A Consistent Regularization Approach to Structured Prediction: Supplementary material

Carlo Ciliberto, Alessandro Rudi and Lorenzo Rosasco

The supplementary material of this work is divided in the following three sections:

A Proofs of Fisher consistency and comparison inequality (Thm. 2).

B Universal Consistency and Generalization Bounds for Alg. 1. (Thm. 4 and 5).

C The characterization of a large family of $\triangle$s satisfying Asm. 1 (Thm. 19).

Mathematical Setting

In the following we will always assume $\mathcal{X}$ and $\mathcal{Y}$ to be Polish spaces, namely separable complete metrizable spaces, equipped with the associated Borel sigma-algebra. When referring to a probability distribution ρ on $\mathcal{X} \times \mathcal{Y}$ we will always assume it to be a Borel probability measure, with $\rho_{\mathcal{X}}$ the marginal distribution on $\mathcal{X}$ and $\rho(\cdot|x)$ the conditional measure on $\mathcal{Y}$ given $x \in \mathcal{X}$. We recall [1] that $\rho(y|x)$ is a regular conditional distribution and its domain, which we will denote $D_{\rho|\mathcal{X}}$ in the following, is a measurable set contained in the support of $\rho_{\mathcal{X}}$ and corresponds to the support of $\rho_{\mathcal{X}}$ up to a set of measure zero.

For convenience, we recall here the main assumption of our work.

Assumption 1. *There exists a separable Hilbert space $\mathcal{H}_{\mathcal{Y}}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\mathcal{Y}}}$, a continuous embedding $\psi : \mathcal{Y} \to \mathcal{H}_{\mathcal{Y}}$ and a bounded linear operator $V : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{Y}}$, such that*

$$\triangle(y, y') = \langle \psi(y), V\psi(y') \rangle_{\mathcal{H}_{\mathcal{Y}}} \qquad \forall y, y' \in \mathcal{Y} \tag{10}$$

Basic notation We recall that a Hilbert space $\mathcal{H}$ is a vector space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, closed with respect to the norm $\|h\|_{\mathcal{H}} = \sqrt{\langle h, h \rangle_{\mathcal{H}}}$ for any $h \in \mathcal{H}$. We denote with $L^2(\mathcal{X}, \rho_{\mathcal{X}}, \mathcal{H})$ the Lebesgue space of square integrable functions on $\mathcal{X}$ with respect to a measure $\rho_{\mathcal{X}}$ and with values in a separable Hilbert space $\mathcal{H}$. We denote with $\langle f, g \rangle_{\rho_{\mathcal{X}}}$ the inner product $\int \langle f(x), g(x) \rangle_{\mathcal{H}} d\rho_{\mathcal{X}}(x)$, for all $f, g \in L^2(\mathcal{X}, \rho_{\mathcal{X}}, \mathcal{H})$. In particular when $\mathcal{H} = \mathbb{R}$ we denote with $L^2(\mathcal{X}, \rho_{\mathcal{X}})$ the space $L^2(\mathcal{X}, \rho_{\mathcal{X}}, \mathbb{R})$.

Given a linear operator $V : \mathcal{H} \to \mathcal{H}'$ between two Hilbert spaces $\mathcal{H}, \mathcal{H}'$, we denote with $\mathrm{Tr}(V)$ the trace of V and with $V^* : \mathcal{H}' \to \mathcal{H}$ the adjoint operator associated to V, namely such that $\langle Vh, h' \rangle'_{\mathcal{H}} = \langle h, V^* h' \rangle_{\mathcal{H}'}$ for every $h \in \mathcal{H}$, $h' \in \mathcal{H}'$. Moreover, we denote with $\|V\| = \sup_{h \in \mathcal{H}} \|Vh\|_{\mathcal{H}'}$ and $\|V\|_{HS} = \sqrt{\mathrm{Tr}(V^* V)}$ respectively the operator norm and Hilbert-Schmidt norm of V. We recall that a linear operator V is continuous if and only if $\|V\| < +\infty$ and we denote $\mathcal{B}(\mathcal{H}, \mathcal{H}')$ the set of all continuous linear operators from $\mathcal{H}$ to $\mathcal{H}'$. Moreover, we denote $\mathcal{B}_2(\mathcal{H}, \mathcal{H}')$ the set of all operators $V : \mathcal{H} \to \mathcal{H}'$ with $\|V\|_{HS} < +\infty$ and recall that $\mathcal{B}_2(\mathcal{H}, \mathcal{H}')$ is isometric to the space $\mathcal{H}' \otimes \mathcal{H}$, with $\otimes$ denoting the tensor product. Indeed, for the sake of simplicity, with some abuse of notation we will not make the distinction between the two spaces.

Note that in most of our results we will require $\mathcal{Y}$ to be non-empty and compact, so that a continuous functional over $\mathcal{Y}$ always attains a minimizer on $\mathcal{Y}$ and therefore the operator $\mathrm{argmin}_{y \in \mathcal{Y}}$ is well defined. Note that for a finite set $\mathcal{Y}$, we will always assume it endowed with the discrete topology, so that $\mathcal{Y}$ is compact and any function $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is continuous.

On the Argmin Notice that for simplicity of notation, in the paper we denoted the minimizer of Alg. 1 as

$$\hat{f}(x) = \operatorname*{argmin}_{y \in \mathcal{Y}} \sum_{i=1}^{n} \alpha_i(x) \triangle (y, y_i). \tag{23}$$

However note that the correct notation should be

$$\hat{f}(x) \in \operatorname*{argmin}_{y \in \mathcal{Y}} \sum_{i=1}^{n} \alpha_i(x) \triangle (y, y_i) \tag{24}$$

since a loss function $\triangle$ can have more than one minimizer in general. In the following we keep this more pedantic, yet correct notation.

Expected Risk Minimization Note that whenever we write an expected risk minimization problem, we implicitly assume the optimization domain to be the space of measurable functions. For instance, Eq. (1) would be written more rigorously as

$$\text{minimize} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} \triangle(f(x), y)\, d\rho(x, y) \;\middle|\; f : \mathcal{X} \to \mathcal{Y} \text{ measurable} \right\} \tag{25}$$

In the next Lemma, following [2] we show that the problem in Eq. (1) admits a measurable pointwise minimizer.

Lemma 6 (Existence of a solution for Eq. (1))**.** *Let $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a continuous function. Then, the expected risk minimization at Eq. (1) admits a measurable minimizer $f^* : \mathcal{X} \to \mathcal{Y}$ such that*

$$f^*(x) \in \operatorname*{argmin}_{y\in\mathcal{Y}} \int_{\mathcal{Y}} \triangle(y, y')d\rho(y'|x) \tag{26}$$

for every $x \in D_{\rho|\mathcal{X}}$. Moreover, the function $m : \mathcal{X} \to \mathbb{R}$ defined as follows, is measurable

$$m(x) = \inf_{y\in\mathcal{Y}} r(x, y), \qquad \textit{with} \qquad r(x, y) = \begin{cases} \int_{\mathcal{Y}} \triangle(y, y')d\rho(y'|x) & \textit{if } x \in D_{\rho|\mathcal{X}} \\ 0 & \textit{otherwise} \end{cases} \tag{27}$$

Proof. Since $\triangle$ is continuous and $\rho(y|x)$ is a regular conditional distribution, then r is a Carathéodory function (see Definition 4.50 (pp. 153) of [3]), namely continuous in y for each $x \in \mathcal{X}$ and measurable in x for each $y \in \mathcal{Y}$. Thus, by Theorem 18.19 (pp. 605) of [3] (or Aumann's measurable selection principle [2, 4]), we have that m is measurable and that there exists a measurable $f^* : \mathcal{X} \to \mathcal{Y}$ such that $r(x, f^*(x)) = m(x)$ for all $x \in \mathcal{X}$. Moreover, by definition of m, given any measurable $f : \mathcal{X} \to \mathcal{Y}$, we have $m(x) \leq r(x, f(x))$. Therefore,

$$\mathcal{E}(f^*) = \int r(x, f^*(x))d\rho_{\mathcal{X}}(x) = \int m(x)d\rho_{\mathcal{X}}(x) \leq \int r(x, f(x))d\rho_{\mathcal{X}}(x) = \mathcal{E}(f). \tag{28}$$

We conclude $\mathcal{E}(f^*) \leq \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f)$ and, since f^* is measurable, $\mathcal{E}(f^*) = \min_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f)$ and f^* is a global minimizer. □

We have an immediate Corollary to Lemma 6.

Corollary 7. *With the hypotheses of Lemma 6, let $\tilde{f} : \mathcal{X} \to \mathcal{Y}$ such that*

$$\tilde{f}(x) \in \operatorname*{argmin}_{y\in\mathcal{Y}} \int_{\mathcal{Y}} \triangle(y, y')d\rho(y'|x)$$

for almost every $x \in D_{\rho|\mathcal{X}}$. Then $\mathcal{E}(\tilde{f}) = \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f)$.

Proof. The result follows directly from Lemma 6 by noting that $r(x, \tilde{f}(x)) = m(x)$ almost everywhere on $D_{\rho|\mathcal{X}}$. Hence, since $D_{\rho|\mathcal{X}}$ is equal to the support of $\rho_{\mathcal{X}}$ up to a set of measure zero, $\mathcal{E}(\tilde{f}) = \int_{\mathcal{X}} m(x)d\rho_{\mathcal{X}}(x) = \mathcal{E}(f^*) = \inf_f \mathcal{E}(f)$. □

With the above basic notation and results, we can proceed to prove the results presented in this work.

A Surrogate Problem, Fisher Consistency and Comparison Inequality

In this section we focus on the surrogate framework introduced in Sec. 3 and prove that it is *Fisher consistent* and that the *comparison inequality*. To do so, we will first characterizes the solution(s) of the surrogate expected risk minimization introduced at Eq. (12). We recall that in our setting, the surrogate risk was defined as the functional $\mathcal{R}(g) = \int_{\mathcal{X}\times\mathcal{Y}} \|\psi(y) - g^*(x)\|^2_{\mathcal{H}_\mathcal{Y}} d\rho(x, y)$, where $\psi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ is continuous (by Asm. 1). In the following, when ψ is bounded, we will denote with $Q = \sup_{y\in\mathcal{Y}} \|\psi(y)\|_{\mathcal{H}_\mathcal{Y}}$. Note that in most our results we will assume $\mathcal{Y}$ to be compact. In these settings we always have $Q = \max_{y\in\mathcal{Y}} \|\psi(y)\|_{\mathcal{H}_\mathcal{Y}}$ by the continuity of ψ.

We start with a preliminary lemma necessary to prove Lemma 1 and Thm. 2.

Lemma 8. *Let $\mathcal{H}_\mathcal{Y}$ a separable Hilbert space and $\psi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ measurable and bounded. Then, the function $g^* : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ such that*

$$g^*(x) = \int_\mathcal{Y} \psi(y) d\rho(y|x) \quad \forall x \in D_{\rho|\mathcal{X}} \tag{29}$$

and $g^*(x) = 0$ otherwise, belongs to $L^2(\mathcal{X}, \rho_\mathcal{X}, \mathcal{H}_\mathcal{Y})$ and is a minimizer of the surrogate expected risk at Eq. (12). Moreover, any minimizer of Eq. (12) is equal to g^* almost everywhere on the domain of $\rho_\mathcal{X}$.

Proof. By hypothesis, $\|\psi\|_{\mathcal{H}_\mathcal{Y}}$ is measurable and bounded. Therefore, since $\rho(y|x)$ is a regular conditional probability, we have that g^* is measurable on $\mathcal{X}$ (see for instance [2]). Moreover, the norm of g^* is dominated by the constant function of value Q, thus g^* is integrable on $\mathcal{X}$ with respect to $\rho_\mathcal{X}$ and in particular it is in $L^2(\mathcal{X}, \rho_\mathcal{X}, \mathcal{H}_\mathcal{Y})$ since $\rho_\mathcal{X}$ is a finite regular measure. Recall that since $\rho(y|x)$ is a regular conditional distribution, for any measurable $g : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$, the functional in Eq. (12) can be written as

$$\mathcal{R}(g) = \int_{\mathcal{X}\times\mathcal{Y}} \|g(x) - \psi(y)\|^2_{\mathcal{H}_\mathcal{Y}} d\rho(x, y) = \int_\mathcal{X} \int_\mathcal{Y} \|g(x) - \psi(y)\|^2_{\mathcal{H}_\mathcal{Y}} d\rho(y|x) d\rho_\mathcal{X}(x). \tag{30}$$

Notice that $g^*(x) = \operatorname{argmin}_{\eta\in\mathcal{H}_\mathcal{Y}} \int_\mathcal{Y} \|\eta - \psi(y)\|^2_{\mathcal{H}_\mathcal{Y}} d\rho(y|x)$ almost everywhere on $D_{\rho|\mathcal{X}}$. Indeed,

$$\int_\mathcal{Y} \|\eta - \psi(y)\|^2_{\mathcal{H}_\mathcal{Y}} d\rho(y|x) = \|\eta\|^2_{\mathcal{H}_\mathcal{Y}} - 2\langle \eta, \left(\int_\mathcal{Y} \psi(y) d\rho(y|x)\right)\rangle + \int_\mathcal{Y} \|\psi(y)\|^2_{\mathcal{H}_\mathcal{Y}} d\rho(y|x) \tag{31}$$

$$= \|\eta\|^2_{\mathcal{H}_\mathcal{Y}} - 2\langle \eta, g^*(x)\rangle_{\mathcal{H}_\mathcal{Y}} + const. \tag{32}$$

for all $x \in D_{\rho|\mathcal{X}}$, which is minimized by $\eta = g^*(x)$ for all $x \in D_{\rho|\mathcal{X}}$. Therefore, since $D_{\rho|\mathcal{X}}$ is equal to the support of $\rho_\mathcal{X}$ up to a set of measure zero, we conclude that $\mathcal{R}(g^*) \leq \inf_{g:\mathcal{X}\to\mathcal{H}_\mathcal{Y}} \mathcal{R}(g)$ and, since g^* is measurable, $\mathcal{R}(g^*) = \min_{g:\mathcal{X}\to\mathcal{H}_\mathcal{Y}} \mathcal{R}(g)$ and g^* is a global minimizer as required.

Finally, notice that for any $g : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ we have

$$\mathcal{R}(g) - \mathcal{R}(g^*) = \int_{\mathcal{X}\times\mathcal{Y}} \|g(x) - \psi(y)\|^2_{\mathcal{H}_\mathcal{Y}} - \|g^*(x) - \psi(y)\|^2_{\mathcal{H}_\mathcal{Y}} d\rho(x, y) \tag{33}$$

$$= \int_\mathcal{X} \|g(x)\|^2_{\mathcal{H}_\mathcal{Y}} - 2\langle g(x), \left(\int_\mathcal{Y} \psi(y) d\rho(y|x)\right)\rangle_{\mathcal{H}_\mathcal{Y}} + \|g^*(x)\|^2_{\mathcal{H}_\mathcal{Y}} d\rho_\mathcal{X}(x) \tag{34}$$

$$= \int_\mathcal{X} \|g(x)\|^2_{\mathcal{H}_\mathcal{Y}} - 2\langle g(x), g^*(x)\rangle_{\mathcal{H}_\mathcal{Y}} + \|g^*(x)\|^2_{\mathcal{H}_\mathcal{Y}} d\rho_\mathcal{X}(x) \tag{35}$$

$$= \int_\mathcal{X} \|g(x) - g^*(x)\|^2_{\mathcal{H}_\mathcal{Y}} d\rho_\mathcal{X}(x) \tag{36}$$

Therefore, for any measurable minimizer $g' : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ of the surrogate expected risk at Eq. (12), we have $\mathcal{R}(g') - \mathcal{R}(g^*) = 0$ which, by the relation above, implies $g'(x) = g^*(x)$ a.e. on $D_{\rho|\mathcal{X}}$. □

Lemma 1. *Let $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfy Asm. 1 with $\psi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ bounded. Then the expected risk in Eq. (1) can be written as*

$$\mathcal{E}(f) = \int_\mathcal{X} \langle \psi(f(x)), V g^*(x)\rangle_{\mathcal{H}_\mathcal{Y}} \, d\rho_\mathcal{X}(x) \tag{11}$$

for all $f : \mathcal{X} \to \mathcal{Y}$, where $g^* : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ minimizes

$$\mathcal{R}(g) = \int_{\mathcal{X}\times\mathcal{Y}} \|g(x) - \psi(y)\|^2_{\mathcal{H}_\mathcal{Y}} \, d\rho(x, y). \tag{12}$$

Proof. By Lemma 8 we know that $g^*(x) = \int_\mathcal{Y} \psi(y) d\rho(y|x)$ almost everywhere on $D_{\rho|\mathcal{X}}$ and is the minimizer of $\mathcal{R}$. Therefore we have

$$\langle \psi(y), V g^*(x)\rangle_{\mathcal{H}_\mathcal{Y}} = \langle \psi(y), V \int_\mathcal{Y} \psi(y') d\rho(y'|x)\rangle_{\mathcal{H}_\mathcal{Y}} \tag{37}$$

$$= \int_\mathcal{Y} \langle \psi(y), V\psi(y')\rangle_{\mathcal{H}_\mathcal{Y}} d\rho(y'|x) = \int_\mathcal{Y} \triangle(y, y') d\rho(y'|x) \tag{38}$$

for almost every $x \in D_{\rho|\mathcal{X}}$. Thus, for any measurable function $f : \mathcal{X} \to \mathcal{Y}$ we have

$$\mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \triangle(f(x), y) d\rho(x, y) = \int_X \int_{\mathcal{Y}} \triangle(f(x), y) d\rho(y|x) d\rho_{\mathcal{X}}(x) \tag{39}$$

$$= \int_X \langle \psi(f(x)), V g^*(x) \rangle_{\mathcal{H}_\mathcal{Y}} d\rho_{\mathcal{X}}(x). \tag{40}$$

□

Theorem 2. *Let $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfy Asm. 1 with $\mathcal{Y}$ a compact set. Then, for every measurable $g : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ and $d : \mathcal{H}_\mathcal{Y} \to \mathcal{Y}$ satisfying Eq. (13), the following holds*

$$\mathcal{E}(d \circ g^*) = \mathcal{E}(f^*) \tag{14}$$

$$\mathcal{E}(d \circ g) - \mathcal{E}(f^*) \leq c_\triangle \sqrt{\mathcal{R}(g) - \mathcal{R}(g^*)}. \tag{15}$$

with $c_\triangle = \|V\| \max_{y\in\mathcal{Y}} \|\psi(y)\|_{\mathcal{H}_\mathcal{Y}}$.

Proof. For the sake of clarity, the result for the fisher consistency and the comparison inequality are proven respectively in Thm. 9, Thm. 12. The two results are proven below. □

Theorem 9 (Fisher Consistency)**.** *Let $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfy Asm. 1 with $\mathcal{Y}$ a compact set. Let $g^* : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ be a minimizer of the surrogate problem at Eq. (12). Then, for any decoding $d : \mathcal{H}_\mathcal{Y} \to \mathcal{Y}$ satisfying Eq. (13)*

$$\mathcal{E}(d \circ g^*) = \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f) \tag{41}$$

Proof. It is sufficient to show that $d \circ g^*$ satisfies Eq. (26) almost everywhere on $D_{\rho|\mathcal{X}}$. Indeed, by directly applying Cor. 7 we have $\mathcal{E}(d \circ g^*) = \mathcal{E}(f^*) = \inf_f \mathcal{E}(f)$ as required.

We recall that a mapping $d : \mathcal{H}_\mathcal{Y} \to \mathcal{Y}$ is a decoding for our surrogate framework if it satisfies Eq. (13), namely

$$d(\eta) \in \operatorname*{argmin}_{y\in\mathcal{Y}} \langle \psi(y), V\eta \rangle_{\mathcal{H}_\mathcal{Y}} \qquad \forall \eta \in \mathcal{H}_\mathcal{Y}. \tag{42}$$

By Lemma 8 we know that $g^*(x) = \int_\mathcal{Y} \psi(y) d\rho(y|x)$ almost everywhere on $D_{\rho|\mathcal{X}}$. Therefore, we have

$$\langle \psi(y), V g^*(x) \rangle_{\mathcal{H}_\mathcal{Y}} = \langle \psi(y), V \int_\mathcal{Y} \psi(y') d\rho(y'|x) \rangle_{\mathcal{H}_\mathcal{Y}} \tag{43}$$

$$= \int_\mathcal{Y} \langle \psi(y), V\psi(y') \rangle_{\mathcal{H}_\mathcal{Y}} d\rho(y'|x) = \int_\mathcal{Y} \triangle(y, y') d\rho(y'|x) \tag{44}$$

for almost every $x \in D_{\rho|\mathcal{X}}$. As a consequence, for any $d : \mathcal{H}_\mathcal{Y} \to \mathcal{Y}$ satisfying Eq. (13), we have

$$d \circ g^*(x) \in \operatorname*{argmin}_{y\in\mathcal{Y}} \langle \psi(y), V g^*(x) \rangle_{\mathcal{H}_\mathcal{Y}} = \operatorname*{argmin}_{y\in\mathcal{Y}} \int_\mathcal{Y} \triangle(y, y') d\rho(y'|x) \tag{45}$$

almost everywhere on $D_{\rho|\mathcal{X}}$. We are therefore in the hypotheses of Cor. 7 with $\tilde{f} = d \circ g^*$, as desired. □

The Fisher consistency of the surrogate problem allows to prove the comparison inequality (Thm. 12) between the excess risk of the structured prediction problem, namely $\mathcal{E}(d \circ g) - \mathcal{E}(f^*)$, and the excess risk $\mathcal{R}(g) - \mathcal{R}(g^*)$ of the surrogate problem. However, before showing such relation, in the following result we prove that for any measurable $g : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ and measurable decoding $d : \mathcal{H}_\mathcal{Y} \to \mathcal{Y}$, the expected risk $\mathcal{E}(d \circ g)$ is well defined.

Lemma 10. *Let $\mathcal{Y}$ be compact and $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfying Asm. 1. Let $g : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ be measurable and $d : \mathcal{H}_\mathcal{Y} \to \mathcal{Y}$ a measurable decoding satisfying Eq. (13). Then $\mathcal{E}(d \circ g)$ is well defined and moreover $|\mathcal{E}(d \circ g)| \leq Q^2 \|V\|$.*

Proof. $\triangle(d \circ g(x), y)$ is measurable in both x and y since $\triangle$ is continuous and $d \circ g$ is measurable by hypothesis (combination of measurable functions). Now, $\triangle$ is pointwise bounded by $Q^2\|V\|$ since

$$|\triangle(y, y')| = |\langle \psi(y), V\psi(y')\rangle_{\mathcal{H}_\mathcal{Y}}| \leq \|\psi(y)\|^2_{\mathcal{H}_\mathcal{Y}}\|V\| \leq Q^2\|V\|. \tag{46}$$

Hence, by Theorem 11.23 pp. 416 in [3] the integral of $\triangle(d \circ g(x), y)$ exists and therefore

$$|\mathcal{E}(d \circ g)| \leq \int_\mathcal{X} |\triangle(d \circ g(x), y)| d\rho(x, y) \leq Q^2\|V\| < +\infty. \tag{47}$$

□

A question introduced by Lemma 10 is whether a *measurable* decoding always exists. The following result guarantees that, under the hypotheses introduced in this work, a decoding $d : \mathcal{H}_\mathcal{Y} \to \mathcal{Y}$ satisfying Eq. (13) always exists.

Lemma 11. *Let $\mathcal{Y}$ be compact and $\psi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ and $V : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ satisfy the requirements in Asm. 1. Define $m : \mathcal{X} \to \mathbb{R}$ as*

$$m(\eta) = \min_{y \in \mathcal{Y}} \langle \psi(y), V\eta\rangle_{\mathcal{H}_\mathcal{Y}} \quad \forall y \in \mathcal{Y}, \eta \in \mathcal{H}_\mathcal{Y}. \tag{48}$$

Then, m is measurable and there exists a measurable decoding $d : \mathcal{H}_\mathcal{Y} \to \mathcal{Y}$ satisfying Eq. (13), namely such that $m(\eta) = \langle \psi(d(\eta)), V\eta\rangle_{\mathcal{H}_\mathcal{Y}}$ for each $\eta \in \mathcal{H}_\mathcal{Y}$.

Proof. Similarly to the proof of Lemma 6, the result is a direct application of Theorem 18.19 (pp. 605) of [3] (or Aumann's measurable selection principle [2, 4]). □

We now prove the *comparison inequality* at Eq. (15).

Theorem 12 (Comparison Inequality)**.** *Let $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfy Asm. 1 with $\mathcal{Y}$ a compact or finite set. Let $f^* : \mathcal{X} \to \mathcal{Y}$ and $g^* : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ be respectively solutions to the structured and surrogate learning problems at Eq. (1) and Eq. (12). Then, for every measurable $g : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ and $d : \mathcal{H}_\mathcal{Y} \to \mathcal{Y}$ satisfying Eq. (13)*

$$\mathcal{E}(d \circ g) - \mathcal{E}(f^*) \leq 2Q\|V\|\sqrt{\mathcal{R}(g) - \mathcal{R}(g^*)}. \tag{49}$$

Proof. Let us denote $f = d \circ g$ and $f_0 = d \circ g^*$. By Thm. 9 we have that $\mathcal{E}(f_0) = \inf_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f)$ and so $\mathcal{E}(f^*) = \mathcal{E}(f_0)$. Now, by combining Asm. 1 with Lemma 8, we have

$$\begin{aligned}
\mathcal{E}(f) - \mathcal{E}(f^*) = \mathcal{E}(f) - \mathcal{E}(f_0) &= \int_{\mathcal{X}\times\mathcal{Y}} \triangle(f(x), y) - \triangle(f_0(x), y) d\rho(x, y) && (50)\\
&= \int_{\mathcal{X}\times\mathcal{Y}} \langle \psi(f(x)) - \psi(f_0(x)), V\psi(y)\rangle_{\mathcal{H}_\mathcal{Y}} d\rho(x, y) && (51)\\
&= \int_\mathcal{X} \langle \psi(f(x)) - \psi(f_0(x)), V\left(\int_\mathcal{Y} \psi(y) d\rho(y|x)\right)\rangle_{\mathcal{H}_\mathcal{Y}} d\rho_\mathcal{X}(x) && (52)\\
&= \int_\mathcal{X} \langle \psi(f(x)) - \psi(f_0(x)), Vg^*(x)\rangle_{\mathcal{H}_\mathcal{Y}} d\rho_\mathcal{X}(x) && (53)\\
&= A + B. && (54)
\end{aligned}$$

where

$$\begin{aligned}
A &= \int_\mathcal{X} \langle \psi(f(x)), V(g^*(x) - g(x))\rangle_{\mathcal{H}_\mathcal{Y}} \, d\rho_\mathcal{X}(x) && (55)\\
B &= \int_\mathcal{X} \langle \psi(f(x)), Vg(x)\rangle_{\mathcal{H}_\mathcal{Y}} d\rho_\mathcal{X}(x) - \int_\mathcal{X} \langle \psi(f_0(x)), Vg^*(x)\rangle_{\mathcal{H}_\mathcal{Y}} d\rho_\mathcal{X}(x) && (56)
\end{aligned}$$

Now, the term A can be minimized by taking the supremum over $\mathcal{Y}$ so that

$$A \leq \int_\mathcal{X} \sup_{y\in\mathcal{Y}} \left|\langle \psi(y), V(g^*(x) - g(x))\rangle_{\mathcal{H}_\mathcal{Y}}\right| d\rho_\mathcal{X}(x). \tag{57}$$

For B, we observe that, by the definition of the decoding d, we have

$$\langle \psi(f_0(x)), Vg^*(x)\rangle_{\mathcal{H}_\mathcal{Y}} = \inf_{y'\in\mathcal{Y}} \langle \psi(y'), Vg^*(x)\rangle_{\mathcal{H}_\mathcal{Y}}, \tag{58}$$

$$\langle \psi(f(x)), Vg(x)\rangle_{\mathcal{H}_\mathcal{Y}} = \inf_{y'\in\mathcal{Y}} \langle \psi(y'), Vg(x)\rangle_{\mathcal{H}_\mathcal{Y}}, \tag{59}$$

for all $x \in \mathcal{X}$l. Therefore,

$$B = \int_\mathcal{X} \inf_{y\in\mathcal{Y}} \langle \psi(y), Vg(x)\rangle_{\mathcal{H}_\mathcal{Y}} - \inf_{y\in\mathcal{Y}} \langle \psi(y), Vg^*(x)\rangle_{\mathcal{H}_\mathcal{Y}}\, d\rho_\mathcal{X}(x) \tag{60}$$

$$\leq \int_\mathcal{X} \sup_{y\in\mathcal{Y}} \Big| \langle \psi(y), V(g(x) - g^*(x))\rangle_{\mathcal{H}_\mathcal{Y}} \Big| d\rho_\mathcal{X}(x) \tag{61}$$

where we have used the fact that for any given two functions $\eta, \zeta : \mathcal{Y} \to \mathbb{R}$ we have

$$|\inf_{y\in\mathcal{Y}} \eta(y) - \inf_{y\in\mathcal{Y}} \zeta(y)| \leq \sup_{y\in\mathcal{Y}} |\eta(y) - \zeta(y)|. \tag{62}$$

Therefore, by combining the bounds on A and B we have

$$\begin{aligned}
\mathcal{E}(f) - \mathcal{E}(f^*) &\leq 2 \int_\mathcal{X} \sup_{y\in\mathcal{Y}} \Big| \langle \psi(y), V(g^*(x) - g(x))\rangle_{\mathcal{H}_\mathcal{Y}} \Big| d\rho_\mathcal{X}(x) \\
&\leq 2 \int_\mathcal{X} \sup_{y\in\mathcal{Y}} \|V^*\psi(y)\|_{\mathcal{H}_\mathcal{Y}} \|g^*(x) - g(x)\|_{\mathcal{H}_\mathcal{Y}} d\rho_\mathcal{X}(x) \\
&\leq 2Q\|V\| \int_\mathcal{X} \|g^*(x) - g(x)\|_{\mathcal{H}_\mathcal{Y}} d\rho_\mathcal{X}(x) \\
&\leq 2Q\|V\| \sqrt{\int_\mathcal{X} \|g^*(x) - g(x)\|^2_{\mathcal{H}_\mathcal{Y}} d\rho_\mathcal{X}(x)},
\end{aligned}$$

where for the last inequality we have used the Jensen's inequality. The proof is concluded by recalling that (see Eq. (33))

$$\mathcal{R}(g) - \mathcal{R}(g^*) = \int_\mathcal{X} \|g(x) - g^*(x)\|^2_{\mathcal{H}_\mathcal{Y}} d\rho_\mathcal{X}(x) \tag{63}$$

□

B Learning Bounds for Structured Prediction

In this section we focus on the analysis of the structured prediction algorithm proposed in this work (Alg. 1). In particular, we will first prove that, given the minimizer $\hat{g} : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ of the empirical risk at Eq. (4), its decoding can be computed in practice according to Alg. 1. Then, we report the proofs for the universal consistency of such approach (Thm. 4) and generalization bounds (Thm. 5).

Notation

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a positive semidefinite function on $\mathcal{X}$, we denote $\mathcal{H}_\mathcal{X}$ the Hilbert space obtained by the completion

$$\mathcal{H}_\mathcal{X} = \overline{\text{span}\{k(x, \cdot) \mid x \in \mathcal{X}\}} \tag{64}$$

according to the norm induced by the inner product $\langle k(x, \cdot), k(x', \cdot)\rangle_{\mathcal{H}_\mathcal{X}} = k(x, x')$. Spaces $\mathcal{H}_\mathcal{X}$ constructed in this way are known as *reproducing kernel Hilbert spaces* and there is a one-to-one relation between a kernel k and its associated RKHS. For more details on RKHS we refer the reader to [5]. Given a kernel k, in the following we will denote with $\varphi : \mathcal{X} \to \mathcal{H}_\mathcal{X}$ the feature map $\varphi(x) = k(x, \cdot) \in \mathcal{H}_\mathcal{X}$ for all $x \in \mathcal{X}$. We say that a kernel is bounded if $\|\varphi(x)\|_{\mathcal{H}_\mathcal{X}} \leq \kappa$ with $\kappa > 0$. Note that k is bounded if and only if $k(x, x') = \langle \varphi(x), \varphi(x')\rangle_{\mathcal{H}_\mathcal{X}} \leq \|\varphi(x)\|_{\mathcal{H}_\mathcal{X}} \|\varphi(x')\| \leq \kappa^2$ for every $x, x' \in \mathcal{X}$. In the following we will always assume k to be continuous and bounded by $\kappa > 0$. The continuity of k with the fact that $\mathcal{X}$ is Polish implies $\mathcal{H}_\mathcal{X}$ to be separable [5].

We introduce here the ideal and empirical operators that we will use in the following to prove the main results of this work.

- $S : \mathcal{H}_\mathcal{X} \to L^2(\mathcal{X}, \rho_\mathcal{X})$ s.t. $f \in \mathcal{H}_\mathcal{X} \mapsto \langle f, \varphi(\cdot)\rangle_{\mathcal{H}_\mathcal{X}} \in L^2(\mathcal{X}, \rho_\mathcal{X})$, with adjoint
- $S^* : L^2(\mathcal{X}, \rho_\mathcal{X}) \to \mathcal{H}_\mathcal{X}$ s.t. $h \in L^2(\mathcal{X}, \rho_\mathcal{X}) \mapsto \int_\mathcal{X} h(x)\varphi(x) d\rho_\mathcal{X}(x) \in \mathcal{H}_\mathcal{X}$,
- $Z : \mathcal{H}_\mathcal{Y} \to L^2(\mathcal{X}, \rho_\mathcal{X})$ s.t. $h \in \mathcal{H}_\mathcal{Y} \mapsto \langle h, g^*(\cdot)\rangle_{\mathcal{H}_\mathcal{Y}} \in L^2(\mathcal{X}, \rho_\mathcal{X})$, with adjoint
- $Z^* : L^2(\mathcal{X}, \rho_\mathcal{X}) \to \mathcal{H}_\mathcal{Y}$ s.t. $h \in L^2(\mathcal{X}, \rho_\mathcal{X}) \mapsto \int_\mathcal{X} h(x) g^*(x) d\rho_\mathcal{X}(x) \in \mathcal{H}_\mathcal{Y}$,
- $C = S^*S : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{X}$ and $L = SS^* : L^2(\mathcal{X}, \rho_\mathcal{X}) \to L^2(\mathcal{X}, \rho_\mathcal{X})$,

with $g^*(x) = \int_\mathcal{Y} \psi(y) d\rho(y|x)$ defined according to Eq. (29), (see Lemma 8).

Given a set of input-output pairs $\{(x_i, y_i)\}_{i=1}^n$ with $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ independently sampled according to ρ on $\mathcal{X} \times \mathcal{Y}$, we define the empirical counterparts of the operators just defined as

- $\hat{S} : \mathcal{H}_\mathcal{X} \to \mathbb{R}^n$ s.t. $f \in \mathcal{H}_\mathcal{X} \mapsto \frac{1}{\sqrt{n}}(\langle \varphi(x_i), f\rangle_{\mathcal{H}_\mathcal{X}})_{i=1}^n \in \mathbb{R}^n$, with adjoint
- $\hat{S}^* : \mathbb{R}^n \to \mathcal{H}_\mathcal{X}$ s.t. $v = (v_i)_{i=1}^n \in \mathbb{R}^n \mapsto \frac{1}{\sqrt{n}} \sum_{i=1}^n v_i \varphi(x_i)$,
- $\hat{Z} : \mathcal{H}_\mathcal{Y} \to \mathbb{R}^n$ s.t. $h \in \mathcal{H}_\mathcal{Y} \mapsto \frac{1}{\sqrt{n}}(\langle \psi(y_i), h\rangle_{\mathcal{H}_\mathcal{Y}})_{i=1}^n \in \mathbb{R}^n$, with adjoint
- $\hat{Z}^* : \mathbb{R}^n \to \mathcal{H}_\mathcal{Y}$ s.t. $v = (v_i)_{i=1}^n \in \mathbb{R}^n \mapsto \frac{1}{\sqrt{n}} \sum_{i=1}^n v_i \psi(y_i)$,
- $\hat{C} = \hat{S}^*\hat{S} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{X}$ and $K = n\hat{S}\hat{S}^* \in \mathbb{R}^{n \times n}$ is the empirical kernel matrix.

In the rest of this section we denote with $A + \lambda$, the operator $A + \lambda I$, for any symmetric linear operator A, $\lambda \in \mathbb{R}$ and I the identity operator.

We recall here a basic result characterizing the operators introduced above.

Proposition 13. *With the notation introduced above,*

$$C = \int_\mathcal{X} \varphi(x) \otimes \varphi(x) d\rho_\mathcal{X}(x) \quad \text{and} \quad Z^*S = \int_{\mathcal{X}\times\mathcal{Y}} \psi(y) \otimes \varphi(x) d\rho(x, y) \tag{65}$$

where $\otimes$ denotes the tensor product. Moreover, when φ and ψ are bounded by respectively κ and Q, we have the following facts

(i) $\text{Tr}(L) = \text{Tr}(C) = \|S\|_{HS}^2 = \int_\mathcal{X} \|\varphi(x)\|_{\mathcal{H}_\mathcal{X}}^2 d\rho_\mathcal{X}(x) \leq \kappa^2$

(ii) $\|Z\|_{HS}^2 = \int_X \|g^*(x)\|^2 d\rho_\mathcal{X}(x) = \|g^*\|_{\rho_\mathcal{X}}^2 < +\infty$.

Proof. By definition of $C = S^*S$, for each $h, h' \in \mathcal{H}_\mathcal{X}$ we have

$$\begin{aligned}
\langle h, Ch'\rangle_{\mathcal{H}_\mathcal{X}} = \langle Sh, Sh'\rangle_{\rho_\mathcal{X}} &= \int_\mathcal{X} \langle h, \varphi(x)\rangle_{\mathcal{H}_\mathcal{X}} \langle \varphi(x), h'\rangle_{\mathcal{H}_\mathcal{X}} d\rho_\mathcal{X}(x) && (66)\\
&= \int_\mathcal{X} \Big\langle h, \Big(\varphi(x)\langle \varphi(x), h'\rangle_{\mathcal{H}_\mathcal{X}}\Big)\Big\rangle_{\mathcal{H}_\mathcal{X}} d\rho_\mathcal{X}(x) && (67)\\
&= \int_\mathcal{X} \Big\langle h, \Big(\varphi(x) \otimes \varphi(x)\Big) h'\Big\rangle d\rho_\mathcal{X}(x) && (68)\\
&= \langle h, \Big(\int_\mathcal{X} \varphi(x) \otimes \varphi(x) d\rho_\mathcal{X}(x)\Big) h'\rangle_{\mathcal{H}_\mathcal{X}} && (69)
\end{aligned}$$

since $\varphi(x) \otimes \varphi(x) : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{X}$ is the operator such that $h \in \mathcal{H}_\mathcal{X} \mapsto \varphi(x)\langle \varphi(x), h\rangle_{\mathcal{H}_\mathcal{X}}$. The characterization for Z^*S is analogous.

Now, (i). The relation $\text{Tr}(L) = \text{Tr}(C) = \text{Tr}(S^*S) = \|S\|_{HS}^2$ holds by definition. Moreover

$$\text{Tr}(C) = \int_\mathcal{X} \text{Tr}(\varphi(x) \otimes \varphi(x)) d\rho_\mathcal{X}(x) = \int_\mathcal{X} \|\varphi(x)\|_{\mathcal{H}_\mathcal{X}}^2 d\rho_\mathcal{X}(x) \tag{70}$$

by linearity of the trace. (ii) is analogous. Note that $\|g^*\|_{\rho_\mathcal{X}}^2 < +\infty$. by Lemma 8 since ψ is bounded by hypothesis.

□

B.1 Reproducing Kernel Hilbert Spaces for Vector-valued Functions

We begin our analysis by introducing the concept of reproducing kernel Hilbert space (RKHS) for vector-valued functions. Here we provide a brief summary of the main properties that will be useful in the following. We refer the reader to [6, 7] for a more in-depth introduction on the topic.

Analogously to the case of scalar functions, a RKHS for vector-valued functions $g : \mathcal{X} \to \mathcal{H}$, with $\mathcal{H}$ a separable Hilbert space, is uniquely characterized by a so-called *kernel of positive type*, which is an operator-valued $\Gamma : \mathcal{X} \times \mathcal{X} \to \mathcal{B}(\mathcal{H}, \mathcal{H})$ generalizing the concept of scalar reproducing kernel.

Definition 14. *Let $\mathcal{X}$ be a set and $\mathcal{H}$ be a Hilbert space, then $\Gamma : \mathcal{X} \times \mathcal{X} \to \mathcal{B}(\mathcal{H}, \mathcal{H})$ is a kernel of positive type if for each $n \in \mathbb{N}$, $x_1, \dots, x_n \in \mathcal{X}$, $c_1, \dots, c_n \in \mathcal{H}$ we have*

$$\sum_{i,j=1}^{n} \langle \Gamma(x_i, x_j)c_i, c_j \rangle_{\mathcal{H}} \geq 0 \tag{71}$$

A kernel of positive type Γ defines an inner product $\langle \Gamma(x, \cdot)c, \Gamma(x', \cdot)c' \rangle_{\mathcal{G}_0} = \langle \Gamma(x, x')c, c' \rangle_{\mathcal{H}}$ on the space

$$\mathcal{G}_0 = \text{span}\{\Gamma(x, \cdot)c \mid x \in \mathcal{X}, c \in \mathcal{H}\}. \tag{72}$$

Then, the completion $\mathcal{G} = \overline{\mathcal{G}_0}$ with respect to the norm induced by $\langle \cdot, \cdot \rangle_{\mathcal{G}_0}$ is known as the reproducing Kernel Hilbert space (RKHS) for vector-valued functions associated to the kernel Γ. Indeed, we have that a reproducing property holds also for RKHS of vector-valued functions, namely for any $x \in \mathcal{X}$, $c \in \mathcal{H}$ and $g \in \mathcal{G}$ we have

$$\langle g(x), c \rangle_{\mathcal{H}} = \langle g, \Gamma(x, \cdot)c \rangle_{\mathcal{G}} \tag{73}$$

and that for each $x \in \mathcal{X}$ the function $\Gamma(x, \cdot) : \mathcal{G} \to \mathcal{H}$ is the evaluation functional in x on $\mathcal{G}$, namely $\Gamma(x, \cdot)(g) = g(x)$.

B.1.1 Separable Vector Valued Kernels

In this work we restrict to the special case of RKHS for vector-valued functions with associated kernel $\Gamma : \mathcal{X} \times \mathcal{X} \to \mathcal{H}_{\mathcal{Y}}$ of the form $\Gamma(x, x') = k(x, x')I_{\mathcal{H}_{\mathcal{Y}}}$ for each $x, x' \in \mathcal{X}$, where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a scalar reproducing kernel and $I_{\mathcal{H}_{\mathcal{Y}}}$ is the identity operator on $\mathcal{H}_{\mathcal{Y}}$. Notice that this choice is not restrictive in terms of the space of functions that can be learned by our algorithm. Indeed, it was proven in [8] (see Example 14) that if k is a universal scalar kernel, then $\Gamma(\cdot, \cdot) = k(\cdot, \cdot)I_{\mathcal{H}_{\mathcal{Y}}}$ is universal. Below, we report a useful characterization of RKHS $\mathcal{G}$ associated to a separable kernel.

Lemma 15. *The RKHS $\mathcal{G}$ associated to the kernel $\Gamma(x, x') = k(x, x')I_{\mathcal{H}_{\mathcal{Y}}}$ is isometric to $\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}$ and for each $g \in \mathcal{G}$ there exists a unique $G \in \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}$ such that*

$$g(x) = G\varphi(x) \in \mathcal{H}_{\mathcal{Y}} \quad \textit{for each } \; x \in \mathcal{X} \tag{74}$$

Proof. We explicitly define the isometry $T : \mathcal{G} \to \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}$ as the linear operator such that $T(\sum_{i=1}^{n} \alpha_i \Gamma(x_i, \cdot)c_i) = \sum_{i=1}^{n} \alpha_i \, c_i \otimes \varphi(x_i)$ for each $n \in \mathbb{N}$, $x_1, \dots, x_n \in \mathcal{X}$ and $c_1, \dots, c_n \in \mathcal{H}_{\mathcal{Y}}$. By construction, $T(\mathcal{G}_0) \subseteq \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}$, with $\mathcal{G}_0$ the linear space defined at Eq. (72) and moreover

$$\langle T(\Gamma(x, \cdot)c), T(\Gamma(x', \cdot)c') \rangle_{HS} = \langle c \otimes \varphi(x)^*, c' \otimes \varphi(x')^* \rangle_{HS} = \langle c, c' \rangle_{\mathcal{H}_{\mathcal{Y}}} \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}_{\mathcal{X}}} \tag{75}$$

$$= \langle k(x, x')c, c' \rangle_{\mathcal{H}_{\mathcal{Y}}} = \langle \Gamma(x, x')c, c' \rangle_{\mathcal{H}_{\mathcal{Y}}} = \langle \Gamma(x, \cdot)c, \Gamma(x', \cdot)c' \rangle_{\mathcal{G}} \tag{76}$$

implying that $\mathcal{G}_0$ is isometrically contained in $\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}$. Since $\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}$ is complete, also $\overline{T(\mathcal{G}_0)} \subseteq \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}$. Therefore $\mathcal{G}$ is isometrically contained in $\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}$, since $T(\mathcal{G}) = T(\overline{\mathcal{G}_0}) = \overline{T(\mathcal{G}_0)}$. Moreover note that

$$\overline{T(\mathcal{G}_0)} = \overline{\text{span}\{c \otimes \varphi(x) \mid x \in \mathcal{X}, c \in \mathcal{H}_{\mathcal{Y}}\}} = \overline{\text{span}\{c \mid c \in \mathcal{H}_{\mathcal{Y}}\} \otimes \text{span}\{\varphi(x) \mid x \in \mathcal{X}\}} \tag{77}$$

$$= \overline{\mathcal{H}_{\mathcal{Y}} \otimes \text{span}\{\varphi(x) \mid x \in \mathcal{X}\}} = \overline{\mathcal{H}_{\mathcal{Y}}} \otimes \overline{\text{span}\{\varphi(x) \mid x \in \mathcal{X}\}} = \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}} \tag{78}$$

from which we conclude that $\mathcal{G}$ is isometric to $\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}$ via T.

To prove Eq. (74), let us consider $x \in \mathcal{X}$ and $g \in \mathcal{G}$ with $G = T(g) \in \mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{X}}$. Then, $\forall c \in \mathcal{H}_{\mathcal{Y}}$ we have that

$$\langle c, g(x) \rangle_{\mathcal{H}_{\mathcal{Y}}} = \langle \Gamma(x, \cdot)c, g \rangle_{\mathcal{G}} = \langle T(\Gamma(x, \cdot)c), G \rangle_{HS} \tag{79}$$

$$= \langle c \otimes \varphi(x), G \rangle_{HS} = \text{Tr}((\varphi(x) \otimes c)G) = \text{Tr}(c^* G\varphi(x)) = \langle c, G\varphi(x) \rangle_{\mathcal{H}_{\mathcal{Y}}}. \tag{80}$$

Since the equation above is true for each $c \in \mathcal{H}_{\mathcal{Y}}$ we can conclude that $g(x) = T(g)\varphi(x)$ as desired. □

The isometry $\mathcal{G} \simeq \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$ allows to characterize the closed form solution for the surrogate risk introduced in Eq. (12). We recall that in Lemma 8, we have shown that $\mathcal{R}$ always attains a minimizer on $L^2(\mathcal{X}, \rho_\mathcal{X}, \mathcal{H}_\mathcal{Y})$. In the following we show that if $\mathcal{R}$ attains a minimum on $\mathcal{G} \simeq \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$, we are able to provide a close form solution for one such element.

Lemma 16. *Let ψ and $\mathcal{H}_\mathcal{Y}$ satisfying Asm. 1 and assume that the surrogate expected risk minimization of $\mathcal{R}$ at Eq. (12) attains a minimum on $\mathcal{G}$, with $\mathcal{G} \simeq \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$. Then the minimizer $g^* \in \mathcal{G}$ of $\mathcal{R}$ with minimal norm $\|\cdot\|_\mathcal{G}$ is of the form*

$$g^*(x) = G\varphi(x), \quad \forall x \in \mathcal{X} \quad \textit{with} \quad G = Z^* S C^\dagger \in \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}. \tag{81}$$

Proof. By hypothesis we have $g^* \in \mathcal{G}$. Therefore, by applying Lemma 15 we have that there exists a unique linear operator $G : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ such that $g^*(x) = G\varphi(x), \forall x \in \mathcal{X}$. Now, expanding the least squares loss on $\mathcal{H}_\mathcal{Y}$, we obtain

$$\mathcal{R}(g) = \int_{\mathcal{X}\times\mathcal{Y}} \|G\varphi(x) - \psi(y)\|^2_{\mathcal{H}_\mathcal{Y}} d\rho(x,y) \tag{82}$$

$$= \mathrm{Tr}(G(\varphi(x)\otimes\varphi(x))G^*)d\rho_\mathcal{X}(x,y) - 2\mathrm{Tr}(G(\varphi(x)\otimes\psi(y))) + \int_{\mathcal{X}\times\mathcal{Y}} \|\psi(y)\|^2_{\mathcal{H}_\mathcal{Y}} \tag{83}$$

$$= \mathrm{Tr}(GCG^*) - 2\mathrm{Tr}(GS^*Z) + const. \tag{84}$$

where we have used Prop. 13 and the linearity of the trace. Therefore $\mathcal{R}$ is a quadratic functional, and is convex since C is positive semidefinite. We can conclude that $\mathcal{R}$ attains a minimum on $\mathcal{G}$ if and only if the range of S^*Z is contained in the range of C, namely $\mathrm{Ran}(S^*Z) \subseteq \mathrm{Ran}(C) \subset \mathcal{H}_\mathcal{X}$ (see [9] Chap. 2). In this case $G = Z^*SC^\dagger \in \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$ exists and is the minimum norm minimizer for $\mathcal{R}$, as desired. □

Analogously to Lemma 16, a closed form solution exists for the regularized surrogate empirical risk minimization problem introduced in Eq. (4). We recall that the associated functional is $\hat{\mathcal{R}}_\lambda : \mathcal{G} \to \mathbb{R}$ defined as

$$\hat{\mathcal{R}}_\lambda(g) = \hat{\mathcal{R}}(g) + \lambda \|g\|^2_\mathcal{G} = \frac{1}{n}\sum_{i=1}^n \|g(x_i) - \psi(y_i)\|^2_{\mathcal{H}_\mathcal{Y}} + \lambda \|g\|^2_{\mathcal{H}_\mathcal{Y}} \tag{85}$$

for all $g \in \mathcal{G}$ and $\{(x_i, y_i)\}_{i=1}^n$ points in $\mathcal{X} \times \mathcal{Y}$. The following result characterizes the closed form solution for the empirical risk minimization when $\mathcal{G} \simeq \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$ and guarantees that such a solution always exists.

Lemma 17. *Let $\mathcal{G} \simeq \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$. For any $\lambda > 0$, the solution $\hat{g}_\lambda \in \mathcal{G}$ of the empirical risk minimization problem at Eq. (4) exists, is unique and is such that*

$$\hat{g}_\lambda(x) = \hat{G}_\lambda \varphi(x), \quad \forall x \in \mathcal{X} \quad \textit{with} \quad \hat{G}_\lambda = \hat{Z}^* \hat{S}(\hat{C} + \lambda)^{-1} \in \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}. \tag{86}$$

Proof. The proof of is analogous to that of Lemma 16 and we omit it. Note that, since $(\hat{C} + \lambda)^{-1}$ is always bounded for $\lambda > 0$, its range corresponds to $\mathcal{H}_\mathcal{X}$ and therefore the range of S^*Z is always contained in it. Then $\hat{G}_\lambda$ exists for any $\lambda > 0$ and is unique since $\|\cdot\|^2_{\mathcal{H}_\mathcal{Y}}$ is strictly convex. □

The closed form solutions provided by Lemma 16 and Lemma 17 will be key in the analysis of the structured prediction algorithm Alg. 1 in the following.

B.2 The Structured Prediction Algorithm

In this section we prove that Alg. 1 corresponds to the decoding of the surrogate empirical risk minimizer $\hat{g}$ (Eq. (4)) via a map $d : \mathcal{H}_\mathcal{Y} \to \mathcal{Y}$ satisfying Eq. (13).

Recall that in Lemma 15 we proved that the vector-valued RKHS $\mathcal{G}$ induced by a kernel $\Gamma(x, x') = k(x, x')I_{\mathcal{H}_\mathcal{Y}}$, for a scalar kernel k on $\mathcal{X}$, is isometric to $\mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$. For the sake of simplicity, in the following, with some abuse of notation, we will not make the distinction between $\mathcal{G}$ and $\mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$ when it is clear from context.

Lemma 3. *Let $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfy Asm. 1 with $\mathcal{Y}$ a compact set. Let $\hat{g} : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ be the minimizer of Eq. (4). Then, for all $x \in \mathcal{X}$*

$$d \circ \hat{g}(x) \;=\; \operatorname*{argmin}_{y \in \mathcal{Y}} \sum_{i=1}^{n} \alpha_i(x) \triangle (y, y_i) \qquad \alpha(x) = (\mathbf{K} + n\lambda I)^{-1} \mathbf{K}_x \in \mathbb{R}^n \tag{16}$$

Proof. From Lemma 17 we know that $\hat{g}(x) = \hat{Z}^* \hat{S} (\hat{C} + \lambda)^{-1} \varphi(x)$ for all $x \in \mathcal{X}$. Recall that $\hat{C} = \hat{S}^* \hat{S}$ and $\mathbf{K} = n \hat{S} \hat{S}^* \in \mathbb{R}^{n \times n}$, is the empirical kernel matrix associated to the inputs, namely such that $\mathbf{K}_{ij} = k(x_i, x_j)$ for each $i, j = 1, \ldots, n$. Therefore we have $\hat{S}(\hat{C} + \lambda)^{-1} = \sqrt{n}(\mathbf{K} + \lambda n)^{-1} \hat{S} : \mathcal{H}_\mathcal{X} \to \mathbb{R}^n$. Now, by denoting $\mathbf{K}_x = \sqrt{n} \hat{S} \varphi(x) = (k(x_i, x))_{i=1}^n \in \mathbb{R}^n$, we have

$$\hat{S}(\hat{C} + \lambda)^{-1} \varphi(x) = (\mathbf{K} + \lambda n I)^{-1} \mathbf{K}_x = \alpha(x) \in \mathbb{R}^n. \tag{87}$$

Therefore, by applying the definition of the operator $\hat{Z} : \mathcal{H}_\mathcal{Y} \to \mathbb{R}^n$, we have

$$\hat{g}(x) = \hat{Z}^* \hat{S} (\hat{C} + \lambda)^{-1} \varphi(x) = \hat{Z}^* \alpha(x) = \sum_{i=1}^{n} \alpha_i(x) \psi(y_i), \quad \forall x \in \mathcal{X} \tag{88}$$

By plugging $\hat{g}(x)$ in the functional minimized by the decoding (Eq. (13)),

$$\langle \psi(y), V \hat{g}(x) \rangle_{\mathcal{H}_\mathcal{Y}} = \langle \psi(y), V \hat{Z}^* \alpha(x) \rangle_{\mathcal{H}_\mathcal{Y}} = \langle \psi(y), \sum_{i=1}^{n} \alpha_i(x) \psi(y_i) \rangle_{\mathcal{H}_\mathcal{Y}} \tag{89}$$

$$= \sum_{i=1}^{n} \alpha_i(x) \langle \psi(y), V \psi(y_i) \rangle_{\mathcal{H}_\mathcal{Y}} = \sum_{i=1}^{n} \alpha_i(x) \triangle (y, y_i), \tag{90}$$

where we have used the bilinear form for $\triangle$ in Asm. 1 for the final equation. We conclude that

$$d \circ \hat{g}(x) \in \operatorname*{argmin}_{y \in \mathcal{Y}} \langle \psi(y), V \hat{g}(x) \rangle_{\mathcal{H}_\mathcal{Y}} = \operatorname*{argmin}_{y \in \mathcal{Y}} \sum_{i=1}^{n} \alpha_i(x) \triangle (y, y_i) \tag{91}$$

as required. □

Lemma 3 focuses on the computational aspects of Alg. 1. In the following we will analyze its statistical properties.

B.3 Universal Consistency

In following, we provide a probabilistic bound on the excess risk $\mathcal{E}(d \circ g) - \mathcal{E}(f^*)$ for any $g \in \mathcal{G} \simeq \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$ that will be key to prove both universal consistency (Thm. 4) and generalization bounds (Thm. 5). To do so, we will make use of the comparison inequality from Thm. 12 to control the structured excess risk by means of the excess risk of the surrogate $\mathcal{R}(g) - \mathcal{R}(g^*)$. Note that the surrogate problem consists in a vector-valued kernel ridge regression estimation. In this setting, the problem of finding a probabilistic bound has been studied (see [10] and references therein). Indeed, our proof will consist of a decomposition of the surrogate excess risk that is similar to the one in [10]. However, note that we cannot direct apply [10] to our setting, since in [10] the operator-valued kernel Γ associated to $\mathcal{G}$ is required to be such that $\Gamma(x, x')$ is trace class $\forall x, x' \in \mathcal{X}$, which does not hold for the kernel used in this work, namely $\Gamma(x, x') = k(x, x') I_{\mathcal{H}_\mathcal{Y}}$ when $\mathcal{H}_\mathcal{Y}$ is infinite dimensional.

In order to express the bound on the excess risk more compactly, here we introduce a measure for the approximation error of the surrogate problem. According to [10], we define the following quantity

$$\mathcal{A}(\lambda) = \lambda \| Z^* (L + \lambda)^{-1} \|_{HS}. \tag{92}$$

Lemma 18. *Let $\mathcal{Y}$ be compact, $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfying Asm. 1 and k a bounded positive definite kernel on $\mathcal{X}$ with $\sup_{x \in \mathcal{X}} k(x, x) \leq \kappa^2$ and associated RKHS $\mathcal{H}_\mathcal{X}$. Let ρ a Borel probability measure on $\mathcal{X} \times \mathcal{Y}$ and $\{(x_i, y_i)\}_{i=1}^n$ independently sampled according to ρ. Let f^* be a solution of the*

problem in Eq. (1), $\hat{f} = d \circ \hat{g}$ as in Alg. 1. Then, for any $\lambda \leq \kappa^2$ and $\delta > 0$, the following holds with probability $1 - \delta$:

$$\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) \leq 8\kappa Q \|V\| \frac{Q + \mathcal{B}(\lambda)}{\sqrt{\lambda n}} \left(1 + \sqrt{\frac{4\kappa^2}{\lambda\sqrt{n}}}\right) \log^2 \frac{8}{\delta} + 2Q\|V\|\mathcal{A}(\lambda), \tag{93}$$

with $\mathcal{B}(\lambda) = \kappa \|Z^* S (C + \lambda)^{-1}\|_{HS}$.

Proof. According to Thm. 12

$$\mathcal{E}(d \circ \hat{g}) - \mathcal{E}(f^*) \leq 2Q\|V\| \sqrt{\mathcal{R}(\hat{g}) - \mathcal{R}(g^*)}. \tag{94}$$

From Lemma 17 we know that $\hat{g}(x) = \hat{G}\varphi(x)$ for all $x \in \mathcal{X}$, with $\hat{G} = \hat{Z}^* \hat{S} (\hat{C} + \lambda)^{-1} \in \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$. By Lemma 8, we know that $g^*(x) = \int_\mathcal{Y} \psi(y) d\rho(y|x)$ almost everywhere on the support of $\rho_\mathcal{X}$. Therefore, a direct application of Prop. 13 leads to

$$\mathcal{R}(\hat{g}) - \mathcal{R}(g^*) = \int \|\hat{g}(x) - g^*(x)\|^2_{\mathcal{H}_\mathcal{Y}} d\rho_\mathcal{X}(x) = \tag{95}$$

$$= \int_\mathcal{X} \|\hat{G}\varphi(x)\|^2_{\mathcal{H}_\mathcal{Y}} - 2\langle \hat{G}\varphi(x), g^*(x)\rangle_{\mathcal{H}_\mathcal{Y}} + \|g^*(x)\|^2_{\mathcal{H}_\mathcal{Y}} d\rho_\mathcal{X}(x) \tag{96}$$

$$= \int_\mathcal{X} \mathrm{Tr}\left(\hat{G}\Big(\varphi(x) \otimes \varphi(x)\Big)\hat{G}^*\right) - 2\mathrm{Tr}\left(\hat{G}\Big(\varphi(x) \otimes g^*(x)\Big)\right) + \mathrm{Tr}(g^*(x) \otimes g^*(x)) d\rho_\mathcal{X}(x) \tag{97}$$

$$= \mathrm{Tr}(\hat{G}S^* S\hat{G}) - 2\mathrm{Tr}(\hat{G}S^* Z) + \mathrm{Tr}(Z^* Z) = \|\hat{G}S^* - Z^*\|^2_{HS} \tag{98}$$

To bound $\|\hat{G}S^* - Z^*\|_{HS}$, we proceed with a decomposition similar to the one in [10]. In particular $\|\hat{G}S^* - Z^*\|_{HS} \leq A_1 + A_2 + A_3$, with

$$A_1 = \|\hat{Z}^* \hat{S} (\hat{C} + \lambda)^{-1} S^* - Z^* S (\hat{C} + \lambda)^{-1} S^*\|_{HS} \tag{99}$$

$$A_2 = \|Z^* S (\hat{C} + \lambda)^{-1} S^* - Z^* S (C + \lambda)^{-1} S^*\|_{HS} \tag{100}$$

$$A_3 = \|Z^* S (C + \lambda)^{-1} S^* - Z^*\|_{HS}. \tag{101}$$

Let $\tau = \delta/4$. Now, for the term A_1, we have

$$A_1 \leq \|\hat{Z}^* \hat{S} - Z^* S\|_{HS} \|(\hat{C} + \lambda)^{-1} S^*\|. \tag{102}$$

To control the term $\|\hat{Z}^* \hat{S} - Z^* S\|_{HS}$, note that $\hat{Z}^* \hat{S} = \frac{1}{n} \sum_{i=1}^n \zeta_i$ with ζ_i the random variable $\zeta_i = \psi(y_i) \otimes \varphi(x_i)$. By Prop. 13, for any $1 \leq i \leq n$ we have

$$\mathbb{E}\zeta_i = \int \psi(y) \otimes \varphi(x) d\rho(x, y) = Z^* S, \tag{103}$$

and

$$\|\zeta_i\|_{HS} \leq \sup_{y \in \mathcal{Y}} \|\psi(y)\|_{\mathcal{H}_\mathcal{Y}} \sup_{x \in \mathcal{X}} \|\varphi(x)\|_{\mathcal{H}_\mathcal{X}} \leq Q\kappa, \tag{104}$$

almost surely on the support of ρ on $\mathcal{X} \times \mathcal{Y}$, and so $\mathbb{E}\|\zeta_i\|^2_{HS} \leq Q^2\kappa^2$. Thus, by applying Lemma 2 of [11], we have

$$\|\hat{Z}^* \hat{S} - Z^* S\|_{HS} \leq \frac{2Q\kappa \log \frac{2}{\tau}}{n} + \sqrt{\frac{2Q^2\kappa^2 \log \frac{2}{\tau}}{n}} \leq \frac{4Q\kappa \log \frac{2}{\tau}}{n} \tag{105}$$

with probability $1 - \tau$, since $\log 2/\tau \geq 1$. To control $\|(\hat{C} + \lambda)^{-1} S^*\|$ we proceed by recalling that $C = S^* S$ and that for any $\lambda > 0$ $\|(\hat{C} + \lambda)^{-1}\| \leq \lambda^{-1}$ and $\|\hat{C}(\hat{C} + \lambda)^{-1}\| \leq 1$. We have

$$\|(\hat{C} + \lambda)^{-1} S^*\| = \|(\hat{C} + \lambda)^{-1} C (\hat{C} + \lambda)^{-1}\|^{1/2} \tag{106}$$

$$\leq \|(\hat{C} + \lambda)^{-1} (C - \hat{C})(\hat{C} + \lambda)^{-1}\|^{1/2} + \|(\hat{C} + \lambda)^{-1} \hat{C} (\hat{C} + \lambda)^{-1}\|^{1/2} \tag{107}$$

$$\leq \|(\hat{C} + \lambda)^{-1}\| \|C - \hat{C}\|^{1/2} + \|(\hat{C} + \lambda)^{-1}\|^{1/2} \|\hat{C}(\hat{C} + \lambda)^{-1}\|^{1/2} \tag{108}$$

$$\leq \lambda^{-1/2} (1 + \lambda^{-1/2} \|C - \hat{C}\|^{1/2}). \tag{109}$$

To control $\|C - \hat{C}\|$, note that $\hat{C} = \frac{1}{n}\sum_{i=1}^{n} \zeta_i$ where ζ_i is the random variable defined as $\zeta_i = \varphi(x_i) \otimes \varphi(x_i)$ for $1 \leq i \leq n$. Note that $\mathbb{E}\zeta_i = C$, $\|\zeta_i\| \leq \kappa^2$ almost surely and so $\mathbb{E}\|\zeta_i\|^2 \leq \kappa^4$ for $1 \leq i \leq n$. Thus we can again apply Lemma 2 of [11], obtaining

$$\|C - \hat{C}\| \leq \|C - \hat{C}\|_{HS} \leq \frac{2\kappa^2 \log\frac{2}{\tau}}{n} + \sqrt{\frac{2\kappa^4 \log\frac{2}{\tau}}{n}} \leq \frac{4\kappa^2 \log\frac{2}{\tau}}{\sqrt{n}}, \tag{110}$$

with probability $1 - \tau$. Thus, by performing an intersection bound, we have

$$A_1 \leq \frac{4Q\kappa \log\frac{2}{\tau}}{\sqrt{\lambda n}} \left(1 + \sqrt{\frac{4\kappa^2 \log\frac{2}{\tau}}{\lambda\sqrt{n}}}\right). \tag{111}$$

with probability $1 - 2\tau$. The term A_2 can be controlled as follows

$$\begin{aligned} A_2 &= \|Z^* S(\hat{C} + \lambda)^{-1} S^* - Z^* S(C + \lambda)^{-1} S^*\|_{HS} && (112)\\ &= \|Z^* S((\hat{C} + \lambda)^{-1} - (C + \lambda)^{-1}) S^*\|_{HS} && (113)\\ &= \|Z^* S(C + \lambda)^{-1}(C - \hat{C})(\hat{C} + \lambda)^{-1} S^*\|_{HS} && (114)\\ &\leq \|Z^* S(C + \lambda)^{-1}\|_{HS} \|C - \hat{C}\| \|(\hat{C} + \lambda)^{-1} S^*\| && (115)\\ &= k^{-1}\mathcal{B}(\lambda)\|C - \hat{C}\| \|(\hat{C} + \lambda)^{-1} S^*\| && (116) \end{aligned}$$

where we have used the fact that for two invertible operators A and B we have $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$. Now, by controlling $\|C - \hat{C}\|,\ \|(\hat{C} + \lambda)^{-1} S^*\|$ as for A_1 and performing an intersection bound, we have

$$A_2 \leq \frac{4\mathcal{B}(\lambda)\kappa \log\frac{2}{\tau}}{\sqrt{\lambda n}} \left(1 + \sqrt{\frac{4\kappa^2 \log\frac{2}{\tau}}{\lambda\sqrt{n}}}\right) \tag{117}$$

with probability $1 - 2\tau$. Finally the term A_3 is equal to

$$\begin{aligned} A_3 &= \|Z^*(S(C + \lambda)^{-1} S^* - I)\|_{HS} = \|Z^*(L(L + \lambda)^{-1} - I)\|_{HS} && (118)\\ &= \|Z^*(L(L + \lambda)^{-1} - (L + \lambda)(L + \lambda)^{-1})\|_{HS} = \lambda\|Z^*(L + \lambda)^{-1}\|_{HS} = \mathcal{A}(\lambda) && (119) \end{aligned}$$

where I denotes the identity operator. Thus, by performing an intersection bound of the events for A_1 and A_2, we have

$$\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) \leq 8\kappa Q\|V\| \frac{Q + \mathcal{B}(\lambda)}{\sqrt{\lambda n}} \left(1 + \sqrt{\frac{4\kappa^2}{\lambda\sqrt{n}}}\right) \log^2\frac{2}{\tau} + 2Q\|V\|\mathcal{A}(\lambda). \tag{120}$$

with probability $1 - 4\tau$. Since $\delta = 4\tau$ we obtain the desired bound. □

Now we are ready to give the universal consistency result.

Theorem 4 (Universal Consistency)**.** *Let $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfy Asm. 1, $\mathcal{X}$ and $\mathcal{Y}$ be compact sets and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous universal reproducing kernel[3]. For any $n \in \mathbb{N}$ and any distribution ρ on $\mathcal{X} \times \mathcal{Y}$ let $\hat{f}_n : \mathcal{X} \to \mathcal{Y}$ be obtained by Alg. 1 with $\{(x_i, y_i)\}_{i=1}^n$ training points independently sampled from ρ and $\lambda_n = n^{-1/4}$. Then,*

$$\lim_{n \to +\infty} \mathcal{E}(\hat{f}_n) = \mathcal{E}(f^*) \qquad \textit{with probability } 1 \tag{17}$$

Proof. By applying Lemma 18, we have

$$\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) \leq 8\kappa Q\|V\| \frac{Q + \mathcal{B}(\lambda)}{\sqrt{\lambda n}} \left(1 + \sqrt{\frac{4\kappa^2}{\lambda\sqrt{n}}}\right) \log^2\frac{8}{\delta} + 2Q\|V\|\mathcal{A}(\lambda), \tag{121}$$

[3]This is a standard assumption for universal consistency (see [2]). An example of continuous universal kernel is the Gaussian $k(x, x') = \exp(-\gamma\|x - x'\|^2)$, with $\gamma > 0$.

with probability $1-\delta$. Note that, since $C = S^*S$, $\|(C+\lambda)^{-1}\| \leq \lambda^{-1}$ and $\|C(C+\lambda)^{-1}\| \leq 1$, we have

$$\kappa^{-1}\mathcal{B}(\lambda) = \|Z^*S(C+\lambda)^{-1}\|_{HS} \leq \|Z\|_{HS}\|S(C+\lambda)^{-1}\| \tag{122}$$

$$\leq \|Z\|_{HS}\|(C+\lambda)^{-1}S^*S(C+\lambda)^{-1}\|^{1/2} \tag{123}$$

$$\leq \|Z\|_{HS}\|(C+\lambda)^{-1}\|^{1/2}\|C(C+\lambda)^{-1}\|^{1/2} \tag{124}$$

$$\leq \|Z\|_{HS}\lambda^{-1/2}. \tag{125}$$

Therefore

$$\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) \leq 8\kappa Q\|V\| \frac{Q + \frac{\kappa}{\sqrt{\lambda}}\|Z\|_{HS}}{\sqrt{\lambda n}} \left(1 + \sqrt{\frac{4\kappa^2}{\lambda\sqrt{n}}}\right) \log^2 \frac{8}{\delta} + 2Q\|V\|\mathcal{A}(\lambda), \tag{126}$$

Now by choosing $\lambda = \kappa^2 n^{-1/4}$, we have

$$\mathcal{E}(\hat{f}_n) - \mathcal{E}(f^*) \leq 24Q\|V\|(Q + \|Z\|_{HS})n^{-1/4} \log^2 \frac{8}{\delta} + 2Q\|V\|\mathcal{A}(\lambda), \tag{127}$$

with probability $1-\delta$. Now we study $\mathcal{A}(\lambda)$, let $L = \sum_{i\in\mathbb{N}} \sigma_i\, u_i \otimes u_i$ be the eigendecomposition of the compact operator L, with $\sigma_i \geq \sigma_j > 0$ for $1 \leq i \leq j \in \mathbb{N}$ and $u_i \in L^2(\mathcal{X}, \rho_\mathcal{X})$. Now, let $w_i^2 = \langle u_i, ZZ^*u_i\rangle_{L^2(\mathcal{X},\rho_\mathcal{X},\mathbb{R})} = \int \langle g^*(x), u_i\rangle^2_{\mathcal{H}_\mathcal{Y}} d\rho_\mathcal{X}(x)$ for $i \in \mathbb{N}$. We need to prove that $(u_i)_{i\in\mathbb{N}}$ is a basis for $L^2(\mathcal{X}, \rho_\mathcal{X})$. Let $W \subseteq \mathcal{X}$ be the support of $\rho_\mathcal{X}$, note that W is compact and Polish since it is closed and subset of the compact Polish space $\mathcal{X}$. Let $\mathcal{L}$ be the RKHS defined by $\mathcal{L} = \overline{\text{span}\{k(x,\cdot) \mid x \in W\}}$, with the same inner product of $\mathcal{H}_\mathcal{X}$. By the fact that W is a compact polish space and k is continuous, then $\mathcal{L}$ is separable. By the universality of k we have that $\mathcal{L}$ is dense in $C(W)$, and, by Corollary 5 of [12], we have $C(W) = \overline{\text{span}\{u_i \mid i \in \mathbb{N}\}}$. Thus, since $C(W)$ is dense in $L^2(\mathcal{X}, \rho_\mathcal{X})$, we have $(u_i)_{i\in\mathbb{N}}$ is a basis of $L^2(\mathcal{X}, \rho_\mathcal{X})$. Thus $\sum_{i\in\mathbb{N}} w_i^2 = \int \|g^*(x)\|^2_{\mathcal{H}_\mathcal{Y}} d\rho_\mathcal{X}(x) = \|Z\|^2_{HS} < \infty$. Therefore

$$\mathcal{A}(\lambda_n)^2 = \lambda_n^2\|Z^*(L+\lambda_n)^{-1}\|^2_{HS} = \sum_{i\in\mathbb{N}} \frac{\lambda_n^2 w_i^2}{(\sigma_i+\lambda)^2}. \tag{128}$$

Let $t_n = n^{-1/8}$, and $T_n = \{i \in \mathbb{N} \mid \sigma_i \geq t_n\} \subset \mathbb{N}$. For any $n \in \mathbb{N}$ we have

$$\mathcal{A}(\lambda_n) = \sum_{i\in T_n} \frac{\lambda_n^2 w_i^2}{(\sigma_i+\lambda_n)^2} + \sum_{i\in\mathbb{N}\setminus T_n} \frac{\lambda_n^2 w_i^2}{(\sigma_i+\lambda_n)^2} \tag{129}$$

$$\leq \frac{\lambda_n^2}{t_n^2} \sum_{i\in T_n} w_i^2 + \sum_{i\in\mathbb{N}\setminus T_n} w_i^2 \leq \kappa^4\|Z\|^2_{HS}\, n^{-1/4} + \sum_{i\in\mathbb{N}\setminus T_n} w_i^2 \tag{130}$$

since $\lambda_n/t_n = \kappa^2 n^{-1/4}/n^{-1/8} = \kappa^2 n^{-1/8}$. We recall that L is a trace class operator, namely $\text{Tr}(L) = \sum_{i=1}^{+\infty} \sigma_i < +\infty$. Therefore $\sigma_i \to 0$ for $i \to +\infty$, from which we conclude

$$0 \leq \lim_{n\to\infty} \mathcal{A}(\lambda_n) \leq \lim_{n\to\infty} \kappa^4\|Z\|^2_{HS} n^{-1/4} + \sum_{i\in\mathbb{N}\setminus T_n} w_i^2 = 0. \tag{131}$$

Now, for any $n \in \mathbb{N}$, let $\delta_n = n^{-2}$ and E_n be the event associated to the equation

$$\mathcal{E}(\hat{f}_n) - \mathcal{E}(f^*) > 24Q\|V\|(Q + \|Z\|_{HS})n^{-1/4} \log^2(8n^2) + 2Q\|V\|\mathcal{A}(\lambda). \tag{132}$$

By Lemma 18, we know that the probability of E_n is at most δ_n. Since $\sum_{n=1}^{+\infty} \delta_n < +\infty$, we can apply the Borel-Cantelli lemma (Theorem 8.3.4. pag 263 of [1]) on the sequence $(E_n)_{n\in\mathbb{N}}$ and conclude that the statement

$$\lim_{n\to\infty} \mathcal{E}(\hat{f}_n) - \mathcal{E}(f^*) > 0, \tag{133}$$

holds with probability 0. Thus, the converse statement

$$\lim_{n\to\infty} \mathcal{E}(\hat{f}_n) - \mathcal{E}(f^*) = 0. \tag{134}$$

holds with probability 1. □

B.4 Generalization Bounds

Finally, we show that under the further hypothesis that g^* belongs to the RKHS $\mathcal{G} \simeq \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$, we are able to prove generalization bounds for the structured prediction algorithm.

Theorem 5 (Generalization Bound)**.** *Let $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfy Asm. 1, $\mathcal{Y}$ be a compact set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a bounded continuous reproducing kernel. Let $\hat{f}_n$ denote the solution of Alg. 1 with n training points and $\lambda = n^{-1/2}$. If the surrogate risk $\mathcal{R}$ defined in Eq. (12) admits a minimizer $g^* \in \mathcal{G}$, then*

$$\mathcal{E}(\hat{f}_n) - \mathcal{E}(f^*) \leq c\tau^2 \, n^{-\frac{1}{4}} \tag{18}$$

holds with probability $1 - 8e^{-\tau}$ for any $\tau > 0$, with c a constant not depending on n and τ.

Proof. By applying 18, we have

$$\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) \leq 8\kappa Q \|V\| \frac{Q + \mathcal{B}(\lambda)}{\sqrt{\lambda n}} \left(1 + \sqrt{\frac{4\kappa^2}{\lambda\sqrt{n}}}\right) \log^2 \frac{8}{\delta} + 2Q\|V\|\mathcal{A}(\lambda), \tag{135}$$

with probability $1 - \delta$. By assumption, $g^* \in \mathcal{G}$ and therefore there exists a $G \in \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$ such that

$$g^*(x) = G\varphi(x), \quad \forall x \in \mathcal{X}. \tag{136}$$

This implies that $Z^* = GS^*$ since, by definition of Z and S, for any $h \in L^2(\mathcal{X}, \rho_\mathcal{X})$,

$$Z^*h = \int_\mathcal{X} g^*(x)h(x)d\rho_\mathcal{X}(x) = \int_\mathcal{X} G\varphi(x)h(x)d\rho_\mathcal{X}(x) = GS^*h. \tag{137}$$

Thus, since $L = SS^*$ and $\|(L + \lambda)^{-1}L\| \leq 1$ and $\|(L + \lambda)^{-1}\| \leq \lambda^{-1}$ for any $\lambda > 0$, we have

$$\begin{aligned}
\mathcal{A}(\lambda) &= \lambda\|Z^*(L + \lambda)^{-1}\|_{HS} = \lambda\|GS^*(L + \lambda)^{-1}\|_{HS} && (138)\\
&\leq \lambda\|G\|_{HS}\|S^*(L + \lambda)^{-1}\| = \lambda\|G\|_{HS}\|(L + \lambda)^{-1}SS^*(L + \lambda)^{-1}\|^{1/2} && (139)\\
&\leq \lambda\|G\|_{HS}\|(L + \lambda)^{-1}L\|^{1/2}\|(L + \lambda)^{-1}\|^{1/2} && (140)\\
&\leq \lambda^{1/2}\|G\|_{HS}. && (141)
\end{aligned}$$

Moreover, since $C = S^*S$, we have

$$\begin{aligned}
\kappa^{-1}\mathcal{B}(\lambda) &= \|Z^*S(C + \lambda)^{-1}\|_{HS} = \|GS^*S(C + \lambda)^{-1}\|_{HS} && (142)\\
&\leq \|G\|_{HS}\|C(C + \lambda)^{-1}\| && (143)\\
&\leq \|G\|_{HS}. && (144)
\end{aligned}$$

Now, let $\lambda = \kappa^2 n^{-1/4}$, we have

$$\begin{aligned}
\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) &\leq 24Q\|V\|(Q + \kappa\|G\|_{HS})n^{-1/4}\log^2\frac{8}{\delta} + 2Q\|V\|\kappa n^{-1/4} && (145)\\
&\leq 24Q\|V\|(Q + \kappa\|G\|_{HS} + \kappa)n^{-1/4}\log^2\frac{8}{\delta} && (146)\\
&= c\tau^2 n^{-1/4} && (147)
\end{aligned}$$

with probability $1 - 8e^{-\tau}$, where we have set $\delta = 8e^{-\tau}$ and $c = 24Q\|V\|(Q + \kappa\|G\|_{HS} + \kappa)$ to obtain the desired inequality. □

C Examples of Loss Functions

In this section we prove Thm. 19 to show that a wide range of functions $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ useful for structured prediction learning satisfies the loss trick (Asm. 1). In the following we state Thm. 19, then we use it to prove that all the losses considered in Example 1 satisfy Asm. 1. Finally we give two lemmas, necessary to prove Thm. 19 and then conclude with its proof.

Theorem 19. *Let $\mathcal{Y}$ be a set. A function $\triangle : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfy Asm. 1 when at least one of the following conditions hold:*

1. *$\mathcal{Y}$ is a finite set, with discrete topology.*

2. *$\mathcal{Y} = [0,1]^d$ with $d \in \mathbb{N}$, and the mixed partial derivative $L(y,y') = \frac{\partial^{2d} \triangle(y_1,\dots,y_d,y'_1,\dots,y'_d)}{\partial y_1,\dots,\partial y_d,\partial y'_1,\dots,\partial y'_d}$ exists almost everywhere, where $y = (y_i)_{i=1}^d, y' = (y'_i)_{i=1}^d \in \mathcal{Y}$, and satisfies*

$$\int_{\mathcal{Y}\times\mathcal{Y}} |L(y,y')|^{1+\epsilon} dy dy' < \infty, \quad \textit{with} \quad \epsilon > 0. \tag{148}$$

3. *$\mathcal{Y}$ is compact and $\triangle$ is a continuous kernel, or $\triangle$ is a function in the RKHS induced by a kernel K. Here K is a continuous kernel on $\mathcal{Y} \times \mathcal{Y}$, of the form*

$$K((y_1,y_2),(y'_1,y'_2)) = K_0(y_1,y'_1)K_0(y_2,y'_2), \quad \forall y_i, y'_i \in \mathcal{Y}, i = 1,2,$$

with K_0 a bounded and continuous kernel on $\mathcal{Y}$.

4. *$\mathcal{Y}$ is compact and*

$$\mathcal{Y} \subseteq \mathcal{Y}_0, \quad \triangle = \triangle_0|_{\mathcal{Y}},$$

that is the restriction of $\triangle_0 : \mathcal{Y}_0 \times \mathcal{Y}_0 \to \mathbb{R}$ on $\mathcal{Y}$, and $\triangle_0$ satisfies Asm. 1 on $\mathcal{Y}_0$,

5. *$\mathcal{Y}$ is compact and*

$$\triangle(y,y') = f(y) \, \triangle_0 \, (F(y), G(y'))g(y'),$$

with F, G continuous maps from $\mathcal{Y}$ to a set $\mathcal{Z}$ with $\triangle_0 : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ satisfying Asm. 1 and $f, g : \mathcal{Y} \to \mathbb{R}$, bounded and continuous.

6. *$\mathcal{Y}$ compact and*

$$\triangle = f(\triangle_1, \dots, \triangle_p),$$

where $f : [-M,M]^d \to \mathbb{R}$ is an analytic function (e.g. a polynomial), $p \in \mathbb{N}$ and $\triangle_1, \dots, \triangle_p$ satisfy Asm. 1 on $\mathcal{Y}$. Here $M \geq \sup_{1\leq i\leq p} \|V_i\| C_i$ where V_i is the operator associated to the loss $\triangle_i$ and C_i is the value that bounds the norm of the feature map ψ_i associated to $\triangle_i$, with $i \in \{1,\dots,p\}$.

Below we expand Example 1 by proving that the considered losses satisfy Asm. 1. The proofs are typically a direct application of Thm. 19 above.

1. *Any loss with, $\mathcal{Y}$ finite.* This is a direct application of Thm. 19, point 1.
2. *Regression and classification loss functions.* Here $\mathcal{Y}$ is an interval on $\mathbb{R}$ and the listed loss functions satisfies Thm. 19, point 2. For example, let $\mathcal{Y} = [-\pi, \pi]$ the mixed partial derivative of the Hinge loss $\triangle(y,y') = \max(0, 1 - yy')$ is defined almost everywhere as $L(y,y') = -1$ when $yy' < 1$ and $L(y,y') = 0$ otherwise. Note that L satisfies Eq. (148), for any $\lambda > 0$.
3. *Robust loss functions.* Here, again $\mathcal{Y}$ is an interval on $\mathbb{R}$. The listed loss functions are: *Cauchy* $\gamma \log(1 + |y - y'|^2/\gamma)$, *German-McLure* $|y - y'|^2/(1 + |y - y'|^2)$ *"Fair"* $\gamma|y - y'| - \gamma^2 \log(1 + |y - y'|/\gamma)$ or the *"*$\mathbf{L_2 - L_1}$*"* $\sqrt{1 + |y - y'|^2} - 1$. They are differentiable on $\mathbb{R}$, hence satisfy Thm. 19, point 2. The *Absolute value* $|y - y'|$ is Lipschitz and satisfies Thm. 19, point 2, as well.
4. *KDE.* When $\mathcal{Y}$ is a compact set and $\triangle$ is a kernel, the point 3 of Thm. 19 is applicable.
5. *Diffusion Distances on Manifolds.* Let $M \in \mathbb{N}$ and $\mathcal{Y} \subset \mathbb{R}^M$ be a compact Reimannian manifold. The *heat kernel* (at time $t > 0$), $k_t : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ induced by the Laplace-Beltrami operator of $\mathcal{Y}$ is a reproducing kernel [13]. The *squared diffusion distance* is defined in terms of k_t as follows $\triangle(y,y') = 1 - k_t(y,y')$. Then, point 3 of Thm. 19 is applicable.
6. *Distances on Histograms/Probabilities.* Let $M \in \mathbb{N}$. A discrete probability distribution (or a normalized histogram) over M entries can be represented as a $y = (y_i)_{i=1}^M \in \mathcal{Y} \subset [0,1]^M$ the M-simplex, namely $\sum_{i=1}^M y_i = 1$ and $y_i \geq 0 \; \forall i = 1,\dots,M$. A typical measure of distance on $\mathcal{Y}$ is the squared *Hellinger (or Bhattacharya)* $\triangle_H(y,y') = (1/\sqrt{2})\|\sqrt{y} - \sqrt{y'}\|_2^2$, with $\sqrt{y} = (\sqrt{y_i})_{i=1}^M$. By Thm. 19, points 4, 6 we have that $\triangle_H$ satisfies Asm. 1.

Indeed, consider the kernel k on $\mathbb{R}$, $k(r,r') = (\sqrt{rr'}+1)^2$ with feature map $\varphi(r) = (r, \sqrt{2r}, 1)^\top \in \mathbb{R}^3$, Then

$$\triangle_0(r,r') = (\sqrt{r}-\sqrt{r'})^2 = r-2\sqrt{rr'}+r' = \varphi(r)^\top V \varphi(r') \quad \text{with} \quad V = \begin{pmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

$\triangle_H$ is obtained by M summations of $\triangle_0$ on $[0,1]$, by Thm. 19, point 6 (indeed $\triangle_H(y,y') = f(\triangle_0(y_1,y_1'),\dots,\triangle_0(y_M,y_M'))$ with $y = (y_i)_{i=1}^M, y' = (y_i')_{i=1}^M \in \mathcal{Y}$ and the function $f : \mathbb{R}^M \to \mathbb{R}$ defined as $f(t_1,\dots,t_M) = \sum_{i=1}^M t_i$, which is analytic on $\mathbb{R}^M$), and then restriction on $\mathcal{Y}$ Thm. 19, point 4. A similar reasoning holds when the loss function is the χ^2 distance on histograms. Indeed the function $(r-r')^2/(r+r')$ satisfies point 2 on $\mathcal{Y} = [0,1]$, then point 6 and 4 are applied.

Hellinger Also the standard Hellinger loss $\triangle_H(y,y') = \sum_{i=1}^M |\sqrt{y_i}-\sqrt{y_i'}|$ satisfies Asm. 1. This can be shown by using Thm. 19 (2) for the (weak) derivative of $\triangle(y,y)$.

To prove Thm. 19 we need the following two Lemmas.

Lemma 20 (multiple Fourier series)**.** *Let $D = [-\pi,\pi]^d$, $(\widehat{f}_h)_{h\in\mathbb{Z}^d} \in \mathbb{C}$ and $f : D \to \mathbb{C}$ with $d \in \mathbb{N}$ defined as*

$$f(y) = \sum_{h\in\mathbb{Z}^d} \widehat{f}_h e^{ih^\top y}, \quad \forall y \in D, \quad \textit{with} \quad \sum_{h\in\mathbb{Z}^d} |\widehat{f}_h| \le B,$$

for a $B < \infty$ and $i = \sqrt{-1}$. Then the function f is continuous and

$$\sup_{y\in D} |f(y)| \le B.$$

Proof. For the continuity, see [14] (pag. 129 and Example 2). For the boundedness we have

$$\sup_{y\in D} |f(y)| \le \sup_{y\in D} \sum_{h\in\mathbb{Z}^d} |\widehat{f}_h||e^{ih^\top y}| \le \sum_{h\in\mathbb{Z}^d} |\widehat{f}_h| \le B.$$

□

Lemma 21. *Let $\mathcal{Y} = [-\pi,\pi]^d$ with $d \in \mathbb{N}$, and $\triangle : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ defined by*

$$\triangle(y,z) = \sum_{h,k\in\mathbb{Z}^d} \widehat{\triangle}_{h,k} e_h(y) e_k(z), \quad \forall y,z \in \mathcal{Y},$$

with $e_h(y) = e^{ih^\top y}$ for any $y \in \mathcal{Y}$, $i = \sqrt{-1}$ and $\widehat{\triangle}_{h,k} \in \mathbb{C}$ for any $h,k \in \mathbb{Z}^d$. The loss $\triangle$ satisfies Asm. 1 when

$$\sum_{h,k\in\mathbb{Z}^d} |\widehat{\triangle}_{h,k}| < \infty.$$

Proof. Note that by applying Lemma 20 with $D = \mathcal{Y}\times\mathcal{Y}$, the function $\triangle$ is bounded continuous. Now we introduce the following sequences

$$\alpha_h = \sum_{k\in\mathbb{Z}^d} |\widehat{\triangle}_{h,k}|, \quad f_h(z) = \frac{1}{\alpha_h} \sum_{h\in\mathbb{Z}^d} \widehat{\triangle}_{hk} e_k(z) \quad \forall h \in \mathbb{Z}^d, z \in \mathcal{Y}.$$

Note that $(\alpha_h)_{h\in\mathbb{Z}^d} \in \ell_1$ and that f_h bounded by 1 and is continuous by Lemma 20 for any $h \in \mathbb{Z}^d$. Therefore

$$\triangle(y,z) = \sum_{h,k\in\mathbb{Z}^d} \widehat{\triangle}_{h,k} e_h(y) e_h(z) = \sum_{h\in\mathbb{Z}^d} \alpha_h e_h(y) f_h(z).$$

Now we define two feature maps $\psi_1, \psi_2 : \mathcal{Y} \to \mathcal{H}_0$, with $\mathcal{H}_0 = \ell_2(\mathbb{Z}^d)$, as

$$\psi_1(y) = (\sqrt{\alpha_h} e_h(y))_{h\in\mathbb{Z}^d}, \quad \psi_2(y) = (\sqrt{\alpha_h} f_h(y))_{h\in\mathbb{Z}^d}, \quad \forall y \in \mathcal{Y}.$$

Now we prove that the two feature maps are continuous. Define $k_1(y,z) = \langle \psi_1(y), \psi_1(z) \rangle_{\mathcal{H}_0}$ and $k_2(y,z) = \langle \psi_2(y), \psi_2(z) \rangle_{\mathcal{H}_0}$ for all $y, z \in \mathcal{Y}$. We have

$$k_1(y,z) = \sum_{h \in \mathbb{Z}^d} \alpha_h \overline{e_h(y)} e_h(z), \tag{149}$$

$$k_2(y,z) = \sum_{h \in \mathbb{Z}^d} \alpha_h \overline{f_h(y)} f_h(z) = \sum_{k,l \in \mathbb{Z}^d} \beta_{k,l} \overline{e_k(y)} e_l(z) \tag{150}$$

with $\beta_{k,l} = \sum_{h \in \mathbb{Z}^d} \frac{\overline{\widehat{\triangle}_{h,k}} \widehat{\triangle}_{h,l}}{\alpha_h}$, for $k, l \in \mathbb{Z}^d$, therefore k_1, k_2 are bounded and continuous by Lemma 20 with $D = \mathcal{Y} \times \mathcal{Y}$, since $\sum_{h \in \mathbb{Z}^d} \alpha_h < \infty$ and $\sum_{k,l \in \mathbb{Z}^d} |\beta_{k,l}| < \infty$. Note that ψ_1 and ψ_2 are bounded, since k_1 and k_2 are. Moreover for any $y, z \in \mathcal{Y}$, we have

$$\|\psi_1(y) - \psi_1(z)\|^2 = \langle \psi_1(z), \psi_1(z) \rangle_{\mathcal{H}_0} + \langle \psi_1(y), \psi_1(y) \rangle_{\mathcal{H}_0} - 2\langle \psi_1(z), \psi_1(y) \rangle_{\mathcal{H}_0} \tag{151}$$

$$= k_1(z,z) + k_1(y,y) - 2k_1(z,y) \leq |k_1(z,z) - k_1(z,y)| + |k_1(z,y) - k_1(y,y)|, \tag{152}$$

and the same holds for ψ_2 with respect to k_2. Thus the continuity of ψ_1 is entailed by the continuity of k_1 and the same for ψ_2 with respect to k_2. Now we define $\mathcal{H}_\mathcal{Y} = \mathcal{H}_0 \oplus \mathcal{H}_0$ and $\psi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ and $V : \mathcal{H}_\mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ as

$$\psi(y) = (\psi_1(y), \psi_2(y)), \ \ \forall y \in \mathcal{Y} \quad \text{and} \quad V = \begin{pmatrix} 0 & I \\ 0 & 0 \end{pmatrix},$$

where $I : \mathcal{H}_0 \to \mathcal{H}_0$ is the identity operator. Note that ψ is bounded continuous, V is bounded and

$$\langle \psi(y), V\psi(z) \rangle_{\mathcal{H}_\mathcal{Y}} = \langle \psi_1(y), \psi_2(z) \rangle_{\mathcal{H}_0} = \sum_{h \in \mathbb{Z}^d} \alpha_h e_h(y) f_h(z) = \triangle(y,z).$$

□

We can now prove Thm. 19.

Proof. (Thm. 19)

1 Let $N = \{1, \ldots, |\mathcal{Y}|\}$ and $q : \mathcal{Y} \to N$ be a one-to-one function. Let $\mathcal{H}_\mathcal{Y} = \mathbb{R}^{|\mathcal{Y}|}$, $\psi(y) : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ defined by $\psi(y) = e_{q(y)}$ for any $y \in \mathcal{Y}$ with $(e_i)_{i=1}^{|\mathcal{Y}|}$ the canonical basis for $\mathcal{H}_\mathcal{Y}$, finally $V \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{Y}|}$ with $V_{i,j} = \triangle(q^{-1}(i), q^{-1}(j))$ for any $i, j \in N$. Then $\triangle$ satisfies Asm. 1, with ψ and V.

2 By Lemma 21, we know that any loss $\triangle$ whose Fourier expansion is absolutely summable, satisfies Asm. 1. The required conditions in points 2 are sufficient (see Theorem 6' pag. 291 of [15]).

3 Let $\mathcal{Y}$ be a compact space. For the first case let $\triangle$ be a bounded and continuous reproducing kernel on $\mathcal{Y}$ and let $\mathcal{H}_\mathcal{Y}$ the associated RKHS, then there exist a bounded and continuous map $\psi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ such that $\triangle(y, y') = \langle \psi(y), \psi(y') \rangle$ for any $y, y' \in \mathcal{Y}$, which satisfies Asm. 1 with $V = I$ the identity on $\mathcal{H}_\mathcal{Y}$. For the second case, let K defined in terms of K_0 as in equation. Let $\mathcal{H}_0$ be the RKHS induced by K_0 and ψ the associated feature map, then, by definition, the $RKHS$ induced by K will be $\mathcal{H} = \mathcal{H}_0 \otimes \mathcal{H}_0$. Since $\triangle$ belongs to $\mathcal{H}$, then there exists a $v \in \mathcal{H}$ such that $\triangle(y, y') = \langle v, \psi(y) \otimes \psi(y') \rangle_{\mathcal{H}_0 \otimes \mathcal{H}_0}$. Now note that $\mathcal{H} = \mathcal{H}_0 \otimes \mathcal{H}_0$ is isomorphic to $B_2(\mathcal{H}_0, \mathcal{H}_0)$, that is the linear space of Hilbert-Schmidt operators from $\mathcal{H}_0$ to $\mathcal{H}_0$, thus, there exist an operator $V \in B_2(\mathcal{H}_0, \mathcal{H}_0)$ such that

$$\triangle(y, y') = \langle v, \psi(y) \otimes \psi(y') \rangle_{\mathcal{H}_0 \otimes \mathcal{H}_0} = \langle \psi(y), V\psi(y') \rangle_{\mathcal{H}_0}, \quad \forall y, y' \in \mathcal{Y}.$$

Finally note that ψ is continuous and bounded, since it is K_0, and V is bounded since it is Hilbert-Schmidt. Thus $\triangle$ satisfies Asm. 1.

4 Since $\triangle_0$ satisfies Asm. 1 we have that there exists a kernel on $\mathcal{Y}_0$ such that Equation 10 holds. Note that the restriction of a kernel on a subset of its domain is again a kernel. Thus, let $\psi = \psi_0|_\mathcal{Y}$, we have that $\triangle|_\mathcal{Y}$ satisfies Equation 10 with ψ and the same bounded operator V as $\triangle$.

5 Let $\triangle_0 : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ satisfy Asm. 1, then $\triangle_0(z, z') = \langle \psi_0(z), V_0 \psi_0(z') \rangle$ for all $z, z' \in \mathcal{Z}$ with $\psi_0 : \mathcal{Y} \to \mathcal{H}_0$ bounded and continuous and $\mathcal{H}_0$ a separable Hilbert space. Now we define two feature maps $\psi_1(y) = f(y)\psi_0(F(y))$ and $\psi_2(y) = g(y)\psi_0(G(y))$. Note that both $\psi_1, \psi_2 : \mathcal{Y} \to \mathcal{H}_0$ are bounded and continuous. We define $\mathcal{H}_\mathcal{Y} = \mathcal{H}_0 \oplus \mathcal{H}_0$, $\psi : \mathcal{Y} \to \mathcal{H}_\mathcal{Y}$ as $\psi(y) = (\psi_1(y), \psi_2(y'))$ for any $y \in \mathcal{Y}$, and $V = \begin{pmatrix} 0 & V_0 \\ 0 & 0 \end{pmatrix}$. Note that now

$$\langle \psi(y), V\psi(y') \rangle_{\mathcal{H}_\mathcal{Y}} = (\psi_1(y)\ \psi_2(y)) \begin{pmatrix} 0 & V_0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \psi_1(y) \\ \psi_2(y) \end{pmatrix} = \langle \psi_1(y), V_0 \psi_2(y') \rangle_{\mathcal{H}_\mathcal{Y}} \quad (153)$$

$$= f(y)\langle \psi(F(y)), V_0 \psi(G(y')) \rangle_{\mathcal{H}_\mathcal{Y}} g(y') = f(y) \triangle_0 (F(y), G(y'))g(y'). \quad (154)$$

6 Let $\mathcal{Y}$ be compact and $\triangle_i$ satisfies Asm. 1 with $\mathcal{H}_i$ the associated RKHS, with continuous feature maps $\psi_i : \mathcal{Y} \to \mathcal{H}_i$ bounded by C_i and with a bounded operator V_i, for $i \in \{1, \dots, p\}$. Since an analytic function is the limit of a series of polynomials, first of all we prove that a finite polynomial in the losses satisfies Asm. 1, then we take the limit. First of all, note that $\alpha \triangle_1 + \beta \triangle_2$, satisfies Asm. 1, for any $\alpha_1, \alpha_2 \in \mathbb{R}$. Indeed we define $\mathcal{H}_\mathcal{Y} = \mathcal{H}_1 \oplus \mathcal{H}_2$, and $\psi(y) = (\sqrt{|\alpha|\|V_1\|}\psi_1(y),\ \sqrt{|\beta|\|V_2\|}\psi_2(y))$ for any $y \in \mathcal{Y}$, so that

$$\alpha_1 \triangle_1(y, y') + \alpha_2 \triangle_2(y, y') = \langle \psi(y), V\psi(y') \rangle_{\mathcal{H}_\mathcal{Y}}, \quad \text{with} \quad V = \begin{pmatrix} \frac{\text{sign}(\alpha_1)}{\|V_1\|} V_1 & 0 \\ 0 & \frac{\text{sign}(\alpha_2)}{\|V_2\|} V_2 \end{pmatrix},$$

for any $y, y' \in \mathcal{Y}$, where ψ is continuous, $\sup_{y \in \mathcal{Y}} \|\psi\|_{\mathcal{H}_\mathcal{Y}} \leq |\alpha_1|\|V_1\|C_1 + |\alpha_2|\|V_2\|C_2$ and $\|V\| \leq 1$. In a similar way we have that $\triangle_1 \triangle_2$ satisfies Asm. 1, indeed, we define $\mathcal{H}_\mathcal{Y} = \mathcal{H}_1 \otimes \mathcal{H}_2$ and ψ to be $\psi(y) = \psi_1(y) \otimes \psi_2(y)$ for any $y \in \mathcal{Y}$, thus

$$\triangle_1(y, y')\ \triangle_2\ (y, y') = \langle \psi(y), V\psi(y) \rangle_{\mathcal{H}_\mathcal{Y}}, \quad \text{with} \quad V = V_1 \otimes V_2,$$

for any $y, y' \in \mathcal{Y}$, where ψ is continuous, $\sup_{y \in \mathcal{Y}} \|\psi\|_{\mathcal{H}_\mathcal{Y}} \leq C_1 C_2$ and $\|V\| \leq \|V_1\|\|V_2\|$. Given a polynomial $P(\triangle)$, with $\triangle = (\triangle_1, \dots, \triangle_p)$ we write it as

$$P(\triangle) = \sum_{t \in \mathbb{N}^p} \alpha_t \triangle^t, \quad \text{with} \quad \triangle^t = \prod_{i=1}^p \triangle_i^{t_i}, \ \ \forall t \in \mathbb{N}^p.$$

where the α's are the coefficents of the polynomial and such that only a finite number of them are non-zero. By applying the construction of the product on each monomial and of the sum on the resulting monomials, we have that $P(\triangle)$ is a loss satisfying Asm. 1 for a continuous ψ and a V such that

$$\sup_{y \in \mathcal{Y}} \|\psi\|_{\mathcal{H}_\mathcal{Y}} \leq \bar{P}(\|V_1\|C_1, \dots, \|V_1\|C_1)$$

and $\|V\| \leq 1$, where $\bar{P}(\triangle) = \sum_{t \in \mathbb{N}^p} |\alpha_t| \triangle^t$ and $\mathcal{H}_\mathcal{Y} = \oplus_{t \in \mathbb{N}^p} \ \otimes_{i=1}^p \mathcal{H}_i^{t_i}$. Note that $\mathcal{H}_\mathcal{Y}$ is again separable. Let now consider

$$f(\triangle) = \sum_{t \in \mathbb{N}^p} \alpha_t \triangle^t, \quad \bar{f}(\triangle) = \sum_{t \in \mathbb{N}^p} |\alpha_t|\ \triangle^t\ .$$

Assume that $\bar{f}(\|V_1\|C_1, \dots, \|V_p\|C_p) < \infty$. Then by repeating the construction for the polynomials, we produce a bounded ψ and a bounded V such that

$$f(\triangle_1(y, y'), \dots, \triangle_p(y, y')) = \langle \psi(y), V\psi(y') \rangle_{\mathcal{H}_\mathcal{Y}}, \quad \forall\ y, y' \in \mathcal{Y},$$

in particular $\mathcal{H}_\mathcal{Y}$ is the same for the polynomial case and $\psi = \oplus_{t \in \mathbb{N}^p} \sqrt{|\alpha_t| v_t} \otimes_{i=1}^p \psi_i^{\otimes t_i}$, with $v_t = \prod_{i=1}^p \|V_i\|^{t_i}$, for any $t \in \mathbb{N}^p$. Now we prove that ψ is continuous on $\mathcal{Y}$. Let ψ_q be the feature map defined for the polynomial $\bar{P}_q(\triangle) = \sum_{t \in \mathbb{N}^p,\ \|t\| \leq q} |\alpha_t| \triangle^t$. We have that

$$\sup_{y \in \mathcal{Y}} \|\psi(y) - \psi_q(y)\|^2_{\mathcal{H}_\mathcal{Y}} = \sup_{y \in \mathcal{Y}} \left\| \oplus_{t \in \mathbb{N}^p, \|t\| > q} \sqrt{|\alpha_t| v^t} \otimes_{i=1}^p \psi_i^{\otimes t_i} \right\|^2_{\mathcal{H}_\mathcal{Y}} \quad (155)$$

$$\leq \sum_{t \in \mathbb{N}^d, \|t\| > q} |\alpha_t| v_t \sup_{y \in \mathcal{Y}} \| \otimes_{i=1}^p \psi_i^{\otimes t_i} \|^2_{\mathcal{H}_\mathcal{Y}} \leq \sum_{t \in \mathbb{N}^d, \|t\| > q} |\alpha_t| v_t c_t, \quad (156)$$

with $c_t = \prod_{i=1}^{p} C_i^{t_i}$ for any $t \in \mathbb{N}^p$. No note that

$$\sum_{t \in \mathbb{N}^d} |\alpha_t| v_t c_t = \bar{f}(\|V_1\| C_1, \ldots, \|V_p\| C_p) < \infty,$$

thus $\lim_{q \to \infty} \sum_{t \in \mathbb{N}^d, \|t\| > q} |\alpha_t| v_t c_t = 0$. Therefore

$$\lim_{q \to \infty} \sup_{y \in \mathcal{Y}} \|\psi(y) - \psi_q(y)\|_{\mathcal{H}_\mathcal{Y}}^2 \leq \lim_{q \to \infty} \sum_{t \in \mathbb{N}^p,\ \|t\| > q} |\alpha_t| v_t c_t = 0. \tag{157}$$

Now, since ψ_q is a sequence of continuous bounded functions, and the sequence converges uniformly to ψ, then ψ is continuous bounded. So $f(\triangle)$ is a loss function satisfying Asm. 1, with a continuous ψ and an operator V such that

$$\sup_{y \in \mathcal{Y}} \|\psi\|_{\mathcal{H}_\mathcal{Y}} \leq \bar{f}(\|V_1\| C_1, \ldots, \|V_1\| C_1),$$

and $\|V\| \leq 1$.

□

References

[1] Richard M Dudley. *Real analysis and probability*, volume 74. Cambridge University Press, 2002.

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

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

[4] Charles Castaing and Michel Valadier. *Convex analysis and measurable multifunctions*, volume 580. Springer, 2006.

[5] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

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

[7] Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4(04):377–408, 2006.

[8] Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

[9] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

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

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

[12] Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *The Journal of Machine Learning Research*, 7:2651–2667, 2006.

[13] Richard Schoen and Shing-Tung Yau. *Lectures on differential geometry*, volume 2. International press Boston, 1994.

[14] J-P Kahane. *Fourier series and wavelets*. Routledge, 1995.

[15] Ferenc Móricz and Antal Veres. On the absolute convergence of multiple fourier series. *Acta Mathematica Hungarica*, 117(3):275–292, 2007.