Proceedings of Machine Learning Research vol 291:1–41, 2025 38th Annual Conference on Learning Theory

Computing Optimal Regularizers for Online Linear Optimization

Khashayar Gatmiry GATMIRY@MIT.EDU
MIT

Jon Schneider JSCHNEI@GOOGLE.COM
Google Research

Stefanie Jegelka STEFJE@CSAIL.MIT.EDU
MIT

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Follow-the-Regularized-Leader (FTRL) algorithms are a popular class of learning algorithms for online linear optimization (OLO) that guarantee sub-linear regret. However, the choice of regularizer can significantly impact dimension-dependent factors in the regret bound. We present an algorithm that takes as input convex and symmetric action sets and loss sets for a specific OLO instance, and outputs a regularizer such that running FTRL with this regularizer guarantees regret within a universal constant factor of the best possible regret bound. In particular, for any choice of (convex, symmetric) action set and loss set we prove that there exists an instantiation of FTRL that achieves regret within a constant factor of the best possible learning algorithm, strengthening the universality result of Srebro et al., 2011.

Our algorithm requires preprocessing time and space exponential in the dimension d of the OLO instance, but can be run efficiently online assuming a membership and linear optimization oracle for the action and loss sets, respectively (and is fully polynomial time for the case of constant dimension d). We complement this with a lower bound showing that even deciding whether a given regularizer is α-strongly-convex with respect to a given norm is NP-hard.

Keywords: List of keywords

1. Introduction

Online Linear Optimization (OLO) is one of the most fundamental problems in the theory of online learning. Here, a learner must repeatedly (for T rounds) select an action x_t from some bounded convex action set $\mathcal{X}$. Simultaneously, an adversary selects a linear loss function ℓ_t from a bounded convex loss set $\mathcal{L}$, and the learner receives loss $\langle x_t, \ell_t \rangle$. The learner would like to minimize their total loss, and more specifically minimize their *regret*: the gap between their total loss and the loss of the best fixed action $x^* \in \mathcal{X}$ in hindsight.

By choosing the action set $\mathcal{X}$ and loss set $\mathcal{L}$ appropriately, online linear optimization captures many other learning-theoretic problems of interest. For example, when $\mathcal{X} = \Delta_d$ (distributions over $\{1, 2, \ldots, d\}$) and $\mathcal{L} = [0, 1]^d$, this captures the classical problem of *learning with experts*. Similarly, when the loss set $\mathcal{L}$ is the ℓ_2 unit ball, this variant of OLO is the core subproblem involved in *online convex optimization* (specifically, of a Lipschitz function with domain $\mathcal{X}$). Even more generally, the works of Gordon et al. (2008) and Abernethy et al. (2011) demonstrate how to reduce the problems of linear ϕ-regret minimization (including swap regret minimization) and Blackwell approachability to different instances of OLO. These problems in turn have many applications extending past learning theory, from designing algorithms for computing correlated equilibria in repeated games, to producing calibrated forecasts, to constructing classifiers satisfying a variety of fairness criteria (Farina et al., 2021; Okoroafor et al., 2024; Chzhen et al., 2021).

For this reason, it is an extremely relevant problem to understand the best possible regret bounds achievable for different instances of OLO. Here, the state-of-the-art leaves something to be desired. It is well-known that learning algorithms such as Follow-The-Regularized-Leader (FTRL) achieve regret that scales with $O(\sqrt{T})$, and that this dependence on T is tight. However, the dependence of the optimal regret on the sets $\mathcal{X}$ and $\mathcal{L}$ (e.g., how the constant factor in the above regret bound depends on the dimension d of these sets) is in general poorly understood.

Moreover, FTRL is not a single algorithm, but a family of algorithms parametrized by a convex function $f : \mathcal{X} \to \mathbb{R}$ called the *regularizer*. The actual regret bounds achieved by FTRL can vary greatly depending how the choice of regularizer interacts with the geometry of $\mathcal{X}$ and $\mathcal{L}$. For example, running FTRL with the quadratic regularizer results in an $O(\sqrt{dT})$ regret algorithm for the learning with experts problem; however, running FTRL with the negative entropy regularizer results in an algorithm with a tight $O(\sqrt{T \log d})$ regret bound, with an exponential improvement in dimension over the quadratic choice of regularizer. On the other hand, there exist other instances (choices of $\mathcal{X}$ and $\mathcal{L}$) where the quadratic regularizer is optimal. Understanding what the optimal choice of regularizer is for a given instance of OLO is a major open problem.

1.1. Our contributions

For any action set $\mathcal{X}$ and loss set $\mathcal{L}$, the optimal possible regret bound (as T goes to infinity) scales as $\mathrm{Rate}(\mathcal{X}, \mathcal{L})\sqrt{T} + o(\sqrt{T})$, for some constant $\mathrm{Rate}(\mathcal{X}, \mathcal{L})$. Our goal in this paper is to design learning algorithms which approximately achieve this optimal regret bound. Specifically, we want to algorithmically construct learning algorithms with worst-case regret at most $C \cdot \mathrm{Rate}(\mathcal{X}, \mathcal{L})\sqrt{T}$ for some universal constant C that holds for any choice of action set and loss set in any dimension. For technical reasons, we restrict our attention in the following results to action sets $\mathcal{X}$ and loss sets $\mathcal{L}$ that are *centrally symmetric* – it is an interesting open direction to extend these results to fully general choices of $\mathcal{X}$ and $\mathcal{L}$.

We begin by showing that the optimal regret bound is achieved by some instantiation of Follow-The-Regularized-Leader. We do so by extending earlier work of Srebro et al. (2011) who, by analyzing the martingale types of Banach spaces, demonstrated that there is always an instance of FTRL which achieves regret $O(\mathrm{Rate}(\mathcal{X}, \mathcal{L})(\log T)\sqrt{T})$. In Theorem 18, we show that a more careful analysis of these martingale types allows us to remove this $\log T$ factor and prove that some variant of FTRL is within a universal constant of optimal.

Although the above argument proves the existence of a near-optimal instance of FTRL, the regularizer for this FTRL instance is not computable. The reason is that the definition of the barrier in Srebro et al. (2011) is a supremum of an infinite sum over infinite sequences of martingales that are defined by the action and loss set. Even worse, it is not clear if any truncation of this sum can approximate the value of the supremum with any given accuracy. In the remainder of the paper we study the following algorithmic question: given sets $\mathcal{X}$ and $\mathcal{L}$ (e.g., via oracle access), how can we compute a regularizer for these sets that guarantees the optimal regret up to universal constants when used with FTRL? Ultimately, we provide an algorithm that takes as input $\mathcal{X}$ and $\mathcal{L}$ (via standard oracle access to both sets), runs in time $\exp(O(d^2 \log d))$, and outputs a regularizer

f with the property that the worst-case regret of FTRL with f is at most a universal constant times $\mathrm{Rate}(\mathcal{X}, \mathcal{L})\sqrt{T}$ (Theorem 4). Note that this regularizer is not an approximation of the regularizer of Srebro et al. (2011), but rather is the solution of an optimization problem that we design.

The main technical ingredient in this algorithm is a new method for optimizing over the set of convex functions that are α-strongly convex with respect to a given norm. This is important for the above problem because one can show that for any regularizer f, the regret of running FTRL with that regularizer is bounded by $O(\sqrt{D\alpha T})$ if the range of f over $\mathcal{X}$ (the maximum value of f minus the minimum value of f) is at most D and if f is α-strongly-convex with respect to the norm induced by the dual set of the loss set $\mathcal{L}$. We can show that this regret-bound is constant-factor-optimal for the near-optimal variant of FTRL in Theorem 18, and hence it suffices to try to minimize $D\alpha$ over all convex functions f.

To do this, we first show that we can approximate any smooth convex function f as a maximum of several "quasi-quadratic" functions: quadratic functions g_{x_0} centered at some point x_0 with a small cubic term which guarantee that that the contribution of g_{x_0} to the Hessian of f decays far from x_0. Note that these are not just approximations of the values of f, but also the gradients and Hessians of f; in particular, if the original function was α-strongly-convex with respect to some norm, our approximation will be similarly strongly-convex.

By restricting our quasi-quadratic functions to be centered at points belonging to a (large but) finite discretization of $\mathcal{X}$, we demonstrate how to optimize over this set of approximations by solving a large convex program with variables for the values, gradients, and Hessians of the quasi-quadratic functions at each point in the discretization. Solving this convex program involves implementing a separation oracle to verify whether a specific approximation is α-strongly-convex with respect to an arbitrary norm.

As stated earlier, this approach takes time exponential in the dimension of the action and loss sets (although is completely independent of the time horizon T, and thus efficient for constant dimension d). We complement this with a lower bound showing that even verifying whether a regularizer f is α-strongly-convex at a specific point $x \in \mathcal{X}$ requires exponentially many oracle queries to $\mathcal{L}$.

2. Related Work

Applications of Online Linear Optimization. The problem of Online Linear Optimization (and its generalization, Online Convex Optimization) are central problems in the field of online learning – we refer the reader to Hazan et al. (2016) for a general-purpose introduction. Traditionally OLO is studied in the case where the action sets and loss sets are unit balls in a standard norm (e.g. the ℓ_1, ℓ_2, or ℓ_∞ norms). However, there are many motivating settings where we wish to minimize regret with less standard sets. Several authors (Takimoto and Warmuth, 2003; Kalai and Vempala, 2005; Koolen et al., 2010; Audibert et al., 2014) study variants of OLO where the action space has some combinatorial structure – for example, $\mathcal{X}$ could be the spanning tree polytope, or the polytope formed by all s-t paths in a graph. Minimizing external regret in extensive form games – one standard method for computing coarse correlated equilibria (Farina et al., 2020) – involves solving an instance of OLO where $\mathcal{X}$ is the sequence form polytope. Finally, as mentioned earlier, the work of Abernethy et al. (2011) and Gordon et al. (2008) allows us to translate any instance of Blackwell approachability or ϕ-regret minimziation to a (usually non-standard) instance of OLO.

Follow-The-Regularized-Leader and Mirror Descent. The Follow-The-Regularized-Leader algorithm can be thought of as a form of *mirror descent*, a family of first-order optimization algorithms that generalize gradient descent by using arbitrary distance-generating functions. Originally, mirror descent was proposed by Nemirovski and Yudin (1978) as an offline optimization algorithm with ℓ_p norm constraints and ℓ_q Lipschitz assumptions, and was shown to have minimax optimal query complexity. Sridharan and Tewari (2010) studied the optimality of mirror descent for online linear optimization when the action and loss vectors are in the unit ball of two Banach spaces dual to each other, proving the existence of a regularizer for mirror descent that almost achieves the minimax rate under an adaptive adversary. Later, Srebro et al. (2011) extended this approach to cases where the action and loss vectors come from independent convex balls in primal and dual Banach spaces. The existence of such strongly convex regularizers is also linked to the Burkholder method introduced by Foster et al. (2018) for more general online learning problems. In particular, the authors propose that given an online learning instance and a target regret bound, the existence of a Burkholder function for that instance guarantees the existence of a prediction strategy that achieves the desired regret. Notably, taking the dual of this Burkholder function for the online linear optimization (OLO) problem results in a strongly convex regularizer that can be used effectively with FTRL (Foster et al., 2018). We survey additional related work in Appendix A.

3. Preliminaries

3.1. Online linear optimization

We begin by defining the problem of *online linear optimization* (OLO). In this problem, every round t (for a total of T rounds) the learner must pick an action x_t from a convex action set $\mathcal{X} \subset \mathbb{R}^d$. The adversary then picks a loss vector ℓ_t from a convex loss set $\mathcal{L}$, after which the learner suffers loss $\langle x_t, \ell_t \rangle$ and observes the loss vector ℓ_t. The learner would like to minimize their total loss, and more specifically minimize their total *regret*: the gap between their loss and the loss of the best action in hindsight. Formally, given a sequence of learner actions $\mathbf{x} = (x_1, x_2, \dots, x_T)$ and losses $\boldsymbol{\ell} = (\ell_1, \ell_2, \dots, \ell_T)$, the regret of the learner is given by

$$\mathrm{Reg}(\mathbf{x}, \boldsymbol{\ell}) = \sum_{t=1}^{T} \langle x_t, \ell_t \rangle - \sum_{t=1}^{T} \min_{x^* \in \mathcal{X}} \langle x^*, \ell_t \rangle.$$

The learner chooses their actions according to some learning algorithm $\mathcal{A}$, which can be thought of as a function $\mathcal{A}$ mapping a sequence of losses $\boldsymbol{\ell} = (\ell_1, \ell_2, \dots, \ell_T)$ to a sequence of actions $\mathbf{x} = (x_1, x_2, \dots, x_T)$ in such a way that x_t depends only on the history of losses $\ell_1, \ell_2, \dots, \ell_{t-1}$ until round $t-1$. We define the T-round regret $\mathrm{Reg}_T(\mathcal{A})$ to be the worst-case regret suffered by algorithm $\mathcal{A}$ against an adversarially chosen sequence of losses, i.e., $\mathrm{Reg}_T(\mathcal{A}) = \sup_{\boldsymbol{\ell} \in \mathcal{L}^T} \mathrm{Reg}(\mathcal{A}(\boldsymbol{\ell}), \boldsymbol{\ell})$.

One of the fundamental results in online learning is that there exist algorithms $\mathcal{A}$ that guarantee $O(\sqrt{T})$ regret (e.g., online gradient descent), which is the best possible dependency one can hope for in terms of T. However, the optimal scaling factor in front of the $\sqrt{T}$ depends on the geometry of the action and loss sets $\mathcal{X}$ and $\mathcal{L}$ and is the primary focus of interest in this paper. To this end, define $\mathrm{Rate}(\mathcal{A}) = \limsup_{T\to\infty} \frac{1}{\sqrt{T}} \cdot \mathrm{Reg}_T(\mathcal{A})$ to be the worst-case scaling factor achieved by the algorithm $\mathcal{A}$, and $\mathrm{Rate}(\mathcal{X}, \mathcal{L}) = \inf_{\mathcal{A}} \mathrm{Rate}(\mathcal{A})$ to be the best possible scaling factor achieved by any algorithm for this action set and loss set. Our goal is to understand how to approximate $\mathrm{Rate}(\mathcal{X}, \mathcal{L})$ and design corresponding optimal algorithms for any choice of action set and loss set.

3.2. Regularizers and Follow-The-Regularized-Leader

One of the most popular classes of learning algorithms for online linear optimization is the class of follow-the-regularized-leader algorithms. *Follow-The-Regularized-Leader (FTRL)* is an algorithm parameterized by a convex function $f : \mathcal{X} \to \mathbb{R}$ (the "regularizer") and a learning rate $\eta > 0$ (which we will generally set equal to $1/\sqrt{T}$). At round t, it plays the action x_t given by

$$x_t = \arg\min_{x \in \mathcal{X}} \left(\eta f(x) + \sum_{s=1}^{t-1} \langle x, \ell_s \rangle \right). \tag{1}$$

Intuitively, FTRL always plays an action that is approximately the best response to the current empirical loss (with the regularizer preventing this action from overfitting too rapidly to the actions of the adversary). The class of FTRL algorithms contains many popular algorithms for special cases of online linear optimization, including online gradient descent and multiplicative weights.

It can be shown that as long as f is strongly convex, FTRL will incur $O(\sqrt{T})$ regret and thus have non-infinite rate – however, the value of $\mathrm{Rate}(\mathcal{X}, \mathcal{L})$ can depend significantly on the choice of f. For example, when $\mathcal{X} = \Delta_d$ and $\mathcal{L} = [0, 1]^d$ (the classic setting for *learning from experts*), it is known that:

- If we use the quadratic regularizer $f(x) = \|x\|^2$, the resulting rate of the FTRL algorithm is $\mathrm{Rate}(\mathcal{A}) = \Theta(\sqrt{d})$. (This corresponds to running online gradient descent).
- If we use the negative entropy regularizer $f(x) = \sum_i x_i \log x_i$, the resulting rate of the FTRL algorithm is $\mathrm{Rate}(\mathcal{A}) = \Theta(\sqrt{\log d})$. (This corresponds to running multiplicative weights / Hedge).

We will soon see that the optimal rate is achieved by some instantiation of FTRL (Theorem 33), and therefore much of our focus will be on computing a suitable regularizer f for a given pair of action set and loss set $(\mathcal{X}, \mathcal{L})$. To this end, it is useful to understand the guarantees the standard analysis of FTRL grants us for a specific choice of regularizer. Before we can state these, we will need to introduce some terminology regarding convex sets and their associated norms.

First, we will make the standard assumption in convex optimization that all of our convex sets are bounded and contain an open ball. In particular, we have the following assumption:

Assumption 1 *We assume the action and loss sets are symmetric*[1]*. We further assume they both include a ball of radius r and are included in a ball of radius R: $B(0, r) \subseteq \mathcal{X}, \mathcal{L} \subseteq B(0, R)$.*

The symmetry assumption allows us to define norms corresponding to $\mathcal{X}$ and $\mathcal{L}$. In general, the norm provided by a bounded symmetric convex set $\mathcal{C}$ is defined as follows:

Definition 1 *Given a bounded symmetric convex subset $\mathcal{C} \subseteq \mathbb{R}^d$, we define the natural norm $\|.\|_{\mathcal{C}}$ corresponding to $\mathcal{C}$ as*

$$\forall v \in \mathbb{R}^d, \|v\|_{\mathcal{C}} \triangleq \inf\{\alpha > 0,\ \tfrac{v}{\alpha} \in \mathcal{C}\}. \tag{2}$$

It is easy to check that $\|.\|_{\mathcal{C}}$ defined in equation 2 is a norm (Leonard and Lewis, 2015).

1. A convex set $S \subset \mathbb{R}^d$ is *(centrally) symmetric* if $x \in S$ implies that $-x \in S$.

Given a symmetric convex set $\mathcal{C}$, we can also define a norm on linear functionals over $\mathcal{C}$ by constructing the appropriate dual convex set.

Definition 2 *Given a symmetric convex set $\mathcal{C} \subseteq \mathbb{R}^d$, the dual set $\mathcal{C}^c$ is defined as $\mathcal{C}^c \triangleq \{x \in \mathbb{R}^d : \forall y \in \mathcal{C}, \langle x, y\rangle \leq 1\}$. Note that if $\mathcal{C}$ is symmetric, bounded, and full-dimensional, the dual set $\mathcal{C}^c$ is symmetric, bounded, and full-dimensional. The dual norm $\|v\|_{\mathcal{C}^c}$ is the norm corresponding to the dual set.*

We also need to define the notion of strong convexity with respect to an arbitrary norm $\|.\|_{\mathcal{C}}$:

Definition 3 *A convex function $f : \mathcal{X} \to \mathbb{R}$ is strongly-convex with respect to norm $\|.\|_{\mathcal{C}}$ if for every $x, y \in \mathcal{X}$ and every sub-gradient g of f at x: $f(y) \geq f(x) + \langle y - x, g\rangle + \frac{\alpha}{2}\|y - x\|_{\mathcal{C}}^2$.*

Now we are ready to state the standard regret bound for FTRL with regularizer f. As we can see, the regret bound depends on both the strong convexity of f with respect to the dual norm of $\mathcal{L}$, and the range of f over $\mathcal{X}$:

Fact 1 *[Theorem 5.2 in Hazan et al. (2016)] Let $\mathrm{FTRL}(f)$ be the FTRL algorithm initialized with regularizer f and learning rate $\eta = 1/\sqrt{T}$. If $0 \leq f(x) \leq C^2$ for all $x \in \mathcal{X}$ and f is α-strongly-convex with respect to $\mathcal{L}^c$ on $\mathcal{X}$ (see Definition 3), then $\mathrm{Reg}(\mathrm{FTRL}(f)) \leq O(C\sqrt{\alpha^{-1}T})$.*

3.3. Convex Optimization and Oracles

We will in general assume that we have *oracle access* (i.e., access to membership oracles, separation oracles, linear optimization oracles) to the sets $\mathcal{X}$ and $\mathcal{L}$. For a more comprehensive definition of these oracles, see Appendix D.

4. Main Result and Overview

Our main contribution is to propose an algorithm for computing a regularizer g such that running FTRL with g achieves the optimal regret of $O\left(\mathrm{Rate}(\mathcal{X}, \mathcal{L})\sqrt{T}\right)$ for the online linear optimization problem, as defined in Section 3.1. In particular, we state our main result in the following theorem.

Theorem 4 (Algorithmic optimal online linear optimization) *Given access to a linear optimization oracle for $\mathcal{L}$, which can minimize any linear function $c^\top x$ over $\mathcal{L}$ up to accuracy δ_{lin} in time $\mathrm{LinO}_{\mathcal{L}}(\delta_{\mathrm{lin}})$, there is a cutting-plane algorithm that runs in time $\left(\frac{dR}{r}\right)^{O(d^2)} \cdot \mathrm{LinO}_{\mathcal{L}}\left(\left(\frac{r}{dR}\right)^{\Theta(d)}\right)$ and calculates a regularizer g which satisfies*

1. $\sup_{x \in \mathcal{X}} |g| = O(\mathrm{Rate}(\mathcal{X}, \mathcal{L})^2)$,
2. *g is 1-strongly convex w.r.t $\|.\|_{\mathcal{L}^c}$.*

Furthermore, given access to a membership oracle to $\mathcal{X}$ and the regularizer g (which can be precomputed and summarized via a $\exp(O(d^2))$-dimensional vector as described in Section 7) there is a cutting-plane algorithm that runs FTRL with regularizer g with running time $O\left(d^2 \ln^{O(1)}(dRT)\right)$ per round and which guarantees regret $O(\mathrm{Rate}(\mathcal{X}, \mathcal{L})\sqrt{T})$.

The starting point of our proof of the above theorem is to demonstrate the existence of a regularizer that enables FTRL to achieve the optimal minimax regret, up to a constant factor.

Theorem 5 *There exists a regularizer f_0 so that running FTRL with f_0 yields a regret of* $\mathrm{Reg}(\mathbf{x}, \boldsymbol{\ell}) \leq O(\mathrm{Rate}(\mathcal{X}, \mathcal{L})\sqrt{T})$.

We prove Theorem 5 in Appendix B, where we eliminate the additional $\log(T)$ factor from the regret analysis of the regularizer in Srebro et al. (2011), proving that it achieves the optimal regret bound of $O\left(\mathrm{Rate}(\mathcal{X}, \mathcal{L})\sqrt{T}\right)$, up to universal constants. This improvement is made possible by a novel analytic estimate for the norm growth of certain martingales. In particular, we prove in Theorem 18 that the regularizer from Srebro et al. (2011) can be chosen to be 1-strongly convex with respect to $\|.\|_{\mathcal{L}^c}$ while being bounded by $O\left(\mathrm{Rate}(\mathcal{X}, \mathcal{L})^2\right)$ on the domain $\mathcal{X}$. Theorem 5 then follows from Theorem 18 and Fact 1.

This allows us to restrict our attention to the problem of finding the optimal regularizer over $\mathcal{X}$ which is 1-strongly-convex with respect to $\|.\|_{\mathcal{L}^c}$. To effectively do this optimization, it is important that the resulting regularizer has not only bounded values, but also bounded *gradients*. Note that this is not a priori achieved by the regularizers guaranteed to exist by Theorem 18, and in fact several optimal regularizers used in practice (e.g. the negative entropy regularizer) do have unbounded gradients. Nonetheless, in Section 5 and Appendix C, we demonstrate how to use Gaussian smoothing to obtain a new regularizer that (1) achieves the same optimal regret when used in FTRL, and (2) has smooth derivatives (Theorem 6).

Our next step is to show that we can effectively optimize over the space of smooth convex functions defined over $\mathcal{X}$. To do so, we show that given a near-optimal smooth regularizer f, we can approximate it using "quasi-quadratic" functions such that the resulting regularizer $\tilde{f}$ remains (1) $\alpha/2$ strongly convex with respect to $\|.\|_{\mathcal{L}^c}$, and (2) bounded by $O\left(\mathrm{Rate}(\mathcal{X}, \mathcal{L})^2\right)$ on $\mathcal{X}$. Notably, the set of quasi-quadratic functions (with a discretized set of centers) is finite-dimensional, and so the optimal regularizer can be encoded by a finite-dimensional vector $\tilde{\mathcal{I}}$. We carry this out in Section 6.

Finally, in Section E, we demonstrate how to optimize over this set by writing an explicit convex program such that $\tilde{f}$ is a feasible solution to this program, but also such that any feasible solution so that any feasible solution $\mathcal{I}$ from this set yields a regularizer $g^{(\mathcal{I})}$ with near optimal regret. Solving this convex program can be done via standard cutting-plane methods, except for one of the constraints that involves checking whether a candidate regularizer g is α-strongly-convex with respect to $\|.\|_{\mathcal{L}^c}$. In Section F, we demonstrate how to construct a separation oracle for this constraint, and finally establish the existence of this algorithm.

As seen in Theorem 4, computing and storing this optimal regularizer takes time that is exponential in the dimension of the problem. In Section 8, we establish a lower bound based on the result of Bhattiprolu et al. (2021) that even checking the strong convexity of the Euclidean norm squared regularizer with respect to $\|.\|_{\mathcal{L}^c}$ requires an exponential number of queries in the dimension.

5. A Smooth Optimal Regularizer

While Theorem 5 promises the existence of an ideal regularizer which achieves the optimal rate, this regularizer is not effective. To obtain a computable regularizer, our goal will be to search over a parametric family of functions. We use f_0 to prove the existence of a good regularizer in our search space. More accurately, we plan to accomplish this by approximating the regularizer f_0 as a maximum over several local approximations to the regularizer at a finite, discrete set $\mathcal{S}$ of N points in $\mathcal{X}$. By Fact 1, doing this requires showing that there exists such an approximation that (1) preserves the strong convexity of f_0, (2) is bounded by $O(\mathrm{Rate}(\mathcal{X}, \mathcal{L})^2)$ on $\mathcal{X}$, ensuring that the resulting regret matches the bound in Theorem 5.

To preserve strong convexity, local first-order approximations of f_0 are insufficient as they flatten the function's curvature (i.e., this would result in a *piecewise-linear* approximation of f_0 once we take the maximum over our local approximations). Therefore, we must produce local second-order approximations of f_0. In order for these approximations to remain close to f_0 locally around each $x_i \in \mathcal{S}$, we need f_0 to have a Lipschitz-continuous Hessian. However, the regularizer from Srebro et al. (2011) does not necessarily even possess smooth derivatives. We side-step this issue by proposing an alternative regularizer that not only achieves the optimal rate of $O(\mathrm{Rate}(\mathcal{X}, \mathcal{L})\sqrt{T})$ but also features smooth derivatives 5. This regularizer can then be approximated by our strategy.

Theorem 6 (Existence of a smooth regularizer) *There exists a regularizer f so that running FTRL with f has regret bound $\mathrm{Reg}(\mathrm{FTRL}(f)) \le O(\mathrm{Rate}(\mathcal{X}, \mathcal{L})\sqrt{T})$. In addition, the derivatives of f are bounded as $|D^k f(x)[v, \ldots, v]| = O\left(\mathrm{Rate}(\mathcal{X}, \mathcal{L})^2 \frac{d^{k/4}}{r^k}\right)$.*

We construct the smooth regularizer f of Theorem 6 by adding Gaussian noise to f_0, and prove that (1) the Gaussian smoothing does not impact performance; running mirror descent with f achieves the same regret bound as running mirror descent with f_0, and (2) the derivatives of f are sufficiently smooth due to the Gaussian smoothing (see Theorem 25). For the remainder of the paper we will let f denote this smooth, optimal regularizer (in contrast to f_0, which we will use to denote the original possibly non-smooth regularizer guaranteed by Theorem 5).

6. Approximating the Smooth Regularizer

Now that we can focus on smooth regularizers, we can explore their approximation using the previously outlined approach; from the derivative bound in Theorem 6, it is easy to show that our smooth regularizer f has an L-Lipschitz Hessian (i.e., satisfies $\left\|\nabla^2 f(x_0) - \nabla^2 f(x_1)\right\| \le L\left\|x_0 - x_1\right\|$ for all $x_0, x_1 \in \mathbb{R}^d$). From this property, we can show that the quadratic approximation of f around x_0 remains close to f, at least locally around x_0.

However, since our final approximation of f is constructed as the maximum over a collection of local approximators, the quadratic approximation at x_0 may be overshadowed by those centered at other points, potentially leading to significant deviations from f in the vicinity of x_0. To address this, we must ensure that each local approximator is designed to maintain its dominance, at least within its own neighborhood. We accomplish this by adding a norm-cubic "decay" term to the quadratic approximation. This ensures that for any two points $x_0, x_1 \in \mathbb{R}^d$, if they are sufficiently distant, the local approximator around x_0 does not dominate the one at x_1 within its own neighborhood, and vice versa. Consequently, our final local approximation of f around a point $x_0 \in \mathbb{R}^d$ takes the following form:

$$f_{x_0}(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{1}{2}(x - x_0)^\top \nabla^2 f(x_0)(x - x_0) - \tfrac{L}{3}\|x - x_0\|^3. \quad (3)$$

We refer to a function of the form in equation 3 as "quasi-quadratic," centered at x_0. The intuition for this approximation is that the norm cubic term adds a decay to the Hessian of the function as we move away from x_0; this decay guarantees that $f_{x_0}(x)$ is always a lower bound for f, and in particular can be estimated by f from above and below with margin $L\left\|x - x_0\right\|^3$. We show this in Lemma 7. On the other hand, this decay is slow enough so that from the L-Hessian smoothness of f we can prove that the Hessian of the approximation remains almost the same as the Hessian of f (at least locally around x_0) and therefore the strong convexity property can be preserved (see Lemma 10). In particular, even though the function in equation 3 is not strongly convex for all $x \in \mathbb{R}^d$ (or even all $x \in \mathcal{X}$), since it is strongly convex locally, if we choose a sufficiently dense discretization set $\mathcal{S}$, our resulting overall approximate regularizer will still be strongly convex (we discuss this more in Section 7).

Lemma 7 (Estimating f by the approximator) *We have the following relation between the value of f and f_{x_0}:*

$$f_{x_0}(x) + \tfrac{L}{6}\|x - x_0\|^3 \le f(x) \le f_{x_0}(x) + \tfrac{L}{2}\|x - x_0\|^3.$$

The proof of Lemma 7 is in Section G.1.

Finally, we combine these local approximations for each $x_i \in \mathcal{S}$ by taking the maximum over all these functions and defining the *piece-wise quasi-quadratic function* $\tilde{f}(x) = \max_{x_i \in \mathcal{S}} f_{x_i}(x)$. Importantly, while $\tilde{f}$ remains strongly convex and suitably bounded on $\mathcal{X}$, it is also efficiently encoded by $f(x_i)$, $\nabla f(x_i)$, and $\nabla^2 f(x_i)$ at discretized points $\mathcal{S} = \{x_i\}_{i=1}^N$, since each $f_{x_i}(x)$ does not use more than zeroth, first, and second order information of f at x_i's. Therefore, we can narrow our search for suitable regularizers from all convex functions on $\mathbb{R}^d$ to the selection of the value, gradient, and Hessian of a piece-wise quasi-quadratic function at a finite set of points. In fact, in the next section we write a convex program to minimize the maximum value of these piecewise quasi-quadratic regularizers.

7. A Convex Program for Calculating an Ideal Regularizer

In the previous section, we showed that it is possible to approximate f with a set of quasi-quadratic approximators which only depend on the value, gradient, and Hessian of f at a finite set of points $\mathcal{S} = \{x_i\}_{i=1}^N$. In this section, we describe how to search the space of such approximators by defining a convex program whose variables are the function's value, gradient and Hessian at $\mathcal{S}$, denoted by $\{r_{x_i}, v_{x_i}, \Sigma_{x_i}\}_{i=1}^N$.

Before rigorously defining the program, we first provide some motivation for its definition. In particular, we want the instance $\tilde{\mathcal{I}} = \Big(\{\tilde{r}_{x_i}\}_{i=1}^N, \{\tilde{v}_{x_i}\}_{i=1}^N, \{\tilde{\Sigma}_{x_i}\}_{i=1}^N\Big)$ where $\tilde{r}_{x_i} \triangleq f(x_i), \tilde{v}_{x_i} \triangleq \nabla f(x_i), \tilde{\Sigma}_{x_i} \triangleq \nabla^2 f(x_i)$, corresponding to the smoothed regularizer f in Theorem 5, to be a feasible point. On the other hand, for any instance $\mathcal{I} = (\mathbf{r}, \mathbf{v}, \mathbf{\Sigma}) = \big(\{r_{x_i}\}_{i=1}^N, \{v_{x_i}\}_{i=1}^N, \{\Sigma_{x_i}\}_{i=1}^N\big)$, we can define a regularizer $g_{x_i}^{(\mathcal{I})}(x)$ as

$$g^{(\mathcal{I})}(x) \triangleq \max_{i \in [N]} g_{x_i}^{(\mathcal{I})}(x), \tag{4}$$

where imitating the approximation that we derived for f in equation 3, $g_{x_i}^{(\mathcal{I})}(x)$ denotes a quasi-quadratic function:

$$g_{x_i}^{(\mathcal{I})}(x) = r_{x_i} + \langle v_{x_i}, x - x_i\rangle + \tfrac{1}{2}(x - x_i)^\top \Sigma_{x_i}(x - x_i) - \tfrac{L}{6}\left\|x - x_i\right\|^3. \tag{5}$$

With this terminology, it is clear that $\tilde{f} = g^{\tilde{\mathcal{I}}}$. Besides having $\tilde{\mathcal{I}}$ as a feasible point of the program, we also want to impose constraints so that for the optimal solution of the program, $\mathcal{I}^*$, the regularizer $g^{(\mathcal{I}^*)}$ is strongly convex and suitably bounded on $\mathcal{X}$. First, note that from Lemma 20, α-strong convexity of f with respect to $\|.\|_{\mathcal{L}^c}$ is equivalent to the condition

$$v^\top \nabla^2 f(x) v \geq \alpha \tag{6}$$

for all $x \in \mathcal{X}$ and $v \in \mathcal{L}$. Hence, we also add the condition $v^\top \Sigma_{x_i} v \geq \alpha,\ \forall v \in \mathcal{L}$ to the program. While this condition asserts strong convexity of $g^{(\mathcal{I})}$ for all feasible instances $\mathcal{I}$ at the discretization points, it does not guarantee strong convexity elsewhere. The reason is that the approximator in equation 5 loses the strongly convexity property for points far from x_i. Therefore, in order to guarantee strong convexity for $g^{(\mathcal{I})}$ everywhere, we need to make sure that at any point $x \in \mathcal{X}$, the maximum in equation 4 is attained by a function $g_{x_i}^{(\mathcal{I})}$ where x_i is sufficiently close to x. Building on this observation, we introduce the concept of "locality" for an arbitrary instance $\mathcal{I}$:

Definition 8 *We define an instance $\mathcal{I} = (\mathbf{r}, \mathbf{v}, \mathbf{\Sigma})$ as ϵ-local if, for every x, $\left\|x_{\hat{i}(x)} - x\right\| = O(\epsilon)$ where $\hat{i}(x) \triangleq \arg\max_{i\in[N]} g_{x_i}^{(\mathcal{I})}(x)$.*

Note that ϵ-locality is guaranteed for $\tilde{f} = g^{(\tilde{\mathcal{I}})}$ by Lemma 7. Specifically, if there is a point $x_i \in \mathcal{S}$ such that $\|x_i - x\| = O(\epsilon)$, then according to Lemma 7, the point $x_{\hat{i}(x)}$ where $g_{x_{\hat{i}(x)}}$ attains its maximum in equation 4 at x, must also be within a distance of $O(\epsilon)$ from x. To ensure that the maximum equation 4 is attained at an $x_{\hat{i}(x)}$ that is close to x, we enforce a slightly relaxed version of the lower bound from Lemma 7 on $g^{(\mathcal{I})}$ at the discretization points:

$$g_{x_i}^{(\mathcal{I})}(x_j) + \tfrac{15L}{96} \|x_j - x_i\|^3 \leq r_{x_j}, i, j = 1, \dots, N. \tag{7}$$

As noted in Lemma 7, $\tilde{f}$ satisfies the inequality $f_{x_0}(x) + \frac{L}{6}|x - x_0|^3 \leq f(x)$. The reason we apply a slightly weaker version of this inequality in equation 7 will become evident when we design a separation oracle for the feasibility set of the convex program. At a high level, this condition ensures that not only is $\tilde{\mathcal{I}}$ a feasible instance for our program, but that a small neighborhood around it also remains feasible. As we will see, even after enforcing the condition in equation 7, an arbitrary feasible instance $\mathcal{I}$ does not achieve $O(\epsilon)$-locality like $\tilde{\mathcal{I}}$. Instead, we can only prove that it is $O(\epsilon^{1/3})$-local (see Lemma 9). The reason is that equation 7 is only enforced at the discretization points, whereas $\tilde{f}$ satisfies it for any $x \in \mathcal{X}$ as shown in Lemma 7.

Finally, we aim to minimize the maximum value of $g^{(\mathcal{I})}$ over $\mathcal{X}$ to obtain a suitable regularizer for FTRL. As mentioned earlier, we smooth the theoretical regularizer f_0 from Srebro et al. (2011) by adding Gaussian noise, resulting in f, which ensures bounded gradients and Hessians. To achieve a similar smoothness condition on the regularizer $g^{(\mathcal{I})}$ that correspond to a feasible instance of our program, we enforce the conditions $\|v_{x_i}\|_\infty \leq c_0$ and $\Sigma_{x_i} \preccurlyeq c_2 I$ for constants c_0, c_2 (we use the infinity norm instead of the 2-norm to maintain a linear constraint.) With the discretization set $\mathcal{S} = \{x_i\}_{i=1}^N$ fixed, the final program is as follows:

$$
\begin{aligned}
&\text{minimize } r && (8)\\
&\text{subject to } r_{x_i} + \langle v_{x_i}, \Delta_{ij}\rangle + \tfrac{1}{2}\Delta_{ij}^\top \Sigma_{x_i}\Delta_{ij} - \tfrac{17L}{96}\left\|\Delta_{ij}\right\|^3 \le r_{x_j} && \forall i,j\in[N], \Delta_{ij} := x_j - x_i\\
&\qquad \left\|v_{x_i}\right\|_\infty \le c_0 && i\in[N]\\
&\qquad \Sigma_{x_i} \preccurlyeq c_2 I && \forall i\in[N]\\
&\qquad v^\top \Sigma_{x_i} v \ge \alpha && \forall v\in\mathcal{L},\ \forall i\in[N]\\
&\qquad r \ge r_{x_i} && \forall i\in[N]\\
&\qquad r, r_{x_i} \le C_0 && \forall i\in[N].
\end{aligned}
$$

Next, to establish the locality property for feasible points of the program, we state in Lemma 9 that for any arbitrary $x\in\mathcal{X}$, the maximum in equation 4 is attained at a discretization point $x_i\in\mathcal{S}$ that is not too far from x. Specifically, given that every point in $\mathcal{X}$ has a discretization point x_i within a distance of ϵ, we show that the maximum in equation 4 is achieved by $x_{\hat{i}}$ which is no further than $O(\epsilon^{1/3})$ from x. Additionally, we prove that the value of $g^{(\mathcal{I})}$ at x is close to $g^{(\mathcal{I})}_{x_i}(x)$.

Lemma 9 (Convex program feasibility → Locality of regularizer g) *Assume that $\mathcal{I} = (\mathbf{r},\mathbf{v},\mathbf{\Sigma})$ is feasible for LP 8, for ϵ satisfying $\epsilon \le \gamma_2 \min\left\{\frac{L}{\sqrt{d}c_0}, \frac{L}{c_0\sqrt{d}c_2^3}, \frac{L}{c_2}, \sqrt{c_0\sqrt{d}}, \frac{c_0\sqrt{d}}{c_2}\right\}$. If for x_i, x_j and $x\in\mathcal{X}$ we have $\|x_i - x\|\le\epsilon$ and $\|x_j - x\| \ge \gamma\left(\frac{\epsilon c_0\sqrt{d}}{L}\right)^{1/3}$ for some universal constant γ, then*

$$g^{(\mathcal{I})}_{x_i}(x) > g^{(\mathcal{I})}_{x_j}(x) + c_0\epsilon\sqrt{d},$$

and if $\|x_j - x\| \le \gamma\left(\frac{\epsilon\sqrt{d}c_0}{L}\right)^{1/3}$, then

$$|g^{(\mathcal{I})}_{x_j}(x_i) - g^{(\mathcal{I})}_{x_j}(x)| \le \gamma_2 c_0\epsilon\sqrt{d},$$

for some constant γ_2.

The proof can be found in Section G.2. To prove strong convexity of $g^{(\mathcal{I})}$ for a feasible point $\mathcal{I}$, we must first establish the strong convexity of the local approximators $g^{(\mathcal{I})}_{x_i}$, defined in equation 5. This is demonstrated in Lemma 10 below. Specifically, we prove that if the quadratic form of the Hessian variable Σ_{x_i} is lower bounded by the norm squared $\|.\|^2_{\mathcal{L}^c}$ in all directions, then $g^{(\mathcal{I})}_{x_i}(x)$ is strongly convex locally around x_i.

Lemma 10 *[Local strong convexity of the approximators] Suppose the PSD matrix Σ is such that for all v, $v^\top\Sigma v \ge \alpha\,\|v\|^2_{\mathcal{L}^c}$. Then, the function*

$$g(x) = r + \langle v, x - x_0\rangle + \tfrac{1}{2}(x-x_0)^\top\Sigma(x-x_0) - \tfrac{L}{6}\|x-x_0\|^3$$

for arbitrary x_0, v, r, L is $\alpha/2$-strongly convex with respect to $\|.\|_{\mathcal{L}^c}$ in the neighborhood $\|x-x_0\|\le\frac{\alpha}{2R^2L}$. Consequently, if f is α-strongly convex with respect to $\|.\|_{\mathcal{L}}$, then $f_{x_0}(x)$ is $\frac{\alpha}{2}$ strongly convex with respect to $\|.\|_{\mathcal{L}}$ for $\|x-x_0\|\le\frac{\alpha}{2R^2L}$.

The proof of Lemma 10 can be found in Section G.3. Finally, by combining Lemmas 35 and 9, we show that the barrier $g^{(\mathcal{I})}$ constructed from a feasible point of the matrix program has a suitable upper bound on $\mathcal{X}$, satisfying the desired strong convexity. Additionally, we prove that the feasible region can be approximated both from the inside and outside by Euclidean balls, a key property necessary for constructing a separation oracle for the feasible set later.

Theorem 11 (Convex program solution → optimal regularizer) *Assume we are given a smooth barrier function $f : \mathbb{R}^d \to \mathbb{R}$ with $|f(x)| \leq C^2, \forall x \in \mathcal{X}$, which is $\tilde{c}_1$ Lipschitz, $\tilde{c}_2$ gradient Lipschitz, $\tilde{L}$ Hessian Lipschitz, and α $\|.\|_{\mathcal{L}^c}$-strongly convex in $\mathcal{X}$. Additionally, if for every two points in the cover $x_i, x_j \in \tilde{\mathcal{X}}$ we have $\|x_i - x_j\| \geq \bar{\epsilon}$, then the convex program in equation 8 with $c_0 = \tilde{c}_1 + L\bar{\epsilon}^3, c_2 = \tilde{c}_2 + L\bar{\epsilon}^3$, $L = \tilde{L}$, $C_0 = C^2 + L\bar{\epsilon}^3$, and discretization parameter $\epsilon \leq \gamma_3 \min\{\frac{L}{\sqrt{d}c_1}, \frac{L}{c_1\sqrt{d}{c_2}^3}, \frac{L}{c_2}, \sqrt{c_1\sqrt{d}}, \frac{c_1\sqrt{d}}{c_2}, \frac{\alpha^3}{512R^6L^2c_1\sqrt{d}}\}$ for all sufficiently small constant γ_3 is feasible. Furthermore, the function $g^{(\mathcal{I}^*)}$, corresponding to the optimal solution $\mathcal{I}^* = (\mathbf{r}^*, \mathbf{v}^*, \mathbf{\Sigma}^*)$ is convex and satisfies the following properties:*

1. *$|g^{(\mathcal{I}^*)}(x)| \leq C^2 + \gamma_2\epsilon\sqrt{d}c_0$ for constant γ_2.*
2. *For any feasible instance $\mathcal{I} \in \mathcal{P}_{\mathcal{I}}$, $g^{(\mathcal{I})}(x)$ is $\frac{\alpha}{2}$ strongly convex with respect to $\|.\|_{\mathcal{L}^c}$.*
3. *$B_{L\bar{\epsilon}^3/288}(\tilde{\mathcal{I}}) \subseteq P_{\mathcal{I}} \subseteq B_{2\sqrt{(N+1){C_0}^2 + Nd\left(c_0^2 + c_2^2\right)}}(\tilde{\mathcal{I}})$.*

The proof of Theorem 11 can be found in Section G.4.

8. Lower Bounds on Membership Oracle Query Complexity for $\mathcal{L}$

In the above sections we demonstrated an algorithm for computing an optimal regularizer that runs in time $\exp(O(d^2))$. In this final section, we show that this is in some sense necessary, by showing that just checking the α-strong convexity of a given regularizer g with respect to $\|.\|_{\mathcal{L}^c}$ at point $x \in \mathcal{X}$ requires an exponential number of queries to a membership oracle $\mathrm{MEM}_{\mathcal{L}}(\delta)$. In particular, even in the simple case where $\nabla^2 g(x) = I$ (i.e., the quadratic regularizer), an exponential number of queries is needed. The lower bound is based on a reduction to Theorem 1.2 in Bhattiprolu et al. (2021).

Theorem 12 (Exponential lower bound) *Given ϵ, for large enough dimension d, there exists a distribution over convex bodies $\mathcal{L}$ such that for every fixed set of queried points $S \subseteq \mathbb{R}^d$,*

1. *$\mathbb{P}_{\mathcal{L}}\left(S \cap \{v| \ \|v\|_{\mathcal{L}} \leq 1\} = S \cap B_1(0)\right) \geq 1 - \epsilon$*
2. *There exists direction $\tilde{v}$ with $\|\tilde{v}\|_{\mathcal{L}^c} = 1$ such that $\|\tilde{v}\|_2 \leq \frac{1}{d^{1-\epsilon}}$,*

where $B_1(0)$ is the Euclidean ball with radius 1.

The proof of Theorem 12 is provided in Section G.5. At a high level, Theorem 12 asserts that there exists a distribution over norm balls $\mathcal{L}$ such that (1) even $\exp(d^{1-\epsilon})$ queries are insufficient to distinguish between $\mathcal{L}$ and the Euclidean unit ball, while (2) the Identity Hessian is not $\alpha = \frac{1}{d^{1-\epsilon}}$ strongly convex with respect to the dual norm $\|.\|_{\mathcal{L}^c}$.

Of course, it is possible that there is a method for computing the optimal regularizer that sidesteps the need to be able to verify the convexity of an arbitrary regularizer – we leave this as an interesting open problem.

References

Jacob Abernethy, Peter L Bartlett, and Elad Hazan. Blackwell approachability and no-regret learning are equivalent. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 27–46. JMLR Workshop and Conference Proceedings, 2011.

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

Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

Pierre-Cyril Aubin-Frankowski, Anna Korba, and Flavien Léger. Mirror descent with relative smoothness in measure spaces, with application to sinkhorn and em. *Advances in Neural Information Processing Systems*, 35:17263–17275, 2022.

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2014.

Santiago Balseiro, Christian Kroer, and Rachitesh Kumar. Online resource allocation under horizon uncertainty. In *Abstract Proceedings of the 2023 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 63–64, 2023.

Nikhil Bansal and Christian Coester. Online metric allocation. *arXiv preprint arXiv:2111.15169*, 2021.

Anastasia Sergeevna Bayandina, Alexander Vladimirovich Gasnikov, Evgenya Vladimirovna Gasnikova, and SV Matsievskii. Primal–dual mirror descent method for constraint stochastic optimization problems. *Computational Mathematics and Mathematical Physics*, 58:1728–1736, 2018.

Vijay Bhattiprolu, Euiwoong Lee, and Assaf Naor. A framework for quadratic form maximization over convex sets through nonconvex relaxations. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 870–881, 2021.

Harold D. Block. The perceptron: A model for brain functioning. *Reviews of Modern Physics*, 34: 123–135, 1962. Reprinted in *Neurocomputing* by Anderson and Rosenfeld.

Sébastien Bubeck, Ronen Eldan, and Nicolò Cesa-Bianchi. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, pages 449–472. JMLR Workshop and Conference Proceedings, 2012.

Nicolò Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 2011. To appear.

Evgenii Chzhen, Christophe Giraud, and Gilles Stoltz. A unified approach to fair online learning via blackwell approachability. *Advances in Neural Information Processing Systems*, 34:18280–18292, 2021.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, pages 345–352, 2008.

John C Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *Colt*, volume 10, pages 14–26. Citeseer, 2010.

Gabriele Farina, Tommaso Bianchi, and Tuomas Sandholm. Coarse correlation in extensive-form games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1934–1941, 2020.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Faster game solving via predictive blackwell approachability: Connecting regret matching and mirror descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5363–5371, 2021.

Dylan J Foster, Alexander Rakhlin, and Karthik Sridharan. Online learning: Sufficient statistics and the burkholder method. In *Conference On Learning Theory*, pages 3028–3064. PMLR, 2018.

Geoffrey J Gordon, Amy Greenwald, and Casey Marks. No-regret learning in convex games. In *Proceedings of the 25th international conference on Machine learning*, pages 360–367, 2008.

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

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

Arun Jambulapati and Kevin Tian. Revisiting area convexity: Faster box-simplex games and spectrahedral generalizations. *Advances in Neural Information Processing Systems*, 36, 2024.

Arun Jambulapati, Jerry Li, and Kevin Tian. Robust sub-gaussian principal component analysis and width-independent schatten packing. *Advances in Neural Information Processing Systems*, 33: 15689–15701, 2020.

Yujia Jin and Aaron Sidford. Efficiently solving mdps with stochastic mirror descent. In *International Conference on Machine Learning*, pages 4890–4900. PMLR, 2020.

Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, http://ttic.uchicago.edu/shai/papers/KakadeShalevTewari09.pdf*, 2010. Technical Report.

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

Jyrki Kivinen and Manfred K Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–64, January 1997.

Wouter M Koolen, Manfred K Warmuth, Jyrki Kivinen, et al. Hedging structured concepts. In *COLT*, pages 93–105. Citeseer, 2010.

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

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

Yunwen Lei and Ke Tang. Stochastic composite mirror descent: Optimal bounds with high probabilities. *Advances in Neural Information Processing Systems*, 31, 2018.

Isaac E Leonard and James Edward Lewis. *Geometry of convex sets*. John Wiley & Sons, 2015.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

Alfonso Lobos, Paul Grigas, and Zheng Wen. Joint online learning and decision-making via dual mirror descent. In *International Conference on Machine Learning*, pages 7080–7089. PMLR, 2021.

Haihao Lu, Santiago Balseiro, and Vahab Mirrokni. Dual mirror descent for online allocation problems. *arXiv preprint arXiv:2002.10421*, 2020.

Arkadi Nemirovski and David Yudin. *Problem complexity and method efficiency in optimization*. Nauka Publishers, Moscow, 1978.

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

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

Princewill Okoroafor, Bobby Kleinberg, and Wen Sun. Faster recalibration of an online predictor via approachability. In *International Conference on Artificial Intelligence and Statistics*, pages 4690–4698. PMLR, 2024.

Francesco Orabona and Dávid Pál. Scale-free algorithms for online linear optimization. In *International Conference on Algorithmic Learning Theory*, pages 287–301. Springer, 2015.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. *Advances in Neural Information Processing Systems*, 23, 2010.

Shahin Shahrampour and Ali Jadbabaie. Distributed online optimization in dynamic environments using mirror descent. *IEEE Transactions on Automatic Control*, 63(3):714–725, 2017.

Jonah Sherman. Area-convexity, linf regularization, and undirected multicommodity flow. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 452–460, 2017.

Nati Srebro, Karthik Sridharan, and Ambuj Tewari. On the universality of online mirror descent. *Advances in neural information processing systems*, 24, 2011.

Karthik Sridharan and Ambuj Tewari. Convex games in banach spaces. In *COLT*, pages 1–13. Citeseer, 2010.

Eiji Takimoto and Manfred K Warmuth. Path kernels and multiplicative updates. *The Journal of Machine Learning Research*, 4:773–818, 2003.

Daniil Tiapkin and Alexander Gasnikov. Primal-dual stochastic mirror descent for mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 9723–9740. PMLR, 2022.

Manfred K. Warmuth and Dima Kuzmin. Randomized online pca algorithms with regret bounds that are logarithmic in the dimension. In *Proceedings of the 20th Annual Conference on Learning Theory (COLT)*, 2007.

Andre Wibisono, Molei Tao, and Georgios Piliouras. Alternating mirror descent for constrained min-max games. *Advances in Neural Information Processing Systems*, 35:35201–35212, 2022.

Deming Yuan, Yiguang Hong, Daniel WC Ho, and Shengyuan Xu. Distributed mirror descent for online composite optimization. *IEEE Transactions on Automatic Control*, 66(2):714–729, 2020.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 928–936, 2003.

Appendix A. Additional Related Work

Many modern learning algorithms are actually variants of mirror descent / FTRL (Block, 1962; Zinkevich, 2003; Kivinen and Warmuth, 1997; Littlestone, 1988; Kakade et al., 2010; Warmuth and Kuzmin, 2007). Recently, Jin and Sidford (2020) used a variant of mirror descent to solve infinite-horizon MDPs, achieving linear runtime in the number of samples. Aubin-Frankowski et al. (2022) extended mirror descent to optimize convex functionals on an infinitesimal space, demonstrating that the primal iterations of Sinkhorn's algorithm for entropic optimal transport in a continuous domain are an instance of mirror descent. Wibisono et al. (2022) studied alternating mirror descent for two-player bilinear zero-sum games, proving a regret bound of $O\left(T^{1/3}\right)$. Abernethy et al. (2014) discusses the equivalence of FTPL (Follow-The-Perturbed-Leader) and FTRL. Orabona and Pál (2015) presents scale-free generalizations of FTRL algorithms (with regret bounds depending on the total magnitude of the losses instead of the time horizon).

Mirror descent has also been used in the context of stochastic optimization Nemirovski et al. (2009). Authors in Duchi et al. (2010) study mirror descent for composite loss functions under both stochastic and online settings. Lei and Tang (2018) relaxed the subgradient boundedness condition from Duchi et al. (2010) and extended their analysis to examine the generalization performance of multi-pass SGD in non-parametric settings. Dani et al. (2008); Cesa-Bianchi and Lugosi (2011); Bubeck et al. (2012) applied mirror descent to address the problem of online linear optimization with bandit feedback. Allen-Zhu and Orecchia (2014) introduced a novel interpretation of mirror descent as optimizing a dual-based lower bound for the objective. Building on this perspective, they proposed a coupling between mirror descent and gradient descent that achieves an accelerated convergence rate. (Yuan et al., 2020; Shahrampour and Jadbabaie, 2017) applied mirror descent in distributed settings. Lobos et al. (2021) utilized mirror descent for a constrained online revenue maximization problem with unknown parameters. Authors in (Bansal and Coester, 2021; Lu et al., 2020; Balseiro et al., 2023) employ mirror descent for online resource allocation problems. Mirror descent has also been instrumental in primal-dual methods for solving structured saddle-point problems (Nesterov, 2009; Tiapkin and Gasnikov, 2022; Bayandina et al., 2018; Sherman, 2017; Jambulapati and Tian, 2024; Jambulapati et al., 2020).

Appendix B. The ideal regularizer and proving stronger martingale type inequalities for $p = 2$

Here, we state the existence of an ideal regularizer such that running FTRL with this regularizer achieves the optimal rate up to a constant. This result is adapted from Srebro et al. (2011), except that they prove the same regularizer results in a regret bound which is off by a logarithmic factor of $\log(T)$; this log factor is indeed not desirable for our purpose as we are interested in long time horizon regimes when T can potentially be exponentially large in dimension. Our contribution here is that we improve the result of Srebro et al. (2011) for $p = 2$ case and remove this log factor. We further demonstrate a continuity condition for this ideal regularizer that we use for our smoothing arguments in Section C.

First, we state the result of Sridharan and Tewari (2010) and Rakhlin et al. (2010) that we build upon; it is known (Sridharan and Tewari, 2010; Rakhlin et al., 2010) that the optimal rate for adversarial online linear optimization translates into a property on the growth of the norm $\|.\|_{\mathcal{X}^c}$ of an arbitrary Rademacher martingale sequence. We state this property rigorously in Theorem 13, which is stated as Theorem 4 in Srebro et al. (2011).

Theorem 13 (Restatement of Theorem 4 in Srebro et al. (2011)) *Given the optimal rate for online linear optimization with action and loss sets $\mathcal{X}, \mathcal{L} \in \mathbb{R}^d$ is $O(C\sqrt{T})$, then for a Rademacher random vector $\epsilon \in \{\pm\}^n$ and any sequence of functions $x_i(\epsilon) : \{\pm\}^i \to \mathbb{R}^d$, where x_i is a function of the first i coordinates in ϵ, we have*

$$\mathbb{E}\left\|\sum_i \epsilon_i x_i(\epsilon)\right\|_{\mathcal{X}^c} \leq O(C) \sup_{0\leq i\leq n} \sup_{\epsilon} \|x_i(\epsilon)\|_{\mathcal{L}}. \tag{9}$$

The main contribution of authors in Srebro et al. (2011) is that they translate equation 9 to the existence of a suitable barrier for mirror descent. In particular, they prove the following key Lemmas 14, 17. We start with Lemma 14 which translates property equation 9 to a more refined argument about the growth of martingale norms that are defined based on the action and loss sets.

Lemma 14 (Restatement of Lemma 12 in Srebro et al. (2011) for $r = 2$) *For $1 < r < 2$, if there exists a constant $C > 0$ such that for any natural number n and any sequence of mappings $(x_i)_{i=1}^n$, $x_i : \{\pm\}^i \to \mathbb{R}^d$ and Rademacher random vector $\epsilon \in \{\pm\}^n$ satisfy*

$$\mathbb{E}\left\|\sum_{i=1}^n \epsilon_i x_i(\epsilon)\right\|_{\mathcal{X}^c} \leq C n^{1/r} \sup_{0\leq i\leq n} \sup_{\epsilon} \|x_i(\epsilon)\|_{\mathcal{L}},$$

then for $p < r$ and $\alpha_p = \frac{20C}{r-p}$, for any sequence $(x_i)_{i=1}^n$ as described above, we have the following inequality:

$$\mathbb{E}\left\|\sum_{i=1}^n \epsilon_i x_i(\epsilon)\right\|_{\mathcal{X}^c} \leq \alpha_p \sup_{\epsilon} \left(\sum_i \|x_i(\epsilon)\|_{\mathcal{L}}^p\right)^{1/p}. \tag{10}$$

The next Lemma states how authors in Srebro et al. (2011) translate the property in Equation equation 10 to the existence of the ideal regularizer:

Lemma 15 (Restatement of Lemma 11 in Srebro et al. (2011)) *For constant $\tilde{C}$, the following statements are equivalent:*

1. *For all n and sequence of mappings $(x_i)_{i=0}^n$ where $x_i : \{\pm\}^{i-1} \to \mathbb{R}^d$:*

$$\mathbb{E}_\epsilon \left\|\sum_{i=1}^n \epsilon_i x_i(\epsilon)\right\|_{\mathcal{X}^c}^p \leq \tilde{C}^p \left(\sum_{i=1}^n \mathbb{E}\|x_n(\epsilon)\|_{\mathcal{L}}^p\right)$$

2. *There exists a 2-homogeneous non-negative convex function f_0 on $\mathbb{R}^d$ which is 1-strongly convex w.r.t $\|.\|_{\mathcal{L}^c}$ and $\forall x$, $\frac{1}{q}\|x\|_{\mathcal{L}^c}^q \leq f_0(x) \leq \frac{\tilde{C}^q}{q}\|x\|_{\mathcal{X}}^q$, where $\frac{1}{p} + \frac{1}{q} = 1$.*

The existence of such regularizer from Lemma 15 then implies a $\tilde{C}T^{1-\frac{1}{p}}$ regret bound for FTRL. Nonetheless, the reason they end up with a $\log(T)$ factor in the regret is that they need to use Lemma 14 with a power $p < 2$ slightly less than two, as the constant α_p reciprocally depends on $2-p$, so p has to be $\Theta(1/\log(T))$ less than 2. We improve Lemma 14 in Lemma 16 below, for the case of $p = 2$, and shave off the α_p factor which is causing the additional $\log(T)$. This enables us to show a tighter upper bound for the regularizer on domain $\mathcal{X}$ in Theorem 18.

Lemma 16 (Improving the Martingale Type for $p=2$) *Suppose for the norm $\|.\|_{\mathcal{X}^c}$ we have*

$$\mathbb{E}\left\| x_0 + \sum_{i=1}^{n} \epsilon_i x_i(\epsilon) \right\|_{\mathcal{X}^c} \leq D(n+1)^{1/2} \sup_{0 \leq i \leq n} \sup_{\epsilon} \|x_i(\epsilon)\|_{\mathcal{L}}, \tag{11}$$

for arbitrary vector valued functions $x_n : \{\pm 1\}^{n-1} \to \mathbb{R}^d$ and Rademacher sequence $(\epsilon_i)_{i=1}^n$, $\epsilon_i \sim \pm 1$. Then, we have

$$\mathbb{E}\left\| x_0 + \sum_{i=1}^{n} \epsilon_i x_i(\epsilon) \right\|_{\mathcal{X}^c} \leq D\left(\sum_{i=1}^{n} \|x_i(\epsilon)\|_{\mathcal{L}}^2 \right)^{1/2}.$$

Proof First, note that if we average equation 11 over x_0 and $-x_0$ and extend the functions $x_i(\epsilon)$ to also depend on a Rademacher variable ϵ_0 at time zero, then we get

$$\mathbb{E}\left\| \sum_{i=0}^{n} \epsilon_i x_i(\epsilon) \right\|_{\mathcal{X}^c} \leq D(n+1)^{1/2} \sup_{0 \leq i \leq n} \sup_{\epsilon} \|x_i(\epsilon)\|_{\mathcal{L}}. \tag{12}$$

Now let $c_i = \|x_i\|_{\mathcal{L}}$. Take a fresh rademacher sequence $(\tilde{\epsilon}_j)_{j=1}^{\infty}$. We will define the sequence $(\epsilon_i)_{i=1}^n$ based on the randomness of $\tilde{\epsilon}_j$'s: define $\hat{\epsilon}_i = 1$ if $\sum_{j=t_i+1}^{t_{i+1}} \tilde{\epsilon}_j \geq \frac{\|x_i\|}{\delta}$ and $\hat{\epsilon}_i = -1$ if $\sum_{j=t_i+1}^{t_{i+1}} \tilde{\epsilon}_j \leq -\frac{\|x_i\|}{\delta}$. From symmetry, it is easy to check that ϵ_i's are indeed i.i.d distributed uniformly on $\{\pm 1\}$. Next, for a given positive $\delta > 0$, define the sequence of indices $(t_i)_{i=1}^n$ and the alternative sequence $(\tilde{x}_i)_{i=0}^m$ such that for all i, $\tilde{x}_{t_i} = \tilde{x}_{t_i+1} = \cdots = \tilde{x}_{t_{i+1}-1} = \frac{x_i}{\|x_i\|_{\mathcal{L}}}\delta$, and t_i is the first index such that $|\sum_{j=t_i+1}^{t_{i+1}} \tilde{\epsilon}_j| \geq \frac{\|x_i\|}{\delta}$. Now from this definition. we have that $\tilde{x}_i$'s satisfy

$$\left\| x_i - \sum_{j=t_i+1}^{t_{i+1}} \tilde{x}_j \right\|_{\mathcal{X}^c} \leq \delta \left\| \frac{x_i}{\|x_i\|_{\mathcal{L}}} \right\|_{\mathcal{X}^c}. \tag{13}$$

But for $t_{\text{sum}} = \sum_{i=0}^n t_i$ equation 13 implies:

$$\left\| \sum_{i=0}^{n} \epsilon_i x_i(\epsilon) - \sum_{j=0}^{t_{\text{sum}}} \tilde{\epsilon}_j \tilde{x}_j \right\|_{\mathcal{X}^c} \leq (n+1)\delta \max_{i=0}^{n} \left\| \frac{x_i}{\|x_i\|_{\mathcal{L}}} \right\|_{\mathcal{X}^c}.$$

The key observation is for all $i \in [n]$, the distribution of t_i is sub-exponential and the sum concentrates around its expectation. In particular,

$$\mathbb{P}\left(t_i \geq \kappa \left(\frac{\|x_i\|_{\mathcal{L}}}{\delta} \right)^2 \right) \leq e^{-O(\kappa)}. \tag{14}$$

It is sufficient for us to show that the sum $\sum_{i=1}^n t_i$ is at most $O\left(\sum_{i=1}^n \left(\frac{\|x_i\|_{\mathcal{L}}}{\delta} \right)^2 \right)$ with at least constant probability p. Call this event $\mathcal{E}$. First, we use Chebyshev inequaility to show $\mathbb{P}(\mathcal{E}) = \Omega(1)$. Note that Equation equation 14 imlies

$$\mathbb{E}t_i^2 = O\left(\frac{\|x_i\|_{\mathcal{L}}}{\delta} \right)^4,$$

which implies

$$Var(\sum_i t_i) = O\left(\sum_i \left(\frac{\|x_i\|_{\mathcal{L}}}{\delta}\right)^4\right).$$

Therefore, from Chebyshev inequality

$$\mathbb{P}\left(\sum_{i=1}^n t_i \geq \sum_{i=1}^n \left(\frac{\|x_i\|_{\mathcal{L}}}{\delta}\right)^2 + l\sqrt{\sum_{i=1}^n \left(\frac{\|x_i\|_{\mathcal{L}}}{\delta}\right)^4}\right) \leq \frac{1}{l^2},$$

which implies

$$\mathbb{P}\left(\sum_{i=1}^n t_i \geq (l+1)\sum_{i=1}^n \left(\frac{\|x_i\|_{\mathcal{L}}}{\delta}\right)^2\right) \leq \frac{1}{l^2},$$

hence we showed that $\mathcal{E}$ happens with at least constant probability. Furthermore, It is easy to check that conditioned on $\mathcal{E}$, ϵ_i's are still Rademacher variables. On the other hand, using equation 12 for sequence $(\tilde{x}_i)$ and $m = \Theta\left(\sum_{i=0}^n \left(\frac{\|x_i\|_{\mathcal{L}}}{\delta}\right)^2\right)$:

$$\mathbb{E}\left\|\sum_{j=1}^m \tilde{\epsilon}_j \tilde{x}_j(\tilde{\epsilon})\right\|_{\mathcal{X}^c} \leq Dm^{1/2} \sup_{0\leq i\leq n} \sup_{\epsilon} \|\tilde{x}_i(\epsilon)\|_{\mathcal{L}}. \tag{15}$$

but from positivity of norm

$$\begin{aligned}
\mathbb{E}\left\|\sum_{j=1}^m \tilde{\epsilon}_j \tilde{x}_j(\tilde{\epsilon})\right\|_{\mathcal{X}^c} &\geq \mathbb{E}\left[\left\|\sum_{j=1}^m \tilde{\epsilon}_j \tilde{x}_j(\tilde{\epsilon})\right\|_{\mathcal{X}^c} \middle| \mathcal{E}\right]\mathbb{P}(\mathcal{E}) \\
&\geq \mathbb{E}\left[\left\|\sum_{i=1}^n \epsilon_i x_i(\epsilon)\right\|_{\mathcal{X}^c} \middle| \mathcal{E}\right]\mathbb{P}(\mathcal{E}) - (n+1)\delta \max_{i=0}^n \left\|\frac{x_i}{\|x_i\|_{\mathcal{L}}}\right\|_{\mathcal{X}^c} \\
&= \mathbb{E}\left\|\sum_{i=1}^n \epsilon_i x_i(\epsilon)\right\|_{\mathcal{X}^c} \mathbb{P}(\mathcal{E}) - (n+1)\delta \max_{i=0}^n \left\|\frac{x_i}{\|x_i\|_{\mathcal{L}}}\right\|_{\mathcal{X}^c} \\
&\geq \frac{1}{2}\mathbb{E}\left\|\sum_{i=1}^n \epsilon_i x_i(\epsilon)\right\|_{\mathcal{X}^c} - (n+1)\delta \max_{i=0}^n \left\|\frac{x_i}{\|x_i\|_{\mathcal{L}}}\right\|_{\mathcal{X}^c}.
\end{aligned}$$

Note that the equality in the third line above is because the size of t_i's is independent of ϵ's. Plugging this back into equation 15

$$\mathbb{E}\left\|\sum_{i=1}^n \epsilon_i x_i(\epsilon)\right\|_{\mathcal{X}^c} \leq \Theta\left(D\sqrt{\sum_{i=0}^n \|x_i\|_{\mathcal{L}}^2}\right) + (n+1)\delta \max_{i=0}^n \left\|\frac{x_i}{\|x_i\|_{\mathcal{L}}}\right\|_{\mathcal{X}^c}.$$

Sending $\delta \to 0$ finishes the proof. ∎

Next we state and prove Lemma 17. This Lemma in similar to Lemma 17 for the case $p = 2$, i.e. it translates the margtingale property to the existence of an ideal regularizer, except that we show an additional useful Lipschitz property for the regularizer which we use for smoothing the regularizer in Section C. The proof of Theorem 18 directly follows from combining Lemmas 17 and 16.

Lemma 17 (Martingale type $\to$ ideal regularizer) *For constant C, the following statements are equivalent:*

1. *For all n and sequence of mappings $(x_i)_{i=0}^n$ where $x_i : \{\pm\}^{i-1} \to \mathbb{R}^d$:*

$$\mathbb{E}_\epsilon \left\| x_0 + \sum_{i=1}^n \epsilon_i x_i(\epsilon) \right\|_{\mathcal{X}^c}^2 \leq C^2 \left(\|x_0\|_{\mathcal{L}}^2 + \sum_{i=1}^n \mathbb{E}\|x_n(\epsilon)\|_{\mathcal{L}}^2 \right)$$

2. *There exists a 2-homogeneous non-negative convex function f on $\mathbb{R}^d$ which is α-strongly convex w.r.t $\|.\|_{\mathcal{L}^c}$ and $\forall x$, $\frac{1}{2}\|x\|_{\mathcal{L}^c}^2 \leq f_0(x) \leq \frac{C^2}{2}\|x\|_{\mathcal{X}}^2$. Furthermore, f is Lipschitz continuous as*

$$|f_0(x_1) - f_0(x_2)| \leq C^2 \|x_1 - x_2\|_{\mathcal{X}} \left(\|x_1\|_{\mathcal{X}} \vee \|x_2\|_{\mathcal{X}}\right).$$

Proof This is Lemma 11 in Srebro et al. (2011), except that we are claiming an additional Lipschitz continuity here for f_0, which we need to show regularity properties for the gaussian smoothed function later on. To show the Lipschitz continuity, we note that from the proof of Lemma 11 in Srebro et al. (2011), f_0 is defined as the Fenchel dual of a barrier f_0^*, i.e. $f_0(x) = \sup\langle x, z\rangle - f_0^*(z)$, where $\frac{1}{C^2}\|x\|_{\mathcal{X}^c}^2 \leq f_0^*(x) \leq \|x\|_{\mathcal{L}}^2$. Therefore, defining $z(x) \triangleq \arg\max_z \langle x, z\rangle - f_0^*(z)$, we have

$$0 \leq f_0(z(x)) \leq \|x\|_{\mathcal{X}} \|z(x)\|_{\mathcal{X}^c} - \frac{1}{C^2}\|z(x)\|_{\mathcal{X}^c}^2,$$

which implies

$$C^2 \|x\|_{\mathcal{X}} \geq \|z(x)\|_{\mathcal{X}^c}.$$

Therefore, for $x_1, x_2 \in \mathcal{X}$ we have

$$\begin{aligned} f_0(x_1) \geq \langle x_1, z(x_2)\rangle - f_0^*(z(x_2)) &\geq \langle x_2, z(x_2)\rangle - f_0^*(z(x_2)) - \|x_1 - x_2\|_{\mathcal{X}} \|z(x_2)\|_{\mathcal{X}^c} \\ &\geq f_0(x_2) - C^2 \|x_1 - x_2\|_{\mathcal{X}} \|x_2\|_{\mathcal{X}}. \end{aligned}$$

Noting the reverse symmetric inequality $f_0(x_2) \geq f_0(x_1) - C^2\|x_1 - x_2\|_{\mathcal{X}} \|x_1\|_{\mathcal{X}}$ completes the proof. ∎

Appendix C. Smoothing the Regularizer

The goal of this section is to show the existence of a regularizer which enables FTRL to achieve the optimal regret for arbitrary pair $(\mathcal{X}, \mathcal{L})$ of action and loss sets which also has smooth derivatives. We achieve this by using Gaussian smoothing of the regularizer f_0 from Srebro et al. (2011). First, we state Theorem in which we prove that FTRL with this regularizer indeed achieves the optimal rate $O\left(\mathrm{Rate}(\mathcal{X}, \mathcal{L})\sqrt{T}\right)$; note that this is a $\log(T)$ improvement over the result of Srebro et al. (2011), and in addition the regularizer satisfies a desirable Lipschitz property. We then proceed to smooth this regularizer by adding Gaussian noise and showing the smoothness properties we want.

Theorem 18 (Existence of an ideal regularizer for mirror descent) *There exists a* 2*-homogeneous continuous regularizer* $f_0 : \mathbb{R}^d \to \mathbb{R}$ *which satisfies*

1. $\max_{x\in\mathcal{X}}|f_0(x)| \leq O(\mathrm{Rate}(\mathcal{X},\mathcal{L})^2)$
2. f_0 *is* 1*-strongly convex w.r.t* $\|.\|_{\mathcal{L}^c}$ *on* $\mathcal{X}$*, where* $\|.\|_{\mathcal{L}^c}$ *is the dual norm of* $\|.\|_{\mathcal{L}}$*.*
3. f_0 *satisfies the following Lipschitz continuity condition:* $\forall x_1, x_2$*:*

$$|f_0(x_1) - f_0(x_2)| \leq O(\mathrm{Rate}(\mathcal{X},\mathcal{L})^2)\, \|x_1 - x_2\|_{\mathcal{X}} \left(\|x_1\|_{\mathcal{X}} \vee \|x_2\|_{\mathcal{X}}\right).$$

Proof Directly from the relation between optimal rate of online optimization and Equation equation 9, which we state in Theorem 13, with Lemmas 17 and 16. ■

For the regularizer f_0 given by Theorem 18, we define the Gaussian smoothed function $f : \mathbb{R}^d \to \mathbb{R}$:

$$f(x) = \mathbb{E}_{y\sim N(x,\sigma^2 I)} f_0(y). \tag{16}$$

We start by showing that strong convexity property with respect to arbitrary norms is inherited for f_0 to f.

Lemma 19 (Strong convexity of the smoothed function) *If* f_0 *is* α *strongly convex w.r.t* $\|.\|_{\mathcal{L}^c}$*, the* f *is also* α *strong convex w.r.t* $\|.\|_{\mathcal{L}^c}$*.*

Proof From α strong convexity of f, for $0 \leq \gamma \leq 1$ we have

$$f_0(\gamma x_1 + (1-\gamma)x_2) \leq \gamma f_0(x_1) + (1-\gamma) f_0(x_2) - \alpha\frac{\gamma(1-\gamma)}{2}\|x_1 - x_2\|^2.$$

Now consider the gaussian random variable $\eta \sim N(0, \sigma^2 I)$ and write $\tilde{f}(x_1) = \mathbb{E}_\eta f(x_1 + \eta)$, $\tilde{f}(x_2) = \mathbb{E}_\eta f(x_2 + \eta)$. Then

$$\begin{aligned}
f(\gamma x_1 + (1-\gamma)x_2) &= \mathbb{E}_\eta f_0(\gamma x_1 + (1-\gamma)x_2 + \eta) \\
&= \mathbb{E} f_0(\gamma(x_1 + \eta) + (1-\gamma)(x_2 + \eta)) \\
&\leq \gamma \mathbb{E} f_0(x_1 + \eta) + (1-\gamma)\mathbb{E} f_0(x_2 + \eta) - \alpha\frac{\gamma(1-\gamma)}{2}\|x_1 - x_2\|_{\mathcal{L}^c}^2 \\
&= \gamma f_0(x_1) + (1-\gamma) f_0(x_2) - \alpha\frac{\gamma(1-\gamma)}{2}\|x_1 - x_2\|_{\mathcal{L}^c}^2.
\end{aligned}$$

■

Lemma 20 (Strong convexity → Hessian lower bound) *If* f *is twice continuously differentiable and* α *strongly convex with respect to* $\|.\|_{\mathcal{L}^c}$*, then for its hessian at arbitrary point* x *and arbitrary direction* v *we have*

$$v^\top \nabla^2 f(x) v \geq \|v\|_{\mathcal{L}^c}^2. \tag{17}$$

Proof From Taylor series around x_1 at points x_2 and $\gamma x_1 + (1-\gamma)x_2$:

$$f(x_2) = f(x_1) + \langle \nabla f(x_1), x_2 - x_1 \rangle + \frac{1}{2}(x_2 - x_1)^\top \nabla^2 f(x_1)(x_2 - x_1) + o(\|x_2 - x_1\|^2),$$

$$f(\gamma x_1 + (1-\gamma)x_2)$$
$$= f(x_1) + \langle \nabla f(x_1), (1-\gamma)(x_2 - x_1)\rangle + \frac{1}{2}(1-\gamma)^2(x_2 - x_1)^\top \nabla^2 f(x_1)(x_2 - x_1) + o(\|x_2 - x_1\|^2).$$

Therefore

$$\gamma f(x_2) + (1-\gamma) f(x_1) - f(\gamma x_1 + (1-\gamma)x_2) = \frac{1}{2}\gamma(1-\gamma)(x_2 - x_1)^\top \nabla^2 f(x_1)(x_2 - x_1) + o(\|x_2 - x_1\|^2).$$

Therefore, α strong convexity is equivalent to equation 17 for all directions v. ■

Lemma 21 (Norm and norm squared Gaussian integral) *Given a two-homogeneous function f_0 satisfying 1 and $\max_{x\in\mathcal{X}}|f_0(x)| \le C^2$, then for f defined in equation 16*

$$|f(x)| \le \frac{C^2}{r^2}\sigma^2 d + C^2 \|x\|_{\mathcal{X}}^2,$$
$$\mathbb{E}_{y\sim N(x,\sigma^2 I)} f_0(y)^2 \le 8C^4 \left(\|x\|_{\mathcal{X}}^4 + \frac{4}{r^4} d\sigma^4 \right).$$

Proof Note that from the property (1) in Theorem 18 and the 2-homogeneity of f_0, we have for all $y \in \mathbb{R}^d$, $f_0(y) \le C^2 \|y\|_{\mathcal{X}}^2$. Now using triangle inequality and Lemma 22, we can write

$$\begin{aligned} |f(x)| &\le \mathbb{E}_{y\sim N(x.\sigma^2 I)} |f_0(y)| \\ &\le \mathbb{E} C^2 \|y\|_{\mathcal{X}}^2 \\ &\le \mathbb{E} C^2 \|y - x\|_{\mathcal{X}}^2 + C^2 \|x\|_{\mathcal{X}}^2 \\ &\le \mathbb{E} C^2 \frac{1}{r^2} \|y - x\|^2 + C^2 \|x\|_{\mathcal{X}}^2 \\ &= \frac{C^2}{r^2}\sigma^2 d + C^2 \|x\|_{\mathcal{X}}^2. \end{aligned}$$

Furthermore

$$\begin{aligned} \mathbb{E}_y f_0(y)^2 \le \mathbb{E}_y C^4 \|y\|_{\mathcal{X}}^4 &\le 8C^4 \mathbb{E}\left(\|x\|_{\mathcal{X}}^4 + \|y - x\|_{\mathcal{X}}^4 \right) \\ &\le 8C^4 \left(\|x\|_{\mathcal{X}}^4 + \frac{1}{r^4}\mathbb{E}\|y - x\|^4 \right) \\ &\le 8C^4 \left(\|x\|_{\mathcal{X}}^4 + \frac{4}{r^4} d\sigma^4 \right). \end{aligned}$$

■

Lemma 22 (Norm comparison) *The $\|.\|_{\mathcal{X}}$ can be upper bounded by the Euclidean norm $\|.\|$ as*

$$\forall y \in \mathbb{R}^d, \frac{1}{R}\|y\| \le \|y\|_{\mathcal{X}} \le \frac{1}{r}\|y\|.$$

Proof Note that for any $y \in \mathbb{R}^d$, for $\alpha = \|y\| / r$ we have $y/\alpha \in \mathcal{X}$. Therefore, from the definition of $\|.\|_{\mathcal{X}}$:

$$\|y\|_{\mathcal{X}} = \inf\{\alpha > 0, \frac{y}{\alpha} \in \mathcal{X}\} \le \frac{\|y\|}{r}.$$

Furthermore, for $\alpha < \frac{\|y\|}{R}$, then $\left\|\frac{y}{\alpha}\right\| > R$, which means $y \notin \mathcal{X}$ (since $\mathcal{X}$ is contained in a ball of radius R). Therefore, $\|y\|_{\mathcal{X}} \ge \frac{\|y\|}{R}$. ■

Lemma 23 (Gaussian smoothing) *For arbitrary unit direction v, given the smooth regularizer defined in equation 16 we have*

$$\begin{aligned}
|Df(x)[v]| &\le \frac{1}{\sigma}\sqrt{\mathbb{E}f_0(y)^2}\\
|D^2 f[v,v]| &\le \frac{4}{\sigma^2}\sqrt{\mathbb{E}f_0(y)^2},\\
D^3 f(x)[v,w,u] &\le \frac{5}{\sigma^3}\sqrt{\mathbb{E}f_0(y)^2}.
\end{aligned}$$

Proof Consider the function $f_0(y)e^{-\frac{(y-x)^2}{2\sigma^2}}$; it is continuous in both y, x due to continuity of f_0 by Lemma 17, and its partial derivative with respect to x in direction v is $f_0(y)\langle\frac{y-x}{\sigma^2}, v\rangle$ which is again continuous wrt x and y. Therefore, from the Leibnitz rule, for arbitrary direction v, $Df(x)[v]$ exists and is equal to

$$Df(x)[v] = \mathbb{E}_y\langle\frac{y-x}{\sigma^2}, v\rangle f_0(y).$$

Therefore, from Cauchy Schwarz

$$|Df(x)[v]| \le \frac{1}{\sigma^2}\sqrt{\mathbb{E}\langle y-x, v\rangle^2}\sqrt{\mathbb{E}f_0(y)^2} = \frac{1}{\sigma}\sqrt{\mathbb{E}f_0(y)^2}.$$

For the second derivative

$$D^2 f(x)[v,w] = \mathbb{E}_y\left(\langle\frac{y-x}{\sigma^2}, v\rangle\langle\frac{y-x}{\sigma^2}, w\rangle f_0(y) - \frac{1}{\sigma^2}\langle v, w\rangle f_0(y)\right)$$

which gives

$$|D^2 f(x)[v,w]| \le \left(\frac{1}{\sigma^2}\sqrt{\mathbb{E}_{\eta\sim N(0,1)}\eta^4} + \frac{1}{\sigma^2}\right)\sqrt{\mathbb{E}f_0(y)^2} = \frac{4}{\sigma^2}\sqrt{\mathbb{E}f_0(y)^2}.$$

where η is normal gaussian with variance one. Similarly for the third derivative

$$D^3 f(x)[v,w,u] = \mathbb{E}_y\left(\langle\frac{y-x}{\sigma^2}, v\rangle\langle\frac{y-x}{\sigma^2}, w\rangle\langle\frac{y-x}{\sigma^2}, u\rangle f_0(y) - \frac{1}{\sigma^2}\sum_{u,v,w}\langle v, w\rangle\langle\frac{y-x}{\sigma^2}, u\rangle f_0(y)\right).$$

Therefore,

$$\begin{aligned}
|D^3 f(x)[v,w,u]| &\le \left(\frac{1}{\sigma^3}\left(\mathbb{E}\eta^6\right)^{1/2} + \frac{1}{\sigma^3}\sqrt{\mathbb{E}\eta^2}\right)\sqrt{\mathbb{E}f_0(y)^2}\\
&= \frac{1}{\sigma^3}(\sqrt{15}+1)\sqrt{\mathbb{E}f_0(y)^2} \le \frac{5}{\sigma^3}\sqrt{\mathbb{E}f_0(y)^2}.
\end{aligned}$$

■

Corollary 24 (Final smoothed derivatives) *For the smoothed barrier defined in Equation equation 16 and $x \in \mathcal{X}$, we have*

$$
\begin{aligned}
&|f(x)| \le C^2(\frac{\sigma^2}{r^2}d + 1)\\
&|Df(x)[v]| \le \frac{C^2}{\sigma}\sqrt{8\left(1 + 4\frac{1}{r^4}d\sigma^4\right)}\\
&|D^2 f[v,v]| \le \frac{4C^2}{\sigma^2}\sqrt{8\left(1 + 4\frac{1}{r^4}d\sigma^4\right)},\\
&D^3 f(x)[v,w,u] \le \frac{5C^2}{\sigma^3}\sqrt{8\left(1 + 4\frac{1}{r^4}d\sigma^4\right)}.
\end{aligned}
$$

Proof Directly by combining Lemmas 21 and 23. ■

Theorem 25 (Existence of a smooth regularizer) *Given that there exists a 2-homogeneous regularizer $f_0 : \mathbb{R}^d \to \mathbb{R}$ that is α-strongly convex w.r.t $\|.\|_{\mathcal{L}^c}$ and that $\max_{x\in\mathcal{X}}|f_0(x)| \le C^2$, then there also exists a smooth regularizer f which is α-strongly convex w.r.t $\|.\|_{\mathcal{L}^c}$ and*

$$
\begin{aligned}
&|f(x)| = O(C^2),\\
&|Df(x)[v]| = O(C^2\frac{d^{1/4}}{r}),\\
&|D^2 f[v,v]| = O(C^2\frac{d^{1/2}}{r^2}),\\
&|D^3 f(x)[v,w,u]| = O(C^2\frac{d^{3/4}}{r^3}).
\end{aligned}
$$

Proof It is enough to set $\sigma = \frac{r}{d^{1/4}}$ in Corollary 24. ■

We can now prove Theorem 6.

Proof [Proof of Theorem 6] The proof follows from combining Theorems 25 and 18 with Fact 1. ■

Appendix D. Calculating the Regularizer

In this section, building upon the properties that we showed for feasible points of the program 8, we show how to compute a suitable regularizer $g^{(\mathcal{I}^o)}$ on $\mathcal{X}$. To do so, we build a separation oracle for $P_{\mathcal{I}}$. We start by defining the notions of separation oracle, as well as membership and linear optimization oracle. Before defining these oracle, we need to state the definition of set neighborhoods.

Definition 26 (Membership Oracle) *For convex set $\mathcal{D} \in \mathbb{R}^d$, a membership oracle receives a vector $y \in \mathbb{R}^d$ and real number $\delta > 0$ and with probability $1-\delta$ asserts $y \in B(\mathcal{D}, \delta)$, or it asserts $y \notin B(\mathcal{D}, -\delta)$. We denote the computational cost of a query to our membership oracle by* $\mathrm{MEM}_{\mathcal{X}}(\delta)$.

Definition 27 (Set neighborhoods) *For a subset $\mathcal{D} \subseteq \mathbb{R}^d$, let $B(\mathcal{D}, \delta)$ be the set of points that are within distance δ of $\mathcal{D}$, and $B(\mathcal{D}, -\delta)$ be the set of points that where a ball of radius δ around them is completely included in $\mathcal{D}$.*

Definition 28 (Separation Oracle) *For a convex set $\mathcal{L} \subseteq \mathbb{R}^d$, a separation oracle receives a vector $y \in \mathbb{R}^d$ and real number $\delta > 0$ and either asserts $y \in B(\mathcal{L}, \delta)$, or it returns a unit vector $c \in \mathbb{R}^d$ such that $c^\top y \le c^\top x + \delta$ for all $x \in B(\mathcal{L}, -\delta)$. We denote the computation time of separation oracle by $\mathrm{SEP}_{\mathcal{L}}(\delta)$.*

Definition 29 (Linear Optimization Oracle) *For a convex set $\mathcal{L} \subset \mathbb{R}^d$, a linear optimization oracle receives a unit vector $c \in \mathbb{R}^d$ and real number δ_{lin} and returns a point $y \in \mathcal{C}$ such that $\forall x \in \mathcal{C}$, $c^\top y \le c^\top x + \delta_{lin}$. We denote the computational cost of calling the linear optimization oracle by $\mathrm{LinO}_{\mathcal{L}}\left(\delta_{lin}\right)$.*

Separation, membership, and linear optimization oracles are known to be equivalent and can be used to implement convex optimization over convex sets (Grötschel et al., 2012). Next, we state a simplified version of Theorem 42 in Lee et al. (2018) (or Theorem 15 in Lee et al. (2015)) on how to build a linear optimization oracle from a separation oracle for a convex set, which we use in the proof of Theorem 31.

Theorem 30 (Theorem 15 in Lee et al. (2018) or Theorem 42 in Lee et al. (2015)) *Let K be a convex set satisfying $B_2(0, r) \subset K \subset B_2(0, 1)$ and let $\kappa = \frac{1}{r}$. For $0 \le \epsilon < 1$, with probability $1 - \epsilon$, we can compute $x \in B(K, \epsilon)$ such that*

$$c^\top x \le \min_{x \in K} c^\top x + \epsilon \|c\|_2$$

with an expected running time of $O\left(nSEP_\delta(K)\log(\frac{n\kappa}{\epsilon}) + n^3 \log^{O(1)}\left(\frac{n\kappa}{\epsilon}\right)\right)$, where $\delta = \left(\frac{\epsilon}{n\kappa}\right)^{\Theta(1)}$.

We now state how we solve the optimization problem in Theorem equation 8 based on a separation oracle that we build for its feasibility set $P_{\mathcal{I}}$ in Section F.

Theorem 31 (Computing the Regularizer - abstract) *In the context of Lemma 11 Then, given arbitrary accuracy parameter $0 < \epsilon_1 < 1$, there is a cutting-plane method that approximately solves the program in equation 8 and obtains an almost feasible instance $\mathcal{I}^o$, in the sense that*

1. $\max_{x \in \mathcal{X}} |g^{(\mathcal{I}^o)}(x)| \le C^2 + \gamma_2 d \tilde{c}_1 \epsilon + \epsilon_1$
2. $g^{(\mathcal{I}^o)}(x)$ *is $\alpha/4$ strongly convex with respect to $\|.\|_{\mathcal{L}^c}$,*

and runs in time (assuming $N \ge d$)

$$O\left(\frac{N\left({C_0}^2 + c_1^2 + c_2^2\right)(c_2 \vee 1) R}{\epsilon_1 \epsilon L r}\right)^{O(d)} \left(\mathrm{LinO}_{\mathcal{L}}\left(\frac{(r \wedge 1)}{R^2 \alpha}\left(\frac{\epsilon_1 \bar{\epsilon} L}{N\left({C_0}^2 + c_0^2 + c_2^2\right)}\right)^{\Theta(1)}\right)\right).$$

Proof The program equation 8 is a linear optimization problem over the convex set $P_{\mathcal{I}}$, for which we can exploit the separation oracle that we constructed in Lemma 34. In particular, the result directly follows from a simplified version of Theorem 42 in Lee et al. (2015) (or Theorem 15 in Lee et al. (2018)), a classical result on how to build a linear optimization oracle from the separation oracle for a convex set. For convenience of the reader, we have restated this result in Theorem 30. According to this theorem, for any $0 < \epsilon_1 < 1$, with probability $1 - \epsilon_1$ we can compute an instance $\mathcal{I}^o$ such that its corresponding barrier $g^{(\mathcal{I}^o)}$ satisfies

1. $\max_{x \in \mathcal{X}} |g^{(\mathcal{I}^o)}(x)| \leq \max_{x \in \mathcal{X}} |g^{(\mathcal{I}^*)}(x)| + \epsilon_1$, where $\mathcal{I}^*$ is the optimal solution to the LP.
2. $\mathcal{I}^o$ is ϵ_1 close to a feasible instance $\mathcal{I}^{(r)}$ in Euclidean distance.

Now applying Lemma 11 we conclude the first argument, namely $\max_{x \in \mathcal{X}} |g^{(\mathcal{I}^o)}(x)| \leq C^2 + \gamma_2 d\tilde{c}_1 \epsilon + \epsilon_1$. Now we need to show that $g^{(\mathcal{I}^o)}$ roughly remains $\Omega(\alpha)$ strongly convex w.r.t $\|.\|_{\mathcal{L}^c}$. For this, note that given $x \in \mathcal{X}$, if $\|x_j - x\| \geq \gamma \left(\frac{\epsilon\sqrt{d}c_0}{L}\right)^{1/3}$ and $\|x_i - x\| \leq \epsilon$, then from Lemma 9 and the feasibility of $\mathcal{I}^{(r)}$ we have $r_{x_i} > g_{x_j}^{(\mathcal{I}^r)}(x) + \sqrt{d}c_0\epsilon$ where r_{x_i} is the variable of the valid instance $\mathcal{I}^r$. But picking $\epsilon_1 \leq \frac{\sqrt{d}c_0\epsilon}{2R^2}$ we get that $g_{x_i}^{(\mathcal{I}^o)}(x) > g_{x_j}^{(\mathcal{I}^o)}(x)$. Therefore, again the maximum at x is achieved by one of the functions $g_{x_j}^{(\mathcal{I}^o)}(x)$ where x_j is not farther than $\gamma \left(\frac{\epsilon\sqrt{d}c_0}{L}\right)^{1/3}$ of x. But then similar to Equation equation 54 in Lemma 35, for all $\hat{i} \in I$ and arbitrary direction v:

$$v^\top \nabla^2 g_{x_{\hat{i}}}^{(\mathcal{I}^r)}(x) v \geq \frac{\alpha}{2} \|v\|_{\mathcal{L}^c}^2 .$$

On the other hand, $\|\mathcal{I}^o - \mathcal{I}^r\| \leq \epsilon_1$ implies $\left\|\nabla^2 g_{x_{\hat{i}}}^{(\mathcal{I}^r)}(x) - \nabla^2 g_{x_{\hat{i}}}^{(\mathcal{I}^o)}(x)\right\|_F \leq \epsilon_1$. Therefore, using $\epsilon_1 \leq \frac{\alpha}{4r^2}$ we conclude

$$v^\top \nabla^2 g_{x_{\hat{i}}}^{(\mathcal{I}^o)}(x) v \geq \frac{\alpha}{4} \|v\|_{\mathcal{L}^c}^2 ,$$

which is the desired property. Finally, using the third argument in Lemma 11, we have the following runtime based on Theorem 30:

$$O\left(N \cdot \mathrm{SEP}_{P_{\mathcal{I}}}(\delta) \log\left(\frac{1}{\delta}\right) + N^3 \log^{O(1)}\left(\frac{1}{\delta}\right)\right),$$

for

$$\delta \triangleq \left(\frac{\epsilon_1 L \bar{\epsilon}^3}{N\sqrt{(N+1){C_0}^2 + Nd\left(c_0^2 + c_2^2\right)}}\right)^{\Theta(1)} = \left(\frac{\epsilon_1 \bar{\epsilon} L}{N\left({C_0}^2 + c_0^2 + c_2^2\right)}\right)^{\Theta(1)} .$$

Note that from Lemma 34, for this choice of δ we have

$$\mathrm{SEP}_{P_{\mathcal{I}}}(\delta) = O\left(\frac{N\left({C_0}^2 + c_1^2 + c_2^2\right)c_2 R}{\epsilon_1 \epsilon L r}\right)^{O(d)} \left(\mathrm{LINO}_{\mathcal{L}}\left(\frac{(r \wedge 1)}{R^2\alpha}\left(\frac{\epsilon_1 \bar{\epsilon} L}{N\left({C_0}^2 + c_0^2 + c_2^2\right)}\right)^{\Theta(1)}\right) + d^2\right) + O\left(N^2 d^2\right),$$

which completes the proof. ■

Next, we appropriately instantiate the constants of the convex program equation 8 based on Theorem 25 and Lemma 11 in Theorem 4 below. We find the running time of our cutting-plane method to solve this program based on Theorem 31.

Theorem 32 (Restatement of Theorem 4) *Assuming $R > 1, r < 1$ for simplicity, given that the best achievable rate for online linear optimization with action and constraint sets $(\mathcal{X}, \mathcal{L})$ is $O(\mathrm{Rate}(\mathcal{X}, \mathcal{L})\sqrt{T})$, there exists an algorithm that runs in time*

$$\left(\frac{dR}{r}\right)^{O(d^2)} \left(\mathrm{LinO}_{\mathcal{L}}\left(\left(\frac{r}{dR}\right)^{\Theta(d)}\right)\right),$$

and calculates a regularizer $g^{(\mathcal{I}^o)}$ given by the representation $(\mathbf{\Sigma}, \mathbf{v}, \mathbf{r})$ as described in Section equation 7, which satisfies

1. $\sup_{x\in\mathcal{X}}|g^{(\mathcal{I}^o)}| \leq 2\mathrm{Rate}(\mathcal{X}, \mathcal{L})^2$
2. $g^{(\mathcal{I}^o)}$ *is* 1*-strongly convex w.r.t* $\|.\|_{\mathcal{L}^c}$.

Proof Let $C \triangleq \mathrm{Rate}(\mathcal{X}, \mathcal{L})$. From Theorem 25 there exists a 2-homogeneous barrier which is $\tilde{c}_1 = O(C^2\frac{d^{1/4}}{r})$ Lipschitz, $\tilde{c}_2 = O(C^2\frac{d^{1/2}}{r^2})$ Gradient Lipschitz, $L = O(C^2\frac{d^{3/4}}{r^3})$ Hessian Lipschitz, and 1-strongly convex w.r.t $\|.\|_{\mathcal{L}^c}$. Therefore, to enjoy the properties of Lemma 11, assuming that we guarantee,

$$\bar{\epsilon}^3 \leq \min\{\frac{\tilde{c}_1}{L}, \frac{\tilde{c}_2}{L}, \frac{C^2}{L}\} \tag{18}$$

then we get that c_0, c_2, C_0 are of the same order as $\tilde{c}_1, \tilde{c}_2, C^2$, respectively (this follows from the definition of c_0, c_2, C_0 which involves the term $L\bar{\epsilon}^3$). Now following the condition of Lemma 11, we consider a cover of accuracy ϵ such that

$$\epsilon \leq \min\{\frac{1}{r^2}, \frac{r^6}{C^6 d^2}, \frac{r}{d^{1/4}}, C\frac{d^{3/8}}{r^{1/2}}, rd^{1/4}, \frac{r^7}{R^6 C^6 d^{11/8}}\}.$$

where we set $L = \gamma_5 C^2\frac{d^{3/4}}{r^3}$ for small enough constant γ_5. For simplicity if either R or C were smaller than one, we upper bound them by one, so we can assume $R, C \geq 1$ without loss of generality. Similarly if $r < 1$, we can take $r = 1$, so without loss of generality we assume $r = 1$. Then, the above bound simplifies to

$$\epsilon \leq \frac{r^6}{R^6 C^6 d^2}. \tag{19}$$

Furthermore we consider the discretization set $\tilde{\mathcal{X}}$ to be points each entry is of the form $k\bar{\epsilon}$ for an integer k. Then, to guarantee equation 18 we should have

$$\bar{\epsilon}^3 \leq \frac{r^2}{d^{1/2}}. \tag{20}$$

On the other hand, rounding every point x to its closest multiple of $\bar{\epsilon}$ in each coordinate implies that the cover has accuracy as small as $\epsilon = \sqrt{d}\bar{\epsilon}$. Hence, to satisfy condition equation 19 we set

$$\bar{\epsilon} \triangleq \frac{\gamma_4 r^6}{R^6 C^6 d^2\sqrt{d}},$$
$$\epsilon \triangleq \frac{\gamma_4 r^6}{R^6 C^6 d^2},$$

for small enough constant γ_4. Then, it is easy to check that condition equation 20 is automatically satisfied. Furthermore, with this choice of $\bar{\epsilon}$ we see that $\gamma_2 d\tilde{c}_1\epsilon \leq \frac{C^2}{2}$ for small enough constant γ_4 (γ_2 is defined in Lemma 31); hence, from the guarantee of Lemma 31

$$\max_{x\in\mathcal{X}}|g^{(\mathcal{I}^o)}(x)| \leq C^2 + \gamma_2 d\tilde{c}_1\epsilon + \epsilon_1 \leq \frac{3}{2}C^2 + \epsilon_1,$$

where recall ϵ_1 is the accuracy parameter for our solver in Lemma 31. Setting

$$\epsilon_1 = \frac{C^2}{2},$$

we conclude

$$\max_{x\in\mathcal{X}}|g^{(\mathcal{I}^o)}(x)| \leq 2C^2.$$

Note that the attained constant two behind C^2 does not matter since the parameter C of the smoothed barrier in Theorem 25 can be off by a universal constant from $\mathrm{Rate}(\mathcal{X},\mathcal{L})$. Now since the regularizer $\tilde{f}$ is $\alpha = 1$ strongly convex, Lemma 31 also guarantees that the regularizer that we find, $g^{(\mathcal{I}^o)}(x)$, is $\frac{1}{4}$ strongly-convex with respect to $\|.\|_{\mathcal{L}^c}$. Finally from the runtime guarantee of Lemma 31, finding such regularizer has runtime

$$O\left(\frac{NR}{r}\right)^{O(d)}\left(\mathrm{LINO}_{\mathcal{L}}\left(\left(\frac{r}{NR}\right)^{\Theta(1)}\right)\right),$$

where we used the fact that $C_0^2 + c_1^2 + c_2^2 = O(C^4R^4d^2)$ and $d \leq N$, and that we can upper bound C by R (Note that we dropped the d in the term $\frac{NRd}{r}$ since N is already exponentially large in d). Furthermore, the cover that we considered has size at most $N = |\tilde{\mathcal{X}}| = O\left(\frac{R}{\epsilon}\right)^d = \left(\frac{dR}{r}\right)^{O(d)}$. Therefore, the overall runtime is

$$\left(\frac{dR}{r}\right)^{O(d^2)}\left(\mathrm{LINO}_{\mathcal{L}}\left(\left(\frac{r}{dR}\right)^{\Theta(d)}\right)\right).$$

■

Appendix E. Online Linear Optimization

Here we show how to run FTRL with regularizer $g^{\mathcal{I}^o}$ that is based on the instance $\mathcal{I}^o$ which we computed in Section D for a general instance of the online linear optimization problem as we defined in Section 3.1; as we mentioned, our approach results in the optimal information theoretic rate up to universal constants.

Theorem 33 (Optimal online optimization) *Consider the problem of online linear optimization with action and loss sets $(\mathcal{X},\mathcal{L})$ as described in Section 3.1. Given access to the regularizer $g^{\mathcal{I}^o}$ for the instance $\mathcal{I}^o$ of the program 8 that we can compute as described in Theorem 4 and a membership oracle for $\mathcal{X}$, there is a cutting-plane algorithm to run FTRL with regularizer $g^{\mathcal{I}^o}$, with running time*

$$O\left(Td^2\ln^{O(1)}(dRT)\left(\mathbf{MEM}_{\mathcal{X}}(\delta)+1\right)\right),$$

which guarantees regret $O(\mathrm{Rate}(\mathcal{X},\mathcal{L})\sqrt{T})$.

Proof We run FTRL with the regularizer $g^{(\mathcal{I}^o)}$; namely, to calculate each step $1 \leq t \leq T$, we solve the following convex optimization using separation oracle for $\mathcal{X}$:

$$x_t = \arg\min_{x \in \mathcal{X}} G_t(x) \tag{21}$$

$$G_t(x) \triangleq \langle x, \sum_{s=1}^{t-1} g_s \rangle + g^{\mathcal{I}^o}(x), \tag{22}$$

up to accuracy $O(\frac{\alpha r}{R^2 T})$, namely for $\tilde{x}_t$ being the output of the algorithm we have

$$G_t(\tilde{x}_t) - G_t(x_t) \leq O(\frac{\alpha r}{R^2 T})|\sup_{x \in \mathcal{X}} G_t(x) - \inf_{x \in \mathcal{X}} G_t(x)| = O(\frac{\alpha r}{R^2 T}\mathrm{Rate}(\mathcal{X}, \mathcal{L})^2). \tag{23}$$

Note that we used the property that for the regularizer $g^{\mathcal{I}^o}$ that we calculate in Theorem 31 we have $\sup_{x \in \mathcal{X}} |g^{(\mathcal{I}^o)}| \leq 2\mathrm{Rate}(\mathcal{X}, \mathcal{L})^2$. Then, from Theorem 1 in Lee et al. (2018), there is a cutting-plane method whose number of queries to a membership oracle for $\mathcal{X}$ is

$$O\left(d^2 \ln^{O(1)}(dRT)\right)$$

in addition to $O\left(d^2 \ln^{O(1)}(dRT)\right)$ arithmetic operations.

But since x_t is the global minimizer of G_t we have $\nabla G_t(x_t) = 0$, and further from $\alpha/4$ strong convexity of G_t w.r.t. $\|.\|_{\mathcal{L}^c}$:

$$G_t(\tilde{x}_t) - G_t(x_t) \geq \frac{\alpha}{4}\|x_t - \tilde{x}_t\|_{\mathcal{L}^c}^2 \geq \frac{\alpha r}{4}\|x_t - \tilde{x}_t\|^2,$$

which combined with equation 23 implies

$$\|x_t - \tilde{x}_t\| \leq \frac{\mathrm{Rate}(\mathcal{X}, \mathcal{L})}{R\sqrt{T}}.$$

Then, from the mirror descent guarantee we have the following regret bound for the sequence x_t

$$\mathbb{E}\left(\max_{x^* \in \mathcal{X}} \sum_{t=1}^{T} \langle x_t, g_t \rangle - \langle x^*, g_t \rangle\right) = O(\mathrm{Rate}(\mathcal{X}, \mathcal{L})\sqrt{T}). \tag{24}$$

On the other hand, using the fact that $\|g_t\| \leq R$ and that $\mathcal{L} \subseteq B_R(0)$,

$$\begin{aligned}
&\mathbb{E}\left(\sum_{t=1}^{T} \langle x_t, g_t \rangle - \langle \tilde{x}_t, g_t \rangle\right) \\
&\mathbb{E}\left(\sum_{t=1}^{T} \|x_t - \tilde{x}_t\| \, \|g_t\|\right) \\
&\leq \mathrm{Rate}(\mathcal{X}, \mathcal{L})\sqrt{T}.
\end{aligned} \tag{25}$$

Combining equation 24 and equation 25 completes the proof for the regret guarantee. ■

Appendix F. Separation Oracle

Here we show a separation oracle for the feasible polytope $P_{\mathcal{I}}$ of program 8.

Lemma 34 (Linear optimization oracle for $\mathcal{L} \to$ Separation Oracle) *The polytope $P_{\mathcal{I}}$ for $\mathcal{I} = (\mathbf{r}, \mathbf{v}, \mathbf{\Sigma})$ defined in equation 8 has a separation oracle with computational cost*

$$\mathrm{SEP}_K(\delta) = O\left(\frac{2c_2 R^3}{\delta r^3}\right)^d \left(\mathrm{L{\small INO}}_{\mathcal{L}}\left(\delta\left(1 \wedge r\right)/(8\alpha R^2)\right) + d^2\right) + O\left(|\tilde{\mathcal{X}}|^2 d^2\right),$$

where $\mathrm{L{\small INO}}_{\mathcal{L}}\left(\delta\left(1 \wedge r\right)/(8\alpha R^2)\right)$ *is the cost of a linear optimization oracle for $\mathcal{L}$ with parameter* $\delta = \left(1 \wedge r\right)/(8\alpha R^2)$.

Proof We can readily check if conditions (1) and (2) hold for the instance $\mathcal{I}$, and if not, that condition defines the direction c for which $\langle \mathcal{I}, c\rangle \geq \langle \tilde{\mathcal{I}}, c\rangle$ for all $\tilde{\mathcal{I}} \in P_{\mathcal{I}}$. To check condition (3) we can do singular value decomposition in $O(d^3)$ Condition (4) is a bit trickier since it might be hard to directly maximize $v^\top \Sigma_{x_i} v$ over $\mathcal{L}$. Therefore, we work with the discretization set $\tilde{S}_d$ of the unit d-dimensional sphere; in particular, for every unit direction $\tilde{v} \in \tilde{S}_d$, we consider condition (5) with a margin δ_m, namely

$$v^\top \Sigma_{x_i} v / \left\|v\right\|_{\mathcal{L}^c}^2 \geq \alpha(1 + \delta_m). \tag{26}$$

This margin allows us to easily obtain a feasible solution in $P_{\mathcal{I}}$ which satisfies $v^\top \Sigma_{x_i} v \geq \alpha$ for all $v \in \mathcal{L}$, using condition in equation 26 which is only for the discretization points; moreover, we check equation 26 with our linear optimization oracle which has error δ_{lin} in calculating $\|v\|_{\mathcal{L}^c}$; namely, suppose equation 26 holds for all $\tilde{v} \in \tilde{S}_d$ given that we substitute $\|v\|_{\mathcal{L}^c}$ in equation 26 with the output of $\mathrm{L{\small INO}}_{\mathcal{L}}\left(\delta_{lin}\right)$. Then, we are guaranteed that for every $\tilde{v} \in \tilde{S}^d$:

$$\tilde{v}^\top \Sigma_{x_i} \tilde{v} / \left(\mathrm{L{\small INO}}_{\mathcal{L}}\left(\delta_{lin}\right)[\tilde{v}]\right)^2 \geq \alpha(1 + \delta_m). \tag{27}$$

Now from the fact that $\|\tilde{v}\|_{\mathcal{L}^c} \geq r$ and $\mathrm{L{\small INO}}_{\mathcal{L}}\left(\delta_{lin}\right)[\tilde{v}] \geq \|\tilde{v}\|_{\mathcal{L}^c} - \delta_{lin}$, picking $\delta_{lin} \leq \frac{r\delta_m}{2}$, we get that

$$\tilde{v}^\top \Sigma_{x_i} \tilde{v} / \left(\left(1 - \delta_{lin}/2\right) \|\tilde{v}\|_{\mathcal{L}^c}\right)^2 \geq \alpha(1 + \delta_m), \tag{28}$$

which using the fact that we picked $\delta_{lin} \leq \delta_m/4$ implies

$$\tilde{v}^\top \Sigma_{x_i} \tilde{v} / \left(\|\tilde{v}\|_{\mathcal{L}^c}\right)^2 \geq \left(1 - \delta_{lin}/2\right)^2 \alpha(1 + \delta_m) \geq \alpha\left(1 + \delta_m/2\right). \tag{29}$$

Now for arbitrary direction $v \in S^d$ on the unit sphere, we bound the value of the quadratic form the closest point in the discretization set: namely for $\tilde{v} \in \tilde{S}^d$ where $\|\tilde{v} - v\| \leq \tilde{\epsilon}$:

$$\begin{aligned} &|v^\top \Sigma_{x_i} v / \|v\|_{\mathcal{L}^c}^2 - \tilde{v}^\top \Sigma_{x_i} \tilde{v} / \|\tilde{v}\|_{\mathcal{L}^c}^2| \\ &= |v^\top \Sigma_{x_i} v / \|v\|_{\mathcal{L}^c}^2 - \tilde{v}^\top \Sigma_{x_i} \tilde{v} / \|v\|_{\mathcal{L}^c}^2| + |\tilde{v}^\top \Sigma_{x_i} \tilde{v} / \|v\|_{\mathcal{L}^c}^2 - \tilde{v}^\top \Sigma_{x_i} \tilde{v} / \|\tilde{v}\|_{\mathcal{L}^c}^2|. \end{aligned} \tag{30}$$

but for the first term, using $\|\Sigma_{x_i}\| \leq c_2$:

$$|v^\top \Sigma_{x_i} v - \tilde{v}^\top \Sigma_{x_i} \tilde{v}| \leq |(v - \tilde{v})^\top \Sigma_{x_i} v| + |(v - \tilde{v})^\top \Sigma_{x_i} \tilde{v}| \qquad \leq 2c_2 \|v - \tilde{v}\| \leq 2c_2 \tilde{\epsilon}$$

and $\|v\|_{\mathcal{L}^c} \geq r$. Hence, from $\tilde{\epsilon} < 1$

$$|v^\top \Sigma_{x_i} v / \|v\|_{\mathcal{L}^c}^2 - \tilde{v}^\top \Sigma_{x_i} \tilde{v} / \|v\|_{\mathcal{L}^c}^2| \leq 2c_2 \frac{\tilde{\epsilon}}{r^2}. \tag{31}$$

For the second term, using the fact that $r \leq \|\tilde{v}\|_{\mathcal{L}^c}, \|v\|_{\mathcal{L}^c} \leq R$ and $\|\tilde{v} - v\|_{\mathcal{L}^c} \leq R \|\tilde{v} - v\|$:

$$\begin{aligned} |\tilde{v}^\top \Sigma_{x_i} \tilde{v} / \|v\|_{\mathcal{L}^c}^2 - \tilde{v}^\top \Sigma_{x_i} \tilde{v} / \|\tilde{v}\|_{\mathcal{L}^c}^2| &\leq c_2 \frac{\|\tilde{v}\|_{\mathcal{L}^c}^2 - \|v\|_{\mathcal{L}^c}^2}{\|\tilde{v}\|_{\mathcal{L}^c}^2 \|v\|_{\mathcal{L}^c}^2} \\ &\leq c_2 \frac{\|\tilde{v} - v\|_{\mathcal{L}^c} (\|v\|_{\mathcal{L}^c} + \|\tilde{v}\|_{\mathcal{L}^c})}{\|\tilde{v}\|_{\mathcal{L}^c}^2 \|v\|_{\mathcal{L}^c}^2} \\ &= c_2 \frac{\|\tilde{v} - v\|_{\mathcal{L}^c}}{\|\tilde{v}\|_{\mathcal{L}^c} \|v\|_{\mathcal{L}^c}^2} + c_2 \frac{\|\tilde{v} - v\|_{\mathcal{L}^c}}{\|\tilde{v}\|_{\mathcal{L}^c}^2 \|v\|_{\mathcal{L}^c}} \\ &\leq \frac{2\tilde{\epsilon} c_2 R}{r^3}. \end{aligned} \tag{32}$$

Combining Equations equation 31 and equation 32 (from $R/r \geq 1$) and plugging into equation 30

$$|v^\top \Sigma_{x_i} v / \|v\|_{\mathcal{L}^c}^2 - \tilde{v}^\top \Sigma_{x_i} \tilde{v} / \|\tilde{v}\|_{\mathcal{L}^c}^2| \leq \frac{4\tilde{\epsilon} c_2 R}{r^3},$$

which combined with equation 29 and triangle inequality

$$v^\top \Sigma_{x_i} v / (\|v\|_{\mathcal{L}^c})^2 \geq \alpha (1 + \delta_m/2) - \frac{4\tilde{\epsilon} c_2 R}{r^3}.$$

Using $\tilde{\epsilon} \leq \frac{\alpha r^3 \delta_m}{c_2 R}$, we get

$$v^\top \Sigma_{x_i} v / (\|v\|_{\mathcal{L}^c})^2 \geq \alpha(1 + \delta_m/4).$$

Recall that v was arbitrary in S^d. Therefore, in the case when all inequalities in equation 27 are satisfied, we showed that $\mathcal{I}$ indeed satisfies condition (4) in equation 8. Finally if any of the inequalities equation 28 are violated, i.e. if $\tilde{v}^\top \Sigma_{x_i} \tilde{v} / (\mathrm{LINO}_{\mathcal{L}}(\delta_{lin})[\tilde{v}])^2 \geq \alpha(1 + \delta_m)$, then similar to equation 28 we get

$$\tilde{v}^\top \Sigma_{x_i} \tilde{v} / ((1 + \delta_{lin}/2) \|\tilde{v}\|_{\mathcal{L}^c})^2 \leq \alpha(1 + \delta_m) \leq \tilde{v}^\top \Sigma_{x_i} \tilde{v} / (\mathrm{LINO}_{\mathcal{L}}(\delta_{lin})[\tilde{v}])^2 \leq \alpha(1 + \delta_m),$$

which implies (from $\delta_{lin} \leq \delta_m/4$)

$$\tilde{v}^\top \Sigma_{x_i} \tilde{v} / (\|\tilde{v}\|_{\mathcal{L}^c})^2 \leq \alpha (1 + \delta_{lin}/2)^2 (1 + \delta_m) \leq \alpha(1 + 2\delta_m).$$

Therefore, we find that the unit direction $\tilde{v}\tilde{v}^\top$ which satisfies

$$\begin{aligned} \langle \tilde{v}\tilde{v}^\top, \Sigma_{x_i} \rangle &\leq \alpha \|\tilde{v}\|_{\mathcal{L}^c}^2 + 2\alpha\delta_m \|\tilde{v}\|_{\mathcal{L}^c}^2 \\ &\leq \alpha \|\tilde{v}\|_{\mathcal{L}^c}^2 + 2\alpha\delta_m R^2, \end{aligned}$$

while for a valid $\mathcal{I} \in \mathcal{P}_{\mathcal{I}}$, we should have $\langle vv^\top, \Sigma_{x_i} \rangle \geq \alpha \|v\|_{\mathcal{L}^c}^2$ for all unit directions v. Hence, we constructed a separation oracle with $2\alpha\delta_m R^2$, which uses $|\tilde{S}^d|$ queries to the linear optimization oracle, and its overal computational cost is $O\left(|\tilde{S}^d|\left(\text{LINO}_{\mathcal{L}}\left(\delta_{lin}\right)+d^2\right)+|\tilde{\mathcal{X}}|^2d^2\right)$. Finally to have a δ-separation oracle, we need to guarantee $2\alpha\delta_m R^2 \leq \delta$, $\delta_{lin} \leq \frac{\delta_m}{4} \wedge \frac{r\delta_m}{2}$, $\tilde{\epsilon} \leq \frac{\alpha r^3 \delta_m}{c_2 R}$, hence we pick

$$\begin{aligned}
\delta_m &\triangleq \frac{\delta}{2\alpha R^2},\\
\delta_{lin} &\triangleq \frac{\delta_m\left(1\wedge r\right)}{4} = \frac{\delta\left(1\wedge r\right)}{8\alpha R^2},\\
\tilde{\epsilon} &\triangleq \frac{r^3\delta}{2c_2R^3}.
\end{aligned}$$

Hence, the overall computational cost is

$$\begin{aligned}
&O\left(1/\tilde{\epsilon}\right)^d\left(\text{LINO}_{\mathcal{L}}\left(\delta\left(1\wedge r\right)/(8\alpha R^2)\right)+d^2\right)+O\left(|\tilde{\mathcal{X}}|^2d^2\right)\\
&=O\left(\frac{2c_2R^3}{\delta r^3}\right)^d\left(\text{LINO}_{\mathcal{L}}\left(\delta\left(1\wedge r\right)/(8\alpha R^2)\right)+d^2\right)+O\left(|\tilde{\mathcal{X}}|^2d^2\right).
\end{aligned}$$

■

Appendix G. Proofs for Sections 5 and 7

G.1. Proof of Lemma 7

For the lower bound, we use the inequality $\nabla^2 f(x_1) \succcurlyeq \nabla^2 f(x_0) - L\left\|x_1 - x_0\right\| I$:

$$\begin{aligned}
f(x) &= f(x_0) + \langle \nabla f(x_0), x - x_0\rangle + \int_0^1\int_0^t (x-x_0)^\top \nabla^2 f(x_0 + s(x-x_0))(x-x_0)dsdt\\
&\geq f(x_0) + \langle \nabla f(x_0), x - x_0\rangle + \int_0^1\int_0^t (x-x_0)^\top\left(\nabla^2 f(x_0) - sL\|x-x_0\|I\right)(x-x_0)dsdt\\
&= f(x_0) + \langle \nabla f(x_0), x - x_0\rangle + \frac{1}{2}(x-x_0)^\top\nabla^2 f(x_0)(x-x_0) - \frac{L}{6}\|x-x_0\|^3\\
&= f_{x_0}(x) + \frac{L}{6}\left\|x-x_0\right\|^3.
\end{aligned}$$

For upper bound, we use the inequality $\nabla^2 f(x_1) \preccurlyeq \nabla^2 f(x_0) + L\left\|x_1 - x_0\right\| I$:

$$\begin{aligned}
f(x) &\leq f(x_0) + \langle \nabla f(x_0), x - x_0\rangle + \int_0^1\int_0^t (x-x_0)^\top(\nabla^2 f(x_0) + sL\|x-x_0\|I)(x-x_0)dsdt\\
&= f(x_0) + \langle \nabla f(x_0), x - x_0\rangle + \frac{1}{2}(x-x_0)^\top\nabla^2 f(x_0)(x-x_0) + \frac{L}{6}\|x-x_0\|^3\\
&= f_{x_0}(x) + \frac{L}{2}\left\|x-x_0\right\|^3.
\end{aligned}$$

G.2. Proof of Lemma 9

We denote $g^{(\mathcal{I})}_{x_i}(x)$ in short by $g_{x_i}(x)$, and without loss of generality let $x_i = x_0$ and $x_j = x_1$. First, note that we can translate the convex program conditions on the norm of v_{x_i} to

$$\|v_{x_i}\| \le c_1,$$

for $c_1 = \sqrt{d}c_0$. From the program constraint we have

$$g_{x_1}(x_0) + \frac{15L}{96}\|x_1 - x_0\|^3 \le r_{x_0}. \tag{33}$$

On the other hand, from $\|x_0 - x\| \le \epsilon$ and the norm bounds on gradient and Hessian

$$|g_{x_1}(x_0) - g_{x_1}(x)| \le |v_{x_1}^\top(x_0 - x)| + |(x_0 - x)^\top \Sigma_{x_0}(x_0 + x - 2x_1)| + \frac{L}{3}|\|x_1 - x_0\|^3 - \|x_1 - x\|^3| \tag{34}$$

$$\le c_1\|x_0 - x\| + c_2\|x_0 - x\|(2\|x_0 - x_1\| + \|x_0 - x\|) \tag{35}$$

$$+ \frac{L}{3}\|x_0 - x\|\Big(\|x_1 - x_0\|^2 + \|x_1 - x\|^2 + \|x_1 - x_0\|\|x_1 - x\|\Big), \tag{36}$$

$$\le c_1\|x_0 - x\| + c_2\|x_0 - x\|(2\|x_0 - x_1\| + \|x_0 - x\|) \tag{37}$$

$$+ \frac{L}{3}\|x_0 - x\|\Big(4\|x_1 - x_0\|^2 + 2\|x_0 - x\|^2\Big), \tag{38}$$

where in the last line we used

$$\|x_1 - x_0\|^2 + \|x_1 - x\|^2 + \|x_1 - x_0\|\|x_1 - x\| \le 2\|x_1 - x_0\|^2 + 2\|x_1 - x\|^2 \tag{39}$$

$$\le 4\|x_1 - x_0\|^2 + 2\|x_0 - x\|^2. \tag{40}$$

Note that picking $\gamma \ge 3$, from the triangle inequality, $\|x - x_1\| \ge 3\left(\frac{\epsilon c_1}{L}\right)^{1/3}$, and the condition that $\frac{\epsilon\sqrt{d}c_0}{L} \le 1$,

$$\|x_0 - x_1\| \ge \|x_1 - x\| - \|x - x_0\| \ge 2\left(\frac{\epsilon c_1}{L}\right)^{1/3}. \tag{41}$$

Now based on equation 41, for the first term in equation 38, we can write

$$c_1\|x_0 - x\| \le c_1\epsilon \le \frac{L}{48}\|x_1 - x_0\|^3, \tag{42}$$

Similarly, also because $\epsilon \le \frac{L}{2000c_1c_2^3}$, for the second term we have

$$2c_2\|x_0 - x\|\|x_0 - x_1\| \le \frac{L}{24}\|x_0 - x_1\|^3, \tag{43}$$

and because $\epsilon \le \frac{8L}{c_2}$,

$$2c_2\|x_0 - x\|^2 \le \frac{L}{24}\|x_0 - x_1\|^3. \tag{44}$$

Finally for the last term, because $\epsilon \leq \sqrt{\frac{c_1}{4096}}$,

$$\frac{4L}{3} \|x_0 - x\| \|x_1 - x_0\|^2 \leq \frac{L}{48} \|x_0 - x_1\|^3 \tag{45}$$

and

$$\frac{4L}{3} \|x_0 - x\|^3 \leq \frac{L}{48} \|x_0 - x_1\|^3 . \tag{46}$$

Therefore, defining

$$\begin{aligned}\psi_{x_0,x}(\|x_0 - x_1\|) \triangleq c_1 \|x_0 - x\| + c_2 \|x_0 - x\| (2 \|x_0 - x_1\| + \|x_0 - x\|) \\ + \frac{L}{3} \|x_0 - x\| \left(4 \|x_1 - x_0\|^2 + 2 \|x_0 - x\|^2\right),\end{aligned}$$

we showed in equation 38 that for arbitrary x_1,

$$|g_{x_1}(x_0) - g_{x_1}(x)| \leq \psi_{x_0,x}(\|x_0 - x_1\|), \tag{47}$$

and for x_1 such that $\|x - x_1\| \geq 3 \left(\frac{\epsilon\sqrt{d}c_0}{L}\right)^{1/3}$, or $\|x_1 - x_0\| \geq 2 \left(\frac{\epsilon c_1}{L}\right)^{1/3}$, Combining equation 42, equation 43, equation 40, equation 45, equation 46 with equation 38:

$$\psi_{x_0,x}(\|x_0 - x_1\|) \leq \frac{3L}{48} \|x_0 - x_1\|^3 . \tag{48}$$

Therefore, for $\|x - x_1\| \geq 4 \left(\frac{\epsilon\sqrt{d}c_0}{L}\right)^{1/3}$,

$$|g_{x_1}(x_0) - g_{x_1}(x)| \leq \frac{7L}{48} \|x_0 - x_1\|^3 ,$$

which combined with Equation equation 33

$$g_{x_1}(x) + \frac{L}{96} \|x_1 - x_0\|^3 \leq r_{x_0}. \tag{49}$$

On the other hand, note that

$$|g_{x_0}(x) - r_{x_0}| \leq |v_{x_0}^\top (x - x_0)| + \frac{1}{2}(x - x_0)^\top \Sigma_{x_0}(x - x_0) \leq c_1\epsilon + \frac{c_2}{2}\epsilon^2 \leq 2c_1\epsilon,$$

where in the last line we used $\epsilon \leq \frac{c_1}{c_2}$. But now picking the constant γ large enough we can guarantee that

$$\frac{L}{96} \|x_0 - x_1\|^3 \geq 3c_1\epsilon.$$

Combining equation 50 with equation 49, we conclude the first argument

$$g_{x_1}(x) + c_1\epsilon \leq g_{x_0}(x).$$

On the other hand, note that $\psi_{x_0,x}(x_1)$ is increasing in $\|x_1 - x_0\|$. Therefore, combining equation 47 and equation 48, for any x_1 such that $\|x_1 - x\| \leq \gamma \left(\frac{\epsilon\sqrt{d}c_0}{L}\right)^{1/3}$

$$|g_{x_1}(x_0) - g_{x_1}(x)| \leq \psi_{x_0,x}(\|x_0 - x_1\|) \leq \psi_{x_0,x}(\gamma \left(\frac{\epsilon\sqrt{d}c_0}{L}\right)^{1/3}) \leq \frac{3L}{48} \left(\left(\gamma\frac{\epsilon\sqrt{d}c_0}{L}\right)^{1/3}\right)^3 \tag{50}$$

$$= \gamma_2\epsilon\sqrt{d}c_0. \tag{51}$$

G.3. Proof of Lemma 10

Note that the Hessian of $\|x - x_0\|^3$ is α strong convexity of f means for v with $\|v\|_{\mathcal{L}} = 1$ we have $v^\top \nabla^2 f(x_0) v \geq \alpha$. But from Assumption equation 1 we get $\|v\| \leq R$. Therefore,

$$\begin{aligned}
v^\top \nabla^2 f_{x_0}(x) v &= v^\top \left(\nabla^2 f(x_0) - L\nabla(\|x - x_0\|(x - x_0))\right) v \\
&= v^\top \left(\nabla^2 f(x_0) - L\nabla(\|x - x_0\|(x - x_0))\right) v \\
&= v^\top \left(\nabla^2 f(x_0) - L\|x - x_0\| I - \frac{L}{\|x - x_0\|}(x - x_0)(x - x_0)^\top\right) v \\
&\geq \alpha - 2R^2 L \|x - x_0\| \\
&\geq \frac{\alpha}{2}.
\end{aligned}$$

G.4. Proof of Theorem 11

Here we prove Theorem 11. Before diving into the proof, we need to state and prove Lemma 35 so that we can obtain an $\alpha/2$ strong convexity property for the approximate regularizer in Theorem 11. In particular, Lemma 35 combines Lemmas 9 and 10 and concludes that the feasibility of $\mathcal{I}$ for the program implies strong convexity of g with respect to $\|.\|_{\mathcal{L}^c}$.

Lemma 35 (Program feasibility → strong convexity) *Suppose $\mathcal{I} = (\mathbf{r}, \mathbf{v}, \mathbf{\Sigma})$ is a feasible solution to LP equation 8 with respect to an ϵ-cover $\{x_i\}_{i=1}^N$ in $\mathcal{X}$ for the Euclidean norm, i.e. $\forall x \in \mathcal{X}$, $\exists x_i$ s.t. $\|x - x_i\| \leq \epsilon$, where ϵ satisfies*

$$\epsilon \leq \frac{\alpha^3}{512 R^6 L^2 c_0 \sqrt{d}}.$$

Then, for any point $x \in \mathcal{X}$, g is second order continuously right and left differentiable with

$$D^{2,l} g(x)[v,v], D^{2,r} g(x)[v,v] \geq \frac{\alpha}{2} \|v\|_{\mathcal{L}^c}^2,$$

where $D^{2,l} g(x)[v,v]$ and $D^{2,r} g(x)[v,v]$ denote the left and right second order directional derivative of f at x in direction v.

Proof For $x \in \mathcal{X}$ let $I(x) = \arg\max_{i \in [N]} g_{x_i}(x)$ be the set of indices for which $g_{x_i}(x)$ achieves its maximum at x. First, note that for the one-dimensional function $h(t) = g^{(\mathcal{I})}(x + tv)$, the subgradient of h zero is exactly

$$[\min_{i \in I(x)} Dg_{x_i}(x)[v], \max_{i \in I(x)} Dg_{x_i}(x)[v]],$$

due to the convexity of g_{x_i}'s. In fact, $h'^{l}(0) = \min_{i \in I(x)} Dg_{x_i}(x)[v]$ and $h'^{r}(0) = \max_{i \in I(x)} Dg_{x_i}(x)[v]$. Now let

$$\begin{aligned}
I^{r,v} &= \arg\max_{i \in I(x)} Dg_{x_i}(x)[v] \\
I^{l,v} &= \arg\min_{i \in I(x)} Dg_{x_i}(x)[v].
\end{aligned}$$

Then the second left and right directional derivatives at point x are given by

$$D^{2,l}g(x)[v,v] = h''^{l}(0) = \max_{i \in I^l(x)} D^2 g_{x_i}(x)[v,v], \tag{52}$$

$$D^{2,l}g(x)[v,v] = h''^{r}(0) = \max_{i \in I^r(x)} D^2 g_{x_i}[v,v]. \tag{53}$$

Furthermore, note that from Lemma 9, for every x_i such that $\|x_i - x\| \geq 4\left(\frac{\epsilon\sqrt{d}c_0}{L}\right)^{1/3}$, we have $g^{(\mathcal{I})}_{x_i}(x) < g^{(\mathcal{I})}_{x_0}(x)$, therefore $i \notin I$. Hence, we should have $\left\|x - x_{\hat{i}(x)}\right\| \leq 4\left(\frac{\epsilon\sqrt{d}c_0}{L}\right)^{1/3}$ for all $\hat{i} \in I$. But using the upper bound given on ϵ we get

$$\left\|x - x_{\hat{i}(x)}\right\| \leq 4\left(\frac{\epsilon\sqrt{d}c_0}{L}\right)^{1/3} \leq \frac{\alpha}{2R^2L}.$$

Hence, From Lemma 10, we have that $g_{x_{\hat{i}}}(x)$ is $\frac{\alpha}{2}$ strongly convex at x, for all $\hat{i} \in I$:

$$v^\top \nabla^2 g_{x_{\hat{i}}}(x) v \geq \frac{\alpha}{2}\|v\|^2_{\mathcal{L}^c}. \tag{54}$$

Finally combining this with equation 53 we conclude

$$D^{2,l}g(x)[v,v], D^{2,r}g(x)[v,v] \geq \frac{\alpha}{2}\|v\|^2_{\mathcal{L}^c}.$$

■

Next, we state the proof of Theorem 11.

Proof [Proof of Theorem 11] Consider the solution $\tilde{\mathcal{I}} = \left(\tilde{\mathbf{r}}, \tilde{\mathbf{v}}, \tilde{\mathbf{\Sigma}}\right)$ where $\forall i \in [N]$

$$\tilde{\Sigma}_{x_i} = \nabla^2 f(x_i),$$
$$\tilde{v}_{x_i} = \nabla f(x_i),$$
$$\tilde{r}_{x_i} = f(x_i).$$

First note that from Lemma 7 we get $f_{x_0}(x) + \frac{1}{6}\|x - x_0\|^3 \leq f(x)$, which implies $g^{(\tilde{\mathcal{I}})}_{x_i}(x_j) + \frac{15L}{96}\|x_j - x_i\|^3 \leq r_{x_j}$ for the above choice for $\tilde{\mathcal{I}}$. Moreover, $r_{x_0} \leq f(x_0) \leq C^2 \leq C_0$, and from $\tilde{c}_1$ Lipschitz and $\tilde{c}_2$ gradient Lipschitz conditions on f, we get $\forall i, \|\tilde{v}_{x_i}\| \leq \tilde{c}_1, \forall i, \tilde{\Sigma}_{x_i} \preccurlyeq \tilde{c}_2 I$, and the $\|.\|_{\mathcal{L}^c_{\tilde{}}} - \alpha$ strong convexity of f shows that $\tilde{\mathcal{I}}$ satisfies the condition $v^\top \Sigma_{x_i} v \geq \alpha, \forall v \in \mathcal{C}, \forall i$. Hence, $\tilde{\mathcal{I}}$ is feasible for the LP. In particular, note that we do not need the additional $L\bar{\epsilon}^3$ terms in the definition of c_0, c_2, C_0 to show the feasibility of $\tilde{\mathcal{I}}$ for the LP; these extra terms are only required for the third argument of Lemma 11 to show that not only $\tilde{\mathcal{I}}$ is feasible, but a ball around it is also feasible. We will prove that shortly. Next, from Lemma 9, we see that the maximum $\max_{i \in [N]} g^{\tilde{\mathcal{I}}}_{x_i}(x)$ at point $x \in \mathcal{X}$ is never achieved by far x_j's from x, farther than $\|x_j - x\| \geq \gamma\left(\frac{\epsilon\sqrt{d}c_0}{L}\right)^{1/3}$, since the value of $g_{x_j}(x)$ is smaller than $g_{x_i}(x)$ for the element of the cover x_i that is ϵ close to x. On the other hand, again from Lemma 9 for x_i such that $\|x_i - x\| \leq \epsilon$ and any x_j such that $\|x_j - x\| \leq \gamma \left(\frac{\epsilon\sqrt{d}c_0}{L}\right)^{1/3}$, we have

$$|g_{x_j}^{(\tilde{\mathcal{I}})}(x_i) - g_{x_j}^{(\tilde{\mathcal{I}})}(x)| \leq \gamma_2\epsilon\sqrt{d}c_0,$$

and from LP feasibility

$$g_{x_j}^{(\tilde{\mathcal{I}})}(x_i) \leq r_{x_i}.$$

Therefore,

$$\begin{aligned}
\max_{i\in[N]}|g_{x_i}^{(\tilde{\mathcal{I}})}(x)| &\leq \max_{i\in[N]}|r_i| + \gamma_2\epsilon\sqrt{d}c_0 \\
&= \max_{i\in[N]}|f(x_i)| + \gamma_2\epsilon\sqrt{d}c_0 \\
&\leq C^2 + \gamma_2\epsilon\sqrt{d}c_0.
\end{aligned}$$

Therefore, the optimal solution $\mathcal{I}^*$ should satisfy $\max_{i\in[N]}|g_{x_i}^{(\mathcal{I}^*)}(x)| \leq C^2+\gamma_2\epsilon\sqrt{d}c_0$ which proves the first argument equation 1. Finally, combining Lemmas 35 and 36 we get the $\alpha/2$ shows strong convexity of $g^{(\mathcal{I})}$ with respect to $\|.\|_{\mathcal{L}^c}$ for argument equation 2.

Next we show the third argument; note that f satisfies a slightly stronger inequality compared to the first condition of the LP equation 8, namely

$$\begin{aligned}
f(x_i) + \langle\nabla f(x_i), x_j - x_i\rangle + \frac{1}{2}(x_j - x_i)^\top\nabla^2 f(x_i)(x_j - x_i) - \frac{L}{3}\|x_j - x_i\|^3 \quad (55)\\
+ \left(\frac{L}{6} - \frac{L}{96}\right)\|x_j - x_i\|^3 + \frac{L}{96}\|x_j - x_i\|^3 \leq f(x_j), \quad (56)
\end{aligned}$$

or, since we constructed instance $\tilde{\mathcal{I}}$ from f,

$$g_{x_i}^{(\tilde{\mathcal{I}})}(x_j) + \frac{15L}{96}\|x_j - x_i\|^3 + \frac{L}{96}\|x_j - x_i\|^3 \leq f(x_j). \quad (57)$$

But if $\left\|\Sigma - \nabla^2 f(x_i)\right\| \leq \frac{L\bar{\epsilon}}{144} \leq \frac{L}{144}\|x_j - x_i\|$, then

$$\begin{aligned}
\frac{1}{2}|(x_j - x_i)^\top\nabla^2 f(x_i)(x_j - x_i) - (x_j - x_i)^\top\Sigma(x_j - x_i)| &\leq \frac{1}{2}\left\|(x_j - x_i)(x_j - x_i)^\top\right\|_F\left\|\nabla^2 f(x_i) - \Sigma\right\|_F \\
&\leq \frac{1}{2}\|x_j - x_i\|^2\left\|\nabla^2 f(x_i) - \Sigma\right\|_F \\
&\leq \frac{L}{288}\|x_j - x_i\|^3.
\end{aligned}$$

Given $\|\nabla f(x_i) - v\| \leq \frac{L\bar{\epsilon}^2}{288} \leq \frac{L}{288}\|x_j - x_i\|^2$, we get

$$|\langle\nabla f(x_i), x_j - x_i\rangle - \langle v, x_j - x_i\rangle| \leq \|\nabla f(x_i) - v\|\|x_j - x_i\| \leq \frac{L}{288}]\|x_i - x_j\|^3.$$

Finally under $|f(x_i) - r| \le \frac{L}{288}\bar{\epsilon}^3 \le \frac{L}{288}\|x_j - x_i\|^3$. Hence, if we assume $\left\|\mathcal{I} - \tilde{\mathcal{I}}\right\| \le \frac{L}{288}\bar{\epsilon}^3$, then combining the above Equations we get

$$|g_{x_i}^{(\mathcal{I})}(x_j) - g_{x_i}^{(\tilde{\mathcal{I}})}(x_j)| = |g_{x_i}^{(\mathcal{I})}(x_j) - f_{x_i}(x_j)| \le \frac{L}{96}\|x_j - x_i\|^3 .$$

But plugging this into equation 57

$$g_{x_i}^{(\mathcal{I})}(x_j) + \frac{15L}{96}\|x_j - x_i\|^3 \le f(x_j), \tag{58}$$

Finally note that $\left\|\mathcal{I} - \tilde{\mathcal{I}}\right\| \le \frac{L}{288}\bar{\epsilon}^3$ also implies $\forall i \in [N]$:

$$\begin{aligned}
&|r_{x_i}| \le |r_{x_i} - \tilde{r}_{x_i}| + |\tilde{r}_{x_i}| \le C^2 + L\bar{\epsilon}^3, \\
&\|v_{x_i}\| \le \|\tilde{v}_{x_i}\|_\infty + \|v_{x_i} - \tilde{v}_{x_i}\| \le \tilde{c}_1 + L\bar{\epsilon}^3, \\
&\Sigma_{x_i} \preccurlyeq \left\|\Sigma_{x_i} - \tilde{\Sigma}_{x_i}\right\| I + \tilde{\Sigma}_{x_i} \preccurlyeq \left(\tilde{c}_2 + L\bar{\epsilon}^3\right) I.
\end{aligned}$$

Therefore, $\tilde{\mathcal{I}}$ is still feasible for the program equation 8 with our choice of parameters c_0, c_2, C_0 here. Hence, we conclude

$$B_{L\bar{\epsilon}^3/288}(\tilde{\mathcal{I}}) \subseteq P_{\mathcal{I}} \subseteq B_{2\sqrt{(N+1){C_0}^2 + Nd\left(c_0^2 + c_2^2\right)}}(\tilde{\mathcal{I}}).$$

Finally note that for arbitrary $\mathcal{I} \in \mathcal{P}_{\mathcal{I}}$ which satisfies the conditions in LP equation 8, we have

$$\begin{aligned}
\|\mathcal{I}\|^2 &\le r^2 + \sum_i |r_{x_i}|^2 + \|v_{x_i}\|^2 + \|\Sigma_{x_i}\|^2 \\
&\le (N+1){C_0}^2 + Ndc_0^2 + Ndc_2^2,
\end{aligned}$$

which implies

$$P_{\mathcal{I}} \subseteq B_{2\sqrt{(N+1){C_0}^2 + Nd\left(c_0^2 + c_2^2\right)}}(\tilde{\mathcal{I}}).$$

■

G.5. Proof of Theorem 12

Consider the random distribution in Theorem 1.2 of Bhattiprolu et al. (2021). Then from property (3), there exists a unit direction v with $\|v\|_{\mathcal{L}} \le \frac{1}{d^{1-\epsilon}}$. Then we claim that $\|v\|_{\mathcal{L}^c} \le \frac{1}{d^{1-\epsilon}}$. This is because $\|v\|_{\mathcal{L}} = \sup_{\|w\|_{\mathcal{L}}^c} \langle v, w\rangle \ge \langle v, \frac{v}{\|v\|_{\mathcal{L}}^c}\rangle = \frac{1}{\|v\|_{\mathcal{L}}^c}$. Hence, $\|v\|_{\mathcal{L}}^c \ge d^{1-\epsilon}$. Hence, for $\tilde{v} = \frac{v}{\|v\|_{\mathcal{L}^c}}$ we have $\|\tilde{v}\|_{\mathcal{L}^c} = 1$ and $\|\tilde{v}\| \le \frac{1}{d^{1-\epsilon}}$.

Appendix H. Strong convexity

Here we show that a lower bound on the second derivative implies strong convexity with respect to arbitrary norms.

Lemma 36 (Lower bound on second derivative → strong convexity) *Suppose for convex function $g : \mathcal{X} \to \mathbb{R}$ which is second order continuously differentiable except in a finite number of points in which it is only left or right second order differentiable. Suppose the second left or right derivatives in arbitrary direction v, which we denote by $D^{2,l}g(x)[v,v], D^{2,r}g(x)[v,v]$ respectively, are at least $\alpha \left\lVert v \right\rVert_{\mathcal{L}^c}^2$. Then, g is strongly convex with respect to $\left\lVert . \right\rVert_{\mathcal{L}}^c$, namely for any $x, y \in \mathcal{X}$ and any subgradient v_x of f at point x:*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \alpha \left\lVert y - x \right\rVert_{\mathcal{L}^c}^2 .$$

Proof Without loss of generality assume $\left\lVert y - x \right\rVert_{\mathcal{L}^c} = 1$ and define the one variable function $h(t) : [0,1] \to \mathbb{R}$: $h(t) = g(x + t(y - x))$, and let $0 \leq t_1 \leq t_2 \leq \cdots \leq t_k \leq 1$ are the non-differentiable points of $h(t)$ on $[0,1]$, which we know are finite from our assumption. But from differentiability of h between these points, we can write (define $t_0 = 0, t_{k+1} = 1$)

$$f(y) = g(1) = \sum_{i=1}^{k} \int_{t_i}^{t_{i+1}} g'(t)dt, \tag{59}$$

where for the integral in $[t_i, t_{i+1}]$ by $h'(t_i)$ and $h'(t_{i+1})$ we mean the right derivative $h'^r(t_i)$ and left derivative $h'^l(t_{i+1})$, respectively. Now we show that for all $t \in [0,1]$

$$h'^l(t), h'^r(t) \geq h'^r(0) + \alpha t. \tag{60}$$

We show this inductively for $t \in (t_i, t_{i+1})$ for $i = 0, \ldots, k$. Particularly, the induction argument for step i is that for $t \in (t_i, t_{i+1})$, $h'(t) \geq \alpha t + h'^r(0)$, and $h'^l(t_{i+1}), h'^r(t_{i+1}) \geq h'^r(0) + t_{i+1}\alpha$. The base trivial since $h'^r(0) \geq h'^r(0)$. For the step of induction from $i-1$ to i, we know

$$g'^r(t_i) \geq \alpha t_i. \tag{61}$$

Now for any $t \in (t_i, t_{i+1})$ we can write

$$h'(t) = \int_{t_i}^{t} h''(s)ds \geq \alpha(t - t_i), \tag{62}$$

and particularly for t_{i+1}:

$$h'^l(t_{i+1}) = \int_{t_i}^{t_{i+1}} h''(s)ds \geq \alpha(t_{i+1} - t_i). \tag{63}$$

On the other hand, from the convexity of g,

$$h'^l(t_{i+1}) \leq h'^r(t_{i+1}). \tag{64}$$

Combining equation 62 equation 63 equation 64 with equation 61 completes the setp of induction.

Finally combining equation 60 with equation 59 and noting the fact that for any subgradient v at point x, $\langle v, y - x, \leq \rangle h'^r(0)$,

$$f(y) \geq h'^r(0) + \int_0^1 \alpha t dt \geq h'^r(0) + \frac{\alpha}{2},$$

which completes the proof. ■

Acknowledgments

We thank a bunch of people and funding agency.