Learning from Biased Data: A Semi-Parametric Approach

Stéphan Clémençon, Patrice Bertail, Yannick Guyonvarch, Nathan Noiry

Learning from Biased Data: A Semi-Parametric Approach

Patrice Bertail [*1] **Stephan Clémençon** [*2] **Yannick Guyonvarch** [*2] **Nathan Noiry** [*2]

Abstract

We consider risk minimization problems where the (source) distribution P_S of the training observations $Z_1, \ldots, Z_n$ differs from the (target) distribution P_T involved in the risk that one seeks to minimize. Under the natural assumption that P_S dominates P_T, *i.e.* $P_T \ll P_S$, we develop a semi-parametric framework in the situation where we *do not* observe any sample from P_T, but rather have access to some auxiliary information at the target population scale. More precisely, assuming that the Radon-Nikodym derivative $dP_T/dP_S(z)$ belongs to a parametric class $\{g(z, \alpha),\ \alpha \in \mathcal{A}\}$ and that some (generalized) moments of P_T are available to the learner, we propose a two-step learning procedure to perform the risk minimization task. We first select $\hat{\alpha}$ so as to match the moment constraints as closely as possible and then reweight each (biased) training observation Z_i by $g(Z_i, \hat{\alpha})$ in the final Empirical Risk Minimization (ERM) algorithm. We establish a $O_{\mathbb{P}}(1/\sqrt{n})$ generalization bound proving that, remarkably, the solution to the weighted ERM problem thus constructed achieves a learning rate of the same order as that attained in absence of any sampling bias. Beyond these theoretical guarantees, numerical results providing strong empirical evidence of the relevance of the approach promoted in this article are displayed.

1. Introduction

In the classic formulation of predictive learning problems, the goal is to find θ in a parameter space Θ with minimum *risk* $\mathcal{R}_P(\theta) = \mathbb{E}_P[\ell(Z, \theta)]$. Here Z is a random variable, taking its values in some measurable space $\mathcal{Z}$, with unknown distribution P and $\ell : \mathcal{Z} \times \Theta \to \mathbb{R}_+$ is a given loss function. In order to solve approximately the *risk minimization* problem

$$\inf_{\theta \in \Theta} \mathcal{R}_P(\theta), \tag{1}$$

one assumes that a training dataset $\mathcal{D}_n = \{Z_1, \ldots, Z_n\}$ composed of $n \geq 1$ independent copies of the generic r.v. Z is available. A natural learning procedure consists in replacing the unknown risk by its empirical version based on the Z_i's and solving next

$$\inf_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(Z_i, \theta). \tag{2}$$

The accuracy of this procedure, referred to as *Empirical Risk Minimization* (ERM in abbreviated form), is usually assessed by establishing upper confidence bounds for the risk excess of empirical minimizers, that is the difference between the risk of solutions $\hat{\theta}_n$ to (2) and the minimum risk attained over the class Θ, under suitable assumptions on the loss function ℓ and the parameter space Θ, see *e.g.* (Devroye et al., 1996). Such results offer statistical guarantees regarding the generalization capacity of the predictive rule encoded by the learned parameter $\hat{\theta}_n$, when applied to a new/test observation Z_{test} with distribution P.

The usual validity framework for ERM crucially relies on the assumption that the distributions of the random variables involved in the training and test/prediction stages are the same. However, this assumption is now highly arguable in a wide variety of situations. Whereas, in the recent past, data collection was expensive and still performed by means of carefully elaborated experimental designs through surveys and questionnaires, practitioners have more and more often poor control over the information acquisition process in the Big Data era. The massive datasets captured by connected devices (*e.g.* IoT sensors) or collected via the Internet may be in significant part corrupted, or non-representative of the target population. For instance, as recently hotly discussed and debated (see *e.g.* (Wang et al., 2019)), certain facial recognition systems may be trained on public databases, possibly very different from the statistical population to which they will be applied when deployed. In other words, the data available for predictive learning may frequently stem from a source distribution P_S that differs from the distribution of interest P_T, referred to as the target distribution here. One

[*]Equal contribution [1]Université Paris-Nanterre, France [2]Télécom Paris, France. Correspondence to: Patrice Bertail <patrice.bertail@parisnanterre.fr>, Stephan Clémençon <stephan.clemenconp@telecom-paris.fr>, Yannick Guyonvarch <yannick.guyonvarch@telecom-paris.fr>, Nathan Noiry <nathan.noiry@telecom-paris.fr>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

may refer to (Quionero-Candela et al., 2009) for an account of the different types of sampling biases or dataset shifts occurring frequently in practice.

In this paper, we consider risk minimization problems in presence of selection/sampling bias. They are now the subject of much attention in the literature, see *e.g.* (Bolukbasi et al., 2016), (Zhao et al., 2017) or (Liu et al., 2016), and can be viewed as specific *transfer learning* problems, *cf* (Pan and Yang, 2010) or (Weiss et al., 2016), insofar as the key idea is to find/approximate a suitable transformation of the source distribution P_S so that relevant information can be transferred from the training dataset to the test population. Most contributions documented in the literature focus on the situation where two samples are observed, one drawn from P_S and another one from P_T. In particular, two very popular transfer-learning algorithms have been proposed in this specific case: the Kullback Leibler Information Estimation Procedure (KLIEP) in (Sugiyama et al., 2007; 2008) and the Kernel Mean Matching (KMM) in (Huang et al., 2007), see also (Gretton et al., 2009). As instances from the target distribution are available to learn a predictive rule, the task performed by these algorithms essentially consists of data augmentation: the reweighted source sample is merely viewed as a tool to increase the size of the target sample.

Here, focus is on a more challenging problem, corresponding to many practical situations, where statistical learning must be based on a single sample drawn from P_S, combined with appropriate assumptions about the relation between distributions P_S and P_T. While it has been seldom considered in transfer learning, see however (Laforgue and Clémençon, 2019; Vogel et al., 2020), this setup has proved useful for various problems in Statistics. It is for instance used in the literature devoted to denoising tasks, where one aims at recovering the distribution of variables that are observed with some corrupting noise. It is generally assumed that uncorrupted observations are not available so that the distribution of interest has to be inferred from either one or several samples of noisy observations, based on structural assumptions on the nature of the noise, see *e.g.* (Chen et al., 2011; Kato and Sasaki, 2019; Kato et al., 2019).

In order to overcome the lack of data drawn from the target distribution P_T, two crucial assumptions are made in the present paper. First, the modelling we propose concerns the "link" function between P_T and P_S: we assume that P_T is absolutely continuous w.r.t. P_S and that the Radon-Nykodim derivative $dP_T/dP_S(z)$ belongs to a parametric family $\mathcal{G}$. This connects the framework we develop for learning from biased data to semi-parametric statistics, see *e.g.* (Gilbert et al., 1999) or (Zhang, 2000): as a matter of fact, even if the elements of class $\mathcal{G}$ are supposed to be characterized by a parameter α in $\mathcal{A} \subset \mathbb{R}^p$ with $p \geq 1$, no parametric assumption is made on the class Θ encoding the decision rule candidates. While tractable, our approach offers a great flexibility: as will be shown in the subsequent analysis, miss-specified models $\mathcal{G}$ can be handled as well. Second, we assume that we have access to several moments under P_T, or estimators thereof. This situation is rather common in practice, as a result of privacy constraints in particular. For instance, in countries where national censuses are conducted, only socio-economic information at the population level is made available to the public. The information furnished by Internet providers is another relevant example: those providers allow users to download freely summary statistics on queries made on the web worldwide but the knowledge of individual-level search data remains confidential.

Using both assumptions, the learning methodology we propose is close in spirit to the so-called calibration method in survey sampling (Deville and Särndal, 1992; Deville, 2000; Guggemos and Tillé, 2010) or its analogue in the econometrics literature (Imbens and Lancaster, 1994). It consists in solving a reweighted version of the ERM problem based on the Z_i's, using some auxiliary information, namely known moments from P_T. It is implemented in two steps. First, the link function $dP_T/dP_S(z)$ is learned via a *generalized-method-of-moments* approach, see *e.g.* (Hansen, 1982). Each (biased) training observation Z_i drawn from P_S can be next reweighted by $g(\alpha, Z_i)$ using functions $g(., \alpha)$ in $\mathcal{G}$ and one searches for a minimizer in $\mathcal{G}$ of the distance between the reweighted empirical moments and the known moments from P_T. We establish nonasymptotic concentration guarantees for this estimator, see Proposition 1. Second, the learned link function is then used to construct a reweighted version of the ERM problem (2). Considering a minimizer of the reweighted empirical risk $\hat{\theta}$, we finally prove a nonasymptotic generalization bound which controls the gap between $\mathcal{R}_{P_T}(\hat{\theta})$ and $\inf_{\theta \in \Theta} \mathcal{R}_{P_T}(\theta)$, see Theorem 1.

The paper is organized as follows. Section 2 presents the theoretical framework for statistical learning based on biased training data considered throughout the article. The algorithmic approach we propose is described in section 3, together with a rate bound analysis proving its accuracy. For illustration purpose, experimental results are displayed in section 4 and some concluding remarks are collected in section 5. The proofs of the main results are sketched in the Appendix section, while additional technical details are deferred to the Supplementary Material.

2. Theoretical Framework

In this section we present at length the framework we consider for statistical learning in presence of sampling bias and introduce the main notations of the paper. We also describe the semi-parametric approach we develop in the subsequent analysis and state the assumptions required to extend Em-

pirical Risk Minimization with statistical guarantees in the form of nonasymptotic generalization bounds. As described in the Introduction, P_S and P_T are two probability measures on a measurable space $\mathcal{Z}$, referred to as the source and target distributions respectively. Here and throughout, the target distribution is assumed to be absolutely continuous w.r.t. the source distribution, by $\|h\|_\infty$ is meant the sup norm of any bounded function $h : \mathcal{Z} \to \mathbb{R}$, the space $\mathbb{R}^q$, $q \geq 1$, is equipped with the usual Euclidean norm $\|.\|$, the unit sphere is denoted by $\mathbb{S}_{q-1} = \{u \in \mathbb{R}^q : \|u\| = 1\}$ and the Dirac mass at any point x by δ_x.

Assumption 1. (Domination) *The probability measure P_T is absolutely continuous with respect to the probability measure P_S. The corresponding Radon-Nikodym derivative is denoted by*

$$w : z \in \mathcal{Z} \mapsto \frac{\mathrm{d}P_T}{\mathrm{d}P_S}(z). \tag{3}$$

Remark 1. (Comparison with existing works) *Having in mind the generic decomposition $Z = (Y, X)$ with Y a label we would like to infer from a feature vector X, our framework encompasses the so-called covariate shift model where $P(y \mid x)$ is fixed and $P(x)$ is allowed to vary. It also covers the question tackled in (Vogel et al., 2020): the authors consider a classification problem where only the class probabilities (of labels) can differ across source and target and the probability of observing a positive label under P_T is assumed to be known. In this very specific case, the ratio $\mathrm{d}P_T^Y/\mathrm{d}P_S^Y(y)$ is partly known and the problem boils down to estimating the class probabilities under the source distribution. Comparing our work to (Laforgue and Clémençon, 2019) is also instructive. Their setup differs from ours: instead of having several biased samples at our disposal with known biasing functions, we assume knowledge of several moments under the target and stipulate a parametric model for the transfer/likelihood function. The framework we develop is more adapted to certain practical situations, when learning from socio-demographic/economic data for instance, as discussed in Remark 2.*

As formulated in the Introduction section, the goal pursued is to approximate a solution θ in Θ to the minimization problem (1) with $P = P_T$, based on i.i.d. training data $Z_1, \ldots, Z_n$ drawn from the distribution P_S, *a priori* different from P_T. Notice that under Assumption 1, the risk minimization problem (1) rewrites

$$\inf_{\theta \in \Theta} \mathbb{E}_{P_S}\left[w(Z)\ell(Z, \theta)\right]. \tag{4}$$

In the (unrealistic) situation where the Radon-Nikodym derivative w is known, one can solve the following weighted Empirical Risk Minimization problem:

$$\inf_{\theta \in \Theta} \frac{1}{n}\sum_{i=1}^{n} w(Z_i)\ell(Z_i, \theta). \tag{5}$$

By means of a direct application of the contraction principle, a generalization bound for the excess of P_T-risk of solutions to (5) can be easily established, showing that the same learning rate as that attained in the case where $P_S = P_T$, *i.e.* $O_{\mathbb{P}}(1/\sqrt{n})$, is achieved, see Lemma 1 in (Vogel et al., 2020).

Transfer learning with marginal constraints. In practice, the weight function $w(z)$ is unknown. The approach we develop here is of *plug-in* type. An estimator $\hat{w}$ of w based on some auxiliary information is first constructed and a version of (5), where w is replaced with $\hat{w}$ is next solved. To make such a strategy tractable, we assume that the weight function w belongs to a parametric class.

Assumption 2. (Parametric link function) *Let $\mathcal{A} \subset \mathbb{R}^p$ be Borel-measurable and let $g : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}_+$ be a measurable function such that $\|g\|_\infty = \sup_{\alpha \in \mathcal{A}} \|g(\cdot, \alpha)\|_\infty < +\infty$ and such that $\sup_{(\alpha_1, \alpha_2) \in \mathcal{A} \times \mathcal{A}} |g(z, \alpha_1) - g(z, \alpha_2)|/\|\alpha_1 - \alpha_2\| \leq L$ for some $L < +\infty$. The function w belongs to the class $\mathcal{G} := \{g(\cdot, \alpha) : \alpha \in \mathcal{A}\}$.*

Among the numerous parametric models that can be considered in practice, the class below shall serve as a running example in this section. To the best of our knowledge, the class presented in Example 1 has not be investigated for estimating Radon derivatives and is thus a nice complement to existing approaches which model w as the exponential of some (semi-)parametric function, see *e.g.* (Gilbert et al., 1999; Zhang, 2000).

Example 1. (Quadratic parametrization) *Let $z \in \mathcal{Z} \mapsto W(z)$ be a measurable function mapping $\mathcal{Z}$ to the set of $p \times p$ symmetric positive matrices. We define:*

$$\mathcal{G}_W := \left\{g(\cdot, \alpha) = \alpha^T W(\cdot)\alpha : \alpha \in \mathbb{R}^p\right\}.$$

The map $z \mapsto W(z)$ can be chosen based on some prior information about the form of the true Radon-Nikodym derivative w. As a simple example, we could pick $W(z)$ as a diagonal matrix with diagonal entries equal to positive-valued functions of z, which capture different characteristics of w, e.g. lightness/heavyness of the tails, occurrence of multiple local extrema, possible irregularities.

In the case where Assumption 2 is satisfied, there exists at least one element $\alpha_* \in \mathcal{A}$ such that $w(\cdot) = g(\cdot, \alpha_*)$. Observe that, for this α_*, the relation $\mathbb{E}_{P_S}[g(Z, \alpha_*)] = 1$ necessarily holds true. However, this relation is far from characterizing fully α_* and w in general, in the sense that many elements in $\mathcal{A}$ and then many elements in $\mathcal{G}$ may satisfy it. This issue is all the more acute as the class $\mathcal{G}$ grows richer. Having access to extra relations linking P_S and P_T is therefore crucial to build a criterion which is able to tell how plausible distinct values in $\mathcal{A}$ are. To design such a criterion, we assume that some extra features of the target distribution P_T are accessible. More precisely we suppose that the following

quantities are known

$$M_l := \mathbb{E}_{P_T}[m_l(Z)],\ l = 1,\ \ldots,\ d, \tag{6}$$

as well as the supposedly P_T-integrable functions $m_l : \mathcal{Z} \to \mathbb{R}$, $l \in \{1,\ \ldots,\ d\}$. If $w(\cdot) = g(\cdot, \alpha_*)$, the relation $\mathbb{E}_{P_S}[g(Z, \alpha_*)m_l(Z)] = M_l$ is satisfied for every $1 \leq l \leq d$.

Remark 2. (Macro-information about the target population) *The assumption that quantities of the form $(M_1,\ \ldots,\ M_d)$ or estimators thereof can be put at the learner's disposal is realistic in many practical situations. In the open data era, national statistical agencies increasingly enable access to a wealth of macro-level summary statistics on numerous topics (e.g. average wage, household composition, health status, life expectancy). For privacy reasons, the micro-level data behind those summary statistics is kept secret by national agencies. For instance, the InFuse portal of the Office for National Statistics in the United Kingdom, or the Quickfacts interface of the US Census Bureau provide macro-information at fairly disaggregated geographical levels (county-level and below in the US). In many fields, exhaustive data collection cannot be conducted at a very high frequency due to the substantial costs induced by this collection. On the other hand, cheap internet surveys allow to collect a huge quantity of data but there is generally little or even no control of the responding population. This highlights the importance of the design of statistical learning methods capable of exploiting biased microeconomic surveys combined with the extra-information brought by macro-level surveys.*

A single theoretical index that incorporates all the accessible information on P_T can be built:

$$\Psi_\infty : \alpha \mapsto \Psi_\infty(\alpha) = \left\| \mathbb{E}_{P_S}\left[g(Z, \alpha)\mathbf{m}(Z) - \mathbf{M}\right] \right\|, \tag{7}$$

with $\mathbf{m}(Z) = (1, m_1(Z), \ldots, m_d(Z))^T$ and $\mathbf{M} = (1, \ldots, M_d)^T$. The *transfer criterion* Ψ_∞ is always positive and equal to zero if and only if α satisfies all the constraints aforementioned. When $w \in \mathcal{G}$, there exists at least one element in $\alpha_* \in \mathcal{A}$ such that $\Psi_\infty(\alpha_*) = 0$. Attention should be paid to the fact that minimization of the criterion is relevant even in the miss-specified case where $w \notin \mathcal{G}$. In this context, it may happen that there is no $\alpha \in \mathcal{A}$ such that all the constraints are satisfied. However, the closer $\Psi_\infty(\alpha)$ to zero, the more plausible $g(\cdot, \alpha)$. As discussed at length in the Supplementary Material, the analysis carried out in Section 3 can be extended to this situation, at the price of an additional approximation term in the generalization bound.

The identifiability assumption below rules out some scenarios that are difficult to handle from a theoretical point of view.

Assumption 3. (Uniqueness) *There exists a unique $\alpha_* \in \mathcal{A}$ such that $\Psi_\infty(\alpha_*) = \min_{\alpha \in \mathcal{A}} \Psi_\infty(\alpha)$.*

Combined with the last point of Assumption 2 (*i.e* $w \in \mathcal{G}$), Assumption 3 ensures that w is the only element in $\mathcal{G}$ that satisfies all the moment constraints on which Ψ_∞ is based. We say that w is well-identified by the moment constraints considered. We maintain this hypothesis throughout the paper to simplify the analysis while conveying the main ideas. In the Supplementary Material, we study the more general case where several minimizers of Ψ_∞ may exist and discuss the implications in terms of identification of w. Existence of a minimizer of Ψ_∞ can be checked in many situations. This is the case for instance when Ψ_∞ is continuous on $\mathcal{A}$ and either $\mathcal{A}$ is compact, or $\mathcal{A} = \mathbb{R}^p$ and Ψ_∞ is coercive (*i.e* $\lim_{\|\alpha\| \to +\infty} \Psi_\infty(\alpha) = +\infty$).

Example 2. (Quadratic parametrization, bis) *We denote by $\|\cdot\|$ the operator norm on matrices, that is $\|W\| = \sup_{\|x\|=1} \|Wx\|$. When $\mathcal{G} = \mathcal{G}_W$, the map Ψ_∞ is continuous and attains its minimum on $\mathcal{A}$ as long as $\mathbb{E}_{P_S}[\,\|W(Z)\|\,] < +\infty$ and $\mathbb{E}_{P_S}[\,\|W(Z)\|\,|m_l(Z)|\,] < +\infty$ for every $1 \leq l \leq d$. This map is also coercive since $\mathbb{E}_{P_S}[W(Z)]$ is symmetric and strictly positive (see the Supplementary Material for a proof). Unfortunately $\Psi_\infty(\alpha_1) = \Psi_\infty(\alpha_2)$ does not imply $\alpha_1 = \alpha_2$, so that uniqueness of the minimizer of Ψ_∞ does not hold in general. We can however uniquely recover $g(\cdot, \alpha_*)$ under additional assumptions, for instance when $W(z)$ is diagonal for every $z \in \mathbb{R}^m$ and $d = p - 1$. Since $w \in \mathcal{G}_W$, it is enough to solve $\Psi_\infty(\alpha)^2 = 0$ which is equivalent to $\mathbb{E}_{P_S}[\alpha^T W(Z)\alpha\mathbf{m}(Z)] = \mathbf{M}$. Letting $\widetilde{\alpha} := (\alpha_1^2, \ldots, \alpha_p^2)^T$, this can be rewritten $\Gamma\widetilde{\alpha} = \mathbf{M}$, where Γ is a $p \times p$ matrix with first line equal to $\mathbb{E}_{P_S}\left[diag(W(Z))\right]^T$, and second to last lines equal to $(\mathbb{E}_{P_S}\left[m_l(Z)diag(W(Z))\right]^T)_{l=1}^{p-1}$. We conclude that $\widetilde{\alpha} = \Gamma^{-1}\mathbf{M}$ as long as Γ is invertible. We can set $\alpha_* = (\sqrt{\widetilde{\alpha}_1}, \ldots, \sqrt{\widetilde{\alpha}_p})^T$ and remark that we can uniquely identify $g(\cdot, \alpha_*) = \alpha_*^T W(\cdot)\alpha_*$.*

3. Semi-Parametric Transfer Learning

In this section, the algorithmic approach we propose to solve the learning task described in the previous section is detailed and a rate bound analysis is next carried out, providing generalization guarantees for the predictive rules built this way.

3.1. A Two-Stage Learning Approach

Here we present the extension of the ERM methodology we promote for statistical learning based on a biased training dataset. It is assumed that auxiliary information about the target population is available to the learner in the form of a vector of P_T-integrals of known functions $m_l(z)$. The learning procedure of *plug-in* type is implemented in two steps that are summarized in Fig. 1. One first minimizes over $\mathcal{A}$ a statistical counterpart of the *transfer criterion* $\Psi_\infty(\alpha)$, obtained by replacing the source distribution P_S in

(7) with its empirical version based on the Z_i's, namely

$$\Psi_n(\alpha) = \left\| \frac{1}{n} \sum_{i=1}^{n} g(Z_i, \alpha)\mathbf{m}(Z_i) - \mathbf{M} \right\|, \quad \alpha \in \mathcal{A}. \tag{8}$$

The minimizer $\hat{\alpha}$ thus obtained is next used to form a reweighted version of the ERM problem (Rw-ERM in abbreviated form), where each observation Z_i is weighted by $g(Z_i, \hat{\alpha})$. This can be seen as a surrogate to (5).

Attention should be paid to the fact that the link function $g(., \hat{\alpha})$ computed in the first step of the procedure sketched above can be used for additional learning tasks, involving other decision spaces and/or loss functions.

Remark 3. (Ease of implementation) *Popular implementations of ERM-like learning procedures such as scikit-learn (Pedregosa et al., 2018) support a weight option which allows to replace the empirical risk* $(1/n)\sum_{i=1}^{n} \ell(Z_i, \theta)$ *with* $(1/n)\sum_{i=1}^{n} \omega_i \ell(Z_i, \theta)$ *when specifying a list of weights* $(\omega_i)_{i=1}^{n}$. *In practice, minimizing the objective function* (10) *for a given ERM task can thus be done by feeding the weight option of existing algorithms with the weights* $(g(Z_i, \widehat{\alpha}))_{i=1}^{n}$ *stemming from* (9). *At a more theoretical level, solving* (10) *amounts to replacing the canonical empirical distribution* $(1/n)\sum_{i=1}^{n} \delta_{Z_i}$ *in the ERM procedure with the measure* $(1/n)\sum_{i=1}^{n} g(Z_i, \hat{\alpha})\delta_{Z_i}$. *We may also re-normalized the weights to 1 to obtain a true probability distribution.*

Weighted ERM through Semi-Parametric Transfer

Input: Biased training dataset $\{Z_1, \ldots, Z_n\}$, function $\mathbf{m}(z)$, parameter $\mathbf{M}$.

1 **Transfer criterion minimization.** Estimate the α-parameter

$$\hat{\alpha} \in \arg\min_{\alpha \in \mathcal{A}} \Psi_n(\alpha). \tag{9}$$

2 **Weighted ERM.** Minimize the resulting reweighted empirical risk

$$\hat{\theta} \in \arg\min_{\alpha \in \Theta} \frac{1}{n} \sum_{i=1}^{n} g(Z_i, \hat{\alpha})\ell(Z_i, \theta). \tag{10}$$

Output: $(\hat{\alpha}, \hat{\theta})$

Figure 1. The Rw-ERM Algorithm

In the subsequent analysis, we focus on the statistical performance of the predictive rules empirically defined by Rw-ERM. The design of practical algorithms (*e.g.* based on gradient descent) for computing approximately minimizers of the transfer and reweighted empirical risk is beyond the scope of the subsequent analysis. A detailed account of the optimization techniques used in the experiments displayed in Section 4 is however given in the Supplementary Material.

3.2. Main Results - Generalization Bounds

Just like for the implementation, two stages are required to establish generalization guarantees for the Rw-ERM methodology. As a first go, we show that the parameter $\hat{\alpha}$ produced in the first step of the algorithm is close to the optimal value α^*. Additional technical assumptions are required for this purpose. To obtain a concentration inequality on $\hat{\alpha}$, it is first crucial to describe the "richness' of the class of functions $\{g(\cdot, \alpha)\mathbf{m}(\cdot) : \alpha \in \mathcal{A}\}$. For this purpose, we consider the Rademacher complexity defined by

$$E_n(\mathcal{A}) := \mathbb{E}_\sigma \left[\sup_{\alpha \in \mathcal{A}} \left\| \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(Z_i, \alpha)\mathbf{m}(Z_i) \right\| \right],$$

where $\sigma_1, \ldots, \sigma_n$ are independent Rademacher variables, independent from the Z_i's.

Assumption 4. *There exist two constants* $K_1, K_2 > 0$ *such that*

$$\sup_{1 \leq l \leq d+1} |g(z, \alpha)m_l(z)| \leq K_1 \quad \text{and} \quad \mathbb{E}_{P_S^{\otimes n}}[E_n(\mathcal{A})] \leq K_2/\sqrt{n}.$$

Assumption 5. *There exist constants* $\varepsilon, R, c > 0$ *such that*
(i) $\forall \alpha \in \mathcal{A}$, $\|\alpha - \alpha_*\| > R \Rightarrow \Psi_\infty(\alpha) > \varepsilon$
(ii) $\forall (v, t) \in \mathbb{S}^{p-1} \times [-R, R]$, $\Psi_\infty(\alpha_* + tv) \geq c|t|$.

Remark 4. *Let us comment on the above assumptions. Assumption 4 is introduced to control the distance between* Ψ_n *and* Ψ_∞ *uniformly over* $\mathcal{A}$ *and with large probability. Its first requirement allows us to resort to the Bousquet-Talagrand inequality (Boucheron et al., 2013)[Theorem 12.5], and its second requirement is a high-level restriction on the complexity of our model. Many concrete examples of functional classes satisfy this second requirement. We illustrate this in the Supplementary Material, where we give simple primitive conditions on* $\mathcal{A}$ *and* g *to ensure* $\mathbb{E}_{P_S^{\otimes n}}[E_n(\mathcal{A})] \leq K_2/\sqrt{n}$. *Assumption 5.(i) implies that the minimum of the function* Ψ_∞ *(which is zero) cannot be arbitrarily approached by a sequence of parameters* $(\alpha_i)_{i \geq 1}$ *that do not converge to* α_*. *Assumption 5.(ii) rules out situations where* Ψ_∞ *is almost flat around its global minimizer. In the Supplementary Material, we give simple sufficient conditions in terms of the Hessian of* Ψ_∞^2.

Under the above assumptions, the functions Ψ_n and Ψ_∞ reach their minimum roughly at the same point, up to a term of order $1/\sqrt{n}$, with high probability.

Proposition 1. *Let* $\hat{\alpha} \in \arg\min \Psi_n(\alpha)$. *Under Assumptions 1 to 5, for every* $\delta \in (0,1)$ *there exists* $C(\delta) < +\infty$ *and* $n_{\delta,\varepsilon}$ *such that for every* $n \geq n_{\delta,\varepsilon}$,

$$\mathbb{P}_{P_S^{\otimes n}}\left(\|\hat{\alpha} - \alpha_*\| > \frac{2C(\delta)}{c\sqrt{n}}\right) \leq \delta. \tag{11}$$

The constant $C(\delta)$ *depends only on* δ, K_1, K_2 *and* d.

Proposition 1 provides theoretical guarantees on the first step (9) of our algorithm. The constants involved in its formulation are described in the Technical proofs and the Supplementary Material. More precisely, the bound (11) controls the level of error incurred when replacing α_* with $\hat{\alpha}$ and is instrumental in deriving our main result, namely Theorem 1. Proposition 1 may be of independent interest as it provides insight into the non-asymptotic behavior of a class of GMM problems (Generalized Method of Moment). In our setting, the GMM criterion is the functional Ψ_n we seek to minimize.

With Proposition 1 in hand, we are in a position to state a theoretical result that guarantees the generalization capacity of the minimizer of the weighted ERM problem (10). For this, we define the Rademacher complexity, this time associated to the class $\{\ell(\cdot,\theta) : \theta \in \Theta\}$:

$$E_n(\Theta) := \mathbb{E}_\sigma\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i \ell(Z_i,\theta)\right|\right].$$

We now state our main theorem, which ensures the solutio to (10) is close to the true risk with high probability.

Theorem 1. *Suppose that* $\|\ell\|_\infty := \sup_{\theta\in\Theta}\|\ell(\cdot,\theta)\|_\infty < +\infty$ *and Assumptions 1 to 5 hold. Let* $\delta \in (0,1)$. *Then, there exist* $C(\delta)$ *and* $n_{\delta,\varepsilon}$ *(defined in Proposition 1) such that for every* $n \geq n_{\delta,\varepsilon}$ *with probability at least* $1-\delta$

$$\begin{aligned}\mathcal{R}_{P_T}(\widehat{\theta}) - \inf_{\theta\in\Theta}\mathcal{R}_{P_T}(\theta) &\leq 4L\|\ell\|_\infty \frac{C(\delta/2)}{c\sqrt{n}} + 4\|g\|_\infty \mathbb{E}_{P_S^{\otimes n}}[E_n(\Theta)] \\ &\quad + \|g\|_\infty\|\ell\|_\infty\left\{\sqrt{\frac{2\ln(2/\delta)}{n}} + \frac{2\ln(2/\delta)}{3n}\right\}. \end{aligned} \tag{12}$$

This theorem shows that, remarkably, the same learning rate as that attained if training observations sampled from the target data generating process were available is achieved by the Rw-ERM method. The generalization bound can be decomposed into two terms. The first one, which depends on $C(\delta/2)$, stems from Proposition 1 and quantifies the error between $g(\cdot,\hat{\alpha})$ and $w(\cdot)$, while the second one comes from a stochastic control of (5). For completeness, notice finally that under standard regularity assumptions on the class $\{\ell(\cdot,\theta) : \theta \in \Theta\}$, the complexity term $\mathbb{E}_{P_S^{\otimes n}}[E_n(\Theta)]$ is of order $O(1/\sqrt{n})$.

4. Numerical Experiments

In this section, we present two numerical experiments which complement the previous theoretical analysis. We start with a simulated dataset where we design both the source and target distributions. We then turn to a more realistic framework: we use the Life Expectancy Dataset and create some distributional shifts on the observations.

Artificial data. The simulation scheme is as follows. We consider a regression model with features $X^S = (X_1^S, X_2^S)^T$ and outcome Y^S. To form the feature vector, we independently draw $X_1^S \sim \mathcal{N}(0,1)$ and $X_2^S \sim \Gamma(2,1)$. The outcome variable is generated as

$$Y^S = 0.5(c_1 X_1^S + c_2 X_2^S + c_3 X_1^S X_2^S) + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0,1) \perp\!\!\!\perp (X_1^S, X_2^S)$. The target distribution is then defined through the following Radon derivative:

$$w(x_1,x_2,y) = \frac{\mathrm{d}P_T}{\mathrm{d}P_S}(x_1,x_2,y) = \alpha_*^T W(x_1,x_2,y)\alpha_*,$$

where $\alpha_* = (1,1,2)^T$ and $W(x_1,x_2,y) = \mathrm{Diag}(x_1^2, 4x_2^2, 2y^2)$.

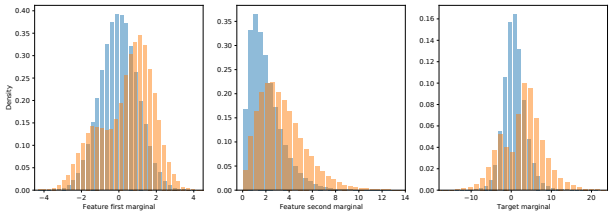

Figure 2. Histograms of the densities of the samples ($n_{obs} = 100,000$). In blue: source. In orange: target.

Notice that this seemingly simple model already displays striking differences between the source and the target distributions. As depicted in Figure 2, the unimodal distributions of X_1^S and Y^S turn into bimodal densities while the density of the second feature is shifted to the right.

Starting from a sample of size n_{obs} $(Z_i^S)_{i=1}^{n_{obs}}$ drawn in P_S, we seek to learn a regressor which generalizes well under P_T implementing our algorithm given by (9) and (10). We assume we know two marginal moments under P_T, namely $\mathbb{E}_{P_T}[X_1]$ and $\mathbb{E}_{P_T}[X_2]$ (these two quantities can be computed from the data generating process described above). To estimate α_*, we implement a gradient descent algorithm to minimize $\Psi_{n_{obs}}$. To avoid getting trapped in potential local minima, we rerun the descent algorithm n_{boot} times using a bootstrapping rationale presented in the Supplementary Material. Among the sequence $(\alpha^{(b)})_{b=1}^{n_{boot}}$ thus constructed, we select $\arg\min_{\alpha\in(\alpha^{(b)})_{b=1}^{n_{boot}}} \Psi_{n_{obs}}(\alpha^{(b)})$ as our final estimator $\hat{\alpha}$. In the last step, we train several regression-type algorithms

(OLS, SVR, RF) on $(Z_i)_{i=1}^{n_{obs}}$ with weights $(g(Z_i, \hat{\alpha}))_{i=1}^{n_{obs}}$. In parallel, we train those regression-type algorithms on the same training sample, but without weights, and on another sample of size n_{obs} drawn from P_T. We finally evaluate the performance of these three approaches through their MSE score computed on n_{test} new observations drawn from P_T. We repeat the whole procedure n_{rep} times. Table 1 presents the mean and standard error of the MSE of the algorithms under consideration for the following choices of parameters: $n_{obs} = 10,000$, $n_{test} = 500$, $n_{rep} = 100$ and $n_{boot} = 100$.

Table 1. Performance evaluation on simulated data using MSE.

ALGORITHM	RW-ERM(P_S)	ERM(P_T)	ERM(P_S)
OLS	**3.8 ± 0.4**	3.8 ± 0.4	6.3 ± 0.7
SVR	**1.5 ± 0.5**	1.2 ± 0.3	2.8 ± 0.8
RF	**1.7 ± 0.2**	1.6 ± 0.2	2.5 ± 0.4

No matter the ERM task carried out (*i.e* OLS, SVR or RF), we remark that the predictive accuracy of our new procedure (Rw-ERM(P_S)) is almost as good as that of an ERM trained on unbiased data (ERM(P_T)). On the other hand, using P_S without reweighting to build a regressor (ERM(P_S)) does not generalize well under P_T.

Experimental results on the Life Expectancy Dataset. Let us now consider the slightly more realistic situation where we do not have access to the weight function $g(\alpha_*, z)$. We use the Life Expectancy Dataset and only keep the *Adult Mortality Rate* (x_1) and the *Alcohol Consumption* (x_2) features in order to predict the *Life Expectancy* (y) output. We drop the observations containing a missing value and normalize the resulting sample (of size 2,735) whose empirical distribution serves as the target. The data are divided into two groups: G_1 containing 90% of the data ($n_{G_1} = 2,462$) and G_2 containing the rest ($n_{G_2} = 273$). G_2 corresponds to the test set. The source sample is artificially built from G_1 with the use of a stratified sampling procedure. We first partition G_1 thanks to a grid on the two-dimensional space spanned by the *Adult Mortality Rate* and the *Life Expectancy* variables, and we assign a probability weight p_{ij} to each box B_{ij}. Then, a source observation is created by selecting a box B_{ij} with probability p_{ij} and drawing uniformly at random an element of $G_1 \cap B_{ij}$. Figure 3 provides a heatmap representation of this procedure: each box B_{ij} is associated with its probability to be sampled in the source representation (left-hand graph), and with the true proportion of observations it contains in the target representation (right-hand graph).

Using this stratified sampling process, we build a sample of n_{G_1} source observations. We then run our two-step procedure on this sample using the parametric class $\{g(z, \alpha) = \alpha^T \text{Diag}(x_1^2, y^2)\alpha,\ \alpha \in \mathbb{R}^2\}$ for the first step, and the OLS and SVR algorithms for the second step. We compare the MSE performance of our approach with the MSE

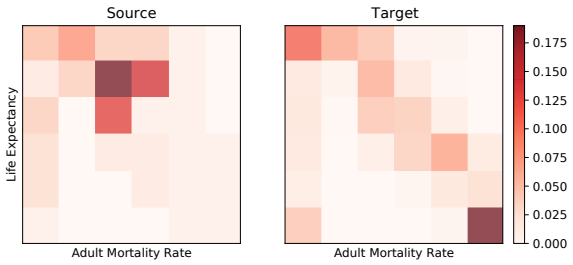

Figure 3. Heatmap representation of the stratified sampling procedure.

performances of the same machine learning algorithms, respectively trained *without weights* on the same source sample and on the initial sample G_1 – whose empirical law is roughly equal to the target distribution. The corresponding scores are displayed in Table 2.

Table 2. Performance evaluation on real data using MSE.

ALGORITHM	RW-ERM(P_S)	ERM(P_T)	ERM(P_S)
OLS	**0.50 ± 0.08**	0.45 ± 0.07	0.71 ± 0.09
SVR(0.01)	**0.71 ± 0.19**	0.40 ± 0.07	0.83 ± 0.11
SVR(0.1)	**0.39 ± 0.08**	0.35 ± 0.07	0.50 ± 0.09
SVR(1)	**0.39 ± 0.08**	0.34 ± 0.07	0.37 ± 0.07

In the OLS case, our approach seems to be efficient: the MSE is very close to the one obtained when training directly on the target, whereas an unweighted training on the source performs poorly. The SVR algorithm is run for three different values of the parameter C (0.01, 0.1 and 1). This corresponds to an increasing amount of complexity in the model learned during the second step of our algorithm. In contrast, the complexity of the first step corresponds to the quality of the chosen parametric class and remains constant. This could explain why the performance of our method draws nearer to that of naive ERM(P_S) as C becomes larger. We conjecture that a solution to alleviate this issue would be to increase the complexity of the first step, by choosing a higher-dimensional parametric class for instance.

5. Conclusion

We considered the problem of building a machine learning algorithm when the training data at hand stems from a biased version of the target distribution we wish to generalize on. We placed ourselves in the context where we do not have access to individual observations sampled from the target, but where some of its moments are at our disposal, a

frequently encountered framework when dealing with sensitive data. We proposed a flexible semi-parametric approach (Rw-ERM) to leverage this auxiliary knowledge, which can be decomposed into two steps: 1) an appropriate link function in a parametric class is first learned by minimizing an empirical criterion involving the known moments, 2) based on the former, a weighted version of the ERM procedure is next performed. Under natural assumptions on our model, we provide non-asymptotic guarantees on this procedure, revealing that the learning rate achieved by Rw-ERM is of the same order as that attained when training a learning algorithm directly on the target. We also provided preliminary experimental results illustrating the relevance of the procedure. Let us finally mention some exciting future lines of investigation. A first interesting question concerns the choice of the parametric class: given some *a priori* knowledge on the shape of the link, are there some natural choices to make? In another direction, our approach could be extended in a non-parametric fashion, where the link function is estimated directly using flexible learning algorithms with loss given by the empirical *transfer criterion* Ψ_n.

Technical Proofs

The proofs of the results stated in Section 3 are sketched below. We start with the proof of Proposition 1 followed by that of Theorem 1. For the sake of completeness, we state two additional lemmas which are instrumental in obtaining our results. The proofs of these lemmas together with all additional details are postponed to the Supplementary Material.

Proof of Proposition 1

The proof of Proposition 1 first consists in introducing for every $\delta \in (0,1)$

$$\Omega_{\delta,n} = \left\{ \sup_{\alpha \in \mathcal{A}} \left| \Psi_n(\alpha) - \Psi_\infty(\alpha) \right| \leq \frac{C(\delta,n)}{\sqrt{n}} \right\}.$$

On the event $\Omega_{\delta,n}$, the functions Ψ_n and Ψ_∞ are *uniformly* close. Combining this fact with Assumption 5, it is then possible to prove that the inequality $\|\hat{\alpha} - \alpha_*\| \leq C(\delta,n)/(c\sqrt{n})$ holds *deterministically* on the event $\Omega_{\delta,n}$. Therefore, the demonstration essentially boils down to obtaining the following lemma, whose proof is postponed to the Supplementary Material.

Lemma 1. *Let $\delta \in (0,1)$. Under Assumption 4, with probability at least $1-\delta$*

$$\sup_{\alpha \in \mathcal{A}} \left| \Psi_n(\alpha) - \Psi_\infty(\alpha) \right| \leq \frac{C(\delta,n)}{\sqrt{n}},$$

where $C(\delta,n)$ is an explicit quantity given in the proof which depends only on δ, n, K_1, K_2 and d that satisfies $\sup_{n\geq 1} C(\delta,n) < +\infty$.

Assuming Lemma 1 is proved, let us give more details on the approach we have just described. Let $\delta \in (0,1)$. We place ourselves on the event $\Omega_{\delta,n}$. In order to lighten notation, we introduce

$$t := C(\delta,n)/\sqrt{n}.$$

Let $\hat{\alpha} \in \arg\min \Psi_n(\alpha)$. We claim that $\|\hat{\alpha} - \alpha_*\| \leq R$. Indeed, if one had $\|\hat{\alpha} - \alpha_*\| > R$, then one would have that $\Psi_\infty(\hat{\alpha}) > \varepsilon$ by Assumption 5.(i). Since we placed ourselves on the event $\Omega_{\delta,n}$, we would then have the following lower bound $\Psi_n(\hat{\alpha}) \geq \Psi_\infty(\hat{\alpha}) - t > \varepsilon - t$. The latter would contradict the fact that Ψ_n attains its minimum at $\hat{\alpha}$, since $\Psi_n(\alpha_*) \leq t$ and since $t < \varepsilon - t$ for large enough n. More precisely, the latter inequality is true whenever $n \geq n_{\delta,\varepsilon} := 4 \sup_{n\geq 1} C(\delta,n)^2/\varepsilon^2$.

Therefore there exists some $v \in \mathbb{S}^{p-1}$ and some $\lambda_0 \in [0,R]$ such that $\hat{\alpha} = \alpha_* + \lambda_0 v$. This prompts us to introduce the functions:

$$\begin{cases} f_n(\lambda) := \Psi_n(\alpha_* + \lambda v), \\ f_\infty(\lambda) := \Psi_\infty(\alpha_* + \lambda v), \end{cases}$$

which are well-defined for every $\lambda \in [0,R]$.

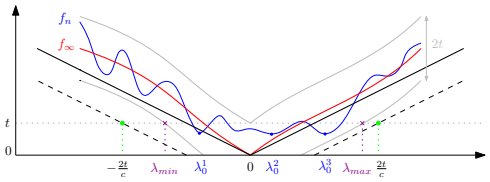

Figure 4. In red: the function f_∞. In blue: the function f_n, which attains its minimum at λ_0^1, λ_0^2 and λ_0^3. The plain gray curves corresponds to the translations by t of f_∞. They delimit the area to which the function f_n belongs on the event $\Omega_{\delta,n}$. The heavy black line is the function $\lambda \mapsto c|\lambda|$, the lower bound on f_∞ given by Assumption 5.(ii) and the dashed black line corresponds to its translation by $-t$. The points that are the most distant from 0 where the function f_n could attain its minimum are λ_{min} and λ_{max}, in magenta. These reals satisfy $-2t/c \leq \lambda_{min}$ and $\lambda_{max} \leq 2c/t$ under Assumption 5.(ii): the worst situation corresponds to $f_\infty(\lambda) = c|\lambda|$, in which case f_n could possibly attain its minimal value at $2t/c$ (or $-2c/t$) which is the abscissa of the (green) point of intersection between the curve $y = t$ and $y = cx - t$ (or $y = -cx - t$).

Finally, we claim that $\lambda_0 \leq \frac{2t}{c}$. Indeed, if one had $\lambda_0 > \frac{2t}{c}$, then $\min_{\alpha\in\mathcal{A}} \Psi_n(\alpha) = \Psi_n(\hat{\alpha}) = f_n(\lambda_0) > t$ by Assumption 5.(ii), which would contradict the fact that $\hat{\alpha} \in \arg\min \Psi_n(\alpha)$ since $\Psi_n(\alpha_*) \leq t$. See Figure 4 for an illustration. Hence, on the event $\Omega_{\delta,n}$ for every $n \geq n_{\delta,\varepsilon}$, $\|\hat{\alpha} - \alpha_*\| \leq 2t/c = (2C(\delta,n))/(c\sqrt{n})$. We set $C(\delta) := \sup_{n\geq 1} C(\delta,n)$ to conclude. □

Proof of Theorem 1

To avoid space-consuming formulas, we introduce the two following notations for the reweighted risk and its empirical

counterpart:

$$\begin{cases}\mathcal{R}^{\alpha}(\theta) := \mathbb{E}_{P_S}[g(Z,\alpha)\ell(Z,\theta)],\\ \mathcal{R}_n^{\alpha}(\theta) := \frac{1}{n}\sum_{i=1}^{n} g(Z_i,\alpha)\ell(Z_i,\theta).\end{cases}$$

We remark that for every $\theta \in \Theta$, $\mathcal{R}_{P_T}(\theta) = \mathcal{R}^{\alpha_*}(\theta)$ so that

$$\mathcal{R}_{P_T}(\hat{\theta}) - \inf_{\theta\in\Theta}\mathcal{R}_{P_T}(\theta) = \mathcal{R}^{\alpha_*}(\hat{\theta}) - \inf_{\theta\in\Theta}\mathcal{R}^{\alpha_*}(\theta) =: (I)$$

Let θ_* be such that $\mathcal{R}^{\alpha_*}(\theta_*) = \inf_{\theta\in\Theta}\mathcal{R}^{\alpha_*}(\theta)$. Using that $\mathcal{R}_n^{\hat{\alpha}}(\hat{\theta}) - \mathcal{R}_n^{\hat{\alpha}}(\theta_*) \leq 0$ by definition of $\hat{\theta}$, we can deduce that

$$(I) \leq 2\sup_{\theta\in\Theta}\left|\mathcal{R}_n^{\hat{\alpha}}(\theta) - \mathcal{R}_n^{\alpha_*}(\theta)\right| + 2\sup_{\theta\in\Theta}\left|\mathcal{R}_n^{\alpha_*}(\theta) - \mathcal{R}^{\alpha_*}(\theta)\right|. \quad (13)$$

Using Assumption 2, boundedness of ℓ and Equation (11), we can bound the first term with probability at least $1-\delta/2$ as follows

$$\begin{aligned}2\sup_{\theta\in\Theta}\left|\mathcal{R}_n^{\hat{\alpha}}(\theta) - \mathcal{R}_n^{\alpha_*}(\theta)\right| &\leq 2L\|\ell\|_\infty\|\hat{\alpha}-\alpha_*\|\\ &\leq 4L\|\ell\|_\infty\frac{C(\delta/2)}{c\sqrt{n}}. \quad (14)\end{aligned}$$

The second term in (13) can be bounded using Lemma 2 which is proved in the Supplementary Material.

Lemma 2. *Let $\delta \in (0,1)$. Under Assumption 2 and $\|\ell\|_\infty < +\infty$, with probability at least $1-\delta$*

$$\begin{aligned}\sup_{\theta\in\Theta}\left|\mathcal{R}_n^{\alpha_*}(\theta) - \mathcal{R}^{\alpha_*}(\theta)\right| \leq& 4\|g\|_\infty\mathbb{E}_{P_S^{\otimes n}}[E_n(\Theta)]\\ &+ \|g\|_\infty\|\ell\|_\infty\sqrt{\frac{2\ln(1/\delta)}{n}}\\ &+ \frac{2\|g\|_\infty\|\ell\|_\infty\ln(1/\delta)}{3n}.\end{aligned}$$

We apply Lemma 2 with $\delta/2$ instead of δ. This ends the proof. □

Acknowledgements

The authors are greatly indebted to the chair DSAIDIS of Telecom Paris for the support.

References

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*, page 4349–4357.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford.

Chen, X., Hong, H., and Nekipelov, D. (2011). Nonlinear models of measurement errors. *Journal of Economic Literature*, 49(4):901–37.

Deville, J.-C. (2000). Generalized calibration and application to weighting for non-response. In Bethlehem, J. G. and van der Heijden, P. G. M., editors, *COMPSTAT*, pages 65–76, Heidelberg. Physica-Verlag HD.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.

Gilbert, P. B., Lele, S. R., and Vardi, Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to aids vaccine trials. *Biometrika*, 86(1):27–43.

Giné, E. and Nickl, R. (2015). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Scholkopf, B. (2009). Covariate shift by kernel mean matching. In *NIPS 2009*.

Guggemos, F. and Tillé, Y. (2010). Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. *Journal of Statistical Planning and Inference*, 140(11):3199 – 3212.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054.

Huang, J., Gretton, A., Borgwardt, K. M., Schölkopf, B., and Smola, A. J. (2007). Correcting sample selection bias by unlabeled data. In *NIPS*, pages 601–608.

Imbens, G. W. and Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *The Review of Economic Studies*, 61(4):655–680.

Kato, K. and Sasaki, Y. (2019). Uniform confidence bands for nonparametric errors-in-variables regression. *Journal of Econometrics*, 213(2):516 – 555.

Kato, K., Sasaki, Y., and Ura, T. (2019). Inference based on kotlarski's identity.

Laforgue, P. and Clémençon, S. (2019). Statistical learning from biased training samples.

Liu, Z., Yang, J.-a., Liu, H., and Wang, W. (2016). Transfer learning by sample selection bias correction and its application in communication specific emitter identification. *JCM*, 11:417–427.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay (2018). Scikit-learn: Machine learning in python.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2009). *Dataset Shift in Machine Learning*. The MIT Press.

Sugiyama, M., Nakajima, S., Kashima, H., Von Buenau, P., and Kawanabe, M. (2007). Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*, volume 7, pages 1433–1440. Citeseer.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., Motoaki, and Kawanabe (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 8(35):985–1005.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence of Empirical Processes: with Applications to Statistics*. Springer-Verlag New York.

Vogel, R., Achab, M., Clémençon, S., and Tillier, C. (2020). Weighted empirical risk minimization: Sample selection bias correction based on importance sampling. In *Proceedings of ICMA*.

Wang, M., Deng, W., Hu, J., Tao, X., and Huang, Y. (2019). Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, pages 692–702. IEEE.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, 3(1):9.

Zhang, B. (2000). M-estimation under a two-sample semiparametric model. *Scandinavian Journal of Statistics*, 27(2):263–280.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*.