Electronic Journal of Statistics
Vol. 16 (2022) 6086–6134
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2084

Statistical learning from biased training samples

Stephan Clémençon

LTCI, Télécom Paris, Institut Polytechnique de Paris, France
e-mail: stephan.clemencon@telecom-paris.fr

Pierre Laforgue

Università degli Studi di Milano, Milan, Italy
e-mail: pierre.laforgue@unimi.it

Abstract: With the deluge of digitized information in the Big Data era, massive datasets are becoming increasingly available for learning predictive models. However, in many practical situations, the poor control of the data acquisition processes may naturally jeopardize the outputs of machine learning algorithms, and selection bias issues are now the subject of much attention in the literature. The present article investigates how to extend Empirical Risk Minimization, the principal paradigm in statistical learning, when training observations are generated from biased models, i.e., from distributions that are different from that in the test/prediction stage, and absolutely continuous with respect to the latter. Precisely, we show how to build a "nearly debiased" training statistical population from biased samples and the related biasing functions, following in the footsteps of the approach originally proposed in [46]. Furthermore, we study from a nonasymptotic perspective the performance of minimizers of an empirical version of the risk computed from the statistical population thus created. Remarkably, the learning rate achieved by this procedure is of the same order as that attained in absence of selection bias. Beyond the theoretical guarantees, we also present experimental results supporting the relevance of the algorithmic approach promoted in this paper.

MSC2020 subject classifications: Primary 62C12; secondary 62D99.
Keywords and phrases: Statistical learning theory, learning under sample selection bias, bias sampling models, nonasymptotic generalization bounds.

Received October 2021.

Contents

1. Introduction

In the standard setting of binary classification, the flagship problem in statistical learning, $Z = (X, Y)$ is a random pair defined on a probability space with unknown probability distribution P. The random vector X, valued in $\mathcal{X} \subset \mathbb{R}^d$, models some information supposedly useful to predict the random binary label Y, taking its values in $\{-1, +1\}$. The objective is to build a Borelian predictive function, i.e., a classifier, $g : \mathcal{X} \to \{-1, +1\}$ that minimizes the error probability, i.e., the risk, of the decision: $L_P(g) = \mathbb{P}\{Y \neq g(X)\}$. It is well-known that the optimal solution is given by the Bayes classifier $g^*(x) = 2\mathbb{I}\{\eta(x) \geq 1/2\} - 1$, where $\eta(X) = \mathbb{P}\{Y = 1 \mid X\}$ denotes the posterior probability, with minimum risk $L_P^* = \mathbb{E}[\min\{\eta(X),\ 1 - \eta(X)\}]$. In practice however, P (and consequently η) is usually unknown, and one generally resorts to a training dataset $\mathcal{D}_n = \{(X_1, Y_1),\ \ldots,\ (X_n, Y_n)\}$, composed of $n \geq 1$ independent copies of (X, Y). Empirical Risk Minimization (ERM in short, see e.g., [12]) consists in solving the minimization problem $\min_{g \in \mathcal{G}} \widehat{L}_n(g)$, where $\widehat{L}_n(g)$ is a statistical estimator of the risk $L_P(g)$, generally obtained by replacing P in L_P with the empirical distribution of the (X_i, Y_i)'s, and $\mathcal{G}$ is a class of predictive rules hopefully rich enough to contain an accurate approximant of g^*. In this case, the empirical risk is the statistical average $\widehat{L}_n(g) = (1/n) \sum_{i=1}^n \mathbb{I}\{Y_i \neq g(X_i)\}$, denoting by $\mathbb{I}\{\mathcal{E}\}$ the indicator function of any event $\mathcal{E}$. Under various assumptions controlling the complexity of the class $\mathcal{G}$ over which the learning task is achieved (e.g., finite VC dimension, metric entropy, or Rademacher complexity), the performance of empirical risk minimizers (i.e., solutions to the ERM problem), measured through the excess of risk $g \mapsto L_P(g) - L_P^*$, can be classically studied by means of concentration inequalities for empirical processes, see e.g., [5]. Although very informative in i.i.d. settings, these generalization results nonetheless crucially rely upon the assumption that training observations are sampled from the true test distribution, which is often violated in practice. Motivated by the poor control of the data acquisition process in many applications (see e.g., [44]), the purpose of the present article is to investigate ERM in the presence of sample

selection bias, that is to say in the situation where the samples at disposal for learning a predictive rule g are not distributed as P, which can be viewed as a very specific case of Transfer Learning, see [3]. As recently highlighted by [4], [51] or [6] among others, representativeness issues do not vanish simply under the effect of the size of the training set. Hence, ignoring selection bias issues may dramatically jeopardize the accuracy of the outputs of machine learning algorithms. The method we propose stands out from previous approaches for two main reasons: (1) it encompasses a wide range of biasing scenarios, (2) it applies to biased training distributions that may not dominate the test distribution.

Selection bias can be due to a wide variety of causes (e.g., the use of a survey scheme to collect observations, censorship, truncation, see for instance [20] or [47]), and the study of its impact on inference methods, as well as techniques to remedy it, have a very long history in statistics. Depending on the nature of the mechanism causing the sample selection bias, and on that of the statistical information available to the learner, particular cases have been considered in the machine learning literature, for which specific approaches have been developed. For instance, the case where some errors occur among the labels of the training data is studied in [24], while in [29] ERM is extended to the framework of survey training data (when inclusion probabilities are known). In [41] and [2], authors consider statistical learning of regression models in the context of right censored training observations. Recently, a very special case of sample selection bias, referred to as *covariate shift*, has been the subject of a good deal of attention (though it had been already considered by [25] in a simplified version). In this case, addressing the sample selection bias issue is made much easier by the hypothesis stipulating that, in supervised problems, only the marginal input distribution may possibly change, the conditional distribution of the output Y given the input observation X being the same in the learning and predictive stages. One may refer to [36], [39] and [22], or to the monographs [32] and [38]. In contrast to the aforementioned settings, the procedure exposed here allows for much more complex biasing mechanisms. Specifically, survey schemes and censorship scenarios can be seen as particular instances of our framework, see Examples 2 and 3. Moreover, bias may apply to covariates, labels, or both at the same time without altering the guarantees. We emphasize that despite focus has been put on supervised learning for the sake of clarity, the presented debiasing approach remains valid for unsupervised tasks, as long as they build upon ERM, see Examples 4 and 5.

Methods dedicated to correct sample bias usually boil down to reweighting the training observations with appropriate weights, based on the Importance Sampling approach, or according to the Inverse Probability Weighting technique (IPW in abbreviated form, see e.g., [13] or [49] in the context of linear regression models), rather than using uniform weights. For instance, these weights are the inverses of the first order inclusion probabilities in the case where data are acquired by means of a survey plan, see [9] and [29], or the inverses of estimates of the probability of not being censored when data suffer from random censorship, see [2] and the references therein. Side information about the cause of the selection may also be used to derive explicit forms for the appropriate

weights, see e.g., [50], [34] in a semi-supervised framework, [14] in the context of maximum entropy density estimation, or [23] for the adaptation of the SVM algorithm to certain selection bias situations. More generally, if the Radon-Nikodym derivative of the test distribution P w.r.t. the training distribution Q (supposedly dominating P) is known, one may simply reweight each training observation z by $(dP/dQ)(z)$ in order to get an unbiased estimate of the true risk. However, this method may be inapplicable, as soon as P is not absolutely continuous w.r.t. Q. To bypass this limitation, several techniques have been developed, based for instance on the discrepancy distance between P and Q, see [26] and references therein, or on their Rényi divergence, see e.g., [11]. We point out that statistical learning based on biased samples can be viewed as a very specific case of *transfer learning*, see [37], but also e.g., [28] and [48]. Several recent works in this area also provide theoretical analyses for particular machine learning tasks without requiring the absolute continuity condition, at the cost of additional restrictive assumptions however. Hence, a no-free-lunch theorem for multitask learning is established in [19], as well as a method to aggregate the datasets if the task distributions have small discrepancies with respect to the target distribution, see the transfer exponent condition therein. In [8, 40] the authors assume that the tasks share an (approximately) common data representation, while [7] analyzes the specific posterior drift model, i.e., it is assumed that the distributions of the covariates remain the same. Finally, [33] studies transfer learning for binary classification under several assumptions on the transfer mechanism, the marginal distributions and their smoothness. We highlight that none of these assumptions is made in the present paper.

The perspective embraced in the present paper is quite different. We consider multiple biased training distributions, none of them being assumed to dominate P. In particular, the variance of the Radon-Nikodym derivatives $(dP/dQ)(Z)$ are not supposed to be bounded, in contrast to [11]. Instead, we leverage training samples drawn from these biased distributions and show how to combine them in order to construct an unbiased estimate of the target distribution P, under mild identifiability hypotheses. The debiasing weights are defined as solutions to a nontrivial system of equations, and do not enjoy any simple closed-form expressions in general, in contrast to those used in the context of survey schemes or censorship models. Precisely, we focus on the case where statistical learning is based on training data sampled from *biased sampling models*, as originally introduced in [46] in the context of asymptotic nonparametric estimation of cumulative distribution functions, see also [16]. This very general selection bias framework accounts for many situations encountered in practice, covering for instance the (far from uncommon) situation where the samples available to learn a binary classifier $g(x)$ are sampled from conditional distributions of (X, Y) given that X lies in specific subsets of the input space $\mathcal{X}$ (assuming that the union of these subsets is equal to X's support). In this setting, we extend ERM to the case of biased training data with nonasymptotic guarantees about their generalization ability. We propose to build an unbiased empirical estimator of the test distribution P by solving a generally nontrivial system of equations, which we use to compute a "nearly unbiased" risk estimate. We then establish

a tail probability bound for the maximal deviations between the risk functional and the estimate thus constructed. Based on this result, we finally prove that minimizers of the "debiased empirical risk" achieve learning rate bounds that are of the same order as those attained by empirical risk minimizers in absence of any bias mechanism. If our approach builds on the distribution estimation procedure for biased sampling models introduced in [16], note that the latter work is restricted to the asymptotic study of cumulative distribution functions. In contrast, we provide the first —to the best of our knowledge— nonasymptotic guarantees for this approach, in a much more general framework. This allows us to devise an extension of the ERM paradigm to biased training datasets with provable finite-sample guarantees, as required in the statistical learning literature. For the sake of completeness, the notion of biased sampling model is recalled at length in Section 2.1 and the slightly stronger assumptions needed to carry out a nonasymptotic analysis are detailed and discussed in Section 2.2. We also present results from various numerical experiments, based on synthetic and real data, that provide strong empirical evidence of the relevance of the approach we propose. If the fact that knowledge of the biasing functions is required can be seen at first glance as a limitation of the framework developed, one should have in mind that absolutely no learning strategy with statistical guarantees can be designed in absence of any understanding of the biasing mechanism at work. Moreover, it is actually far from uncommon in practice that the latter is known (e.g., one may know the types of images that are more easily collected, or the profiles of individuals who most likely answer a questionnaire). Yet, the situation where the biasing mechanism is only approximately known is of considerable interest in practice, and investigating to which extent the statistical guarantees established in this paper are preserved will be the subject of future research.

The rest of the article is structured as follows. In Section 2, basics on biased sampling models are briefly recalled, and the framework for statistical learning based on biased training samples is described at length, as well as the algorithmic approach extending the ERM methodology to this setting. In Section 3, the main theoretical results of this paper, guaranteeing the generalization capacity of ERM under selection bias, are stated. Illustrative experiments are displayed in Section 4, while technical details are deferred to the Appendix section.

2. Background and preliminaries

We first recall in Section 2.1 the *biased sampling models* framework developed in [46] and [16] for asymptotic estimation of cumulative distribution functions. Next, we present in Section 2.2 our approach to generalize ERM to the case where training data samples are drawn from such models. Here and throughout, we denote by δ_a the Dirac mass at any point a, by $\|U\|_{\sup}$ the essential supremum of any real-valued random variable (r.v.) U, and by $\textsc{Supp}(P)$ the support of any probability distribution P. Vectors are denoted by bold characters, e.g., $\boldsymbol{v} \in \mathbb{R}^K = (v_1, \ldots, v_K)$ for $K \in \mathbb{N}$. The Euclidean and sup norms are denoted by $\|\cdot\|_2$ and $\|\cdot\|_\infty$, such that $\|\boldsymbol{v}\|_2^2 = \sum_{k=1}^K v_k^2$, and $\|\boldsymbol{v}\|_\infty = \max_{k \leq K} |v_k|$.

2.1. Biased sampling models – The statistical framework

Let Z be a random vector, taking its values in $\mathcal{Z} \subset \mathbb{R}^q$, where $q \in \mathbb{N}$, with unknown probability distribution P. If independent copies $Z_1, \ldots, Z_n$ of Z were at disposal, a natural estimator of P would be the raw empirical distribution $(1/n)\sum_{i=1}^n \delta_{Z_i}$. In *biased sampling models*, as defined in [46], one cannot rely on such observations. Instead, statistical inference must be based on $K \geq 1$ independent biased i.i.d. samples $\mathcal{D}_k = \{Z_{k,1}, \ldots, Z_{k,n_k}\}$, of size $n_k \geq 1$. We denote by $n = \sum_{k=1}^K n_k$ the size of the pooled sample, and by $\hat{\lambda}_k = n_k/n$ the proportion of each sample among the total population. For $k \leq K$, the distribution P_k of the $Z_{k,i}$ is assumed to be absolutely continuous w.r.t. the test distribution P, and related to it through a known nonnegative biasing function ω_k such that

$$\forall k \leq K,\ \forall z \in \mathcal{Z}, \qquad \frac{dP_k}{dP}(z) = \frac{\omega_k(z)}{\Omega_k}\,, \tag{2.1}$$

where $\Omega_k = \mathbb{E}_P[\omega_k(Z)] = \int \omega_k(z) dP(z)$. We emphasize that, just like P, the Ω_k's are unknown. Note that in the case of interest where $\omega_k(Z) = \mathbb{I}\{Z \in \mathcal{Z}_k\}$ for $\mathcal{Z}_k \subset \mathcal{Z}$, see Example 2 below for instance, it is much easier to know —or guess— the biasing functions ω_k (or equivalently the subsets $\mathcal{Z}_k$ in which the observations lie), rather than having access to the Ω_k's. Estimating the Ω_k's is incidentally at the core of our debiasing procedure, see e.g., Proposition 3. We further emphasize that, unlike the Ω_k's, knowing the biasing functions ω_k does not provide any information about the target distribution P. In particular, knowing a stratum $\mathcal{Z}_k$ which the observations belong to does not imply in any way that one has access to the conditional distribution $P_k = P(\cdot \mid Z \in \mathcal{Z}_k)$.

The statistical framework defined by Equations (2.1) has been considered in [16] for nonparametric estimation of a univariate cumulative distribution function (cdf). Under mild assumptions, it is shown therein that a consistent and asymptotically normal estimator of P can be constructed from the biased samples $\mathcal{D}_k$ and the knowledge of the biasing functions ω_k, for $k \leq K$. The first fundamental assumption, referred to as Assumption S in [16], guarantees identifiability. It can be formulated as follows.

Assumption 1. *The union of the supports of the biased distributions P_k is equal to the support of distribution P:*

$$\bigcup_{k=1}^K \Big\{ z \in \mathcal{Z} :\ \omega_k(z) > 0 \Big\} = \textsc{Supp}(P)\,.$$

Of course, we have by definition $\bigcup_{k=1}^K \textsc{Supp}(P_k) \subset \textsc{Supp}(P)$. If this inclusion is strict, some parts of $\textsc{Supp}(P)$ shall never be covered by observations sampled from the P_k. As may be the case, one may only hope to estimate the restriction of P to $\bigcup_{k=1}^K \textsc{Supp}(P_k)$, and estimation on the entire support is impossible in absence of prior knowledge. From now on, Assumption 1 is thus supposed to be satisfied. One should pay attention to the fact that Assumption 1 does not require that the support of a single biased distribution P_k entirely covers that of

the target distribution P. In particular, each likelihood ratio (2.1) may vanish on a certain measurable subset weighted by P here, i.e., for all $k \leq K$, one may have: $\mathbb{P}\{\omega_k(Z) = 0\} > 0$. As discussed in the introduction, this significantly differs from the biased learning framework developed in other works, see [1] and the references therein, where the biased distributions are generally assumed to dominate the test distribution as in the usual Importance Sampling setting. Note also that Assumption 1 prevents biasing functions to vanish all at the same time. This condition is key to invert the likelihood ratio (2.2) and be able to recover the distribution P statistically based on samples drawn from the P_k's. When $K = 1$, this means $\omega_1(Z) > 0$ and IPW debiasing is then immediate of course: one simply weights each observation by means of $1/\omega_1$. In this work, focus is naturally on situations where $K \geq 2$. As shall be seen, the difficulty caused by the possibly vanishing biasing functions can be bypassed by combining appropriately the biased datasets, so as to compute nearly debiasing weights through the resolution of a system of equations, see (2.6). The generic setting described by Assumption 1 encompasses many estimation/learning problems, ranging from stratified sampling to censorship and clustering, see Examples 2, 3, and 4.

The second assumption required is standard in a multi-sample setting. It stipulates that the sample sizes n_k all tend to infinity as $n \to \infty$, in a way such that the fractions $\hat{\lambda}_k$ converge towards fixed values $\lambda_k > 0$.

Assumption 2. *There exist* $(\lambda_1, \ldots, \lambda_K) \in (0,1)^K$ *satisfying* $\sum_{k=1}^K \lambda_k = 1$ *such that for all* $k \leq K$ *it holds* $\hat{\lambda}_k \to \lambda_k$ *as* $n \to +\infty$.

Ignoring the bias selection issue, one may compute the empirical distribution based on the pooled sample

$$\widehat{P}_n = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \delta_{Z_{k,i}} = \sum_{k=1}^{K} \hat{\lambda}_k \widehat{P}_k \, ,$$

where $\widehat{P}_k = (1/n_k) \sum_{i \leq n_k} \delta_{Z_{k,i}}$ is the raw empirical distribution based on the (biased) sample $\mathcal{D}_k$, for $k \leq K$. This discrete random measure is a natural estimator of the linear convex combination of the P_k given by $\bar{P} = \sum_k \lambda_k P_k$. Since $\bar{P}$ is different from P in general, it is then easy to see why minimizing the raw empirical risk over the pooled sample may lead to decision rules that generalize poorly. However, observe that $\bar{P}$ is absolutely continuous w.r.t. P, with likelihood ratio

$$\frac{d\bar{P}}{dP}(z) = \sum_{k=1}^{K} \frac{\lambda_k \omega_k(z)}{\Omega_k} \, .$$

Under Assumption 1, the latter is strictly positive on the whole support of Z, and we have:

$$dP(z) = \left(\sum_{k=1}^{K} \frac{\lambda_k \omega_k(z)}{\Omega_k} \right)^{-1} d\bar{P}(z) \, . \tag{2.2}$$

Hence, if estimates $\widehat{\Omega}_k$ of the unknown expectations $\mathbb{E}_P[\omega_k(Z)]$ were at our disposal, one could immediately form a plug-in estimator of P by replacing $\bar{P}$, the Ω_k and the λ_k in Equation (2.2) with their statistical versions, namely $\widehat{P}_n$, the $\widehat{\Omega}_k$ and the $\hat{\lambda}_k$

$$d\widetilde{P}_n(z) = \left(\sum_{k=1}^{K} \frac{\hat{\lambda}_k \omega_k(z)}{\widehat{\Omega}_k} \right)^{-1} d\widehat{P}_n(z)\,. \tag{2.3}$$

In order to estimate the vector $\boldsymbol{\Omega} = (\Omega_1, \ldots, \Omega_K)$, note that Equation (2.2) immediately implies that $\boldsymbol{\Omega}$ is a solution (in $\boldsymbol{W} \in \mathbb{R}^K$) to the system of equations

$$\mathbf{1} = \big(\Gamma_1(\boldsymbol{W}), \ldots, \Gamma_K(\boldsymbol{W})\big)\,, \tag{2.4}$$

where $\mathbf{1}$ means the K-dimensional vector with all components equal to 1, and for any $k \leq K$, and all $\boldsymbol{W} = (W_1, \ldots, W_K) \in (\mathbb{R}_+)^K$, the notation

$$\Gamma_k(\boldsymbol{W}) = \frac{1}{W_k} \int \frac{\omega_k(z)}{\sum_{l=1}^{K} \frac{\lambda_l \omega_l(z)}{W_l}} d\bar{P}(z)\,. \tag{2.5}$$

A natural way to approximately recover $\boldsymbol{\Omega}$ thus consists in solving a statistical version of Equation (2.4), namely

$$\mathbf{1} = \left(\widehat{\Gamma}_1(\boldsymbol{W}), \ldots, \widehat{\Gamma}_K(\boldsymbol{W})\right)\,, \tag{2.6}$$

where the $\widehat{\Gamma}_l(\boldsymbol{W})$ are built by replacing λ_l and $\bar{P}$ in Equation (2.5) with $\hat{\lambda}_l$ and $\widehat{P}_n$ respectively. It is important to notice that the $\widehat{\Gamma}_k$ (just like the Γ_k) are homogeneous of degree 0. Hence, it is only possible to solve Systems (2.4) and (2.6) up to a multiplicative factor. Hopefully, $\boldsymbol{\Omega}$ can be recovered from any solution $\boldsymbol{W}^*$ to System (2.4). Indeed, for all $k \leq K$ it holds:

$$\Omega_k = \frac{W_k^*}{\int \left(\sum_{l=1}^{K} \frac{\lambda_l \omega_l(z)}{W_l^*} \right)^{-1} d\bar{P}(z)}\,, \tag{2.7}$$

refer to Appendix A for technical details. Similarly, for any solution $\widehat{\boldsymbol{W}}_n$ to System (2.6) and any $k \leq K$, we define:

$$\widehat{\Omega}_{n,k} = \frac{\widehat{W}_{n,k}}{\int \left(\sum_{l=1}^{K} \frac{\hat{\lambda}_l \omega_l(z)}{\widehat{W}_{n,l}} \right)^{-1} d\widehat{P}_n(z)}\,. \tag{2.8}$$

Plugging estimators (2.8) into Equation (2.3), the debiased estimate $\widetilde{P}_n$ is

$$\widetilde{P}_n = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(\frac{\left(\sum_{l=1}^{K} \frac{\hat{\lambda}_l \omega_l(Z_{k,i})}{\widehat{W}_{n,l}} \right)^{-1}}{\sum_{m=1}^{K} \sum_{j=1}^{n_m} \left(\sum_{l'=1}^{K} \frac{\hat{\lambda}_{l'} \omega_{l'}(Z_{m,j})}{\widehat{W}_{n,l'}} \right)^{-1}} \right) \delta_{Z_{k,i}}\,. \tag{2.9}$$

The next assumption now aims at ensuring that the solution to System (2.6) is asymptotically unique. The mapping from distribution P to the family of biased distributions $(P_k)_{k\leq K}$, is then one-to-one. It is expressed as a graph connectivity hypothesis, cf. Assumption C in [16].

Assumption 3. *Let G be the (undirected) graph with vertices in $\{1,\ldots,K\}$, and edges between vertices k and l $(k \neq l)$ if and only if*

$$\int \mathbb{I}\{\omega_k(z) > 0\} \cdot \mathbb{I}\{\omega_l(z) > 0\} dP(z) > 0.$$

The graph G is connected.

In the one-dimensional case ($q = 1$), and under Assumptions 1 to 3, the limit behavior (i.e., consistency, asymptotic normality) of the univariate cdf estimator of Equation (2.9), namely $z \in \mathbb{R} \mapsto \widetilde{P}_n(]-\infty,\ z])$, has been investigated in [16]. It is the purpose of the subsequent analysis to show that this approach can be successfully applied to statistical learning via ERM in the presence of selection bias, by deriving nonasymptotic guarantees under slightly stronger assumptions. Incidentally, the arguments which this analysis relies upon permit to establish an exponential tail bound for the cdf estimator mentioned above, extending the Dvoretzky-Kiefer-Wolfowitz inequality, and completing the results of [16], see Theorem 2. In the next subsection, the notion of biased sampling model is used in order to develop a framework for statistical learning based on biased training examples. Before showing rigorously in the next subsection how the ideas previously sketched permit to extend the ERM principle to this framework, a few remarks are in order.

Remark 1 (Covariate shift)**.** Let $Z = (X, Y)$ be a random pair taking its values in $\mathcal{X} \times \mathcal{Y}$ with distribution P and defining a supervised predictive problem, where X models some input information, useful to predict the output r.v. Y. In the very specific so-called *covariate shift* situation, for each sampling distribution P_k involved in the biasing model, the conditional distribution of Y given X is the same and is thus independent from k. However, the X-marginals are not necessarily the same and can be possibly supported on different subsets $\mathcal{X}_k \subset \mathcal{X}$. Note that, in the dedicated covariate shift literature, Assumption 3 is not stipulated in general, insofar as solving the predictive problem statistically only requires to recover the conditional distribution. However, this assumption is of course necessary to emulate the whole distribution P and accomplish other tasks, unsupervised for instance, even in such a specific context, see Example 4.

Remark 2 (Truncation, missing values)**.** We point out that, because of Assumption 3, the biased sampling models analyzed here do not cover the case of truncated observations, nor certain settings of missing variables. The latter may instead be treated by different methods, such as (multiple) imputation techniques, see e.g., [35].

We now exhibit a simple example supporting the need for a general approach and showing in particular that solving System (2.6) cannot be avoided in general.

Example 1 (MULTIVARIATE LENGTH BIASED SAMPLES)**.** *The bias sampling model where the probability of sampling an observation is proportional to its length is referred to as* length bias*. Its use is motivated by various applications, such as estimating the distribution of the number of children with a rare anomaly in families with proneness to engender such children [18], or correcting visibility bias during wildlife population estimation from aerial data [10] for instance. Refer to e.g., [30] for an overview of its applications. In the univariate case, it corresponds to* $\omega(z) = z$*, with* $z \in \mathbb{R}_+$*. When the learner can access two samples (one unbiased, one length biased), an approach to recover the nonparametric maximum likelihood estimator of* P *is proposed in [45]. Precisely, let* $\mathcal{D}_1 = \{Z_{1,1}, \ldots, Z_{1,n_1}\}$ *be an i.i.d. sample drawn from* P*, and* $\mathcal{D}_2 = \{Z_{2,1}, \ldots, Z_{2,n_2}\}$ *be an i.i.d. sample drawn from* P_2*, the length biased version of* P *such that* $dP_2(z) = zdP(z)/\int_{\mathbb{R}_+} zdP(z)$*. Let* $Z_1 < \ldots < Z_n$ *be the observations of the pooled sample* $\mathcal{D}_1 \cup \mathcal{D}_2$ *sorted in increasing order (we assume that ties cannot occur for simplicity), and let* $\xi_i = \mathbb{I}\{Z_i \in \mathcal{D}_1\}$ *indicate whether* Z_i *comes from* $\mathcal{D}_1$ *or not for* $i = 1, \ldots, n$*. It is immediate to see that the cdf* P *that maximizes*

$$\prod_{i=1}^{n} \big(dP(Z_i)\big)^{\xi_i} \left(\frac{Z_i \, dP(Z_i)}{\int_{\mathbb{R}_+} z \, dP(z)} \right)^{1-\xi_i} \tag{2.10}$$

has positive jumps only at the Z_i*'s, so that estimating the jumps* $dP(Z_i)$ *in* (2.10) *is sufficient. Simple computations show that*

$$dP(Z_i) = \frac{\hat{\mu}}{n_2 Z_i + n_1 \hat{\mu}}, \quad \text{where } \hat{\mu} \text{ satisfies} \quad \sum_{i=1}^{n} \frac{Z_i}{n_2 Z_i + n_1 \hat{\mu}} = 1. \tag{2.11}$$

Hence, maximizing (2.10) *requires to solve the equation on the right hand side of* (2.11)*. The approach can be straightforwardly extended to the multivariate case. Assume that the random variables observed are now valued in* $\mathbb{R}_+^K$ *and consider* $K+1$ *datasets such that* $\mathcal{D}_0$ *is composed of i.i.d. realizations drawn from* P *and* $\mathcal{D}_k$ *is drawn from* P_k *such that* $dP_k(z) = z^{(k)} dP(z)/\int_{\mathbb{R}^+} z^{(k)} dP(z)$*, where* $z^{(k)}$ *denotes the* k*-th coordinate of* $z = (z^{(1)}, \ldots, z^{(K)})$*. In other words, all datasets (except* $\mathcal{D}_0$*) are length biased according to different dimensions. Equipped with the notations* $\{Z_1, \ldots, Z_n\} = \mathcal{D}_0 \cup \ldots \cup \mathcal{D}_K$ *and* $\xi_{i,k} = \mathbb{I}\{Z_i \in \mathcal{D}_k\}$ *for* $i = 1, \ldots, n$ *and* $k = 1, \ldots, K$*, the quantity that must be maximized writes*

$$\prod_{i=1}^{n} \big(dP(Z_i)\big)^{\xi_{i,0}} \prod_{k=1}^{K} \left(\frac{Z_i^{(k)} \, dP(Z_i)}{\int_{\mathbb{R}_+} z^{(k)} \, dP(z)} \right)^{\xi_{i,k}}.$$

As in the scalar case, we have: $dP(Z_i) = 1/\left(n_0 + \sum_{k=1}^{K} n_k Z_i^{(k)}/\hat{\mu}_k\right)$*, where the* $\hat{\mu}_k$*'s satisfy*

$$\hat{\mu}_l = \sum_{i=1}^{n} \frac{Z_i^{(l)}}{n_0 + \sum_{k=1}^{K} \frac{n_k Z_i^{(k)}}{\hat{\mu}_k}}, \quad 1 \le l \le K.$$

Thus, we need to solve a system of K equations in order to recover the $\hat{\mu}_k$'s, which is actually a specific case of (2.6). Hence, except in certain simplistic situations, see e.g., Appendix B, solving System (2.6) cannot be avoided to debias biased samples in general. Although the approach based on bias sampling models encompasses IPW, we highlight that it is much more general than the latter.

2.2. Learning from biased samples – Extending the ERM approach

Recall that Z is a random vector valued in $\mathcal{Z} \subset \mathbb{R}^q$, $q \geq 1$, with probability distribution P. Let Θ be a decision space and consider some loss function $\psi : \mathbb{R}^q \times \Theta \to \mathbb{R}_+$, that is P-integrable for any decision rule $\theta \in \Theta$. The goal pursued here is to solve the risk minimization problem

$$\min_{\theta \in \Theta} \quad L_P(\theta) = \mathbb{E}_P[\psi(Z, \theta)] \tag{2.12}$$

where L_P is called the risk function. As recalled in introduction, if independent copies $Z_1, \ldots, Z_n$ of Z are available, the unknown risk L_P is classically replaced with $\widehat{L}_n = L_{\widehat{P}_n}$, where $\widehat{P}_n = (1/n)\sum_{i=1}^n \delta_{Z_i}$ is the empirical distribution. Here, the training data is composed of K biased samples $\mathcal{D}_k$, as defined in Section 2.1. Note that it is a strict generalization of the standard ERM setting, insofar as the latter can be recovered as the special case $K = 1$ and $\omega_1 \equiv 1$. This general framework encompasses a wide variety of situations encountered in practice, as illustrated by the following examples.

Example 2 (Binary classification under stratified sampling)**.** *We place ourselves in the context of binary classification, i.e., we have $Z = (X, Y)$, $\mathcal{Z} = \mathcal{X} \times \{-1, +1\}$, $q = d + 1$, $\Theta = \mathcal{G}$, and $\psi((X, Y), g) = \mathbb{I}\{Y \neq g(X)\}$. Consider $K \geq 1$ subsets $\mathcal{X}_1, \ldots, \mathcal{X}_K$ of the input space $\mathcal{X}$, such that $\mu(\mathcal{X}_k) > 0$ for all $k \leq K$, μ denoting X's marginal distribution. The case where only labeled examples with input observations in $\mathcal{X}_k$ can be collected to form sample $\mathcal{D}_k$ corresponds to the situation where $\omega_k(Z) = \mathbb{I}\{X \in \mathcal{X}_k\}$. In this case, the P_k are the conditional distributions of Y given that $X \in \mathcal{X}_k$.*

In Example 2, selection bias is due to stratified sampling schemes, where observations are sampled in strata of interest (the $\mathcal{X}_k$'s namely). Note that in this case the $\mathcal{X}_k$'s, and therefore the ω_k's, are controlled and known by the learner. This setting also covers many practical situations, where the training dataset is constructed by the aggregation of different sources, and naturally applies to other learning tasks. In such scenarios, having access to the ω_k is natural, as the learner may know the conditions in which the data have been collected, e.g., the part of the world in which the photos have been taken, or the profile of a user answering the questionary.

Example 3 (Regression under right censorship)**.** *Let the distribution-free regression framework where T is a bounded random duration (i.e., a non-negative r.v. such that $\|T\|_{\sup} < +\infty$), and X is a random vector valued in $\mathcal{X} \subset \mathbb{R}^d$, defined on the same probability space, and supposedly useful to predict*

T. The goal is to learn a regression function $h : \mathcal{X} \to \mathbb{R}$ in a class $\mathcal{H}$ of bounded functions with minimum quadratic risk. This corresponds to $Z = (X, T)$, $\mathcal{Z} = \mathcal{X} \times \mathbb{R}_+$, $q = d + 1$, $\Theta = \mathcal{H}$, and $\psi((X, T), h) = (T - h(X))^2$. Let $K \geq 1$, and $0 < \tau_1 < \ldots < \tau_{K-1} < \tau_K = \|T\|_{\sup}$. Consider the case where, the samples $\mathcal{D}_k$ are composed of censored observations with a deterministic right censorship, i.e., of copies of $(X, \min\{T, \tau_k\})$. This is equivalent to the case $\omega_k(Z) = \mathbb{I}\{T \leq \tau_k\}$, and the P_k are the conditional distributions of (X, T) given that $T \leq \tau_k$.

Example 3 typically arises in longitudinal experiments (e.g., medical trials, customer behavior evaluations) that must be stopped at some point (by lack of means for instance). Instead of discarding every observation for which the event of interest has not occurred yet, one may register the time at which the experiment has been stopped (the τ_k) and leverage this information to debias the population.

Example 4 (Clustering)**.** *Consider an unsupervised variant of Example 2, where $Z = X \in \mathbb{R}^q$ has distribution P and $\omega_k(X) = \mathbb{I}\{X \in \mathcal{X}_k\}$ where $\mathcal{X}_k \subset \mathcal{X}$ for $k = 1, \ldots, K$ is the observable strata of the population of interest. A popular approach to clustering consists in assuming that distribution P is an unknown mixture of $\mathcal{K} \geq 2$ Gaussian distributions with means $\mu_1, \ldots, \mu_{\mathcal{K}}$ in $\mathbb{R}^q$ and same covariance matrix Σ, see e.g. [15, Chapter 14]: $P = \sum_{m=1}^{\mathcal{K}} \pi_m \mathcal{N}(\mu_m, \Sigma)$, with $(\pi_1, \ldots, \pi_{\mathcal{K}}) \in [0, 1]^{\mathcal{K}}$ s.t. $\sum_{m=1}^{\mathcal{K}} \pi_m = 1$. Denoting by θ the parameter encoding the Gaussian mixture model and $p_\theta(x)$ its likelihood, the Expectation-Maximization algorithm (EM algorithm, see e.g., [15, Section 8.5]) computes the optimal θ by maximizing the (log-)likelihood over the observed datapoints. The latter can be seen as minimizing an empirical version of problem (2.12) with $\psi(X, \theta) = -\log p_\theta(X)$, and is therefore another particular case in which our debiasing approach applies.*

As illustrated by Examples 2, 3, and 4, the framework developed in this paper applies to a wide variety of statistical learning problems, indifferently supervised or unsupervised, sampling bias being determined by the covariates and/or the output. We also point out that the vast majority of statistical techniques for correcting sampling/selection bias relies on some Inverse Probability Weighting approaches, e.g., Beran, Kaplan-Meier methods, Horvitz-Thompson techniques, propensity score matching. The sole difference in these variations is the form of the biasing functions and their arguments. The main advantage of the general framework developed here consists in encompassing all these situations, diverse in appearance only. It allows to derive generalization guarantees for any risk minimization problem in the presence of selection bias. The price to pay for our unifying framework is that the debiasing weights cannot be computed trivially but requires the solving of System (2.6). We recall that Example 1 shows that this step is unavoidable in general. In this context, we prove in Section 3 that, under mild assumptions, minimizing $L_{\widetilde{P}_n}$, where $\widetilde{P}_n$ is defined in (2.9), allows to attain learning rates that are of the same order, $O_{\mathbb{P}}(1/\sqrt{n})$ namely, as those achieved in absence of any selection bias, i.e., when $\omega_k \equiv 1$ for all $k \leq K$. The minimization of the functional $L_{\widetilde{P}_n}$ boils down to a weighted ERM procedure,

- **Input.** Samples $\mathcal{D}_k = \{Z_{k,i},\ i \leq n_k\}$, coefficients $\hat{\lambda}_k = n_k/n$, and biasing functions ω_k for $k \leq K$.
- **Debiasing the raw empirical distribution.** Form the raw empirical distribution based on the pooled sample

$$\widehat{P}_n = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\delta_{Z_{k,i}},$$

(i) for $k \leq K$, compute the functions given by: $\forall \boldsymbol{W} \in (\mathbb{R}_+)^K$,

$$\widehat{\Gamma}_k(\boldsymbol{W}) = \frac{1}{W_k}\int \frac{\omega_k(z)}{\sum_{l=1}^{K}\frac{\hat{\lambda}_l\omega_l(z)}{W_l}}d\widehat{P}_n(z);$$

(ii) solve (2.6), i.e., find $\widehat{\boldsymbol{W}}_n = (\widehat{W}_{n,1},\ \ldots,\ \widehat{W}_{n,K}) \in (\mathbb{R}_+)^K$ satisfying

$$\max_{1\leq k\leq K}\ \widehat{W}_{n,k}/\hat{\lambda}_k = 1; \tag{2.13}$$

(iii) for $k \leq K$ and $i \leq n_k$, compute the weights

$$\pi_{k,i} = \frac{\left(\sum_{l=1}^{K}(\hat{\lambda}_l/\widehat{W}_{n,l})\omega_l(Z_{k,i})\right)^{-1}}{\sum_{m=1}^{K}\sum_{j=1}^{n_m}\left(\sum_{l'=1}^{K}(\hat{\lambda}_{l'}/\widehat{W}_{n,l'})\omega_{l'}(Z_{m,j})\right)^{-1}},$$

so as to form the "debiased" distribution estimator given by

$$\widetilde{P}_n = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\pi_{k,i}\delta_{Z_{k,i}}.$$

- **ERM.** Solve the ERM problem $\min_{\theta\in\Theta}\widetilde{L}_n(\theta)$, to produce the solution $\tilde{\theta}_n$, with $\widetilde{L}_n(\theta)$ given by

$$\widetilde{L}_n(\theta) \overset{\text{def}}{=} L_{\widetilde{P}_n}(\theta) = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\pi_{k,i}\psi(Z_{k,i},\ \theta). \tag{2.14}$$

Fig 1: ERM BASED ON BIASED TRAINING SAMPLES

with debiasing weights depending on the solution to System (2.6). The learning procedure can thus be implemented in three steps as summarized in Figure 1.

1. First, we use the raw empirical distribution $\widehat{P}_n$ to form System (2.6) by computing the estimates $\widehat{\Gamma}_k$ of the Γ_k;
2. Next, we solve System (2.6) to build the "debiased" estimate $\widetilde{P}_n$ of P;
3. Finally, we obtain the decision rule by solving the statistical version of Problem (2.12), in which P is replaced with $\widetilde{P}_n$.

Discussing how to perform in practice the minimization of the nearly debiased empirical risk estimate of Equation (2.14), or of a smooth/penalized version of it, is beyond the scope of the present paper. However, observe that most machine learning libraries offer the option to reweight the training observations involved in the learning stage in a simple plug-in fashion (e.g., the `sample_weight` option for scikit-learn [31]). We also highlight the generality of the above approach, insofar as it may be straightforwardly combined with any ERM-like learning algorithm, for a wide range of biasing scenarios. We point out however that this generality goes along with the solving of System (2.6). This step cannot be avoided in general, and yields nontrivial solutions, except for simplistic cases such as that discussed in Appendix B. Finally, note that the computational cost induced by the debiasing procedure is low, the unique difference with standard methods lying in the computation of the weights involved in the risk functional, which can be tackled efficiently by means of a Gradient Descent strategy. Indeed, as can be seen in the proof of Proposition 1, solving System (2.4) is equivalent to minimize the strongly convex function $\bar{D}$, defined for all $\boldsymbol{u} = (u_1, \ldots, u_K) \in \mathbb{R}^K$ by

$$\bar{D}(\boldsymbol{u}) = \int \log \left[\sum_{l=1}^{K} e^{u_l} \omega_l(z) \right] d\bar{P}(z) - \sum_{l=1}^{K} \lambda_l u_l \,,$$

and with Hessian matrix $\bar{D}'' \in \mathbb{R}^{K \times K}$ such that

$$\left[\bar{D}''(\boldsymbol{u}) \right]_{k,k'} = \int \left[\frac{e^{u_k} \omega_k(z) \delta_{kk'}}{\sum_{l=1}^{K} e^{u_l} \omega_l(z)} - \frac{e^{u_k} \omega_k(z) e^{u_{k'}} \omega_{k'}(z)}{\left(\sum_{l=1}^{K} e^{u_l} \omega_l(z) \right)^2} \right] d\bar{P}(z) \,. \tag{2.15}$$

By characterizing the curvature of $\bar{D}$, the eigenvalues of $\bar{D}''$ thus influence the convergence of the solution to System (2.4). Bounding away from 0 the second smallest eigenvalue of $\bar{D}''$ is actually required in the subsequent nonasymptotic analysis, see Assumption 7 for more details. We conclude this section with two remarks, on the normalization (2.13) and about the possibility to use a sampling approach instead of the reweighting, and a numerical illustration of the benefits of the approach presented here.

Remark 3. (On normalization (2.13)) As highlighted in Section 2.1, recall that System (2.6) is homogeneous of degree 0. Hence, normalization (2.13) is just a way to select one $\widehat{\boldsymbol{W}}_n$ among all possible solutions. In [16] for instance, the normalization $\widehat{W}_{n,K} = 1$ is used instead. Normalization (2.13) happens to be more suited to our nonasymptotic analysis. In particular, it ensures that $\widehat{\boldsymbol{W}}_n$ is unique and bounded away from 0 with high probability, see Proposition 1.

Remark 4. (Plug-in *vs* Sampling) From a practical perspective, modifying the objective function using the weights computed at step (*iii*) in the above scheme is not the only option. An alternative to learn the predictive rule would be to sample observations from the distribution (2.9), given the original data. This would generate a new (unique and nearly debiased) dataset, from which any ERM-based learning algorithm can be run in a standard fashion.

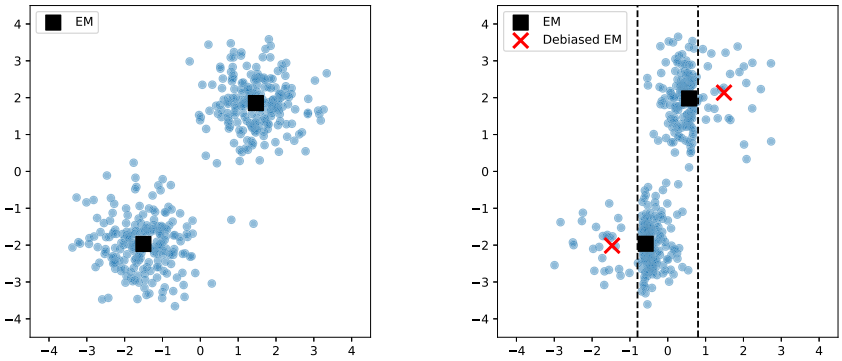


Fig 2: (Clustering by vanilla and debiased EM algorithms). In absence of sampling bias (left), the vanilla EM algorithm finds the right centroids. When the data are biased, over-represented towards the center region (right), the centroids obtained through the vanilla EM algorithm are also attracted towards the center region. In contrast, the debiased EM algorithm produces nearly the same centroids than those obtained from the unbiased dataset.

Example 5. (Clustering, bis) *Let distribution P be a mixture of two 2-d Gaussian distributions, centered at $(X_1, X_2) = (-1.5, -2)$ and $(X_1, X_2) = (1.5, 2)$ respectively. On the left of Figure 2, an unbiased dataset is displayed, on which the vanilla EM algorithm finds the centroids accurately. The second dataset (on the right) is actually composed of 3 samples: that on the left, for which $\omega_1(X) = \mathbb{I}\{X_1 \leq -0.5\}$, that in the middle, for which $\omega_2(X) = \mathbb{I}\{-0.7 \leq X_1 \leq 0.7\}$, and that on the right for which $\omega_3(X) = \mathbb{I}\{0.5 \leq X_1\}$. The dataset in the middle is of size $n_2 = 300$ observations, while the left/right ones are composed of $n_1 = n_3 = 30$ observations. As expected, the centroids found by means of the vanilla version of the EM algorithm are heavily shifted towards the center, whereas the debiased variant of the EM algorithm is able to leverage the bias functions information so as to recover nearly the correct centroids.*

3. Empirical risk minimization in biased sampling models

In this section, we provide theoretical guarantees for the extension of ERM to biased training samples we have introduced in Section 2.2. Unsurprisingly, the subsequent nonasymptotic analysis requires slightly more stringent assumptions than those involved in the asymptotic study carried out in [16], and listed in Section 2.1. In particular, Assumption 4 strengthens Assumption 2 in order to control the fluctuations of the sample sizes, so as to establish finite-sample learning rate bounds. In the same spirit, additional parameters are introduced to guarantee that crucial quantities are bounded away from critical values. Hence, expectations involved in Assumption 3 are supposed to be greater than $\kappa > 0$, while the minimal positive value of the ω_k is lower bounded by $\varepsilon > 0$ (see

Assumption 6). Notice that this lower bound does not prevent the ω_k to vanish, preserving the generality of the approach.

Assumption 4. *There exist* $(\lambda_1, \ldots, \lambda_K) \in (0,1)^K$ *satisfying* $\sum_{k=1}^K \lambda_k = 1$, *and* $C_\lambda, \underline{\lambda} > 0$ *such that for all* $k \leq K$ *and* $n \geq K$ *it holds*

$$\underline{\lambda} \leq \lambda_k\,, \qquad \underline{\lambda} \leq \hat{\lambda}_k\,, \qquad \textit{and} \qquad \left|\hat{\lambda}_k - \lambda_k\right| \leq \frac{C_\lambda}{\sqrt{n}}\,. \tag{3.1}$$

Observe that the control of the order of magnitude of the sample sizes and that of their fluctuations in Assumption 4 cannot be avoided, since the goal here is to establish nonasymptotic (learning) rate bounds, see Lemma 2 in particular.

Remark 5. We point out that, in the situation where the vector of sample sizes $(n_1, \ldots, n_K)$ is random, distributed as a multinomial of size n with parameters $(\lambda_1, \ldots, \lambda_k)$, the last bounds in Equation (3.1) simultaneously hold true for any k and an appropriate constant C_λ with overwhelming probability. Indeed, using Hoeffding's inequality (see [21]) combined with the union bound for instance, one obtains that, for any $\delta \in (0,1)$, all these conditions are fulfilled with probability larger than $1-\delta$ with $C_\lambda = \sqrt{\log(K/\delta)/2}$, and that $\underline{\lambda} \geq \min_k \lambda_k - C_\lambda/\sqrt{n}$, provided that $n > C_\lambda^2 / \min_k \lambda_k$. Note that for simplicity, we restrict our analysis to the situation where the sample sizes are deterministic, the random case being a straightforward extension.

Assumption 5. *For* $\kappa > 0$, *define* G_κ *the (undirected) graph with vertices in* $\{1, \ldots, K\}$, *and edge between* k *and* l $(k \neq l)$ *if and only if*

$$\int \mathbb{I}\{\omega_k(z) > 0\} \cdot \mathbb{I}\{\omega_l(z) > 0\} dP(z) \geq \kappa.$$

There exists $\kappa > 0$ *such that* G_κ *is connected.*

From an algebraic viewpoint, one may classically check whether Assumption 5 is fulfilled or not by means of a breadth-first search algorithm, or by examining the spectrum of the Laplacian matrix of G_κ for instance, see e.g., [17]. Note that such a verification would require to have access to P, which is unknown in general. However, we highlight that in practice the connectivity property that must be checked concerns $\widehat{G}_n$, the empirical counterpart of G defined in Equation (3.2), which only depends on the observed empirical distributions $\widehat{P}_k$.

Assumption 6. *There exists* $\varepsilon > 0$ *such that*

$$\forall z \in \mathcal{Z},\ \forall k \leq K, \qquad \varepsilon \cdot \mathbb{I}\{\omega_k(z) > 0\} \leq \omega_k(z) \leq 1.$$

In particular this implies $\omega_k(z_i) \geq \varepsilon$ *for all* $z_i \in \mathcal{D}_k$, *and* $\Omega_k \leq 1$ *for all* $k \leq K$.

Note that parameters κ and ε allow to quantify the overlap between two biasing functions, in a way that extends the simple overlap/non overlap condition of [16] (recovered here by $\mathbb{I}\{\kappa\varepsilon \neq 0\}$). The results derived below typically hold true with probability $1 - e^{-(\kappa\varepsilon)^2 n}$, see e.g., Proposition 1, confirming that no learning is possible without overlap, but also providing the new insight that performances improve with the overlapping.

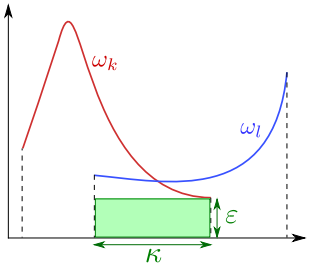


Fig 3: Overlap control.

Remark 6. We point out that, in Example 2, Assumption 1 simply means that $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_K$. Assumption 6 is fulfilled with $\varepsilon = 1$, and Assumption 5 can be checked in a simple manner, insofar as we have: $\forall 1 \leq k \neq l \leq K$, $e_{k,l} = 1 \Leftrightarrow \mu(\mathcal{X}_k \cap \mathcal{X}_l) \geq \kappa$. In Example 3, Assumption 1 is directly fulfilled, just like Assumption 6 with $\varepsilon = 1$.

As discussed at the end of Section 2.2, we also introduce an assumption on the second smallest eigenvalue of the Hessian matrix $\bar{D}''$ defined in Equation (2.15).

Assumption 7. *Let $U = \log(K/\varepsilon) \sum_{t=1}^{K-1} 2^t (\underline{\lambda}\kappa\varepsilon)^{-t}$, $\mathcal{U} = [0, U]^K \subset \mathbb{R}^K$, and $\sigma > 0$. See Proposition 4 in Appendix C.2 for more insights about the first two values. For all $\boldsymbol{u} \in \mathcal{U}$, $\sigma_2(\bar{D}''(\boldsymbol{u})) \geq \sigma$, where $\sigma_2(A)$ denotes the second smallest eigenvalue of a matrix A.*

As revealed by the proof of Proposition 2, Assumption 7 is required to control the deviation $\|\hat{\boldsymbol{u}}_n - \boldsymbol{u}^*\|_2$ in terms of $\left\|\widehat{D}'_n(\boldsymbol{u}^*) - \bar{D}'_n(\boldsymbol{u}^*)\right\|_2$, using the curvature of D in the non-flat parts of the optimization landscape. Note that Assumption 7 is not needed in the asymptotic analysis as even the smallest possible curvature (and we know it is strictly positive by the proof of Proposition 1 in [16]) is still sufficient when n goes to infinity. On the opposite, to establish finite sample bounds, we have to bound away from zero the second smallest eigenvalue of $\bar{D}''(\boldsymbol{u})$, uniformly over $\mathcal{U}$, in an explicit manner. Although such a lower bound is always attained, as $\bar{D}''$ is continuous on the compact set $\mathcal{U}$, its dependence with respect to the problem instance (i.e., the distribution P, the biasing functions ω_k, the sample proportions λ_k) is non-trivial. For this reason, we rather opted for explicitly introducing a parameter σ to materialize this lower bound. The results subsequently derived then depend on σ in a more interpretable fashion.

Equipped with these assumptions, we now carry out a detailed rate bound analysis. The first step, described in Section 3.1, consists in showing that with overwhelming probability the solution to System (2.6) exists, is unique, and belongs to a compact set bounded away from 0 (Proposition 1). This crucial property then allows to derive nonasymptotic concentration bounds for $\widehat{\boldsymbol{W}}_n$ (Proposition 2) and next for $\widehat{\boldsymbol{\Omega}}_n$ (Proposition 3). The generalization results are finally stated in Section 3.2, under a standard complexity assumption. The guarantees for the minimizers of the debiased risk version are established in Theorem 1, and a corollary about the excess risk is discussed. When the concept

class is composed of indicator functions of subsets, a tighter analysis is presented (Theorem 2), which provides an extension of the Dvoretsky-Kiefer-Wolfowitz inequality under biased sampling models.

3.1. Existence, uniqueness, and concentration of the solution

As detailed in Section 2.2, our debiasing ERM procedure critically relies on solving System (2.6). It is shown in [16] (Theorem 1.1 therein) that the latter admits a unique solution if and only if a directed and statistical (i.e., with $\widehat{P}_k$ instead of P_k) version of graph G in Assumption 3, denoted by $\widehat{G}_n$ thereafter, is strongly connected. From a limit perspective, the strong law of large numbers suffices to guarantee that, with probability 1, the edges of $\widehat{G}_n$ are asymptotically the same as those of G. Assumption 3 then allows to conclude that $\widehat{G}_n$ is strongly connected and that System (2.6) admits a unique solution. This result is stated as Corollary 1.1 in [16]. The proposition below refines this assertion from a nonasymptotic angle. It shows that existence and uniqueness actually occur with overwhelming probability. Uniqueness is of course understood up to the homogeneity property. To avoid any ambiguity, $\widehat{\boldsymbol{W}}_n$ now refers to the solution to System (2.6) satisfying $\max_{k\le K} \widehat{W}_{n,k}/\hat{\lambda}_k = 1$, see Equation (2.13). Similarly, $\boldsymbol{W}^*$ is assumed to verify $\max_{k\le k} W_k^*/\lambda_k = 1$. Proposition 1 also shows that both $\widehat{\boldsymbol{W}}_n$ and $\boldsymbol{W}^*$ belong to a compact set bounded away from 0. This property is key in the subsequent analysis, as $\widehat{\boldsymbol{W}}_n$ is often present in denominators, see e.g., Equation (2.9). Note that to keep notation simple, we use generic constants in the statements of the results, that may have different values from one proposition to the other. For completeness, we provide their exact values in the technical proofs of the Appendix section. Importantly, they only depend on parameters $K, C_\lambda, \underline{\lambda}, \kappa, \varepsilon, \sigma$ introduced in Assumptions 4, 5, 6, and 7. Although K is treated as a constant here, note that our results remain meaningful as long as $K = o(n)$. If K grows linearly with n, it is immediate to see that the dataset sizes n_k are then necessarily bounded, making a consistent recovery of the $\widehat{P}_k$'s impossible, and the debiasing procedure bound to fail.

Proposition 1. *Suppose that Assumptions 4, 5, and 6 are satisfied. Then, there exist $M, c, \rho > 0$, depending only on $K, \underline{\lambda}, \kappa, \varepsilon$, such that for all $n \ge \log(M)/c$, it holds with probability at least $1 - M\exp(-cn)$:*

- *the solution $\widehat{\boldsymbol{W}}_n$ to System (2.6) exists and is unique,*
- *for all $k \le K$, $\quad \rho \le \widehat{W}_{n,k} \le 1, \quad$ and $\quad \rho \le W_k^* \le 1$.*

The rationale behind the proof is similar to that used to establish Corollary 1.1 in [16]. Rather than simply establishing that the edges of $\widehat{G}_n$ asymptotically match those of G, we bound the probability that they differ from those of G_κ, defined in Assumption 5.

Proof. First, define the directed graph $\widehat{G}_n$ with vertices in $\{1, \dots, K\}$ and edge

$k \to l$ if and only if

$$\int \mathbb{I}\{\omega_k(z) > 0\} d\widehat{P}_l(z) > 0. \tag{3.2}$$

The graph $\widehat{G}_n$ is said to be strongly connected if, for any pair of vertices (k, l), there exist a directed path from k to l and a directed path from l to k. It is proved in [46] (see also Theorem 1.1 in [16]) that this is a necessary and sufficient condition for System (2.6) to have a unique solution. To show that $\widehat{G}_n$ is strongly connected, we prove that the left hand side in Equation (3.2) is sufficiently close to (a weighted version of) the link condition in Assumption 5 with overwhelming probability. Let (k, l) be an edge in G_κ. Using Assumptions 5 and 6 we get:

$$\begin{aligned}\int \mathbb{I}\{\omega_k(z) > 0\} dP_l(z) &= \int \mathbb{I}\{\omega_k(z) > 0\} \frac{\omega_l(z)}{\Omega_l} dP(z), \\ &\geq \varepsilon \int \mathbb{I}\{\omega_k(z) > 0\} \cdot \mathbb{I}\{\omega_l(z) > 0\} dP(z), \\ &\geq \kappa\varepsilon.\end{aligned}$$

Now, observe that the left-hand side in Equation (3.2) is the empirical version of the above term. By Hoeffding's inequality, for every $t > 0$ it holds:

$$\begin{aligned}\mathbb{P}\left\{\int \mathbb{I}\{\omega_k(z) > 0\} d\widehat{P}_l(z) - \int \mathbb{I}\{\omega_k(z) > 0\} dP_l(z) \leq -t\right\} &\leq \exp(-2n_l t^2), \\ &\leq \exp(-2\underline{\lambda} n t^2).\end{aligned}$$

In particular, setting $\delta = \exp\left(-\frac{\underline{\lambda}(\kappa\varepsilon)^2 n}{2}\right)$ it holds with probability at least $1-\delta$:

$$\int \mathbb{I}\{\omega_k(z) > 0\} d\widehat{P}_l(z) \geq \int \mathbb{I}\{\omega_k(z) > 0\} dP_l(z) - \frac{\kappa\varepsilon}{2} \geq \frac{\kappa\varepsilon}{2}, \tag{3.3}$$

so that $k \to l$ in $\widehat{G}_n$. The exact same reasoning can be applied after having switched k and l. The union bound then gives that with probability at least $1 - 2\delta$ it holds: $k \to l$ and $l \to k$ in $\widehat{G}_n$. Now, $G_\kappa = (V, E)$ being connected, we know that there exists a set of edges $E_{\min} \subset E$ of cardinal $K - 1$ such that $G_{\min} = (V, E_{\min})$ is connected. Applying the above method to every edge in $E_{\min}$, we get that with probability at least $1 - 2(K-1)\delta$, every pair (k, l) linked in $G_{\min}$ is linked both ways in $\widehat{G}_n$. Since $G_{\min}$ is connected, this means that $\widehat{G}_n$ is strongly connected. The proof is concluded by setting $M = 2(K-1)$, and $c = \underline{\lambda}(\kappa\varepsilon)^2/2$. As a lengthy technical analysis is required to identify ρ, the proof of the second claim of Proposition 1 is postponed to Appendix C.1. □

The identification of a compact set, bounded away from 0 and containing $\widehat{\boldsymbol{W}}_n$ and $\boldsymbol{W}^*$ with high probability, is essential to carry out a nonasymptotic analysis. In particular, it permits to derive the following exponential concentration bound.

Proposition 2. *Suppose that Assumptions 4, 5, 6, and 7 are satisfied. Then, there exist $M, M', c, c', \gamma, n_0 > 0$, depending only on $K, C_\lambda, \underline{\lambda}, \kappa, \varepsilon, \sigma$, such that for all $t > 0$ and $n \geq n_0$ it holds:*

$$\mathbb{P}\left\{ \left\| \widehat{\boldsymbol{W}}_n - \boldsymbol{W}^* \right\|_2 > \frac{\gamma}{\sqrt{n}} + t \right\} \leq M e^{-cn} + M' e^{-c' n t^2}.$$

Note that the sup norm $\|\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^*\|_\infty$ is upper bounded by the Euclidean norm $\|\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^*\|_2$, so that the inequality in Proposition 2 is simultaneously satisfied by $|\widehat{W}_{n,k} - W_k^*|$ for all $k \leq K$. The proof uses the same reparameterization as for the second claim of Proposition 1. It involves some notion of curvature related to System (2.4), that characterizes the hardness of the problem. The compactness derived in Proposition 1 is then used to uniformly lower bound this curvature, see Assumption 7. Technical details are provided in Appendix C.2.

Using Equations (2.7) and (2.8), one can finally control the deviation of $\widehat{\boldsymbol{\Omega}}_n$ with respect to $\boldsymbol{\Omega}$, as revealed by the following proposition, whose proof is detailed in Appendix C.3.

Proposition 3. *Suppose that Assumptions 4, 5, 6, and 7 are satisfied. Then, there exist $M, M', c, c', \gamma, n_0$, depending only on $K, C_\lambda, \underline{\lambda}, \kappa, \varepsilon, \sigma$, such that for all $t > 0$ and $n \geq n_0$ it holds:*

$$\mathbb{P}\left\{ \left\| \widehat{\boldsymbol{\Omega}}_n - \boldsymbol{\Omega} \right\|_\infty > \frac{\gamma}{\sqrt{n}} + t \right\} \leq M e^{-cn} + M' e^{-c' n t^2}.$$

Hence, we have shown that, using the procedure described in Section 2.2, we can compute an estimate $\widehat{\boldsymbol{\Omega}}_n$ of $\boldsymbol{\Omega}$ with good nonasymptotic concentration properties. The last step consists in analyzing the performance of the minimizers of the debiased risk (2.14) when $\widetilde{P}_n$ is built using Equation (2.3) and $\widehat{\boldsymbol{\Omega}}_n$.

3.2. Generalization ability of minimizers of the debiased risk

As a first go, we introduce the following standard complexity assumption on the class $\mathcal{F} = \mathcal{F}_\Theta = \{\psi(\cdot, \theta) \colon \theta \in \Theta\}$, see e.g., Equation (2.14.6) in [43].

Assumption 8. *The collection of functions $\mathcal{F}_\Theta = \{\psi(\cdot, \theta) \colon \theta \in \Theta\}$ satisfies $|\psi(z, \theta)| \leq 1$ for all z, θ, and is a uniform Donsker class (relative to L_2) with polynomial uniform covering numbers, i.e., there exist constants $C_\Theta > 0$ and $r \geq 1$ such that for all $\zeta > 0$*

$$\sup_Q \mathcal{N}(\zeta,\ \mathcal{F}_\Theta,\ L_2(Q)) \leq (C_\Theta / \zeta)^r$$

where the supremum is taken over the set of probability measures Q on $\mathcal{Z}$, and $\mathcal{N}(\zeta, \mathcal{F}_\Theta, L_2(Q))$ is the minimum number of $L_2(Q)$ balls of radius ζ needed to cover $\mathcal{F}_\Theta$.

Remark 7. The above hypothesis is a classic complexity assumption. Of course, the subsequent rate bound analysis can be straightforwardly extended to settings involving alternative complexity conditions, such as e.g., finite VC dimension, Rademacher averages. Recall that a collection of functions $\mathcal{F}_\Theta$ of finite VC dimension $V < +\infty$, and with envelope function $F \equiv 1$, satisfies Assumption 8 with $r = 2V - 2$, and C_Θ depending only on V, see e.g., Theorem 2.6.7 in [43].

The main argument of the subsequent analysis then consists in showing that the uniform deviation between the nearly debiased risk and the true risk

$$\sup_{\theta\in\Theta} \left|\widetilde{L}_n(\theta) - L(\theta)\right| \tag{3.4}$$

is small with high probability. This is however far from being as straightforward as in the unbiased situation, since the set of random variables $\{\widetilde{L}_n(\theta)-L(\theta)\}_{\theta\in\Theta}$ is not an empirical process (i.e., a collection of i.i.d. averages), and the standard concentration inequalities therefore do not apply. Indeed, we recall that $\widetilde{L}_n(\theta)$ depends on $\widehat{\mathbf{\Omega}}_n$, which is obtained from the solution to System (2.6), and for which no closed analytical form is available in general, see Section 2.2. To bypass this difficulty, we decompose the excess of risk $|\widetilde{L}_n(\theta) - L(\theta)|$ as follows. Let

$$\hat{h}_{n,\theta}(z) = \psi(z,\theta)\left(\sum_{k=1}^K \frac{\hat{\lambda}_k\omega_k(z)}{\widehat{\Omega}_{n,k}}\right)^{-1}, \quad \text{and} \quad h_\theta(z) = \psi(z,\theta)\left(\sum_{k=1}^K \frac{\lambda_k\omega_k(z)}{\Omega_k}\right)^{-1}.$$

We have (see Lemma 2 for details):

$$\begin{aligned}
&\left|\widetilde{L}_n(\theta) - L(\theta)\right| \\
&= \left|\int \psi(z,\theta)\left(\sum_{k=1}^K \frac{\hat{\lambda}_k\omega_k(z)}{\widehat{\Omega}_{n,k}}\right)^{-1} d\widehat{P}_n(z) - \int \psi(z,\theta)\left(\sum_{k=1}^K \frac{\lambda_k\omega_k(z)}{\Omega_k}\right)^{-1} d\bar{P}(z)\right| \\
&= \left|\int \hat{h}_{n,\theta}(z)d\widehat{P}_n(z) - \int h_\theta(z)d\bar{P}(z)\right| \\
&\leq \left\|\hat{h}_{n,\theta} - h_\theta\right\|_\infty + \|h_\theta\|_\infty \sum_{k=1}^K \left|\hat{\lambda}_k - \lambda_k\right| + \sum_{k=1}^K \hat{\lambda}_k \left|\int h_\theta d\widehat{P}_k - \int h_\theta dP_k\right|
\end{aligned} \tag{3.5}$$

As it is assumed that $|\psi(z,\theta)| \leq 1$, the first term of the right hand side of (3.5) actually depends on $\|\widehat{\mathbf{\Omega}}_n - \mathbf{\Omega}\|_\infty$ only, and can thus be bounded uniformly over Θ using Proposition 3. Similarly, the second term depends on the $|\hat{\lambda}_k - \lambda_k|$, and can be uniformly upper bounded using Assumption 4. Finally, the last term writes as the sum of empirical processes, indexed by Θ, for which standard arguments apply. This leads to the following theorem, whose complete proof can be found in Appendix C.4.

Theorem 1. *Suppose that Assumptions 4, 5, 6, 7, and 8 are satisfied. Then, there exist* $M, M', M'', c, c', c'', \gamma, n_0$*, depending only on* $K, C_\lambda, \underline{\lambda}, \kappa, \varepsilon, \sigma, C_\Theta, r$*, such that for all* $t > 0$ *and* $n \geq n_0$ *it holds:*

$$\mathbb{P}\left\{\sup_{\theta\in\Theta}\left|\widetilde{L}_n(\theta) - L(\theta)\right| > \frac{\gamma}{\sqrt{n}} + t\right\} \leq Me^{-cn} + M'e^{-c'nt^2} + (\sqrt{n}t)^r M''e^{-c''nt^2}.$$

An important corollary of Theorem 1 is obtained by combining the above bound with the following argument. Let $\tilde{\theta}_n = \operatorname{argmin}_\Theta \widetilde{L}_n(\theta)$, we have

$$L(\tilde{\theta}_n) - \inf_{\theta\in\Theta} L(\theta) \leq 2\sup_{\theta\in\Theta}\left|\widetilde{L}_n(\theta) - L(\theta)\right|.$$

This immediately results in Corollary 1, which reveals that minimizers of the "debiased" version of the empirical risk achieve exactly the same learning rate as minimizers of an (unbiased) empirical risk based on $n \geq 1$ independent observations $Z_1, \ldots, Z_n$ drawn from the test distribution P. Notice that an analogous bound for the expectation of the risk excess of Equation (2.14)'s minimizers can be proved using the same argument.

Corollary 1. *Suppose that the assumptions of Theorem 1 are satisfied, and keep the same values for* $M, M', M'', c, c', c'', \gamma, n_0$*. Let* $\tilde{\theta}_n$ *be any minimizer of the debiased risk* $\widetilde{L}_n$ *defined in* (2.14)*. Then, for all* $t > 0$ *and* $n \geq n_0$ *it holds:*

$$\mathbb{P}\left\{L(\tilde{\theta}_n) - \inf_{\theta\in\Theta} L(\theta) > \frac{2\gamma}{\sqrt{n}} + 2t\right\} \leq Me^{-cn} + M'e^{-c'nt^2} + (\sqrt{n}t)^r M''e^{-c''nt^2}.$$

Another application of particular interest is the case where $q = 1$ (univariate case), and where the class $\mathcal{F}$ is the set composed of all indicator functions $z \in \mathbb{R} \mapsto \mathbb{I}\{z \leq \tau\}$, for $\tau \in \mathbb{R}$. Recall that in the i.i.d. case, the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality, see e.g., [27], then yields: $\forall t \geq 0$,

$$\mathbb{P}\left\{\sup_{z\in\mathbb{R}}\left|(\widehat{P}_n - P)((-\infty,\ z])\right| \geq t\right\} \leq 2e^{-2nt^2} \tag{3.6}$$

where $z \in \mathbb{R} \mapsto \widehat{P}_n((-\infty,\ z]) = (1/n)\sum_{i=1}^n \mathbb{I}\{Z_i \leq z\}$ denotes the empirical cumulative distribution function based on i.i.d. observations $Z_1, \ldots, Z_n$ drawn from the univariate probability distribution P. Analogously, under the sample biasing models, the quantity (3.4) then corresponds to the maximal deviation $\sup_{z\in\mathbb{R}}\left|(\widetilde{P}_n - P)((-\infty,\ z])\right|$. While a functional central limit theorem for this cdf estimator is established in [16], the application of Theorem 1 allows to refine this statement from a nonasymptotic perspective. However, recalling that the class composed of half-lines is of VC dimension 2, and thus satisfies Assumption 8 with $r = 2V - 2 = 2$, the bound obtained contains a term of order $nt^2e^{-nt^2}$, which does not match (3.6). A sharper analysis, leveraging the fact that $\mathcal{F}$ is a class of indicator functions, is necessary, see Appendix C.5. The refined rate thus achieved (Theorem 2) then matches (3.6), and provides an exact extension of the DKW inequality under biased sampling models.

Theorem 2. *Suppose that Assumptions 4, 5, 6, and 7 are satisfied. Then, there exist $M, M', c, c', \gamma, n_0$, depending only on $K, C_\lambda, \underline{\lambda}, \kappa, \varepsilon, \sigma$, such that for all $t > 0$ and $n \geq n_0$ it holds:*

$$\mathbb{P}\left\{\sup_{z\in\mathbb{R}}\left|(\widetilde{P}_n - P)((-\infty,\ z])\right| > \frac{\gamma}{\sqrt{n}} + t\right\} \leq Me^{-cn} + M'e^{-c'nt^2}.$$

Finally, we point out that the finite sample analysis carried out in this paper can be used in the context of M-estimation as well. Indeed, under the additional hypothesis that there exists a unique minimizer θ^* of the true risk $L(\cdot)$ in the state space $\Theta \subset \mathbb{R}^d$ combined with usual smoothness/coercivity assumptions related to the risk functional, nonasymptotic bounds for the estimation error $||\tilde{\theta}_n - \theta||$ can be classically deduced from the excess of risk bounds proved here.

4. Numerical experiments

In this section, we display numerical results illustrating the performance of the extension of the ERM approach we propose when training data suffer from selection bias. First, observe that the procedure is by no means computationally expensive, the sole difference with standard methods lying in the computation of the weights involved in the risk functional. In addition, it can be readily implemented in a plug-in manner with most machine learning libraries, using e.g., scikit-learn's `sample_weight` option during the learning stage, see [31].

Consider first the *Boston* housing dataset problem. It is a regression problem, where one has to predict the price of a house on the range $[0, 50]$, based on 14 attributes such as the number of rooms or statistics about the neighborhood. One can easily imagine that such a dataset is actually composed of two samples: one dataset taken from a local estate agency, large but containing cheap houses as the neighborhood is not very trendy, and a second one, national and unbiased but smaller. Of course, running ERM on the pooled sample without debiasing procedure should result in a global underestimation of the prices. To replicate this framework, we have implemented the following protocol. From the 500 available observations, 100 are kept for the testing phase. From the remaining 400 observations, two samples are extracted: a first one of size 200, sampled among the cheapest houses, i.e., with prices lower than 22 (see Figure 4), and a second unbiased of size 100 (i.e., sampled uniformly at random). The biasing functions are therefore $\omega_1(z) = \mathbb{I}\{y \leq 22\}$, and $\omega_2(z) \equiv 1$. Then, we have trained several ERM-based algorithms, namely Ridge Regression (RR), Support Vector Regressors (SVR), and Random Forest (RF), on the total sample of size 300, with and without debiasing. A third model is trained on the small unbiased sample only. All algorithms have been run with several choices of hyperparameters around the default value. Results in terms of Mean Square Error (MSE) on the test sample of size 100, averaged over 100 runs, are displayed in Table 1 (top). Except for SVR with very small regularization, the debiased procedure (db-ERM) outperforms standard ERM and ERM on the unbiased sample (ub-ERM).

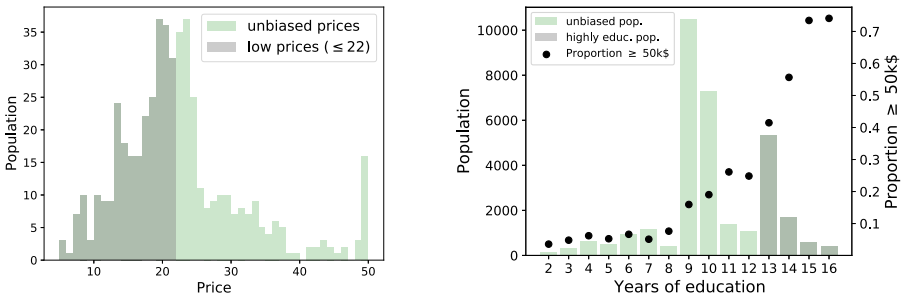


Fig 4: House prices in *Boston* dataset (left), and years of education against the proportion of people earning more than $50k$\$ yearly in *Adult* dataset (right).

Note that the previous example cannot be treated as a Covariate Shift (CS) problem, since sample bias applies to the output, breaking the CS assumption. However, one might argue that biasing directly the output favors too much our procedure. In the next example, we propose a binary classification problem where sample bias applies on the covariates. Consider the machine learning problem associated to the *Adult* dataset, also known as the *Census Income* dataset. It is a binary classification task, where the goal is to predict whether a person's income exceeds $50,000$\$ a year, based on census data. As can be seen in Figure 4 (left), the proportion of persons having an income exceeding 50k\$ a year substantially depends on the number of years of education. If highly educated people happen to be over-represented in the dataset (it is for instance more convenient to poll people concentrated in big cities, who have usually studied longer than people living in the countryside), it should deteriorate the predictions in absence of a debiasing procedure. In order to highlight the interest of our debiasing procedure, we have implemented the following experimental protocol. From the whole dataset, 1500 observations are kept for the testing phase. From the rest are sampled two subgroups: one of 12+ years of education people of size 5900, and one unbiased (i.e., sampled uniformly from the entire population) of size 100. Then, logistic regression models (LogReg) and RFs are trained on the concatenation of the 6000 observations, with standard and debiased ERM, as well as on the small second sample of size 100. Numerical results are displayed in Table 1 (bottom) in terms of test prediction scores. Again, debiased ERM shows the best performances. Another comment that can be made is that the advantage brought by the debiasing decreases with the capacity of the model class (i.e., small λ, big C, or large number of trees).

Hence, we have presented two learning examples, one regression task and one classification task, which cannot be tackled through ordinary CS (either bias applies to the target, or the conditional laws obviously change), empirically endorsing the soundness of our debiased ERM approach. Additional experiments leading to similar conclusions are presented in Appendix D. Notice finally that the code used to compute the debiasing weights is publicly available as a Python package at the following GitHub repository: plaforgue/db_learn.

TABLE 1
MSEs on Boston *and prediction scores on* Adult*, averaged over 100 runs.*

		ERM	db-ERM	ub-ERM
Boston	RR ($\lambda = 0$)	27.41 ± 8.83	**25.62 ± 6.63**	28.38 ± 7.99
	RR ($\lambda = 0.1$)	27.46 ± 8.90	**25.59 ± 6.67**	28.11 ± 7.73
	RR ($\lambda = 1$)	27.94 ± 9.25	**25.72 ± 6.84**	28.05 ± 7.61
	SVR ($C = 0.1$)	99.63 ± 21.55	**86.29 ± 18.79**	86.60 ± 18.67
	SVR ($C = 1$)	100.02 ± 21.87	**85.66 ± 19.04**	86.04 ± 18.67
	SVR ($C = 10$)	97.27 ± 22.38	88.37 ± 21.68	**82.35 ± 18.60**
	RF (trees=10)	19.83 ± 7.13	**19.11 ± 6.97**	20.46 ± 6.50
	RF (trees=100)	18.20 ± 6.46	**17.93 ± 6.58**	18.71 ± 6.10
	RF (trees=1000)	18.11 ± 6.61	**17.69 ± 6.59**	18.54 ± 6.16
Adult	LogReg ($C = 0.1$)	63.87 ± 1.58	**79.25 ± 1.67**	78.24 ± 1.97
	LogReg ($C = 1$)	63.81 ± 1.67	**79.51 ± 1.80**	77.79 ± 2.25
	LogReg ($C = 10$)	63.87 ± 1.65	**79.53 ± 1.78**	78.01 ± 2.45
	RF (trees=10)	39.00 ± 3.74	**40.27 ± 4.16**	18.48 ± 6.52
	RF (trees=100)	44.37 ± 3.28	**45.36 ± 3.89**	23.81 ± 5.71
	RF (trees=1000)	44.92 ± 3.24	**46.03 ± 3.61**	24.42 ± 5.51

5. Conclusion

In this article, we have provided a sound methodology to address selection bias issues in statistical learning. We have extended the paradigmatic ERM approach to the situation where learning is based on several biased training samples. In contrast to alternative techniques previously documented in the literature, the method proposed covers a wide range of sample bias scenarios, and applies to any ERM-like learning algorithm. It relies on a preliminary debiasing of the raw empirical risk functional in the spirit of the procedure introduced in [46] for cumulative distribution function estimation. The nonasymptotic theoretical analysis carried out under mild assumptions shows that the rate achieved is the same as that attained in absence of any selection bias. Numerical experiments are also documented, validating our theoretical findings. A natural direction for future research is now to extend the statistical learning approach promoted in this article to situations were the biasing models at work are only partially known.

Appendix A: Derivation of Equation (2.7)

Some computations, omitted in the core text for the sake of readability, are detailed below. For all $k \leq K$, we have:

$$\Omega_k = \int \omega_k(z) dP(z) = \frac{\int \omega_k(z) dP(z)}{\int dP(z)} = \frac{\int \omega_k(z) \left(\sum_{l=1}^K \frac{\lambda_l}{\Omega_l} \omega_l(z) \right)^{-1} d\bar{P}(z)}{\int \left(\sum_{l=1}^K \frac{\lambda_l}{\Omega_l} \omega_l(z) \right)^{-1} d\bar{P}(z)}$$

$$= \frac{\int \omega_k(z) \left(\sum_{l=1}^K \frac{\lambda_l}{W_l^*} \omega_l(z) \right)^{-1} d\bar{P}(z)}{\int \left(\sum_{l=1}^K \frac{\lambda_l}{W_l^*} \omega_l(z) \right)^{-1} d\bar{P}(z)}$$
$$= \frac{W_k^*}{\int \left(\sum_{l=1}^K \frac{\lambda_l}{W_l^*} \omega_l(z) \right)^{-1} d\bar{P}(z)},$$

where we have successively used the fact that $\int dP = 1$, Equation (2.2), the fact that $\boldsymbol{W}^* \propto \boldsymbol{\Omega}$ and Equation (2.4).

Appendix B: A simplistic example

Here, we exhibit a simple example where the training data samples are biased, but no system solving is required to form a debiased empirical distribution. The flagship problem in supervised learning is multi-class classification and consists in the simplest situation, where $Z = (X, Y)$, Y being a discrete random variable valued in $\{1, \ldots, Q\}$ with $Q \geq 1$ say, and the r.v. X takes its values in a measurable space $\mathcal{X}$ and models some information hopefully useful to predict Y. The parameter space Θ is a set $\mathcal{G}$ of measurable mappings (i.e., classifiers) $g : \mathcal{X} \to \{1, \ldots, Q\}$ and the loss function is given by $\ell(g, (x, y)) = \mathbb{I}\{g(x) \neq y\}$ for all g in $\mathcal{G}$ and any $(x, y) \in \mathcal{X} \times \{1, \ldots, Q\}$. The distribution P of the random pair (X, Y) can be either described by X's marginal distribution μ and the posterior probability $\eta(x) = (\eta_1(x), \ldots, \eta_Q(x))$, where $\eta_q(x) = \mathbb{P}\{Y = q \mid X = x\}$ for $q \in \{1, \ldots, Q\}$, or else by the $((p_1, F_1), \ldots, (p_Q, F_Q))$ where $p_q = \mathbb{P}\{Y = q\}$ and F_q is X's conditional distribution given $Y = q$ with $q \in \{1, \ldots, Q\}$. Observe that $p_1 + \ldots + p_Q = 1$, we assume that $p_q \in (0, 1)$ for all $q \in \{1, \ldots, Q\}$. It is very common that the class probabilities in the training datasets are significantly different from those in the test stage, the p_q's namely. We thus consider the case where, for all $k \in \{1, \ldots, K\}$, the distribution P_k of the k-th training dataset $\mathcal{D}_k = \{(X_{k,1}, Y_{k,1}), \ldots, (X_{k,n_k}, Y_{k,n_k})\}$ is described by $((p_{k,1}, F_1), \ldots, (p_{k,Q}, F_Q))$, where the vector of class probabilities $\boldsymbol{p}_k = (p_{k,1}, \ldots, p_{k,Q}) \in [0, 1]^Q$ (note incidentally that $p_{k,1} + \ldots + p_{k,Q} = 1$) may differ from $\boldsymbol{p} = (p_1, \ldots, p_Q)$. We point out that it may happen that certain class probabilities $p_{k,q}$ are equal to zero, so that some labels cannot be observed among certain data samples. The likelihood function takes the form

$$\forall (x, y) \in \mathcal{X} \times \{1, \ldots, Q\}, \quad \frac{dP_k}{dP}(x, y) = \sum_{q=1}^{Q} \mathbb{I}\{y = q\}(p_{k,q}/p_q),$$

which reveals that it depends on the label y solely. Hence, in this very simple case, we have $\omega_k(x, y) = p_{k,y}/p_y$ and $\Omega_k = 1$ for all $(y, k) \in \{1, \ldots, Q\} \times \{1, \ldots, K\}$ and there is no need for solving any system to compute a nearly debiased empirical distribution. Observe that, for all $\kappa > 0$, vertices k and l in $\{1, \ldots, K\}$ are connected in the graph G_κ iff $\sum_{q \leq Q} p_{k,q} p_{l,q}/p_q \geq \kappa$. However, in this situation the biasing functions ω_k can be directly estimated from the

data samples, replacing $p_{k,q}$ by $n_{k,q}/n_k$ with $n_{k,q} = \sum_{i=1}^{n_k} \mathbb{I}\{Y_{k,i} = q\}$ for all $(k,q) \in \{1, \ldots, K\} \times \{1, \ldots, Q\}$ and computing

$$\widehat{\omega}_k(x,y) = \widehat{p}_{k,y}/p_y,$$

for all $(y,k) \in \{1, \ldots, Q\} \times \{1, \ldots, K\}$. One may then consider the estimator

$$\left(\sum_{k=1}^{K} \frac{n_k}{n} \widehat{\omega}_k(x,y)\right)^{-1} \widehat{P}_n$$

of the distribution P.

Appendix C: Technical proofs

We now provide the technical proofs of the results stated in the paper. Recall that for notation simplicity we used universal constants $M, M', c, c', \gamma, n_0$, in the core text. For the sake of clarity, we now index them by propositions, such that $M_i, M_i', c_i, c_i', \gamma_i, n_{0,i}$ correspond to $M, M', c, c', \gamma, n_0$ for Proposition i.

C.1. Proof of Proposition 1

First, we introduce the following notation. Let $\bar{D}$ and $\widehat{D}_n$ be the two functions from $\mathbb{R}^K$ to $\mathbb{R}$ such that for all $\boldsymbol{u} = (u_1, \ldots, u_K) \in \mathbb{R}^K$:

$$\bar{D}(\boldsymbol{u}) = \int \log\left[\sum_{l=1}^{K} e^{u_l} \omega_l(z)\right] d\bar{P}(z) - \sum_{l=1}^{K} \lambda_l u_l,$$

$$\widehat{D}_n(\boldsymbol{u}) = \int \log\left[\sum_{l=1}^{K} e^{u_l} \omega_l(z)\right] d\widehat{P}_n(z) - \sum_{l=1}^{K} \hat{\lambda}_l u_l.$$

Let $\boldsymbol{u}^* = \operatorname{argmin}_{\boldsymbol{u}} \bar{D}(\boldsymbol{u})$, and similarly $\hat{\boldsymbol{u}}_n = \operatorname{argmin}_{\boldsymbol{u}} \widehat{D}_n(\boldsymbol{u})$. We also compute the gradients $\bar{D}'$, $\widehat{D}_n'$ and the Hessian matrices $\bar{D}''$, $\widehat{D}_n''$, of these two smooth functions. For all $\boldsymbol{u} = (u_1, \ldots, u_K) \in \mathbb{R}^K$ and all $k, k' \leq K$, we have:

$$\left[\bar{D}'(\boldsymbol{u})\right]_k = \int \frac{e^{u_k}\omega_k(z)}{\sum_{l=1}^{K} e^{u_l}\omega_l(z)} d\bar{P}(z) - \lambda_k,$$

$$\left[\widehat{D}_n'(\boldsymbol{u})\right]_k = \int \frac{e^{u_k}\omega_k(z)}{\sum_{l=1}^{K} e^{u_l}\omega_l(z)} d\widehat{P}_n(z) - \hat{\lambda}_k,$$

$$\left[\bar{D}''(\boldsymbol{u})\right]_{k,k'} = \int \left[\frac{e^{u_k}\omega_k(z)\delta_{kk'}}{\sum_{l=1}^{K} e^{u_l}\omega_l(z)} - \frac{e^{u_k}\omega_k(z)e^{u_{k'}}\omega_{k'}(z)}{\left(\sum_{l=1}^{K} e^{u_l}\omega_l(z)\right)^2}\right] d\bar{P}(z),$$

$$\left[\widehat{D}_n''(\boldsymbol{u})\right]_{k,k'} = \int \left[\frac{e^{u_k}\omega_k(z)\delta_{kk'}}{\sum_{l=1}^K e^{u_l}\omega_l(z)} - \frac{e^{u_k}\omega_k(z)e^{u_{k'}}\omega_{k'}(z)}{\left(\sum_{l=1}^K e^{u_l}\omega_l(z)\right)^2} \right] d\widehat{P}_n(z).$$

Observe now that Systems (2.4) and (2.6) are equivalent to $\bar{D}'(\boldsymbol{u}^*) = \mathbf{0}$ and $\widehat{D}_n'(\hat{\boldsymbol{u}}_n) = \mathbf{0}$ respectively, with $\boldsymbol{u}^* = \log(\boldsymbol{\lambda}/\boldsymbol{W}^*)$ and $\hat{\boldsymbol{u}}_n = \log(\hat{\boldsymbol{\lambda}}/\widehat{\boldsymbol{W}}_n)$, where the division and the logarithm are meant componentwise. Recall that Systems (2.4) and (2.6) are homogeneous of degree 0. Equivalently, $\bar{D}$ and $\widehat{D}_n$ are invariant under translation of vectors that are colinear with $\mathbf{1}$, i.e., $\bar{D}(\boldsymbol{u}+c\mathbf{1}) = \bar{D}(\boldsymbol{u})$ and $\widehat{D}_n(\boldsymbol{u}+c\mathbf{1}) = \widehat{D}_n(\boldsymbol{u})$ for all $\boldsymbol{u} \in \mathbb{R}^K$ and $c \in \mathbb{R}$. To ensure uniqueness of the solutions to Systems (2.4) and (2.6), we consider $\boldsymbol{W}^*$ and $\widehat{\boldsymbol{W}}_n$ such that $\max_{k\le K} W_k^*/\lambda_k = 1$, and $\max_{k\le K} \widehat{W}_{n,k}/\hat{\lambda}_k = 1$. In terms of $\boldsymbol{u}^*$ and $\hat{\boldsymbol{u}}_n$, this normalization writes $\min_{k\le K} u_k^* = \min_{k\le K} \hat{u}_{n,k} = 0$. We now show that there exists $\rho > 0$ such that for all $k \le K$ it holds:

$$\rho \le \widehat{W}_{n,k} \le 1 \qquad \text{and} \qquad \rho \le W_k^* \le 1.$$

The upper bounds above are immediate, insofar as for, all $k \le K$, we have:

$$\widehat{W}_{n,k} \le \hat{\lambda}_k \le 1 \qquad \text{and similarly} \qquad W_k^* \le \lambda_k \le 1.$$

To derive ρ, we show that there exist $U > 0$ such that:

$$\forall k \le K, \qquad \hat{u}_{n,k} \le U \qquad \text{and} \qquad u_k^* \le U.$$

The proof mechanism is as follows. First, we derive a lower bound of $\widehat{D}_n(\boldsymbol{u})$, that depends linearly on u_{k_0} and u_{k_1}, for any couple (k_0, k_1) being an edge in $\widehat{G}_n$. Next, we apply this lower bound at point $\hat{\boldsymbol{u}}_n$, with k_0 such that $\hat{u}_{n,k_0} = 0$ (such an index exists by the normalization we impose). Combining this lower bound with the observation that $\widehat{D}_n(\hat{\boldsymbol{u}}_n) \le \widehat{D}_n(\mathbf{0})$, we obtain an upper bound on $\hat{u}_{n,k_1}$. Finally, this approach is used recursively to bound the neighbors of k_1, and so on and so forth. The graph $\widehat{G}_n$ being connected by the first claim of Proposition 1, every component $\hat{u}_{n,k}$ is attained after at most $K-1$ iterations.

Let (k_0, k_1) be an edge in $\widehat{G}_n$. Using the definition of $\widehat{P}_n$ it holds:

$$\begin{aligned}\widehat{D}_n(\boldsymbol{u}) &= \int \log\left[\sum_{l=1}^K e^{u_l}\omega_l(z)\right] d\widehat{P}_n(z) - \sum_{l=1}^K \hat{\lambda}_l u_l \\ &= \sum_{k=1}^K \hat{\lambda}_k \left(\int \log\left[\sum_{l=1}^K e^{u_l}\omega_l(z)\right] d\widehat{P}_k(z) - u_k \right).\end{aligned}$$

For $k \ne k_0$, it holds:

$$\int \log\left[\sum_{l=1}^K e^{u_l}\omega_l(z)\right] d\widehat{P}_k(z) - u_k \ge \int \log(e^{u_k}\varepsilon) d\widehat{P}_k(z) - u_k = \log(\varepsilon).$$

For $k = k_0$, we have:

$$\begin{aligned}\int \log\left[\sum_{l=1}^{K} e^{u_l}\omega_l(z)\right] d\widehat{P}_{k_0}(z) &\geq \int \log(\varepsilon)(1 - \mathbb{I}\{\omega_{k_1}(z) > 0\})d\widehat{P}_{k_0}(z) \\ &\quad + \int \log(e^{u_{k_1}}\varepsilon)\mathbb{I}\{\omega_{k_1}(z) > 0\}d\widehat{P}_{k_0}(z) \\ &= \log(\varepsilon) + u_{k_1}\int \mathbb{I}\{\omega_{k_1}(z) > 0\}d\widehat{P}_{k_0}(z).\end{aligned}$$

From the proof of the first claim of Proposition 1 (see Equation (3.3)), we know that it holds:

$$\int \mathbb{I}\{\omega_{k_1}(z) > 0\}d\widehat{P}_{k_0}(z) \geq \frac{\kappa\varepsilon}{2} \tag{C.1}$$

so that one gets:

$$\begin{aligned}\widehat{D}_n(\boldsymbol{u}) &\geq (1 - \hat{\lambda}_{k_0})\log(\varepsilon) + \hat{\lambda}_{k_0}\left(\log(\varepsilon) + \frac{\kappa\varepsilon}{2}u_{k_1} - u_{k_0}\right) \\ &\geq \log(\varepsilon) + \frac{\underline{\lambda}\kappa\varepsilon}{2}u_{k_1} - u_{k_0}. \end{aligned} \tag{C.2}$$

Observe also that we have:

$$\widehat{D}_n(\hat{\boldsymbol{u}}_n) \leq \widehat{D}_n(\mathbf{0}) = \int \log\left[\sum_{l=1}^{K}\omega_l(z)\right] d\widehat{P}_n(z) \leq \log(K). \tag{C.3}$$

Combining Equation (C.2) evaluated at point $\hat{\boldsymbol{u}}_n$ and Equation (C.3), we obtain:

$$\hat{u}_{n,k_1} \leq \frac{2(\log(K/\varepsilon) + \hat{u}_{n,k_0})}{\underline{\lambda}\kappa\varepsilon}. \tag{C.4}$$

The last step consists in extending this bound to every $\hat{u}_{n,k}$. To do so, we first set (without loss of generality) $\hat{u}_{n,k_0} = \min_{k\leq K}\hat{u}_{n,k} = 0$. Recall also the definition of graph $G_{\min}$, as introduced in the proof of the first claim of Proposition 1. We can then apply Equation (C.4) to all k_1 that are neighbors of k_0 in $G_{\min}$. Next, notice that this method can be used in a recursive fashion, with now the k_1 as anchor points. Eventually, every $\hat{u}_{n,k}$ is attained, as $G_{\min}$ is connected. Equation (C.4) becoming looser and looser as it is applied, the last question is *how many recursive steps are required?* The minimum number of recursive steps needed is the biggest (among $k \leq K$) shortest path (in $G_{\min}$) between k_0 and k, denoted $\mathrm{diam}(G_{\min}, k_0)$. Combining all the arguments, we get:

$$\begin{aligned}\forall k \leq K, \qquad \hat{u}_{n,k} &\leq \frac{\left(\frac{2}{\underline{\lambda}\kappa\varepsilon}\right)^{\mathrm{diam}(G_{\min},k_0)+1} - 1}{\frac{2}{\underline{\lambda}\kappa\varepsilon} - 1}\log(K/\varepsilon) \\ &\leq \frac{\left(\frac{2}{\underline{\lambda}\kappa\varepsilon}\right)^{K} - 1}{\frac{2}{\underline{\lambda}\kappa\varepsilon} - 1}\log(K/\varepsilon).\end{aligned}$$

Therefore, for all $k \leq K$ it holds:

$$\widehat{W}_{n,k} = \hat{\lambda}_k e^{-\hat{u}_{n,k}} \geq \underline{\lambda} e^{-U} \coloneqq \rho \tag{C.5}$$

with

$$U = \frac{\left(\frac{2}{\underline{\lambda}\kappa\varepsilon}\right)^K - 1}{\frac{2}{\underline{\lambda}\kappa\varepsilon} - 1} \log(K/\varepsilon). \tag{C.6}$$

Finally, note that the exact same method can be applied to $\boldsymbol{u}^*$, by substituting $\widehat{P}_k$ with P_k in the computations. □

C.2. Proof of Proposition 2

First, we prove the following lemma, ensuring that the deviation $\left\|\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^*\right\|_2$ is upper bounded by the deviation $\|\hat{\boldsymbol{u}}_n - \boldsymbol{u}^*\|_2$.

Lemma 1. *Suppose that Assumption 4 is satisfied. Then it holds:*

$$\left\|\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^*\right\|_2 \leq \|\hat{\boldsymbol{u}}_n - \boldsymbol{u}^*\|_2 + C_\lambda \sqrt{\frac{K}{n}}.$$

Proof. For all $k \leq K$ it holds:

$$\begin{aligned}
\left|\widehat{W}_{n,k} - W_k^*\right| &= \left|\hat{\lambda}_k e^{-\hat{u}_{n,k}} - \lambda_k e^{-u_k^*}\right| \\
&\leq \left|\hat{\lambda}_k e^{-\hat{u}_{n,k}} - \hat{\lambda}_k e^{-u_k^*}\right| + \left|\hat{\lambda}_k e^{-u_k^*} - \lambda_k e^{-u_k^*}\right| \\
&\leq \left|e^{-\hat{u}_{n,k}} - e^{-u_k^*}\right| + \left|\hat{\lambda}_k - \lambda_k\right| \\
&\leq |\hat{u}_{n,k} - u_k^*| + \frac{C_\lambda}{\sqrt{n}}
\end{aligned}$$

where we have used the definition of $\boldsymbol{u}^*$ and $\hat{\boldsymbol{u}}_n$, the triangle inequality, the fact that $\hat{\lambda}_k \leq 1$, and that $u_k^* \geq 0$, the mean value theorem on $u \mapsto e^{-u}$ with $\hat{u}_{n,k} \geq 0$, and Assumption 4. Applying again the triangle inequality finally yields:

$$\begin{aligned}
\left\|\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^*\right\|_2 &= \sqrt{\sum_{k=1}^K \left|\hat{W}_{n,k} - W_k^*\right|^2} \\
&\leq \sqrt{\sum_{k=1}^K \left(|\hat{u}_{n,k} - u_k^*| + \frac{C_\lambda}{\sqrt{n}}\right)^2} \\
&= \left\||\hat{\boldsymbol{u}}_n - \boldsymbol{u}^*| + \frac{C_\lambda}{\sqrt{n}}\mathbf{1}\right\|_2 \\
&\leq \|\hat{\boldsymbol{u}}_n - \boldsymbol{u}^*\|_2 + C_\lambda \sqrt{\frac{K}{n}}.
\end{aligned}$$

□

Next, we show that Proposition 1 allows to bound the deviation $\|\hat{\boldsymbol{u}}_n - \boldsymbol{u}^*\|_2$ in terms of the deviation $\left\|\widehat{D}'_n(\boldsymbol{u}^*) - \bar{D}'(\boldsymbol{u}^*)\right\|_2$.

Proposition 4. *Suppose that Assumptions 4, 5, 6, and 7 are satisfied. Then, there exist $M_4, c_4, n_{0,4}, L$, depending only on $K, C_\lambda, \underline{\lambda}, \kappa, \varepsilon, \sigma$, such that for all $n \geq n_{0,4}$ it holds with probability at least $1 - M_4 \exp(-c_4 n)$:*

$$\|\hat{\boldsymbol{u}}_n - \boldsymbol{u}^*\|_2 \leq L \left\|\widehat{D}'_n(\boldsymbol{u}^*) - \bar{D}'(\boldsymbol{u}^*)\right\|_2 .$$

Proof. Define $F : [0,1] \to \mathbb{R}^K$ such that $F(t) = \widehat{D}'_n\left(\hat{\boldsymbol{u}}_n + t(\boldsymbol{u}^* - \hat{\boldsymbol{u}}_n)\right)$. It holds:

$$\begin{aligned}
F(1) - F(0) &= \left(\int_0^1 F'(t)dt\right) \\
\widehat{D}'_n(\boldsymbol{u}^*) - \widehat{D}'_n(\hat{\boldsymbol{u}}_n) &= \left(\int_0^1 \left[\widehat{D}''_n(\hat{\boldsymbol{u}}_n + t(\boldsymbol{u}^* - \hat{\boldsymbol{u}}_n))\right](\boldsymbol{u}^* - \hat{\boldsymbol{u}}_n)dt\right) \\
\widehat{D}'_n(\boldsymbol{u}^*) - \bar{D}'(\boldsymbol{u}^*) &= \left(\int_0^1 \left[\widehat{D}''_n(\hat{\boldsymbol{u}}_n + t(\boldsymbol{u}^* - \hat{\boldsymbol{u}}_n))\right] dt\right)(\boldsymbol{u}^* - \hat{\boldsymbol{u}}_n) \qquad \text{(C.7)}
\end{aligned}$$

where the integral over matrices must be understood componentwise. The key point to relate $\left\|\widehat{D}'_n(\boldsymbol{u}^*) - \bar{D}'(\boldsymbol{u}^*)\right\|_2$ to $\|\hat{\boldsymbol{u}}_n - \boldsymbol{u}^*\|_2$ is then to study the smallest eigenvalues of $\int_0^1 \left[\widehat{D}''_n(\hat{\boldsymbol{u}}_n + t(\boldsymbol{u}^* - \hat{\boldsymbol{u}}_n))\right]dt$. From the definition of $\widehat{D}''_n$, one can see that its smallest eigenvalue is 0, associated to $\mathbf{1}$. Hopefully, thanks to the normalization, $\hat{\boldsymbol{u}}_n - \boldsymbol{u}^*$ is not collinear to $\mathbf{1}$ unless $\hat{\boldsymbol{u}}_n = \boldsymbol{u}^*$. Let $\hat{\boldsymbol{u}}_n - \boldsymbol{u}^* = c\mathbf{1} + \boldsymbol{w}$ be the decomposition of $\hat{\boldsymbol{u}}_n - \boldsymbol{u}^*$ on $\mathrm{Span}(\mathbf{1}) \otimes \mathrm{Span}(\mathbf{1})^\perp$, such that $\mathbf{1}^\top \boldsymbol{w} = 0$. One can check that $\|\boldsymbol{w}\|_\infty \geq \|c\mathbf{1}\|_\infty$, so that it holds

$$\|\hat{\boldsymbol{u}}_n - \boldsymbol{u}^*\|_2^2 = \|\boldsymbol{w}\|_2^2 + \|c\mathbf{1}\|_2^2 \leq \|\boldsymbol{w}\|_2^2 + K\|c\mathbf{1}\|_\infty^2 \leq (K+1)\|\boldsymbol{w}\|_2^2. \qquad \text{(C.8)}$$

Combining Equations (C.7) and (C.8), one gets

$$\begin{aligned}
\left\|\widehat{D}'_n(\boldsymbol{u}^*) - \bar{D}'(\boldsymbol{u}^*)\right\|_2 &\geq \sigma_2\left(\int_0^1 \left[\widehat{D}''_n(\hat{\boldsymbol{u}}_n + t(\boldsymbol{u}^* - \hat{\boldsymbol{u}}_n))\right] dt\right)\|\boldsymbol{w}\|_2 \\
&\geq \inf_{\boldsymbol{v} \in [\hat{\boldsymbol{u}}_n, \boldsymbol{u}^*]} \sigma_2\left(\widehat{D}''_n(\boldsymbol{v})\right) \frac{\|\hat{\boldsymbol{u}}_n - \boldsymbol{u}^*\|_2}{\sqrt{K+1}}
\end{aligned}$$

where $\sigma_2(A)$ denotes the second smallest eigenvalue of a matrix A. We have now to find a lower bound of $\inf_{\boldsymbol{v} \in [\hat{\boldsymbol{u}}_n, \boldsymbol{u}^*]} \sigma_2\big(\widehat{D}''_n(\boldsymbol{v})\big)$. Let $\boldsymbol{v} \in [\hat{\boldsymbol{u}}_n, \boldsymbol{u}^*]$. For notation simplicity, we omit the $\boldsymbol{v}$ in the following, and use $\widehat{D}''_n$ and $\bar{D}''$. It holds:

$$\begin{aligned}
\sigma_2(\widehat{D}''_n) &= \inf_{\substack{\|\boldsymbol{u}\|=1, \\ \mathbf{1}^\top \boldsymbol{u} = 0}} \boldsymbol{u}^\top \widehat{D}''_n \boldsymbol{u} \\
&= \inf_{\substack{\|\boldsymbol{u}\|=1, \\ \mathbf{1}^\top \boldsymbol{u} = 0}} \boldsymbol{u}^\top (\widehat{D}''_n - \bar{D}'')\boldsymbol{u} + \boldsymbol{u}^\top \bar{D}'' \boldsymbol{u}
\end{aligned}$$

$$\geq -\|\widehat{D}''_n - \bar{D}''\|_{\sigma_\infty} + \sigma_2(\bar{D}'')$$

where $\|A\|_{\sigma_\infty}$ denotes the Schatten ∞-norm such that $\|A\|_{\sigma_\infty} = \|\sigma(A)\|_\infty$, with $\sigma(A)$ the vector of singular values of a matrix A. For $A \in \mathbb{R}^{K\times K}$ it holds:

$$\|A\|_{\sigma_\infty} \leq \|A\|_{\mathrm{Fr}} \leq K \sup_{i,j} A_{i,j}$$

so that we get:

$$\sigma_2(\widehat{D}''_n) \geq \sigma_2(\bar{D}'') - K \sup_{i,j} \left|[\widehat{D}''_n]_{i,j} - [\bar{D}'']_{i,j}\right|.$$

Now, define the compact set $\mathcal{U} = [0, U]^K$, with U defined in Equation (C.6). We know from Proposition 1 that with probability at least $1 - M_1 \exp(-c_1 n)$ both $\hat{\boldsymbol{u}}_n$ and $\boldsymbol{u}^*$ belong to $\mathcal{U}$, so that $[\hat{\boldsymbol{u}}_n, \boldsymbol{u}^*] \subset \mathcal{U}$. We can then use Assumption 7 to lower bound $\sigma_2(\bar{D}''(\boldsymbol{v}))$ by $\sigma > 0$ uniformly on $[\hat{\boldsymbol{u}}_n, \boldsymbol{u}^*]$.

Focus now on the term $K \sup_{i,j} \left|[\widehat{D}''_n]_{i,j} - [\bar{D}'']_{i,j}\right|$. From the definition of $\widehat{D}''_n$ and $\bar{D}''$, we can see that their entries (k, k') are the integrals of some function comprised in $[-1, 1]$, according to $\widehat{P}_n$ and $\bar{P}$ respectively. For all $i, j \leq K$, Corollary 2 gives that for all $t > 0$ and $n \geq 2\log(2K)/(\underline{\lambda} t^2)$ it holds with probability at least $1 - 2K \exp\left(-\frac{\underline{\lambda} n t^2}{2}\right)$:

$$\left|[\widehat{D}''_n]_{i,j} - [\bar{D}'']_{i,j}\right| \leq \frac{C_\lambda K}{\sqrt{n}} + t.$$

The union bound then gives that with probability $1 - 2K^3 \exp\left(-\frac{\underline{\lambda} n t^2}{2}\right)$ it holds:

$$K \sup_{i,j} \left|[\widehat{D}''_n]_{i,j} - [\bar{D}'']_{i,j}\right| \leq \frac{C_\lambda K^2}{\sqrt{n}} + Kt.$$

Thus, for $n \geq \max\left(\frac{16 C_\lambda^2 K^4}{\sigma^2}, \frac{96 K^2}{\underline{\lambda}\sigma^2}\log(2K)\right)$, it holds with probability at least $1 - 2K^3 \exp\left(-\frac{\underline{\lambda}\sigma^2}{32K^2} n\right)$:

$$K \sup_{i,j} \left|[\widehat{D}''_n]_{i,j} - [\bar{D}'']_{i,j}\right| \leq \frac{\sigma}{4} + \frac{\sigma}{4} = \frac{\sigma}{2}$$

and consequently

$$\sigma_2\left(\widehat{D}''_n(\boldsymbol{v})\right) \geq \frac{\sigma}{2}.$$

The last step consists in extending this bound uniformly over the line segment $[\hat{\boldsymbol{u}}_n, \boldsymbol{u}^*]$. To do so, we adopt an entropic point of view: we cover the set $\mathcal{U}$ (in which the line segment $[\hat{\boldsymbol{u}}_n, \boldsymbol{u}^*]$ is contained with high probability) with balls, apply the union bound for the centers of these balls, and show that within a ball, the second smallest eigenvalue is relatively stable. By definition, note that $\mathcal{U}$ can be covered with $\mathcal{N}_\epsilon = U^K/(2\epsilon)^K$ $\|\cdot\|_\infty$-balls of radius ϵ. Now, let $(\boldsymbol{u}, \boldsymbol{v}) \in \mathcal{U}^2$

such that $\|\boldsymbol{u}-\boldsymbol{v}\|_\infty \leq \epsilon$. What is the value of $\left|\sigma_2(\widehat{D}_n''(\boldsymbol{u})) - \sigma_2(\widehat{D}_n''(\boldsymbol{v}))\right|$? As noticed in [16], for any $\boldsymbol{a} \in \mathbb{R}^K$ it holds:

$$\boldsymbol{a}^\top \widehat{D}_n''(\boldsymbol{u})\boldsymbol{a} = \int_z \sum_{k=1}^K p_k(z)a_k^2 - \left(\sum_{k=1}^K a_k p_k(z)\right)^2 d\widehat{P}_n(z)$$

with $p_k(z) = e^{u_k}\omega_k(z)/\sum_{l=1}^K e^{u_l}\omega_l(z)$. Define $q_k(z) = e^{v_k}\omega_k(z)/\sum_{l=1}^K e^{v_l}\omega_l(z)$, and assume that $\|\boldsymbol{a}\|_2 = 1$. It holds:

$$\begin{aligned}
&\left|\boldsymbol{a}^\top \widehat{D}_n''(\boldsymbol{u})\boldsymbol{a} - \boldsymbol{a}^\top \widehat{D}_n''(\boldsymbol{v})\boldsymbol{a}\right| \\
&= \left| \int \sum_{k=1}^K p_k(z)a_k^2 - \left(\sum_{k=1}^K a_k p_k(z)\right)^2 - \sum_{k=1}^K q_k(z)a_k^2 + \left(\sum_{k=1}^K a_k q_k(z)\right)^2 d\widehat{P}_n(z)\right| \\
&\leq \int \sum_{k=1}^K \left|p_k(z) - q_k(z)\right| a_k^2 \; d\widehat{P}_n(z) \\
&\quad + \int \left|\sum_{k=1}^K (p_k(z)+q_k(z))a_k\right| \cdot \left|\sum_{k=1}^K (p_k(z)-q_k(z))a_k\right| d\widehat{P}_n(z) \\
&\leq \int \|\boldsymbol{p}(z)-\boldsymbol{q}(z)\|_\infty \; d\widehat{P}_n(z) + \int \|\boldsymbol{p}(z)+\boldsymbol{q}(z)\|_2 \cdot \|\boldsymbol{p}(z)-\boldsymbol{q}(z)\|_2 \; d\widehat{P}_n(z) \\
&\leq 3\sqrt{K} \int \|\boldsymbol{p}(z)-\boldsymbol{q}(z)\|_2 \; d\widehat{P}_n(z).
\end{aligned}$$

Furthermore, notice that $\boldsymbol{p}(z)$ is exactly the integrand in $\widehat{D}_n''(\boldsymbol{u})$, while $\boldsymbol{q}(z)$ is the integrand in $\widehat{D}_n''(\boldsymbol{v})$. Using the same integral calculus as in the beginning of the proof, and bounding the biggest eigenvalue of the matrices by K (as it is an upper bound of the trace), we get that for all z it holds $\|\boldsymbol{p}(z)-\boldsymbol{q}(z)\|_2 \leq K\|\boldsymbol{u}-\boldsymbol{v}\|_2$. Therefore, we get for all $\boldsymbol{a} \in \mathbb{R}^K$ such that $\|\boldsymbol{a}\|_2 = 1$:

$$\left|\boldsymbol{a}^\top \widehat{D}_n''(\boldsymbol{u})\boldsymbol{a} - \boldsymbol{a}^\top \widehat{D}_n''(\boldsymbol{v})\boldsymbol{a}\right| \leq 3K^2\|\boldsymbol{u}-\boldsymbol{v}\|_\infty,$$

and consequently

$$\left|\sigma_2\left(\widehat{D}_n''(\boldsymbol{u})\right) - \sigma_2\left(\widehat{D}_n''(\boldsymbol{v})\right)\right| \leq 3K^2\|\boldsymbol{u}-\boldsymbol{v}\|_\infty.$$

Now, let $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{\mathcal{N}_\epsilon})$ be an ϵ-coverage of $\mathcal{U}$. Applying the union bound, we get that with probability at least $1 - \frac{2K^3U^K}{(2\epsilon)^K}\exp\left(-\frac{\lambda\sigma^2}{32K^2}n\right)$ for any $i \leq \mathcal{N}_\epsilon$ it holds:

$$\sigma_2\left(\widehat{D}_n''(\boldsymbol{u}_i)\right) \geq \frac{\sigma}{2}.$$

Let $\boldsymbol{v} \in [\hat{\boldsymbol{u}}_n, \boldsymbol{u}^*] \subset \mathcal{U}$. By definition, there exists $i \leq \mathcal{N}_\epsilon$ such that $\|\boldsymbol{v}-\boldsymbol{u}_i\|_\infty \leq \epsilon$. Therefore, we get:

$$\sigma_2\left(\widehat{D}_n''(\boldsymbol{v})\right) \geq \sigma_2\left(\widehat{D}_n''(\boldsymbol{u}_i)\right) - 3K^2\epsilon.$$

Taking the infimum, we have with probability $1 - \frac{2K^3U^K}{(2\epsilon)^K}\exp\left(-\frac{\underline{\lambda}\sigma^2}{32K^2}n\right)$

$$\inf_{\boldsymbol{v}\in[\hat{\boldsymbol{u}}_n,\boldsymbol{u}^*]} \sigma_2\left(\widehat{D}_n''(\boldsymbol{v})\right) \geq \frac{\sigma}{2} - 3K^2\epsilon.$$

Choosing $\epsilon = \frac{\sigma}{12K^2}$, we have with probability $1-2K^{2K+3}\left(\frac{6U}{\sigma}\right)^K\exp\left(-\frac{\underline{\lambda}\sigma^2}{32K^2}n\right)$

$$\inf_{\boldsymbol{v}\in[\hat{\boldsymbol{u}}_n,\boldsymbol{u}^*]} \sigma_2\left(\widehat{D}_n''(\boldsymbol{v})\right) \geq \frac{\sigma}{4}.$$

Collecting all arguments, for all $n \geq 16C_\lambda^2K^4/\sigma^2$ it holds with probability $1 - M_1\exp(-c_1n) - 2K^{2K+3}\left(\frac{6U}{\sigma}\right)^K\exp\left(-\frac{\underline{\lambda}\sigma^2}{32K^2}n\right)$:

$$\left\|\widehat{D}_n'(\boldsymbol{u}^*) - \bar{D}'(\boldsymbol{u}^*)\right\|_2 \geq \frac{\sigma}{4\sqrt{K+1}}\,\|\hat{\boldsymbol{u}}_n - \boldsymbol{u}^*\|_2.$$

The proof is finally concluded by setting $M_4 = 2\max\left(M_1; 2K^{2K+3}\left(\frac{6U}{\sigma}\right)^K\right)$, $c_4 = \min\left(c_1; \underline{\lambda}\sigma^2/(32K^2)\right)$, $n_{0,4} = \max\left(16C_\lambda^2K^4/\sigma^2; \log(M_4/c_4)\right)$, and $L = 4\sqrt{K+1}/\sigma$. □

The following key lemma allows to decompose the deviation $\left|\int \hat{h}_n d\widehat{P}_n - \int h d\bar{P}\right|$ into different pieces that are more easily controllable. It is used for instance to bound $\left|\widetilde{L}_n(\theta) - L(\theta)\right|$, see Equation (3.5).

Lemma 2. *Let $\hat{h}_n : \mathcal{Z} \to \mathbb{R}$, $h : \mathcal{Z} \to \mathbb{R}$ be two real-valued functions. We have*

$$\left|\int \hat{h}_n d\widehat{P}_n - \int h d\bar{P}\right| \leq \left\|\hat{h}_n - h\right\|_\infty + \|h\|_\infty \sum_{k=1}^K \left|\hat{\lambda}_k - \lambda_k\right| + \sum_{k=1}^K \hat{\lambda}_k \left|\int h d\widehat{P}_k - \int h dP_k\right|.$$

Proof. It holds

$$\begin{aligned}
&\left|\int \hat{h}_n(z) d\widehat{P}_n(z) - \int h(z) d\bar{P}(z)\right| \\
&\quad\leq \left|\int \hat{h}_n(z) d\widehat{P}_n(z) - \int h(z) d\widehat{P}_n(z)\right| + \left|\int h(z) d\widehat{P}_n(z) - \int h(z) d\bar{P}(z)\right| \\
&\quad\leq \sup_z \left|\hat{h}_n(z) - h(z)\right| + \left|\sum_{k=1}^K \hat{\lambda}_k \int h(z) d\widehat{P}_k(z) - \sum_{k=1}^K \lambda_k \int h(z) dP_k(z)\right| \\
&\quad\leq \sup_z \left|\hat{h}_n(z) - h(z)\right| + \left|\sum_{k=1}^K \hat{\lambda}_k \int h(z) d\widehat{P}_k(z) - \sum_{k=1}^K \hat{\lambda}_k \int h(z) dP_k(z)\right| \\
&\qquad\qquad + \left|\sum_{k=1}^K \hat{\lambda}_k \int h(z) dP_k(z) - \sum_{k=1}^K \lambda_k \int h(z) dP_k(z)\right|
\end{aligned}$$

$$\leq \sup_z \left|\hat{h}_n(z) - h(z)\right| + \sup_z |h(z)| \sum_{k=1}^{K} \left|\hat{\lambda}_k - \lambda_k\right| + \sum_{k=1}^{K} \hat{\lambda}_k \left| \int h(z) d\widehat{P}_k(z) - \int h(z) dP_k(z) \right|$$

□

Corollary 2. *Let $\hat{h}_n : \mathcal{Z} \to \mathbb{R}$ and $h : \mathcal{Z} \to \mathbb{R}$ be two real-valued functions. Assume that there exist $a, b \in \mathbb{R}^2$ such that: $a \leq h(z) \leq b$ for all $z \in \mathcal{Z}$. If Assumption 4 is satisfied, then for all $t > 0$ and $n \geq (b-a)^2 \log(2K)/(2\underline{\lambda}t^2)$, it holds with probability at least $1 - 2K \exp\left(-\frac{2\underline{\lambda}nt^2}{(b-a)^2}\right)$:*

$$\left| \int \hat{h}_n(z) d\widehat{P}_n(z) - \int h(z) d\bar{P}(z) \right| \leq \sup_z \left|\hat{h}_n(z) - h(z)\right| + \frac{C_\lambda K \sup_z |h(z)|}{\sqrt{n}} + t.$$

Proof. Using Lemma 2 and Assumption 4, we have

$$\left| \int \hat{h}_n(z) d\widehat{P}_n(z) - \int h(z) d\bar{P}(z) \right| \leq \sup_z \left|\hat{h}_n(z) - h(z)\right| + \frac{C_\lambda K \sup_z |h(z)|}{\sqrt{n}} + \sum_{k=1}^{K} \hat{\lambda}_k \left| \int h d\widehat{P}_k - \int h dP_k \right|.$$

Now, applying Hoeffding's inequality gives that, for all $t > 0$ and all $k \leq K$,

$$\mathbb{P}\left\{ \left| \int h d\widehat{P}_k - \int h dP_k \right| > t \right\} \leq 2 \exp\left(-\frac{2n_k t^2}{(b-a)^2}\right) \leq 2 \exp\left(-\frac{2\underline{\lambda}nt^2}{(b-a)^2}\right).$$

The proof is concluded by applying the union bound. □

Proposition 5. *Suppose that Assumption 4 is verified. Then, for all $t > 0$ and $n \geq \log(2K^2)/(2\underline{\lambda}t^2)$, it holds with probability at least $1 - 2K^2 \exp\left(-2\underline{\lambda}nt^2\right)$:*

$$\left\| \widehat{D}'_n(\boldsymbol{u}^*) - \bar{D}'(\boldsymbol{u}^*) \right\|_2 \leq \frac{2C_\lambda K^{3/2}}{\sqrt{n}} + \sqrt{K}t.$$

Proof. Apply Corollary 2 for every component k of $\widehat{D}'_n(\boldsymbol{u}^*) - \bar{D}'(\boldsymbol{u}^*)$ with $\hat{h}_n = e^{u_k^*}\omega_k/(\sum_l e^{u_l^*}\omega_l) - \hat{\lambda}_k$ and $h = e^{u_k^*}\omega_k/(\sum_l e^{u_l^*}\omega_l) - \lambda_k$, and the union bound permits to conclude. □

Proof of Proposition 2. Combining Lemma 1, Propositions 4 and 5, we have that it holds with probability at least $1 - M_4 \exp(-c_4 n) - 2K^2 \exp(-2\underline{\lambda}nt^2)$:

$$\left\| \widehat{\boldsymbol{W}}_n - \boldsymbol{W}^* \right\|_2 \leq \|\hat{\boldsymbol{u}}_n - \boldsymbol{u}^*\|_2 + C_\lambda \sqrt{\frac{K}{n}} \leq L \left\| \widehat{D}'_n(\boldsymbol{u}^*) - \bar{D}'(\boldsymbol{u}^*) \right\|_2 + C_\lambda \sqrt{\frac{K}{n}}$$

$$\leq L\left(\frac{2C_\lambda K^{3/2}}{\sqrt{n}}+\sqrt{K}t\right)+C_\lambda\sqrt{\frac{K}{n}}$$
$$=L\sqrt{K}t+\frac{C_\lambda\sqrt{K}(2LK+1)}{\sqrt{n}}.$$

The proof is concluded by setting $\gamma_2 = C_\lambda\sqrt{K}(2LK+1)$, $M_2 = M_4$, $c_2 = c_4$, $M_2' = 2K^2$, $c_2' = 2\underline{\lambda}/(L^2K)$, and $n_{0,2} = n_{0,4}$. □

C.3. Proof of Proposition 3

Recall that by Proposition 1, it holds with probability $1 - M_1\exp(-c_1 n)$:

$$\forall k \leq K, \qquad \rho \leq \widehat{W}_{n,k} \leq 1, \quad \text{and} \quad \rho \leq W_k^* \leq 1.$$

This implies for all $z \in \mathcal{Z}$:

$$\rho \leq \left(\sum_{l=1}^K \frac{\hat{\lambda}_l\omega_l(z)}{\widehat{W}_{n,l}}\right)^{-1} \leq \frac{1}{\varepsilon\underline{\lambda}}, \quad \text{and} \quad \rho \leq \left(\sum_{l=1}^K \frac{\lambda_l\omega_l(z)}{W_l^*}\right)^{-1} \leq \frac{1}{\varepsilon\underline{\lambda}}.$$

Using the above inequalities and the mean value theorem on $t \mapsto 1/t$, we get for all $k \leq K$:

$$\left|\widehat{\Omega}_{n,k} - \Omega_k\right| = \left|\frac{\widehat{W}_{n,k}}{\int\left(\sum_{l=1}^K \frac{\hat{\lambda}_l\omega_l}{\widehat{W}_{n,l}}\right)^{-1} d\widehat{P}_n} - \frac{W_k^*}{\int\left(\sum_{l=1}^K \frac{\lambda_l\omega_l}{W_l^*}\right)^{-1} d\bar{P}}\right|$$

$$\leq \frac{1}{\int\left(\sum_{l=1}^K \frac{\hat{\lambda}_l\omega_l}{\widehat{W}_{n,l}}\right)^{-1} d\widehat{P}_n}\left|\widehat{W}_{n,k} - W_k^*\right|$$

$$+ W_k^*\left|\frac{1}{\int\left(\sum_{l=1}^K \frac{\hat{\lambda}_l\omega_l}{\widehat{W}_{n,l}}\right)^{-1} d\widehat{P}_n} - \frac{1}{\int\left(\sum_{l=1}^K \frac{\lambda_l\omega_l}{W_l^*}\right)^{-1} d\bar{P}}\right|$$

$$\leq \frac{1}{\rho}\left|\widehat{W}_{n,k} - W_k^*\right| + \frac{1}{\rho^2}\left|\int\left(\sum_{l=1}^K \frac{\hat{\lambda}_l\omega_l}{\widehat{W}_{n,l}}\right)^{-1} d\widehat{P}_n - \int\left(\sum_{l=1}^K \frac{\lambda_l\omega_l}{W_l^*}\right)^{-1} d\bar{P}\right|. \tag{C.9}$$

The first term in Equation (C.9) can be bounded using Proposition 2. For the second, we can use Corollary 2. First we must compute:

$$\left|\left(\sum_{l=1}^K \frac{\hat{\lambda}_l\omega_l(z)}{\widehat{W}_{n,l}}\right)^{-1} - \left(\sum_{l=1}^K \frac{\lambda_l\omega_l(z)}{W_l^*}\right)^{-1}\right| \leq \left(\frac{1}{\varepsilon\underline{\lambda}}\right)^2\left|\sum_{l=1}^K \frac{\hat{\lambda}_l\omega_l(z)}{\widehat{W}_{n,l}} - \sum_{l=1}^K \frac{\lambda_l\omega_l(z)}{W_l^*}\right|$$

$$\leq \left(\frac{1}{\varepsilon\underline{\lambda}}\right)^2 \sum_{l=1}^{K} \frac{\left|\hat{\lambda}_l - \lambda_l\right|}{\widehat{W}_{n,l}} + \lambda_l \left|\frac{1}{\widehat{W}_{n,l}} - \frac{1}{W_l^*}\right|$$
$$\leq \left(\frac{1}{\varepsilon\underline{\lambda}}\right)^2 \left(\frac{C_\lambda K}{\rho\sqrt{n}} + \frac{1}{\rho^2}\sum_{l=1}^{K} \lambda_l \left|\widehat{W}_{n,l} - W_l^*\right|\right).$$

Proposition 2 then allows to bound the last term with overwhelming probability. Next, applying Corollary 2 with $\hat{h}_n = \left(\sum_l \frac{\hat{\lambda}_l \omega_l}{\widehat{W}_{n,l}}\right)^{-1}$ and $h = \left(\sum_l \frac{\lambda_l \omega_l}{W_l^*}\right)^{-1}$, we obtain that for all $t_1, t_2 > 0$ with probability at least $1 - M_2\exp(-c_2 n) - M_2'\exp(-c_2' n t_1^2) - 2K\exp(-2\varepsilon^2\underline{\lambda}^3 n t_2^2)$ it holds for all $k \leq K$:

$$\left|\widehat{\Omega}_{n,k} - \Omega_k\right|$$
$$\leq \frac{1}{\rho}\left(t_1 + \frac{\gamma_2}{\sqrt{n}}\right) + \frac{1}{\rho^2}\left(\frac{C_\lambda K}{\varepsilon^2\underline{\lambda}^2\rho\sqrt{n}} + \frac{1}{\varepsilon^2\underline{\lambda}^2\rho^2}\left(t_1 + \frac{\gamma_2}{\sqrt{n}}\right) + \frac{C_\lambda K}{\varepsilon\underline{\lambda}\sqrt{n}} + t_2\right)$$
$$= t_1\left(\frac{1}{\rho} + \frac{1}{\varepsilon^2\underline{\lambda}^2\rho^4}\right) + \frac{t_2}{\rho^2} + \left(\frac{\gamma_2}{\rho} + \frac{C_\lambda K}{\varepsilon^2\underline{\lambda}^2\rho^3} + \frac{\gamma_2}{\varepsilon^2\underline{\lambda}^2\rho^4} + \frac{C_\lambda K}{\varepsilon\underline{\lambda}\rho^2}\right)\frac{1}{\sqrt{n}}.$$

The proof is concluded by setting $M_3 = M_2$, $c_3 = c_2$, $M_3' = 2\max\left(M_2'; 2K\right)$,

$$c_3' = \max\left(\frac{c_2'}{4\left(\frac{1}{\rho} + \frac{1}{\varepsilon^2\underline{\lambda}^2\rho^4}\right)^2}; \frac{\varepsilon^2\underline{\lambda}^3\rho^4}{2}\right),$$
$$\gamma_3 = \frac{\gamma_2}{\rho} + \frac{C_\lambda K}{\varepsilon^2\underline{\lambda}^2\rho^3} + \frac{\gamma_2}{\varepsilon^2\underline{\lambda}^2\rho^4} + \frac{C_\lambda K}{\varepsilon\underline{\lambda}\rho^2},$$

and $n_{0,3} = n_{0,2}$. □

C.4. Proof of Theorem 1

Let $\theta \in \Theta$. The first step of the proof consists in using Lemma 2 with the choices $\hat{h}_{n,\theta}(z) = \psi(z,\theta)\left(\sum_{k=1}^K \frac{\hat{\lambda}_k\omega_k(z)}{\widehat{\Omega}_{n,k}}\right)^{-1}$, and $h_\theta(z) = \psi(z,\theta)\left(\sum_{k=1}^K \frac{\lambda_k\omega_k(z)}{\Omega_k}\right)^{-1}$. We obtain

$$\left|\widetilde{L}_n(\theta) - L(\theta)\right|$$
$$= \left|\int \psi(z,\theta)\left(\sum_{k=1}^K \frac{\hat{\lambda}_k\omega_k(z)}{\widehat{\Omega}_{n,k}}\right)^{-1} d\widehat{P}_n(z) - \int \psi(z,\theta)\left(\sum_{k=1}^K \frac{\lambda_k\omega_k(z)}{\Omega_k}\right)^{-1} d\bar{P}(z)\right|$$
$$= \left|\int \hat{h}_n(z,\theta)d\widehat{P}_n(z) - \int h(z,\theta)d\bar{P}(z)\right|$$

$$\begin{aligned}
&\leq \sup_z \left|\hat{h}_{n,\theta}(z) - h_\theta(z)\right| + \sup_z |h_\theta(z)| \sum_{k=1}^K |\hat{\lambda}_k - \lambda_k| + \sum_{k=1}^K \hat{\lambda}_k \left|\int h_\theta \, d(\widehat{P}_k - P)\right| \\
&\leq \sup_z \left|\hat{h}_{n,\theta}(z) - h_\theta(z)\right| + \frac{C_\lambda K \sup_z |h_\theta(z)|}{\sqrt{n}} + \max_{k \leq K} \left|\int h_\theta \, d(\widehat{P}_k - P)\right|.
\end{aligned} \tag{C.10}$$

Now, by the definitions of $\widehat{\mathbf{\Omega}}_n$ and $\mathbf{\Omega}$, for all $k \leq K$, we have

$$\varepsilon\underline{\lambda}\rho \leq \widehat{\Omega}_{n,k} \leq \frac{1}{\rho}, \qquad \text{and} \qquad \varepsilon\underline{\lambda}\rho \leq \Omega_k \leq 1.$$

Hence, for all $z \in \mathcal{Z}$ it holds

$$0 \leq \left(\sum_{k=1}^K \frac{\hat{\lambda}_k \omega_k(z)}{\widehat{\Omega}_{n,k}}\right)^{-1} \leq \frac{1}{\varepsilon\underline{\lambda}\rho}, \quad \text{and} \quad 0 \leq \left(\sum_{k=1}^K \frac{\lambda_k \omega_k(z)}{\Omega_k}\right)^{-1} \leq \frac{1}{\varepsilon\underline{\lambda}}.$$

As $|\psi(z,\theta)| \leq 1$, this implies that $\sup_{z,\theta} |h_\theta(z)| \leq 1/(\varepsilon\underline{\lambda})$. And we also have

$$\begin{aligned}
\left|\hat{h}_{n,\theta}(z) - h_\theta(z)\right| &= \left|\psi(z,\theta)\left(\sum_{k=1}^K \frac{\hat{\lambda}_k \omega_k(z)}{\widehat{\Omega}_{n,k}}\right)^{-1} - \psi(z,\theta)\left(\sum_{k=1}^K \frac{\lambda_k \omega_k(z)}{\Omega_k}\right)^{-1}\right| \\
&\leq \left(\frac{1}{\varepsilon\underline{\lambda}\rho}\right)^2 \left|\sum_{k=1}^K \frac{\hat{\lambda}_k \omega_k(z)}{\widehat{\Omega}_{n,k}} - \sum_{k=1}^K \frac{\lambda_k \omega_k(z)}{\Omega_k}\right|, \\
&\leq \left(\frac{1}{\varepsilon\underline{\lambda}\rho}\right)^3 \frac{C_\lambda K}{\sqrt{n}} + \left(\frac{1}{\varepsilon\underline{\lambda}\rho}\right)^4 \sum_{k=1}^K \lambda_k \left|\widehat{\Omega}_{n,k} - \Omega_k\right|.
\end{aligned}$$

Plugging into Equation (C.10), and taking the supremum over $\theta \in \Theta$, we obtain

$$\begin{aligned}
\sup_{\theta\in\Theta} \left|\widetilde{L}_n(\theta) - L(\theta)\right| \leq{}& \left(\frac{1}{\varepsilon\underline{\lambda}\rho}\right)^3 \frac{C_\lambda K}{\sqrt{n}} + \left(\frac{1}{\varepsilon\underline{\lambda}\rho}\right)^4 \sum_{k=1}^K \lambda_k \left|\widehat{\Omega}_{n,k} - \Omega_k\right| \\
&+ \frac{C_\lambda K}{\varepsilon\underline{\lambda}\sqrt{n}} + \max_k \sup_{\theta\in\Theta} \left|\int h_\theta(z) d(\widehat{P}_k - P)(z)\right|
\end{aligned} \tag{C.11}$$

Thus, we have bounded $\sup_{\theta\in\Theta} |\widetilde{L}_n(\theta) - L(\theta)|$ by a sum involving: (1) non-random terms scaling as $\mathcal{O}(1/\sqrt{n})$, (2) random terms independent from θ which can be controlled using Proposition 3, and (3) the supremum of an empirical process. We can now use standard arguments such as chaining [42, 43] to bound this last term. Let $k \leq K$, and $t > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\theta\in\Theta} \left|\int h_\theta(z) d(\widehat{P}_k - P)(z)\right| > t\right\} \\
&\qquad = \mathbb{P}\left\{\sup_{\theta\in\Theta} \left|\int \varepsilon\underline{\lambda} h(z,\theta) d(\hat{P}_k - P)(z)\right| > \varepsilon\underline{\lambda} t\right\}.
\end{aligned} \tag{C.12}$$

By applying Theorem 2.14.9 in [43] to the class

$$\mathcal{G}_\Theta = \Big\{\varepsilon\underline{\lambda} h(\cdot,\theta)\colon \theta\in\Theta\Big\} = \left\{\varepsilon\underline{\lambda}\left(\sum_{k=1}^K \frac{\lambda_k\omega_k(\cdot)}{\Omega_k}\right)^{-1}\psi(\cdot,\theta)\colon \theta\in\Theta\right\}$$

which also satisfies Assumption 8, since it is a pointwise multiplication of $\mathcal{F}_\Theta = \big\{\psi(\cdot,\theta)\colon \theta\in\Theta\big\}$ by a function with values in $[0,1]$, we obtain that (C.12) is upper bounded by

$$\left(\frac{\Delta\varepsilon\underline{\lambda}\sqrt{n_k}t}{\sqrt{r}}\right)^r e^{-2(\varepsilon\underline{\lambda})^2 n_k t^2} \leq \left(\frac{\Delta\varepsilon\underline{\lambda}}{\sqrt{r}}\right)^r n^{\frac{r}{2}} t^r e^{-2\varepsilon^2\underline{\lambda}^3 n t^2} \tag{C.13}$$

where Δ is a constant that depends only on C_Θ. Finally, plugging (C.13) with the union bound and Proposition 3 into (C.11), we get that with probability at least $1 - M_3 e^{-c_3 n} - M_3' e^{-c_3' n t_1^2} - K\left(\frac{\Delta\varepsilon\underline{\lambda}}{\sqrt{r}}\right)^r n^{\frac{r}{2}} t_2^r\, e^{-2\varepsilon^2\underline{\lambda}^3 n t_2^2}$

$$\sup_{\theta\in\Theta}\left|\widetilde{L}_n(\theta) - L(\theta)\right| \leq \left(\frac{1}{\varepsilon\underline{\lambda}\rho}\right)^3 \frac{C_\lambda K}{\sqrt{n}} + \left(\frac{1}{\varepsilon\underline{\lambda}\rho}\right)^4\left(\frac{\gamma_3}{\sqrt{n}} + t_1\right) + \frac{C_\lambda K}{\varepsilon\underline{\lambda}\sqrt{n}} + t_2\,,$$

or again, we have with probability at least $1 - M_3 e^{-c_3 n} - M_3' \exp\left(-\frac{c_3'(\varepsilon\underline{\lambda}\rho)^8 n t^2}{4}\right) - K\left(\frac{\Delta\varepsilon\underline{\lambda}}{2\sqrt{r}}\right)^r n^{\frac{r}{2}} t^r \exp\left(-\frac{\varepsilon^2\underline{\lambda}^3 n t^2}{2}\right)$

$$\sup_{\theta\in\Theta}\left|\widetilde{L}_n(\theta) - L(\theta)\right| \leq \frac{\gamma}{\sqrt{n}} + t\,,$$

with

$$\gamma = \frac{C_\lambda K}{(\varepsilon\underline{\lambda}\rho)^3} + \frac{\gamma_3}{(\varepsilon\underline{\lambda}\rho)^4} + \frac{C_\lambda K}{\varepsilon\underline{\lambda}}\,.$$

The proof is concluded by setting $M = M_3$, $c = c_3$, $M' = M_3'$, $M'' = K\left(\frac{\Delta\varepsilon\underline{\lambda}}{2\sqrt{r}}\right)^r$, $c' = c_3'(\varepsilon\underline{\lambda}\rho)^8/4$, $c'' = (\varepsilon^2\underline{\lambda}^3)/2$, and $n_0 = n_{0,3}$. □

C.5. Proof of Theorem 2

Let $\mathcal{F} = \big\{z \mapsto \mathbb{I}\{z\leq\tau\}\colon \tau\in\mathbb{R}\big\}$. As discussed in the main body of the paper, applying Theorem 1 to $\mathcal{F}$ in a straightforward fashion yields a bound that does not match the standard DKW inequality. To match the rate of the DKW inequality, we have to develop a refined analysis, specifically tailored to classes which are composed of indicator functions. We introduce the following complexity assumption [43, Chapter 14], that strengthens Assumption 8.

Assumption 9. *The class $\mathcal{F} = \mathcal{F}_\mathcal{C}$ is composed of indicator functions of sets, i.e., $\mathcal{F}_\mathcal{C} = \{z\mapsto\mathbb{I}\{z\in C\}\colon C\in\mathcal{C}\}$, with $\mathcal{C}$ a collection of sets of satisfying for some constants $C_\mathcal{C} > 0$ and $r \geq 1$*

$$\sup_Q \mathcal{N}(\zeta,\ \mathcal{C},\ L_1(Q)) \leq (C_\mathcal{C}/\zeta)^r.$$

Furthermore, for $k \leq K$ and $\delta > 0$, let $\mathcal{C}_{k,\delta} = \{C \in \mathcal{C} : |P_k(C) - 1/2| < \delta\}$. We also assume the existence of $C'_{\mathcal{C}}, r', r''$ such that for every $\delta \geq \zeta > 0$

$$\sup_{k \leq K} \mathcal{N}(\zeta,\ \mathcal{C}_{k,\delta},\ L_1(P_k)) \leq C'_{\mathcal{C}}\ \delta^{r'}\ \zeta^{-r''}.$$

Note that a class $\mathcal{F}_{\mathcal{C}}$ of finite VC dimension $V < +\infty$ verifies the first inequality with $r = V - 1$, and $C_{\mathcal{C}}$ that depends only on V, see e.g., Theorem 2.6.4 in [43]. Under Assumption 9, a tighter control of the empirical processes in decomposition (3.5) is possible, yielding the following theorem.

Theorem 3. *Suppose that Assumptions 4, 5, 6, 7, and 9 are satisfied. Then, there exist $M, M', M'', c, c', c'', \gamma, n_0$, depending only on $K, C_\lambda, \underline{\lambda}, \kappa, \varepsilon, \sigma, C_{\mathcal{C}}, C'_{\mathcal{C}}$, r, r', and r'' such that for all $t > 0$ and $n \geq n_0$ it holds:*

$$\mathbb{P}\left\{\sup_{\theta \in \Theta} \left|\widetilde{L}_n(\theta) - L(\theta)\right| > \frac{\gamma}{\sqrt{n}} + t\right\} \leq Me^{-cn} + M'e^{-c'nt^2} + (nt^2)^{r''-r'} M''e^{-c''nt^2}.$$

Proof. The proof follows the same path as that of Theorem 1. In particular, we start from the same decomposition (C.11), but Assumption 9 now allows a better control on the empirical processes that compose the last term. Specifically, for every $k \leq K$, Theorem 2.14.14 in [43] gives that

$$\begin{aligned}\mathbb{P}\left\{\sup_{C \in \mathcal{C}} \left|\int h_C(z) d(\widehat{P}_k - P)(z)\right| > t\right\} &\leq \Delta(\varepsilon\underline{\lambda}\sqrt{n_k}t)^{2r''-2r'} e^{-2\varepsilon^2\underline{\lambda}^2 n_k t^2} \\ &\leq \Delta(\varepsilon\underline{\lambda}\sqrt{n}t)^{2r''-2r'} e^{-2\varepsilon^2\underline{\lambda}^3 nt^2}\end{aligned}$$

where Δ is a constant that depends only on $C_{\mathcal{C}}, C'_{\mathcal{C}}, r, r'$, and r''. Plugging into Equation (C.11), we get that $\sup_{C \in \mathcal{C}} \left|\widetilde{L}_n(C) - L(C)\right| \leq \frac{\gamma}{\sqrt{n}} + t$ with probability at least $1 - M_3 e^{-c_3 n} - M'_3 \exp\left(-\frac{c'_3(\varepsilon\underline{\lambda}\rho)^8 nt^2}{4}\right) - K\Delta(\varepsilon\underline{\lambda}\sqrt{n}t)^{2r''-2r'}\ e^{-\frac{\varepsilon^2\underline{\lambda}^3 nt^2}{2}}$. We conclude by setting $M = M_3$, $c = c_3$, $M' = M'_3$, $c' = c'_3(\varepsilon\underline{\lambda}\rho)^8/4$, $M'' = K\Delta(\varepsilon\underline{\lambda})^{2r''-2r'}$, $c'' = (\varepsilon^2\underline{\lambda}^3)/2$, and $n_0 = n_{0,3}$. □

Theorem 2 is actually a corollary of Theorem 3, applied to the class $\mathcal{F} = \{z \mapsto \mathbb{I}\{z \leq \tau\} : \tau \in \mathbb{R}\}$.

Proof of Theorem 2. The class $\mathcal{F} = \{z \mapsto \mathbb{I}\{z \leq \tau\} : \tau \in \mathbb{R}\}$ satisfies Assumption 9 with $r = r' = r'' = 1$, see [43, page 247]. Applying Theorem 3, we obtain

$$\begin{aligned}\mathbb{P}\left\{\sup_{z \in \mathbb{R}} \left|(\widetilde{P}_n - P)((-\infty,\ z])\right| > \frac{\gamma}{\sqrt{n}} + t\right\} &\\ \leq Me^{-cn} + M'e^{-c'nt^2} + M''e^{-c''nt^2} &\\ \leq Me^{-cn} + 2\max(M'; M'')e^{-\min(c';c'')nt^2}.&\end{aligned}$$

□

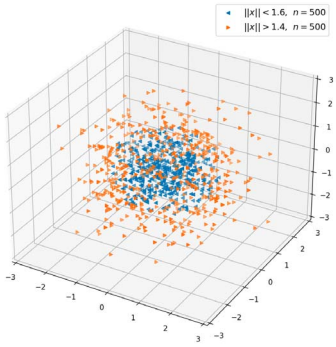
Appendix D: Additional experiments

In this section we provide additional experimental results, both on a synthetic estimation problem (Appendix D.1) and real data learning applications, see Appendix D.2.

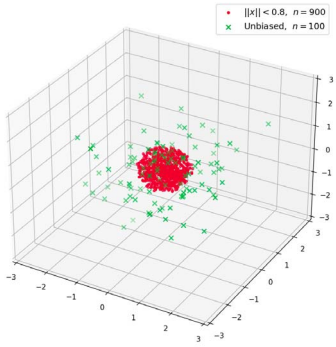
D.1. Estimation experiments

Recall that the synthetic data here consist of 1000 train and 300 test realizations of a 3-dimensional Gaussian random vector. The goal is to predict the norm of the realizations through different learning algorithms: Linear Regression (LR), Kernel Ridge Regression (KRR), Support Vector Regression (SVR) and Random Forest (RF). They are implemented with default hyperparameters, as focus is not on the performances *per se*, but rather on the impact of the debiasing for a given model. The biasing functions ω_k used here are indicator functions of subspaces of $\mathbb{R}^3$. These functions (or equivalently the subsets) are chosen according to twelve different scenarios, so as to contrast the debiasing effects. When one biasing function is the identity (i.e., one subspace is $\mathbb{R}^3$), the algorithm is also trained on the sole unbiased sample. However, this approach does not benefit from the whole dataset, and performances reported compare unfavorably to debiased ERM. Numerical results are gathered in Tables 2 and 3. For scenarios in which no subspace is $\mathbb{R}^3$, two lines are displayed: the upper one corresponds to the standard ERM (ERM), while the second one is achieved through the debiased approach we promote (db-ERM). When one subspace is $\mathbb{R}^3$, a third line is displayed, which corresponds to the result obtained with training on the sole unbiased sample (ub-ERM).

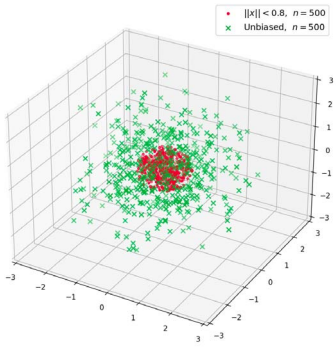
We now thoroughly describe the first six scenarios, that depict situations where selection bias applies directly to the norm of the realizations, and whose visualizations are available in Figure 5. To understand scenario a), one must have in mind that 1.5 is approximately the median value of $\|x\|$ when $x \sim \mathcal{N}(\mathbf{0}_3, \mathbf{I}_3)$ (see the $\chi^2(3)$ law). Hence, partitioning the whole space using $\mathbb{I}\{\|x\| \leq 1.6\}$ and $\mathbb{I}\{\|x\| \geq 1.4\}$ (the two subspaces must intersect) divides $\mathbb{R}^3$ into parts of roughly equal importance. Considering two samples of equal size, each associated to one of these biasing functions, should therefore be almost equivalent to considering blindly the concatenated sample. Consequently, debiasing ERM in this scenario should not lead to any particular improvement, what is verified empirically. As no subset is the full space, no third line is provided. On the contrary, if the samples were of different sizes, one should expect an improvement when using debiasing ERM. In order to emphasize this effect, scenario b) considers even strongly concentrated points around 0, with $\mathbb{I}\{\|x\| \leq 0.8\}$. A sample of size 900 is drawn from this part of the space, which usually represents 10% of the distribution, while a 100 long unbiased sample completes the scenario. As expected, the debiasing ERM appears to be less fooled by the outnumbered examples with small norm, and induces a significant improvement compared to the naive ERM. Furthermore, ERM based the sole unbiased sample is also



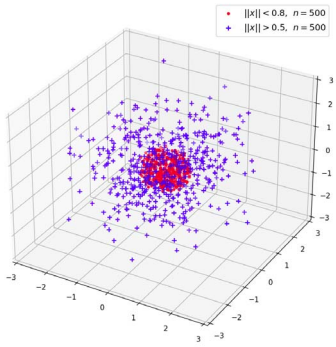
(a) Scenario a)



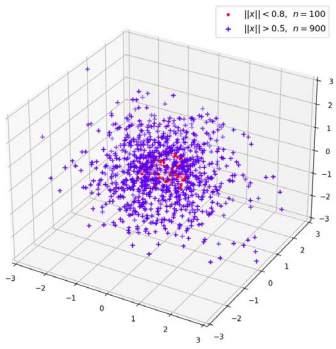
(b) Scenario b)



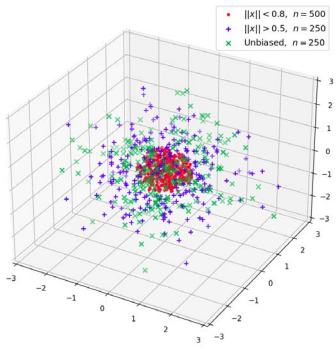
(c) Scenario c)



(d) Scenario d)



(e) Scenario e)



(f) Scenario f)

Fig 5: Different scenarios when selection bias applies to the vector's norm

TABLE 2
Mean Squared Errors by 4 Algorithms on the 6 Norm Biased Scenarios.

		LR	KRR	SVR	RF
a)	ERM	**4.6e-1 ± 4.0e-2**	6.8e-2 ± 2.9e-2	6.6e-3 ± 2.7e-3	**3.4e-2 ± 6.7e-3**
	db-ERM	4.6e-1 ± 4.0e-2	**6.3e-2 ± 2.8e-2**	**6.5e-3 ± 2.6e-3**	3.4e-2 ± 6.6e-3
b)	ERM	1.3e+0 ± 9.8e-2	3.2e-1 ± 7.5e-2	3.8e-2 ± 1.2e-2	1.5e-1 ± 3.2e-2
	db-ERM	**4.8e-1 ± 4.8e-2**	**1.8e-1 ± 5.6e-2**	4.4e-2 ± 1.3e-2	**1.2e-1 ± 2.8e-2**
	ub-ERM	4.8e-1 ± 4.9e-2	3.4e-1 ± 7.8e-2	**3.0e-2 ± 9.7e-3**	1.3e-1 ± 2.8e-2
c)	ERM	7.2e-1 ± 6.6e-2	1.1e-1 ± 3.7e-2	**1.0e-2 ± 4.0e-3**	5.2e-2 ± 1.1e-2
	db-ERM	**4.6e-1 ± 3.8e-2**	**7.7e-2 ± 3.1e-2**	1.0e-2 ± 3.7e-3	**4.5e-2 ± 9.0e-3**
	ub-ERM	**4.6e-1 ± 3.8e-2**	1.0e-1 ± 3.7e-2	1.1e-2 ± 4.1e-3	4.6e-2 ± 8.9e-3
d)	ERM	7.0e-1 ± 6.6e-2	1.0e-1 ± 3.6e-2	**9.8e-3 ± 3.8e-3**	5.1e-2 ± 1.0e-2
	db-ERM	**4.6e-1 ± 3.8e-2**	**7.5e-2 ± 3.1e-2**	9.9e-3 ± 3.6e-3	**4.4e-2 ± 8.5e-3**
e)	ERM	4.6e-1 ± 4.0e-2	6.2e-2 ± 2.7e-2	6.2e-3 ± 2.5e-3	3.4e-2 ± 6.7e-3
	db-ERM	**4.6e-1 ± 3.8e-2**	**6.0e-2 ± 2.7e-2**	**6.2e-3 ± 2.4e-3**	**3.3e-2 ± 6.3e-3**
f)	ERM	7.1e-1 ± 6.8e-2	1.0e-1 ± 3.6e-2	**9.7e-3 ± 3.6e-3**	5.1e-2 ± 1.1e-2
	db-ERM	**4.6e-1 ± 3.9e-2**	**7.4e-2 ± 3.0e-2**	9.9e-3 ± 3.4e-3	**4.4e-2 ± 8.8e-3**
	ub-ERM	4.7e-1 ± 4.1e-2	1.7e-1 ± 5.1e-2	1.7e-2 ± 5.8e-3	6.9e-2 ± 1.5e-2

globally outperformed. Scenario c) is similar to scenario b), with less imbalanced samples. Debiasing ERM remains the most successful approach, but by expected lower margins. What happens if one attempts to fight the selection bias towards $\mathbf{0}_3$ by considering a second sample biased towards great norms, rather than an unbiased one? It is the purpose of scenarios d) and e) to investigate this option, using $\mathbb{I}\{\|x\| \geq 0.5\}$ as a second biasing function. Almost no change can be acknowledged when the sample sizes are the same as in scenario c) (see scenario d)). However, the advantage of debiasing ERM decreases with the proportion of small norm points, as illustrated by scenario e). Finally, scenario f) illustrates that the number of samples is of low importance. If the sample biased towards small norms is large enough, debiasing ERM outperforms all other methods, even if two additional samples are considered, one biased towards large norms, and one unbiased. All numerical results can be found in Table 2 and attest that: 1) ignoring selection bias may have dramatic consequences 2) discarding some data and learning only on the unbiased sample – when it exists – is not a viable solution either, thus endorsing the debiased approach we promote.

One may however argue that results presented in Table 2 overestimate the debiasing effect, as bias occur precisely on the problem's target. We now present similar results obtained when selection bias applies on on one component of the Gaussian only, and not on the norm itself. Again, six different scenarios have been investigated, and depicted in Figure 6, while complete numerical results are gathered in Table 3. Scenarios g) and h) are analogous to scenarios b) and c), except that only one component, x_0, is now biased towards small values using $\mathbb{I}\{|x_0| < 0.1\}$. The improvements induced by debiasing ERM remains substantial, and decrease expectedly as the unbiased sample becomes larger

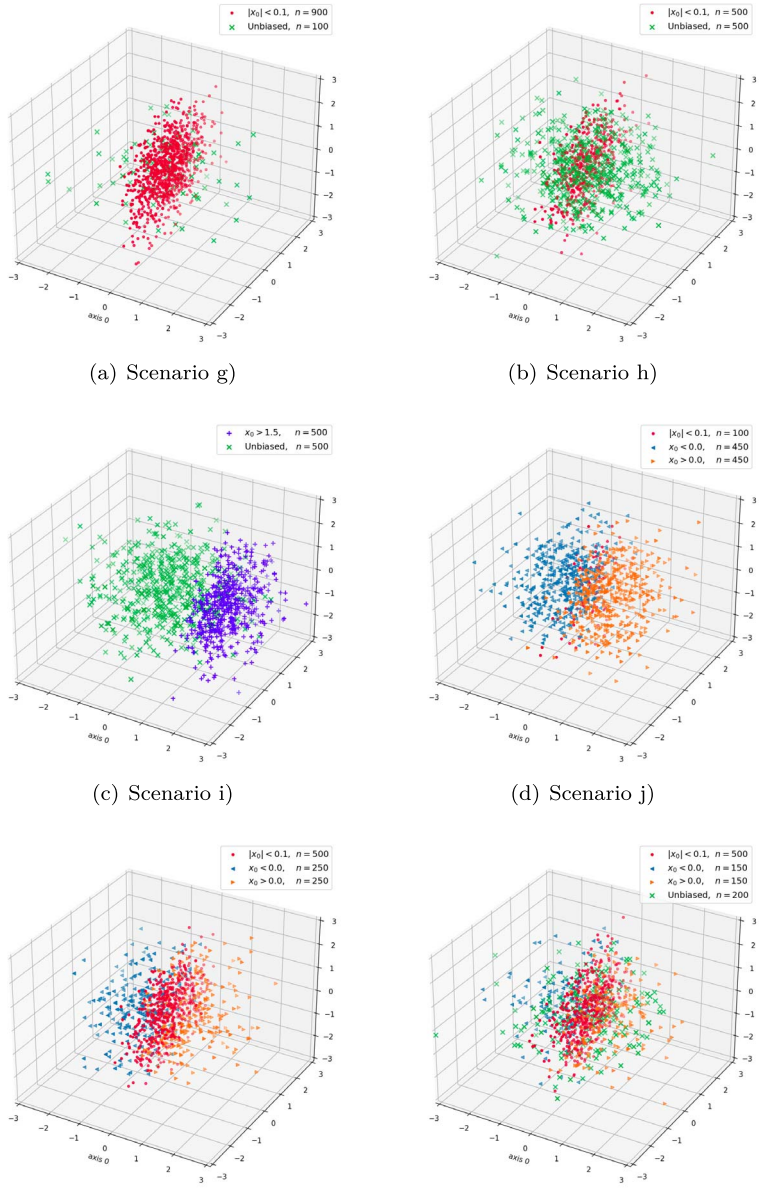
(scenario h)). Scenario i) illustrates that debiasing ERM may improve the results even if a bias applies on large values, using $\mathbb{I}\{x_0 > 1.5\}$ for instance. However, this bias does not distort the predictions towards small norm values, inducing smaller squared norm errors, hence the smaller benefit of debiasing. Scenario j) is analogous to scenario a), but with 3 samples, and leads to similar conclusions: when the blind concatenated sample is very similar to an unbiased sample (the interval $|x_0| < 0.1$ indeed represents 10% of the distribution), debiased ERM is of lower interest. But when the proportions are not respected anymore, as in scenario k), it significantly increases the performances. Finally, scenario l) involves 4 samples, with similar conclusions as above. Again, and although bias does not apply on the target itself, but rather on one simple covariate, the debiasing approach naturally yields improvements, both upon the standard and the unbiased methods.

TABLE 3
Mean Squared Errors by 4 *Algorithms on the* 6 *First Component Biased Scenarios*

		LR	KRR	SVR	RF
Sc. g)	ERM	5.6e-1 ± 5.7e-2	2.0e-1 ± 5.8e-2	**1.5e-2 ± 5.4e-3**	1.4e-1 ± 3.2e-2
	db-ERM	**4.8e-1 ± 4.5e-2**	**1.6e-1 ± 5.3e-2**	3.8e-2 ± 1.3e-2	**8.6e-2 ± 2.1e-2**
	ub-ERM	4.8e-1 ± 4.6e-2	3.4e-1 ± 8.1e-2	3.0e-2 ± 1.0e-2	1.3e-1 ± 3.0e-2
Sc. h)	ERM	4.9e-1 ± 4.7e-2	8.7e-2 ± 3.4e-2	**8.3e-3 ± 3.2e-3**	4.4e-2 ± 9.1e-3
	db-ERM	**4.6e-1 ± 4.0e-2**	**7.6e-2 ± 3.1e-2**	1.0e-2 ± 3.5e-3	**4.1e-2 ± 8.1e-3**
	ub-ERM	4.6e-1 ± 4.0e-2	1.0e-1 ± 3.7e-2	1.1e-2 ± 3.9e-3	4.6e-2 ± 9.3e-3
Sc. i)	ERM	5.5e-1 ± 4.8e-2	6.7e-2 ± 2.9e-2	**6.7e-3 ± 2.3e-3**	3.9e-2 ± 7.9e-3
	db-ERM	**4.6e-1 ± 3.8e-2**	**6.7e-2 ± 2.9e-2**	8.7e-3 ± 3.0e-3	**3.8e-2 ± 7.8e-3**
	ub-ERM	4.6e-1 ± 3.9e-2	1.0e-1 ± 3.7e-2	1.1e-2 ± 3.9e-3	4.6e-2 ± 9.0e-3
Sc. j)	ERM	**4.6e-1 ± 4.0e-2**	6.4e-2 ± 2.9e-2	**6.4e-3 ± 2.6e-3**	3.3e-2 ± 6.9e-3
	db-ERM	4.6e-1 ± 4.0e-2	**6.3e-2 ± 2.9e-2**	6.5e-3 ± 2.6e-3	**3.3e-2 ± 6.8e-3**
Sc. k)	ERM	4.9e-1 ± 4.6e-2	8.7e-2 ± 3.5e-2	**8.3e-3 ± 3.4e-3**	4.4e-2 ± 9.2e-3
	db-ERM	**4.6e-1 ± 4.0e-2**	**7.6e-2 ± 3.3e-2**	1.0e-2 ± 3.7e-3	**4.1e-2 ± 8.6e-3**
Sc. l)	ERM	4.9e-1 ± 4.7e-2	8.6e-2 ± 3.3e-2	**8.2e-3 ± 3.2e-3**	4.4e-2 ± 8.8e-3
	db-ERM	**4.6e-1 ± 4.0e-2**	**7.5e-2 ± 3.1e-2**	9.9e-3 ± 3.5e-3	**4.1e-2 ± 8.3e-3**
	ub-ERM	4.7e-1 ± 4.2e-2	2.0e-1 ± 5.8e-2	2.0e-2 ± 7.0e-3	8.1e-2 ± 1.7e-2

D.2. Second experiments on the Adult *dataset*

In this subsection, we present another experiment on the *Adult* dataset showing the benefit of the debiasing approach we promote. Following a similar reasoning as that of Section 4, first notice that the age of the subject has a strong impact on his/her probability to earn more than 50k$ a year (see Figure 7(a)). Moreover, and as for the example based on years of education, this scenario cannot be cast as a Covariate Shift problem. Indeed, the conditional laws cannot be assumed to remain identical. Figure 7(b) illustrates this phenomenon by showing the dependence of the income with respect to the years of education by age group.

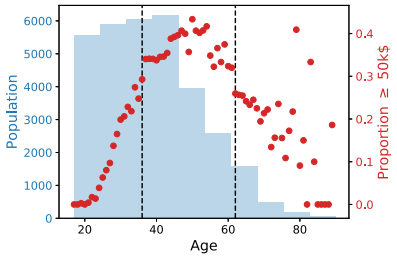


(a) Scenario g) (b) Scenario h)

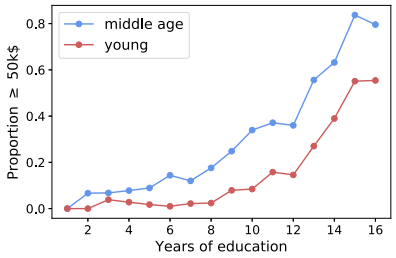
(c) Scenario i) (d) Scenario j)

(e) Scenario k) (f) Scenario l)

Fig 6: Different scenarios when selection bias applies to vector's first dimension



(a) Persons earning more than 50k$ w.r.t. age



(b) Persons earning more than 50k$ w.r.t. education by age group

Fig 7: Proportion of people earning more than 50k with respect to age (left), and with respect to years of education by age group (right)

TABLE 4
Prediction errors on the Adult *dataset, bias on age, averaged over 100 runs.*

	LogReg	RF
Standard ERM	21.26 ± 1.24	16.48 ± 0.52
Debiased ERM	**19.10 ± 1.09**	**15.91 ± 0.62**
Unbiased Sample	22.04 ± 1.96	19.54 ± 1.17

Clearly, middle age people take more advantage of their education than younger people, which is totally normal as they are working for a longer period. This observation makes simple covariate shift impossible to consider here. If middle age people happen to be over-represented in the training dataset, it should induce a general over-estimation of the probability, unless our general debiasing procedure is used. This setting has been simulated as follows. From the initial observations, 5 000 are kept for the testing phase. From the rest are sampled two subgroups: one of middle age people of size 9 900, and one unbiased (i.e., sampled from the entire population) of size 100. A Logistic Regression (LogReg) and a Random Forest (RF) are then trained on the concatenation of the 10 000 observations, with and without debiasing procedure, as well as on the small second sample of size 100 only. Numerical results are summarized in Table 4 in terms of prediction errors. Again, the debiased version of the ERM yields the best performances, and for both algorithms. The gaps are however less spectacular than that presented in Section 4. It is probably due to a softer biasing effect than the one achieved when it applies to the years of education. The less striking difference between conditional laws (Figure 4 and Figure 7(b)) is another marker that the debiasing effect expected in this latter example is less important.

References

[1] M. Achab, S. Clémençon, C. Tillier, and R. Vogel. Weighted empirical risk minimization: Sample selection bias correction based on importance sampling. In *Proceedings of the International Conference on Machine Learning, Artificial Intelligence and Applications*, 2020.

[2] G. Ausset, S. Clémençon, and F. Portier. Empirical Risk Minimization under Random Censorship: Theory and Practice. *Submitted, available at* https://arxiv.org/abs/1906.01908, 2019.

[3] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1), 2010. MR3108150

[4] T. Bolukbasi, K. Chang, J. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems (NIPS)*, page 4349–4357, 2016.

[5] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005. MR2182250

[6] K. Burns, L. Hendricks, K. Saenko, T. Darrell, and A. Rohrbach. Women also snowboard: Overcoming bias in captioning models. *arXiv preprint arXiv:1803.09797*, 2018.

[7] T. T. Cai and H. Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100–128, 2021. MR4206671

[8] K. Chua, Q. Lei, and J. D. Lee. How fine-tuning allows for effective meta-learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[9] S. Clémençon, P. Bertail, and E. Chautru. Sampling and empirical risk minimization. *Statistics*, 51(1):30–42, 2017. MR3600460

[10] R. D. Cook and F. B. Martin. A model for quadrat sampling with "visibility bias". *Journal of the American Statistical Association*, 69(346):345–349, 1974.

[11] C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450, 2010.

[12] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996. MR1383093

[13] J. Dubin and D. Rivers. Selection bias in linear regression, logit and probit models. *Sociological Methods & Research*, 18(2-3):360–390, 1989.

[14] M. Dudík, S. Phillips, and R. Schapire. Correcting sample selection bias in maximum entropy density estimation. In *Advances in neural information processing systems*, pages 323–330, 2006.

[15] J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning: Data-Mining, Inference and Prediction. Second edition*. Springer, New York, 2009. MR2722294

[16] R. Gill, Y. Vardi, and J. Wellner. Large sample theory of empirical distribu-

tions in biased sampling models. *The Annals of Statistics*, 16(3):1069–1112, 1988. MR0959189

[17] C. Godsil and G. Royle. *Algebraic Graph Theory*. Springer-Verlag, 2001. MR1829620

[18] J. Haldane. The estimation of the frequencies of recessive conditions in man. *Annals of Eugenics*, 8(3):255–262, 1938.

[19] S. Hanneke and S. Kpotufe. A no-free-lunch theorem for multitask learning. *arXiv preprint arXiv:2006.15785*, 2020.

[20] J. Heckman. Varieties of selection bias. *The American Economic Review*, 80(2):313, 1990.

[21] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963. MR0144363

[22] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. Smola. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608, 2007.

[23] Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in nonstandard situations. *Machine learning*, 46(1-3):191–202, 2002. MR2841444

[24] G. Lugosi. Learning with an unreliable teacher. *Pattern Recognition*, 25(1):79–87, 1992. MR1145407

[25] C. Manski and S. Lerman. The estimation of choice probabilities from choice based samples. *Econometrica: Journal of the Econometric Society*, pages 1977–1988, 1977. MR0501708

[26] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

[27] P. Massart. The tight constant in the dvoretsky-kiefer-wolfowitz inequality. *The Annals of Probability*, 18:1269–1283, 1990. MR1062069

[28] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct 2010.

[29] G. Papa, S. Clémençon, and P. Bertail. Learning from Survey Training Samples: Rate Bounds for Horvitz-Thompson Risk Minimizers. In *Proceedings of ACML*, 2016.

[30] G. Patil and C. Rao. The weighted distributions: a survey of their applications. *Applications of statistics*, pages 383–405, 1977.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. MR2854348

[32] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

[33] H. W. Reeve, T. I. Cannings, and R. J. Samworth. Adaptive transfer learning. *The Annals of Statistics*, 49(6):3618–3649, 2021. MR4352543

[34] S. Rosset, J. Zhu, H. Zou, and T. Hastie. A method for inferring label sampling mechanisms in semi-supervised learning. In *Advances in neural*

information processing systems, pages 1161–1168, 2005.
[35] D. B. Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004. MR2117498
[36] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000. MR1795598
[37] A. Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, pages 3–28, 2009.
[38] M. Sugiyama and M. Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. The MIT Press, 2012.
[39] M. Sugiyama and K. Müller. Input-dependent estimation of generalization error under covariate shift. *Statistics & Decisions*, 23(4/2005):249–279, 2005. MR2255627
[40] N. Tripuraneni, C. Jin, and M. Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434–10443. PMLR, 2021.
[41] V. Van Belle, K. Pelckmans, J. Suykens, and S. Van Huffel. Learning transformation models for ranking and survival analysis. *Journal of machine learning research*, page 44, 2011. MR2786912
[42] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.
[43] A. van der Vaart and J. Wellner. *Weak convergence and empirical processes*. Springer-Verlag, 1996. MR1385671
[44] E. van Miltenburg. Stereotyping and bias in the flickr30k dataset. In *Workshop on Multi-modal Corpora: Computer vision and language processing*, 2016.
[45] Y. Vardi. Nonparametric estimation in the presence of length bias. *The Annals of Statistics*, 10(2):616–620, 1982. MR0653536
[46] Y. Vardi. Empirical distributions in selection bias models. *Ann. Statist.*, 13:178–203, 1985. MR0773161
[47] F. Vella. Estimating models with sample selection bias: a survey. *Journal of Human Resources*, pages 127–169, 1998.
[48] K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, May 2016.
[49] C. Winship and R. Mare. Models for sample selection bias. *Annual review of sociology*, 18(1):327–350, 1992.
[50] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114. ACM, 2004.
[51] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.