The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Online Learning for Structured Loss Spaces

Siddharth Barman, Aditya Gopalan, Aadirupa Saha
Indian Institute of Science
Bangalore 560012
{barman, aditya, aadirupa}@iisc.ac.in

Abstract

We consider prediction with expert advice when the loss vectors are assumed to lie in a set described by the sum of atomic norm balls. We derive a regret bound for a general version of the online mirror descent (OMD) algorithm that uses a combination of regularizers, each adapted to the constituent atomic norms. The general result recovers standard OMD regret bounds, and yields regret bounds for new structured settings where the loss vectors are (i) noisy versions of vectors from a low-dimensional subspace, (ii) sparse vectors corrupted with noise, and (iii) sparse perturbations of low-rank vectors. For the problem of online learning with structured losses, we also show lower bounds on regret in terms of rank and sparsity of the loss vectors, which implies lower bounds for the above additive loss settings as well.

1 Introduction

Online learning problems, such as prediction with expert advice (Cesa-Bianchi and Lugosi 2006) and online convex optimization (Zinkevich 2003), involve a learner who sequentially makes decisions from a decision set. The learner seeks to minimize her total loss over a sequence of loss functions, unknown at the beginning, but which is revealed causally. Specifically, she attempts to achieve low regret, for each sequence in a class of loss sequences, with respect to the best single decision point in hindsight.

The theory of online learning, by now, has yielded flexible and elegant algorithmic techniques that enjoy provably sublinear regret in the time horizon of plays. Regret bounds for online learning algorithms typically hold across inputs (loss function sequences) that have little or no structure. For instance, for the prediction with experts problem, the exponentially weighted forecaster (Cesa-Bianchi and Lugosi 2006) is known to achieve an expected regret of $O(\sqrt{T \ln N})$ over any sequence of N-dimensional loss vectors with coordinates bounded in $[0, 1]$; here T is the number of rounds of play.

There is often, however, more structure in the inputs of online learning problems beyond elementary ℓ_∞-type constraints, which a learner with a priori knowledge can hope to exploit and improve her performance. A notable example is when the loss vectors for the prediction with experts problem come from a low-dimensional subspace (Hazan et al. 2016). This is often the case in recommender systems based on latent factor models (Koren, Bell, and Volinsky 2009), where users and items are represented in terms of their features or attribute vectors, typically of small dimension. Under a bilinear model for the utility of a user-item pair, each user's utility across all items becomes a vector from a subspace of dimension at most the size of the feature vectors. (Hazan et al. 2016) show that in this setup the learner can limit her regret to $O(d\sqrt{T})$ when each loss vector comes from a d-dimensional subspace of $\mathbb{R}^N$. If $d \ll N$ (in fact, $d = o\left(\sqrt{\ln N}\right)$), then this is potentially advantageous over a more general best-experts algorithm like Exponential Weights.

This example is interesting not only because it shows that geometric/structural properties known in advance can help the learner achieve order-wise better regret, but also because it opens up the possibility of studying whether other, arguably more realistic, forms of structure can be exploited, such as sparsity in the input (or more generally small norm) and, more importantly, "additive" combinations of such structures, e.g., low-rank losses added with losses of small ℓ_2-norm, which expresses losses that are noisy perturbations of a low-dimensional subspace. In this paper, we take a step in this direction and develop a framework for online learning problems with structured losses.

Our Results and Techniques: We consider the prediction with experts problem with loss sequences in which each element (loss vector) belongs to a set that respects structural constraints. Specifically, we assume that the loss vectors belong to a sum of atomic norm balls[1] (Chandrasekaran et al. 2012), say $A + B$, where the sum of sets, A and B, is in the Minkowski sense.[2] For this setup—which we call online learning with *additive loss* spaces—we show a general regret guarantee for an online mirror descent (OMD) algorithm that uses a combination of regularizer functions, each

A full version of this paper is available at https://arxiv.org/abs/1706.04125

[1] centrally symmetric, convex, compact sets with their centroids at the origin.

[2] $A + B = \{a + b : a \in A, b \in B\}$.

of which is adapted to a constituent atomic norms of A and B, respectively.

Specializing this result for a variety of loss-function sets recovers standard OMD regret guarantees for strongly convex regularizers (Shalev-Shwartz 2012a), and subsumes a result for the online low-rank problem (Hazan et al. 2016). But more importantly, this allows us to obtain "new results from old"—regret guarantees for settings such as noisy low rank (where losses are perturbations from a low-dimensional subspace), noisy sparse (where losses are perturbations of sparse vectors), and sparse low-rank (where losses are sparse perturbations from a low-dimensional subspace); see Tables 1 and 2.

Another contribution of this work is to show lower bounds on regret for online learning with structured losses. We derive a generic lower bound on regret, for any algorithm for the prediction with experts problem, using structured (in terms of sparsity and dimension) loss vectors. This result allows us to derive regret lower bounds in a variety of individual and additive loss space settings including sparse, noisy, low rank, noisy low-rank, and noisy sparse losses.

Related work. The work that is perhaps closest in spirit to ours is that of (Hazan et al. 2016), who study the best experts problem when the loss vectors all come from a low-dimensional subspace of the ambient space. A key result of theirs is that the online mirror descent (OMD) algorithm, used with a suitable regularization, improves the regret to depend only on the low rank and not the ambient dimension. More broadly, OMD theory provides regret bounds depending on properties of the regularizer and the geometry of the loss and decision spaces (Shalev-Shwartz 2012b). We are able to notably generalize this to the more flexible setting of additive losses.

Online learning with structure has been studied in the recent past from the point of view of overall sequence complexity or "hardness" (learning with "easy data"). This includes work that shows algorithms enjoying first- and second-order regret bounds (Cesa-Bianchi, Mansour, and Stoltz 2007), and with performance depending on the quadratic variation of the inputs (Hazan and Kale 2010; Steinhardt and Liang 2014). There is also recent work on achieving regret scaling with the covering number of the sequence of observed loss vectors (Cohen and Mannor 2017), which is another notion of easy data.

Our problem formulation explores a different formulation of learning with easy data, in which the adversary, instead of being constrained to choose loss sequences with low total magnitude or variation, chooses loss vectors from sets with enough geometric structure (atomic norm balls).

2 Notation and Preliminaries

For an integer $n \in \mathbb{Z}_+$, we use $[n]$ to denote the set $\{1, 2, \dots n\}$. For a vector $\mathbf{x} \in \mathbb{R}^n$, x_i denotes the ith component of $\mathbf{x}$. The p-norm of $\mathbf{x}$ is defined as $\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$, $1 \le p < \infty$. Write $\|\mathbf{x}\|_\infty := \max_{i=1}^n |x_i|$ and $\|\mathbf{x}\|_0 := |\{i \mid x_i \neq 0\}|$. If $\|\cdot\|$ is a norm defined on a closed convex set $\Omega \subseteq \mathbb{R}^n$, then its corresponding *dual norm* is defined as

$$\|\mathbf{u}\|^* = \sup_{\mathbf{x}\in\Omega:\|\mathbf{x}\|\le 1} \mathbf{x}\cdot\mathbf{u},$$

where $\mathbf{x}\cdot\mathbf{u} = \sum_i x_i u_i$ is the standard inner product. It follows that the dual of the standard p-norm ($p \ge 1$) is the q-norm, where q is the Hölder conjugate of p, i.e., $\frac{1}{p} + \frac{1}{q} = 1$. The *$n$-probability simplex* is defined as $\Delta_n = \{\mathbf{x} \in [0,1]^n \mid \sum_{i=1}^n x_i = 1\}$. Given any set $\mathcal{A} \subseteq \mathbb{R}^n$, we denote the convex hull of $\mathcal{A}$ as $\mathrm{conv}(\mathcal{A})$. Clearly, when $\mathcal{A} = \{\mathbf{e}_1, \mathbf{e}_2, \dots \mathbf{e}_n\}$, $\mathrm{conv}(\mathcal{A}) = \Delta_n$, where $\mathbf{e}_i \in [0,1]^n$ denotes ith standard basis vector of $\mathbb{R}^n$.

2.1 Atomic Norm and its Dual (Chandrasekaran et al. 2012)

The notion of an atomic norm along with its dual will provide us with a unified framework for addressing structured loss spaces, and will be used extensively in the paper. Let $\mathcal{A} \subseteq \mathbb{R}^n$ be a set which is convex, compact, and centrally symmetric about the origin (i.e., $\mathbf{a} \in \mathcal{A}$ if and only if $-\mathbf{a} \in \mathcal{A}$).

The atomic norm induced by the set $\mathcal{A}$ is defined as

$$||\mathbf{x}||_{\mathcal{A}} := \inf\{t > 0 \mid \mathbf{x} \in t\mathcal{A}\}, \quad \text{for } \mathbf{x} \in \mathbb{R}^n.$$

The dual of the atomic norm induced by $\mathcal{A}$ becomes the *support function* of $\mathcal{A}$ (Boyd and Vandenberghe 2004); formally,

$$||\mathbf{x}||^*_{\mathcal{A}} := \sup\{\mathbf{x}.\mathbf{z} \mid \mathbf{z} \in \mathcal{A}\}, \quad \text{for } \mathbf{x} \in \mathbb{R}^n.$$

For example, if the set $\mathcal{A}$ is the convex hull of all unit-norm one-sparse vectors, i.e., $\mathcal{A} := \mathrm{conv}\left(\{\pm\mathbf{e}_i\}_{i=1}^n\right)$, then the corresponding atomic norm is the standard ℓ_1-norm $\|\mathbf{x}\|_1 = \sum_i |x_i|$.

2.2 Problem setup

We consider the online learning problem of learning with expert advice from a collection of N experts (Cesa-Bianchi and Lugosi 2006). In each round $t = 1, 2, \dots, T$, the learner receives advice from each of the N experts, following which the learner selects an expert from a distribution $\mathbf{p}_t \in \Delta_N$, maintained over the experts, whose advice is to be followed. Upon this, the adversary reveals the losses incurred by the N experts, $\boldsymbol{l}_t = \big(l_t(1), l_t(2), \dots, l_t(N)\big) \in [0,1]^N$, $l_t(i)$ being the loss incurred by the i^{th} expert. The learner suffers an expected loss of $\mathbf{E}_{I_t\sim\mathbf{p}_t}[l_t(I_t)] = \sum_{i=1}^N p_t(i) l_t(i)$. If the game is played for a total of T rounds, then the objective of the learner is to minimize the expected cumulative regret defined as:

$$\mathbf{E}\left[\text{Regret}_T\right] = \sum_{t=1}^T \mathbf{p}_t.\boldsymbol{l}_t - \min_{i\in[N]} \sum_{t=1}^T l_t(i).$$

It is well-known that without any further assumptions over the losses $\boldsymbol{l}_t$, the best achievable regret for this problem is $\Theta(\sqrt{T \ln N})$. Indeed, the exponential weights algorithm or the Hedge algorithm achieves regret $O(\sqrt{T \ln N})$ (Arora, Hazan, and Kale 2012, Theorem 2.3), and a matching lower

bound exists as well (Cesa-Bianchi and Lugosi 2006, Theorem 3.7).

Now, a very natural question to ask is: can a better (smaller) regret be achieved if the loss sequence has more structure? Suppose the loss vectors $(\boldsymbol{l}_t)_{t=1}^T$ all belong to a common *structured loss space* $\mathcal{L} \subseteq [0,1]^N$, such as:

1. Sparse loss space: $\mathcal{L} = \{\boldsymbol{l} \in [0,1]^N \mid \|\boldsymbol{l}\|_0 = s\}$. Here, $s \in [N]$ is the sparsity parameter.
2. Spherical loss space: $\mathcal{L} = \{\boldsymbol{l} \in [0,1]^N \mid \|\boldsymbol{l}\|_{\mathbf{A}}^2 = \boldsymbol{l}^\top \mathbf{A} l \leq \epsilon\}$, where $\mathbf{A}$ is a positive definite matrix and $\epsilon > 0$.
3. Noisy loss space: $\mathcal{L} = \{\boldsymbol{l} \in [0,1]^N \mid \|\boldsymbol{l}\|_2^2 = \epsilon\},\ \epsilon > 0\}$. Note that *noisy losses* are a special class of *spherical losses* where $\mathbf{A} = \mathbf{I}_N$, the identity matrix.
4. Low-rank loss space: $\mathcal{L} = \{\boldsymbol{l} \in [0,1]^N \mid \boldsymbol{l} = \mathbf{U}\mathbf{v}\ \}$, here the rank of matrix $\mathbf{U} \in \mathbb{R}^{N \times d}$ is equal to $d \in [N]$ and vector $\mathbf{v} \in \mathbb{R}^d$ (as mentioned previously, such loss vectors were considered by (Hazan et al. 2016)).
5. Additive loss space: $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$ (Minkowski Sum). More formally, $\mathcal{L} = \{\boldsymbol{l} = \boldsymbol{l}_1 + \boldsymbol{l}_2 \mid \boldsymbol{l}_1 \in \mathcal{L}_1 \text{ and } \boldsymbol{l}_2 \in \mathcal{L}_2\}$, where $\mathcal{L}_1 \subseteq [0,1]^N$ and $\mathcal{L}_2 \subseteq [0,1]^N$ are structured loss spaces themselves.[3] Examples include any combination of the previously described loss spaces, such as the low-rank + noisy space.

The Exponential Weight or Hedge algorithm achieves $O(\sqrt{T \ln N})$ regret (Cesa-Bianchi and Lugosi 2006; Shalev-Shwartz 2012b) in all of the above settings. The relevant question is whether the geometry of such loss spaces can be exploited in a principled fashion to achieve improved regret guarantees (possibly independent of $\ln N$). In other words, can we come up with algorithms for above cases such that the regret is $O(\sqrt{\omega T})$, where $\omega < \ln N$?

We will show that, for all of the above loss spaces, we can obtain a regret factor ω which is order-wise better than $\ln N$. In particular, we will establish these regret bounds by employing the Online Mirror Descent algorithm (described below) with a right choice of atomic norms. Furthermore, using this algorithm, we will also develop a framework to obtain new regret bounds from old. That is, we show that if we have an online mirror descent setup for $\mathcal{L}_1$ and $\mathcal{L}_2$, then we can in fact obtain a low-regret algorithm for the additive loss space $\mathcal{L}_1 + \mathcal{L}_2$.

2.3 Online Mirror Descent

In this section, we give a brief introduction to the Online Mirror Descent (OMD) algorithm (Bubeck 2011; Shalev-Shwartz 2012b), which is a subgradient descent based method for online convex optimization with a suitably chosen regularizer. A reader well-versed with the analysis of OMD may skip the statement of Theorem 3 and proceed to the next section.

OMD generalizes the basic mirror descent algorithm used for offline optimization problems (see, e.g., (Beck and Teboulle 2003)). Before detailing the algorithm, we will recall a few relevant definitions:

Definition 1. ***Bregman Divergence.*** *Let $\Omega \in \mathbb{R}^n$ be a convex set, and $f : \Omega \to \mathbb{R}$ be a strictly convex and differentiable function. Then the* Bregman divergence *associated with f, denoted by $B_f : \Omega \times \Omega \to \mathbb{R}$, is defined as*

$$B_f(\mathbf{u}, \mathbf{v}) = f(\mathbf{u}) - f(\mathbf{v}) - (\mathbf{u} - \mathbf{v}) \cdot \nabla f(\mathbf{v}),\ \forall \mathbf{u}, \mathbf{v} \in \Omega\,.$$

Definition 2. Strong Convexity *Let $\Omega \in \mathbb{R}^n$ be a convex set, and $f : \Omega \to \mathbb{R}$ be a differentiable function. Then f is called α-strongly convex over Ω with respect to the norm $\|\cdot\|$ iff for all $\mathbf{x}, \mathbf{y} \in \Omega$,*

$$f(\mathbf{x}) - f(\mathbf{y}) - (\nabla f(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \geq \frac{\alpha}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

Equivalently, a continuous twice differentiable function, f, over Ω is said to be α-strongly convex iff for all $\mathbf{x}, \mathbf{w} \in \Omega$ we have

$$\mathbf{x}^T \nabla^2 f(\mathbf{w}) \mathbf{x} \geq \alpha \|\mathbf{x}\|^2.$$

We now describe the OMD algorithm for the online learning problem setup (Sec. 2.2).

Algorithm 1 Online Mirror Descent (OMD)

Parameters: Learning rate $\eta > 0$.
  Convex set $\Omega \subseteq \mathbb{R}^N$, such that $\Delta_N \subseteq \Omega$
  Strictly convex, differentiable function $R : \Omega \to \mathbb{R}$
Initialize: $\mathbf{p}_1 = \operatorname{argmin}_{\mathbf{p} \in \Delta_N} R(\mathbf{p})$
for $t = 1, 2, \cdots T$ **do**
  Play $\mathbf{p}_t \in \Delta_N$
  Receive loss vector $\boldsymbol{l}_t \in [0,1]^N$
  Incur loss $\mathbf{p}_t.\boldsymbol{l}_t$
  Update:
    $\nabla R(\tilde{\mathbf{p}}_{t+1}) \leftarrow \nabla R(\mathbf{p}_t) - \eta \boldsymbol{l}_t$ (Assuming $\tilde{\mathbf{p}}_{t+1} \in \Omega$)
    $\mathbf{p}_{t+1} \leftarrow \operatorname{argmin}_{\mathbf{p} \in \Delta_N} B_R(\mathbf{p}, \tilde{\mathbf{p}}_{t+1})$
end for

The following regret guarantee for the above algorithm is well-known.

Theorem 3 (**OMD regret bound** (**Theorem** 5.2**, (Bubeck 2011))).** *Let the loss vectors, $\{\boldsymbol{l}_t\}_{t=1}^T$, belong to a loss space $\mathcal{L} \subseteq [0,1]^N$, which is bounded with respect to a (arbitrary) norm $\|\cdot\|$; in particular, for any $\boldsymbol{l} \in \mathcal{L}$ we have $\|\boldsymbol{l}\| \leq G$. Furthermore, let $\Omega \supseteq \Delta_N$ be a convex set, and $R : \Omega \to \mathbb{R}$ be a strictly convex, differentiable function that satisfies $R(\mathbf{p}) - R(\mathbf{p}_1) \leq D^2$ for parameter $D \in \mathbb{R}$ and all $\mathbf{p} \in \Delta_N$; where $\mathbf{p}_1 := \operatorname{argmin}_{p \in \Delta_N} R(\mathbf{p})$. Also, let the restriction of R to Δ_N be α-strongly convex with respect to $\|\cdot\|^*$, the dual norm of $\|\cdot\|$.*

Then, the regret of OMD algorithm with set Ω, regularizer function R, and learning rate $\eta > 0$, for T rounds satisfies

$$\begin{aligned} \textit{Regret}_T(\text{OMD}(\eta)) &= \sum_{t=1}^T \mathbf{p}_t.\boldsymbol{l}_t - \min_{i=1}^N \sum_{t=1}^T l_t(i) \\ &\leq \frac{1}{\eta}\left(D^2 + \frac{\eta^2 G^2 T}{2\alpha}\right), \end{aligned}$$

[3] Note that, in the problem setup at hand the learner observes only the loss vectors l_t, and does not have access to the loss components l_{1t} or l_{2t}.

Loss Space	Regret Bound	Atomic Norm	Regularizer
s-Sparse	$2\sqrt{\ln(s+1)T}$	$\frac{1}{\sqrt{2}}\|\cdot\|_p$ $(p = 2\ln(s+1))$	$\|\mathbf{x}\|_q^2$ $(q = \frac{p}{p-1})$
Spherical	$\sqrt{\epsilon\lambda_{\max}(\mathbf{A}^{-1})T}$	$\frac{1}{\sqrt{\epsilon}}\|\cdot\|_A$	$\epsilon\mathbf{x}^\top\mathbf{A}^{-1}\mathbf{x}$
ϵ-Noise	$\sqrt{\epsilon T}$	$\frac{1}{\sqrt{\epsilon}}\|\cdot\|_2$	$\epsilon\mathbf{x}^\top\mathbf{x}$

Table 1: OMD Regret Bounds for Structured Loss Spaces

where $\mathbf{p}_1, \mathbf{p}_2, \ldots \mathbf{p}_T$ denotes the sequential predictions of the algorithm in T rounds. Moreover, setting $\eta^* = \frac{D}{G}\sqrt{\frac{2\alpha}{T}}$ (i.e., minimizing the right-hand-side of the above bound), we have

$$\textit{Regret}_T(\text{OMD}(\eta^*)) \;\leq\; DG\sqrt{\frac{2T}{\alpha}}\,.$$

3 Online Mirror Descent for Structured Losses

This section shows that, for specific structured loss spaces, instantiating the OMD algorithm—with a right choice of the norm $\|\cdot\|$ and regularizer R—leads to improved (over the standard $O(\sqrt{T\ln N})$ bound) regret guarantees. Proofs of these results appear in the full version of the paper.

1. Sparse loss space: $\mathcal{L} = \{\boldsymbol{l} \in [0,1]^N \mid \|\boldsymbol{l}\|_0 = s\}$, $s \in [N]$ being the loss sparsity parameter. Then using the q-norm, $R(\mathbf{x}) = \|\mathbf{x}\|_q^2 = \left(\sum_{i=1}^N x_i^q\right)^{\frac{2}{q}}$, where $q = \frac{\ln s'}{\ln s' - 1}$, $s' = (s+1)^2$, as the regularizer, we get,

$$\text{Regret}_T \;\leq\; 2\sqrt{\ln(s+1)T}\,.$$

2. Spherical loss space: $\mathcal{L} = \{\boldsymbol{l} \in [0,1]^N \mid \|\boldsymbol{l}\|_{\mathbf{A}}^2 = \boldsymbol{l}^\top\mathbf{A}l \leq \epsilon\}$, where $\mathbf{A}$ is a positive definite matrix, $\epsilon > 0$. Then using the square of the ellipsoidal norm as the regularizer, $R(\mathbf{x}) = \epsilon\|\mathbf{x}\|_{\mathbf{A}^{-1}}^2 = \epsilon\mathbf{x}^\top\mathbf{A}^{-1}\mathbf{x}$, we get,

$$\text{Regret}_T \;\leq\; \sqrt{\lambda_{\max}(\mathbf{A}^{-1})\epsilon T}\,,$$

where $\lambda_{\max}(\mathbf{A}^{-1})$ denotes the maximum eigenvalue of $\mathbf{A}^{-1}$.

3. Noisy loss sapce: $\mathcal{L} = \{\boldsymbol{l} \in [0,1]^N \mid \|\boldsymbol{l}\|_2^2 \leq \epsilon\}$, $\epsilon > 0$. Then using the square of the standard Euclidean norm as the regularizer, $R(\mathbf{x}) = \epsilon\|\mathbf{x}\|_2^2$, we get,

$$\text{Regret}_T \;\leq\; \sqrt{\epsilon T}\,.$$

Note that the *noisy-loss* case is a special case of the *spherical-loss* setting, where $\mathbf{A} = \mathbf{A}^{-1} = \mathbf{I}_N$.

(Hazan et al. 2016) have also used OMD to address the loss vectors that belong to a low-dimensional subspace. Specifically, if the loss space $\mathcal{L} = \{\boldsymbol{l} \in [0,1]^N \mid \boldsymbol{l} = \mathbf{U}\mathbf{v}\}$, with $\mathbf{U} \in \mathbb{R}^{N\times d}$ being a rank d matrix and vector $\mathbf{v} \in \mathbb{R}^d$. They have shown that the regularizer $R(\mathbf{x}) = \|\mathbf{x}\|_{\mathbf{H}}^2 = \mathbf{x}^\top\mathbf{H}\mathbf{x}$ (where $\mathbf{H} = \mathbf{I}_N + \mathbf{U}^\top\mathbf{M}\mathbf{U}$, $\mathbf{M}$ is the matrix corresponding to the Löwner-John ellipsoid (Hazan et al. 2016) of $\mathcal{L}$ and $\mathbf{I}_N$ is the identity matrix) leads to the following regret bound:

$$\text{Regret}_T \;\leq\; 4\sqrt{dT}\,.$$

In addition, for the standard loss space $\mathcal{L} = [0,1]^n$, one can execute the OMD algorithm with the unnormalized negative entropy, $R(\mathbf{x}) = \sum_{i=1}^N x_i \log x_i - \sum_{i=1}^N x_i$, as the regularizer, to obtain:

$$\text{Regret}_T \;\leq\; \sqrt{2T\ln N}\,.$$

Note that the above regret bound is the same as that for the Hedge algorithm. In fact, it can be verified that, with the above choice of regularizer, the OMD algorithm exactly reduces to Hedge (Bubeck 2011).

4 Online Learning for Additive Losses

We now present a key result of this paper, which enables us to obtain new regret bounds from old. In particular, we will develop a framework that provides a low-regret OMD algorithm for an additive loss space $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$, using the OMD setup of the constituent loss spaces $\mathcal{L}_1$ and $\mathcal{L}_2$. Specifically, we detail how to choose an appropriate regularizer for losses from $\mathcal{L}$ and, hence, construct a low-regret OMD algorithm.

Theorem 4. ***(Main Result)*** *Let $\mathcal{L}_1, \mathcal{L}_2 \subseteq [0,1]^N$ be two loss spaces, such that $\mathcal{L}_1 \subseteq \mathcal{A}_1$, $\mathcal{L}_2 \subseteq \mathcal{A}_2$, where $\mathcal{A}_1,\ \mathcal{A}_2 \in \mathbb{R}^N$ are two centrally symmetric, convex, compact sets. We observe a sequence of loss vectors $\{\boldsymbol{l}_t\}_{t=1}^T$, such that in any round $t \in [T]$, $\boldsymbol{l}_t = \boldsymbol{l}_{1t} + \boldsymbol{l}_{2t}$, where $\boldsymbol{l}_{1t} \in \mathcal{L}_1$ and $\boldsymbol{l}_{2t} \in \mathcal{L}_2$. Consider two differentiable, strictly convex functions $R_1 : \Omega_1 \mapsto \mathbb{R}$, $R_2 : \Omega_2 \mapsto \mathbb{R}$, where $\Omega_1, \Omega_2 \supseteq \Delta_N$ are two convex sets. The restrictions of R_1 and R_2 to Δ_N are, respectively, α_1- and α_2-strongly convex with respect to the norms $||\cdot||_{\mathcal{A}_1}^*$ and $||\cdot||_{\mathcal{A}_2}^*$.*

Also, let parameters D_1 and D_2 be such that $R_1(\mathbf{p}) - R_1(\mathbf{p}_1) \leq D_1^2$ and $R_2(\mathbf{p}) - R_2(\mathbf{p}_1) \leq D_2^2$, for all $\mathbf{p} \in \Delta_N$; where $\mathbf{p}_1 := \operatorname{argmin}_{\mathbf{p}\in\Delta_N}(R_1(\mathbf{p}) + R_2(\mathbf{p}))$.

Then (with learning rate $\eta^* = \sqrt{\frac{(D_1^2+D_2^2)\min(\alpha_1,\alpha_2)}{T}}$, regularizer $R := R_1 + R_2$, and $\mathbf{p}_1$ as the initial prediction) the regret of the OMD algorithm is bounded as

$$\textit{Regret}_T \;\leq\; 2\sqrt{\frac{(D_1^2 + D_2^2)\,T}{\min(\alpha_1,\alpha_2)}}\,.$$

A proof of the above theorem appears in Section 4.2.

Loss Space	Regret Bound	Atomic Norm	Regularizer
d-Low Rank + ϵ-Noise	$\sqrt{2(16d+\epsilon)T}$	$\|\cdot\|_{\mathcal{A}},\ \mathcal{A}=\mathcal{A}_1+\mathcal{A}_2$, where $\mathcal{A}_1=\left\{\mathbf{x}\in\mathbb{R}^N \mid \sqrt{\mathbf{x}^\top\mathbf{H}^{-1}\mathbf{x}}\le 1\right\}$, $\mathcal{A}_2=\left\{\mathbf{x}\in\mathbb{R}^N \mid \frac{1}{\sqrt{\epsilon}}\sqrt{\mathbf{x}^T\mathbf{x}}\le 1\right\}$.	$\|\mathbf{x}\|_{\mathbf{H}}^2+\epsilon\|\mathbf{x}\|_2^2$
s-Sparse + ϵ-Noise	$2\sqrt{2(1+\epsilon)\ln(s+1)T}$	$\|\cdot\|_{\mathcal{A}},\ \mathcal{A}=\mathcal{A}_1+\mathcal{A}_2$, where $\mathcal{A}_1=\left\{\mathbf{x}\in\mathbb{R}^N \mid \frac{1}{\sqrt{2}}\|\mathbf{x}\|_p\le 1\right\}$, $\mathcal{A}_2=\left\{\mathbf{x}\in\mathbb{R}^N \mid \frac{1}{\sqrt{\epsilon}}\sqrt{\mathbf{x}^T\mathbf{x}}\le 1\right\}$.	$\|\mathbf{x}\|_q^2+\epsilon\|\mathbf{x}\|_2^2$
d-Low Rank + s-Sparse	$2\sqrt{2(16d+1)\ln(s+1)T}$	$\|\cdot\|_{\mathcal{A}},\ \mathcal{A}=\mathcal{A}_1+\mathcal{A}_2$, where $\mathcal{A}_1=\left\{\mathbf{x}\in\mathbb{R}^N \mid \sqrt{\mathbf{x}^\top\mathbf{H}^{-1}\mathbf{x}}\le 1\right\}$, $\mathcal{A}_2=\left\{\mathbf{x}\in\mathbb{R}^N \mid \frac{1}{\sqrt{2}}\|\mathbf{x}\|_p\le 1\right\}$.	$\|\mathbf{x}\|_{\mathbf{H}}^2+\|\mathbf{x}\|_q^2$

Table 2: Our Results for Additive Loss Spaces

Remark 5. *There exist loss spaces $\mathcal{L}_1$ and $\mathcal{L}_2$ such that OMD algorithm obtained via Theorem 4 provides an order-wise optimal regret bound for the additive loss space $\mathcal{L}=\mathcal{L}_1+\mathcal{L}_2$; see Appendix A for specific examples.*

The above theorem leads to the following corollary.

Corollary 6. ***(New Regret Bounds from Old)*** *Suppose $\mathcal{L}_1,\mathcal{L}_2\subseteq[0,1]^N$ are two loss spaces such that $\|\boldsymbol{l}\|_{\mathcal{A}_1}\le 1,\ \forall \boldsymbol{l}\in\mathcal{L}_1$, and $\|\boldsymbol{l}\|_{\mathcal{A}_2}\le 1,\ \forall \boldsymbol{l}\in\mathcal{L}_2$, where $\mathcal{A}_1,\mathcal{A}_2\in\mathbb{R}^N$ are two centrally symmetric, convex, compact sets. Also, suppose there exists two strictly convex, differentiable functions $R_1:\Omega_1\mapsto\mathbb{R}$ and $R_2:\Omega_2\mapsto\mathbb{R}$, ($\Omega_1,\Omega_2\supseteq\Delta_N$, convex) such that OMD with regularizer functions R_1 and R_2 gives the regret bounds of $D_1\sqrt{\frac{2T}{\alpha_1}}$ and $D_2\sqrt{\frac{2T}{\alpha_2}}$ over loss spaces $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively. Here, α_1 (α_2) is the strong convexity parameter of R_1 (R_2) over Δ_N, with respect to the atomic norm $||\cdot||_{\mathcal{A}_1}^*$ ($||\cdot||_{\mathcal{A}_2}^*$).*

In addition, let D_1 and D_2 be parameters such that, for all $\mathbf{p}\in\Delta_N$,

$$R_1(\mathbf{p})-R_1(\mathbf{p}_1')\le D_1^2 \quad \textit{with } \mathbf{p}_1'=\operatorname{argmin}_{\mathbf{q}\in\Delta_N} R_1(\mathbf{q}),$$
$$R_2(\mathbf{p})-R_2(\mathbf{p}_2')\le D_2^2 \quad \textit{with } \mathbf{p}_2'=\operatorname{argmin}_{\mathbf{q}\in\Delta_N} R_2(\mathbf{q}).$$

Then, for the additive loss space $\mathcal{L}=\mathcal{L}_1+\mathcal{L}_2$, the OMD algorithm with regularizer function $R=R_1+R_2$, initial prediction $\mathbf{p}_1=\operatorname{argmin}_{\mathbf{p}\in\Delta_N}(R_1(\mathbf{p})+R_2(\mathbf{p}))$ and learning rate $\eta^*=\sqrt{\frac{(D_1^2+D_2^2)\min(\alpha_1,\alpha_2)}{T}}$) enjoys the following regret guarantee:

$$\textit{Regret}_T\ \le\ 2\sqrt{\frac{(D_1^2+D_2^2)T}{\min(\alpha_1,\alpha_2)}}\,.$$

Note that we can prove this corollary—using Theorem 4—by simply verifying the following inequalities: $R_1(\mathbf{p})-R_1(\mathbf{p}_1)\le D_1^2$ and $R_2(\mathbf{p})-R_2(\mathbf{p}_1)\le D_2^2$, for all $\mathbf{p}\in\Delta_N$ and $\mathbf{p}_1:=\operatorname{argmin}_{\mathbf{q}\in\Delta_N}(R_1(\mathbf{q})+R_2(\mathbf{q}))$. This follows, since $R_1(\mathbf{p}_1')\le R_1(\mathbf{p}_1)$ and $R_2(\mathbf{p}_2')\le R_2(\mathbf{p}_1)$; recall that $\mathbf{p}_1':=\operatorname{argmin}_{\mathbf{q}\in\Delta_N}R_1(\mathbf{q})$ and $\mathbf{p}_2':=\operatorname{argmin}_{\mathbf{q}\in\Delta_N}R_2(\mathbf{q})$.

4.1 Applications of the main result

In this section, we will derive novel regret bounds for additive loss spaces ($\mathcal{L}=\mathcal{L}_1+\mathcal{L}_2$) wherein the individual components ($\mathcal{L}_1$ and $\mathcal{L}_2$) are the loss spaces which were considered in Section 3. These results are derived by applying Theorem 4; detailed proofs appear in the full version of the paper.

Corollary 7 (Noisy Low Rank Losses)**.** *Suppose $\mathcal{L}_1=\{\boldsymbol{l}\in[0,1]^N \mid \boldsymbol{l}=\mathbf{U}\mathbf{v}\}$ is a d-dimensional loss space ($1\le d\le\ln N$), perturbed with noisy losses $\mathcal{L}_2=\{\boldsymbol{l}\in[0,1]^N \mid \|\boldsymbol{l}\|_2^2\le\epsilon,\ \epsilon>0\}$. Then, the regret of the OMD algorithm over the loss space $\mathcal{L}=\mathcal{L}_1+\mathcal{L}_2$—with regularizer $R(\mathbf{x})=\mathbf{x}^\top\mathbf{H}\mathbf{x}+\epsilon\|\mathbf{x}\|_2^2$ and learning rate $\eta^*=\sqrt{\frac{2(16d+\epsilon)}{T}}$—is upper bounded as follows*

$$\textit{Regret}_T\ \le\ \sqrt{2(16d+\epsilon)T}\,.$$

Corollary 8 (Noisy Sparse Losses)**.** *Suppose $\mathcal{L}_1=\{\boldsymbol{l}\in[0,1]^N \mid \|\boldsymbol{l}\|_0=s\}$ is an s-sparse loss space ($s\in[N]$), perturbed with noisy losses from $\mathcal{L}_2=\{\boldsymbol{l}\in[0,1]^N \mid \|\boldsymbol{l}\|_2^2\le\epsilon,\ \epsilon>0\}$. Then, the regret of the OMD algorithm over the loss space $\mathcal{L}=\mathcal{L}_1+\mathcal{L}_2$—with regularizer $R(\mathbf{x})=\|\mathbf{x}\|_q^2+\epsilon\|\mathbf{x}\|_2^2$ and learning rate $\eta^*=\sqrt{\frac{1+\epsilon}{(2\ln(s+1)-1)T}}$—is upper bounded as follows*

$$\textit{Regret}_T\ \le\ 2\sqrt{2(1+\epsilon)\ln(s+1)T}\,.$$

Corollary 9 (Low Rank losses with Sparse noise)**.** *Suppose $\mathcal{L}_1=\{\boldsymbol{l}\in[0,1]^N \mid \boldsymbol{l}=\mathbf{U}\mathbf{v}\}$ is a d rank loss space ($1\le d\le\ln N$), perturbed with s-sparse losses $\mathcal{L}_2=\{\boldsymbol{l}\in[0,1]^N \mid \|\boldsymbol{l}\|_0=s\},\ s\in[N]$. Then, the regret of the OMD algorithm over the loss space $\mathcal{L}=\mathcal{L}_1+\mathcal{L}_2$—with regularizer $R(\mathbf{x})=\mathbf{x}^\top\mathbf{H}\mathbf{x}+\|\mathbf{x}\|_q^2$ and learning rate $\eta^*=\sqrt{\frac{16d+1}{(2\ln(s+1)-1)T}}$—is upper bounded as follows*

$$\textit{Regret}_T\ \le\ 2\sqrt{2(16d+1)\ln(s+1)T}\,.$$

4.2 Proof of Theorem 4

Before proceeding to prove the theorem, we will establish the following useful lemmas. Let $\mathcal{A}_1, \mathcal{A}_2$ be any two convex, compact, centrally symmetric subsets of $\mathbb{R}^n$ and $\mathcal{A} = \mathcal{A}_1 + \mathcal{A}_2$ (Minkowski Sum). Then, note that $\mathcal{A}$ is also convex, compact, and centrally symmetric. This follows from the fact that $\text{conv}(\mathcal{X}) + \text{conv}(\mathcal{Y}) = \text{conv}(\mathcal{X} + \mathcal{Y})$ for any $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^n$. In addition, we have

Lemma 10. $||\mathbf{x}||_{\mathcal{A}} \leq \max\{||\mathbf{x}_1||_{\mathcal{A}_1}, ||\mathbf{x}_2||_{\mathcal{A}_2}\}$, *where* $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$, $\mathbf{x}_1 \in \mathcal{A}_1$, $\mathbf{x}_2 \in \mathcal{A}_2$.

Proof. Recall the definition of atomic norm $\|\cdot\|_{\mathcal{A}}$ from Section 2.1. Suppose for any $\mathbf{x} = (\mathbf{x}_1 + \mathbf{x}_2) \in \mathbb{R}^n$, $t_1 = ||\mathbf{x}_1||_{\mathcal{A}_1}$ and $t_2 = ||\mathbf{x}_2||_{\mathcal{A}_2}$. Clearly, $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2 \in (t_1\mathcal{A}_1 + t_2\mathcal{A}_2) \subseteq t(\mathcal{A}_1 + \mathcal{A}_2)$, where $t = \max\{t_1, t_2\}$. The proof now follows directly from the definition of atomic norm, $\|\mathbf{x}\|_{\mathcal{A}}$. □

Lemma 11. $||\mathbf{x}||^*_{\mathcal{A}} = ||\mathbf{x}||^*_{\mathcal{A}_1} + ||\mathbf{x}||^*_{\mathcal{A}_2}$, *for all* $\mathbf{x} \in \mathbb{R}^n$.

Proof. Consider any $\mathbf{x} \in \mathbb{R}^n$,

$$\begin{aligned}
||\mathbf{x}||^*_{\mathcal{A}} &= \sup\{\mathbf{x}.\mathbf{z} \mid \mathbf{z} \in \mathcal{A}\} \\
&= \sup\{\mathbf{x}.(\mathbf{z}_1 + \mathbf{z}_2) \mid \mathbf{z}_1 \in \mathcal{A}_1,\ \mathbf{z}_2 \in \mathcal{A}_2\} \\
&= \sup\{\mathbf{x}.\mathbf{z}_1 \mid \mathbf{z}_1 \in \mathcal{A}_1\} + \sup\{\mathbf{x}.\mathbf{z}_2 \mid \mathbf{z}_2 \in \mathcal{A}_2\} \\
&= ||\mathbf{x}||^*_{\mathcal{A}_1} + ||\mathbf{x}||^*_{\mathcal{A}_2}\,.
\end{aligned}$$

□

Lemma 12. *Let* $\tilde{\Omega} \subseteq \mathbb{R}^n$ *be a convex set. Consider two differentiable functions* $R_1 : \mathbb{R}^n \mapsto \mathbb{R}$ *and* $R_2 : \mathbb{R}^n \mapsto \mathbb{R}$, *that are respectively* α_1 *and* α_2*-strongly convex with respect to* $||\cdot||^*_{\mathcal{A}_1}$ *and* $||\cdot||^*_{\mathcal{A}_2}$ *over* $\tilde{\Omega}$. *Then,* $R = R_1 + R_2$ *is* $\alpha = \frac{1}{2}\min(\alpha_1, \alpha_2)$*-strongly convex with respect to* $||\cdot||^*_{\mathcal{A}}$ *over* $\tilde{\Omega}$.

Proof. For any $\mathbf{x}, \mathbf{y} \in \tilde{\Omega}$,

$$\begin{aligned}
&R(\mathbf{x}) - R(\mathbf{y}) - \nabla R(\mathbf{y})(\mathbf{x} - \mathbf{y}) \\
&= R_1(\mathbf{x}) - R_1(\mathbf{y}) + R_2(\mathbf{x}) - R_2(\mathbf{y}) - \nabla R_1(\mathbf{y})^T(\mathbf{x} - \mathbf{y}) \\
&\qquad + \nabla R_2(\mathbf{y})(\mathbf{x} - \mathbf{y}) \\
&\geq \frac{\alpha_1}{2}\|\mathbf{x} - \mathbf{y}\|^{*2}_{\mathcal{A}_1} + \frac{\alpha_2}{2}\|\mathbf{x} - \mathbf{y}\|^{*2}_{\mathcal{A}_2} \\
&\geq \frac{\alpha}{2}(2\|\mathbf{x} - \mathbf{y}\|^{*2}_{\mathcal{A}_1} + 2\|\mathbf{x} - \mathbf{y}\|^{*2}_{\mathcal{A}_2}),\ \ (\alpha = \frac{1}{2}\min(\alpha_1, \alpha_2)) \\
&\geq \frac{\alpha}{2}(\|\mathbf{x} - \mathbf{y}\|^*_{\mathcal{A}_1} + \|\mathbf{x} - \mathbf{y}\|^*_{\mathcal{A}_2})^2, \\
&\qquad \big(\text{since } 2(a^2 + b^2) > (a + b)^2,\ \forall a, b \in \mathbb{R}\big) \\
&= \frac{\alpha}{2}(\|\mathbf{x} - \mathbf{y}\|^{*2}_{\mathcal{A}})\ \ (\text{via Lemma 11})\,.
\end{aligned}$$

Hence, $R = R_1 + R_2$ is $\alpha = \frac{1}{2}\min(\alpha_1, \alpha_2)$-strongly convex with respect to $||\cdot||^*_{\mathcal{A}}$ over $\tilde{\Omega}$.

□

Proof. **of Theorem 4** Consider the norm $\|\cdot\| = \|\cdot\|_{\mathcal{A}}$, and its dual norm $\|\cdot\|^* = \|\cdot\|^*_{\mathcal{A}}$. Note that:

1. Lemma 10 along with the bounds $\|\boldsymbol{l}_1\|_{\mathcal{A}_1} \leq 1$ and $\|\boldsymbol{l}_2\|_{\mathcal{A}_2} \leq 1$ imply that $\|\boldsymbol{l}\|_{\mathcal{A}} \leq 1$, for any $\boldsymbol{l} = \boldsymbol{l}_1 + \boldsymbol{l}_2 \in \mathcal{L}$. Hence, $\mathcal{L} \subseteq \mathcal{A}$.
2. For any $\mathbf{p} \in \Delta_N$, $R(\mathbf{p}) - R(\mathbf{p}_1) = (R_1(\mathbf{p}) - R_1(\mathbf{p}_1)) + (R_2(\mathbf{p}) - R_2(\mathbf{p}_1)) \leq D_1^2 + D_2^2$. Hence, $D = \sqrt{D_1^2 + D_2^2}$.
3. $R(\mathbf{x}) = R_1(\mathbf{x}) + R_2(\mathbf{x})$ is $\frac{\min\{\alpha_1, \alpha_2\}}{2}$-strongly convex with respect to $\|\cdot\|^*_{\mathcal{A}}$, $\forall \mathbf{x} \in \Delta_N$ (Lemma 12). Hence, $\alpha = \frac{\min\{\alpha_1, \alpha_2\}}{2}$.

The result now follows by applying Theorem 3. □

5 Lower Bounds

In this section we will derive lower bounds for the problem of learning with expert advice, for various structured loss spaces. Relevant definitions and missing proofs can be found in the full version of the paper. We first state the lower bound for a general loss space $\mathcal{L} \subseteq \mathbb{R}^N$ (Theorem 13) based on a lower-bound result of (Ben-David, Pál, and Shalev-Shwartz 2009) for online learning of binary hypotheses classes in terms of Littlestone's dimension.[4]

Theorem 13 (A Generic Lower Bound). *Given parameters* $V > 0$ *and* $s > 0$ *along with any online learning algorithm, there exists a sequence of* V*-dimensional loss vectors* $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_T \in \{0, \pm s\}^N$ *of sparsity* $2^V \leq N$ *(i.e., rank* $([\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_T]) = V$ *and* $\|\boldsymbol{l}_t\|_0 = 2^V$*, for all* $t \in [T]$*) such that*

$$\textit{Regret}_T \ \geq\ 2s\sqrt{\frac{VT}{8}}\,.$$

The following corollary is a direct consequence of Theorem 13.

Corollary 14. *Given parameters* $V \in [\ln N]$ *and* $s > 0$ *along with any online learning algorithm, there exists a sequence of loss vectors* $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_T \in [-s, s]^N$ *of VC-dimension* V *(i.e.,* $VC(\{\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_T\}) = V$*), such that*

$$\textit{Regret}_T \ \geq\ 2s\sqrt{\frac{VT}{8}}\,.$$

Proof. Consider the set of loss vectors $L = \{\boldsymbol{l} \in \{0, \pm s\}^N \mid \|\boldsymbol{l}\|_0 \leq 2^V\}$. From the definition of VC dimension,[5] it follows that $VC(L) = V$. Hence, Theorem 13 implies the stated claim. □

Next we instantiate Theorem 13 to derive the regret lower bounds for the structured loss spaces introduced in Section 3. In particular, we begin by stating a lower bound for sparse loss vectors.

Corollary 15. *(Lower Bound for Sparse losses) Given* $k \in [N]$ *and* $s > 0$ *along with any online learning algorithm, there exists a sequence of loss vectors* $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_T \in [-s, s]^N$ *of sparsity* $k \in N$ *(i.e.* $\|\boldsymbol{l}_t\|_0 = k$ *for all* $t \in [T]$*) such that*

$$\textit{Regret}_T \ \geq\ 2s\sqrt{\frac{\lfloor \ln k \rfloor T}{8}}\,.$$

[4] For online learning problems, Littlestone's dimension is used as a characterization of complexity of a hypothesis class, learnable in an online fashion. Further details can be found in the full version of the paper.

[5] The full version of the paper provides a definition of VC dimension along with relevant references.

Along the same lines, Theorem 13 leads to a lower bound for losses with small ℓ_p norm.

Corollary 16. *(Lower Bound for ℓ_p losses) Given $p \leq [\ln N]$ and $s > 0$ along with any online learning algorithm, there exists a sequence of loss vectors $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_T \in [-s, s]^N$ of ℓ_p norm at most s (i.e., $\|\boldsymbol{l}_t\|_p \leq s$) such that*

$$\textit{Regret}_T \;\geq\; s\sqrt{\frac{pT}{8}}\,.$$

Proof. Consider the set of all 2^p-sparse loss vectors in $[-\frac{s}{2}, \frac{s}{2}]^N$. Any such loss vector $\boldsymbol{l} \in [-\frac{s}{2}, \frac{s}{2}]^N$ has $\|\boldsymbol{l}\|_p \leq s$. The stated claim now follows by applying Theorem 13 with parameters $\frac{s}{2}$ and $V = p$. □

Corollary 17. *(Lower Bound for Noisy Losses) Given $\epsilon > 0$ and any online learning algorithm, there exists a sequence of ϵ-noisy loss vectors $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_T \in [-\epsilon, \epsilon]^N$ (i.e., $\|\boldsymbol{l}_t\|_2^2 \leq \epsilon$) such that*

$$\textit{Regret}_T \;\geq\; \sqrt{\frac{\epsilon T}{4}}\,.$$

Proof. Consider the set of all 2-sparse loss vectors in $[-\sqrt{\frac{\epsilon}{2}}, \sqrt{\frac{\epsilon}{2}}]^N$. Clearly any such loss vector $\boldsymbol{l} \in [-\sqrt{\frac{\epsilon}{2}}, \sqrt{\frac{\epsilon}{2}}]^N$ has $\|\boldsymbol{l}\|_2^2 \leq \epsilon$. Hence with parameters $s = \sqrt{\frac{\epsilon}{2}}$ and $V = 1$, the result follows directly from theorem 13. □

Remark 18. *Note that Theorem 13 (with parameter $V = d,\ s = 1$) recovers the lower bound for low rank loss spaces as established by (Hazan et al. 2016): given $1 \leq d \leq \ln N$ and any online learning algorithm, there exists a sequence of d-rank loss vectors $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_T \in [-1, 1]^N$ such that*

$$\textit{Regret}_T \;\geq\; 2\sqrt{\frac{dT}{8}}\,.$$

We next derive the regret lower bounds for few instances of additive loss spaces.

Corollary 19. *(Lower Bound for Noisy Low Rank) Given parameters $\epsilon > 0$ and $d \in [\ln N]$ along with any online learning algorithm, there exists a sequence of loss vectors $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_T \in [-(1+\epsilon), (1+\epsilon)]^N$, where $\boldsymbol{l}_t = \boldsymbol{l}_{t1} + \boldsymbol{l}_{t2}$, with $\boldsymbol{l}_{t1} \in \{\boldsymbol{l} \in [-1, 1]^N \mid \boldsymbol{l} = \mathbf{U}\mathbf{v}\}$ ($\mathbf{U} \in \mathbb{R}^{N \times d}$ is a rank d matrix), and $\|\boldsymbol{l}_{t2}\|_2^2 \leq \epsilon$, such that*

$$\textit{Regret}_T \;\geq\; 2\left(1 + \sqrt{\frac{\epsilon}{2^d}}\right)\sqrt{\frac{dT}{8}}\,.$$

Proof. Let $N = 2^d$. Consider the matrix $\mathbf{H} \in \{-1, 1\}^{N \times d}$ where 2^d rows of $\mathbf{H}$ represent 2^d vertices of the d-hypercube in $[-1, 1]^N$. Let, $\mathcal{L}_1 = \{\mathbf{H}(:, 1), \ldots \mathbf{H}(:, d)\}$, and $\mathcal{L}_2 = \left\{\boldsymbol{l} \in \left\{-\sqrt{\frac{\epsilon}{2^d}}, \sqrt{\frac{\epsilon}{2^d}}\right\}^N \mid \|l\|_2^2 = \epsilon\right\}$. Note that any loss vectors in $\mathcal{L}_2$ is 2^d-sparse. Consider $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$. The result now follows from Theorem 13, noting that with $s = \left(1 + \sqrt{\frac{\epsilon}{2^d}}\right)$ and $V = d$, the lower-bounding loss vectors assured in Theorem 13, $\boldsymbol{l}_1, \ldots, \boldsymbol{l}_T$, are contained in $\mathcal{L}$. □

Corollary 20. *(Lower Bound for Noisy Sparse) Given parameters $\epsilon > 0$ and $k \in [N]$ along with any online learning algorithm, there exists a sequence of loss vectors $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_T \in [-(1+\epsilon), (1+\epsilon)]^N$, where $\boldsymbol{l}_t = \boldsymbol{l}_{t1} + \boldsymbol{l}_{t2}$, with $\boldsymbol{l}_{t1} \in \{\boldsymbol{l} \in [-1, 1]^N \mid \|\boldsymbol{l}\|_0 \leq k\}$, and $\|\boldsymbol{l}_{t2}\|_2^2 = \epsilon$, such that*

$$\textit{Regret}_T \;\geq\; 2\left(1 + \sqrt{\frac{\epsilon}{k}}\right)\sqrt{\frac{\lfloor \ln k \rfloor T}{8}}\,.$$

Proof. Consider the loss spaces: $\mathcal{L}_1 = \{l \in \{-1, 1\}^N \mid \|l\|_0 = k\}$, and $\mathcal{L}_2 = \left\{l \in \left\{-\sqrt{\frac{\epsilon}{k}}, \sqrt{\frac{\epsilon}{k}}\right\}^N \mid \|l\|_2^2 = \epsilon\right\}$. Note that any loss vectors in $\mathcal{L}_2$ is k-sparse. Write $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$. The corollary now follows from Theorem 13, noting that—with $s = \left(1 + \sqrt{\frac{\epsilon}{k}}\right)$ and $V = \lfloor \ln k \rfloor$—the lower-bounding loss vectors assured in Theorem 13, $\boldsymbol{l}_1, \ldots, \boldsymbol{l}_T$, are contained in $\mathcal{L}$. □

6 Conclusion

In this paper, we have developed a theoretical framework for online learning with structured losses, namely the broad class of problems with additive loss spaces. The framework yields both algorithms that generalize standard online mirror descent and also novel regret upper bounds for relevant settings such as noisy + sparse, noisy + low-rank, and sparse + low-rank losses. In addition, we have derived lower bounds—i.e., fundamental limits—on regret for a variety of online learning problems with structured loss spaces. In light of these results, tightening the gap between the upper and lower bounds for structured loss spaces is a natural, open problem.

Another relevant thread of research is to study settings wherein the learner knows that the loss space is structured, but is oblivious to the exact instantiation of the loss space, e.g., the losses might be perturbations of vectors from a low-dimensional subspace, but, a priori, the learning algorithm might not know the underlying subspace.[6] Addressing structured loss spaces in bandit settings also remains an interesting direction for future work.

Appendix

A Tight Examples for Theorem 4

In this section, we present loss spaces $\mathcal{L}_1$ and $\mathcal{L}_2$ such that OMD algorithm obtained via Theorem 4 provides an order-wise optimal regret guarantee for the additive loss space $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$.

Composition of Low Ranks: Let $\mathcal{L}_1 = \{\boldsymbol{l} \in [0, 1]^N \mid \boldsymbol{l} = \mathbf{U}_1\mathbf{v}\}$ and $\mathcal{L}_2 = \{\boldsymbol{l} \in [0, 1]^N \mid \boldsymbol{l} = \mathbf{U}_2\mathbf{v}\}$ be loss spaces of rank d_1 and d_2, respectively (i.e., rank of the matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ are respectively d_1 and d_2). Here $(d_1 + d_2) \leq \ln N$. Consider the regularizer $R(\mathbf{x}) = \mathbf{x}^\top(\mathbf{H}_1 + \mathbf{H}_2)\mathbf{x}$, where $\mathbf{H}_1 = \mathbf{I}_N + \mathbf{U}_1^\top \mathbf{M}_1 \mathbf{U}_1$, and $\mathbf{H}_2 = \mathbf{I}_N + \mathbf{U}_2^\top \mathbf{M}_2 \mathbf{U}_2$, $\mathbf{M}_1$ and $\mathbf{M}_2$ being the Löwner John ellipsoid matrix for $\mathcal{L}_1$ and $\mathcal{L}_2$.

[6] The results of (Hazan et al. 2016) address the noiseless version of this problem.

That is, $R(\mathbf{x}) = R_1(\mathbf{x}) + R_2(\mathbf{x})$, where $R_1(\mathbf{x})$ and $R_2(\mathbf{x})$ are the regularizers for $\mathcal{L}_1$ and $\mathcal{L}_2$ respectively.

Theorem 4 assets that the OMD algorithm, with regularizer R, for the loss space $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$ achieves the following regret bound:

$$\text{Regret}_T \;\leq\; 4\sqrt{2(d_1 + d_2)T}\,.$$

This regret guarantee is tight, since $Rank(\mathcal{L})$ can be as high as $(d_1 + d_2)$ and, hence, we get a nearly matching lower bound by applying the result of (Hazan et al. 2016); see also Remark 18 in Section 5.

Composition of Noise Let loss spaces $\mathcal{L}_1 = \{\boldsymbol{l} \in [0,1]^N \mid \|\boldsymbol{l}\|_2^2 \leq \epsilon_1\}$ and $\mathcal{L}_2 = \{\boldsymbol{l} \in [0,1]^N \mid \|\boldsymbol{l}\|_2^2 \leq \epsilon_2\}$. Then, via an instantiation of Theorem 4, we get that the regret of the OMD algorithm over the loss space $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$, with regularizer $R(\mathbf{x}) = (\epsilon_1 + \epsilon_2)\|\mathbf{x}\|_2^2$ (and $\eta^* = \sqrt{\frac{2(\epsilon_1+\epsilon_2)}{T}}$) is upper bounded as follows:

$$\text{Regret}_T \;\leq\; \sqrt{2(\epsilon_1 + \epsilon_2)T}\,.$$

Again, modulo constants, this is the best possible regret guarantee for $\mathcal{L}$; see Corollary 17.

References

Arora, S.; Hazan, E.; and Kale, S. 2012. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing* 8(1):121–164.

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

Ben-David, S.; Pál, D.; and Shalev-Shwartz, S. 2009. Agnostic online learning. In *COLT*.

Boyd, S., and Vandenberghe, L. 2004. *Convex Optimization*. New York, NY, USA: Cambridge University Press.

Bubeck, S. 2011. Introduction to online optimization. Lecture Notes, Princeton University.

Cesa-Bianchi, N., and Lugosi, G. 2006. *Prediction, learning, and games*. Cambridge university press.

Cesa-Bianchi, N.; Mansour, Y.; and Stoltz, G. 2007. Improved second-order bounds for prediction with expert advice. *Machine Learning* 66(2):321–352.

Chandrasekaran, V.; Recht, B.; Parrilo, P. A.; and Willsky, A. S. 2012. The convex geometry of linear inverse problems. *Foundations of Computational mathematics* 12(6):805–849.

Cohen, A., and Mannor, S. 2017. Online learning with many experts. *arXiv preprint arXiv:1702.07870*.

Hazan, E., and Kale, S. 2010. Extracting certainty from uncertainty: regret bounded by variation in costs. *Machine Learning* 80(2):165–188.

Hazan, E.; Koren, T.; Livni, R.; and Mansour, Y. 2016. Online learning with low rank experts. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, 1096–1114.

Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *Computer* 42(8):30–37.

Shalev-Shwartz, S. 2012a. Online learning and online convex optimization. *Found. Trends Mach. Learn.* 4(2).

Shalev-Shwartz, S. 2012b. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning* 4(2):107–194.

Steinhardt, J., and Liang, P. 2014. Adaptivity and optimism: An improved exponentiated gradient algorithm. In *International Conference on Machine Learning*, 1593–1601.

Zinkevich, M. 2003. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*.