Expert advice problem with noisy low rank loss

Yaxiong Liu YAXIONG.LIU@INF.KYUSHU-U.AC.JP
Department of Informatics, Kyushu University/ RIKEN AIP

Xuanke Jiang JIANG.XUANKE.290@S.KYUSHU-U.AC.JP
Department of Informatics, Kyushu Univeristy

Kohei Hatano HATANO@INF.KYUSHU-U.AC.JP
Faculty of Art and Science, Kyushu University/ RIKEN AIP

Eiji Takimoto EIJI@INF.KYUSHU-U.AC.JP
Department of Informatics, Kyushu University

Editors: Vineeth N Balasubramanian and Ivor Tsang

Abstract

We consider the expert advice problem with a low rank but noisy loss sequence, where a loss vector $\boldsymbol{l}_t \in [-1,1]^N$ in each round t is of the form $\boldsymbol{l}_t = U\boldsymbol{v}_t + \boldsymbol{\epsilon}_t$ for some fixed but unknown $N \times d$ matrix U called the kernel, some d-dimensional seed vector $\boldsymbol{v}_t \in \mathbb{R}^d$, and some additional noisy term $\boldsymbol{\epsilon}_t \in \mathbb{R}^N$ whose norm is bounded by ϵ. This is a generalization of the works of Hazan et al. and Barman et al., where the former only treats noiseless loss and the latter assumes that the kernel is known in advance. In this paper, we propose an algorithm, where we re-construct the kernel under the assumptions, that the low rank loss is noised and there is no prior information about kernel. In this algorithm, we approximate the kernel by choosing a set of loss vectors with a high degree of independence from each other, and we give a regret bound of $O(d\sqrt{T} + d^{4/3}(N\epsilon)^{1/3}\sqrt{T})$. Moreover, even if in experiment, the proposed algorithm performs better than Hazan's algorithm and Hedge algorithm.

Keywords: Online expert advice, Low rank loss, Noise, Minimum volume ellipsoid enclosing

1. Introduction

The expert advice problem with low rank loss (Hazan et al., 2016) is an extension of the standard expert problem by considering a latent structure in losses. In this problem, we model the expert advice with d-rank loss as follows: the environment chooses a full rank $N \times d$ matrix U, called kernel. On each round $t \in [T]$, the algorithm picks a prediction $\boldsymbol{w}_t$ in an N-dimensional simplex over the set of N experts. Then the environment gives a d-rank loss vector $\boldsymbol{l}_t \in [-1,1]^N$, where $\boldsymbol{l}_t = U\boldsymbol{v}_t$ to the algorithm. At last the algorithm suffers the loss $\boldsymbol{w}_t \cdot \boldsymbol{l}_t$. We measure the performance of this algorithm based difference between the cumulative loss and the best expert strategy in hindsight, which is defined as regret in the following equation:

$$\text{Regret}_T = \sum_{t=1}^{T} \boldsymbol{w}_t \cdot \boldsymbol{l}_t - \min_{i \in [N]} \sum_{t=1}^{T} \boldsymbol{l}_t(i). \tag{1}$$

This low rank loss setting is popular in recommendation system, especially when experts give losses based on a latent structure (Koren et al., 2009). For instance, these experts share some common information, or their prediction methods are similar and depended on only few factors. As a consequence, the loss vectors given by experts are in fact in a relatively lower dimensional space, which implies that $d \leq N$.

Compared with the original expert advice problem, whose regret bound is $\Theta(\sqrt{T \ln N})$ with Hedge algorithm (Freund and Schapire, 1997; Cesa-Bianchi and Lugosi, 2006), the expert advice problem with d-rank loss incurs an upper bound as $O(d\sqrt{T})$ (Hazan et al., 2016), and $O(\sqrt{dT \ln T})$ (Koren and Livni, 2017), respectively. Moreover, the optimal bound is shown to be $\Theta(\sqrt{dT})$ if the kernel is known to the algorithm (Hazan et al., 2016). The work of Hazan et al. (2016) combines two parts, recovering the kernel U and predicting the $\boldsymbol{w}_t$. Hence, this algorithm not only provides a more precise upper bound, if $d \leq o(\sqrt{\ln N})$, but also re-constructs the kernel while running the algorithm. Equivalently, kernel U is explored and exploited in the learning process. Actually, in the algorithm of Hazan et al. (2016), Online mirror descent (OMD) is designed to predict $\boldsymbol{w}_t$ utilising the recovered U. Without U, OMD can give only a loose bound as $O(\sqrt{NT})$, even if the loss is d-ranked. In contrast, if the kernel is acquired in advance, OMD achieves the optimal bound $O(\sqrt{dT})$.

After this pioneering work (Hazan et al., 2016), Barman et al. (2018) considered a case that, the d-rank loss $\boldsymbol{l}_t$ is corrupted by a L_2-norm bounded $\boldsymbol{\epsilon}_t$ noise, i.e., the environment gives the loss vector $\tilde{\boldsymbol{l}}_t = \boldsymbol{l}_t + \boldsymbol{\epsilon}_t = U\boldsymbol{v}_t + \boldsymbol{\epsilon}_t$, which is near to the d-dimensional space. In their work, Barman et al. (2018) gave a regret bound $O\left(\sqrt{(d+\epsilon)T}\right)$ and a lower bound as $\Omega\left(\sqrt{T(d+\frac{\epsilon}{2^d})}\right)$, where $\|\boldsymbol{\epsilon}_t\|_2^2 \leq \epsilon$. However, there is a strong assumption in their work that the algorithm needs know both U and ϵ.

In this paper, we release the strong assumption in Barman et al. (2018), and assume that the kernel is unknown to the algorithm, while the low rank loss is corrupted by $\boldsymbol{\epsilon}_t$. This problem setting is more realistic in two points: One is the loss vectors are not always strictly well-structured. Although the experts share some common prediction methods, there is still some slight disturbance, which damages the low rank structure loss. The other is the specific structure is unknown to the algorithm before it receives the loss vectors. It is natural that the recommendation system needs to confront new users who are recently registered without any prior information. Thus, the system is required to explore the characteristics of users while recommending their goods.

From the above works, we can see how the low rank structure plays the important role in the algorithms. Either, the kernel is known to the algorithm, then the algorithm can overcome the noised loss vectors easily; or, the loss is precisely low ranked, then the algorithm can extract the kernel from the received losses, even if the kernel is unknown. However, if we apply the algorithms of Hazan et al. (2016), and Koren and Livni (2017) in our problem setting directly, it is equivalent to run OMD on expert advice directly, and the regret bound is $O(\sqrt{NT})$, since the low rank loss no longer exists. Hence, one of the essential problems is how to deal with the "noisy low rank loss" and recover the underlying kernel U. Although recovering a low rank matrix is not a novel concept for machine learning (Goldberg et al., 2010; Foygel and Srebro, 2011; Balcan and Zhang, 2016; Zhang et al., 2018), to the best of the authors knowledge, there is no such research for online expert advice problem with noise attached loss vectors. The most tricky obstacle is

that, it is impossible to re-construct the kernel U exactly with the corrupted loss vectors. Thus, in our proposed algorithm, we attempt to approximate this kernel from the received loss vectors.

In our algorithm, we assume that there is no prior information about the kernel. Thus, we select at most d highly-independent loss vectors (see details in latter section) to construct a pseudo-kernel $\widetilde{U}$ in our learning process. Unlike the selection criteria in work (Balcan and Zhang, 2016), which recovers the kernel with a randomised algorithm stochastically, in our paper, each selected vector is supposed to be not only "long" enough, but also far enough to the current sub-space, spanned by all previously selected vectors. Therefore, our algorithm can approximate the kernel of the low rank loss deterministically and obtains a worst case guarantee. Instead of the unknown U in (Hazan et al., 2016), we next utilise the pseudo-kernel, which is spanned by the selected vectors, in OMD to update the predictions. Our algorithm obtains an upper bound with respect to $\tilde{\boldsymbol{l}}_t$ as $O(\sqrt{T}(d + d^{4/3}(N\epsilon)^{1/3}))$. In practice, we can consider our algorithm and the Hedge algorithm as two meta-experts and run another Hedge algorithm on top of them, which performs nearly as well as our algorithm and the Hedge algorithm with additional negligeable ($O(\sqrt{T})$) regret.

The following table shows the relationship between our work and previous work.

Table 1: Comparison with noisy and non-noisy low rank loss

	known kernel	unknown kernel
without noise	$\Theta(\sqrt{dT})$ (Hazan et al., 2016)	$O(d\sqrt{T})$ (Hazan et al., 2016) $O(\sqrt{dT\ln T})$ (Koren and Livni, 2017)
ϵ-noise case	$O(\sqrt{(d+\epsilon)T})$ $\Omega\left(\sqrt{T(d+\frac{\epsilon}{2^d})}\right)$ (Barman et al., 2018)	$O((1+\epsilon)\sqrt{T}(d + d^{4/3}(N\epsilon)^{1/3}))$ (Our work)

This paper is composed as follows. In the second section, we define our expert advice with noisy d-rank loss formally and some necessary notations. In the third section, the proposed algorithm is given and we show the upper bound. In section 4, we compare the proposed algorithm and previous work under synthetic environments[1].

2. Preliminaries

We denote the set $\{1, \cdots, T\}$ by $[T]$. We denote the N-dimensional vector with all 1 elements by $\mathbb{I}_N$ and $N \times N$ identity matrix by I_N. We define a norm $\|\boldsymbol{x}\|_H$ with respect to a positive definite matrix H, as $\|\boldsymbol{x}\|_H = \sqrt{\boldsymbol{x}^T H \boldsymbol{x}}$, for a vector $\boldsymbol{x}$, and the dual norm of $\|\cdot\|_H$ is defined as $\|\boldsymbol{x}\|_H^* = \sqrt{\boldsymbol{x}^T H^{-1} \boldsymbol{x}}$. We define the angle between two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ as $\theta(\boldsymbol{u}, \boldsymbol{v}) = \arccos \frac{\langle \boldsymbol{u}, \boldsymbol{v} \rangle}{\|\boldsymbol{u}\|_2 \|\boldsymbol{v}\|_2}$. If $\boldsymbol{V}$ is a subspace then $\theta(\boldsymbol{u}, \boldsymbol{V}) = \min_{\boldsymbol{v} \in \boldsymbol{V} \setminus \{0\}} \theta(\boldsymbol{u}, \boldsymbol{v})$, and $\theta(\boldsymbol{U}, \boldsymbol{V}) = \max_{\boldsymbol{u} \in U} \theta(\boldsymbol{u}, \boldsymbol{V})$. At last, we denote the orthogonal projection from a vector $\boldsymbol{l}$ to a subspace $\boldsymbol{U}$ as $\mathcal{P}_{\boldsymbol{U}} \boldsymbol{l}$. Moreover, the Euclidean distance from a vector $\boldsymbol{l}$ to a sub-space $\boldsymbol{U}$ is defined as $\|\boldsymbol{l} - \mathcal{P}_{\boldsymbol{U}} \boldsymbol{l}\|_2 = \|\boldsymbol{l}\|_2 \sin \theta(\boldsymbol{l}, \boldsymbol{U}) = \|\boldsymbol{l}\|_2 \sin \theta(\boldsymbol{l}, \mathcal{P}_{\boldsymbol{U}} \boldsymbol{l})$. A linear space spanned by all columns

1. The code is available at https://github.com/2015211217/LowRankStructureC-.git

from a matrix $V \in \mathbb{R}^{N\times k}$ can be represented as $\mathrm{span}(V)$. In the following part, we simplify $\mathrm{span}(V)$ as $\boldsymbol{V}$, when it leads no ambiguity.

2.1. Problem Setting

Firstly we define the d-rank loss: For a sequence of $\boldsymbol{l}_t \in [-1,1]^N$, for $t \in [T]$ we define $L_T \in \mathbb{R}^{N\times T}$ as follows:

$$L_T = [\boldsymbol{l}_1, \boldsymbol{l}_2, \cdots, \boldsymbol{l}_T], \tag{2}$$

where each loss vector $\boldsymbol{l}_t$ is a column in L_T. We define that the sequence $\{\boldsymbol{l}_1, \cdots, \boldsymbol{l}_T\}$ is d-rank loss if and only if $\mathrm{rank}(L_T) = d$. Equivalently, for the sequence $\{\boldsymbol{l}_1, \cdots, \boldsymbol{l}_t\}$, there exists a kernel $U \in \mathbb{R}^{N\times d}$ such that $\boldsymbol{l}_t = U\boldsymbol{v}_t$, where $\mathrm{rank}(U) = d$. If $d \ll N$, we can call $\boldsymbol{l}_t$ as low rank loss vector.

From d-rank loss $\boldsymbol{l}_t$, we define the noisy d-rank loss $\tilde{\boldsymbol{l}}_t : \tilde{\boldsymbol{l}}_t = \boldsymbol{l}_t + \boldsymbol{\epsilon}_t$, where $\|\boldsymbol{\epsilon}_t\|_2 \le \epsilon, \forall t \in [T]$. We call $\boldsymbol{\epsilon}_t$ as the noise vector.

In this paper we consider the following learning problem of online expert advice with noisy d-rank loss. On round $t = 1, \cdots, T$, an algorithm gives a prediction $\boldsymbol{w}_t \in \Delta(N)$; Then an environment gives a loss vector $\tilde{\boldsymbol{l}}_t \in [-1-\epsilon, 1+\epsilon]^N$, note that $\tilde{\boldsymbol{l}}_t = \boldsymbol{l}_t + \boldsymbol{\epsilon}_t$. It implies that there exists an underlying $\boldsymbol{l}_t \in [-1,1]^N$ as d-rank loss and $\boldsymbol{\epsilon}_t$ as ϵ-noise. However, $\boldsymbol{l}_t, U$ and $\boldsymbol{\epsilon}_t$ are unknown to the algorithm. At last the algorithm suffers the loss as $\boldsymbol{w}_t \cdot \tilde{\boldsymbol{l}}_t$. Next we define the regret as follows:

$$\mathrm{Regret}_T = \sum_{t=1}^{T} \boldsymbol{w}_t \cdot \tilde{\boldsymbol{l}}_t - \min_{i\in[N]} \sum_{t=1}^{T} \tilde{\boldsymbol{l}}_t(i). \tag{3}$$

2.2. Online mirror descent

Online mirror descent (OMD) is a basic algorithm utilised in both Hazan et al. (2016) and Barman et al. (2018), as well as in this paper. We give OMD with the time-varying matrix norms as follows:

Algorithm 1 Online Mirror descent

Initialization: $H_0, \cdots, H_{T-1} \succ 0$, $\{\eta_t\}_{t=1}^T$, and $\boldsymbol{w}_1 \in \Delta(N)$.
  for $t = 1, \cdots, T$ **do**
    Receive $\boldsymbol{l}_t$,
    Suffer cost $\boldsymbol{w}_t \cdot \boldsymbol{l}_t$,
    Update $\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w}\in\Delta(N)} \boldsymbol{l}_t \cdot \boldsymbol{w} + \eta_t^{-1}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2_{H_{t-1}}$.
  end for

Theorem 1 (Orabona et al. (2015)) *The T-round regret of OMD is bounded as follow:*

$$\sum_{t=1}^{T} \boldsymbol{w}_t \cdot \boldsymbol{l}_t - \sum_{t=1}^{T} \boldsymbol{l}_t \cdot \boldsymbol{w}^* \le \frac{1}{\eta_T}\|\boldsymbol{w}_1 - \boldsymbol{w}^*\|^2_{H_T} + \frac{1}{2}\sum_{t=1}^{T}(\eta_t\|\boldsymbol{l}_t\|^*_{H_t})^2 + \sum_{t=1}^{T}\eta_t^{-1}(\|\boldsymbol{w}_t\|^2_{H_{t-1}} - \|\boldsymbol{w}_t\|^2_{H_t}).$$

2.3. Ellipsoid approximation of convex bodies

Given a positive semi-definite matrix $M \succeq 0$, we define the ellipsoid with respect to M as follows:

$$\mathcal{E}(M) = \{\boldsymbol{x} : \boldsymbol{x}^T M^\dagger \boldsymbol{x} \leq 1\}, \tag{4}$$

where $M^\dagger$ is the Moore-Penrose pseudo-inverse matrix of M. In the following theorem we give a result for minimum volume enclosing ellipsoid (MVEE) to a central symmetric convex body.

Theorem 2 (John's Theorem (Ball et al., 1997)) *Let K be a convex body in $\mathbb{R}^d$ that is symmetric around zero. Let $\mathcal{E}$ be an ellipsoid with minimum volume enclosing K. Then:*

$$\frac{1}{\sqrt{d}}\mathcal{E} \subseteq K \subseteq \mathcal{E}. \tag{5}$$

Moreover for a polytope defined as $P_A = \{\boldsymbol{x} : \|A\boldsymbol{x}\|_\infty \leq 1, \boldsymbol{x} \in \mathbb{R}^d\}$, corresponding to a given matrix $A \in \mathbb{R}^{N\times d}$, we have the following theorem.

Theorem 3 (Grötschel et al. (2012)) *There exists a poly-time procedure* MVEE(A) *that receives as input a matrix $A \in \mathbb{R}^{N\times d}$ and returns a matrix M such that*

$$\frac{1}{\sqrt{2d}}\mathcal{E}(M) \subseteq P_A \subseteq \mathcal{E}(M).$$

3. Algorithm for no prior information about kernel

In this section, we consider the case that there is no prior information about the kernel in the online expert advice with noisy d-rank loss problem. Under this assumption, we construct a pseudo kernel, $\widetilde{U}$, in our learning process, when the d-rank structure is corrupted by $\boldsymbol{\epsilon}_t$. Note that in this algorithm, we construct this pseudo kernel cumulatively. We denote it as $\widetilde{U}^k \in \mathbb{R}^{N\times k}$ for convenience.

Concisely, our algorithm will select some loss vectors to construct the pseudo kernel and run OMD with respect to from $\tilde{U}^k$ processed matrix H^k. The criteria to select $\tilde{\boldsymbol{l}}_t$ is twofold: firstly, the L_2-norm of $\tilde{\boldsymbol{l}}_t$ is supposed to be large enough; then, the Euclidean distance from $\tilde{\boldsymbol{l}}_t$ to the current pseudo kernel spanned space $\text{span}(\widetilde{U}^k)$ should be far enough. Thus, for some $k \in [d]$ and corresponding parameters s_k, γ_k, the basic idea of our algorithm is as follows:

1. $\|\tilde{\boldsymbol{l}}_t\|_2 \leq 2s_k \rightarrow$ Do OMD with respect to matrix H^k.
2. $\|\tilde{\boldsymbol{l}}_t\|_2 \geq 2s_k$ and $\|\tilde{\boldsymbol{l}}_t - \mathcal{P}_{\text{span}(\widetilde{U}^k)}\tilde{\boldsymbol{l}}_t\|_2 \leq 2\epsilon + \gamma_k(\|\tilde{\boldsymbol{l}}_t\|_2 + \epsilon) \rightarrow$ Do OMD with respect to H^k.
3. $\|\tilde{\boldsymbol{l}}_t\|_2 \geq 2s_k$ and $\|\tilde{\boldsymbol{l}}_t - \mathcal{P}_{\text{span}(\widetilde{U}^k)}\tilde{\boldsymbol{l}}_t\|_2 \geq 2\epsilon + \gamma_k(\|\tilde{\boldsymbol{l}}_t\|_2 + \epsilon) \rightarrow$ Adding $\tilde{\boldsymbol{l}}_t$ as a column to $\widetilde{U}^k$, and reset k as $k+1$, then do OMD with respect to H^k.

Remark 4 *If we have that $\|\tilde{\boldsymbol{l}}_t\| \geq 2s_k > 2\epsilon$, then we obtain the following result: Since $\|\tilde{\boldsymbol{l}}_t - \boldsymbol{l}_t\|_2 \leq \epsilon$, $\theta(\tilde{\boldsymbol{l}}_t, \boldsymbol{l}_t) \leq 2\sin\theta(\tilde{\boldsymbol{l}}_t, \boldsymbol{l}_t) \leq \frac{\epsilon}{s_k}$*

According to this criteria, we can observe that we need to guarantee that $K = \max k \leq d$, to ensure that our algorithm effective, since the algorithm from (Hazan et al., 2016) updates N-times for low rank noisy loss vector. It implies that the previous algorithm in (Hazan et al., 2016) is OMD and obtains $O(\sqrt{(N+\epsilon)T})$ actually.

We confirm s_k, γ_k according to the following proposition.

Proposition 5 *Define a sequence s_i and non-decreasing sequence γ_i such that $\gamma_i \in (0, \frac{\pi}{2})$ and*

$$\frac{\gamma_i}{\gamma_{i-1} - (i-1)\frac{\beta\epsilon}{s_{i-1}\gamma_{i-1}}} \cdot \frac{\pi}{2\beta} + \frac{s_i\gamma_i}{s_i\gamma_{i-1}}(i-1) \leq i \quad \forall i \in \{2, \cdots, k\}, \tag{6}$$

and

$$\gamma_i - \frac{\beta\epsilon}{s_i\gamma_i} i > 0 \quad \forall i \in \{1, \cdots, k\}. \tag{7}$$

Given any space $\boldsymbol{W} \subseteq \mathbb{R}^N$: Let $\boldsymbol{U}^k = \mathrm{span}\{\boldsymbol{W}, \boldsymbol{l}_1, \cdots, \boldsymbol{l}_k\}$ and $\widetilde{\boldsymbol{U}}^k = \mathrm{span}\{\boldsymbol{W}, \tilde{\boldsymbol{l}}_1, \cdots, \tilde{\boldsymbol{l}}_k\}$ be two subspace such that $\theta(\boldsymbol{l}_i, \tilde{\boldsymbol{l}}_i) \leq \epsilon/s_i$ for all $i \in \{2, \cdots, k\}$.

If $\theta(\tilde{\boldsymbol{l}}_i, \widetilde{\boldsymbol{U}}^{i-1}) > \gamma_i$, and $\beta \geq \frac{\pi}{2}$ then we have that

$$\theta(\boldsymbol{U}^k, \widetilde{\boldsymbol{U}}^k) < \beta k \frac{\epsilon}{s_k\gamma_k}. \tag{8}$$

Proof For $i = 1$, due to Lemma 10, by setting that $\theta(l_1, \emptyset) = \frac{\pi}{2}$, it is trivial that

$$\theta(\boldsymbol{U}^1, \widetilde{\boldsymbol{U}}^1) \leq \frac{\pi\epsilon}{2s_1\gamma_1} \leq \beta\frac{\epsilon}{s_1\gamma_1}. \tag{9}$$

Here $\theta(\tilde{\boldsymbol{l}}_1, \boldsymbol{l}_1) \leq \frac{\epsilon}{s_1}$, and $\theta(\tilde{\boldsymbol{l}}_1, \widetilde{\boldsymbol{U}}^0) = \theta(\tilde{\boldsymbol{l}}_1, \boldsymbol{W}) \geq \gamma_1$. If $\boldsymbol{W} = \emptyset$ then we have that

$$\theta(\boldsymbol{U}^1, \widetilde{\boldsymbol{U}}^1) = \theta(\boldsymbol{l}_1, \tilde{\boldsymbol{l}}_1) \leq \frac{\epsilon}{s_1} \leq \frac{\pi}{2} \times \frac{\epsilon}{\gamma_1 s_1},$$

if $\gamma_1 \leq \frac{\pi}{2}$.

Then if it holds for $i-1$: Let us involve $\boldsymbol{U}_0^i = \mathrm{span}\{\boldsymbol{U}^{i-1}, \tilde{\boldsymbol{l}}_i\}$, note that

$$\theta(\boldsymbol{U}_0^i, \widetilde{\boldsymbol{U}}^i) = \theta(\mathrm{span}\{\boldsymbol{U}^{i-1}, \tilde{\boldsymbol{l}}_i\}, \mathrm{span}\{\widetilde{\boldsymbol{U}}^{i-1}, \tilde{\boldsymbol{l}}_i\}) \tag{10}$$

Since the induction hypothesis so we have that

$$\theta(\mathrm{span}\{\boldsymbol{W}, \boldsymbol{l}_1, \cdots, \boldsymbol{l}_{i-1}\}, \mathrm{span}\{\boldsymbol{W}, \tilde{\boldsymbol{l}}_1 \cdots, \tilde{\boldsymbol{l}}_{i-1}\}) \leq (i-1)\frac{\beta\epsilon}{s_{i-1}\gamma_{i-1}}. \tag{11}$$

then we have that

$$\begin{aligned}\theta(\boldsymbol{U}^k, \widetilde{\boldsymbol{U}}^k) &\leq \theta(\boldsymbol{U}^k, \boldsymbol{U}_0^k) + \theta(\boldsymbol{U}_0^k, \widetilde{\boldsymbol{U}}^k) \\ &\leq \frac{\pi}{2}\frac{\theta(\tilde{\boldsymbol{l}}_k, \boldsymbol{l}_k)}{\theta(\tilde{\boldsymbol{l}}_k, \boldsymbol{U}^{k-1})} + \theta(\boldsymbol{U}_0^i, \widetilde{\boldsymbol{U}}^i)\end{aligned}$$

The first inequality is from Remark 13. The second inequality is due to Lemma 10.

Now due to the Lemma 12 we have that

$$\theta(\boldsymbol{U}^k, \widetilde{\boldsymbol{U}}^k) \leq \frac{\pi}{2} \frac{\theta(\boldsymbol{l}_i, \tilde{\boldsymbol{l}}_i)}{\theta(\tilde{\boldsymbol{l}}_i, \widetilde{\boldsymbol{U}}^{i-1}) - \theta(\boldsymbol{U}^{i-1}, \widetilde{\boldsymbol{U}}^{i-1})} + \theta(\boldsymbol{U}_0^i, \widetilde{\boldsymbol{U}}^i) \tag{12}$$

Since the fact that $\theta(\tilde{\boldsymbol{l}}_k, \widetilde{\boldsymbol{U}}^{k-1}) > \gamma_k$, and the induction hypothesis we obtain that

$$\begin{aligned} \theta(\boldsymbol{U}^i, \widetilde{\boldsymbol{U}}^i) &< \frac{\pi}{2} \frac{\frac{\epsilon}{s_i}}{\gamma_{i-1} - \beta(i-1)\frac{\epsilon}{s_{i-1}\gamma_{i-1}}} + (i-1)\frac{\beta\epsilon}{s_{i-1}\gamma_{i-1}} \\ &= \frac{\epsilon\beta}{s_i\gamma_i} \left(\frac{\gamma_i}{\gamma_{i-1} - (i-1)\frac{\beta\epsilon}{s_{i-1}\gamma_{i-1}}} \times \frac{\pi}{2\beta} + (i-1)\frac{s_i\gamma_i}{s_{i-1}\gamma_{i-1}} \right) \end{aligned} \tag{13}$$

Since Equation 6 we have that

$$\frac{\epsilon\beta}{s_i\gamma_i} \left(\frac{\gamma_i}{\gamma_{i-1} - (i-1)\frac{\beta\epsilon}{s_{i-1}\gamma_{i-1}}} \times \frac{\pi}{2\beta} + (i-1)\frac{s_i\gamma_i}{s_{i-1}\gamma_{i-1}} \right) \leq \frac{\beta\epsilon}{s_i\gamma_i} i \tag{14}$$

Thus we have our conclusion. ■

The description of the algorithm is shown in Algorithm 2.

Algorithm 2 Mirror descent with l_2 noise

Initialization: $\boldsymbol{w}_1 = \frac{1}{N}\mathbb{I}_N$, $\tau = 0$, $k = 0$, $\widetilde{U} = \{\}$, a sequence of $\{s_k\}$ and a non decreasing sequence $\{\gamma_k\}$ for all $\gamma_k \in (0, \frac{\pi}{2})$ such that γ_k and s_k satisfies the conditions in Equation (6) and (7). Setting $m_k = \max\{2s_k, 6\epsilon + \gamma_k\sqrt{N}\}$.
for $t = 1, \cdots, T$ **do**
  Observe $\tilde{\boldsymbol{l}}_t$, suffer loss $\boldsymbol{w}_t \cdot \tilde{\boldsymbol{l}}_t$.
  if $\|\tilde{\boldsymbol{l}}_t\|_2 \geq 2s_k$ **then**
    if $\|\tilde{\boldsymbol{l}}_t - \mathcal{P}_{\widetilde{\boldsymbol{U}}^k}\tilde{\boldsymbol{l}}_t\|_2 \geq 2\epsilon + \gamma_k(\|\tilde{\boldsymbol{l}}_t\|_2 + \epsilon)$ **then**
      Add $\tilde{\boldsymbol{l}}_t$ as a new column of $\widetilde{U}^k$, reset $\tau = 0$ and set $k \leftarrow k + 1$.
      Compute $M = \text{MVEE}(\widetilde{U}^k)$ and $H^k = I_N + \widetilde{U}^k M (\widetilde{U}^k)^T$.
    end if
  end if
  let $\tau \leftarrow \tau + 1$ and $\eta_t = \sqrt{8k/(1+m_k)^2\tau}$ and set:

$$\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w} \in \Delta(N)} \tilde{\boldsymbol{l}}_t \cdot \boldsymbol{w} + \eta_t^{-1} \|\boldsymbol{w} - \boldsymbol{w}_t\|_{H^k}^2.$$

end for

According to our proposed algorithm, we see that our updating rule divides the T rounds into K epoches, where the final size of pseudo kernel, $K \leq d$ (will be show in the following Theorem). In each epoch, OMD procees with respect to H^k, a fixed matrix. Before we give the regret bound, we need some useful lemmata.

Lemma 6 *Let* $\|\tilde{\boldsymbol{l}}_t - \boldsymbol{l}_t\|_2 \leq \epsilon$, *if* $\theta(\boldsymbol{l}_t, \mathrm{span}(\widetilde{U}^k)) \leq \gamma_k$ *then we have*

$$\|\tilde{\boldsymbol{l}}_t - \mathcal{P}_{\widetilde{\boldsymbol{U}}^k}\tilde{\boldsymbol{l}}_t\|_2 \leq 2\epsilon + \gamma_k(\|\tilde{\boldsymbol{l}}_t\|_2 + \epsilon). \tag{15}$$

Proof Firstly we obtain that

$$\|\tilde{\boldsymbol{l}}_t - \mathcal{P}_{\widetilde{\boldsymbol{U}}^k}\tilde{\boldsymbol{l}}_t\|_2 \leq \|\tilde{\boldsymbol{l}}_t - \boldsymbol{l}_t\| + \|\boldsymbol{l}_t - \mathcal{P}_{\widetilde{\boldsymbol{U}}^k}\boldsymbol{l}_t\| + \|\mathcal{P}_{\widetilde{\boldsymbol{U}}^k}\boldsymbol{l}_t - \mathcal{P}_{\widetilde{\boldsymbol{U}}^k}\tilde{\boldsymbol{l}}_t\|.$$

Since the fact that $\|\tilde{\boldsymbol{l}}_t - \boldsymbol{l}_t\| \leq \epsilon$, and $\|\boldsymbol{l}_t - \mathcal{P}_{\widetilde{\boldsymbol{U}}^k}\boldsymbol{l}_t\| \leq \gamma_k\|\boldsymbol{l}_t\| \leq \gamma_k(\|\tilde{\boldsymbol{l}}_t\| + \epsilon)$, we have our conclusion. ■

Lemma 6 states the criteria of distance between $\tilde{\boldsymbol{l}}_t$ and current space $\widetilde{\boldsymbol{U}}^k$.

Lemma 7 *Given* $\widetilde{U}^k \in \mathbb{R}^{N\times k}$, *and* $\mathrm{span}(\widetilde{U}^k) \subset \mathbb{R}^N$, *for any* $\tilde{\boldsymbol{l}}_t$ *such that* $\|\tilde{\boldsymbol{l}}_t\|_2 \leq 2s_k$ *or* $\|\tilde{\boldsymbol{l}}_t - \mathcal{P}_{\widetilde{\boldsymbol{U}}^k}\tilde{\boldsymbol{l}}_t\|_2 \leq 2\epsilon + \gamma_k(\|\tilde{\boldsymbol{l}}_t\|_2 + \epsilon)$, *there exists a unique* $\boldsymbol{v}_t \in \mathbb{R}^k$, *and* $\boldsymbol{e}_t$ *such that* $\tilde{\boldsymbol{l}}_t = \widetilde{U}^k\boldsymbol{v}_t + \boldsymbol{e}_t$, *where* $\theta(\boldsymbol{e}_t, \widetilde{\boldsymbol{U}}^k) = \frac{\pi}{2}$. *Therefore we have that*

$$\|\boldsymbol{e}_t\|_2 \leq \max\{2s_k, 2\epsilon + \gamma_k(\|\tilde{\boldsymbol{l}}_t\|_2 + \epsilon)\}. \tag{16}$$

In particular, we set that $\forall t \in [T]$

$$m_k = \max\{2s_k, 6\epsilon + \gamma_k\sqrt{N}\} \geq \max\{2s_k, 2\epsilon + \gamma_k(\|\tilde{\boldsymbol{l}}_t\|_2 + \epsilon)\}. \tag{17}$$

Proof Since the fact that $\|\boldsymbol{e}_t\|_2 \leq \|\tilde{\boldsymbol{l}}_t\|_2$ and the updating rule, we have that

$$\|\boldsymbol{e}_t\|_2 \leq \max\{2s_k, 2\epsilon + \gamma_k(\|\tilde{\boldsymbol{l}}_t\|_2 + \epsilon)\}. \tag{18}$$

Due to our setting that $\gamma_k \leq \frac{\pi}{2} \leq 2$, and $\|\tilde{\boldsymbol{l}}_t\|_2 \leq \sqrt{N}+\epsilon$, thus we obtain Equation (17). ■

Theorem 8 *Running Algorithm 2 on* T*-rounds, denoting that* K *is the final size of pseudo kernel, and* $K \leq d$, *we have*

$$\mathrm{Regret}_T \leq O\left((1+\epsilon)\sqrt{T}\left(K + \sqrt{\sum_{k=1}^{K} k m_k^2}\right)\right). \tag{19}$$

Proof Firstly let us prove that $K \leq d$. Compared with $\widetilde{U}^k$ with k loss vectors of $\tilde{\boldsymbol{l}}_t$, we define a matrix U^k with respect to $\boldsymbol{l}_t$ corresponding to $\tilde{\boldsymbol{l}}_t$.

Due to our updating rule in algorithm, we have that

$$\theta(\tilde{\boldsymbol{l}}_t, \mathcal{P}_{\widetilde{\boldsymbol{U}}^k}\tilde{\boldsymbol{l}}_t) \geq \gamma_k,$$

since if $\|\tilde{\boldsymbol{l}}_t - \mathcal{P}_{\widetilde{\boldsymbol{U}}^k}\tilde{\boldsymbol{l}}_t\|_2 \geq 2\epsilon + \gamma_k(\|\tilde{\boldsymbol{l}}_t\|_2 + \epsilon)$, then we have that

$$\theta(\tilde{\boldsymbol{l}}_t, \mathcal{P}_{\widetilde{\boldsymbol{U}}^k}\tilde{\boldsymbol{l}}_t) \geq \sin\theta(\tilde{\boldsymbol{l}}_t, \mathcal{P}_{\widetilde{\boldsymbol{U}}^k}\tilde{\boldsymbol{l}}_t) \geq \frac{2\epsilon + \gamma_k(\|\tilde{\boldsymbol{l}}_t\|_2 + \epsilon)}{\|\tilde{\boldsymbol{l}}_t\|_2} \geq \gamma_k.$$

Therefore by Proposition 5 we have that $\theta(\text{span}(U^k), \text{span}(\widetilde{U}^k)) \leq \frac{\beta k \epsilon}{\gamma_k s_k}$. So we have that

$$\theta(\boldsymbol{l}_t, \text{span}(U^k)) \geq \theta(\boldsymbol{l}_t, \text{span}(\widetilde{U}^k)) - \theta(\text{span}(U^k), \text{span}(\widetilde{U}^k)) \geq \gamma_k - \frac{\beta\epsilon}{s_k\gamma_k}k > 0.$$

The first inequality is from Lemma 12. Since Lemma 6 we have that $\theta(\boldsymbol{l}_t, \text{span}(\widetilde{U}^k)) \geq \gamma_k$. It implies that if $\|\tilde{\boldsymbol{l}}_t - \mathcal{P}_{\widetilde{\boldsymbol{U}}^k}\tilde{\boldsymbol{l}}_t\|_2 \geq 2\epsilon + \gamma_k(\|\tilde{\boldsymbol{l}}_2\|_2 + \epsilon)$, then we have $\boldsymbol{l}_t \notin \text{span}(U^k)$. So we have that $\text{rank}(\widetilde{U}^K) \leq \text{rank}(L_T) = d$.

Now let us prove the second part. Due to above result we can divide the total learning rounds T into K episodes. $T = \sum_{k=1}^{K} |T_k|$ and $T_k \cup T_{k'} = \emptyset$, if $k \neq k'$. For a given k, if $t \in T_k$, we obtain that $\|\tilde{\boldsymbol{l}}_t - \mathcal{P}_{\widetilde{\boldsymbol{U}}^k}\tilde{\boldsymbol{l}}_t\|_2 \leq 2\epsilon + \gamma_k(\|\tilde{\boldsymbol{l}}_2\|_2 + \epsilon)$ or $\|\tilde{\boldsymbol{l}}_t\|_2 \leq 2s_k$. Thus we need give upper bound of $(\|\tilde{\boldsymbol{l}}_t\|_{H^k}^*)^2$, and $\|\boldsymbol{w}_1 - \boldsymbol{w}^*\|_{H^k}$.

Given a k-dimensional polytope as

$$P = \{\boldsymbol{v} \in \mathbb{R}^k : \|\widetilde{U}^k\boldsymbol{v}\|_\infty \leq 1 + \epsilon\},$$

we are able to have

$$\mathcal{E}(\frac{1}{2k}M) \subseteq P \subseteq \mathcal{E}(M), \tag{20}$$

due to the MVEE procedure, and Theorem 3 in previous section. For each $\tilde{\boldsymbol{l}}_t$ we give a unique orthogonal decomposition of $\tilde{\boldsymbol{l}}_t$ upon $\text{span}(\widetilde{U}^k)$. Thus we obtain $\tilde{\boldsymbol{l}}_t = \widetilde{U}^k\boldsymbol{v}_t + \boldsymbol{e}_t$. Since $\tilde{\boldsymbol{l}}_t \in [0,1]^N$, and we have that $\boldsymbol{v}_t \in P$, and $\theta(\boldsymbol{e}_t, \widetilde{\boldsymbol{U}}^k) = \frac{\pi}{2}$. Therefore we have following inequality:

$$(\|\tilde{\boldsymbol{l}}_t\|_{H^k}^*)^2 = (\|\widetilde{U}^k\boldsymbol{v}_t + \boldsymbol{e}_t\|_{H^k}^*)^2 \leq (\|\widetilde{U}^k\boldsymbol{v}_t\|_{H^k}^* + \|\boldsymbol{e}_t\|_{H^k}^*)^2 \leq 2(\|\widetilde{U}^k\boldsymbol{v}_t\|_{H^k}^*)^2 + 2(\|\boldsymbol{e}_t\|_{H^k}^*)^2.$$

Firstly we give upper bound of $(\|\widetilde{U}^k\boldsymbol{v}_t\|_{H^k}^*)^2$.

$$\begin{aligned}(\|\widetilde{U}^k\boldsymbol{v}_t\|_{H^k}^*)^2 &= (\widetilde{U}^k\boldsymbol{v}_t)^T(I_N + \widetilde{U}^kM(\widetilde{U}^k)^T)^{-1}(\widetilde{U}^k\boldsymbol{v}_t)\\ &\leq (\widetilde{U}^k\boldsymbol{v}_t)^T(\widetilde{U}^kM(\widetilde{U}^k)^T)^+(\widetilde{U}^k\boldsymbol{v}_t) = \boldsymbol{v}_t^TM^{-1}\boldsymbol{v}_t \leq 1.\end{aligned}$$

The last equality is due to Lemma 11 in Appendix.

Next we try to bound $(\|\boldsymbol{e}_t\|_{H^k}^*)^2$. We obtain that

$$(\|\boldsymbol{e}_t\|_{H^k}^*)^2 = \boldsymbol{e}_t^TH^{-1}\boldsymbol{e}_t = \boldsymbol{e}_t^T(I_N + \widetilde{U}^kM(\widetilde{U}^k)^T)^{-1}\boldsymbol{e}_t \leq \boldsymbol{e}_t^TI_n\boldsymbol{e}_t \leq \|\boldsymbol{e}_t\|_2^2. \tag{21}$$

Based on previous discussion we have that $\|\boldsymbol{e}_t\|_2 \leq \max\{2s_k, 2\epsilon + \gamma_k(\|\tilde{\boldsymbol{l}}_t\|_2 + \epsilon)\}$. Therefore we have that

$$\|\boldsymbol{e}_t\|_2 \leq \max\{2s_k, 2\epsilon + \gamma_k(\sqrt{N} + \epsilon)\}. \tag{22}$$

Due to above Lemma 7, we have that $\|\boldsymbol{e}\|_2 \leq m_k$. In conclusion we have:

$$(\|\tilde{\boldsymbol{l}}_t\|_{H^k}^*)^2 \leq 2(1 + m_k)^2. \tag{23}$$

Now we show the bound of $\|\boldsymbol{w}_1 - \boldsymbol{w}^*\|_{H^k}$. Since $\|\boldsymbol{w}_1 - \boldsymbol{w}^*\|_{H^k} \leq 2\max_{\boldsymbol{w}\in\Delta(N)}\|\boldsymbol{w}\|_{H^k}$, it suffices to bound that $\max_{\boldsymbol{w}\in\Delta(N)}\|\boldsymbol{w}\|_{H^k}$. Since $\|\boldsymbol{w}\|_{H^k}^2 \leq 1 + 2k\|\boldsymbol{w}\|_{H(k)'}^2$ with $H(k)' = \frac{1}{2k}\widetilde{U}^kM(\widetilde{U}^k)^T$.

Given a convex set P in $\mathbb{R}^k$, so the dual set P^* is defined as

$$P^* = \{\boldsymbol{x} : \sup_{\boldsymbol{p}\in P} |\boldsymbol{x}\cdot\boldsymbol{p}| \le 1\}.$$

The dual of an ellipsoid $\mathcal{E}(M)$ is given by $(\mathcal{E}(M))^* = \mathcal{E}(M^{-1})$ and since Equation (20) it is standard to show that

$$(\mathcal{E}(M))^* \subseteq P^* \subseteq (\mathcal{E}(\frac{1}{2k}M))^*. \tag{24}$$

It implies that $P^* \subseteq \mathcal{E}(2kM^{-1})$. Note that due to the definition of P^* we have that for all $i \in [N]$, $(1+\epsilon)^{-1}\tilde{\boldsymbol{u}}_i \cdot \boldsymbol{v} \le 1$, where $\tilde{\boldsymbol{u}}_i$ is the i-th row of $\widetilde{U}^k$. So each row of $\widetilde{U}^k$ are in P^*, thus we have for each $\tilde{\boldsymbol{u}}_i$:

$$\begin{aligned}(1+\epsilon)^{-1}\tilde{\boldsymbol{u}}_i^T \left(2kM^{-1}\right)^{-1}(1+\epsilon)\tilde{\boldsymbol{u}}_i \le 1 &\Leftrightarrow (1+\epsilon)^{-1}\tilde{\boldsymbol{u}}_i^T M (1+\epsilon)^{-1}\tilde{\boldsymbol{u}}_i \le 2k \\ &\Leftrightarrow \|(1+\epsilon)^{-1}\tilde{\boldsymbol{u}}_i\|_M^2 \le 2k\end{aligned} \tag{25}$$

Since $\boldsymbol{w} \in \Delta(N)$, we have that

$$\|\boldsymbol{w}\|_{H(k)'}^2 = \frac{1}{2k}\|\widetilde{U}\boldsymbol{w}\|_M^2 \le \frac{1}{2k}\max_i \|\boldsymbol{u}_i\|_M^2 \le (1+\epsilon)^2.$$

Therefore we have that $\|\boldsymbol{w}\|_{H^k}^2 \le 1 + (1+\epsilon)^2 2k$. Moreover we obtain that $\|\boldsymbol{w}\|_{H^k} \le 2(1+\epsilon)\sqrt{k}$. Given $\boldsymbol{w}^* = \arg\min_{t=1}^T \sum_{t=1}^T \boldsymbol{w}\cdot\tilde{\boldsymbol{l}}_t$, then we have that

$$\begin{aligned}\text{Regret}_{T_k} &= \sum_{t\in T_k} \boldsymbol{w}_t\cdot\tilde{\boldsymbol{l}}_t - \sum_{t\in T_k}\tilde{\boldsymbol{l}}_t\cdot\boldsymbol{w}^* \\ &\le \frac{1}{\eta_{T_k}}\|\boldsymbol{w}_{T_{k-1}} - \boldsymbol{w}^*\|_{H^k}^2 + \frac{1}{2}\sum_{t\in T_k}\eta_t(\|\tilde{\boldsymbol{l}}_t\|_{H^k}^*)^2 \\ &\le \frac{4}{\eta_{T_k}}\max_{\boldsymbol{w}\in\Delta(N)}\|\boldsymbol{w}\|_{H^k}^2 + \frac{1}{2}\sum_{t\in T_k}\eta_t(\|\tilde{\boldsymbol{l}}_t\|_{H^k}^*)^2 \\ &\le \frac{8k(1+\epsilon)^2}{\eta_{T_k}} + \sum_{t\in T_k}\eta_t(1+m_k)^2,\end{aligned}$$

where we set that $\boldsymbol{w}_{T_{k-1}}$ as last term in episode T_{k-1}, the first inequality is due to Theorem 1.

Particularly, we can choose $\eta_t = \sqrt{8k(1+\epsilon)^2/(1+m_k^2)t}$, then we get the regret bound as

$$\text{Regret}_{\text{T}_\text{k}} \le O(\sqrt{(8k(1+\epsilon)^2)(1+m_k)^2 T_k}).$$

Thus, we have the regret as

$$\text{Regret}_T \le O\left((1+\epsilon)\sqrt{T}\left(K + \sqrt{\sum_{k=1}^K k m_k^2}\right)\right) \tag{26}$$

■

3.1. Parameter optimization

In this sub-section we are going to give an optimal setting of parameter γ_i, m_i and s_i. Since Equation (17), we can simplify

$$m_i = \max\{2s_i, 6\epsilon + \gamma_i\sqrt{N}\} \leq 2s_i + 2\gamma_i\sqrt{N}, \tag{27}$$

if we assume that $6\epsilon \leq \gamma_i\sqrt{N}, \forall i \in [K]$. Then we need to solve the following optimal problem:

$$\begin{aligned} &\min s_i + \gamma_i\sqrt{N} \\ &\text{s.t.}\gamma_i - \frac{\beta\epsilon}{s_i\gamma_i}i > 0 \quad \wedge \quad \beta \geq \frac{\pi}{2} \geq \gamma_i > 0 \quad \forall i \in \{1, \cdots, K\}. \end{aligned} \tag{28}$$

where the latter inequality is due to Proposition 5.

It is equivalent to solve:

$$\begin{aligned} &\min s_i + \gamma_i\sqrt{N} \\ &\text{s.t.}s_i \geq (1+x)\frac{\beta\epsilon i}{\gamma_i} \quad \wedge \quad \beta \geq \frac{\pi}{2} \geq \gamma_i > 0 \quad \forall i \in \{1, \cdots, K\}, \forall x > 0. \end{aligned} \tag{29}$$

With simple calculation we have that

$$\begin{cases} s_i = (N(1+x)\beta i\epsilon)^{1/3}, \\ \gamma_i = ((1+x)\beta i\epsilon)^{1/3} N^{-1/6}, \end{cases} \tag{30}$$

for any $x > 0$.

Let us re-consider the constraint in Equation (6)

$$\frac{\gamma_i}{\gamma_{i-1} - (i-1)\frac{\beta\epsilon}{s_{i-1}\gamma_{i-1}}} \cdot \frac{\pi}{2\beta} + \frac{s_i\gamma_i}{s_i\gamma_{i-1}}(i-1) \leq i \quad \forall i \in \{2, \cdots, K\},$$

with above s_i and γ_i.

We can simplify this equation as

$$\frac{\pi}{2\beta} \leq i^{1/3}(i-1)^{1/3}(i^{1/3} - (i-1)^{1/3}) \cdot x, \forall i \in \{2, \cdots, K\}. \tag{31}$$

Without loss the generality, we can set that $x = 1$ and $\beta = 10$, according to Lemma 14.

Therefore if we set that $s_k = (20kN\epsilon)^{1/3}$ and $\gamma_k = \sqrt{20k\epsilon/s_k}$, in Algorithm 2, we have the following bound:

$$\text{Regret}_T \leq O\left((1+\epsilon)\sqrt{T}(K + K^{4/3}(N\epsilon)^{1/3})\right), \tag{32}$$

with the constraint that $6\epsilon \leq \gamma_k\sqrt{N} \Leftrightarrow \epsilon^2 \leq \frac{20kN}{216}$, $\gamma_k \leq \frac{\pi}{2} \Leftrightarrow \epsilon^2 \leq \frac{64\pi^6 N}{400k^2}$ and $s_k \geq \epsilon \Leftrightarrow \epsilon \leq 20kN$ for all $k \in [K]$, where the last constraint is due to Remark 4.

Corollary 9 *If* $\epsilon^2 \leq \min\{\frac{64\pi^6 N}{400k^2}, 20kN, \frac{20kN}{216}\}, \forall k \in [d]$, *setting* $s_k = (20kN\epsilon)^{1/3}$ *and* $\gamma_k = \sqrt{20k\epsilon/s_k}$, *running our algorithm for* T *times we have the regret bound as*

$$\text{Regret}_T \leq O\left((1+\epsilon)\sqrt{T}(K + K^{4/3}(N\epsilon)^{1/3})\right) \leq O\left((1+\epsilon)\sqrt{T}(d + d^{4/3}(N\epsilon)^{1/3})\right). \tag{33}$$

4. Experiments

We perform preliminary experiments using a synthetic environment. In our experiments, we construct the environment as follows: Firstly, we produce a $N \times d$ matrix U, if the rank of U is not d, then produce another one, until we obtain a d-rank matrix U as kernel.

To approach the maximal regret in our experiment, before the algorithms start, we randomly produce $T \times M$ seed vectors $\boldsymbol{v}_t^j$, where $\boldsymbol{v}_t^j(i) \in [-1, 1]$, $\forall i \in [d]$, $t \in [T]$ and $j \in [M]$. Next we denote $\boldsymbol{l}_t^j = (U\boldsymbol{v}_t^j)/\|U\boldsymbol{v}_t^j\|_\infty$, and the noise vector $\boldsymbol{\epsilon}_t = s\tilde{\boldsymbol{\epsilon}}_t/(\|\tilde{\boldsymbol{\epsilon}}_t\|_2)$, where $\tilde{\boldsymbol{\epsilon}}_t(i)$ is randomly produced between $[-1, +1]$. We define $\tilde{\boldsymbol{l}}_t^j = \boldsymbol{l}_t^j + \boldsymbol{\epsilon}_t$. Then, we process Hedge, Hazan's and our algorithm(Algorithm 2) with the same input sequences accordingly, and by going this procedure for M sequence respectively, we can obtain M regret results for each algorithm, then choosing the maximum value among those M as follows:

On round t, for any algorithm $\mathcal{A}$, we record maximum of the regret with respect to M loss sequence as $\mathrm{Regret}_{\mathcal{A}}(t)$

$$\mathrm{Regret}_{\mathcal{A}}(t) = \max_{j \in [M]} \left\{ \sum_{s=1}^{t} \boldsymbol{w}_s^j \cdot \tilde{\boldsymbol{l}}_s^j - \min_{i \in [N]} \sum_{s=1}^{t} \tilde{\boldsymbol{l}}_s^j \right\}, \tag{34}$$

where we denote $\boldsymbol{w}_s^j$ as output of $\mathcal{A}$ with j-th loss sequence on round s.

Firstly, we set $T = 1000, M = 5, N = 300, d = 2$. and plot $\mathrm{Regret}_{\mathcal{A}}(T)$ of each algorithm as a function of ϵ for $T = 1000$ in Figure 1.

As can be seen in Figure 1, our algorithm performs more robustly than others for different choices of noise ϵ. In particular, we set ϵ(i.e., setting s) as $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$, respectively, and other parameters remain, and detailed graphs for each ϵ is in the supplementary material.

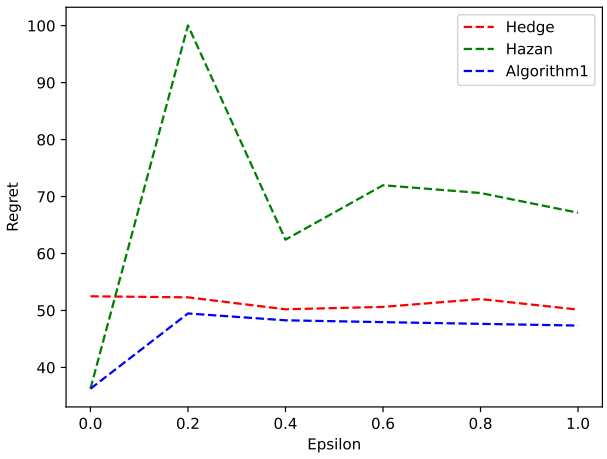
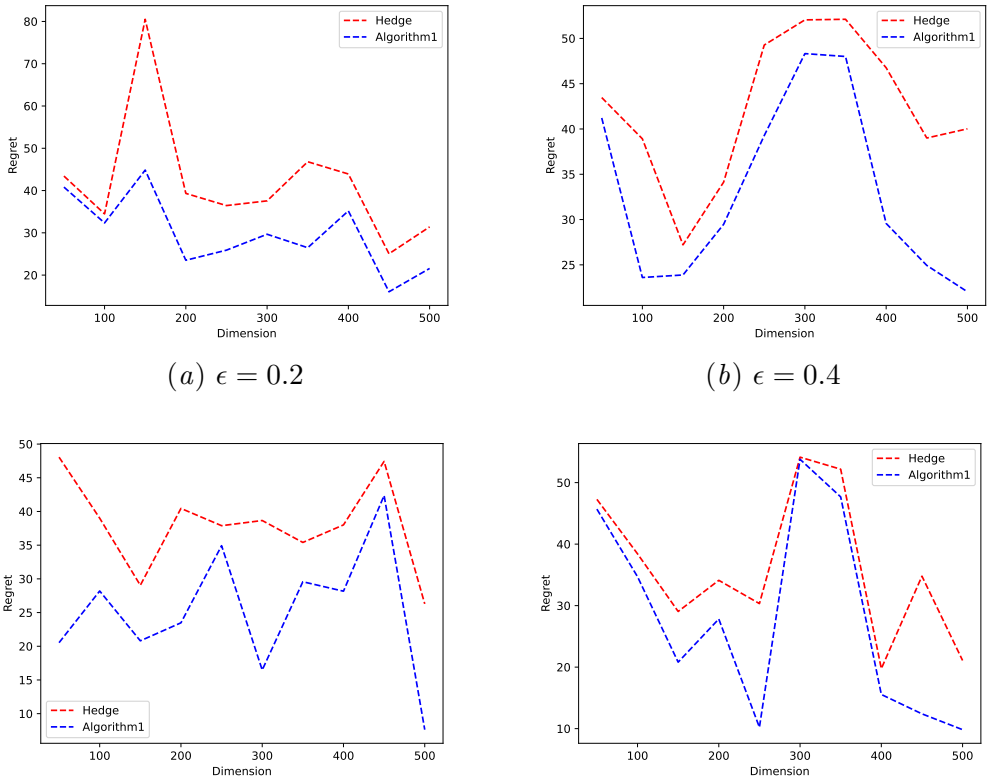


Figure 1: $\mathrm{Regret}_{\mathcal{A}}(T)$ as a function of ϵ for $T = 1000$

Secondly, in order to see the results based on the different dimensions, we construct the experiments and set $N = 50, 100, 150, ..., 500$, and $\epsilon = 0.2, 0.4, 0.6, 0.8$, since the unsatisfying

performance of Hazan's algorithm shown in Figure 1, and rapidly increased running time with respect to dimension, we discard it to make sure the experiments can be done in an acceptable amount of time.



(a) $\epsilon = 0.2$ (b) $\epsilon = 0.4$

(c) $\epsilon = 0.6$ (d) $\epsilon = 0.8$

Figure 2: Results for different choices of N and ϵ

5. Concluding remarks

In this paper, we construct an algorithm for expert advice with noisy low rank loss. This algorithm is designed for the problem where the algorithm obtains no prior information about the low rank structure but only the noise bound ϵ. Theoretically, we achieve a regret bound as $O(\sqrt{T}(d + d^{4/3}(N\epsilon)^{1/3}))$, however, in the experiments, our algorithm performs better even if $\epsilon \geq \Omega\left(\frac{1}{N}\right)$, which indicates that there might be a gap between our bound and the optimal one.

6. Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers JP19H04174 and JP19H04067, respectively. We thank all reviewers for the helpful comments and suggestions. Liu thank Guangsheng Ma for the beneficial discussion about some mathematical details.

References

Maria-Florina F Balcan and Hongyang Zhang. Noise-tolerant life-long matrix completion via adaptive sampling. In *Advances in Neural Information Processing Systems*, pages 2955–2963, 2016.

Keith Ball et al. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31:1–58, 1997.

Siddharth Barman, Aditya Gopalan, and Aadirupa Saha. Online learning for structured loss spaces. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

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

Rina Foygel and Nathan Srebro. Concentration-based guarantees for low-rank matrix reconstruction. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 315–340, 2011.

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.

Andrew Goldberg, Ben Recht, Junming Xu, Robert Nowak, and Jerry Zhu. Transduction with matrix completion: Three birds with one stone. In *Advances in neural information processing systems*, pages 757–765, 2010.

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

Elad Hazan, Tomer Koren, Roi Livni, and Yishay Mansour. Online learning with low rank experts. In *Conference on Learning Theory*, pages 1096–1114, 2016.

Tomer Koren and Roi Livni. Affine-invariant online optimization and the low-rank experts problem. *Advances in Neural Information Processing Systems*, 30:4747–4755, 2017.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Francesco Orabona, Koby Crammer, and Nicolo Cesa-Bianchi. A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99(3):411–435, 2015.

Xiao Zhang, Lingxiao Wang Wang, and Quanquan Gu. A unified framework for nonconvex low-rank plus sparse matrix recovery. In *International Conference on Artificial Intelligence and Statistics*, 2018.

7. Appendix A. Necessary Lemmata

Lemma 10 ((Balcan and Zhang, 2016)) *Let* $\boldsymbol{W} = \text{span}\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{k-1}\}$, $\boldsymbol{U} = \text{span}\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{k-1}, \boldsymbol{u}\}$, *and* $\widetilde{\boldsymbol{U}} = \text{span}\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{k-1}, \tilde{\boldsymbol{u}}\}$ *be subspaces spanned by vectors in* $\mathbb{R}^m$. *Then*

$$\theta(\boldsymbol{U}, \widetilde{\boldsymbol{U}}) \leq \frac{\pi}{2} \frac{\theta(\tilde{\boldsymbol{u}}, \boldsymbol{u})}{\theta(\tilde{\boldsymbol{u}}, \boldsymbol{W})}$$

Lemma 11 ((Hazan et al., 2016) Lemma 11) *Let* $M \in \mathbb{R}^{k \times k}$, $U \in \mathbb{R}^{N \times d}$ *such that* $M \succ 0$ *and* U. *Then*

$$U^T (UMU^T)^+ U = M^{-1}. \tag{35}$$

Lemma 12 *For any* $\boldsymbol{l}_t \in \mathbb{R}^N$, *and two sub-spaces* $\widetilde{\boldsymbol{U}}^k, \boldsymbol{U}^k \subseteq \mathbb{R}^N$. *We have that*

$$\theta(\boldsymbol{l}_t, \boldsymbol{U}^k) \geq \theta(\boldsymbol{l}_t, \widetilde{\boldsymbol{U}}^k) - \theta(\boldsymbol{U}^k, \widetilde{\boldsymbol{U}}^k). \tag{36}$$

Proof Naturally we obtain that $\theta(l_t, \tilde{x}) \leq \theta(x, \tilde{x}) + \theta(l_t, x)$, $\forall x, \tilde{x}$.

Given $\tilde{x}^* = \arg\min_{\tilde{x} \in \widetilde{U}^k} \theta(x, \tilde{x})$, then we have that

$$\min_{\tilde{x} \in \widetilde{U}^k} \theta(l_t, \tilde{x}) \leq \theta(l_t, \tilde{x}^*) \leq \theta(x, \tilde{x}^*) + \theta(l_t, x) \tag{37}$$

Rearranging the terms we obtain that

$$\min_{\tilde{x} \in \widetilde{U}^k} \theta(l_t, \tilde{x}) - \min_{\tilde{x} \in \widetilde{U}^k} \theta(x, \tilde{x}) \leq \theta(l_t, x) \quad \forall x. \tag{38}$$

Since the definition we have that $\max_{x \in U^k} \min_{\tilde{x} \in \widetilde{U}^k} \theta(x, \tilde{x}) = \theta(U^k, \widetilde{U}^k)$, we have

$$\min_{\tilde{x} \in \widetilde{U}^k} \theta(l_t, \tilde{x}) - \theta(U^k, \widetilde{U}^k) \leq \min_{\tilde{x} \in \widetilde{U}^k} \theta(l_t, \tilde{x}) - \min_{\tilde{x} \in \widetilde{U}^k} \theta(x, \tilde{x}) \leq \theta(l_t, x) \quad \forall x. \tag{39}$$

At last, setting $x = \arg\min_{x \in U^k} \theta(l_t, x)$, we have that

$$\theta(l_t, \widetilde{U}^k) - \theta(U^k, \widetilde{U}^k) = \min_{\tilde{x} \in \widetilde{U}^k} \theta(l_t, \tilde{x}) - \theta(U^k, \widetilde{U}^k) \leq \min_{x \in U^k} \theta(l_t, x) = \theta(l_t, U^k). \tag{40}$$

■

Remark 13 *For any* $\boldsymbol{l}_t \in \boldsymbol{U}^k \subseteq \mathbb{R}^N$ *and two spaces* $\boldsymbol{U}_0^k, \widetilde{\boldsymbol{U}}^k \subseteq \mathbb{R}^N$. *We have*

$$\theta(\boldsymbol{U}^k, \boldsymbol{U}_0^k) + \theta(\boldsymbol{U}_0^k, \widetilde{\boldsymbol{U}}^k) \geq \theta(\boldsymbol{U}^k, \widetilde{\boldsymbol{U}}^k). \tag{41}$$

Proof According to Lemma 12 we have that

$$\theta(l_t, U_0^k) \geq \theta(l_t, \widetilde{U}^k) - \theta(U_0^k, \widetilde{U}^k). \tag{42}$$

Rearranging the equation we have that

$$\theta(l_t, U_0^k) + \theta(U_0^k, \widetilde{U}^k) \geq \theta(l_t, \widetilde{U}^k). \tag{43}$$

Letting $l'_t = \arg\max_{l_t \in U^k} \theta(l_t, \widetilde{U}^k)$, we obtain that $\theta(l'_t, U_0^k) + \theta(U_0^k, \widetilde{U}^k) \geq \theta(U^k, \widetilde{U}^k)$. Since $\theta(U^k, U_0^k) = \max_{l_t \in U^k} \theta(l_t, U_0^k) \geq \theta(l'_t, U_0^k)$, we have our conclusion that

$$\theta(U^k, U_0^k) + \theta(U_0^k, \widetilde{U}^k) \geq \theta(U^k, \widetilde{U}^k). \tag{44}$$

■

Lemma 14 *For any $k \geq 2$, we have that*

$$\left(\frac{k}{k-1}\right)^{2/3} \left(\pi \cdot \left(\frac{k-1}{k}\right)^{1/3} + 10(k-1)\right) \leq 10k. \tag{45}$$

Proof First we have that:

$$\begin{aligned}
&\left(\frac{k}{k-1}\right)^{2/3} \left(\pi \cdot \left(\frac{k-1}{k}\right)^{1/3} + 10(k-1)\right) \leq 10k \\
&\Leftrightarrow \left(\frac{k}{k-1}\right)^{1/3} \pi + 10k^{2/3}(k-1)^{1/3} \leq 10k \\
&\Leftrightarrow \pi \leq (10k - 10k^{2/3}(k-1)^{1/3}) \cdot \left(\frac{k-1}{k}\right)^{1/3} \\
&\Leftrightarrow \pi \leq 10k^{2/3}(k^{1/3} - (k-1)^{1/3}) \cdot k^{-1/3} \cdot (k-1)^{1/3} \\
&\Leftrightarrow \pi \leq 10k^{1/3}(k-1)^{1/3}(k^{1/3} - (k-1)^{1/3}).
\end{aligned}$$

We set that $g(k) = k^{1/3}(k-1)^{1/3}(k^{1/3} - (k-1)^{1/3})$, then we have that

$$g'(k) = \frac{-3k(k-1)^{1/3} + 3k^{4/3} - 2k^{1/3} + (k-1)^{1/3}}{3k^{2/3}(k-1)^{2/3}}. \tag{46}$$

Next we can show that

$$\begin{aligned}
&-3k(k-1)^{1/3} + 3k^{4/3} - 2k^{1/3} + (k-1)^{1/3} > 0 \\
&\Leftrightarrow k^{1/3}(3x-1) > (k-1)^{1/3}(3x-1) \\
&\Leftrightarrow k(3x-2)^3 > (k-1)(3k-1)^3 \\
&\Leftrightarrow 2k - 1 > 0.
\end{aligned}$$

Above equation implies that $\forall k \geq 2$, then we have $g'(k) \geq 0$, and $g(k)$ is an increasing function. Meanwhile since $\pi \leq 10 \cdot g(2)$, we obtain our conclusion. ■