Exploiting geometric structure in mixture proportion estimation with generalised Blanchard-Lee-Scott estimators

Reeve, Henry W. J.; Kaban, Ata

Exploiting geometric structure in mixture proportion estimation with generalised Blanchard-Lee-Scott estimators

Henry W J Reeve H.W.J.REEVE@BHAM.AC.UK
School of Computer Science, University of Birmingham, Edgbaston, B15 2TT, Birmingham, UK

Ata Kabán A.KABAN@CS.BHAM.AC.UK
School of Computer Science, University of Birmingham, Edgbaston, B15 2TT, Birmingham, UK

Abstract

Mixture proportion estimation is a building block in many weakly supervised classification tasks (missing labels, label noise, anomaly detection). Estimators with finite sample guarantees help analyse algorithms for such tasks, but so far only exist for Euclidean and Hilbert space data. We generalise the framework of Blanchard, Lee and Scott to allow extensions to other data types, and exemplify its use by deducing novel estimators for metric space data, and for randomly compressed Euclidean data – both of which make use of favourable geometry to tighten guarantees. Finally we demonstrate a theoretical link with the state of the art estimator specialised for Hilbert space data.

Keywords: Mixture proportion estimation, metric spaces, covering dimension, random projections, Gaussian width.

1. Introduction and background

The problem of mixture proportion estimation (MPE) is as follows: Suppose that there are three Borel probability measures F, G and H supported on a topological space $\mathcal{X}$ and some $\kappa \in [0,1]$ with

$$F = (1-\kappa) \cdot G + \kappa \cdot H.$$

The learner is given access to i.i.d. samples from H and F only:

$$X_0^1, \cdots, X_0^{n_0} \overset{\text{i.i.d.}}{\sim} H; \qquad X_1^1, \cdots, X_1^{n_1} \overset{\text{i.i.d.}}{\sim} F.$$

and the goal is to estimate κ.

This problem was studied by Blanchard, Lee and Scott in their seminal work on semi-supervised novelty detection Blanchard et al. (2010), and more recently by Blanchard et al. (2016) for classification with label noise. As noted there, the problem is ill-defined without assumptions – indeed, if for some $\kappa \in [0,1]$ there exists a distribution G such that $F = (1-\kappa) \cdot G + \kappa \cdot H$ then for any $\tilde{\kappa} \in (0, \kappa)$ there exists another distribution $\tilde{G}$ such that $F = (1-\tilde{\kappa}) \cdot \tilde{G} + \tilde{\kappa} \cdot H$. Since the learner is only in possession of data from F and H (and not G) they are unable to distinguish between κ and $\tilde{\kappa}$. The authors then introduce a minimal assumption (irreducibility condition) that allows them to devise a consistent estimator. However they also prove that the convergence of $\hat{\kappa}_{\text{BLS}}$ to κ may be arbitrarily slow (Blanchard et al., 2010, Corollary 10) without stronger assumptions.

Irreducibility condition 1 *We say that (G, H) satisfies the irreducibility condition if for any $\gamma \in [0,1]$ and distribution I with $G = \gamma \cdot H + (1-\gamma) \cdot I$ we must have $\gamma = 0$.*

The construction of $\hat{\kappa}_{\text{BLS}}$ is as follows. Take a sequence of subsets of $\mathcal{X}$ $(\mathcal{S}_k)_{k\in\mathbb{N}}$ with respective VC dimensions $(V_k)_{k\in\mathbb{N}}$. The Blanchard-Lee-Scott estimator $\hat{\kappa}_{\text{BLS}}$ for κ is defined by

$$\hat{\kappa}_{\text{BLS}} \doteq \min\left\{\inf_{k\in\mathbb{N}}\left\{\inf_{S\in\mathcal{S}_k}\left\{\frac{\hat{F}_{n_1}(S)+\epsilon_{\text{VC}}(V_k,\delta,n_1)}{\left(\hat{H}_{n_0}(S)-\epsilon_{\text{VC}}(V_k,\delta,n_0)\right)_+}\right\}\right\},1\right\}.$$

where

$$\epsilon_{\text{VC}}(q,\delta,n) \doteq \sqrt{\frac{8q\log(2en/q)+8\log(4/\delta)}{n}}.$$

is a quantity constructed to be a high probability $(1-\delta)$ uniform upper bound on the deviation between the probability of a set of VC dimension q and its empirical measure under an i.i.d. sample (see (Mohri et al., 2012, Chapter 3)).

The first condition that allows a rate of convergence of $\hat{\kappa}_{\text{BLS}}$ to κ is due to Scott (2015).

Irreducibility condition 2 *We say that (G,H) satisfies Scott's irreducibility condition if supp$(H)\not\subset$ supp(G).*

In particular, there must be some open ball $B\subset\mathbb{R}^d$ with $H(B)>0$ and $G(B)=0$. Let $\mathcal{B}_d$ denote the set of all open Euclidean balls in $\mathbb{R}^d$. In the presence of condition 2 it suffices to consider the following simplified variant of the Blanchard-Lee-Scott estimator due to Scott (2015),

$$\hat{\kappa}_{\text{Scott}} \doteq \min\left\{\inf_{B\in\mathcal{B}_d}\left\{\frac{\hat{F}_{n_1}(B)+\epsilon_{\text{VC}}(d+1,\delta,n_1)}{\left(\hat{H}_{n_0}(B)-\epsilon_{\text{VC}}(d+1,\delta,n_0)\right)_+}\right\},1\right\}.$$

Theorem 1 (Scott (2015)) *Suppose that (G,H) satisfies Scott's irreducibility condition (2), with $\nu \doteq \sup_{B\in\mathcal{B}_d}\{H(B):G(B)=0\}$. Then, with probability at least $1-\delta$, we have*

$$\kappa \le \hat{\kappa}_{Scott} \le \kappa + \frac{8}{\nu}\cdot\epsilon_{VC}\left(d+1,\delta,\min\{n_0,n_1\}\right).$$

The work of Scott (2015) concerned convergence rates, treating d as constant. This is of course impractical for large d. A more practical estimator with the same convergence rate, specialised for reproducing kernel Hilbert space, was proposed in Ramaswamy et al. (2016) under quite different condition – we shall discuss this in a later section. Other approaches to MPE type problems assume the availability of some additional information or data, these are outside our scope here.

From a theoretical perspective, we observe that the BLS estimator lends itself to extensions to other probability spaces / data types in principle – including types that would incur too much distortion to embed into a Hilbert space, for instance metric space data that has applications in computer vision (Gottlieb and Kontorovich, 2014). We shall generalise the framework in Section 2, and instantiate it in Sections 2 and 3, where we develop dimension-free uniform deviation bounds exploiting favourable geometry. The final section will establish a connection with the approach in Ramaswamy et al. (2016).

2. Generalised Blanchard-Lee-Scott estimators

This section gives a construction to convert any uniform deviation bound into an BLS-type mixture proportion estimator with a quantified estimation error, in a generic abstract setting. Let $\mathcal{X}$ be a topological space. Let G, H, F be Borel probability measures on $\mathcal{X}$, and $\kappa \in [0,1]$ such that $F = (1-\kappa)\cdot G + \kappa \cdot H$.

Given a Borel probability measure Q on $\mathcal{X}$, we let Q^n denote the corresponding product measure on $\mathcal{X}^n$. Equivalently, Q^n denotes the probability measure over random samples $\{X^i\}_{i\in[n]}$ where X_i are sampled i.i.d. from Q. Let $\hat{Q}_n$ denote the corresponding empirical measure defined for Borel sets $S \subset \mathcal{X}$, that is $\hat{Q}_n(S) \doteq \frac{1}{n}\sum_{i\in[n]} \mathbb{1}\left\{X^i \in S\right\}$.

Definition 1 (Uniform deviation) *Given a set Ω of real-valued Borel functions on $\mathcal{X}$, we define for each $\delta \in (0,1)$ and $n \in \mathbb{N}$,*

$$\epsilon_{uni}(\Omega, \delta, n) \doteq \inf\left\{\epsilon > 0 : Q^n\left[\sup_{\omega\in\Omega}\left\{\left|\int \omega d\hat{Q}_n - \int \omega dQ\right|\right\} > \epsilon\right] < \delta/2\right\}.$$

Theorem 2 (Generalised Blanchard-Lee-Scott estimators) *Let $(\Omega_q)_{q\in\mathbb{N}}$ be a sequence of disjoint classes of Borel functions, along with a sequence $(\delta_q)_{q\in\mathbb{N}} \subset (0,1)$. Suppose that $\epsilon_* : \mathbb{N}\times\mathbb{N} \to (0,\infty)$ is a function such that for every $q, n \in \mathbb{N}$ we have $\epsilon_{uni}(\Omega_q, \delta_q, n) \leq \epsilon_*(q,n)$. Fix $\mu \in [0,1], \nu \in (0,1]$ and $q_* \in \mathbb{N}$. Suppose we have two i.i.d. samples $\{X_0^i\}_{i\in[n_0]} \sim H^{n_0}$ and $\{X_1^i\}_{i\in[n_1]} \sim F^{n_1}$ and let $\Omega_* \subseteq [0,\infty)^{\mathcal{X}} \cap \bigcup_{q\in\mathbb{N}} \Omega_q$ be a random (possibly data dependent) subset of $[0,\infty)^{\mathcal{X}} \cap \bigcup_{q\in\mathbb{N}} \Omega_q$. Suppose that with probability at least $1-\delta_*$ there exists $\omega_* \in \Omega_* \cap \Omega_{q_*}$ that satisfies $\int \omega_* dG \leq \mu$ and $\int \omega_* dH \geq \nu$. We define the following estimator:*

$$\hat{\kappa}_{GBLS} \doteq \min\left\{\inf_{q\in\mathbb{N}}\left\{\inf_{\omega\in\Omega_*\cap\Omega_q}\left\{\frac{\int \omega d\hat{F}_{n_1} + \epsilon_*(q,n_1)}{\left(\int \omega d\hat{H}_{n_0} - \epsilon_*(q,n_0)\right)_+}\right\}\right\}, 1\right\}.$$

Then, with probability at least $1 - \delta_* - \sum_{q\in\mathbb{N}}\delta_q$ we have

$$\kappa \leq \hat{\kappa}_{GBLS} \leq \kappa + \frac{\mu}{\nu} + \left(1 + \frac{\mu}{\nu}\right)\cdot\frac{8}{\nu}\cdot\max\left\{\epsilon_*(q_*, n_0), \epsilon_*(q_*, n_1)\right\}.$$

To use this theorem, one needs to choose the space $\mathcal{X}$, specify the (sequence of) function class(es) $(\Omega_q)_{q\in\mathbb{N}}$ together with a convergent sequence of failure probabilities $(\delta_q)_{q\in\mathbb{N}}$ that allow uniform deviation bounds $\epsilon_*(q,n)$ and then choose Ω_*. In contexts where functions are represented through a sample, we may specify a nonzero value for δ_*.

Before giving the proof, let us illustrate the working of Theorem 2 by deducing from it a version of Theorem 1 that shows that it exhibits graceful degradation with the violation of its condition. To this end, consider the following relaxation of Scott's irreducibility condition (2):

Irreducibility condition 3 *We say that (G,H) satisfies the relaxed irreducibility condition with constants $\mu \in [0,1], \nu \in (0,1]$, if there exists an open ball $B \in \mathcal{B}_d$ such that $G(B) \leq \mu$ and $H(B) \geq \nu$.*

Corollary 3 (to Theorem 2) *Suppose that (G, H) satisfies the relaxed irreducibility condition (3), with constants μ, ν. Then, with probability at least $1 - \delta$, we have*

$$\kappa \leq \hat{\kappa}_{Scott} \leq \kappa + \frac{\mu}{\nu} + \left(1 + \frac{\mu}{\nu}\right) \cdot \frac{8}{\nu} \cdot \epsilon_{VC}\left(d + 1, \delta, \min\{n_0, n_1\}\right).$$

Proof [Proof of Corollary 3] We let $\Omega_1 = \Omega_*$ be $\mathcal{B}_d$, the set of indicator functions for open balls in $\mathbb{R}^d$. This set has VC dimension $d + 1$ (Dudley, 1979). For $q > 1$ we let $\Omega_q = \emptyset$ and $\delta_q = 0$. Hence, we may take $\epsilon_*(q, n) \doteq \epsilon_{\text{VC}}(d + 1, \delta, n)$. By standard results in VC theory we have $\epsilon_{\text{uni}}(\Omega_1, \delta_1, n) \leq \epsilon_*(1, n)$ for all $\omega \in \Omega_*$ (Mohri et al., 2012, Chapter 3). Let $\delta_* = 0$. Under these conditions $\hat{\kappa}_{\text{Scott}} = \hat{\kappa}_{\text{GBLS}}$. Moreover, the relaxed irreducibility condition (3) entails that the assumptions of Theorem 2 apply. This completes the proof. ■

2.1. Proof of Theorem 2

To prove Theorem 2 we require the following lemma.

Lemma 4 *Suppose that we have functions $f, g, h : \mathcal{W} \to [0, \infty)$ such that for some $\kappa \in [0, 1]$ we have $f(w) = (1 - \kappa) \cdot g(w) + \kappa \cdot h(w)$. Suppose further that there exits error functions $\epsilon_f, \epsilon_h : \mathcal{W} \to (0, \infty)$ along with approximation functions $\hat{f}, \hat{h} : \mathcal{W} \to [0, 1]$ such that for all $w \in \mathcal{W}$ we have*

$$\left|\hat{f}(w) - f(w)\right| \leq \epsilon_f(w); \qquad \left|\hat{h}(w) - h(w)\right| \leq \epsilon_h(w).$$

In addition we define,

$$\hat{\kappa}\left(\hat{f}, \hat{h}\right) \doteq \min\left\{\inf_{w \in \mathcal{W}}\left\{\frac{\hat{f}(w) + \epsilon_f(w)}{\left(\hat{h}(w) - \epsilon_h(w)\right)_+}\right\}, 1\right\}.$$

Then letting $\epsilon(w) = \max\{\epsilon_f(w), \epsilon_h(w)\}$,

$$\kappa \leq \hat{\kappa}\left(\hat{f}, \hat{h}\right) \leq \inf_{w \in \mathcal{W}}\left\{\frac{f(w)}{h(w)} + \max\left\{1, \frac{f(w)}{h(w)}\right\} \cdot \frac{8\epsilon(w)}{h(w)}\right\}.$$

Proof We begin by showing that $\kappa \leq \hat{\kappa}\left(\hat{f}, \hat{h}\right)$. Take $w \in \mathcal{W}$. If $\hat{h}(w) > \epsilon_h(w)$ then

$$\frac{\hat{f}(w) + \epsilon_f(w)}{\left(\hat{h}(w) - \epsilon_h(w)\right)_+} = \frac{\hat{f}(w) + \epsilon_f(w)}{\hat{h}(w) - \epsilon_h(w)} \geq \frac{f(w)}{h(w)} = \kappa + (1 - \kappa) \cdot \frac{g(w)}{h(w)} \geq \kappa.$$

On the other hand, if $\hat{h}(w) \leq \epsilon_h(w)$ then

$$\frac{\hat{f}(w) + \epsilon_f(w)}{\left(\hat{h}(w) - \epsilon_h(w)\right)_+} = +\infty \geq \kappa.$$

Hence, given that $\kappa \leq 1$ we have

$$\hat{\kappa}\left(\hat{f},\hat{h}\right) \doteq \min\left\{\inf_{w\in\mathcal{W}}\left\{\frac{\hat{f}(w)+\epsilon_f(w)}{\left(\hat{h}(w)-\epsilon_h(w)\right)_+}\right\},1\right\} \geq \kappa.$$

To prove the second inequality we first fix $w \in \mathcal{W}$. Suppose first that $h(w) \leq 4\cdot\epsilon(w)$. Then we have

$$\hat{\kappa} \leq 1 \leq \frac{4\epsilon(w)}{h(w)} \leq \frac{f(w)}{h(w)} + \max\left\{1,\frac{f(w)}{h(w)}\right\}\cdot\frac{8\epsilon(w)}{h(w)}.$$

since $f(w), h(w) \geq 0$. Now suppose $h(w) > 4\cdot\epsilon(w)$. Then since $|\hat{h}(w)-h(w)| < \epsilon_h(w) \leq \epsilon(w)$ we must have

$$\hat{h}(w)-\epsilon_h(w) \geq h(w) - 2\cdot\epsilon(w) \geq \frac{h(w)}{2} > 0.$$

Thus, we have

$$\begin{aligned}
\hat{\kappa} &\leq \frac{\hat{f}(w)+\epsilon_f(w)}{\left(\hat{h}(w)-\epsilon_h(w)\right)_+} = \frac{\hat{f}(w)+\epsilon_f(w)}{\hat{h}(w)-\epsilon_h(w)} \\
&= \frac{f(w)}{h(w)} + \frac{\hat{f}(w)+\epsilon_f(w)-f(w)}{\hat{h}(w)-\epsilon_h(w)} + \frac{f(w)}{h(w)}\cdot\left(\frac{h(w)}{\hat{h}(w)-\epsilon_h(w)}-1\right) \\
&\leq \frac{f(w)}{h(w)} + \max\left\{1,\frac{f(w)}{h(w)}\right\}\cdot\frac{8\epsilon(w)}{h(w)}.
\end{aligned}$$

Taking the infimum over all $w \in \mathcal{W}$ completes the proof of the lemma. ■

We shall now prove Theorem 2 by deploying Lemma 4.

Proof [Proof of Theorem 2] By Definition 1 we see that for each $q \in \mathbb{N}$ with probability at least $1-\delta_q$ the following holds uniformly for all $\omega \in \Omega_q$

$$\begin{aligned}
\left|\int \omega d\hat{H}_{n_0} - \int \omega dH\right| &\leq \epsilon_{\text{uni}}\left(\Omega_q,\delta_q,n_0\right) \leq \epsilon_*(q,n_0) \\
\left|\int \omega d\hat{F}_{n_1} - \int \omega dF\right| &\leq \epsilon_{\text{uni}}\left(\Omega_q,\delta_q,n_1\right) \leq \epsilon_*(q,n_1).
\end{aligned} \tag{1}$$

Let's assume that (1) holds uniformly over all $\omega \in \Omega_* \subseteq \bigcup_{q\in\mathbb{N}}\Omega_q$, and $\omega_* \in \Omega_* \cap \Omega_{q_*}$ satisfies $\int \omega_* dG \leq \mu$ and $\int \omega_* dH \geq \nu$. By the union bound this holds with probability at least $1-\delta_*-\sum_{q\in\mathbb{N}}\delta_q$. We can then apply Lemma 4 with $\mathcal{W}=\Omega_*$, $f(\omega)=\int\omega dF$, $g(\omega)=\int\omega dG$, $h(\omega)=\int\omega dH$ and $\hat{f}(\omega)=\int\omega d\hat{F}_{n_1}$, $\hat{h}(\omega)=\int\omega d\hat{H}_{n_0}$. In addition for $\omega\in\Omega_q$, $\epsilon_h(\omega)=\epsilon_*(q,n_0)$, $\epsilon_f(\omega)=\epsilon_*(q,n_1)$. Note that ϵ_h, ϵ_f are well defined since the classes $\{\Omega_q\}_{q\in\mathbb{N}}$ are pair-wise disjoint. It follows that the conditions of Lemma 4 are satisfied with $\hat{\kappa}_{\text{GBLS}} = \hat{\kappa}(\hat{f},\hat{h})$. Hence, it follows from Lemma 4 that

$$\kappa \leq \hat{\kappa}_{\text{GBLS}} \leq \inf_{q\in\mathbb{N}}\left\{\inf_{\omega\in\Omega_*\cap\Omega_q}\left\{\frac{\int\omega dF}{\int\omega dH} + \max\left\{1,\frac{\int\omega dF}{\int\omega dH}\right\}\cdot\frac{8\max\{\epsilon_*(q,n_0),\epsilon_*(q,n_1)\}}{\int\omega dH}\right\}\right\}. \tag{2}$$

Moreover $\omega_* \in \Omega_* \cap \Omega_{q_*}$ satisfies $\int \omega_* dG \leq \mu$ and $\int \omega_* dH \geq \nu$. It follows that

$$\frac{\int \omega_* dF}{\int \omega_* dH} = \kappa + (1-\kappa) \cdot \frac{\int \omega_* dG}{\int \omega_* dH} \leq 1 + \frac{\mu}{\nu}.$$

Plugging this into eq. (2), with q_* selected for q and ω_* for ω, we have w.p. $1 - \delta_* - \sum_{q \in \mathbb{N}} \delta_q$ that:

$$\begin{aligned}\kappa \leq \hat{\kappa}_{\text{GBLS}} &\leq \frac{\int \omega_* dF}{\int \omega_* dH} + \max\left\{1, \frac{\int \omega_* dF}{\int \omega_* dH}\right\} \cdot \frac{8 \max\{\epsilon_*(q_*, n_0), \epsilon_*(q_*, n_1)\}}{\int \omega_* dH} \\ &\leq \kappa + \frac{\mu}{\nu} + \left(1 + \frac{\mu}{\nu}\right) \cdot \frac{8}{\nu} \cdot \max\left\{\epsilon_*(q_*, n_0), \epsilon_*(q_*, n_1)\right\}.\end{aligned}$$

■

The next sections instantiate Theorem 2 in more interesting scenarios, where we obtain new estimators that exploit favourable geometry. To this end we introduce the following definition.

Irreducibility condition 4 *We say that (G, H) satisfies the relaxed irreducibility condition with margin, if there exist $\mu \in [0,1], \nu, \lambda \in (0,1]$, and $x_0 \in \mathcal{X}$, $r_0 > 0$ such that $G\left(B\left(x_0, r_0 + \lambda\right)\right) \leq \mu$ and $H\left(B\left(x_0, r_0\right)\right) \geq \nu$. If so, we say the ordered pair $(x_0, r_0) \in \mathcal{X} \times (0, \infty)$* witnesses *the relaxed irreducibility condition with margin.*

We point out that the newly introduced parameter λ, a latent margin, is crucial for our purposes, while μ is inessential and can be simply put to 0 throughout if one is only interested in consistent estimators. In addition, μ will facilitate the link to Ramaswamy et al. (2016) in Section 5.

3. Example: MPE on a metric space with finite covering dimension

Consider a metric space $(\mathcal{X}, \rho)$ with finite diameter $\text{diam}(\mathcal{X}) < \infty$ with finite covering dimension i.e. there exists constants $\dim_{\text{cov}}(\mathcal{X})$, $C_{\mathcal{X}} > 0$ such that for all $\epsilon > 0$ we have,

$$M_\rho(\mathcal{X}, \epsilon) \leq C_{\mathcal{X}} \cdot \epsilon^{-\dim_{\text{cov}}(\mathcal{X})}, \tag{3}$$

where $M_\rho(\mathcal{X}, \epsilon) \doteq \min\left\{m \in \mathbb{N} : \exists \{x_j\}_{j \in [m]} \text{ such that } \mathcal{X} \subset \bigcup_{j \in [m]} B_\rho(x_j, \epsilon)\right\}$.

Note that the assumption of finite covering dimension (3) is strictly weaker than the doubling assumption which has received a lot of interest in statistical learning (Gottlieb and Kontorovich (2014)). Examples of metric spaces which are doubling (and hence satisfy eq. (3)) include data whose natural metric is the edit distance, or the earth-mover distance. In this section we obtain the first MPE for general metric data. As we shall see, it exhibits the same convergence rate as Scott's estimator did for Euclidean data, but with the covering dimension $\dim_{\text{cov}}(\mathcal{X})$ in place of the Euclidean dimension d. In particular, this means that if F is known to be supported on a subset $\mathcal{X} \subset \mathbb{R}^d$ with low covering dimension $\dim_{\text{cov}}(\mathcal{X}) \ll d$, such as a low-dimensional manifold in a high-dimensional feature space, then the bound takes advantage of this low geometric complexity.

The uniform deviation bound required for applying our Theorem 2 will be obtained by Rademacher analysis of carefully designed function class, as follows. For each $x_0 \in \mathcal{X}$, $r_0 > 0$ define $B_\rho(x_0, r_0) \doteq \{z \in \mathcal{X} : \rho(z, x_0) < r_0\}$ to be the open ball and $g_{x_0, r_0} : \mathcal{X} \to \mathbb{R}$ as the signed

distance to the surface of the metric ball $g_{x_0,r_0}(x) \doteq \rho(x,x_0) - r_0$.
Further, let $\Phi_1 : \mathbb{R} \to [0,1]$ be the following 1-Lipschitz function:

$$\Phi_1(z) \doteq \begin{cases} 1 & \text{if } z \leq 0 \\ 1-z & \text{if } 0 \leq z \leq 1 \\ 0 & \text{if } z \geq 1. \end{cases}$$

Now, for each $q \in \mathbb{N}$ we define

$$\Omega_q \doteq \{\Phi_1(q \cdot g_{x_0,r_0}) : x_0 \in \mathcal{X}, r_0 \in [0, \text{diam}(\mathcal{X})]\}.$$

to form our sequence of function classes. These functions are q-Lipschitz in g, so we are able to bound the Rademacher complexity of each class Ω_q:

Lemma 5 *For each $q \in \mathbb{N}$, $\delta \in (0,1)$, $n \in \mathbb{N}$, we have*

$$\epsilon_{uni}(\Omega_q, \delta, n) \leq 43 \cdot (2C_{\mathcal{X}} diam(\mathcal{X}))^{\frac{1}{\dim_{cov}(\mathcal{X})+1}} \cdot q \cdot \sqrt{\frac{\dim_{cov}(\mathcal{X})+1}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}.$$

The proof uses Talagrand's contraction lemma, and Dudley inequality (see Appendix B).

Remarkably, this deviation bound has a convergence rate of order $n^{-1/2}$ – we note that for learning on metric spaces, previous work only considered the larger class of all q-Lipschitz functions, for which such rate was only obtained in the large sample regime through fat shattering based analysis (Gottlieb and Kontorovich (2014)), which is not applicable here as we need both upper and lower bounds on the uniform deviation in the estimator.

We now construct our estimator $\hat{\kappa}_{\mathcal{X}}$ as follows. For each $\delta \in (0,1)$ and $q, n \in \mathbb{N}$ we define

$$\epsilon_{\mathcal{X}}(q,\delta,n) \doteq 43 \cdot (2C_{\mathcal{X}} \text{diam}(\mathcal{X}))^{\frac{1}{\dim_{\text{cov}}(\mathcal{X})+1}} \cdot q \cdot \sqrt{\frac{\dim_{\text{cov}}(\mathcal{X})+1}{n}} + \sqrt{\frac{q+\log(4/\delta)}{2n}},$$

so that, by Lemma 5 and a union bound we have $\epsilon_{\text{uni}}\left(\Omega_q, \delta \cdot 2^{-q-1}, n\right) \leq \epsilon_{\mathcal{X}}(q,\delta,n)$ for all $q \in \mathbb{N}$. We define $\hat{\kappa}_{\mathcal{X}}$ by

$$\hat{\kappa}_{\mathcal{X}} \doteq \min\left\{ \inf_{x_0 \in \{X_1^j\}_{j\in[n_1]}, r_0 \in [0,\text{diam}(\mathcal{X})], q \in \mathbb{N}} \left\{ \frac{\int \Phi_1(q \cdot g_{x_0,r_0})\, d\hat{F}_{n_1} + \epsilon_{\mathcal{X}}(q,\delta,n_1)}{\left(\int \Phi_1(q \cdot g_{x_0,r_0})\, d\hat{H}_{n_0} - \epsilon_{\mathcal{X}}(q,\delta,n_0)\right)_+} \right\}, 1 \right\}.$$

Combining Lemma 5 with Theorem 2 gives the following bound.

Theorem 6 *Suppose that we have a bounded metric space $(\mathcal{X}, \rho)$ with finite covering dimension (3). Suppose further that (G,H) satisfy the relaxed irreducibility with margin condition (4), with constants $\mu \in [0,1], \nu, \lambda \in (0,1]$ and witnessed by (x_0, r_0). Suppose that $n_1 \geq \log(2/\delta)/F(B(x_0, \lambda/3))$. Then, with probability at least $1-\delta$, we have*

$$\kappa \leq \hat{\kappa}_{GBLS} \leq \kappa + \frac{\mu}{\nu} + \left(1 + \frac{\mu}{\nu}\right) \cdot \frac{8}{\nu} \cdot \epsilon_{\mathcal{X}}(\lceil 3/\lambda \rceil, \delta, \min\{n_0, n_1\}).$$

Proof [Proof of Theorem 6] For each $q \in \mathbb{N}$ we define $\delta_q = \delta \cdot 2^{-q-1}$ and we defined $\epsilon_* : \mathbb{N}^2 \to (0, \infty)$ by $\epsilon_*(q, n) = \epsilon_{\mathcal{X}}(q, \delta, n)$. Lemma 5 implies that for every $\omega \in \Omega_q$, $\epsilon_{\text{uni}}(\Omega_q, \delta_q, n) \leq \epsilon_*(q, n)$. To complete the proof we must show that it suffices to consider functions parameterised by points within the sample $\left\{X_1^j\right\}_{j \in [n_1]}$. To see this, let $\delta_* = \delta/2$, and let $\Omega_* \subset \bigcup_{q \in \mathbb{N}} \Omega_q$ be the set

$$\Omega_* \doteq \left\{\Phi_1\left(q \cdot g_{X_1^j, r_0}\right) : x_0 \in \mathcal{X}, r_0 \in [0, \text{diam}(\mathcal{X})], q \in \mathbb{N}, j \in [n_1]\right\}$$

Now take $x_0 \in \mathcal{X}$, $r_0 > 0$ such that $G\left(B\left(x_0, r_0 + \lambda\right)\right) \leq \mu$ and $H\left(B\left(x_0, r_0\right)\right) \geq \nu$ and $n_1 \geq \log(2/\delta)/F\left(B(x_0, \lambda/3)\right)$. If $\mu = 1$ the theorem is trivial, so we may assume that $\mu < 1$, which in turn implies that $r_0 + \lambda < \text{diam}(\mathcal{X})$ since $G\left(B\left(x_0, r_0 + \lambda\right)\right) \leq \mu$. It follows that with probability at least $1 - \delta/2 = 1 - \delta_*$, that for some $j \in [n_1]$ we have $X_1^j \in B(x_0, \lambda/3)$. This in turn implies that $G\left(B\left(X_1^j, (r_0 + \lambda/3) + \lambda/3\right)\right) \leq \mu$ and $H\left(B\left(X_1^j, r_0 + \lambda/3\right)\right) \geq \nu$. Hence, $\omega_* = \Phi_1\left(\lceil 3/\lambda \rceil \cdot g_{X_1^j, r_0 + \lambda/3}\right)$ satisfies $\int \omega_* dG \leq \mu$ and $\int \omega_* dH \geq \nu$. Thus, we may deduce the result from Theorem 2. ■

4. Example: MPE with random projections

For this section, we consider the special case of $\mathcal{X} \subseteq \mathbb{R}^d$, where $\mathcal{X}$ need not be bounded. We develop a compressive approach to MPE, whereby the data is only assumed to be available in randomly projected form. Random projections (RP) are a powerful tool for overcoming the computational challenges of high dimensional data, with a natural ability to exploit fortuitous geometry. In the same spirit as the covering dimension replaced the ambient dimension d in the previous section, here the squared Gaussian width Liaw et al. (2017) of an appropriate bounded subset of $\mathcal{X}$ will play this role. Moreover, the estimator in this section can work without prior knowledge of this quantity. For a set T the Gaussian width is defined as:

$$w\left(T\right) \doteq \mathbb{E}_{g \sim N(0, I_d)}\left[\sup_{x \in T}\left\{\langle g, x\rangle\right\}\right].$$

We also let $\text{rad}\left(T\right) \doteq \sup_{x \in T}\left\{\|x\|_2\right\}$ denote the radius of T.

Recall that a random variable Z is said to be sub-Gaussian if it has finite Orlicz norm:

$$||Z||_{\psi_2} \doteq \inf\left\{K > 0 : \mathbb{E}\left[\exp\left(|Z|^2/K^2\right)\right] \leq 2\right\}.$$

The Orlicz norm of a random vector $W \in \mathbb{R}^d$ is $||W||_{\psi_2} \doteq \sup\left\{||\langle W, u\rangle||_{\psi_2} : u \in S^{d-1}\right\}$, and for random matrices $M \in \mathbb{R}^{k \times d}$ it is defined as $||M||_{\psi_2} \doteq \max_{i \in [k]}\left\{\left|\left|M_{i:}^T\right|\right|_{\psi_2}\right\}$, where $M_{i:}$ denotes the ith row of M. A random matrix $M \in \mathbb{R}^{k \times d}$ is said to be isotropic if every row $M_{i:}$ of M satisfies $\mathbb{E}\left[M_{i:}^T M_{i:}\right] = I_d$ and $\left|\left|M_{i:}^T\right|\right|_{\psi_2} \leq K$.

Conditional on the RP matrix $M \in \mathbb{R}^{k \times d}$, we have two i.i.d. compressive samples :

$$M(X_0^1), \cdots, M(X_0^{n_0}) \overset{\text{i.i.d.}}{\sim} H \circ M^{-1}; \qquad M(X_1^1), \cdots, M(X_1^{n_1}) \overset{\text{i.i.d.}}{\sim} F \circ M^{-1}.$$

With these preliminaries in place, we first give Lemma 8, which shows that the relaxed irreducibility with margin condition (4) in the original space ensures the relaxed irreducibility condition (3) in the compressed space. We then construct the compressive MPE in Theorem 9 by an appropriate specialisation of Theorem 2.

Definition 2 (Projection onto a ball) *Given $x \in \mathbb{R}^d$, $r > 0$ we define $\mathfrak{P}_{x,r} : \mathbb{R}^d \to \overline{B(x,r)}$ to be the map from points $z \in \mathbb{R}^d$ to the closest point within $\overline{B(x,r)}$, i.e.*

$$\mathfrak{P}_{x,r}(z) = x + \min\left\{\frac{r}{\|z-x\|_2}, 1\right\} \cdot (z - x)\,.$$

We extend the definition to sets $A \subset \mathbb{R}^d$ by $\mathfrak{P}_{x,r}(A) = \{\mathfrak{P}_{x,r}(z) : z \in A\}$.

We will also need the following result of Liaw et al. Liaw et al. (2017).

Theorem 7 (Liaw et al. (2017)) *There exists a universal constant $C_{Liaw} \geq 1$ such that the following holds. Given any random projection matrix M, which is both isotropic and sub-Gaussian with $||M||_{\psi_2} \leq K$, along with a set $T \subseteq \mathbb{R}^d$ and some $\delta > 0$, with probability at least $1 - \delta$ we have*

$$\sup_{x \in T}\left\{\left|\|Mx\|_2 - \sqrt{k} \cdot \|x\|_2\right|\right\} \leq C_{Liaw} \cdot K^2\left(w(T) + \sqrt{\log(1/\delta)} \cdot rad\,(T)\right).$$

Lemma 8 *Suppose that (G, H) satisfies the relaxed irreducibility with margin condition (4), with constants μ, ν, λ, witnessed by $(x_0, r_0) \in \mathcal{X} \times (0, \infty)$. Suppose that $k \in [d]$ satisfies*

$$k \geq \left(\frac{2}{\lambda} \cdot C_{Liaw} \cdot K^2 \cdot \left(w\left(\mathfrak{P}_{x_0, r_0+\lambda}\left(supp(F)\right)\right) + \sqrt{\log(2/\delta)} \cdot (r_0 + \lambda)\right)\right)^2, \tag{4}$$

where C_{Liaw} is a universal constant. Suppose that $M : \mathbb{R}^d \to \mathbb{R}^k$ is a random projection which is both isotropic and sub-Gaussian with $||M||_{\psi_2} \leq K$. Then, with probability at least $1 - \delta$, $\left(G \circ M^{-1}, H \circ M^{-1}\right)$ satisfies the relaxed irreducibility condition (3), with constants μ, ν.

Proof [Proof of Lemma 8] Firstly, since (G, H) satisfies the relaxed irreducibility with margin condition (4), with constants μ, ν, λ, witnessed by $(x_0, r_0) \in \mathcal{X} \times (0, \infty)$, we have $G\left(B_d\left(x_0, r_0 + \lambda\right)\right) \leq \mu$ and $H\left(B_d\left(x_0, r_0\right)\right) \geq \nu$.

Let $T \doteq \{z = y - x_0 : y \in \mathfrak{P}_{x_0, r_0+\lambda}\left(\text{supp}(F)\right)\}$ so that $w(T) = w\left(\mathfrak{P}_{x_0, r_0+\lambda}\left(\text{supp}(F)\right)\right)$ and $\text{rad}\,(T) \leq r_0 + \lambda$. Hence, by Liaw's theorem (Theorem 7) with probability at least $1 - \delta$ over M we have

$$\begin{aligned}\sup_{z \in \mathfrak{P}_{x_0, r_0+\lambda}(\text{supp}(F))}\left\{\left|\|Mz - Mx_0\|_2 - \sqrt{k} \cdot \|z - x_0\|_2\right|\right\} &= \sup_{z \in T}\left\{\left|\|Mz\|_2 - \sqrt{k} \cdot \|z\|_2\right|\right\} \\ &\leq C_{\text{Liaw}} \cdot K^2\left(w(T) + \sqrt{\log(1/\delta)} \cdot \text{rad}\,(T)\right) \leq \sqrt{k} \cdot \frac{\lambda}{2},\end{aligned} \tag{5}$$

whenever k satisfies eq. (4).

Let us suppose that eq. (5) holds and take $x_1 = M(x_0)$ and $r_1 = \sqrt{k} \cdot \left(r_0 + \frac{\lambda}{2}\right)$. Now suppose $z \in M^{-1}\left(B_k(x_1, r_1)\right) \cap \text{supp}(G)$, so $M(z) \in B_k(x_1, r_1)$. Now, take $z' = \mathfrak{P}_{x_0, r_0+\lambda}(z) \in \mathfrak{P}_{x_0, r_0}(\text{supp}(F))$ and we have:

$$\begin{aligned}\sqrt{k} \cdot \|z' - x_0\|_2 &\leq \|Mz' - Mx_0\|_2 + \sqrt{k} \cdot \frac{\lambda}{2} \\ &\leq \|Mz - Mx_0\|_2 + \sqrt{k} \cdot \frac{\lambda}{2} < \sqrt{k} \cdot (r_0 + \lambda),\end{aligned}$$

where the second inequality uses the fact that $z' - x_0 = c \cdot (z - x_0)$ for some $c \in [0, 1]$ and the third inequality uses the definition of r_1. It follows that $z' = \mathfrak{P}_{x_0, r_0+\lambda}(z) \in B_d(x_0, r_0 + \lambda)$, so $z \in B_d(x_0, r_0 + \lambda)$. Hence, we have

$$G\left(M^{-1}\left(B(x_1, r_1)\right)\right) \leq G\left(B(x_0, r_0 + \lambda)\right) \leq \mu.$$

To show the condition on the H measure, take $z \in B(x_0, r_0) \cap \text{supp}(H) \subset \mathfrak{P}_{x, r_0+\lambda}(\text{supp}(F))$, then we have

$$\|Mz - Mx_0\|_2 \leq \sqrt{k} \cdot \|z - x_0\|_2 + \sqrt{k} \cdot \frac{\lambda}{2} < \sqrt{k} \cdot \left(r_0 + \frac{\lambda}{2}\right) = r_1.$$

Thus, $H\left(M^{-1}\left(B_k(x_1, r_1)\right)\right) \geq H\left(B_d(x_0, r_0)\right) \geq \nu$. Thus, provided (5) holds, $G \circ M^{-1}$ satisfies the relaxed irreducibility condition (3) with respect to $H \circ M^{-1}$, with constants μ, ν. Since (5) holds with probability at least $1 - \delta$ this completes the proof. ■

We are now ready to construct our compressive MPE. Consider an ensemble $\mathcal{M} = (M_k)_{k \in [d]}$ consisting of random projection matrices $M_k : \mathbb{R}^d \to \mathbb{R}^k$. We assume that for some $K > 0$ each random matrix M_k is both isotropic and sub-Gaussian with $||M_k||_{\psi_2} \leq K$. We define a random ensemble estimator by

$$\hat{\kappa}_{\text{RPE}}(\mathcal{M}) \doteq \min\left\{\min_{k \in [d]}\left\{\inf_{B \in \mathcal{B}_k}\left\{\frac{\hat{F} \circ (M_k)^{-1}(B) + \epsilon_{\text{VC}}(k+1, \delta \cdot 2^{-k-1}, n_1)}{(\hat{H} \circ (M_k)^{-1}(B) - \epsilon_{\text{VC}}(k+1, \delta \cdot 2^{-k-1}, n_0))_+}\right\}\right\}, 1\right\}.$$

Theorem 9 *Suppose that (G, H) satisfies the relaxed irreducibility with margin condition (4), with constants $\mu, \nu, \lambda \in [0, 1]$. We let $\mathcal{M} = (M_k)_{k \in [d]}$ where $M_k : \mathbb{R}^d \to \mathbb{R}^k$ is isotropic and sub-Gaussian with $||M_k||_{\psi_2} \leq K$, where $K \geq 1$. Then, with probability at least $1 - \delta$, we have*

$$\begin{aligned}\kappa \leq \hat{\kappa}_{RPE}(\mathcal{M}) \leq \kappa + \frac{\mu}{\nu} + \left(1 + \frac{\mu}{\nu}\right) \cdot \frac{64}{\nu \cdot \lambda} \cdot C_{Liaw} \cdot K^2 \cdots \\ \cdot \left(w\left(\mathfrak{P}_{x_0, r+\lambda}(supp(F))\right) + 2\sqrt{\log(4/\delta)} \cdot (r_0 + \lambda + 1)\right) \cdot \sqrt{\frac{\log(2e \min\{n_0, n_1\})}{\min\{n_0, n_1\}}}.\end{aligned}$$

Proof [Proof of Theorem 9] For each $k \in [d]$ we let

$$\Omega_k = \{x \mapsto \mathbb{1}\{M_k(x) \in B\} : B \in \mathcal{B}_k\}.$$

Observe that the sets Ω_k depend upon the random projections M_k, but not the data. Let k^* be as in eq. (4), then by Lemma 8 with probability at least $1 - \delta/2$, $G \circ (M_{k_*})^{-1}$ satisfies the relaxed

irreducibility condition (3) with respect to $H \circ (M_{k_*})^{-1}$, with constants μ, ν. Thus, with probability at least $1 - \delta/2$ there exists $\omega_* \in \Omega_{k_*}$ such that $\int \omega_* dG \leq \mu$ and $\int \omega_* dH \geq \nu$. Let us assume that such an $\omega_* \in \Omega_{k_*}$ exists.

For each $k \in [d]$ we let $\delta_k = \delta \cdot 2^{-k-1}$ and for $k > d$ we let $\delta_k = 0$ and $\Omega_k = \emptyset$. Thus, we have

$$\epsilon_* (k, n) = \begin{cases} \epsilon_{\text{VC}} \left(k + 1, \delta \cdot 2^{-k-1}, n\right) & \text{for } k \in [d] \\ 0 & \text{for } k > d. \end{cases} \tag{6}$$

For each $k \in [d]$, the set Ω_k has VC dimension $k + 1$ Dudley (1979), so

$$\epsilon_* (k, n) = \epsilon_{\text{VC}} \left(k + 1, \delta \cdot 2^{-k-1}, n\right) \geq \epsilon_{\text{uni}}(\Omega_k, \delta_k, n).$$

For $k > d$, $\Omega_k = \emptyset$, so $\epsilon_* (k, n) \geq \epsilon_{\text{uni}}(\Omega_k, \delta_k, n)$ holds vacuously. Observe that with these choices of $(\Omega_k)_{k \in \mathbb{N}}$, $(\delta_k)_{k \in \mathbb{N}}$ and ϵ_*, $\hat{\kappa}_{\text{GBLS}}$ is equal to $\hat{\kappa}_{\text{RPE}}$.

We let $\Omega_* = \bigcup_{k \in \mathbb{N}} \Omega_k$ and $\delta_* = 0$. Thus, by Theorem 2, provided $\omega_* \in \Omega_{k_*}$ exists, with probability at least $1 - \sum_{k \in \mathbb{N}} \delta_k \geq 1 - \delta/2$ we have

$$\begin{aligned} \kappa \leq \hat{\kappa}_{\text{RPE}} &\leq \kappa + \frac{\mu}{\nu} + \left(1 + \frac{\mu}{\nu}\right) \cdot \frac{8}{\nu} \cdot \max \left\{\epsilon_*(k_*, n_0), \epsilon_*(k_*, n_1)\right\} \\ &\leq \kappa + \frac{\mu}{\nu} + \left(1 + \frac{\mu}{\nu}\right) \cdot \frac{8}{\nu} \cdot \epsilon_*(k_*, n_*), \end{aligned}$$

where $n_* = \min\{n_0, n_1\}$. Thus, under the assumption that $\omega_* \in \Omega_{k_*}$ exists, w.p. $1 - \delta/2$, we have

$$\begin{aligned} \epsilon_*(k_*, n_*) &= \epsilon_{\text{VC}} \left(k_* + 1, \delta \cdot 2^{-k_*-1}, n_*\right) \qquad (7) \\ &\leq \frac{8}{\lambda} \cdot \sqrt{\frac{\log(2en_*)}{n_*}} \cdot C_{\text{Liaw}} \cdot K^2 \cdot \left(w\left(\mathfrak{P}_{x_0, r+\lambda}\left(\text{supp}(F)\right)\right) + 2\sqrt{\log(4/\delta)} \cdot (r_0 + \lambda + 1)\right), \end{aligned}$$

where the inequality follows from plugging in eq. (4) for k_*. Moreover, the existence of $\omega_* \in \Omega_{k_*}$ with $\int \omega_* dG \leq \mu$ and $\int \omega_* dH \geq \nu$ holds with probability at least $1 - \delta/2$. Thus, by the union bound (7) holds w.p. $1 - \delta$. Substituting back in to the bound on $\hat{\kappa}_{\text{RPE}}$ above completes the proof. ■

Finally, for a known value of k that satisfies eq. (4), the estimator simplifies to Scott's estimator, applied to the randomly projected data:

Corollary 10 *Suppose that (G, H) satisfies the relaxed irreducibility with margin condition (4), with constants μ, ν, λ, witnessed by $(x_0, r_0) \in \mathcal{X} \times (0, \infty)$. Fix k so that eq. (4) holds, and let $M : \mathbb{R}^d \to \mathbb{R}^k$ be a random projection which is both isotropic and sub-Gaussian with $||M||_{\psi_2} \leq K$. Let*

$$\hat{\kappa}_{RP}(M) \doteq \min \left\{ \inf_{B \in \mathcal{B}_k} \left\{ \frac{\hat{F}_{n_1} \circ M^{-1}(B) + \epsilon_{VC}(k + 1, \delta/2, n_1)}{\left(\hat{H}_{n_0} \circ M^{-1}(B) - \epsilon_{VC}(k + 1, \delta/2, n_0)\right)_+} \right\}, 1 \right\}.$$

Then, with probability at least $1 - \delta$, we have

$$\kappa \leq \hat{\kappa}_{RP}(M) \leq \kappa + \frac{\mu}{\nu} + \left(1 + \frac{\mu}{\nu}\right) \cdot \frac{8}{\nu} \cdot \epsilon_{VC}\left(k + 1, \delta/2, \min\{n_0, n_1\}\right).$$

5. BLS-type methods under Ramaswamy's Condition

We conclude by demonstrating a theoretical link between the BLS framework and the existing state of the art approach of Ramaswamy et al. (2016) specialised to reproducing kernel Hilbert spaces. Specifically, we derive a BLS-type estimator under the assumption of Ramaswamy et al. (2016), achieving the same guarantees as the specialised method. We refer to our estimator as being of BLS-type since, whilst it isn't quite a generalised BLS estimator (Section 2), it is similar in spirit.

In this section we assume the existence of a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which is both continuous and positive semi-definite. We let $\mathcal{H}_k$ denote the associated reproducing kernel Hilbert space consisting of continuous real valued functions and let $\|\cdot\|_{\mathcal{H}_k}$ denote the corresponding norm (see (Mohri et al., 2012, Chapter 5)). In this setting we shall consider the following irreducibility condition introduced by Ramaswamy et al. (2016).

Irreducibility condition 5 *We say that (G, H) satisfies the relaxed Hilbert space irreducibility condition with kernel k and constants $\alpha > 0, \beta \geq 0$ if there exists $h \in \mathcal{H}_k$ with $\|h\|_{\mathcal{H}_k} \leq 1$ such that*

$$\int h dG \leq \inf_{x \in \mathcal{X}} \{h(x)\} + \beta \leq \int h dH - \alpha.$$

For each $(i_1, i_2) \in \{0, 1\}^2$ we define a matrix $K_{i_a i_b} \in \mathbb{R}^{n_{i_a} \times n_{i_b}}$ by

$$K_{i_a i_b} = \left(k\left(X_{i_a}^{j_a}, X_{i_b}^{j_b}\right)\right)_{(j_a, j_b) \in [n_{i_a}] \times [n_{i_b}]}.$$

In addition we define a matrix $K \in \mathbb{R}^{(n_0+n_1)\times(n_0+n_1)}$ by

$$\mathbf{K} = \begin{bmatrix} K_{00} & K_{01} \\ K_{10} & K_{11} \end{bmatrix}.$$

In addition, we let $1_n \in \mathbb{R}^n$ denote an $n \times 1$ vector consisting entirely of 1s. Let us suppose that for all $x \in \mathcal{X}$ we have $k(x, x) \leq B^2$. We define an estimator $\hat{\kappa}_{\text{HS}}$ for κ as follows. Firstly, we let $\xi(\delta) = 2 \sup_{x \in \mathcal{X}} \left\{\sqrt{k(x, x)}\right\} + \sqrt{\log(10/\delta)/2}$, and let $\mathcal{A}(\boldsymbol{K}) \subset \mathbb{R}^{n_0+n_1}$ denote the set of vectors $\theta \in \mathbb{R}^{n_0+n_1}$ satisfying both $\theta^T \boldsymbol{K} \theta \leq 1$ and

$$n_0^{-1} 1_{n_0}^T [K_{00}\ K_{01}]\theta - \min(\mathbf{K}\theta) \geq \xi(\delta)/\sqrt{n_0} + (\log n_1)^{-\frac{1}{2}},$$

where $\min(\mathbf{K}\theta)$ denotes the minimum element of the column vector $\mathbf{K}\theta$. Finally, we define $\hat{\kappa}_{\text{HS}}$ by

$$\hat{\kappa}_{\text{HS}} \doteq \min\left\{\inf_{\theta \in \mathcal{A}(K)} \left\{\frac{n_1^{-1} 1_{n_1}^T [K_{10}\ K_{11}]\theta - \min(\mathbf{K}\theta) + \xi(\delta)/\sqrt{n_1}}{n_0^{-1} 1_{n_0}^T [K_{00}\ K_{01}]\theta - \min(\mathbf{K}\theta) - \xi(\delta)/\sqrt{n_0}}\right\}, 1\right\}. \tag{8}$$

Theorem 11 *Suppose that we have a bounded, continuous and positive semi-definite kernel k and (G, H) satisfies the relaxed Hilbert space irreducibility condition (5) with constants $\alpha > 0, \beta \geq 0$. Suppose further that $n_0 \geq \left(\frac{8\xi(\delta)}{\alpha}\right)^2$ and*

$$n_1 \geq \max\left\{8 \log(5/\delta) / \min\left\{(1 - \kappa), \log\left(\frac{\alpha + 2\beta}{\kappa \cdot \alpha + 2\beta}\right)\right\}, \exp(16/\alpha^2)\right\}.$$

Then, with probability at least $1-\delta$ *we have*

$$\kappa - \xi(\delta)\cdot\sqrt{\frac{2\log n_1}{n_1}} \leq \hat{\kappa}_{HS} \leq \kappa + \frac{2\beta}{\alpha} + \left(1+\frac{\beta}{\alpha}\right)\cdot\frac{16\cdot\xi(\delta)}{\alpha\sqrt{\min\{n_0,n_1\}}}. \tag{9}$$

In the interest of space, the proof is deferred to Appendix A. Unlike Theorems 6 and 9, Theorem 11 is not a corollary to Theorem 2, owing to the data dependent minimum in the construction of $\hat{\kappa}_{\text{HS}}$ (9). Two observations are worth noting. Firstly, by comparing the guarantees in (9) with those of the original estimator of Ramaswamy et al. (2016), we see they are of the same order. Secondly, from the form of the obtained estimator, eq. (8) we can see that the parameters α and β play similar roles to those of ν and μ of the previous sections respectively. This suggests that the BLS framework lends itself to further extensions beyond those obtainable by its original irreducibility condition.

A worthwhile avenue for future work is to extend and instantiate the framework to MPE for structured data types such as graphs, trees, through semimetric spaces.

References

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11(Nov):2973–3009, 2010.

Gilles Blanchard, Marek Flaska, Gregory Handy, Sara Pozzi, and Clayton Scott. Classification with asymmetric label noise: Consistency and maximal denoising. *Electron. J. Statist.*, 10(2): 2780–2824, 2016.

R Dudley. Sizes of compact subsets of hilbert space and continuity of gaussian processes. *J. Funct. Anal.*, 1:290–330, 1967.

Richard M Dudley. Balls in rk do not cut all subsets of k+ 2 points. *Advances in Mathematics*, 31 (3):306–308, 1979.

Lee-Ad Gottlieb and Aryeh Kontorovich. Efficient classification for metric data. *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014.

Christopher Liaw, Abbas Mehrabian, Yaniv Plan, and Roman Vershynin. A simple tool for bounding the deviation of random matrices on geometric sets. In *Geometric aspects of functional analysis*, pages 277–299. Springer, 2017.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *International Conference on Machine Learning*, pages 2052–2060, 2016.

Clayton Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Artificial Intelligence and Statistics*, pages 838–846, 2015.

Acknowledgments

This work is funded by EPSRC Grant EP/P004245/1.

Appendix A. Proof of Theorem 11

Proof [Proof of Theorem 11] Firstly, we let $\phi : \mathcal{X} \to \mathcal{H}_k$ denote the canonical embedding given by $\phi(x_1)(x_2) = k(x_1, x_2)$ for $x_1, x_2 \in \mathcal{X}$. We define a data dependent mapping $w : \mathbb{R}^{n_0+n_1} \to \mathcal{H}_k$ by

$$w_{\boldsymbol{K}}(\theta) \doteq \sum_{i\in\{0,1\}} \sum_{j\in[n_i]} \theta_{i\cdot n_0+j} \cdot \phi\left(X_i^j\right).$$

For all $\theta \in \mathbb{R}^{n_0+n_1}$ we have $\|w_{\boldsymbol{K}}(\theta)\|_{\mathcal{H}_k} = \sqrt{\theta^T \boldsymbol{K}\theta}$. We also have

$$\int w_{\boldsymbol{K}}(\theta) d\hat{F}_{n_1} = n_1^{-1} \cdot 1_{n_1}^T [K_{10}\ K_{11}]\theta$$
$$\int w_{\boldsymbol{K}}(\theta) d\hat{H}_{n_0} = n_0^{-1} \cdot 1_{n_0}^T [K_{00}\ K_{01}]\theta.$$

The Rademacher complexity $\mathfrak{R}_n\left(\mathcal{H}_k^1\right)$ of $\mathcal{H}_k^1 \doteq \{h \in \mathcal{H}_k : \|h\|_{\mathcal{H}_k} \leq 1\}$ of an average sample of size n is bounded above by

$$\mathfrak{R}_n\left(\mathcal{H}_k^1\right) \leq \frac{\sup_{x\in\mathcal{X}}\left\{\sqrt{k(x,x)}\right\}}{\sqrt{n}}.$$

See, for example, (Mohri et al., 2012, Theorem 5.5). Hence, it follows standard Rademacher theory (Mohri et al., 2012, Theorem 3.1) with probability at least $1 - \delta/5$ the following holds $h \in \mathcal{H}_k^1$,

$$\left|\int h d\hat{F}_{n_1} - \int h dF\right| \leq \frac{2\sup_{x\in\mathcal{X}}\left\{\sqrt{k(x,x)}\right\}}{\sqrt{n_1}} + \sqrt{\frac{\log(10/\delta)}{n_1}} = \frac{\xi(\delta)}{\sqrt{n_1}}. \tag{10}$$

In particular, for all $\theta \in \mathcal{A}(\boldsymbol{K})$,

$$\left|n_1^{-1} \cdot 1_{n_1}^T [K_{10}\ K_{11}]\theta - \int w_{\boldsymbol{K}}(\theta) dF\right| \leq \frac{\xi(\delta)}{\sqrt{n_1}}.$$

Similarly, with probability of $1 - \delta/5$,

$$\left|\int h d\hat{H}_{n_0} - \int h dH\right| \leq \frac{\xi(\delta)}{\sqrt{n_0}}. \tag{11}$$

In particular, for all $\theta \in \mathcal{A}(\boldsymbol{K})$ we have

$$\left|n_0^{-1} \cdot 1_{n_0}^T [K_{00}\ K_{01}]\theta - \int w_{\boldsymbol{K}}(\theta) dH\right| \leq \frac{\xi(\delta)}{\sqrt{n_0}}.$$

Suppose that we have a sample $\left\{\left(\tilde{X}^j, \tilde{Z}\right)\right\}_{j\in[n_1]}$ generated i.i.d. with each $\tilde{Z}^j \in \{0,1\}$ with $\mathbb{E}\left[\tilde{Z}^j\right] = \kappa$ and

$$\mathbb{P}\left[\tilde{X}^j | \tilde{Z}^j\right] = \begin{cases} G(\tilde{X}^j) & \text{if } \tilde{Z}^j = 0 \\ H(\tilde{X}^j) & \text{if } \tilde{Z}^j = 1. \end{cases}$$

By the multiplicative Chernoff bound, given the assumption that $n_1 \geq 8\log(5/\delta)/(1-\kappa)$, with probability at least $1-\delta/5$ we have $\#\left\{j \in [n_1] : \tilde{Z}^j = 0\right\} \geq n_1(1-\kappa)/2$. Moreover, conditioned on this event then with probability at least $1-\delta/5$, for all $\theta \in \mathcal{A}(\boldsymbol{K})$ we have

$$\begin{aligned}\min\left(\left\{w_{\boldsymbol{K}}(\theta)\left(\tilde{X}^j\right)\right\}_{j\in[n_1]}\right) &\leq \frac{\sum_{j\in[n_1]}\mathbb{1}\left\{\tilde{Z}^j=0\right\}\cdot w_{\boldsymbol{K}}(\theta)\left(\tilde{X}^j\right)}{\sum_{j\in[n_1]}\mathbb{1}\left\{\tilde{Z}^j=0\right\}} \\ &\leq \int w_{\boldsymbol{K}}(\theta)dG + \xi(\delta)\cdot\sqrt{\frac{2}{(1-\kappa)n_1}}.\end{aligned}$$

By the union bound this event holds with probability at least $1-2\delta/5$. Now note that $F = (1-\kappa)\cdot G + \kappa\cdot H$, so $\{X_1^j\}_{j\in[n_1]}$ and $\{\tilde{X}_1^j\}_{j\in[n_1]}$ share the same distribution. Thus, with probability at least $1-2\delta/5$ we have,

$$\begin{aligned}\min(\boldsymbol{K}\theta) &\leq \min\left(\left\{w_{\boldsymbol{K}}(\theta)\left(\tilde{X}^j\right)\right\}_{j\in[n_1]}\right) \\ &\leq \int w_{\boldsymbol{K}}(\theta)dG + \xi(\delta)\cdot\sqrt{\frac{2}{(1-\kappa)n_1}}.\end{aligned} \tag{12}$$

Take $h^* \in \mathcal{H}_k$ with $\|h\|_{\mathcal{H}_k} \leq 1$ and

$$\int h^* dG \leq \inf_{x\in\mathcal{X}}\{h^*(x)\} + \beta \leq \int h^* dH - \alpha.$$

Let $\underline{h}^*(z) = h^*(z) - \inf_{x\in\mathcal{X}}\{h^*(x)\}$. It follows that

$$\begin{aligned}\mathbb{P}_{X\sim F}\left[h^*(X) \geq \inf_{x\in\mathcal{X}}\{h^*(X)\} + \beta + \frac{\alpha}{2}\right] &= \mathbb{P}_{X\sim F}\left[\underline{h}^*(X) \geq \beta + \frac{\alpha}{2}\right] \\ &= 1 - \mathbb{P}_{X\sim F}\left[\underline{h}^*(X) < \beta + \frac{\alpha}{2}\right] \\ &\leq 1 - (1-\kappa)\cdot\mathbb{P}_{X\sim G}\left[\underline{h}^*(X) < \beta + \frac{\alpha}{2}\right] \\ &= \kappa + (1-\kappa)\cdot\mathbb{P}_{X\sim G}\left[\underline{h}^*(X) \geq \beta + \frac{\alpha}{2}\right] \\ &\leq \kappa + (1-\kappa)\cdot\frac{\int h^* dG - \inf_{x\in\mathcal{X}}\{h^*(X)\}}{\beta + \frac{\alpha}{2}} \\ &\leq \frac{\alpha\cdot\kappa + 2\beta}{\alpha + 2\beta}.\end{aligned}$$

Since $n_1 \geq \log(5/\delta)/\log\left(\frac{\alpha+2\beta}{\kappa\cdot\alpha+2\beta}\right)$, with probability at least $1-\delta/5$ we have

$$\min\left(\left\{h^*\left(X_1^j\right)\right\}_{j\in[n_1]}\right) \leq \inf_{x\in\mathcal{X}}\{h^*(X)\} + \beta + \frac{\alpha}{2}. \tag{13}$$

By the union bound, (10), (11), (12), (13) hold simultaneously with probability at least $1-\delta$. Henceforth, we assume that (10), (11), (12), (13) all hold.

By (10), (11), (12), for any $\theta \in \mathcal{A}(\boldsymbol{K})$ we have

$$\frac{n_1^{-1} 1_{n_1}^T [K_{10}\ K_{11}]\theta - \min(\mathbf{K}\theta) + \xi(\delta)/\sqrt{n_1}}{n_0^{-1} 1_{n_0}^T [K_{00}\ K_{01}]\theta - \min(\mathbf{K}\theta) - \xi(\delta)/\sqrt{n_0}}$$

$$\begin{aligned}
&\geq \frac{\int w_{\boldsymbol{K}}(\theta) dF - \min(\mathbf{K}\theta)}{\int w_{\boldsymbol{K}}(\theta) dH - \min(\mathbf{K}\theta)} \\
&\geq \kappa + (1-\kappa) \cdot \frac{\int w_{\boldsymbol{K}}(\theta) dG - \min(\mathbf{K}\theta)}{\int w_{\boldsymbol{K}}(\theta) dH - \min(\mathbf{K}\theta)} \\
&\geq \kappa - (1-\kappa) \cdot \xi(\delta) \cdot \sqrt{\frac{2\log n_1}{(1-\kappa) n_1}} \geq \kappa - \xi(\delta) \cdot \sqrt{\frac{2\log n_1}{n_1}}.
\end{aligned}$$

Hence,

$$\hat{\kappa}_{\text{HS}} \geq \kappa - \xi(\delta) \cdot \sqrt{\frac{2\log n_1}{n_1}}.$$

We now prove the upper bound. We start by following the method of the representer theorem (Mohri et al., 2012, Theorem 5.4) and choose $\theta^* \in \mathbb{R}^{n_0+n_1}$ and $f^* \in \text{span}(\{w_{\boldsymbol{K}}(\theta) : \theta \in \mathbb{R}^{n_0+n_1}\})^\perp$ so that we can write $h^* = w_{\boldsymbol{K}}(\theta^*) + f^*$. Since $f^* \in \text{span}(\{w_{\boldsymbol{K}}(\theta) : \theta \in \mathbb{R}^{n_0+n_1}\})^\perp$ we have $h^*\left(X_i^j\right) = w_{\boldsymbol{K}}(\theta^*)\left(X_i^j\right)$ for $i \in \{0,1\}$, $j \in [n_i]$. Hence,

$$\begin{aligned}
n_1^{-1} 1_{n_1}^T [K_{10}\ K_{11}]\theta^* &= \int h^* d\hat{F}_{n_1} \leq \int h^* dF + \xi(\delta)/\sqrt{n_1} \\
n_0^{-1} 1_{n_0}^T [K_{00}\ K_{01}]\theta^* &= \int h^* d\hat{H}_{n_0} \geq \int h^* dH - \xi(\delta)/\sqrt{n_0}.
\end{aligned}$$

Note also that

$$\begin{aligned}
0 &\leq \min(\boldsymbol{K}\theta^*) - \inf_{x \in \mathcal{X}} \{h^*(x)\} \\
&= \min\left(\left\{h^*\left(X_1^j\right)\right\}_{j \in [n_1]}\right) - \inf_{x \in \mathcal{X}} \{h^*(x)\} \leq \beta + \frac{\alpha}{2}.
\end{aligned}$$

Combining this with the definition of h^* gives

$$\begin{aligned}
\frac{\int h^* dF - \min(\mathbf{K}\theta^*)}{\int h^* dH - \min(\mathbf{K}\theta^*)} &= \kappa + \frac{\int h^* dG - \min(\mathbf{K}\theta^*)}{\int h^* dH - \min(\mathbf{K}\theta^*)} \\
&\leq \kappa + \frac{\int h^* dG - \inf_{x \in \mathcal{X}} \{h^*(x)\}}{(\alpha+\beta) - (\beta + \alpha/2)} \leq \kappa + \frac{2\beta}{\alpha}.
\end{aligned}$$

In addition, given that $n_0 \geq \left(\frac{8\xi(\delta)}{\alpha}\right)^2$ we have

$$\begin{aligned}
&n_0^{-1} 1_{n_0}^T [K_{00}\ K_{01}]\theta^* - \min(\boldsymbol{K}\theta^*) - \xi(\delta)/\sqrt{n_0} \\
&\quad \geq \int h^* dH - 2\xi(\delta)/\sqrt{n_0} - \min(\boldsymbol{K}\theta^*) \\
&\quad \geq \frac{\alpha}{2} - 2\xi(\delta)/\sqrt{n_0} \geq \frac{\alpha}{4}.
\end{aligned}$$

In particular, since $n_1 \geq \exp(16/\alpha^2)$ we have

$$n_0^{-1} 1_{n_0}^T [K_{00}\ K_{01}]\theta - \min(\mathbf{K}\theta) \geq \xi(\delta)/\sqrt{n_0} + (\log n_1)^{-\frac{1}{2}},$$

so $\theta^* \in \mathcal{A}(\boldsymbol{K}^*)$.

Moreover, piecing the above together we have

$$\begin{aligned}
&\frac{n_1^{-1} 1_{n_1}^T [K_{10}\ K_{11}]\theta^* - \min(\mathbf{K}\theta^*) + \xi(\delta)/\sqrt{n_1}}{n_0^{-1} 1_{n_0}^T [K_{00}\ K_{01}]\theta^* - \min(\mathbf{K}\theta^*) - \xi(\delta)/\sqrt{n_0}} \\
&\leq \frac{\int h^* dF - \min(\mathbf{K}\theta^*) + 2\xi(\delta)/\sqrt{n_1}}{\int h^* dH - \min(\mathbf{K}\theta^*) - 2\xi(\delta)/\sqrt{n_0}} \\
&\leq \frac{\int h^* dF - \min(\mathbf{K}\theta^*)}{\int h^* dH - \min(\mathbf{K}\theta^*)} + \left(1 + \frac{\int h^* dF - \min(\mathbf{K}\theta^*)}{\int h^* dH - \min(\mathbf{K}\theta^*)}\right) \cdot \frac{2\xi(\delta)/\sqrt{\min\{n_0, n_1\}}}{\int h^* dH - \min(\mathbf{K}\theta^*) - 2\xi(\delta)/\sqrt{n_0}} \\
&\leq \frac{\int h^* dF - \min(\mathbf{K}\theta^*)}{\int h^* dH - \min(\mathbf{K}\theta^*)} + \left(1 + \frac{\int h^* dF - \min(\mathbf{K}\theta^*)}{\int h^* dH - \min(\mathbf{K}\theta^*)}\right) \cdot \frac{8\xi(\delta)}{\alpha\sqrt{\min\{n_0, n_1\}}} \\
&\leq \kappa + \frac{2\beta}{\alpha} + \left(1 + \frac{\beta}{\alpha}\right) \cdot \frac{16 \cdot \xi(\delta)}{\alpha\sqrt{\min\{n_0, n_1\}}}.
\end{aligned}$$

Therefore, with probability at least $1 - \delta$ we have

$$\kappa - \xi(\delta) \cdot \sqrt{\frac{2\log n_1}{n_1}} \leq \hat{\kappa}_{\text{HS}} \leq \kappa + \frac{2\beta}{\alpha} + \left(1 + \frac{\beta}{\alpha}\right) \cdot \frac{16 \cdot \xi(\delta)}{\alpha\sqrt{\min\{n_0, n_1\}}}.$$

■

Appendix B. Proof of Lemma 5

Proof [Proof of Lemma 5] We begin by bounding the ϵ-covering number of Ω_q as follows

$$M_{\|\cdot\|_\infty}(\Omega_q, \epsilon) \leq (2C_{\mathcal{X}}\text{diam}(\mathcal{X})) \cdot ((2q)/\epsilon)^{-(\dim_{\text{cov}}(\mathcal{X})+1)}. \tag{14}$$

By assumption for each $\epsilon > 0$ we have $M_\rho(\mathcal{X}, \epsilon) \leq C_{\mathcal{X}} \cdot \epsilon^{-\dim_{\text{cov}}(\mathcal{X})}$. Moreover, for each $\epsilon > 0$ we have $M_{|\cdot|}([0, \text{diam}(\mathcal{X})], |\cdot|, \epsilon) \leq 2\text{diam}(\mathcal{X})/\epsilon$. Hence, by the triangle inequality we have

$$M_{\|\cdot\|_\infty}\left(\{g_{x_0,r_0}\}_{x_0\in\mathcal{X}, r\in[0,\text{diam}(\mathcal{X})]}, \epsilon\right) \leq 2C_{\mathcal{X}}\text{diam}(\mathcal{X}) \cdot (2/\epsilon)^{\dim_{\text{cov}}(\mathcal{X})+1}$$

The bound (14) now follows since $z \mapsto \Phi_1(q \cdot z)$ is q-Lipschitz. We now apply Dudley's inequality Dudley (1967) to bound the Rademacher complexity,

$$\begin{aligned}
\mathfrak{R}_n(\Omega_q) &\leq \frac{12}{\sqrt{n}} \cdot \int_0^1 \sqrt{\log_e\left(M_{\|\cdot\|_\infty}(\Omega_q, \epsilon)\right)} d\epsilon \\
&\leq \frac{12\sqrt{\dim_{\text{cov}}(\mathcal{X})+1}}{\sqrt{n}} \cdot \int_0^1 \sqrt{\log_e\left((2C_{\mathcal{X}}\text{diam}(\mathcal{X}))^{\frac{1}{\dim_{\text{cov}}(\mathcal{X})+1}} \cdot (2q)/\epsilon\right)} d\epsilon \\
&= 24\,(2C_{\mathcal{X}}\text{diam}(\mathcal{X}))^{\frac{1}{\dim_{\text{cov}}(\mathcal{X})+1}} \cdot q \cdot \sqrt{\frac{\dim_{\text{cov}}(\mathcal{X})+1}{n}} \cdot \int_0^1 \sqrt{\log_e(1/\theta)} d\theta \\
&\leq 21.5\,(2C_{\mathcal{X}}\text{diam}(\mathcal{X}))^{\frac{1}{\dim_{\text{cov}}(\mathcal{X})+1}} \cdot q \cdot \sqrt{\frac{\dim_{\text{cov}}(\mathcal{X})+1}{n}}.
\end{aligned}$$

The result now follows immediately from the Rademacher concentration bound (Mohri et al., 2012, Theorem 3.1). ■