Proceedings of Machine Learning Research vol 291:1–45, 2025 38th Annual Conference on Learning Theory

Online Convex Optimization with a Separation Oracle

Zakaria Mhammedi MHAMMEDI@GOOGLE.COM
Google Research, NYC

Editors: Nika Haghtalab and Ankur Moitra

Abstract

In this paper, we introduce a new projection-free algorithm for Online Convex Optimization (OCO) with a state-of-the-art regret guarantee among separation-based algorithms. Existing projection-free methods based on the classical Frank-Wolfe algorithm achieve a suboptimal regret bound of $O(T^{3/4})$, while more recent separation-based approaches guarantee a regret bound of $O(\kappa\sqrt{T})$, where κ denotes the asphericity of the feasible set, defined as the ratio of the radii of the containing and contained balls. However, for ill-conditioned sets, κ can be arbitrarily large, potentially leading to poor performance. Our algorithm achieves a regret bound of $\widetilde{O}(\sqrt{dT} + \kappa d)$, while requiring only $\widetilde{O}(1)$ calls to a separation oracle per round. Crucially, the main term in the bound, $\widetilde{O}(\sqrt{dT})$, is independent of κ, addressing the limitations of previous methods. Additionally, as a by-product of our analysis, we recover the $O(\kappa\sqrt{T})$ regret bound of existing OCO algorithms with a more straightforward analysis and improve the regret bound for projection-free online exp-concave optimization. Finally, for constrained stochastic convex optimization, we achieve a state-of-the-art convergence rate of $\widetilde{O}(\sigma/\sqrt{T} + \kappa d/T)$, where σ represents the noise in the stochastic gradients, while requiring only $\widetilde{O}(1)$ calls to a separation oracle per iteration.

Keywords: Projection-free, online convex optimization, stochastic convex optimization, separation oracle.

1. Introduction

Convex optimization is a foundational tool in computer science and machine learning, underpinning many modern techniques in these fields. Although classical algorithms like interior-point and cutting-plane methods are effective (Grötschel et al., 2012; Bubeck, 2015; Lee et al., 2018; Lee and Sidford, 2019), they become computationally prohibitive as problem sizes and dimensions grow. Given the high-dimensional nature of many contemporary problems, there is a growing demand for alternative algorithms that preserve strong theoretical guarantees while being computationally efficient enough to tackle large-scale optimization tasks.

Online Gradient Descent (OGD) (Zinkevich, 2003) is a popular first-order optimization method that trades a higher number of iterations for lower memory usage and per-step computational cost, making it widely used in practice. However, in constrained convex optimization, OGD requires a Euclidean projection onto the feasible set at every step in the worst-case, which can be computationally expensive, especially for complex feasible sets. This drawback often offsets the benefits of first-order methods. To address this, projection-free optimization methods have been developed (Hazan, 2008; Jaggi, 2013; Lacoste-Julien and Jaggi, 2015; Garber and Hazan, 2016), with the most well-known being the Frank-Wolfe algorithm (Frank et al., 1956), which replaces costly Euclidean projections with more efficient *linear optimization* over the feasible set. More recently, another class of projection-free algorithms has emerged that uses *membership* or *separation* oracles instead of linear optimization (Mhammedi, 2022; Garber and Kretzu, 2022; Lu et al., 2023; Grimmer, 2024), providing greater flexibility and efficiency in handling constraints.

Most modern projection-free algorithms are designed for Online Convex Optimization (OCO) (Hazan, 2016), a framework that generalizes classical offline convex optimization. In OCO, at each round t, the algorithm selects a vector $\boldsymbol{w}_t$ from the feasible set $\mathcal{K} \subset \mathbb{R}^d$ and incurs a loss $f_t(\boldsymbol{w}_t)$, where f_t is a convex function that may be adversarially chosen. The objective is to ensure that the regret $\mathsf{Reg}_T \coloneqq \sup_{\boldsymbol{w}\in\mathcal{K}} \sum_{t=1}^T (f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}))$ grows sublinearly in T. Existing projection-free algorithms, such as those in (Hazan and Kale, 2012; Mhammedi, 2022; Garber and Kretzu, 2022), achieve sublinear regret while requiring only $\widetilde{O}(1)$ calls per round to a linear optimization or separation oracle. Sublinear regret in OCO translates into guarantees for offline and stochastic convex optimization via standard online-to-batch conversion techniques (Cesa-Bianchi and Lugosi, 2006; Shalev-Shwartz et al., 2011; Cutkosky, 2019), where smaller regret yields better convergence rates. While (projected) OGD achieves an optimal, dimension-free regret of $O(\sqrt{T})$, there remains a significant gap between this bound and the regret bounds of state-of-the-art projection-free algorithms (Hazan and Kale, 2012; Mhammedi, 2022; Garber and Kretzu, 2022; Lu et al., 2023). In this paper, we aim to close that gap.

The current state-of-the-art regret bound for linear optimization-based algorithms (e.g., Frank-Wolfe-style algorithms) is $O(T^{3/4})$ (Hazan and Kale, 2012). Since this result was first established, no improvements have been made without introducing additional structural assumptions on the objective function or feasible set. It remains an open question whether this is the best achievable regret bound for online algorithms that make only a constant number of calls to a linear optimization oracle per round. More recently, Mhammedi (2022); Garber and Kretzu (2022) introduced a new class of projection-free algorithms that use separation oracles instead of linear optimization oracles, guaranteeing a regret bound of $O(\kappa\sqrt{T})$, where $\kappa \coloneqq \frac{R}{r}$ represents the *asphericity* of the set $\mathcal{K}$, defined as the ratio between the radii of the containing and contained balls. Achieving this $O(\kappa\sqrt{T})$ regret bound, which provides optimal dependence on the number of rounds T, was somewhat surprising given the difficulty of improving the $O(T^{3/4})$ bound for linear optimization-based algorithms. However, the asphericity factor κ can be arbitrarily large for ill-conditioned feasible sets, which poses a challenge. While previous work has shown that for many sets of interest, κ is $O(d^\alpha)$ with $\alpha < 1$ (Mhammedi, 2022), and any convex set can be pre-processed (e.g., put in isotropic position) to ensure $\kappa \le d$ (Flaxman et al., 2005), this results in a potentially high pre-processing computational cost and a worst-case regret bound of $O(d\sqrt{T})$. In this paper, we show that the dependence on κ and the dimension d in the regret bounds for separation-based algorithms can be further improved.

Contributions. In this paper, we introduce a separation-based projection-free algorithm for OCO that improves upon the state-of-the-art guarantees for such algorithms. Specifically, our method achieves a regret bound of $\widetilde{O}(\sqrt{dT} + \kappa d)$ while making only $\widetilde{O}(1)$ calls to a separation oracle per round. Crucially, the main term of this bound, $\widetilde{O}(\sqrt{dT})$, is *independent* of the asphericity κ. As discussed earlier, existing separation-based algorithms can incur regret as large as $\widetilde{O}(d\sqrt{T})$ in the worst case, even after preprocessing the feasible set. Our bound improves this by a factor of $\sqrt{d}$, without any need for preprocessing.

As a by-product of our analysis, we provide an improved regret bound for projection-free online exp-concave optimization, which we required in an intermediate step of our analysis. Additionally, we recover the $O(\kappa\sqrt{T})$ regret bound for OCO with a simplified analysis.

By applying a standard online-to-batch conversion, our new $\widetilde{O}(\sqrt{dT})$ regret translates to a $\widetilde{O}(\sqrt{d/T})$ convergence rate in offline and stochastic optimization settings. Additionally, by leveraging our intermediate results on projection-free exp-concave optimization (which are of independent

Table 1: Comparison of projection-free regret bounds for OCO (see Section 2.1) and Stochastic Convex Optimization (SCO) (see Section 5). Here, $\kappa \coloneqq R/r$ represents the asphericity of the feasible set $\mathcal{K}$, where r and R are such that $\mathbb{B}(r) \subseteq \mathcal{K} \subseteq \mathbb{B}(R)$. For ill-conditioned sets, κ can be arbitrarily large. Unlike existing regret bounds, our new bound places κ in a lower-order term, rather than having it multiply $\sqrt{T}$. In SCO, σ^2 represents the variance of the stochastic gradients. In the offline setting ($\sigma = 0$), we achieve the fast rate of $O(\frac{\kappa d}{T})$.

Papers	Regret bound in OCO	Convergence rate in SCO	Oracle type	Number of oracle calls per round
(Hazan and Kale, 2012)	$O(T^{3/4})$	$O\left(\frac{1}{T^{1/3}}\right)$	Linear optimization	1
(Mhammedi, 2022) (Garber and Kretzu, 2022)	$O(\boldsymbol{\kappa}\sqrt{T})$	$O\left(\frac{\boldsymbol{\kappa}}{\sqrt{T}}\right)$	Separation	$O(1)$-$\widetilde{O}(1)$
This paper	$\widetilde{O}(\sqrt{dT} + \boldsymbol{\kappa} d)$	$\widetilde{O}\left(\boldsymbol{\sigma}\sqrt{\frac{d}{T}} + \frac{\boldsymbol{\kappa} d}{T}\right)$	Separation	$\widetilde{O}(1)$

interest), we achieve a faster convergence rate of $\widetilde{O}(\sigma\sqrt{d/T} + \kappa d/T)$. Notably, this rate simplifies to $\widetilde{O}(\kappa d/T)$ when $\sigma = 0$ (i.e., in the offline setting). Our results are summarized in Table 1.

Related works. Our approach builds on the projection-free reduction method introduced by Mhammedi (2022), which transforms any OCO problem over a feasible set $\mathcal{K}$ into an OCO problem over a ball $\mathbb{B}(R)$ containing $\mathcal{K}$ (i.e., $\mathcal{K} \subseteq \mathbb{B}(R)$), where Euclidean projections can be computed at a cost of $O(d)$. This method is closely related to the earlier "constrained-to-unconstrained" reduction by Cutkosky and Orabona (2018), but instead of relying on potentially expensive Euclidean projections onto $\mathcal{K}$, it uses Gauge projections, which can be performed efficiently with a logarithmic number of calls to a separation oracle for the feasible set. The reduction in Mhammedi (2022) has also been successfully applied outside the OCO setting, providing global non-asymptotic superlinear rates for a quasi-Newton method (Jiang et al., 2023; Jiang and Mokhtari, 2024), and for the design of extension functions in bandit convex optimization (Fokkema et al., 2024).

Our approach integrates the projection-free reduction from (Mhammedi, 2022) with the efficient exp-concave optimization algorithm introduced by Mhammedi and Gatmiry (2023). The latter is an Online Newton Step (`ONS`) method that implicitly tracks specific Follow-The-Regularized-Leader (`FTRL`) iterates, where the associated regularizer is the log-barrier for a Euclidean ball. By exploiting the unique structure of the log-barrier for the Euclidean ball, Mhammedi and Gatmiry (2023) show that the *generalized projections*, typically required by the classical `ONS` algorithm (Hazan et al., 2007) and often computationally expensive even for a ball (Koren, 2013), can be completely bypassed. Additionally, they prove that the algorithm requires only $\widetilde{O}(1)$ calls to a separation oracle per round, along with $\widetilde{O}(1)$ matrix-vector multiplications. While our primary focus is on OCO, we incorporate several of the techniques in Mhammedi and Gatmiry (2023) and, somewhat unexpectedly, achieve state-of-the-art guarantees in OCO by taking a detour through online exp-concave optimization.

Outline. In Section 2, we describe the OCO setup and introduce the necessary notation and definitions. In Section 3, we present `Barrier-ONS`, an efficient projection-free algorithm for exp-concave optimization over a ball, which forms a key component of our final method for OCO. In Section 4, we present our results for OCO and show how we reduce the problem to online exp-concave optimization over a ball. In Section 5, we extend these results to the stochastic and offline convex optimization settings, achieving a state-of-the-art convergence rate for projection-free stochastic convex optimization. We discuss the limitations our approach and future work in Section 6.

2. Preliminaries

In Section 2, we formally introduce the OCO setup and the notation used. In Section 2.2, we present key convex analysis notations and preliminary results that are used throughout the paper.

2.1. Setup and Notation

Throughout, let $\mathcal{K}$ denote a closed convex subset of the Euclidean space $\mathbb{R}^d$. We consider the standard OCO setup over $\mathcal{K}$, where an algorithm produces iterates within $\mathcal{K}$ over multiple rounds. At the start of each round $t \geq 1$, the algorithm outputs $\boldsymbol{w}_t$ and incurs a loss $f_t(\boldsymbol{w}_t)$, where $f_t : \mathcal{K} \to \mathbb{R}$ is a convex function, potentially chosen adversarially based on the history and $\boldsymbol{w}_t$. As is typical in OCO literature, we assume that the algorithm observes only a subgradient $\boldsymbol{g}_t \in \partial f_t(\boldsymbol{w}_t)$, rather than the full function. The algorithm's performance is evaluated in terms of regret after $T \geq 1$ rounds:

$$\mathtt{Reg}_T \coloneqq \sum_{t=1}^T f_t(\boldsymbol{w}_t) - \inf_{\boldsymbol{w} \in \mathcal{K}} \sum_{t=1}^T f_t(\boldsymbol{w}).$$

By the convexity of (f_t), $\mathtt{Reg}_T$ is bounded from above by the *linearized regret*: $\sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle - \inf_{\boldsymbol{w} \in \mathcal{K}} \sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{w} \rangle$. Therefore, to bound $\mathtt{Reg}_T$, it is sufficient to bound the linearized regret.

Our goal in this paper is to design an efficient OCO algorithm that achieves sublinear regret while requiring only a logarithmic number of calls per round to a separation oracle for the feasible set $\mathcal{K}$, rather than relying on Euclidean projections.

Definition 2.1 (Separation oracle)**.** *A* Separation oracle $\mathrm{Sep}_{\mathcal{C}}$ *for a set $\mathcal{C}$ is an oracle that given $\boldsymbol{w} \in \mathbb{R}^d$ returns $(b, \boldsymbol{v}) \in \{0,1\} \times \mathbb{B}(1)$ (where $\mathbb{B}(1)$ denotes the unit Euclidean ball in $\mathbb{R}^d$), such that*

- *$b = 0$ and $\boldsymbol{v} = \mathbf{0}$, if $\boldsymbol{w} \in \mathcal{C}$; and otherwise,*
- *$b = 1$ and $\langle \boldsymbol{v}, \boldsymbol{w} \rangle > \langle \boldsymbol{v}, \boldsymbol{u} \rangle$, for all $\boldsymbol{u} \in \mathcal{K}$.*

We denote by $C_{\mathrm{sep}}(\mathcal{C})$ the computational cost of one call to this oracle.

We consider the OCO problem described above and additionally assume that the functions (f_t) are G-Lipschitz, for some $G > 0$, and that $\mathcal{K}$ is "sandwiched" between two Euclidean balls with radii r and R. To formalize these assumptions, let $\|\cdot\|$ denote the Euclidean norm, and $\mathbb{B}(\gamma) \subset \mathbb{R}^d$ represent the Euclidean ball of radius $\gamma > 0$.

Assumption 2.1. *The set $\mathcal{K} \subseteq \mathbb{R}^d$ is a closed and convex and there are some $r, R > 0$ such that*

$$\mathbb{B}(r) \subseteq \mathcal{K} \subseteq \mathbb{B}(R).$$

Assumption 2.2. *There is some $G > 0$, such that for all $t \geq 1$, the function $f_t : \mathcal{K} \to \mathbb{R}$ is convex and for all $\boldsymbol{w} \in \mathcal{K}$ and $\boldsymbol{g}_t \in \partial f_t(\boldsymbol{w})$, we have $\|\boldsymbol{g}_t\| \leq G$.*

Additional notation. We denote by $\mathcal{K}^\circ \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq 1, \forall \boldsymbol{y} \in \mathcal{K}\}$ the *polar* set of $\mathcal{K}$ (Hiriart-Urruty and Lemaréchal, 2004). We denote by $\operatorname{int} \mathcal{K}$ the interior of a set $\mathcal{K}$. Given a function $f : \mathcal{C} \to \mathbb{R}$ on a compact set $\mathcal{C}$, we let $\arg\min_{\boldsymbol{x} \in \mathcal{C}} f(\boldsymbol{x})$ denote the subset of points in $\mathcal{C}$ that minimize the function f. We use $\widetilde{O}(\cdot)$ to denote a bound up to factors polylogarithmic in parameters appearing in the expression.

Algorithm 1 `GaugeDist`$(\boldsymbol{w}; \mathcal{C}, \varepsilon, r)$: Approximate value and subgradient of the Gauge distance function.

require: Separation oracle $\mathrm{Sep}_{\mathcal{C}}$ for $\mathcal{C}$, input vector $\boldsymbol{w} \in \mathbb{R}^d$, and parameters $\varepsilon, r > 0$.
returns $S \approx S_{\mathcal{C}}(\boldsymbol{w})$ and $\boldsymbol{s} \approx \partial S_{\mathcal{C}}(\boldsymbol{w})$, where $S_{\mathcal{C}}(\boldsymbol{u}) := \inf_{\boldsymbol{x} \in \mathcal{C}} \gamma_{\mathcal{C}}(\boldsymbol{u} - \boldsymbol{x})$ is the Gauge distance.

```
Set (b, v) ← Sep_C(w).                              // b = 1 if w ∈ C and 0, otherwise.
if b = 1 then                                       // This corresponds to the case where w ∈ C.
    Set (S, s) ← (0, 0).
    return (S, s).
Set α ← 0, β ← 1, and μ ← (α + β)/2.
while β − α > r²ε / (2‖w‖²) do
    Set (b, v) ← Sep_C(μw).                         // b = 1 if μw ∈ C and 0, otherwise.
    Set α ← μ if b = 1; and β ← μ otherwise.
    Set μ ← (α + β)/2.
Set S ← α⁻¹ − 1 and s ← v / (β·vᵀw).
return (S, s).
```

2.2. Gauge Distance and Projection

We now introduce some convex analysis concepts and preliminary results that will be used throughout the paper, beginning with the notion of a Gauge function (a.k.a. Minkowski functional (Hiriart-Urruty and Lemaréchal, 2004)). In this section, let $\mathcal{C} \subseteq \mathbb{R}^d$ represent a closed convex set that contains the origin in its interior.

Definition 2.2. *The Gauge function $\gamma_{\mathcal{C}} : \mathbb{R}^d \to \mathbb{R}$ of the set $\mathcal{C}$ is defined as*

$$\gamma_{\mathcal{C}}(\boldsymbol{u}) := \inf\{\lambda \in \mathbb{R}_{\geq 0} \mid \boldsymbol{u} \in \lambda \mathcal{C}\}.$$

The Gauge function $\gamma_{\mathcal{C}}$ can be viewed as a "pseudo" norm induced by the convex set $\mathcal{C}$; it becomes a true norm when $\mathcal{C}$ is centrally symmetric (i.e., $\mathcal{C} = -\mathcal{C}$). With the Gauge function defined, we can introduce the Gauge distance (Mhammedi, 2022), a key concept in the approach of this paper.

Definition 2.3 (Gauge distance)**.** *The Gauge distance function $S_{\mathcal{C}}$ corresponding to the set $\mathcal{C}$ is*

$$S_{\mathcal{C}}(\boldsymbol{u}) := \inf_{\boldsymbol{x} \in \mathcal{C}} \gamma_{\mathcal{C}}(\boldsymbol{u} - \boldsymbol{x}), \quad \boldsymbol{u} \in \mathbb{R}^d.$$

This naturally leads to the concept of the Gauge projection (Mhammedi, 2022).

Definition 2.4. *The Gauge projection operator $\Pi^{\mathrm{gau}}_{\mathcal{C}}$ induced by $\mathcal{C}$ is the set-valued mapping:*

$$\Pi^{\mathrm{gau}}_{\mathcal{C}}(\boldsymbol{u}) := \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathcal{C}} \gamma_{\mathcal{C}}(\boldsymbol{u} - \boldsymbol{x}), \quad \boldsymbol{u} \in \mathbb{R}^d.$$

Our projection-free OCO approach in this paper uses Gauge projections instead of Euclidean projections. As we will see shortly, Gauge projections can be performed efficiently using a separation oracle. To understand this, we first need the following result from (Mhammedi, 2022), which allows us to express both the Gauge distance and its subgradients in terms of the Gauge function $\gamma_{\mathcal{C}}$.

Lemma 2.1. *Suppose that the set $\mathcal{C}$ satisfies $\mathbf{0} \in \operatorname{int} \mathcal{C}$. Then, for any $\boldsymbol{w} \in \mathbb{R}^d$, we have*

$$\Pi_{\mathcal{C}}^{\mathrm{gau}}(\boldsymbol{w}) = \begin{cases} \frac{\boldsymbol{w}}{\gamma_{\mathcal{C}}(\boldsymbol{w})}, & \text{if } \boldsymbol{w} \notin \mathcal{C}; \\ \boldsymbol{w}, & \text{otherwise}. \end{cases}$$

and the Gauge distance function satisfies:

$$S_{\mathcal{C}}(\boldsymbol{w}) = \max(0, \gamma_{\mathcal{C}}(\boldsymbol{w}) - 1); \text{ and } \partial S_{\mathcal{C}}(\boldsymbol{w}) = \begin{cases} \partial \gamma_{\mathcal{C}}(\boldsymbol{w}) = \arg\max_{\boldsymbol{x} \in \mathcal{C}^\circ} \langle \boldsymbol{x}, \boldsymbol{w} \rangle, & \text{if } \boldsymbol{w} \notin \mathcal{C}; \\ \{\mathbf{0}\}, & \text{if } \boldsymbol{w} \in \mathcal{C}. \end{cases} \tag{1}$$

The key implication of Lemma 2.1 is that, to compute the Gauge distance and its subgradients—both of which are needed for our OCO algorithm—it is sufficient to compute or approximate the Gauge function $\gamma_{\mathcal{C}}$ and its subgradients. This is accomplished through Algorithm 1, whose guarantee we now present (with the proof provided in Appendix G).

Lemma 2.2. *Let $\varepsilon, r > 0$ and $\boldsymbol{w} \in \mathbb{R}^d$ be given, and suppose that $\mathbb{B}(r) \subseteq \mathcal{C}$. Consider a call to Algorithm 1 with input $(\mathcal{C}, \boldsymbol{w}, \varepsilon, r)$. Then, the output $(S, \boldsymbol{s})$ of Algorithm 1 satisfies*

$$\|\boldsymbol{s}\| \le 1/r, \quad S_{\mathcal{C}}(\boldsymbol{w}) \le S \le S_{\mathcal{C}}(\boldsymbol{w}) + \varepsilon, \text{ and } \forall \boldsymbol{u} \in \mathbb{R}^d, \quad S_{\mathcal{C}}(\boldsymbol{u}) \ge S_{\mathcal{C}}(\boldsymbol{w}) + (\boldsymbol{u} - \boldsymbol{w})^\top \boldsymbol{s} - \varepsilon. \tag{2}$$

Algorithm 1 makes at most $1 + \log_2(\frac{4\|\boldsymbol{w}\|^2}{r^2 \varepsilon})$ calls to the separation oracle $\mathrm{Sep}_{\mathcal{C}}$ in Definition 2.1.

3. Barrier-Regularized Online Newton Steps over a Ball

In this section, we present `Barrier-ONS` (Algorithm 2), a key component of our projection-free OCO approach detailed in the next section. `Barrier-ONS` generates online Newton iterates that implicitly track specific Follow-The-Regularized-Leader (`FTRL`) iterates, where the associated regularizer is the log-barrier for a Euclidean ball.

Algorithm 2 `Barrier-ONS`: Barrier-regularized `ONS` over a Euclidean ball. Pseudocode of Algorithm 4.

---

inputs: Number of rounds $T \ge 1$, and parameters $\eta, \nu, c > 0$.

Set $\boldsymbol{z}_1 \leftarrow \mathbf{0}$, $\boldsymbol{u}_1 \leftarrow \mathbf{0}$, and $m \leftarrow \mathfrak{c} \cdot \log_c(dT)$, with $\mathfrak{c}$ a sufficiently large universal constant.

for $t = 1, \dots, T$ **do**

  Play $\boldsymbol{u}_t$ and observe $\widetilde{\boldsymbol{g}}_t$. // $\widetilde{g}_t$ may be chosen adversarially as in OCO setting—see Remark 3.1

  Set $\Sigma_t \leftarrow \left(\frac{2\nu I}{R^2 - \|\boldsymbol{z}_t\|^2} + \frac{4\nu \boldsymbol{u}_t \boldsymbol{u}_t^\top}{(R^2 - \|\boldsymbol{u}_t\|^2)^2} + \eta \sum_{s=1}^t \widetilde{\boldsymbol{g}}_s \widetilde{\boldsymbol{g}}_s^\top\right)^{-1}$.

  // Σ_t can be computed in $O(d^2)$ most rounds—see Algorithm 4

  Set $\gamma_t \leftarrow \frac{2\nu}{R^2 - \|\boldsymbol{z}_t\|^2} - \frac{2\nu}{R^2 - \|\boldsymbol{u}_t\|^2}$.

  Let Φ_{t+1} be as in (3) with t replaced by $t+1$.

  Set $H_t \leftarrow \sum_{k=1}^{m+1} \gamma_t^{k-1} \Sigma_t^k$. // Taylor approximation of $\nabla^{-2}\Phi_{t+1}(\boldsymbol{u}_t)$

  Set $\boldsymbol{u}_{t+1} \leftarrow \boldsymbol{u}_t - H_t \nabla \Phi_{t+1}(\boldsymbol{u}_t)$. // Approximate Newton step.

  /* Check if the Taylor approximation point needs to be updated */

  if $|\|\boldsymbol{u}_{t+1}\|^2 - \|\boldsymbol{z}_t\|^2| \le c \cdot (R^2 - \|\boldsymbol{z}_t\|^2)$ **then**

    Set $\boldsymbol{z}_{t+1} \leftarrow \boldsymbol{z}_t$.

  else

    Set $\boldsymbol{z}_{t+1} \leftarrow \boldsymbol{u}_{t+1}$.

`Barrier-ONS` was originally proposed by Mhammedi and Gatmiry (2023) in the context of online and stochastic exp-concave optimization as a more computationally efficient alternative to the classical Online Newton Step (`ONS`) algorithm (Hazan et al., 2007). Unlike `ONS`, `Barrier-ONS` eliminates the need for *generalized projections* onto the feasible set, which can be computationally expensive, even for a Euclidean ball. This is achieved by generating iterates that remain close to the `FTRL` iterates, which stay within the interior of the feasible set due to the log-barrier. In this paper, we adopt `Barrier-ONS` from (Mhammedi and Gatmiry, 2023) with only minor modifications, while improving its analysis and guarantees for the online exp-concave optimization setting.

We now provide a brief description of `Barrier-ONS` (Algorithm 2); for further details, the reader may refer to (Mhammedi and Gatmiry, 2023). Algorithm 2 offers a simplified version of Algorithm 4, written for ease of understanding. While Algorithm 2 is technically equivalent to Algorithm 4, it abstracts away the details of how certain steps can be implemented efficiently. Algorithm 4, on the other hand, makes these efficiency considerations explicit. For now, we will focus on Algorithm 2.

3.1. `Barrier-ONS`: Algorithm Description

To describe `Barrier-ONS` (Algorithm 2), we first need to introduce a series of objective functions (Φ_t). Given parameters $\eta, \nu, R > 0$ and the history of observed loss vectors $(\widetilde{\boldsymbol{g}}_s)_{s<t}$ before round $t \geq 1$, the objective function Φ_t is defined as:

$$\Phi_t(\boldsymbol{x}) \coloneqq \Psi(\boldsymbol{x}) + \frac{\eta}{2}\sum_{s=1}^{t-1}\langle \widetilde{\boldsymbol{g}}_s, \boldsymbol{x} - \boldsymbol{u}_s\rangle^2 + \boldsymbol{x}^\top \sum_{s=1}^{t-1} \widetilde{\boldsymbol{g}}_s, \quad \text{where} \quad \Psi(\boldsymbol{x}) \coloneqq -\nu \log(R^2 - \|\boldsymbol{x}\|^2). \tag{3}$$

Remark 3.1. *In this section, the sequence $(\widetilde{\boldsymbol{g}}_t)$ is arbitrary and may be chosen adversarially, as in the standard OCO setting. We use the notation $(\widetilde{\boldsymbol{g}}_t)$ instead of $(\boldsymbol{g}_t)$ to reserve the latter for the subgradients observed by our final algorithm (Algorithm 3) in Section 4.*

Note that Ψ/ν is the standard log-barrier for $\mathbb{B}(R)$. The iterates $(\boldsymbol{u}_t)$ of `Barrier-ONS` are approximate online Newton iterates with respect to the objective functions (Φ_t) in the sense that

$$\boldsymbol{u}_{t+1} \approx \boldsymbol{u}_t - \nabla^{-2}\Phi_{t+1}(\boldsymbol{u}_t)\nabla\Phi_{t+1}(\boldsymbol{u}_t), \quad \text{for all } t \in [T]. \tag{4}$$

The iterate $\boldsymbol{u}_{t+1}$ in `Barrier-ONS` is only an approximation of the right-hand side of (4) because `Barrier-ONS` does not compute the inverse Hessian $\nabla^{-2}\Phi_{t+1}(\boldsymbol{u}_t)$ exactly at each iteration. Instead, it approximates it using a Taylor expansion around a neighboring point (see Line 8 of Algorithm 2).

The motivation behind this approach is that computing the Taylor expansion is significantly cheaper than calculating the exact inverse Hessian. Instead of performing expansions around a fixed point, the algorithm updates the current expansion point $\boldsymbol{z}_t$ whenever the next iterate $\boldsymbol{u}_{t+1}$ drifts too far from $\boldsymbol{z}_t$; see Line 9-Line 12. A full inverse Hessian is computed only when the Taylor expansion point is updated. A key insight from (Mhammedi and Gatmiry, 2023) is that the iterates of `Barrier-ONS` are stable enough to ensure that the Taylor expansion point needs to be updated only $O(\sqrt{T})$ times over T rounds, as the next lemma states (the proof can be found in Appendix D.2).

Lemma 3.1 (Stability). *Let $\eta, \nu, \widetilde{G}, R > 0$, $T \geq 1$, and $c \in (0,1)$ be given. Consider a call to Algorithm 2 with input parameters (T, η, ν, c), and let $(\boldsymbol{z}_t)$ be the Taylor expansion points in Algorithm 4. Further, suppose that $\widetilde{\boldsymbol{g}}_t \in \mathbb{B}(\widetilde{G})$, for all $t \in [T]$; $10\widetilde{G}R \leq \nu \leq 10d\widetilde{G}RT$; and $\eta \leq \frac{1}{5\widetilde{G}R}$. Then, it holds that $\sum_{t=1}^{T-1}\mathbb{I}\{\boldsymbol{z}_{t+1} \neq \boldsymbol{z}_t\} \leq \frac{52}{c}\sqrt{T \cdot \left(1 + \frac{d}{\nu\eta}\log(1 + \frac{T}{d})\right)}$.*

Computational cost. From a computational perspective, this result is highly promising because it implies that, as long as η and ν are chosen such that $\eta\nu \geq d$, a full Hessian inverse is only required in a $\widetilde{O}(T^{-1/2})$ fraction of the rounds. For the remaining rounds, the computational cost per round is $\widetilde{O}(d^2)$ due to matrix-vector multiplications. As a result, the total computational cost after T rounds is $\widetilde{O}(d^2 T + d^{\omega}\sqrt{T})$. Therefore, the average per-round computational cost of `Barrier-ONS` is $\widetilde{O}(d^2)$, assuming $\eta\nu \geq d$ and $T \geq d$(and a matrix multiplication exponent $\omega \leq 5/2$).

Feasibility of the iterates. As we show in the analysis, the fact that $(\boldsymbol{u}_t)$ are approximate online Newton iterates (in the sense of (4)) ensures that $(\boldsymbol{u}_t)$ remain within $\mathbb{B}(R)$; it is known (see e.g. Abernethy et al. (2012)) that the exact online Newton iterates with respect to (Φ_t) are guaranteed to stay within $\operatorname{int}\mathbb{B}(R)$ due to the self-concordance properties of Ψ in the definition of (Φ_t).

3.2. Regret Guarantee

The fact that $(\boldsymbol{u}_t)$ satisfy (4) essentially means that it suffices to bound the regret of the Newton iterates. This is advantageous because the (exact) Newton iterates are known to be close to the `FTRL` iterates $(\boldsymbol{w}_t)$ with respect to (Φ_t):

$$\boldsymbol{w}_t \in \underset{\boldsymbol{w}\in\mathbb{R}^d}{\arg\min}\, \Phi_t(\boldsymbol{w}). \tag{5}$$

Due to the curvature of the log-barrier Ψ in the definition of Φ_t in (3), a standard `FTRL` analysis leads to the following regret bound for the iterates $(\boldsymbol{w}_t)$ (the proof is in Appendix D.4).

Lemma 3.2 (Regret of `FTRL`)**.** *Let $\eta, \nu \in (0,1)$, $R > 0$, and $\widetilde{G} > 0$ be such that $\eta \leq \frac{1}{5\widetilde{G}R}$ and $\nu \geq 10\widetilde{G}R$. Further, let $(\widetilde{\boldsymbol{g}}_t) \subset \mathbb{R}^d$ be a sequence of vectors such that $\|\widetilde{\boldsymbol{g}}_t\| \leq \widetilde{G}$, for all $t \geq 1$. Then, the `FTRL` iterates $(\boldsymbol{w}_t)$ in (5) satisfy, for all $\boldsymbol{w} \in \operatorname{int}\mathbb{B}(R)$,*

$$\sum_{t=1}^{T}\langle \boldsymbol{w}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle \leq \sum_{t=1}^{T}\frac{\eta}{2}\langle \boldsymbol{w}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle^2 + \Psi(\boldsymbol{w}) - \Psi(\boldsymbol{0}) + \frac{3d\log(1+T/d)}{\eta}.$$

So far, we have outlined that the `Barrier-ONS` iterates are approximate Newton iterates, which in turn approximate the `FTRL` iterates. Using these insights, along with the regret bound for `FTRL` in Lemma 3.2, we can derive the following regret bound for `Barrier-ONS` (the proof is in Appendix D.5).

Theorem 3.1 (Regret of `Barrier-ONS`)**.** *Let $c \in (0,1)$, $T \in \mathbb{N}$, and $\widetilde{G} > 0$ be given and consider a call to Algorithm 2 with input parameters (T, η, ν, c) such that $\eta \leq \frac{1}{5\widetilde{G}R}$ and $10\widetilde{G}R \leq \nu \leq 10dT\widetilde{G}R$. If the loss vectors $(\widetilde{\boldsymbol{g}}_t)$ in Algorithm 2 satisfy $(\widetilde{\boldsymbol{g}}_t) \subset \mathbb{B}(\widetilde{G})$, then the iterates $(\boldsymbol{u}_t)$ of Algorithm 2 satisfy $(\boldsymbol{u}_t) \subset \mathbb{B}(R)$ and for all $\boldsymbol{w} \in \operatorname{int}\mathbb{B}(R)$,*

$$\sum_{t=1}^{T}\left(\langle \boldsymbol{u}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle - \frac{\eta}{2}\langle \boldsymbol{u}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle^2\right) \leq \widetilde{G}R - \nu\log(1 - \tfrac{\|\boldsymbol{w}\|^2}{R^2}) + \frac{18d\log(1+T/d)}{5\eta}. \tag{6}$$

Further, there is an implementation of Algorithm 2, which we display in Algorithm 4, with a total computational cost bounded by $\widetilde{O}\left(d^2 T\log_c(dT) + c^{-1}d^{\omega}\sqrt{\frac{d}{\nu\eta}T}\right)$.

3.3. Link to Online Exp-Concave Optimization

The bound in Theorem 3.1 immediately implies logarithmic regret for online exp-concave optimization. In this setting, $(\widetilde{\boldsymbol{g}}_t)$ are the subgradients of exp-concave functions (ℓ_t), where a function $\ell : \mathcal{C} \to \mathbb{R}^d$ over a convex set $\mathcal{C}$ is α-exp-concave if the mapping $\boldsymbol{x} \mapsto e^{-\alpha \ell(\boldsymbol{x})}$ is concave over $\mathcal{C}$. It is well known (see e.g., Hazan et al. (2007)) that for a $\widetilde{G}$-Lipschitz, α-exp-concave function ℓ, and as long as $\mathcal{C} \subseteq \mathbb{B}(R)$, we have the following for all $\eta \le \frac{1}{2}\min\left(\frac{1}{4R\widetilde{G}}, \alpha\right)$:

$$\forall \boldsymbol{u}, \boldsymbol{w} \in \mathcal{C}, \forall \widetilde{\boldsymbol{g}} \in \partial \ell(\boldsymbol{u}), \quad \ell(\boldsymbol{u}) - \ell(\boldsymbol{w}) \le \langle \boldsymbol{u} - \boldsymbol{w}, \widetilde{\boldsymbol{g}} \rangle - \frac{\eta}{2} \cdot \langle \boldsymbol{u} - \boldsymbol{w}, \widetilde{\boldsymbol{g}} \rangle^2.$$

This implies that for α-exp-concave losses (ℓ_t), the corresponding regret $\sum_{t=1}^T (\ell_t(\boldsymbol{u}_t) - \ell_t(\boldsymbol{w}))$ can be bounded by the left-hand side of (6), and thus Theorem 3.1 implies that `Barrier-ONS` achieves logarithmic regret in this case. Additionally, the term $\sum_{t=1}^T \left(\langle \boldsymbol{u}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t \rangle - \frac{\eta}{2}\langle \boldsymbol{u}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t \rangle^2\right)$ on the left-hand side of (6) can itself be interpreted as the regret corresponding to the losses

$$\ell_t : \boldsymbol{x} \mapsto \langle \boldsymbol{x}, \widetilde{\boldsymbol{g}}_t \rangle + \frac{\eta}{2}\langle \boldsymbol{u}_t - \boldsymbol{x}, \widetilde{\boldsymbol{g}}_t \rangle^2, \tag{7}$$

which are exp-concave for a certain range of η's.

Regret improvement over prior work. Compared to (6), the regret bound in (Mhammedi and Gatmiry, 2023) includes additional terms like $\widetilde{O}(\widetilde{G}/\eta)$ and $O(\widetilde{G}^2)$, with the removal of the latter left as an open problem. As discussed in the next section, in the context of projection-free OCO, $\widetilde{G}$ is set to κG, where $\kappa \coloneqq R/r$, with r and R defined in Assumption 2.1, and G is the Lipschitz constant of the losses (see Assumption 2.2). Consequently, having κ multiply $1/\eta$ (as in the bound from (Mhammedi and Gatmiry, 2023)) would hinder effective tuning of η to achieve our desired $\widetilde{O}(\sqrt{dT})$ regret bound for OCO. The improved regret bound for `Barrier-ONS` in Theorem 3.1 resolves this issue, as we will see in the next section.

Algorithm 3 OCO reduction with Gauge projections.

require: Number of rounds T, and an OCO algorithm $\mathcal{B}$ over $\mathbb{R}^d$.
Set $\varepsilon = 1/T$ and $\boldsymbol{u}_1 = \boldsymbol{0}$.
Initialize $\mathcal{B}$, and set $\boldsymbol{u}_1$ to $\mathcal{B}$'s first output.
for $t = 1, \dots, T$ **do**
  Set $(S_t, \boldsymbol{s}_t) \leftarrow$ `GaugeDist`$(\boldsymbol{u}_t; \mathcal{K}, \varepsilon, r)$. // $S_t \approx S_{\mathcal{K}}(\boldsymbol{u}_t)$ and $\boldsymbol{s}_t \approx \partial S_{\mathcal{K}}(\boldsymbol{u}_t)$.
  Play $\boldsymbol{w}_t = \frac{\boldsymbol{u}_t}{1+S_t}$. // $\boldsymbol{w}_t$ represents an approximate Gauge projection of $\boldsymbol{u}_t$ onto $\mathcal{K}$.
  Observe subgradient $\boldsymbol{g}_t \in \partial f_t(\boldsymbol{w}_t)$.
  Set $\widetilde{\boldsymbol{g}}_t = \boldsymbol{g}_t - \mathbb{I}\{\langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle < 0\} \cdot \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle \cdot \boldsymbol{s}_t$.
  Send $\widetilde{\boldsymbol{g}}_t$ to $\mathcal{B}$ as the tth loss vector.
  Set $\boldsymbol{u}_{t+1} \in \mathbb{R}^d$ to $\mathcal{B}$'s $(t+1)$th output given the history $(\boldsymbol{u}_s, \widetilde{\boldsymbol{g}}_s)_{s \le t}$.

4. Projection-Free OCO via Exp-Concave Optimization

In this section, we show how, through Algorithm 3, we can effectively reduce an OCO problem over $\mathcal{K}$ to an online exp-concave problem over a Euclidean ball that contains the feasible set $\mathcal{K}$, allowing us to apply `Barrier-ONS` from Section 3. Crucially, this reduction only requires Gauge projections (Definition 2.4), which are inexpensive to approximate using a separation oracle; see Section 2.2. We now provide an overview of our reduction and will elaborate on some of the steps in the sequel.

4.1. Overview of Reduction

To reduce OCO over $\mathcal{K}$ to online exp-concave optimization over $\mathbb{B}(R) \supseteq \mathcal{K}$, we use Algorithm 3. This is a wrapper algorithm, which we refer to as $\mathcal{A}$ in the sequel, that wraps around a base algorithm $\mathcal{B}$. We will instantiate $\mathcal{B}$ as `Barrier-ONS` over the ball $\mathbb{B}(R) \supseteq \mathcal{K}$.

The outputs $(\boldsymbol{w}_t)$ of $\mathcal{A}$ are the Gauge projections of the outputs $(\boldsymbol{u}_t)$ of the base algorithm $\mathcal{B}$. By definition of the Gauge projection, these iterates are guaranteed to be feasible, i.e., $(\boldsymbol{w}_t) \subset \mathcal{K}$.

At each round t, the wrapper algorithm $\mathcal{A}$ observes a subgradient $\boldsymbol{g}_t \in \partial f_t(\boldsymbol{w}_t)$ and constructs a surrogate subgradient $\widetilde{\boldsymbol{g}}_t$ (Line 7 of Algorithm 3), which is then passed to $\mathcal{B}$ as a loss vector. As we elaborate in the sequel, the surrogate vectors $(\widetilde{\boldsymbol{g}}_t)$ are constructed to satisfy the bound $\|\widetilde{\boldsymbol{g}}_t\| \leq 2G\kappa$ for all t, where $\kappa \coloneqq \frac{R}{r}$ (with $r, R > 0$ as defined in Assumption 2.1), and for all $\boldsymbol{w} \in \mathcal{K}$,

$$\sum_{t=1}^{T}\left(\langle \boldsymbol{w}_t - \boldsymbol{w}, \boldsymbol{g}_t\rangle - \frac{\eta}{2}\langle \boldsymbol{w}_t - \boldsymbol{w}, \boldsymbol{g}_t\rangle^2\right) \leq \sum_{t=1}^{T}\left(\langle \boldsymbol{u}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle - \frac{\eta}{2}\langle \boldsymbol{u}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle^2\right) + O(GR); \tag{8}$$

Note that the sum on the right-hand side of (8) corresponds to the regret of the base algorithm $\mathcal{B}$ with respect to the exp-concave losses (ℓ_t) defined in (7). Eq. (8) thus shows that the regret of the wrapper algorithm $\mathcal{A}$ (the first sum on the left-hand side) can be controlled by the regret of $\mathcal{B}$ in an exp-concave setting. This effectively reduces OCO over $\mathcal{K}$ to online exp-concave optimization over the ball $\mathbb{B}(R)$ containing $\mathcal{K}$.

By setting $\mathcal{B}$ to `Barrier-ONS`, and combining (8) with the regret bound of `Barrier-ONS` in Theorem 3.1 and the fact that $(\widetilde{\boldsymbol{g}}_t) \subseteq \mathbb{B}(2\kappa G)$, we essentially obtain

$$\forall \boldsymbol{w} \in \mathcal{K}, \quad \sum_{t=1}^{T}\langle \boldsymbol{w}_t - \boldsymbol{w}, \boldsymbol{g}_t\rangle \leq \frac{\eta}{2}\sum_{t=1}^{T}\langle \boldsymbol{w}_t - \boldsymbol{w}, \boldsymbol{g}_t\rangle^2 + \frac{d}{\eta}\log\left(1 + \frac{T}{d}\right) + O(GR). \tag{9}$$

Finally, tuning $\eta \propto \min\left(\frac{1}{\kappa GR}, \frac{1}{GR}\sqrt{\frac{d}{T}}\right)$ gives the desired $\widetilde{O}(\sqrt{dT} + \kappa d)$ regret bound.

The construction of the surrogate subgradients $(\widetilde{\boldsymbol{g}}_t)$ in Line 7, which ensures that (8) holds, is inspired by the approach of Mhammedi (2022). We now provide further details on the choice of $(\widetilde{\boldsymbol{g}}_t)$ and defer a proof sketch of (8) to Section 4.2.

Choice of surrogate subgradients. At round $t \geq 1$, given $\boldsymbol{u}_t$, $\boldsymbol{w}_t$, and $\boldsymbol{g}_t$, Algorithm 3 sets the surrogate subgradient as $\widetilde{\boldsymbol{g}}_t \approx \boldsymbol{g}_t^\star$, where

$$\boldsymbol{g}_t^\star \in \boldsymbol{g}_t - \mathbb{I}\{\langle \boldsymbol{g}_t, \boldsymbol{u}_t\rangle < 0\} \cdot \langle \boldsymbol{g}_t, \boldsymbol{w}_t\rangle \cdot \partial S_{\mathcal{K}}(\boldsymbol{u}_t), \tag{10}$$

and $S_{\mathcal{K}}(\boldsymbol{u}) \coloneqq \max(0, \gamma_{\mathcal{K}}(\boldsymbol{u}) - 1)$ is the Gauge distance function. Algorithm 3 computes an approximate subgradient $\boldsymbol{s}_t$ of $S_{\mathcal{K}}$ at $\boldsymbol{u}_t$ using the `GaugeDist` base algorithm. By Lemma 2.2, we have that $S_{\mathcal{K}}(\boldsymbol{w}) \geq S_{\mathcal{K}}(\boldsymbol{u}_t) + \langle \boldsymbol{w} - \boldsymbol{u}_t, \boldsymbol{s}_t\rangle - \varepsilon$ for all $\boldsymbol{w} \in \mathbb{R}^d$ (ε is set to $1/T$ in Algorithm 3), which essentially shows that $\boldsymbol{s}_t$ is an approximate subgradient of $S_{\mathcal{K}}$ at $\boldsymbol{u}_t$. This, in turn, implies that $\widetilde{\boldsymbol{g}}_t \approx \boldsymbol{g}_t^\star$ by Line 7 of Algorithm 3 and (10). Moreover, as noted in Lemma 2.2, computing $\widetilde{\boldsymbol{g}}_t$ is computationally inexpensive, requiring only $\widetilde{O}(1)$ calls to the separation oracle.

Approximate Gauge projections. To ensure feasible iterates, Algorithm 3 sets $(\boldsymbol{w}_t)$ as approximate Gauge projections of the outputs $(\boldsymbol{u}_t)$ from $\mathcal{B}$ onto $\mathcal{K}$; i.e., $\boldsymbol{w}_t \approx \boldsymbol{w}_t^\star$, where $\boldsymbol{w}_t^\star \in \arg\min_{\boldsymbol{x}\in\mathcal{K}} \gamma_{\mathcal{K}}(\boldsymbol{u}_t - \boldsymbol{x})$, $\forall t \geq 1$. By Lemma 2.1, the exact Gauge projection points $(\boldsymbol{w}_t^\star)$ satisfy

$$\boldsymbol{w}_t^\star = \frac{\boldsymbol{u}_t}{1 + S_{\mathcal{K}}(\boldsymbol{u}_t)}, \tag{11}$$

since $S_{\mathcal{K}}(\boldsymbol{u}_t) = \max(0, \gamma_{\mathcal{K}}(\boldsymbol{u}_t) - 1)$; see (1). In Algorithm 3, we approximate $\gamma_{\mathcal{K}}(\boldsymbol{u}_t)$ using the `GaugeDist` base algorithm (Algorithm 1). By Lemma 2.2, we have that S_t in Algorithm 3 satisfies $S_{\mathcal{K}}(\boldsymbol{u}_t) \le S_t \le S_{\mathcal{K}}(\boldsymbol{u}_t) + \varepsilon$. Thus, by Line 5 of Algorithm 3 and (11), we indeed have that $\boldsymbol{w}_t \approx \boldsymbol{w}_t^\star$ and $\boldsymbol{w}_t \in \mathcal{K}$ for all $t \ge 1$ (the details are in the proof of Lemma 4.1 in Appendix E.1). As noted in Lemma 2.2, a call to `GaugeDist` requires only $\widetilde{O}(1)$ calls to the separation oracle $\mathrm{Sep}_{\mathcal{K}}$. Therefore, Algorithm 3 ensures feasibility with only a logarithmic number of oracle calls.

4.2. Guarantees of Reduction

The specific choice of surrogate subgradients in (10) ensures that $\langle \boldsymbol{g}_t, \boldsymbol{w}_t^\star - \boldsymbol{w} \rangle \le \langle \boldsymbol{g}_t^\star, \boldsymbol{u}_t - \boldsymbol{w} \rangle$ for all $\boldsymbol{w} \in \mathcal{K}$; this follows from the analysis in (Mhammedi, 2022). Taking into account the approximation errors $\boldsymbol{w}_t \approx \boldsymbol{w}_t^\star$ and $\widetilde{\boldsymbol{g}}_t \approx \boldsymbol{g}_t^\star$, we obtain the following result (the proof is in Appendix E.1).

Lemma 4.1 (Key reduction result)**.** *Let $T \ge 1$ be given, and suppose that Assumption 2.1 and Assumption 2.2 hold. Further, let $(\boldsymbol{w}_t)$, $(\boldsymbol{u}_t)$, $(\boldsymbol{g}_t)$, and $(\widetilde{\boldsymbol{g}}_t)$ be as in Algorithm 3 with input T. If the iterates $(\boldsymbol{u}_t)$ of the base algorithm $\mathcal{B}$ satisfy $(\boldsymbol{u}_t) \subset \mathbb{B}(R)$, then we have $\|\widetilde{\boldsymbol{g}}_t\| \le 2\kappa \cdot G$, where $\kappa \coloneqq \frac{R}{r}$, and for all $t \in [T]$: $\boldsymbol{w}_t \in \mathcal{K}$ and for all $\boldsymbol{w} \in \mathcal{K}$, $\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w} \rangle \le \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w} \rangle + \frac{2GR}{T}$.*

By summing the last inequality in Lemma 4.1 over $t = 1, \dots, T$, we get that

$$\forall \boldsymbol{w} \in \mathcal{K}, \quad \sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w} \rangle \le \sum_{t=1}^{T} \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w} \rangle + 2GR. \tag{12}$$

Remark 4.1 (Recovering existing regret bounds)**.** *At this point, we can already recover the $O(\kappa\sqrt{T})$ regret bound of existing separation-based projection-free algorithms (Mhammedi, 2022; Garber and Kretzu, 2022). The right-hand side of (12) represents the regret of the base algorithm $\mathcal{B}$ (with respect to the losses $(\boldsymbol{w} \mapsto \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{w} \rangle)$). If we instantiate $\mathcal{B}$ as projected gradient descent over the ball $\mathbb{B}(R)$ (where Euclidean projections cost $O(d)$) and use the fact that $\|\widetilde{\boldsymbol{g}}_t\| \le 2\kappa G$ for all $t \ge 1$ (by Lemma 4.1), we obtain $\sum_{t=1}^{T} \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w} \rangle \le O(\kappa\sqrt{T})$. Combining this with (12) results in an overall $O(\kappa\sqrt{T})$ regret bound for Algorithm 2, which importantly makes only $\widetilde{O}(1)$ calls to a separation oracle per round. Note that `Barrier-ONS` was not needed for this part.*

Returning to our reduction, we note that the inequality in (12) is similar but does not exactly match the inequality we seek in (9), as the terms $\frac{\eta}{2}\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w} \rangle^2$ and $\frac{\eta}{2}\sum_{t=1}^{T} \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w} \rangle^2$ are missing from (12). However, we can still derive the inequality in (9), starting from Lemma 4.1. The full details are in the proof of Proposition 4.1 in Appendix E.2, but here we provide a sketch.

Suppose that $(\boldsymbol{u}_t) \subset \mathbb{B}(R)$ (which holds if $\mathcal{B}$ is set to `Barrier-ONS` by Theorem 3.1). Since the function $x \mapsto x - \eta x^2/2$ is non-decreasing for $x \le 1/\eta$, by choosing $\eta \le \frac{1}{10\kappa GR}$, setting x to $\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w} \rangle$ and $\langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w} \rangle + \frac{2GR}{T}$, and applying Lemma 4.1, we get for all $t \in [T]$ and $\boldsymbol{w} \in \mathcal{K}$:

$$\begin{aligned}
&\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w} \rangle - \tfrac{\eta}{2} \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w} \rangle^2 \\
&\le \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w} \rangle + \tfrac{2GR}{T} - \tfrac{\eta}{2} \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w} \rangle^2 - \tfrac{2\eta GT}{T} \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w} \rangle - \tfrac{2\eta G^2 R^2}{T^2}, \\
&\le \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w} \rangle - \tfrac{\eta}{2} \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w} \rangle^2 + \tfrac{3GR}{T},
\end{aligned} \tag{13}$$

where the last inequality follows by the facts that for all $t \in [T]$ and $\boldsymbol{w} \in \mathcal{K}$:

- $|\langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w} \rangle| \le \|\widetilde{\boldsymbol{g}}_t\| \cdot \|\boldsymbol{u}_t - \boldsymbol{w}\|$ by Cauchy Schwarz;

- $\|\widetilde{\boldsymbol{g}}_t\| \le 2\kappa G$ by Lemma 4.1;
- $\|\boldsymbol{u}_t - \boldsymbol{w}\| \le 2R$, since $\boldsymbol{u}_t, \boldsymbol{w} \in \mathbb{B}(R)$; and
- $\eta \le \frac{1}{10\kappa G R}$.

Summing (13) over $t = 1, \dots, T$ yields (8). As noted in Section 4.1, the sum on the right-hand side of (8) represents the regret of the base algorithm $\mathcal{B}$ with respect to the exp-concave losses $(\ell_t : \boldsymbol{x} \mapsto \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{x}\rangle + \frac{\eta}{2}\langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{x}\rangle^2)$. Thus, by setting $\mathcal{B}$ to `Barrier-ONS` and using the regret bound of `Barrier-ONS` in Theorem 3.1, we recover the regret bound in (9). We now state this result (the proof can be found in Appendix E.2).

Proposition 4.1. *Let $c \in (0,1)$ and $T \ge 1$ be given. Suppose that Assumption 2.1 and Assumption 2.2 hold. Consider a call to Algorithm 3 with input T and where the base algorithm $\mathcal{B}$ is an instance of* `Barrier-ONS` *(Algorithm 2) with input parameters (T, η, ν, c) satisfying $\eta \le \frac{1}{10\kappa R G}$, and $20\kappa G R \le \nu \le 20 d\kappa G R$, where κ is as in Lemma 4.1. Then, the sequences $(\boldsymbol{w}_t)$ and $(\boldsymbol{g}_t)$ in Algorithm 3 satisfy:*

$$\forall \boldsymbol{w} \in \operatorname{int} \mathcal{K}, \quad \sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}\rangle \le \frac{\eta}{2}\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}\rangle^2 + 5\kappa G R - \nu \log\left(1 - \frac{\|\boldsymbol{w}\|^2}{R^2}\right) + \frac{4d\log(1+\frac{T}{d})}{\eta}.$$

To get our main $\widetilde{O}(\sqrt{dT})$ regret bound for Algorithm 3, we instantiate Proposition 4.1 with: $c = \frac{1}{2}$, $\eta = \frac{1}{GR}\cdot \min\left(\frac{1}{10\kappa}, \sqrt{2T^{-1} d \log(1+\frac{T}{d})}\right)$, and $\nu = GR \cdot \max\left(20\kappa d, \sqrt{dT \log(1+\frac{T}{d})^{-1}}\right)$.

Theorem 4.1 (Main guarantee)**.** *Let $T \ge 1$ be given, and suppose that Assumption 2.1 and Assumption 2.2 hold. Consider a call to Algorithm 3 with input T and where the base algorithm $\mathcal{B}$ is an instance of* `Barrier-ONS` *(Algorithm 2) with input parameters (η, ν, c) as specified above. Then,*

$$\forall \boldsymbol{w} \in \mathcal{K}, \quad \sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}\rangle \le 5GR\sqrt{2dT\log(1+\tfrac{T}{d})} + 66GR\kappa d \log(1+\tfrac{T}{d}),$$

with $(\boldsymbol{w}_t)$ and $(\boldsymbol{g}_t)$ as in Algorithm 3. The computation cost of the current instance of Algorithm 3 is bounded by $\widetilde{O}\left(C_{\text{sep}}(\mathcal{K}) \cdot T + d^2 \cdot T + d^{\omega}\sqrt{T}\right)$, where $C_{\text{sep}}(\mathcal{K})$ is the cost of one call to $\mathrm{Sep}_{\mathcal{K}}$.

The proof of Theorem 4.1 is deferred to Appendix E.3. Here, we sketch why the computational cost of the instance of Algorithm 3 in Theorem 4.1 is bounded by $\widetilde{O}\left(C_{\text{sep}}(\mathcal{K}) \cdot T + d^2 \cdot T + d^{\omega}\sqrt{T}\right)$.

Proof sketch of cost bound. By Theorem 3.1, the computational cost of the `Barrier-ONS` base algorithm within Algorithm 3 is bounded by $\widetilde{O}\left(d^2 T + d^{\omega}\sqrt{\frac{dT}{\nu\eta}}\right)$. Given the choice of η and ν in Proposition 4.1, we have $\eta\nu \ge d$, implying that the computational cost of the `Barrier-ONS` base algorithm is bounded by $\widetilde{O}\left(d^2 T + d^{\omega}\sqrt{T}\right)$. In addition to the cost of the `Barrier-ONS` base algorithm, Algorithm 3 incurs $O(d + C_{\texttt{GaugeDist}}(\mathcal{K}))$ per round, where $C_{\texttt{GaugeDist}}(\mathcal{K})$ represents the cost of calling the `GaugeDist` base algorithm to approximate the gauge distance $S_{\mathcal{K}}$ and its subgradient. By Lemma 2.2, we have $C_{\texttt{GaugeDist}}(\mathcal{K}) \le \widetilde{O}(1) \cdot C_{\text{sep}}(\mathcal{K})$. This implies the desired cost bound. □

Computational cost. If we take the matrix exponent ω to be $\omega \le 5/2$, the average per-round computational cost of our approach is bounded by $\widetilde{O}(C_{\text{sep}}(\mathcal{K}) + d^2)$ as long as $T \ge d$. In contrast, existing separation-based projection-free algorithms, such as those in (Mhammedi, 2022; Garber and Kretzu, 2022; Lu et al., 2023), have a per-round computational cost bounded by $\widetilde{O}(C_{\text{sep}}(\mathcal{K}) + d)$

but guarantee a regret bound of $O(\kappa\sqrt{T})$. However, as discussed in the introduction, κ can be arbitrarily large for ill-conditioned sets. As shown in (Flaxman et al., 2005), one can apply an affine transformation to $\mathcal{K}$ (e.g., to put it in isotropic position) to ensure that κ is at most d. Doing so, however, incurs an additional $O(d^2)$ computational cost per round, as it requires multiplying the incoming subgradients by the matrix corresponding to the affine transformation at each round. This adjustment brings the overall computational cost in line with the cost of Algorithm 3 in Proposition 4.1.

Thus, one can think of the $O(d^2)$ cost of our approach as the "price" for adapting to the ill-conditioning of the set without needing to compute an affine transformation, while still ensuring a $\widetilde{O}(\sqrt{dT})$ regret bound—improving on the worst-case regret bounds of previous separation-based projection-free algorithms by a factor of $\sqrt{d}$ (because $\kappa\sqrt{T}$ can be as large as $d\sqrt{T}$ even after pre-processing). Finally, we note that the $\widetilde{O}(d^2)$ in our final algorithm cost comes from matrix-vector multiplications, which are easily parallelizable.

5. Application to Offline and Stochastic Optimization

In this section, we leverage the results from the previous sections to achieve a state-of-the-art convergence rate for projection-free stochastic convex optimization. We begin by presenting our results for stochastic convex optimization, then specialize them to the offline convex optimization setting for finding a near-optimal point. To proceed, we now state our assumption for the stochastic optimization setting.

Assumption 5.1. *There is a function $f : \mathcal{K} \to \mathbb{R}$ and parameters $\sigma \geq 0$ and $G > 0$ such that the loss vector $\boldsymbol{g}_t$ that the algorithm receives at round $t \geq 1$ is of the form $\boldsymbol{g}_t = \bar{\boldsymbol{g}}_t + \boldsymbol{\xi}_t$, where*

- *For all $t \geq 1$, $\bar{\boldsymbol{g}}_t \in \partial f(\boldsymbol{w}_t)$, where $\boldsymbol{w}_t$ is the output of the algorithm at round t;*
- *$(\boldsymbol{\xi}_t) \subset \mathbb{R}^d$ are i.i.d. noise vectors such that $\mathbb{E}[\boldsymbol{\xi}_t] = \mathbf{0}$ and $\mathbb{E}[\|\boldsymbol{\xi}_t\|^2] \leq \sigma^2$, for all $t \geq 1$; and*
- *For all $t \geq 1$, $\|\bar{\boldsymbol{g}}_t\| \leq G$.*

5.1. Convergence Rate in Stochastic Convex Optimization

Under Assumption 5.1, we now state our main guarantee for the stochastic convex optimization setting. As in the OCO setting, we use Algorithm 3 with the base algorithm $\mathcal{B}$ set as `Barrier-ONS`. Here, we set the parameters of `Barrier-ONS` as

$$c = \tfrac{1}{2}, \quad \eta = \tfrac{1}{R} \cdot \min\left(\tfrac{1}{10G\kappa}, \tfrac{1}{\sigma}\sqrt{\tfrac{2d\log(1+\frac{T}{d})}{T}}\right), \quad \text{and} \quad \nu = R \cdot \max\left(20G\kappa d, \sigma\sqrt{\tfrac{dT}{\log(1+\frac{T}{d})}}\right), \tag{14}$$

where $\kappa \coloneqq R/r$ and $r, R > 0$ are as in Assumption 2.1.

Theorem 5.1. *Let $T > 0$ be given. Suppose that Assumption 2.1 and Assumption 5.1 (with $\sigma, G \geq 0$) hold and consider a call to Algorithm 3 with input T, where the base algorithm $\mathcal{B}$ is an instance of* `Barrier-ONS` *(Algorithm 2) with input parameters (η, ν, c) as in (14). Then, we have*

$$\mathbb{E}[f(\widehat{\boldsymbol{w}}_T)] - \inf_{\boldsymbol{w}\in\mathcal{K}} f(\boldsymbol{w}) \leq 16R\sigma\sqrt{\frac{d\log(1+\frac{T}{d})}{T}} + \frac{74GR\kappa d\log(1+\frac{T}{d})}{T}, \tag{15}$$

where $\widehat{\boldsymbol{w}}_T \coloneqq \frac{1}{T}\sum_{t=1}^T \boldsymbol{w}_t$ and $(\boldsymbol{w}_t)$ are the iterates of Algorithm 3. The computational cost is bounded by

$$\widetilde{O}\left(C_{\mathrm{sep}}(\mathcal{K})\cdot T + d^2\cdot T + d^\omega\sqrt{T}\right), \tag{16}$$

The proof of the theorem, which follows from an application of Proposition 4.1, is in Appendix F. Extending the result in Theorem 5.1 to a high-probability guarantee is possible using martingale concentration bounds.

5.2. Computational Cost of Finding a Near-Optimal Point

We now consider the computational cost for finding a near-optimal point in offline convex optimization; that is, when $\sigma = 0$. In this case, from (15), if we set

$$T = \mathfrak{c}\cdot\frac{GR\kappa d}{\varepsilon},$$

with $\mathfrak{c} = \mathrm{polylog}(d, 1/\varepsilon)$ sufficiently large, we get that

$$f(\widehat{\boldsymbol{w}}_T) - \inf_{\boldsymbol{w}\in\mathcal{K}} f(\boldsymbol{w}) \le \varepsilon,$$

where $\widehat{\boldsymbol{w}}_T \coloneqq \frac{1}{T}\sum_{t=1}^T \boldsymbol{w}_t$ and $(\boldsymbol{w}_t)$ are the iterates of Algorithm 3. Thus, $\widehat{\boldsymbol{w}}_T$ represents an ε-optimal point for the objective function f. Now, by Theorem 5.1, the computational cost of the instance of Algorithm 3 in Theorem 5.1 is bounded by (16). Instantiating (16) with the choice of $T = \mathfrak{c}\cdot\frac{GR\kappa d}{\varepsilon}$, which implies that $d^2 T \gg \Omega(d^\omega\sqrt{T})$ (as long as $\omega \le 5/2$), the cost of finding an ε-optimal point in offline convex optimization using our approach is bounded by

$$\frac{\kappa d}{\varepsilon}\cdot(C_{\mathrm{sep}}(\mathcal{K}) + d^2).$$

6. Limitations and Future Work

While we have relegated the asphericity parameter κ to a lower-order term (in terms of T) in the regret bound, this term can still be arbitrarily large for ill-conditioned sets, potentially rendering the bound vacuous. Although any convex set $\mathcal{K}$ can, in principle, be transformed into isotropic position to ensure $\kappa = O(d)$ (Flaxman et al., 2005), this reparametrization does not fully resolve the issue: the ℓ_2-norm of the gradients may become inflated in the transformed space. As a result, the regret bound can still exhibit a polynomial dependence on κ—a limitation also present in the work of (Mhammedi, 2022). Additionally, computing an approximate isotropic transformation may require up to $\widetilde{O}(d^4)$ calls to a separation oracle (Lovász and Vempala, 2006), which is often computationally prohibitive in practice. Removing the dependence on κ in the regret bounds of separation-oracle-based algorithms—or at least reducing it to logarithmic factors—remains an open and important challenge.

We also note that, in addition to the $\widetilde{O}(1)$ separation oracle calls per round, our algorithm incurs $O(d^2)$ computational and memory overhead. This is a factor of $O(d)$ higher than the costs incurred by the algorithm in (Mhammedi, 2022). Determining whether this $O(d^2)$ overhead is inherent to achieving our improved regret guarantees is an interesting direction for future work.

References

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.

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

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

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

Ashok Cutkosky and Francesco Orabona. Black-Box Reductions for Parameter-free Online Learning in Banach Spaces. *Conference on Learning Theory*, 2018.

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.

Hidde Fokkema, Dirk van der Hoeven, Tor Lattimore, and Jack J Mayo. Online newton method for bandit convex optimisation. *arXiv preprint arXiv:2406.06506*, 2024.

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

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. New projection-free algorithms for online convex optimization with adaptive regret guarantees. In *Conference on Learning Theory*, pages 2326–2359. PMLR, 2022.

Benjamin Grimmer. Radial duality part i: foundations. *Mathematical Programming*, 205(1):33–68, 2024.

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 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, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

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.

Ruichen Jiang and Aryan Mokhtari. Accelerated quasi-newton proximal extragradient: Faster rate for smooth convex optimization. *Advances in Neural Information Processing Systems*, 36, 2024.

Ruichen Jiang, Qiujiang Jin, and Aryan Mokhtari. Online learning guided curvature approximation: A quasi-newton method with global non-asymptotic superlinear convergence. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1962–1992. PMLR, 2023.

Tomer Koren. Open problem: Fast stochastic exp-concave optimization. In *Conference on Learning Theory*, pages 1073–1075. PMLR, 2013.

Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of frank-wolfe optimization variants. *Advances in neural information processing systems*, 28, 2015.

Yin Tat Lee and Aaron Sidford. Solving linear programs with sqrt(rank) linear system solves. *arXiv preprint arXiv:1910.08033*, 2019.

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.

László Lovász and Santosh Vempala. Simulated annealing in convex bodies and an o*(n4) volume algorithm. *Journal of Computer and System Sciences*, 72(2):392–417, 2006. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2005.08.004. URL https://www.sciencedirect.com/science/article/pii/S0022000005000966. JCSS FOCS 2003 Special Issue.

Zhou Lu, Nataly Brukhim, Paula Gradu, and Elad Hazan. Projection-free adaptive regret with membership oracles. In *International Conference on Algorithmic Learning Theory*, pages 1055–1073. PMLR, 2023.

Zakaria Mhammedi. Efficient projection-free online convex optimization with membership oracle. In *Conference on Learning Theory*, pages 5314–5390. PMLR, 2022.

Zakaria Mhammedi and Khashayar Gatmiry. Quasi-newton steps for efficient online exp-concave optimization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4473–4503. PMLR, 2023.

Zakaria Mhammedi and Alexander Rakhlin. Damped online newton step for portfolio selection. In *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 5561–5595. PMLR, 2022.

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

Arkadi S Nemirovski and Michael J Todd. Interior-point methods for optimization. *Acta Numerica*, 17:191–234, 2008.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

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

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, pages 928–936, 2003.

Contents of Appendix

Appendix A. Organization of the Appendix

This appendix is organized as follows:

- In Appendix B, we present the full version of the `Barrier-ONS` algorithm.
- Appendix C provides background on self-concordant functions, highlighting key properties used in our analysis.
- In Appendix D, we present the proof for the regret guarantee of `Barrier-ONS` in Theorem 3.1.
- In Appendix D, we present the proof of our main OCO result in Theorem 4.1.
- In Appendix F, we provide the proof of Theorem 5.1 for the convergence rate of our algorithm in the stochastic convex optimization setting.
- Finally, Appendix G includes the proof of Lemma 2.1 on the approximate computation of the Gauge distance and its subgradients.

Appendix B. Full Version of `Barrier-ONS`

Algorithm 4 Efficient implementation of `Barrier-ONS` (Algorithm 2). The current algorithm is technically equivalent to Algorithm 2 in that both algorithms produce the same outputs.

input Parameters $\eta, \nu, c > 0$.
Set $\boldsymbol{z}_1 \leftarrow \boldsymbol{0}$, $\boldsymbol{u}_1 \leftarrow \boldsymbol{0}$, $V_0 \leftarrow 0$, $\Sigma'_0 \leftarrow \frac{1}{2\nu} I$, $S_0 \leftarrow \boldsymbol{0}$, $G_0 \leftarrow \boldsymbol{0}$, and $m = \mathfrak{c} \log_c(dT)$, with $\mathfrak{c}$ a sufficiently large universal constant.
for $t = 1, \dots, T$ **do**
  Play $\boldsymbol{u}_t$ and observe $\boldsymbol{g}_t \in \partial \ell_t(\boldsymbol{u}_t)$.
  Set $G_t \leftarrow G_{t-1} + \boldsymbol{g}_t$, $S_t \leftarrow S_{t-1} + \boldsymbol{g}_t \boldsymbol{g}_t^\top \boldsymbol{u}_t$, and $V_t \leftarrow V_{t-1} + \boldsymbol{g}_t \boldsymbol{g}_t^\top$.
  Set $\nabla_t \leftarrow \frac{2\nu \boldsymbol{u}_t}{1 - \|\boldsymbol{u}_t\|^2} + \eta V_t \boldsymbol{u}_t - \eta S_t + G_t$. // $\nabla_t = \nabla \Phi_{t+1}(\boldsymbol{u}_t)$
  Set $\Sigma'_t \leftarrow \Sigma'_{t-1} - \frac{\eta \Sigma'_{t-1} \boldsymbol{g}_t \boldsymbol{g}_t^\top \Sigma'_{t-1}}{1 + \eta \boldsymbol{g}_t^\top \Sigma'_{t-1} \boldsymbol{g}_t}$ and $\Sigma_t \leftarrow \Sigma'_t - \frac{4\nu \Sigma'_t \boldsymbol{u}_t \boldsymbol{u}_t^\top \Sigma'_t}{(R^2 - \|\boldsymbol{u}_t\|^2)^2 + 4\nu \boldsymbol{u}_t^\top \Sigma'_t \boldsymbol{u}_t}$.
  /* Computing the Taylor expansion */
  Set $\boldsymbol{\delta}_t \leftarrow \Sigma_t \nabla_t$ and $\tilde{\boldsymbol{\delta}}_t \leftarrow \Sigma_t \boldsymbol{\delta}_t$.
  for $k = 1, \dots, m$ **do**
    Set $\tilde{\boldsymbol{\delta}}_t \leftarrow \left(\frac{2\nu}{R^2 - \|\boldsymbol{z}_t\|^2} - \frac{2\nu}{R^2 - \|\boldsymbol{u}_t\|^2} \right) \Sigma_t \tilde{\boldsymbol{\delta}}_t$.
    Set $\boldsymbol{\delta}_t \leftarrow \boldsymbol{\delta}_t + \tilde{\boldsymbol{\delta}}_t$.
  /* Perform approximate Newton step */
  Set $\boldsymbol{u}_{t+1} \leftarrow \boldsymbol{u}_t - \boldsymbol{\delta}_t$. // $\boldsymbol{u}_{t+1} \approx \boldsymbol{u}_t - \nabla^{-2} \Phi_{t+1}(\boldsymbol{u}_t) \nabla \Phi_{t+1}(\boldsymbol{u}_t)$.
  /* Check if the Taylor expansion point needs to be updated */
  if $\left| \|\boldsymbol{u}_{t+1}\|^2 - \|\boldsymbol{z}_t\|^2 \right| \le c \cdot (R^2 - \|\boldsymbol{z}_t\|^2)$ **then**
    Set $\boldsymbol{z}_{t+1} \leftarrow \boldsymbol{z}_t$.
  else
    Set $\boldsymbol{z}_{t+1} \leftarrow \boldsymbol{u}_{t+1}$.
    Set $\Sigma'_t = \left(\frac{2\nu I}{R^2 - \|\boldsymbol{u}_{t+1}\|^2} + \eta V_t \right)^{-1}$. // $\Sigma'_t \leftarrow \left(\nabla^2 \Phi_{t+1}(\boldsymbol{u}_{t+1}) - \frac{4\nu \boldsymbol{u}_{t+1} \boldsymbol{u}_{t+1}^\top}{(R^2 - \|\boldsymbol{u}_{t+1}\|^2)^2} \right)^{-1}$

Appendix C. Self-Concordant Functions

In this section, we introduce the concept of self-concordant functions and outline several key properties that play an important role in our proofs (the results presented in this section are taken from (Mhammedi and Gatmiry, 2023)). We begin by defining a self-concordant function. For the rest of this section, let $\mathcal{C}$ represent a convex, compact set with a non-empty interior, denoted by $\operatorname{int}\mathcal{C}$. For a function that is twice differentiable [or thrice differentiable], we denote its Hessian as $\nabla^2 f(\boldsymbol{u})$ [and its third derivative tensor as $\nabla^3 f(\boldsymbol{u})$] at $\boldsymbol{u}$.

Definition C.1. *A convex function $f : \operatorname{int}\mathcal{C} \to \mathbb{R}$ is called* self-concordant *with constant $M_f \geq 0$, if f is C^3 and satisfies*

- $f(\boldsymbol{x}_k) \to +\infty$ *for* $\boldsymbol{x}_k \to \boldsymbol{x} \in \partial\mathcal{C}$*; and*
- *For all* $\boldsymbol{x} \in \operatorname{int}\mathcal{C}$ *and* $\boldsymbol{u} \in \mathbb{R}^d$*:*

$$|\nabla^3 f(\boldsymbol{x})[\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}]| \leq 2M_f \|\boldsymbol{u}\|^3_{\nabla^2 f(\boldsymbol{x})}.$$

By definition, if f is self-concordant with a constant $M_f \geq 0$, it remains self-concordant for any constant $M \geq M_f$. For a self-concordant function f and a point $\boldsymbol{x} \in \operatorname{dom} f$, the quantity $\lambda(\boldsymbol{x}, f) \coloneqq \|\nabla f(\boldsymbol{x})\|_{\nabla^{-2} f(\boldsymbol{x})}$, referred to as the *Newton decrement*, plays a key role in our proofs. The following two lemmas summarize properties of the Newton decrement and the Hessians of self-concordant functions, which will be utilized frequently in the proofs of `Barrier-ONS` (see, for example, Nemirovski and Todd (2008); Nesterov et al. (2018)).

Lemma C.1. *Let $f : \operatorname{int}\mathcal{C} \to \mathbb{R}$ be a self-concordant function with constant $M_f \geq 1$. Further, let $\boldsymbol{x} \in \operatorname{int}\mathcal{C}$ and $\boldsymbol{x}_f \in \arg\min_{\boldsymbol{x}\in\mathcal{C}} f(\boldsymbol{x})$. Then,*

- *Whenever $\lambda(\boldsymbol{x}, f) < 1/M_f$, we have*

$$\|\boldsymbol{x} - \boldsymbol{x}_f\|_{\nabla^2 f(\boldsymbol{x}_f)} \vee \|\boldsymbol{x} - \boldsymbol{x}_f\|_{\nabla^2 f(\boldsymbol{x})} \leq \lambda(\boldsymbol{x}, f)/(1 - M_f \lambda(\boldsymbol{x}, f));$$

- *For any $M \geq M_f$, the* Newton step *$\boldsymbol{x}_+ \coloneqq \boldsymbol{x} - \nabla^{-2} f(\boldsymbol{x}) \nabla f(\boldsymbol{x})$ satisfies $\boldsymbol{x}_+ \in \operatorname{int}\mathcal{C}$ and*

$$\lambda(\boldsymbol{x}_+, f) \leq M\lambda(\boldsymbol{x}, f)^2/(1 - M\lambda(\boldsymbol{x}, f))^2.$$

Lemma C.2. *Consider a self-concordant function $f : \operatorname{int}, \mathcal{C} \to \mathbb{R}$ with constant M_f and a point $\boldsymbol{x} \in \operatorname{int}, \mathcal{C}$. For any $\boldsymbol{y}$ such that $r \coloneqq \|\boldsymbol{y} - \boldsymbol{x}\|_{\nabla^2 f(\boldsymbol{x})} < 1/M_f$, the following holds:*

$$(1 - M_f r)^2 \nabla^2 f(\boldsymbol{y}) \preceq \nabla^2 f(\boldsymbol{x}) \preceq (1 - M_f r)^{-2} \nabla^2 f(\boldsymbol{y}).$$

The following result from (Nesterov et al., 2018, Theorem 5.1.5) will be helpful in showing that the iterates of our algorithms consistently remain within the feasible set.

Lemma C.3. *Let $f : \operatorname{int}\mathcal{C} \to \mathbb{R}$ be a self-concordant function with constant $M_f \geq 1$ and $\boldsymbol{x} \in \operatorname{int}\mathcal{C}$. Then, $\mathcal{E}_{\boldsymbol{x}} \coloneqq \{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w} - \boldsymbol{x}\|_{\nabla^2 f(\boldsymbol{x})} < 1/M_f\} \subseteq \operatorname{int}\mathcal{C}$. Furthermore, for all $\boldsymbol{w} \in \mathcal{E}_{\boldsymbol{x}}$, we have*

$$\|\boldsymbol{w} - \boldsymbol{x}\|_{\nabla^2 f(\boldsymbol{w})} \leq \frac{\|\boldsymbol{w} - \boldsymbol{x}\|_{\nabla^2 f(\boldsymbol{x})}}{1 - M_f \|\boldsymbol{w} - \boldsymbol{x}\|_{\nabla^2 f(\boldsymbol{x})}}.$$

Finally, we also need the following result due to Mhammedi and Rakhlin (2022).

Lemma C.4. *Let $f : \operatorname{int} \mathcal{C} \to \mathbb{R}$ be a self-concordant function with constant $M_f > 0$. Then, for any $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{int} \mathcal{C}$ such that $r := \|\boldsymbol{x} - \boldsymbol{y}\|_{\nabla^2 f(\boldsymbol{x})} < 1/M_f$, we have*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2_{\nabla^{-2} f(\boldsymbol{x})} \leq \frac{1}{(1 - M_f r)^2} \|\boldsymbol{y} - \boldsymbol{x}\|^2_{\nabla^2 f(\boldsymbol{x})}.$$

Appendix D. ONS Analysis: Proof of Theorem 3.1

This appendix provides a proof of Theorem 3.1. Each subsection provides results and proofs of intermediate results outlined in Section 3.

D.1. Taylor Expansion of Inverse Hessians

In this section, we prove that the Taylor expansions used in Algorithm 2 indeed approximate the inverse Hessians $(\nabla^{-2}\Phi_{t+1}(\boldsymbol{u}_t))$. This result we state next is a slight modification of a result in Mhammedi and Gatmiry (2023).

Lemma D.1. *Let $\eta, \nu > 0$, $c \in (0, 1)$, and $T \geq 1$ be given. Further, let (T, γ_t) and (Σ_t) be as in Algorithm 2 with parameters (η, ν, c). Then, for $t \in [T]$ such that $\boldsymbol{u}_t \in \operatorname{int} \mathbb{B}(R)$ and for any $m \geq 1$, we have*

$$\left\| \nabla^{-2}\Phi_{t+1}(\boldsymbol{u}_t) - \sum_{k=1}^{m+1} \gamma_t^{k-1} \Sigma_t^k \right\| \leq \frac{R^2 c^m}{2\nu \cdot (1 - c)}.$$

Proof. Fix $m \geq 1$ and let $\alpha_t := \frac{\|\boldsymbol{u}_t\|^2 - \|\boldsymbol{z}_t\|^2}{R^2 - \|\boldsymbol{u}_t\|^2}$. We have

$$\alpha_t = \frac{R^2 - \|\boldsymbol{z}_t\|^2}{R^2 - \|\boldsymbol{u}_t\|^2} - 1 = -\frac{R^2 - \|\boldsymbol{z}_t\|^2}{2\nu} \gamma_t,$$

where we recall that $\gamma_t = \frac{2\nu}{R^2 - \|\boldsymbol{z}_t\|^2} - \frac{2\nu}{R^2 - \|\boldsymbol{u}_t\|^2}$. Note that Σ_t in Algorithm 2 satisfies

$$\begin{aligned}
\Sigma_t^{-1} &= \frac{2\nu I}{R^2 - \|\boldsymbol{z}_t\|^2} + \frac{4\nu \boldsymbol{u}_t \boldsymbol{u}_t^\top}{(R^2 - \|\boldsymbol{u}_t\|^2)^2} + \eta \sum_{s=1}^{t-1} \widetilde{\boldsymbol{g}}_s \widetilde{\boldsymbol{g}}_s^\top, \qquad (17) \\
&= \nabla^2 \Phi_{t+1}(\boldsymbol{u}_t) - \frac{2\nu I}{R^2 - \|\boldsymbol{u}_t\|^2} + \frac{2\nu I}{R^2 - \|\boldsymbol{z}_t\|^2}, \\
&= \nabla^2 \Phi_{t+1}(\boldsymbol{u}_t) - \frac{2\nu I}{R^2 - \|\boldsymbol{z}_t\|^2} \cdot \left(\frac{R^2 - \|\boldsymbol{z}_t\|^2}{R^2 - \|\boldsymbol{u}_t\|^2} - 1 \right), \\
&= \nabla^2 \Phi_{t+1}(\boldsymbol{u}_t) - \frac{2\nu \alpha_t I}{R^2 - \|\boldsymbol{z}_t\|^2}.
\end{aligned}$$

Therefore, if we let $U_t \coloneqq (R^2 - \|\boldsymbol{z}_t\|^2)H_t^{-1}/(2\nu)$, we have

$$\begin{aligned}
\nabla^{-2}\Phi_{t+1}(\boldsymbol{u}_t) &= \left(\frac{2\nu\alpha_t I}{R^2 - \|\boldsymbol{z}_t\|^2} + H_t^{-1}\right)^{-1}, \\
&= \frac{R^2 - \|\boldsymbol{z}_t\|^2}{2\nu}\left(\alpha_t I + \frac{R^2 - \|\boldsymbol{z}_t\|^2}{2\nu} H_t^{-1}\right)^{-1}, \\
&= \frac{R^2 - \|\boldsymbol{z}_t\|^2}{2\nu}(\alpha_t I + U_t)^{-1}, \\
&= \frac{R^2 - \|\boldsymbol{z}_t\|^2}{2\nu} U_t^{-1}(I + \alpha_t U_t^{-1})^{-1}.
\end{aligned} \tag{18}$$

Now, by (17), we have $U_t \succeq I$ and so $\|U_t^{-1}\| \leq 1$. Using this and that $|\alpha_t| \leq c < 1$ (this is an invariant of Algorithm 2—see Line 9 of Algorithm 2), we have

$$(1 + \alpha_t U_t^{-1})^{-1} = \sum_{k=0}^{\infty} (-\alpha_t)^k U_t^{-k}, \quad \text{and} \quad \left\|(1 + \alpha_t U_t)^{-1} - \sum_{k=0}^{m} (-\alpha_t)^k U_t^{-k}\right\| \leq \frac{c^m}{1-c}.$$

Therefore, by (18) and the fact that $\|U_t^{-1}\| \leq 1$ we have

$$\left\|\nabla^{-2}\Phi_{t+1}(\boldsymbol{u}_t) - \frac{1 - \|\boldsymbol{z}_t\|^2}{2d\eta}\sum_{k=1}^{m+1}(-\alpha_t)^{k-1}U_t^{-k}\right\| \leq \frac{(R^2 - \|\boldsymbol{z}_t\|^2)\cdot c^m}{2\nu\cdot(1-c)}.$$

Now, the fact that $\frac{R^2 - \|\boldsymbol{z}_t\|^2}{2d\eta}\sum_{k=1}^{m+1}(-\alpha_t)^{k-1}U_t^{-k} = \sum_{k=1}^{m+1}\gamma_t^{k-1}\Sigma_t^k$ completes the proof. □

The key to the computational efficiency of our approach lies in the fact that we only need to compute the full inverse of a matrix in a small fraction of the rounds, as demonstrated by the following lemma. The proof of this lemma leverages the stability of the Newton iterates, which is ensured by the non-linear terms in (Φ_t).

D.2. Number of Taylor Expansion Points (Proof of Lemma 3.1)

For the proof of Lemma 3.1, we need the following elementary result.

Lemma D.2. *Let $\nu > 0$ and $R > 0$ be given and define $\Psi(\boldsymbol{x}) \coloneqq -\nu\log(R^2 - \|\boldsymbol{x}\|^2)$. For any $\boldsymbol{u}, \boldsymbol{w} \in \mathbb{B}(R)$, we have*

$$\frac{1}{\nu}\|\boldsymbol{w} - \boldsymbol{u}\|^2_{\nabla^2\Psi(\boldsymbol{w})} \geq \frac{(\|\boldsymbol{w}\|^2 - \|\boldsymbol{u}\|^2)^2}{(R^2 - \|\boldsymbol{w}\|^2)^2}.$$

Proof. Fix $\boldsymbol{u}, \boldsymbol{w} \in \mathbb{B}(R)$. We have

$$\begin{aligned}
\frac{1}{2\nu}\|\boldsymbol{w} - \boldsymbol{u}\|^2_{\nabla^2\Psi(\boldsymbol{w})} &= (\boldsymbol{w} - \boldsymbol{u})^\top\left(\frac{I}{R^2 - \|\boldsymbol{w}\|^2} + \frac{2\boldsymbol{w}\boldsymbol{w}^\top}{(R^2 - \|\boldsymbol{w}\|^2)^2}\right)(\boldsymbol{w} - \boldsymbol{u}), \\
&= \frac{\|\boldsymbol{w}\|^2 + \|\boldsymbol{u}\|^2 - 2\boldsymbol{w}^\top\boldsymbol{u} - 2\|\boldsymbol{w}\|^2\boldsymbol{w}^\top\boldsymbol{u} + \|\boldsymbol{w}\|^4 + 2(\boldsymbol{w}^\top\boldsymbol{u})^2 - \|\boldsymbol{w}\|^2\|\boldsymbol{u}\|^2}{(R^2 - \|\boldsymbol{w}\|^2)^2}, \\
&= \frac{2(\|\boldsymbol{w}\|^2 - \boldsymbol{w}^\top\boldsymbol{u})(R^2 - \boldsymbol{w}^\top\boldsymbol{u})}{(R^2 - \|\boldsymbol{w}\|^2)^2} + \frac{\|\boldsymbol{u}\|^2 - \|\boldsymbol{w}\|^2}{R^2 - \|\boldsymbol{w}\|^2}, \\
&= \frac{2(\|\boldsymbol{w}\|^2 - \boldsymbol{w}^\top\boldsymbol{u})(R^2 - \|\boldsymbol{w}\|^2)}{(R^2 - \|\boldsymbol{w}\|^2)^2} + \frac{2(\|\boldsymbol{w}\|^2 - \boldsymbol{w}^\top\boldsymbol{u})^2}{(R^2 - \|\boldsymbol{w}\|^2)^2} + \frac{\|\boldsymbol{u}\|^2 - \|\boldsymbol{w}\|^2}{R^2 - \|\boldsymbol{w}\|^2}.
\end{aligned}$$

Now using that $-\boldsymbol{w}\boldsymbol{u} = 2^{-1}(\|\boldsymbol{w}-\boldsymbol{u}\|^2 - \|\boldsymbol{w}\|^2 - \|\boldsymbol{u}\|^2)$, we get that

$$\begin{aligned}\frac{1}{2\nu}\|\boldsymbol{w}-\boldsymbol{u}\|^2_{\nabla^2\Psi(\boldsymbol{w})} &= \frac{\|\boldsymbol{w}-\boldsymbol{u}\|^2 + \|\boldsymbol{w}\|^2 - \|\boldsymbol{u}\|^2}{R^2 - \|\boldsymbol{w}\|^2} + \frac{2(\|\boldsymbol{w}\|^2 - \boldsymbol{w}^\top\boldsymbol{u})^2}{(R^2-\|\boldsymbol{w}\|^2)^2} + \frac{\|\boldsymbol{u}\|^2 - \|\boldsymbol{w}\|^2}{R^2-\|\boldsymbol{w}\|^2},\\ &= \frac{\|\boldsymbol{w}-\boldsymbol{u}\|^2}{R^2-\|\boldsymbol{w}\|^2} + \frac{(\|\boldsymbol{w}-\boldsymbol{u}\|^2 + \|\boldsymbol{w}\|^2 - \|\boldsymbol{u}\|^2)^2}{2(R^2-\|\boldsymbol{w}\|^2)^2}. \end{aligned} \tag{19}$$

Now, consider the function

$$f : X \to \frac{X}{R^2-\|\boldsymbol{w}\|^2} + \frac{(X + \|\boldsymbol{w}\|^2 - \|\boldsymbol{u}\|^2)^2}{2(R^2-\|\boldsymbol{w}\|^2)^2}.$$

Note that $\operatorname{sgn}(f'(X)) = \operatorname{sgn}(X - \|\boldsymbol{u}\|^2 + 1)$. Thus, since $\|\boldsymbol{u}\|^2 \le 1$, the function f is non-decreasing over $\mathbb{R}_{\ge 0}$, and so $f(\|\boldsymbol{w}-\boldsymbol{u}\|^2) \ge f(0)$. Using this with (19), we get

$$\frac{1}{\nu}\|\boldsymbol{w}-\boldsymbol{u}\|^2_{\nabla^2\Psi(\boldsymbol{w})} \ge \frac{(\|\boldsymbol{w}\|^2 - \|\boldsymbol{u}\|^2)^2}{(R^2-\|\boldsymbol{w}\|^2)^2}.$$

□

Proof of Lemma 3.1. Let $i_1, \dots, i_n$ be the rounds t where $\boldsymbol{z}_t \ne \boldsymbol{z}_{t-1}$, and note that by Line 9 of Algorithm 2, we have

$$\left|\|\boldsymbol{z}_{i_{k+1}}\|^2 - \|\boldsymbol{z}_{i_k}\|^2\right| > c \cdot (R^2 - \|\boldsymbol{z}_{i_k}\|^2), \quad \forall k \in [n-1]. \tag{20}$$

Further, let

$$\alpha_t \coloneqq \frac{\|\boldsymbol{u}_{t+1}\|^2 - \|\boldsymbol{u}_t\|^2}{R^2 - \|\boldsymbol{u}_{t+1}\|^2}, \quad \text{and} \quad \mu_t \coloneqq \frac{\|\boldsymbol{u}_t\|^2 - \|\boldsymbol{u}_{t+1}\|^2}{R^2 - \|\boldsymbol{u}_t\|^2}.$$

Fix $k \in [n-1]$. Suppose that $(\sum_{t=i_k}^{i_{k+1}-1}\alpha_t) \vee (\sum_{t=i_k}^{i_{k+1}-1}\mu_t) \le 1/2$ and let $m_k \coloneqq i_{k+1} - i_k$. In this case, by (20) we have that

$$\begin{aligned}\log(1+c) &\le \left(\log\frac{R^2 - \|\boldsymbol{z}_{i_k}\|^2}{R^2 - \|\boldsymbol{z}_{i_{k+1}}\|^2}\right) \vee \left(\log\frac{R^2 - \|\boldsymbol{z}_{i_{k+1}}\|^2}{R^2 - \|\boldsymbol{z}_{i_k}\|^2}\right),\\ &\le \left(\log\prod_{t=i_k}^{i_{k+1}-1}(1+\alpha_t)\right) \vee \left(\log\prod_{t=i_k}^{i_{k+1}-1}(1+\mu_t)\right),\\ &= \left(\sum_{t=i_k}^{i_{k+1}-1}\log(1+\alpha_t)\right) \vee \left(\sum_{t=i_k}^{i_{k+1}-1}\log(1+\mu_t)\right),\\ &\le \log\left(1 + \frac{1}{m_k}\sum_{t=i_k}^{i_{k+1}-1}\alpha_t\right)^{m_k} \vee \log\left(1 + \frac{1}{m_k}\sum_{t=i_k}^{i_{k+1}-1}\mu_t\right)^{m_k}, \quad \text{(Jensen)}\\ &\le \log\left(1 + 2\sum_{t=i_k}^{i_{k+1}-1}\alpha_t\right) \vee \log\left(1 + 2\sum_{t=i_k}^{i_{k+1}-1}\mu_t\right),\end{aligned}$$

where the last inequality follows by the facts that $(\sum_{t=i_k}^{i_{k+1}-1}\alpha_t)\vee(\sum_{t=i_k}^{i_{k+1}-1}\mu_t)\le 1/2$ and $(1+x)^r\le 1+\frac{rx}{1-(r-1)x}$, for all $x\in(-1,\frac{1}{r-1}]$ and $r\ge 1$. Now, using that $\log(1+x)\le x$ for $x\ge 0$ and $\log(1+x)\ge x/2$, for $x\in(0,1)$, we get that

$$\frac{c}{2}\le\log(1+c)\le\left(2\sum_{t=i_k}^{i_{k+1}-1}\alpha_t\right)\vee\left(2\sum_{t=i_k}^{i_{k+1}-1}\mu_t\right), \tag{21}$$

$$\le 2\left(\sqrt{m_k\sum_{t=i_k}^{i_{k+1}-1}\alpha_t^2}\right)\vee\left(\sqrt{m_k\sum_{t=i_k}^{i_{k+1}-1}\mu_t^2}\right),\quad\text{(Jensen)}$$

$$\le 2\sqrt{m_k\sum_{t=i_k}^{i_{k+1}-1}\alpha_t^2+m_k\sum_{t=i_k}^{i_{k+1}-1}\mu_t^2}. \tag{22}$$

So far, we have assumed that $(\sum_{t=i_k}^{i_{k+1}-1}\alpha_t)\vee(\sum_{t=i_k}^{i_{k+1}-1}\mu_t)\le 1/2$. If this does not hold, then we have $(\sum_{t=i_k}^{i_{k+1}-1}\alpha_t)\vee(\sum_{t=i_k}^{i_{k+1}-1}\mu_t)\ge 1/2$. This implies (21) from which (22) follows. Now, (22) implies

$$\sum_{t=i_k}^{i_{k+1}-1}\alpha_t^2+\sum_{t=i_k}^{i_{k+1}-1}\mu_t^2\ge\frac{c^2}{16m_k}.$$

Thus, by summing over $k=1,\dots,n-1$, and using Lemma D.2 and Lemma D.7 (in particular (31)), we get

$$\frac{1}{2}+\frac{13^2 d\log(1+T/d)}{\nu\eta}\ge\sum_{t=1}^{T}(\alpha_t^2+\mu_t^2)\ge\sum_{k=1}^{n}\frac{c^2}{16m_k}=\sum_{k=1}^{n}\frac{c^2}{16(i_{k+1}-i_k)}\ge\frac{c^2n^2}{16T},$$

where the last inequality follows by the fact that $x\mapsto 1/x$ is convex and Jensen's inequality. By taking the square-root on both sides and rearranging, we get that

$$n\le\frac{52}{c}\sqrt{T\cdot\left(1+\frac{d\log(1+T/d)}{\nu\eta}\right)}.$$

□

D.3. Structural Results for `Barrier-ONS` and `FTRL`

Most of the structural results we present in this section are small modifications of existing results in (Mhammedi and Gatmiry, 2023).

Lemma D.3. *For any $t\ge 1$, the functions Ψ and Φ_t in (3) are self-concordant with constant $1/\sqrt{\nu}$.*

Proof. Fix $t\ge 1$. First, we note that $\boldsymbol{x}\mapsto\Psi(\boldsymbol{x})/\nu=-\log(R^2-\|\boldsymbol{x}\|^2)$ is self-concordant with constant 1 (see e.g. (Nesterov et al., 2018, Exampled 5.1.1)). Thus, Ψ is a self-concordant function with constant $1/\sqrt{\nu}$; this follows by the fact that if a function f is self-concordant with constant M_f, then αf, for $\alpha>0$, it is self-concordant with constant $1/\sqrt{\alpha}$ (see e.g. (Nesterov et al., 2018, Corollary 5.1.3)). On the other hand, since $\Phi_t(\boldsymbol{x})$ is equal to $\Psi(\boldsymbol{x})$ plus a quadratic in $\boldsymbol{x}$, then Φ_t is self-concordant with the same constant as Ψ (see e.g. (Nesterov et al., 2018, Corollary 5.1.2)). □

Lemma D.4. *Let $\eta, \nu \in (0,1)$, $R > 0$, and $\widetilde{G} > 0$ be such that $\nu \geq 10\widetilde{G}R$ and $\eta \leq \frac{1}{5\widetilde{G}R}$. Further, let $(\widetilde{\boldsymbol{g}}_t) \subset \mathbb{R}^d$ be a sequence of vectors such that $\|\widetilde{\boldsymbol{g}}_t\| \leq \widetilde{G}$, for all $t \geq 1$. Then, for any sequence $(\boldsymbol{y}_t) \subset \operatorname{int} \mathbb{B}(R)$, the potential functions (Φ_t) in (3) satisfy*

$$\forall t \geq 1, \quad \|\widetilde{\boldsymbol{g}}_t\|^2_{\nabla^{-2}\Phi_t(\boldsymbol{y}_t)} \leq \frac{\widetilde{G}^2 R^2}{2\nu}, \tag{23}$$

and

$$\forall T \geq 1, \quad \sum_{t=1}^{T} \|\widetilde{\boldsymbol{g}}_t\|^2_{\nabla^{-2}\Phi_t(\boldsymbol{y}_t)} \leq \frac{d \log(1 + T/d)}{\eta}. \tag{24}$$

Proof. Fix the sequence $(\boldsymbol{y}_t)$. First, note that the Hessian of Ψ in (3) satisfies

$$\forall t \geq 1, \quad \nabla^2 \Psi(\boldsymbol{y}_t) = \frac{2\nu}{R^2 - \|\boldsymbol{y}_t\|^2} I + \frac{4\nu \boldsymbol{y}_t \boldsymbol{y}_t^\top}{(R^2 - \|\boldsymbol{y}_t\|^2)^2} I. \tag{25}$$

Therefore, we have

$$\begin{aligned}
\forall t \geq 1, \quad \|\widetilde{\boldsymbol{g}}_t\|^2_{\nabla^{-2}\Phi_t(\boldsymbol{y}_t)} &\leq \widetilde{\boldsymbol{g}}_t^\top \left(\nabla^2 \Psi(\boldsymbol{y}_t)\right)^{-1} \widetilde{\boldsymbol{g}}_t, \\
&\leq \widetilde{\boldsymbol{g}}_t^\top \left(\nabla^2 \Psi(\boldsymbol{y}_t)\right)^{-1} \widetilde{\boldsymbol{g}}_t, \\
&\leq \frac{R^2 \widetilde{G}^2}{2\nu},
\end{aligned}$$

where in the last step we used (25) and $(\widetilde{\boldsymbol{g}}_t) \subset \mathbb{B}(\widetilde{G})$. This shows (23).

We now show (24). Since $\eta \leq \frac{1}{5\widetilde{G}R}$, $\nu \geq 10\widetilde{G}R$, and $(\widetilde{\boldsymbol{g}}_t) \subset \mathbb{B}(\widetilde{G})$, we have

$$\forall t \geq 1, \quad \eta \widetilde{\boldsymbol{g}}_t \widetilde{\boldsymbol{g}}_t^\top \leq \frac{\widetilde{G}}{5R} I \leq \frac{\nu}{5R^2} I \leq \frac{\nu}{5(R^2 - \|\boldsymbol{y}_t\|^2)} I.$$

Combining this with (25) implies that

$$\forall t \geq 1, \quad \eta \widetilde{\boldsymbol{g}}_t \widetilde{\boldsymbol{g}}_t^\top \leq \frac{1}{10} \nabla^2 \Psi(\boldsymbol{y}_t). \tag{26}$$

Note that (25) also implies that

$$\forall t \geq 1, \quad \frac{\nu}{R^2} I \leq \frac{1}{2} \nabla^2 \Psi(\boldsymbol{y}_t). \tag{27}$$

Therefore, we have

$$\begin{aligned}
\forall t \geq 1, \quad \|\widetilde{\boldsymbol{g}}_t\|^2_{\nabla^{-2}\Phi_t(\boldsymbol{y}_t)} &= \widetilde{\boldsymbol{g}}_t^\top \left(\nabla^2 \Psi(\boldsymbol{y}_t) + \eta \sum_{s=1}^{t-1} \widetilde{\boldsymbol{g}}_s \widetilde{\boldsymbol{g}}_s^\top\right)^{-1} \widetilde{\boldsymbol{g}}_t, \\
&\leq \widetilde{\boldsymbol{g}}_t^\top \left(\frac{1}{2} \nabla^2 \Psi(\boldsymbol{y}_t) + \eta \sum_{s=1}^{t} \widetilde{\boldsymbol{g}}_s \widetilde{\boldsymbol{g}}_s^\top\right)^{-1} \widetilde{\boldsymbol{g}}_t, \quad \text{(by (26))} \\
&\leq \widetilde{\boldsymbol{g}}_t^\top \left(\frac{\nu}{R^2} I + \eta \sum_{s=1}^{t} \widetilde{\boldsymbol{g}}_s \widetilde{\boldsymbol{g}}_s^\top\right)^{-1} \widetilde{\boldsymbol{g}}_t, \quad \text{(by (27))} \\
&\leq \frac{1}{\eta} \widetilde{\boldsymbol{g}}_t^\top Q_t^{-1} \widetilde{\boldsymbol{g}}_t.
\end{aligned} \tag{28}$$

where $Q_t \coloneqq \frac{\nu}{R^2\eta} I + \sum_{s=1}^{t} \widetilde{\boldsymbol{g}}_s \widetilde{\boldsymbol{g}}_s^\top$. Thus, by (28) and (Hazan et al., 2007, Lemma 11), we have

$$
\begin{aligned}
\forall t \in [T], \quad \|\widetilde{\boldsymbol{g}}_t\|^2_{\nabla^{-2}\Phi_t(\boldsymbol{y}_t)} &\leq \frac{1}{\eta} \sum_{t=1}^{T} \widetilde{\boldsymbol{g}}_t^\top Q_t^{-1} \widetilde{\boldsymbol{g}}_t, \\
&\leq \frac{1}{\eta} \log \frac{\det Q_T}{\det Q_0}, \\
&= \frac{1}{\eta} \log\det\left(I + Q_0^{-1} \sum_{t=1}^{T} \widetilde{\boldsymbol{g}}_t \widetilde{\boldsymbol{g}}_t^\top \right), \\
&\leq \frac{d}{\eta} \log \frac{\operatorname{Tr}\left(Q_0^{-1} \sum_{t=1}^{T} \widetilde{\boldsymbol{g}}_t \widetilde{\boldsymbol{g}}_t^\top \right)}{d}, \quad \text{(Jensen's inequality)} \\
&\leq \frac{d \log(1 + \frac{\eta T R^2 \widetilde{G}^2}{\nu d})}{\eta}, \quad \text{(using the expression of } Q_0 \text{ and } (\widetilde{\boldsymbol{g}}_t) \subset \mathbb{B}(\widetilde{G})) \\
&\leq \frac{d \log(1 + \frac{T}{d})}{\eta},
\end{aligned}
$$

where in the last step uses that $\nu \geq \widetilde{G}R$ and $\eta \leq \frac{1}{\widetilde{G}R}$. This completes the proof. $\square$

Lemma D.5. *Let $\eta, R, \widetilde{G}, \nu > 0$, and $T \geq 1$ be given, and suppose that $\nu \leq 10 d\widetilde{G}RT$ and that $\boldsymbol{u}_t \in \mathbb{B}(R)$ and $\widetilde{\boldsymbol{g}}_t \in \mathbb{B}(\widetilde{G})$ for all $t \in [T]$. Then, for any $t \in [T]$, the* FTRL *iterate $\boldsymbol{w}_t$ in (5) satisfies:*

$$
\frac{\nu R}{R^2 - \|\boldsymbol{w}_t\|^2} \leq 20 d\widetilde{G} \cdot (1 + 2\eta \widetilde{G}R) T.
$$

Proof. Fix $t \in [T]$. Since $\Psi(\boldsymbol{x})$ is a self-concordant barrier, we have $\boldsymbol{w}_t \in \operatorname{int} \mathbb{B}(R)$. Thus, by the first-order optimality condition involving $\boldsymbol{w}_t$, we have

$$
\frac{2\nu \boldsymbol{w}_t}{R^2 - \|\boldsymbol{w}_t\|^2} + \eta \sum_{s=1}^{t-1} \widetilde{\boldsymbol{g}}_s \widetilde{\boldsymbol{g}}_s^\top (\boldsymbol{w}_t - \boldsymbol{u}_s) + \sum_{s=1}^{t-1} \widetilde{\boldsymbol{g}}_s = \boldsymbol{0}.
$$

Thus, using that $\boldsymbol{w}_s, \boldsymbol{u}_s \in \mathbb{B}(R)$ and $\widetilde{\boldsymbol{g}}_s \in \mathbb{B}(\widetilde{G})$ for all $s \in [t-1]$, we get

$$
\frac{2\nu \|\boldsymbol{w}_t\|}{R^2 - \|\boldsymbol{w}_t\|^2} \leq \widetilde{G} \cdot (1 + 2\eta \widetilde{G}R) \cdot T. \tag{29}
$$

If $\|\boldsymbol{w}_t\| \leq R/2$, then we are done since in this case $\frac{1}{R^2 - \|\boldsymbol{w}_t\|^2} \leq \frac{4}{3R^2} \leq \frac{2}{R^2}$, and so

$$
\frac{\nu R}{R^2 - \|\boldsymbol{w}_t\|^2} \leq \frac{2\nu}{R} \overset{(a)}{\leq} 20 d\widetilde{G} \cdot T \leq 20 d\widetilde{G} \cdot (1 + 2\eta \widetilde{G}R) \cdot T,
$$

where (a) follows from the assumption that $\nu \leq 10 d\widetilde{G}RT$. Now, suppose that $\|\boldsymbol{w}_t\| > R/2$. Plugging this into (29) directly implies that

$$
\frac{\nu R}{R^2 - \|\boldsymbol{w}_t\|^2} \leq \widetilde{G} \cdot (1 + 2\eta \widetilde{G}R) T,
$$

which completes the proof. $\square$

Lemma D.6. *Let $\eta, R, \widetilde{G}, B, \nu > 0$, and $T \geq 1$ be given, and suppose that $\nu \leq 10d\widetilde{G}RT$ and that $\boldsymbol{u}_t \in \mathbb{B}(R)$ and $\widetilde{\boldsymbol{g}}_t \in \mathbb{B}(\widetilde{G})$ for all $t \in [T]$. Further, for $t \in [T]$, let $\boldsymbol{w}_t$ be the* FTRL *iterate in (5); that is, $\boldsymbol{w}_t \in \arg\min_{\boldsymbol{w} \in \mathbb{B}(R)} \Phi_t(\boldsymbol{w})$. Then, for any $t \in [T]$ and $\boldsymbol{z} \in \operatorname{int} \mathbb{B}(R)$ such that $\|\boldsymbol{z} - \boldsymbol{w}_t\|^2_{\nabla^2 \Psi(\boldsymbol{z})} \leq \nu B^2$, where $\Psi(\boldsymbol{z}) \coloneqq -\nu \cdot \log(R^2 - \|\boldsymbol{z}\|^2)$, we have*

$$\|\nabla \Phi_t(\boldsymbol{z})\| \leq C_T,$$

where $C_T \coloneqq d\widetilde{G} \cdot (41 + 40B) \cdot (1 + 2\eta\widetilde{G}R)T$. Furthermore,

$$\nabla^2 \Phi_t(\boldsymbol{z}) \preceq \left(\eta \widetilde{G}^2 T + C_T/R + 2\nu^{-1} C_T^2\right) \cdot I.$$

Proof. Fix $t \in [T]$ and $\boldsymbol{z} \in \operatorname{int} \mathbb{B}(R)$ such that $\|\boldsymbol{z} - \boldsymbol{w}_t\|^2_{\nabla^2 \Psi(\boldsymbol{z})} \leq \nu B^2$. By Lemma D.2, we have

$$B^2 \geq \left(\frac{\|\boldsymbol{z}\|^2 - \|\boldsymbol{w}_t\|^2}{R^2 - \|\boldsymbol{z}\|^2}\right)^2 = \left(\frac{R^2 - \|\boldsymbol{w}_t\|^2}{R^2 - \|\boldsymbol{z}\|^2} - 1\right)^2.$$

This implies that

$$\frac{2\nu}{R^2 - \|\boldsymbol{z}\|^2} \leq \frac{2\nu \cdot (1 + B)}{R^2 - \|\boldsymbol{w}_t\|^2} \leq 40 d R^{-1} \widetilde{G}(1 + B)(1 + 2\eta\widetilde{G}R) \cdot T, \tag{30}$$

where the last inequality follows by Lemma D.5. Therefore, by the expression of $\nabla \Phi_t(\boldsymbol{z})$ and the triangle inequality, we have

$$\begin{aligned}
\|\nabla \Phi_t(\boldsymbol{z})\| &\leq \left\|\frac{2\nu \boldsymbol{z}}{R^2 - \|\boldsymbol{z}\|^2}\right\| + \left\|\eta \sum_{s=1}^{t-1} \widetilde{\boldsymbol{g}}_s \widetilde{\boldsymbol{g}}_s^\top (\boldsymbol{z} - \boldsymbol{u}_s) + \sum_{s=1}^{t-1} \widetilde{\boldsymbol{g}}_s\right\|, \\
&\leq \frac{2\nu R}{R^2 - \|\boldsymbol{z}\|^2} + \widetilde{G} \cdot (1 + 2\eta \widetilde{G} R) T, \\
&\leq C_T, \quad (C_T \text{ as in the lemma statement})
\end{aligned}$$

where in the last inequality we used (30). On the other hand, we have

$$\begin{aligned}
\nabla^2 \Phi_t(\boldsymbol{z}) &= \frac{2\nu I}{R^2 - \|\boldsymbol{z}\|^2} + \frac{4\nu \boldsymbol{z}\boldsymbol{z}^\top}{(R^2 - \|\boldsymbol{z}\|^2)^2} + \eta \sum_{s=1}^{t-1} \widetilde{\boldsymbol{g}}_s \widetilde{\boldsymbol{g}}_s^\top, \\
&\preceq \left(\eta \widetilde{G}^2 T + C_T/R + 2\nu^{-1} C_T^2\right) \cdot I.
\end{aligned}$$

where the last inequality follows from (30) and the facts that $\boldsymbol{z} \in \mathbb{B}(R)$ and $\widetilde{\boldsymbol{g}}_s \in \mathbb{B}(\widetilde{G})$, for all $s \in [t-1]$. $\square$

Lemma D.7 (Master lemma)**.** *Let $\eta, \nu, c \in (0,1)$, $T \in \mathbb{N}$, and $\widetilde{G} > 0$ be given. Further, let $(\boldsymbol{u}_t)$ be the iterates of Algorithm 2 with parameters (T, η, ν, c) and suppose that*

- *$\widetilde{\boldsymbol{g}}_t \in \mathbb{B}(\widetilde{G})$, for all $t \in [T]$;*
- *$10\widetilde{G}R \leq \nu \leq 10d\widetilde{G}RT$; and*
- *$\eta \leq \frac{1}{5\widetilde{G}R}$.*

Then, we have $(\boldsymbol{u}_t) \subset \operatorname{int} \mathbb{B}(R)$ *and*

$$\forall t \in [T], \quad \frac{\sqrt{\nu}}{4}\left(\|\boldsymbol{u}_t - \boldsymbol{w}_t\|_{\nabla^2 \Phi_t(\boldsymbol{u}_t)} - 2\xi\right) \le \frac{\sqrt{\nu}}{2}\left(\lambda(\boldsymbol{u}_t, \Phi_t) - \xi\right) \le \lambda(\boldsymbol{u}_{t-1}, \Phi_t)^2 \le \frac{2\widetilde{G}^2 R^2}{\nu},$$

where $\xi \coloneqq \sqrt{\frac{\widetilde{G}R}{30T}}$ *and* $\lambda(\cdot,\cdot)$ *is the Newton decrement (see Appendix C). Further, we have*

$$\sum_{t=1}^{T} \|\boldsymbol{u}_t - \boldsymbol{w}_t\|_{\nabla^2 \Phi_t(\boldsymbol{u}_t)} \le \frac{3\sqrt{\nu}}{32} + \frac{6d \log(1 + T/d)}{\eta\sqrt{\nu}}.$$

and

$$\sum_{t=1}^{T} \|\boldsymbol{u}_t - \boldsymbol{u}_{t-1}\|^2_{\nabla^2 \Psi(\boldsymbol{u}_t)} + \sum_{t=1}^{T} \|\boldsymbol{u}_t - \boldsymbol{u}_{t-1}\|^2_{\nabla^2 \Psi(\boldsymbol{u}_{t-1})} \le \frac{\nu}{2} + \frac{13^2 d \log(1 + T/d)}{\eta}. \tag{31}$$

Proof of Lemma D.7. Define

$$\widetilde{\boldsymbol{u}}_{t+1} \coloneqq \boldsymbol{u}_t - \nabla^{-2}\Phi_{t+1}(\boldsymbol{u}_t)\nabla\Phi_{t+1}(\boldsymbol{u}_t), \quad \text{and} \quad \widetilde{\nabla}_t \coloneqq \sum_{k=1}^{m+1}\left(\frac{2\nu}{R^2 - \|\boldsymbol{z}_t\|^2} - \frac{2\nu}{R^2 - \|\boldsymbol{u}_t\|^2}\right)^{k-1} \Sigma_t^k \nabla_t,$$

and note that $\boldsymbol{u}_{t+1} = \boldsymbol{u}_t - \widetilde{\nabla}_t$ from Line 8 of Algorithm 2. We will show by induction that for all $s \ge 1$,

$$\boldsymbol{u}_s \in \operatorname{int} \mathbb{B}(R),$$

and

$$\frac{\sqrt{\nu}}{4}(\|\boldsymbol{u}_s - \boldsymbol{w}_s\|_{\nabla^2 \Phi_s(\boldsymbol{u}_s)} - 2\xi) \le \frac{\sqrt{\nu}}{2}(\lambda(\boldsymbol{u}_s, \Phi_s) - \xi) \le \lambda(\boldsymbol{u}_{s-1}, \Phi_s)^2 \le \frac{2\widetilde{G}^2 R^2}{\nu}, \tag{32}$$

where $\xi = \sqrt{\frac{\widetilde{G}R}{30T}}$ and $\boldsymbol{u}_0 = \mathbf{0}$ by convention. The base case follows trivially since $\nabla\Phi_1(\boldsymbol{u}_0) = \nabla\Phi_1(\boldsymbol{u}_1) = \mathbf{0}$ and $\boldsymbol{u}_1 = \boldsymbol{w}_1$. Suppose that (32) holds for $s = t$. We will show that it holds for $s = t + 1$. By the expression of Φ_{t+1} in (3), we have $\nabla\Phi_{t+1}(\boldsymbol{u}_t) = \widetilde{\boldsymbol{g}}_t + \nabla\Phi_t(\boldsymbol{u}_t)$, and so by the fact that $(a + b)^2 \le 2a^2 + 2b^2$, we get

$$\begin{aligned}
\lambda(\boldsymbol{u}_t, \Phi_{t+1})^2 &= \|\nabla\Phi_{t+1}(\boldsymbol{u}_t)\|^2_{\nabla^{-2}\Phi_{t+1}(\boldsymbol{u}_t)}, \\
&\le 2\|\nabla\Phi_t(\boldsymbol{u}_t)\|^2_{\nabla^{-2}\Phi_t(\boldsymbol{u}_t)} + 2\|\widetilde{\boldsymbol{g}}_t\|^2_{\nabla^{-2}\Phi_t(\boldsymbol{u}_t)}, \quad (\nabla^2\Phi_{t+1}(\cdot) \succeq \nabla^2\Phi_t(\cdot)) \\
&= 2\lambda(\boldsymbol{u}_t, \Phi_t)^2 + 2\|\widetilde{\boldsymbol{g}}_t\|^2_{\nabla^{-2}\Phi_t(\boldsymbol{u}_t)}, && (33) \\
&\le 2\cdot\left(16\frac{\widetilde{G}^4 R^4}{\nu^3} + 8\xi\frac{\widetilde{G}^2 R^2}{\nu^{3/2}} + \xi^2\right) + \frac{\widetilde{G}^2 R^2}{\nu}, && (34) \\
&\le \frac{2\widetilde{G}^2 R^2}{\nu}, && (35)
\end{aligned}$$

where in (34) we used the induction hypothesis in (32) for $s = t$ and the bound on $\|\widetilde{\boldsymbol{g}}_t\|^2_{\nabla^{-2}\Phi_t(\boldsymbol{u}_t)}$ from Lemma D.4; and (35) uses that $10\widetilde{G}R \le \nu \le 10d\widetilde{G}RT$ and $\xi = \sqrt{\frac{\widetilde{G}R}{30T}}$ (which implies that $\xi^2 \le \frac{\widetilde{G}^2 R^2}{3\nu}$).

By taking the square-root in (35), we get that

$$\lambda(\boldsymbol{u}_t, \Phi_{t+1}) \le \frac{\sqrt{2}\widetilde{G}R}{\sqrt{\nu}}. \tag{36}$$

Using this with Lemma C.1 and the facts that $\widetilde{\boldsymbol{u}}_{t+1}$ is the standard Newton step and Φ_{t+1} is self-concordant with constant $1/\sqrt{\nu}$, we get that

$$\begin{aligned}
\lambda(\widetilde{\boldsymbol{u}}_{t+1}, \Phi_{t+1}) &\le \frac{\nu^{-1/2} \cdot \lambda(\boldsymbol{u}_t, \Phi_{t+1})^2}{(1 - \nu^{-1/2}\lambda(\boldsymbol{u}_t, \Phi_{t+1}))^2}, \\
&\le \frac{\nu^{-1/2} \cdot \lambda(\boldsymbol{u}_t, \Phi_{t+1})^2}{(1 - \sqrt{2}\nu^{-1}\widetilde{G}R)^2}, \quad \text{(by (35))} \\
&\le \frac{2}{\sqrt{\nu}} \cdot \lambda(\boldsymbol{u}_t, \Phi_{t+1})^2, \qquad (37)
\end{aligned}$$

where the last inequality follows by the fact that $\nu \ge 10\widetilde{G}R$. Using (37) together with (37) and (36) implies

$$\lambda(\widetilde{\boldsymbol{u}}_{t+1}, \Phi_{t+1}) \le \frac{4\widetilde{G}^2R^2}{\nu^{3/2}} \le \frac{2}{3\sqrt{6}}\sqrt{\widetilde{G}R} \le \frac{\sqrt{\nu}}{9}, \tag{38}$$

where we used the fact that $\nu \ge 10\widetilde{G}R$ again. Using this with Lemma C.1 and the facts that $\boldsymbol{w}_{t+1}$ is the minimizer of Φ_{t+1} and Φ_{t+1} is self-concordant with constant $1/\sqrt{\nu}$ (see Lemma D.3), we have

$$\|\widetilde{\boldsymbol{u}}_{t+1} - \boldsymbol{w}_{t+1}\|_{\nabla^2\Phi_{t+1}(\widetilde{\boldsymbol{u}}_{t+1})} \le 2\lambda(\widetilde{\boldsymbol{u}}_{t+1}, \Phi_{t+1}).$$

Combining this with (38) implies that

$$\|\widetilde{\boldsymbol{u}}_{t+1} - \boldsymbol{w}_{t+1}\|^2_{\nabla^2\Phi_{t+1}(\widetilde{\boldsymbol{u}}_{t+1})} \le \frac{4\nu}{81}$$

Thus, Lemma D.6 instantiated with $B = 2/9$ implies that

$$\nabla^2\Phi_{t+1}(\widetilde{\boldsymbol{u}}_{t+1}) \preceq (\eta\widetilde{G}^2T + C_T/R + 2\nu^{-1}C_T^2) \cdot I, \tag{39}$$

where

$$C_T \coloneqq 51d\widetilde{G} \cdot (1 + 2\eta\widetilde{G}R)T. \tag{40}$$

On the other hand, since $\boldsymbol{\nabla}_t = \nabla\Phi_{t+1}(\boldsymbol{z}_t)$ we have,

$$\begin{aligned}
\|\widetilde{\boldsymbol{u}}_{t+1} - \boldsymbol{u}_{t+1}\| &= \left\| \nabla^{-2}\Phi_{t+1}(\boldsymbol{u}_t)\nabla\Phi_{t+1}(\boldsymbol{u}_t) - \sum_{k=1}^{m+1} \left(\frac{2\nu}{R^2 - \|\boldsymbol{z}_t\|^2} - \frac{2\nu}{R^2 - \|\boldsymbol{u}_t\|^2} \right)^{k-1} \Sigma_t^k \boldsymbol{\nabla}_t \right\|, \\
&= \left\| \left(\nabla^{-2}\Phi_{t+1}(\boldsymbol{u}_t) - \sum_{k=1}^{m+1} \left(\frac{2\nu}{R^2 - \|\boldsymbol{z}_t\|^2} - \frac{2\nu}{R^2 - \|\boldsymbol{u}_t\|^2} \right)^{k-1} \Sigma_t^k \right) \nabla\Phi_{t+1}(\boldsymbol{u}_t) \right\|, \\
&\le \left\| \nabla^{-2}\Phi_{t+1}(\boldsymbol{u}_t) - \sum_{k=1}^{m+1} \left(\frac{2\nu}{R^2 - \|\boldsymbol{z}_t\|^2} - \frac{2\nu}{R^2 - \|\boldsymbol{u}_t\|^2} \right)^{k-1} \Sigma_t^k \right\| \cdot \|\nabla\Phi_{t+1}(\boldsymbol{u}_t)\|. \qquad (41)
\end{aligned}$$

Now, by the induction hypothesis (i.e., (32)), we have $\|\boldsymbol{u}_t - \boldsymbol{w}_t\|_{\nabla^2\Phi_t(\boldsymbol{u}_t)} \leq 8\nu^{-1/2}\xi + \frac{8\widetilde{G}^2R^2}{\nu^{3/2}} \leq \frac{2\sqrt{\nu}}{9}$, where the last inequality uses that $\nu \geq 10\widetilde{G}R$ and $\xi = \sqrt{\frac{\widetilde{G}R}{30T}}$. Thus, by Lemma D.6 instantiated with $B = 2/9$, we have

$$\|\nabla\Phi_t(\boldsymbol{u}_t)\| \leq C_T,$$

where C_T is as in (40). Therefore, by the triangle inequality and $\widetilde{\boldsymbol{g}}_t \in \mathbb{B}(\widetilde{G})$ (by assumption), we have

$$\|\nabla\Phi_{t+1}(\boldsymbol{u}_t)\| = \|\nabla\Phi_t(\boldsymbol{u}_t) + \widetilde{\boldsymbol{g}}_t\| \leq C_T + \widetilde{G}. \tag{42}$$

On the other hand, by the fact that $\boldsymbol{u}_t \in \operatorname{int}\mathbb{B}(R)$ (by the induction hypothesis), Lemma D.1 implies that

$$\left\|\nabla^{-2}\Phi_{t+1}(\boldsymbol{u}_t) - \sum_{k=1}^{m+1}\left(\frac{2\nu}{R^2 - \|\boldsymbol{z}_t\|^2} - \frac{2\nu}{R^2 - \|\boldsymbol{u}_t\|^2}\right)^{k-1}\Sigma_t^k\right\| \leq \frac{R^2c^m}{2\nu\cdot(1-c)}.$$

Plugging this and (42) into (41) implies that

$$\|\widetilde{\boldsymbol{u}}_{t+1} - \boldsymbol{u}_{t+1}\| \leq \frac{R^2c^m\cdot(C_T + \widetilde{G})}{2\nu\cdot(1-c)}. \tag{43}$$

Combining (43) with (39) implies that

$$\|\widetilde{\boldsymbol{u}}_{t+1} - \boldsymbol{u}_{t+1}\|_{\nabla^2\Phi_{t+1}(\widetilde{\boldsymbol{u}}_{t+1})} \leq \tilde{\xi} \coloneqq \frac{R^2c^m\cdot(C_T + \widetilde{G})\cdot\sqrt{\eta\widetilde{G}^2T + C_T/R + 2\nu^{-1}C_T^2}}{2\nu\cdot(1-c)}. \tag{44}$$

We now show that this implies that $\boldsymbol{u}_{t+1} \in \mathbb{B}(R)$. First, by (36) and the fact that $\nu \geq 10\widetilde{G}R$, we have

$$\|\widetilde{\boldsymbol{u}}_{t+1} - \boldsymbol{u}_t\|_{\nabla^2\Phi_{t+1}(\boldsymbol{u}_t)} = \lambda(\boldsymbol{u}_t, \Phi_{t+1}) \leq \frac{\sqrt{2}\widetilde{G}R}{\sqrt{\nu}} \leq \sqrt{\frac{\widetilde{G}R}{5}} < \frac{\sqrt{\nu}}{7}. \tag{45}$$

Combining this with Lemma C.3 and the facts that Φ_{t+1} is self-concordant with constant $1/\sqrt{\nu}$ (Lemma D.3) and $\boldsymbol{u}_t \in \operatorname{int}\mathbb{B}(R)$ (induction hypothesis), we get that

$$\widetilde{\boldsymbol{u}}_{t+1} \in \operatorname{int}\mathbb{B}(R).$$

Now, by our choice of m in Algorithm 2 (i.e., $m = \mathfrak{c}\cdot\log_c(dT)$ with $\mathfrak{c}$ a sufficiently large constant) and the facts that $\nu \geq 10\widetilde{G}R$ and $\eta \leq \frac{1}{5\widetilde{G}R}$, we have

$$\tilde{\xi} < \frac{1}{5}\sqrt{\frac{\widetilde{G}R}{30T}} = \frac{\xi}{5}, \quad \left(\text{since } \xi = \sqrt{\frac{\widetilde{G}R}{30T}}\right) \tag{46}$$

$$\leq \frac{\sqrt{\nu}}{86}, \tag{47}$$

and so (44) implies that

$$\|\widetilde{\boldsymbol{u}}_{t+1} - \boldsymbol{u}_{t+1}\|_{\nabla^2\Phi_{t+1}(\widetilde{\boldsymbol{u}}_{t+1})} < \sqrt{\nu}/4. \tag{48}$$

We now use this to bound the Newton decrement $\lambda(\boldsymbol{u}_{t+1}, \Phi_{t+1})$. By Lemma C.3, (48), and the fact that Φ_{t+1} is self-concordant with constant $1/\sqrt{\nu}$ (Lemma D.3), we have

$$\|\widetilde{\boldsymbol{u}}_{t+1} - \boldsymbol{u}_{t+1}\|_{\nabla^2\Phi_{t+1}(\boldsymbol{u}_{t+1})} \le 2\|\widetilde{\boldsymbol{u}}_{t+1} - \boldsymbol{u}_{t+1}\|_{\nabla^2\Phi_{t+1}(\widetilde{\boldsymbol{u}}_{t+1})} \le 2\tilde{\xi}, \quad \text{(by (44))} \tag{49}$$

$$< \sqrt{\nu}/2. \quad \text{(by (47))} \tag{50}$$

Using this and the triangle inequality, we get

$$\begin{aligned}
&\lambda(\boldsymbol{u}_{t+1}, \Phi_{t+1}) \\
&= \|\nabla\Phi_t(\boldsymbol{u}_{t+1})\|_{\nabla^{-2}\Phi_{t+1}(\boldsymbol{u}_{t+1})} \\
&\le \|\nabla\Phi_t(\widetilde{\boldsymbol{u}}_{t+1})\|_{\nabla^{-2}\Phi_{t+1}(\boldsymbol{u}_{t+1})} + \|\nabla\Phi_{t+1}(\boldsymbol{u}_{t+1}) - \nabla\Phi_{t+1}(\widetilde{\boldsymbol{u}}_{t+1})\|_{\nabla^{-2}\Phi_{t+1}(\boldsymbol{u}_{t+1})}, \quad \text{(triangle inequality)} \\
&\le \|\nabla\Phi_t(\widetilde{\boldsymbol{u}}_{t+1})\|_{\nabla^{-2}\Phi_{t+1}(\boldsymbol{u}_{t+1})} + 2\|\widetilde{\boldsymbol{u}}_{t+1} - \boldsymbol{u}_{t+1}\|_{\nabla^2\Phi_{t+1}(\boldsymbol{u}_{t+1})}, \quad \text{(by (50) and Lemma C.4)} \\
&\le (1 - 2\tilde{\xi}/\sqrt{\nu})^{-1}\|\nabla\Phi_t(\widetilde{\boldsymbol{u}}_{t+1})\|_{\nabla^{-2}\Phi_{t+1}(\widetilde{\boldsymbol{u}}_{t+1})} + 2\|\widetilde{\boldsymbol{u}}_{t+1} - \boldsymbol{u}_{t+1}\|_{\nabla^2\Phi_{t+1}(\boldsymbol{u}_{t+1})}, \quad \text{(by (49) and Lemma C.2)} \\
&\le (1 - 2\tilde{\xi}/\sqrt{\nu})^{-1}\|\nabla\Phi_t(\widetilde{\boldsymbol{u}}_{t+1})\|_{\nabla^{-2}\Phi_{t+1}(\widetilde{\boldsymbol{u}}_{t+1})} + 4\tilde{\xi}, \quad \text{(by (49))} \\
&\le \lambda(\widetilde{\boldsymbol{u}}_{t+1}, \Phi_{t+1}) + 4\tilde{\xi}\cdot\lambda(\widetilde{\boldsymbol{u}}_{t+1}, \Phi_{t+1})/\sqrt{\nu} + 4\tilde{\xi}, \quad \text{((47) and } \tfrac{1}{1-x} \le 1 + 2x, \forall x \le \tfrac{1}{2}) \\
&\le \lambda(\widetilde{\boldsymbol{u}}_{t+1}, \Phi_{t+1}) + 5\tilde{\xi},
\end{aligned} \tag{51}$$

where the last inequality follows by (38). By combining (51) with (37) and (35), we get that

$$\begin{aligned}
\lambda(\boldsymbol{u}_{t+1}, \Phi_{t+1}) &\le 5\tilde{\xi} + \frac{2}{\sqrt{\nu}}\lambda(\boldsymbol{u}_t, \Phi_{t+1})^2, \\
&\le 5\tilde{\xi} + \frac{4\widetilde{G}^2R^2}{\nu^{3/2}},
\end{aligned}$$

and so by (46) and the fact that $\nu \ge 10\widetilde{G}R$ implies that

$$\lambda(\boldsymbol{u}_{t+1}, \Phi_{t+1}) \le \xi + \frac{2}{\sqrt{\nu}}\lambda(\boldsymbol{u}_t, \Phi_{t+1})^2 \quad \text{and} \quad \lambda(\boldsymbol{u}_{t+1}, \Phi_{t+1}) \le \frac{\sqrt{\nu}}{4}. \tag{52}$$

Now, by Lemma C.1 and the facts that $\boldsymbol{w}_{t+1}$ is the minimizer of Φ_{t+1} and $\lambda(\boldsymbol{u}_{t+1}, \Phi_{t+1}) \le \frac{\sqrt{\nu}}{2}$ (by (52)), we have $\|\boldsymbol{u}_{t+1} - \boldsymbol{w}_{t+1}\|_{\nabla^2\Phi_{t+1}(\boldsymbol{u}_t)} \le 2\lambda(\boldsymbol{u}_{t+1}, \Phi_{t+1})$. Combining this with the inequality on the left-hand side of (52), implies (32) for $s = t + 1$, which concludes the induction.

We now use (32) together with (33) to bound the sums

$$S \coloneqq \sum_{t=1}^{T} \|\boldsymbol{u}_t - \boldsymbol{w}_t\|_{\nabla^2\Phi_t(\boldsymbol{u}_t)}, \quad S' \coloneqq \sum_{t=1}^{T} \|\boldsymbol{u}_t - \boldsymbol{u}_{t-1}\|^2_{\nabla^2\Psi(\boldsymbol{u}_t)}, \quad \text{and} \quad S'' \coloneqq \sum_{t=1}^{T} \|\boldsymbol{u}_t - \boldsymbol{u}_{t-1}\|^2_{\nabla^2\Psi(\boldsymbol{u}_{t-1})}.$$

Bounding S. We first bound the sum $\sum_{t=1}^{T}\lambda(\boldsymbol{u}_t, \Phi_t)^i$, for $i = 1, 2$. By (33) and the fact that $\lambda(\boldsymbol{u}_{t+1}, \Phi_{t+1}) \le \frac{2}{\sqrt{\nu}}\lambda(\boldsymbol{u}_t, \Phi_{t+1})^2 + \xi$ (see (52)), we have

$$\lambda(\boldsymbol{u}_{t+1}, \Phi_{t+1}) \le \frac{4}{\sqrt{\nu}}\lambda(\boldsymbol{u}_t, \Phi_t)^2 + \frac{4}{\sqrt{\nu}}\|\widetilde{\boldsymbol{g}}_t\|^2_{\nabla^{-2}\Phi_t(\boldsymbol{u}_t)} + \xi. \tag{53}$$

Summing (53), for $t = 1, \dots, T$, rearranging, and using that $\lambda(\boldsymbol{u}_{T+1}, \Phi_{T+1}) \ge 0$, we get

$$\sum_{t=2}^{T}\left(\lambda(\boldsymbol{u}_t, \Phi_t) - \frac{4}{\sqrt{\nu}}\lambda(\boldsymbol{u}_t, \Phi_t)^2\right) \le \frac{4}{\sqrt{\nu}}\lambda(\boldsymbol{u}_1, \Phi_1)^2 + \frac{4}{\sqrt{\nu}}\sum_{t=1}^{T}\|\widetilde{\boldsymbol{g}}_t\|^2_{\nabla^{-2}\Phi_t(\boldsymbol{u}_t)} + T\xi.$$

Using (32) (the induction hypothesis), we have for all $t \in [T]$:

$$0 \leq \frac{4}{\sqrt{\nu}} \lambda(\boldsymbol{u}_t, \Phi_t) \leq \frac{16\widetilde{G}^2 R^2}{\nu^2} + \frac{4\xi}{\sqrt{\nu}} \leq \frac{1}{4}, \tag{54}$$

where the last inequality follows by the fact that $\nu \geq 10\widetilde{G}R$ and $\xi = \sqrt{\frac{\widetilde{G}R}{30T}}$. Therefore, we have

$$\begin{aligned}
\frac{3}{4}\sum_{t=1}^{T} \lambda(\boldsymbol{u}_t, \Phi_t) &\leq \lambda(\boldsymbol{u}_1, \Phi_1) + \frac{4}{\sqrt{\nu}} \sum_{t=1}^{T} \|\widetilde{\boldsymbol{g}}_t\|^2_{\nabla^{-2}\Phi_t(\boldsymbol{u}_t)}, \\
&\leq \frac{\sqrt{\nu}}{16} + \frac{4}{\sqrt{\nu}} \sum_{t=1}^{T} \|\widetilde{\boldsymbol{g}}_t\|^2_{\nabla^{-2}\Phi_t(\boldsymbol{u}_t)}, \\
&\leq \frac{\sqrt{\nu}}{16} + \frac{4d\log(1+T/d)}{\eta\sqrt{\nu}},
\end{aligned} \tag{55}$$

where the last inequality follows by Lemma D.4 and the range assumption on η. Now, by Lemma C.1, (54), and the facts that $\boldsymbol{w}_t$ is the minimizer of Φ_t and Φ_t is self-concordant with constant $1/\sqrt{\nu}$, we have:

$$\|\boldsymbol{u}_t - \boldsymbol{w}_t\|_{\nabla^2\Phi_t(\boldsymbol{u}_t)} \leq 2\lambda(\boldsymbol{u}_t, \Phi_t).$$

Combining this with (55) implies that

$$S = \sum_{t=1}^{T} \|\boldsymbol{u}_t - \boldsymbol{w}_t\|_{\nabla^2\Phi_t(\boldsymbol{u}_t)} \leq \frac{3\sqrt{\nu}}{32} + \frac{6d\log(1+T/d)}{\eta\sqrt{\nu}}.$$

Bounding S' and S''. We now bound S' and S''. By Lemma C.2 and the facts that $\|\boldsymbol{u}_{t+1} - \widetilde{\boldsymbol{u}}_{t+1}\|_{\nabla^2\Psi(\boldsymbol{u}_{t+1})} \leq \sqrt{\nu}/2$ (which follows from (50) since $\nabla^2\Psi(\cdot) \leq \nabla^2\Phi_{t+1}(\cdot)$) and Ψ is self-concordant with constant $1/\sqrt{\nu}$ (Lemma D.3), we have

$$\begin{aligned}
\|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\|_{\nabla^2\Psi(\boldsymbol{u}_{t+1})} &\leq 2\|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\|_{\nabla^2\Psi(\widetilde{\boldsymbol{u}}_{t+1})}, \\
&\leq 2\|\boldsymbol{u}_{t+1} - \widetilde{\boldsymbol{u}}_{t+1}\|_{\nabla^2\Psi(\widetilde{\boldsymbol{u}}_{t+1})} + 2\|\widetilde{\boldsymbol{u}}_{t+1} - \boldsymbol{u}_t\|_{\nabla^2\Psi(\widetilde{\boldsymbol{u}}_{t+1})}, \quad \text{(triangle inequality)} \\
&\leq 2\tilde{\xi} + 2\|\widetilde{\boldsymbol{u}}_{t+1} - \boldsymbol{u}_t\|_{\nabla^2\Phi_{t+1}(\widetilde{\boldsymbol{u}}_{t+1})} \quad \text{(by (50) and } \nabla^2\Phi_{t+1}(\widetilde{\boldsymbol{u}}_{t+1}) \geq \nabla^2\Psi(\widetilde{\boldsymbol{u}}_{t+1})\text{)},
\end{aligned}$$

and so using Lemma C.3 and the facts that $\|\widetilde{\boldsymbol{u}}_{t+1} - \boldsymbol{u}_t\|_{\nabla^2\Phi_{t+1}(\boldsymbol{u}_t)} \leq \sqrt{\nu}/2$ (by (45)) and Φ_{t+1} is self-concordant with constant $1/\sqrt{\nu}$, we have

$$\begin{aligned}
\|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\|_{\nabla^2\Psi(\boldsymbol{u}_{t+1})} &\leq 2\tilde{\xi} + 4\|\widetilde{\boldsymbol{u}}_{t+1} - \boldsymbol{u}_t\|_{\nabla^2\Phi_{t+1}(\boldsymbol{u}_t)}, \\
&\leq 2\tilde{\xi} + 4\lambda(\boldsymbol{u}_t, \Phi_{t+1}), \quad \text{(by (45))} \\
&\leq \frac{3\sqrt{\nu}}{5},
\end{aligned} \tag{56}$$

where the last inequality follows by (47) and (45). Thus, since Ψ is self-concordant with constant $1/\sqrt{\nu}$, Lemma C.3 implies that

$$\|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\|_{\nabla^2\Psi(\boldsymbol{u}_t)} \leq \frac{5}{2}\|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\|_{\nabla^2\Psi(\boldsymbol{u}_{t+1})}. \tag{57}$$

From this, it suffices to bound the sum $S' \coloneqq \sum_{t=1}^T \|\boldsymbol{u}_t - \boldsymbol{u}_{t-1}\|^2_{\nabla^2\Psi(\boldsymbol{u}_t)}$. Using (56) and the fact that $(a+b)^2 \le 5a^2 + (5/4)b^2$, for all $a, b \in \mathbb{R}$, we have

$$
\begin{aligned}
&\sum_{t=1}^T \|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\|^2_{\nabla^2\Psi(\boldsymbol{u}_{t+1})} \\
&\le 20T\tilde{\xi}^2 + 20\sum_{t=1}^T \lambda(\boldsymbol{u}_t, \Phi_{t+1})^2, \\
&\le 20T\tilde{\xi}^2 + 40\sum_{t=1}^T \lambda(\boldsymbol{u}_t, \Phi_t)^2 + 40\sum_{t=1}^T \|\widetilde{\boldsymbol{g}}_t\|^2_{\nabla^{-2}\Phi_t(\boldsymbol{u}_t)}, \quad \text{(by (33))} \\
&\le 20T\tilde{\xi}^2 + 40\sum_{t=1}^T \lambda(\boldsymbol{u}_t, \Phi_t)^2 + \frac{40d\log(1+T/d)}{\eta}, \quad \text{(by Lemma D.4)} \\
&\le 20T\tilde{\xi}^2 + \frac{5\sqrt{\nu}}{2}\sum_{t=1}^T \lambda(\boldsymbol{u}_t, \Phi_t) + \frac{40d\log(1+T/d)}{\eta}, \quad (\lambda(\boldsymbol{u}_t, \Phi_t) \le \tfrac{\sqrt{\nu}}{16} \text{ by (54))}, \\
&\le 20T\tilde{\xi}^2 + \frac{\nu}{12} + \frac{16d\log(1+T/d)}{3\eta} + \frac{40d\log(1+T/d)}{\eta}, \quad \text{(by (55))} \\
&\le \frac{\nu}{8} + \frac{46d\log(1+T/d)}{\eta},
\end{aligned}
$$

where the last inequality follows by the bound on $\tilde{\xi}$ in (47). Combining this with (57) implies (31). $\square$

D.4. Regret of FTRL (Proof of Lemma 3.2)

Proof. Fix $\boldsymbol{w} \in \operatorname{int}\mathbb{B}(R)$. For any $t \ge 1$, define $\phi_t(\boldsymbol{x}) \coloneqq \boldsymbol{x}^\top \widetilde{\boldsymbol{g}}_t + \eta\langle\widetilde{\boldsymbol{g}}_t, \boldsymbol{x} - \boldsymbol{u}_t\rangle^2/2$ and $\phi_0(\boldsymbol{x}) \coloneqq \Psi(\boldsymbol{x})$, and note that $\Phi_t(\boldsymbol{x}) = \sum_{s=0}^{t-1}\phi_s(\boldsymbol{x})$ and $\boldsymbol{u}_t \in \arg\min_{\boldsymbol{x}\in\mathbb{B}(R)} \sum_{s=0}^{t-1}\phi_s(\boldsymbol{x})$. By (Cesa-Bianchi and Lugosi, 2006, Lemma 3.1), we have

$$
\sum_{t=0}^T \phi_t(\boldsymbol{u}_{t+1}) \le \sum_{t=0}^T \phi_t(\boldsymbol{w}),
$$

which implies that

$$
\begin{aligned}
\sum_{t=1}^T \langle \boldsymbol{u}_{t+1} - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle &\le \phi_0(\boldsymbol{w}) - \phi_0(\boldsymbol{u}_1) + \frac{\eta}{2}\sum_{t=1}^T \langle \boldsymbol{u}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle^2, \\
&\le \Psi(\boldsymbol{w}) - \Psi(\boldsymbol{0}) + \frac{\eta}{2}\sum_{t=1}^T \langle \boldsymbol{u}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle^2,
\end{aligned} \tag{58}
$$

where the last inequality follows by the fact that $\boldsymbol{u}_1 = \boldsymbol{0} \in \arg\min_{\boldsymbol{x}\in\mathbb{B}(R)} -\log(R^2 - \|\boldsymbol{x}\|^2)$. Now, it suffices to bound the sum $\sum_{t=1}^T \langle \boldsymbol{u}_t - \boldsymbol{u}_{t+1}, \widetilde{\boldsymbol{g}}_t\rangle$. By Taylor's theorem, the exists $\boldsymbol{y}_t$ in the segment $[\boldsymbol{u}_t, \boldsymbol{u}_{t+1}]$ such that

$$
\begin{aligned}
\Phi_{t+1}(\boldsymbol{u}_t) - \Phi_{t+1}(\boldsymbol{u}_{t+1}) &\ge \nabla\Phi_{t+1}(\boldsymbol{u}_{t+1})^\top(\boldsymbol{u}_t - \boldsymbol{u}_{t+1}) + \frac{1}{2}\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|^2_{\nabla^2\Phi_{t+1}(\boldsymbol{y}_t)}, \\
&\ge \frac{1}{2}\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|^2_{\nabla^2\Phi_{t+1}(\boldsymbol{y}_t)},
\end{aligned} \tag{59}
$$

where the last inequality uses the fact that $\boldsymbol{u}_{t+1} \in \arg\min_{\boldsymbol{x}\in\mathbb{B}(R)} \Phi_{t+1}(\boldsymbol{x})$ is in the interior of $\mathbb{B}(R)$ by self-concordance of Φ_{t+1}. On the other hand, using the convexity of Φ_{t+1} and the fact that $\nabla\Phi_{t+1}(\boldsymbol{u}_t) = \nabla\phi_{t+1}(\boldsymbol{u}_t) + \nabla\Phi_t(\boldsymbol{u}_t) = \nabla\phi_{t+1}(\boldsymbol{u}_t)$ (by optimality of $\boldsymbol{u}_t$), we get that

$$\begin{aligned}\Phi_{t+1}(\boldsymbol{u}_t) - \Phi_{t+1}(\boldsymbol{u}_{t+1}) &\le \langle \boldsymbol{u}_t - \boldsymbol{u}_{t+1}, \nabla\phi_{t+1}(\boldsymbol{u}_t)\rangle,\\ &= \langle \boldsymbol{u}_t - \boldsymbol{u}_{t+1}, \widetilde{\boldsymbol{g}}_t\rangle(1 + \eta\langle\widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{u}_{t+1}\rangle),\\ &\le \|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_{\nabla^2\Phi_{t+1}(\boldsymbol{y}_t)} \cdot \|\widetilde{\boldsymbol{g}}_t\|_{\nabla^{-2}\Phi_{t+1}(\boldsymbol{y}_t)} \cdot (1 + \eta\langle\widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{u}_{t+1}\rangle),\\ &\le \frac{3}{2}\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_{\nabla^2\Phi_{t+1}(\boldsymbol{y}_t)} \cdot \|\widetilde{\boldsymbol{g}}_t\|_{\nabla^{-2}\Phi_{t+1}(\boldsymbol{y}_t)},\end{aligned}$$

where the last inequality follows by the fact that $\eta \le \frac{1}{5\widetilde{G}R}$, $(\widetilde{\boldsymbol{g}}_t) \subset \mathbb{B}(\widetilde{G})$, and $(\boldsymbol{u}_t) \subset \mathbb{B}(R)$. Combining this and (59), we get

$$\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_{\nabla^2\Phi_{t+1}(\boldsymbol{y}_t)} \le 3\|\widetilde{\boldsymbol{g}}_t\|_{\nabla^{-2}\Phi_{t+1}(\boldsymbol{y}_t)}.$$

Using this and Hölder's inequality leads to

$$\langle\widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{u}_{t+1}\rangle \le \|\widetilde{\boldsymbol{g}}_t\|_{\nabla^{-2}\Phi_{t+1}(\boldsymbol{y}_t)}\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_{\nabla^2\Phi_{t+1}(\boldsymbol{y}_t)} \le 3\|\widetilde{\boldsymbol{g}}_t\|^2_{\nabla^{-2}\Phi_{t+1}(\boldsymbol{y}_t)}.$$

Thus, by summing this inequality for $t = 1, \dots, T$, we get that

$$\sum_{t=1}^T \langle\widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{u}_{t+1}\rangle \le 3\sum_{t=1}^T \|\widetilde{\boldsymbol{g}}_t\|^2_{\nabla^{-2}\Phi_{t+1}(\boldsymbol{y}_t)} \le \frac{3d\log(d + T/d)}{\eta},$$

where the last inequality follows by Lemma D.4 and $\nabla^2\Phi_{t+1} \succeq \nabla^2\Phi_t$, for all $t \ge 1$. Combining this with (58), we get the desired bound. □

D.5. Regret of `Barrier-ONS` (Proof of Theorem 3.1)

Proof. First, the fact that $(\boldsymbol{u}_t) \subset \operatorname{int}\mathbb{B}(R)$ follows from Lemma D.7.

We now show (6). Fix $\boldsymbol{w} \in \operatorname{int}\mathbb{B}(R)$ and let $(\boldsymbol{w}_t)$ be the `FTRL` iterates in (5). We have

$$\begin{aligned}&\sum_{t=1}^T \left(\langle \boldsymbol{u}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle - \frac{\eta}{2}\langle \boldsymbol{u}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle^2\right)\\ &= \sum_{t=1}^T \left(\langle \boldsymbol{w}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle - \frac{\eta}{2}(\langle \boldsymbol{w}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle + \langle \boldsymbol{u}_t - \boldsymbol{w}_t, \widetilde{\boldsymbol{g}}_t\rangle)^2\right) + \sum_{t=1}^T \langle \boldsymbol{u}_t - \boldsymbol{w}_t, \widetilde{\boldsymbol{g}}_t\rangle,\\ &= \sum_{t=1}^T \left(\langle \boldsymbol{w}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle - \frac{\eta}{2}\langle \boldsymbol{w}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle^2\right) + \sum_{t=1}^T (1 - \eta\langle \boldsymbol{w}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle) \cdot \langle \boldsymbol{u}_t - \boldsymbol{w}_t, \widetilde{\boldsymbol{g}}_t\rangle - \frac{\eta}{2}\sum_{t=1}^T \langle \boldsymbol{u}_t - \boldsymbol{w}_t, \widetilde{\boldsymbol{g}}_t\rangle^2,\\ &\le \sum_{t=1}^T \left(\langle \boldsymbol{w}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle - \frac{\eta}{2}\langle \boldsymbol{w}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle^2\right) + \sum_{t=1}^T (1 - \eta\langle \boldsymbol{w}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle) \cdot \langle \boldsymbol{u}_t - \boldsymbol{w}_t, \widetilde{\boldsymbol{g}}_t\rangle,\\ &\le \sum_{t=1}^T \left(\langle \boldsymbol{w}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle - \frac{\eta}{2}\langle \boldsymbol{w}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle^2\right) + \frac{7}{5}\sum_{t=1}^T \|\boldsymbol{u}_t - \boldsymbol{w}_t\|_{\nabla^2\Phi_t(\boldsymbol{u}_t)}\|\widetilde{\boldsymbol{g}}_t\|_{\nabla^{-2}\Phi_t(\boldsymbol{u}_t)}, \qquad (60)\end{aligned}$$

where the last step follows by Hölder's inequality, and the facts that $(\widetilde{\boldsymbol{g}}_t) \subset \mathbb{B}(\widetilde{G})$, $(\boldsymbol{w}_t) \subset \mathbb{B}(R)$, and $\eta \le \frac{1}{5\widetilde{G}R}$ (which implies that $\eta|\langle \boldsymbol{w}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle| \le 2/5$, for all $t \in [T]$).

Now, by Lemma D.4, we have $\|\widetilde{\boldsymbol{g}}_t\|_{\nabla^2\Phi_t(\boldsymbol{u}_t)} \le \frac{\widetilde{G}R}{\sqrt{2\nu}}$ for all $t \in [T]$, and by Lemma D.7, we have

$$\sum_{t=1}^{T} \|\boldsymbol{u}_t - \boldsymbol{w}_t\|_{\nabla^2\Phi_t(\boldsymbol{u}_t)} \le \frac{3\sqrt{\nu}}{32} + \frac{6d\log(1+T/d)}{\eta\sqrt{\nu}}.$$

Therefore, we have

$$\begin{aligned}\sum_{t=1}^{T} \|\boldsymbol{u}_t - \boldsymbol{w}_t\|_{\nabla^2\Phi_t(\boldsymbol{u}_t)} \|\widetilde{\boldsymbol{g}}_t\|_{\nabla^{-2}\Phi_t(\boldsymbol{u}_t)} &\le \frac{3\widetilde{G}R}{32\sqrt{2}} + \frac{6\widetilde{G}Rd\log(1+T/d)}{\sqrt{2}\eta\nu}, \\ &\le \frac{3\widetilde{G}R}{32\sqrt{2}} + \frac{3d\log(1+T/d)}{7\eta}, \end{aligned} \tag{61}$$

where the last inequality follows by $\nu \ge 10\widetilde{G}R$. On the other hand, by Lemma 3.2, we have

$$\begin{aligned}\sum_{t=1}^{T}\left(\langle \boldsymbol{w}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle - \frac{\eta}{2}\langle \boldsymbol{w}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle^2\right) &\le \Psi(\boldsymbol{w}) - \Psi(\boldsymbol{0}) + \frac{3d\log(1+T/d)}{\eta}, \\ &= -\nu\log(1 - \tfrac{\|\boldsymbol{w}\|^2}{R^2}) + \frac{3d\log(1+T/d)}{\eta}. \end{aligned} \tag{62}$$

Plugging (61) and (62) into (60), we get

$$\sum_{t=1}^{T}\left(\langle \boldsymbol{u}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle - \frac{\eta}{2}\langle \boldsymbol{u}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t\rangle^2\right) \le \frac{21\widetilde{G}R}{160\sqrt{2}} - \nu\log(1 - \tfrac{\|\boldsymbol{w}\|^2}{R^2}) + \frac{18d\log(1+T/d)}{5\eta}.$$

Combining this with $\frac{21}{160\sqrt{2}} \le 1$ implies (6).

Computational cost. Now, we analyze the computational complexity of Algorithm 4, which is equivalent to Algorithm 2 in that both algorithms produce the same outputs. The most computationally expensive step in Algorithm 4 occurs in Line 16, where a full matrix inverse is required when $\boldsymbol{z}_t \neq \boldsymbol{z}_{t-1}$. However, by Lemma 3.1, the matrix inverse only needs to be computed at most $O(c^{-1}\sqrt{\frac{d}{\nu\eta}T\log(1+T/d)})$ times over T rounds. In the rounds where $\boldsymbol{z}_t = \boldsymbol{z}_{t-1}$, Algorithm 4 performs at most $O(m)$ matrix-vector multiplications (see Line 7-Line 11), which results in a cost of $O(md^2)$. Therefore, the total computational complexity of Algorithm 4 follows from the fact that $m = \mathfrak{c}\cdot\log_c(dT)$, where $\mathfrak{c}$ is a universal constant. $\square$

Appendix E. OCO Analysis: Proof of Theorem 4.1

This appendix provides the proof of Theorem 4.1. We begin in Appendix E.1 by proving the main reduction result in Lemma 4.1, which bounds the instantaneous regret of Algorithm 3 by that of its base algorithm $\mathcal{B}$. In Appendix E.2, we prove the regret bound for Algorithm 3 when $\mathcal{B}$ is set as the `Barrier-ONS`, prior to parameter tuning (Proposition 4.1). Finally, in Appendix E.3, we present the complete proof of Theorem 4.1.

E.1. OCO Reduction with Gauge Projections (Proof of Lemma 4.1)

Proof. Fix $t \in [T]$, and let S_t, $\boldsymbol{s}_t$, $\boldsymbol{u}_t, \widetilde{\boldsymbol{g}}_t$, and $\boldsymbol{w}_t$ be as in Algorithm 3. We first show that $\boldsymbol{w}_t \in \mathcal{K}$. By definition of $\boldsymbol{w}_t$, we have $\boldsymbol{w}_t = \frac{\boldsymbol{u}_t}{1+S_t}$. Therefore, by the homogeneity of the Gauge function (see Lemma G.1), we have

$$\begin{aligned}\gamma_{\mathcal{K}}(\boldsymbol{w}_t) &= \frac{\gamma_{\mathcal{K}}(\boldsymbol{u}_t)}{1+S_t}, \\ &\le \frac{1+S_{\mathcal{K}}(\boldsymbol{u}_t)}{1+S_t}, \quad \text{(since } S_{\mathcal{K}}(\boldsymbol{u}_t) = \max(0, \gamma_{\mathcal{K}}(\boldsymbol{u}_t) - 1) \text{ by Lemma 2.1)} \\ &\le 1, \end{aligned} \tag{63}$$

where the last inequality follows from $S_{\mathcal{K}}(\boldsymbol{u}_t) \le S_t$ by Lemma 2.2. Eq. (63) implies that $\boldsymbol{w}_t \in \mathcal{K}$ by definition of the Gauge function (see Definition 2.2).

We now prove that

$$\forall t \in [T], \forall \boldsymbol{w} \in \mathcal{K}, \quad \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w} \rangle \le \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w} \rangle + \frac{2GR}{T}. \tag{64}$$

For this, define the surrogate loss function ℓ_t:

$$\forall \boldsymbol{w} \in \mathbb{R}^d, \quad \ell_t(\boldsymbol{w}) \coloneqq \langle \boldsymbol{g}_t, \boldsymbol{w} \rangle - \mathbb{I}\{\langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle < 0\} \cdot \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle \cdot S_{\mathcal{K}}(\boldsymbol{w}).$$

Since the pair $(S_t, \boldsymbol{s}_t)$ is the output of $\texttt{GaugeDist}(\mathcal{K}, \boldsymbol{u}_t, \varepsilon, r)$ with $\varepsilon = 1/T$, we have by Lemma 2.2:

$$\forall \boldsymbol{u} \in \mathbb{R}^d, \quad S_{\mathcal{K}}(\boldsymbol{u}) \ge S_{\mathcal{K}}(\boldsymbol{u}_t) + (\boldsymbol{u} - \boldsymbol{u}_t)^\top \boldsymbol{s}_t - \frac{1}{T}. \tag{65}$$

Now, since $\boldsymbol{w}_t = \boldsymbol{u}_t/(1+S_t)$ (see Algorithm 3) and $S_t \ge S_{\mathcal{K}}(\boldsymbol{u}_t) \ge 0$ (by Lemma 2.2), we have that $-\mathbb{I}\{\langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle < 0\} \cdot \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle \ge 0$. And so, using (65) and the definition of ℓ_t, we get

$$\begin{aligned}\forall \boldsymbol{u} \in \mathbb{R}^d, \quad \ell_t(\boldsymbol{u}_t) - \ell_t(\boldsymbol{u}) &\le \langle \boldsymbol{g}_t - \mathbb{I}\{\langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle < 0\} \cdot \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle \cdot \boldsymbol{s}_t, \boldsymbol{u}_t - \boldsymbol{u} \rangle + |\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle| \cdot \frac{1}{T}, \\ &= \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{u} \rangle + |\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle| \cdot \frac{1}{T}, \quad \text{(by definition of } \widetilde{\boldsymbol{g}}_t \text{ in Algorithm 3)} \\ &\le \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{u} \rangle + \frac{GR}{T}, \end{aligned} \tag{66}$$

where the last inequality uses that $\boldsymbol{w}_t \in \mathcal{K} \subseteq \mathbb{B}(R)$ and $\|\boldsymbol{g}_t\| \le G$.

It remains to show that $\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{u} \rangle \le \ell_t(\boldsymbol{u}_t) - \ell_t(\boldsymbol{u})$, for all $\boldsymbol{u} \in \mathcal{K}$. First, note that for all $\boldsymbol{u} \in \mathcal{K}$, we have $S_{\mathcal{K}}(\boldsymbol{u}) = \max(0, \gamma_{\mathcal{K}}(\boldsymbol{u}) - 1) = 0$ (by Lemma 2.1 and the definition of the Gauge function), and so

$$\ell_t(\boldsymbol{u}) = \langle \boldsymbol{g}_t, \boldsymbol{u} \rangle, \quad \forall \boldsymbol{u} \in \mathcal{K}. \tag{67}$$

We will now compare $\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle$ to $\ell_t(\boldsymbol{u}_t)$ by considering cases. Suppose that $S_t = 0$. In this case, we have $\boldsymbol{w}_t = \boldsymbol{u}_t$ and so $\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle = \langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle = \ell_t(\boldsymbol{u}_t)$. Now suppose that $S_t > 0$ and $\langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle \ge 0$. In this case, since $\boldsymbol{w}_t = \frac{\boldsymbol{u}_t}{1+S_t}$, we immediately have

$$\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle \le \langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle = \ell_t(\boldsymbol{u}_t). \quad [\text{case where } \langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle \ge 0] \tag{68}$$

Now suppose that $S_t > 0$ and $\langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle < 0$. Again, using that $\boldsymbol{w}_t = \frac{\boldsymbol{u}_t}{1+S_t}$, we have

$$
\begin{aligned}
&\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle + \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle \cdot S_{\mathcal{K}}(\boldsymbol{u}_t) \\
&= \langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle \cdot \frac{1 + S_{\mathcal{K}}(\boldsymbol{u}_t)}{1 + S_t}, \\
&\leq \langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle \cdot \frac{1 + S_t - \frac{1}{T}}{1 + S_t}, \quad \text{(since } \langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle < 0 \text{ and } S_{\mathcal{K}}(\boldsymbol{u}_t) \geq S_t - \tfrac{1}{T} \text{ by Lemma 2.2)} \\
&\leq \langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle + |\langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle| \cdot \frac{1}{T}, \\
&\leq \langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle + \frac{GR}{T},
\end{aligned}
$$

where the last inequality follows from the fact that $\boldsymbol{u}_t \in \mathbb{B}(R)$ and $\|\boldsymbol{g}_t\| \leq G$. Rearranging this, we get

$$
\langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle - \frac{GR}{T} \leq \langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle - \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle \cdot S_{\mathcal{K}}(\boldsymbol{u}_t) = \ell_t(\boldsymbol{u}_t). \quad [\text{case where } \langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle < 0] \tag{69}
$$

By combining (66), (67), (68), and (69), we obtain

$$
\forall \boldsymbol{u} \in \mathcal{K}, \quad \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{u} \rangle - \frac{GR}{T} \leq \ell_t(\boldsymbol{u}_t) - \ell_t(\boldsymbol{u}) \leq \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{u} \rangle + \frac{GR}{T},
$$

which shows the inequality in (64).

Bounding the surrogate subgradients. It remains to bound $\|\widetilde{\boldsymbol{g}}_t\|$ in terms of $\|\boldsymbol{g}_t\|$. Using that $\widetilde{\boldsymbol{g}}_t = \boldsymbol{g}_t - \mathbb{I}\{\langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle < 0\} \cdot \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle \cdot \boldsymbol{s}_t$ and $\boldsymbol{w}_t = \frac{\boldsymbol{u}_t}{1+S_t}$, we have

$$
\begin{aligned}
\|\widetilde{\boldsymbol{g}}_t\| &= \|\boldsymbol{g}_t - \mathbb{I}\{\langle \boldsymbol{g}_t, \boldsymbol{u}_t \rangle < 0\} \cdot \langle \boldsymbol{g}_t, \boldsymbol{w}_t \rangle \cdot \boldsymbol{s}_t\|, \\
&\leq \|\boldsymbol{g}_t\| + \|\boldsymbol{g}_t\| \cdot \frac{\|\boldsymbol{u}_t\|}{1 + S_t} \cdot \|\boldsymbol{s}_t\|, \quad \text{(by the triangle inequality and Cauchy Schwarz)} \\
&\leq \|\boldsymbol{g}_t\| \cdot \left(1 + \frac{\|\boldsymbol{u}_t\|}{r}\right), \quad \text{(since } S_t \geq 0 \text{ and } \|\boldsymbol{s}_t\| \leq 1/r \text{ by Lemma 2.2)} \\
&\leq (1 + \kappa) \cdot \|\boldsymbol{g}_t\|,
\end{aligned}
$$

where the last step follows by the assumption that $\boldsymbol{u}_t \in \mathbb{B}(R)$. This completes the proof. □

E.2. OCO Regret Bound Pre-Tuning of Parameters (Proof of Proposition 4.1)

Proof. Fix $\boldsymbol{w} \in \operatorname{int} \mathcal{K}$. By Lemma 4.1, the sequence of loss vectors $(\widetilde{\boldsymbol{g}}_t)$ that the `Barrier-ONS` base algorithm receives satisfies $(\widetilde{\boldsymbol{g}}_t) \subset \mathbb{B}(\widetilde{G})$ with $\widetilde{G} = 2\kappa G$. Thus, by invoking the guarantee of `Barrier-ONS` in Theorem 3.1, we get $(\boldsymbol{u}_t) \subset \mathbb{B}(R)$ and

$$
\sum_{t=1}^{T} \left(\langle \boldsymbol{u}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t \rangle - \frac{\eta}{2} \langle \boldsymbol{u}_t - \boldsymbol{w}, \widetilde{\boldsymbol{g}}_t \rangle^2 \right) \leq \widetilde{G} R - \nu \log(1 - \tfrac{\|\boldsymbol{w}\|^2}{R^2}) + \frac{4d \log(1 + \frac{T}{d})}{\eta}, \tag{70}
$$

where we used that $18/5 \leq 4$. We now prove that

$$\sum_{t=1}^{T}\left(\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}\rangle - \frac{\eta}{2}\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}\rangle^2\right) \leq \sum_{t=1}^{T}\left(\langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w}\rangle - \frac{\eta}{2}\langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w}\rangle^2\right) + 3GR, \tag{71}$$

which together with (70) would complete the proof. Using that $(\boldsymbol{u}_t) \subset \mathbb{B}(R)$, $(\widetilde{\boldsymbol{g}}_t) \subset \mathbb{B}(2\kappa G)$, and Assumption 2.1, we obtain

$$\forall t \in [T], \forall \boldsymbol{u} \in \mathcal{K}, \quad |\langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{u}\rangle| \leq 4\kappa RG. \tag{72}$$

Combining this with the facts that:

- $\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}\rangle \leq \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w}\rangle + \frac{2GR}{T}$, for all $t \in [T]$ (by Lemma 4.1);
- $x \to x - \frac{\eta}{2}x^2$ in non-decreasing for all $x \leq \frac{1}{\eta}$ (we instantiate this with $x = \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}\rangle$ and $x = \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w}\rangle + \frac{2GR}{T}$); and
- $\eta \leq \frac{1}{10\kappa GR}$;

we get that for all $t \in [T]$

$$\begin{aligned}
&\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}\rangle - \frac{\eta}{2}\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}\rangle^2 \\
&\leq \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w}\rangle + \frac{2GR}{T} - \frac{\eta}{2}\langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w}\rangle^2 - \frac{2\eta GR}{T}\langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w}\rangle - \frac{2\eta G^2R^2}{T^2}, \\
&\leq \langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w}\rangle - \frac{\eta}{2}\langle \widetilde{\boldsymbol{g}}_t, \boldsymbol{u}_t - \boldsymbol{w}\rangle^2 + \frac{3GR}{T},
\end{aligned}$$

where the last step follows by (72) and $\eta \leq \frac{1}{10\kappa GR}$. Summing this over $t = 1, \dots, T$, we obtain (71). Combining (71) with (70) we get the desired result. □

E.3. Main OCO Regret Bound (Proof of Theorem 4.1)

We recall the choice of parameters in Theorem 4.1:

$$c = \tfrac{1}{2}, \quad \eta = \tfrac{1}{GR}\cdot \min\left(\tfrac{1}{10\kappa}, \sqrt{\tfrac{2d\log(1+\frac{T}{d})}{T}}\right), \quad \text{and} \quad \nu = GR\cdot\max\left(20\kappa d, \sqrt{\tfrac{dT}{\log(1+\frac{T}{d})}}\right), \tag{73}$$

Proof of Theorem 4.1. Fix $\boldsymbol{w} \in \mathcal{K}$ and let $\widetilde{\boldsymbol{w}} \coloneqq \boldsymbol{w}\cdot(1-1/T)$. Note that $\widetilde{\boldsymbol{w}} \in \operatorname{int}\mathcal{K}$. By Assumption 2.2 (boundness of $(\boldsymbol{g}_t)$) and Assumption 2.1 (boundness of $\mathcal{K}$), we have that

$$\begin{aligned}
\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}\rangle &\leq -\sum_{t=1}^{T}\frac{1}{T}\langle \boldsymbol{g}_t, \boldsymbol{w}\rangle + \sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}}\rangle, \\
&\leq GR + \sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}}\rangle.
\end{aligned} \tag{74}$$

Thus, it suffices bound the (linearized) regret relative to $\widetilde{\boldsymbol{w}}$. By instantiating the bound in Proposition 4.1 with comparator $\widetilde{\boldsymbol{w}} \in \operatorname{int} \mathcal{K}$, we get:

$$\begin{aligned}\sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}} \rangle &\leq \frac{\eta}{2} \sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}} \rangle^2 + 5\kappa G R - \nu \log(1 - \tfrac{\|\widetilde{\boldsymbol{w}}\|^2}{R^2}) + \frac{4d \log(1 + T/d)}{\eta}, \\ &\leq \frac{\eta}{2} \sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}} \rangle^2 + 5\kappa G R + \nu \log T + \frac{4d \log(1 + T/d)}{\eta}, \end{aligned} \tag{75}$$

where in the last inequality, we used that

$$\|\widetilde{\boldsymbol{w}}\| \leq \|\boldsymbol{w}\| \cdot \left(1 - \tfrac{1}{T}\right) \leq R \cdot \left(1 - \tfrac{1}{T}\right) \quad \text{and} \quad - \log\left(1 - \left(1 - \tfrac{1}{T}\right)^2\right) = - \log\left(\tfrac{2}{T} - \tfrac{1}{T^2}\right) \leq \log T.$$

Now, using that $(\boldsymbol{g}_t) \subset \mathbb{B}(G)$, $(\boldsymbol{w}_t) \subset \mathbb{B}(R)$, and Assumption 2.1 ($\mathcal{K} \subseteq \mathbb{B}(R)$), we have $|\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}} \rangle| \leq 2GR$, for all $t \in [T]$. Combining this with (75), we get

$$\sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}} \rangle \leq 2\eta G^2 R^2 T + 5\kappa G R + \nu \log T + \frac{4d \log(1 + \frac{T}{d})}{\eta}. \tag{76}$$

We now use (76) to show the desired result. First, note that the optimal tuning of η in (76) is given by

$$\eta^\star = \sqrt{\frac{2d \log(1 + \frac{T}{d})}{T G^2 R^2}}.$$

We now consider cases.

Case where $\eta^\star \leq \frac{1}{10\kappa G R}$. First, note that this implies that $\eta = \eta^\star$. Now, using that $\eta^\star \leq \frac{1}{10\kappa G R}$ and the expression of $\eta^\star$, we have that

$$10\sqrt{2} \cdot \kappa \leq \sqrt{\frac{T}{d \log(1 + \frac{T}{d})}}.$$

This implies that $\nu \leq \sqrt{\frac{2dT}{d\log(1+\frac{T}{d})}}$ (see definition of ν in (73)). Using this together with $\eta = \eta^\star$ and (76) implies

$$\begin{aligned}(\text{case } \eta^\star \leq \tfrac{1}{10\kappa G R}) \quad \sum_{t=1}^T \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}} \rangle &\leq 4GR\sqrt{2dT \log\left(1 + \tfrac{T}{d}\right)} + 5\kappa G R + GR\sqrt{\frac{2dT}{\log(1 + \frac{T}{d})}} \cdot \log T, \\ &\leq 5GR\sqrt{2dT \log\left(1 + \tfrac{T}{d}\right)} + 5\kappa G R. \end{aligned} \tag{77}$$

Case where $\eta^\star \geq \frac{1}{10\kappa G R}$. In this case, we have $\eta = \frac{1}{10\kappa G R}$. Now, using that $\eta^\star \geq \frac{1}{10\kappa G R}$ and the expression of $\eta^\star$, we have that

$$10\sqrt{2} \cdot \kappa \geq \sqrt{\frac{T}{d \log(1 + \frac{T}{d})}}. \tag{78}$$

This implies that $\nu \le 20GR\kappa d$ (see definition of ν in (73)). Plugging this and $\eta = \frac{1}{10\kappa GR}$ into (76), we get

$$\begin{aligned}
(\text{case } \eta^\star \ge \tfrac{1}{10\kappa GR}) \quad \sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}} \rangle &\le \frac{TGR}{5\kappa} + 5\kappa GR + 20GR\kappa d \log T + 40 d\kappa GR \log(1 + \tfrac{T}{d}), \\
&\le \frac{TGR}{5\kappa} + 65 d\kappa GR \log(1 + \tfrac{T}{d}), \\
&\le 2GR\sqrt{2dT \log(1 + \tfrac{T}{d})} + 65GR\kappa d \log(1 + \tfrac{T}{d}), \qquad (79)
\end{aligned}$$

where the last inequality follows by (78). Thus, combining (77) and (79), we get

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}} \rangle \le 5GR\sqrt{2dT \log(1 + \tfrac{T}{d})} + 65GR\kappa d \log(1 + \tfrac{T}{d}).$$

Using this together with (74) implies the desired result.

Computational cost. By Theorem 3.1 and the choice of $c = 1/2$, the computational cost of the `Barrier-ONS` base algorithm with Algorithm 3 is bounded by $\widetilde{O}\left(d^2 T + d^\omega \sqrt{\frac{dT}{\nu\eta}}\right)$. Now, by the choice of η and ν in (73), we have $\eta\nu \ge d$. This implies that the computation of the `Barrier-ONS` base algorithm is bounded by

$$\widetilde{O}\left(d^2 T + d^\omega \sqrt{T}\right).$$

Now, in addition to the computational cost of the `Barrier-ONS` base algorithm, Algorithm 3 incurs $O(d) + C_{\texttt{GaugeDist}}(\mathcal{K})$ per round, where $C_{\texttt{GaugeDist}}(\mathcal{K})$ is the cost of one call to the `GaugeDist` base algorithm (Algorithm 1) for approximating the gauge distance $S_{\mathcal{K}}$ and its the subgradients. By Lemma 2.2, we have

$$C_{\texttt{GaugeDist}}(\mathcal{K}) \le \widetilde{O}(1) \cdot C_{\text{sep}}(\mathcal{K}).$$

□

Appendix F. Stochastic Convex Optimization (Proof of Theorem 5.1)

Proof. Let $\boldsymbol{w}^\star \in \arg\min_{\boldsymbol{u} \in \mathcal{K}} f(\boldsymbol{u})$. Further, let $\widetilde{\boldsymbol{w}}^\star \coloneqq \boldsymbol{w}^\star \cdot (1 - T^{-1})$. Note that $\widetilde{\boldsymbol{w}}^\star \in \operatorname{int} \mathcal{K}$. By Assumption 5.1 (boundness of $(\boldsymbol{g}_t)$), we have that

$$\begin{aligned}
\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}^\star \rangle &\le -\sum_{t=1}^{T} \frac{1}{T} \langle \boldsymbol{g}_t, \boldsymbol{w}^\star \rangle + \sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}}^\star \rangle, \\
&\le GR + \sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}}^\star \rangle. \qquad (80)
\end{aligned}$$

Now, using Jensen's inequality, we get

$$\begin{aligned}
\mathbb{E}[f(\widehat{\boldsymbol{w}}_T)] - f(\boldsymbol{w}^\star) &\leq \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} f(\boldsymbol{w}_t) - f(\boldsymbol{w}^\star)\right], \\
&\leq \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \langle \bar{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{w}^\star \rangle\right], \quad \text{(by convexity and } \bar{\boldsymbol{g}}_t \in \partial f(\boldsymbol{w}_t)) \\
&= \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}^\star \rangle\right]. \quad (\mathbb{E}[\boldsymbol{\xi}_t] = \mathbf{0} \text{ by Assumption 5.1}) \qquad (81)
\end{aligned}$$

Now, by instantiating the bound in Proposition 4.1 with comparator $\widetilde{\boldsymbol{w}} \in \operatorname{int} \mathcal{K}$ and parameters (η, ν, c) as in (14), we get:

$$\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}}^\star \rangle \leq \frac{\eta}{2}\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}}^\star \rangle^2 + 5\kappa G R + \nu \log T + \frac{d \log(1 + \frac{T}{d})}{\eta}, \qquad (82)$$

where in the last inequality, we used that

$$\|\widetilde{\boldsymbol{w}}^\star\| \leq \|\boldsymbol{w}\| \cdot \left(1 - \tfrac{1}{T}\right) \leq R \cdot \left(1 - \tfrac{1}{T}\right) \quad \text{and} \quad -\log\left(1 - \left(1 - \tfrac{1}{T}\right)^2\right) = -\log\left(\tfrac{2}{T} - \tfrac{1}{T^2}\right) \leq \log T.$$

Now, by Assumption 5.1 (in particular, the fact that $\boldsymbol{g}_t = \bar{\boldsymbol{g}}_t + \boldsymbol{\xi}_t$) together with the fact that $(a+b)^2 \leq 2a^2 + 2b^2$ and $\boldsymbol{w}_t, \widetilde{\boldsymbol{w}}^\star \in \mathbb{B}(R)$, we have

$$\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}}^\star \rangle^2 \leq 2\langle \bar{\boldsymbol{g}}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}}^\star \rangle^2 + 8R^2 \|\boldsymbol{\xi}_t\|^2. \qquad (83)$$

On the other hand, by definition of $\widetilde{\boldsymbol{w}}^\star$ and the facts that $\boldsymbol{w}, \boldsymbol{w}_1, \boldsymbol{w}_2, \dots \in \mathbb{B}(R)$, we have for all $t \in [T]$:

$$\begin{aligned}
2GR \geq \langle \bar{\boldsymbol{g}}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}}^\star \rangle &\geq \langle \bar{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{w}^\star \rangle - \frac{GR}{T}, \\
&\geq f(\boldsymbol{w}_t) - f(\boldsymbol{w}^\star) - \frac{GR}{T}, \quad \text{(by convexity of } f \text{ and } \bar{\boldsymbol{g}}_t \in \partial f(\boldsymbol{w}_t)) \\
&\geq -\frac{GR}{T}, \qquad (84)
\end{aligned}$$

where the last inequality follows by the fact that $\boldsymbol{w}_t \in \mathcal{K}$ and that $\boldsymbol{w}^\star$ is the minimizer of f within $\mathcal{K}$. Note that (84) implies that for all $t \in [T]$,

$$|\langle \bar{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{w}^\star \rangle| \leq \langle \bar{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{w}^\star \rangle + \frac{2GR}{T}. \qquad (85)$$

Picking up from (83), we get

$$\begin{aligned}
\sum_{t=1}^{T} \langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}^\star \rangle^2 &\leq 2\sum_{t=1}^{T} \langle \bar{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{w}^\star \rangle^2 + 8R^2 \sum_{t=1}^{T} \|\boldsymbol{\xi}_t\|^2, \\
&\leq 4GR \sum_{t=1}^{T} |\langle \bar{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{w}^\star \rangle| + 8R^2 \sum_{t=1}^{T} \|\boldsymbol{\xi}_t\|^2, \quad \text{(by the left-hand side inequality in (84))} \\
&\leq 4GR \sum_{t=1}^{T} \langle \bar{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{w}^\star \rangle + 8G^2R^2 + 8R^2 \sum_{t=1}^{T} \|\boldsymbol{\xi}_t\|^2, \quad \text{(by (85))}
\end{aligned}$$

Plugging this into (82) and rearranging, we get

$$\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}}^\star\rangle - 2GR\eta\sum_{t=1}^{T}\langle \bar{\boldsymbol{g}}_t, \boldsymbol{w}_t - \boldsymbol{w}^\star\rangle \leq 4\eta G^2R^2 + 4\eta R^2\sum_{t=1}^{T}\|\boldsymbol{\xi}_t\|^2 + 5\kappa GR + \nu\log T + \frac{d\log(1+\frac{T}{d})}{\eta}.$$

Taking the expectation on both sides and using that $\mathbb{E}[\boldsymbol{g}_t] = \bar{\boldsymbol{g}}_t$ and $\mathbb{E}[\|\boldsymbol{\xi}_t\|^2] \leq \sigma^2$, we get

$$\begin{aligned}
&4\eta R^2T\sigma^2 + 4\eta G^2R^2 + 5\kappa GR + \nu\log T + \frac{d\log(1+\frac{T}{d})}{\eta}\\
&\geq \mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \widetilde{\boldsymbol{w}}^\star\rangle\right] - 2GR\eta\cdot\mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}^\star\rangle\right],\\
&\geq (1-2GR\eta)\cdot\mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{g}_t, \boldsymbol{w}_t - \boldsymbol{w}^\star\rangle\right] - GR, \quad \text{(by (80))}\\
&\geq \frac{T}{2}\cdot(\mathbb{E}[f(\widehat{\boldsymbol{w}}_T)] - f(\boldsymbol{w}^\star)),
\end{aligned}$$

where the last inequality follows by the fact that $\eta \leq \frac{1}{4GR}$ and (81). Now, dividing by $\frac{T}{2}$ on both sides and rearranging, we get

$$\begin{aligned}
\mathbb{E}[f(\widehat{\boldsymbol{w}}_T)] - f(\boldsymbol{w}^\star) &\leq 8\eta R^2\sigma^2 + \frac{8\eta G^2R^2}{T} + \frac{12\kappa GR}{T} + \frac{2\nu\log T}{T} + \frac{2d\log(1+\frac{T}{d})}{\eta T},\\
&\leq 8\eta R^2\sigma^2 + \frac{2d\log(1+\frac{T}{d})}{\eta T} + \frac{14\kappa GR}{T} + \frac{2\nu\log T}{T}, \qquad (86)
\end{aligned}$$

where the last inequality follows by $\eta \leq \frac{1}{10\kappa GR}$. Note that the optimal tuning of η in (86) is given by

$$\eta^\star = \sqrt{\frac{d\log(1+\frac{T}{d})}{4R^2\sigma^2T}}.$$

We now consider cases.

Case where $\eta^\star \leq \frac{1}{10\kappa GR}$. First, note that this implies that $\eta = \eta^\star$. Now, using that $\eta^\star \leq \frac{1}{10\kappa GR}$ and the expression of $\eta^\star$, we have that

$$5G\kappa \leq \sigma\sqrt{\frac{T}{d\log(1+\frac{T}{d})}}.$$

This implies that $\nu \leq 4\sigma\sqrt{\frac{dT}{\log(1+\frac{T}{d})}}$ (see definition of ν in (14)). Using this together with $\eta = \eta^\star$ and (86) implies that

$$\begin{aligned}
(\text{case } \eta^\star \leq \tfrac{1}{10\kappa GR}) \quad \mathbb{E}[f(\widehat{\boldsymbol{w}}_T)] - f(\boldsymbol{w}^\star) &\leq 8R\sigma\cdot\sqrt{\frac{d}{T}} + \frac{14\kappa GR}{T} + 8R\sigma\sqrt{\frac{d}{T\log(1+\frac{T}{d})}}\cdot\log T,\\
&\leq 16R\sigma\cdot\sqrt{\frac{d\log(1+\frac{T}{d})}{T}} + \frac{14\kappa GR}{T}. \qquad (87)
\end{aligned}$$

Case where $\eta^\star \geq \frac{1}{10\kappa GR}$. In this case, we have $\eta = \frac{1}{10\kappa GR}$. Now, using that $\eta^\star \geq \frac{1}{10\kappa GR}$ and the expression of $\eta^\star$, we have

$$5G\kappa \geq \sigma\sqrt{\frac{T}{d\log(1+\frac{T}{d})}}. \tag{88}$$

This implies that $\nu \leq 20GR\kappa d$ (see definition of ν in (14)). Plugging this and $\eta = \frac{1}{10\kappa GR}$ into (86), we get

$$\begin{aligned}
(\text{case } \eta^\star \geq \tfrac{1}{10\kappa GR}) \quad \mathbb{E}[f(\widehat{\boldsymbol{w}}_T)] - f(\boldsymbol{w}^\star) &\leq \frac{4R\sigma^2}{5\kappa G} + \frac{20GR\kappa d\log(1+\frac{T}{d})}{T} + \frac{14\kappa GR}{T} + \frac{40GR\kappa d\log T}{T}, \\
&\leq \frac{4R\sigma^2}{5\kappa G} + \frac{74GR\kappa d\log(1+\frac{T}{d})}{T}, \\
&\leq 4R\sigma\sqrt{\frac{d\log(1+\frac{T}{d})}{T}} + \frac{74GR\kappa d\log(1+\frac{T}{d})}{T}, \qquad (89)
\end{aligned}$$

where the last inequality follows by (88). Thus, combining (87) and (89), we get

$$\mathbb{E}[f(\widehat{\boldsymbol{w}}_T)] - f(\boldsymbol{w}^\star) \leq 16R\sigma\sqrt{\frac{d\log(1+\frac{T}{d})}{T}} + \frac{74GR\kappa d\log(1+\frac{T}{d})}{T}.$$

This proves the desired convergence rate.

Computational cost. By Theorem 3.1 and the choice of $c = 1/2$, the computational cost of the `Barrier-ONS` subroutine with Algorithm 3 is bounded by $\widetilde{O}\left(d^2T + d^\omega\sqrt{\frac{dT}{\nu\eta}}\right)$. Now, by the choice of η and ν in (73), we have $\eta\nu \geq d$. This implies that the computation of the `Barrier-ONS` subroutine is bounded by

$$\widetilde{O}\left(d^2T + d^\omega\sqrt{T}\right).$$

Now, in addition to the computational cost of the `Barrier-ONS` subroutine, Algorithm 3 incurs $O(d) + C_{\texttt{GaugeDist}}(\mathcal{K})$ per round, where $C_{\texttt{GaugeDist}}(\mathcal{K})$ is the cost of one call to the `GaugeDist` subroutine (Algorithm 1) for approximating the gauge distance $S_\mathcal{K}$ and its the subgradients. By Lemma 2.2, we have

$$C_{\texttt{GaugeDist}}(\mathcal{K}) \leq \widetilde{O}(1)\cdot C_{\text{sep}}(\mathcal{K}).$$

This implies the desired computational cost. □

Appendix G. Computing the Gauge Distance (Proof of Lemma 2.2)

For the proof of Lemma 2.2, we need the following properties of the Gauge function (see e.g. Molinaro (2020) for a proof).

Lemma G.1. *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 $\mathbb{B}(r) \subseteq \mathcal{C} \subseteq \mathbb{B}(R)$. Then, the following properties hold:*

a. $\gamma_{\mathcal{C}}(\boldsymbol{w}) = \sigma_{\mathcal{C}^\circ}(\boldsymbol{w}) = \sup_{\boldsymbol{x}\in\mathcal{C}^\circ} \boldsymbol{x}^\top \boldsymbol{w}$ *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* $\mathbb{B}(1/R) \subseteq \mathcal{C}^\circ \subseteq \mathbb{B}(1/r)$.

With this, we now prove Lemma 2.2.

Proof of Lemma 2.2. Fix $\boldsymbol{w} \in \mathbb{R}^d$. We consider cases. If $\boldsymbol{w} \in \mathcal{C}$, then the 'if' condition in Line 2 of Algorithm 1 evaluates to 'true', and so the algorithm returns the pair $(S, \boldsymbol{s}) = (0, \boldsymbol{0})$. Since $\boldsymbol{w} \in \mathcal{C}$, we have $\gamma_{\mathcal{C}}(\boldsymbol{w}) \leq 1$, and so by Lemma 2.1, we have for all $\boldsymbol{u} \in \mathbb{R}^d$:

$$S_{\mathcal{C}}(\boldsymbol{u}) = \max(0, \gamma_{\mathcal{C}}(\boldsymbol{u}) - 1) \geq 0 = \max(0, \gamma_{\mathcal{C}}(\boldsymbol{w}) - 1) = S_{\mathcal{C}}(\boldsymbol{w}).$$

This implies the desired result since $\boldsymbol{0} \in \partial S_{\mathcal{C}}(\boldsymbol{w})$ by Lemma 2.1.

Now, consider the case where $\boldsymbol{w} \notin \mathcal{C}$, and let α, β, μ, $\boldsymbol{v}$, and $\boldsymbol{s}$ be as in Algorithm 1 when the algorithm returns. Then, by design, when Algorithm 1 returns, we have

$$\alpha\boldsymbol{w} \in \mathcal{C}, \quad \beta\boldsymbol{w} \notin \mathcal{C}, \quad \text{and} \quad |\beta - \alpha| \leq \frac{r^2\varepsilon}{2\|\boldsymbol{w}\|^2}. \tag{90}$$

Since $\gamma_{\mathcal{C}}(\boldsymbol{w}) = \inf\{\lambda > 0 \mid \boldsymbol{w} \in \lambda\mathcal{C}\}$, we have that

$$\frac{1}{\beta} \leq \gamma_{\mathcal{C}}(\boldsymbol{w}) \leq \frac{1}{\alpha}. \tag{91}$$

Now, since $\gamma_{\mathcal{C}}(\boldsymbol{w}) \leq \|\boldsymbol{w}\|/r$ (by Lemma G.1.c), the left-hand side inequality in (91) implies that

$$\beta \geq \frac{r}{\|\boldsymbol{w}\|}. \tag{92}$$

Note that $\|\boldsymbol{w}\| > 0$ since $\boldsymbol{w} \notin \mathcal{C}$ and $\mathbb{B}(r) \subseteq \mathcal{C}$. Using (92) and the fact that $|\beta - \alpha| \leq \frac{r^2\varepsilon}{2\|\boldsymbol{w}\|^2}$ (see (90)), we have

$$\begin{aligned}
\frac{1}{\alpha} &\leq \frac{1}{\beta - \frac{r^2\varepsilon}{2\|\boldsymbol{w}\|^2}}, \\
&\leq \frac{1}{\beta} + \frac{r^2\varepsilon}{\beta^2\|\boldsymbol{w}\|^2}, \quad \text{(see below)} \qquad (93)\\
&\leq \frac{1}{\beta} + \varepsilon, \quad \text{(by (92))} \qquad (94)
\end{aligned}$$

where (93) follows by (92) and the fact that $\frac{1}{1-x} \leq 1 + 2x$, for all $x \leq \frac{1}{2}$; we instantiate the latter with $x = \frac{r^2\varepsilon}{2\beta\|\boldsymbol{w}\|^2}$ which satisfies $x \leq 1/2$ since $\|\boldsymbol{w}\| \geq r$ (because $\boldsymbol{w} \notin \mathcal{C}$ and $\mathbb{B}(r) \subseteq \mathcal{C}$). Combining (94) with (91) and using that $S = \alpha^{-1} - 1$ (see Algorithm 1), we get

$$\gamma_{\mathcal{C}}(\boldsymbol{w}) \leq S + 1 \leq \gamma_{\mathcal{C}}(\boldsymbol{w}) + \varepsilon.$$

This together with the facts that $S_{\mathcal{C}}(\boldsymbol{w}) = \max(0, \gamma_{\mathcal{C}}(\boldsymbol{w}) - 1)$ and $\gamma_{\mathcal{C}}(\boldsymbol{w}) \geq 1$ (since $\boldsymbol{w} \notin \mathcal{C}$) implies that $S_{\mathcal{C}}(\boldsymbol{w}) \leq S \leq S_{\mathcal{C}}(\boldsymbol{w}) + \varepsilon$, as desired.

Approximate subgradient. We now show that the second output $\boldsymbol{s}$ of Algorithm 1 is an approximate subgradient of the gauge distance function at $\boldsymbol{w}$.

Since $\boldsymbol{v}$ is the separating hyperplane returned by the call to $\operatorname{Sep}_{\mathcal{C}}(\mu\boldsymbol{w})$, we have that

$$\forall \boldsymbol{u} \in \mathcal{C}, \quad \boldsymbol{u}^\top \boldsymbol{v} \leq \mu \boldsymbol{w}^\top \boldsymbol{v}.$$

Thus, since the vector $\boldsymbol{s}$ returned by Algorithm 1 satisfies $\boldsymbol{s} = \frac{\boldsymbol{v}}{\beta \cdot \boldsymbol{w}^\top \boldsymbol{v}}$ and $\mu = \frac{\alpha+\beta}{2} \leq \beta$, we have

$$\forall \boldsymbol{u} \in \mathcal{C}, \quad \boldsymbol{u}^\top \boldsymbol{s} \leq \mu \boldsymbol{w}^\top \boldsymbol{s} \leq 1. \tag{95}$$

This implies that $\boldsymbol{s} \in \mathcal{C}^\circ$ (by definition of the polar set) and so by Lemma G.1.a, this implies that

$$\forall \boldsymbol{u} \in \mathbb{R}^d, \quad \boldsymbol{s}^\top \boldsymbol{u} \leq \sup_{\boldsymbol{x} \in \mathcal{C}^\circ} \boldsymbol{x}^\top \boldsymbol{u} = \gamma_{\mathcal{C}}(\boldsymbol{u}). \tag{96}$$

On the other hand, combining (94) with (91), we get

$$\gamma_{\mathcal{C}}(\boldsymbol{w}) - \varepsilon \leq \frac{1}{\beta} = \boldsymbol{s}^\top \boldsymbol{w}, \tag{97}$$

where the equality uses the expression of $\boldsymbol{s}$. Combining (96) and (97) implies that

$$\forall \boldsymbol{u} \in \mathbb{R}^d, \quad \boldsymbol{s}^\top (\boldsymbol{u} - \boldsymbol{w}) + \gamma_{\mathcal{C}}(\boldsymbol{w}) - \varepsilon \leq \gamma_{\mathcal{C}}(\boldsymbol{u}).$$

Thus, subtracting 1 from both sides and using that $S_{\mathcal{C}}(\boldsymbol{w}) = \gamma_{\mathcal{C}}(\boldsymbol{w}) - 1$ (since $\boldsymbol{w} \notin \mathcal{C}$), we get that

$$\begin{aligned}
\forall \boldsymbol{u} \in \mathbb{R}^d, \quad \boldsymbol{s}^\top (\boldsymbol{u} - \boldsymbol{w}) + S_{\mathcal{C}}(\boldsymbol{w}) - \varepsilon &\leq \gamma_{\mathcal{C}}(\boldsymbol{u}) - 1, \\
&\leq \max(0, \gamma_{\mathcal{C}}(\boldsymbol{u}) - 1), \\
&= S_{\mathcal{C}}(\boldsymbol{u}).
\end{aligned}$$

This shows the inequality on the right-hand side of (2). Now, as mentioned earlier, (95) implies that $\boldsymbol{s} \in \mathcal{C}^\circ$. And since $\mathbb{B}(r) \subseteq \mathcal{C}$, we have $\mathcal{C}^\circ \subseteq \mathbb{B}(1/r)$ by Lemma G.1. Therefore, $\|\boldsymbol{s}\| \leq 1/r$.

Number of oracle calls. The number of oracle calls is bounded by the number of iterations of the 'while' loop in Line 6. Since Algorithm 1 implements a bisection, the number of iteration is at most $\log_2(\frac{4\|\boldsymbol{w}\|^2}{r^2 \varepsilon})$. □