\|)\\ &\leq (1 + d\Delta_t + d\kappa)\|\boldsymbol{g}_t\|,\end{aligned}$$

where the last inequality follows from the fact that $\mathbb{I}_{\{\hat{\gamma}_t \geq 1\}} \|\hat{\boldsymbol{s}}_t\| \leq d \cdot (\Delta_t/R + 1/r)$, by Lemma 32. From the penultimate inequality in the above display and Lemma 13, we also have

$$\mathbb{E}_{t-1}[\|\tilde{\boldsymbol{g}}_t\|^2] \leq 2\|\boldsymbol{g}_t\|^2(1 + R^2\mathbb{E}_{t-1}[\|\hat{\boldsymbol{s}}_t\|^2]) \leq 2\|\boldsymbol{g}_t\|^2(1 + d \cdot (\kappa + \delta_t)^2).$$

Finally, the fact that $\boldsymbol{x}_t \subset \mathcal{C}$ follows from Theorem 15. In fact, looking at the proof of Theorem 15 it is clear that the fact that $(\boldsymbol{x}_t) \subset \mathcal{C}$ only uses the fact that the subroutine A outputs iterates in $(\boldsymbol{x}_t)$ in $\mathcal{B}(R)$. ■

We now present the proof of Theorem 17. We note that the proof follows from that of (Cutkosky and Orabona, 2018, Theorem 7) with small modifications to account for the differences between their constrained-to-unconstrained reduction and our projection-free reduction (Algorithm 1), as well as the differences described in Section C.2 to achieve scale-invariance.

Proof of Theorem 17. First of all, note that $(\boldsymbol{x}_t) \subset \mathcal{C}$ follows directly from Lemma 34. Next, we will instantiate Theorem 35 below with $\delta_t = \delta/t^2$ for all $t \geq 1$ and some $\delta \in (0, 1/3)$. By Lemma 34 and the fact that the losses are B-Lipschitz, the variables ζ_T, Z_T, M_T, and N_T in Theorem 35 below satisfy

- $\mathbb{E}[\zeta_T] \leq 1 + d\kappa + 2d\delta_T$, where we used the fact that $\mathbb{E}_{t-1}[\Delta_t] \leq \delta_t$ and $\sum_{t=1}^{\infty} 1/t^2 \leq 2$.
- $\zeta_T \leq 16^2 d^5 \kappa^3/\delta_T^2$, where we used the fact that $\Delta_t \leq 15^2 d^4 \kappa^3/\delta_t^2$, $\forall t$.
- $Z_T \leq \epsilon^2 + B^2(16^4 d^{10}\kappa^6/\delta_T^4)(1+T) \leq B^2(16^4 d^{10}\kappa^6/\delta_T^4)(2+T)$.
- $M_T \leq 2\sqrt{\ln_+\left(2(16^6 d^{15}\kappa^9/\delta_T^6)(2T+1)^2 RB^2/\epsilon^2\right)}$.
- $\mathbb{E}[N_T] \leq \mathbb{E}[4(1+d\kappa+2d\delta)B \ln_+\left(4(16^6 d^{15}\kappa^9/\delta_T^6)(2T+1)^2 RB^2/\epsilon^2\right) + \epsilon/R + (1+2d\kappa+8d\delta)B]$. This follows by the facts that $\mathbb{E}[\zeta_T] \leq 1 + d\kappa + 2d\delta$ and $\zeta_T \leq 16^2 d^5\kappa^3/\delta_T^2$.
- $\mathbb{E}_{t-1}[\|\tilde{\boldsymbol{g}}_t\|^2] \leq 2(1 + d(\kappa + \delta_t)^2)\|\boldsymbol{g}_t\|^2$ by Lemma 34.

Thus, for $\nu := 1/(R \wedge 1) + \kappa + 2\delta$, there exists $W_T = O\left(\ln(e + \kappa dRTB/(\delta\epsilon))\right)$ such that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle\right] &\leq W_T\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\|\tilde{\boldsymbol{g}}_t\|^2\|\boldsymbol{x}_t - \boldsymbol{x}\|^2\right]} + d\nu RBW_T^2, \\
&= W_T\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{t-1}[\|\tilde{\boldsymbol{g}}_t\|^2\|\boldsymbol{x}_t - \boldsymbol{x}\|^2]\right]} + d\nu RBW_T^2, \\
&\leq W_T\sqrt{\mathbb{E}\left[2d\nu^2\sum_{t=1}^{T}\|\boldsymbol{g}_t\|^2\|\boldsymbol{x}_t - \boldsymbol{x}\|^2\right]} + d\nu RBW_T^2, \\
&\leq U_T\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\|\boldsymbol{g}_t\|^2\|\boldsymbol{x}_t - \boldsymbol{x}\|^2\right]} + RBU_T^2/\nu, \qquad (111)
\end{aligned}$$

where $U_T = O\left(\nu d^{1/2} \ln(e + \kappa dRTB/(\delta\epsilon))\right)$. Thus, for general convex functions, we have, in expectation

$$\sum_{t=1}^{T} (\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{x})) \leq U_T RB\sqrt{T} + RBU_T^2/\nu.$$

For μ-strongly convex functions (ℓ_t), we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} (\ell_t(\boldsymbol{x}_t) - \ell_t(\boldsymbol{x}))\right] \\
\leq &\mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle\right] - \mathbb{E}\left[\frac{\mu}{2}\sum_{t=1}^{T} \|\boldsymbol{x}_t - \boldsymbol{x}\|^2\right], \\
\leq &U_T \sqrt{\mathbb{E}\left[\sum_{t=1}^{T} \|\boldsymbol{g}_t\|^2 \|\boldsymbol{x}_t - \boldsymbol{x}\|^2\right]} + RBU_T^2/\nu - \mathbb{E}\left[\frac{\mu}{2}\sum_{t=1}^{T} \|\boldsymbol{x}_t - \boldsymbol{x}\|^2\right], \qquad (112) \\
\leq &\inf_{\eta > 0} \left\{\mathbb{E}\left[\sum_{t=1}^{T} \|\boldsymbol{g}_t\|^2 \|\boldsymbol{x}_t - \boldsymbol{x}\|^2/\eta\right] + \eta U_T^2/4\right\} - \mathbb{E}\left[\mu \sum_{t=1}^{T} \|\boldsymbol{x}_t - \boldsymbol{x}\|^2/2\right] + \nu RBU_T^2/\nu, \\
\leq &\frac{\mu}{2}\mathbb{E}\left[\sum_{t=1}^{T} \frac{\|\boldsymbol{g}_t\|^2}{B^2} \|\boldsymbol{x}_t - \boldsymbol{x}\|^2\right] + \frac{B^2 U_T^2}{2\mu} - \mathbb{E}\left[\frac{\mu}{2}\sum_{t=1}^{T} \|\boldsymbol{x}_t - \boldsymbol{x}\|^2\right] + RBU_T^2/\nu, \qquad (113) \\
\leq &B^2 U_T^2/(2\mu) + RBU_T^2/\nu,
\end{aligned}$$

where (112) follows by (111) and (113) follows by setting $\eta = 2B^2/\mu$. ■

Theorem 35 *Let $\delta \in (0, 1/3)$ and $\kappa := R/r$, with r and R as in (1). Suppose that Algorithm 1 is run with $\mathcal{O}_{\mathcal{C}^\circ} \equiv \mathsf{OPT}_{1\mathrm{d},\mathcal{C}^\circ}$; $\delta_t = \delta/t^2$, $\forall t \geq 1$; and sub-routine A set to Alg. 5 with parameter $\epsilon > 0$. Then, for any adversarial sequence of convex losses (ℓ_t) on $\mathcal{C}$, the iterates $(\boldsymbol{x}_t)$ of Alg. 1 in response to (ℓ_t) satisfy,*

$$\forall T \geq 1, \forall \boldsymbol{x} \in \mathcal{C}, \quad \sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \leq M_T \sqrt{2\left[(\epsilon^2 + \zeta_T B_T^2)\|\boldsymbol{x}\|^2 + \sum_{t=1}^{T} \|\tilde{\boldsymbol{g}}_t\|^2 \|\boldsymbol{x}_t - \boldsymbol{x}\|^2\right] \cdot \ln \frac{Z_T}{\epsilon^2}} + 2RN_T \cdot \left(1 + \ln \frac{Z_T}{\epsilon^2}\right) + \Xi_T(\boldsymbol{x}),$$

where $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}_t)$, $\forall t$, $B_T := \epsilon \vee \max_{t\in[T]} \|\boldsymbol{g}_t\|$, and $\zeta_T, Z_T, M_T, \Xi_T(\boldsymbol{x})$ and N_T are such that:

- *$\zeta_T := 1 + d\kappa + d \max_{t\in[T]} \Delta_t$, with $\Delta_t \geq 0$ a non-negative random variable satisfying $\mathbb{E}_{t-1}[\Delta_t] \leq \delta_t$.*
- *$Z_T \leq \epsilon^2 + \tilde{B}_T^2 + \sum_{t=1}^{T} \|\tilde{\boldsymbol{g}}_t\|^2 \leq \epsilon^2 + \zeta_T^2 B_T^2 + \zeta_T^2 \sum_{t=1}^{T} \|\boldsymbol{g}_t\|^2$.*
- *$M_T := 2\sqrt{\ln_+\left(2\zeta_t^2 (R + 2\zeta_t RT) Q_T/\epsilon^2\right)}$, and $Q_t := \epsilon^2 + \sum_{i=1}^{t} \|\boldsymbol{g}_i\|^2$.*

- $\Xi_T(\boldsymbol{x}) \coloneqq \sum_{t=1}^T \xi_t(\boldsymbol{x})$, *where* $\{\xi_t(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{C}\}$ *is a family of random variables satisfying for all* $\boldsymbol{x} \in \mathcal{C}$, $\mathbb{E}_{t-1}[\xi_t(\boldsymbol{x})] = 0$.
- $N_T \coloneqq 4\zeta_t B_T \ln_+ \left(4\zeta_t^2 B_T(R + 2\zeta_t RT)\sqrt{Q_T}/\epsilon^2\right) + B_T + \frac{\epsilon}{R} + \sum_{t=1}^T (2\delta_t + \Delta_t)\|\boldsymbol{g}_t\| + 2d(\kappa + \delta)B_T$.

Proof Define $\zeta_t \coloneqq 1 + d\kappa + d\max_{s\le t}\Delta_s$, where $\kappa = R/r$ and Δ_t is the same random variable as in Lemma 34. Note that $\zeta_t \ge 0$ and $\mathbb{E}[\zeta_t] \le 1 + d\kappa + d\sum_{s=1}^t \delta_s$. For any $t \ge 1$, consider the random vector $\boldsymbol{X}_t$ that takes value $\boldsymbol{x}_i$ for $s \le t$ with probability proportional to $\|\hat{\boldsymbol{g}}_i\|^2$, and value $\mathbf{0}$ with probability proportional to $\epsilon^2 + \tilde{B}_t^2$, where $\tilde{B}_t \coloneqq \epsilon \vee \max_{i\in[t]}\|\hat{\boldsymbol{g}}_i\|$ and $\hat{\boldsymbol{g}}_t \coloneqq \tilde{\boldsymbol{g}}_t \cdot \tilde{B}_{t-1}/\tilde{B}_t$ (see Algorithm 5). Moving forward, we define

$$z_t \coloneqq \|\hat{\boldsymbol{g}}_t\|^2, \quad \text{and} \quad z_{0,t} \coloneqq \epsilon^2 + \tilde{B}_t^2,$$

so that $Z_t \coloneqq z_{0,t} + \sum_{i=1}^t z_i = \epsilon^2 + \tilde{B}_t^2 + \sum_{i=1}^t \|\hat{\boldsymbol{g}}_i\|^2$. We make the following definitions/observations:

- We define $V_T(\boldsymbol{x}) = z_{0,T}\|\boldsymbol{x}\|^2 + \sum_{t=1}^T z_t\|\boldsymbol{x}_t - \boldsymbol{x}\|^2 = Z_T \cdot \mathbb{E}[\|\boldsymbol{X}_T - \boldsymbol{x}\|^2]$.
- $Z_t \le \zeta_T^2 B_T^2 \cdot (T+1) + \epsilon^2$, for all $t \in [T]$, which follows from Lemma 34.
- $\overline{\boldsymbol{x}}_t = \mathbb{E}[\boldsymbol{X}_t] = Z_t^{-1} \cdot \sum_{i=1}^t z_i \cdot \boldsymbol{x}_i = \boldsymbol{v}_t/Z_t$, where $(\boldsymbol{v}_i)$ are as in Algorithm 5.
- $\sigma_t^2 \coloneqq (z_{0,t}\|\overline{\boldsymbol{x}}_t\|^2 + \sum_{i=1}^t z_t \cdot \|\boldsymbol{x}_i - \overline{\boldsymbol{x}}_t\|^2)/Z_t$ so that $\sigma_t^2 = \mathbb{E}[\|\boldsymbol{X}_t - \overline{\boldsymbol{x}}_t\|^2]$.

To prove the theorem, we are going to show for any $\boldsymbol{x} \in \mathcal{C}$,

$$\sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \le M_T\sqrt{Z_T \cdot \|\boldsymbol{x} - \boldsymbol{x}_T\|^2} + M_T\sqrt{\sigma_T^2 \cdot Z_T \cdot \ln\frac{Z_T}{\epsilon^2}} + 2RN_T \cdot \left(1 + \ln\frac{Z_T}{\epsilon^2}\right) + \Xi_T(\boldsymbol{x}), \tag{114}$$

where M_T and N_T are as in (116) and (117) below, respectively, which implies the desired bound by a bias-variance decomposition:

$$Z_T \cdot \|\boldsymbol{x} - \overline{\boldsymbol{x}}_T\|^2 + Z_T \cdot \sigma_T^2 = Z_T \cdot \mathbb{E}[\|\boldsymbol{X}_T - \boldsymbol{x}\|^2] = V_T(\boldsymbol{x}).$$

In particular, combining this with (114) and using the fact that $\sqrt{x} + \sqrt{y} \le \sqrt{2(x+y)}$ for all $x, y > 0$, we get

$$\sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \le M_T\sqrt{2V_T(\boldsymbol{x}) \cdot \|\boldsymbol{x} - \boldsymbol{x}_T\|^2 \cdot \ln\frac{Z_T}{\epsilon^2}} + 2RN_T \cdot \left(1 + \ln\frac{Z_T}{\epsilon^2}\right) + \Xi_T(\boldsymbol{x}).$$

To get to (114), first observe that for all $\boldsymbol{x} \in \mathcal{C}$, the linearized regret of Algorithm 5 satisfies

$$\sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle - \sum_{t=1}^T \langle \bar{\boldsymbol{g}}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle = \sum_{t=1}^T \langle \boldsymbol{g}_t - \bar{\boldsymbol{g}}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle \overset{(*)}{\le} 2R\sum_{t=1}^T (B_t - B_{t-1}) \le 2RB_T,$$

where $(*)$ following by Cauchy Schwarz. Using this together with (107) (multiplied by $\tilde{B}_{t-1}/\tilde{B}_t$) and the clipped regret of FreeGrad in Proposition 26, we have, for all $\boldsymbol{x} \in \mathcal{C}$,

$$\begin{aligned}
&\sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle - 2RB_T \\
\leq &\sum_{t=1}^T \langle \bar{\boldsymbol{g}}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle, \\
\leq &\sum_{t=1}^T \langle \hat{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle + \sum_{t=1}^T (2\delta_t + \Delta_t) R\|\boldsymbol{g}_t\| + \sum_{t=1}^T \xi_t(\boldsymbol{x}), \\
= &\sum_{t=1}^T \langle \hat{\boldsymbol{g}}_t, \boldsymbol{u}_t - (\boldsymbol{x} - \overline{\boldsymbol{x}}_T)\rangle + \sum_{t=1}^T \langle \hat{\boldsymbol{g}}_t, \overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_T\rangle + \sum_{t=1}^T (\xi_t(\boldsymbol{x}) + (2\delta_t + \Delta_t) R\|\boldsymbol{g}_t\|), \\
= &M_T \sqrt{Z_T \cdot \|\boldsymbol{x} - \overline{\boldsymbol{x}}_T\|^2} + \sum_{t=1}^T \langle \hat{\boldsymbol{g}}_t, \overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_T\rangle + 2RN_T + \Xi_T(\boldsymbol{x}),
\end{aligned} \tag{115}$$

where $\Xi_T(\boldsymbol{x}) \coloneqq \sum_{t=1}^T \xi_t(\boldsymbol{x})$. Note that the first term on the RHS of (115) is exactly what we want, so we only have to derive an upper bound on the second one. This is readily done through Lemma 36 that immediately gives us the stated result. ■

Lemma 36 *Under the hypotheses of Theorem 35, we have*

$$\sum_{t=1}^T \langle \hat{\boldsymbol{g}}_t, \overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_T\rangle \leq M_T \sigma_T \sqrt{Z_T \ln \frac{Z_T}{\epsilon^2}} + 2RN_T \cdot \ln \frac{Z_T}{\epsilon^2}, \quad \textit{where}$$

$$M_T \coloneqq 2\sqrt{\ln_+ \left(\frac{2\zeta_t^2 (R + 2\zeta_t RT) Q_T}{\epsilon^2}\right)}; \quad Q_T \coloneqq \epsilon^2 + \sum_{t=1}^T \|\boldsymbol{g}_t\|^2; \quad \textit{and} \tag{116}$$

$$N_T \coloneqq 4\zeta_t B_T \ln_+ \left(\frac{4\zeta_t^2 B_T (R + 2\zeta_t RT)\sqrt{Q_T}}{\epsilon^2}\right) + B_T + \frac{\epsilon}{R} + \sum_{t=1}^T (2\delta_t + \Delta_t)\|\boldsymbol{g}_t\| + 2d(\kappa + \delta)B_T. \tag{117}$$

Proof We have that

$$\sum_{i=1}^t \langle \hat{\boldsymbol{g}}_i, \overline{\boldsymbol{x}}_{i-1} - \overline{\boldsymbol{x}}_t\rangle - \sum_{i=1}^{t-1} \langle \hat{\boldsymbol{g}}_i, \overline{\boldsymbol{x}}_{i-1} - \overline{\boldsymbol{x}}_{t-1}\rangle = \left\langle \sum_{i=1}^t \hat{\boldsymbol{g}}_i, \overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_t \right\rangle .$$

The telescoping sum gives us

$$\sum_{t=1}^T \langle \hat{\boldsymbol{g}}_t, \overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_T\rangle = \sum_{t=1}^T \left\langle \sum_{i=1}^t \hat{\boldsymbol{g}}_i, \overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_t \right\rangle \leq \sum_{t=1}^T \left\| \sum_{i=1}^t \hat{\boldsymbol{g}}_i \right\| \|\overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_t\| .$$

So in order to bound $\sum_{t=1}^T \langle \hat{\boldsymbol{g}}_t, \overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_T\rangle$, it suffices to bound $\left\|\sum_{i=1}^t \hat{\boldsymbol{g}}_i\right\| \|\overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_t\|$ by a sufficiently small value. First, we will tackle $\left\|\sum_{i=1}^t \hat{\boldsymbol{g}}_i\right\|$. By Lemma 34 (in particular (108)) and the

fact that $\hat{\boldsymbol{g}}_t = \tilde{\boldsymbol{g}}_t \cdot B_{t-1}/B_t$, we have, for all $x > 0$,

$$\begin{aligned}
\sum_{i=1}^{t} -2R\|\hat{\boldsymbol{g}}_i\| + \left\|\sum_{i=1}^{t}\hat{\boldsymbol{g}}_i\right\| x &\le \sum_{i=1}^{t}\langle \hat{\boldsymbol{g}}_i, \boldsymbol{x}_i - \overline{\boldsymbol{x}}_{i-1}\rangle + \left\|\sum_{i=1}^{t}\hat{\boldsymbol{g}}_i\right\| x, \quad (\text{since } \boldsymbol{x}_i, \overline{\boldsymbol{x}}_{i-1} \in \mathcal{C}) \\
&\le \sum_{i=1}^{t}\langle \hat{\boldsymbol{g}}_i, \boldsymbol{w}_i - \overline{\boldsymbol{x}}_{i-1}\rangle + \left\|\sum_{i=1}^{t}\hat{\boldsymbol{g}}_i\right\| x + \sum_{i=1}^{t}\xi_i(\boldsymbol{x}_i) + (2\delta_i + \Delta_i)R\|\boldsymbol{g}_i\|), \\
&= \sum_{i=1}^{t}\langle \hat{\boldsymbol{g}}_i, \boldsymbol{u}_i\rangle + \left\|\sum_{i=1}^{t}\hat{\boldsymbol{g}}_i\right\| x + \sum_{i=1}^{t}(2\delta_i + \Delta_i)R\|\boldsymbol{g}_i\| + \sum_{i=1}^{t}\xi_i(\boldsymbol{x}_i), \\
&\le 2x\sqrt{\left(\epsilon^2 + \sum_{i=1}^{t}\|\hat{\boldsymbol{g}}_i\|^2\right)\ln_+\left(\frac{2\zeta_t^2 x Q_t}{\epsilon^2}\right)} + \sum_{i=1}^{t}\xi_i(\boldsymbol{x}_i) \\
&\quad + \sum_{i=1}^{t}(2\delta_i + \Delta_i)R\|\boldsymbol{g}_i\| + 4\zeta_t' x \ln\left(\frac{4\zeta_t^2 B_T x\sqrt{Q_t}}{\epsilon^2}\right) + \epsilon, \qquad (118)
\end{aligned}$$

where $\boldsymbol{u}_i$ is the ith output of FreeGrad, $Q_t \coloneqq \epsilon^2 + \sum_{i=1}^{t}\|\boldsymbol{g}_i\|^2 \le \epsilon^2 + tB_t^2$, and $\zeta_t' \coloneqq \max_{s\in[t]}\{(1 + d\kappa + d\Delta_s)B_s\}$. The passage to (118) follows from the regret bound FreeGrad (see Proposition 26) and the fact that $\|\hat{\boldsymbol{g}}_t\| \le \zeta_t\|\boldsymbol{g}_t\|$ (see Lemma 34). Moving forward, we let $G_t \coloneqq \epsilon + \sum_{i=1}^{t}(2\delta_i + \Delta_i)R\|\boldsymbol{g}_i\|$. Dividing by x in (118) and solving for $\left\|\sum_{i=1}^{t}\hat{\boldsymbol{g}}_i\right\|$, we get

$$\begin{aligned}
\left\|\sum_{i=1}^{t}\hat{\boldsymbol{g}}_i\right\| &\le 2\sqrt{\left(\epsilon^2 + \sum_{i=1}^{t}\|\hat{\boldsymbol{g}}_i\|^2\right)\ln_+\left(\frac{2\zeta_t^2 x Q_t}{\epsilon^2}\right)} \\
&\quad + 4\zeta_t' \ln\left(\frac{4\zeta_t^2 B_t x\sqrt{Q_t}}{\epsilon^2}\right) + \frac{G_t}{x} + \frac{2R}{x}\sum_{i=1}^{t}\|\hat{\boldsymbol{g}}_i\| + \frac{1}{x}\sum_{i=1}^{t}\xi_i(\boldsymbol{x}_i)\,.
\end{aligned}$$

Set $x = R + 2R\sum_{i=1}^{t}(\|\hat{\boldsymbol{g}}_i\| \vee \|\boldsymbol{g}_i\|)/B_T$ and using the facts that $\|\hat{\boldsymbol{g}}_t\| \le \zeta_t\|\boldsymbol{g}_t\|$ and $|\xi_t(\boldsymbol{x})| \le 4d(\kappa + \delta_t)R\|\boldsymbol{g}_t\|$, for all $\boldsymbol{x} \in \mathcal{C}$ (see Lemma 34), we conclude that:

$$\begin{aligned}
\left\|\sum_{i=1}^{t}\hat{\boldsymbol{g}}_i\right\| &\le M_t\sqrt{\epsilon^2 + \sum_{i=1}^{t}\|\hat{\boldsymbol{g}}_i\|^2} + N_t, \quad \text{where} \\
M_t &\coloneqq 2\sqrt{\ln_+\left(\frac{2\zeta_t^2(R + 2\zeta_t Rt)Q_t}{\epsilon^2}\right)}, \\
\text{and} \quad N_t &\coloneqq 4\zeta_t' \cdot \ln_+\left(\frac{4\zeta_t^2 B_t(R + 2\zeta_t Rt)\sqrt{Q_t}}{\epsilon^2}\right) + B_t + \frac{G_t}{R} + 2d(\kappa + \delta)B_t.
\end{aligned}$$

We recall that $Q_t \coloneqq \epsilon^2 + \sum_{i=1}^{t}\|\boldsymbol{g}_i\|^2 \le \epsilon^2 + tB_t^2$. With this in hand, we have

$$\begin{aligned}
\sum_{t=1}^{T}\langle \hat{\boldsymbol{g}}_t, \overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_T\rangle &\le \sum_{t=1}^{T}\left\|\sum_{i=1}^{t}\hat{\boldsymbol{g}}_i\right\|\|\overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_t\|, \\
&\le M_T\sum_{t=1}^{T}\sqrt{\epsilon^2 + \sum_{i=1}^{t}\|\hat{\boldsymbol{g}}_i\|^2}\|\overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_t\| + N_T\sum_{t=1}^{T}\|\overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_t\|\,.
\end{aligned}$$

Now, we relate $\|\overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_t\|$ to $\|\boldsymbol{x}_t - \overline{\boldsymbol{x}}_t\|$:

$$\overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_t = \overline{\boldsymbol{x}}_{t-1} - \frac{Z_{t-1}\overline{\boldsymbol{x}}_{t-1} + \|\hat{\boldsymbol{g}}_t\|^2 \boldsymbol{x}_t}{Z_t} = \frac{\|\hat{\boldsymbol{g}}_t\|^2}{Z_t}(\overline{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_t) = \frac{\|\hat{\boldsymbol{g}}_t\|^2}{Z_t}(\overline{\boldsymbol{x}}_t - \boldsymbol{x}_t) + \frac{\|\hat{\boldsymbol{g}}_t\|^2}{Z_t}(\overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_t),$$

that implies

$$Z_t \cdot (\overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_t) = \|\hat{\boldsymbol{g}}_t\|^2(\boldsymbol{x}_t - \overline{\boldsymbol{x}}_t) + \|\hat{\boldsymbol{g}}_t\|^2(\overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_t),$$

that is

$$\overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_t = \frac{\|\hat{\boldsymbol{g}}_t\|^2}{Z_{t-1}}(\boldsymbol{x}_t - \overline{\boldsymbol{x}}_t)\ . \tag{119}$$

Hence, we have

$$M_T \sum_{t=1}^T \sqrt{\epsilon^2 + \sum_{i=1}^t \|\hat{\boldsymbol{g}}_i\|^2}\|\overline{\boldsymbol{x}}_t - \overline{\boldsymbol{x}}_{t-1}\| \leq M_T \sum_{t=1}^T \sqrt{Z_t}\frac{\|\hat{\boldsymbol{g}}_t\|^2}{Z_{t-1}}\|\boldsymbol{x}_t - \overline{\boldsymbol{x}}_t\|,$$

and

$$N_T \sum_{t=1}^T \|\overline{\boldsymbol{x}}_t - \overline{\boldsymbol{x}}_{t-1}\| \leq N_T \sum_{t=1}^T \frac{\|\hat{\boldsymbol{g}}_t\|^2}{Z_{t-1}}\|\boldsymbol{x}_t - \overline{\boldsymbol{x}}_t\| \leq 2RN_T \sum_{t=1}^T \frac{\|\hat{\boldsymbol{g}}_t\|^2}{Z_{t-1}}\ .$$

Using Cauchy–Schwarz inequality, we have

$$M_T \sum_{t=1}^T \sqrt{Z_t}\frac{\|\hat{\boldsymbol{g}}_t\|^2}{Z_{t-1}}\|\boldsymbol{x}_t - \overline{\boldsymbol{x}}_t\| \leq M_T \sqrt{\sum_{t=1}^T \frac{\|\hat{\boldsymbol{g}}_t\|^2}{Z_{t-1}}} \sqrt{\sum_{t=1}^T \frac{Z_t}{Z_{t-1}}\|\hat{\boldsymbol{g}}_t\|^2\|\boldsymbol{x}_t - \overline{\boldsymbol{x}}_t\|^2}\ .$$

So, putting together the last inequalities, we have

$$\sum_{t=1}^T \langle \hat{\boldsymbol{g}}_t, \overline{\boldsymbol{x}}_{t-1} - \overline{\boldsymbol{x}}_T\rangle \leq M_T \sqrt{\sum_{t=1}^T \frac{\|\hat{\boldsymbol{g}}_t\|^2}{Z_{t-1}}} \sqrt{\sum_{t=1}^T \frac{Z_t}{Z_{t-1}}\|\hat{\boldsymbol{g}}_t\|^2\|\boldsymbol{x}_t - \overline{\boldsymbol{x}}_t\|^2} + 2RN_T \sum_{t=1}^T \frac{\|\hat{\boldsymbol{g}}_t\|^2}{Z_{t-1}}\ .$$

We now focus on the the term $\sum_{t=1}^T \frac{\|\hat{\boldsymbol{g}}_t\|^2}{Z_{t-1}}$ that is easily bounded:

$$\sum_{t=1}^T \frac{\|\hat{\boldsymbol{g}}_t\|^2}{Z_{t-1}} \leq \sum_{t=1}^T \frac{\|\hat{\boldsymbol{g}}_t\|^2}{\epsilon^2 + \sum_{i=1}^t \|\hat{\boldsymbol{g}}_i\|^2} \leq \ln \frac{Z_T}{\epsilon^2},$$

where the first inequality follows by the fact that $\tilde{B}_{t-1} \geq \|\hat{\boldsymbol{g}}_t\|$, and in the last inequality we used the inequality

$$\sum_{t=1}^T \frac{a_t}{\sum_{i=0}^t a_i} \leq \ln\left(\frac{\sum_{t=0}^T a_t}{a_0}\right),$$

for all $a_t \geq 0$. To bound the term $\sum_{t=1}^{T} \frac{Z_t}{Z_{t-1}} \|\tilde{\boldsymbol{g}}_t\|^2 \|\boldsymbol{x}_t - \overline{\boldsymbol{x}}_t\|^2$ from above, observe that

$$\begin{aligned}
\sigma_T^2 Z_T &= (\epsilon^2 + \tilde{B}_T^2)\|\overline{\boldsymbol{x}}_T\|^2 + \sum_{t=1}^{T} \|\hat{\boldsymbol{g}}_t\|^2 \|\boldsymbol{x}_t - \overline{\boldsymbol{x}}_T\|^2, \\
&\geq (\epsilon^2 + \tilde{B}_{T-1}^2)\|\overline{\boldsymbol{x}}_T\|^2 + \sum_{t=1}^{T-1} \|\hat{\boldsymbol{g}}_t\|^2 \|\boldsymbol{x}_t - \overline{\boldsymbol{x}}_T\|^2 + \|\hat{\boldsymbol{g}}_T\|^2 \|\boldsymbol{x}_T - \overline{\boldsymbol{x}}_T\|^2, \\
&= Z_{T-1} \cdot (\sigma_{T-1}^2 + \|\overline{\boldsymbol{x}}_T - \overline{\boldsymbol{x}}_{T-1}\|^2) + \|\hat{\boldsymbol{g}}_T\|^2 \|\boldsymbol{x}_T - \overline{\boldsymbol{x}}_T\|^2, \\
&= Z_{T-1}\sigma_{T-1}^2 + \|\hat{\boldsymbol{g}}_T\|^2 \left(1 + \frac{\|\hat{\boldsymbol{g}}_T\|^2}{Z_{T-1}}\right) \|\boldsymbol{x}_T - \overline{\boldsymbol{x}}_T\|^2, \\
&= Z_{T-1}\sigma_{T-1}^2 + \|\hat{\boldsymbol{g}}_T\|^2 \frac{Z_T}{Z_{T-1}} \|\boldsymbol{x}_T - \overline{\boldsymbol{x}}_T\|^2,
\end{aligned}$$

where the third equality comes from bias-variance decomposition and the fourth one comes from (119). Hence, we have

$$\sum_{t=1}^{T} \frac{Z_t}{Z_{t-1}} \|\hat{\boldsymbol{g}}_t\|^2 \|\boldsymbol{x}_t - \overline{\boldsymbol{x}}_t\|^2 = \sum_{t=1}^{T} (\sigma_t^2 Z_t - \sigma_{t-1}^2 Z_{t-1}) \leq \sigma_T^2 Z_T \,.$$

Putting all together, we have the stated bound. ■

I.6. Proof of Theorem 18 (The Smooth Stochastic Case)

We will need to use the result of Lemma 31 (in particular (73)) to show the desired convergence rate. In order for the result of Lemma 31 to be valid in the setting of Theorem 18, we need to show that the iterates $(\boldsymbol{w}_t)$ in Algorithm 6 are in $\mathcal{B}(R)$, which is what we do next:

Lemma 37 *In the setting of Theorem 18, we have* $(\boldsymbol{w}_t) \subset \mathcal{B}(R)$.

Proof Let γ_t be as in Algorithm 6. For $t = 1$, we have $\boldsymbol{w}_1 = \boldsymbol{0}$. For $t > 1$, we have $\boldsymbol{w}_t = (1 - \mu_t)(\boldsymbol{x}_{t-1} - \eta_{t-1}\boldsymbol{g}_{t-1}) + \mu_t \boldsymbol{u}_t$ and $\|\boldsymbol{u}_t\| \leq R'$ (since it is the output of FTRL-prox with parameter R'), and so

$$\begin{aligned}
\|\boldsymbol{w}_t\| \leq (1 - \mu_t)\|\boldsymbol{x}_{t-1}\| + \mu_t R' + |\eta_{t-1}|\|\boldsymbol{g}_{t-1}\| &= (1 - \mu_t)\|\boldsymbol{x}_{t-1}\| + \mu_t R' + \frac{\nu R' \|\boldsymbol{g}_{t-1}\|}{\sqrt{Z_{t-1}}}, \\
&\leq (1 - \mu_t)\|\boldsymbol{x}_{t-1}\| + (\mu_t + \nu)R'.
\end{aligned}$$

Thus, $\boldsymbol{w}_t \in \mathcal{B}(R)$ if $\boldsymbol{x}_t \in \mathcal{C}$ (which implies $\|\boldsymbol{x}_t\| \leq R'$). We will now show that $\boldsymbol{x}_t \in \mathcal{C}$. We consider two cases. Suppose that $\gamma_t < 1$. In this case, $\boldsymbol{x}_t = \boldsymbol{w}_t$. If $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \leq 9/16$, then by definition of the Gauge function we have $\boldsymbol{w}_t \in \mathcal{C}$ and the same holds for $\boldsymbol{x}_t$ (since $\boldsymbol{x}_t = \boldsymbol{w}_t$). On the other hand, if $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \geq 9/16$, then by Lemma 10, we have $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \leq \gamma_t < 1$, and so $\boldsymbol{x}_t = \boldsymbol{w}_t \in \mathcal{C}$.

Now suppose that $\gamma_t \geq 1$. In this case, we have $\boldsymbol{x}_t = \boldsymbol{w}_t/\gamma_t$ and so $\gamma_{\mathcal{C}}(\boldsymbol{x}_t) = \gamma_{\mathcal{C}}(\boldsymbol{w}_t/\gamma_t) = \gamma_{\mathcal{C}}(\boldsymbol{w}_t)/\gamma_t \leq \gamma_{\mathcal{C}}(\boldsymbol{w}_t)/\gamma_{\mathcal{C}}(\boldsymbol{w}_t) = 1$ (since $\gamma_{\mathcal{C}}(\boldsymbol{w}_t) \leq \gamma_t$ by Lemma 10), and so $\boldsymbol{x}_t \in \mathcal{C}$. ■

The following lemma will also be useful to us in the proof of Theorem 18:

Lemma 38 *Let $\beta > 0$, $t \geq 1$, $\boldsymbol{\xi}_t \in \mathbb{R}^d$ be a random vector satisfying (23) for $\sigma > 0$, and ℓ_t be as in (22). Further, suppose that $\mathcal{C}$ satisfies (26). When f is β-smooth on $\mathcal{C}$, we have,*

$$\mathbb{E}[\|\boldsymbol{g}_t\|^2] \leq \sigma^2 + 2R'\beta, \quad \textit{for any } \boldsymbol{x} \in \mathcal{C} \textit{ and } \boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x}).$$

Proof Let $\boldsymbol{x} \in \mathcal{C}$ and $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{x})$. First, since f is differentiable, ℓ_t is also differentiable. Thus, $\boldsymbol{g}_t = \nabla f(\boldsymbol{x}) + \boldsymbol{\xi}_t$, and so

$$\mathbb{E}[\|\boldsymbol{g}_t\|^2] = \mathbb{E}[\|\nabla f(\boldsymbol{x})\|^2 + 2\langle \nabla f(\boldsymbol{x}), \boldsymbol{\xi}_t \rangle + \|\boldsymbol{\xi}_t\|^2] = \|\nabla f(\boldsymbol{x})\|^2 + \mathbb{E}[\|\boldsymbol{\xi}_t\|^2] \leq 2R'\beta + \sigma^2,$$

where the last inequality follows (23) and (25). ■

We now present the proof of Theorem 18. We note that the proof is very similar to that of (Cutkosky, 2019, Theorem 4) with modifications to account for the application of our projection-free reduction (Algorithm 1).

Proof of Theorem 18. Throughout this proof, we let $\lambda_t = t$, $\forall t$, and ν and R' be as in (26). Note that Λ_t in Algorithm 6 satisfies $\Lambda_t = \sum_{s=1}^t \lambda_s$. By a standard convexity argument, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T \lambda_t (f(\boldsymbol{x}_t) - f(\boldsymbol{x}))\right] &\leq \mathbb{E}\left[\sum_{t=1}^T \lambda_t \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}\rangle\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^T \lambda_t \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{x}\rangle\right] + \mathbb{E}\left[\sum_{t=1}^T 3\lambda_t \delta_t R \|\boldsymbol{g}_t\|\right], \quad \text{(by Lemma 31)} \\
&= \mathbb{E}\left[\sum_{t=1}^T \lambda_t \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{x}\rangle + \sum_{t=1}^T \lambda_t \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{u}_t\rangle\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^T 3\lambda_t \delta_t R \cdot \mathbb{E}[\|\boldsymbol{g}_t\| \mid \boldsymbol{x}_t]\right], \\
&\leq \mathbb{E}\left[\sum_{t=1}^T \lambda_t \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{x}\rangle + \sum_{t=1}^T \lambda_t \langle \tilde{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{u}_t\rangle\right] \\
&\quad + 3R\sqrt{\sigma^2 + 2\beta R'} \sum_{t=1}^T \lambda_t \delta_t, \qquad (120)
\end{aligned}$$

where the last inequality follows by Lemma 38 and Jensen's inequality. By letting $\boldsymbol{y}_s \coloneqq \boldsymbol{x}_s - \eta_s \boldsymbol{g}_s$, we have by Line 8 of Algorithm 6

$$\lambda_s \cdot (\boldsymbol{w}_s - \boldsymbol{u}_s) = \Lambda_{s-1} \cdot (\boldsymbol{y}_{s-1} - \boldsymbol{w}_s), \quad \text{for all } s \geq 1. \qquad (121)$$

Moreover, the first term on the RHS of (120) is the regret of the FTRL-prox instance (Algorithm A) within Algorithm 6. Thus, by Proposition 25 and Lemma 31, this term is bounded from above by

$$\mathcal{R}_T^{\mathsf{FTRL}}(\boldsymbol{x}) \coloneqq 4(1+\kappa)R' \sqrt{\sum_{t=1}^T \lambda_t^2 \|\boldsymbol{g}_t\|^2} + 4R' \sum_{t=1}^T \lambda_t \Delta_t \|\boldsymbol{g}_t\|, \quad \forall \boldsymbol{x} \in \mathcal{C}. \qquad (122)$$

Plugging (121) and (122) into (120), yields

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\lambda_t(f(\boldsymbol{x}_t)-f(\boldsymbol{x}))\right]\\
\leq&\mathbb{E}\left[\mathcal{R}_T^{\mathsf{FTRL}}(\boldsymbol{x})\right]+\mathbb{E}\left[\sum_{t=1}^{T}\Lambda_{t-1}\langle\tilde{\boldsymbol{g}}_t,\boldsymbol{y}_{t-1}-\boldsymbol{w}_t\rangle\right]+3R\sqrt{\sigma^2+2\beta R'}\sum_{t=1}^{T}\lambda_t\delta_t,\\
\leq&\mathbb{E}\left[\mathcal{R}_T^{\mathsf{FTRL}}(\boldsymbol{x})\right]+\mathbb{E}\left[\sum_{t=1}^{T}\Lambda_{t-1}\langle\boldsymbol{g}_t,\boldsymbol{y}_{t-1}-\boldsymbol{x}_t\rangle\right]\\
&\quad+\mathbb{E}\left[\sum_{t=1}^{T}3\Lambda_{t-1}\delta_t R\|\boldsymbol{g}_t\|\right]+3R\sqrt{\sigma^2+2\beta R'}\sum_{t=1}^{T}\lambda_t\delta_t, \qquad (123)\\
=&\mathbb{E}\left[\mathcal{R}_T^{\mathsf{FTRL}}(\boldsymbol{x})\right]+\mathbb{E}\left[\sum_{t=1}^{T}\Lambda_{t-1}\langle\boldsymbol{g}_t,\boldsymbol{y}_{t-1}-\boldsymbol{x}_t\rangle\right]\\
&\quad+\mathbb{E}\left[\sum_{t=1}^{T}3\delta_t\Lambda_{t-1}R\cdot\mathbb{E}[\|\boldsymbol{g}_t\|\mid\boldsymbol{x}_t]\right]+3R\sqrt{\sigma^2+2\beta R'}\sum_{t=1}^{T}\lambda_t\delta_t,\\
\leq&\mathbb{E}\left[\mathcal{R}_T^{\mathsf{FTRL}}(\boldsymbol{x})\right]+\mathbb{E}\left[\sum_{t=1}^{T}\Lambda_{t-1}\langle\boldsymbol{g}_t,\boldsymbol{y}_{t-1}-\boldsymbol{x}_t\rangle\right]+3R\sqrt{\sigma^2+2\beta R'}\sum_{t=1}^{T}(\lambda_t+\Lambda_{t-1})\delta_t,
\end{aligned}$$

where (123) follows by Lemma 31, and the last inequality follows by Lemma 38 and Jensen's inequality. Next, we use convexity again to argue

$$\mathbb{E}[\langle\boldsymbol{g}_t,\boldsymbol{y}_{t-1}-\boldsymbol{x}_t\rangle]\leq\mathbb{E}[f(\boldsymbol{y}_{t-1})-f(\boldsymbol{x}_t)],$$

and then we subtract $\mathbb{E}[\sum_{t=1}^{T}\lambda_t f(\boldsymbol{x}_t)]$ from both sides:

$$\mathbb{E}[-\Lambda_T f(\boldsymbol{x})]\leq\mathbb{E}\left[\mathcal{R}_T^{\mathsf{FTRL}}(\boldsymbol{x})\right]+\sum_{t=1}^{T}\left(\Lambda_{t-1}f(\boldsymbol{y}_{t-1})-\Lambda_t f(\boldsymbol{x}_t)\right)+3R\sqrt{\sigma^2+2\beta R'}\sum_{t=1}^{T}(\lambda_t+\Lambda_{t-1})\delta_t.$$

Now we use smoothness to relate $f(\boldsymbol{y}_t)$ to $f(\boldsymbol{x}_t)$. Let $\nu\coloneqq 4\sqrt{2}(1+\kappa)$ so that $\eta_t=\nu R'/\sqrt{Z_t}$, where Z_t is as in Algorithm 6; i.e. $Z_t\coloneqq\varepsilon^2+\sum_{s=1}^{t}\Lambda_s\|\boldsymbol{g}_s\|^2$. With this, we have:

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{y}_t)]&\leq\mathbb{E}[f(\boldsymbol{x}_t)+\nabla f(\boldsymbol{x}_t)(\boldsymbol{y}_t-\boldsymbol{x}_t)+\frac{\beta}{2}\|\boldsymbol{x}_t-\boldsymbol{y}_t\|^2],\\
&\leq\mathbb{E}\left[f(\boldsymbol{x}_t)-\eta_t\|\boldsymbol{g}_t\|^2+\eta_t\langle\boldsymbol{\xi}_t,\boldsymbol{g}_t\rangle+\frac{\beta\eta_t^2\|\boldsymbol{g}_t\|^2}{2}\right].
\end{aligned}$$

Then multiply by Λ_t:

$$\mathbb{E}[\Lambda_t(f(\boldsymbol{y}_t)-f(\boldsymbol{x}_t))]\leq\mathbb{E}\left[-\frac{\nu R'\Lambda_t\|\boldsymbol{g}_t\|^2}{\sqrt{\varepsilon^2+\sum_{i=1}^{t}\Lambda_i\|\boldsymbol{g}_i\|^2}}+\frac{\beta\eta_t^2\Lambda_t\|\boldsymbol{g}_t\|^2}{2}+\eta_t\Lambda_t\langle\boldsymbol{\xi}_t,\boldsymbol{g}_t\rangle\right].$$

Next, we make use of the following facts (see e.g. (Cutkosky, 2019; Levy et al., 2018)): for positive numbers $\alpha_0, \dots, \alpha_n$,

$$\sqrt{\sum_{i=1}^n \alpha_i} \leq \sum_{i=1}^n \frac{\alpha_i}{\sqrt{\sum_{j=1}^i \alpha_j}} \leq 2\sqrt{\sum_{i=1}^n \alpha_i} \quad \text{and} \quad \sum_{i=1}^n \frac{\alpha_i}{\alpha_0 + \sum_{j=1}^i \alpha_j} \leq \ln\left(\alpha_0 + \sum_{i=1}^n \alpha_i\right) - \ln \alpha_0.$$

Using this, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \Lambda_t (f(\boldsymbol{y}_t) - f(\boldsymbol{x}_t))\right] && (124)\\
\leq &\mathbb{E}\left[-\nu R' \sqrt{\varepsilon^2 + \sum_{t=1}^T \Lambda_t \|\boldsymbol{g}_t\|^2} + \frac{\nu^2 (R')^2 \beta}{2} \ln\left(1 + \frac{\sum_{t=1}^T \Lambda_t \|\boldsymbol{g}_t\|^2}{\varepsilon^2}\right)\right] \\
&\quad + \nu R' \varepsilon + \mathbb{E}\left[\sum_{t=1}^T \eta_t \langle \boldsymbol{\xi}_t, \Lambda_t \boldsymbol{g}_t\rangle\right], \\
\leq &\mathbb{E}\left[-\nu R' \sqrt{\varepsilon^2 + \sum_{t=1}^T \Lambda_t \|\boldsymbol{g}_t\|^2}\right] + \nu R' \varepsilon + \mathbb{E}\left[\sum_{t=1}^T \eta_t \langle \boldsymbol{\xi}_t, \Lambda_t \boldsymbol{g}_t\rangle\right] \\
&\quad + \frac{\nu^2 (R')^2 \beta}{2} \ln\left(1 + \frac{\sum_{t=1}^T \Lambda_t \mathbb{E}\left[\|\boldsymbol{g}_t\|^2\right]}{\varepsilon^2}\right), && (125)\\
\leq &\mathbb{E}\left[-\nu R' \sqrt{\varepsilon^2 + \sum_{t=1}^T \Lambda_t \|\boldsymbol{g}_t\|^2}\right] + \nu R' \varepsilon + \frac{\nu^2 (R')^2 \beta}{2} \ln\left(1 + \frac{(\sigma^2 + 2R'\beta)\sum_{t=1}^T \Lambda_t}{\varepsilon^2}\right) \\
&\quad + \mathbb{E}\left[\sum_{t=1}^T \eta_t \langle \boldsymbol{\xi}_t, \Lambda_t \boldsymbol{g}_t\rangle\right], && (126)
\end{aligned}$$

where (125) follows by Jensen's inequality (the log is concave) and the triangular inequality, and (126) follows by Lemma 38. Using Cauchy-Schwarz, we obtain:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T \eta_t \langle \boldsymbol{\xi}_t, \Lambda_t \boldsymbol{g}_t\rangle\right] &\leq \mathbb{E}\left[\sqrt{\sum_{t=1}^T \Lambda_t \|\boldsymbol{\xi}_t\|^2} \sqrt{\sum_{t=1}^T \eta_t^2 \Lambda_t \|\boldsymbol{g}_t\|^2}\right], \\
&\leq \mathbb{E}\left[\nu R' \sqrt{\sum_{t=1}^T \Lambda_t \|\boldsymbol{\xi}_t\|^2 \cdot \ln\left(1 + \varepsilon^{-2} \sum_{s=1}^T \Lambda_s \|\boldsymbol{g}_s\|^2\right)}\right], \\
&\leq \mathbb{E}\left[\nu R' \sqrt{\sum_{t=1}^T \Lambda_t \|\boldsymbol{\xi}_t\|^2 \cdot \ln\left(1 + 2\varepsilon^{-2} \sum_{s=1}^T \Lambda_s (\|\boldsymbol{\xi}_s\|^2 + \|\nabla f(\boldsymbol{x}_s)\|^2)\right)}\right], \\
&\leq \mathbb{E}\left[\nu R' \sqrt{\sum_{t=1}^T \Lambda_t \|\boldsymbol{\xi}_t\|^2 \cdot \ln\left(1 + 2\varepsilon^{-2} \sum_{s=1}^T \Lambda_s (\|\boldsymbol{\xi}_s\|^2 + 2R'\beta)\right)}\right], \quad \text{(by (25))},
\end{aligned}$$

$$\leq \mathbb{E}\left[\nu R' \sqrt{\sum_{t=1}^{T} \Lambda_t \|\boldsymbol{\xi}_t\|^2 \cdot \ln\left(1 + 2\varepsilon^{-2} \sum_{s=1}^{T} \Lambda_s (\|\boldsymbol{\xi}_s\|^2 + 2R'\beta)\right)}\right], \quad (127)$$

$$\leq \nu\sigma R' \sqrt{\sum_{t=1}^{T} \Lambda_t \cdot \ln\left(1 + \frac{2(\sigma^2 + 2R'\beta)\sum_{s=1}^{T}\Lambda_s}{\varepsilon^2}\right)},$$

where (127) follows by the concavity of the function $x \mapsto \sqrt{x \cdot \ln(a + bx)}$, for any $a \geq 1$, $b > 0$, and $x > 0$ (see Lemma 41) and Jensen's inequality, and the last inequality follows by the fact that $\mathbb{E}[\|\boldsymbol{\xi}_t\|^2] \leq \sigma^2$.

Combining everything, we get

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} -\lambda_t f(\boldsymbol{x})\right] \leq \mathbb{E}\Bigg[&\mathcal{R}_T^{\mathsf{FTRL}}(\boldsymbol{x}) + \sum_{t=1}^{T}\left(\Lambda_{t-1} f(\boldsymbol{y}_{t-1}) - \Lambda_t f(\boldsymbol{y}_t)\right)\Bigg] \\
&+ \frac{\nu^2\beta(R')^2}{2} \ln\left(1 + \frac{\sum_{t=1}^{T}\Lambda_t(\sigma^2 + 2R'\beta)}{\varepsilon^2}\right) - \nu R' \sqrt{\varepsilon^2 + \sum_{t=1}^{T}\Lambda_t\|\boldsymbol{g}_t\|^2} \\
&+ \nu R'\varepsilon + \nu R'\sigma \sqrt{\sum_{t=1}^{T} \Lambda_t \ln\left(1 + \frac{2(\sigma^2 + 2R'\beta)\sum_{s=1}^{T}\Lambda_s}{\varepsilon^2}\right)} \\
&+ 3R\sqrt{\sigma^2 + 2\beta R'} \sum_{t=1}^{T} (\lambda_t + \Lambda_{t-1})\delta_t.
\end{aligned}$$

Now observe that $t^2 > \Lambda_t > \lambda_t^2/2$ and recall $\mathcal{R}_T^{\mathsf{FTRL}}(\boldsymbol{x}) = 4(1+\kappa)R'\sqrt{\sum_{t=1}^{T}\lambda_t^2\|\boldsymbol{g}_t\|^2} + 4R'\sum_{t=1}^{T}\lambda_t\Delta_t\|\boldsymbol{g}_t\|$. Therefore, since $\nu = 4\sqrt{2}(1+\kappa)$ we have:

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{R}_T^{\mathsf{FTRL}}(\boldsymbol{x}) - \nu R' \sqrt{\varepsilon^2 + \sum_{t=1}^{T}\Lambda_t\|\boldsymbol{g}_t\|^2}\right] \\
\leq &\mathbb{E}\left[4(1+\kappa)R'\sqrt{\sum_{t=1}^{T}\lambda_t^2\|\boldsymbol{g}_t\|^2} - 4\sqrt{2}(1+\kappa)R'\sqrt{\sum_{t=1}^{T}\lambda_t^2\|\boldsymbol{g}_t\|^2/2}\right] + \mathbb{E}\left[4R'\sum_{t=1}^{T}\lambda_t\Delta_t\|\boldsymbol{g}_t\|\right], \\
= &\mathbb{E}\left[4R'\sum_{t=1}^{T}\lambda_t\Delta_t \cdot \mathbb{E}[\|\boldsymbol{g}_t\| \mid \boldsymbol{x}_t]\right] \leq 4R'\sqrt{\sigma^2 + 2\beta R}\sum_{t=1}^{T}\lambda_t\delta_t,
\end{aligned}$$

where the last inequality follows by Lemma 38. Also, observe that $\sum_{t=1}^{T}\Lambda_t \leq \sum_{t=1}^{T} t^2 \leq T^3$. Thus, we telescope the sum to obtain:

$$\begin{aligned}
&\mathbb{E}[\Lambda_T(f(\boldsymbol{y}_T) - f(\boldsymbol{x}))] \\
\leq &\nu R'\varepsilon + \frac{\nu^2(R')^2\beta}{2}\ln\left(1 + \frac{(\sigma^2 + 2R'\beta)T^3}{\varepsilon^2}\right) + \nu R' T^{3/2}\sigma\sqrt{\ln\left(1 + \frac{2(\sigma^2 + 2R'\beta)T^3}{\varepsilon^2}\right)} \\
&+ 3R\sqrt{\sigma^2 + 2\beta R'}\sum_{t=1}^{T}(\Lambda_{t-1} + 3\lambda_t)\delta_t,
\end{aligned}$$

$$\leq \nu R' \varepsilon + \frac{\nu^2 (R')^2 \beta}{2} \ln\left(1 + \frac{(\sigma^2 + 2R'\beta) T^3}{\varepsilon^2}\right) + \nu R' T^{3/2} \sigma \sqrt{\ln\left(1 + \frac{2(\sigma^2 + 2R'\beta) T^3}{\varepsilon^2}\right)}$$
$$+ 3\delta R (\ln T + 6) \sqrt{\sigma^2 + 2\beta R'}, \tag{128}$$

where the last inequality follows by the fact that $\delta_t = \delta / t^3$, $\sum_{t=1}^T 1/t \leq \ln T$; and $\sum_{t=1}^T 1/t^2 \leq 2$. Dividing (128) by $\Lambda_T = \sum_{t=1}^T \lambda_t = \frac{T(T+1)}{2} > T^2/2$ shows the inequality of theorem. Finally, the fact that $(\boldsymbol{x}_t) \subset \mathcal{C}$ follows by Lemma 31. ■

Appendix J. Technical Lemmas

This section contains some technical lemmas we need to prove our results.

Lemma 39 *Let $(Y_t) \subset \mathbb{R}_{\geq 0}$ be a sequence of random variable satisfying $\mathbb{E}[Y_t \mid \mathcal{G}_{t-1}] \leq \delta_t$, for all $t \geq 1$, for some sequence $(\delta_t) \subset \mathbb{R}_{\geq 0}$. Then, for any $\rho \in (0,1)$ and $T \geq 1$, we have with probability at least $1 - \rho$,*

$$\mathbb{P}\left[\sum_{t=1}^T Y_t \geq (1 + 1/\rho) \sum_{t=1}^T \delta_t\right] \leq \rho.$$

Proof Let $X_t \coloneqq \sum_{i=1}^t (Y_i - \bar{\delta}_i)$, where $\bar{\delta}_i \coloneqq \mathbb{E}[Y_i \mid \mathcal{G}_{i-1}] \leq \delta_i$. The process (X_t) is a martingale; that is, for all $i \geq 1$, we have, for all $i < t$,

$$\mathbb{E}[X_t \mid \mathcal{G}_i] = \sum_{s=1}^i (Y_s - \bar{\delta}_s) = X_i.$$

Thus, by Doob's martingale inequality (Durrett, 2019, Theorem 4.4.2), we have, for any $\rho \in (0,1)$, and $T \geq 1$

$$\mathbb{P}\left[\sum_{t=1}^T Y_t \geq (1 + 1/\rho) \sum_{t=1}^T \delta_t\right] \leq \mathbb{P}\left[X_T \geq \sum_{t=1}^T \delta_t / \rho\right] \leq \mathbb{P}\left[\max_{t \leq T} X_t \geq \sum_{t=1}^T \delta_t / \rho\right] \leq \frac{\rho \mathbb{E}\left[X_T \vee 0\right]}{\sum_{t=1}^T \delta_t}$$
$$\leq \rho.$$

■

Theorem 40 *Let $\mathcal{F}_1, \ldots, \mathcal{F}_n$ be a filtration, and $X_1, \ldots, X_n$ be real random variables such that X_i is $\mathcal{F}_i$-measurable, $\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] = 0$, $|X_i| \leq b$, and $\sum_{i=1}^n \mathbb{E}[X_i^2 \mid \mathcal{F}_{i-1}] \leq V$ for some $b, V \geq 0$. Then, for any $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$\sum_{i=1}^n X_i \leq 2\sqrt{V_n \ln(1/\delta)} + b \ln(1/\delta).$$

Lemma 41 *For any $a \geq 1$, $b > 0$, the map $f : x \mapsto \sqrt{x \cdot \ln(a + bx)}$ is concave for $x > 0$.*

Proof Let $a \geq 1$ and $b > 0$. We will show that the second derivative of f is negative for all $x > 0$. We have

$$\begin{aligned}\forall x > 0, \quad f''(x) &= \frac{-(a+bx)^2 \log^2(a+bx) + 2abx \log(a+bx) - b^2x^2}{4(a+bx)^2(x\log(a+bx))^{3/2}}, \\ &\leq \frac{-a^2 \log^2(a+bx) + 2abx \log(a+bx) - b^2x^2}{4(a+bx)^2(x\log(a+bx))^{3/2}}, \\ &= -\frac{-(a\log(a+bx) - bx)^2}{4(a+bx)^2(x\log(a+bx))^{3/2}} \leq 0.\end{aligned}$$

■

Lemma 42 *Let $\boldsymbol{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ and $0 < r \leq R$. Further, let $\mathcal{C}$ be a closed convex set such that $\mathcal{B}(r) \subseteq \mathcal{C} \subseteq \mathcal{B}(R)$. Then, the following properties hold:*

(a) $\sigma_{\mathcal{C}^\circ}(\boldsymbol{w}) = \gamma_{\mathcal{C}}(\boldsymbol{w})$ *and* $(\mathcal{C}^\circ)^\circ = \mathcal{C}$.

(b) $\sigma_{\mathcal{C}}(\alpha\boldsymbol{w}) = \alpha\sigma_{\mathcal{C}}(\boldsymbol{w})$ *and* $\partial\sigma_{\mathcal{C}}(\alpha\boldsymbol{w}) = \partial\sigma_{\mathcal{C}}(\boldsymbol{w}) = \arg\max_{\boldsymbol{u}\in\mathcal{C}}\langle \boldsymbol{u}, \boldsymbol{w}\rangle$, *for all* $\alpha \geq 0$.

(c) $r\|\boldsymbol{w}\| \leq \sigma_{\mathcal{C}}(\boldsymbol{w}) \leq R\|\boldsymbol{w}\|$, $\|\boldsymbol{w}\|/R \leq \gamma_{\mathcal{C}}(\boldsymbol{w}) \leq \|\boldsymbol{w}\|/r$, *and* $\mathcal{B}(1/R) \subseteq \mathcal{C}^\circ \subseteq \mathcal{B}(1/r)$.

(d) $\langle \boldsymbol{w}, \boldsymbol{u}\rangle \leq \sigma_{\mathcal{C}}(\boldsymbol{w}) \cdot \gamma_{\mathcal{C}}(\boldsymbol{u})$, *for all* $\boldsymbol{u} \in \mathbb{R}^d$. *(Cauchy Schwarz)*

(e) $\sigma_{\mathcal{C}}(\boldsymbol{w} + \boldsymbol{u}) \leq \sigma_{\mathcal{C}}(\boldsymbol{w}) + \sigma_{\mathcal{C}}(\boldsymbol{u})$, *for all* $\boldsymbol{u} \in \mathbb{R}^d$. *(Sub-additivity)*

Proof Points (a), (b), and (e) follow from standard results in convex analysis, see e.g. (Molinaro, 2020, Lemma 2) for point (a) and (Hiriart-Urruty and Lemaréchal, 2004) for points (b) and (e). Point *(d)* follows from (Friedlander et al., 2014, Equation 2.3 & Proposition 2.3) and Point *(a)*. We now show point (c). The first inequality in Point (c) follows by the fact that

$$R\|\boldsymbol{w}\| \overset{(i)}{\geq} \sup_{\boldsymbol{u}\in\mathcal{C}}\langle \boldsymbol{u}, \boldsymbol{w}\rangle = \sigma_{\mathcal{C}}(\boldsymbol{w}) \geq \inf_{\boldsymbol{u}\in\mathcal{C}}\langle \boldsymbol{u}, \boldsymbol{w}\rangle \overset{(ii)}{\geq} r\|\boldsymbol{w}\|,$$

where (i) and (ii) follow by the assumption that $\mathcal{B}(r) \subseteq \mathcal{C} \subseteq \mathcal{B}(R)$. We now show that $\mathcal{B}(1/R) \subseteq \mathcal{C}^\circ \subseteq \mathcal{B}(1/r)$. For any $\boldsymbol{x} \in \mathcal{B}(1/R)$, we have $\langle \boldsymbol{x}, \boldsymbol{w}\rangle \leq 1$ for all $\boldsymbol{w} \in \mathcal{C}$, since $\mathcal{C} \subseteq \mathcal{B}(R)$. By definition of the polar set, this implies that $\mathcal{B}(1/R) \subseteq \mathcal{C}^\circ$. Now, let $\boldsymbol{x} \in \mathcal{C}^\circ$. This implies that $\langle \boldsymbol{x}, \boldsymbol{w}\rangle \leq 1$ for all $\boldsymbol{w} \in \mathcal{C}$. For $\boldsymbol{w} = r\boldsymbol{x}/\|\boldsymbol{x}\|$ (which is guaranteed to be in $\mathcal{C}$ since $\mathcal{B}(r) \subseteq \mathcal{C}$) this inequality implies that $\|\boldsymbol{x}\| \leq 1/r$, and so $\mathcal{C}^\circ \subseteq \mathcal{B}(1/r)$. Finally, $\|\boldsymbol{w}\|/R \leq \gamma_{\mathcal{C}}(\boldsymbol{w}) \leq \|\boldsymbol{w}\|/r$ follows by point (a) and the facts that $\mathcal{B}(1/R) \subseteq \mathcal{C}^\circ \subseteq \mathcal{B}(1/r)$ and that $(\mathcal{B}(r) \subseteq \mathcal{C} \subseteq \mathcal{B}(R) \implies r\|\boldsymbol{w}\| \leq \sigma_{\mathcal{C}}(\boldsymbol{w}) \leq R\|\boldsymbol{w}\|)$, which we just showed. ■