Complexity of Vector-Valued Prediction: From Linear Models to Stochastic Convex Optimization

Matan Schliserman[1*] **Tomer Koren**[12]

[1]Blavatnik School of Computer Science, Tel Aviv University
[2]Google Research, Tel Aviv

Abstract

We study the problem of learning vector-valued linear predictors: these are prediction rules parameterized by a matrix that maps an m-dimensional feature vector to a k-dimensional target. We focus on the fundamental case with a convex and Lipschitz loss function, and show several new theoretical results that shed light on the complexity of this problem and its connection to related learning models. First, we give a tight characterization of the sample complexity of Empirical Risk Minimization (ERM) in this setting, establishing that $\widetilde{\Omega}(k/\varepsilon^2)$ examples are necessary for ERM to reach ε excess (population) risk; this provides for an exponential improvement over recent results by Magen and Shamir (2023) in terms of the dependence on the target dimension k, and matches a classical upper bound due to Maurer (2016). Second, we present a black-box conversion from general d-dimensional Stochastic Convex Optimization (SCO) to vector-valued linear prediction, showing that any SCO problem can be embedded as a prediction problem with $k = \Theta(d)$ outputs. These results portray the setting of vector-valued linear prediction as bridging between two extensively studied yet disparate learning models: linear models (corresponds to $k = 1$) and general d-dimensional SCO (with $k = \Theta(d)$).

1 Introduction

Prediction problems, such as classification and regression, lie at the core of both practical applications and theoretical research in machine learning. Within this framework, learning vector-valued predictors (VVPs), characterized by functions of the form:

$$x \;\to\; \ell(Wx)\,,$$

mapping vectors $x \in \mathbb{R}^m$ through a linear transformation parameterized by a matrix $W \in \mathbb{R}^{k\times m}$ followed by a loss function $\ell : \mathbb{R}^k \mapsto \mathbb{R}$, constitutes a rich learning framework that captures a wide range of problems in machine learning, from classical to modern. For instance, the scenario where $k = 1$ corresponds to the extensively studied domain of generalized linear models (e.g., Bartlett and Mendelson, 2002; Shalev-Shwartz and Ben-David, 2014). When $k > 1$, this setting encompasses multi-class problems, where W acts as a matrix of predictors and ℓ corresponds to a specific loss function, such as the cross-entropy loss or the multiclass hinge loss (Crammer and Singer, 2001; Mohri et al., 2018).

Another example of VVPs arises in feed-forward neural networks, where a composition of such transformations occurs, each of which corresponds to a layer with a weight matrix W and an activation function ℓ. Motivated by this connection, a recent line of work studied VVP in the regime where ℓ is Lipschitz continuous and W is constrained within a unit ball, relative to some matrix norm $\|\cdot\|$, centered around an "initialization", or reference matrix W_0 (Daniely and Granot, 2019, 2022; Vardi

*Corresponding author; schliserman@mail.tau.ac.il.

Mathematics of Modern Machine Learning Workshop at NeurIPS 2024.

et al., 2022; Magen and Shamir, 2023). These studies have yielded a range of sample complexity results depending on the particular choice of a matrix norm and properties of the initialization W_0.

In this work, we discuss an arguably more basic and fundamental case of the VVP framework: where ℓ is a *convex* (and Lipschitz) loss function and the domain is restricted to a simple unit ball, with respect to the Frobenius norm, centered around a given reference matrix W_0. In this scenario, recent work by Magen and Shamir (2023) reveals an interesting finding: while learning is possible within this framework using a specific algorithm (namely stochastic gradient descent, SGD), there exist problem instances where generic empirical risk minimization (ERM) fails. As they mention in their work, this finding is analogous to a series of studies within the more general context of Stochastic Convex Optimization (SCO), which established that learnability in SCO is algorithmic dependent and in general, learning through ERM could fail when the problem dimension is sufficiently large (Shalev-Shwartz et al., 2010; Feldman, 2016).

1.1 Our contributions

In this work, we present several findings that contribute to a better understanding of the complexity of learning vector-valued predictors (VVPs) with convex loss functions and its connection to stochastic convex optimization (SCO). Our main contributions of this paper are summarized as follows:

(i) We characterize the exact sample complexity of ERMs within the framework of convex and Lipschitz VVPs, demonstrating a lower bound of $\widetilde{\Omega}(k/\varepsilon^2)$. Together with a classic result of Maurer (2016), this implies that the sample complexity of ERM in the VVP setting is $\widetilde{\Theta}(k/\varepsilon^2)$. In particular, our lower bound provides for an exponential improvement as compared to the lower bound of Magen and Shamir (2023), that scaled poly-logarithmically with the target dimension k, and further includes the tight dependence on ε.

(ii) We present a black-box transformation from general SCO to VVP, that converts a given SCO instance in d-dimensions to a convex VVP problem with $k = \Theta(d)$ outputs. We show that using any algorithm for the VVP setting to solve the converted problem instance to within ε excess risk using a training sample of size n, we can directly recover a solution to the original SCO problem with excess risk $O(\varepsilon + 1/\sqrt{n})$.

Put together, the two results indicate that, in terms of its complexity, VVP bridges between two extreme models: generalized linear models, namely the case $k = 1$, and general d-dimensional SCO, that roughly correspond to $k = \Theta(d)$. First, our sample complexity bounds for ERM can be seen as interpolating between the classical $\Theta(1/\varepsilon^2)$ sample complexity rate of generalized linear models (Bartlett and Mendelson, 2002) and the analogous bound in d-dimensional SCO, which is linear in d (Feldman, 2016; Carmon et al., 2023). Second, from a more structural perspective, our transformation from SCO to VVP suggests that in the extreme where $k = \Theta(d)$, vector-valued prediction becomes rich enough so as to encompass generic SCO problems.

The revealed connection between linear models and SCO through the lens of VVP is perhaps somewhat surprising, since the two are extensively-studied problems that traditionally differ from one another in terms of techniques and results. Further, it partially addresses a common conceptual criticism of SCO as a learning framework: in SCO, there is no apparent concept of "prediction" and losses are rather implicitly assigned to model parameters, whereas VVP is naturally a supervised learning model that explicitly defines a rule $x \to Wx$ from which predictions are generated and losses are induced.

1.2 Additional related work

Upper bounds for the sample complexity of VVPs. As alluded to in the introduction, closely related to our setting is the work of Maurer (2016) that gives an upper bound that scales like $O(k)$, for convex and Lipschitz predictors with bounded Frobenius norm. In another work, Daniely and Granot (2022) achieved an upper bound that is similar to Maurer (2016) for non-convex predictors and with bounded difference from a reference matrix W_0. For the specific case of $W_0 = 0$, Vardi et al. (2022) shows an upper bound that scales like $O(\log k)$ and Magen and Shamir (2023) proved an upper bound independent on k. The works of Lei et al. (2019); Zhang and Zhang (2024) studied VVPs with arbitrary initialization under an ℓ_∞-Lipschitz condition, a stronger assumption compared to our setting, and showed upper bounds that are logarithmic in k. In addition, Zhang and Zhang (2024)

state a similar upper bound in the ℓ_2-Lipschitz case. In a personal communication with the authors of this paper, it was pointed out that the improved bound, which will appear in an upcoming journal version of their work, in fact holds under an additional smoothness assumption. This is compatible with our results, which indicate that some additional assumptions (though not necessarily smoothness) and potentially new techniques are indeed required to obtain such an improved dependence on k in the standard ℓ_2-Lipschitz case. Whether smoothness is particularly necessary for obtaining improved bounds in the ℓ_2- Lipchitz case remains an interesting open problem for further investigation.

Lower bounds for the sample complexity of VVPs. For one-hidden-layer neural networks, which is a special case of valued predictors with non-convex loss, Daniely and Granot (2019) provide a fat-shattering lower bound for the case where $k = m$, crucially rely on the inputs have norm that scales with k. Then, Vardi et al. (2022) referred to the case $k \neq m$ and showed a lower bound of $\Omega(k/\varepsilon^2)$ for the sample complexity of a class of non-convex predictors, where the initialization matrix is $W_0 = 0$ and the ℓ_2 norm of the prediction matrix is bounded by a constant. Then, Daniely and Granot (2022) refer to the class of predictors with bounded Frobenius norm and showed that this class can shatter a training set of $\Omega(k)$ examples, assuming that the inputs have norm $\sqrt{m}$ and that $k = O(2^m)$. In a recent work, Magen and Shamir (2023) generalized this bound to the case where the Frobenius norm of the distance from an arbitrary initialization matrix is bounded. In the same work, they discussed convex vector valued predictors and showed a lower bound of $\Omega(\log k)$ for the sample complexity of convex predictors. In this work, we improve their lower bound for convex predictors as we achieve an exponential increase in the dependence in k.

Generalized Linear Models. In the landscape of learning theory literature, the Generalized Linear Models (GLM) framework stands as one of the most basic and extensively explored settings (e.g., Shalev-Shwartz and Ben-David, 2014), as it captures some fundamental problems like logistic regression and support-vector machines. In this setting, due to dimension-independent uniform convergence, it is guaranteed that constrained ERM learns with optimal sample complexity of $O(1/\varepsilon^2)$ examples (Bartlett and Mendelson, 2002). A more recent work by Amir et al. (2022) show that unregularized gradient methods, such as full-batch Gradient Descent achieve the same sample sample complexity when learning GLMs.

Stochastic Convex optimization. Stochastic convex optimization (SCO) is a fundamental theoretical framework widely used for studying common optimization algorithms. This is often justified by the simplicity of the framework and the possibility of a rigorous analysis that can hint at the pros and cons of various optimization techniques in practical setups arising in machine learning. The works of Shalev-Shwartz et al. (2010); Feldman (2016); Carmon et al. (2023) demonstrated that, although learnability in this setting is possible (e.g., by Stochastic Gradient Descent) ERM may not learn in this setting since uniform convergence does not generally hold. Specifically, Carmon et al. (2023) recently established that the sample complexity of ERM in d-dimensional SCO is $\Theta(d/\varepsilon + 1/\varepsilon^2)$. We note that since our lower bound requires that the number of columns of the vector-valued predictor matrix is $m = \Theta(n)$ and the total number of parameters is $\Omega(mn)$, this lower bound does not contradict their upper bound. Beyond generic ERM, several specific and natural ERM algorithms, which are also frequently used in practice, such as full batch Gradient Descent have been shown to fail in learning this setting (Amir et al., 2021; Schliserman et al., 2024; Livni, 2024).

2 Problem Setup and Basic Definitions

Notations. For every vector $x \in \mathbb{R}^d$, we denote its ith entry by $x[i]$ and the vector in $\mathbb{R}^{j-i+1}$ which is achieved by taking the entries with index $i \leq k \leq j$ by $x[i:j]$. For every $n \in \mathbb{N}$, we denote $[n] = \{1, \ldots, n\}$. We denote the Frobenius norm of a matrix M by $\|M\|_F = (\sum_{i,j} M_{i,j}^2)^{1/2}$, and denote a unit ball with respect to $\|\cdot\|_F$ centered at W_0 by $\mathbb{B}_{W_0}^{k\times m}$. Moreover, for every dimension d, we denote the d-dimensional unit ball around the origin by $\mathbb{B}^d$, the d-dimensional standard basis by $\{e_1, \ldots, e_d\}$.

Vector-valued prediction. Our main setting of interest in this paper is vector-valued prediction with a convex and Lipschitz loss function. Let $\mathcal{D}$ be a distribution supported over vectors $x \in \mathbb{R}^m$ such that $\|x\| \leq 1$. Given a convex and G-Lipschitz loss function ℓ defined over the k-dimensional

unit ball $\mathbb{B}^k \subseteq \mathbb{R}^k$, and an initialization matrix W_0, the objective is to find a matrix $W \in \mathbb{B}^{k\times m}_{W_0}$ with low population loss, defined as the expected value of the loss function over the distribution $\mathcal{D}$, namely

$$L(W) = \mathbb{E}_{x\sim\mathcal{D}}[\ell(Wx)].$$

To find such a model W, the learner uses a set of n training examples $S = \{z_1, \ldots, z_n\}$, drawn i.i.d. from the unknown distribution $\mathcal{D}$. Given the sample S, the corresponding *empirical loss* (or *risk*) of W, denoted $\widehat{L}(W)$, is defined as its average loss over samples in S:

$$\widehat{L}(W) = \frac{1}{n}\sum_{i=1}^{n} L(Wx_i).$$

A population minimizer in this context is any W^* that minimizes the population risk, namely such that $W^* \in \arg\min_{W\in\mathbb{B}^{k\times m}_{W_0}} L(W)$, and an empirical risk minimizer (ERM) is any $\widehat{W}_*$ that minimizes the empirical risk, namely such that $\widehat{W}_* \in \arg\min_{W\in\mathbb{B}^{k\times m}_{W_0}} \widehat{L}(W)$.

Stochastic Convex Optimization. Another learning model we discuss is Stochastic Convex Optimization (SCO), which is a more general framework that includes (convex) vector-valued prediction as a special case. In this problem, there is a population distribution $\mathcal{D}$ over an arbitrary instance set Z and a loss function $f : \mathcal{W} \times Z \to \mathbb{R}$ which is convex and G-Lipschitz (for some $G > 0$) with respect to its first argument over a domain $\mathcal{W}$. For simplicity, we fix in this paper the domain $\mathcal{W}$ to be the d-dimensional unit ball around the origin, denoted $\mathbb{B}^d$. Analogously to vector-valued prediction, the population loss with respect to f, denoted by F, is defined as,

$$F(w) = \mathbb{E}_{z\sim\mathcal{D}}[f(w, z)].$$

and the empirical loss, denoted by $\hat{F}$, is defined as,

$$\widehat{F}(w) = \frac{1}{n}\sum_{i=1}^{n} f(w, z_i).$$

and corresponding minimizers as w_* and $\widehat{w}_*$, respectively.

3 Sample Complexity of ERM in Convex Vector-Valued Prediction

In this section, we present a tight characterization of the sample complexity of ERM in the setting of convex vector-valued prediction. This result is stated as follows.

Theorem 1. *Let $k,n \in \mathbb{N}$. There exist $m = \Theta(n)$, a reference matrix $W_0 \in \mathbb{R}^{k\times m}$, a convex and 1-Lipschitz loss function $\ell \in \mathbb{R}^k \to \mathbb{R}$ and a distribution $\mathcal{D}$ such that in the VVP parameterized by $W_0, \mathcal{D}$ and ℓ, with constant probability over the choice of the training set $S \sim \mathcal{D}^n$, there exists an ERM $\widehat{W}_*$ with $L(\widehat{W}_*) - L(W^*) = \widetilde{\Omega}\big(\sqrt{k/n}\big)$.*

We further show that this lower bound is tight up to logarithmic factors, using a vector-contraction inequality for Rademacher complexity due to Maurer (2016), that implies an $O(k/\varepsilon^2)$ sample complexity upper bound. We defer details on this standard derivation to Appendix A, and below focus on proving the lower bound in Theorem 1.

For the case of $k \le O(\log n)$, Theorem 1 follows from the well-known lower bound of $\Omega(1/\sqrt{n})$ for learning scalar-valued predictors with convex and $O(1)$-Lipschitz losses (for completeness, we provide a proof in Appendix A). Thus, henceforth we focus on the case where $k \ge \Omega(\log n)$. Our proof approach in this case is to show that, for large enough column dimension m and for a certain reference matrix W_0, the class of predictors parameterized by matrices in a unit Frobenius-norm ball centered at W_0 can shatter $\widetilde{\Omega}(k/\varepsilon^2)$ examples with margin ε.[2] This is formalized in the following.

Lemma 1. *Let $10000 \le k \in \mathbb{N}$ and $\frac{1}{12} \ge \varepsilon \ge \sqrt{k}2^{-k/312}$. There exists column dimension $m_0 = \Theta(k/\varepsilon^2)$, such that for any $m \ge m_0$, there exist a matrix $W_0 \in \mathbb{R}^{k\times m}$ and a loss function $\ell : \mathbb{R}^k \to \mathbb{R}$, such that the class of vector-valued predictors*

$$\mathcal{F}^{\ell,W_0}_{k,m} := \left\{x \to \ell(Wx) \;:\; W \in \mathbb{R}^{k\times m}, \|W - W_0\|_F \le 1\right\}$$

can shatter $\Omega(k/\varepsilon^2)$ examples with margin ε.

[2] A class of functions $\mathcal{F}$ on an input domain $\mathcal{X}$ *shatters* m points $x_1, \ldots, x_m \in \mathcal{X}$ with margin ε, if for all $y \in \{0,1\}^m$ we can find $f \in \mathcal{F}$ such that for all $i \in [m]$, it holds $f(x_i) \le -\varepsilon$ if $y_i = 0$ and $f(x_i) \ge \varepsilon$ if $y_i = 1$.

Lemma 1 implies Theorem 1 via standard arguments; we defer this proof to Appendix A and below focus on proving the lemma, which forms our main contribution in this section.

Before we formally prove Lemma 1, let us first outline the main steps and challenges in constructing the lower bound instance. Our general approach is analogous to the arguments of Magen and Shamir (2023). They show that for every $n \in \mathbb{N}$ there exists a data set $\{x_1, \ldots, x_n\}$ and labeling $y \in \mathbb{R}^n$, there exists a matrix W_y with $k = \Omega(2^n)$ such that for every i, $\ell(W_y x_i) = \varepsilon$ if $y_i = 1$ and $\ell(W_y x_i) = -\varepsilon$ if $y_i = 0$. Their approach is to use the exponentially-sized set $\{e_i\}_{i=1}^{2^n}$ of standard basis vectors and associate every possible labeling $y \in \{0, 1\}^n$ with a vector e_y in this set and a matrix W_y with $n + 1$ columns which its first n columns are the identity matrix and its last column is e_y. Then, they used a convex loss function ℓ constructed such that the predictor $\hat{y}_i = W_y x_i$ output the prediction according to the corresponding label of $x_i = e_i$.

Our main challenge, however, is to shatter a training set $\{x_1, \ldots, x_n\}$ using matrices $\{W_y\}_{y \in 2^n}$ with only $k = \Theta(n)$ rows rather than $k = O(2^n)$ rows, as in the construction above. For this, we employ a construction of a set U of approximately orthogonal vectors in $\mathbb{R}^{O(k)}$ with size which is exponential in k, adapted from Feldman (2016). (The specific construction appears in Lemma 7 in Appendix A.) In our construction of hard instance, for we use this construction twice. First, as the columns of the initialization matrix (replacing of the standard basis vectors in Magen and Shamir (2023)) and second, by identifying every possible labeling y with a vector u_y in this set (instead of e_y in Magen and Shamir (2023)) and using the matrices W_y which their last columns are u_y.

Finally, for getting the correct dependency with respect to ε, we add $\Theta(1/\varepsilon^2)$ columns to the prediction matrix, and modify the previous construction such that every possible labeling y is identified not just with a single vector u_y, but rather with a sequence of vectors in the set U alluded to in Lemma 7. This adding of the matrix enables the class $\mathcal{F}_{k,m}^{\ell,W_0}$ to shatter a larger amount of possible labelings.

The proof of Lemma 1 appears in Appendix A.

4 Black-box Transformation from SCO to VVP

In this section we provide our second main result which constitutes a black-box conversion between SCO and learning of vector-valued predictors. Namely, we show that there exists a initialization matrix $W_0 \in \mathbb{R}^{k \times m}$, such that any d-dimensional stochastic optimization problem, with loss function f and distribution $\mathcal{D}$, can be converted to a vector-valued predictions problem over $\mathbb{B}_{W_0}^{k \times m}$ with $k = O(d)$.

4.1 The transformation

Let us first outline our transformation. Consider any SCO instance in d-dimensions characterized by a distribution $\mathcal{D}$ over sample space Z, and a convex and 1-Lipschitz loss function $f : \mathbb{B}^d \times Z \to \mathbb{R}$, and consider an algorithm $\mathcal{A}$ with a guarantee that for every VVP problem, using any training set S' with n examples that are sampled i.i.d from the corresponding distribution, denoted as $\mathcal{D}'$, outputs a model $W(S')$ which has $\varepsilon(n)$-sub optimal population loss. The conversion uses a training set $S = \{z_1 \ldots z_{2n}\}$ of $2n$ examples sampled i.i.d. from $\mathcal{D}$, and takes the following form:

(i) Construct a VVP instance $\mathcal{P}$ as follows:
- The dimensions of the VVP problem are $m = 2n + 1$, $k = d + 2$.
- The reference matrix $W_0 \in \mathbb{R}^{k \times m}$ is as follows,

$$W_0 = c \left(\begin{array}{c|c|c|c|c} \phi(1) & \phi(2) & \ldots & \phi(2n) & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{array} \right).$$

  where $c > 0$ is a parameter and ϕ is a mapping $\phi : [2n] \to \mathbb{R}^2$ defined below.
- The distribution $\mathcal{D}'$ is the uniform distribution on $\{x_1, \ldots, x_{2n}\}$ where $x_i = e_i + e_{2n+1}$ for all i.
- The loss function $\ell : \mathbb{R}^k \to \mathbb{R}$ is defined as

$$\ell(\hat{y}) = \max_{j \in [2n]} \left\{ \left\langle \hat{y}[1:2], \phi(j) \right\rangle + f(\hat{y}[3:k], z_j) \right\}, \tag{1}$$

(ii) Sample a training set S' with n examples drawn i.i.d. from $\mathcal{D}'$, and use $\mathcal{A}$ with S' to solve $\mathcal{P}$ and obtain a solution matrix $W(S') \in \mathbb{B}_{W_0}^{k\times m}$.

(iii) Return the vector $w(S')$ formed by the d last entries of the $(2n+1)$th column of $W(S')$.

Here, the mapping ϕ is an embedding of the integers $1, \ldots, 2n$ into the unit sphere in two dimensions, via $\phi(j) = (\sin(\pi j/4n), \cos(\pi j/4n))^T$. Note that, since the loss function ℓ defined in Eq. (1) is convex and 2-Lipschitz (as the maximum of convex functions is also a convex function), $\mathcal{P}$ is a valid VVP problem which $\mathcal{A}$ can be used to learn.

We show that when running the algorithm $\mathcal{A}$ on $\mathcal{P}$, the solution $w(S')$ emitted by the conversion satisfies the following.

Theorem 2. *Consider any SCO instance in d-dimensions characterized by a distribution $\mathcal{D}$ over sample space Z, and a convex and 1-Lipschitz loss function $f : \mathbb{B}^d \times Z \to \mathbb{R}$ that further satisfies $|f(w, z)| \le b$ for every w, z. Let $\mathcal{P}$ be the corresponding VVP problem as defined by the conversion above for $\delta > 0$ given by Lemma 2 and $c = 4b/\delta$. Let $\mathcal{A}$ be an algorithm with a guarantee that for every VVP problem, using any training set S' with n examples that are sampled i.i.d from the corresponding distribution, outputs a model $W(S')$ with*

$$\mathbb{E}[L(W(S')) - L(W^*)] \le \varepsilon(n).$$

Then, when running the algorithm $\mathcal{A}$ on $\mathcal{P}$, the solution $w(S')$ emitted by the conversion satisfies,

$$\mathbb{E}[F(w(S')) - F(w^*)] \le 2\varepsilon(n) + \frac{10}{\sqrt{n}}.$$

The proof of Theorem 2 appears in Appendix B. Here we review the main ideas that we used for constructing this conversion.

First, we aim to represent an arbitrary unknown distribution $\mathcal{D}$ using a distribution $\mathcal{D}'$ over the unit ball $\mathbb{B}^m$ and a finite set of samples $S = \{z_1, \ldots, z_n\}$ sampled i.i.d. from $\mathcal{D}$. To achieve this, we model not $\mathcal{D}$ directly, but rather its empirical distribution, denoted as $\hat{\mathcal{D}}$, which is the uniform distribution over S and, when taking expectation over S, approximates $\mathcal{D}$. To implement this, we associate each example z_i with a standard basis vector e_i, and define the distribution $\mathcal{D}'$ as the uniform distribution over the set $\{e_1, \ldots, e_n\}$.

Second, we show how to utilize a one-parameter loss function $\ell : \mathbb{R}^k \to \mathbb{R}$ to model the two-parameter loss function $f : \mathbb{R}^d \times Z \to \mathbb{R}$, where f receive a model w and an example z as an input. Our aim is, given $w \in \mathbb{R}^d$ which is a proposed solution for the SCO problem, to construct a function $\ell : \mathbb{R}^k \to \mathbb{R}$ and a matrix $W \in \mathbb{R}^{k\times m}$ such that $k = O(d)$ and

$$\forall i \in [n] \ : \quad \ell(Wx_i) \approx f(w, z_i). \tag{2}$$

To achieve this, we use the embedding ϕ that was defined above. This embedding satisfies the following lemma,

Lemma 2. *Let $a \ge 2$. Let $\phi : [a] \to \mathbb{R}^2$ be the embedding such that for every $j \in [a]$, $\phi(j) = (\sin(\pi j/2a), \cos(\pi j/2a))^T$. Then, $\|\phi(i)\| = 1$ and there exist $\delta > 0$ such that for every $i \ne j \in [a]$, it holds that $\langle \phi(i), \phi(j)\rangle \le 1 - \delta$.*

Specifically, using δ and ϕ defined in Lemma 2 for $a = n$, we set $k = d + 2$ and utilize the first two entries of $\hat{y}_i := Wx_i \in \mathbb{R}^{d+2}$ to encode the corresponding index i using $\phi(i)$. Then, by incorporating a max term over all $i \in [n]$ into ℓ and set $c = 4b/\delta$ (where b is a bound on the values of f), this index can be decoded and the corresponding loss function $f(\cdot, z_i)$ can be applied. Finally, for constructing the matrix W we add another column with index $n + 1$ to the matrix and modify $\mathcal{D}'$ to represent the uniform distribution over $\{e_i + e_{n+1}\}_{i=1}^n$. This change of the distribution makes the last d entries of $\hat{y}_i$ equal to the last d entries of the added column (for every i). Then, when the latter is used as placeholder for w, we get that Eq. (2) holds.

Third, we relate the population loss of the two problems. For this, we employ the technique of double sampling. Specifically, we use a set $S = \{z_1, \ldots, z_{2n}\}$ sampled i.i.d. from $\mathcal{D}$ and our conversion samples only n examples from $\mathcal{D}'$. This ensures that at least half of the examples will not appear in the training set of the prediction problem. By Eq. (2), when taking the expectation over S, the expected prediction loss on such samples will be equal to the population loss in the convex optimization problem. Finally, by bounding the loss $\ell(\hat{y}_i)$ for examples that appears in the training set of the prediction problem, we get a population guarantee for the SCO problem.

Acknowledgments

Funded by the European Union (ERC, OPTGEN, 101078075). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work received additional support from the Israel Science Foundation (ISF, grant numbers 2549/19 and 3174/23), from the Len Blavatnik and the Blavatnik Family foundation, and from the Adelis Foundation.

References

I. Amir, T. Koren, and R. Livni. SGD generalizes better than GD (and regularization doesn't help). In *Conference on Learning Theory*, pages 63–92. PMLR, 2021.

I. Amir, R. Livni, and N. Srebro. Thinking outside the ball: Optimal learning with gradient descent for generalized linear stochastic convex optimization. *Advances in Neural Information Processing Systems*, 35:23539–23550, 2022.

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

D. Carmon, R. Livni, and A. Yehudayoff. The sample complexity of ERMs in stochastic convex optimization. *arXiv preprint arXiv:2311.05398*, 2023.

K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292, 2001.

A. Daniely and E. Granot. Generalization bounds for neural networks via approximate description length. *Advances in Neural Information Processing Systems*, 32, 2019.

A. Daniely and E. Granot. On the sample complexity of two-layer networks: Lipschitz vs. element-wise lipschitz activation. *CoRR*, abs/2211.09634, 2022. doi: 10.48550/arXiv.2211.09634.

V. Feldman. Generalization of ERM in stochastic convex optimization: The dimension strikes back. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

T. Koren, R. Livni, Y. Mansour, and U. Sherman. Benign underfitting of stochastic gradient descent. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 19605–19617. Curran Associates, Inc., 2022.

Y. Lei, Ü. Dogan, D.-X. Zhou, and M. Kloft. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65(5):2995–3021, 2019.

R. Livni. The sample complexity of gradient descent in stochastic convex optimization. *arXiv preprint arXiv:2404.04931*, 2024.

R. Magen and O. Shamir. Initialization-dependent sample complexity of linear predictors and neural networks. *arXiv preprint arXiv:2305.16475*, 2023.

A. Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pages 3–17. Springer, 2016.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.

M. Schliserman, U. Sherman, and T. Koren. The dimension strikes back with gradients: Generalization of gradient methods in stochastic convex optimization. *arXiv preprint arXiv:2401.12058*, 2024.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014. ISBN 978-1-10-705713-5.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

G. Vardi, O. Shamir, and N. Srebro. The sample complexity of one-hidden-layer neural networks. *CoRR*, abs/2202.06233, 2022.

Y. Zhang and M.-L. Zhang. Generalization analysis for multi-label learning. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 60220–60243. PMLR, 21–27 Jul 2024.

A Proofs of Section 3

A.1 Proof of Upper Bound

First, we use the result of Maurer (2016) to show an upper bound for the sample complexity of vector valued predictors using Rademacher Complexity. The result that we show is,

Theorem 3. *Let $k,n \in \mathbb{N}$. For every $m \in N$, $W_0 \in \mathbb{R}^{k\times m}$, convex and G-Lipschitz loss function $\ell : \mathbb{B}^{k\times m}_{W_0} \to \mathbb{R}$, distribution $\mathcal{D}$ over $\mathbb{B}^m$, it holds that,*

$$\mathbb{E}_{S\sim\mathcal{D}^n}\left[L(\widehat{W}_*) - L(W^*)\right] = O\left(\frac{\sqrt{k}}{\sqrt{n}}\right).$$

In the proof we use the standard bound of the generalization error, via the Rademacher complexity of the class (e.g. Bartlett and Mendelson (2002)), we have that:

$$\mathbb{E}_{S\sim\mathcal{D}^n}\left[\sup_{W\in\mathbb{B}^{k\times m}_{W_0}}\left\{L(W) - \hat{L}(W)\right\}\right] \leq 2\mathbb{E}_{S\sim\mathcal{D}^n}\left[R_S(\ell\circ\mathbb{B}^{k\times m}_{W_0})\right],$$

Where we notate the function class:

$$\ell\circ\mathbb{B}^{k\times m}_{W_0} = \{x \to \ell(Wx) : W \in \mathbb{B}^{k\times m}_{W_0}\}.$$

and $R_S(\ell\circ\mathbb{B}^{k\times m}_{W_0})$ is the Rademacher complexity of the class $\ell\circ\mathbb{B}^{k\times m}_{W_0}$. Namely:

$$R_S(\ell\circ\mathbb{B}^{k\times m}_{W_0}) := \mathbb{E}_\sigma\left[\sup_{h\in\ell\circ\mathbb{B}^{k\times m}_{W_0}}\frac{1}{n}\sum_{x_i\in S}\sigma_i h(x_i)\right], \tag{3}$$

and $\sigma_1,\ldots,\sigma_n$ are i.i.d. Rademacher random variables. Now we use the contraction lemma for vector valued predictors given in Maurer (2016).

Lemma 3 (Maurer (2016), Corollary 4). *Let $k \in \mathbb{N}$ and $\mathcal{X}$ be any set, $(x_1,\ldots,x_n) \in \mathcal{X}^n$, let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}^m$ and let $h_i : \mathbb{R}^k \to \mathbb{R}$ be G-Lipschitz functions. Then,*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\sum_i \sigma_i h_i\left(f\left(x_i\right)\right) \leq \sqrt{2}G\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i,j}\sigma_{ij}f_j\left(x_i\right),$$

where σ_{ij} is an independent doubly indexed Rademacher sequence and $f_j\left(x_i\right)$ is the j-th component of $f\left(x_i\right)$.

We derive the following lemma

Lemma 4. *Let $k,n \in \mathbb{N}$. For every $m \in N$, $W_0 \in \mathbb{R}^{k\times m}$, convex and G-Lipschitz loss function $\ell : \mathbb{B}^{k\times m}_{W_0} \to \mathbb{R}$, distribution $\mathcal{D}$ over $\mathbb{B}^m$, it holds that,*

$$\mathbb{E}_{S\sim\mathcal{D}^n}\left[\sup_{W\in\mathbb{B}^{k\times m}_{W_0}}\left\{L(W) - \hat{L}(W)\right\}\right] \leq \frac{2\sqrt{2}L\sqrt{k}}{\sqrt{n}}.$$

Proof. Let $S = \{x_1, \dots x_n\}$. First, for $\mathcal{F} = \{Ax_i \mid \|A\|_F \leq 1\}$, denoting the j-th row of any matrix A as A_j, and defining $h_i(w) = \ell(W_0 x_i + w)$, by Lemma 3 it holds that,

$$
\begin{aligned}
R_S(\ell \circ \mathbb{B}_{W_0}^{k\times m}) &= \mathbb{E}_\sigma \left[\sup_{W \in \mathbb{B}_{W_0}^{k\times m}} \frac{1}{n} \sum_{x_i \in S} \sigma_i \ell(W x_i) \right] \\
&= \mathbb{E}_\sigma \left[\sup_{W \in \mathbb{B}_{W_0}^{k\times m}} \frac{1}{n} \sum_{x_i \in S} \sigma_i \ell(W x_i - W_0 x_i + W_0 x_i) \right] \\
&= \mathbb{E}_\sigma \left[\sup_{W \in \mathbb{B}_{W_0}^{k\times m}} \frac{1}{n} \sum_{x_i \in S} \sigma_i h_i((W - W_0) x_i) \right] \\
&= \mathbb{E}_\sigma \left[\sup_{\|A\|_F \leq 1} \frac{1}{n} \sum_{x_i \in S} \sigma_i h_i(A x_i) \right] \\
&\leq \frac{\sqrt{2}G}{n} \mathbb{E}_\sigma \left[\sup_{\|A\|_F \leq 1} \sum_{i,j} \sigma_{ij} A_j^T x_i \right] \\
&\leq \left[\frac{\sqrt{2}G}{n} \mathbb{E}_\sigma \sup_{\|A\|_F \leq 1} \sum_j \sum_i \sigma_{ij} A_j^T x_i \right]
\end{aligned}
$$

Now, if D is the matrix that its jth column is $\sum_i \sigma_{ij} A_j^T x_i$, we get,

$$
\begin{aligned}
R_S(\ell \circ \mathbb{B}_{W_0}^{k\times m}) &\leq \frac{\sqrt{2}G}{n} \mathbb{E}_\sigma \sup_{\|A\|_F \leq 1} Tr(AD) \\
&\leq \frac{\sqrt{2}G}{n} \mathbb{E}_\sigma \sup_{\|A\|_F \leq 1} \|A\|_F \mathbb{E}_\sigma \|D\|_F \\
&\leq \frac{\sqrt{2}G}{n} \mathbb{E}_\sigma \|D\|_F \\
&= \frac{\sqrt{2}G}{n} \mathbb{E}_\sigma \sqrt{\sum_j \| \sum_i \sigma_{ij} x_i \|^2} \\
&\leq \frac{\sqrt{2}G}{n} \sqrt{\sum_j \sum_i \|x_i\|^2} \\
&\leq \frac{\sqrt{2}G\sqrt{k}}{\sqrt{n}}.
\end{aligned}
$$

The lemma follows by combining everything together. ■

For finalizing the proof of Theorem 3 we use the following lemma from Koren et al. (2022).

Lemma 5 (Koren et al. (2022), Lemma 1). *Let $W \subseteq \mathbb{R}^d$ with diameter D, $\mathcal{Z}$ any distribution over Z, and $f : W \times Z \to \mathbb{R}$ convex and G-Lipschitz in the first argument. For every sample set $S = \{z_1, \dots, z_n\}$ sampled i.i.d from $\mathcal{Z}$, let $w_S^\star = \arg\min \hat{F}(w)$ the empirical risk minimizer. Then*

$$
\mathbb{E}_S[\hat{F}(w^*) - \hat{F}(w_S^*)] \leq \frac{4GD}{\sqrt{n}}.
$$

Now, we can derive Theorem 3.

Proof of Theorem 3. By Lemma 5 and Lemma 4, we know that

$$\begin{aligned}
&\mathbb{E}_{S\sim\mathcal{D}^n}\left[L(\widehat{W}_*) - L(W^*)\right] \\
&= \mathbb{E}_{S\sim\mathcal{D}^n}\left[L(\widehat{W}_*) - \hat{L}(\widehat{W}_*)\right] + \mathbb{E}_{S\sim\mathcal{D}^n}\left[\hat{L}(\widehat{W}_*) - \hat{L}(W^*)\right] \\
&\leq \frac{2\sqrt{2}G\sqrt{k}}{\sqrt{n}} + \frac{4G}{\sqrt{n}} \\
&\leq \frac{10G\sqrt{k}}{\sqrt{n}}.
\end{aligned}$$

■

A.2 Proof of Lower Bound

First, we prove the following lemma, which implies the lower bound for the case of $k = O(\log n)$,

Lemma 6. *Let $k \in \mathbb{N}$ and $0 \leq \varepsilon \leq \frac{1}{\sqrt{n}}$. Then exists a dimension m_0 and a matrix W_0 such that for any $m \geq m_0 = \Theta\left(\frac{1}{\varepsilon^2}\right)$, $\mathcal{F}_{k,m}^{\ell,W_0}$ can shatter $\Omega\left(\frac{1}{\varepsilon^2}\right)$ examples with margin ε.*

Proof. Let $m = \frac{1}{\varepsilon^2}$ and $W_0 = \mathbf{0}_{k\times m}$. Now, for every possible labeling for S, $y \in \{\pm\varepsilon\}^{\frac{1}{\varepsilon^2}}$, we define the matrix W_y to be the matrix which its first row is u_y and the rest of the rows are 0. Note that $\|W_y - W_0\|_F = 1$. Moreover, we define $\ell : \mathbb{R}^k \to \mathbb{R}$ as $\ell(\hat{y}) = e_1\hat{y}$. This function is convex and 1-Lipschitz. For every $i \in \left[\frac{1}{\varepsilon^2}\right]$ we define $x_i = e_i$. It is left to show that the set $S = \{x_1 \ldots, x_{\frac{1}{\varepsilon^2}}\}$ can be shattered. It holds since for every y,

$$\ell(W_y x_i) = e_1 W_y x_i = y e_i = y_i.$$

■

Lemma 7. *Let $d \geq 100$. There exists a set $U_d \subseteq \mathbb{R}^d$, with $|U_d| \geq 2^{d/12}$, such that all $u \in U_d$ are of unit length $\|u\| = 1$, and for all $u, v \in U_d, u \neq v$, it holds that $\langle u, v\rangle \leq \frac{1}{2}$.*

Proof of Lemma 7. Let $r = 2^{\frac{d}{12}}$. For every $1 \leq i \leq r$ and $1 \leq j \leq d$ let the u_i^j the random variable which is $\frac{1}{\sqrt{d}}$ with probability $\frac{1}{2}$ and $-\frac{1}{\sqrt{d}}$ with probability $\frac{1}{2}$. Then, for every $1 \leq i \leq r$, we define the vector $u_i = (u_i^1, \cdots, u_i^d)$ and look at the set $U = \{u_1, u_2, ...u_r\}$. We now show that U satisfies the required property with positive probability. By Hoeffding's inequality, it holds that,

$$Pr(\langle u_i, u_k\rangle \geq \frac{1}{2}) \leq e^{\frac{-2\left(\frac{1}{2}\right)^2}{d\cdot\frac{4}{d^2}}} = e^{-\frac{d}{8}}.$$

Then, by union bound on the $\binom{r}{2}$ pairs of vectors in U,

$$Pr(\exists i, k\ \langle u_i, u_k\rangle \geq \frac{1}{2}) \leq e^{-\frac{d}{8}} \cdot \binom{r}{2} < e^{-\frac{d}{8}} \cdot \frac{1}{2}r^2 \leq 1.$$

■

Proof of Lemma 1. We denote $J = 12/\varepsilon^2$. We use the set $U := U_{k/13}$ of $(k/13)$-dimensional nearly-orthogonal vectors, given in Lemma 7 with $|U| = 2^{k/156}$, and use an arbitrary enumeration of this set $U = \{u_1, \ldots, u_{|U|}\}$. For every $u \in U \subseteq \mathbb{R}^{\frac{k}{13}}$, we define the vector $u' \in \mathbb{R}^k$ which is for $1 \leq i \leq \frac{k}{13}$, $\tilde{u}[i] = u[i]$ and other entries equal zero. Let $m = (\frac{k}{13} + \frac{1}{12})J$ and $W_0 \in \mathbb{R}^{k\times m}$ be the following matrix (note that by the lower bound for ε, it holds that $\frac{kJ}{13} \leq 2^{\frac{k}{156}}$):

$$W_0 = \varepsilon\left(\begin{array}{c|c|c|c|c} u_1 & u_2 & \ldots & u_{\frac{mJ}{7}} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{array}\right).$$

Now, for every $i \in [\frac{12k}{13}]$ and $j \in \left[\frac{J}{12}\right]$ we define $x_{i,j} = \frac{1}{\sqrt{2}}e_{\frac{12k}{13}(j-1)+i} + \frac{1}{\sqrt{2}}e_{\frac{kJ}{13}+j}$. We show that the set $S = \{x_{i,j} : i \in [\frac{12k}{13}], j \in \left[\frac{J}{12}\right]\}$ can be shattered by $\mathcal{F}_{k,m}^{\ell,W_0}$.

For this, we use the set $\hat{U} := U_{\frac{12k}{13}}$, of $\frac{12k}{13}$-dimensional nearly-orthogonal vectors, given in Lemma 7 with $|\hat{U}| = 2^{\frac{k}{13}}$ and use an arbitrary enumeration of this set $\hat{U} = \{\hat{u}_1, \dots \hat{u}_{|\hat{U}|}\}$. For the rest of the proof, we refer to every vector $z \in \{0,1\}^{\frac{kJ}{13}}$ as a sequence of $\frac{J}{12}$ vectors in $\{0,1\}^{\frac{12k}{13}}$, $z^{(1)}, \dots, z^{(\frac{J}{12})}$, where for every $r \in \frac{J}{12}$, $z^{(r)} = z[(r-1)(\frac{12k}{13}) + 1 : \frac{12kr}{13}]$. Now, since we refer to every possible labeling for S, $y \in \{0,1\}^{\frac{kJ}{13}}$ as a sequence of $\frac{J}{12}$ vectors, $y^{(1)}, \dots, y^{(\frac{J}{12})}$, it is possible to identify such labeling y with a sequence $(\hat{u}_{y^{(1)}}, \hat{u}_{y^{(2)}} \dots \hat{u}_{y^{(\frac{J}{12})}}) \in \hat{U}^{\frac{J}{12}}$. For every such y, we define the following matrix:

$$W_y = \varepsilon \left(\begin{array}{c|c|c|c|c} \mathbf{0} & \mathbf{0} & & & \\ \hline \mathbf{0} & \hat{u}_{y^{(1)}} & \hat{u}_{y^{(2)}} & \dots & \hat{u}_{y^{(\frac{J}{12})}} \end{array} \right).$$

By the definition of J, it holds that $\|W_y\|_F \leq 1$. Now, for every $\hat{u} \in \hat{U}$, we define the vector $\tilde{u} \in \mathbb{R}^k$ which is zero in its first $\frac{k}{13}$ entries and for the rest of the entries, $\frac{k}{13} + 1 \leq i \leq k$, it holds that, $\tilde{u}[i] = \hat{u}[i - \frac{k}{13}]$. We turn to define the loss function $\ell : \mathbb{R}^k \to \mathbb{R}$. For this, we define the following set

$$A = \left\{ z \in \{0,1\}^{\frac{12k}{13}}, \hat{z} \in \{0,1\}^{\frac{kJ}{13}}, r \in [\frac{12k}{13}], p \in [\frac{J}{12}] \quad \middle| \quad \forall 1 \leq b \leq \frac{J}{12} \ : \ \hat{z}^{(b)} = z, \hat{z}^{(p)}(r) = \varepsilon \right\}.$$

The the loss function ℓ is defined as following,

$$\ell(\hat{y}) = 2\sqrt{8} \max_{(\hat{z},z,r,p) \in A} \left\{ \frac{3}{\sqrt{8}}\varepsilon, \max\{\frac{\varepsilon}{\sqrt{8}}, \tilde{u}_z^T \hat{y}\} + \max\{\frac{\varepsilon}{\sqrt{8}}, u'^T_{r+\frac{12k}{13}(p-1)} \hat{y}\} \right\} - 7\varepsilon.$$

The function is 1-Lipschitz and convex as a maximum over linear 1-Lipschitz functions. For every $y \in \{0,1\}^{\frac{kJ}{13}}$ we define $W'_y = W_0 + W_y$. Let $x_{i,j} \in S$. Then,

$$W'_y x_{i,j} = \frac{1}{\sqrt{2}} \varepsilon u'_{\frac{12k}{13}(j-1)+i} + \frac{1}{\sqrt{2}} \varepsilon \tilde{u}_{y^{(j)}},$$

and

$$\begin{aligned} &\ell(W'_y x_{i,j}) = \\ &\quad = 2\sqrt{8} \max_{(\hat{z},z,r,p) \in A} \left\{ \frac{3}{\sqrt{8}}\varepsilon, \max\{\frac{\varepsilon}{\sqrt{8}}, \tilde{u}_z^T W'_y x_{i,j}\} + \max\{\frac{\varepsilon}{\sqrt{8}}, u'^T_{r+\frac{12k}{13}(p-1)} W'_y x_{i,j}\} \right\} - 7\varepsilon \\ &\quad = 2\sqrt{8}\varepsilon \max_{(\hat{z},z,r,p) \in A} \left\{ \frac{3}{\sqrt{8}}, \max\{\frac{1}{\sqrt{8}}, \frac{1}{\sqrt{2}} \hat{u}_z^T \hat{u}_{y^{(j)}}\} + \max\{\frac{1}{\sqrt{8}}, \frac{1}{\sqrt{2}} u^T_{r+\frac{12k}{13}(p-1)} u_{\frac{12k}{13}(j-1)+i}\} \right\} - 7\varepsilon \\ &\quad = 2\varepsilon \max_{(\hat{z},z,r,p) \in A} \left\{ 3, \max\{1, 2\hat{u}_z^T \hat{u}_{y^{(j)}}\} + \max\{1, 2u^T_{r+\frac{12k}{13}(p-1)} u_{\frac{12k}{13}(j-1)+i}\} \right\} - 7\varepsilon. \end{aligned}$$

If $y_{i,j} = y^{(j)}[i] = 1$, the maximum of the first term is attained at $z = y^{(j)}$ and the maximum of the sum of the terms is attained at $\hat{z}$ such that for every b, $\hat{z}^{(b)} = y^{(j)}$. Moreover, since $y^{(j)}[i] = 1$, it holds that $\hat{z}^{(b)}[i] = 1$ for every b, and particularly, for $p = j$, $r = i$, $\hat{z}^{(p)}[r] = 1$, thus, since $p = j, r = i$ gives the maximal inner product and the condition of the max holds, the maximum of the second term is attained at $p = j, r = i$, and,

$$\begin{aligned} \ell(W'_y x_{i,j}) &= 2\varepsilon \max_{(\hat{z},z,r,p) \in A} \left\{ 3, \max\{1, 2\max \hat{u}_z^T \hat{u}_{y^{(j)}}\} + \max\{1, 2\max u^T_{r+\frac{12k}{13}(p-1)} u_{\frac{12k}{13}(j-1)+i}\} \right\} - 7\varepsilon \\ &= 2\varepsilon \max_{(\hat{z},z,r,p) \in A} \left\{ 3, 2\hat{u}^T_{y^{(j)}} \hat{u}_{y^{(j)}} + 2u^T_{i+\frac{12k}{13}(j-1)} u_{\frac{12k}{13}(j-1)+i} \right\} - 7\varepsilon \\ &= 8\varepsilon - 7\varepsilon \\ &= \varepsilon. \end{aligned}$$

If $y_{i,j} = y^{(j)}[i] = 0$, and there exists r such that $y^{(j)}[r] = 1$, the maximum of the first term is attained at $z = y^{(j)}$ and the maximum of the sum of the terms is attained at $\hat{z}$ such that for every b, $\hat{z}^{(b)} = y^{(j)}$. Moreover, since $y^{(j)}[i] = 0$, it holds that $\hat{z}^{(b)}[i] = 0$ for every b, thus, for every r, p

such that $\hat{z}^{(p)}[r] = 1$, it holds that $r + \frac{12k}{13}(p-1) \neq i + \frac{12k}{13}(j-1)$ and $u^T_{r+\frac{12k}{13}(p-1)} u_{\frac{12k}{13}(j-1)+i} \leq \frac{1}{2}$. Then,

$$\begin{aligned}
\ell(W'_y x_{i,j}) &= 2\varepsilon \max_{(\hat{z},z,r,p)\in A} \left\{3, \max\{1, 2\max \hat{u}_z^T \hat{u}_{y^{(j)}}\} + \max\{1, 2\max u^T_{r+\frac{12k}{13}(p-1)} u_{\frac{12k}{13}(j-1)+i}\}\right\} - 7\varepsilon \\
&= 2\varepsilon \max_{(\hat{z},z,r,p)\in A} \left\{3, 2\hat{u}^T_{y^{(j)}} \hat{u}_{y^{(j)}} + 1\right\} - 7\varepsilon \\
&= 6\varepsilon - 7\varepsilon \\
&= -\varepsilon.
\end{aligned}$$

If $y_{r,j} = y^{(j)}[r] = 0$ for every r, for every $\hat{z}$ such that for every b, $\hat{z}^{(b)} = y^{(j)}$ it holds that $\hat{z} = \{0\}^{\frac{kJ}{13}}$. Then, the set where the maximum is applied is empty and

$$\ell(W'_y x_{i,j}) = 2\varepsilon \cdot 3 - 7\varepsilon = -\varepsilon.$$

We showed that S can be shattered by $\mathcal{F}^{\ell,W_0}_{k,m}$, which implies the lemma. ■

Proof of Theorem 1. We first prove the theorem for the case of $k = \Omega(\log n)$. Let ε which satisfy the condition of Lemma 1. Let m_0, W_0 and ℓ be as defined in Lemma 1. By Lemma 1, there exists a constant C and a set of examples $S = \{x_i\}_{i=1}^{\frac{Ck}{\varepsilon^2}}$ such that for every labeling $y \in \{0,1\}^{\frac{Ck}{\varepsilon^2}}$, there exists a matrix W_y with $\|W_y - W_0\|_F \leq 1$ such that for every $i \in [\frac{Ck}{\varepsilon^2}]$, $\ell(W_y x_i) = \varepsilon$ if $y_i = 1$ and $\ell(W_y x_i) = -\varepsilon$ if $y_i = 0$. Now, let $y^* = \{0\}^{\frac{Ck}{\varepsilon^2}}$ and W^* be the corresponding W_{y^*} and let D' be the uniform distribution over S. We prove that for every data set S' such that $|S'| \leq \frac{Ck}{2\varepsilon^2}$ sampled i.i.d from D', there exists an ERM $\widehat{W}_*$ with $L(\widehat{W}_*) - L(W^*) \geq \varepsilon$, this will prove Theorem 1. Let $S' = \{x_{i_1} \ldots x_{i_{|S'|}}\}$ be such a data set. Let $y^S \in \{0,1\}^{\frac{Ck}{\varepsilon^2}}$ be a labeling as following

$$y^S = \begin{cases} 0 & x_i \in S \\ 1 & x_i \notin S, \end{cases}$$

and $W_S := W_{y^S}$ be the corresponding matrix. First, by the definition of W_{y^S} it follows that W_S is a ERM since it holds that

$$\hat{L}(W_S) = \frac{1}{|S'|}\sum_{j=1}^{|S'|} \ell(W_S x_{i_j}) = \frac{1}{|S|}\sum_{j=1}^{|S'|} -\varepsilon = -\varepsilon.$$

Moreover, since at least $\frac{|S|}{2}$ of the examples in S are not in S', it also holds that

$$\begin{aligned}
L(W_S) - L(W^*) &= \frac{1}{|S|}\sum_{i=1}^{|S|} \ell(W_S x_i) - \ell(W^* x_i) \\
&= \frac{1}{|S|}\sum_{i=1}^{|S|} \ell(W_S x_i) + \varepsilon \\
&\geq \frac{1}{|S|}\sum_{i\notin S'} \ell(W_S x_i) + \varepsilon \\
&\geq \frac{1}{2}\cdot 2\varepsilon \\
&= \varepsilon.
\end{aligned}$$

The proof for the case of $k = O(\log n)$ is analogous and can be implied by using Lemma 6 instead of Lemma 1. ■

B Proofs of Section 4

Proof of Lemma 2. Let $\phi : [a] \to \mathbb{R}^2$, $\phi(i) = \left(\sin\left(\frac{\pi i}{2a}\right), \cos\left(\frac{\pi i}{2a}\right)\right)^T$ and $\delta = 1 - \cos\left(\frac{\pi}{2a}\right)$. We notice that $0 < \delta < 1$. Then, as a result, for every i it holds that

$$\|\phi(i)\| = \sqrt{\sin\left(\frac{\pi i}{2a}\right)^2 + \cos\left(\frac{\pi i}{2a}\right)^2} = 1,$$

and if $i \neq j$,

$$\begin{aligned}
\langle \phi(i), \phi(j) \rangle &= \sin\left(\frac{\pi i}{2a}\right)\sin\left(\frac{\pi j}{2a}\right) + \cos\left(\frac{\pi i}{2a}\right)\cos\left(\frac{\pi j}{2a}\right) \\
&= \cos\left(\frac{\pi(i-j)}{2a}\right) \\
&\leq \cos\left(\frac{\pi}{2a}\right) && \text{(cos is monotonic decreasing in } [0, \pi/2]) \\
&= 1 - \delta.
\end{aligned}$$

■

Proof of Theorem 2. First, defining $W := W(S')$ and denoting the columns of any matrix $M \in \mathbb{R}^{k \times m}$ as $M_1, \ldots M_m$. By the definitions of $w(S')$, ℓ and W_0, it holds that

$$\begin{aligned}
L(W) &= \frac{1}{2n}\sum_{i=1}^{2n} \max_{j \in [2n]} \left(\langle W_i[1:2], \phi(j)\rangle + f(w(S'), z_j)\right) \\
&\geq \frac{1}{2n}\sum_{i=1}^{2n} \langle W_i[1:2], \phi(i)\rangle + f(w(S'), z_i).
\end{aligned}$$

Now, we define the following matrix,

$$\tilde{W} = \left(\begin{array}{c|c} W_0 & \mathbf{0} \\ \hline \mathbf{0} & w^* \end{array}\right).$$

By Lemma 2, for every $i \neq j \in [2n]$,

$$\begin{aligned}
&(\langle c\phi(i), \phi(i)\rangle + f(w^*, z_i)) - \left(\langle c\phi(i), \phi(j)\rangle + f(w^*, z_j)\right) \\
&\quad = c\langle \phi(i), \phi(i)\rangle - c\langle \phi(i), \phi(j)\rangle + f(w^*, z_i) - f(w^*, z_j) \\
&\quad \geq c\delta - 2b \\
&\quad > 0.
\end{aligned}$$

As a result,

$$\begin{aligned}
L(W^*) &\leq L(\tilde{W}) \\
&= \frac{1}{2n}\sum_{i=1}^{2n} \max_{j \in [2n]} \left(\langle \tilde{W}_i[1:2], \phi(j)\rangle + f(w^*, z_j)\right) \\
&= \frac{1}{2n}\sum_{i=1}^{2n} \max_{j \in [2n]} \left(\langle c\phi(i), \phi(j)\rangle + f(w^*, z_j)\right) \\
&= \frac{1}{2n}\sum_{i=1}^{2n} \langle W_{0_i}[1:2], \phi(i)\rangle + f(w^*, z_i).
\end{aligned}$$

Now, combining with the Cauchy-Schwartz and Jensen inequalities we get that,

$$\begin{aligned}
L(W) - L(W^*) &\geq \frac{1}{2n}\sum_{i=1}^{2n} f(w(S'), z_i) - f(w_*, z_i) + \frac{1}{2n}\sum_{i=1}^{2n} \langle W_i[1:2] - W_{0_i}[1:2], \phi(i)\rangle \\
&\geq \frac{1}{2n}\sum_{i=1}^{2n} f(w(S'), z_i) - f(w_*, z_i) - \frac{1}{2n}\sum_{i=1}^{2n} \|W_i[1:2] - W_{0_i}[1:2]\| \\
&\geq \frac{1}{2n}\sum_{i=1}^{2n} f(w(S'), z_i) - f(w_*, z_i) - \sqrt{\frac{1}{2n}\sum_{i=1}^{2n} \|W_i[1:2] - W_{0_i}[1:2]\|^2} \\
&\geq \frac{1}{2n}\sum_{i=1}^{2n} f(w(S'), z_i) - f(w_*, z_i) - \frac{1}{\sqrt{2n}},
\end{aligned}$$

where in the last inequality we used the fact that $\frac{1}{2n}\sum_{i=1}^{2n}\|W_i[1:2]-W_{0_i}[1:2]\|^2 \le \|W-W_0\|_F^2 \le 1$. Then, we denote $I=\{i_1,\dots,i_p\}$ the set of indices $i\in[2n]$ that sampled from $\mathcal{D}'$ as the data set of the VVP problem. Since $p\le n$, we can add $n-p$ additional items from $[2n]\setminus I$ to I to create a set $\tilde{I}=\{i_1,\dots i_n\}$. Fixing S' and taking expectation on $S=\{z_1,\dots,z_{2n}\}$ (note that $w(S')$ and samples z_i are independent if $i\notin\tilde{I}$), we get

$$
\begin{aligned}
&\mathbb{E}_{S'}\left[L(W-L(W^*)\right]+\frac{1}{\sqrt{2n}}\\
&\quad\ge \mathbb{E}_{\{z_i:i\in\tilde{I}\}}\frac{1}{2n}\sum_{i\in\tilde{I}} f(w(S'),z_i)-f(w_*,z_i)+\mathbb{E}_{\{z_i:i\in S\}}\frac{1}{2n}\sum_{i\notin I'} f(w(S'),z_i)-f(w_*,z_i)\\
&\quad\ge \mathbb{E}_{z_{i_1},\dots,z_{i_n}}\left[\frac{1}{2n}\sum_{j=1}^{n} f(w(S'),z_{i_j})-f(w_*,z_{i_j})\right]+\mathbb{E}_{z_{i_1},\dots,z_{i_p}}\left[\frac{1}{2}F(w(S'))-\frac{1}{2}F(w_*)\right].
\end{aligned}
$$

Now, taking expectation over S', we get by Lemma 1 of Koren et al. (2022) (see Lemma 5 in Appendix A),

$$
\begin{aligned}
&\mathbb{E}\left[L(W)-L(W^*)\right]+\frac{1}{\sqrt{2n}}\\
&\quad\ge \mathbb{E}\left[\frac{1}{2n}\sum_{j=1}^{n} f(w(S'),z_{i_j})-f(w_*,z_{i_j})\right]+\mathbb{E}\left[\frac{1}{2}F(w(S'))-\frac{1}{2}F(w_*)\right]\\
&\quad\ge -\frac{2}{\sqrt{n}}+\frac{1}{2}\mathbb{E}\left[F(w(S'))-\frac{1}{2}F(w_*)\right]. \qquad \text{(Lemma 5)}
\end{aligned}
$$

The theorem follows by the guarantee on $\mathcal{A}$ and arranging the inequality. ■