JMLR: Workshop and Conference Proceedings vol (2012) 1–20

Online Optimization with Gradual Variations

Chao-Kai Chiang[1,2] CHAOKAI@IIS.SINICA.EDU.TW
Tianbao Yang[3] YANGTIA1@MSU.EDU
Chia-Jung Lee[1] LEECJ@IIS.SINICA.EDU.TW
Mehrdad Mahdavi[3] MAHDAVIM@CSE.MSU.EDU
Chi-Jen Lu[1] CJLU@IIS.SINICA.EDU.TW
Rong Jin[3] RONGJIN@CSE.MSU.EDU
Shenghuo Zhu[4] ZSH@SV.NEC-LABS.COM
[1] *Institute of Information Science, Academia Sinica, Taipei, Taiwan.*
[2] *Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.*
[3] *Department of Computer Science and Engineering Michigan State University, East Lansing, MI, 48824, USA*
[4] *NEC Laboratories America Cupertino, CA, 95014, USA*

Abstract

We study the online convex optimization problem, in which an online algorithm has to make repeated decisions with convex loss functions and hopes to achieve a small regret. We consider a natural restriction of this problem in which the loss functions have a small deviation, measured by the sum of the distances between every two consecutive loss functions, according to some distance metrics. We show that for the linear and general smooth convex loss functions, an online algorithm modified from the gradient descend algorithm can achieve a regret which only scales as the square root of the deviation. For the closely related problem of prediction with expert advice, we show that an online algorithm modified from the multiplicative update algorithm can also achieve a similar regret bound for a different measure of deviation. Finally, for loss functions which are strictly convex, we show that an online algorithm modified from the online Newton step algorithm can achieve a regret which is only logarithmic in terms of the deviation, and as an application, we can also have such a logarithmic regret for the portfolio management problem.

Keywords: Online Learning, Regret, Convex Optimization, Deviation.

1. Introduction

We study the online convex optimization problem in which a player has to make decisions iteratively for a number of rounds in the following way. In round t, the player has to choose a point x_t from some convex feasible set $\mathcal{X} \subseteq \mathbb{R}^N$, and after that the player receives a convex loss function f_t and suffers the corresponding loss $f_t(x_t) \in [0, 1]$. The player would like to have an online algorithm that can minimize its regret, which is the difference between the total loss it suffers and that of the best fixed point in hindsight. It is known that when playing for T rounds, a regret of $O(\sqrt{TN})$ can be achieved, using the gradient

descend algorithm (Zinkevich, 2003). When the loss functions are restricted to be linear, it becomes the well-known online linear optimization problem. Another related problem is the prediction with expert advice problem, in which the player in each round has to choose one of N actions to play, possibly in a probabilistic way. This can be seen as a special case of the online linear optimization problem, with the feasible set being the set of probability distributions over the N actions, and a regret of $O(\sqrt{T \ln N})$ can be achieved using the multiplicative update algorithm (Littlestone and Warmuth, 1994; Freund and Schapire, 1997). There have been many wonderful results and applications for these problems, and more information can be found in (Cesa-Bianchi and Lugosi, 2006). The regrets achieved for these problems are in fact optimal since matching lower bounds are also known (see e.g., (Cesa-Bianchi and Lugosi, 2006; Abernethy et al., 2008)). On the other hand, when the loss functions satisfy some nice property, a smaller regret becomes possible. Hazan et al. (2007) showed that a regret of $O(N \ln T)$ can be achieved for loss functions satisfying some strong convexity properties, which includes functions arising from the portfolio management problem (Cover, 1991).

Most previous works, including those discussed above, considered the most general setting in which the loss functions could be arbitrary and possibly chosen in an adversarial way. However, the environments around us may not always be adversarial, and the loss functions may have some patterns which can be exploited for achieving a smaller regret. One work along this direction is that of Hazan and Kale (2008). For the online linear optimization problem, in which each loss function is linear and can be seen as a vector, they considered the case in which the loss functions have a small variation, defined as $V = \sum_{t=1}^{T} \|f_t - \mu\|_2^2$, where $\mu = \sum_{t=1}^{T} f_t/T$ is the average of the loss functions and $\|\cdot\|_p$ denotes the L_p-norm. For this, they showed that a regret of $O(\sqrt{V})$ can be achieved, and they also have an analogous result for the prediction with expert advice problem. In another paper, Hazan and Kale (2009) considered the portfolio management problem in which each loss function has the form $f_t(x) = -\ln \langle v_t, x \rangle$ with $v_t \in [\delta, 1]^N$ for some constant $\delta \in (0, 1)$, where $\langle v_t, x \rangle$ denotes the inner product of the vectors v_t and x, and they showed how to achieve a regret of $O(N \log Q)$, with $Q = \sum_{t=1}^{T} \|v_t - \mu\|_2^2$ and $\mu = \sum_{t=1}^{T} v_t/T$. Note that according to their definition, a small V means that most of the loss functions center around some fixed loss function μ, and similarly for the case of small Q. This seems to model a stationary environment, in which all the loss functions are produced according to some fixed distribution.

The variation introduced in (Hazan and Kale, 2008) is defined in terms of total difference between individual linear cost vectors to their mean. In this paper we introduce a new measure, which we call L_p-deviation, for the loss functions, defined as

$$D_p = \sum_{t=1}^{T} \max_{x \in \mathcal{X}} \|\nabla f_t(x) - \nabla f_{t-1}(x)\|_p^2, \tag{1}$$

which is defined in terms of sequential difference between individual loss function to its previous one, where we use the convention that f_0 is the all-0 function. The motivation of defining gradual variation (i.e., L_p-deviation) stems from two observations: one is practical and the other one is technical raised by the limitation of extending the results in (Hazan and Kale, 2008) to general convex functions. From practical point of view, we are interested in a more general scenario, in which the environment may be evolving but

in a somewhat gradual way. For example, the weather condition or the stock price at one moment may have some correlation with the next and their difference is usually small, while abrupt changes only occur sporadically. Obviously, L_p-deviation easily models these situations. In order to understand the limitation of extending the results in (Hazan and Kale, 2008), let us apply the results in (Hazan and Kale, 2008) to general convex loss functions as follows. Since the results in (Hazan and Kale, 2008) were developed for linear loss functions, a straightforward approach is to use the first order approximation for convex loss functions, i.e., $f_t(x) \simeq f_t(x_t) + \langle \nabla f_t(x_t), x - x_t \rangle$, and replace the linear loss vector with the gradient of the loss function $f_t(x)$ at x_t. Using the convexity of loss function $f_t(x)$, we have

$$\sum_{t=1}^{T} f_t(x_t) - \min_{\pi \in \mathcal{X}} \sum_{t=1}^{T} f_t(\pi) \leq \sum_{t=1}^{T} \langle \nabla f_t(x_t), x_t \rangle - \min_{\pi \in \mathcal{X}} \sum_{t=1}^{T} \langle \nabla f_t(x_t), \pi \rangle . \tag{2}$$

By assuming $\|\nabla f_t(x)\|_2 \leq 1, \forall t \in [T]$ and $\forall x \in \mathcal{X}$, we can apply Hazan and Kale's variation based bound to bound the regret in (2) by the variation of the gradients as

$$V = \sum_{t=1}^{T} \|\nabla f_t(x_t) - \mu\|_2^2 = \sum_{t=1}^{T} \left\| \nabla f_t(x_t) - \frac{1}{T} \sum_{\tau=1}^{T} \nabla f_\tau(x_\tau) \right\|_2^2 . \tag{3}$$

To better understand V in (3), we rewrite it as

$$\begin{aligned} V &= \sum_{t=1}^{T} \left\| \nabla f_t(x_t) - \frac{1}{T} \sum_{\tau=1}^{T} \nabla f_\tau(x_\tau) \right\|_2^2 = \frac{1}{2T} \sum_{t,\tau=1}^{T} \|\nabla f_t(x_t) - \nabla f_\tau(x_\tau)\|_2^2 \\ &\leq \frac{1}{T} \sum_{t=1}^{T} \sum_{\tau=1}^{T} \|\nabla f_t(x_t) - \nabla f_t(x_\tau)\|_2^2 + \frac{1}{T} \sum_{t=1}^{T} \sum_{\tau=1}^{T} \|\nabla f_t(x_\tau) - \nabla f_\tau(x_\tau)\|_2^2 = V_1 + V_2. \end{aligned}$$

We see that the variation V is bounded by two parts: V_1 essentially measures the smoothness of the individual loss functions, while V_2 measures the variation in the gradients of loss functions. As a result, even when all the loss functions are identical, V_2 vanishes, while V_1 still exists, and therefore the regret of the algorithm in (Hazan and Kale, 2008) for online convex optimization may still be bounded by $O(\sqrt{T})$ regardless of the smoothness of the cost function. To address above mentioned challenges, the bounds in this paper are developed in terms of L_p-deviation. We note that for linear functions, L_p-deviation becomes $D_p = \sum_{t=1}^{T} \|f_t - f_{t-1}\|_p^2$. It can be shown that $D_2 \leq O(V)$ while there are loss functions with $D_2 \leq O(1)$ and $V = \Omega(T)$. Thus, one can argue that our constraint of a small deviation is strictly easier to satisfy than that of a small variation in (Hazan and Kale, 2008). For the portfolio management problem, a natural measure of deviation is $\sum_{t=1}^{T} \|v_t - v_{t-1}\|_2^2$, and one can show that $D_2 \leq O(N) \cdot \sum_{t=1}^{T} \|v_t - v_{t-1}\|_2^2 \leq O(NQ)$, so one can again argue that our constraint is easier to satisfy than that of (Hazan and Kale, 2009).

In this paper, we consider loss functions with such deviation constraints and obtain the following results. First, for the online linear optimization problem, we provide an algorithm which, when given loss functions with L_2-deviation D_2, can achieve a regret of $O(\sqrt{D_2})$. This is in fact optimal as a matching lower bound can be shown. Since $D_2 \leq O(TN)$, we immediately recover the result of Zinkevich (2003). Furthermore, as

discussed before, since one can upper-bound D_2 in terms of V but not vice versa, our result is arguably stronger than that of Hazan and Kale (2008); interestingly, our analysis even looks simpler than theirs. A similar bound was given by Rakhlin et al. (2011) in a game-theoretical setting, but they did not discuss any algorithm. Next, for the prediction with expert advice problem, we provide an algorithm such that when given loss functions with L_∞-deviation D_∞, it achieves a regret of $O(\sqrt{D_\infty \ln N})$, which is also optimal with a matching lower bound. Note that since $D_\infty \leq O(T)$, we also recover the $O(\sqrt{T \ln N})$ regret bound of Freund and Schapire (1997), but our result seems incomparable to that of Hazan and Kale (2008). We then establish variation bound for general convex loss functions aiming to take one step further along the work done in (Hazan and Kale, 2008). Our results shows that for general smooth convex functions, the proposed algorithm attains $O(\sqrt{D_2})$ bound. We show that smoothness assumptions is unavoidable for general convex loss functions. Finally, we provide an algorithm for the online convex optimization problem studied by Hazan et al. (2007), in which the loss functions are strictly convex. Our algorithm achieves a regret of $O(N \ln T)$ which matches that of an algorithm in (Hazan et al., 2007), and when the loss functions have L_2-deviation D_2, for a large enough D_2, and satisfy some smoothness condition, our algorithm achieves a regret of $O(N \ln D_2)$. This can be applied to the portfolio management problem considered by Hazan and Kale (2009) as the corresponding loss functions in fact satisfy our smoothness condition, and we can achieve a regret of $O(N \ln D)$ when $\sum_{t=1}^{T} \|v_t - v_{t-1}\|_2^2 \leq D$. As discussed before, one can again argue that our result is stronger than that of Hazan and Kale (2009).

All of our algorithms are based on the following idea, which we illustrate using the online linear optimization problem as an example. For general linear functions, the gradient descent algorithm is known to achieve an optimal regret, which plays in round t the point $x_t = \Pi_{\mathcal{X}}(x_{t-1} - \eta f_{t-1})$, the projection of $x_{t-1} - \eta f_{t-1}$ to the feasible set $\mathcal{X}$. Now, if the loss functions have a small deviation, f_{t-1} may be close to f_t, so in round t, it may be a good idea to play a point which moves further in the direction of $-f_{t-1}$ as it may make its inner product with f_t (which is its loss with respect to f_t) smaller. In fact, it can be shown that if one could play the point $x_{t+1} = \Pi_{\mathcal{X}}(x_t - \eta f_t)$ in round t, a very small regret could be achieved, but in reality one does not have f_t available before round t to compute x_{t+1}. On the other hand, if f_{t-1} is a good estimate of f_t, the point $\hat{x}_t = \Pi_{\mathcal{X}}(x_t - \eta f_{t-1})$ should be a good estimate of x_{t+1} too. The point $\hat{x}_t$ can actually be computed before round t since f_{t-1} is available, so our algorithm plays $\hat{x}_t$ in round t. Our algorithms for the prediction with expert advice problem and the online convex optimization problem use the same idea. We unify all our algorithms by a meta algorithm, which can be seen as a type of mirror descent algorithm (Nemirovski and Yudin, 1978; Beck and Teboulle, 2003), using the notion of Bregman divergence with respect to some function $\mathcal{R}$. Then we derive different algorithms for different settings simply by substantiating the meta algorithm with different choices for the function $\mathcal{R}$. For the linear and general smooth online convex optimization problems, the prediction with expert advice problem, and the online strictly convex optimization problem, respectively, the algorithms we derive can be seen as modified from the gradient descent algorithm of (Zinkevich, 2003), the multiplicative algorithm of (Littlestone and Warmuth, 1994; Freund and Schapire, 1997), and the online Newton step of (Hazan et al., 2007), with the modification based on the idea of moving further in the direction of $-f_{t-1}$ discussed above.

2. Preliminaries

For a positive integer n, let $[n]$ denote the set $\{1, 2, \cdots, n\}$. Let $\mathbb{R}$ denote the set of real numbers, $\mathbb{R}^N$ the set of N-dimensional vectors over $\mathbb{R}$, and $\mathbb{R}^{N\times N}$ the set of $N \times N$ matrices over $\mathbb{R}$. We will see a vector x as a column vector and see its transpose, denoted by $x^\top$, as a row vector. For a vector $x \in \mathbb{R}^N$ and an index $i \in [N]$, let $x(i)$ denote the i'th component of x. For $x, y \in \mathbb{R}^N$, let $\langle x, y\rangle = \sum_{i=1}^N x(i)y(i)$ and let $\mathrm{RE}\,(x\|y) = \sum_{i=1}^N x(i) \ln \frac{x(i)}{y(i)}$. All the matrices considered in this paper will be symmetric and we will assume this without stating it later. For two matrices A and B, we write $A \succeq B$ if $A-B$ is a positive semidefinite (PSD) matrix. For $x \in \mathbb{R}^N$, let $\|x\|_p$ denote the L_p-norm of x, and for a PSD matrix $H \in \mathbb{R}^{N\times N}$, define the norm $\|x\|_H$ by $\sqrt{x^\top Hx}$. Note that if H is the identity matrix, then $\|x\|_H = \|x\|_2$. We will need the following simple fact, which will be proved in Appendix A.

Proposition 1 *For any $y, z \in \mathbb{R}^N$ and any PSD $H \in \mathbb{R}^{N\times N}$, $\|y + z\|_H^2 \le 2\|y\|_H^2 + 2\|z\|_H^2$.*

We will need the notion of Bregman divergence and the projection according to it.

Definition 2 *Let $\mathcal{R} : \mathbb{R}^N \to \mathbb{R}$ be a differentiable function and $\mathcal{X} \subseteq \mathbb{R}^N$ a convex set. Define the Bregman divergence of $x, y \in \mathbb{R}^N$ with respect to $\mathcal{R}$ by $\mathcal{B}^{\mathcal{R}}(x, y) = \mathcal{R}(x) - \mathcal{R}(y) - \langle \nabla\mathcal{R}(y), x - y\rangle$. Define the projection of $y \in \mathbb{R}^N$ onto $\mathcal{X}$ according to $\mathcal{B}^{\mathcal{R}}$ by $\Pi_{\mathcal{X},\mathcal{R}}(y) = \arg\min_{x\in\mathcal{X}} \mathcal{B}^{\mathcal{R}}(x, y)$.*

We consider the *online convex optimization problem*, in which an online algorithm must play in T rounds in the following way. In each round $t \in [T]$, it plays a point $x_t \in \mathcal{X}$, for some convex feasible set $\mathcal{X} \subseteq \mathbb{R}^N$, and after that, it receives a loss function $f_t : \mathcal{X} \to \mathbb{R}$ and suffers a loss of $f_t(x_t)$. The goal is to minimize its *regret*, defined as

$$\sum_{t=1}^T f_t(x_t) - \arg\min_{\pi\in\mathcal{X}} \sum_{t=1}^T f_t(\pi),$$

which is the difference between its total loss and that of the best offline algorithm playing a single point $\pi \in \mathcal{X}$ for all T rounds. We study four special cases of this problem. The first is the *online linear optimization problem*, in which each loss function f_t is linear. The second case is the *prediction with expert advice problem*, which can be seen as a special case of the online linear optimization problem with the set of probability distributions over N actions as the feasible set $\mathcal{X}$. The third case is when the loss functions are smooth which is a generalizeiton of linear optimization setting. Finally, we consider the case when the loss functions are strictly convex in the sense defined as follows.

Definition 3 *For $\beta > 0$, we say that a function $f : \mathcal{X} \to \mathbb{R}$ is β-convex, if for all $x, y \in \mathcal{X}$,*

$$f(x) \ge f(y) + \langle \nabla f(y), x - y\rangle + \beta \langle \nabla f(y), x - y\rangle^2.$$

As shown in (Hazan et al., 2007), all the convex functions considered there are in fact β-convex, and thus our result also applies to those convex functions.

For simplicity of presentation, we will assume throughout the paper that the feasible set $\mathcal{X}$ is a closed convex set contained in the unit ball $\{x \in \mathbb{R}^n : \|x\|_2 \le 1\}$; the extension to the general case is straightforward.

Algorithm 1 META algorithm

1: Initially, let $x_1 = \hat{x}_1 = (1/N, \ldots, 1/N)^\top$.
2: In round $t \in [T]$:
2(a): Play $\hat{x}_t$.
2(b): Receive f_t and compute $\ell_t = \nabla f_t(\hat{x}_t)$.
2(c): Update

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \left(\langle \ell_t, x \rangle + \mathcal{B}^{\mathcal{R}_t}(x, x_t) \right),$$
$$\hat{x}_{t+1} = \arg\min_{\hat{x} \in \mathcal{X}} \left(\langle \ell_t, \hat{x} \rangle + \mathcal{B}^{\mathcal{R}_{t+1}}(\hat{x}, x_{t+1}) \right).$$

3. Meta Algorithm

All of our algorithms in the coming sections are based on the META algorithm, given in Algorithm 1, which has the parameter $\mathcal{R}_t$ for $t \in [T]$. For different types of problems, we will have different choices of $\mathcal{R}_t$, which will be specified later in the respective sections. Here we allow $\mathcal{R}_t$ to depend on t, although we do not need this freedom for linear functions and general convex functions; we only need this for strictly convex functions. Note that we define x_{t+1} using $\mathcal{R}_t$ instead of $\mathcal{R}_{t+1}$ for some technical reason which will be discussed soon and will become clear in the proof Lemma 6.

Our META algorithm is related to the mirror descent algorithm, as it can be shown to have the following equivalent form, which will be proved in Appendix B.1.

Lemma 4 *Suppose y_{t+1} and $\hat{y}_{t+1}$ satisfy the conditions $\nabla \mathcal{R}_t(y_{t+1}) = \nabla \mathcal{R}_t(x_t) - \ell_t$ and $\nabla \mathcal{R}_{t+1}(\hat{y}_{t+1}) = \nabla \mathcal{R}_{t+1}(x_{t+1}) - \ell_t$, respectively, for a strictly convex $\mathcal{R}_t$. Then the update in Step 2(c) of the* META *algorithm is identical to*

$$x_{t+1} = \Pi_{\mathcal{X}, \mathcal{R}_t}(y_{t+1}) = \arg\min_{x \in \mathcal{X}} \mathcal{B}^{\mathcal{R}_t}(x, y_{t+1}),$$
$$\hat{x}_{t+1} = \Pi_{\mathcal{X}, \mathcal{R}_{t+1}}(\hat{y}_{t+1}) = \arg\min_{\hat{x} \in \mathcal{X}} \mathcal{B}^{\mathcal{R}_{t+1}}(\hat{x}, \hat{y}_{t+1}).$$

Note that a typical mirror descent algorithm plays in round t a point roughly corresponding to our x_t, while we move one step further along the direction of $-\ell_{t-1}$ and play $\hat{x}_t = \arg\min_{\hat{x} \in \mathcal{X}} \left(\langle \ell_{t-1}, \hat{x} \rangle + \mathcal{B}^{\mathcal{R}_t}(\hat{x}, x_t) \right)$ instead. The intuition behind our algorithm is the following. It can be shown that if one could play $x_{t+1} = \arg\min_{x \in \mathcal{X}} \left(\langle \ell_t, x \rangle + \mathcal{B}^{\mathcal{R}_t}(x, x_t) \right)$ in round t, then a small regret could be achieved, but in reality one does not have f_t available to compute x_{t+1} before round t. Nevertheless, if the loss vectors have a small deviation, ℓ_{t-1} is likely to be close to ℓ_t, and so is $\hat{x}_t$ to x_{t+1}, which is made possible by defining x_{t+1} and $\hat{x}_t$ both using $\mathcal{R}_t$. Based on this idea, we let our algorithm play $\hat{x}_t$ in round t.

Now let us see how to bound the regret of the algorithm. Consider any $\pi \in \mathcal{X}$ taken by the offline algorithm. Then for a β-convex function f_t, we know from the definition that

$$f_t(\hat{x}_t) - f_t(\pi) \leq \langle \ell_t, \hat{x}_t - \pi \rangle - \beta \|\hat{x}_t - \pi\|_{h_t}^2, \text{ where } h_t = \ell_t \ell_t^\top, \tag{4}$$

while for a linear or a general convex f_t, the above still holds with $\beta = 0$. Thus, the key is to bound $\langle \ell_t, \hat{x}_t - \pi \rangle$, which is given by the following lemma. We give the proof in Appendix B.2.

Lemma 5 *Let* $S_t = \langle \ell_t - \ell_{t-1}, \hat{x}_t - x_{t+1} \rangle$, $A_t = \mathcal{B}^{\mathcal{R}_t}(\pi, x_t) - \mathcal{B}^{\mathcal{R}_t}(\pi, x_{t+1})$ *and* $B_t = \mathcal{B}^{\mathcal{R}_t}(x_{t+1}, \hat{x}_t) + \mathcal{B}^{\mathcal{R}_t}(\hat{x}_t, x_t)$. *Then*

$$\langle \ell_t, \hat{x}_t - \pi \rangle \leq S_t + A_t - B_t.$$

The following lemma provides an upper bound for S_t.

Lemma 6 *Suppose* $\|\cdot\|$ *is a norm, with dual norm* $\|\cdot\|_*$, *such that* $\frac{1}{2}\|x - x'\|^2 \leq \mathcal{B}^{\mathcal{R}_t}(x, x')$ *for any* $x, x' \in \mathcal{X}$. *Then,*

$$S_t = \langle \ell_t - \ell_{t-1}, \hat{x}_t - x_{t+1} \rangle \leq \left\| \ell_t - \ell_{t-1} \right\|_*^2.$$

Proof By a generalized Cauchy-Schwartz inequality,

$$S_t = \langle \ell_t - \ell_{t-1}, \hat{x}_t - x_{t+1} \rangle \leq \left\| \ell_t - \ell_{t-1} \right\|_* \|\hat{x}_t - x_{t+1}\|.$$

Then we need the following, which will be proved in Appendix B.3.

Proposition 7 $\|\hat{x}_t - x_{t+1}\| \leq \|\nabla \mathcal{R}_t(\hat{y}_t) - \nabla \mathcal{R}_t(y_{t+1})\|_*$.

From this proposition, we have

$$\|\hat{x}_t - x_{t+1}\| \leq \left\| \left(\nabla \mathcal{R}_t(x_t) - \ell_{t-1}\right) - \left(\nabla \mathcal{R}_t(x_t) - \ell_t\right) \right\|_* = \left\| \ell_t - \ell_{t-1} \right\|_*. \tag{5}$$

This is why we define x_{t+1} and y_{t+1} using $\mathcal{R}_t$ instead of $\mathcal{R}_{t+1}$. Finally, by combining these bounds together, we have the lemma. ■

Taking the sum over t of the bounds in Lemma 5 and Lemma 6, we obtain a general regret bound for the Meta algorithm. In the following sections, we will make different choices of $\mathcal{R}_t$ and the norms for different types of loss functions, and we will derive the corresponding regret bounds.

4. Linear Loss Functions

In this section, we consider the case that each loss function f_t is linear, which can be seen as an N-dimensional vector in $\mathbb{R}^N$ with $f_t(x) = \langle f_t, x \rangle$ and $\nabla f_t(x) = f_t$. We measure the deviation of the loss functions by their L_p-deviation, defined in (1), which becomes $\sum_{t=1}^T \|f_t - f_{t-1}\|_p^2$ for linear functions. To bound the regret suffered in each round, we can use the bound in (4) with $\beta = 0$ and we drop the term B_t from the bound in Lemma 5. By summing the bound over t, we have

$$\sum_{t=1}^T f_t(\hat{x}_t) - f_t(\pi) \leq \sum_{t=1}^T S_t + \sum_{t=1}^T A_t, \tag{6}$$

where $S_t = \langle f_t - f_{t-1}, \hat{x}_t - x_{t+1} \rangle$ and $A_t = \mathcal{B}^{\mathcal{R}_t}(\pi, x_t) - \mathcal{B}^{\mathcal{R}_t}(\pi, x_{t+1})$. In the following two subsections, we will consider the online linear optimization problem and the prediction with expert advice problem, respectively, in which we will have different choices of $\mathcal{R}_t$ and use different measures of deviation.

4.1. Online Linear Optimization Problem

In this subsection, we consider the online linear optimization problem, and we consider loss functions with L_2-deviation D_2. To instantiate the Meta algorithm for such loss functions, we choose

- $\mathcal{R}_t(x) = \frac{1}{2\eta}\|x\|_2^2$, for every $t \in [T]$,

where η is the learning rate to be determined later; in fact, it can also be adjusted in the algorithm using the standard doubling trick by keeping track of the deviation accumulated so far. It is easy to show that with this choice of $\mathcal{R}_t$,

- $\nabla\mathcal{R}_t(x) = \frac{x}{\eta}$, $\mathcal{B}^{\mathcal{R}_t}(x,y) = \frac{1}{2\eta}\|x-y\|_2^2$, and $\Pi_{\mathcal{X},\mathcal{R}_t}(y) = \arg\min_{x\in\mathcal{X}}\|x-y\|_2^2$.

Then, according to Lemma 4, the update in Step 2(c) of Meta algorithm becomes:

- $x_{t+1} = \arg\min_{x\in\mathcal{X}}\|x-y_{t+1}\|_2^2$, with $y_{t+1} = x_t - \eta f_t$,
  $\hat{x}_{t+1} = \arg\min_{\hat{x}\in\mathcal{X}}\|\hat{x}-\hat{y}_{t+1}\|_2^2$, with $\hat{y}_{t+1} = x_{t+1} - \eta f_t$.

The regret achieved by our algorithm is guaranteed by the following.

Theorem 8 *When the L_2-deviation of the loss functions is D_2, the regret of our algorithm is at most $O(\sqrt{D_2})$.*

Proof We start by bounding the first sum in (6). Note that we can apply Lemma 6 with the norm $\|\cdot\| = \frac{1}{\sqrt{\eta}}\|\cdot\|_2$, since $\frac{1}{2}\|x-x'\|^2 = \frac{1}{2\eta}\|x-x'\|_2^2 = \mathcal{B}^{\mathcal{R}_t}(x,x')$ for any $x, x' \in \mathcal{X}$. As the dual norm is $\|\cdot\|_* = \sqrt{\eta}\|\cdot\|_2$, Lemma 6 gives us

$$\sum_{t=1}^{T} S_t \le \sum_{t=1}^{T}\|f_t - f_{t-1}\|_*^2 \le \sum_{t=1}^{T}\eta\|f_t - f_{t-1}\|_2^2 \le \eta D_2.$$

Next, note that $A_t = \frac{1}{2\eta}\|\pi - x_t\|_2^2 - \frac{1}{2\eta}\|\pi - x_{t+1}\|_2^2$, so the second sum in (6) is

$$\sum_{t=1}^{T} A_t = \frac{1}{2\eta}\left(\|\pi - x_1\|_2^2 - \|\pi - x_{T+1}\|_2^2\right) \le \frac{2}{\eta},$$

by telescoping and then using the fact that $\|\pi - x_1\|_2^2 \le 4$ and $\|\pi - x_{T+1}\|_2^2 \ge 0$. Finally, by substituting these two bounds into (6), we have

$$\sum_{t=1}^{T}(f_t(\hat{x}_t) - f_t(\pi)) \le \eta D_2 + \frac{2}{\eta} \le O\left(\sqrt{D_2}\right),$$

by choosing $\eta = \sqrt{2/D_2}$, which proves the theorem. ■

Let us make three remarks about Theorem 8. First, as mentioned in the introduction, one can argue that our result is strictly stronger than that of (Hazan and Kale, 2008) as our deviation bound is easier to satisfy. This is because by Proposition 1, we have

$\|f_t - f_{t-1}\|_2^2 \leq 2(\|f_t - \mu\|_2^2 + \|\mu - f_{t-1}\|_2^2)$ and thus $D_2 \leq 4V + O(1)$, while, for example, with $N = 1$, $f_t = 0$ for $1 \leq t \leq T/2$ and $f_t = 1$ for $T/2 < t \leq T$, we have $D_2 \leq O(1)$ and $V \geq \Omega(T)$. Next, we claim that the regret achieved by our algorithm is optimal. This is because a matching lower bound can be shown by simply setting the loss functions of all but the first $r = D_2$ rounds to be the all-0 function, and then applying the known $\Omega(\sqrt{r})$ regret lower bound on the first r rounds. Finally, our algorithm can be seen as a modification of the gradient descent (GD) algorithm of (Zinkevich, 2003), which plays x_t, instead of our $\hat{x}_t$, in round t. Then one may wonder if GD already performs as well as our algorithm does. The following lemma, to be proved in Appendix C.1, provides a negative answer, which means that our modification is in fact necessary.

Lemma 9 *The regret of the GD algorithm is at least* $\Omega(\min\{D_2, \sqrt{T}\})$.

4.2. Prediction with Expert Advice

In this subsection, we consider the prediction with expert advice problem. Now, the feasible set $\mathcal{X}$ is the set of probability distributions over N actions, which can also be represented as N-dimensional vectors. Although this problem can be seen as a special case of that in Subsection 4.1 and Theorem 8 there also applies here, we would like to obtain a stronger result. More precisely, now we consider L_∞-deviation instead of L_2-deviation, and we assume that the loss functions have L_∞-deviation D_∞. Note that with $D_2 \leq D_\infty N$, Theorem 8 only gives a regret bound of $O(\sqrt{D_\infty N})$. To obtain a smaller regret, we instantiate the Meta algorithm with

- $\mathcal{R}_t(x) = \frac{1}{\eta}\sum_{i=1}^N x(i)\,(\ln x(i) - 1)$, for every $t \in [T]$,

where η is the learning rate to be determined later and recall that $x(i)$ denotes the i'th component of the vector x. It is easy to show that with this choice,

- $\nabla\mathcal{R}_t(x) = \frac{1}{\eta}(\ln x(1), \ldots, \ln x(N))^\top$, $\mathcal{B}^{\mathcal{R}_t}(x, y) = \frac{1}{\eta}\mathrm{RE}\,(x\|y)$, and $\Pi_{\mathcal{X},\mathcal{R}_t}(y) = y/Z$ with the normalization factor $Z = \sum_{j=1}^N y(j)$.

Then, according to Lemma 4, the update in Step 2(c) of the Meta algorithm becomes:

- $x_{t+1}(i) = x_t(i)e^{-\eta f_t(i)}/Z_{t+1}$, for each $i \in [N]$, with $Z_{t+1} = \sum_{j=1}^N x_t(j)e^{-\eta f_t(j)}$,
  $\hat{x}_{t+1}(i) = x_{t+1}(i)e^{-\eta f_t(i)}/\hat{Z}_{t+1}$, for each $i \in [N]$, with $\hat{Z}_{t+1} = \sum_{j=1}^N x_{t+1}(j)e^{-\eta f_t(j)}$.

Note that our algorithm can be seen as a modification of the multiplicative updates algorithm (Littlestone and Warmuth, 1994; Freund and Schapire, 1997) which plays x_t, instead of our $\hat{x}_t$, in round t. The regret achieved by our algorithm is guaranteed by the following, which we will prove in Appendix C.2.

Theorem 10 *When the L_∞-deviation of the loss functions is D_∞, the regret of our algorithm is at most* $O(\sqrt{D_\infty \ln N})$.

We remark that the regret achieved by our algorithm is also optimal. This is because a matching lower bound can be shown by simply setting the loss functions of all but the first $r = D_\infty$ rounds to be the all-0 function, and then applying the known $\Omega(\sqrt{r \ln N})$ regret lower bound on the first r rounds.

5. General Convex Loss Functions

In this section, we consider general convex loss functions. We measure the deviation of loss functions by their L_2-deviation defined in (1), which is $\sum_{t=1}^{T} \max_{x\in\mathcal{X}} \|\nabla f_t(x) - \nabla f_{t-1}(x)\|_2^2$. Our algorithm for such loss functions is the same algorithm for linear functions. To bound its regret, now we need the help of the term B_t in Lemma 5, and we have

$$\sum_{t=1}^{T} \left(f_t\left(\hat{x}_t\right) - f_t\left(\pi\right)\right) \leq \sum_{t=1}^{T} S_t + \sum_{t=1}^{T} A_t - \sum_{t=1}^{T} B_t. \tag{7}$$

From the proof of Theorem 8, we know that $\sum_{t=1}^{T} A_t \leq \frac{2}{\eta}$ and $\sum_{t=1}^{T} S_t \leq \sum_{t=1}^{T} \eta \left\|\ell_t - \ell_{t-1}\right\|_2^2$ which, unlike in Theorem 8, can not be immediately bounded by L_2-deviation. This is because $\left\|\ell_t - \ell_{t-1}\right\|_2^2 = \|\nabla f_t(\hat{x}_t) - \nabla f_{t-1}(\hat{x}_{t-1})\|_2^2$, where the two gradients are taken at different points. To handle this issue, we further assume that each gradient ∇f_t satisfies the following λ-smoothness condition:

$$\|\nabla f_t(x) - \nabla f_t(y)\|_2 \leq \lambda \|x - y\|_2, \text{ for any } x, y \in \mathcal{X}. \tag{8}$$

We emphasize that our assumption about the smoothness of loss functions is necessary to achieve the desired variation bound. To see this, consider the special case of $f_1(x) = \cdots = f_T(x) = f(x)$. If the variation bound $O(\sqrt{D_2})$ holds for any sequence of convex functions, then for the special case where all loss functions are identical, we will have

$$\sum_{t=1}^{T} f(\hat{x}_t) \leq \min_{\pi\in\mathcal{X}} \sum_{t=1}^{T} f(\pi) + O(1),$$

implying that $(1/T)\sum_{t=1}^{T} \hat{x}_t$ approaches the optimal solution at the rate of $O(1/T)$. This contradicts the lower complexity bound (i.e. $\Omega(1/\sqrt{T})$) for any first order optimization method (Nesterov, 2004, Theorem 3.2.1) and therefore smoothness assumption is necessary to extend our results to general convex loss functions.

Our main result of this section is the following theorem which establishes the variation bound for general smooth convex loss functions applying Meta algorithm.

Theorem 11 *When the loss functions have L_2-deviation D_2 and the gradient of each loss function is λ-smooth, with $\lambda \leq 1/\sqrt{8D_2}$, the regret of our algorithm is at most $O(\sqrt{D_2})$.*

The proof of Theorem 11 immediately results from the following two lemmas. First, we need the following to bound $\sum_{t=1}^{T} S_t$ in terms of D_2.

Lemma 12 $\sum_{t=1}^{T} \left\|\ell_t - \ell_{t-1}\right\|_2^2 \leq 2D_2 + 2\lambda^2 \sum_{t=1}^{T} \|\hat{x}_t - \hat{x}_{t-1}\|_2^2$.

Proof $\left\|\ell_t - \ell_{t-1}\right\|_2^2 = \|\nabla f_t(\hat{x}_t) - \nabla f_{t-1}(\hat{x}_{t-1})\|_2^2$, which by Proposition 1 is at most

$$2\|\nabla f_t(\hat{x}_t) - \nabla f_{t-1}(\hat{x}_t)\|_2^2 + 2\|\nabla f_{t-1}(\hat{x}_t) - \nabla f_{t-1}(\hat{x}_{t-1})\|_2^2,$$

where the second term above is at most $2\lambda^2 \|\hat{x}_t - \hat{x}_{t-1}\|_2^2$ by the λ-smoothness condition. By summing the bound over t, we have the lemma. ■

To eliminate the undesirable term $2\lambda^2 \sum_{t=1}^{T} \|\hat{x}_t - \hat{x}_{t-1}\|_2^2$ in the lemma, we use the help from the sum $\sum_{t=1}^{T} B_t$, which has the following bound.

Lemma 13 $\sum_{t=1}^{T} B_t \geq \frac{1}{4\eta} \sum_{t=1}^{T} \|\hat{x}_t - \hat{x}_{t-1}\|_2^2 - O(1)$.

Proof Recall that $B_t = \frac{1}{2\eta} \|x_{t+1} - \hat{x}_t\|_2^2 + \frac{1}{2\eta} \|\hat{x}_t - x_t\|_2^2$, so we can write $\sum_{t=1}^{T} B_t$ as

$$\begin{aligned} \frac{1}{2\eta} \sum_{t=2}^{T+1} \|x_t - \hat{x}_{t-1}\|_2^2 + \frac{1}{2\eta} \sum_{t=1}^{T} \|\hat{x}_t - x_t\|_2^2 &\geq \frac{1}{2\eta} \sum_{t=2}^{T} \left(\|x_t - \hat{x}_{t-1}\|_2^2 + \|\hat{x}_t - x_t\|_2^2 \right) \\ &\geq \frac{1}{4\eta} \sum_{t=2}^{T} \|\hat{x}_t - \hat{x}_{t-1}\|_2^2 , \end{aligned}$$

by Proposition 1, with H being the identity matrix so that $\|x\|_H^2 = \|x\|_2^2$. Then the lemma follows as $\|\hat{x}_2 - \hat{x}_1\|_2^2 \leq O(1)$. ■

According to the bounds obtained so far, the regret of our algorithm is at most

$$2\eta D_2 + 2\eta\lambda^2 \sum_{t=1}^{T} \|\hat{x}_t - \hat{x}_{t-1}\|_2^2 - \frac{1}{4\eta} \sum_{t=1}^{T} \|\hat{x}_t - \hat{x}_{t-1}\|_2^2 + O(1) + \frac{2}{\eta} \leq O\left(\eta D_2 + \frac{1}{\eta}\right) \leq O\left(\sqrt{D_2}\right),$$

when $\lambda \leq 1/\sqrt{8\eta^2}$ and $\eta = 1/\sqrt{D_2}$.

6. Strictly Convex Loss Functions

In this section, we consider convex functions which are strictly convex. More precisely, suppose for some $\beta > 0$, each loss function is β-convex, so that

$$f_t(\hat{x}_t) - f_t(\pi) \leq \langle \ell_t, \hat{x}_t - \pi \rangle - \beta \|\pi - \hat{x}_t\|_{h_t}^2 , \text{ where } h_t = \ell_t \ell_t^\top . \tag{9}$$

Again, we measure the deviation of loss functions by their L_2-deviation, defined in (1). To instantiate the Meta algorithm for such loss functions, we choose

- $\mathcal{R}_t(x) = \frac{1}{2} \|x\|_{H_t}^2$, with $H_t = I + \beta\gamma^2 I + \beta \sum_{\tau=1}^{t-1} \ell_\tau \ell_\tau^\top$,

where I is the $N \times N$ identity matrix, and γ is an upper bound of $\|\ell_t\|_2$, for every t, so that $\gamma^2 I \succeq \ell_t \ell_t^\top$. It is easy to show that with this choice,

- $\nabla \mathcal{R}_t(x) = H_t x$, $\mathcal{B}^{\mathcal{R}_t}(x, y) = \frac{1}{2} \|x - y\|_{H_t}^2$, and $\Pi_{\mathcal{X}, \mathcal{R}_t}(y) = \arg\min_{x \in \mathcal{X}} \|x - y\|_{H_t}^2$.

Then, according to Lemma 4, the update in Step 2(c) of the Meta algorithm becomes:

- $x_{t+1} = \arg\min_{x \in \mathcal{X}} \|x - y_{t+1}\|_{H_t}^2$, with $y_{t+1} = x_t - H_t^{-1} \ell_t$,
  $\hat{x}_{t+1} = \arg\min_{\hat{x} \in \mathcal{X}} \|\hat{x} - \hat{y}_{t+1}\|_{H_{t+1}}^2$, with $\hat{y}_{t+1} = x_{t+1} - H_{t+1}^{-1} \ell_t$.

We remark that our algorithm is related to the online Newton step algorithm in (Hazan et al., 2007), except that our matrix H_t is slightly different from theirs and we play $\hat{x}_t$ in round t while they play a point roughly corresponding to our x_t. It is easy to verify that the update of our algorithm can be computed at the end of round t, because we have $\ell_1, \ldots, \ell_t$ available to compute H_t and H_{t+1}.

To bound the regret of our algorithm, note that by substituting the bound of Lemma 5 into (9) and then taking the sum over t, we obtain

$$\sum_{t=1}^{T} \left(f_t\left(\hat{x}_t\right) - f_t\left(\pi\right)\right) \leq \sum_{t=1}^{T} S_t + \sum_{t=1}^{T} A_t - \sum_{t=1}^{T} B_t - \sum_{t=1}^{T} C_t, \tag{10}$$

with $S_t = \left\langle \ell_t - \ell_{t-1}, \hat{x}_t - x_{t+1} \right\rangle$, $A_t = \frac{1}{2}\left\|\pi - x_t\right\|_{H_t}^2 - \frac{1}{2}\left\|\pi - x_{t+1}\right\|_{H_t}^2$, $B_t = \frac{1}{2}\left\|x_{t+1} - \hat{x}_t\right\|_{H_t}^2 + \frac{1}{2}\left\|\hat{x}_t - x_t\right\|_{H_t}^2$, and $C_t = \beta\left\|\pi - \hat{x}_t\right\|_{h_t}^2$. Then our key lemma is the following, which will be proved in Appendix D.1.

Lemma 14 *Suppose the loss functions are β-convex for some $\beta > 0$. Then*

$$\sum_{t=1}^{T} S_t + \sum_{t=1}^{T} A_t - \sum_{t=1}^{T} C_t \leq O(1 + \beta\gamma^2) + \frac{8N}{\beta} \ln\left(1 + \frac{\beta}{4}\sum_{t=1}^{T}\left\|\ell_t - \ell_{t-1}\right\|_2^2\right).$$

Note that the lemma does not use the nonnegative sum $\sum_{t=1}^{T} B_t$ but it already provides a regret bound matching that in (Hazan et al., 2007). To bound $\sum_{t=1}^{T}\left\|\ell_t - \ell_{t-1}\right\|_2^2$ in terms of L_2-deviation, we again assume that each gradient ∇f_t satisfies the λ-smoothness condition defined in (8), and we will also use the help from the sum $\sum_{t=1}^{T} B_t$. To get a cleaner regret bound, let us assume without loss of generality that $\lambda \geq 1$ and $\beta \leq 1$, because otherwise we can set $\lambda = 1$ and $\beta = 1$ and the inequalities in (8) and (9) still hold. Our main result of this section is the following, which we will prove in Appendix D.3.

Theorem 15 *Suppose the loss functions are β-convex and their L_2-deviation is D_2, with $\beta \leq 1$ and $D_2 \geq 1$. Furthermore, suppose the gradient of each loss function is λ-smooth, with $\lambda \geq 1$, and has L_2-norm at most γ. Then the regret of our algorithm is at most $O(\beta\gamma^2 + (N/\beta)\ln(\lambda N D_2))$, which becomes $O((N/\beta)\ln D_2)$ for a large enough D_2.*

An immediate application of Theorem 15 is to the portfolio management problem considered in (Hazan and Kale, 2009). In the problem, the feasible set $\mathcal{X}$ is the N-dimensional probability simplex and each loss function has the form $f_t(x) = -\ln\langle v_t, x\rangle$, with $v_t \in [\delta, 1]^N$ for some $\delta \in (0, 1)$. A natural measure of deviation, extending that of (Hazan and Kale, 2009), for such loss functions is $D = \sum_{t=1}^{T}\left\|v_t - v_{t-1}\right\|_2^2$. By applying Theorem 15 to this problem, we have the following, which will be proved in Appendix D.4.

Corollary 16 *For the portfolio management problem described above, there is an online algorithm which achieves a regret of $O((N/\delta^2)\ln((N/\delta)D))$.*

Acknowledgments

The authors T. Yang, M. Mahdavi, and R. Jin acknowledge support from the National Science Foundation (IIS-0643494) and Office of Navy Research (for ONR award N00014-09-1-0663 and N00014-12-1-0431). The authors C.-K. Chiang, C.-J. Lee, and C.-J. Lu acknowledge support from Academia Sinica and National Science Council (NSC 100-2221-E-001-008-MY3) of Taiwan.

References

Jacob Abernethy, Peter L. Bartlett, Alexander Rakhlin, and Ambuj Tewari. Optimal strategies and minimax lower bounds for online convex games. In *COLT*, pages 415–424, 2008.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. ISBN 0521833787.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge Univerity Press, New York, 2006.

Thomas Cover. Universal portfolios. *Mathematical Finance*, 1:1–19, 1991.

Yoav Freund and Robert E. Schapire. A decision theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Elad Hazan and Satyen Kale. Extracting certainty from uncertainty: regret bounded by variation in costs. In *COLT*, pages 57–68, 2008.

Elad Hazan and Satyen Kale. On stochastic and worst-case models for investing. In *NIPS*, pages 709–717, 2009.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Journal of Computer and System Sciences*, 69(2-3):169–192, 2007.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Nauka Publishers, Moscow, 1978.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course (Applied Optimization)*. Springer Netherlands, 1 edition, 2004.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: stochastic, constrained, and smoothed adversaries. In *NIPS*, pages 1764–1772, 2011.

Nati Srebro, Karthik Sridharan, and Ambuj Tewari. On the universality of online mirror descent. In *NIPS*, pages 2645–2653, 2011.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003.

Appendix A. Proof of Proposition 1 in Section 2

By definition,

$$2\|y\|_H^2 + 2\|z\|_H^2 - \|y+z\|_H^2 = \|y\|_H^2 + \|z\|_H^2 - 2y^\top Hz = \|y-z\|_H^2 \geq 0,$$

which implies that $2\|y\|_H^2 + 2\|z\|_H^2 \geq \|y+z\|_H^2$.

Appendix B. Proofs in Section 3

B.1. Proof of Lemma 4

The lemma follows immediately from the following known fact (see e.g. (Beck and Teboulle, 2003; Srebro et al., 2011)); we give the proof for completeness.

Proposition 17 *Suppose $\mathcal{R}$ is strictly convex and differentiable, and y satisfies the condition $\nabla\mathcal{R}(y) = \nabla\mathcal{R}(u) - \ell$. Then*

$$\arg\min_{x\in\mathcal{X}} \left(\langle \ell, x\rangle + \mathcal{B}^{\mathcal{R}}(x,u)\right) = \arg\min_{x\in\mathcal{X}} \mathcal{B}^{\mathcal{R}}(x,y).$$

Proof Since $\mathcal{R}$ is strictly convex, the minimum on each side is achieved by a unique point. Next, note that $\mathcal{B}^{\mathcal{R}}(x,y) = \mathcal{R}(x) - \mathcal{R}(y) - \langle \nabla\mathcal{R}(y), x-y\rangle = \mathcal{R}(x) - \langle \nabla\mathcal{R}(y), x\rangle + c$, where $c = -\mathcal{R}(y) + \langle \nabla\mathcal{R}(y), y\rangle$ does not depend on the variable x. Thus, using the condition that $\nabla\mathcal{R}(y) = \nabla\mathcal{R}(u) - \ell$, we have

$$\begin{aligned}\arg\min_{x\in\mathcal{X}} \mathcal{B}^{\mathcal{R}}(x,y) &= \arg\min_{x\in\mathcal{X}} \left(\mathcal{R}(x) - \langle \nabla\mathcal{R}(u) - \ell, x\rangle\right) \\ &= \arg\min_{x\in\mathcal{X}} \left(\langle \ell, x\rangle + \mathcal{R}(x) - \langle \nabla\mathcal{R}(u), x\rangle\right).\end{aligned}$$

On the other hand, $\mathcal{B}^{\mathcal{R}}(x,u) = \mathcal{R}(x) - \mathcal{R}(u) - \langle \nabla\mathcal{R}(u), x-u\rangle = \mathcal{R}(x) - \langle \nabla\mathcal{R}(u), x\rangle + c'$, where $c' = -\mathcal{R}(u) + \langle \nabla\mathcal{R}(u), u\rangle$ does not depend on the variable x. Thus, we have

$$\arg\min_{x\in\mathcal{X}} \left(\langle \ell, x\rangle + \mathcal{B}^{\mathcal{R}}(x,u)\right) = \arg\min_{x\in\mathcal{X}} \left(\langle \ell, x\rangle + \mathcal{R}(x) - \langle \nabla\mathcal{R}(u), x\rangle\right) = \arg\min_{x\in\mathcal{X}} \mathcal{B}^{\mathcal{R}}(x,y).$$

■

B.2. Proof of Lemma 5

Let us write $\langle \ell_t, \hat{x}_t - \pi\rangle = \langle \ell_t, \hat{x}_t - x_{t+1}\rangle + \langle \ell_t, x_{t+1} - \pi\rangle$ which in turn equals

$$\left\langle \ell_t - \ell_{t-1}, \hat{x}_t - x_{t+1}\right\rangle + \left\langle \ell_{t-1}, \hat{x}_t - x_{t+1}\right\rangle + \langle \ell_t, x_{t+1} - \pi\rangle. \tag{11}$$

To bound the second and third terms in (11), we rely on the following.

Proposition 18 *Suppose $\ell \in \mathbb{R}^n$, $v = \arg\min_{x\in\mathcal{X}} \left(\langle \ell, x\rangle + \mathcal{B}^R(x,u)\right)$, and $w \in \mathcal{X}$. Then*

$$\langle \ell, v - w\rangle \leq \mathcal{B}^{\mathcal{R}}(w,u) - \mathcal{B}^{\mathcal{R}}(w,v) - \mathcal{B}^{\mathcal{R}}(v,u).$$

Proof We need the following well-known fact; for a proof, see e.g. pages 139–140 of (Boyd and Vandenberghe, 2004).

Fact 1 *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set and $x = \arg\min_{z\in\mathcal{X}} \phi(z)$ for some continuous and differentiable function $\phi : \mathcal{X} \to \mathbb{R}$. Then for any $w \in \mathcal{X}$, $\langle \nabla\phi(x), w - x\rangle \geq 0$.*

Let ϕ be the function defined by $\phi(x) = \langle \ell, x\rangle + \mathcal{B}^{\mathcal{R}}(x, u)$. Since $\mathcal{X}$ is a convex set and v is the minimizer of $\phi(x)$ over $x \in \mathcal{X}$, it follows from Fact 1 that $\langle \nabla\phi(v), w - v\rangle \geq 0$. Since $\nabla\phi(v) = \ell + \nabla\mathcal{R}(v) - \nabla\mathcal{R}(u)$, we have $\langle \ell, v - w\rangle \leq \langle \nabla\mathcal{R}(v) - \nabla\mathcal{R}(u), w - v\rangle$. Then, by the definition of Bregman divergence, we obtain

$$\begin{aligned}\mathcal{B}^{\mathcal{R}}(w, u) - \mathcal{B}^{\mathcal{R}}(w, v) - \mathcal{B}^{\mathcal{R}}(v, u) &= -\langle \nabla\mathcal{R}(u), w - v\rangle + \langle \nabla\mathcal{R}(v), w - v\rangle \\ &= \langle \nabla\mathcal{R}(v) - \nabla\mathcal{R}(u), w - v\rangle .\end{aligned}$$

As a result, we have

$$\langle \ell, v - w\rangle \leq \langle \nabla\mathcal{R}(v) - \nabla\mathcal{R}(u), w - v\rangle = \mathcal{B}^{\mathcal{R}}(w, u) - \mathcal{B}^{\mathcal{R}}(w, v) - \mathcal{B}^{\mathcal{R}}(v, u).$$

■

From Proposition 18 and the definitions of $\hat{x}_t$ and x_{t+1}, we have

$$\langle \ell_{t-1}, \hat{x}_t - x_{t+1}\rangle \leq \mathcal{B}^{\mathcal{R}_t}(x_{t+1}, x_t) - \mathcal{B}^{\mathcal{R}_t}(x_{t+1}, \hat{x}_t) - \mathcal{B}^{\mathcal{R}_t}(\hat{x}_t, x_t), \text{ and} \tag{12}$$

$$\langle \ell_t, x_{t+1} - \pi\rangle \leq \mathcal{B}^{\mathcal{R}_t}(\pi, x_t) - \mathcal{B}^{\mathcal{R}_t}(\pi, x_{t+1}) - \mathcal{B}^{\mathcal{R}_t}(x_{t+1}, x_t). \tag{13}$$

Combining the bounds in (11), (12), (13) together, we have the lemma.

B.3. Proof of Proposition 7

To simplify the notation, let $\mathcal{R} = \mathcal{R}_t$, $x = \hat{x}_t$, $x' = x_{t+1}$, $y = \hat{y}_t$, and $y' = y_{t+1}$. Then, from the property of the norm, we know that

$$\frac{1}{2}\left\|x - x'\right\|^2 \leq \mathcal{R}(x) - \mathcal{R}(x') - \left\langle \nabla\mathcal{R}(x'), x - x'\right\rangle ,$$

and also

$$\frac{1}{2}\left\|x' - x\right\|^2 \leq \mathcal{R}(x') - \mathcal{R}(x) - \left\langle \nabla\mathcal{R}(x), x' - x\right\rangle .$$

Adding these two bounds, we obtain

$$\left\|x - x'\right\|^2 \leq \left\langle \nabla\mathcal{R}(x) - \nabla\mathcal{R}(x'), x - x'\right\rangle . \tag{14}$$

Next, we show that

$$\left\langle \nabla\mathcal{R}(x) - \nabla\mathcal{R}(x'), x - x'\right\rangle \leq \left\langle \nabla\mathcal{R}(y) - \nabla\mathcal{R}(y'), x - x'\right\rangle . \tag{15}$$

For this, we need Fact 1 in Appendix B.2. By letting $\phi(z) = \mathcal{B}^{\mathcal{R}}(z, y)$, we have $x = \arg\min_{z\in\mathcal{X}} \phi(z)$, $\nabla\phi(x) = \nabla\mathcal{R}(x) - \nabla\mathcal{R}(y)$, and

$$\left\langle \nabla\mathcal{R}(x) - \nabla\mathcal{R}(y), x' - x\right\rangle \geq 0.$$

On the other hand, by letting $\phi(z) = \mathcal{B}^{\mathcal{R}}(z, y')$, we have $x' = \arg\min_{z\in\mathcal{X}} \phi(z)$, $\nabla\phi(x') = \nabla\mathcal{R}(x') - \nabla\mathcal{R}(y')$, and

$$\left\langle \nabla\mathcal{R}(x') - \nabla\mathcal{R}(y'), x - x' \right\rangle \geq 0.$$

Combining these two bounds, we have

$$\left\langle \left(\nabla\mathcal{R}(y) - \nabla\mathcal{R}(y')\right) - \left(\nabla\mathcal{R}(x) - \nabla\mathcal{R}(x')\right), x - x' \right\rangle \geq 0,$$

which implies the inequality in (15).

Finally, by combining (14) and (15), we obtain

$$\left\|x - x'\right\|^2 \leq \left\langle \nabla\mathcal{R}(y) - \nabla\mathcal{R}(y'), x - x' \right\rangle \leq \left\|\nabla\mathcal{R}(y) - \nabla\mathcal{R}(y')\right\|_* \left\|x - x'\right\|,$$

by a generalized Cauchy-Schwartz inequality. Dividing both sides by $\|x - x'\|$, we have the proposition.

Appendix C. Proofs in Section 4

C.1. Proof of Lemma 9 in Section 4

One may wonder if the GD algorithm can also achieve the same regret as our algorithm's by choosing the learning rate η properly. We show that no matter what the learning rate η the GD algorithm chooses, there exists a sequence of loss vectors which can cause a large regret. Let f be any unit vector passing through x_1. Let $s = \lfloor 1/\eta \rfloor$, so that if we use $f_t = f$ for every $t \leq s$, each such $y_{t+1} = x_1 - t\eta f$ still remains in $\mathcal{X}$ and thus $x_{t+1} = y_{t+1}$. Next, we analyze the regret by considering the following three cases depending on the range of s.

First, when $s \geq \sqrt{T}$, we choose $f_t = f$ for t from 1 to $\lfloor s/2 \rfloor$ and $f_t = 0$ for the remaining t. Clearly, the best strategy of the offline algorithm is to play $\pi = -f$. On the other hand, since the learning rate η is too small, the strategy x_t played by GD, for $t \leq \lfloor s/2 \rfloor$, is far away from π, so that $\langle f_t, x_t - \pi \rangle \geq 1 - t\eta \geq 1/2$. Therefore, the regret is at least $\lfloor s/2 \rfloor\,(1/2) = \Omega(\sqrt{T})$.

Second, when $0 < s < \sqrt{T}$, the learning rate is high enough so that GD may overreact to each loss vector, and we make it pay by flipping the direction of loss vectors frequently. More precisely, we use the vector f for the first s rounds so that $x_{t+1} = x_1 - t\eta f$ for any $t \leq s$, but just as x_{s+1} moves far enough in the direction of $-f$, we make it pay by switching the loss vector to $-f$, which we continue to use for s rounds. Note that $x_{s+1+r} = x_{s+1-r}$ but $f_{s+1+r} = -f_{s+1-r}$ for any $r \leq s$, so $\sum_{t=1}^{2s} \langle f_t, x_t - x_1 \rangle = \langle f_{s+1}, x_{s+1} - x_1 \rangle \geq \Omega(1)$. As x_{2s+1} returns back to x_1, we can see the first $2s$ rounds as a period, which only contributes $\|2f\|_2^2 = 4$ to the deviation. Then we repeat the period for τ times, where $\tau = \lfloor D_2/4 \rfloor$ if there are enough rounds, with $\lfloor T/(2s) \rfloor \geq \lfloor D_2/4 \rfloor$, to use up the deviation D_2, and $\tau = \lfloor T/(2s) \rfloor$ otherwise. For any remaining round t, we simply choose $f_t = 0$. As a result, the total regret is at least $\Omega(1) \cdot \tau = \Omega(\min\{D_2/4, T/(2s)\}) = \Omega(\min\{D_2, \sqrt{T}\})$.

Finally, when $s = 0$, the learning rate is so high that we can easily make GD pay by flipping the direction of the loss vector in each round. More precisely, by starting with $f_1 = -f$, we can have x_2 on the boundary of $\mathcal{X}$, which means that if we then alternate between f and $-f$, the strategies GD plays will alternate between x_3 and x_2 which have a constant distance from each other. Then following the analysis in the second case, one can show that the total regret is at least $\Omega(\min\{D_2, T\})$.

C.2. Proof of Theorem 10

We start by bounding the first sum in (6). Note that we can apply Lemma 6 with the norm $\|\cdot\| = \frac{1}{\sqrt{\eta}}\|\cdot\|_1$, since for any $x, x' \in \mathcal{X}$, $\frac{1}{2}\|x - x'\|^2 = \frac{1}{2\eta}\|x - x'\|_1^2 \le \frac{1}{\eta}\mathrm{RE}(x\|x') = \mathcal{B}^{\mathcal{R}_t}(x, x')$, by Pinsker's inequality. As the dual norm is $\|\cdot\|_* = \sqrt{\eta}\|\cdot\|_\infty$, Lemma 6 gives us

$$\sum_{t=1}^{T} S_t \le \sum_{t=1}^{T} \|f_t - f_{t-1}\|_*^2 \le \sum_{t=1}^{T} \eta \|f_t - f_{t-1}\|_\infty^2 \le \eta D_\infty.$$

Next, note that $A_t = \frac{1}{\eta}\mathrm{RE}(\pi\|x_t) - \frac{1}{\eta}\mathrm{RE}(\pi\|x_{t+1})$, so the second sum in (6) is

$$\sum_{t=1}^{T} A_t = \frac{1}{\eta}\left(\mathrm{RE}(\pi\|x_1) - \mathrm{RE}(\pi\|x_{T+1})\right) \le \frac{1}{\eta}\ln N,$$

by telescoping and then using the fact that $\mathrm{RE}(\pi\|x_1) \le \ln N$ and $\mathrm{RE}(\pi\|x_{T+1}) \ge 0$. Finally, by substituting these two bounds into (6), we have

$$\sum_{t=1}^{T}\left(f_t(\hat{x}_t) - f_t(\pi)\right) \le \eta D_\infty + \frac{1}{\eta}\ln N \le O\left(\sqrt{D_\infty \ln N}\right),$$

by choosing $\eta = \sqrt{(\ln N)/D_\infty}$, which proves the theorem.

Appendix D. Proofs in Section 6

D.1. Proof of Lemma 14

We start by bounding the first sum in (10). Note that we can apply Lemma 6 with the norm $\|\cdot\| = \|\cdot\|_{H_t}$, since $\frac{1}{2}\|x - x'\|^2 = \frac{1}{2}\|x - x'\|_{H_t}^2 = \mathcal{B}^{\mathcal{R}_t}(x, x')$ for any $x, x' \in \mathcal{X}$. As the dual norm is $\|\cdot\|_* = \|\cdot\|_{H_t^{-1}}$, Lemma 6 gives us

$$\sum_{t=1}^{T} S_t \le \sum_{t=1}^{T} \left\|\ell_t - \ell_{t-1}\right\|_*^2 \le \sum_{t=1}^{T} \left\|\ell_t - \ell_{t-1}\right\|_{H_t^{-1}}^2.$$

Next, we bound the second sum $\sum_{t=1}^{T} A_t$ in (10), which can be written as

$$\frac{1}{2}\|\pi - x_1\|_{H_1}^2 - \frac{1}{2}\|\pi - x_{T+1}\|_{H_{T+1}}^2 + \frac{1}{2}\sum_{t=1}^{T}\left(\|\pi - x_{t+1}\|_{H_{t+1}}^2 - \|\pi - x_{t+1}\|_{H_t}^2\right).$$

Since $\|\pi - x_1\|_{H_1}^2 = O\left(1 + \beta\gamma^2\right)$, $\|\pi - x_{T+1}\|_{H_{T+1}}^2 \ge 0$, and $H_{t+1} - H_t = \beta h_t$, we have

$$\sum_{t=1}^{T} A_t \le O\left(1 + \beta\gamma^2\right) + \frac{\beta}{2}\sum_{t=1}^{T}\|\pi - x_{t+1}\|_{h_t}^2.$$

Note that unlike in the case of linear functions, here the sum does not telescope and hence we do not have a small bound for it. The last sum $\sum_{t=1}^{T} C_t$ in (10) now comes to help. Recall that $C_t = \beta \|\pi - \hat{x}_t\|_{h_t}^2$, so by Proposition 1,

$$\frac{\beta}{2}\|\pi - x_{t+1}\|_{h_t}^2 - C_t \le \beta\|\pi - \hat{x}_t\|_{h_t}^2 + \beta\|\hat{x}_t - x_{t+1}\|_{h_t}^2 - C_t = \beta\|\hat{x}_t - x_{t+1}\|_{h_t}^2,$$

which, by the fact that $H_t \succeq \beta\gamma^2 I \succeq \beta h_t$ and the bound in (5), is at most

$$\|\hat{x}_t - x_{t+1}\|_{H_t}^2 \le \left\|\ell_t - \ell_{t-1}\right\|_{H_t^{-1}}^2.$$

Combining the bounds derived so far, we obtain

$$\sum_{t=1}^{T} S_t + \sum_{t=1}^{T} A_t - \sum_{t=1}^{T} C_t \le O(1+\beta\gamma^2) + 2\sum_{t=1}^{T}\left\|\ell_t - \ell_{t-1}\right\|_{H_t^{-1}}^2. \tag{16}$$

Finally, to complete our proof of Lemma 14, we rely on the following, which provides a bound for the last term in (16) and will be proved in Appendix D.2.

Lemma 19 $\sum_{t=1}^{T}\left\|\ell_t - \ell_{t-1}\right\|_{H_t^{-1}}^2 \le \frac{4N}{\beta}\ln\left(1 + \frac{\beta}{4}\sum_{t=1}^{T}\left\|\ell_t - \ell_{t-1}\right\|_2^2\right)$.

D.2. Proof of Lemma 19

We need the following lemma from (Hazan et al., 2007):

Lemma 20 *Let $u_t \in \mathbb{R}^N$, for $t \in [T]$, be a sequence of vectors. Define $V_t = I + \sum_{\tau=1}^{t} u_\tau u_\tau^\top$. Then,*

$$\sum_{t=1}^{T} u_t^\top V_t^{-1} u_t \le N \ln\left(1 + \sum_{t=1}^{T}\|u_t\|_2^2\right).$$

To prove our Lemma 19, first note that for any $t \in [T]$,

$$H_t = I + \beta\gamma^2 I + \beta\sum_{\tau=1}^{t-1}\ell_\tau\ell_\tau^\top \succeq I + \beta\sum_{\tau=1}^{t}\ell_\tau\ell_\tau^\top \succeq I + \frac{\beta}{2}\sum_{\tau=1}^{t}\left(\ell_\tau\ell_\tau^\top + \ell_{\tau-1}\ell_{\tau-1}^\top\right),$$

since $\gamma^2 I \succeq \ell_t\ell_t^\top$ and ℓ_0 is the the all-0 vector. Next, we claim that

$$\ell_\tau\ell_\tau^\top + \ell_{\tau-1}\ell_{\tau-1}^\top \succeq \frac{1}{2}\left(\ell_\tau - \ell_{\tau-1}\right)\left(\ell_\tau - \ell_{\tau-1}\right)^\top.$$

This is because by subtracting the right-hand side from the left-hand side, we have

$$\frac{1}{2}\ell_\tau\ell_\tau^\top + \frac{1}{2}\ell_\tau\ell_{\tau-1}^\top + \frac{1}{2}\ell_{\tau-1}\ell_\tau^\top + \frac{1}{2}\ell_{\tau-1}\ell_{\tau-1}^\top = \frac{1}{2}\left(\ell_\tau + \ell_{\tau-1}\right)\left(\ell_\tau + \ell_{\tau-1}\right)^\top \succeq 0.$$

Thus, with $K_t = I + \frac{\beta}{4}\sum_{\tau=1}^{t}\left(\ell_\tau - \ell_{\tau-1}\right)\left(\ell_\tau - \ell_{\tau-1}\right)^\top$, we have $H_t \succeq K_t$ and $K_t^{-1} \succeq H_t^{-1}$. This implies that

$$\sum_{t=1}^{T}\left\|\ell_t - \ell_{t-1}\right\|_{H_t^{-1}}^2 \le \sum_{t=1}^{T}\left\|\ell_t - \ell_{t-1}\right\|_{K_t^{-1}}^2 = \frac{4}{\beta}\sum_{t=1}^{T}\left\|\sqrt{\frac{\beta}{4}}\left(\ell_t - \ell_{t-1}\right)\right\|_{K_t^{-1}}^2,$$

which by Lemma 20 is at most $\frac{4N}{\beta}\ln\left(1 + \frac{\beta}{4}\sum_{t=1}^{T}\left\|\ell_t - \ell_{t-1}\right\|_2^2\right)$.

D.3. Proof of Theorem 15

To bound the last term in the bound of Lemma 14 in terms of our deviation bound D_2, we use Lemma 12 in Section 5. Combining this with (10), we can upper-bound $\sum_{t=1}^{T}(f_t(\hat{x}_t) - f_t(\pi))$ by

$$O(1+\beta\gamma^2) + \frac{8N}{\beta}\ln\left(1 + \frac{\beta}{2}D_2 + \frac{\beta\lambda^2}{2}\sum_{t=1}^{T}\|\hat{x}_t - \hat{x}_{t-1}\|_2^2\right) - \sum_{t=1}^{T} B_t. \tag{17}$$

To eliminate the undesirable last term inside the parenthesis above, we need the help from the sum $\sum_{t=1}^{T} B_t$, which has the following bound.

Lemma 21 $\sum_{t=1}^{T} B_t \geq \frac{1}{4}\sum_{t=1}^{T}\|\hat{x}_t - \hat{x}_{t-1}\|_2^2 - O(1)$.

Proof Recall that $B_t = \frac{1}{2}\|x_{t+1} - \hat{x}_t\|_{H_t}^2 + \frac{1}{2}\|\hat{x}_t - x_t\|_{H_t}^2$, so we can write $\sum_{t=1}^{T} B_t$ as

$$\frac{1}{2}\sum_{t=2}^{T+1}\|x_t - \hat{x}_{t-1}\|_{H_{t-1}}^2 + \frac{1}{2}\sum_{t=1}^{T}\|\hat{x}_t - x_t\|_{H_t}^2 \geq \frac{1}{2}\sum_{t=2}^{T}\|x_t - \hat{x}_{t-1}\|_{H_{t-1}}^2 + \frac{1}{2}\sum_{t=2}^{T}\|\hat{x}_t - x_t\|_{H_{t-1}}^2$$

since $H_t \succeq H_{t-1}$ and $\|x_{T+1} - \hat{x}_T\|_{H_T}^2, \|\hat{x}_1 - x_1\|_{H_1}^2 \geq 0$. By Proposition 1, this is at least

$$\frac{1}{4}\sum_{t=2}^{T}\|\hat{x}_t - \hat{x}_{t-1}\|_{H_{t-1}}^2 \geq \frac{1}{4}\sum_{t=2}^{T}\|\hat{x}_t - \hat{x}_{t-1}\|_I^2 = \frac{1}{4}\sum_{t=2}^{T}\|\hat{x}_t - \hat{x}_{t-1}\|_2^2$$

as $H_{t-1} \succeq I$ and I is the identity matrix. Then the lemma follows as $\|\hat{x}_2 - \hat{x}_1\|_2^2 \leq O(1)$. ■

Applying this lemma to (17), we obtain a regret bound of the form

$$O(1+\beta\gamma^2) + \frac{8N}{\beta}\ln\left(1 + \frac{\beta}{2}D_2 + \frac{\beta\lambda^2}{2}W\right) - \frac{1}{4}W$$

where $W = \sum_{t=1}^{T}\|\hat{x}_t - \hat{x}_{t-1}\|_2^2$. Observe that the combination of the last two terms above become negative when $W \geq (\lambda N D_2)^c/\beta$ for some large enough constant c, as we assume $\beta \leq 1$ and $\lambda, D_2 \geq 1$. Thus, the regret bound is at most $O(\beta\gamma^2 + (N/\beta)\ln(\lambda N D_2))$, which completes the proof of Theorem 15.

D.4. Proof of Corollary 16

Recall that each loss function has the form $f_t(x) = -\ln\langle v_t, x\rangle$ for some $v_t \in [\delta, 1]^N$ with $\delta \in (0,1)$, and note that $\nabla f_t(x) = -v_t/\langle v_t, x\rangle$. To apply Theorem 15, we need to determine the parameters $\beta, \gamma, \lambda, D_2$.

First, by a Taylor expansion, we know that for any $x, y \in \mathcal{X}$, there is some ξ_t on the line between x and y such that

$$f_t(x) = f_t(y) + \langle \nabla f_t(y), x - y\rangle + \frac{1}{2\langle v_t, \xi_t\rangle^2}(x-y)^\top v_t v_t^\top (x-y),$$

where the last term above equals

$$\frac{1}{2\langle v_t, \xi_t\rangle^2}\langle v_t, x-y\rangle^2 = \frac{\langle v_t, y\rangle^2}{2\langle v_t, \xi_t\rangle^2}\langle \nabla f_t(y), x-y\rangle^2 \geq \frac{\delta^2}{2}\langle \nabla f_t(y), x-y\rangle^2 .$$

Thus, we can choose $\beta = \delta^2/2$. Next, since $\|\nabla f_t(x)\|_2 = \|v_t\|_2 / \langle v_t, x\rangle \leq \sqrt{N}/\delta$, we can choose $\gamma = \sqrt{N}/\delta$. Third, note that

$$\|\nabla f_t(x) - \nabla f_t(y)\|_2 = \left\| \frac{v_t}{\langle v_t, x\rangle} - \frac{v_t}{\langle v_t, y\rangle} \right\|_2 = \frac{\|v_t\|_2 |\langle v_t, x-y\rangle|}{\langle v_t, x\rangle \langle v_t, y\rangle},$$

which by a Cauchy-Schwarz inequality is at most

$$\frac{\|v_t\|_2^2}{\langle v_t, x\rangle \langle v_t, y\rangle}\|x-y\|_2 \leq \frac{N}{\delta^2}\|x-y\|_2 .$$

Thus, we can choose $\lambda = N/\delta^2$. Finally, note that for any $x \in \mathcal{X}$,

$$\|\nabla f_t(x) - \nabla f_{t-1}(x)\|_2 = \left\| \frac{v_t}{\langle v_t, x\rangle} - \frac{v_{t-1}}{\langle v_{t-1}, x\rangle} \right\|_2 = \left\| \frac{v_t - v_{t-1}}{\langle v_t, x\rangle} + \frac{v_{t-1}\left(\langle v_{t-1}, x\rangle - \langle v_t, x\rangle\right)}{\langle v_t, x\rangle \langle v_{t-1}, x\rangle} \right\|_2,$$

which by a triangle inequality and then a Cauchy-Schwarz inequality is at most

$$\frac{\|v_t - v_{t-1}\|_2}{\langle v_t, x\rangle} + \frac{\|v_{t-1}\|_2 |\langle v_{t-1} - v_t, x\rangle|}{\langle v_t, x\rangle \langle v_{t-1}, x\rangle} \leq \frac{\|v_t - v_{t-1}\|_2}{\langle v_t, x\rangle} + \frac{\|v_{t-1}\|_2 \|x\|_2 \|v_t - v_{t-1}\|_2}{\langle v_t, x\rangle \langle v_{t-1}, x\rangle},$$

which in turn is at most $\left(\frac{1}{\delta} + \frac{\sqrt{N}}{\delta^2}\right)\|v_t - v_{t-1}\|_2 \leq \frac{2\sqrt{N}}{\delta^2}\|v_t - v_{t-1}\|_2$. This implies

$$\sum_{t=1}^{T} \max_{x\in\mathcal{X}} \|\nabla f_t(x) - \nabla f_{t-1}(x)\|_2^2 \leq \sum_{t=1}^{T} \left(\frac{2\sqrt{N}}{\delta^2}\right)^2 \|v_t - v_{t-1}\|_2^2 \leq \left(\frac{4N}{\delta^4}\right) D.$$

Thus, we can choose $D_2 = (4N/\delta^4)D$. Using these bounds in Theorem 15, we have the corollary.