Classification with Low Rank and Missing Data

Elad Hazan EHAZAN@CS.PRINCETON.EDU
Princeton University and Microsoft Research, Herzliya

Roi Livni ROI.LIVNI@MAIL.HUJI.AC.IL
The Hebrew University of Jerusalem and Microsoft Research, Herzliya

Yishay Mansour MANSOUR.YISHAY@GMAIL.COM
Microsoft Research, Hertzelia and Tel Aviv University

Abstract

We consider classification and regression tasks where we have missing data and assume that the (clean) data resides in a low rank subspace. Finding a hidden subspace is known to be computationally hard. Nevertheless, using a non-proper formulation we give an efficient agnostic algorithm that classifies as good as the best linear classifier coupled with the best low-dimensional subspace in which the data resides. A direct implication is that our algorithm can linearly (and non-linearly through kernels) classify provably as well as the best classifier that has access to the full data.

1. Introduction

The importance of handling correctly missing data is a fundamental and classical challenge in machine learning. There are many reasons why data might be missing. For example, consider the medical domain, some data might be missing because certain procedures were not performed on a given patient, other data might be missing because the patient choose not to disclose them, and even some data might be missing due to malfunction of certain equipment. While it is definitely much better to have always complete and accurate data, this utopian desire is, in many cases, unfulfilled. For this reason we need to utilize the available data even if some of it is missing.

Another, very different motivation for missing data are recommendations. For example, a movie recommendations dataset might have users opinions on certain movies, which is the case, for example, in the Netflix motion picture dataset. Clearly, no user has seen or reviewed all movies, or even close to it. In this respect recommendation data is an extreme case: the vast majority is usually missing (i.e., it is sparse to the extreme).

Many times we can solve the missing data problem since the data resides on a lower dimension manifold. In the above examples, if there are prototypical users (or patients) and any user is a mixture of the prototypical users, then this implicitly suggests that the data is *low rank*. Another way to formalize this assumption is to consider the data in a matrix form, say, the users are rows and movies are columns, then our assumption is that the true complete matrix has a low rank.

Our starting point is to consider the low rank assumption, but to avoid any explicit matrix completion, and instead directly dive in to the classification problem. At the end of the introduction we show that matrix completion is neither sufficient and/or necessary.

We consider perhaps the most fundamental data analysis technique of the machine learning toolkit: linear (and kernel) classification, as applied to data where some (or even most) of the attributes in an example might be missing. **Our main result is an efficient algorithm for linear and kernel classification that performs as well as the best classifier that has access to all data**, under low rank assumption with natural non-degeneracy conditions.

We stress that our result is worst case, we do not assume that the missing data follows any probabilistic rule other than the underlying matrix having low rank. This is a clear contrast to most existing matrix completion algorithms. We also cast our results in a distributional setting, showing that the classification error that we achieve is close to the best classification using the subspace of the examples (and with no missing data). Notably, many variants of the problem of finding a hidden subspace are computationally hard (see e.g. (Berthet & Rigollet, 2013)), yet as we show, learn-

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

ing a linear classifier on a hidden subspace is non-properly learnable.

At a high level, we assume that a sample is a triplet $(\mathbf{x}, \mathbf{o}, y)$, where $\mathbf{x} \in \mathbb{R}^d$ is the complete example, $\mathbf{o} \subset \{1, \ldots, d\}$ is the set of observable attributes and $y \in \mathcal{Y}$ is the label. The learner observes only $(\mathbf{x}_o, y)$, where $\mathbf{x}_o$ omits any attribute not in $\mathbf{o}$. Our goal is given a sample $S = \{(\mathbf{x}_o^{(i)}, y^{(i)})\}_{i=1}^m$ to output a classifier f_S such that w.h.p.:

$$\mathbb{E}\left[\ell(f_S(\mathbf{x}_\mathbf{o}), y)\right] \leq \min_{\substack{\mathbf{w} \in \mathbb{R}^d \\ \|\mathbf{w}\| \leq 1}} \mathbb{E}\left[\ell(\mathbf{w} \cdot \mathbf{x}, y)\right] + \epsilon,$$

where ℓ is the loss function. Namely, we like our classifier f_S to compete with the best linear classifier for the completely observable data.

Our main result is achieving this task (under mild regularity conditions) using a computationally efficient algorithm for any convex Lipschitz-bounded loss function. Our basic result requires a sample size which is quasi-polynomial, but we complement it with a kernel construction which can guarantee efficient learning under appropriate large margin assumptions. Our kernel depends only on the intersection of observable values of two inputs, and is efficiently computable. (We give a more detailed overview of our main results in Section 2.)

We complement our theoretical contributions with experimental findings that show superior classification performance both on synthetic data and on publicly-available recommendation data.

Previous work. Classification with missing data is a well studied subject in statistics with numerous books and papers devoted to its study, (see, e.g., (Little & Rubin, 2002)). The statistical treatment of missing data is broad, and to a fairly large extent assumes parametric models both for the data generating process as well as the process that creates the missing data. One of the most popular models for the missing data process is *Missing Completely at Random (MCAR)* where the missing attributes are selected independently from the values.

We outline a few of the main approaches handling missing data in the statistics literature. The simplest method is simply to discard records with missing data, even this assumes independence between the examples with missing values and their labels. In order to estimate simple statistics, such as the expected value of an attribute, one can use importance sampling methods, where the probability of an attribute being missing can depend on it value (e.g., using the Horvitz-Thompson estimator (Horvitz & Thompson, 1952)). A large body of techniques is devoted to *imputation* procedures which complete the missing data. This can be done by replacing a missing attribute by its mean (mean imputation), or using a regression based on the observed value (regression imputation), or sampling the other examples to complete the missing value (hot deck).[1] The imputation methodologies share a similar goal as matrix completion, namely reduce the problem to one with complete data, however their methodologies and motivating scenarios are very different. Finally, one can build a complete Bayesian model for both the observed and unobserved data and use it to perform inference (e.g. (**?**)). As with almost any Bayesian methodology, its success depends largely on selecting the right model and prior, this is even ignoring the computational issues which make inference in many of those models computationally intractable.

In the machine learning community, missing data was considered in the framework of limited attribute observability (Ben-David & Dichterman, 1998) and its many refinements (Dekel et al., 2010; Cesa-Bianchi et al., 2010; 2011; Hazan & Koren, 2012). However, to the best of our knowledge, the low-rank property is not captured by previous work, nor is the extreme amount of missing data. More importantly, much of the research is focused on selecting *which attributes to observe* or on missing attributes at test or train time (see also (Eban et al., 2014; Globerson & Roweis, 2006)). In our case the learner has no control which attributes are observable in an example and the domain is fixed. The latter case is captured in the work of (Chechik et al., 2008), who rescale inner-products according to the amount of missing data. Their method, however, does not entail theoretical guarantees on reconstruction in the worst case, and gives rise to non-convex programs.

A natural and intuitive methodology to follow is to treat the labels (both known and unknown) as an additional column in the data matrix and complete the data using a matrix completion algorithm, thereby obtaining the classification. Indeed, this exactly was proposed in the innovative work of (Goldberg et al., 2010), who connected the methodology of matrix completion to prediction from missing data. Although this is a natural approach, we show that completion is neither necessary nor sufficient for classification. Furthermore, the techniques for provably completing a low rank matrix are only known under probabilistic models with restricted distributions (Srebro, 2004; Candes & Recht, 2009; Lee et al., 2010; Salakhutdinov & Srebro, 2010; Shamir & Shalev-Shwartz, 2011). Agnostic and non-probabilistic matrix completion algorithms such as (Srebro et al., 2004; Hazan et al., 2012) we were not able to use for our purposes.

Is matrix completion sufficient and/or necessary? We demonstrate that classification with missing data is prov-

[1] We remark that our model implicitly includes mean-imputation or 0-imputation method and therefore will always outperform them.

ably different from that of matrix completion. We start by considering a learner that tries to complete the missing entries in an unsupervised manner and then performs classification on the completed data, this approach is close akin to imputation techniques, generative models and any other two step – unsupervised/supervised algorithm. Our example shows that even under realizable assumptions, such an algorithm may fail. We then proceed to analyze the approach previously mentioned – to treat the labels as an additional column.

To see that unsupervised completion is insufficient for prediction, consider the example in Figure 1: the original data is represented by filled red and green dots and it is linearly separable. Each data point will have one of its two coordinates missing (this can even be done at random). In the figure the arrow from each instance points to the observed attribute. However, the rank-one completion of projection onto the pink hyperplane is possible, and admits no separation. The problem is clearly that the mapping to a low dimension is independent from the labels, and therefore we should not expect that properties that depend on the labels, such as linear separability, will be maintained.

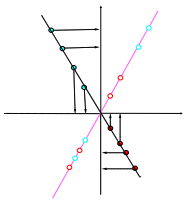

Figure 1. Linearly separable data, for which certain completions make the data non-separable.

Next, consider a learner that treats the labels as an additional column. (Goldberg et al., 2010) Considered the following problem:

$$\begin{aligned} \underset{Z}{\text{minimize}} \quad & \text{rank}(Z) \\ \text{subject to:} \quad & Z_{i,j} = \mathbf{x}_{i,j}, \quad (i,j) \in \Omega \ , . \end{aligned} \tag{G}$$

where Ω is the set of observed attributes (or observed labels for the corresponding columns). Now assume that we always see one of the following examples: $[1,\ *,\ 1,\ *]$, $[*,\ -1,\ *,\ -1]$, or $[1,\ ,-1,\ 1,\ -1]$. The observed labels are respectively $1,-1$ and 1. A typical data matrix with one test point might be of the form:

$$M = \left[\begin{array}{cccc||c} 1 & * & 1 & * & 1 \\ * & -1 & * & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & * & 1 & * & * \end{array}\right] \tag{1}$$

First note that there is no 1-rank completion of this matrix. On the other hand, there is more than one 2-rank completion each lead to a different classification of the test point. The first two possible completions are to complete odd columns to a constant one vector, and even column vectors to a constant -1 vector. Then complete the labeling whichever way you choose. We can also complete the first and last rows to a constant 1 vector, and the second row to a constant -1 vector. All possible completions lead to an optimal solution w.r.t Problem G but have different outcome w.r.t classification. We stress that this is not a sample complexity issue. Even if we observe abundant amount of data, the completion task is still ill-posed.

Finally, matrix completion is also not necessary for prediction. Consider movie recommendation dataset with two separate populations, French and Chinese, where each population reviews a different set of movies. Even if each population has a low rank, performing successful matrix completion, in this case, is impossible (and intuitively it does not make sense in such a setting). However, linear classification in this case is possible via a single linear classifier, for example by setting all non-observed entries to zero. For a numerical example, return to the matrix M in Eq. 1. Note that we observe only three instances hence the classification task is easy but does not lead to reconstruction of the missing entries.

2. Problem Setup and Main Result

We begin by presenting the general setting: A vector with missing entries can be modeled as a tuple $\mathbf{x} \times \mathbf{o}$, where $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{o} \in 2^d$ is a subset of indices. The vector $\mathbf{x}$ represents the *full data* and the set $\mathbf{o}$ represents the *observed attributes*. Given such a tuple, let us denote by $\mathbf{x_o}$ a vector in $(\mathbb{R} \cup \{*\})^d$ such that

$$(\mathbf{x_o})_i = \begin{cases} \mathbf{x}_i & i \in \mathbf{o} \\ * & \text{else} \end{cases}$$

The task of learning a linear classifier with missing data is to return a target function over $\mathbf{x_o}$ that competes with best linear classifier over $\mathbf{x}$. Specifically, a sequence of triplets $\{(\mathbf{x}^{(i)}, \mathbf{o}^{(i)}, y_i)\}_{i=1}^m$ is drawn iid according to some distribution D. An algorithm is provided with the sample $S = \{(\mathbf{x}^i_{\mathbf{o}^i}, y_i)\}_{i=1}^m$ and should return a target function f_S over missing data such that w.h.p:

$$\mathbb{E}\left[\ell(f_S(\mathbf{x_o}), y))\right] \leq \min_{\mathbf{w} \in B_d(1)} \mathbb{E}\left[\ell(\mathbf{w} \cdot \mathbf{x}, y))\right] + \epsilon, \tag{2}$$

where $\ell : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ is the loss function and $B_d(r)$ denotes the Euclidean ball in dimension d of radius $\sqrt{r}$. For brevity, we will say that a target function f_S is ϵ-good if Eq. 2 holds.

Without any assumptions on the distribution D, the task is ill-posed. One can construct examples where the learner over missing data does not have enough information to compete with the best linear classifier. Such is the case when, e.g., y_i is some attribute that is constantly concealed and independent of all other features. Therefore, certain assumptions on the distribution must be made.

One reasonable assumption is to assume that the marginal distribution D over $\mathbf{x}$ is supported on a small dimensional linear subspace E and that for every set of observations, we can linearly reconstruct the vector $\mathbf{x}$ from the vector $P_{\mathbf{o}}\mathbf{x}$, where $P_{\mathbf{o}} : \mathbb{R}^d \to \mathbb{R}^{|\mathbf{o}|}$ is the projection on the observed attributes. In other words, we demand that the mapping $P_{\mathbf{o}|E} : E \to P_{\mathbf{o}}E$, which is the restriction of $P_{\mathbf{o}}$ to E, is full-rank. As the learner doesn't have access to the subspace E, the learning task is still far from trivial.

We give a precise definition of the last assumption in Assumption 1. Though our results hold under the low rank assumption the convergence rates we give depend on a certain regularity parameter. Roughly, we parameterize the "distance" of $P_{\mathbf{o}|E}$ from singularity, and our results will quantitatively depend on this distance. Again, we defer all rigorous definitions to Section 3.2.

2.1. Main Result

Our first result is a an upper bound on the sample complexity of the problem. We then proceed to a more general statement that entails an efficient kernel-based algorithm. Proofs are available in the supplamentary material (see (**?**) for a full version).

Theorem 1 (Main Result)**.** *Assume that ℓ is a L-Lipschitz convex loss function, let D be a λ-regular distribution (see Definition 1), let $\gamma(\epsilon) \geq \frac{\log 2L/(\lambda\epsilon)}{\lambda}$ and*

$$\Gamma(\epsilon) = \frac{d^{\gamma(\epsilon)+1} - d}{d-1}.$$

There exists an algorithm (independent of D) that receives a sample $S = \{(\mathbf{x}^i_{\mathbf{o}^i}, y_i)\}_{i=1}^m$ of size $m \in \Omega\left(\frac{L^2\Gamma(\epsilon)^2 \log 1/\delta}{\epsilon^2}\right)$ and returns a target function f_S that is ϵ-good with probability at least $(1-\delta)$. The algorithm runs in time $\mathrm{poly}(|S|)$.

As the sample complexity in Theorem 1 is quasipolynomial, the result has limited practical value in many situations. However, as the next theorem states, f_S can actually be computed by applying a kernel trick. Thus, under further *large margin* assumptions we can significantly improve performance.

Theorem 2. *For every $\gamma \geq 0$, there exists an embedding over missing data*

$$\phi_\gamma : \mathbf{x}_{\mathbf{o}} \to \mathbb{R}^\Gamma,$$

such that $\Gamma = \sum_{k=1}^{\gamma} d^k = \frac{d^{\gamma+1}-d}{d-1}$, and the scalar product between two samples $\phi_\gamma(\mathbf{x}^1_{\mathbf{o}^1})$ and $\phi_\gamma(\mathbf{x}^2_{\mathbf{o}^2})$ can be efficiently computed, specifically it is given by the formula:

$$k_\gamma(\mathbf{x}^1_{\mathbf{o}^1}, \mathbf{x}^2_{\mathbf{o}^2}) := \frac{|\mathbf{o}^{(1)} \cap \mathbf{o}^{(2)}|^\gamma - 1}{|\mathbf{o}^{(1)} \cap \mathbf{o}^{(2)}| - 1} \sum_{i \in \mathbf{o}^{(1)} \cap \mathbf{o}^{(2)}} \mathbf{x}^{(1)}_i \cdot \mathbf{x}^{(2)}_i.$$

In addition, let ℓ be an L-Lipschitz loss function and $S = \{(\mathbf{x}^i_{\mathbf{o}^i}, y_i)\}_{i=1}^m$ a sample drawn iid according to a distribution D. We make the assumption that $\|P_{\mathbf{o}}\mathbf{x}\| \leq 1$ a.s. The followings hold:

1. *At each iteration of Alg. 1 we can efficiently compute $\mathbf{v}_t^\top \phi_\gamma(\mathbf{x}^t_{\mathbf{o}^t})$ for any new example $\mathbf{x}^t_{\mathbf{o}^t}$. Specifically it is given by the formula*

$$\mathbf{v}_t^\top \phi_\gamma(\mathbf{x}^t_{\mathbf{o}^t}) := \sum_{i=1}^{t-1} \alpha_i^{(t-1)} k(\mathbf{x}^i_{\mathbf{o}^i}, \mathbf{x}^t_{\mathbf{o}^t}).$$

   Hence Alg. 1 runs in $\mathrm{poly}(T)$ time and sequentially produces target functions $f_t(\mathbf{x}_{\mathbf{o}}) = \mathbf{v}_t^\top \phi_\gamma(\mathbf{x}_{\mathbf{o}})$ that can be computed at test time in $\mathrm{poly}(T)$ time.

2. *Run Alg. 1 with fixed $T > 0$, $0 < \rho < \frac{1}{4}$ and $\eta_t = \frac{1}{\rho t}$. Let $\bar{\mathbf{v}} = \frac{1}{T}\sum_{t=1}^T \mathbf{v}_t$. For every $\mathbf{v}$ let*

$$\hat{F}_\rho(\mathbf{v}) = \frac{\rho}{2}\|\mathbf{v}\|^2 + \frac{1}{m}\sum_{i=1}^m \ell(\mathbf{v}^\top \phi_\gamma(\mathbf{x}^i_{\mathbf{o}^i}), y_i)$$

   then with probability $(1-\delta)$:

$$\hat{F}_\rho(\bar{\mathbf{v}}) \leq \min \hat{F}_\rho(\mathbf{v}) + O\left(\frac{L^2\Gamma \ln T/\delta}{\rho T}\right). \qquad (3)$$

3. *For any $\epsilon > 0$, if D is a λ-regular distribution and $\gamma \geq \frac{\log 2L/(\lambda\epsilon)}{\lambda}$ then for some $\mathbf{v}^* \in B_\Gamma(\Gamma)$*

$$\mathbb{E}\left[\ell(\mathbf{v}^* \cdot \phi_\gamma(\mathbf{x}_{\mathbf{o}}), y)\right] \leq \min_{\mathbf{w} \in B_d(1)} \mathbb{E}\left[\ell(\mathbf{w} \cdot \mathbf{x}, y)\right] + \epsilon.$$

To summarize, Theorem 2 states that we can embed the sample points with missing attributes in a high dimensional, finite, Hilbert space of dimension Γ, such that:

- The scalar product between embedded points can be computed efficiently. Hence, due to the conventional representor argument, the task of empirical risk minimization is tractable.
- Following the conventional analysis of kernel methods: Under large margin assumptions in the ambient space, we can compute a predictor with scalable sample complexity and computational efficiency.

- Finally, the best linear predictor over embedded sample points in a $\sqrt{\Gamma}$–ball is comparable to the best linear predictor over fully observed data. Taken together, we can learn a predictor with sample complexity $\Omega(\Gamma^2(\epsilon)/\epsilon^2 \log \frac{1}{\delta})$ and Theorem 1 holds.

For completeness we present the method together with an efficient algorithm that optimizes the RHS of Eq. 3 via an SGD method. The optimization analysis is derived in a straightforward manner from the work of (Shalev-Shwartz et al., 2011). Other optimization algorithms exist in the literature, and we chose this optimization method as it allows us to also derive regret bounds which are formally stronger (see Section 2.2). We point out that the main novelty of this paper is in the introduction of a new kernel and our guarantees do not depend on a specific optimization algorithm.

Finally, note that ϕ_1 induces the same scalar product as a 0-imputation. In that respect, by considering different $\gamma = 1, 2, \ldots$ and using a holdout set we can guarantee that our method will outperform the 0-imputation method.

2.2. Regret minimization for joint subspace learning and classification

A significant technical contribution of this manuscript is the agnostic learning of a subspace coupled with a linear classifier. A subspace is represented by a projection matrix $Q \in \mathbb{R}^{d\times d}$, which satisfies $Q^2 = Q$. Denote the following class of target functions

$$\mathcal{F}_0 = \{f_{\mathbf{w},Q} : \mathbf{w} \in B_d,\ Q \in M_{d\times d},\ Q^2 = Q\}$$

where $f_{\mathbf{w},Q}(\mathbf{x_o})$ is the linear predictor defined by $\mathbf{w}$ over subspace defined by the matrix Q, as formally defined in definition 2.

Given the aforementioned efficient kernel mapping ϕ_γ, we consider the following kernel-gradient-based online algorithm for classification called KARMA (Kernelized Algorithm for Risk-minimization with Missing Attributes):

Our main result for the fully adversarial online setting is given next, and proved in the supplamentary material and in the full version. Notice that the subspace E^* and associated projection matrix Q^* are chosen by an adversary and unknown to the algorithm.

Theorem 3. *For any $\gamma > 1, \lambda > 0, X > 0, \rho > 0, B > 0$, L-Lipschitz convex loss function ℓ, and λ-regular sequence $\{(\mathbf{x}^t, \mathbf{o}^t, y_t)\}$ w.r.t subspace E^* and associated projection matrix Q^* such that $\|\mathbf{x}^t\|_\infty < X$, Run Algorithm 1 with $\{\eta_t = \frac{1}{\rho t}\}$, sequentially outputs $\{\mathbf{v}_t \in R^t\}$ that satisfy*

$$\sum_t \ell(\mathbf{v}_t^\top \phi_\gamma(\mathbf{x}^t_{\mathbf{o}_t}), y_t) - \min_{\|\mathbf{w}\|\le 1} \sum_t \ell(f_{\mathbf{w},Q^*}(\mathbf{x}^t_{\mathbf{o}_t}), y_t) \le \frac{2L^2X^2\Gamma}{\rho}(1+\log T) + \frac{\rho}{2}T\cdot B + \frac{e^{-\lambda\gamma}}{\lambda}LT$$

Algorithm 1 KARMA: Kernelized Algorithm for Risk-minimization with Missing Attributes

Input: parameters $\gamma > 1,\ \{\eta_t > 0\}, 0 < \rho < 1$
Initialize: $\mathbf{v}_1 = 0,\ \alpha^{(0)} = 0$.
for $t = 1$ to T **do**
  Observe example $(\mathbf{x}^t_{\mathbf{o}^t}, y_t)$, suffer loss $\ell(\mathbf{v}_t^\top \phi_\gamma(\mathbf{x}^t_{\mathbf{o}^t}), y_t)$
  Update (ℓ' denotes the derivative w.r.t. the first argument)

$$\alpha_i^{(t)} = \begin{cases} (1-\eta_t\rho)\cdot\alpha_i^{(t-1)} & i < t \\ -\eta_t \ell'(\mathbf{v}_t^\top \phi_\gamma(\mathbf{x}^t_{\mathbf{o}^t}), y_t) & i = t \\ 0 & \text{else} \end{cases}$$

$$\mathbf{v}_{t+1} = \sum_{i=1}^t \alpha_i^{(t)} \phi_\gamma(\mathbf{x}^i_{\mathbf{o}^i})$$

end for

In particular, taking $\rho = \frac{LX\sqrt{\Gamma}}{\sqrt{BT}}$, $\gamma = \frac{1}{\lambda}\log T$ we obtain for every $\|\mathbf{w}\| \le 1$:

$$\sum_t \ell(\mathbf{v}_t^\top \phi_\gamma(\mathbf{x}^t_{\mathbf{o}_t}), y_t) - \ell(f_{\mathbf{w},Q^*}(\mathbf{x}^t_{\mathbf{o}_t}), y_t) \in O(XL\sqrt{\Gamma BT})$$

3. Preliminaries and Notations

3.1. Notations

As discussed, we consider a model where a distribution D is fixed over $\mathbb{R}^d \times \mathcal{O} \times \mathcal{Y}$, where $\mathcal{O} = 2^d$ consists of all subsets of $\{1, \ldots, d\}$. We will generally denote elements of $\mathbb{R}^d$ by $\mathbf{x}, \mathbf{w}, \mathbf{v}, \mathbf{u}$ and elements of $\mathcal{O}$ by $\mathbf{o}$. We denote by B_d the unit ball of $\mathbb{R}^d$, and by $B_d(r)$ the ball of radius $\sqrt{r}$.

Given a subset $\mathbf{o}$ we denote by $P_{\mathbf{o}} : \mathbb{R}^d \to \mathbb{R}^{|\mathbf{o}|}$ the projection onto the indices in $\mathbf{o}$, i.e., if $i_1 \le i_2 \le \cdots \le i_k$ are the elements of $\mathbf{o}$ in increasing order then $(P_{\mathbf{o}}\mathbf{x})_j = \mathbf{x}_{i_j}$. Given a matrix A and a set of indices $\mathbf{o}$, we let

$$A_{\mathbf{o},\mathbf{o}} = P_{\mathbf{o}} A P_{\mathbf{o}}^\top.$$

Finally, given a subspace $E \subseteq \mathbb{R}^d$ we denote by $P_E : \mathbb{R}^d \to \mathbb{R}^d$ the projection onto E.

3.2. Model Assumptions

Definition 1 (λ-regularity)**.** *We say that D is λ-regular with associated subpsace E if the following happens with probability 1 (w.r.t the joint random variables $(\mathbf{x}, \mathbf{o})$):*

1. $\|P_{\mathbf{o}}\mathbf{x}\| \le 1$.
2. $\mathbf{x} \in E$.

3. $\ker(P_{\mathbf{o}}P_E) = \ker(P_E)$
4. *If $\lambda_{\mathbf{o}} > 0$ is a strictly positive singular value of the matrix $P_{\mathbf{o}}P_E$ then $\lambda_{\mathbf{o}} \geq \lambda$.*

Assumption 1 (Low Rank Assumption)**.** *We say that D satisfies the low rank assumption with associated subspace E if it is λ-regular with associated subspace E for some $\lambda > 0$.*

Let $\mathbf{o}$ be a set of probable observables, and let $X_{\mathbf{o}}$ be the matrix obained by taking only columns in $\mathbf{o}$ from the data matrix X. If $\mathrm{rank}(X_{\mathbf{o}}) < \mathrm{rank}(X)$ then Assumption 1 is not satisfied. Since $\mathrm{rank}(X_{\mathbf{o}}) < |\mathbf{o}|$ we have that assumption 1 is met only if $\mathrm{rank}(X)$ is upperbounded minimal number of observations. Previous example, one can show that without some further restrictive assumptions, if the rank was larger than is required in the assumption, the problem of finding an ϵ-good function becomes ill-posed.

4. Learning under low rank assumption and λ-regularity.

Definition 2 (The class $\mathcal{F}_0$)**.** *We define the following class of target functions*

$$\mathcal{F}_0 = \{f_{\mathbf{w},Q} : \mathbf{w} \in B_d(1),\ Q \in M_{d\times d},\ Q^2 = Q\}$$

where

$$f_{\mathbf{w},Q}(\mathbf{x}_{\mathbf{o}}) = (P_{\mathbf{o}}\mathbf{w}) \cdot Q^{\dagger}_{\mathbf{o},\mathbf{o}} \cdot (P_{\mathbf{o}}\mathbf{x}).$$

(Here $M^{\dagger}$ denotes the pseudo inverse of M.)

The following Lemma states that, under the low rank assumption, the problem of linear learning with missing data is reduced to the problem of learning the class $\mathcal{F}_0$, in the sense that the hypothesis class $\mathcal{F}_0$ is not less-expressive.

Lemma 1. *Let D be a distribution that satisfies the low rank assumption. For every $\mathbf{w}^* \in \mathbb{R}^d$ there is $f^*_{\mathbf{w},Q} \in \mathcal{F}_0$ such that a.s:*

$$f^*_{\mathbf{w},Q}(\mathbf{x}_{\mathbf{o}}) = \mathbf{w}^* \cdot \mathbf{x}.$$

In particular $Q = P_E$ and $\mathbf{w} = P_E^{\top}\mathbf{w}^*$.

4.1. Approximating $\mathcal{F}_0$ under regularity

We next define a surrogate class of target functions that approximates $\mathcal{F}_0$

Definition 3 (The classes $\mathcal{F}^{\gamma}$)**.** *For every γ we define the following class*

$$\mathcal{F}^{\gamma} = \{f^{\gamma}_{\mathbf{w},Q} : \mathbf{w} \in B_d(1),\ Q \in \mathbb{R}^{d\times d},\ Q^2 = Q\}$$

where,

$$f^{\gamma}_{\mathbf{w},Q}(\mathbf{x}_{\mathbf{o}}) = (P_{\mathbf{o}}\mathbf{w}) \cdot \sum_{j=0}^{\gamma-1} (Q_{\mathbf{o},\mathbf{o}})^j \cdot (P_{\mathbf{o}}\mathbf{x})$$

Lemma 2. *Let $(\mathbf{x}, \mathbf{o})$ be a sample drawn according to a λ-regular distribution D with associated subspace E. Let $Q = P_E$ and $\|\mathbf{w}\| \leq 1$ then a.s:*

$$\|f^{\gamma}_{\mathbf{w},I-Q}(\mathbf{x}_{\mathbf{o}}) - f_{\mathbf{w},Q}(\mathbf{x}_{\mathbf{o}})\| \leq \frac{(1-\lambda)^{\gamma}}{\lambda}.$$

Corollary 1. *Let ℓ be a L-Lipschitz function. Under λ-regularity, for every $\gamma \geq \frac{\log L/\lambda\epsilon}{\lambda}$ the class $\mathcal{F}^{\gamma}$ contains an ϵ-good target function.*

4.2. Improper learning of $\mathcal{F}^{\gamma}$ and a kernel trick

Let $\mathbb{G}$ be the set of all finite, non empty, sequences of length at most γ over d. For each $\mathbf{s} \in \mathbb{G}$ denote $|\mathbf{s}|$– the length of the sequence and $\mathbf{s}_{\mathrm{end}}$ the last element of the sequence. Given a set of observations $\mathbf{o}$ we write $\mathbf{s} \subseteq \mathbf{o}$ if all elements of the sequence $\mathbf{s}$ belong to $\mathbf{o}$. We let

$$\Gamma = \sum_{j=1}^{\gamma} d^j = |\mathbb{G}| = \frac{d^{\gamma+1} - d}{d-1}$$

and we index the coordinates of $\mathbb{R}^{\Gamma}$ by the elements of $\mathbb{G}$:

Definition 4. *We let $\phi_{\gamma} : (\mathbb{R}^d \times \mathcal{O}) \to \mathbb{R}^{\Gamma}$ be the embedding:*

$$(\phi_{\gamma}(\mathbf{x}_{\mathbf{o}}))_{\mathbf{s}} = \begin{cases} \mathbf{x}_{\mathbf{s}_{\mathrm{end}}} & \mathbf{s} \subseteq \mathbf{o} \\ 0 & \textit{else} \end{cases}$$

Lemma 3. *For every Q and $\mathbf{w}$ we have:*

$$f^{\gamma}_{\mathbf{w},Q}(\mathbf{x}_{\mathbf{o}}) = \sum_{\mathbf{s}_1 \in \mathbf{o}} \mathbf{w}_{\mathbf{s}_1}\mathbf{x}_{\mathbf{s}_1} +$$

$$\sum_{\{\mathbf{s}:\mathbf{s}\subseteq\mathbf{o},\ 2\leq|\mathbf{s}|\leq t\}} \mathbf{w}_{\mathbf{s}_1} \cdot Q_{\mathbf{s}_1,\mathbf{s}_2} \cdot Q_{\mathbf{s}_2,\mathbf{s}_3} \cdots Q_{\mathbf{s}_{|\mathbf{s}|-1},\mathbf{s}_{\mathrm{end}}} \cdot \mathbf{x}_{\mathbf{s}_{\mathrm{end}}}$$

Corollary 2. *For every $f^{\gamma}_{\mathbf{w},Q} \in \mathcal{F}^{\gamma}$ there is $\mathbf{v} \in B_{\Gamma}(\Gamma)$, such that:*

$$f^{\gamma}_{\mathbf{w},Q}(\mathbf{x}_{\mathbf{o}}) = \mathbf{v} \cdot \phi_{\gamma}(\mathbf{x}_{\mathbf{o}}).$$

As a corollary, for every loss function ℓ and distribution D we have that:

$$\min_{\mathbf{v}\in B_{\Gamma}(\Gamma)} \mathbb{E}\left[\ell(\mathbf{v}\cdot\phi(\mathbf{x}_{\mathbf{o}}), y)\right] \leq \min_{f^{\gamma}_{\mathbf{w},Q}\in\mathcal{F}^{\gamma}} \mathbb{E}\left[\ell(f^{\gamma}_{\mathbf{w},Q}(\mathbf{x}_{\mathbf{o}}), y)\right]$$

Due to Corollary 2, learning $\mathcal{F}^{\gamma}$ can be improperly done via learning a linear classifier over the embedded sample set $\{\phi_{\gamma}(\mathbf{x}_{\mathbf{o}})\}_{i=1}^{m}$. The ambient space $\mathbb{R}^{\Gamma}$ may be very large. However, as we next show, the scalar product between two embedded points can be computed efficiently:

Theorem 4.

$$\phi_\gamma(\mathbf{x}^{(1)}_{\mathbf{o}_1}) \cdot \phi_\gamma(\mathbf{x}^{(2)}_{\mathbf{o}_2}) = \frac{|\mathbf{o}_1 \cap \mathbf{o}_2|^\gamma - 1}{|\mathbf{o}_1 \cap \mathbf{o}_2| - 1} \sum_{k \in \mathbf{o}_1 \cap \mathbf{o}_2} \mathbf{x}^{(1)}_k \mathbf{x}^{(2)}_k.$$

(We use the convention that $\frac{1^\gamma - 1}{1-1} = \lim_{x \to 1} \frac{x^\gamma - 1}{x - 1} = \gamma$*)*

5. Experiments

We've conducted both toy and real-data experiments to compare the performance of our new method vs. other existing methods: 0-**imputation**, **mcb**, **mc1** and **geom**. We've conducted experiments on binary classification tasks, multi-class classification and regression tasks. We begin by describing how each method was implemented for each scenario

- **karma**: We've applied our method in all experiments. We evaluated the loss with $\gamma = \{1, 2, 3, 4\}$ and $C = \{10^{-5}, 10^{-4}, \ldots, 10^4, 10^5\}$. We chose γ and λ using a holdout set. A constant feature was added to allow bias and the data was normalized. For binary classification we let ℓ be the Hinge loss, for multi-class we used the multiclass Hinge loss as described in (Crammer & Singer, 2002) and finally for regression tasks we used squared loss (which was also used at test time).
- 0-**imputation**: 0-**imputation** is simply reduced to $\gamma = 1$. A constant feature was added to allow bias and data was normalized so the mean of each feature was 0.
- **mcb/ mc1**: **mcb** and **mc1** are methods suggested by (Goldberg et al., 2010). They are inspired by a convex relaxation of Problem G and differ by the way a bias term is introduced.

  When applying **mcb** and **mc1** in binary classification tasks, we used the algorithm as suggested there. We ran their iterative algorithm until the incremental change was smaller than $1e^{-5}$ and used the logistic loss function to fit the entries. In multi-class tasks the label matrix was given by a $\{0, 1\}$ matrix with a 1 entry at the correct label. In regression tasks we've used the squared loss to fit the labels (as this is the loss we tested the results against).
- **geom**: Finally we've applied **geom** (Chechik et al., 2008). This is an algorithm that tries to maximize the margin of the classifier but the margin is computed with respect to the revealed entries subspace. We've applied this method only for binary classification as this method was designed specifically to correct the hinge loss methods. Again we used the iterative algorithm as suggested in there with 5 iterations. Regularization parameters were chosen using a holdout set.

5.1. Toy Examples

5.1.1. MCAR- Model

The first toy data we've experimented on was a regression task. We randomly picked $m = 10^3$ examples (normal distribution), in an $r = 10$ dimensional subspace embedded in a $d = 20$-dimensional linear space. We then randomly picked a classifier $\mathbf{w}$ and let $\hat{y}_i = \mathbf{w} \cdot \mathbf{x}^{(i)}$. The labels where normalized to have standard deviation 1 and mean 0 (i.e., $y_i = \frac{\hat{y}_i - \mathbb{E}_i[\hat{y}_i]}{\sqrt{\frac{1}{m}\sum_i (\hat{y}_i - \mathbb{E}_i[\hat{y}_i])^2}}$). We then randomly chose a subset of observed features using iid Bernoulli distribution. Each feature had probability $p = 0.1, 0.2, \ldots, 0.9, 1$ to be observed. The results appear in Figure 2: Square loss vs. fraction of observed features.

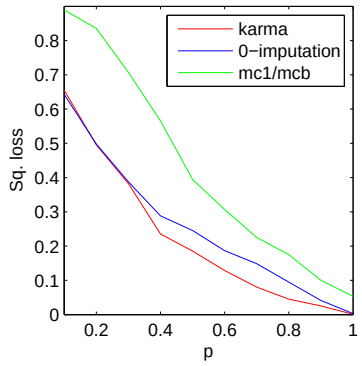

Figure 2. Missing Completely At Random features.

5.1.2. Toy Data with Large Margin

As discussed earlier, under large margin assumption our algorithm enjoys strong guarantees. For this we considered a different type of noise model – one that guarantees large margin. First we constructed fully observed sample of size $m = 10^4$: The data resides in $r = 20$ dimension subspace in a $d = 200$ dimension linear space. We then divided the features into three types- $\{A_1, A_2, A_3\}$. Each subset contained "enough information" and we made sure that for each subset A_i there is a classifier $\mathbf{w}^{A_i}$ with large margin that correctly classify $\mathbf{x}_\mathbf{o}$ when $\mathbf{o} = A_i$ (in fact a random division of the features led to this). Thus, if at each turn $\mathbf{o} = A_i$ for some i, the 0-**imputation** method would guarantee to perform well. However, in our noise model, for each sample, each type had probability $1/3$ to appear (independently, conditioned on the event that at least one type of feature appears). Thus there is no guarantee that 0-**imputation** will work well.

Such a noise distribution can model, for example, a scenario where we collect our data by letting people answer a survey. A person is likely to answer a certain type of question by and not answer a different type (say for example, he

is disinclined to answer questions on his financial matters but would not mind answering questions on his recreational preferences).

Denote:

$\chi(\mathbf{o}; A) = \begin{cases} 1 & A \subseteq \mathbf{o} \\ 0 & \text{else} \end{cases}$. One can show that the class $\{\mathbf{v} \cdot \phi_\gamma(\mathbf{x_o}) : \mathbf{v} \in \mathbb{R}^\Gamma\}$ can express the following target function whenever $\gamma \geq 3$:

$$h(\mathbf{x_o}) = \sum_i \chi(\mathbf{o}; A_i)\mathbf{w}^{A_i}\mathbf{x_o} -$$

$$\frac{1}{2}\sum_{i_1 \neq i_2} \chi(\mathbf{o}; A_{i_1} \cup A_{i_2})(\mathbf{w}^{A_{i_1}} + \mathbf{w}^{A_{i_2}}) \cdot \bar{\mathbf{x}} +$$

$$\chi(\mathbf{o}; A_1 \cup A_2 \cup A_3)(\mathbf{w}^{A_1} + \mathbf{w}^{A_2} + \mathbf{w}^{A_3}) \cdot \bar{\mathbf{x}}$$

Where $\bar{\mathbf{x}}$ is the vector received by filling zeros at non observed features. One can also verify that $h(\mathbf{x_o})$ will have zero expected loss. Also we can make sure that the corresponding $\mathbf{v}$ will have norm $\frac{7}{3}\sum \|\mathbf{w}^{A_i}\|$.

Our method, indeed, picks a good classifier with a comparably small sample size. Also, due to higher expressiveness it significantly outperforms 0-**imputation**.

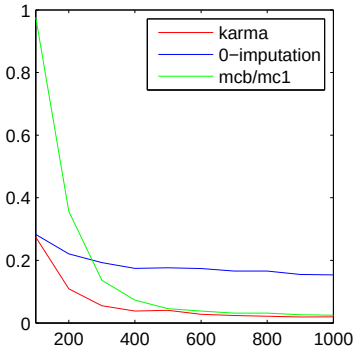

Figure 3. Block type revealed entries. Mean square error, Vs. sample size

5.2. Real Data

Finally, we tested our method on real data sets. We emphasize that the missing data was inherent in the real data. The comparison appears in Table 1. The data sets were collected primarily from (Alcal-Fdez et al.). The Jester data set was collected from (Goldberg et al.), The books dataset was collected from (Ziegler et al.) and we've also used the MovieLens data set[2]. We took from the Jester data set around 1000 users that rated at least 36 Jokes out of 100. One joke that was rated by almost all users was used as a label (specifically joke number 5). The movielens data set includes users who rated various movies, and includes met-data such as age, gender and occupation. We used the meta-data to construct three different tasks. In the appendix we add the details as to data set sizes and percentage of missing values in each task. Throughout regression tasks were normalized to have mean zero and standard deviation 1.

[2]available here: http://grouplens.org/datasets/movielens/

Table 1. Experiment Results

(a) Multiclass Labeling

name	**karma**	**0–imp**	**mcb**	**mc1**
Cleveland	0.44	**0.42**	0.48	0.42
dermatology	**0.03**	0.04	0.04	0.04
marketing	**0.70**	0.71	**0.70**	**0.70**
movielens(occupation)	**0.81**	0.87	0.86	0.87

(b) Regression

name	**karma**	**0–imp**	**mcb**	**mc1**
jester	**0.23**	0.24	0.27	0.27
books	0.25	0.25	0.25	0.25
movielens (age)	**0.16**	0.22	0.25	0.25

(c) Binary Labeling

name	**karma**	**0–imp**	**mcb**	**mc1**	**geom**
mammographic	**0.17**	**0.17**	**0.17**	0.18	**0.17**
bands	**0.24**	0.34	0.41	0.40	0.35
hepatitis	0.23	**0.17**	0.23	0.21	0.22
Wisconsin	**0.03**	**0.03**	**0.03**	0.04	0.04
horses	**0.35**	0.36	0.55	0.37	0.36
movielens (gender)	**0.22**	0.26	0.28	0.28	0.25

6. Discussion and future work

We have described the first theoretically-sound method to cope with low rank missing data, giving rise to a classification algorithm that attains competitive error to that of the optimal linear classifier that has access to all data. Our non-proper agnostic framework for learning a hidden low-rank subspace comes with provable guarantees, whereas heuristics based on separate data reconstruction and classification are shown to fail for certain scenarios.

Our technique is directly applicable to classification with low rank missing data and polynomial kernels via kernel (polynomial) composition. General kernels can be handled by polynomial approximation, but it is interesting to think about a more direct approach.

It is possible to derive all our results for a less stringent condition than λ-regularity: instead of bounding the smallest eigenvalue of the hidden subspace, it is possible to bound only the ratio of largest-to-smallest eigenvalue. This results

in better bounds in a straightforward plug-and-play into our analysis, but was omitted for simplicity.

Acknowledgements: EH: This research was supported in part by the European Research Council project SUBLRN and the Israel Science Foundation grant 810/11.

RL is a recipient of the Google Europe Fellowship in Learning Theory, and this research is supported in part by this Google Fellowship.

YM: This research was supported in part by The Israeli Centers of Research Excellence (I-CORE) program, (Center No. 4/11), by a grant from the Israel Science Foundation, by a grant from United States-Israel Binational Science Foundation (BSF).

References

Alcal-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., Garca, S., Snchez, L., and Herrera., F. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework.

Ben-David, S. and Dichterman, E. Learning with restricted focus of attention. *Journal of Computer and System Sciences*, 56(3):277–298, 1998.

Berthet, Q. and Rigollet, P. Complexity theoretic lower bounds for sparse principal component detection. *J. Mach. Learn. Res., W&CP*, 30:1046–1066 (electronic), 2013.

Candes, E. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, 2009.

Cesa-Bianchi, N., Shalev-Shwartz, S., and Shamir, O. Efficient learning with partially observed attributes. In *Proceedings of the 27th international conference on Machine learning*, 2010.

Cesa-Bianchi, N., Shalev-Shwartz, S., and Shamir, O. Online learning of noisy data. *IEEE Transactions on Information Theory*, 57(12):7907 –7931, dec. 2011. ISSN 0018-9448. doi: 10.1109/TIT.2011.2164053.

Chechik, Gal, Heitz, Geremy, Elidan, Gal, Abbeel, Pieter, and Koller, Daphne. Max-margin classification of data with absent features. *J. Mach. Learn. Res.*, 9:1–21, 2008. ISSN 1532-4435.

Crammer, Koby and Singer, Yoram. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2: 265–292, 2002.

Dekel, Ofer, Shamir, Ohad, and Xiao, Lin. Learning to classify with missing and corrupted features. *Mach. Learn.*, 81(2):149–178, November 2010. ISSN 0885-6125.

Eban, Elad, , Mezuman, Elad, and Globerson, Amir. Discrete chebyshev classifiers. 2014.

Globerson, Amir and Roweis, Sam. Nightmare at test time: robust learning by feature deletion. In *Proceedings of the 23rd international conference on Machine learning*, pp. 353–360. ACM, 2006.

Goldberg, Andrew B., Zhu, Xiaojin, Recht, Ben, Xu, Jun-Ming, and Nowak, Robert D. Transduction with matrix completion: Three birds with one stone. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems 2010.*, pp. 757–765, 2010.

Goldberg, Ken, Roeder, Theresa, Gupta, Dhruv, and Perkins., Chris. Eigentaste: A constant time collaborative filtering algorithm.

Hazan, Elad. *Introduction to Online Convex Optimization*. 2014. URL http://ocobook.cs.princeton.edu/.

Hazan, Elad and Koren, Tomer. Linear regression with limited observation. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

Hazan, Elad, Kale, Satyen, and Shalev-Shwartz, Shai. Near-optimal algorithms for online matrix prediction. In *COLT*, pp. 38.1–38.13, 2012.

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952.

Lee, J., Recht, B., Salakhutdinov, R., Srebro, N., and Tropp, J. A. Practical large-scale optimization for max-norm regularization. In *NIPS*, pp. 1297–1305, 2010.

Little, Roderick J. A. and Rubin, Donald B. *Statistical Analysis with Missing Data, 2nd Edition*. Wiley-Interscience, 2002.

Salakhutdinov, R. and Srebro, N. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *NIPS*, pp. 2056–2064, 2010.

Shalev-Shwartz, Shai, Singer, Yoram, Srebro, Nathan, and Cotter, Andrew. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

Shamir, O. and Shalev-Shwartz, S. Collaborative filtering with the trace norm: Learning, bounding, and transducing. *JMLR - Proceedings Track*, 19:661–678, 2011.

Srebro, Nathan. *Learning with Matrix Factorizations*. PhD thesis, Massachusetts Institute of Technology, 2004.

Srebro, Nathan, Rennie, Jason, and Jaakkola, Tommi S. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pp. 1329–1336, 2004.

Sridharan, Karthik, Shalev-Shwartz, Shai, and Srebro, Nathan. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems*, pp. 1545–1552, 2009.

Ziegler, Cai-Nicolas, McNee, Sean M., Konstan, Joseph A., and Lausen., Georg. Improving recommendation lists through topic diversification.

A. Proofs of theorems and lemmas from main text

A.1. Technical Claims

Claim 1. *Let $Q \in M_{d\times d}$ be a square projection matrix and $P \in M_{k\times d}$ a matrix. Recall that:*

$$\mathrm{Im}(A) = \{\mathbf{v} : \exists \mathbf{u}\ A\mathbf{u} = \mathbf{v}\}, \quad \text{and} \quad \ker(A) = \{\mathbf{v} : A\mathbf{v} = 0\}.$$

And that $\mathrm{rank}(A)$ *is the size of the largest collection of linearly independent columns of A.*

The following statements are equivalent:

1. $\ker(PQ) = \ker(Q)$.
2. $\mathrm{rank}(PQ) = \mathrm{rank}(QP^\top) = \mathrm{rank}(PQP^\top) = \mathrm{rank}(Q)$.
3. $\mathrm{Im}(QP^\top) = \mathrm{Im}(Q)$.

Proof.

1 ⇒ 2 Clearly $\mathrm{rank}(PQ) \le \mathrm{rank}(Q)$. If $\mathrm{rank}(PQ) < \mathrm{rank}(Q)$ we must have some collection of linearly independent columns of Q that are linearly dependent in PQ this implies that there is $\mathbf{v}$ such that $PQ\mathbf{v} = 0$ but $Q\mathbf{v} \neq 0$. Hence $\ker(PQ) \neq \ker(Q)$ and thus a contradiction, we conclude that $\mathrm{rank}(PQ) = \mathrm{rank}(Q)$.

That $\mathrm{rank}(PQ) = \mathrm{rank}(QP^\top) = \mathrm{rank}(PQP^\top)$ follows from the fact that $\mathrm{rank}(A) = \mathrm{rank}(A^\top) = \mathrm{rank}(AA^\top)$ and using the fact that $Q^2 = Q$ since Q is a projection matrix.

2⇒ 3 We have that $\mathrm{Im}(QP^\top) \subseteq \mathrm{Im}(Q)$. The two subspaces, $\mathrm{Im}(QP^\top)$ and $\mathrm{Im}(Q)$, are in fact the linear span of the columns of $QP^\top$ and Q respectively.

Since $\mathrm{rank}(QP^\top) = \mathrm{rank}(Q)$ we conclude that the dimension of the two subspaces is equal. It follows that $\mathrm{Im}(QP^\top) = \mathrm{Im}(Q)$.

3 ⇒ 1 Since $\mathrm{Im}(QP^\top) = \mathrm{Im}(Q)$ we also have $\mathrm{rank}(QP^\top) = \mathrm{rank}(Q)$ and as a corollary $\mathrm{rank}(PQ) = \mathrm{rank}(Q)$.

Now by the rank-nullity Theorem, for every $A \in M_{k\times d}$, $\dim(\ker(A)) = d - \mathrm{rank}(A)$.

Hence $\dim(\ker(PQ)) = \dim(\ker(Q))$. Since $\ker(PQ) \subseteq \ker(Q)$ we must have . $\ker(PQ) = \ker(Q)$.

□

Claim 2. *Let $\mathbf{o} \in 2^d$ be drawn according to a distribution D that satisfies the low rank assumption. If $Q = P_E$ then:*

$$\mathrm{Im}(Q_{\mathbf{o},\mathbf{o}}) = \mathrm{Im}(P_{\mathbf{o}}Q)$$

Proof. $\ker(P_{\mathbf{o}}Q) = \ker(Q)$ holds by assumption (assumption 3 in Definition 1). $\mathrm{Im}(Q) = \mathrm{Im}(QP_{\mathbf{o}}^\top)$ then follows from item 3. In particular $\mathrm{Im}(P_{\mathbf{o}}Q) = \mathrm{Im}(P_{\mathbf{o}}QP_{\mathbf{o}}{}^\top) = \mathrm{Im}(Q_{\mathbf{o},\mathbf{o}})$. □

A.2. proof of Lemma 1

By definition, if $P_{\mathbf{o}}\mathbf{x} \in \mathrm{Im}(Q_{\mathbf{o},\mathbf{o}})$ then $Q_{\mathbf{o},\mathbf{o}}\,(Q_{\mathbf{o},\mathbf{o}})^\dagger\, P_{\mathbf{o}}\mathbf{x} = P_{\mathbf{o}}\mathbf{x}$. We claim that due to the low rank assumption, $P_{\mathbf{o}}\mathbf{x} \in \mathrm{Im}(Q_{\mathbf{o},\mathbf{o}})$.

Indeed, recall that $Q = P_E$ and $\mathbf{x} \in E$ hence $Q\mathbf{x} = \mathbf{x}$ and $P_{\mathbf{o}}\mathbf{x} \in \mathrm{Im}(P_{\mathbf{o}}Q)$. By Claim 2 we have $\mathrm{Im}(Q_{\mathbf{o},\mathbf{o}}) = \mathrm{Im}(P_{\mathbf{o}}Q)$, hence $P_{\mathbf{o}}\mathbf{x} \in (\mathrm{Im}Q_{\mathbf{o},\mathbf{o}})$.

Next, we have that

$$P_{\mathbf{o}}QP_{\mathbf{o}}^\top\,(Q_{\mathbf{o},\mathbf{o}})^\dagger\, P_{\mathbf{o}}\mathbf{x} = Q_{\mathbf{o},\mathbf{o}}\,(Q_{\mathbf{o},\mathbf{o}})^\dagger\, P_{\mathbf{o}}\mathbf{x} = P_{\mathbf{o}}\mathbf{x}$$

Alternatively

$$P_{\mathbf{o}}(QP_{\mathbf{o}}^{\top}Q_{\mathbf{o},\mathbf{o}}^{\dagger}P_{\mathbf{o}}\mathbf{x}-\mathbf{x})=0. \tag{4}$$

Again, since $Q\mathbf{x}=\mathbf{x}$ we have that:

$$P_{\mathbf{o}}Q(P_{\mathbf{o}}^{\top}Q_{\mathbf{o},\mathbf{o}}^{\dagger}P_{\mathbf{o}}\mathbf{x}-\mathbf{x})=0. \tag{5}$$

The low rank assumption implies that $P_{\mathbf{o}}Q\mathbf{v}=0$ if and only if $Q\mathbf{v}=0$. Apply this to $\mathbf{v}=P_{\mathbf{o}}^{\top}Q_{\mathbf{o}}^{\dagger}P_{\mathbf{o}}\mathbf{x}-\mathbf{x}$ and get:

$$QP_{\mathbf{o}}^{\top}Q_{\mathbf{o},\mathbf{o}}^{\dagger}P_{\mathbf{o}}\mathbf{x}=Q\mathbf{x}=\mathbf{x}.$$

Finally we have that

$$f_{\mathbf{w},Q}(\mathbf{x}_{\mathbf{o}})=(P_{\mathbf{o}}Q^{\top}\mathbf{w}^{*})\cdot Q_{\mathbf{o},\mathbf{o}}^{\dagger}P_{\mathbf{o}}\mathbf{x}=\mathbf{w}^{*}\cdot QP_{\mathbf{o}}^{\top}Q_{\mathbf{o},\mathbf{o}}^{\dagger}P_{\mathbf{o}}\mathbf{x}=\mathbf{w}^{*}\cdot\mathbf{x}.$$

A.3. proof of Lemma 2

Let I denote the identity matrix in $M_{d\times d}$. First note that $(I_{\mathbf{o},\mathbf{o}}-Q_{\mathbf{o},\mathbf{o}})=(I-Q)_{\mathbf{o},\mathbf{o}}$ and that $I_{\mathbf{o},\mathbf{o}}$ is the identity matrix in $\mathbb{R}^{|\mathbf{o}|\times|\mathbf{o}|}$.

Let $\mathbf{v}_1,\ldots,\mathbf{v}_k$ be the normalized and orthogonal eigen-vectors of $Q_{\mathbf{o},\mathbf{o}}$ with strictly positive eigenvalues $\lambda_1\geq\ldots,\geq\lambda_k$. By λ-regularity we have that $\lambda_k\geq\lambda$ and since the spectral norm of $Q_{\mathbf{o},\mathbf{o}}$ is smaller than the spectral norm of Q we have that $\lambda_1\leq 1$.

Note that for every $\mathbf{v}_j$ we have $Q_{\mathbf{o},\mathbf{o}}^{\dagger}\mathbf{v}_j=\frac{1}{\lambda_j}\mathbf{v}_j$. Next, recall that $Q=P_E$ and $\mathbf{x}\in E$ hence $Q\mathbf{x}=\mathbf{x}$ and $P_{\mathbf{o}}\mathbf{x}\in\mathrm{Im}(P_{\mathbf{o}}Q)$. By Claim 2 we have $P_{\mathbf{o}}\mathbf{x}\in\mathrm{Im}(Q_{\mathbf{o},\mathbf{o}})$. Since $\mathrm{Im}(Q_{\mathbf{o},\mathbf{o}})=\mathrm{span}(\mathbf{v}_1,\ldots,\mathbf{v}_k)$, we may write $P_{\mathbf{o}}\mathbf{x}=\sum\alpha_i\mathbf{v}_i$. Since $\|P_{\mathbf{o}}\mathbf{x}\|\leq 1$ and $\{\mathbf{v}_1,\ldots,\mathbf{v}_k\}$ is an orthonormal system we have $\sum\alpha_i^2\leq 1$.

Hence

$$\|\left(\sum_{j=0}^{\gamma-1}(I_{\mathbf{o},\mathbf{o}}-Q_{\mathbf{o},\mathbf{o}})^{j}-Q_{\mathbf{o},\mathbf{o}}^{\dagger}\right)P_{\mathbf{o}}\mathbf{x}\|=\|\sum_{i}\alpha_i\left(\sum_{j=0}^{\gamma-1}(1-\lambda)_i^{j}-\frac{1}{\lambda_i}\right)\mathbf{v}_i\|\leq\max_i\left|\sum_{j=0}^{\gamma-1}(1-\lambda_i)^{j}-\frac{1}{\lambda_i}\right|\leq$$

$$\max_i\left|\frac{1-(1-\lambda_i)^{\gamma}}{\lambda_i}-\frac{1}{\lambda_i}\right|\leq\frac{(1-\lambda)^{\gamma}}{\lambda}.$$

Finally since $\|P_{\mathbf{o}}\mathbf{w}\|\leq 1$ we get that

$$\|f_{\mathbf{w},I-Q}^{\gamma}(\mathbf{x}_{\mathbf{o}})-f_{\mathbf{w},Q}(\mathbf{x}_{\mathbf{o}})\|\leq\frac{(1-\lambda)^{\gamma}}{\lambda}$$

A.4. Proof of Lemma 3

Let $\mathbf{o}_1\leq\mathbf{o}_2,\leq\ldots\leq\mathbf{o}_{|\mathbf{o}|}$ be the elements of $\mathbf{o}$ ordered in increasing order. First by definition we have that:

$$f_{\mathbf{w},Q}^{\gamma}(\mathbf{x}_{\mathbf{o}})=\sum_{j=0}^{\gamma-1}\sum_{n,k=1}^{|\mathbf{o}|}\mathbf{w}_{\mathbf{o}_n}((Q_{\mathbf{o},\mathbf{o}})^{j})_{n,k}\mathbf{x}_{\mathbf{o}_k}=\sum_{i\in\mathbf{o}}\mathbf{w}_i\mathbf{x}_i+\sum_{j=1}^{\gamma-1}\sum_{n,k=1}^{|\mathbf{o}|}\mathbf{w}_{\mathbf{o}_n}((Q_{\mathbf{o},\mathbf{o}})^{j})_{n,k}\mathbf{x}_{\mathbf{o}_k} \tag{6}$$

We also have by definition that for $j\geq 1$:

$$((Q_{\mathbf{o},\mathbf{o}})^{j})_{n,k}=\sum_{s=1}^{|\mathbf{o}|}((Q_{\mathbf{o},\mathbf{o}})^{j-1})_{n,s}((Q_{\mathbf{o},\mathbf{o}}))_{s,k}=\sum_{s=1}^{|\mathbf{o}|}((Q_{\mathbf{o},\mathbf{o}})^{j-1})_{n,s}Q_{\mathbf{o}_s,\mathbf{o}_k}$$

By induction we can show that:

$$((Q_{\mathbf{o},\mathbf{o}})^j)_{n,k} = \sum_{\mathbf{s}_1 \in \mathbf{o}} Q_{\mathbf{o}_n,\mathbf{s}_1} \left(\sum_{\mathbf{s}_2 \in \mathbf{o}} Q_{\mathbf{s}_1,\mathbf{s}_2} \left(\sum \cdots \left(\sum_{\mathbf{s}_{j-1} \in \mathbf{o}} Q_{\mathbf{s}_{j-2},\mathbf{s}_{j-1}} Q_{\mathbf{s}_{j-1},\mathbf{o}_k} \right) \cdots \right) \right).$$

Reordering the elements we get for $j \geq 1$:

$$((Q_{\mathbf{o},\mathbf{o}})^j)_{n,k} = \sum_{\{\mathbf{s}:|\mathbf{s}|=j+1, \mathbf{s}_1=\mathbf{o}_n, \mathbf{s}_{j+1}=\mathbf{o}_k\}} Q_{\mathbf{s}_1,\mathbf{s}_2} \cdot Q_{\mathbf{s}_2,\mathbf{s}_3} \cdots Q_{\mathbf{s}_j,\mathbf{s}_{j+1}} \tag{7}$$

The result now follows from Eq. 6 and Eq. 7 by a change of indexes.

A.5. Proof of Corollary 2

Choose

$$\mathbf{v}_{\mathbf{s}} = \begin{cases} \mathbf{w}_{\mathbf{s}_1} & |\mathbf{s}| = 1 \\ \mathbf{w}_{\mathbf{s}_1} \cdot Q_{\mathbf{s}_1,\mathbf{s}_2} \cdot Q_{\mathbf{s}_2,\mathbf{s}_3} \cdots Q_{\mathbf{s}_{|\mathbf{s}|-1},\mathbf{s}_{\text{end}}} & |\mathbf{s}| > 1 \end{cases}$$

It is clear from Lemma 3 that $f^{\gamma}_{\mathbf{w},Q}(\mathbf{x}_{\mathbf{o}}) = \mathbf{v} \cdot \phi_{\gamma}(\mathbf{x}_{\mathbf{o}})$. We only need to show that $\|\mathbf{v}\| \leq \sqrt{\Gamma}\|\mathbf{w}\|$.

Note that since $Q^2 = Q$ we have $\max(|Q_{i,j}|) < 1$. Hence $|\mathbf{v}_{\mathbf{s}}| \leq |\mathbf{w}_{\mathbf{s}_1}|$ and:

$$\|\mathbf{v}\|^2 = \sum_{\mathbf{s} \in \mathbb{G}} \mathbf{v}_{\mathbf{s}}^2 \leq \sum_{\mathbf{s} \in \mathbb{G}} \mathbf{w}_{\mathbf{s}_1}^2 \leq \Gamma \|\mathbf{w}\|^2$$

A.6. Proof of Theorem 4

By definition of ϕ_{γ} we have:

$$\phi_{\gamma}(\mathbf{x}^{(1)}_{\mathbf{o}_1}) \cdot \phi_{\gamma}(\mathbf{x}^{(2)}_{\mathbf{o}_2}) = \sum_{\mathbf{s} \subseteq \mathbf{o}_1 \cap \mathbf{o}_2} \mathbf{x}^{(1)}_{\mathbf{s}_{\text{end}}} \cdot \mathbf{x}^{(2)}_{\mathbf{s}_{\text{end}}} = \sum_{l=1}^{\gamma} \sum_{k \in \mathbf{o}_1 \cap \mathbf{o}_2} \sum_{\mathbf{s} \subseteq \mathbf{o}_1 \cap \mathbf{o}_2, \mathbf{s}_{\text{end}}=k, |\mathbf{s}|=l} \mathbf{x}^{(1)}_k \cdot \mathbf{x}^{(2)}_k$$

$$= \sum_{l=1}^{1} \sum_{|\mathbf{s}|=l-1, \mathbf{s} \subset \mathbf{o}_1 \cap \mathbf{o}_2} \sum_{k \in \mathbf{o}_1 \cap \mathbf{o}_2} \mathbf{x}^{(1)}_k \cdot \mathbf{x}^{(2)}_k$$

$$= \sum_{l=1}^{1} |\mathbf{s} : |\{|\mathbf{s}| = l-1,\ \mathbf{s} \subset \mathbf{o}_1 \cap \mathbf{o}_2\}| \sum_{k \in \mathbf{o}_1 \cap \mathbf{o}_2} \mathbf{x}^{(1)}_k \cdot \mathbf{x}^{(2)}_k = \sum_{l=1}^{\gamma} |\mathbf{o}_1 \cap \mathbf{o}_2|^{l-1} \cdot \sum_{k \in \mathbf{o}_1 \cap \mathbf{o}_2} \mathbf{x}^{(1)}_k \cdot \mathbf{x}^{(2)}_k =$$

$$\frac{1 - |\mathbf{o}_1 \cap \mathbf{o}_2|^{\gamma}}{1 - |\mathbf{o}_1 \cap \mathbf{o}_2|} \sum_{k \in \mathbf{o}_1 \cap \mathbf{o}_2} \mathbf{x}^{(1)}_k \mathbf{x}^{(2)}_k$$

A.7. Proof of Theorem 2

We take ϕ_{γ} as in Definition 4. That $\phi_{\gamma}(\mathbf{x}^1_{\mathbf{o}^1}) \cdot \phi_{\gamma}(\mathbf{x}^2_{\mathbf{o}^2}) = \frac{1-|\mathbf{o}^{(1)} \cap \mathbf{o}^{(2)}|^{\gamma}}{1-|\mathbf{o}^{(1)} \cap \mathbf{o}^{(2)}|} \sum_{i \in \mathbf{o}^{(1)} \cap \mathbf{o}^{(2)}} \mathbf{x}^{(1)}_i \cdot \mathbf{x}^{(2)}_i$ is shown in Theorem 4.

The analysis of sub-gradient descent methods to optimize problems of the form:

$$\frac{\rho}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m}(\ell(\mathbf{w}^{\top}\phi(\mathbf{x}_i), y_i)$$

was studied in (Shalev-Shwartz et al., 2011) and the detailed analysis can be found there (with generalization to mercer kernels and general losses). We mention that since ℓ is L-Lipschitz and $\|\phi_{\gamma}(\mathbf{x}_{\mathbf{o}})\| \leq \sqrt{\Gamma}$ a bound on the gradient of $\nabla \ell(\mathbf{v}^{\top}\phi_{\gamma}(\mathbf{x}_{\mathbf{o}}), y) = \ell'(\mathbf{v}^{\top}\phi_{\gamma}(\mathbf{x}_{\mathbf{o}}), y)\phi_{\gamma}(\mathbf{x}_{\mathbf{o}})$ is given by $L\sqrt{\Gamma}$.

This establishes items 1 and 2.

Next we let ℓ be an L-Lipschitz loss function and D a λ-regular distribution and we assume that $\gamma \geq \frac{\log 2L/(\lambda\epsilon)}{\lambda}$.

Due to Corollary 2, for some $\mathbf{v}^* \in B_\Gamma(\Gamma)$

$$\mathbb{E}\left[\ell(\mathbf{v}^* \cdot \phi_\gamma(\mathbf{x_o}), y)\right] \leq \min_{f^\gamma_{\mathbf{w},I-Q} \in \mathcal{F}^\gamma} \mathbb{E}\left[\ell(f^\gamma_{\mathbf{w},I-Q}(\mathbf{x_o}), y)\right]$$

Applying Lemma 2 and L-Lipschitness, for every $f^*_{\mathbf{w},Q} \in \mathcal{F}_0$ we have:

$$\mathbb{E}\left[\ell(\mathbf{v}^* \cdot \phi_\gamma(\mathbf{x_o}), y)\right] \leq \mathbb{E}\left[\ell(f^*_{\mathbf{w},Q}(\mathbf{x_o}, y))\right] + L\frac{(1-\lambda)^\gamma}{\lambda}.$$

The result follows Lemma 1 and choice of γ:

$$\frac{(1-\lambda)^{\frac{\log 2L/(\lambda\epsilon)}{\lambda}}}{\lambda} \leq \frac{(1-\lambda)^{\frac{\log(\lambda\epsilon)/(2L)}{\log(1-\lambda)}}}{\lambda} = \frac{\epsilon}{2L}.$$

A.8. Proof of Theorem 1

Fix a sample $S = \{\mathbf{x}^i_{\mathbf{o}^i}\}_{i=1}^m$ and $\gamma \geq \frac{\log 2L/\lambda\epsilon}{\lambda}$. Let

$$\mathcal{L}(\mathbf{v}) = \mathbb{E}(\ell(\mathbf{v}^\top \phi_\gamma(\mathbf{x_o}), y) \quad \hat{\mathcal{L}}(\mathbf{v}) = \frac{1}{m}\sum_{i=1}^m \ell(\mathbf{v}^\top \phi_\gamma(\mathbf{x}^i_{\mathbf{o}^i}),$$

the expected and empirical losses of the vector $\mathbf{v}$.

Further denote by

$$F_\rho(\mathbf{v}) = \frac{\rho}{2}\|\mathbf{v}\|^2 + \mathcal{L}(\mathbf{v}) \quad \hat{F}_\rho(\mathbf{v}) = \frac{\rho}{2}\|\mathbf{v}\|^2 + \hat{\mathcal{L}}(\mathbf{v})$$

Set $\rho(m) \in O\left(\sqrt{\frac{\log 1/\delta}{m}}\right)$. Run Alg. 1 with $T = m$ and let $\bar{\mathbf{v}} = \frac{1}{T}\sum_{i=1}^T \mathbf{v}_t$. By Theorem 2, item 2 we get:

$$\hat{F}_{\rho(m)}(\bar{\mathbf{v}}) \leq \min \hat{F}_{\rho(m)}(\mathbf{v}) + O\left(\frac{L^2\Gamma(\epsilon)}{\rho m}\right)$$

Note that $\|\phi_\gamma(\mathbf{x_o})\| \leq \sqrt{\frac{d^\gamma - 1}{d-1}}\|P_{\mathbf{o}}\mathbf{x}\| \leq \sqrt{\Gamma(\epsilon)}$. We now apply Corollary 4. in (Sridharan et al., 2009) with $B = \sqrt{\Gamma(\epsilon)}$ to obtain the following bound (with probability $1-\delta$) for every $\mathbf{w}$:

$$\mathcal{L}(\bar{\mathbf{v}}) \leq \mathcal{L}(\mathbf{w}) + O\left(\sqrt{\frac{L^2\Gamma(\epsilon)\|\mathbf{w}\|^2\log(1/\delta)}{m}}\right)$$

In particular for every $\|\mathbf{w}\| \leq \sqrt{\Gamma(\epsilon)}$ we have

$$\mathcal{L}(\bar{\mathbf{v}}) \leq \mathcal{L}(\mathbf{w}) + O\left(\sqrt{\frac{L^2\Gamma(\epsilon)^2\log(1/\delta)}{m}}\right).$$

From Theorem 2, item 3 we have that for some $\|\mathbf{w}\| \leq \sqrt{\Gamma(\epsilon)}$:

$$\mathcal{L}(\mathbf{w}) \leq \min_{\|\mathbf{w}\|\leq 1} \mathbb{E}(\ell(\mathbf{w}^\top \mathbf{x}, y) + \epsilon.$$

The result now follows from the choice of m.

A.9. Proof of Theorem 3

Before proving the theorem, we formally define the sequences for which the algorithm applies: a λ-regular sequence is one such that the uniform distribution over the sequence elements is λ-regular with associated subspace E.

Proof of Theorem 3. Let E^* denote the adversarially chosen subspace and Q^* The projection associated with it. Since the sequence $\{(\mathbf{x}^t, \mathbf{o}^t, y_t)$ is λ-regular w.r.t. subspace E^*, we have by Lemma 2,

$$\forall \|\mathbf{w}\| \leq 1 \,.\, \|f^{\gamma}_{\mathbf{w},I-Q^*}(\mathbf{x_o}) - f_{\mathbf{w},Q^*}(\mathbf{x_o})\| \leq \frac{(1-\lambda)^{\gamma}}{\lambda} \leq \frac{1}{\lambda} e^{-\lambda\gamma}$$

Thus, taking $f^{\gamma}_{\mathbf{w}^*,I-Q^*} \in \mathcal{F}^{\gamma}$ we have

$$\begin{aligned}
&\min_{\mathbf{w}\in B_d} \textstyle\sum_t \ell(f_{\mathbf{w},Q^*}(\mathbf{x}^t_{\mathbf{o}_t}), y_t) - \sum_t \ell(f^{\gamma}_{\mathbf{w}^*,I-Q^*}(\mathbf{x}^t_{\mathbf{o}_t}), y_t) \\
&= \textstyle\sum_t \ell(f^*_{\mathbf{w},Q}(\mathbf{x}^t_{\mathbf{o}_t}), y_t) - \sum_t \ell(f^{\gamma}_{\mathbf{w}^*,I-Q^*}(\mathbf{x}^t_{\mathbf{o}_t}), y_t) \\
&\leq \textstyle\sum_t L\|f^*_{\mathbf{w},Q}(\mathbf{x}^t_{\mathbf{o}_t}) - f^{\gamma,*}_{\mathbf{w},Q}(\mathbf{x}^t_{\mathbf{o}_t})\| && \ell \text{ is } L\text{-Lipschitz} \\
&\leq TL\tfrac{1}{\lambda} e^{-\lambda\gamma} && \text{Lemma 2}
\end{aligned}$$

Hence it suffices to show that

$$\begin{aligned}
&\textstyle\sum_t \ell(\mathbf{v}_t^{\top}\phi_{\gamma}(\mathbf{x}^t_{\mathbf{o}_t}), y_t) - \sum_t \ell(f^{\gamma}_{\mathbf{w}^*,I-Q^*}(\mathbf{x}^t_{\mathbf{o}_t}), y_t) \\
&\leq \textstyle\sum_t \ell(\mathbf{v}_t^{\top}\phi_{\gamma}(\mathbf{x}^t_{\mathbf{o}_t}), y_t) - \min_{f_{\mathbf{w},Q}\in\mathcal{F}^{\gamma}} \sum_t \ell(f^{\gamma}_{\mathbf{w},Q}(\mathbf{x}^t_{\mathbf{o}_t}), y_t) = O(\sqrt{T})
\end{aligned}$$

Corollary 2 asserts that

$$f^{\gamma}_{\mathbf{w},Q}(\mathbf{x_o}) = \mathbf{v} \cdot \phi_{\gamma}(\mathbf{x_o})$$

Thus, the theorem statement can be further reduced to

$$\sum_t \ell(\mathbf{v}_t^{\top}\phi_{\gamma}(\mathbf{x}^t_{\mathbf{o}_t}), y_t) - \min_{\mathbf{v}_*\in B_{\Gamma}(\Gamma)} \sum_t \ell(\mathbf{v}_*^{\top}\phi_{\gamma}(\mathbf{x}^t_{\mathbf{o}_t}), y_t) = O(\sqrt{T}) \tag{8}$$

We proceed to prove equation Eq. 8 above.

Algorithm 1 applies the following update rule

$$\mathbf{v}_{t+1} = \sum_{i=1}^{t} \alpha_i^{(t)} \phi_{\gamma}(\mathbf{x}^i_{\mathbf{o}^i})$$

where $\mathbf{w}_{t+1}$ can be re-written as:

$$\begin{aligned}
\mathbf{v}_{t+1} \quad &= (1-\eta_t\rho)\mathbf{v}_t - \eta_t \ell'(\mathbf{v}_t^{\top}\phi_{\gamma}(\mathbf{x}^t_{\mathbf{o}_t}))\phi_{\gamma}(\mathbf{x}^t_{\mathbf{o}^t}) \\
&= \mathbf{v}_t - \eta_t \nabla \tilde{\ell}_t(\mathbf{v}_t)
\end{aligned} \tag{9}$$

where

$$\tilde{\ell}_t(\mathbf{v}) = \ell(\mathbf{v}^{\top}\phi_{\gamma}(\mathbf{x}^t_{\mathbf{o}_t})) + \frac{\rho}{2}\|\mathbf{v}\|^2$$

The above implies a bound on the norm of the gradients of $\tilde{\ell}_t$, as given by the following lemma:

Lemma 4. *For all iterations $t \in [T]$ we have*

$$\|\mathbf{v}_t\| \leq LX\sqrt{\Gamma}\,,\ \|\nabla\tilde{\ell}_t(\mathbf{v}_t)\| \leq 2LX\sqrt{\Gamma}$$

Equation Eq. 9 implies that KARMA applies the online gradient descent algorithm on the functions $\tilde{\ell}$ which are ρ-strongly-convex. Hence, the bound of Theorem 3.3 in (Hazan, 2014), with appropriate learning rates η_t and with $\alpha = \rho$, $G = 2LX\sqrt{\Gamma}$) by lemma 4, gives

$$\sum_t \tilde{\ell}_t(\mathbf{v}_t) - \min_{\mathbf{v}^*} \sum_t \tilde{\ell}_t(\mathbf{v}^*) \leq \frac{2L^2X^2\Gamma}{\rho}(1+\log T)$$

This directly implies our theorem since (recall that $\|\mathbf{v}^*\| \leq B$ by assumption):

$$\begin{aligned}
&\textstyle\sum_t \ell(\mathbf{v}_t^\top \phi_\gamma(\mathbf{x}^t_{\mathbf{o}_t}), y_t) - \min_{\|\mathbf{w}\|\leq 1} \sum_t \ell(f_{\mathbf{w},Q^*}(\mathbf{x}^t_{\mathbf{o}_t}), y_t) \\
&= \textstyle\sum_t \tilde{\ell}_t(\mathbf{v}_t) - \min_{\mathbf{v}^*} \sum_t \tilde{\ell}_t(\mathbf{v}^*) + \frac{\rho}{2}(\sum_t \|\mathbf{v}^*\|^2 - \|\mathbf{v}_t\|^2) \\
&\leq \tfrac{2L^2X^2\Gamma}{\rho}(1+\log T) + \tfrac{\rho}{2}T \cdot B
\end{aligned}$$

□

Proof of Lemma 4. First, notice that the norms of the gradients of the loss functions ℓ can be bounded by

$$\|\nabla \ell(\mathbf{v}_t^\top \phi_\gamma(\mathbf{x}^t_{\mathbf{o}_t}), y_t)\| = |\ell'(\mathbf{v}_t^\top \phi_\gamma(\mathbf{x}^t_{\mathbf{o}^t}), y_t)| \cdot \|\phi_\gamma(\mathbf{x}^t_{\mathbf{o}^t})\| \leq LX\sqrt{\Gamma}$$

where the last inequality follows from the Lipschitz property of ℓ and the fact that $\phi_\gamma(\mathbf{x}^t_{\mathbf{o}^t})$ is a vector in $\mathbb{R}^\Gamma$, with coordinates from the vector $\mathbf{x}^t$, and the bound $\|\mathbf{x}^t\|_\infty \leq X$.

Next, we prove by induction that $\|\mathbf{v}_t\| \leq LX\sqrt{\Gamma}$. For $t = 0$ we have $\mathbf{v}_1 = 0$. Equation Eq. 9 implies that $\mathbf{v}_{t+1}$ is a convex combination of two vectors:

$$\begin{aligned}
\|\mathbf{v}_{t+1}\| &= \|(1-\eta_t\rho)\mathbf{v}_t - \eta_t \ell'(\mathbf{v}_t^\top \phi_\gamma(\mathbf{x}^t_{\mathbf{o}_t}))\phi_\gamma(\mathbf{x}^t_{\mathbf{o}^t})\| \\
&\leq \max\left\{\rho\|\mathbf{v}_t\|\,,\ \|\nabla \ell(\mathbf{v}_t^\top \phi_\gamma(\mathbf{x}^t_{\mathbf{o}_t}))\|\right\} \\
&\leq \max\left\{\rho LX\sqrt{\Gamma}\,,\ \|\nabla \ell(\mathbf{v}_t^\top \phi_\gamma(\mathbf{x}^t_{\mathbf{o}_t}))\|\right\} && \text{induction hypothesis} \\
&\leq \max\left\{\rho LX\sqrt{\Gamma}\,,\ LX\sqrt{\Gamma}\right\} && \text{above bound on } \nabla\ell \\
&\leq LX\sqrt{\Gamma} && \rho < 1
\end{aligned}$$

We can now conclude with the lemma, by definition of $\tilde{\ell}_t$

$$\|\nabla \tilde{\ell}_t(\mathbf{v}_t)\| \leq \|\nabla \ell(\mathbf{v}_t^\top \phi_\gamma(\mathbf{x}^t_{\mathbf{o}_t}))\| + \frac{\rho}{2}\|\mathbf{v}_t\| \leq LX\sqrt{\Gamma} + \frac{\rho}{2}LX\sqrt{\Gamma} \leq 2LX\sqrt{\Gamma}$$

□

B. Experiments – Data size and details

Data Set	Dimesnion	no. Examples	Label (Binary/Real/Multiclass)	percentage of Missing Values
mamographic	5	961	B	13.63
bands	19	539	B	32.3
clevland	13	303	$M(5)$	1.98
dermatology	34	366	$M(6)$	2.19
hepatitis	19	155	B	48.39
marketing	13	8993	$M(9)$	23.54
wisconsin	9	699	B	2.29
jester	94	1000	R	29
movieLens(Age)	3952	6040	R	95.8
movieLens(Gender)	3952	6040	B	95.8
movieLens(Occupation)	3952	6040	$M(21)$	95.8
Books	2494	19651	R	99.8
horses	23	368	B	22.7