Learning from Biased Data: A Semi-Parametric Approach

Patrice Bertail, Stéphan Clémençon, Yannick Guyonvarch, Nathan Noiry

Objective

Learning a decision rule $\theta \in \Theta$ on a space $\mathcal{Z}$ from source observations

$$Z_1 = (Y_1, X_1), \ldots, Z_n = (Y_n, X_n) \overset{\text{i.i.d.}}{\sim} P_S,$$

able to generalize under a different target distribution P_T. We measure the quality of the estimator according to the risk functional

$$\mathcal{R}_{P_T}(\theta) := \mathbb{E}_{P_T}[\ell(Z, \theta)],$$

where $\ell : \mathcal{Z} \times \Theta \to \mathbb{R}_+$ is a given loss function.

Parametric Transfer

We assume that P_T is absolutely continuous with respect to P_S and denote by

$$w(z) := \frac{\mathrm{d}P_T}{\mathrm{d}P_S}(z)$$

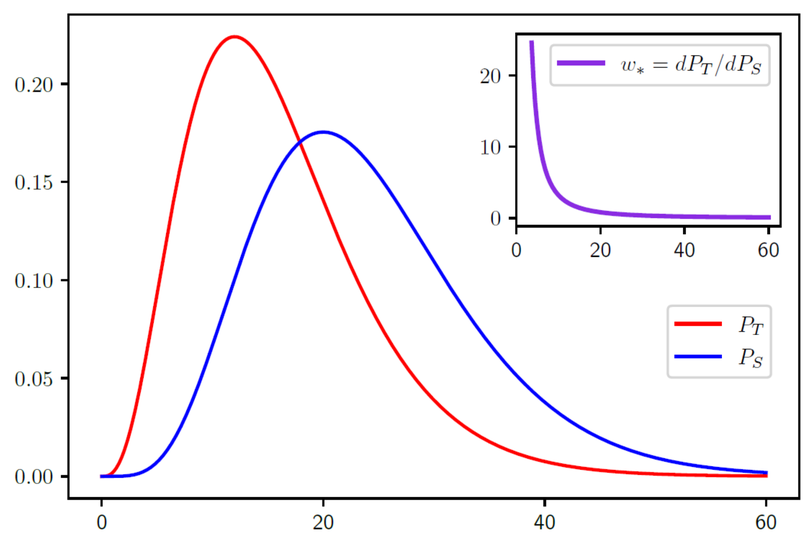
the corresponding Radon-Nikodym derivative. Furthermore, we assume that w belongs to the class

$$\mathcal{G} := \{g(\cdot, \alpha) : \alpha \in \mathcal{A}\},$$

where $\mathcal{A}$ is a compact subspace[a] of $\mathbb{R}^p$, $p \geq 1$.

[a] In this poster, we do not aim at the greatest possible generality and refer the interesting reader to the conference paper for more details.

Illustration



A Priori Knowledge on the Target

In contrast with many papers, we do have access to a sample under P_T and thus importance sampling is ruled out.

We rather assume a priori knowledge on some characteristics of P_T. More precisely, we suppose we know some functions $m_1, \ldots, m_d : \mathcal{Z} \to \mathbb{R}$ and their corresponding expectations under P_T:

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

Ubiquity of Auxiliary Macro-Information

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 portal of the Office for National Statistics in the United Kingdom, or the interface of the US Census Bureau provide macro-information at fairly disaggregated geographical levels (county-level and below in the US).

A Two Steps Procedure

1. Transfer criterion minimization. Estimate the α-parameter

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

with $\mathbf{m} = (m_1, \ldots, m_d)$ and $\mathbf{M} = (M_1, \ldots, M_d)$. Minimizing (1) amounts to finding the empirical measure which matches the moments $\mathbf{M}$ the best.

2. Weighted ERM. Minimize the resulting reweighted empirical risk

$$\hat{\theta} \in \arg\min_{\alpha \in \Theta} \frac{1}{n} \sum_{i=1}^{n} g(Z_i,\ \hat{\alpha})\ell(Z_i,\ \theta). \tag{2}$$

Statistical guarantees

Let

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

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

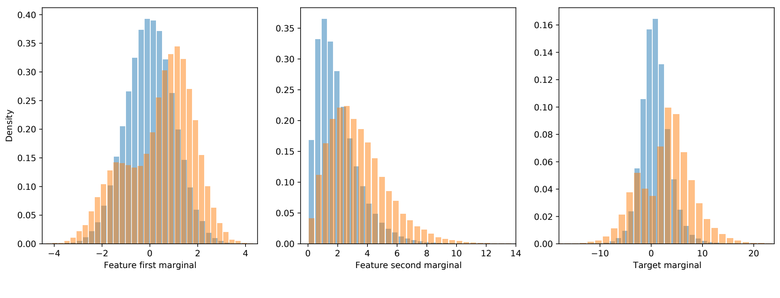
Let $\delta \in (0, 1)$. Then, there exist $C(\delta)$ and $n_{\delta,\varepsilon}$ such that for every $n \geq n_{\delta,\varepsilon}$ with probability at least $1 - \delta$

$$\mathcal{R}_{P_T}(\widehat{\theta}) - \inf_{\theta \in \Theta} \mathcal{R}_{P_T}(\theta) \leq \frac{C(\delta)}{\sqrt{n}}. \tag{3}$$

Numerical Experiments

Synthetic data: regression in dimension two.

- Source (in blue): $Z^S = (Y^S, X_1^S, X_2^S)$, with $X_1^S \sim \mathcal{N}(0, 1)$, $X_2^S \sim \Gamma(2, 1)$ and $Y^S = 0.5(c_1 X_1^S + c_2 X_2^S + c_3 X_1^S X_2^S) + \varepsilon$ with a noise term $\varepsilon \sim \mathcal{N}(0, 1)$ and $c_1 = 2$, $c_2 = 0.8$, $c_3 = 1.3$.
- Target (in orange) defined by the link function $w(y, x_1, x_2) = \mathrm{d}P_T(y, x_1, x_2)/\mathrm{d}P_S(y, x_1, x_2) = \alpha_*^T W(x_1, x_2, y)\alpha_*$ with $\alpha_* = (1, 1, 2)^T$ and $W(x_1, x_2, y) = \mathrm{Diag}(x_1^2, 4x_2^2, 2y^2)$.



We assume knowledge of $10,000$ observations drawn from P_S, and of the target moments $\mathbb{E}_{P_T}[X_1]$ and $\mathbb{E}_{P_T}[X_2]$. We apply our two steps methodology for three machine learning decision rules and evaluate the performances through MSE scores computed on 500 new observations drawn from P_T. The table below presents the mean and standard error of the MSE of the algorithms under consideration for 100 repetitions of the whole procedure.

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

We see that our procedure (Rw-ERM(P_S)) has similar performances compared to an algorithm directly trained on target observations (ERM(P_T)), while an estimator trained on the source without reweighting (ERM(P_S)) does not generalize well under P_T.