On the Capacity Limits of Privileged ERM

Michal Sharoni
Department of Computer Science,
Ben-Gurion University of the Negev
Beer-Sheva, Israel

Sivan Sabato
Department of Computer Science,
Ben-Gurion University of the Negev
Beer-Sheva, Israel

Abstract

We study the supervised learning paradigm called *Learning Using Privileged Information*, first suggested by Vapnik and Vashist (2009). In this paradigm, in addition to the examples and labels, additional (privileged) information is provided only for training examples. The goal is to use this information to improve the classification accuracy of the resulting classifier, where this classifier can only use the non-privileged information of new example instances to predict their label. We study the theory of privileged learning with the zero-one loss under the natural Privileged ERM algorithm proposed in Pechyony and Vapnik (2010a). We provide a counter example to a claim made in that work regarding the VC dimension of the loss class induced by this problem; We conclude that the claim is incorrect. We then provide a correct VC dimension analysis which gives both lower and upper bounds on the capacity of the Privileged ERM loss class. We further show, via a generalization analysis, that worst-case guarantees for Privileged ERM cannot improve over standard non-privileged ERM, unless the capacity of the privileged information is similar or smaller to that of the non-privileged information. This result points to an important limitation of the Privileged ERM approach. In our closing discussion, we suggest another way in which Privileged ERM might still be helpful, even when the capacity of the privileged information is large.

1 INTRODUCTION

The classical paradigm of supervised machine learning considers the following setting: given a set of labeled training examples, try to find in a given set of functions, the one with the smallest generalization error on the unknown test examples. In this work, we study an augmentation of this setting, first proposed by Vapnik and Vashist (2009), called *Learning Using Privileged Information*, or simply, *privileged learning*. In this paradigm, during the training stage, additional information about the training examples is provided to the learner. This information, called *privileged information*, is available only for training examples during the training stage. The goal is to use this information to improve the classification accuracy of the resulting classifier. The classifier itself can use only non-privileged information of new examples to predict their label. Thus, the privileged information is only helpful inasmuch as it helps to obtain a better classifier.

A classical motivating example to this paradigm (see Vapnik and Vashist, 2009) considers a case where the goal is to find a rule that predicts the outcome of a surgery after three months, based on information about the patient which is available before the surgery. However, for previous patients, there is additional information collected during and after the surgery. Although this information is not available during classification of new patients, it does exist in historical data and thus can be used as privileged information during training.

In this work, we study the natural ERM algorithm proposed in Pechyony and Vapnik (2010a), called Privileged ERM. This algorithm minimizes a joint loss of the non-privileged and the privileged information. We provide new results which point to the limitations of this approach when the privileged information is high-dimensional, or more generally, when the associated privileged loss class has a high capacity. High-dimensional privileged information is natural in many settings where offline measurements collected for training have a higher bandwidth or sensitivity than measurements during test time. For instance, consider a learning problem in which the goal is to classify images, in which the non-privileged information provides a low-resolution image, and the privileged information provides a high-resolution image. This would be the case if during training the training samples can be scanned using advanced equipment, while

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

the classifier is deployed in a low-resource environment in the field, in which only low-quality images can be obtained. A similar application was studied in Lee et al. (2020). We show here that the Privileged ERM approach with the zero-one loss cannot guarantee successes in this regime without additional assumptions.

We provide a VC dimension analysis for the loss class induced by the Privileged ERM algorithm. Our analysis includes a counter example to a claim previously made in Pechyony and Vapnik (2010a); The mistake can be traced to an error in the proof of that claim. We provide a correct analysis with both lower and upper bounds on the VC dimension. Thereafter, we study the regimes in which it is possible to provide a guarantee that Privileged ERM will result in an improved error bound over standard ERM, in which the privileged information is not used at all. We conclude that such worst-case guarantees must rely on a low-capacity privileged information class. Lastly, we suggest a possible way in which Privileged ERM can still be helpful, even when the capacity of the privileged information is large.

2 RELATED WORK

The paradigm of privileged learning was first proposed by Vapnik and Vashist (2009). This work introduced the SVM+ algorithm, which demonstrated how privileged information can be used in SVM-type algorithms, by changing their goal such that it will incorporate the privileged information. In addition to introducing the SVM+ algorithm, Vapnik and Vashist (2009) derived results showing an improvement in the rate of convergence that can be achieved when utilizing privileged information in those types of algorithms, when the privileged-information class is low-dimensional. Pechyony and Vapnik (2010a) proposed an empirical risk minimization algorithm called Privileged ERM and generalized of the privileged learning optimization problem to other losses. They provide several theoretical claims regarding the convergence rate of this algorithm.

Since its inception, privileged learning has been applied in various domains. In Lapin et al. (2014) the connection between SVM+ and weighted SVM is studied. It is shown that privileged information can be encoded by weights associated with every training example. In addition, it is shown that weighted SVM can always replicate an SVM+ solution, while the converse is not true. In Vapnik and Izmailov (2015), two mechanisms related to knowledge transfer between the instance space and the privileged information space are described. These mechanisms can be used for accelerating the speed of learning. In Qi et al. (2015), a semi-supervised learning approach using privileged information is proposed. This approach can exploit both the distribution information in unlabeled data and privileged information, to improve the efficiency of the learning. In Yang et al. (2016), a metric-learning algorithm is proposed, which exploits privileged information to relax a previous method for metric-learning, under the ERM framework. In Vrigkas et al. (2016), a probabilistic approach is described, that combines learning using privileged information and active learning. In Pasunuri et al. (2016), an algorithm for learning decision trees using privileged information is proposed. In Vapnik and Izmailov (2017), a mechanism of knowledge transfer from the privileged information space to the features space is proposed. It is shown that this mechanism is applicable to a neural network framework as well as to SVM. Recent works study privileged learning in vision domains (e.g., Yuan et al., 2019; Gao et al., 2019; Li et al., 2019). Lee et al. (2020) considers an application in which the privileged information is high-dimensional. However, the theory of privileged learning has not addressed the capacity limits of privileged information under its basic methodologies.

3 PRELIMINARIES AND SETTING

We start by describing the privileged learning setting for general losses, as defined in Pechyony and Vapnik (2010a). Let $\mathcal{X}$ be the domain of elements that we wish to label. Let $\mathcal{X}^*$ be the domain of the privileged information that is available for training examples. Let $\mathcal{Y}$ be the set of possible labels. The input to the learner consists of a sequence of i.i.d. triplets:

$$\begin{aligned} &S = (x_1, x_1^*, y_1), ..., (x_m, x_m^*, y_m), \\ &x_i \in \mathcal{X}, \quad x_i^* \in \mathcal{X}^*, \quad y_i \in \mathcal{Y}, \end{aligned} \tag{1}$$

generated according to a fixed but unknown probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{X}^* \times \mathcal{Y}$. Let $\ell_{\mathcal{X}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ be a bounded loss function over the non-privileged example domain. The goal of privileged learning is to find a hypothesis that obtains a low loss on $\mathcal{D}$, by using the sample S that includes the privileged information.

Assume a bounded loss for privileged information, $\ell_{\mathcal{X}^*} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$. Let $C > 0$ be a constant, and denote $[t]_+ = \max(t, 0)$. Given a classifier $h : \mathcal{X} \to \mathcal{Y}$ that uses only non-privileged information, and a privileged-information function $\phi : \mathcal{X}^* \to \mathcal{Y}$, Pechyony and Vapnik (2010a) define the loss of the composite hypothesis (h, ϕ) on the example (x, x^*, y) by:

$$\ell'_C(h, \phi, (x, x^*, y)) = \frac{1}{C}\ell_{\mathcal{X}^*}(\phi(x^*), y) + [\ell_{\mathcal{X}}(h(x), y) - \ell_{\mathcal{X}^*}(\phi(x^*), y)]_+.$$

The function ϕ is thought of as a "correcting function" for the loss induced by h on the example. Given a function class over the non-privileged information $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, and a function class over the privileged information $\Phi \subseteq \mathcal{Y}^{\mathcal{X}^*}$, Pechyony and Vapnik (2010a) defined the Privileged ERM minimization problem as the following optimization prob-

lem:

$$\min_{h\in\mathcal{H},\phi\in\Phi}\sum_{i=1}^{m}\ell'_C((h,\phi),(x_i,x_i^*,y_i)). \tag{2}$$

In this work, we focus on the setting above in the important special case of binary labels ($\mathcal{Y}=\{0,1\}$) and binary loss functions, with $C=1$. In this case, we have

$$\ell'_C(h,\phi,(x,x^*,y))=\max\{\ell_{\mathcal{X}}(h(x),y),\ell_{\mathcal{X}^*}(\phi(x^*),y)\}.$$

This leads to the following optimization problem:

$$\min_{h\in\mathcal{H},\phi\in\Phi}\sum_{i=1}^{m}\max\{\ell_{\mathcal{X}}(h(x_i),y_i),\ell_{\mathcal{X}^*}(\phi(x_i^*),y_i)\}. \tag{3}$$

It is instructive to think of ϕ as indicating which training examples should be taken into account when minimizing the loss over h, where $\phi(x_i^*)=1$ indicates that example x_i should be ignored in the minimization. For instance, this could be relevant if the privileged information allows identifying the reliability of the labeling, as in a case of crowd-sourced labels. We thus assume that $\ell_{\mathcal{X}}$ is the standard loss on the non-privileged information, defined by

$$\ell^{01}_{\mathcal{X}}(\hat{y},y):=\mathbf{1}[\hat{y}\neq y]$$

and that $\ell_{\mathcal{X}^*}$ is an "ignoring" loss on the privileged information, defined by:

$$\ell^{\text{ig}}_{\mathcal{X}^*}(z,y):=\mathbf{1}[z=1].$$

Denote the error of h with respect to the distribution $\mathcal{D}$ by $\text{err}(h,\mathcal{D}):=\mathbb{P}_{(X,Y)\sim\mathcal{D}}[h(X)\neq Y]=\mathbb{E}[\ell_{\mathcal{X}}(h(X),Y)]$. Let $\text{err}(h,S)$ be the empirical error of h over the uniform distribution on S. The goal of privileged learning is thus to find a hypothesis from $\mathcal{H}$ that obtains a low error on $\mathcal{D}$, using the sample S. In the paradigm of Privileged ERM that we study here, this is attempted by solving the optimization problem in Eq. (3).

4 VC-DIMENSION ANALYSIS

In this section, we study the VC-dimension of the relevant function class for the minimization problem defined in Eq. (3). Denote the VC dimension of a class of functions by $\text{VC}(\cdot)$. Denote $d:=\text{VC}(\mathcal{H})$ and $d^*:=\text{VC}(\Phi)$. Define

$$f_{(h,\phi)}((x,x^*),y):=\max(\ell^{01}_{\mathcal{X}}(h(x),y),\ell^{\text{ig}}_{\mathcal{X}^*}(\phi(x^*),y)),$$

and let

$$\mathcal{F}_{(\mathcal{H},\Phi)}=\{f_{(h,\phi)}\mid h\in\mathcal{H},\phi\in\Phi\}.$$

We write $\mathcal{F}$ for $\mathcal{F}_{(\mathcal{H},\Phi)}$ when the subscripts are clear from context. Eq. (3) is equivalent to running an ERM on S with the hypothesis class $\mathcal{F}$. Thus, the generalization behavior of Privileged ERM is characterized by the VC dimension of $\mathcal{F}$.

What is the relationship between $\text{VC}(\mathcal{F})$ and the values of $\text{VC}(\mathcal{H}),\text{VC}(\Phi)$? This question was seemingly answered in Pechyony and Vapnik (2010a); They defined the following loss classes:

$$\mathcal{L}(\mathcal{H}):=\{\ell_{\mathcal{X}}(h(\cdot),\cdot)\mid h\in\mathcal{H}\},$$
$$\mathcal{L}(\Phi):=\{\ell_{\mathcal{X}^*}(\phi(\cdot),\cdot)\mid\phi\in\Phi\},$$
$$\mathcal{L}(\mathcal{H},\Phi)=\{\ell'_C((h,\phi),(\cdot,\cdot,\cdot))\mid h\in\mathcal{H},\phi\in\Phi\}.$$

and claimed that the following equality holds:[1]

Claim of Pechyony and Vapnik (2010a):

$$\text{VC}(\mathcal{L}(\mathcal{H},\Phi))=\text{VC}(\mathcal{L}(\mathcal{H}))+\text{VC}(\mathcal{L}(\Phi)). \tag{4}$$

The equality was then used to prove generalization upper bounds for the Privileged ERM optimization problem.

For $\ell_{\mathcal{X}}:=\ell^{01}_{\mathcal{X}}$ and $\ell_{\mathcal{X}^*}:=\ell^{\text{ig}}_{\mathcal{X}^*}$, we have $\text{VC}(\mathcal{H})=\text{VC}(\mathcal{L}(\mathcal{H}))$, $\text{VC}(\Phi)=\text{VC}(\mathcal{L}(\Phi))$ and $\text{VC}(\mathcal{L}(\mathcal{H},\Phi))=\text{VC}(\mathcal{F}_{(\mathcal{H},\Phi)})$. Therefore, if Eq. (4) were true, it would imply that $\text{VC}(\mathcal{F})=\text{VC}(\mathcal{H})+\text{VC}(\Phi)$. However, we now show that this equality in fact *does not hold*.[2] Theorem 4.1 below provides a counter example to the claimed Eq. (4).

Theorem 4.1. *For any integer $d>0$, and any two domains $\mathcal{X},\mathcal{X}^*$ such that $|\mathcal{X}|,|\mathcal{X}^*|\geq 3d$, there exist hypothesis classes $\mathcal{H}_d\subseteq\{0,1\}^{\mathcal{X}}$ and $\Phi_d\subseteq\{0,1\}^{\mathcal{X}^*}$ such that $\text{VC}(\mathcal{H}_d)=\text{VC}(\Phi_d)=d$ while $\text{VC}(\mathcal{F}_{(\mathcal{H}_d,\Phi_d)})=3d$.*

Proof. First, consider the case of $d=1$. Let $X_3=\{x_1,x_2,x_3\}\subseteq\mathcal{X}$ be a set of size three of domain examples from $\mathcal{X}$. We describe a hypothesis h from X_3 to $\{0,1\}$ via the triplet $(h(x_1),h(x_2),h(x_3))$. Similarly, let $X_3^*=\{x_1^*,x_2^*,x_3^*\}\subseteq\mathcal{X}^*$, and describe a hypothesis ϕ from X_3^* to $\{0,1\}$ via the triplet $(\phi(x_1^*),\phi(x_2^*),\phi(x_3^*))$. Define the following hypothesis classes over X_3 and X_3^*:

$$\mathcal{H}_1:=\{(0,0,0),(0,0,1),(1,0,0),(1,1,0)\},$$
$$\Phi_1:=\{(0,0,0),(0,0,1),(0,1,0),(1,0,1)\}.$$

It is easy to see that $\text{VC}(\mathcal{H}_1)=\text{VC}(\Phi_1)=1$. On the other hand, when restricting $\mathcal{F}_{(\mathcal{H}_1,\Phi_1)}$ to the set $\tilde{X}_3:=\{((x_i,x_i^*),0)\}_{i\in[3]}$, we get that $\text{VC}(\mathcal{F}_{(\mathcal{H}_1,\Phi_1)})\leq 3$, since the domain is of size 3. Moreover, the VC dimension is exactly 3, since $\mathcal{F}_{(\mathcal{H}_1,\Phi_1)}$ induces all possible labelings on $\tilde{X}_3$: Any labeling $h\in\mathcal{H}_1$ can be obtained by f_{h,ϕ_0}, where ϕ_0 is the all-zero function in Φ_1. Similarly, all labelings in Φ_1 can be obtained using the all-zero $h_0\in\mathcal{H}_1$. The two additional missing labelings are $(0,1,1)$ and $(1,1,1)$. The first can be obtained using $h=(0,0,1)$ and $\phi=(0,1,0)$,

[1] The original claim includes real-valued losses, which requires generalizing the definition of VC-dimension; Here we state it for the special case of losses that map into $\{0,1\}$

[2] We traced the issue to an application of quantifiers in the wrong order in the proof of Eq. (4) in Pechyony and Vapnik (2010a), which is available in the full version (Pechyony and Vapnik, 2010b).

and the second can be obtained using $h = (1, 1, 0)$ and $\phi = (0, 0, 1)$. Thus, $\mathrm{VC}(\mathcal{F}_{(\mathcal{H}_1,\Phi_1)}) = 3$, as claimed.

Next, consider $d > 1$. Let $X_{3d} = \{x_1, ..., x_{3d}\} \subseteq \mathcal{X}$ be a set of $3d$ different domain points from $\mathcal{X}$. Partition these points into d triplets, denoted $t_1 := (x_1, x_2, x_3), ..., t_d := (x_{3d-2}, x_{3d-1}, x_{3d})$. We describe a hypothesis h over X_{3d} as a sequence of d functions from $\mathcal{H}_1$ that are applied to the examples in the triplets $t_1, \ldots, t_d$. A description of ϕ over X^*_{3d} is analogous. Define the following hypothesis classes of functions from X_{3d} to $\{0, 1\}$ and from X^*_{3d} to $\{0, 1\}$:

$$\mathcal{H}_d := \{(h_1, ..., h_d) \mid h_1, ..., h_d \in \mathcal{H}_1\},$$
$$\Phi_d := \{(\phi_1, ..., \phi_d) \mid \phi_1, ..., \phi_d \in \Phi_1\}.$$

We first prove that $\mathrm{VC}(\mathcal{H}_d) = d$. Suppose for contradiction that $\mathrm{VC}(\mathcal{H}_d) > d$. Then there exists a shattered set with $d + 1$ points. Since the predictors are defined on d triplets, by the pigeonhole principle, there must be two points of the shattered set that are from the same triplet. From this we can conclude that $\mathcal{H}_1$ shatters a set of size two, in contradiction to $\mathrm{VC}(\mathcal{H}_1) = 1$. Therefore, $\mathrm{VC}(\mathcal{H}_d) \leq d$. Next, we prove that $\mathrm{VC}(\mathcal{H}) \geq d$, by showing that there exists a shattered set of size d. Since $\mathrm{VC}(\mathcal{H}_1) = 1$, in each of the d domain triplets there is a point that $\mathcal{H}_1$ shatters. The set of all of these points is a set of size d which is shattered by $\mathcal{H}_d$, as needed. We conclude that $\mathrm{VC}(\mathcal{H}_d) = d$. By analogous arguments, $\mathrm{VC}(\Phi_d) = d$.

Lastly, we show that $\mathrm{VC}(\mathcal{F}_{(\mathcal{H}_d,\Phi_d)}) = 3d$. Define the set

$$\mathcal{F}_d := \{(f_1, ..., f_d) \mid f_1, ..., f_d \in \mathcal{F}_{(\mathcal{H}_1,\Phi_1)}\}.$$

We first claim that $\mathcal{F}_d \subseteq \mathcal{F}_{(\mathcal{H}_d,\Phi_d)}$: Let $(f_{(h_1,\phi_1)}, ..., f_{(h_d,\phi_d)}) \in \mathcal{F}_d$. From the definition of $\mathcal{F}_{(\mathcal{H}_1,\Phi_1)}$, $\forall i \in [d]$ we have

$$f_{(h_i,\phi_i)}((x, x^*), y) = \max(\ell^{01}_{\mathcal{X}}(h_i(x), y), \ell^{\mathrm{ig}}_{\mathcal{X}^*}(\phi_i(x^*), y)).$$

Therefore,

$$(f_{(h_1,\phi_1)}, ..., f_{(h_d,\phi_d)}) = f_{((h_1,...,h_d),(\phi_1,...,\phi_d))} \in \mathcal{F}_{(\mathcal{H}_d,\Phi_d)}.$$

From this we conclude that $\mathrm{VC}(\mathcal{F}_{(\mathcal{H}_d,\Phi_d)}) \geq \mathrm{VC}(\mathcal{F}_d)$. Now, consider the set $\widetilde{X}_{3d} = \{((x_i, x^*_i), 0)\}_{i\in[3d]}$. Restricting $\mathcal{F}_d$ to the set $\widetilde{X}_{3d}$ results in the set of all possible functions over each triplet in $\widetilde{X}_{3d}$, as in the case of $d = 1$. Thus, $\mathcal{F}_d$ is shattered by $\widetilde{X}_{3d}$. Therefore, $\mathrm{VC}(\mathcal{F}_d) \geq 3d$ as needed. □

We provide a correct upper bound for $\mathrm{VC}(\mathcal{F}_{(\mathcal{H},\Phi)})$ in the following theorem.

Theorem 4.2. *Let d, d^* be integers and let $\mathcal{X}$ be some domain. Let $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}, \Phi \subseteq \{0, 1\}^{\mathcal{X}^*}$ be hypothesis classes such that $\mathrm{VC}(\mathcal{H}) = d$ and $\mathrm{VC}(\Phi) = d^*$. Then*

$$\mathrm{VC}(\mathcal{F}_{(\mathcal{H},\Phi)}) \leq 4\log_2(4e)(d+d^*+1) \approx 13.77(d+d^*+1).$$

To prove the theorem, we first provide a tight upper bound on the VC dimension of the union of two hypothesis classes over the same domain.

Lemma 4.3. *Let $\mathcal{Y} = \{0, 1\}$. Let $\mathcal{J}$, $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be two hypothesis class over the same domain. Then*

$$\mathrm{VC}(\mathcal{H} \cup \mathcal{J}) \leq \mathrm{VC}(\mathcal{H}) + \mathrm{VC}(\mathcal{J}) + 1.$$

This bound is tight: For any $d, d^* \in \mathbb{N}$, there exist $\mathcal{J}$ and $\mathcal{H}$ such that $\mathrm{VC}(\mathcal{H}) = d$, $\mathrm{VC}(\mathcal{J}) = d^*$ and $\mathrm{VC}(\mathcal{H} \cup \mathcal{J}) = d + d^* + 1$.

Proof. For a hypothesis class $\mathcal{F}$, denote the growth function of $\mathcal{F}$ by

$$\Pi_{\mathcal{F}}(m) := \max\{\, |\mathcal{F}|_S \mid S \subseteq \mathcal{X} \times \mathcal{Y}, |S| = m \,\}.$$

Clearly, for any S, $|\mathcal{H} \cup \mathcal{J}|_S \leq |\mathcal{H}|_S + |\mathcal{J}|_S$. Therefore $\Pi_{\mathcal{H}\cup\mathcal{J}}(m) \leq \Pi_{\mathcal{H}}(m)+\Pi_{\mathcal{J}}(m)$. By Sauer's Lemma (Sauer, 1972) and using the identity $\binom{m}{k} = \binom{m}{m-k}$, denoting $d := \mathrm{VC}(\mathcal{H})$ and $d^* := \mathrm{VC}(\mathcal{J})$, we get

$$\Pi_{\mathcal{H}\cup\mathcal{J}}(m) \leq \sum_{i=0}^{d}\binom{m}{i} + \sum_{i=0}^{d^*}\binom{m}{i}$$
$$= \sum_{i=0}^{d}\binom{m}{i} + \sum_{i=m-d^*}^{m}\binom{m}{i}. \tag{5}$$

If $m > d + d^* + 1$ then $m - d^* \geq d + 2$, so:

$$\sum_{i=0}^{d}\binom{m}{i} + \sum_{i=m-d^*}^{m}\binom{m}{i} \leq \sum_{i=0}^{m}\binom{m}{i} = 2^m. \tag{6}$$

Combining Eq. (5) and Eq. (6), we conclude that for $m > d + d^* + 1$, $\Pi_{\mathcal{H}\cup\mathcal{J}}(m) < 2^m$. Therefore,

$$\mathrm{VC}(\mathcal{H} \cup \mathcal{J}) \leq d + d^* + 1.$$

This proves the upper bound.

To show that this bound is tight, let $d, d^* \in \mathbb{N}$ and consider a domain $\mathcal{X}$ of size $d + d^* + 1$. Let $\mathcal{H}$ be the set of all the functions that map at most d examples from $\mathcal{X}$ to 1 and let $\mathcal{J}$ be the set of all the functions that map at most d^* examples from $\mathcal{X}$ to 0. Then, $\mathrm{VC}(\mathcal{H}) = d$ and $\mathrm{VC}(\mathcal{J}) = d^*$. To show that $\mathrm{VC}(\mathcal{H} \cup \mathcal{J}) \geq d + d^* + 1$, consider two cases for a labeling of $\mathcal{X}$:

- If the labeling includes at most d positive labels, then there is a function in $\mathcal{H}$ that provides this labeling.
- If the labeling includes more than d positive labels, then since the domain is of size $d+d^*+1$, the labeling contains at most d^* negative labels. Therefore, there is a function in $\mathcal{J}$ that provides this labeling.

Thus, $\mathcal{H}\cup\mathcal{J} = \mathcal{Y}^{\mathcal{X}}$, hence $\mathrm{VC}(\mathcal{H}\cup\mathcal{J}) = |\mathcal{X}| = d+d^*+1$, as claimed. □

In our proof of Theorem 4.2 we further use a theorem from Blumer et al. (1989).[3]

Theorem 4.4 (Blumer et al. 1989)**.** *Let $(\mathcal{X}, \mathcal{R})$ be a set system, where $\mathcal{X}$ is a set of elements and $\mathcal{R}$ is a set of subsets of $\mathcal{X}$. For an integer $k \geq 2$ and a set of sets $\mathcal{R}$, define $\mathcal{R}^{k\cup} := \{R_1 \cup ... \cup R_k \mid R_1, ..., R_k \in \mathcal{R}\}$, the k-fold union of $\mathcal{R}$. Then*

$$\mathrm{VC}(\mathcal{R}^{k\cup}) \leq \mathrm{VC}(\mathcal{R}) \cdot 2k \log_2(2ek).$$

We now prove Theorem 4.2.

Proof of Theorem 4.2. Let $\mathcal{H}, \Phi$ be hypothesis classes as stated in the theorem. Define $\mathcal{H}', \Phi' \subseteq \{0,1\}^{\mathcal{X}\times\mathcal{X}^*\times\{0,1\}}$ as follows:

$$\mathcal{H}' := \{(x, x^*, y) \mapsto \ell^{01}_{\mathcal{X}}(h(x), y) \mid h \in \mathcal{H}\}$$

and

$$\Phi' := \{(x, x^*, y) \mapsto \ell^{\mathrm{ig}}_{\mathcal{X}^*}(\phi(x^*), y) \mid \phi \in \Phi\}.$$

In addition, let

$$\mathcal{F}'_{\mathcal{H},\Phi} := \{(x, x^*, y) \mapsto f_{(h,\phi)}((x, x^*), y) \mid h \in \mathcal{H}, \phi \in \Phi\}.$$

It is easy to see that $\mathrm{VC}(\mathcal{H}') = \mathrm{VC}(\mathcal{H})$, $\mathrm{VC}(\Phi') = \mathrm{VC}(\Phi)$, and $\mathrm{VC}(\mathcal{F}'_{\mathcal{H},\Phi}) = \mathrm{VC}(\mathcal{F}_{\mathcal{H},\Phi})$. In addition, for any $h \in \mathcal{H}, \phi \in \Phi$, we have

$$\begin{aligned}\{(x, x^*, y) \mid f_{(h,\phi)}((x, x^*), y) = 1\} = \\ \{(x, x^*, y) \mid \ell^{01}_{\mathcal{X}}(h(x), y) = 1\} \\ \cup \{(x, x^*, y) \mid \ell^{\mathrm{ig}}_{\mathcal{X}^*}(\phi(x^*), y) = 1\}.\end{aligned}$$

Therefore, treating functions in $\{0,1\}^{\mathcal{X}\times\mathcal{X}^*\times\{0,1\}}$ as sets, we have $f_{(h,\phi)} = h' \cup \phi'$, where $h' \in \mathcal{H}'$ and $\phi' \in \Phi'$. Letting $\mathcal{R} := \mathcal{H}' \cup \Phi'$, it follows that $\mathcal{F}_{(\mathcal{H},\Phi)} \subseteq \mathcal{R}^{2\cup}$. Therefore, $\mathrm{VC}(\mathcal{F}) \leq \mathrm{VC}(\mathcal{R}^{2\cup})$. By Theorem 4.4 with $k = 2$, it follows that $\mathrm{VC}(\mathcal{F}) \leq \mathrm{VC}(\mathcal{R}) \cdot 4\log_2(4e)$. In addition, by Lemma 4.3, $\mathrm{VC}(\mathcal{R}) = \mathrm{VC}(\mathcal{H} \cup \Phi) \leq d + d^* + 1$. Therefore, $\mathrm{VC}(\mathcal{F}) \leq 4\log_2(4e)(d + d^* + 1)$, as claimed. □

5 CAPACITY LIMITS OF PRIVILEGED ERM

We now turn to study the convergence rate of the solution to the optimization problem in Eq. (3), and derive conditions on the VC dimension values that allow this bound to be better than the known bound for regular ERM. We show that for guaranteed generalization improvement, the VC dimension of the privileged information cannot be much larger than the VC dimension of the non-privileged information, thus limiting the usefulness of this approach in the case of zero-one losses.

A guaranteed generalization improvement occurs if the error guarantee of the optimization problem in Eq. (3) is smaller than that of standard ERM generalization bounds. As shown in previous works (e.g., Vapnik and Vashist, 2009), this requires bounds that take into account the error with respect to the hypothesis class: The advantage of privileged information, when it exists, comes from the possibility of faster convergence due to a smaller error rate. Boucheron et al. (2005) provide tight error bounds that take into account the error. Fixing $m \in \mathbb{N}$, denote for $d \in \mathbb{N}, x \in [0,1]$,

$$\begin{aligned}R_{\mathrm{f}}(d) &:= \frac{8d\log(m+1) + 4\log(\frac{4}{\delta})}{m} \\ R_{\mathrm{s}}(x, d) &:= \sqrt{x \cdot R_{\mathrm{f}}(d)},\end{aligned}$$

where R_{f} stands for a fast rate an R_{s} stands for a slow rate. By Boucheron et al. (2005, Corollary 5.2), denoting by $\hat{h}_{\mathrm{ERM}}$ the output of a standard ERM algorithm for the hypothesis class $\mathcal{H}$ on the sample S, and its training error by $\hat{\varepsilon}_{\mathrm{ERM}} := \mathrm{err}(\hat{h}_{\mathrm{ERM}}, S)$, we have that with a probability at least $1 - \delta$,

$$\mathrm{err}(\hat{h}_{\mathrm{ERM}}, \mathcal{D}) \leq \hat{\varepsilon}_{\mathrm{ERM}} + R_{\mathrm{s}}(\hat{\varepsilon}_{\mathrm{ERM}}, d) + R_{\mathrm{f}}(d) := \mathrm{B}_{\mathrm{ERM}}.$$

Based on this result, we derive an analogous upper bound for the case of Privileged ERM. To provide the bound, we first define some notations. Define an auxiliary loss function

$$\ell^a_{(h,\phi)}(x, x^*, y) := \mathbf{1}[h(x) \neq y \wedge \phi(x^*) = 0]$$

and the loss class $\mathcal{L}^a_{(\mathcal{H},\Phi)} := \{\ell^a_{(h,\phi)} \mid h \in \mathcal{H}, \phi \in \Phi\}$. Denote $d_a := \mathrm{VC}(\mathcal{L}^a_{(\mathcal{H},\Phi)})$. By Theorem 4.2, $d_a \leq 13.77(d + d^* + 1)$. The following lemma provides a lower bound for d_a, leading to the conclusion that $d_a = \Theta(d + d^*)$.

Lemma 5.1. *For $\mathcal{H}$, Φ such that $d, d^* > 1$,*

$$d_a \geq d + d^* - 2.$$

Proof. Let $C_{\mathcal{H}} = \{x_1, ..., x_d\} \subseteq \mathcal{X}$ be a set of size d that is shattered by $\mathcal{H}$ and $C_\Phi = \{x^*_1, ..., x^*_{d^*}\} \subseteq \mathcal{X}^*$ be a set of size d^* that is shattered by Φ. Define

$$\begin{aligned}C_1 &:= \{(x_i, x^*_{d^*}, 0) \mid 1 \leq i \leq d-1\}, \\ C_2 &:= \{(x_d, x^*_j, 0) \mid 1 \leq j \leq d^*-1\}, \\ C_{\mathcal{L}^a_{(\mathcal{H},\Phi)}} &:= C_1 \cup C_2.\end{aligned}$$

Note that $|C_{\mathcal{L}^a_{(\mathcal{H},\Phi)}}| = |C_1| + |C_2| = d + d^* - 2$. We now show that $C_{\mathcal{L}^a_{(\mathcal{H},\Phi)}}$ is shattered by $\mathcal{L}^a$.

Let $L = (l_1, ..., l_{d+d^*-2}) \in \{0,1\}^{d+d^*-2}$ be a potential labeling of $\mathcal{L}^a$. For every $p = (x, x^*_{d^*}, 0) \in C_1$, let $l(p)$ be the label of p according to L. Let $h \in \mathcal{H}$ be such that

[3] In Blumer et al. (1989), the dependence on k was not specified in the theorem statement; we extracted the exact constants from the proof.

for every $p = (x, x^*_{d^*}, 0) \in C_1$, $h(x) = l(p)$, and also $h(x_d) = 1$. Such an $h \in \mathcal{H}$ exists since $C_{\mathcal{H}}$ is shattered by $\mathcal{H}$. Similarly, for every $p^* = (x_d, x^*, 0) \in C_2$, let $l(p^*)$ be the label of p^* according to L. Let $\phi \in \Phi$ be such that for every $p^* = (x_d, x^*, 0) \in C_2$, $\phi(x^*) = 1 - l(p^*)$, and also $\phi(x^*_{d^*}) = 0$. Such a $\phi \in \Phi$ exists since C_Φ is shattered by Φ. We now claim that $\ell^a_{(h,\phi)}$ obtains the labeling L for $C_{\mathcal{L}^a_{(\mathcal{H},\Phi)}}$: For every $p = (x, x^*_{d^*}, 0) \in C_1$,

$$\begin{aligned}&\ell^a_{(h,\phi)}(x, x^*_{d^*}, 0) = \mathbf{1}[h(x) \neq 0 \wedge \phi(x^*_{d^*}) = 0] \\ &= \mathbf{1}[h(x) \neq 0] = l(p).\end{aligned}$$

In addition, for every $p^* = (x_d, x^*, 0) \in C_2$,

$$\begin{aligned}&\ell^a_{(h,\phi)}(x_d, x^*, 0) = \mathbf{1}[h(x_d) \neq 0 \wedge \phi(x^*) = 0] \\ &= \mathbf{1}[\phi(x^*) = 0] = l(p^*).\end{aligned}$$

We conclude that $C_{\mathcal{L}^a_{(\mathcal{H},\Phi)}}$ is shattered by $\mathcal{L}^a$. Since $|C_{\mathcal{L}^a_{(\mathcal{H},\Phi)}}| = d + d^* - 2$, $\mathrm{VC}(\mathcal{L}^a_{(\mathcal{H},\Phi)}) \geq d + d^* - 2$ as claimed. □

Denote by $(\hat{h}, \hat{\phi})$ some assignment that obtains the minimum of the optimization problem in Eq. (3). Denote the *empirical ignored weight* by

$$\widehat{\varepsilon}_{ig} := \frac{1}{m} \sum_{(x,x^*,y) \in S} \mathbf{1}[\hat{\phi}(x^*) = 1].$$

This is the fraction of training examples that are ignored due to the privileged information when minimizing the error over $\mathcal{H}$. Denote the *empirical unexplained error* by

$$\widehat{\varepsilon}_u := \frac{1}{m} \sum_{(x,x^*,y) \in S} \mathbf{1}[\ell^a_{(\hat{h},\hat{\phi})}(x, x^*, y) = 1].$$

This is the fraction of training examples that were not ignored by ϕ, but were still classified incorrectly by $\hat{h}$. The following theorem gives a generalization error bound for Privileged ERM.

Theorem 5.2. *With a probability $1 - 2\delta$ over the random choice of $S \sim \mathcal{D}^m$,*

$$\begin{aligned}&\mathrm{err}(\hat{h}, \mathcal{D}) \leq \widehat{\varepsilon}_{ig} + \widehat{\varepsilon}_u + R_{\mathrm{s}}(\widehat{\varepsilon}_{ig}, d^*) + R_{\mathrm{s}}(\widehat{\varepsilon}_u, d_a) \\ &+ R_{\mathrm{f}}(d^*) + R_{\mathrm{f}}(d_a) := \mathrm{B}_{\mathrm{PR}}.\end{aligned}$$

Proof. Let $\mathcal{D}'$ be a distribution over $(\mathcal{X} \times \mathcal{X}^* \times \mathcal{Y}) \times \{0\}$ such that the marginal over $(\mathcal{X} \times \mathcal{X}^* \times \mathcal{Y})$ is $\mathcal{D}$. Recall that $(\hat{h}, \hat{\phi})$ are minimizers of Eq. (3). Decompose the error of $\hat{h}$ as follows:

$$\begin{aligned}&\mathrm{err}(\hat{h}, \mathcal{D}) = \\ &= \mathbb{P}[\hat{h}(X) \neq Y \wedge \hat{\phi}(X^*) = 1] \\ &\quad + \mathbb{P}[\hat{h}(X) \neq Y \wedge \hat{\phi}(X^*) = 0] \\ &\leq \mathbb{P}[\hat{\phi}(X^*) = 1] + \mathbb{P}[\hat{h}(X) \neq Y \wedge \hat{\phi}(X^*) = 0] \\ &= \mathrm{err}(\hat{\phi}, \mathcal{D}') + \mathrm{err}(\ell^a_{(\hat{h},\hat{\phi})}, \mathcal{D}'), \qquad (7)\end{aligned}$$

Where we treat $\hat{\phi}$ as equivalent to $(x, x^*, y) \mapsto \hat{\phi}(x^*)$.

We will bound each of the terms on the RHS separately.

Given $S = (\,(x_i, x^*_i, y_i)\,)_{i \in [m]} \sim \mathcal{D}^m$, let $S_0 := (\,((x_i, x^*_i, y_i), 0)\,)_{i \in [m]}$, so that S_0 is distributed as an i.i.d. sample from $\mathcal{D}'$. Then $\mathrm{err}(\ell^a_{(\hat{h},\hat{\phi})}, S_0) = \widehat{\varepsilon}_u$ and $\mathrm{err}(\hat{\phi}, S_0) = \widehat{\varepsilon}_{ig}$.

By Boucheron et al. (2005), Given a sample $\tilde{S} = ((a_1, b_1), \ldots, (a_m, b_m))$ generated according to a distribution $\tilde{\mathcal{D}}$ over $\mathcal{A} \times \{0, 1\}$ and a hypothesis class $\mathcal{J} \subseteq \{0, 1\}^{\mathcal{A}}$ with VC dimension d', with probability at least $1 - \delta$, for all $g \in \mathcal{J}$, if $\hat{\varepsilon} = \mathrm{err}(g, \tilde{S})$, then

$$\mathrm{err}(g, \tilde{\mathcal{D}}) \leq \hat{\varepsilon} + R_{\mathrm{s}}(\hat{\varepsilon}) + R_{\mathrm{f}}(d'). \qquad (8)$$

Assigning $\tilde{\mathcal{D}} := \mathcal{D}'$, $\mathcal{J} := \mathcal{L}^a_{\mathcal{H},\Phi}$, $\tilde{S} := S_0$, it follows that

$$\mathrm{err}(\ell^a_{(\hat{h},\hat{\phi})}, \mathcal{D}') \leq \widehat{\varepsilon}_u + R_{\mathrm{s}}(\widehat{\varepsilon}_u, d_a) + R_{\mathrm{f}}(d_a). \qquad (9)$$

In addition, assigning $\tilde{\mathcal{D}} := \mathcal{D}'$, $\mathcal{F} := \Phi$, $\tilde{S} := S_0$, Eq. (8), we get

$$\mathrm{err}(\hat{\phi}, \mathcal{D}') \leq \widehat{\varepsilon}_{ig} + R_{\mathrm{s}}(\widehat{\varepsilon}_{ig}, d^*) + R_{\mathrm{f}}(d^*). \qquad (10)$$

Combining Eq. (9) and Eq. (10) with Eq. (7) and using the union bound, with probability at least $1 - 2\delta$,

$$\begin{aligned}\mathrm{err}(\hat{h}, \mathcal{D}) \leq\ &\widehat{\varepsilon}_{ig} + R_{\mathrm{s}}(\widehat{\varepsilon}_{ig}, d^*) + R_{\mathrm{f}}(d^*) \\ &+ \widehat{\varepsilon}_u + R_{\mathrm{s}}(\widehat{\varepsilon}_u, d_a) + R_{\mathrm{f}}(d_a),\end{aligned}$$

as claimed. □

The upper bound in Theorem 5.2 is derived using known upper bounds for ERM under bounded agnostic error. While these upper bounds are known to be tight for the zero-one loss, there does not exist an equivalent result for the loss we use for the predictions of Φ. The following theorem shows that nonetheless, the classical agnostic uniform convergence upper bound is tight also for this loss. The proof is provided in Appendix A.1.

Theorem 5.3. *Let Φ be a hypothesis class with $\mathrm{VC}(\Phi) = d^*$. For all $\varepsilon \in (0, 1)$ and $\delta < 1/128$, if the sample size is $m < (d^* - 1)/(1280 \cdot \varepsilon^2)$, then there exists a distribution $\mathcal{D}$ such that with a probability larger than δ, $\exists \phi \in \Phi$ such that $\mathbb{P}[\phi(X) = 1] - \hat{\mathbb{P}}[\phi(X) = 1] > \varepsilon$, where $\hat{\mathbb{P}}$ denotes the empirical probability based on a random i.i.d. sample of size m.*

We wish to derive conditions under which $\mathrm{B}_{\mathrm{PR}} < \mathrm{B}_{\mathrm{ERM}}$. However, each of these bounds uses different empirical measures. Our next lemma links the two sets of measures, by showing that regardless of the set of examples that are ignored, the empirical error of an ERM algorithm is smaller than the value of the minimization of Eq. (3).

Lemma 5.4. *For any $S' \subseteq S$ and for any $h \in \mathcal{H}$,*

$$m \cdot \hat{\varepsilon}_{\mathrm{ERM}} \leq \sum_{(x_i,y_i)\in S\setminus S'} \mathbf{1}[h(x_i) \neq y_i] + |S'|.$$

It follows that $\hat{\varepsilon}_{\mathrm{ERM}} \leq \widehat{\varepsilon}_u + \widehat{\varepsilon}_{ig}$.

Proof. Let $h \in \mathcal{H}$ and $S' \subseteq S$. By definition, we have $\hat{\varepsilon}_{\mathrm{ERM}} \leq \mathrm{err}(h, S)$. In addition,

$$\begin{aligned} m \cdot \mathrm{err}(h,S) &= \sum_{(x_i,y_i)\in S} \mathbf{1}[h(x_i) \neq y_i] \\ &= \sum_{(x_i,y_i)\in S\setminus S'} \mathbf{1}[h(x_i) \neq y_i] + \sum_{(x_i,y_i)\in S'} \mathbf{1}[h(x_i) \neq y_i] \\ &\leq \sum_{(x_i,y_i)\in S\setminus S'} \mathbf{1}[h(x_i) \neq y_i] + |S'|. \end{aligned}$$

This proves the claim. □

This lemma is crucial for the comparison of B_{PR} and $\mathrm{B}_{\mathrm{ERM}}$, as it implies that the only way to get $\mathrm{B}_{\mathrm{PR}} < \mathrm{B}_{\mathrm{ERM}}$ is to have smaller convergence terms in B_{PR} compared to $\mathrm{B}_{\mathrm{ERM}}$.

Next, we derive a sufficient condition for having $\mathrm{B}_{\mathrm{PR}} \leq \mathrm{B}_{\mathrm{ERM}}$. This condition considers is a best-case scenario, in the sense that it requires the privileged information to cause the ERM to ignore exactly the examples that cannot be classified correctly using $\mathcal{H}$. Under this scenario, the privileged learning bound is smaller than the ERM bound if the unexplained error is sufficiently small. This theorem is proved in Appendix A.2.

Theorem 5.5. *Suppose that $\hat{\varepsilon}_{\mathrm{ERM}} = \widehat{\varepsilon}_{ig} + \widehat{\varepsilon}_u$. Assume that δ is fixed and d, d^*, d_a are large. Then, if*

$$\sqrt{\widehat{\varepsilon}_u} \leq \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} \cdot \Theta\left(\frac{\sqrt{d} - \sqrt{d^*}}{\sqrt{d_a}}\right) - \sqrt{\frac{\log(m)}{m}} \cdot \Theta\left(\frac{d_a + d^* - d}{\sqrt{d_a}}\right),$$

then $\mathrm{B}_{\mathrm{PR}} \leq \mathrm{B}_{\mathrm{ERM}}$.

Note that by Lemma 5.1, the second term is necessarily positive.

The sufficient condition above is stricter when the empirical error is smaller. In particular, in the realizable case, where $\hat{\varepsilon}_{\mathrm{ERM}} = 0$, this sufficient condition never holds. In addition, the sufficient condition can only hold if $d^* \leq d$ (in addition to a small sample size). Therefore, this does not allow a privileged class Φ of a large capacity. Indeed, the following result shows that in general, d^* cannot be much larger than d while still allowing $\mathrm{B}_{\mathrm{PR}} \leq \mathrm{B}_{\mathrm{ERM}}$.

Theorem 5.6. *For any fixed $\delta \in (0,1)$, if $\mathrm{B}_{\mathrm{PR}} \leq \mathrm{B}_{\mathrm{ERM}}$ then*

$$d^* \leq 2.25 \cdot d + o(1).$$

The convergence of the last term is with respect to the growth of d, d^* together and/or of m.

This theorem is proved in Appendix A.3.

6 DISCUSSION

Our work shows that the Privileged ERM approach for privileged learning suffers an inherent capacity limit on the privileged information class in the case of the zero-one loss. This analysis is relevant when ERM can be accurately executed and pertains to the *statistical* benefits of privileged learning. However, when surrogate losses are used, the situation may be quite different. In these cases, privileged learning may have a *computational* advantage, as it may be possible to use privileged information to make the computational problem of minimizing the true loss easier. For instance, if privileged information allows identifying outliers, and thus helps to ignore some training examples in a way that would make the optimization objective of the surrogate loss closer to that of the true target loss, the resulting training error could be lower, leading to a lower true error. We plan to study this promising direction in future work.

We further note that our analysis only provides a limitation on the dimension of the privileged information under worst-case analysis and within a specific privileged-ERM framework. Studying other variants of this framework may lead to less restrictive results.

References

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

Zhifan Gao, Sitong Wu, Zhi Liu, Jianwen Luo, Heye Zhang, Mingming Gong, and Shuo Li. Learning the implicit strain reconstruction in ultrasound elastography using privileged information. *Medical image analysis*, 58: 101534, 2019.

Maksim Lapin, Matthias Hein, and Bernt Schiele. Learning using privileged information: Svm+ and weighted svm. *Neural Networks*, 53:95–108, 2014.

Wonkyung Lee, Junghyup Lee, Dohyung Kim, and Bumsub Ham. Learning with privileged information for efficient image super-resolution. In *European Conference on Computer Vision*, pages 465–482. Springer, 2020.

Yan Li, Fanqing Meng, and Jun Shi. Learning using privileged information improves neuroimaging-based cad of alzheimer's disease: a comparative study. *Medical & biological engineering & computing*, 57(7):1605–1616, 2019.

Rahul Pasunuri, Phillip Odom, Tushar Khot, Kristian Kersting, and Sriraam Natarajan. Learning with privileged information: Decision-trees and boosting. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) Workshop*, 2016.

Dmitry Pechyony and Vladimir Vapnik. On the theory of learnining with privileged information. In *Advances in neural information processing systems*, pages 1894–1902, 2010a.

Dmitry Pechyony and Vladimir Vapnik. On the theory of learning with privileged information (full version), 01 2010b. URL `https://www.researchgate.net/publication/228565062_On_the_Theory_of_Learning_with_Privileged_Information_Full_version`.

Zhiquan Qi, Yingjie Tian, Lingfeng Niu, and Bo Wang. Semi-supervised classification with privileged information. *International Journal of Machine Learning and Cybernetics*, 6(4):667–676, 2015.

Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *Journal of machine learning research*, 16(2023-2049):2, 2015.

Vladimir Vapnik and Rauf Izmailov. Knowledge transfer in svm and neural networks. *Annals of Mathematics and Artificial Intelligence*, 81(1-2):3–19, 2017.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris. Active privileged learning of human activities from weakly labeled samples. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3036–3040. IEEE, 2016.

Xun Yang, Meng Wang, Luming Zhang, and Dacheng Tao. Empirical risk minimization for metric learning using privileged information. In *IJCAI*, pages 2266–2272, 2016.

Shanxin Yuan, Bjorn Stenger, and Tae-Kyun Kim. 3d hand pose estimation from rgb using privileged learning with depth data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

A DEFERRED PROOFS

Below, we provide deferred proofs for theorems stated above. Appendix A.1 provides the proof of the lower bound, Theorem 5.3, Appendix A.2 provides the proof of the sufficient condition, Theorem 5.5 and Appendix A.3 provides the proof of the necessary condition, Theorem 5.6.

A.1 Proof Of The Lower Bound For The Privileged Learning Loss

Proof of Theorem 5.3. This proof is an adaptation of the proof of the lower bound for the zero-one loss given in Anthony and Bartlett (2009, Theorem 5.2) to our setting. The main challenge is constructing a set of distributions that can only be distinguished using a worst-case number of samples. This is achieved using the following new construction.

Since Φ has VC-dimension d^* , there is a set $C = \{x_1^*, ..., x_{d^*}^*\}$ of d^* examples that is shattered by Φ. For simplicity, assume that d^* is an even number. If d^* is odd, then the proof below holds for $d^* - 1$. We partition the set C into $d^*/2$ pairs $\{(a_i, b_i)\}_{i\in[d^*/2]}$. Consider the class of all distributions $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ with the following properties:

- $\mathcal{D}$ assigns a zero probability to all sets not intersecting $C \times \{0, 1\}$.
- For $x \in \mathcal{X}$, denote $\mathcal{D}(x) = \mathbb{P}_{(X,Y)\sim\mathcal{D}}[X = x]$. Set $\alpha := \frac{8\varepsilon}{(1-8\delta)}$. For each $i = 1, 2, ..., d^*/2$ and a pair (a_i, b_i) in the partition of C, either:
  - $\mathcal{D}(a_i) = \frac{1+\alpha}{d^*}$ and $\mathcal{D}(b_i) = \frac{1-\alpha}{d^*}$, or
  - $\mathcal{D}(b_i) = \frac{1+\alpha}{d^*}$ and $\mathcal{D}(a_i) = \frac{1-\alpha}{d^*}$.

Let $\Phi' \subseteq \Phi$ be the set including all hypotheses ϕ such that for each pair (a_i, b_i), ϕ maps one of the elements in the pair to 0 and the other to 1.

Given $\mathcal{D}$, Let $\phi^* \in \Phi'$ be the function such that for each pair (a_i, b_i), $\phi^*(a_i) = 1$ if and only if $\mathcal{D}(a_i) = \frac{1-\alpha}{d^*}$. Then,

$$\mathbb{P}[\phi^*(X) = 1] = \sum_{i=1}^{d^*/2} \frac{1-\alpha}{d^*} = \frac{1-\alpha}{2}.$$

Furthermore, for any $\phi \in \Phi'$, we have

$$\mathbb{P}[\phi(X) = 1] = \sum_{i=1}^{d^*/2} (\frac{1+\alpha}{d^*}\mathbf{1}[\phi(a_i) \neq \phi^*(a_i)] + \frac{1-\alpha}{d^*}\mathbf{1}[\phi(a_i) = \phi^*(a_i)]) = \mathbb{P}[\phi^* = 1] + \frac{2\alpha}{d^*}\sum_{i=1}^{d^*/2} \mathbf{1}[\phi(a_i) \neq \phi^*(a_i)].$$

For any sample $S \in S^m$, let $N(S) = (N_1(S), ..., N_{d^*/2}(S))$, where $N_i(S)$ is the number of occurrences of either a_i or b_i in S. Then, letting L be a learning algorithm for Φ', we have that for $\hat{\phi} := L(S)$,

$$\begin{aligned}\mathbb{E}[\frac{2}{d^*}\sum_{i=1}^{d^*/2} \mathbf{1}[\hat{\phi}(a_i) \neq \phi^*(a_i)]] &= \frac{2}{d^*}\mathbb{E}[\sum_{i=1}^{d^*/2} \mathbf{1}[\hat{\phi}(a_i) \neq \phi^*(a_i)]] \\ &= \frac{2}{d^*}\sum_N \sum_{i=1}^{d^*/2} \mathbb{P}[\hat{\phi}(a_i) \neq \phi^*(a_i) \mid N(S) = N] \cdot \mathbb{P}[N(S) = N].\end{aligned}$$

where $N = (N_1, ..., N_{d^*/2})$ ranges over the set of $d^*/2$-tuples of positive integers with $\sum_{i=1}^{d^*/2} N_i = m$.

Similarly to the proof of Anthony and Bartlett (2009, Theorem 5.2), we can conclude that if $m < \frac{d^*}{320\cdot\varepsilon^2}$, then with a probability larger than $1/64$ over samples $S \sim \mathcal{D}^m$, $\mathbb{P}[\hat{\phi}(X) = 1] - \mathbb{P}[\phi^*(X) = 1] > \varepsilon$. In particular, this holds for $\hat{\phi} = \text{argmin}_{\phi\in\Phi'} \hat{\mathbb{P}}[\phi(X) = 1]$.

Let $m < \frac{d^*}{1280\cdot\varepsilon^2}$. By the conclusion above, we have that with a probability larger than δ over samples $S \sim \mathcal{D}^m$, $\mathbb{P}[\hat{\phi}(X) = 1] - \mathbb{P}[\phi^*(X) = 1] > 2\varepsilon$.

We now claim that at least one of the following holds with a probability larger than δ:

- $|\mathbb{P}[\hat{\phi}(X)=1]-\hat{\mathbb{P}}[\hat{\phi}(X)=1]|>\varepsilon$;
- $|\mathbb{P}[\phi^*(X)=1]-\hat{\mathbb{P}}[\phi^*(X)=1]|>\varepsilon$.

Assume in contradiction that each of these inequalities holds with a probability at most δ. Then, with a probability at least $1-2\delta$,

$$\mathbb{P}[\hat{\phi}(X)=1]-\varepsilon<\hat{\mathbb{P}}[\hat{\phi}(X)=1$$

and

$$\hat{\mathbb{P}}[\phi^*(X)=1]<\mathbb{P}[\phi^*(X)=1]+\varepsilon.$$

Also, from the definition of $\hat{\phi}$, we have $\hat{\mathbb{P}}[\hat{\phi}(X)=1]\leq\hat{\mathbb{P}}[\phi^*(X)=1]$. We get that with a probability at least $1-2\delta$, $\mathbb{P}[\hat{\phi}(X)=1]-\mathbb{P}[\phi^*(X)=1]<2\varepsilon$. Since $\delta<1/128$ and $m<\frac{d^*}{1280\cdot\varepsilon^2}$, this contradicts the lower bound above. It follows that at least one of the assumed inequalities above holds, which proves the claim. □

A.2 Proof Of The Sufficient Condition

We now prove Theorem 5.5. We derive a sufficient condition for the following inequality to hold:

$$\mathrm{B}_{\mathrm{PR}}=\widehat{\varepsilon}_{ig}+\widehat{\varepsilon}_u+R_{\mathrm{s}}(\widehat{\varepsilon}_{ig},d^*)+R_{\mathrm{s}}(\widehat{\varepsilon}_u,d_a)+R_{\mathrm{f}}(d^*)+R_{\mathrm{f}}(d_a)\leq\hat{\varepsilon}_{\mathrm{ERM}}+R_{\mathrm{s}}(\hat{\varepsilon}_{\mathrm{ERM}},d)+R_{\mathrm{f}}(d)=\mathrm{B}_{\mathrm{ERM}}.$$

Under the assumption that $\hat{\varepsilon}_{\mathrm{ERM}}=\widehat{\varepsilon}_{ig}+\widehat{\varepsilon}_u$, it suffices to have

$$R_{\mathrm{s}}(\widehat{\varepsilon}_{ig},d^*)+R_{\mathrm{s}}(\widehat{\varepsilon}_u,d_a)+R_{\mathrm{f}}(d^*)+R_{\mathrm{f}}(d_a)\leq R_{\mathrm{s}}(\hat{\varepsilon}_{\mathrm{ERM}},d)+R_{\mathrm{f}}(d).$$

This is equivalent to

$$\sqrt{\widehat{\varepsilon}_{ig}}R_{\mathrm{s}}(1,d^*)+\sqrt{\widehat{\varepsilon}_u}R_{\mathrm{s}}(1,d_a)+R_{\mathrm{f}}(d^*)+R_{\mathrm{f}}(d_a)\leq\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}}R_{\mathrm{s}}(1,d)+R_{\mathrm{f}}(d).$$

Since $\widehat{\varepsilon}_{ig}\leq\hat{\varepsilon}_{\mathrm{ERM}}$, it suffices to have

$$\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}}R_{\mathrm{s}}(1,d^*)+\sqrt{\widehat{\varepsilon}_u}R_{\mathrm{s}}(1,d_a)+R_{\mathrm{f}}(d^*)+R_{\mathrm{f}}(d_a)\leq\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}}R_{\mathrm{s}}(1,d)+R_{\mathrm{f}}(d),$$

which is equivalent to

$$\sqrt{\widehat{\varepsilon}_u}\leq\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}}\cdot\frac{R_{\mathrm{s}}(1,d)-R_{\mathrm{s}}(1,d^*)}{R_{\mathrm{s}}(1,d_a)}+\frac{R_{\mathrm{f}}(d)-R_{\mathrm{f}}(d^*)-R_{\mathrm{f}}(d_a)}{R_{\mathrm{s}}(1,d_a)}.$$

For a fixed δ and large $d,d*,d_a$, this is equivalent to

$$\sqrt{\widehat{\varepsilon}_u}\leq\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}}\cdot\Theta(\frac{\sqrt{d}-\sqrt{d^*}}{\sqrt{d_a}})-\sqrt{\frac{\log(m)}{m}}\cdot\Theta(\frac{d_a+d^*-d}{\sqrt{d_a}}),$$

as claimed.

A.3 Proof Of The Necessary Condition

We now prove Theorem 5.6. First, we prove an additional lemma that provides a necessary condition for $\mathrm{B}_{\mathrm{PR}}\leq\mathrm{B}_{\mathrm{ERM}}$.

Lemma A.1. *For any fixed $\delta\in(0,1)$, if $\mathrm{B}_{\mathrm{PR}}\leq\mathrm{B}_{\mathrm{ERM}}$ then*

$$\sqrt{\widehat{\varepsilon}_u}\leq\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}}\cdot\frac{\sqrt{d}}{\sqrt{d_a}}-\sqrt{\widehat{\varepsilon}_{ig}}\cdot\frac{\sqrt{d^*}}{\sqrt{d_a}}+o(1),$$

The convergence of the last term is with respect to the growth of d,d^* together and/or of m.

Proof. By the assumption of the lemma, $\mathrm{B}_{\mathrm{PR}}\leq\mathrm{B}_{\mathrm{ERM}}$. Thus, by definition,

$$\widehat{\varepsilon}_{ig}+\widehat{\varepsilon}_u+R_{\mathrm{s}}(\widehat{\varepsilon}_{ig},d^*)+R_{\mathrm{s}}(\widehat{\varepsilon}_u,d_a)+R_{\mathrm{f}}(d^*)+R_{\mathrm{f}}(d_a)\leq\hat{\varepsilon}_{\mathrm{ERM}}+\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}}R_{\mathrm{s}}(1,d)+R_{\mathrm{f}}(d).$$

This is equivalent to:

$$\widehat{\varepsilon}_{ig} + \widehat{\varepsilon}_u + \sqrt{\widehat{\varepsilon}_{ig}} R_{\mathrm{s}}(1, d^*) + \sqrt{\widehat{\varepsilon}_u} R_{\mathrm{s}}(1, d_a) + R_{\mathrm{f}}(d^*) + R_{\mathrm{f}}(d_a) \leq \hat{\varepsilon}_{\mathrm{ERM}} + \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} R_{\mathrm{s}}(1, d) + R_{\mathrm{f}}(d).$$

Therefore,

$$\begin{aligned}\sqrt{\widehat{\varepsilon}_u} R_{\mathrm{s}}(1, d_a) &\leq \hat{\varepsilon}_{\mathrm{ERM}} - (\widehat{\varepsilon}_{ig} + \widehat{\varepsilon}_u) + \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} R_{\mathrm{s}}(1, d) - \sqrt{\widehat{\varepsilon}_{ig}} R_{\mathrm{s}}(1, d^*) + R_{\mathrm{f}}(d) - R_{\mathrm{f}}(d_a) - R_{\mathrm{f}}(d^*) \\ &\leq \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} R_{\mathrm{s}}(1, d) - \sqrt{\widehat{\varepsilon}_{ig}} R_{\mathrm{s}}(1, d^*) + R_{\mathrm{f}}(d) - R_{\mathrm{f}}(d^*) - R_{\mathrm{f}}(d_a).\end{aligned}$$

The last inequality follows since by Lemma 5.4, $\hat{\varepsilon}_{\mathrm{ERM}} \leq \widehat{\varepsilon}_{ig} + \widehat{\varepsilon}_u$.

Now, from the definition of R_{f}, we have

$$R_{\mathrm{f}}(d) - R_{\mathrm{f}}(d^*) - R_{\mathrm{f}}(d_a) = 8\frac{\log(m+1)}{m}(d - d^* - d_a) - \frac{4\log(\frac{4}{\delta})}{m}.$$

By Lemma 5.1, $d_a \geq d + d^* - 2$. Thus, $d - d^* - d_a < 0$. It follows that $R_{\mathrm{f}}(d) - R_{\mathrm{f}}(d^*) - R_{\mathrm{f}}(d_a) < 0$. Therefore,

$$\sqrt{\widehat{\varepsilon}_u} R_{\mathrm{s}}(1, d_a) \leq \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} R_{\mathrm{s}}(1, d) - \sqrt{\widehat{\varepsilon}_{ig}} R_{\mathrm{s}}(1, d^*).$$

It follows that

$$\sqrt{\widehat{\varepsilon}_u} \leq \frac{\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} R_{\mathrm{s}}(1, d) - \sqrt{\widehat{\varepsilon}_{ig}} R_{\mathrm{s}}(1, d^*)}{R_{\mathrm{s}}(1, d_a)} = \frac{\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}(d + A)} - \sqrt{\widehat{\varepsilon}_{ig}(d^* + A)}}{\sqrt{d_a + A}},$$

Where $A := \log(4/\delta)/(2\log(m+1))$. Thus,

$$\sqrt{\widehat{\varepsilon}_u} \leq \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} \cdot \frac{\sqrt{d}}{\sqrt{d_a}} - \sqrt{\widehat{\varepsilon}_{ig}} \cdot \frac{\sqrt{d^*}}{\sqrt{d_a}} + o(1),$$

where convergence of the last term is with respect to the growth of d, d^* together and/or of m. □

Next, we prove the theorem using the two lemmas above.

Proof of Theorem 5.6. Assume that $\mathrm{B}_{\mathrm{PR}} \leq \mathrm{B}_{\mathrm{ERM}}$. By Lemma A.1,

$$\sqrt{\widehat{\varepsilon}_u} \leq \sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} \cdot \frac{\sqrt{d}}{\sqrt{d_a}} - \sqrt{\widehat{\varepsilon}_{ig}} \cdot \frac{\sqrt{d^*}}{\sqrt{d_a}} + o(1).$$

Denote $\alpha := d^*/d$. Suppose that $\alpha \geq 1$ (otherwise the statement in the theorem clearly holds). We have

$$\sqrt{\widehat{\varepsilon}_u} \leq \frac{\sqrt{d}}{\sqrt{d_a}} \cdot (\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} - \sqrt{\alpha} \cdot \sqrt{\widehat{\varepsilon}_{ig}}) + o(1).$$

Here, the convergence is under a fixed α with growing d, d^* or m. Since $d_a \geq d^* + d - 2 = (1+\alpha) \cdot d - 2$, we have

$$\sqrt{\widehat{\varepsilon}_u} \leq \frac{1}{\sqrt{1+\alpha}} \cdot (\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} - \sqrt{\alpha} \cdot \sqrt{\widehat{\varepsilon}_{ig}}) + o(1). \tag{11}$$

Since $\sqrt{\widehat{\varepsilon}_u} \geq 0$, we have

$$\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} - \sqrt{\alpha} \cdot \sqrt{\widehat{\varepsilon}_{ig}} + o(1) \geq 0.$$

Thus, $\widehat{\varepsilon}_{ig} \leq \hat{\varepsilon}_{\mathrm{ERM}}/\alpha + o(1)$. Combining with Lemma 5.4, we get

$$\hat{\varepsilon}_{\mathrm{ERM}} \leq \widehat{\varepsilon}_{ig} + \widehat{\varepsilon}_u \leq \hat{\varepsilon}_{\mathrm{ERM}}/\alpha + \widehat{\varepsilon}_u + o(1).$$

Combining this with Eq. (11), it follows that

$$\sqrt{(1 - \frac{1}{\alpha}) \cdot \hat{\varepsilon}_{\mathrm{ERM}}} \leq \sqrt{\widehat{\varepsilon}_u} \leq \frac{1}{\sqrt{1+\alpha}} \cdot (\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}} - \sqrt{\alpha} \cdot \sqrt{\widehat{\varepsilon}_{ig}}) + o(1).$$

Rearranging, we get

$$\sqrt{1+\alpha}\cdot(1-\frac{1}{\alpha})\leq 1-\sqrt{\alpha}\cdot\frac{\sqrt{\widehat{\varepsilon}_{ig}}}{\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}}}+o(1).$$

This leads to

$$\sqrt{\alpha}\frac{\sqrt{\widehat{\varepsilon}_{ig}}}{\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}}}\leq 1-\sqrt{1+\alpha}\cdot(1-\frac{1}{\alpha})+o(1).$$

Since $0\leq\frac{\sqrt{\widehat{\varepsilon}_{ig}}}{\sqrt{\hat{\varepsilon}_{\mathrm{ERM}}}}$, it must hold that $\sqrt{1+\alpha}\cdot(1-\frac{1}{\alpha})\leq 1+o(1)$. Solving for α, we obtain that $\alpha\leq 2.25+o(1)$. Since $\alpha=d^*/d$, we conclude that $d^*\leq 2.25\cdot d+o(1)$, as claimed. □