Localized Structured Prediction

Carlo Ciliberto, Francis Bach, Alessandro Rudi

Localized Structured Prediction

Carlo Ciliberto *
c.ciliberto@ucl.ac.uk

Francis Bach †
francis.bach@inria.fr

Alessandro Rudi †
alessandro.rudi@inria.fr

December 18, 2018

Abstract

Key to structured prediction is exploiting the problem structure to simplify the learning process. A major challenge arises when data exhibit a local structure (e.g., are made by "parts") that can be leveraged to better approximate the relation between (parts of) the input and (parts of) the output. Recent literature on signal processing, and in particular computer vision, has shown that capturing these aspects is indeed essential to achieve state-of-the-art performance. While such algorithms are typically derived on a case-by-case basis, in this work we propose the first theoretical framework to deal with part-based data from a general perspective. We derive a novel approach to deal with these problems and study its generalization properties within the setting of statistical learning theory. Our analysis is novel in that it explicitly quantifies the benefits of leveraging the part-based structure of the problem with respect to the learning rates of the proposed estimator.

1 Introduction

Structured prediction deals with supervised learning problems where the output space is not endowed with a canonical linear metric but has a rich semantic or geometric structure [1, 2]. Typical examples are settings in which the outputs correspond to strings (e.g., captioning [3]), images (segmentation [4]), ordered sequences [5] or protein foldings [6] to name a few.

The lack of linearity on the output space poses several modeling and computational challenges when designing a learning algorithm for structured prediction. However, this additional complexity comes with a potential significant advantage. Indeed, if suitably incorporated within the learning model, knowledge about the structure could capture key properties of the data. This could potentially lower the (sample) complexity of the problem, attaining better generalization performance with less training examples. In this sense, a natural scenario is the case where both input and output data are organized into "parts" that can interact with one another according to a specific structure. This arises typically in applications such as computer vision (e.g., segmentation [4], localization [7, 8], pixel-wise

*University College London, WC1E 6BT London, United Kingdom.
†INRIA - Département d'informatique de l'ENS, Ecole normale supérieure, CNRS, INRIA, PSL Research University, 75005 Paris, France

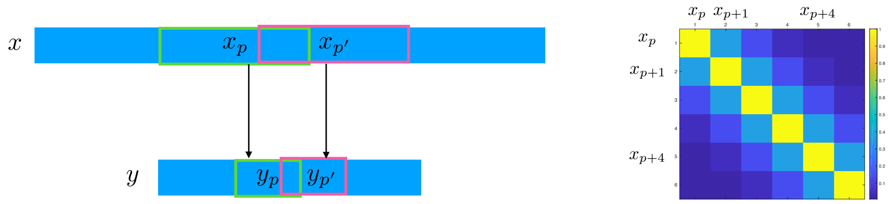

Figure 1: Locality on a sequence-to-sequence prediction setting. (Left) Inter-locality between parts of the input sequence x to the corresponding overlapping parts of the output sequence y. The output y_p depends only on the input x_p for every part $p \in P$. (Right) Intra-locality in terms of the covariance between different parts of the input. The covariance between parts decreases as the parts become further apart (off-diagonal entries).

classification [9]), speech recognition [10, 11], natural language processing [12], trajectory planing [13] or hierarchical classification [14].

Recent literature on the topic has shown that if correctly handled, the local structure in the data can lead to significantly better predictions over more global approaches [15, 16]. On the applicative side, these problems are typically addressed on a case-by-case basis, deriving algorithms that are ad-hoc for the individual learning problem. On the theoretical side, few works have considered less specific part-based factorizations [17] and a comprehensive theory analyzing the effect of local interactions between parts within the context of supervised learning is still missing.

In this paper, we propose (1) a novel theoretical framework that can be applied to a wide family of structured prediction settings able to capture potential local structure in the data, and (2) a structured prediction algorithm, based on this framework for which we prove universal consistency and generalization rates. A key aspect of our analysis is to quantify the impact of the part-based structure of the problem on the learning rates of the proposed estimator. In particular, we show that under natural assumptions on the local behavior of the data, our algorithm naturally benefits from this underlying structure.

2 Motivation: Learning with *Inter-locality* and *Intra-locality*

In this work we assume that data points have a natural characterization in terms of "parts". Practical examples of this setting often arise in image/audio or language processing, where the signal has a natural factorization in patches or sub-sequences. Following these guiding examples, we assume that any $x \in X$ and $y \in Y$ can be interpreted as a collection of (possibly overlapping) parts, and denote x_p (respectively y_p) its corresponding p-th part, with $p \in P$ a set of parts identifiers (e.g., possible patch positions and sizes).

To investigate the role of the parts in the learning process, in the following we introduce two key assumptions which are illustrated in Fig. 1. Their purpose is to formalize the intuition that the learning problem should interact well with the structure of parts of both input and output. Inspired by the motivating example of image processing, where parts (i.e., patches) capture the local properties of the data, we refer to these assumptions as *inter-locality* and *intra-locality* since they characterize respectively the interplay *between* corresponding input-output parts and the correlation of parts *within* the same input.

Assumption 1 (Inter-locality). *y_p is conditionally independent from x, given x_p, moreover the probability of y_p given x_p is the same as $y_{p'}$ given $x_{p'}$, for any $p, p' \in P$.*

Inter-locality formalizes the intuition that the p-th part of the output $y \in Y$ depends only on the p-th part of the input $x \in X$, see Fig. 1 (Left) for an intuition of this. A natural setting where this assumption is verified is for instance the case of pixel-wise classification, where the class y_p of a pixel p on image can be determined only based on the sub-image depicted in the corresponding patch x_p (e.g., a smaller window around the pixel p). Note that our assumption, although based on conditional independence, is weaker than assuming a joint graphical model on all parts of x and all parts y, where y_p is only connected to x_p, and connections among the parts x_p are arbitrary.

Assumption 1 suggests that we can solve a "simpler" learning problem, by focusing on the parts of X and the corresponding parts of Y. This motivates the adoption of learning approaches that directly learn the relation between parts, which have been observed to be remarkably effective in computer vision applications [8, 15, 16].

Inter-locality however offers a significant benefit only when the input parts are not too highly correlated. For instance, in the extreme case where parts are all identical copies, there is no advantage in solving the learning problem locally. In this sense, intra-locality measures the amount of "covariance" between two parts p and q of an input x as

$$C_{p,q} = \mathbb{E}_x\ S(x_p, x_q) - \mathbb{E}_{x,x'}\ S(x_p, x'_q) \tag{1}$$

for $S(x_p, x_q)$ a suitable measure of similarity between parts (if $S(x_p, x_q) = x_p x_q$, with x_p and x_q scalars random variables, then $C_{p,q}$ is the p, q-th entry of the covariance matrix of the vector $(x_1, \dots, x_{|P|})$). In particular note that $\mathbb{E}_x S(x_p, x_q)$ and $\mathbb{E}_{x,x'} S(x_p, x'_q)$ measure the similarity between the p-th and the q-th part of, respectively, the *same* input x, and two *independent* inputs x, x'. So if the p-th part of an input is independent of the q-th part, then two expectations correspond exactly and we have $\mathbb{E}_x S(x_p, x_q) = \mathbb{E}_{x,x'} S(x_p, x'_q)$, so $C_{p,q} = 0$. In many contexts, when there is a notion of distance on P, it is safe assume that $C_{p,q}$ between the p-th and the q-th part decays with the distance between p and q.

Assumption 2 (Intra-locality). *There exists a distance d over P, and $\gamma \geq 0$ such that*

$$|C_{p,q}| \ \leq\ r^2\ e^{-\gamma d(p,q)}, \tag{2}$$

with $r = \sup_{x,x'} |S(x, x')|$.

Note in particular that the intra-locality condition is always satisfied with $\gamma = 0$. However when x_p is independent of x_q, it holds with $\gamma = \infty$ and $d(p, q) = \delta_{p,q}$. Exponential decays of correlation are typically observed when the distribution of the parts of x factorizes in a graphical model that connects parts which are close in terms of the distance d: although all parts depend on each other, the long-range dependence typically goes to zero exponentially fast in the distance (see, e.g., [18] for mixing properties of Markov chains). Fig. 1 (Right) illustrate a potential decay of the relation $|C_{p,q}|$ between two parts x_p and x_q of an sequence, proportional to their distance $d(p, q)$.

A main contribution of this work is to show that the structured prediction estimator we will introduce in Sec. 4 has generalization properties that match those of the state of the art (see Thm. 2, 4 in Sec. 5). More importantly, we prove that if the problem satisfies the locality assumptions introduced in this section, the generalization properties of our estimator improve proportionally to the number of the parts. Here we give an informal version of this key result, which is reported in Thm. 7 in detail. Below we denote by $\widehat{f}$ the proposed structured prediction estimator and by $\mathcal{E}(f)$ the expected error of a predictor $f : X \to Y$. We will denote by n the number of examples and P the number of parts.

Theorem 1 (Informal - Learning Rates & Locality). *Under mild assumptions on the loss and data distribution. If the learning problem is local (Asm. 1, 2), then*

$$\mathbb{E}\, \mathcal{E}(\widehat{f}\,) - \inf_f \mathcal{E}(f) \;\; \leq \;\; c_0 \left(\frac{1}{nP}\right)^{1/4} \; \left(1 + \frac{\sum_{p,q} e^{-\gamma d(p,q)}}{P}\right)^{1/4}. \tag{3}$$

In the worst-case scenario where $\gamma = 0$ (no exponential decay of the covariance between parts) the overall bound will scale as $1/n^{1/4}$, which recovers the result of [19] where no structure among parts is assumed. However, as soon as γ increases, then the bound will scale as $1/(Pn)^{1/4}$, as if all parts were totally independent. Note that in this paper we assumed the exponential decay model for the intra-locality of Assumption 2. Clearly, also longer-range dependencies capturing more refined behaviors between the parts can be considered.

3 Problem Formulation

We denote by X, Y and Z respectively the *input space*, *label space* and *output space* of a learning problem. Let ρ be a probability measure on $X \times Y$ and $\triangle : Z \times Y \times X \to \mathbb{R}$ be a loss function measuring prediction errors between a label $y \in Y$ and a output $z \in Z$, possibly parametrized by an input $x \in X$. To stress this interpretation in the following we adopt the notation $\triangle(z, y|x)$. The structure of $\triangle$ is a key aspect of this work and we will discuss it further in the rest of this section.

In structured prediction settings, the goal is to estimate the function $f^* : X \to Z$ defined as a minimizer of the *expected risk*

$$\min_{f:X\to Z} \mathcal{E}(f), \quad \text{with} \quad \mathcal{E}(f) = \int \triangle(f(x), y|x) \; d\rho(x, y), \tag{4}$$

over the set of measurable functions $f : X \to Z$. In practice, the distribution ρ is given but unknown and only $(x_i, y_i)_{i=1}^n$ independently and identically distributed according to ρ are accessible.

Loss Made by Parts. We formalize the intuition introduced in Sec. 2 that data are decomposable into "parts" and denote with $[X], [Y]$ and $[Z]$ the sets of *parts* associated to respectively X, Y and Z. We consider a set P of part indices and define the operator from $X \times P \to [X]$ as the map sending the pair (x, p) to a point in $[X]$ that we denote $[x]_p$ for any $x \in X$ and $p \in P$ (analogously for Y and Z). The concept of "part" is introduced here in a rather abstract sense and allows to describe a wide range of possible structures. For a more concrete example consider the case where $X = \mathbb{R}^D$ and the set P identifies all sets of subsequence indexes of dimension $d \in \mathbb{N}$. Then, $[X] = \mathbb{R}^d$ and for any $x \in X$ and $p \in P$

such that $p = \{i, \dots, i+d\}$ with $i < D - d - 1$, we have that $[x]_p = (x_i, \dots, x_{i+p}) \in \mathbb{R}^d$ is the orthogonal projection of x onto its coordinates indexed by p. As mentioned in Sec. 2, practical examples of this setting arise often in image and audio processing settings, where the signal, for instance an image, has a natural factorization into overlapping patches or windows [8]. In the following, when it is clear from context, we will adopt the shorthand notation $x_p = [x]_p$, which however should not be confused with the p-th coordinate of a vector x as in the previous example (since in general X is not necessarily be vector space).

For simplicity, in the following we will assume P to be a finite set, however our analysis generalizes naturally to infinite and possibly dense sets of parts P (see supplementary material). Let $\pi(\cdot|x)$ be a probability distribution over the set of parts, conditioned with respect to an input $x \in X$. In this work we study the family of loss functions $\triangle$ that can be represented as

$$\triangle(z, y|x) = \sum_{p \in P} \pi(p|x)\ L_p(z_p, y_p|\ x_p). \tag{5}$$

The collection of $(L_p)_{p \in P}$ is a family of loss functions $L_p : [Z] \times [Y] \times [X] \to \mathbb{R}$, each comparing the p-th part of a label y and output z. For instance, in an image processing scenario, L_p could measure the similarity between the two images at different locations and scales, indexed by p. In this sense, the distribution $\pi(p|x)$ allows to weigh each L_p differently depending on the application (e.g., mistakes at large scales could be more relevant than at lower scales). Note that we adopted the non-standard notation $L_p(\cdot|x)$ to stress dependency of the prediction errors *given* the observed input.

Remark 1 (Examples Loss Functions by Parts). *Several loss functions used in machine learning have a natural formulation by parts in terms of Eq.* (5). *Notable examples are the Hamming distance [20–22], used in settings such as hierarchical classification [14], computer vision [2, 9, 16] or trajectory planning [13] to name a few. Also, loss functions used in natural language processing, such as the precision/recall and $F1$ score can be written in this form. Finally, we point out that multi-task learning settings [23] can be seen as problem by parts, with the loss corresponding to the sum of standard regression/classification loss functions (least-squares, logistic, etc.) over the tasks/parts.*

4 Algorithm

In this section we introduce our estimator for structured prediction problems with parts. Our learning strategy is preceded by an auxiliary step for dataset generation that explicitly extracts the parts from the data.

Auxiliary Dataset Generation. The locality assumptions introduced in Sec. 2 motivate us to learn the local relations between individual parts $p \in P$ of each input-output pair. In this sense, given a training dataset $\mathcal{D} = (x_i, y_i)_{i=1}^n$ a first step would be to extract a new, part-based dataset $\{(x_p, p, y_p) \mid (x, y) \in \mathcal{D},\ p \in P\}$. However in real scenarios the cardinality $|P|$ of the set of parts can be very large (possibly infinite as we discuss in the Appendix) and so generating such part-based dataset would be infeasible. Instead, we generate an *auxiliary dataset* by randomly sub-sampling $m \in \mathbb{N}$ elements from the part-based dataset.

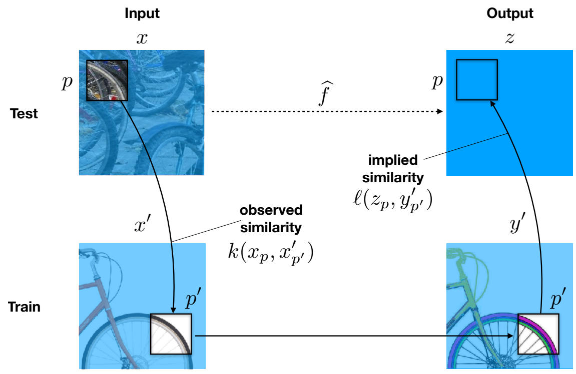

Figure 2: Illustration of the prediction process for the estimator $\widehat{f}$ considered in this work (see Eq. (6)) in an computer vision application: for a given test image x, the α scores detect a similarity between the p-th patch of x (Top-left) and the p_j-th patch of the training input x_{i_j} (Bottom-left). As a consequence, the estimator will enforce the p-th patch of the output z (Top-right) to be similar to the p_j-th patch of the training label y_{i_j} (Bottom-right).

Concretely, for $j \in \{1, \ldots, m\}$, we first sample i_j uniformly on $\{1, \ldots, n\}$, then we choose $\chi_j = x_{i_j}$, sample $p_j \sim \pi(\cdot \mid \chi_j)$ and finally choose $\eta_j = [y_{i_j}]_{p_j}$. Then the auxiliary dataset results in $\mathcal{D}' = (\chi_j, p_j, \eta_j)_{j=1}^m$. This procedure is summarized in the GENERATE routine of Alg. 1.

Estimator. Given the auxiliary dataset, we consider an estimator $\widehat{f} : X \to Z$, such that for any $x \in X$

$$\widehat{f}(x) = \operatorname*{argmin}_{z \in Z} \sum_{p \in P} \sum_{j=1}^{m} \alpha_j(x, p) \left[\pi(p|x) \ L_p(z_p, \eta_j | x_p)\right]. \tag{6}$$

The functions $\alpha_j : X \times P \to \mathbb{R}$ are *learned* from the auxiliary dataset and are the fundamental components allowing the estimator to capture the part-based structure of the learning problem. Indeed, for any test point $x \in X$ and part $p \in P$, the value $\alpha_j(x, p)$ can be interpreted as a measure of how similar x_p is to the p_j-th part of the auxiliary training point χ_j. For instance, assume $\alpha_j(x, p)$ to be an approximation of the delta function that is 1 when $x_p = [\chi_j]_{p_j}$ and 0 otherwise. Then, the terms in the objective functional in Eq. (6) become

$$\alpha_j(x, p) \ L_p(z_p, \eta_j | x_p) \quad \approx \quad \delta(x_p, [\chi_j]_{p_j}) \ L_p(z_p, \eta_j | x_p), \tag{7}$$

implying essentially that

$$x_p \approx \chi_{j_{p_j}} \quad \Longrightarrow \quad z_p \approx \eta_j, \tag{8}$$

that is, if a similarity is observed between the p-th part of test input x and the p_j-th part of the auxiliary training input χ_j (i.e. the p_j-th part of the training input x_{i_j}), then the p-th part of the test output z will be chosen to be similar to the auxiliary part η_j. This process is depicted in Fig. 2 for an illustrative computer vision scenario: for a given test image x, the α scores detect a similarity between the p-th patch of x and the p_j-th patch of the training input x_{i_j}. As a consequence, the estimator will enforce the p-th patch of the output z to be similar to the p_j-th patch of the training label y_{i_j}.

Learning α. In line with previous work on structured prediction [19], in this work we learn the function α_j by solving a linear system for a problem akin to kernel ridge regression (see Sec. 5 for the theoretical motivation). In particular, let $k : (X \times P) \times (X \times P) \to \mathbb{R}$ be a positive definite kernel, we define

$$(\alpha_1(x,p), \dots, \alpha_m(x,p))^\top = (K + m\lambda I)^{-1} v(x,p), \tag{9}$$

where $K \in \mathbb{R}^{m \times m}$ is the empircal kernel matrix with entries $K_{jh} = k((\chi_j, p_j), (\chi_h, p_h))$ and $v(x,p) \in \mathbb{R}^m$ is the vector with entries $v(x,p)_j = k((\chi_j, p_j), (x,p))$. Training the proposed algorithm, consists essentially in precomputing $C = (K + m\lambda I)^{-1}$, that is necessary to evaluate the coefficients α as detailed by the LEARNING routine in Alg. 1. Note that if we compute C with direct methods, the total computational cost amounts to $O(m^3)$, however it is possible to exploit low rank approximation methods, to achieve essentially the same accuracy with complexity $O(m\sqrt{m})$ (see [24, 25]).

We care to point out that the proposed estimator can be seen as a refinement of the one in [19], which is not able to capture the structure-based nature of the problem in terms of its parts. Indeed, we recover this method when no explicit decomposition into parts is assumed on $\triangle$ (i.e. P is a singleton), as detailed in Appendix I.

Remark 2 (Evaluating $\widehat{f}$). *According to* (6)*, evaluating $\widehat{f}$ on a test point $x \in X$ consists in solving an optimization problem over the output space Z. This design of the test phase is standard in structured prediction settings [2], where a corresponding optimization protocol is derived on a case-by-case basis depending on the loss and the space Z (see e.g. [2]). However, the specific form of the objective functional in our setting allows also to suggest a general stochastic meta-algorithm. In particular,* (6) *can interpreted as the problem of minimizing an expectation*

$$\widehat{f}(x) = \underset{z \in Z}{\operatorname{argmin}} \ \mathbb{E}_{j,p} \ h_{j,p}(z|x) \tag{10}$$

with p sampled according to π, $j \in \{1, \dots, m\}$ sampled according to the relevance weights α_j and $h_{j,p}$ defined accordingly in terms of L_p. When the $h_{j,p}$ are (sub)differentiable, problems of the form of (10) *are effectively addressed by stochastic gradient methods (SGM). In Alg. 3 in the supplementary material we give an example of this strategy.*

5 Generalization Properties of Structured Prediction with Parts

In this section we study the statistical properties for the proposed algorithm. We prove that under mild assumptions on the loss, the approach is universally consistent. We further derive learning rates. Our analysis leverages the assumption that the loss function $\triangle$ is a *Structure Encoding Loss Function (SELF) by Parts*.

Definition 1 (SELF by Parts)**.** *A function $\triangle : Z \times Y \times X \to \mathbb{R}$ is a* Structure Encoding Loss Function (SELF) by Parts *if it admits a factorization in the form of* (5) *with functions*

Algorithm 1 Learning $\widehat{f}$

Input: training set $(x_i, y_i)_{i=1}^n$, distributions $\pi(\cdot|x)$ a reproducing kernel k on $X \times P$, hyperparameter $\lambda > 0$, auxiliary dataset size $m \in \mathbb{N}$.

GENERATE the auxiliary dataset $(\eta_j, \chi_j, p_j)_{j=1}^m$:
Sample i_j uniformly from $\{1, \dots, n\}$. Set $\chi_j = x_{i_j}$
Sample $p_j \sim \pi(\cdot|\chi_j)$.
$\eta_j = [y_{i_j}]_{p_j}$.

LEARN the coefficients for the score function α:
$K \in \mathbb{R}^{m \times m}$ with entries $K_{jj'} = k((\chi_j, p_j), (\chi_{j'}, p_{j'}))$
$A = (K + m\lambda I)^{-1}$

Return $\alpha : X \times P \to \mathbb{R}^m$ such that $\alpha(x, p) = A\ v(x, p)$ with $v(x, p) \in \mathbb{R}^m$ is the vector with entries $v(x, p)_j = k((\chi_j, p_j), (x, p))$.

$L_p : [Z] \times [Y] \times [X] \to \mathbb{R}$, *and there exists a separable Hilbert space* $\mathcal{H}$ *and two bounded maps* $\psi : [Z] \times [X] \times P \to \mathcal{H}$, $\varphi : [Y] \to \mathcal{H}$ *such that for any* $\zeta \in [Z]$, $\eta \in [Y]$, $\xi \in [X]$, $p \in P$

$$L_p(\zeta, \eta|\xi) = \langle \psi(\zeta, \xi, p), \varphi(\eta) \rangle_{\mathcal{H}}. \tag{11}$$

The definition of "SELF by Parts" specializes the definition of SELF in [26] and in the following we will always assume $\triangle$ to satisfy it. Indeed, Def. 2 is always satisfied when the spaces of parts involved are discrete sets and it is rather mild in the general case. For instance if Y, Z are bounded subsets of the Euclidean space, then the condition holds for any absolutely continuous loss function (see [19], in particular Example 1 and Thm. 19 of the same paper for an exhaustive list of examples). Since we will not make use of the original definition of SELF, for simplicity, in this work we will refer to a function satisfying Def. 1 as SELF.

Now we are ready to prove the universal consistency of the estimator in Eq. (6).

Theorem 2 (Universal Consistency)**.** *Let* $\triangle$ *be SELF and* Z *a compact set. Let* K *be a bounded continuous universal kernel on* $X \times P$. *Let* $\widehat{f}_n$ *as in Eq.* (6) *with i.i.d. training set and auxiliary dataset sampled according to Sec. 4, with* $m \propto n$. *Then*

$$\lim_{n\to\infty} \mathcal{E}(\widehat{f}_n) = \inf_{f:X\to Z} \mathcal{E}(f) \quad \textit{with probability } 1. \tag{12}$$

The proof of the theorem above is in Appendix H. Note that the requirement of universal kernel is a standard assumption for universal consistency (see [27]). An example of continuous universal kernel on $X \times P$ is

$$K((x, p), (x', p')) = K_0(x, x')\ \delta_{p,p'} \tag{13}$$

where K_0 is any unversal kernel on X, e.g. the Gaussian $K_0(x, x') = \exp(-\|x - x'\|^2)$. While the proposed estimator is consistent with the kernel described above, it is not able to benefit from the effect of locality. At the end of Sec. 5.1 we will provide a kernel that guarantees consistency and benefits from locality at the same time.

Learning Rates (General Case). The analysis for learning rates starts from the observation that when the loss function is SELF the solution of the learning problem in Eq. (4) is completely characterized in terms of the *conditional expectation* or *conditional mean embedding* of $\varphi(y_p)$ given x, denoted by $g^* : X \times P \to \mathcal{H}$ [28–30] and defined as follows

$$g^*(x, p) = \int_Y \varphi(y_p) d\rho(y|x). \tag{14}$$

Lemma 3. *Let $\triangle$ be SELF and Z a compact set, then the solution of Eq. (4) is characterized by*

$$f^*(x) = \underset{z \in Z}{\operatorname{argmin}} \sum_{p \in P} \pi(p|x) \left\langle \psi(z_p, x_p, p), g^*(x, p) \right\rangle_{\mathcal{H}}, \tag{15}$$

almost everywhere with respect to the input distribution ρ_X.

To show Lemma 3 we make use of Berge's maximum theorem (see Appendix C for the details of the proof). The result characterizes the optimal solution f^* of the structured prediction problem in terms of the conditional expectation g^*. In this sense it should not come as surprising that the "regularity" of g^* will play a key role in controlling the learning rates. In particular we consider the quite standard assumption in the context of non-parametric estimation [19, 28, 31], that $g^* \in \mathcal{G} = \mathcal{H} \otimes \mathcal{F}$, where $\mathcal{F}$ is the reproducing kernel Hilbert space associated to the chosen kernel in Eq. (9). The learning rate of the estimator depends on the following constants $g, r, c_\triangle, q$, where the first three are defined as

$$g = \|g^*\|_{\mathcal{G}}, \quad r = \sup_{x \in X, p \in P} K((x,p),(x,p)), \quad c_\triangle^2 = \sup_{z \in Z, x \in X} \mathbb{E}_{p|x} \|\psi(z, x, p)\|_{\mathcal{H}}^2, \tag{16}$$

Note that the quantities above are rather natural. Indeed g characterizes the complexity of the conditional distribution ρ in terms of the hypothesis space induced by the kernel k on the input. This quantity is related to the inter-locality assumption as discussed in Lemma 5. r is the bound of the kernel. $c_\triangle$ measures the "complexity" of learning with the loss $\triangle$. Finally, q is defined as

$$q = \mathbb{E}_{x,x'} \mathbb{E}_{p,q|x,r|x'} \left[K((x,p),(x,q))^2 - K((x,p),(x',r))^2 \right] \tag{17}$$

where $\mathbb{E}_{p,q|x}[\cdot]$ is a shorthand for $\sum_{p,q \in P} \pi(p|x)\pi(q|x)[\cdot]$ (analogously for $\mathbb{E}_{r|x}$). This latter quantity will be key in Sec. 5.1 to capture and leverage intra-locality of the learning problem. In particular it will allow us to explicitly characterize the benefit of using the locality-aware estimator considered in this work, from a statistical viewpoint.

With the notation introduced above, we have the following general result (the proof is in Appendix F).

Theorem 4. *Let $\widehat{f}$ as in Eq. (6) with i.i.d. training set and auxiliary dataset sampled according to Alg. 1. If the output space Z is compact, the loss function $\triangle$ is SELF, $g^* \in \mathcal{G}$ and $\lambda \geq (r^2/m + q/n)^{1/2}$, then*

$$\mathbb{E}\, \mathcal{E}(\widehat{f}\,) - \mathcal{E}(f^*) \;\; \leq \;\; 12\, c_\triangle\, g \left(\frac{r^2}{\lambda m} + \frac{q}{\lambda n} + \lambda \right)^{1/2}. \tag{18}$$

Thm. 4 above characterizes the learning rates of $\widehat{f}$ under standard regularity assumption on the problem. This result is general in that it does not rely on the locality assumptions introduced in Sec. 2. In particular, we note that when $m \propto n$ and $\lambda \propto n^{-1/2}$, the bound in Thm. 4 recovers the excess risk bounds of structure prediction *without parts* [19, 26] of order $O(n^{-1/4})$.

In the following we show that under the locality assumptions the result in Thm. 4 can be improved significantly.

5.1 Main Result: Statistical Properties of Learning with locality

In this section we present the main result of this work (Thm. 7). In particular, we further investigate the bound of Thm. 4 in light of the two assumptions of inter and intra locality introduced in Sec. 2. To this end, we first study the direct effects of these two assumptions on the learning framework introduced in this work.

The Effect of Inter-locality. We start by observing that the inter-locality between parts of the inputs and parts of the output allows for a refined characterization of the conditional mean g^*.

Lemma 5. *Let* g^* *be defined as in Eq.* (14)*. Under Asm. 1, there exists* $\bar{g}^* : [X] \to \mathcal{H}$ *such that*

$$g^*(x, p) = \bar{g}^*(x_p) \qquad\qquad \forall x \in X,\ p \in P. \tag{19}$$

Lemma 5 above shows that we can learn g^* by focusing on a "simpler" problem, identified by the function $\bar{g}^*$ acting only the parts $[X]$ of X rather than on the whole input directly (for a proof see Lemma 23 in Appendix G). This motivates the adoption of the restriction kernel [7], namely a function $K : (X \times P) \times (X \times P) \to \mathbb{R}$ such that

$$K((x, p), (x', q)) = \bar{K}(x_p, x_q), \tag{20}$$

which, for any pair of inputs $x, x' \in X$ and parts $p, q \in P$, measures the similarity between the p-part of x and the q-th part of q via a kernel $\bar{K} : [X] \times [X] \to \mathbb{R}$ on the parts of X. The restriction kernel is a well-established tool in structured prediction settings [7] and it has indeed been observed to be remarkably effective in computer vision applications [8, 15, 16].

The effect of Intra-locality. We recall that intra-locality characterizes the statistical correlation between two different parts of the input (see Asm. 2). Below we show that this quantity is tightly related to constant q introduced in Eq. (17). To this end we consider the simplified scenario where the parts are sampled from the uniform distribution on P. While more general situations can be considered, this setting is useful to illustrate the effect we are interested in this work.

Lemma 6. *Under the same assumptions of Thm. 4, let* K *denote the restriction kernel defined in Eq.* (20) *in terms of* $\bar{K} : [X] \times [X] \to \mathbb{R}$*. Let* $\pi(p|x) = \frac{1}{|P|}$ *for any* $x \in X$ *and* $p \in P$*. Then, the constant* q *in Eq.* (17) *can be factorized as*

$$q = \frac{1}{|P|^2} \sum_{p,q \in P} C_{p,q}, \quad \textit{with} \quad C_{p,q} = \mathbb{E}_{x,x'}\left[\bar{K}(x_p, x_q)^2 - \bar{K}(x_p, x'_q)^2\right]. \tag{21}$$

For a proof of this result see Lemma 25 in Appendix G. It is clear that the $C_{p,q}$ in Eq. (21) correspond to the measure of correlation introduced in Eq. (1) when the similarity function S is replaced by the squared kernel on the parts $\bar{K}^2$.

We are now ready to specialize Thm. 4 in terms of the locality assumptions. In particular let $\bar{K}$ be a reproducing kernel on $[X]$, $\widehat{f}$ be the structured prediction estimator in Eq. (6) learned using the restriction kernel in Eq. (20) based on $\bar{K}$, and denote by $\bar{\mathcal{G}}$ the space of functions $\bar{\mathcal{G}} = \mathcal{H} \otimes \bar{\mathcal{F}}$ with $\bar{\mathcal{F}}$ the RKHS associated to $\bar{K}$.

Theorem 7 (Learning Rates & Locality)**.** *Under Assumption 1 and Assumption 2 with* $S = \bar{K}$*, let* $\bar{g}^*$ *satisfying Lemma 5, with* $\bar{g} = \|\bar{g}^*\|_{\bar{\mathcal{G}}} < \infty$*. When* $\lambda = (r^2/m + q/n)^{1/2}$*, then*

$$\mathbb{E}\, \mathcal{E}(\widehat{f}\,) - \mathcal{E}(f^*) \quad \leq \quad 12\, c_{\triangle}\, \bar{g}\, r^{1/2} \left(\frac{1}{m} + \frac{1}{|P|n} + \frac{\sum_{p \neq q} e^{-\gamma d(p,q)}}{|P|^2 n} \right)^{1/4}. \tag{22}$$

The proof of the theorem above can be found in Appendix G.2. We can see that inter and intra locality allow to refine (and potentially improve) the bound in Thm. 4 with terms that depend on the number of parts. In particular, we observe that the adoption of the restriction kernel in Thm. 7 allows the structured prediction estimator to leverage the intra-locality, gaining a benefit proportional to the magnitude of the parameter γ. More precisely, if $\gamma = 0$ (e.g. all parts are identical copies) then we recover the rate of $O(n^{-1/4})$ of Thm. 4, while if γ is large (the parts are almost not correlated) we can take $m \propto n|P|$ achieving a rate of the order of $O\big((n|P|)^{-1/4}\big)$. We clearly see that depending on the amount of intra-locality in the learning problem, the proposed estimator is able to gain significantly in terms of finite sample bounds.

A natural question is how to design a structured prediction estimator that is both able to leverage the locality assumptions, when they hold, and be universally consistent even when there is no locality. The following remark addresses this questions and concludes our theoretical analysis.

Remark 3 (Universal and Local Kernels)**.** *By construction, the restriction kernel allows to learn only functions* $g^* : X \times P \to \mathcal{H}$ *such that* $g^*(x, p) = \bar{g}^*(x_p)$*. Consequently, the corresponding structured prediction estimator is not universal. However, in Thm. 7 we have observed that under the locality assumptions, the restriction kernel achieves significantly faster rates with respect to universal kernels of the form of Eq.* (13)*.*

Interestingly, it is possible to design a kernel able to take the best of both worlds, leading to an estimator that is universal but also able to leverage the parts-based structure of a learning problem when possible. We obtain this kernel as the sum $K_B = K_U + K_L$ *of a universal kernel* K_U *on* $X \times P$ *and a restriction (or "local") kernel* K_L*. Indeed, as shown in Appendix I.3, the kernel* K_B *is universal, hence Thm. 2 applies to the corresponding estimator* $\widehat{f}$*. Moreover, under the locality assumptions, a result identical to Thm. 7 holds for the estimator trained with* K_B*.*

6 Empirical Evaluation

We report here on the empirical performance of the proposed estimator on simulations and preliminary experimental results. The goal is to highlight the role played by the parts in achieving better generalization performance even when only few training examples are available.

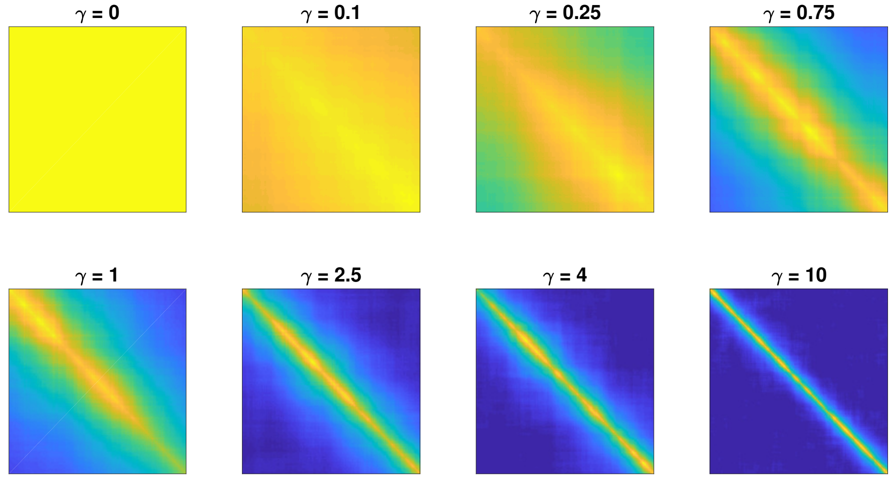

Figure 3: Empirical intra-locality matrix (with entries C_{pq} defined in Eq. (21)) for varying values of γ and linear restriction kernel. Data generated according to the protocol in Sec. 6.1 with $|P| = 200$ parts and $n = 100$ points. The intra-locality matrices are normalized between 0 (Blue) and 1 (Yellow).

6.1 Simulation - Intra Locality

The coefficient γ in Asm. 2 characterizes the "amount" of intra-locality in a learning problem. To clearly appreciate the role played by this parameter in combination with the number of parts we studied a simplified scenario with simulated data. In particular we adopted a data generation protocol in which it is possible to control the parameter γ directly.

We considered a setting where input data is a vector $x \in \mathbb{R}^{k|P|}$ composed of $|P|$ parts, with each part corresponding to a vector in $\mathbb{R}^k$. For all our experiments we used $k = 1000$. The input points $x \in \mathbb{R}^{k|P|}$ are then sampled according to a normal distribution with zero mean and covariance $\Sigma(\gamma) = M(\gamma) \otimes I$, where $I \in \mathbb{R}^{k \times k}$ denotes the identity matrix and $M(\gamma) \in \mathbb{R}^{|P| \times |P|}$ the matrix with entries

$$M(\gamma)_{pq} = e^{-\gamma d(p,q)}, \tag{23}$$

with $d(p, q) = |p - q|/|P|$. To verify that this generation protocol allows us to control the amount of intra-locality in the data, in Fig. 3 we report the empirical estimation of the intra-locality matrix C, with entries C_{pq} defined as in Eq. (21), for different values of γ. We used $n = 100$ points, $|P| = 200$ parts and the linear restriction kernel. As intended, when the parameter γ increases from 0 to infinity, the intra-locality matrix varies from being rank-one (all parts identical copies of each other) to diagonal (all parts independently sampled).

To isolate the specific effect of intra-locality on the learning rates, we evaluated the estimator introduced in Eq. (6) on a linear multitask (in particular vector-valued) regression problem with $\triangle$ the least-squares loss. To guarantee *inter*-locality, we generated a regression vector $w \in \mathbb{R}^{k|P|}$ by first sampling $\bar{w} \in \mathbb{R}^k$ uniformly on the radius one ball and then taking

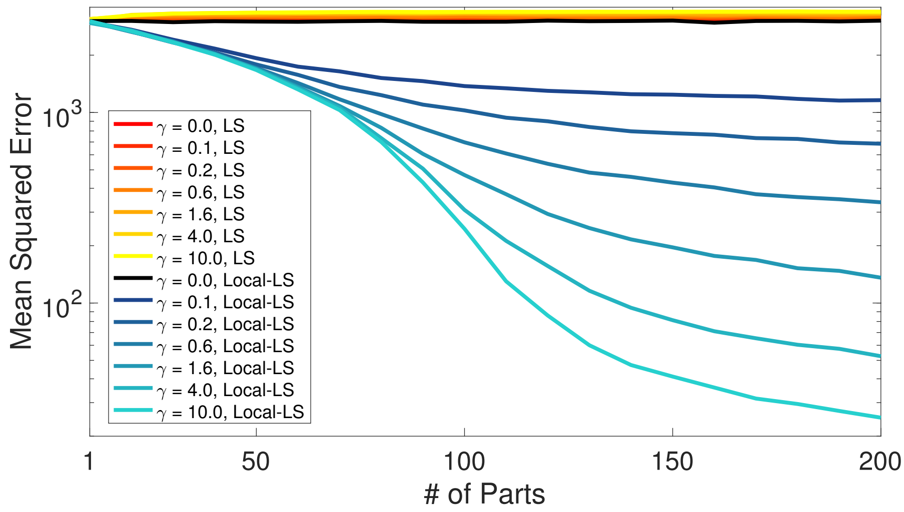

Figure 4: Standard linear regression *LS* is compared with linear regression using the parts *Local-LS* as described in Sec. 6.1 with respect to the Mean Squared Error (log scale), for different values γ of intra-locality and increasing number of parts $|P|$, and fixed numnber of examples $n = 100$.

$w = [\bar{w}, \ldots, \bar{w}]$ the vector concatenating $|P|$ copies of $\bar{w}$. We generated datasets $(x_i, y_i)_{i=1}^n$ of size $n = 100$ for training and $n = 1000$ for testing, with x_i sampled according to the procedure described above and $y_i = w^\top x_i + \epsilon$ with noise $\epsilon \in \mathbb{R}^{k|P|}$ sampled from an isotropic gaussian with standard deviation 0.5.

We perfomed regression with linear restriction kernel (denoted in Fig. 4 by *Local-LS*) on the auxiliary dataset $([x_i]_p, [y_i]_p)_{i=1,p=1}^{n,|P|}$, and compared with standard linear regression (denoted in the figure by *LS*) on the original dataset $(x_i, y_i)_{i=1}^n$. The parameter λ was chosen by hold-out cross-validation in the range $[10^{-6}, 10]$ (logarithmically spaced). For each experimental condition, tests performances have been averaged over 100 runs to account for statistical variability.

Fig. 4 reports the performance of the estimator $\widehat{f}$ for different intra-locality values γ as the number of parts increases. As predicted by Thm. 7 we observe that when input data is intra-local (large values of γ) and the number of parts is large, Local-LS offers a remarkable advantage in terms of generalization error with respect to standard LS, which does not benefit from the local properties of the problem. When γ is close to zero, this advantage is less prominent even if the number of parts is large. Indeed, we do not observe any significant variation in the prediction error when $\gamma = 0$, since every input point corresponds to the concatenation of $|P|$ identical copies of a vector in $\mathbb{R}^k$.

6.2 Learning the Direction of Ridges in Fingerprint Images

We considered a learning problem inspired by the one in [32] where the goal is to recover the pointwise direction of ridges in a fingerprint image. We used the FVC04 dataset[1] which consists in 80 grayscale 640×480 pictures of fingerprints in input, with the corresponding

[1] `http://bias.csr.unibo.it/fvc2004`, DB1_B. The output is obtained by applying 7×7 Sobel filtering.

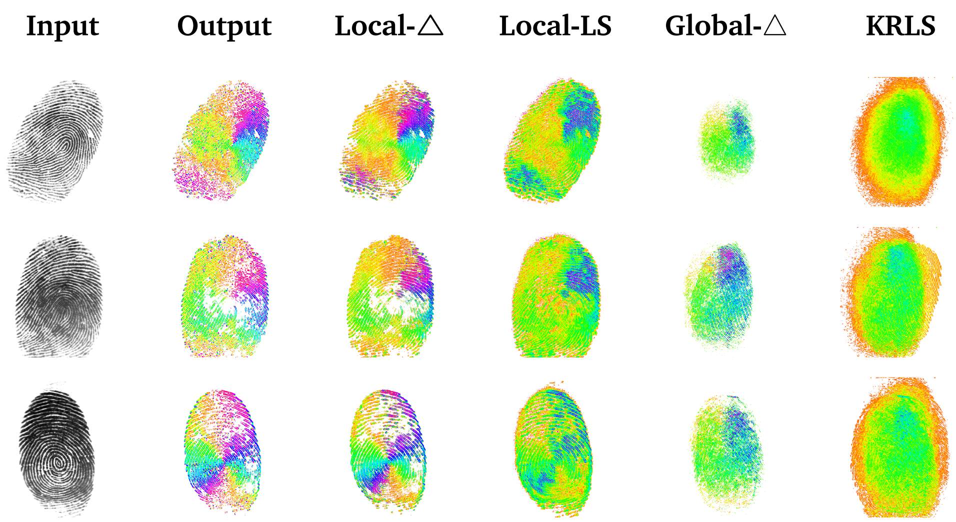

Figure 5: **Learning the direction of ridges in fingerprint images - Examples**. From left to right: (Input) fingerprint images; test ground truth (Output), with pixels' color corresponding to the local direction of ridges between $[-\pi, \pi]$; predictions of Alg. 1 with loss in Eq. (25) (Local-△) and squared distance (Local-LS); Predictions of the algorithm in [19] with loss in Eq. (24) (Global-LS); Predictions of kernel ridge regression (KRLS). See discussion in Sec. 6.2.

output pictures encoding the direction (from $-\pi$ to π) associated to each pixel in the ridges of the input fingerprint. In Fig. 5 (First and Second columns) we report three input-output examples in the dataset. The color of individual pixels on output images encodes the local orientation of the ridge.

By denoting with $[\cdot]_{ij}$ the i, j element of a matrix, the natural loss function associated to this problem is

$$\triangle(z, y) = \frac{1}{640 \times 480} \sum_{i=1}^{640} \sum_{j=1}^{480} g([z]_{ij}, [y]_{ij})^2, \quad g(\alpha, \beta) = |\sin(\alpha - \beta)|, \tag{24}$$

where g is a natural distance between two directions (discarding the orientation) and $\alpha, \beta \in [-\pi, \pi]$. In particular, to apply the proposed algorithm, we consider the following representation of $\triangle$ in term of parts. Let P be the collection of patches of dimension 20×20 and equispaced each 5×5 pixels[2], then each pixel belongs exactly to 16 patches and so the loss $\triangle$ in the equation above is characterized by

$$\triangle(z, y) = \frac{16}{|P|} \sum_{p \in P} G(z_p, y_p), \quad G(\zeta, \eta) = \frac{1}{20 \times 20} \sum_{i,j=1}^{20} g([\zeta]_{ij}, [\eta]_{ij})^2, \tag{25}$$

where $\eta, \zeta \in [-\pi, \pi]^{20 \times 20}$ are the extracted patches.

Results. We compared the approach proposed in this work with competitors that do not take into account the local structure of the problem. In particular, denote by *Local-△* the

[2]Assume the picture to be circular e.g. $[x]_{i,j} = [x]_{(i \mod 640),(j \mod 480)}$, to avoid technicalities on the boundary.

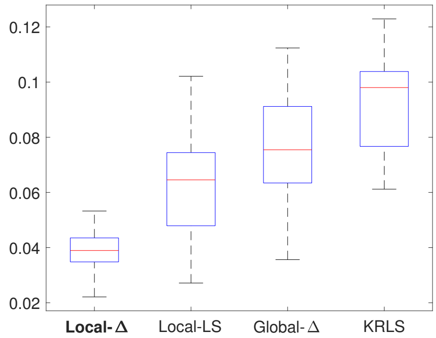

Figure 6: Learning the direction of ridges in fingerprint images: test error according to $\triangle$ in Eq. (24).

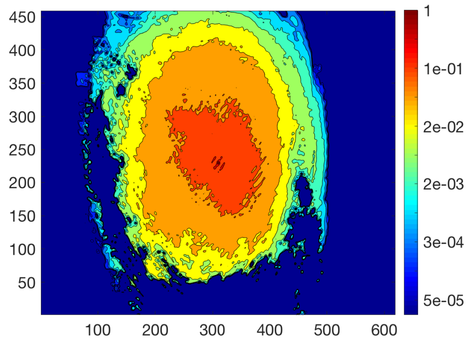

Figure 7: Empirical estimation of the *intra-locality* for the central patch of the fingerprints dataset.

proposed Alg. 1 with loss in Eq. (25); by *Local-LS* the same algorithm, but using loss in Eq. (25) with $g^2(\alpha, \beta) = (\alpha - \beta)^2$; by *Global-*$\triangle$ the structured prediction algorithm in [19] with loss in Eq. (24); by *KRLS*, the vector valued Kernel Ridge Regression estimator [28].

The above methods were trained on 50 examples and tested on the remaining 30 examples, the Gaussian kernel $K(u, u') = e^{-\frac{1}{2\sigma^2}\|u-u'\|^2}$ has been used for all the methods (in particular the restriction kernel in Eq. (20) with $\bar{K}$ Gaussian, for Alg. 1) and σ together with the regularization parameter λ have been choosen via cross validation, finally for Local-$\triangle$ and Local-LS we built and used an auxiliary set with $m = 30000$ (as described in Sec. 4), based on the 50 examples in the training set.

Fig. 5 reports three examples of the predictions on the test set, provided by the methods considered. It can be noticed that the learning process is remarkably improved when leveraging the parts in the data. Indeed, although provided with only 50 training examples, the predictions of our algorithm are remarkably similar to those of the desired output, while the other methods produce less accurate approximations. This is consistent with the result of Thm. 7, showing that when using part-based structured prediction the generalization error is reduced by a factor depending on the number of parts $|P|$, if the locality assumptions hold (Asm. 1, Asm. 2). This effect is evident in Fig. 6, which quantify the test error performed by the algorithms (in terms of the loss function in Eq. (24)), showing that part-based structured prediction Alg. 1 consistently outperforms the other methods.

Finally we stress the fact that both part-based learning and a structured approach seem to be crucial for reducing the learning error. Indeed from Fig. 6 it is clear that using the right loss Eq. (24), without exploiting the parts, is suboptimal (see Global-$\triangle$ in the figure), as using the parts without the right loss (see Local-LS in the figure).

Intra-locality. In Fig. 7 we visualize the intra-locality properties of one patch. In particular, denoting by p the central patch of the image, the figure shows the coefficient $C_{p,q}$ (defined in Lemma 6), with $q \in P$, and estimated on the whole dataset (The point i, j in the plot corresponds to $C_{p,q}$ with q the 20×20 patch centered in i, j). As it is possible to observe, there is a fast decay of the values depending on the distance from the patch p, suggesting that the intra-locality condition is well suited for this problem.

7 Conclusion

We presented a novel approach to structured prediction in presence of locality in the data. Our approach specializes the one in [19] in that it allows to incorporate knowledge about the parts directly within the learning model. We investigate the benefits provided by this model under assumptions on the unknown local relation between parts. In particular, by imposing a natural conditional independence assumption on the relation between input-output parts, our analysis provides a natural justification to the adoption of the so-called "restriction kernel", previously proposed in the literature, as a mean to lower the sample complexity of the problem. Furthermore, by imposing a low-correlation assumption on the parts of the input, we observe that the learning rates of our estimator can be significantly improved proportionally to the number of parts of the problem.

As a complementary result, we show that under mild assumptions on the problem the proposed estimator is also universally consistent and characterize its learning rates. This guarantees that while the proposed estimator is able to efficiently capture the local structure in the data, it is still able to solve the learning problem when the problem does not satisfy our assumptions.

References

[1] GH Bakir, T Hofmann, B Schölkopf, AJ Smola, B Taskar, and SVN Vishwanathan. Predicting structured data. neural information processing, 2007.

[2] Sebastian Nowozin, Christoph H Lampert, et al. Structured learning and prediction in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 2011.

[3] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[4] Karteek Alahari, Pushmeet Kohli, and Philip HS Torr. Reduce, reuse & recycle: Efficiently solving multi-label mrfs. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[5] John C Duchi, Lester W Mackey, and Michael I Jordan. On the consistency of ranking algorithms. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 327–334, 2010.

[6] Thorsten Joachims, Thomas Hofmann, Yisong Yue, and Chun-Nam Yu. Predicting structured objects with support vector machines. *Communications of the ACM*, 52(11):97–104, 2009.

[7] Matthew B Blaschko and Christoph H Lampert. Learning to localize objects with structured output regression. In *European conference on computer vision*, pages 2–15. Springer, 2008.

[8] Christoph H Lampert, Matthew B Blaschko, and Thomas Hofmann. Efficient subwindow search: A branch and bound framework for object localization. *IEEE transactions on pattern analysis and machine intelligence*, 31(12):2129–2142, 2009.

[9] Martin Szummer, Pushmeet Kohli, and Derek Hoiem. Learning crfs using graph cuts. In *European conference on computer vision*, pages 582–595. Springer, 2008.

[10] Lalit Bahl, Peter Brown, Peter De Souza, and Robert Mercer. Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.*, volume 11, pages 49–52. IEEE, 1986.

[11] Charles Sutton, Andrew McCallum, et al. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373, 2012.

[12] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. In *Journal of Machine Learning Research*, 2005.

[13] Nathan D Ratliff, J Andrew Bagnell, and Martin A Zinkevich. Maximum margin planning. In *Proceedings of the 23rd international conference on Machine learning*, pages 729–736. ACM, 2006.

[14] Devis Tuia, Jordi Muñoz-Marí, Mikhail Kanevski, and Gustavo Camps-Valls. Structured output svm for remote sensing image classification. *Journal of signal processing systems*, 65(3):301–310, 2011.

[15] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[16] Andrea Vedaldi and Andrew Zisserman. Structured output regression for detection with partial truncation. In *Advances in neural information processing systems*, pages 1928–1936, 2009.

[17] Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. Structured prediction theory based on factor graph complexity. In *Advances in Neural Information Processing Systems*, pages 2514–2522, 2016.

[18] Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, 2012.

[19] Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 4412–4420, 2016.

[20] Michael Collins. Parameter estimation for statistical parsing models: Theory and practice of distribution-free methods. In *New developments in parsing technology*, pages 19–55. Springer, 2004.

[21] Ben Taskar, Carlos Guestrin, and Daphne Koller. Max-margin markov networks. In *Advances in neural information processing systems*, pages 25–32, 2004.

[22] Corinna Cortes, Vitaly Kuznetsov, and Mehryar Mohri. Ensemble methods for structured prediction. In *International Conference on Machine Learning*, pages 1134–1142, 2014.

[23] Charles A Micchelli and Massimiliano Pontil. Kernels for multi–task learning. In *Advances in Neural Information Processing Systems*, pages 921–928, 2004.

[24] Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1):3520–3570, 2017.

[25] Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems*, pages 3891–3901, 2017.

[26] Carlo Ciliberto, Alessandro Rudi, Lorenzo Rosasco, and Massimiliano Pontil. Consistent multitask learning with nonlinear output relations. In *Advances in Neural Information Processing Systems*, pages 1983–1993, 2017.

[27] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Information Science and Statistics. Springer New York, 2008.

[28] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[29] Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, Massimiliano Pontil, and Steffen Grünewälder. Conditional mean embeddings as regressors. In *International Conference on Machine Learing (ICML)*, volume 5, 2012.

[30] Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

[31] H. Kadri, M. Ghavamzadeh, and P. Preux. A generalized kernel approach to structured output learning. *Proc. International Conference on Machine Learning (ICML)*, 2013.

[32] Florian Steinke, Matthias Hein, and Bernhard Schölkopf. Nonparametric regression between general riemannian manifolds. *SIAM Journal on Imaging Sciences*, 3(3):527–563, 2010.

[33] Charalambos D Aliprantis and Kim Border. *Infinite dimensional analysis: a hitchhiker's guide*. Springer Science & Business Media, 2006.

[34] VV Yurinskii. Exponential inequalities for sums of random vectors. *Journal of multivariate analysis*, 6(4):473–499, 1976.

[35] Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

[36] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

[37] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Supplementary Material:
Localized Structured Prediction

In this appendix we provide further background to the main discussion and results in the main sections of the current work. In particular:

- Appendix A introduces a generalization of the proposed framework to account for a larger family of structured prediction problems where locality can be exploited.
- Appendix B introduces the notation and auxiliary results that will be useful to prove the results discussed in this work.
- Appendix C discusses the derivation of the structured prediction estimator proposed and studied in this work.
- Appendix D extends the Comparison inequality for the SELF estimator in [19] to the case where the locality of the problem can be exploited.
- Appendix E provides an analytical decomposition of a bound for the excess risk of the proposed estimator that is then used to prove the learning rates of the proposed estimator without (Appendix F) and with parts (Appendix G) and also the universal consistency (Appendix H).
- Appendix I compares the proposed framework with structured prediction (without parts) in [19].
- Appendix J provides more details on the problem of learning and evaluating the estimator proposed in this work.
- Appendix K discusses in more detail loss functions considered in the literature that can be decomposed into "parts".

A Generalization of the Model by Parts

In this section we introduce a slight generalization of the model considered in this work and that will be used in the rest of the appendixes. In particular we consider the case where P is not necessarily finite and, possibly, the observed parts of y are not necessarily deterministic.

A.1 When the Parts don't correspond exactly

In general, y_p (the p-th part of y) could not be univocally determined given $p \in P$. For instance, consider a speech recognition problem where the goal is to predict the sentence pronounced by a speaker from an audio signal. In this setting the input space X is the set of all audio signals and $Y = Z$ is the set of all strings that can be produced in the speaker's language. In principle, for any part x_p of an input signal $x \in X$ it is possible to identify the corresponding part y_p of the target string. In practice, such a procedure would require significant preprocessing (e.g. using hidden markov models) and would however not be guaranteed to be error-free.

In general, given an input $x \in X$ a label $y \in Y$ and a part $p \in P$, observations for the p-th part of y can be distributed according to some probability $\mu(w|y, x, p)$ over the set $[Y]$ of parts of Y. A possible way to model this situation is to consider a characterization of L in terms of a further function $\ell : Z \times [Y] \times X \times P \to \mathbb{R}$ such that

$$\triangle(z, y|x) = \int_P L(z, y|x, p) d\pi(p|x), \quad \text{where} \tag{26}$$

$$L(z, y|x, p) = \int_{[Y]} \ell(z, \eta; x, p) \; d\mu(\eta|y, x, p). \tag{27}$$

In this sense, the distribution μ can be interpreted as characterizing how likely it is for the part p of an input x with associated label y to correspond to $\eta \in [Y]$. It is possible to recover the standard characterization by selecting μ to be the Dirac de

$$\mu(\eta|y, x, p) = \delta(\eta, y_p).$$

Remark 4 (Connection with standard Structured Prediction). *Note that the loss above generalizes the standard structured prediction framework as in [2, 12, 19]. Indeed, it is always possible to formulate a structured prediction loss $\triangle$ in the proposed setting, by taking $\ell = \triangle$ and $P = \{0\}$, $[Y] = Y$, $\pi(0|x) = 1$ and $\mu(w|y, x, 0) = \delta_y$. However, if there exists a non-trivial characterization of $\triangle$ in terms of these objects, then the algorithm proposed in this work is able to exploit this additional structure to achieve improved generalization performance.*

Here we give the extended defintion of the SELF assumption, given the definition of loss in Eq. (26)

Definition 2 (SELF by Parts (Extended)). *A function $\triangle : Z \times Y \times X \to \mathbb{R}$ is a* Structure Encoding Loss Function (SELF) by Parts *if it admits a factorization in the form of* (26) *with functions $\ell : Z \times [Y] \times X \times P \to \mathbb{R}$, and there exists a separable Hilbert space $\mathcal{H}$ and two bounded continuous maps $\psi : [Z] \times [X] \times P \to \mathcal{H}$, $\varphi : [Y] \to \mathcal{H}$ such that for any $z \in Z$, $\eta \in [Y]$, $x \in X$, $p \in P$*

$$\ell(z, \eta|x, p) = \langle \psi(z, x, p), \varphi(\eta) \rangle_{\mathcal{H}} . \tag{28}$$

Remark 5 (Def. 2 is more general than Def. 1). *Given a loss $\triangle$ satisfying Def. 1 for some $\psi', \phi, \mathcal{H}'$, then it satisfy Def. 2, with $\psi(z, x, p) = \psi'(z_p, z_p, p)$, with $\phi = \phi'$ with $\mathcal{H} = \mathcal{H}'$.*

B Notation and Main Definitions

Let $L^2(X \times P, \pi\rho_X)$ be the Lebesgue function space with norm

$$\|\beta\|^2_{L^2(X \times P, \pi\rho_X)} = \int_{X \times P} \beta(x, p)^2 \; d\pi(p|x) d\rho_X(x)$$

with $\beta : X \times P \to \mathbb{R}$. Analogously, $L^2(X \times P, \pi\rho_X, \mathcal{H})$ be the Lebesgue function space with norm

$$\|\beta\|^2_{L^2(X \times P, \pi\rho_X, \mathcal{H})} = \int_{X \times P} \|\beta(x, p)\|^2_{\mathcal{H}} \; d\pi(p|x) d\rho_X(x)$$

with $\beta : X \times P \to \mathcal{H}$. Let $\big((x_i, y_i)\big)_{i=1}^n$ be the training set and let $\big((x_{i_j}, y_{i_j}, p_j, w_j)\big)_{j=1}^m$. Denote with $\widehat{\rho}_X$ the probability measure $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. We define $L^2(X \times P, \pi\widehat{\rho}_X, \mathcal{H})$ the Lebesgue function space with norm

$$\|\beta\|^2_{L^2(X\times P, \pi\widehat{\rho}_X, \mathcal{H})} = \frac{1}{n}\sum_{i=1}^n \int_P \|\beta(x_i, p)\|^2_{\mathcal{H}} \; d\pi(p|x_i).$$

with $\beta : X \times P \to \mathcal{H}$.

Let $k : (X \times P) \times (X \times P) \to \mathbb{R}$ be a reproducing kernel with associated reproducing kernel Hilbert space (RKHS) $\mathcal{F}$. For any $(x, p) \in X \times P$ we denote $k_{x,p} = k\big((x, p), \cdot\big) \in \mathcal{F}$.

We introduce the following objects:

- $S : \mathcal{F} \to L^2(X \times P, \pi\rho_X)$ the operator such that, for any $f \in \mathcal{F}$,
$$(Sf)(\cdot, \cdot) = \Big\langle f, k_{(\cdot,\cdot)} \Big\rangle_{\mathcal{F}}.$$
- $S^* : L^2(X \times P, \pi\rho_X) \to \mathcal{F}$ the operator such that, for any $\beta \in L^2(X \times P, \pi\rho_X)$,
$$S^*\beta = \int_{X\times P} k_{x,p}\beta(x, p) \; d\pi(p|x)d\rho_X(x).$$
- $C : \mathcal{F} \to \mathcal{F}$ the operator $C = \int_{X\times P} k_{x,p} \otimes k_{x,p} \; d\pi(p|x)d\rho_X(x)$.
- $\widetilde{C} : \mathcal{F} \to \mathcal{F}$ the operator $\widetilde{C} = \frac{1}{n}\sum_{i=1}^n \int_P k_{x_i,p} \otimes k_{x_i,p} \; d\pi(p|x_i)$.
- $\widehat{C} : \mathcal{F} \to \mathcal{F}$ the operator $\widehat{C} = \frac{1}{m}\sum_{j=1}^m k_{x_{i_j},p_j} \otimes k_{x_{i_j},p_j}$.
- $L : L^2(X \times P, \pi\rho_X) \to L^2(X \times P, \pi\rho_X)$ the operator such that for any $\beta \in L^2(X \times P, \pi\rho_X)$, we have that $(L\beta)(\cdot) = \int_{X\times P} k\big((x, p), \cdot\big)\beta(x, p) \; d\pi(p|x)d\rho_X(x)$.
- $B : \mathcal{H} \to \mathcal{F}$ the operator $B = \int_{P\times X} k_{x,p} \otimes \varphi(w) \; d\mu(w|y, x, p)d\pi(p|x)d\rho(y, x)$. Note that by definiton $B = \int k_{x,p} \otimes g^*(x, p) \; d\pi(p|x)d\rho_X(x)$ with g^* defined as in Eq. (14).
- $\widehat{B} : \mathcal{H} \to \mathcal{F}$ the operator $\widehat{B} = \frac{1}{m}\sum_{j=1}^m k_{x_{i_j},p_j} \otimes \varphi(w_j)$.
- $G : \mathcal{H} \to L^2(X \times P, \pi\rho_X)$ the operator such that, for any $h \in \mathcal{H}$ is such that $(Gh)(\cdot) = \langle g^*(\cdot), h\rangle_{\mathcal{H}}$ for any $h \in \mathcal{H}$, with g^* defined as in Eq. (14).

Further Notation. Let $\mathcal{H}$ and $\mathcal{F}$ be two Hilbert spaces and let $h \in \mathcal{H}$ and $f \in \mathcal{F}$, we denote with $h \otimes f$ the bounded linear operator from $\mathcal{F} \to \mathcal{H}$ such that, for any $g \in \mathcal{F}$, we have $(h \otimes f)g = h\,\langle f, g\rangle_{\mathcal{F}}$. Note that $h \otimes f \in \mathcal{H} \otimes \mathcal{F}$, where $\mathcal{H} \otimes \mathcal{F}$ is the tensor product between the Hilbert spaces $\mathcal{H}, \mathcal{F}$ and is isometric to the the space of Hilbert-Schmidt operators from $\mathcal{F}$ to $\mathcal{H}$, denoted by $\mathrm{HS}(\mathcal{F}, \mathcal{H})$, namely the bounded linear operators $G : \mathcal{F} \to \mathcal{H}$ with finite Hilbert-Schmidt norm $\|G\|_{HS} = \sqrt{\mathrm{Tr}(G^*G)}$.

B.1 Auxiliary Results

Lemma 8. *With the notation introduced above, the following equations hold.*

$L = SS^*$.

$C = S^*S$.

$SC_\lambda^{-1}S^* = LL_\lambda^{-1} = I - \lambda L_\lambda^{-1}$.

$C_\lambda^{-1}S^* = S^*L_\lambda^{-1}$.

$\|C_\lambda^{-1/2}S^*\| = \|S^*L_\lambda^{-1/2}\| \leq 1$ *for any* $\lambda > 0$

The proof of the result above are well known and we refer to Appendix B in [19] for a proof with same notation as the one adopted in this paper. Below we show two further results that we will need

Lemma 9. *with the notation introduced above we have*

$$B = S^*G. \tag{29}$$

Proof. By applying the definition of the two operators S and G we have that for any $h \in \mathcal{H}$,

$$\begin{aligned} S^*Gh &= S^*((Gh)(\cdot)) && (30)\\ &= S^*(\langle g^*(\cdot), h\rangle_{\mathcal{H}}) && (31)\\ &= \int k_{x,p}\, \langle g^*(x,p), h\rangle_{\mathcal{H}}\ \, d\pi(p|x)d\rho_X(x) = \int (k_{x,p} \otimes g^*(x,p))h\ d\pi(p|x)d\rho_X(x) = Bh && (32) \end{aligned}$$

Hence $B = S^*G$ as required. □

C Derivation of the algorithm

In this section we show how the algorithm naturally derives from the definition of the problem and in particular we prove Lemma 3. Our analysis starts from the observation that when the loss function is SELF the solution of the learning problem in Eq. (4) is completely characterized in terms of the *conditional expectation* of $\varphi(y_p)$ given x, denoted by $g^* : X \times P \to \mathcal{H}$, with

$$g^*(x,p) = \int \varphi(\eta) d\mu(\eta|x,y,p) d\rho(y|x). \tag{33}$$

Note that since $\varphi(\cdot)$ is bounded and continuous, we have that $g^* \in L^2(X, \pi\rho_X, \mathcal{H})$.

Now we prove Lemma 3, in the extended version

Proof of Lemma 3. By Berge maximum theorem [33] (see also [19]), since Z is compact, we have that the solution of the learning problem in Eq. 4 is characterized by

$$f^*(x) = \operatorname*{argmin}_{z \in Z} \int \triangle(z, y|x) d\rho(y|x).$$

The result is obtained by expanding the definition of $\triangle$ with respect to SELF (Def. 2) and the linearity of the inner product and the integral

$$\int \triangle(z, y|x) d\rho(y|x) = \int \ell(z, \eta|x, p) d\mu(\eta|y, x, p) d\pi(p|x) d\rho(y|x) \tag{34}$$

$$= \int \langle \psi(z, x, p), \varphi(\eta) \rangle_{\mathcal{H}} \, d\mu(\eta|y, x, p) d\pi(p|x) d\rho(y|x) \tag{35}$$

$$= \int \left\langle \psi(z, x, p), \int \varphi(\eta) d\mu(\eta|y, x, p) d\rho(y|x) \right\rangle_{\mathcal{H}} d\pi(p|x) \tag{36}$$

$$= \int \langle \psi(z, x, p), g^*(x, p) \rangle_{\mathcal{H}} \, d\pi(p|x). \tag{37}$$

□

Since g^* depends on the unknown distribution ρ, we substitute it in Eq. (15) with an approximation $\widehat{g}$. In particular, since g^* is the conditional expectation induced by $\rho(y|x)$, a viable choice for $\widehat{g}$ is the *empirical risk minimizer* of the squared loss, which is a well known estimator for the conditional expectation [28], namely

$$\widehat{g} = \operatorname*{argmin}_{g \in \mathcal{G}} \frac{1}{m} \sum_{j=1}^{m} \|\psi(\eta_j) - g(\chi_j, p_j)\|_{\mathcal{H}}^2 + \lambda \|g\|_{\mathcal{G}}^2, \tag{38}$$

where $\mathcal{G}$ is a normed space of functions from $X \times P$ to $\mathcal{H}$. In this work we will consider $\mathcal{G} = \mathcal{H} \otimes \mathcal{F}$ where $\mathcal{F}$ is the space of functions associated to a kernel K on $X \times P$. In this case $\widehat{g}$ can be obtained in closed form in terms of the auxiliary dataset and, when plugged in Eq. (15), the resulting estimator corresponds exactly to the one in Eq. (6), as shown in next Lemma.

Lemma 10. *Let $\triangle$ be SELF, Z a compact set and K be a positive definite kernel on $X \times P$ and $\widehat{f}$ defined as in Eq.* (6) *with weights as in Eq.* (9) *computed using kernel K. Then $\widehat{f}$ is characterized by*

$$\widehat{f}(x) = \operatorname*{argmin}_{z \in Z} \sum_{p \in P} \pi(p|x) \, \langle \psi(z_p, x_p, p), \widehat{g}(x, p) \rangle_{\mathcal{H}}, \tag{39}$$

with $\widehat{g}$ the solution of Eq. (38) *computed using kernel K.*

Proof. We recall (see [28]) that the least-squares solution of Eq. (38) can be obtained in close form solution as

$$\widehat{g}(x, p) = \sum_{j=1}^{m} \alpha_j(x, p) \varphi(y_{p_j})$$

for any $x \in X$ and $p \in P$, where the weights α are defined as in Eq. (9). By linearity of the inner product we have

$$\sum_{p \in P} \pi(p|x) \, \langle \psi(z_p, x_p, p), \widehat{g}(x, p) \rangle_{\mathcal{H}} = \sum_{j=1}^{m} \sum_{p \in P} \pi(p|x) \alpha_j(x, p) \left\langle \psi(z_p, x_p, p), \varphi(y_{p_j}) \right\rangle_{\mathcal{H}} \tag{40}$$

$$= \sum_{j=1}^{m} \sum_{p \in P} \pi(p|x) \alpha_j(x, p) L_p(z_p, y_p|x_p) \tag{41}$$

where the last step follows from the assumption that the loss is SELF. □

An interesting consequence of the lemma above is that $\psi, \varphi, \widehat{g}, g^*, \mathcal{H}$ are only needed for theoretical purposes – i.e. to establish the connection between the estimator $\widehat{f}$ and the ideal solution f^* – and are not needed for the evaluation of $\widehat{f}$ which is done in terms of known objects, via Eq. (6).

D Comparison Inequality

In this we derive a result, Thm. 11, that is crucial to prove the statistical properties of the proposed algorithm. Note that it is analogous to the Comparison Inequality of [19] and of independent interest for the proposed framework. First we define the following estimator, that is a more general version of the one presented in the paper

$$\widehat{f}(x) = \operatorname*{argmin}_{z \in Z} \int_P \langle \psi(z, x, p), \widehat{g}(x, p) \rangle_{\mathcal{H}} \, \pi(p|x). \tag{42}$$

Note that the estimator presented in the main paper which is characterized by (39), Lemma 10 can be written like (42), applying Remark 5 in Appendix A.1.

Theorem 11. *When Z is a compact set and $\triangle$ satisfies Def. 2, for any measurable $\widehat{g} : X \times P \to \mathcal{H}$ and $\widehat{f} : X \to Z$ defined in terms of $\widehat{g}$ as in* (42). *Then*

$$\mathcal{E}(\widehat{f}\,) - \mathcal{E}(f^*) \leq c_{\triangle} \|\widehat{g} - g^*\|_{L^2(X \times P, \pi\rho_X, \mathcal{H})} \tag{43}$$

and $c_{\triangle}$ is a constant depending only on $\triangle$ and defined at the end of the proof.

Proof. For any $x \in X$ and $z \in Z$, let

$$A(z|x) = \int_P \langle \psi(z, x, p), g^*(x, p) \rangle_{\mathcal{H}} \; d\pi(p|x), \tag{44}$$

$$\widehat{A}(z|x) = \int_P \langle \psi(z, x, p), \widehat{g}(x, p) \rangle_{\mathcal{H}} \; d\pi(p|x). \tag{45}$$

By the SELF assumption $\ell(z, w|x, p) = \langle \psi(z, x, p), \varphi(w) \rangle_{\mathcal{H}}$ and the definition of g^* as in (14) we have the following alternative characterization for $A(z|x)$ as shown in Lemma 3

$$A(z|x) = \int_{[Y] \times Y \times P} \ell(z, w|x, p) \; d\mu(w|y, x, p) d\rho(y|x) d\pi(p|x). \tag{46}$$

Then, $\mathcal{E}(f) = \int_X A(f(x)|x) \; d\rho_X(x)$ for any $f : X \to Z$ and we have the following decomposition of the excess risk

$$\mathcal{E}(\widehat{f}) - \mathcal{E}(f^*) = \int_X A(\widehat{f}(x)|x) - A(f^*(x)|x) \; d\rho_X(x) \tag{47}$$

$$= \int_X A(\widehat{f}(x)|x) - \widehat{A}(\widehat{f}(x)|x) + \underbrace{\widehat{A}(\widehat{f}(x)|x) - \widehat{A}(f^*(x)|x)}_{\leq 0} \tag{48}$$

$$+ \int_X \widehat{A}(f^*(x)|x) - A(f^*(x)|x) \; d\rho_X(x) \tag{49}$$

$$\leq 2 \int_X \sup_{z \in Z} \left| \widehat{A}(z|x) - A(z|x) \right| \; d\rho_X(x) \tag{50}$$

where we have used the fact that $\widehat{A}(\widehat{f}(x)|x) - \widehat{A}(f^*(x)|x) \leq 0$ since, by definition, $\widehat{f}(x)$ is the minimizer of $\widehat{A}(\cdot|x)$ (see Eq. (42)).

Now, note that by the linearity of the inner product we have

$$\left|\widehat{A}(z|x) - A(z|x)\right| = \left|\int_P \langle \psi(z,x,p), \widehat{g}(x,p) - g^*(x,p)\rangle_{\mathcal{H}} \; d\pi(p|x)\right| \tag{51}$$

$$\leq \int_P \|\psi(z,x,p)\|_{\mathcal{H}} \; \|g^*(x,p) - \widehat{g}(x,p)\|_{\mathcal{H}} \; d\pi(p|x) \tag{52}$$

$$\leq \sqrt{\int_P \|\psi(z,x,p)\|^2_{\mathcal{H}} \; d\pi(p|x)} \sqrt{\int_P \|g^*(x,p) - \widehat{g}(x,p)\|^2_{\mathcal{H}} \; d\pi(p|x)} \tag{53}$$

$$= q(x,z) \; \sqrt{\int_P \|g^*(x,p) - \widehat{g}(x,p)\|^2_{\mathcal{H}} \; d\pi(p|x)} \tag{54}$$

where we applied Cauchy-Schwartz for each of the two inequalities, with $q(x,z) = \sqrt{\int_P \|\psi(z,x,p)\|^2_{\mathcal{H}} \; d\pi(p|x)}$.

Denote with $\|\cdot\|_{L^2(X\times P, \pi\rho_X, \mathcal{H})}$ the norm such that

$$\|g\|^2_{L^2(X\times P, \pi\rho_X, \mathcal{H})} = \int_{X\times P} \|g(x,p)\|^2_{\mathcal{H}} \; d\pi(p|x) d\rho_X(x), \tag{55}$$

for any $g : X \times P \to \mathcal{H}$. Then, plugging the inequality above in (50), we obtain

$$2\int_X \sup_{z\in Z} \left|\widehat{A}(z|x) - A(z|x)\right| \; d\rho_X(x) \tag{56}$$

$$\leq 2\int_X \sup_{z\in Z} \; \left[q(x,z) \; \sqrt{\int_P \|g^*(x,p) - \widehat{g}(x,p)\|^2_{\mathcal{H}} \; d\pi(p|x)} \;\right] \; d\rho_X(x) \tag{57}$$

$$= 2\int_X \sup_{z\in Z} \; \left[q(x,z)\right] \; \sqrt{\int_P \|g^*(x,p) - \widehat{g}(x,p)\|^2_{\mathcal{H}} \; d\pi(p|x)} \; d\rho_X(x) \tag{58}$$

$$\leq 2\sqrt{\int_X \left(\sup_{z\in Z} \; q(x,z)\right)^2 \; d\rho_X(x)} \sqrt{\int_{X\times P} \|g^*(x,p) - \widehat{g}(x,p)\|^2_{\mathcal{H}} \; d\pi(p|x) d\rho_X(x)} \tag{59}$$

$$= c_{\triangle} \|\widehat{g} - g^*\|_{L^2(X\times P, \pi\rho_X, \mathcal{H})} \tag{60}$$

where the last inequality follows from Cauchy-Schwartz and

$$c_{\triangle} = 2\;\sqrt{\int_X \left(\sup_{z\in Z} q(x,z)\right)^2 \; d\rho_X(x)} \tag{61}$$

$$= 2\;\sqrt{\int_X \; \sup_{z\in Z} \left[\int_P \|\psi(z,x,p)\|^2_{\mathcal{H}} \; d\pi(p|x)\right] \; d\rho_X(x)} \tag{62}$$

□

Remark 6 (Remove the dependency of $c_\triangle$ from ρ_X)**.** *Note that it is always possible to remove the dependency of* $c_\triangle$ *from* ρ_X *by bounding it with*

$$c_\triangle \leq 2\left(\sup_{\substack{z\in Z\\ x\in X}}\int_P \|\psi(z,x,p)\|^2_{\mathcal{H}}\, d\pi(p|x)\right)^{1/2} \tag{63}$$

E Analytical Decomposition

According to the comparison inequality Eq. (43) it is sufficient to bound the quantity $\|\widehat{g}-g^*\|_{L^2(X\times P,\pi\rho_X,\mathcal{H})}$ in order to control the excess risk of the estimator $\widehat{f}$. Equipped with the notation introduced above, we can now focus on studying this quantity. In particular in Thm. 13 we provide an analytical decomposition of $\|\widehat{g}-g^*\|_{L^2(X\times P,\pi\rho_X,\mathcal{H})}$ in terms of basic quantities that can be controlled in expectation (or probability, for the universal consistency).

Proposition 12. *Let* $\widehat{g}, g^*$ *be defined as in Eq. 38 and Eq. 33, then the following holds*

$$\|\widehat{g}-g^*\|_{L^2(X\times P,\pi\rho_X,\mathcal{H})} = \|S\widehat{C}_\lambda^{-1}\widehat{B} - G\|_{HS(\mathcal{H},L^2(X\times P,\pi\rho_X))} \tag{64}$$

Proof. First of all we recall that the space $L^2(X\times P,\pi\rho_X,\mathcal{H})$ is isometric to $\mathcal{H}\otimes L^2(X\times P,\pi\rho_X)$ which is isometric to the space of linear Hilbert-Schmidt operators from $\mathcal{H}\to L^2(X\times P,\pi\rho_X)$, denoted by $HS(\mathcal{H},L^2(X\times P,\pi\rho_X))$. Now note that G is the operator in $HS(\mathcal{H},L^2(X\times P,\pi\rho_X))$, that is isometric to $g^*\in L^2(X\times P,\pi\rho_X,\mathcal{H})$, indeed $Gv = \langle g^*(\cdot,\cdot),v\rangle_{\mathcal{H}}$, for any $v\in\mathcal{H}$.

Now note that is the solution of the problem in Eq. (38). Indeed, first note that the functional $\widehat{R}_\lambda(W)$, defining the problem in Eq. (38), is smooth and strongly convex ($W\in\mathcal{H}\otimes\mathcal{F}$, $\lambda>0$). Then we find the solution by equating the derivative of $\widehat{R}_\lambda(W)$ to 0. First note that for any $W\in\mathcal{H}\otimes\mathcal{F}$, the functional $\widehat{R}_\lambda(W)$, is equivalent to

$$\widehat{R}_\lambda(W) := \frac{1}{m}\sum_{j=1}^m \|\phi(w_j) - Wk_{(x_{i_j},p_j)}\|^2_{\mathcal{H}} + \lambda\|W\|_{\mathcal{H}\otimes\mathcal{F}} \tag{65}$$

$$= \mathrm{Tr}\Big[W\left(\frac{1}{m}\sum_{j=1}^m k_{(x_{i_j},p_j)}\otimes k_{(x_{i_j},p_j)} \ +\ \lambda I\right)W^* \tag{66}$$

$$-2\left(\frac{1}{m}\sum_{j=1}^m k_{(x_{i_j},p_j)}\otimes\phi(w_j)\right)W^* + \frac{1}{m}\sum_{j=1}^m \phi(w_j)\otimes\phi(w_j)\Big] \tag{67}$$

$$= \mathrm{Tr}\Big[W\left(\widehat{C}+\lambda I\right)W^* - 2\widehat{B}W + \frac{1}{m}\sum_{j=1}^m \phi(w_j)\otimes\phi(w_j)\Big], \tag{68}$$

where for the last step we applied the defintion of $\widehat{C}$ and $\widehat{B}$. By taking the derivative of $\widehat{R}_\lambda(W)$ in W and equating it to 0 the following minimizer is obtained $\widehat{W} = \widehat{B}^*\widehat{C}_\lambda^{-1}$.

Moreover note that, $S\widehat{C}_\lambda^{-1}\widehat{B}$ is the operator in $HS(\mathcal{H},L^2(X\times P,\pi\rho_X))$, that is isometric to $\widehat{g}\in L^2(X\times P,\pi\rho_X,\mathcal{H})$, indeed by definition of S

$$S\widehat{C}_\lambda^{-1}\widehat{B}v = \left\langle k_{(\cdot,\cdot)},\widehat{W}^*v\right\rangle_{\mathcal{F}} = \left\langle \widehat{W}k_{(\cdot,\cdot)},v\right\rangle_{\mathcal{H}} = \langle\widehat{g}(\cdot,\cdot),v\rangle_{\mathcal{H}}, \quad \forall v\in\mathcal{H}.$$

□

Theorem 13. *Let* $\lambda > 0$. *With the definitions in Sec. B, we have*

$$\|\widehat{g} - g^*\|_{L^2(X\times P,\pi\rho_X,\mathcal{H})} \leq \left(\frac{1}{\sqrt{\lambda}} + \frac{\beta_1^{1/2}}{\lambda}\right)\left(\beta_1\mathcal{A}_{1/2}(\lambda) + \beta_2\right) + \lambda\mathcal{A}_1(\lambda). \tag{69}$$

where $\beta_1 = \|C - \widehat{C}\|$, $\beta_2 = \|\widehat{B} - B\|_{HS}$ *and* $\mathcal{A}_r(\lambda) = \|L_\lambda^{-r}G\|_{HS}$ *for* $r > 0$.

Proof. By Prop. 12 and by adding and subtracting $S\widehat{C}_\lambda^{-1}B$ and $SC_\lambda^{-1}B$ we have

$$\|\widehat{g} - g^*\|_{L^2(X\times P,\pi\rho_X,\mathcal{H})} = \|S\widehat{C}_\lambda^{-1}\widehat{B} - G\|_{HS(\mathcal{H},L^2)} \leq A_1 + A_2 + A_3 \tag{70}$$

with

$$A_1 = \|S\widehat{C}_\lambda^{-1}\widehat{B} - S\widehat{C}_\lambda^{-1}B\|_{HS(\mathcal{H},L^2)} \tag{71}$$

$$A_2 = \|S\widehat{C}_\lambda^{-1}B - SC_\lambda^{-1}B\|_{HS(\mathcal{H},L^2)} \tag{72}$$

$$A_3 = \|SC_\lambda^{-1}B - G\|_{HS(\mathcal{H},L^2)}. \tag{73}$$

Bounding A_1. Now, by dividing and multiplying by $C_\lambda^{1/2}$, we have

$$A_1 = \|S\widehat{C}_\lambda^{-1}(\widehat{B} - B)\|_{HS(\mathcal{H},L^2)} \leq \|S\widehat{C}_\lambda^{-1}\|\|\widehat{B} - B\|_{HS(\mathcal{H},\mathcal{F})} \tag{74}$$

Bounding A_2. By using the identity $R^{-1} - T^{-1} = R^{-1}(T - R)T^{-1}$ holding for any invertible operators $R, T : \mathcal{F} \to \mathcal{F}$, we have

$$A_2 = \|S(\widehat{C}_\lambda^{-1} - C_\lambda^{-1})B\|_{HS(\mathcal{H},L^2)} \tag{75}$$

$$= \|S\widehat{C}_\lambda^{-1}(C_\lambda - \widehat{C}_\lambda)C_\lambda^{-1}B\|_{HS(\mathcal{H},L^2)} \tag{76}$$

$$= \|S\widehat{C}_\lambda^{-1}(C - \widehat{C})C_\lambda^{-1}B\|_{HS(\mathcal{H},L^2)} \tag{77}$$

$$\leq \|S\widehat{C}_\lambda^{-1}\|\|C - \widehat{C}\|\|C_\lambda{}^{-1}B\|_{HS(\mathcal{H},\mathcal{F})}. \tag{78}$$

$$\tag{79}$$

We further apply Lemma 8 to have $\|C_\lambda^{-1/2}S^*\| = \|S^*L_\lambda^{-1/2}\| \leq 1$ and $C_\lambda^{-1}S = S^*L_\lambda^{-1}$. Then,

$$\|C_\lambda^{-1}B\|_{HS(\mathcal{H},\mathcal{F})} = \|C_\lambda^{-1}S^*G\|_{HS(\mathcal{H},\mathcal{F})} = \|S^*L_\lambda^{-1}G\|_{HS(\mathcal{H},\mathcal{F})} \tag{80}$$

$$\leq \|S^*L_\lambda^{-1/2}\|\|L_\lambda^{-1/2}G\|_{HS(\mathcal{H},L^2)} \leq \|L_\lambda^{-1/2}G\|_{HS(\mathcal{H},L^2)}. \tag{81}$$

Bounding A_3. From Lemma 8 we have $B = S^*G$ and $SC_\lambda^{-1}S^* = LL_\lambda^{-1} = I - \lambda L_\lambda^{-1}$. Then,

$$A_3 = \|SC_\lambda^{-1}S^*G - G\|_{HS(\mathcal{H},L^2)} = \|(I - \lambda L_\lambda^{-1})G - G\|_{HS(\mathcal{H},L^2)} = \lambda\|L_\lambda^{-1}G\|_{HS(\mathcal{H},L^2)}. \tag{82}$$

To conclude, we control the term $\|S\widehat{C}_\lambda^{-1}\|$ by

$$\|S\widehat{C}_\lambda^{-1}\|^2 = \|\widehat{C}_\lambda^{-1}C\widehat{C}_\lambda^{-1}\| \leq \|\widehat{C}_\lambda^{-1}(C - \widehat{C})\widehat{C}_\lambda^{-1}\| + \|\widehat{C}_\lambda^{-1}\widehat{C}\widehat{C}_\lambda^{-1}\| \tag{83}$$

$$\leq \|\widehat{C}_\lambda^{-1}\|^2\|C - \widehat{C}\| + \frac{1}{\lambda} \tag{84}$$

$$\leq \frac{1}{\lambda^2}\|C - \widehat{C}\| + \frac{1}{\lambda} \tag{85}$$

Therefore

$$\|S\widehat{C}_\lambda^{-1}\| \leq \sqrt{\frac{\|C-\widehat{C}\|}{\lambda^2}+\frac{1}{\lambda}} \leq \frac{1}{\sqrt{\lambda}}+\frac{\sqrt{\|C-\widehat{C}\|}}{\lambda} \tag{86}$$

Combining the bounds for A_1, A_2 and A_3 we obtain the desired result.

□

F Learning Rates

In this section we focus on the analysis of the learning rates of the proposed estimator in expectation with respect to the sample of a training dataset. The main result of this section is Thm. 22, from which Thm. 4 is a corollary. Building on the analytic decomposition of Thm. 13 we observe that the key quantities to study in this setting are the $\mathbb{E}\|\widehat{C}-C\|^2$ and $\mathbb{E}\|\widehat{B}-B\|_{HS}^2$ as discussed below. In particular the following theorem further decomposes the quantities from Thm. 13, and $\mathbb{E}\|\widehat{C}-C\|^2$ and $\mathbb{E}\|\widehat{B}-B\|_{HS}^2$, are bounded in Appendices F.1 and F.2. Finally Thm. 22 is given in Appendix F.3.

Theorem 14. *Let $\lambda > 0$. With the definitions in Sec. B and Thm. 13, we have*

$$\mathbb{E}\|\widehat{g}-g^*\|_{L^2(X\times P,\pi\rho_X,\mathcal{H})} \leq 2\left(1+\frac{\sqrt{\mathbb{E}\beta_1^2}}{\lambda}\right)^{1/2}\left(\frac{\mathcal{A}_{1/2}(\lambda)^2\mathbb{E}\beta_1^2}{\lambda}+\frac{\mathbb{E}\beta_2^2}{\lambda}\right)^{1/2}+\lambda\mathcal{A}_1(\lambda). \tag{87}$$

Proof. Let $a=\frac{1}{\sqrt{\lambda}}$, $b=\frac{1}{\lambda}$, $c=\|L_\lambda^{-1/2}G\|_{HS}$ and $d=\lambda\|L_\lambda^{-1}G\|_{HS}$. Then,

$$\begin{aligned}
\mathbb{E}\|\widehat{g}-g^*\|_{L^2(X\times P,\pi\rho_X,\mathcal{H})} &\leq \mathbb{E}(a+b\beta_1^{1/2})(c\beta_1+\beta_2)+d && (88)\\
&\leq \sqrt{\mathbb{E}(a+b\beta_1^{1/2})^2\mathbb{E}(c\beta_1+\beta_2)^2}+d && (89)\\
&\leq \sqrt{4(a^2+b^2\mathbb{E}\beta_1)(c^2\mathbb{E}\beta_1^2+\mathbb{E}\beta_2^2)}+d && (90)\\
&\leq 2\sqrt{(a^2+b^2\sqrt{\mathbb{E}\beta_1^2})(c^2\mathbb{E}\beta_1^2+\mathbb{E}\beta_2^2)}+d && (91)
\end{aligned}$$

□

The rest of this section will be devoted to characterizing the behavior of $\mathbb{E}\beta_1^2$ and $\mathbb{E}\beta_2^2$ in order to obtain a more interpretable learning rates for the estimator proposed in this work.

F.1 Bounding $\mathbb{E}\beta_1^2$

Denote $\zeta_{x_{i_j},p_j}=k_{x_{i_j},p_j}\otimes k_{x_{i_j},p_j}-C$. First, we show that $\mathbb{E}\zeta_{x_{i_j},p_j}=0$.

Lemma 15. *With the definition above, when $x_1,\dots,x_n$ are identically distributed, we have*

$$\mathbb{E}\,\zeta_{x_{i_j},p_j}=0$$

Proof. Since $x_1, \ldots, x_n$ are identically distributed, for any $j = 1, \ldots, m$, we have

$$\mathbb{E}\, k_{x_{i_j},p_j} \otimes k_{x_{i_j},p_j} = \frac{1}{n} \sum_{i_j=1}^{n} \int_{P\times X} k_{x_{i_j},p_j} \otimes k_{x_{i_j},p_j}\ d\pi(p_j|x_{i_j}) d\rho_X(x_{i_j}) \tag{92}$$

$$= \int_{P\times X} k_{x,p} \otimes k_{x,p}\ d\pi(p|x) d\rho_X(x) \tag{93}$$

$$= C. \tag{94}$$

□

Lemma 16. *With the definitions of Section B let* $Q_1 = \mathbb{E}\|\zeta_{x,p}\|_{HS}^2$ *and*

$$\mathfrak{C} = \int_{P\times X} \zeta_{x,p}\zeta_{x,p'}\ d\pi(p|x) d\pi(p'|x) d\rho_X(x) \tag{95}$$

$$\mathbb{E}\|\widehat{C} - C\|_{HS}^2 = \frac{Q_1}{m} + \frac{(m-1)}{m}\frac{\mathrm{Tr}(\mathfrak{C})}{n}. \tag{96}$$

Proof. From the definition of $\widehat{C}$, we have

$$\mathbb{E}\|\widehat{C} - C\|_{HS}^2 = \mathbb{E}\|\frac{1}{m}\sum_{j=1}^{m} \zeta_{x_{i_j},p_j}\|_{HS}^2 = \frac{1}{m^2}\sum_{j,h=1}^{m} \mathbb{E}\ \mathrm{Tr}\left(\ \zeta_{x_{i_j},p_j}\zeta_{x_{i_h},p_h}\right) \tag{97}$$

We consider separately the elements in the sum that correspond to the case $j = h$ and $j \neq h$.

1. Case $j = h$**.** We have

$$\mathbb{E}\ \mathrm{Tr}\left(\zeta_{x_{i_j},p_j}\zeta_{x_{i_h},p_h}\right) = \mathbb{E}\|\zeta_{x_{i_j},p_j}\|_{HS}^2 = Q_1 \tag{98}$$

2. Case $j \neq h$ We have $\mathbb{E}\ \mathrm{Tr}\left(\zeta_{x_{i_j},p_j}\zeta_{x_{i_h},p_h}\right) = \frac{1}{n^2}\sum_{i_j,i_h=1}^{n} R_{i_j,i_h}^{j,h}$ where

$$R_{u,v}^{j,h} = \int_{P\times X} \mathrm{Tr}(\zeta_{x_u,p_j}\zeta_{x_v,p_h})\ d\pi(p_j|x_u) d\pi(p_h|x_v) d\rho_X(x_1)\cdots d\rho_X(x_n). \tag{99}$$

We consider separately the case $i_j = i_h$ and $i_j \neq i_h$.

2.1 Case $j \neq h$ **and** $i_j = i_h$**.** We have that

$$R_{i_j,i_j}^{j,h} = \int_{P\times X} \mathrm{Tr}\left(\zeta_{x_{i_j},p_j}\zeta_{x_{i_j},p_h}\right)\ d\pi(p_j|x_{i_j}) d\pi(p_h|x_{i_j}) d\rho_X(x_{i_j}) \tag{100}$$

$$= \int_{P\times X} \mathrm{Tr}\left(\zeta_{x,p}\zeta_{x,p'}\right)\ d\pi(p|x) d\pi(p'|x) d\rho_X(x) = \mathrm{Tr}(\mathfrak{C}). \tag{101}$$

2.2 Case $j \neq h$ **and** $i_j \neq i_h$**.** We have that

$$R^{j,h}_{i_j,i_h} = \int \operatorname{Tr}\left(\zeta_{x_{i_j},p_j}\zeta_{x_{i_h},p_h}\right)\, d\pi(p_j|x_{i_j})d\pi(p_h|x_{i_h})d\rho_X(x_{i_j})d\rho_X(x_{i_h}) \tag{102}$$

$$= \int \operatorname{Tr}\left(\zeta_{x,p}\zeta_{x',p'}\right)\, d\pi(p|x)d\pi(p'|x')d\rho_X(x)d\rho_X(x') \tag{103}$$

$$= \operatorname{Tr}\left(\int \zeta_{x,p}\, d\pi(p|x)d\rho_X(x) \int \zeta_{x',p'}\, d\pi(p'|x')d\rho_X(x')\right) \tag{104}$$

$$= \|\mathbb{E}\, \zeta_{x,p}\|^2_{HS} = 0 \tag{105}$$

where the last equality follows from the fact that the $\zeta_{x,p}$ have zero mean according to Lemma 15.

Combining the above cases. Note that in (97), Case 1 occurs m times and Case 2 occurs the remaining $m(m-1)$ times. Therefore, we have

$$\mathbb{E}\|\widehat{C} - C\|^2_{HS} = \frac{Q_1}{m} + \frac{m-1}{m}\frac{1}{n^2}\sum_{i_j,i_h=1}^{n} R^{j,h}_{i_j,i_h} \tag{106}$$

Now, for the second term on the right hand side, Case 2.1 occurs n times while Case 2.2 occurs the remaining $n(n-1)$ times, leading to the desired result. □

Lemma 17. *With the notation of Lemma 16 and the definition of* q *in* (17)*, we have*

$$\operatorname{Tr}(\mathfrak{C}) = \mathfrak{c}_1 - \mathfrak{c}_2 = q, \tag{107}$$

where

$$\mathfrak{c}_1 = \int k\big((x,p),(x,p')\big)^2\, d\pi(p|x)d\pi(p'|x)d\rho_X(x) \tag{108}$$

$$\mathfrak{c}_2 = \int k\big((x,p),(x',p')\big)^2\, d\pi(p|x)d\pi(p'|x')d\rho_X(x)d\rho_X(x'). \tag{109}$$

Proof. Note that by definition of ζ and the reproducing property of the kernel k, for any $x, x' \in X$ and $p, p' \in P$ the following holds

$$\operatorname{Tr}(\zeta_{x,p}\zeta_{x',p'}) = k\big((x,p),(x',p')\big)^2 - \operatorname{Tr}\left(C\left(k_{x,p}\otimes k_{x,p}\right)\right) \tag{110}$$

$$- \operatorname{Tr}\left(C\left(k_{x',p'}\otimes k_{x',p'}\right)\right) + \operatorname{Tr}(C^2). \tag{111}$$

Then, by definition of $C = \mathbb{E}\, k_{x,p}\otimes k_{x,p}$, we have

$$\operatorname{Tr}(\mathfrak{C}) = \int \operatorname{Tr}\left(\zeta_{x,p}\zeta_{x,p'}\right)\, d\pi(p|x)d\pi(p'|x)d\rho_X(x) \tag{112}$$

$$= -\operatorname{Tr}(C^2) + \int k\big((x,p),(x,p')\big)^2\, d\pi(p|x)d\pi(p'|x)d\rho_X(x) \tag{113}$$

$$= -\operatorname{Tr}(C^2) + \int k\big((x,p),(x,p')\big)^2\, d\pi(p|x)d\pi(p'|x)d\rho_X(x) \tag{114}$$

$$= \mathfrak{c}_1 - \operatorname{Tr}(C^2). \tag{115}$$

To conclude,

$$\operatorname{Tr}(C^2) = \operatorname{Tr}\left(\left(\int k_{x,p} \otimes k_{x,p} \, d\pi(p|x) d\rho_X(x)\right)\left(\int k_{x',p'} \otimes k_{x',p'} \, d\pi(p'|x') d\rho_X(x')\right)\right) \tag{116}$$

$$= \int k((x,p),(x',p'))^2 \, d\pi(p|x) d\pi(p'|x') d\rho_X(x) d\rho_X(x') \tag{117}$$

$$= \mathfrak{c}_2. \tag{118}$$

The last step consists in noting that $\mathfrak{c}_1 - \mathfrak{c}_2$ is exactly the definition of q in (17).

□

F.2 Bounding $\mathbb{E}\beta_2^2$

The analysis for $\mathbb{E}\beta_2^2$ is analogous to that of $\mathbb{E}\beta_1^2$. For completeness we report it below.

Denote $\eta_{x_{i_j},p_j,w_j} = k_{x_{i_j},p_j} \otimes \varphi(w_j) - B$. We show that $\mathbb{E}\, \eta_{x_{i_j},p_j,w_j} = 0$.

Lemma 18. *With the definition above, when $x_1, \dots, x_n$ are identically distributed, we have*

$$\mathbb{E}\, \eta_{x_{i_j},p_j,w_j} = 0$$

Proof. Since $x_1, \dots, x_n$ are identically distributed, for any $j = 1, \dots, m$, we have

$$\mathbb{E}\, k_{x_{i_j},p_j} \otimes \varphi(w_j) = \frac{1}{n}\sum_{i_j=1}^{n} \int k_{x_{i_j},p_j} \otimes \varphi(w_j) \, d\mu(w_j|y_{i_j}, x_{i_j}, p_j) d\pi(p_j|x_{i_j}) d\rho(y_{i_j}, x_{i_j}) \tag{119}$$

$$= \int k_{x,p} \otimes \varphi(w) \, d\mu(w|y,x,p) d\pi(p|x) d\rho(y,x) \tag{120}$$

$$= B. \tag{121}$$

□

Lemma 19. *Let $Q_2 = \mathbb{E}\|\eta_{x,p,w}\|_{HS}^2$ and*

$$\mathfrak{B} = \int \eta^*_{x,p,w}\eta_{x,p',w'} \, d\mu(w|y,x,p) d\mu(w'|y,x,p') d\pi(p|x) d\pi(p'|x) d\rho(y,x) \tag{122}$$

$$\mathbb{E}\|\widehat{B} - B\|_{HS}^2 = \frac{Q_2}{m} + \frac{(m-1)}{m}\frac{\operatorname{Tr}(\mathfrak{B})}{n}. \tag{123}$$

Proof. From the definition of $\widehat{B}$, we have

$$\mathbb{E}\|\widehat{B} - B\|_{HS}^2 = \mathbb{E}\|\frac{1}{m}\sum_{j=1}^{m} \eta_{x_{i_j},p_j,w_j}\|_{HS}^2 = \frac{1}{m^2}\sum_{j,h=1}^{m} \mathbb{E}\operatorname{Tr}\left(\eta^*_{x_{i_j},p_j,w_j}\eta_{x_{i_h},p_h,w_h}\right) \tag{124}$$

We consider separately the elements in the sum that correspond to the case $j = h$ and $j \neq h$.

1. Case $j = h$. We have

$$\mathbb{E}\ \mathrm{Tr}\left(\eta^*_{x_{i_j},p_j,w_j}\eta_{x_{i_h},p_h,w_h}\right) = \mathbb{E}\|\eta_{x_{i_j},p_j,w_j}\|^2_{HS} = Q_2. \tag{125}$$

2. Case $j \neq h$ We have $\mathbb{E}\ \mathrm{Tr}\left(\eta^*_{x_{i_j},p_j,w_j}\eta_{x_{i_h},p_h,w_h}\right) = \frac{1}{n^2}\sum_{i_j,i_h=1}^n Z^{j,h}_{i_j,i_h}$ where

$$Z^{j,h}_{u,v} = \int \mathrm{Tr}(\eta^*_{x_u,p_j,w_j}\eta_{x_v,p_h,w_h})\ d\mu(w_j|y_{i_j},x_{i_j},p_j)d\mu(w_h|y_{i_h},x_{i_h},p_h)\times \tag{126}$$

$$\times\ d\pi(p_j|x_u)d\pi(p_h|x_v)d\rho(y_1,x_1)\cdots d\rho(y_n,x_n). \tag{127}$$

We consider separately the case $i_j = i_h$ and $i_j \neq i_h$.

2.1 Case $j \neq h$ **and** $i_j = i_h$. We have that

$$Z^{j,h}_{i_j,i_j} = \int \mathrm{Tr}\left(\eta^*_{x_{i_j},p_j,w_j}\eta_{x_{i_j},p_h,w_h}\right)\ d\mu(w_j|y_{i_j},x_{i_j},p_j)d\mu(w_h|y_{i_j},x_{i_j},p_h)\times \tag{128}$$

$$\times\ d\pi(p_j|x_{i_j})d\pi(p_h|x_{i_j})d\rho(y_{i_j},x_{i_j}) \tag{129}$$

$$= \int \mathrm{Tr}\left(\eta^*_{x,p,w}\eta_{x,p',w'}\right)\ d\mu(w|y,x,p)d\mu(w'|y,x,p')d\pi(p|x)d\pi(p'|x)d\rho(y,x) \tag{130}$$

$$= \mathrm{Tr}(\mathfrak{B}). \tag{131}$$

2.2 Case $j \neq h$ **and** $i_j \neq i_h$. We have that

$$Z^{j,h}_{i_j,i_h} = \int \mathrm{Tr}\left(\eta^*_{x_{i_j},p_j,w_j}\eta_{x_{i_h},p_h,w_h}\right)\ d\mu(w_j|y_{i_j},x_{i_j},p_j)d\mu(w_h|y_{i_h},x_{i_h},p_h)\times \tag{132}$$

$$\times\ d\pi(p_j|x_{i_j})d\pi(p_h|x_{i_h})d\rho(y_{i_j},x_{i_j})d\rho(y_{i_h},x_{i_h}) \tag{133}$$

$$= \int \mathrm{Tr}\left(\eta^*_{x,p,w}\eta_{x',p',w'}\right)\ d\mu(w|y,x,p)d\mu(w'|y',x',p')\times \tag{134}$$

$$\times\ d\pi(p|x)d\pi(p'|x')d\rho(y,x)d\rho(y',x') \tag{135}$$

$$= \mathrm{Tr}\Bigg(\Big(\int \eta^*_{x,p,w}\ d\mu(w|y,x,p)d\pi(p|x)d\rho(y,x)\times \tag{136}$$

$$\times \int \eta_{x',p',w'}\ d\mu(w'|y',x',p')d\pi(p'|x')d\rho(y',x')\Big)\Bigg) \tag{137}$$

$$= \|\mathbb{E}\ \eta_{x,p,w}\|^2_{HS} = 0, \tag{138}$$

where the last equality follows from the fact that the $\eta_{x,p,w}$ have zero mean according to Lemma 18.

Combining the above cases. Note that in (124), Case 1 occurs m times and Case 2 occurs the remaining $m(m-1)$ times. Therefore, we have

$$\mathbb{E}\|\widehat{B} - B\|^2_{HS} = \frac{Q_2}{m} + \frac{m-1}{m}\frac{1}{n^2}\sum_{i_j,i_h=1}^{n} Z^{j,h}_{i_j,i_h} \tag{139}$$

Now, for the second term on the right hand side, Case 2.1 occurs n times while Case 2.2 occurs the remaining $n(n-1)$ times, leading to the desired result. □

Lemma 20. *With the notation of Lemma 19, we have*

$$\mathrm{Tr}(\mathfrak{B}) = \mathfrak{b}_1 - \mathfrak{b}_2 \tag{140}$$

where

$$\mathfrak{b}_1 = \int \langle g^*(x,p), g^*(x,p')\rangle_{\mathcal{H}} \; k\big((x,p),(x,p')\big) \; d\pi(p|x)d\pi(p'|x)d\rho_X(x) \tag{141}$$

$$\mathfrak{b}_2 = \int \langle g^*(x,p), g^*(x',p')\rangle_{\mathcal{H}} \; k\big((x,p),(x',p')\big) \; d\pi(p|x)d\pi(p'|x')d\rho_X(x)d\rho_X(x'). \tag{142}$$

Proof. Note that by definition of η and the reproducing property of the kernel k, for any $x, x' \in X$, $p, p' \in P$ and $w, w' \in [Y]$ the following holds

$$\mathrm{Tr}(\eta^*_{x,p,w}\eta_{x',p',w'}) = \langle \varphi(w), \varphi(w')\rangle_{\mathcal{H}} k\big((x,p),(x',p')\big) - \mathrm{Tr}\left(B^*\left(k_{x,p}\otimes\varphi(w)\right)\right) \tag{143}$$

$$- \mathrm{Tr}\left(B^*\left(k_{x',p'}\otimes\varphi(w')\right)\right) + \mathrm{Tr}(B^*B). \tag{144}$$

Then, by definition of $B = \mathbb{E}\, k_{x,p}\otimes\varphi(w)$, we have

$$\mathrm{Tr}(\mathfrak{B}) = \int \mathrm{Tr}\left(\eta^*_{x,p,w}\eta_{x,p',w'}\right) \; d\mu(w|y,x,p)d\mu(w'|y,x,p')d\pi(p|x)d\pi(p'|x)d\rho(y,x) \tag{145}$$

$$= -\mathrm{Tr}(B^*B) + \int \langle\varphi(w),\varphi(w')\rangle_{\mathcal{H}} k\big((x,p),(x',p')\big) \; d\mu(w|y,x,p)d\mu(w'|y,x,p')\times \tag{146}$$

$$\times\; d\pi(p|x)d\pi(p'|x)d\rho(y,x) \tag{147}$$

$$= -\mathrm{Tr}(B^*B) + \int \langle g^*(x,p), g^*(x,p')\rangle_{\mathcal{H}} \; k\big((x,p),(x,p')\big) \; d\pi(p|x)d\pi(p'|x)d\rho_X(x) \tag{148}$$

$$= \mathfrak{b}_1 - \mathrm{Tr}(B^*B), \tag{149}$$

where in the third equality we used the definition of $g^*(x,p) = \int \varphi(w)\; d\mu(w|y,x,p)d\rho(y|x)$. Moreover, since B can be written in terms of g^* as

$$B = \int k_{x,p}\otimes g^*(x,p)\; d\pi(p|x)d\rho_X(x) \tag{150}$$

we have

$$\mathrm{Tr}(B^*B) = \int \langle g^*(x,p), g^*(x',p')\rangle_{\mathcal{H}} \; k\big((x,p),(x',p')\big) \; d\pi(p|x)d\pi(p'|x')d\rho_X(x)d\rho_X(x') \tag{151}$$

$$= \mathfrak{b}_2. \tag{152}$$

□

F.3 Learning bound in expectation

We introduce here the assumption that the target function g^* of the learning problem belongs to the RKHS where we are performing the optimization.

Assumption 3. *There exists a* $G \in \mathcal{H} \otimes \mathcal{F}$*, such that almost everywhere on* $X \times P$*,*

$$Gk_{x,p} = g^*(x,p).$$

The following results will leverage the assumption above.

Lemma 21. *Under Assumption 3,*

$$\mathrm{Tr}(\mathfrak{B}) \quad \leq \quad \|G\|^2 \mathrm{Tr}(\mathfrak{C}), \tag{153}$$

Proof. We begin first observing that $\mathfrak{C}$ is positive semidefinite since

$$\mathfrak{C} = \int \zeta_{x,p}\zeta_{x,p'} \; d\pi(p|x)d\pi(p'|x)d\rho_X(x) = \mathbb{E}\; \zeta_x\zeta_x \tag{154}$$

is the expectation of the random variable $\zeta_x\zeta_x$, where $\zeta_x = \int \zeta_{x,p} \; d\pi(p|x)$ is positive semidefinite. Moreover, by the definition of $\mathfrak{C}$ in terms of $\zeta_{x,p} = k_{x,p} \otimes k_{x,p} - C$, we have

$$\mathfrak{C} = \int \left(k_{x,p} \otimes k_{x,p'}\right) k((x,p),(x,p')) - \left(k_{x,p} \otimes k_{x,p}\right) C \; d\pi(p|x)\pi(p'|x)\rho_X(x) \tag{155}$$

$$+ \int C^2 - C\left(k_{x,p'} \otimes k_{x,p'}\right) d\pi(p|x)\pi(p'|x)\rho_X(x) \tag{156}$$

$$= -C^2 + \int \left(k_{x,p} \otimes k_{x,p'}\right) k((x,p),(x,p')) \; d\pi(p|x)\pi(p'|x)\rho_X(x) \tag{157}$$

where we have used the definition of $C = \mathbb{E}\; k_{x,p} \otimes k_{x,p}$.

Now note that under Assumption 3, for any $x, x' \in X$ and $p, p' \in P$

$$\langle g^*(x,p), g^*(x',p')\rangle_{\mathcal{H}} = \langle Gk_{x,p}, Gk_{x',p'}\rangle_{\mathcal{H}} = \mathrm{Tr}\left(G^*G\left(k_{x,p} \otimes k_{x',p'}\right)\right). \tag{158}$$

Therefore, substituting the above equation in $\mathfrak{b}_1$ and $\mathfrak{b}_2$ defined in Lemma 20, we have

$$\mathrm{Tr}(\mathfrak{B}) = \mathfrak{b}_1 - \mathfrak{b}_2 \tag{159}$$

$$= \mathrm{Tr}\left(G^*G\left[\left[\int \left(k_{x,p} \otimes k_{x,p'}\right) k((x,p),(x,p')) \; d\pi(p|x)\pi(p'|x)\rho_X(x) - C^2\right]\right]\right) \tag{160}$$

$$= \mathrm{Tr}(G^*G\;\mathfrak{C}) \tag{161}$$

$$\leq \|G\|^2 \mathrm{Tr}(\mathfrak{C}) \tag{162}$$

where the last inequality follows from the fact that both G^*G and $\mathfrak{C}$ are positive semidefinite. □

Theorem 22. *When* $\triangle$ *is SELF and* Z *is a compact set, under Assumption 3, and the notation in Eqs.* (16) *and* (17) *we have*

$$\mathbb{E}\;\mathcal{E}(\widehat{f}\,) - \mathcal{E}(f^*) \quad \leq \quad 2\, c_{\triangle} g \left[\lambda^{1/2} + 2\sqrt{2}\left(1 + \left(\frac{r^2}{\lambda^2 m} + \frac{q}{\lambda^2 n}\right)^{1/2}\right)^{1/2}\left(\frac{r^2}{\lambda m} + \frac{q}{\lambda n}\right)^{1/2}\right].$$

In particular when $\lambda \geq \sqrt{\frac{r^2}{m} + \frac{q}{n}}$, *then*

$$\mathbb{E}\, \mathcal{E}(\widehat{f}\,) - \mathcal{E}(f^*) \quad \leq \quad 12\, c_{\triangle} g \, \left(\frac{r^2}{\lambda m} + \frac{q}{\lambda n} + \lambda\right)^{1/2}.$$

Proof. By the comparison inequality in Thm. 11, we have that

$$\mathbb{E}\, \mathcal{E}(\widehat{f}\,) - \mathcal{E}(f^*) \leq 2c_{\triangle}\, \mathbb{E}\|\widehat{g} - g^*\|_{L^2(X\times P, \pi\rho_X, \mathcal{H})}.$$

To bound $\mathbb{E}\|\widehat{g} - g^*\|_{L^2(X\times P, \pi\rho_X, \mathcal{H})}$ we need to control some auxiliary quantities. With the notation of Thm. 13 and Lemmas 16, 19 and 21, we have

$$\mathbb{E}\beta_1^2 \leq \frac{Q_1}{m} + \frac{\mathrm{Tr}(\mathfrak{C})}{n} =: V, \quad \mathbb{E}\beta_2^2 \leq \|G\|V.$$

In particular note that $\mathrm{Tr}(\mathfrak{C}) = q$, by Lemma 17 and that by definition of Q_1, r and C we have

$$\begin{aligned}
Q_1 &:= \mathbb{E} k_{x,p} \otimes k_{x,p} - C^2_{HS} && (163)\\
&= \mathrm{Tr}\left(\mathbb{E}\; k((x,p),(x,p))(k_{x,p} \otimes k_{x,p}) - 2C(k_{x,p} \otimes k_{x,p}) + C^2\right) && (164)\\
&= \mathrm{Tr}\left(\mathbb{E}\; k((x,p),(x,p))(k_{x,p} \otimes k_{x,p}) - C^2\right) \leq r\mathrm{Tr}\left(\mathbb{E}\;(k_{x,p} \otimes k_{x,p})\right) \leq r^2. && (165)
\end{aligned}$$

Moreover, by Assumption 3 we have that $G = SG$ and so

$$\mathcal{A}_{1/2}(\lambda) = \|L_\lambda^{-1/2} G\|_{HS(\mathcal{H}, L^2)} = \|L_\lambda^{-1/2} SG\|_{HS(\mathcal{H}, L^2)} \leq \|L_\lambda^{-1/2} S\|\|G\|_{HS(\mathcal{H},\mathcal{F})} \leq \|G\|_{HS(\mathcal{H},\mathcal{F})}.$$

Analogously

$$\mathcal{A}_1(\lambda) = \|L_\lambda^{-1} G\|_{HS(\mathcal{H}, L^2)} \leq \|L_\lambda^{-1/2}\|\|L_\lambda^{-1/2} G\|_{HS(\mathcal{H}, L^2)} = \lambda^{-1/2}\mathcal{A}_{1/2}(\lambda) \leq \lambda^{-1/2}\|G\|_{HS(\mathcal{H},\mathcal{F})}.$$

By plugging the bounds above in the result of Thm. 14, we have

$$\mathbb{E}\|\widehat{g} - g^*\|_{L^2(X\times P, \pi\rho_X, \mathcal{H})} \leq 2\sqrt{2}\|G\|_{HS(\mathcal{H},\mathcal{F})}\sqrt{1 + \frac{V^{1/2}}{\lambda}}\sqrt{\frac{V}{\lambda}} + \|G\|_{HS(\mathcal{H},\mathcal{F})}\lambda^{1/2}.$$

By selecting $\lambda \geq V^{1/2}$, we have

$$\begin{aligned}
\mathbb{E}\|\widehat{g} - g^*\|_{L^2(X\times P, \pi\rho_X, \mathcal{H})} &\leq 4\|G\|_{HS(\mathcal{H},\mathcal{F})}\sqrt{\frac{V}{\lambda}} + \|G\|_{HS(\mathcal{H},\mathcal{F})}\lambda^{1/2} && (166)\\
&\leq 4\|G\|_{HS(\mathcal{H},\mathcal{F})}\left(\sqrt{\frac{V}{\lambda}} + \lambda^{1/2}\right) && (167)\\
&\leq 4\sqrt{2}\|G\|_{HS(\mathcal{H},\mathcal{F})}\left(\frac{V}{\lambda} + \lambda\right)^{1/2}, && (168)
\end{aligned}$$

since $a^{1/2} + b^{1/2} \leq \sqrt{2(a+b)}$ for any $a, b > 0$. □

F.4 Proof of Theorem 4

Proof. Theorem 4 corresponds to the second statement of Theorem 22. □

G Learning Rates with the effect of parts

In this part we start from the results of Thm. 4 and study the effect of interlocality and intra-locality. Lemma 6 is essentially a corollary of Lemma 25 and it is proven in Section G.1. Finally the proof of Thm 7, is given in Section G.2 and it is based on Thm. 4 and results from this section.

We consider here the natural generalization of inter-locality Asm. 1 to the case where the parts of y are sampled non-deterministically from μ.

Assumption 4. *There exist two spaces* $[X]$ *and* $[Y]$ *of parts on* X *and* Y *respectively and a conditional probability distribution* $\bar{\mu}$ *on* $[Y]$ *with respect to* $[X]$*, such that*

$$\bar{\mu}(w|x_p) = \int \mu(w|y,x,p)d\rho(y|x) \tag{169}$$

Clearly, Asm. 4 formalizes the concept of inter-locality and recovers it when μ corresponds to

$$\mu(\cdot|y,x,p) = \delta_{y_p}(\cdot) \tag{170}$$

where δ denotes the Dirac's delta on the point $y_p \in [Y]$. Indeed, in this case we are requiring $w = y_p$ to depend exclusively on x_p for any $p \in P$, hence to be conditionally independent with respect to x. Moreover, we are requiring such distribution $\bar{\mu}$ to be the same for any $p \in P$, hence recovering Asm. 1. The following result is therefore a generalization of Lemma 5, which is recovered as a corollary.

Lemma 23. *Under Assumption 4,* g^* *is such that* $g^*(x,p) = \bar{g}^*(x_p)$ *for any* $x \in X$ *and* $p \in P$*, where* $\bar{g}^* : [X] \to \mathcal{H}$ *is such that*

$$\bar{g}^*(\xi) = \int \varphi(w)\ d\bar{\mu}(w|\xi) \tag{171}$$

almost surely on $[X]$.

Proof. The result follows directly from Assumption 4 and the definition of g^*

$$g^*(x,p) = \int \varphi(w)\ d\mu(w|y,x,p)d\rho(y|x) = \int \varphi(w)\ d\bar{\mu}(w|x_p) = \bar{g}^*(x_p). \tag{172}$$

□

Assumption 5. *Denote by* $\bar{k} : [X] \times [X] \to \mathbb{R}$ *the reproducing kernel on* $[X]$ *with associated rkhs denoted by* $\overline{\mathcal{G}}$*, defined as for all* $x, x' \in X$ *and* $p, p' \in P$

$$k((x,p),(x',p')) = \bar{k}(x_p, x'_{p'}) \tag{173}$$

Assumption 6. *There exists* $A_0 \in \mathcal{H} \otimes \overline{\mathcal{G}}$ *such that the function* $\bar{g}^* : [X] \to \mathcal{H}$ *can be written as*

$$\bar{g}^*(\eta) = A_0 \bar{k}_\eta.$$

Lemma 24. *Under Assumption 5, we have that* $\mathcal{F} = \{g \circ i_X \mid g \in \overline{\mathcal{G}}\}$*, with inner product* $\langle g \circ i_X, g' \circ i_X \rangle_{\mathcal{F}} = \langle g, g' \rangle_{\overline{\mathcal{G}}}$ *is a reproducing kernel Hilbert space on* $X \times P$*, with kernel* $k((x,p),(x',p')) = \bar{k}(x_p, x'_{p'})$*. Moreover there exists a linear unitary operator* $U : \overline{\mathcal{G}} \to \mathcal{F}$ *such that* $Ug = g \circ i_X \in \mathcal{F}$ *for any* $g \in \overline{\mathcal{G}}$*.*

In particular under Assumptions 4 to 6, we have that Assumption 3 is satisfied for $G = A_0 U^*$*, and*

$$\|g^*\|_{\mathcal{H}\otimes\mathcal{F}} := \|G\|_{\mathrm{HS}(\mathcal{F},\mathcal{H})} = \|A_0\|_{\mathrm{HS}(\bar{\mathcal{G}},\mathcal{H})} = \|\bar{g}^*\|_{\mathcal{H}\otimes\bar{\mathcal{G}}}.$$

Proof. By definition $\overline{\mathcal{G}}$ is the RKHS associated to the kernel $\bar{k}$ on $[X]$, where the scalar product $\langle \cdot, \cdot \rangle_{\overline{\mathcal{G}}}$ is defined such that $\left\langle \bar{k}_\eta, \bar{k}_\zeta \right\rangle_{\overline{\mathcal{G}}} = \bar{k}(\eta, \zeta)$, for any $\eta, \zeta \in [X]$ and $\overline{\mathcal{G}}$ is the closure of $\overline{\mathcal{G}}_0 = \mathrm{span}\{\bar{k}(\eta, \cdot) \mid \eta \in [X]\}$ w.r.t. $\langle \cdot, \cdot \rangle_{\overline{\mathcal{G}}}$. Similarly $\mathcal{F}$ is the RKHS associated to the kernel k such that the scalar product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ is defined as $\langle k_{x,p}, k_{x',p'} \rangle_{\mathcal{F}} = \bar{k}(i_X(x,p), i_X(x',p'))$, for all $(x,p),(x',p') \in X \times P$. Note that by definition of $\mathcal{F}$, we have that $\mathcal{F}$ is the closure of $\mathcal{F}_0$ w.r.t. $\langle \cdot, \cdot \rangle_{\mathcal{F}}$, with

$$\begin{aligned}
\mathcal{F}_0 &= \mathrm{span}\{k((x,p),(\cdot,\cdot)) \mid (x,p) \in X \times P\} && (174)\\
&= \mathrm{span}\{\bar{k}(i_X(x,p), i_X(\cdot,\cdot)) \mid (x,p) \in X \times P\} && (175)\\
&= \mathrm{span}\{\bar{k}(\eta, i_X(\cdot,\cdot)) \mid \eta \in [X]\} && (176)\\
&= \overline{\mathcal{G}}_0 \circ i_X. && (177)
\end{aligned}$$

Now, since for any $\eta, \zeta \in [X]$ there exist $(x,p),(x',p') \in [X]$ such that $\eta = i_X(x,p), \zeta = i_X(x',p')$, we have that,

$$\begin{aligned}
\left\langle \bar{k}(\eta, i_X(\cdot,\cdot)), \bar{k}(\zeta, i_X(\cdot,\cdot)) \right\rangle_{\mathcal{F}} &= \left\langle \bar{k}(i_X(x,p), i_X(\cdot,\cdot)), \bar{k}(i_X(x',p'), i_X(\cdot,\cdot)) \right\rangle_{\mathcal{F}} && (178)\\
&= \bar{k}(i_X(x,p), i_X(x',p')) = \bar{k}(\eta,\zeta) = \left\langle \bar{k}_\eta, \bar{k}_\zeta \right\rangle_{\overline{\mathcal{G}}}. && (179)
\end{aligned}$$

So, let $f, f' \in \mathcal{F}_0$, by definition we have $f = g \circ i_X$ and $f' = g' \circ i_X$ with $g, g' \in \overline{\mathcal{G}}_0$. Moreover by definition of g, g' there exist $n, m \in \mathbb{N}$ and $\eta_1, \dots, \eta_n, \zeta_1, \dots, \zeta_m \in [X]$ and $\alpha_1, \dots, \alpha_n, \beta_1, \dots, \beta_m \in \mathbb{R}$ such that $g(\cdot) = \sum_{i=1}^n \alpha_i \bar{k}(\eta_i, \cdot)$ and analogously $g'(\cdot) = \sum_{j=1}^m \beta_j \bar{k}(\zeta_j, , \cdot)$.

Now we show that $\langle g \circ i_X, g' \circ i_X \rangle_{\mathcal{F}} = \langle g, g' \rangle_{\overline{\mathcal{G}}}$ for $g, g' \in \overline{\mathcal{G}}_0$ and then we extend it to $\overline{\mathcal{G}}$. First we recall that the composition on the right is linear, indeed

$$(\alpha f + \beta g) \circ h = \alpha (f \circ h) + \beta (g \circ h),$$

for any $\alpha, \beta \in \mathbb{R}$, any function $f, g : A \to \mathbb{R}$ and $h : B \to A$, and A, B two sets. Then we have

$$\langle f, f' \rangle_{\mathcal{F}} = \langle g \circ i_X, g' \circ i_X \rangle_{\mathcal{F}} = \left\langle \left(\sum_{i=1}^{n} \alpha_i \bar{k}(\eta_i, \cdot) \right) \circ i_X, \left(\sum_{j=1}^{m} \beta_j \bar{k}(\zeta_j, \cdot) \right) \circ i_X \right\rangle \tag{180}$$

$$= \left\langle \sum_{i=1}^{n} \alpha_i \bar{k}(\eta_i, i_X(\cdot, \cdot)), \sum_{j=1}^{m} \beta_j \bar{k}(\zeta_j, i_X(\cdot, \cdot)) \right\rangle \tag{181}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j \left\langle \bar{k}(\eta_i, i_X(\cdot, \cdot)), \bar{k}(\zeta_j, i_X(\cdot, \cdot)) \right\rangle_{\mathcal{F}} \tag{182}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j \left\langle \bar{k}_{\eta_i}, \bar{k}_{\zeta_j} \right\rangle_{\overline{\mathcal{G}}} = \left\langle \sum_{i=1}^{n} \alpha_i \bar{k}_{\eta_i}, \sum_{j=1}^{m} \beta_j \bar{k}_{\zeta_j} \right\rangle_{\overline{\mathcal{G}}} \tag{183}$$

$$= \langle g, g' \rangle_{\overline{\mathcal{G}}} . \tag{184}$$

By noting that

$$\|g_n \circ i_X - g_m \circ i_X\|_{\mathcal{F}} = \|(g_n - g_m) \circ i_X\|_{\mathcal{F}} = \|g_n - g_m\|_{\overline{\mathcal{G}}}$$

for any Cauchy sequence $(g_n)_{n \in \mathbb{N}}$ in $\overline{\mathcal{G}}_0$, and the fact that $\mathcal{F}_0 = \overline{\mathcal{G}} \circ i_X$ and that $\langle g \circ i_X, g \circ i_X \rangle_{\mathcal{F}} = \langle g, g' \rangle_{\overline{\mathcal{G}}}$, for $g, g' \in \overline{\mathcal{G}}_0$, then we have that $\mathcal{F} = \overline{\mathcal{G}} \circ i_X$, and that $\langle g \circ i_X, g \circ i_X \rangle_{\mathcal{F}} = \langle g, g' \rangle_{\overline{\mathcal{G}}}$, for $g, g' \in \overline{\mathcal{G}}$.

Now denote by $U : \overline{\mathcal{G}} \to \mathcal{F}$ the operator such that $Ug = g \circ i_X$. First note that U is linear, indeed

$$U(\alpha g + \beta h) = (\alpha g + \beta h) \circ i_X = \alpha(g \circ i_X) + \beta(h \circ i_X) = \alpha U g + \beta U h,$$

for any $g, h \in \overline{\mathcal{G}}$ and $\alpha, \beta \in \mathbb{R}$. Moreover we show that U is a partial isometry, indeed

$$\|Ug\|_{\mathcal{F}}^2 = \|g \circ i_X\|_{G}^2 = \langle g \circ i_X, g \circ i_X \rangle_{\mathcal{F}} = \langle g, g \rangle_{\overline{\mathcal{G}}} = \|g\|_{\overline{\mathcal{G}}}^2.$$

Finally by applying the result above to g^* and $\bar{g}^*$, under Assumptions 4 to 6, we have that $G = A_0 U^*$ and so, by using the isomorphism between $\mathcal{H} \otimes \mathcal{F}$ and $\mathrm{HS}(\mathcal{F}, \mathcal{H})$, we have

$$\|g^*\|_{\mathcal{H} \otimes \mathcal{F}} := \|G\|_{\mathrm{HS}(\mathcal{F}, \mathcal{H})} = \|A_0\|_{\mathrm{HS}(\bar{\mathcal{G}}, \mathcal{H})} = \|\bar{g}^*\|_{\mathcal{H} \otimes \bar{\mathcal{G}}}.$$

□

G.1 Proof of Lemma 6

Assumption 7. *The distribution $\pi(\cdot|x) = \pi(\cdot|x')$ for any $x, x' \in X$. For the sake of simplicity we will denote it by $\pi(\cdot)$.*

Lemma 25. *Under Assumption 7, the following hold*

$$q = \mathbb{E}_{pq} \, \mathfrak{c}_{pq}, \tag{185}$$

where, for $p, q \in P$

$$\mathfrak{c}_{pq} = \mathbb{E}_{x,x'} \left[k((x,p),(x,q))^2 - k((x,p),(x',q))^2 \right]. \tag{186}$$

Proof. First note that with the definitions of Lemma 17, we have

$$\mathfrak{q} = \mathfrak{c}_1 - \mathfrak{c}_2$$

by Lemma 17 . Under Assumption 7 we can denote $\pi(\cdot|x) = \pi(\cdot)$ without ambiguity. Then with the notation of Lemma 17, we have

$$\mathfrak{c}_1 = \int k((x,p),(x,q))^2 \, d\pi(p)d\pi(q)d\rho_X(x) \tag{187}$$

$$= \mathbb{E}_{p,q} \int k((x,p),(x,q))^2 \, d\rho_X(x) \tag{188}$$

$$= \mathbb{E}_{p,q}\mathbb{E}_x k((x,p),(x,q))^2 \tag{189}$$

Analogously for $\mathfrak{c}_2$

$$\mathfrak{c}_2 = \int k((x,p),(x',q))^2 \, d\pi(p)d\pi(q) \, d\rho_X(x)\rho_X(x') \tag{190}$$

$$= \mathbb{E}_{p,q} \int k((x,p),(x,q))^2 \, d\rho_X(x)\rho_X(x') \tag{191}$$

$$= \mathbb{E}_{p,q}\mathbb{E}_{x,x'} k((x,p),(x,q))^2 \tag{192}$$

□

G.2 Proof of Theorem 7

Proof. This proof consists in applying Theorem 4 with $\lambda = \sqrt{r^2/m + q/n}$, and taking into account inter-locality and intra-locality.

First, under the inter-locality condition formalized in our measure theoretic setting as Assumption 4, there exists a $\bar{g}^* : [X] \to \mathcal{H}$ such that $g^*(x,p) = \bar{g}^*(x_p)$ for any $x \in X$ and $p \in P$ as proven by Lemma 23. So the restriction kernel can learn $\bar{g}^*$ if it is rich enough, that is $\bar{g}^* \in \mathcal{H} \otimes \bar{\mathcal{F}}$ (here formalized as Assumption 6, with $\bar{\mathcal{F}}$ denoted by $\bar{\mathcal{G}}$). Then we can apply Lemma 24, that guarantees the applicability of Theorem 4.

Second, by the assumption on the fact that $\pi(p|x) = 1/|P|$, we can apply Lemma 6 and then the intra-locality condition of Assumption 2, obtaining the desired result. □

H Universal Consistency

In this section we prove universal consistency for the proposed algorithm. In particular this consists in the same proof of Thm 4, Section B.3, but using our Comparison inequality (Thm 11) and our bound in high probability of the distance between $\widehat{g}$ and g^*, that is the following Thm 29. First, we recall and specify the Pinelis inequality [34–36] to our setting.

Proposition 26. *Let $\delta \in (0,1]$ and $m \in \mathbb{N}$. Let $\mathcal{H}$ be a separable Hilbert space. Let $\zeta_1, \dots, \zeta_m$ be independently distributed $\mathcal{H}$-valued random variables. Let $R > 0$ be such that* ess sup $\|\zeta_j\|_{\mathcal{H}} \le R$ *for every $j = 1, \dots, m$. Then,*

$$\left\| \frac{1}{m} \sum_{j=1}^{m} \left[\zeta_j - \mathbb{E}\, \zeta_j \right] \right\|_{\mathcal{H}} \le \frac{4R \log \frac{3}{\delta}}{\sqrt{m}} \tag{193}$$

with probability at least $1-\delta$*.*

Proof. By applying Lemma 2 of [36] with constants $\widetilde{M}=R$ and $\sigma^2=\sup_j \mathbb{E}\|\zeta_j\|^2 \leq R^2$, we obtain

$$\left\|\frac{1}{m}\sum_{j=1}\left[\zeta_j-\mathbb{E}\zeta_j\right]\right\|_{HS} \leq \frac{2R\log\frac{2}{\delta}}{m}+\sqrt{\frac{2R^2\log\frac{2}{\delta}}{m}} \tag{194}$$

with probability at least $1-\delta$. Now, $\log\frac{2}{\delta}\leq\log\frac{3}{\delta}$ and $\log 3\delta \geq 1$ for any $\delta\in(0,1]$. Then, we can bound the above inequality by

$$\frac{2R\log\frac{2}{\delta}}{m}+\sqrt{\frac{2R^2\log\frac{2}{\delta}}{m}} \leq \frac{4R\log\frac{3}{\delta}}{\sqrt{m}}. \tag{195}$$

□

Remark 7 (Pinelis Inequality for Hilbert-Schmidt Operators)**.** *We recall that the space of Hilbert-Schmidt operators between two separable Hilbert spaces is itself a separable Hilbert space with the Hilbert-Schmidt norm. Therefore, Pinelis inequality in Prop. 26 is directly applicable.*

Lemma 27. *Let* C *and* $\widehat{C}$ *and* $\kappa=\sup_{x,p}\|k_{x,p}\|_{\mathcal{F}}$ *defined as Sec. B. Let* $\delta\in(0,1]$*. Then*

$$\|\widehat{C}-C\| \leq 4\kappa^2\left(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt{n}}\right)\log\frac{6}{\delta} \tag{196}$$

with probability at least $1-\delta$*.*

Proof. Given a dataset $(x_i)_{i=1}^n$, we introduce the operator $\widetilde{C}:\mathcal{F}\to\mathcal{F}$ defined as

$$\widetilde{C}=\frac{1}{n}\sum_{i=1}^n\int_P k_{x_i,p}\otimes k_{x_i,p}\ d\pi(p|x_i). \tag{197}$$

and consider the following decomposition

$$\|\widehat{C}-C\| \leq \|\widehat{C}-\widetilde{C}\|+\|\widetilde{C}-C\|. \tag{198}$$

Let $\tau=\delta/2$, in the following we separately bound the terms above in probability and then take the intersection bound.

Bounding $\|\widehat{C}-\widetilde{C}\|$. For any $j=1,\dots,m$ let $\zeta_j=k_{x_{i_j},p_j}\otimes k_{x_{i_j},p_j}$ with i_j and p_j independently sampled respectively from: the uniform distribution on $\{1,\dots,n\}$ and the conditional probability $\pi(\cdot|x_{i_j})$. Therefore, for any $j=1,\dots,m$

$$\widehat{C}=\frac{1}{m}\sum_{j=1}^m\zeta_j, \qquad \widetilde{C}=\mathbb{E}\ \zeta_j=\frac{1}{n}\sum_{i=1}^n\int_P k_{x_i,p}\otimes k_{x_i,p}\ d\pi(p|x_i) \tag{199}$$

and

$$\text{ess sup } \|\zeta_j\|_{HS} \leq \sup_{x\in X, p\in P} \langle k_{x,p}, k_{x,p}\rangle_{\mathcal{F}} \leq \sup_{x\in X, p\in P} \|k_{x,p}\|_{\mathcal{F}}^2 \leq \kappa^2$$

We apply Pinelis inequality (see Remark 7), leading to

$$\|\widehat{C} - \widetilde{C}\| \leq \|\widehat{C} - \widetilde{C}\|_{HS} = \left\| \frac{1}{m} \sum_{j=1} [\zeta_j - \mathbb{E}\zeta_j] \right\|_{HS} \leq \frac{4\kappa^2 \log \frac{3}{\tau}}{\sqrt{m}} \tag{200}$$

with probability at least $1 - \tau$.

Bounding $\|\widetilde{C} - C\|$. For $i = 1, \dots, n$ let $\eta_i = \int_P k_{x_i,p} \otimes k_{x_i,p}\, d\pi(p|x_i)$ with x_i independently sampled from ρ_X. Therefore, for every $i = 1, \dots, n$,

$$\widetilde{C} = \frac{1}{n} \sum_{i=1}^n \eta_i, \qquad C = \mathbb{E}\, \eta_i = \int_{X\times P} k_{x,p} \otimes k_{x,p}\, d\pi(p|x) d\rho_X(x) \tag{201}$$

and

$$\text{ess sup } \|\eta_i\|_{HS} \leq \sup_{x\in X, p\in P} \|k_{x,p}\|_{\mathcal{F}}^2 \leq \kappa^2.$$

We apply again Pinelis inequality, obtaining

$$\|\widetilde{C} - C\| \leq \|C - \widetilde{C}\|_{HS} = \left\| \frac{1}{n} \sum_{i=1}^n [\eta_i - \mathbb{E}\eta_i] \right\|_{HS} \leq \frac{4\kappa^2 \log \frac{3}{\tau}}{\sqrt{n}} \tag{202}$$

with probability at least $1 - \tau$.

By taking the intersection bound of the two events above, we obtain

$$\|\widehat{C} - C\|_{HS} \leq \frac{4\kappa^2 \log \frac{3}{\tau}}{\sqrt{m}} + \frac{4\kappa^2 \log \frac{3}{\tau}}{\sqrt{n}} \tag{203}$$

with probability at least $1 - 2\tau$. By recalling $\tau = \frac{\delta}{2}$ we obtain the desired result. □

Lemma 28. *Let* B, $\widehat{B}$, $\kappa = \sup_{x,p} \|k_{x,p}\|_{\mathcal{F}}$ *and* $q = \sup_w \|\varphi(w)\|_{\mathcal{H}}$ *defined as Sec. B. Let* $\delta \in (0, 1]$. *Then*

$$\|\widehat{B} - B\|_{HS} \leq 4\kappa q \left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} \right) \log \frac{6}{\delta} \tag{204}$$

with probability at least $1 - \delta$.

Proof. Given $(x_i, y_i)_{i=1}^n$ a dataset, we introduce the operator $\widetilde{B} : \mathcal{H} \to \mathcal{F}$ defined as

$$\widetilde{B} = \frac{1}{n} \sum_{i=1}^n \int_P k_{x_i,p} \otimes \varphi(w)\, d\mu(w|y_i, x_i, p) d\pi(p|x_i). \tag{205}$$

and consider the following decomposition

$$\|\widehat{B} - B\|_{HS} \leq \|\widehat{B} - \widetilde{B}\|_{HS} + \|\widetilde{B} - B\|_{HS}. \tag{206}$$

Let $\tau = \delta/2$, in the following we separately bound the terms above in probability and then take the intersection bound.

Bounding $\|\widehat{B} - \widetilde{B}\|_{HS}$. For any $j = 1, \dots, m$ let $\xi_j = k_{x_{i_j}, p_j} \otimes \varphi(w_j)$ with i_j, p_j and w_j independently sampled respectively from: the uniform distribution on $\{1, \dots, n\}$, the conditional probability $\pi(\cdot|x_{i_j})$ and the conditional probability $\mu(\cdot|y_{i_j}, x_{i_j}, p_j)$. Therefore, for any $j = 1, \dots, m$

$$\widehat{B} = \frac{1}{m} \sum_{j=1}^{m} \xi_j, \qquad \widetilde{B} = \mathbb{E}\ \xi_j = \frac{1}{n} \sum_{i=1}^{n} \int_{[Y] \times P} k_{x_i, p} \otimes \varphi(w)\ d\mu(w|x_i, y_i, p) d\pi(p|x_i), \tag{207}$$

moreover

$$\text{ess sup } \|\xi_j\|_{HS} \leq \sup_{x,p,w} \|k_{x,p} \otimes \varphi(w)\|_{HS} = \sup_{x,p,w} \|k_{x,p}\|_{\mathcal{F}} \|\varphi(w)\|_{\mathcal{H}} \leq \kappa q. \tag{208}$$

We apply Pinelis inequality (see Remark 7), leading to

$$\|\widehat{B} - \widetilde{B}\|_{HS} = \left\| \frac{1}{m} \sum_{j=1}^{m} [\xi_j - \mathbb{E}\ \xi_j] \right\|_{HS} \leq \frac{4\kappa q \log \frac{3}{\tau}}{\sqrt{m}} \tag{209}$$

with probability at least $1 - \tau$.

Bounding $\|B - \widetilde{B}\|_{HS}$. For any $i = 1, \dots, n$, let $v_i = \int_{[Y] \times P} k_{x_i, p} \otimes \varphi(w)\ d\mu(w|y_i, x_i, p) d\pi(p|x_i)$ with (x_i, y_i) independently sampled from ρ. Then, for any $i = 1, \dots, n$

$$\begin{aligned}
\mathbb{E}\ v_i &= \int_{[Y] \times Y \times X \times P} k_{x,p} \otimes \varphi(w)\ d\mu(w|y_i, x_i, p) d\pi(p|x_i) d\rho(y, x) && (210)\\
&= \int_{X \times P} k_{x,p} \otimes \left[\int_{[Y] \times Y} \varphi(w)\ d\mu(w|y_i, x_i, p) d\rho(y|x) \right] d\pi(p|x_i) d\rho_X(x) && (211)\\
&= \int_{X \times P} k_{x,p} \otimes g^*(x, p)\ d\pi(p|x_i) d\rho_X(x) && (212)\\
&= B && (213)
\end{aligned}$$

and $\widetilde{B} = \frac{1}{n} \sum_{i=1}^{n} v_i$. Moreover,

$$\begin{aligned}
\text{ess sup } \|v_i\|_{HS} &\leq \sup_{x,y} \int_{[Y] \times P} \|k_{x,p} \otimes \varphi(w)\|_{HS}\ d\mu(w|y, x, p) d\pi(p|x) && (214)\\
&= \sup_{x,y} \int_{[Y] \times P} \|k_{x,p}\|_{\mathcal{F}} \|\varphi(w)\|_{\mathcal{H}}\ d\mu(w|y, x, p) d\pi(p|x) && (215)\\
&\leq \kappa q \sup_{x,y} \int_{[Y] \times P} d\mu(w|y, x, p) d\pi(p|x) && (216)\\
&= \kappa q && (217)\\
& && (218)
\end{aligned}$$

Therefore, applying again Pinelis inequality,

$$\|B - \widetilde{B}\|_{HS} = \left\| \frac{1}{n} \sum_{i=1}^{n} [\nu_i - \mathbb{E}\nu_i] \right\|_{HS} \leq \frac{4\kappa q \log \frac{3}{\tau}}{\sqrt{n}} \tag{219}$$

with probability at least $1 - \tau$.

By taking the intersection bound of the two events above, we obtain

$$\|\widehat{B} - B\|_{HS} \leq \frac{4\kappa q \log \frac{3}{\tau}}{\sqrt{m}} + \frac{4\kappa q \log \frac{3}{\tau}}{\sqrt{n}} \tag{220}$$

with probability at least $1 - 2\tau$ as desired.

□

Theorem 29. *Let* $\delta \in (0, 1]$. *Let* $Q > 0$, $n \in \mathbb{N}$, $c_Q = 1 + 1/\sqrt{Q}$ *and* $m = Qn$. *Then*

$$\|\widehat{g} - g^*\|_{L^2(X\times P, \pi\rho_X, \mathcal{H})} \leq \frac{4\kappa^2 c_Q(\|L_\lambda^{-1/2} G\|_{HS} + \frac{q}{\kappa}) \log \frac{12}{\delta}}{\sqrt{\lambda n}} \left(1 + 2\kappa \sqrt{\frac{c_Q \log \frac{12}{\delta}}{\lambda \sqrt{n}}}\right) + \lambda \|L_\lambda^{-1} G\|_{HS} \tag{221}$$

with probability at least $1 - \delta$.

Proof. In Thm. 13 we have bounded $\|\widehat{g} - g^*\|_{L^2(X\times P, \pi\rho_X, \mathcal{H})}$ in terms of an analytic expression of $\|C - \widehat{C}\|$ and $\|B - \widehat{B}\|_{HS}$. We bound these two terms with probability $1 - \tau$ with $\tau = \delta/2$ via Lemma 27 and Lemma 28. We further take the intersection bound to obtain the desired result. □

H.1 Proof of Theorem 2

Proof. The proof is exactly the same as in Theorem 4 Section B.3 of the supplementary materials of [19], where instead of using their comparison inequality (their Thm. 2) we use our Thm. 11 and instead using their Lemma 18 we use our Theorem 29. □

I Equivalence between SELF and SELF by Parts without assumptions

I.1 SELF without Parts

We begin by briefly recalling the SELF framework in [19]. We will see that this is a special case of the setting proposed in this work for a special choice of the kernel on $X \times P$.

We recall the definition of SELF introduced in [19] and consider the formulation in [26].

Definition 3. *A function* $\triangle : Z \times Y \to \mathbb{R}$ *is a Structure Encoding Loss Function (SELF) if there exist a Hilbert space* $\bar{\mathcal{H}}$ *and two maps* $\bar{\psi} : Z \to \mathcal{H}$ *and* $\bar{\varphi} : Y \to \mathcal{H}$ *such that*

$$\triangle(z, y) = \left\langle \bar{\psi}(z), \bar{\varphi}(y) \right\rangle_{\bar{\mathcal{H}}} \tag{222}$$

for all $z \in Z, y \in Y$.

Below we show that the definition of SELF by parts introduced in this work is a refinement of the original one. Since the original definition of SELF did not account for the possibility of $\triangle$ do depend also on the input, below we consider only the case $\triangle(z, y|x) = \triangle(z, y)$. In particular we will assume in Def. 1 that $\pi(p|x) = \pi(p|x')$ for any $x, x' \in X$, $p \in P$ and denote it $\pi(p)$.

Lemma 30. *Let $\triangle : Z \times Y \to \mathbb{R}$ satisfy Def. 1 with*

$$\triangle(z, y) = \sum_{p \in P} \ell(z, y|p)\pi(p) = \sum_{p \in P} \langle \psi(z, p), \varphi(y_p) \rangle_{\mathcal{H}} \tag{223}$$

Then $\triangle$ satisfies the original SELF definition Def. 3, with $\bar{\mathcal{H}} = \mathcal{H} \otimes \mathbb{R}^P$ and maps $\bar{\psi} : Z \to \bar{\mathcal{H}}$ and $\bar{\varphi} : Y \to \bar{\mathcal{H}}$ such that

$$\bar{\psi}(z) = (\sqrt{\pi(p)}\psi(z, p))_{p \in P} \qquad \textit{and} \qquad (\sqrt{\pi(p)}\varphi(y_p))_{p \in P} \tag{224}$$

In particular, we have that the constant $c_\triangle$ is

$$c_\triangle = \sqrt{\sup_{z \in Z} \sum_{p \in P} \pi(p) \|\psi(z, p)\|^2_{\mathcal{H}}} = \sup_{z \in Z} \|\bar{\psi}(z)\|_{\bar{\mathcal{H}}}. \tag{225}$$

Proof. Recall that by construction $\bar{\mathcal{H}} = \mathcal{H} \otimes \mathbb{R}^P = \bigoplus_{p \in P} \mathcal{H}$. Therefore, any vector $\eta \in \bar{\mathcal{H}}$ is the collection $(\eta_p)_{p \in P}$ with $\eta_1, \dots, \eta_P \in \mathcal{H}$ and the corresponding inner product with a $\zeta = (\zeta_p)_{p \in P} \in \bar{\mathcal{H}}$ is

$$\langle \eta, \zeta \rangle_{\bar{\mathcal{H}}} = \sum_{p \in P} \langle \eta_p, \zeta_p \rangle_{\mathcal{H}}. \tag{226}$$

Plugging the definition of $\bar{\psi}$ and $\bar{\varphi}$ in the definition of SELF by parts, we have

$$\begin{aligned} \triangle(z, y) &= \sum_{p \in P} \pi(p) \langle \varphi(z, p), \psi(y_p) \rangle_{\mathcal{H}} & (227) \\ &= \sum_{p \in P} \left\langle \sqrt{\pi(p)}\varphi(z, p), \sqrt{\pi(p)}\psi(y_p) \right\rangle_{\mathcal{H}} & (228) \\ &= \left\langle \bar{\psi}(z), \bar{\varphi}(y) \right\rangle_{\bar{\mathcal{H}}} & (229) \end{aligned}$$

as required. □

I.1.1 SELF Solution

Given a loss $\triangle$ that is a SELF by parts, we have already observed that the solution $f^* : X \to Z$ of the structured prediction problem in (4), can be characterized in terms of a function $g^* : X \times P \to \mathcal{H}$ introduced in (14). Based on the relation highlighted by Lemma 30, we have the following equivalent characterization

$$f^*(x) = \operatorname*{argmin}_{z \in Z} \left\langle \bar{\psi}(z), h^*(x) \right\rangle_{\bar{\mathcal{H}}} \tag{230}$$

where now $h^* : X \to \bar{\mathcal{H}}$ is conditional mean embedding of $\bar{\varphi}(y)$ in $\bar{\mathcal{H}}$ with respect to the conditional distribution $\rho(y|x)$. In particular, let $e_p \in \mathbb{R}^P$ denote the p-th element of the canonical basis in $\mathbb{R}^P$. Then, for any $\eta \in \mathcal{H}$, $x \in X$ and $p \in P$, we have

$$\langle h^*(x), \eta \otimes e_p \rangle_{\bar{\mathcal{H}}} = \left\langle \int \bar{\varphi}(y) \, d\rho(y|x), \eta \otimes e_p \right\rangle \tag{231}$$

$$= \sqrt{\pi(p)} \left\langle \int \varphi(y_p) \, d\rho(y|x), \eta \right\rangle_{\mathcal{H}} \tag{232}$$

$$= \sqrt{\pi(p)} \left\langle g^*(x, p), \eta \right\rangle_{\mathcal{H}}, \tag{233}$$

and in particular,

$$h^*(x) = (\sqrt{\pi(p)} g^*(x, p))_{p \in P}. \tag{234}$$

We conclude that

$$\|h^*\|^2_{L^2(X, \rho_X, \bar{\mathcal{H}})} = \int \langle h^*(x), h^*(x) \rangle_{\bar{\mathcal{H}}} \, d\rho_X(x) \tag{235}$$

$$= \int \sum_{p \in P} \left\langle \sqrt{\pi(p)} g^*(x, p), \sqrt{\pi(p)} g^*(x, p) \right\rangle_{\mathcal{H}} \, d\rho_X(x) \tag{236}$$

$$= \int \sum_{p \in P} \pi(p) \langle g^*(x, p), g^*(x, p) \rangle_{\mathcal{H}} \, d\rho_X(x) \tag{237}$$

$$= \|g^*\|^2_{L^2 X, \pi\rho_X, \mathcal{H}}. \tag{238}$$

I.2 If g^* is "simple" (e.g. Asm. 1 holds)

Let $\bar{K}$ be a kernel on X with RKHS $\mathcal{F}$. Let K be a kernel on $X \times P$ defined as $K((x, p), (x', p')) = \bar{K}(x, x')\delta_{p,p'}$, for $x, x' \in X$, $p, p' \in P$. Note that the RKHS associated to K is $\mathcal{F} \otimes \mathbb{R}^P$ with $K_{x,p} = \bar{K}_x \otimes e_p$ and $e_p \in \mathbb{R}^P$ the p-th element of the canonical basis of $\mathbb{R}^P$.

Lemma 31. *Let $G \in \mathcal{H} \otimes \mathcal{F} \otimes \mathbb{R}^P$ be such that $g^*(x, p) = GK_{x,p}$ for any $x \in X$ and $p \in P$. Let $G_1, \ldots, G_P \in \mathcal{H} \otimes \mathcal{F}$ the operator such that $G_p \eta = G(\eta \otimes e_p)$ for any $p \in P$ and $\eta \in \mathcal{F}$. Then,*

- $G = \sum_{p \in P} G_p \otimes e_p$.
- *For any $x \in X$, $h^*(x) = H\bar{K}_x$ with $H = \sum_{p \in P} e_p \otimes \sqrt{\pi(p)} G_p \in \mathbb{R}^P \otimes \mathcal{H} \otimes \mathcal{F}$.*

In particular

$$\|G\|^2_{HS(\mathcal{F} \otimes \mathbb{R}^P, \mathcal{H})} = \sum_{p \in P} \|G_p\|^2_{HS(\mathcal{F}, \mathcal{H})} \qquad \textit{and} \qquad \|H\|^2_{HS(\mathcal{F}, \mathcal{H} \otimes \mathbb{R}^P)} = \sum_{p \in P} \pi(p) \|G_p\|^2_{HS(\mathcal{F}, \mathcal{H})}. \tag{239}$$

Lemma 32. *Let $G \in \mathcal{H} \otimes (\mathcal{F} \otimes \mathbb{R}^P)$ be such that $g^*(x, p) = GK_{x,p}$ for any $x \in X$ and $p \in P$. Let $G_1, \ldots, G_P \in \mathcal{H} \otimes \mathcal{F}$ the operator such that $G_p \eta = G(\eta \otimes e_p)$ for any $p \in P$ and $\eta \in \mathcal{F}$. Then, there exists an operator $H \in (\mathcal{H} \otimes \mathbb{R}^P) \otimes \mathcal{F}$, such that*

- $H\bar{K}_x = h^*(x)$ *for all* $x \in X$.
- $\|G\|^2_{HS(\mathcal{F}\otimes\mathbb{R}^P,\mathcal{H})} = \sum_{p\in P} \|G_p\|^2_{HS(\mathcal{F},\mathcal{H})}$.
- $\|H\|^2_{HS(\mathcal{F},\mathcal{H}\otimes\mathbb{R}^P)} = \sum_{p\in P} \pi(p)\|G_p\|^2_{HS(\mathcal{F},\mathcal{H})}$.

Proof. Note that since e_p form a basis of $\mathbb{R}^P$, we can write $G = \sum_{p\in P} G_p \otimes e_p$ and therefore

$$\|G\|^2_{HS(\mathcal{F}\otimes\mathbb{R}^P,\mathcal{H})} = \sum_{p\in P} \|G_p\|^2_{HS(\mathcal{F},\mathcal{H})} \tag{240}$$

as required.

Now, by definition of h^* and the relation with g^*, we have that

$$\begin{aligned} h^*(x) &= (\sqrt{\pi(p)}\ g^*(x,p))_{p\in P} && (241)\\ &= (\sqrt{\pi(p)}\ GK_{x,p})_{p\in P} && (242)\\ &= (\sqrt{\pi(p)}\ G(\bar{K}_x \otimes e_p))_{p\in P} && (243)\\ &= (\sqrt{\pi(p)}G_p\bar{K}_x)_{p\in P} && (244)\\ &= H\bar{K}_x, && (245)\end{aligned}$$

where we have denoted with $H \in (\mathcal{H}\otimes\mathbb{R}^P)\otimes\mathcal{F}$, the operator from $\mathcal{F}$ to $\mathcal{H}\otimes\mathbb{R}^P$, such that for any $\eta \in \mathcal{F}$ we have $H = (\sqrt{\pi(p)}G_p\eta)_{p\in P}$. The required results follow directly from the construction of both G and H in terms of the G_p for $p \in P$. □

We can therefore conclude the equivalence between the original SELF estimator with kernel $\bar{K}$ and the SELF estimator by parts considered in this work, with kernel K, under the assumption that g^* (and equivalently h^*) belong to the corresponding RKHS.

Theorem 33. *The SELF estimator with kernel* $\bar{K}$ *has same rates as the SELF by parts with kernel* K

For simplicity, assume $\pi(p|x) = \frac{1}{|P|}$ for every $x \in X$ and $p \in P$. From (5) and the SELF assumption, we have

$$\triangle(z,y|x) = \frac{1}{|P|}\sum_{p\in P} \langle \psi(z_p,x_p,p), \varphi(y_p)\rangle_{\mathcal{H}}. \tag{246}$$

Denote $\bar{\psi} : Z \times X \to \mathcal{H}\otimes\mathbb{R}^P$ and $\bar{\varphi} : Y \to \mathcal{H}\otimes\mathbb{R}^P$ the maps such that

$$\bar{\psi}(z,x) = \Big(\psi(z_p,x_p,p)\Big)_{p\in P} \qquad \bar{\varphi}(y) = \Big(\varphi(y_p)\Big)_{p\in P} \tag{247}$$

which can be interpreted as the concatenation of the different ψ and φ for $p \in P$. Then we can rewrite $\triangle$ in terms of the canonical inner product of $\mathcal{H}\otimes\mathbb{R}^P$,

$$\triangle(z,y|x) = \frac{1}{|P|}\left\langle \bar{\psi}(z,x), \bar{\varphi}(y)\right\rangle_{\mathcal{H}\otimes\mathbb{R}^P}. \tag{248}$$

We can now apply the approach proposed in this work to the case of a problem *with one single part* (or equivalently apply the SELF approach in [19]). The target function of this problem is $h^* : X \to \mathcal{H} \otimes \mathbb{R}^P$ defined as

$$h * (x) = \frac{1}{|P|} \int \bar{\varphi}(y) \, d\rho(y|x) = \frac{1}{|P|} \Big(\int \varphi(y_p) \, d\rho(y|x) \Big)_{p \in P} = \frac{1}{|P|} (g^*(x,p))_{p \in P} \in \mathcal{H} \otimes \mathbb{R}^P \tag{249}$$

and is the concatenation of all functions $g^*(\cdot, p)$ for $p \in P$.

Now, let us consider a rkhs $\mathcal{F}$ of functions $h : X \to \mathbb{R}$ with associated kernel $k : X \times X \to \mathbb{R}$. Assume that h^* belongs to the space of vector valued functions $\mathcal{F} \otimes (\mathcal{H} \otimes \mathbb{R}^P)$. In other words, there exists an Hilbert-Schmidt operator $H : \mathcal{F} \to \mathcal{H} \otimes \mathbb{R}^P$ such that $Hk_x = h^*(x)$ for any $p \in P$. Note that this is equivalent to require that the function g^* belongs to the space $(\mathcal{F} \otimes \mathbb{R}^P) \otimes \mathcal{H}$, namely that there exists an Hilbert-Schmidt operator, such that $G : \mathcal{F} \otimes \mathbb{R}^P \to \mathcal{H}$, such that, $G(k_x \otimes e_p) = g^*(x,p)$ for any $x \in X$ and $p \in P$, with $e_p \in \mathbb{R}^P$ denoting the p-th element of the canonical basis of $\mathbb{R}^P$. In particular, note that, for any $\eta \in \mathcal{H}$, $p \in P$ and $x \in X$, we have

$$\langle Hk_x, \eta \otimes e_p \rangle_{\mathcal{H}} = \langle h^*(x), \eta \otimes e_p \rangle = \langle h^*(x)_p, \eta \rangle_{\mathcal{H}} = \langle g^*(x,p), \eta \rangle_{\mathcal{H}} = \frac{1}{|P|} \langle G(k_x \otimes e_p), \eta \rangle . \tag{250}$$

We conclude that $H = \frac{1}{|P|} G$ and $\|H\|_{HS} = \frac{1}{\sqrt{|P|}} \|G\|_{HS}$. In particular, note that since $G \in (\mathcal{F} \otimes \mathbb{R}^P) \otimes \mathcal{H}$, we have that for any $p \in P$, the function $g(\cdot, p) : X \to \mathcal{H}$ is such that $g(\cdot, p) \in \mathcal{F} \otimes \mathcal{H}$. Therefore we have

$$\|G\|_{HS} = \sqrt{\sum_{p \in P} \|g^*(\cdot, p)\|^2_{\mathcal{F} \otimes \mathcal{H}}}. \tag{251}$$

Interestingly, if all the functions $g^*(\cdot, p)$ have same norm $\mathfrak{g} = \|g^*(\cdot, p)\|_{\mathcal{F} \otimes \mathcal{H}}$ in $\mathcal{F} \otimes \mathcal{H}$, we have

$$\|H\|_{HS} = \frac{1}{|P|} \|G\|_{HS} = \frac{1}{\sqrt{P}} \sqrt{\sum_{p \in P} \mathfrak{g}^2} = \mathfrak{g}. \tag{252}$$

I.3 The best of both worlds

Here we formalize the comment in Remark 3, where we introduced the kernel $K_B = K_U + K_L$ that is sum of a bounded universal continuous kernel K_U over $X \times P$ and a bounded restriction (or "local") kernel K_L, satisfying Eq. (20). In particular we show that K_B is universal but at the same time allows to train a structured prediction estimator $\widehat{f}$ that is able to leverage the locality of the learning problem, when available. For simplicity, we assume the input space X to be compact and the set of parts indices P to be finite.

Let $\mathcal{F}_B, \mathcal{F}_U$ and $\mathcal{F}_L$ denote the RKHSs of respectively K_B, K_U and K_L. According to [37], we know that $\mathcal{F}_B \supseteq \mathcal{F}_U \cup \mathcal{F}_L$ and moreover that for any $h \in \mathcal{F}_B$, the norm is such that

$$\|h\|^2_{\mathcal{F}_B} = \min_{h = h_U + h_L} \|h_U\|^2_{\mathcal{F}_U} + \|h_L\|^2_{\mathcal{F}_L}, \tag{253}$$

with $h_U \in \mathcal{F}_U, h_L \in \mathcal{F}_L$. We immediately see that K_B is universal. Indeed, since K_U is universal, $\mathcal{F}_U$ is dense in the space of continuous functions on X and consequently also $\mathcal{F}_B \supseteq \mathcal{F}_U$ is.

The following result is analogous to Lemma 6 and shows that the kernel K_B is not only universal but also equivalent to K_L in capturing the locality of the learning problem.

Lemma 34. *Denote by* $K = K_B = K_U + K_L$ *the sum kernel, where* K_U *and* K_L *are the universal and restriction kernels on* $X \times P$ *characterized by Eqs.* (13) *and* (20) *in terms of respectively* $\bar{K} : [X] \times [X] \to \mathbb{R}$ *and* $K_0 : X \times X \to \mathbb{R}$*. Let* $\bar{r} = \sup_{\chi \in [X]} \bar{K}(\chi, \chi)$ *and* $r_0 = \sup_{x \in X} K_0(x, x)$*.*

Let $\pi(p|x) = \frac{1}{|P|}$ *for any* $x \in X$ *and* $p \in P$*. Denote with* $\bar{C}_{pq}$ *the constant defined in Eq.* (21) *associated to the restriction kernel* K_L*. Then, the constant* q *in Eq.* (17) *associated to* K_B *can be factorized as*

$$q = \frac{1}{|P|^2} \sum_{p,q \in P} C_{p,q}, \quad \textit{with} \quad C_{p,q} \leq \bar{C}_{p,q} \ + \ (4\bar{r} + r_0) r_0 \ \delta_{p,q}. \tag{254}$$

Proof. The proof of the result above follows by noting that, since π is uniform, by Lemma 25, for any $p, q \in P$, $C_{p,q}$ is characterized by

$$C_{p,q} = \mathbb{E}_{x,x'} \left[(\bar{K}(x_p, x_q) + K_0(x, x)\delta_{p,q})^2 - (\bar{K}(x_p, x'_q) + K_0(x, x')\delta_{p,q})^2 \right] \tag{255}$$

$$= \bar{C}_{p,q} + \mathbb{E}_{x,x'} \left[K_0(x, x)^2 - K_0(x, x')^2 \right] \delta_{p,q} + \tag{256}$$

$$- 2\mathbb{E}_{x,x'} \left[\bar{K}(x_p, x_q) K_0(x, x) - \bar{K}(x_p, x'_q) K_0(x, x') \right] \delta_{p,q} \tag{257}$$

$$\leq \bar{C}_{p,q} + \delta_{p,q} \sup_{x \in X} K_0(x, x)^2 + 4\delta_{p,q} \left[\sup_{\chi \in [X]} \bar{K}(\chi, \chi) \sup_{x \in X} K_0(x, x) \right] \tag{258}$$

$$\leq \bar{C}_{p,q} + (4\bar{r} + r_0) \ r_0 \ \delta_{p,q} \tag{259}$$

as desired. Note that the first inequality follows from the fact that $\bar{K}$ and K_0 are positive definite symmetric kernels. □

Interestingly, Lemma 34 shows that the proposed sum kernel inherits the ability of the restriction kernel to capture the intra and inter-locality of the learning problem. Combining this with the learning rates of Thm. 4, we obtain a result analogous to that of Thm. 7.

Theorem 35 (Learning Rates & Locality)**.** *With the same notation of Lemma 34 let* K_U *be a bounded continouous universal kernel on* X*,* K_L *be the restriction kernel based on the reproducing kernel* $\bar{K}$ *on* $[X]$ *and let* $\bar{\mathcal{F}}$ *be the RKHS associated to* $\bar{K}$*. Let* $\widehat{f}$ *be the structured prediction estimator of Eq.* (6) *learned with kernel* $K = K_B = K_U + K_L$*. Then*

1. $\widehat{f}$ *is universally consistent,*
2. *Under Assumptions 1 and 2 and* $\pi(p|x) = \frac{1}{|P|}$ *for* $x \in X, p \in P$*, let* $\bar{g}^*$ *be defined as in Lemma 5 and* $\bar{g}^* \in \mathcal{H} \otimes \bar{\mathcal{F}}$*. Denote by* $\bar{g}$ *the norm* $\bar{g} = \|\bar{g}^*\|_{\mathcal{H} \otimes \bar{\mathcal{F}}}$*. When* $\lambda = (r^2/m + q/n)^{1/2}$*, then*

$$\mathbb{E} \ \mathcal{E}(\widehat{f}\ ) - \mathcal{E}(f^*) \ \leq \ 12 \ c_\triangle \ \bar{g} \ r^{1/2} \left(\frac{1}{m} + \frac{c_1}{|P|n} + \frac{\sum_{p \neq q} e^{-\gamma d(p,q)}}{|P|^2 n} \right)^{1/4}, \tag{260}$$

where $r = r_0 + \bar{r}$*, with* $r_0, \bar{r}$ *defined as in Lemma 34 and* $c_1 = 1 + (4\bar{r} + r_0) \ r_0/r^2$*.*

Proof. Let $\mathcal{F}_B, \mathcal{F}_U$ and $\mathcal{F}_L$ denote the RKHSs of respectively K_B, K_U and K_L.

First, as discussed at the beginning of this section, the kernel $K = K_B := K_U + K_L$ is universal, since $\mathcal{F}_U \subseteq \mathcal{F}_B$ (see [37]) and $\mathcal{F}_U$ is dense in the continuous functions on $X \times P$. Then we can directly apply Thm. 2 obtaining the unversal consistency for $\widehat{f}$.

Second, under Assumption 1, by Lemma 5, we have that there exists $\bar{g}^* : [X] \to \mathcal{H}$ such that g^*, defined as in Eq. (14), is characterized by $g^*(x,p) = \bar{g}^*(x_p)$. S ince we assume that $\bar{g}^* \in \mathcal{H} \otimes \bar{\mathcal{F}}$ and we are using a restriction kernel under inter-locality, we can apply Lemma 24 (where we used $\bar{\mathcal{G}}$ to denote $\bar{\mathcal{F}}$ and $\mathcal{F}$ to denote $\mathcal{F}_L$ and $\bar{g}^* \in \mathcal{H} \otimes \bar{\mathcal{F}}$ is expressed more formally by Asm. 6), then $g^* \in \mathcal{H} \otimes \mathcal{F}_L$ and $\|g^*\|_{\mathcal{H}\otimes\mathcal{F}_L} = \|\bar{g}^*\|_{\mathcal{H}\otimes\bar{\mathcal{F}}}$. Now, according to Eq. (253) (see [37]), for any function $h \in \mathcal{F}_L$ we have

$$\|h\|_{\mathcal{F}_B} := \min\{\|h_U\|_{\mathcal{F}_U} + \|h_L\|_{\mathcal{F}_L} \mid h = h_U + h_L, h_U \in \mathcal{F}_U, h_L \in \mathcal{F}_L\} \leq \|h\|_{\mathcal{F}_L},$$

since h can be always decomposed as $h = h_L + h_U$ with $h_L = h$ and $h_U = 0$, then $\|g^*\|_{\mathcal{H}\otimes\mathcal{F}_B} \leq \|g\|_{\mathcal{H}\otimes\mathcal{F}_L}$. So

$$\|g^*\|_{\mathcal{H}\otimes\mathcal{F}_B} \leq \|\bar{g}^*\|_{\mathcal{H}\otimes\bar{\mathcal{F}}}.$$

Now we are ready to apply Thm. 4, with $\lambda = \sqrt{r^2/m + q/n}$ obtaining

$$\mathbb{E}\,\mathcal{E}(\widehat{f}\,) - \mathcal{E}(f^*) \quad \leq \quad 12\, c_\triangle\, \bar{g}\, \left(\frac{r^2}{m} + \frac{q}{n}\right)^{1/4}. \tag{261}$$

Finally note that since $\pi(p|x) = \frac{1}{|P|}$ for $p \in P, x \in X$,we can apply Lemma 34

$$\frac{q}{n} = \frac{r^2 c_1}{|P|n} + \frac{r^2 \sum_{p\neq q} e^{-\gamma d(p,q)}}{|P|^2 n},$$

obtaining the desired result. □

The discussion above implies that under the locality assumptions, the rates in Thm. 35 are essentially equivalent to the ones of the estimator trained with only the restriction kernel in Thm. 7.

J Additional details on evaluating $\widehat{f}$

According to (6), evaluating $\widehat{f}$ on a test point $x \in X$ consists in solving an optimization problem over the output space Z. This is a standard procedure in structured prediction settings [2], where a corresponding optimization method is derived on a case-by-case basis depending on the loss and the space Z ([2]). However, the specific form of the objective functional characterizing $\widehat{f}$ in our setting allows to devise a general stochastic meta-algorithm to solve such problem. We observe that (6) can be rewritten as

$$\widehat{f}(x) = \operatorname*{argmin}_{z\in Z}\ \mathbb{E}_{(j,p)}\ \ell_{j,p}(z|x) \tag{262}$$

where for any $p \in P$ and $j \in \{1, \ldots, m\}$ we have introduced the functions $\ell_{j,p} : Z \to \mathbb{R}$, such that

$$h_{j,p}(\cdot|x) = \big(\mathrm{sign}(\alpha_j(x,p))\ A(x,p)\big)\ L(\cdot, w_j|x,p) \tag{263}$$

Algorithm 2 Learning $\widehat{f}$

Input: training set $(x_i, y_i)_{i=1}^n$, distributions $\pi(\cdot|x)$ and $\mu(\cdot|y, x, p)$, reproducing kernel k on $X \times P$, hyperparameter $\lambda > 0$, auxiliary dataset size $m \in \mathbb{N}$.

Generate auxiliary dataset $(w_j, x_{i_j}, p_j)_{j=1}^m$:
Sample i_j uniformly from $\{1, \dots, n\}$
Sample $p_j \sim \pi(\cdot|x_{i_j})$
Sample $w_j \sim \mu(\cdot|y_{i_j}, x_{i_j}, p_j)$

Compute the coefficients for the score function α:
$K \in \mathbb{R}^{m \times m}$ with entries $K_{jj'} = k((x_{i_j}, p_j), (x_{i_{j'}}, p_{j'}))$
$C = (K + m\lambda I)^{-1}$

Return $\alpha : X \times P \to \mathbb{R}^m$ such that $\alpha(x, p) = C\, v(x, p)$ with $v(x, p) \in \mathbb{R}^m$ is the vector with entries $v(x, p)_j = k((x_{i_j}, p_j), (x, p))$.

with $A(x, p) = \sum_{j=1}^m |\alpha_j(x, p)|$. In the expectation above, the variable p is sampled according to $\pi(\cdot|x)$ and j is sampled from the set $\{1, \dots, m\}$ with probability $\frac{|\alpha_j(x,p)|}{A(x,p)}$. When the $h_{j,p}$ are (sub)differentiable, problems of the form of (10) can be addressed by stochastic gradient methods (SGM). In Alg. 3 in the supplementary material we provide an example of such strategy.

K Additional examples of Loss Functions by Parts

Several structured prediction settings are recovered within the setting considered in this work and the associated loss functions have the form of Eq. (5). Below recall some of the most relevant examples.

Hamming. A standard loss function used in structured prediction is the Hamming loss [20–22], which for any factorization by parts can be written as in (5) with $L_p(z_p, y_p|x_p) = \delta(z_p \neq y_p)$, the function equal to 0 if $z_p = y_p$ and 1 otherwise.

- **Computer Vision**. The Hamming loss is often used in computer vision [2, 16]. For instance, in image segmentation [9] the goal is to label each pixel p of an input image x, as background ($y_p = 0$) or foreground ($y_p = 1$). Errors are measured as total number of mistakes $z_p \neq y_p$ over the total number of pixels.

- **Hierarchical Classification**. In classification settings with a hierarchy [14], errors are weighted according to the semantic distance between two classes (e.g. classifying the image of a "dog" as a "bus" is worse than classifying it as a "cat"). Assuming the hierarchy between classes to be represented as a tree, these loss functions can be written as the Hamming loss between the parts of a class $y = (y_{\text{root}}, \dots, y_{\text{leaf}})$ seens as the collection of all the nodes in its hierarchy (e.g. "cat", "feline", "mammal", "animate object", "entity").

Algorithm 3 Evaluating $\widehat{f}$

Input: input $x \in X$, distribution $\pi(\cdot|x)$, auxiliary dataset $(w_j, x_{i_j}, p_j)_{j=1}^m$, score functions $\alpha : X \times P \to \mathbb{R}$, number of iterations T, step sizes $\{\gamma_t\}_{t\in\mathbb{N}}$.
Initialization $z_0 = 0$

For $t = 1$ to T
Sample $p \sim \pi(\cdot|x)$
$A(x,p) = \sum_{j=1}^m |\alpha_j(x,p)|$
Sample j from $\{1,\dots,m\}$ with $\mathbb{P}(j = k) = |\alpha_k(x,p)|/A(x,p)$
$h_{j,p} = \mathrm{sign}(\alpha_j(x,p))\ A(x,p)\ \ell(z, w_j|x,p)$
Choose $u \in \partial h_{j,p}(\cdot|x)(z_{t-1})$
$z_t = \mathrm{proj}_Z(z_{t-1} - \gamma_t u)$

Return z_T

- **Planning**. In learning-to-plan applications [13], the goal is to predict a trajectory z closest to a ground truth trajectory (typically provided by an expert). A trajectory is represented as a sequence of contiguous states $y = (y_{\text{start}}, \dots, y_{\text{end}})$ and errors with respect to a predicted trajectory z are measured in terms of the number of states that do not coincide, namely the hamming loss between the two sequences.

This loss has been extensively used in computer vision for applications such as pixel-wise classification [9] or image segmentation [4].

Precision/Recall, F1 Score. The precision/recall and F1 score are loss functions often adopted in natural language processing [12]. They are used to measure the similarity between two binary sequences. Given two binary sequences $z, y \in \{0,1\}^k$ of length k, we have $\triangle(z,y) = \triangle(z^\top y, \|z\|^2, \|y\|^2)$. In particular, the precision correponds to $\triangle(z,y) = z^\top y/\|z\|^2$, the recall to $\triangle(z,y) = z^\top y/\|y\|^2$ and the F1 score to $\triangle(z,y) = z^\top y/(\|z\|^2 + \|y\|^2)$. These functions are in the form of (5) if taking $|P| = k$ and $i_Y(y,p) = (y_p, \|y\|)$, $i_Y(z,p) = (z_p, \|z\|)$. Note that the number of elements in y and z can vary depending on the cardinality $|x|$ of each input x, (see e.g. [12]). In this sense the $\triangle(z,y|x)$ is necessarily parametrized by x and in particular the set P is a set $P(x) = \{1,\dots,|x|\}$.

Multitask Learning Multitask learning settings have a natural decomposition into parts: the output and label spaces Z and Y are subset of $\mathbb{R}^T$, and $\triangle(z,y) = \frac{1}{T}\sum_{t=1}^T L(z_t, y_t)$, with L any loss function commonly used in standard supervised learning problems (e.g. least-squares for regression, hinge or logistic for classification). In settings where Z is not a linear space but a *constraint set*, our model recovers the non-linear multitask learning framework considered in [26].

Learning sequences. Let $X = A^k$, $Y = Z = B^k$ for two sets A, B and $k \in \mathbb{N}$ a fixed length. We consider a set of structures $P \subseteq \mathbb{N}^2$ such that any pair $p = (s,l) \in P$ indicates the starting element and the length of a subsequence. In particular, we choose the set of parts $\mathcal{X} = \cup_{t=1}^k A^t$ and $\mathcal{Y} = \mathcal{Z} = \cup_{t=1}^k B^t$ with

$$x_p = (x^{(s)}, \dots, x^{(s+l-1)}) \in \mathcal{X} \qquad \forall\, x \in X,\ \ \forall\, (s,l) \in P \tag{264}$$

where we have denoted $x^{(s)}$ the s-th entry of the sequence $x \in X$. Analogously $y_p = (y^{(s)}, \dots, y^{(s+l-1)})$ for $y \in Y$. Finally, we choose the loss L_0 to be the (normalized) edit distance between two strings of same length

$$L_0(z, y; x, (s, l)) = \frac{1}{l} \sum_{i=1}^{l} \mathbf{1}(z^{(i)} \neq y^{(i)}) \tag{265}$$

where $\mathbf{1}(z^{(i)} \neq y^{(i)}) = 0$ if $z^{(i)} = y^{(i)}$ and 1 otherwise (clearly a generic loss function $h(z^{(i)} \neq y^{(i)})$ and weight w_i can be used instead of $\mathbf{1}$ and $1/l$). Finally, we can choose the uniform distribution $\pi(p|x) = 1/|P|$ (but clearly also less symmetric weighting strategy can be adopted).

Pixelwise classification on images. Consider the problem of assigning each pixel of an image to one of T separate classes. In this setting $X = \mathbb{R}^{d \times d}$ is the set of images (with fixed width and height equal to $d \in \mathbb{N}$) and $Y = Z = \mathbb{R}^{T \times d \times d}$ is the set of all possible ways to label an image. We choose the set of parts $\mathcal{X} = \cup_{w,h=1}^{d} \mathbb{R}^{w \times h}$ to be the set of all possible patches of $d \times d$ image and the set of structures to be a $P \subset \mathbb{N}^4$ such that for any image $x \in X$ and $p = (u, l, w, h) \in P$ the selectors $x_p \in \mathbb{R}^{w \times h}$ and $y_p, z_p \in \mathbb{R}^{T \times w \times h}$ correspond to the patch of the image x or the labeling y and z with width w, height h and upper-left corner at the pixel (u, l).

We choose the loss L_0 to be a function comparing the class "statistics" in a given patch: e.g.

$$L_0(z_p, y_p; x_p, p) = \|\sigma(z_p) - \sigma(y_p)\|^2 \qquad \sigma(\zeta) = \frac{\sum_{i=1}^{\mathrm{width}(\zeta)} \sum_{j=1}^{\mathrm{height}(\zeta)} \zeta_{:,i,j}}{\mathrm{width}(\zeta)\mathrm{height}(\zeta)}. \tag{266}$$

Since it is more likely to have larger values for L_0 at higher scales (the object patch overlaps other classes), we choose a weighting $\pi(p|x)$ that is decreasing with respect to the size of the patch $p = (u, l, w, h)$. For instance we can choose $\pi(p|x) = \frac{\exp(-\gamma w h)}{\sum_{p'=(u',l',w',h') \in P} \exp(-\gamma w' h')}$, for $\gamma > 0$.