Sketch In, Sketch Out: Accelerating both Learning and Inference for Structured Prediction with Kernels

Tamim El Ahmad, Luc Brogat-Motte, Pierre Laforgue, Florence d'Alché-Buc

Sketch In, Sketch Out: Accelerating both Learning and Inference for Structured Prediction with Kernels

Tamim El Ahmad
Télécom Paris

Luc Brogat-Motte
CentraleSupélec

Pierre Laforgue
University of Milan

Florence d'Alché-Buc
Télécom Paris

Abstract

Leveraging the kernel trick in both the input and output spaces, surrogate kernel methods are a flexible and theoretically grounded solution to structured output prediction. If they provide state-of-the-art performance on complex data sets of moderate size (e.g., in chemoinformatics), these approaches however fail to scale. We propose to equip surrogate kernel methods with sketching-based approximations, applied to both the input and output feature maps. We prove excess risk bounds on the original structured prediction problem, showing how to attain close-to-optimal rates with a reduced sketch size that depends on the eigendecay of the input/output covariance operators. From a computational perspective, we show that the two approximations have distinct but complementary impacts: sketching the input kernel mostly reduces training time, while sketching the output kernel decreases the inference time. Empirically, our approach is shown to scale, achieving state-of-the-art performance on benchmark data sets where non-sketched methods are intractable.

1 INTRODUCTION

Ubiquitous in real-world applications, structured objects have attracted a great deal of attention in machine learning (Bakir et al., 2007; Gärtner, 2008; Nowozin and Lampert, 2011; Deshwal et al., 2019). Depending on their role, i.e., either as input or output variables, they raise distinct challenges. Classification and regression from structured *inputs* generally rely on a continuous representation learned by a deep neural network (Defferrard et al., 2016), or implicitly defined through a dedicated kernel (Collins and Duffy, 2001; Borgwardt et al., 2020). In contrast, structured *output* prediction calls for a more involved approach, since the discrete nature of the outputs impacts the definition of the loss function (Nowak et al., 2019; Ciliberto et al., 2020; Cabannes et al., 2021), and therefore the learning problem itself.

To handle this problem, several methods have been developed to relax the combinatorial problems that appear both at training and inference. Energy-based approaches convert structured prediction into learning a scalar score function (Tsochantaridis et al., 2005; LeCun et al., 2007; Belanger and McCallum, 2016; Deshwal et al., 2019). End-to-end learning typically exploits a differentiable model, together with a differentiable loss, to run gradient descent (Long et al., 2015; Niculae et al., 2018; Berthet et al., 2020). Surrogate methods (Ciliberto et al., 2020) solve a regression problem in a Hilbert space where outputs have been implicitly embedded, shortcutting the inference during learning.

Rare are the methods that enjoy both scalability at learning/inference steps and statistical guarantees (Osokin et al., 2017; Cabannes et al., 2021). In this work, we focus on surrogate approaches and their implementation as kernel methods, i.e., the input output kernel regression framework (Cortes et al., 2005; Brouard et al., 2016b). Recent works Ciliberto et al. (2016, 2020) have shown that they enjoy consistency, their excess risk being governed by that of the surrogate regression. Moreover, they are well appropriate to make prediction from one structured modality to another, since kernels can be leveraged in both the input and output spaces. Overall, they offer a general, theoretically grounded, and simple-to-implement solution to structured prediction, providing state-of-the-art results in applications such as molecule identification (Schymanski et al., 2017).

However, contrary to deep neural networks, they do not scale neither in memory nor in time without further approximation. The aim of this paper is to equip these methods with kernel approximations to obtain a drastic

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

complexity reduction while maintaining their statistical properties. Several works have highlighted the power of kernel approximations, from Random Fourier Features (Rahimi and Recht, 2007; Brault et al., 2016; Rudi and Rosasco, 2017; Li et al., 2021), to general low-rank approaches (Bach, 2013; Meanti et al., 2020).

In this work we focus on sketching (Mahoney et al., 2011; Woodruff, 2014), a general dimension reduction method based on linear random projections. Applied to kernel approximation, sketching has been widely studied through Nyström's sub-sampling approximation (Williams and Seeger, 2001; Alaoui and Mahoney, 2015; Rudi et al., 2015), and further explored using Gaussian or Randomized Orthogonal Systems (Yang et al., 2017; Lacotte and Pilanci, 2020). Interpreted as a way to provide data-dependent random features (Williams and Seeger, 2001; Yang et al., 2012; Kpotufe and Sriperumbudur, 2020), this approach has allowed to scale up kernel PCA (Sterge and Sriperumbudur, 2022), kernel mean embedding (Chatalic et al., 2022a,b) or independence tests (Kalinke and Szabó, 2023) while enjoying statistical guarantees. However, sketching has been limited so far to scalar kernel machines. No current approach covers both sides of the coin, i.e., applying approximations to both the input and output kernels. Motivated by surrogate structured prediction, we close this gap and make the following contributions:

- We apply sketching to the vector-valued kernel regression problem solved in structured prediction, both on inputs and outputs, which accelerates respectively learning and inference.
- We derive excess risk bounds controlled by the properties of the sketched projection operators.
- We prove that sub-Gaussian sketches provide close-to-optimal rates with small sketch sizes.
- We empirically show that our algorithms maintain good accuracy on moderate-size datasets while enabling kernel surrogate methods on large datasets where the standard approach is simply intractable.

Notations. We introduce now generic notations for the input (output) space and kernel, detailed in Appendix A. If $\mathcal{Z}$ denotes a generic Polish space, $\mathrm{k}_{\mathcal{Z}}$ is a positive definite kernel over $\mathcal{Z}$ and $\psi_{\mathcal{Z}}(z) \coloneqq \mathrm{k}_{\mathcal{Z}}(\cdot, z)$ is the canonical feature map of $\mathrm{k}_{\mathcal{Z}}$. $\mathcal{H}_{\mathcal{Z}}$ denotes the Reproducing Kernel Hilbert Space (RKHS) associated to $\mathrm{k}_{\mathcal{Z}}$. $\mathrm{S}_{\mathrm{Z}} : f \in \mathcal{H}_{\mathcal{Z}} \mapsto (1/\sqrt{n})(f(z_1), \dots, f(z_n))^\top$ is the sampling operator over $\mathcal{H}_{\mathcal{Z}}$ (Smale and Zhou, 2007).

2 BACKGROUND

We now recall the structured prediction setting based on a kernel-induced loss, and a state-of-the-art surrogate approach to solve it. We also provide reminders about sketching as a way to scale-up kernel methods.

Structured prediction with surrogate kernel methods. Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ a structured output space. In general, $\mathcal{Y}$ is finite and extremely large. Define a positive definite kernel $\mathrm{k}_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, that measures how close two objects from $\mathcal{Y}$ are. We consider the loss function induced by $\mathrm{k}_{\mathcal{Y}}$, defined as $\ell : (y, y') \to \| \psi_{\mathcal{Y}}(y) - \psi_{\mathcal{Y}}(y') \|^2_{\mathcal{H}_{\mathcal{Y}}}$. Note that it can be computed using the kernel trick. Given an unknown joint probability distribution ρ defined on $\mathcal{X} \times \mathcal{Y}$, the goal of structured prediction is to approximate

$$f^* = \operatorname*{arg\,min}_{f : \mathcal{X} \to \mathcal{Y}} \mathcal{R}(f)\,, \tag{1}$$

where $\mathcal{R}(f) = \mathbb{E}_{(x,y)\sim\rho}\left[\| \psi_{\mathcal{Y}}(y) - \psi_{\mathcal{Y}}(f(x)) \|^2_{\mathcal{H}_{\mathcal{Y}}} \right]$, using only an i.i.d. sample $\{(x_1, y_1), \dots, (x_n, y_n)\}$ drawn from ρ. Estimating directly f^* is not tractable, such that many works (Cortes et al., 2005; Geurts et al., 2006; Brouard et al., 2011; Ciliberto et al., 2016) have proposed instead the following two-step approach:

1. Surrogate Regression: Find an estimator $\hat{h}$ of the surrogate target $h^* : x \mapsto \mathbb{E}_y[\psi_{\mathcal{Y}}(y)|x]$ such that

$$h^* = \operatorname*{arg\,min}_{h} \; \mathbb{E}_{(x,y)} \left[\| h(x) - \psi_{\mathcal{Y}}(y) \|^2_{\mathcal{H}_{\mathcal{Y}}} \right] .$$

2. Pre-image: Define $\hat{f}$ by decoding $\hat{h}$, i.e.,

$$\hat{f}(x) = d(\hat{h}(x)) \coloneqq \operatorname*{arg\,min}_{y \in \mathcal{Y}} \left\| \hat{h}(x) - \psi_{\mathcal{Y}}(y) \right\|^2_{\mathcal{H}_{\mathcal{Y}}} .$$

The surrogate regression in Step 1 is much easier to handle than the initial structured prediction problem: it avoids learning f through the composition with the implicit feature map $\psi_{\mathcal{Y}}$, and relegates the difficulty of handling structured objects to Step 2, i.e. at inference. In addition, vector-valued regression into infinite-dimensional spaces is a well-studied problem, that can be solved by using the kernel trick in the output space. This two-step approach belongs to the general framework of SELF (Ciliberto et al., 2016) and ILE (Ciliberto et al., 2020) and enjoys valuable theoretical guarantees. It is Fisher consistent, i.e., h^* yields f^* after decoding, and the excess risk of $\hat{f}$ is controlled by that of $\hat{h}$.

Input Output ridge Kernel Regression. A common choice to tackle in practice the surrogate regression problem consists in solving a *kernel ridge regression problem*, leveraging kernels in both input and output spaces. The hypothesis space is chosen as a vector-valued Reproducing Kernel Hilbert Space (vv-RKHS) (Senkene and Tempel'man, 1973; Micchelli and Pontil, 2005; Carmeli et al., 2006, 2010). In the same way

that RKHS are based on positive symmetric definite kernels, vv-RKHS are based on Operator-Valued Kernels (OVK). In our setting, we define an OVK $\mathcal{K}$, as a mapping $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{H}_\mathcal{Y})$, where $\mathcal{L}(\mathcal{H}_\mathcal{Y})$ is the set of bounded linear operators on $\mathcal{H}_\mathcal{Y}$, and that satisfies the properties recalled in Appendix B. An OVK $\mathcal{K}$ is uniquely associated with a vv-RKHS $\mathcal{H}$, i.e. a Hilbert space of functions from $\mathcal{X}$ to $\mathcal{H}_\mathcal{Y}$ that enjoys the reproducing kernel property (see Appendix B).

In what follows, we opt for the identity decomposable OVK $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{H}_\mathcal{Y})$, defined as: $\mathcal{K}(x, x') = \mathrm{k}_\mathcal{X}(x, x') I_{\mathcal{H}_\mathcal{Y}}$, where $\mathrm{k}_\mathcal{X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a p.d. scalar-valued kernel on $\mathcal{X}$. In *Input Output Kernel Ridge Regression* (IOKR for short, Brouard et al. 2011; Kadri et al. 2013; Brouard et al. 2016b; Ciliberto et al. 2020, also introduced as Kernel Dependency Estimation by Weston et al. (2003)), the estimator of the surrogate regression is obtained by solving the following Ridge regression problem within $\mathcal{H}$, given a regularisation penalty $\lambda > 0$,

$$\hat{h} = \operatorname*{arg\,min}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \|\psi_\mathcal{Y}(y_i) - h(x_i)\|^2_{\mathcal{H}_\mathcal{Y}} + \lambda \|h\|^2_\mathcal{H} . \quad (2)$$

Interestingly, the unique solution to the above problem can be expressed in different ways. From one hand, we can derive from the representer theorem in vv-RKHSs (Micchelli and Pontil, 2005) the following expression:

$$\hat{h}(x) = \sum_{i=1}^{n} \hat{\alpha}_i(x)\, \psi_\mathcal{Y}(y_i), \quad (3)$$

with $\hat{\alpha}(x) = (\mathrm{K_X} + n\lambda I_n)^{-1} \mathrm{k_X^x} := \widehat{\Omega}\, \mathrm{k_X^x}$, where $\mathrm{K_X} = (\mathrm{k}_\mathcal{X}(x_i, x_j))_{i,j=1}^n$ and $\mathrm{k_X^x} = \big(\mathrm{k}_\mathcal{X}(x, x_1), \ldots, \mathrm{k}_\mathcal{X}(x, x_n)\big)$. On the other hand, using an operator view one obtains

$$\hat{h}(x) = \widehat{H}\, \psi_\mathcal{X}(x)\,, \quad (4)$$

where $\widehat{H} = \mathrm{S_Y}^\# \mathrm{S_X} \left(\widehat{\mathrm{C}}_\mathrm{X} + \lambda I\right)^{-1}$. The latter expression can be seen as a re-writing of the first (Ciliberto et al., 2016), echoing the KDE equations with finite-dimensional feature maps (Cortes et al., 2005). It can also be related to the conditional kernel empirical mean embedding (Grünewälder et al., 2012).

The final estimator $\hat{f}$ is computed using the expression in (3), in order to benefit from the kernel trick:

$$\hat{f}(x) = \operatorname*{arg\,min}_{y \in \mathcal{Y}} \ \mathrm{k}_\mathcal{Y}(y, y) - 2\mathrm{k_X^x}^T \widehat{\Omega} \mathrm{k_Y^y}\,, \quad (5)$$

where $\mathrm{k_Y^y} = \left(\mathrm{k}_\mathcal{Y}(y, y_1), \ldots, \mathrm{k}_\mathcal{Y}(y, y_n)\right)^\top$. The training phase thus involves the inversion of a $n \times n$ matrix, whose cost without any approximation is $\mathcal{O}(n^3)$. Besides, it implies storing n^2 values in memory, which induces a heavy space complexity as well. In practice, decoding is performed by searching in a candidate set $\mathcal{Y}_c \subseteq \mathcal{Y}$ of size n_c. Hence, performing predictions on a test set X_te of size n_te mainly implies computing

$$\underbrace{\mathrm{K_X}^{\mathrm{te,tr}}}_{n_\mathrm{te} \times n} \underbrace{\widehat{\Omega}}_{n \times n} \underbrace{\mathrm{K_Y}^{\mathrm{tr},c}}_{n \times n_c}, \quad (6)$$

where $\mathrm{K_X}^{\mathrm{te,tr}} = (\mathrm{k}_\mathcal{X}(x_i^\mathrm{te}, x_j))_{1 \le i \le n_\mathrm{te}, 1 \le j \le n} \in \mathbb{R}^{n_\mathrm{te} \times n}$, and $\mathrm{K_Y}^{\mathrm{tr},c} = \big(\mathrm{k}_\mathcal{Y}(y_i, y_j^c)\big)_{1 \le i \le n, 1 \le j \le n_c} \in \mathbb{R}^{n \times n_c}$. The complexity of the decoding part is $\mathcal{O}(n_\mathrm{te} n n_c)$, considering $n_\mathrm{te} < n \le n_c$. IOKR thus suffers from both heavy time and space computational costs. To cope with this limitation, we develop a general sketching approach that applies to both input and output feature spaces, accelerating both training and decoding.

Sketching for kernel methods. Applied to kernel methods to reduce their dependency in n, sketching can be seen as linear projections induced by a random matrix R (the sketching matrix) drawn from a probability distribution over $\mathbb{R}^{m \times n}$, where $m \ll n$. Classic examples include Nyström's approximation, where each row of R is randomly drawn from the rows of the identity matrix I_n, and Gaussian sketches, where all entries of R are i.i.d. Gaussian random variables. Nyström's approximation acts as a random training data subsampler, but it can be interpreted in many ways. In Drineas et al. (2005); Bach (2013), it is shown to generate a low-rank approximation of the Gram matrix, while in Williams and Seeger (2001); Yang et al. (2012), it is seen as a way to construct data-dependent finite-dimensional random features. In Rudi et al. (2015), instead, it is presented as a projection onto a small subspace of the RKHS. For other sketching schemes such as Gaussian or Randomized Orthogonal Systems, most of the works adopt an optimization viewpoint, where a variable substitution is operated after the application of a Representer theorem (Yang et al., 2017; Lacotte and Pilanci, 2020). An interesting view provided in Kpotufe and Sriperumbudur (2020) explores the construction of random features based on Gaussian sketching. All these works are however limited to sketching the *input* kernel, in scalar regression problems. In this work: (1) we generalize input sketching to vector-valued problems, (2) we sketch the outputs, which is critical to scale-up surrogate methods with kernelized outputs.

3 SKETCHED INPUT SKETCHED OUTPUT KERNEL REGRESSION

The goal of this section is to construct a low-rank estimator of $\hat{h}$ by using sketching on both the input and output kernels. Note that sketching the feature maps is not desirable here: if we replace the output features $\psi_\mathcal{Y}(y_i) \in \mathcal{H}_\mathcal{Y}$ with some sketch-dependent approximations $\tilde{\psi}_\mathcal{Y}(y_i) \in \mathbb{R}^m$ we become unable to compare the

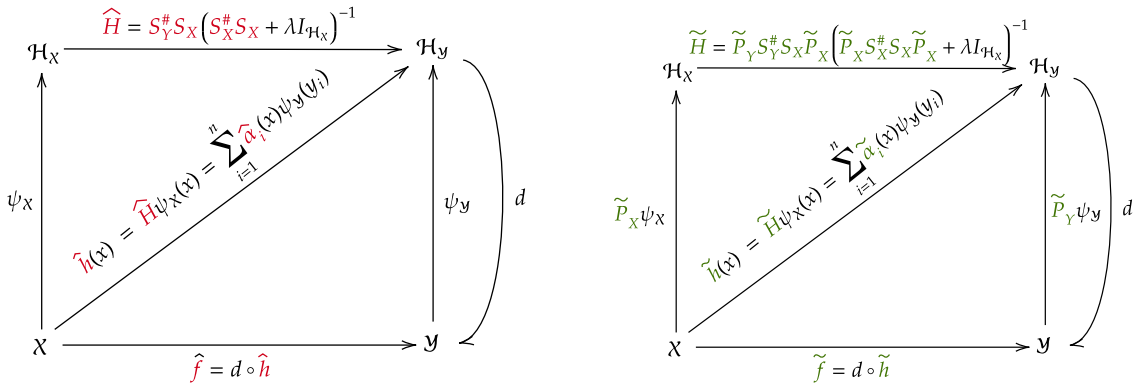

Figure 1: IOKR (left) and SISOKR (right) in the KDE setting. Note that SISOKR consists in IOKR when kernels $\mathrm{k}_{\mathcal{Z}}$ are replaced with their projected versions $\tilde{k}_{\mathcal{Z}}(\cdot,\cdot) = \langle \psi_{\mathcal{Z}}(\cdot), \widetilde{\mathrm{P}}_{\mathrm{Z}}\, \psi_{\mathcal{Z}}(\cdot) \rangle_{\mathcal{H}_{\mathcal{Z}}}$. However, this new output kernel changes the pre-image problem, and consequently the estimator $\tilde{f}$. In the paper, we modify $\widetilde{H}$ (and not the kernels) in order to use the comparison inequality from Ciliberto et al. (2020), see the proof of Corollary 1.

resulting $\tilde{h}$ to the target h^*. Indeed, $\tilde{h}$ is an approximation of $x \mapsto \mathbb{E}_y[\tilde{\psi}_{\mathcal{Y}}(y)|x]$, which is a biased version of h^* due to the sketch realization. Instead, as we show below, seeing sketching as orthogonal projections provides a natural way to solve our problem. Ultimately, this gives rise to an estimator $\tilde{f}$ for structured prediction which is versatile, easy-to-implement, theoretically-based and scalable to large data sets.

Low-rank estimator. Given two orthogonal projection operators $\widetilde{\mathrm{P}}_{\mathrm{X}}$ and $\widetilde{\mathrm{P}}_{\mathrm{Y}}$, we start from (4) and replace the sampling operators on both sides, S_{X} and S_{Y}, by their projected counterparts, $\mathrm{S}_{\mathrm{X}}\, \widetilde{\mathrm{P}}_{\mathrm{X}}$ and $\mathrm{S}_{\mathrm{Y}}\, \widetilde{\mathrm{P}}_{\mathrm{Y}}$, so as to encode dimension reduction. The proposed low-rank estimator is expressed as follows:

$$\tilde{h}(x) = \widetilde{\mathrm{P}}_{\mathrm{Y}}\, \mathrm{S}_{\mathrm{Y}}{}^{\#}\, \mathrm{S}_{\mathrm{X}}\, \widetilde{\mathrm{P}}_{\mathrm{X}} \left(\widetilde{\mathrm{P}}_{\mathrm{X}}\, \widehat{\mathrm{C}}_{\mathrm{X}}\, \widetilde{\mathrm{P}}_{\mathrm{X}} + \lambda I_{\mathcal{H}_{\mathcal{X}}} \right)^{-1} \psi_{\mathcal{X}}(x)\,.$$

We now show how to design the projection operators using sketching and then derive the novel expression of the low-rank estimator in terms of a weighted combination of the training outputs: $\tilde{h}(x) = \sum_{i=1}^{n} \tilde{\alpha}_i\, \psi_{\mathcal{Y}}(y_i)$, yielding a reduced computational cost. IOKR and SISOKR approaches are illustrated on Figure 1.

Sketching. In this work, we chose to leverage sketching to obtain random projectors within the input and output feature spaces. Indeed, sketching consists of approximating a feature map $\psi_{\mathcal{Z}} : \mathcal{Z} \to \mathcal{H}_{\mathcal{Z}}$ by projecting it thanks to a random projection operator $\widetilde{\mathrm{P}}_{\mathrm{Z}}$ defined as follows. Given a random matrix $\mathrm{R}_{\mathcal{Z}} \in \mathbb{R}^{\mathrm{m}_{\mathcal{Z}} \times n}$, n data $(z_i)_{i=1}^n \in \mathcal{Z}$ and $\mathrm{m}_{\mathcal{Z}} \ll n$, the linear subspace defining $\widetilde{\mathrm{P}}_{\mathrm{Z}}$ is constructed as the linear subspace generated by the span of the following $\mathrm{m}_{\mathcal{Z}}$ random vectors

$$\sum_{j=1}^{n} (\mathrm{R}_{\mathcal{Z}})_{ij}\, \psi_{\mathcal{Z}}(z_j) \in \mathcal{H}_{\mathcal{Z}}, \quad i = 1, \ldots, \mathrm{m}_{\mathcal{Z}}\,.$$

One can show (Proposition 2 in Appendix C) that the corresponding orthogonal projector writes

$$\widetilde{\mathrm{P}}_{\mathrm{Z}} = (\mathrm{R}_{\mathcal{Z}}\, \mathrm{S}_{\mathrm{Z}})^{\#} \left(\mathrm{R}_{\mathcal{Z}}\, \mathrm{S}_{\mathrm{Z}} (\mathrm{R}_{\mathcal{Z}}\, \mathrm{S}_{\mathrm{Z}})^{\#} \right)^{\dagger} \mathrm{R}_{\mathcal{Z}}\, \mathrm{S}_{\mathrm{Z}}\,. \quad (7)$$

Sketched Input Sketched Output Kernel Regression (SISOKR). The SISOKR estimator is the low-rank estimator $\tilde{h}$, where both $\widetilde{\mathrm{P}}_{\mathrm{X}}$ and $\widetilde{\mathrm{P}}_{\mathrm{Y}}$ have been chosen as (7), for some random sketches $\mathrm{R}_{\mathcal{X}}$ and $\mathrm{R}_{\mathcal{Y}}$. It also admits the following expression based on a linear combination of the $\psi_{\mathcal{Y}}(y_i)$. The proof of the following proposition is given in Appendix C.

Proposition 1 (Expression of SISOKR). *$\forall\, x \in \mathcal{X}$,*

$$\tilde{h}\,(x) = \sum_{i=1}^{n} \tilde{\alpha}_i\,(x)\, \psi_{\mathcal{Y}}\,(y_i)\,,$$

where $\tilde{\alpha}\,(x) = \mathrm{R}_{\mathcal{Y}}{}^{\top}\, \widetilde{\Omega}\, \mathrm{R}_{\mathcal{X}}\, \mathrm{k}_{\mathrm{X}}^{\mathrm{x}}$ *and*

$$\widetilde{\Omega} = \widetilde{\mathrm{K}}_{\mathrm{Y}}{}^{\dagger}\, \mathrm{R}_{\mathcal{Y}}\, \mathrm{K}_{\mathrm{Y}}\, \mathrm{K}_{\mathrm{X}}\, \mathrm{R}_{\mathcal{X}}{}^{\top} (\mathrm{R}_{\mathcal{X}}\, \mathrm{K}_{\mathrm{X}}{}^{2}\, \mathrm{R}_{\mathcal{X}}{}^{\top} + n\lambda\, \widetilde{\mathrm{K}}_{\mathrm{X}})^{\dagger}\,,$$

with $\widetilde{\mathrm{K}}_{\mathrm{X}} = \mathrm{R}_{\mathcal{X}}\, \mathrm{K}_{\mathrm{X}}\, \mathrm{R}_{\mathcal{X}}{}^{\top}$ *and* $\widetilde{\mathrm{K}}_{\mathrm{Y}} = \mathrm{R}_{\mathcal{Y}}\, \mathrm{K}_{\mathrm{Y}}\, \mathrm{R}_{\mathcal{Y}}{}^{\top}$.

Note that the matrix quantity that we recover above, $\mathrm{K}_{\mathrm{X}}\, \mathrm{R}_{\mathcal{X}}{}^{\top} \left(\mathrm{R}_{\mathcal{X}}\, \mathrm{K}_{\mathrm{X}}{}^{2}\, \mathrm{R}_{\mathcal{X}}{}^{\top} + n\lambda\, \mathrm{R}_{\mathcal{X}}\, \mathrm{K}_{\mathrm{X}}\, \mathrm{R}_{\mathcal{X}}{}^{\top} \right)^{\dagger} \mathrm{R}_{\mathcal{X}}\, \mathrm{k}_{\mathrm{X}}^{\mathrm{x}}$, is typical to sketched kernel Ridge regression (Rudi et al., 2015; Yang et al., 2017). It allows to reduce the size of the matrix to invert, which is now an $\mathrm{m}_{\mathcal{X}} \times \mathrm{m}_{\mathcal{X}}$ matrix. This is the main reason for the reduction of the learning step's complexity and is due to the input sketching. Nonetheless, we still need to perform matrix multiplication $\mathrm{R}_{\mathcal{X}}\, \mathrm{K}_{\mathrm{X}}$, whose efficiency depends on the sketch used). Note that output sketching also requires additional operations, but the overall cost of computing $\tilde{\alpha}$ remains negligible compared to $\mathcal{O}(n^3)$,

Table 1: Time and space complexities at training and inference for the IOKR and SISOKR algorithms with sub-sampling, p-sparsified ($p \in (0,1]$) or Gaussian sketching, for a test set of size n_{te} and a candidate set of size n_c, such that $n_{te} \leq \mathrm{m}_\mathcal{X}, \mathrm{m}_\mathcal{Y} < n \leq n_c$. For the sake of simplicity, we omit the $\mathcal{O}(\cdot)$ in the following.

Method	Training Time	Training Space	Inference Time	Inference Space
IOKR	n^3	n^2	$n_{te} n n_c$	$n n_c$
SISOKR (sub-sampling)	$\max(\mathrm{m}_\mathcal{X}, \mathrm{m}_\mathcal{Y}) n$	$\max(\mathrm{m}_\mathcal{X}, \mathrm{m}_\mathcal{Y}) n$	$n_{te}\, \mathrm{m}_\mathcal{Y}\, n_c$	$\mathrm{m}_\mathcal{Y}\, n_c$
SISOKR (p-sparsified)	$\max(\mathrm{m}_\mathcal{X}, \mathrm{m}_\mathcal{Y})^2 pn$	$\max(\mathrm{m}_\mathcal{X}, \mathrm{m}_\mathcal{Y}) pn$	$\max(n_{te}, n\, \mathrm{m}_\mathcal{Y}\, p)\, \mathrm{m}_\mathcal{Y}\, n_c$	$np\, \mathrm{m}_\mathcal{Y}\, n_c$
SISOKR (Gaussian)	$\max(\mathrm{m}_\mathcal{X}, \mathrm{m}_\mathcal{Y}) n^2$	n^2	$n\, \mathrm{m}_\mathcal{Y}\, n_c$	$n n_c$

see "training time" column in Table 1. As an example, with input/output Gaussian sketching which is the less efficient one, the time complexity is of order $\max(\mathrm{m}_\mathcal{X}, \mathrm{m}_\mathcal{Y}) n^2$, where $\mathrm{m}_\mathcal{X}, \mathrm{m}_\mathcal{Y} \ll n$. We obtain the corresponding structured prediction estimator $\tilde{f}$ by decoding $\tilde{h}$, i.e., by replacing $\widehat{\Omega}$ by $\widetilde{\Omega}$ in (5). In fact, the main quantity we have to compute for prediction is now

$$\underbrace{\mathrm{K_X}^{\mathrm{te,tr}}\, \mathrm{R_\mathcal{X}}^\top}_{n_{\mathrm{te}} \times \mathrm{m}_\mathcal{X}} \underbrace{\widetilde{\Omega}}_{\mathrm{m}_\mathcal{X} \times \mathrm{m}_\mathcal{Y}} \underbrace{\mathrm{R_\mathcal{Y}}\, \mathrm{K_Y}^{\mathrm{tr},c}}_{\mathrm{m}_\mathcal{Y} \times n_c}. \tag{8}$$

The time complexity of this operation is $\mathcal{O}(n_{\mathrm{te}} m_\mathcal{Y} n_c)$ if $n_{\mathrm{te}} \leq \mathrm{m}_\mathcal{X}, \mathrm{m}_\mathcal{Y} < n \leq n_c$, which is a significant complexity reduction (the dependence in n vanishes), governed by the output sketch size $\mathrm{m}_\mathcal{Y}$, see Table 1 for more details.

4 THEORETICAL ANALYSIS

In this section, we present a statistical analysis of the proposed estimators $\tilde{h}$ and $\tilde{f}$. After introducing the assumptions on the learning task, we upper bound the excess-risk of the sketched kernel ridge estimator, highlighting the approximation errors due to sketching. We then provide bounds for these approximation error terms. Finally, we study under which setting the proposed estimators $\tilde{h}$ and $\tilde{f}$ obtain substantial computational gains, while still benefiting from a close-to-optimal learning rates. We consider the following set of common assumptions in the kernel literature (Bauer et al., 2007; Steinwart et al., 2009; Rudi et al., 2015; Pillaud-Vivien et al., 2018; Fischer and Steinwart, 2020; Ciliberto et al., 2020; Brogat-Motte et al., 2022).

Assumption 1 (Attainability). *We assume that $h^* \in \mathcal{H}$, i.e., that there is a linear operator $H : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$, with $\|H\|_{\mathrm{HS}} < +\infty$, s.t. $h^*(x) = H\, \psi_\mathcal{X}(x)$, $\forall\, x \in \mathcal{X}$.*

This is a standard assumption in the context of least-squares regression (Caponnetto and De Vito, 2007), making the target h^* belong to the hypothesis space. Note that relaxing this assumption is possible, although it would add a bias term that still requires some knowledge about h^* to be bounded. For instance, if h^* is supposed to be square-integrable, one usually chooses a RKHS associated with a universal operator-valued kernel, which is dense in the space of the square-integrable functions (Carmeli et al., 2010, Section 4). We now describe a set of generic assumptions that have to be satisfied by both input and output kernels $\mathrm{k}_\mathcal{X}$ and $\mathrm{k}_\mathcal{Y}$.

Assumption 2 (Bounded kernel). *There exists $\kappa_\mathcal{Z} > 0$ such that $\mathrm{k}_\mathcal{Z}(z,z) \leq {\kappa_\mathcal{Z}}^2$, $\forall\, z \in \mathcal{Z}$. We note $\kappa_\mathcal{X}, \kappa_\mathcal{Y} > 0$ for the input and output kernels $\mathrm{k}_\mathcal{X}$ and $\mathrm{k}_\mathcal{Y}$ respectively.*

Assumption 3 (Capacity condition). *There exists $\gamma_\mathcal{Z} \in [0,1]$ such that $\mathrm{Q}_\mathcal{Z} := \mathrm{Tr}({\mathrm{C}_\mathcal{Z}}^{\gamma_\mathcal{Z}}) < +\infty$.*

Note that Assumption 3 is always verified for $\gamma_\mathcal{Z} = 1$, as $\mathrm{Tr}(\mathrm{C}_\mathcal{Z}) = \mathbb{E}[\|\psi_\mathcal{Z}(z)\|^2_{\mathcal{H}_\mathcal{Z}}] < +\infty$ from Assumption 2, and that the smaller $\gamma_\mathcal{Z}$ the faster the eigendecay of $\mathrm{C}_\mathcal{Z}$, with $\gamma_\mathcal{Z} = 0$ when $\mathrm{C}_\mathcal{Z}$ is of finite rank. More generally, this assumption is for instance verified for a Sobolev kernel and a marginal distribution whose density is upper-bounded (Ciliberto et al., 2020, Assumption 2).

Assumption 4 (Embedding property). *There exist $\mathrm{b}_\mathcal{Z} > 0$ and $\mu_\mathcal{Z} \in [0,1]$ such that $\psi_\mathcal{Z}(z) \otimes \psi_\mathcal{Z}(z) \preceq \mathrm{b}_\mathcal{Z}\, {\mathrm{C}_\mathcal{Z}}^{1-\mu_\mathcal{Z}}$ almost surely.*

Note that Assumption 4 is always verified for $\mu_\mathcal{Z} = 1$, as $\psi_\mathcal{Z}(z) \otimes \psi_\mathcal{Z}(z) \preceq {\kappa_\mathcal{Z}}^2\, I_{\mathcal{H}_\mathcal{Z}}$ by Assumption 2, and that the smaller $\mu_\mathcal{Z}$, the stronger the assumption, with $\mu_\mathcal{Z} = 0$ when $\mathrm{C}_\mathcal{Z}$ is of finite. It allows to control the regularity of the functions in $\mathcal{H}_\mathcal{Z}$ with respect to the L^∞-norm, as it implies $\|h\|_{L_\infty} \leq {\mathrm{b}_\mathcal{Z}}^{1/2}\, \|h\|^\mu_{\mathcal{H}_\mathcal{Z}}\, \mathbb{E}[h(z)^2]^{(1-\mu)/2}$ (Pillaud-Vivien et al., 2018). For instance, an absolutely continuous distribution whose density is lower-bounded almost everywhere and a Matérn kernel verifies Assumption 4 (Pillaud-Vivien et al., 2018, Example 2).

SISOKR Excess-Risk. We can now provide a bound on the excess-risk of SISOKR.

Theorem 1 (SISOKR excess-risk bound). *Let $\delta \in (0,1]$, $n \in \mathbb{N}$ such that $\lambda = n^{-1/(1+\gamma_\mathcal{X})} \geq \frac{9\, {\kappa_\mathcal{X}}^2}{n} \log(\frac{n}{\delta})$. Under Assumptions 1 to 4, with probability $1-\delta$ we have*

$$\mathbb{E}_x \left[\|\tilde{h}(x) - h^*(x)\|^2_{\mathcal{H}_\mathcal{Y}} \right]^{\frac{1}{2}} \leq S(n,\delta) + c_2 A^{\psi_\mathcal{X}}_{\rho_\mathcal{X}}(\widetilde{P}_X) + A^{\psi_\mathcal{Y}}_{\rho_\mathcal{Y}}(\widetilde{P}_Y), \tag{9}$$

where $S(n,\delta) = c_1 \log(4/\delta)\, n^{-\frac{1}{2(1+\gamma_{\mathcal{X}})}}$ *and*

$$A^{\psi_{\mathcal{Z}}}_{\rho_{\mathcal{Z}}}(\widetilde{P}_Z) = \mathbb{E}_z \left[\|(\widetilde{\mathrm{P}}_{\mathrm{Z}} - I_{\mathcal{H}_{\mathcal{Z}}})\, \psi_{\mathcal{Z}}(z)\|^2_{\mathcal{H}_{\mathcal{Z}}} \right]^{\frac{1}{2}},$$

with $c_1, c_2 > 0$ *constants independent of* n *and* δ.

Proof sketch. The proof relies on a decomposition of the operator $\widetilde{H}$ such that $\tilde{h}(x) = \widetilde{H}\psi_{\mathcal{X}}(x)$, see (44). The first term in (9) corresponds to the non-sketched kernel Ridge regression error, and the second term to the input sketching error. The latter extends both the results of Ciliberto et al. (2020) to sketched estimators, and that of Rudi et al. (2015) to the vector vector-valued case. The third term, i.e., the output sketching error is specific to our framework and derives from the expression of h^* and Jensen's inequality. □

The learning rate of the first term, i.e., the non-sketched kernel Ridge regression error, has been shown to be optimal under our set of assumptions in a minimax sense (Caponnetto and De Vito, 2007). The second and the third terms are approximation errors due to the sketching of the input and the output kernels, respectively. In particular, they write as *reconstruction errors* (Blanchard et al., 2007) associated to the random projection $\widetilde{\mathrm{P}}_{\mathrm{X}}$ and $\widetilde{\mathrm{P}}_{\mathrm{Y}}$ of the feature maps $\psi_{\mathcal{X}}$ and $\psi_{\mathcal{Y}}$ through the input and output marginal distributions.

Sketching Reconstruction Error. In Theorem 2, we give bounds on the sketching reconstruction error for the family of sub-Gaussian sketches, enlarging the scope of sketching distributions whose reconstruction error's bound is known —it was previously limited to uniform and approximate leverage scores sub-sampling sketches (Rudi et al., 2015). More generally, note that are admissible in our theoretical framework all sketching distributions for which concentration bounds on the induced empirical covariance operators can be derived, since quantity $A^{\psi_{\mathcal{Z}}}_{\rho_{\mathcal{Z}}}(\widetilde{P}_Z)$ is then easily controlled. We now recall the definition of sub-Gaussian sketches, and show how to bound their reconstruction error.

Definition 1. *A sub-Gaussian sketch* $\mathrm{R}_{\mathcal{Z}} \in \mathbb{R}^{\mathrm{m}_{\mathcal{Z}} \times n}$ *is composed of i.i.d. entries such that* $\mathbb{E}\left[\mathrm{R}_{\mathcal{Z}_{\mathrm{ij}}}\right] = 0$, $\mathbb{E}\left[{\mathrm{R}_{\mathcal{Z}_{\mathrm{ij}}}}^2\right] = 1/\mathrm{m}_{\mathcal{Z}}$ *and* $\mathrm{R}_{\mathcal{Z}_{\mathrm{ij}}}$ *is* $\frac{{\nu_{\mathcal{Z}}}^2}{\mathrm{m}_{\mathcal{Z}}}$*-sub-Gaussian, for all* $1 \le i \le \mathrm{m}_{\mathcal{Z}}$ *and* $1 \le j \le n$, *where* $\nu_{\mathcal{Z}} \ge 1$.

Recall that a standard normal r.v. is 1-sub-Gaussian. Moreover, by Hoeffding's lemma, any r.v. taking values in a bounded interval $[a,b]$ is $(b-a)^2/4$-sub-Gaussian. Hence, any sketch matrix composed of i.i.d. Gaussian or bounded r.v. is a sub-Gaussian sketch. Finally, note that p-sparsified sketches (El Ahmad et al., 2023) are sub-Gaussian with ${\nu_{\mathcal{Z}}}^2 = 1/p$, with $p \in]0,1]$.

Theorem 2 (sub-Gaussian sketching reconstruction error)**.** *For* $\delta \in (0, 1/e]$, $n \in \mathbb{N}$ *sufficiently large such that* $\frac{9}{n}\log(n/\delta) \le n^{-\frac{1}{1+\gamma_{\mathcal{Z}}}} \le \|\,\mathrm{C}_{\mathcal{Z}}\,\|_{\mathrm{op}}/2$, *then if*

$$\mathrm{m}_{\mathcal{Z}} \ge c_4 \max\left(\nu^2_{\mathcal{Z}}\, n^{\frac{\gamma_{\mathcal{Z}}+\mu_{\mathcal{Z}}}{1+\gamma_{\mathcal{Z}}}}, \nu^4_{\mathcal{Z}} \log(1/\delta)\right), \qquad (10)$$

with probability $1-\delta$ *we have*

$$\mathbb{E}_z \left[\|(\widetilde{\mathrm{P}}_{\mathrm{Z}} - I_{\mathcal{H}_{\mathcal{Z}}})\, \psi_{\mathcal{Z}}(z)\|^2_{\mathcal{H}_{\mathcal{Z}}} \right] \le c_3\, n^{-\frac{1-\gamma_{\mathcal{Z}}}{1+\gamma_{\mathcal{Z}}}}, \qquad (11)$$

where $c_3, c_4 > 0$ *are constants independents of* $n, \mathrm{m}_{\mathcal{Z}}, \delta$.

Proof sketch. The proof essentially consists in bounding the difference between the empirical covariance operator and its sketched counterpart in operator norm, see (89). The latter rewrites as a sum of sub-Gaussian random variables in a separable Hilbert space, and we invoke Koltchinskii and Lounici (2017, Theorem 9). □

Hence, depending on the regularity of the distribution (defined through our set of assumptions), one can obtain a small reconstruction error even with a small sketching size. For instance, if $\mu_{\mathcal{Z}} = \gamma_{\mathcal{Z}} = 1/3$, one obtains a reconstruction error of order $n^{-1/2}$ by using a sketching size of order $n^{1/2} \ll n$. As a limiting case, when $\mu_{\mathcal{Z}} = \gamma_{\mathcal{Z}} = 0$, one obtains a reconstruction error of order n^{-1} when using a constant sketching size.

Remark 1 (Comparison to Nyström's approximation)**.** *Note that the rate in Theorem 2 is the same as that obtained with Nyström's approximation. However, our lower bound on the sketching size is slightly better. Recall that for uniform Nyström it is of order* $\max\left(n^{\frac{\gamma_{\mathcal{Z}}+\mu_{\mathcal{Z}}}{1+\gamma_{\mathcal{Z}}}}, 1\right)\left(\log(n) + \log\left(4\,{\kappa_{\mathcal{Z}}}^2/\delta\right)\right)$.

Remark 2 (Relaxation of Assumption 4)**.** *Assumption 4 allows to derive an upper bound of* $\mathcal{N}^{\infty}_{\mathcal{Z}}(t)$, *with* $t = n^{-\frac{1}{1+\gamma_{\mathcal{Z}}}}$, *that appears in the lower bound of the sketching size* $\mathrm{m}_{\mathcal{Z}}$, *see Lemma 12 in Appendix G and the proof of Theorem 2 in Appendix E. However, we also have that* $\mathcal{N}^{\infty}_{\mathcal{Z}}(t) \le t^{-1}$, *hence, if* $\mu_{\mathcal{Z}} + \gamma_{\mathcal{Z}} \ge 1 + \frac{\log(\mathrm{b}_{\mathcal{Z}}\,\mathrm{Q}_{\mathcal{Z}})(1+\gamma_{\mathcal{Z}})}{\log(n)}$, *we can relax Assumption 4 and rather obtain*

$$\mathrm{m}_{\mathcal{Z}} \ge c_4 \max\left(\nu^2_{\mathcal{Z}}\, n^{\frac{1}{1+\gamma_{\mathcal{Z}}}}, \nu^4_{\mathcal{Z}} \log(1/\delta)\right), \qquad (12)$$

as a lower bound.

Learning rates for SISOKR with sub-Gaussian sketches. For the sake of presentation, we use $\lesssim$ to keep only the dependencies in $n, \delta, \nu, \gamma, \mu$. We note $a \vee b \coloneqq \max(a,b)$.

Corollary 1 (SISOKR learning rates)**.** *Consider the Assumptions of Theorems 1 and 2, that* $\|\,\psi_{\mathcal{Y}}(y)\|_{\mathcal{H}_{\mathcal{Y}}} =$

$\kappa_{\mathcal{Y}}$ for all $y \in \mathcal{Y}$, and $n \in \mathbb{N}$ such that $\frac{9}{n}\log(n/\delta) \leq n^{-\frac{1}{1+\gamma_{\mathcal{Z}}}} \leq \|\mathrm{C}_{\mathcal{Z}}\|_{\mathrm{op}}/2$ for $\mathcal{Z} \in \{\mathcal{X}, \mathcal{Y}\}$. Set

$$\mathrm{m}_{\mathcal{Z}} \gtrsim \max\left(\nu_{\mathcal{Z}}^2\, n^{\frac{\gamma_{\mathcal{Z}}+\mu_{\mathcal{Z}}}{1+\gamma_{\mathcal{Z}}}}, \nu_{\mathcal{Z}}^4 \log(1/\delta)\right) \tag{13}$$

for $\mathcal{Z} \in \{\mathcal{X}, \mathcal{Y}\}$. Then with probability $1-\delta$

$$\mathcal{R}(\tilde{f}) - \mathcal{R}(f^*) \lesssim \log(4/\delta)\, n^{-\frac{1-\gamma_{\mathcal{X}} \vee \gamma_{\mathcal{Y}}}{2(1+\gamma_{\mathcal{X}} \vee \gamma_{\mathcal{Y}})}}. \tag{14}$$

Proof. Using Theorems 1 and 2 to bound $A^{\psi_{\mathcal{X}}}_{\rho_{\mathcal{X}}}(\widetilde{P}_X)$ and $A^{\psi_{\mathcal{Y}}}_{\rho_{\mathcal{Y}}}(\widetilde{P}_Y)$ gives that with probability $1-\delta$ it holds $\mathbb{E}_x\left[\|\tilde{h}(x) - h^*(x)\|^2_{\mathcal{H}_{\mathcal{Y}}}\right]^{\frac{1}{2}} \lesssim \log(4/\delta)\, n^{-\frac{1-\gamma_{\mathcal{X}} \vee \gamma_{\mathcal{Y}}}{2(1+\gamma_{\mathcal{X}} \vee \gamma_{\mathcal{Y}})}}$. We then apply the comparison inequality (Ciliberto et al., 2020) to the loss $\Delta(y, y') = \|\psi_{\mathcal{Y}}(y) - \psi_{\mathcal{Y}}(y')\|^2_{\mathcal{H}_{\mathcal{Y}}}$. □

This corollary shows that under strong enough regularity assumptions, the proposed estimators benefit from a close-to-optimal learning rate, even with small input and output sketching sizes. For instance, if $\mu_{\mathcal{X}} = \mu_{\mathcal{Y}} = \gamma_{\mathcal{X}} = \gamma_{\mathcal{Y}} = 1/3$, one obtains a learning rate of $\mathcal{O}(n^{-1/4})$, instead of the optimal rate of $\mathcal{O}(n^{-3/8})$ under the same assumptions, but only requiring sketching sizes $\mathrm{m}_{\mathcal{X}}, \mathrm{m}_{\mathcal{Y}}$ of order $n^{1/2} \ll n$. As a limiting case, when $\mu_{\mathcal{X}} = \mu_{\mathcal{Y}} = \gamma_{\mathcal{X}} = \gamma_{\mathcal{Y}} = 0$, one attains the optimal $\mathcal{O}(n^{-1/2})$ learning rate using constant sketching sizes.

Remark 3 (Other Sketches)**.** *Although we focused on sub-Gaussian sketches, any sketching distribution admitting concentration bounds for operators on separable Hilbert spaces allows to bound the quantity $A^{\psi_{\mathcal{Z}}}_{\rho_{\mathcal{Z}}}(\widetilde{P}_Z)$ and is then admissible for our theoretical framework. For instance, as showed in Rudi et al. (2015), uniform and approximate leverage scores sub-sampling schemes fit into the presented theory.*

5 EXPERIMENTS

In this section, we present experiments on synthetic and real-world data sets. SIOKR and ISOKR denote the models with sketching leveraged only on the inputs (resp. outputs). Results are averaged over 30 replicates, unless for the metabolite's experiments (5 replicates).

On the choice of the sketching types and its hyper-parameters. We focus on uniform sub-sampling (Rudi et al., 2015) and p-sparsified (p-SR/SG) (El Ahmad et al., 2023) sketches, which are covered by our theory. Sub-sampling is the most efficient approach computationally, but we empirically observe that p-SR/SG sketching is more accurate statistically. For SIOKR/ISOKR, we privilege accuracy and p-SR/SG sketching, as it is already providing substantial training/inference accelerations. Regarding SISOKR, we want the method to be the fastest both in training and inference. However, since output sketching adds training computations, we compensate and use input sub-sampling to remain faster in training than SIOKR. Regarding the input/output sketching sizes $\mathrm{m}_{\mathcal{X}}$ and $\mathrm{m}_{\mathcal{Y}}$, the first way consists of leveraging the theoretical lower bounds derived for $\mathrm{m}_{\mathcal{X}}$ and $\mathrm{m}_{\mathcal{Y}}$, see Equation (10). Indeed, by computing the Singular Value Decomposition of the input/output Gram matrix, one may determine their eigendecay (i.e., $\gamma_{\mathcal{Z}}$, $\mu_{\mathcal{Z}}$, $\nu_{\mathcal{Z}}$) and set $\mathrm{m}_{\mathcal{X}}$ and $\mathrm{m}_{\mathcal{Y}}$ accordingly. However, computing the SVD is very expensive, hence one can rather compute the approximate leverage scores as in Alaoui and Mahoney (2015) for instance. In the following, we instead adopt an empirical routine. Given training and/or inference time budgets (corresponding e.g., to IOKR's training/inference times or the hardware limitations), we start from small $\mathrm{m}_{\mathcal{X}}$ and $\mathrm{m}_{\mathcal{Y}}$, which we progressively increase to maximize accuracy while respecting the budget. For the p-SR/SG sketches, we always set $p = 20/n$.

Synthetic Least Squares Regression. We generate a synthetic data set of least-squares regression, with $n = 10\,000$ training data points, $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, $d = 300$, and use input and output linear kernels, hence $\mathcal{H}_{\mathcal{X}} = \mathcal{H}_{\mathcal{Y}} = \mathbb{R}^d$. We construct covariance matrices $\mathrm{C}_{\mathcal{X}}$ and E by drawing randomly their eigenvectors such that their eigenvalues are $\sigma_k(\mathrm{C}_{\mathcal{X}}) = k^{-3/2}$ and $\sigma_k(E) = 0.2\, k^{-1/10}$. We draw $H_0 \in \mathbb{R}^{d\times d}$ with i.i.d. coefficients from the standard normal distribution and set $H = \mathrm{C}_{\mathcal{X}} H_0$. For $i \leq n$, we generate inputs $x_i \sim \mathcal{N}(0, \mathrm{C}_{\mathcal{X}})$, noise $\epsilon_i \sim \mathcal{N}(0, E)$ and outputs $y_i = Hx_i + \epsilon_i$. We generate validation and test sets of $n_{\mathrm{val}} = n_{\mathrm{te}} = 1000$ points in the same way. Such choices for $\mathrm{C}_{\mathcal{X}}$ (with a polynomial eigenvalue decay), E (with very low eigenvalues and eigenvalue decay), and $H = \mathrm{C}_{\mathcal{X}} H_0$ enforce a high eigenvalue decay for $\mathrm{C}_{\mathcal{Y}}$ (since it will have a similar eigendecay as $\mathrm{C}_{\mathcal{X}}$) while being a favorable setting to deploy sketching, as the true regression function H is low rank. We select the regularisation penalty λ via 1-fold cross-validation. We learn the SISOKR model for different values of $\mathrm{m}_{\mathcal{X}}$ and $\mathrm{m}_{\mathcal{Y}}$ (from 10 to 295) and $(2 \cdot 10^{-3})$-SR input and output sketches. Note that for such a problem where $\mathcal{Y} = \mathcal{H}_{\mathcal{Y}}$, no decoding step is needed for inference. We still perform an artificial pre-image problem to illustrate the computational benefit of sketching during this phase.

Figure 2 (left and center) presents computational training (solid lines) and inference (dotted lines) time (as a percentage of IOKR's training/inference time) w.r.t. $\mathrm{m}_{\mathcal{X}}$ (resp. $\mathrm{m}_{\mathcal{Y}}$) for two values of $\mathrm{m}_{\mathcal{Y}}$ (resp. $\mathrm{m}_{\mathcal{X}}$). First, since $\mathrm{m}_{\mathcal{X}}, \mathrm{m}_{\mathcal{Y}} \leq 295 \ll n = 10\,000$, note that SISOKR's training and inference times are significantly

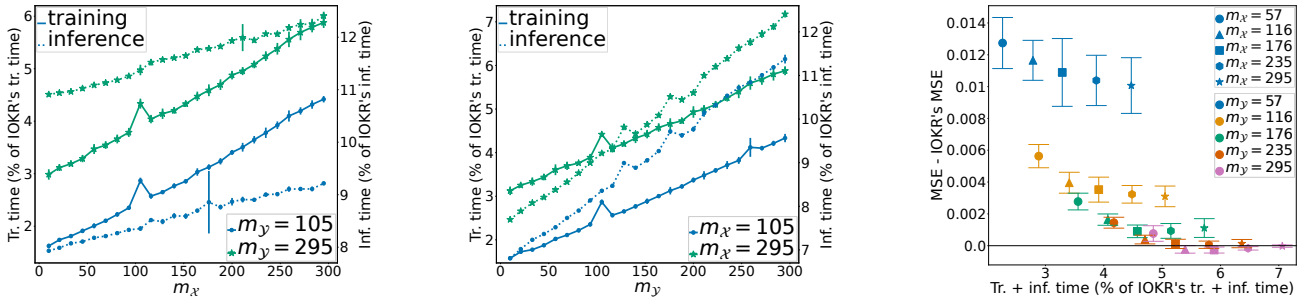

Figure 2: Variation of training and inference time w.r.t. $\mathrm{m}_\mathcal{X}$ and $\mathrm{m}_\mathcal{Y}$ (left and center), and trade-off performance against computational time (right) for SISOKR with $(2 \cdot 10^{-3})$-SR input/output sketches on synthetic data.

Table 2: F_1 scores on tag prediction from text data.

Method	Bibtex	Bookmarks	Mediamill
LR	37.2	30.7	NA
SPEN	42.2	34.4	NA
PRLR	44.2	34.9	NA
DVN	44.7	37.1	NA
SISOKR	44.1 ± 0.07	$\mathbf{39.3} \pm 0.61$	57.26 ± 0.04
ISOKR	44.8 ± 0.01	NA	58.02 ± 0.01
SIOKR	44.7 ± 0.09	39.1 ± 0.04	57.33 ± 0.04
IOKR	**44.9**	NA	**58.17**

smaller than IOKR's (between 2 and 6% of IOKR's training time and 8 and 12% IOKR's inference time). On Figure 2 (left) the slopes of the training time's lines are higher than the inference time's ones, while the opposite happens on Figure 2 (center). This confirms that training complexity is more sensitive to $\mathrm{m}_\mathcal{X}$, while inference complexity is governed by $\mathrm{m}_\mathcal{Y}$. Figure 2 (right) presents the difference with IOKR's test errors, in terms of Mean Squared Error (MSE), for some choices of $\mathrm{m}_\mathcal{X}$ and $\mathrm{m}_\mathcal{Y}$, as a function of the sum of the training and inference times. The MSE decreases as the sketch sizes increase and at a faster rate with respect to $\mathrm{m}_\mathcal{X}$. This might be due to the fact that we directly control the eigendecay of $\mathrm{C}_\mathcal{X}$, whereas $\mathrm{C}_\mathcal{Y} = \mathrm{C}_\mathcal{X}\, H_0\, \mathrm{C}_\mathcal{X}\, H_0^\top\, \mathrm{C}_\mathcal{X} + E$, such that its range is not totally controlled by $\mathrm{C}_\mathcal{X}$. SISOKR obtains better MSE performance than IOKR for $\mathrm{m}_\mathcal{X} \geq 116$ and $\mathrm{m}_\mathcal{Y} = 295$, which is consistent with the results obtained when applying sketching to the input (resp. output) kernel only, see Appendix I.1.

Multi-Label Classification. We compare our models to state-of-the-art multi-label and structured prediction methods, namely IOKR (Brouard et al., 2016b), logistic regression (LR) trained independently for each label (Lin et al., 2014), the multi-label approach Posterior-Regularized Low-Rank (PRLR) (Lin et al., 2014), the energy-based model Structured Prediction Energy Networks (SPEN) (Belanger and McCallum, 2016) and Deep Value Networks (DVN) (Gygli et al., 2017). Results are taken from the cited articles. Data sets Bibtex and Bookmarks are tag recommendation problems, in which the objective is to propose a relevant set of tags (e.g., url, description, journal volume) to users when they add a new Bookmark or Bibtex entry to the social bookmarking system Bibsonomy. The MediaMill Challenge (Snoek et al., 2006) is a multi-label classification problem, where the goal is to detect the presence of semantic concepts in a video. They contain respectively $n = 4880$, $n = 60\,000$ and $n = 30\,993$ training points, see Appendix I.2 for details. We use the train-test splits available at `https://mulan.sourceforge.net/datasets-mlc.html`.

For all multi-label experiments, we use Gaussian input and output kernels with widths σ^2_{in} and σ^2_{out}. We use p-SG input (resp. output) sketches for SIOKR (resp. ISOKR), uniform sub-sampling input sketches and p-SG output sketches for SISOKR. For Bibtex experiments, we choose $\mathrm{m}_\mathcal{X} = 2250$ and $\mathrm{m}_\mathcal{Y} = 200$, for Bookmarks experiments, $\mathrm{m}_\mathcal{X} = 13\,000$ and $\mathrm{m}_\mathcal{Y} = 750$, and for Mediamill experiments, $\mathrm{m}_\mathcal{X} = 8\,000$ and $\mathrm{m}_\mathcal{Y} = 500$. All the training data are used as candidate sets. The performance is measured by example-based F1 score, and hyper-parameters are selected on logarithmic grids by 5-fold cross-validation. The results in Table 2 show that surrogate methods (last four columns) compete with SOTA methods, including deep-learning-based methods such as SPEN or DVN. On Bibtex, sketched models preserve good performance compared to IOKR (which performs best) while being faster to train (SIOKR and SISOKR) and significantly faster for inference (ISOKR and SISOKR), see Table 3. Since the Bookmarks data set is too large, storing the whole n^2-Gram matrix K_X exceeds CPU's space limitations, which highlights the necessity of efficient sketching approximations such that sub-sampling or p-SR/SG sketches for kernel methods. Hence, we can only test SIOKR and SISOKR models on this data set, which outperforms other methods. SISOKR's inference phase is notably faster than SIOKR's (20 seconds vs. 5 minutes). Similarly, on the Mediamill problem, our approximated approaches are shown to be significantly

Table 3: Training/inference times (in seconds).

Method	Bibtex	Bookmarks	Mediamill
SISOKR	$\mathbf{1.41 \pm 0.03}$ / $\mathbf{0.46 \pm 0.01}$	$\mathbf{118 \pm 1.5}$ / $\mathbf{20 \pm 0.2}$	$\mathbf{66 \pm 0.1}$ / $\mathbf{4 \pm 0.01}$
ISOKR	2.51 ± 0.06 / 0.58 ± 0.01	NA	636 ± 3.7 9 ± 0.2
SIOKR	1.99 ± 0.07 / 1.22 ± 0.03	354 ± 2.1 / 297 ± 2.1	199 ± 0.1 / 121 ± 0.02
IOKR	2.54 / 1.18	NA	621 / 204

Table 4: Standard errors for the metabolite identification problem and computation times (in seconds).

Method	kernel loss	Top-1 | 5 | 10 accuracies	training	inference
SPEN	0.537 ± 0.008	25.9% | 54.1% | 64.3%	NA	NA
SISOKR	0.566 ± 0.007	25.1% | 54.2% | 64.7%	4.05 ± 0.05	$\mathbf{1112 \pm 29}$
ISOKR	0.509 ± 0.009	28.0% | 58.9% | 68.9%	6.25 ± 50.31	1133 ± 32
SIOKR	0.492 ± 0.008	29.5% | 61.3% | 70.9%	$\mathbf{1.25 \pm 0.02}$	1179 ± 37
IOKR	$\mathbf{0.486 \pm 0.008}$	**29.6%** | **61.6%** | **71.4%**	3.54 ± 0.15	1191 ± 38

faster to run while suffering a minimal reduction in F_1 score. Note that, with the same sketch matrix $\mathrm{R}_{\mathcal{X}}$, SIOKR's training is faster than SISOKR's as there is no additional computation on Gram matrix K_{Y}. In Table 3, SISOKR is faster to train as it uses a more efficient input sketching (sub-sampling vs. p-SG).

Metabolite Identification. Metabolite identification consists here of predicting small molecules, called metabolites, from their tandem mass spectrum. The metabolite structure is represented as a binary vector of length $d = 7593$, called a fingerprint. Each entry of the fingerprint encodes the presence or absence of a molecular property. IOKR is the SOTA method for this problem (Brouard et al., 2016a). The data set consists of $n = 6974$ training labeled mass spectra, the median size of the candidate sets is 292 and the largest candidate set contains 36 918 fingerprints. This metabolite identification problem thus involves high-dimensional complex outputs, for which the choice of the output kernel is crucial, and a large number of candidates, making the inference step long.

Our experimental protocol is similar to that of Brouard et al. (2016a) (5-CV Outer / 4-CV Inner loops). We use probability product input kernel for mass spectra and Gaussian-Tanimoto output kernel (Ralaivola et al., 2005) – with width σ^2 – for the molecular fingerprints. We select hyper-parameters λ and σ^2 in logarithmic grids based on MSE in $\mathcal{H}_{\mathcal{Y}}$ (hence no decoding is needed during selection). For the sketched models, we use p-SR input (resp. output) sketches for SIOKR (resp. ISOKR), and uniform sub-sampling input sketches and p-SR output sketches for SISOKR, with $\mathrm{m}_{\mathcal{X}} = 1500$, and $\mathrm{m}_{\mathcal{Y}} = 800$.

We compare our sketched models with IOKR and SPEN, see Table 4. Results for SPEN are taken from Brogat-Motte et al. (2022). SIOKR obtains results similar to IOKR while being slightly faster in both the training and inference phases. ISOKR is slightly less accurate, but outperforms (S)IOKR in terms of inference time, while SISOKR has the fastest inference phase and still competes with SPEN statistically. We observe here that it is difficult to reduce significantly the inference time while keeping a good accuracy and to reduce both the training and inference time. This is due to the particular setting of the metabolite data set. Indeed, each molecule is associated with a specific candidate set, so when performing predictions one has to run through each element one by one to pick its candidate set. When performing predictions, one has to compute the matrix multiplication (8), which has a smaller complexity than (6), given that $\mathrm{R}_{\mathcal{Y}} \mathrm{K}_{\mathrm{Y}}{}^{\mathrm{tr},c}$ is already known. However, in the case of metabolite identification, one has to perform it for each test data, which takes most of the inference for both ISOKR and SISOKR models. As an example, for the 1133 (resp. 1112) seconds-long ISOKR's (resp. SISOKR) inference phase, computing $\mathrm{R}_{\mathcal{Y}} \mathrm{K}_{\mathrm{Y}}{}^{\mathrm{tr},c}$ takes 940 (resp. 917) seconds. Since we have access to all candidate sets for each molecule, one could pre-process these data beforehand and perform these matrix multiplications during training, leading to a high training time, but a very small inference time, which could be of interest according to the practitioner's wish. When candidate sets are known and fixed (e.g., in multi-label prediction), sketching the output kernel is of particular interest as no additional operation is needed for each prediction.

6 CONCLUSION

In this paper, we scale-up surrogate methods for structured prediction based on kernel Ridge regression by using random projections for both inputs and outputs. An interesting avenue for future work is the study of non-parametric estimators with kernelized outputs that do not benefit from the Ridge regression closed-form.

Acknowledgments

This work was supported by the Télécom Paris research chair on Data Science and Artificial Intelligence for Digitalized Industry and Services (DSAIDIS) and the French National Research Agency (ANR) through ANR-18-CE23-0014 APi (Apprivoiser la Pré-image). Funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them. This work received funding from the European Union's Horizon Europe research and innovation program under grant agreement 101120237 (ELIAS).

References

Alaoui, A. and Mahoney, M. W. (2015). Fast randomized kernel ridge regression with statistical guarantees. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28.

Bach, F. (2013). Sharp analysis of low-rank kernel matrix approximations. In *Proc. of the 26th annual Conference on Learning Theory*, pages 185–209. PMLR.

Bakir, G., Hofmann, T., Smola, A. J., Schölkopf, B., and Taskar, B. (2007). *Predicting structured data*. The MIT Press.

Bauer, F., Pereverzev, S., and Rosasco, L. (2007). On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72.

Belanger, D. and McCallum, A. (2016). Structured prediction energy networks. In *International Conference on Machine Learning*, pages 983–992.

Berthet, Q., Blondel, M., Teboul, O., Cuturi, M., Vert, J.-P., and Bach, F. (2020). Learning with differentiable perturbed optimizers.

Blanchard, G., Bousquet, O., and Zwald, L. (2007). Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2):259–294.

Borgwardt, K., Ghisu, E., Llinares-López, F., O'Bray, L., and Rieck, B. (2020). Graph kernels: State-of-the-art and future challenges. *Foundations and Trends in Machine Learning*, 13(5-6):531–712.

Brault, R., Heinonen, M., and Buc, F. (2016). Random fourier features for operator-valued kernels. In *Asian Conference on Machine Learning*, pages 110–125. PMLR.

Brogat-Motte, L., Rudi, A., Brouard, C., Rousu, J., and d'Alché Buc, F. (2022). Vector-valued least-squares regression under output regularity assumptions. *Journal of Machine Learning Research*, 23(344):1–50.

Brouard, C., d'Alché-Buc, F., and Szafranski, M. (2011). Semi-supervised penalized output kernel regression for link prediction. In *International Conference on Machine Learning (ICML)*, pages 593–600.

Brouard, C., Shen, H., Dührkop, K., d'Alché-Buc, F., Böcker, S., and Rousu, J. (2016a). Fast metabolite identification with input output kernel regression. *Bioinformatics*, 32(12):28–36.

Brouard, C., Szafranski, M., and D'Alché-Buc, F. (2016b). Input output kernel regression: supervised and semi-supervised structured output prediction with operator-valued kernels. *The Journal of Machine Learning Research*, 17(1):6105–6152.

Cabannes, V. A., Bach, F., and Rudi, A. (2021). Fast rates for structured prediction. In *conference on learning theory*, pages 823–865. PMLR.

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

Carmeli, C., De Vito, E., and Toigo, A. (2006). Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4(04):377–408.

Carmeli, C., De Vito, E., Toigo, A., and Umanitá, V. (2010). Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61.

Chatalic, A., Carratino, L., De Vito, E., and Rosasco, L. (2022a). Mean nyström embeddings for adaptive compressive learning. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 9869–9889. PMLR.

Chatalic, A., Schreuder, N., Rosasco, L., and Rudi, A. (2022b). Nyström kernel mean embeddings. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 3006–3024. PMLR.

Ciliberto, C., Rosasco, L., and Rudi, A. (2016). A consistent regularization approach for structured prediction. In *Advances in Neural Information Processing Systems (NIPS) 29*, pages 4412–4420.

Ciliberto, C., Rosasco, L., and Rudi, A. (2020). A general framework for consistent structured prediction with implicit loss embeddings. *J. Mach. Learn. Res.*, 21(98):1–67.

Collins, M. and Duffy, N. (2001). Convolution kernels for natural language. *Advances in neural information processing systems*, 14.

Cortes, C., Mohri, M., and Weston, J. (2005). A general regression technique for learning transductions. In *International Conference on Machine Learning (ICML)*, pages 153–160.

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29.

Deshwal, A., Doppa, J. R., and Roth, D. (2019). Learning and inference for structured prediction: A unifying perspective. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*.

Drineas, P., Mahoney, M. W., and Cristianini, N. (2005). On the nyström method for approximating a gram matrix for improved kernel-based learning. *JMLR*, 6(12).

El Ahmad, T., Laforgue, P., and d'Alché Buc, F. (2023). Fast kernel methods for generic lipschitz losses via p-sparsified sketches. *Transactions on Machine Learning Research*.

Fischer, S. and Steinwart, I. (2020). Sobolev norm learning rates for regularized least-squares algorithms. *J. Mach. Learn. Res.*, 21:205–1.

Gärtner, T. (2008). *Kernels for Structured Data*, volume 72 of *Series in Machine Perception and Artificial Intelligence*. WorldScientific.

Geurts, P., Wehenkel, L., and d'Alché Buc, F. (2006). Kernelizing the output of tree-based methods. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 345–352, New York, NY, USA. Association for Computing Machinery.

Grünewälder, S., Lever, G., Baldassarre, L., Patterson, S., Gretton, A., and Pontil, M. (2012). Conditional mean embeddings as regressors. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1803–1810.

Gygli, M., Norouzi, M., and Angelova, A. (2017). Deep value networks learn to evaluate and iteratively refine structured outputs. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1341–1351. JMLR.org.

Kadri, H., Ghavamzadeh, M., and Preux, P. (2013). A generalized kernel approach to structured output learning. In *International Conference on Machine Learning*, pages 471–479. PMLR.

Kalinke, F. and Szabó, Z. (2023). Nystr\" om m-hilbert-schmidt independence criterion. *arXiv preprint arXiv:2302.09930*.

Koltchinskii, V. and Lounici, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110 – 133.

Kpotufe, S. and Sriperumbudur, B. K. (2020). Gaussian sketching yields a J-L lemma in RKHS. In Chiappa, S. and Calandra, R., editors, *AISTATS 2020*, volume 108 of *Proceedings of Machine Learning Research*, pages 3928–3937. PMLR.

Lacotte, J. and Pilanci, M. (2020). Adaptive and oblivious randomized subspace methods for high-dimensional optimization: Sharp analysis and lower bounds. *arXiv preprint arXiv:2012.07054*.

LeCun, Y., Chopra, S., Ranzato, M., and Huang, F.-J. (2007). Energy-based models in document recognition and computer vision. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 1, pages 337–341. IEEE.

Li, Z., Ton, J.-F., Oglic, D., and Sejdinovic, D. (2021). Towards a unified analysis of random fourier features. *Journal of Machine Learning Research*, 22(108):1–51.

Lin, X. V., Singh, S., He, L., Taskar, B., and Zettlemoyer, L. (2014). Multi-label learning with posterior regularization.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Mahoney, M. W. et al. (2011). Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224.

Meanti, G., Carratino, L., Rosasco, L., and Rudi, A. (2020). Kernel methods through the roof: Handling billions of points efficiently. *Advances in Neural Information Processing Systems (NeurIPS)*, 33.

Micchelli, C. A. and Pontil, M. (2005). On learning vector-valued functions. *Neural computation*, 17(1):177–204.

Niculae, V., Martins, A., Blondel, M., and Cardie, C. (2018). Sparsemap: Differentiable sparse structured inference. In *International Conference on Machine Learning (ICML)*, pages 3799–3808. PMLR.

Nowak, A., Bach, F., and Rudi, A. (2019). Sharp analysis of learning with discrete losses. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 1920–1929.

Nowozin, S. and Lampert, C. H. (2011). Structured learning and prediction in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 6(3-4):185–365.

Osokin, A., Bach, F. R., and Lacoste-Julien, S. (2017). On structured prediction theory with calibrated convex surrogate losses. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems (NeurIPS) 30:*, pages 302–313.

Pillaud-Vivien, L., Rudi, A., and Bach, F. (2018). Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems*, 31.

Rahimi, A. and Recht, B. (2007). Random features for large scale kernel machines. *NIPS*, 20:1177–1184.

Ralaivola, L., Swamidass, S. J., Saigo, H., and Baldi, P. (2005). Graph kernels for chemical informatics. *Neural Networks*, 18(8):1093–1110. Neural Networks and Kernel Methods for Structured Domains.

Rudi, A., Camoriano, R., and Rosasco, L. (2015). Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, 28.

Rudi, A., Canas, G. D., and Rosasco, L. (2013). On the sample complexity of subspace learning. In *Advances in Neural Information Processing Systems*, pages 2067–2075.

Rudi, A. and Rosasco, L. (2017). Generalization properties of learning with random features. In *Advances on Neural Information Processing Systems (NeurIPS)*, pages 3215–3225.

Schymanski, E., Ruttkies, C., Krauss, M., Brouard, C., Kind, T., Dührkop, K., Allen, F., Vaniya, A., Verdegem, D., Böcker, S., Rousu, J., Shen, H., Tsugawa, H., Sajed, T., Fiehn, O., Ghesquiere, B., and Neumann, S. (2017). Critical assessment of small molecule identification 2016: automated methods. *Journal of Cheminformatics*, 9:22.

Senkene, E. and Tempel'man, A. (1973). Hilbert spaces of operator-valued functions. *Lithuanian Mathematical Journal*, 13(4):665–670.

Smale, S. and Zhou, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172.

Snoek, C. G. M., Worring, M., van Gemert, J. C., Geusebroek, J.-M., and Smeulders, A. W. M. (2006). The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th ACM International Conference on Multimedia*, MM '06, page 421–430, New York, NY, USA. Association for Computing Machinery.

Steinwart, I., Hush, D. R., Scovel, C., et al. (2009). Optimal rates for regularized least squares regression. In *COLT*, pages 79–93.

Sterge, N. and Sriperumbudur, B. K. (2022). Statistical optimality and computational efficiency of nystrom kernel pca. *Journal of Machine Learning Research*, 23(337):1–32.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of machine learning research*, 6(Sep):1453–1484.

Weston, J., Chapelle, O., Vapnik, V., Elisseeff, A., and Schölkopf, B. (2003). Kernel dependency estimation. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 897–904. MIT Press.

Williams, C. and Seeger, M. (2001). Using the nyström method to speed up kernel machines. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13, pages 682–688. MIT Press.

Woodruff, D. P. (2014). Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1-2):1–157.

Yang, T., Li, Y.-f., Mahdavi, M., Jin, R., and Zhou, Z.-H. (2012). Nyström method vs random fourier features: A theoretical and empirical comparison. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Yang, Y., Pilanci, M., Wainwright, M. J., et al. (2017). Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics*, 45(3):991–1023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] See Sections 2 to 4 and Appendices A to C.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] See Sections 2, 3 and 5 and Table 1.

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] See Section 4.
   (b) Complete proofs of all theoretical results. [Yes] See Appendices C to G.
   (c) Clear explanations of any assumptions. [Yes] See Section 4.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] See Section 5.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] See Section 5 and Appendix I.2.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] See Section 5.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes] See Section 5.
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A NOTATIONS AND DEFINITIONS

In this section, we remind some important notations and definitions.

Setting. In the following, we consider $\mathcal{X}$ and $\mathcal{Y}$ to be Polish spaces. We denote by ρ the unknown data distribution on $\mathcal{X} \times \mathcal{Y}$. We denote by $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$ the marginal distributions of the inputs and outputs, respectively.

Linear algebra notation. For an operator A, $A^{\#}$ is its adjoint, $\sigma_{\max}(A)$ its largest eigenvalue, and $\sigma_k(A)$ its k^{th} largest eigenvalue (if A admits an eigendecomposition). Let $\mathcal{B}(E)$ be the space of bounded linear operators in a separable Hilbert space E, given positive semi-definite operators $A, B \in \mathcal{B}(E)$, $A \preceq B$ if $B - A$ is positive semidefinite. For any $t > 0$ and $A : E \to E$, $A_t = A + tI_E$. Let M be a matrix, $M_{i:}$ denotes its i^{th} row and $M_{:j}$ its j^{th} column, and $M^{\dagger}$ denotes its Moore-Penrose inverse.

Notation for simplified bounds. To keep the dependencies of a bound only in the parameters of interest, for $a, b \in \mathbb{R}$ we note $a \lesssim b$ as soon as there exists a constant $c > 0$ independents of the parameters of interest such that $a \leq c \times b$.

Least-squares notation. For any function $h : \mathcal{X} \to \mathcal{H}_{\mathcal{Y}}$, its least-squares expected risk is given by

$$\mathcal{E}(h) = \mathbb{E}_{\rho}\left[\|h(x) - \psi_{\mathcal{Y}}(y)\|^2_{\mathcal{H}_{\mathcal{Y}}}\right]. \tag{15}$$

The measurable minimizer of $\mathcal{E}$ is given by $h^*(x) = \mathbb{E}_{\rho(y|x)}[\psi_{\mathcal{Y}}(y)]$ (Ciliberto et al., 2020, Lemma A.2).

RKHS notation. We denote by $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ the RKHSs associated to the input $\mathrm{k}_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and output $\mathrm{k}_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ kernels, respectively. We denote by $\psi_{\mathcal{X}} : \mathcal{X} \to \mathcal{H}_{\mathcal{X}}$ and $\psi_{\mathcal{Y}} : \mathcal{Y} \to \mathcal{H}_{\mathcal{Y}}$ the canonical feature maps $\psi_{\mathcal{X}}(x) = \mathrm{k}_{\mathcal{X}}(x, .)$ and $\psi_{\mathcal{Y}}(y) = \mathrm{k}_{\mathcal{Y}}(y, .)$, respectively. We denote by $\mathcal{H}$ the vv-RKHS associated to the operator-valued kernel $\mathcal{K} = kI_{\mathcal{H}_{\mathcal{Y}}}$. We denote $\hat{h} \in \mathcal{H}$ the KRR estimator trained with n couples $(x_i, y_i)_{i=1}^n$ i.i.d. from ρ.

Kernel ridge operators. We define the following operators.

- $S : f \in \mathcal{H}_{\mathcal{X}} \mapsto \langle f, \psi_{\mathcal{X}}(\cdot)\rangle_{\mathcal{H}_{\mathcal{X}}} \in L^2(\mathcal{X}, \rho_{\mathcal{X}})$
- $T : f \in \mathcal{H}_{\mathcal{Y}} \mapsto \langle f, h^*(\cdot)\rangle_{\mathcal{H}_{\mathcal{Y}}} \in L^2(\mathcal{X}, \rho_{\mathcal{X}})$
- $\mathrm{C}_{\mathcal{X}} = \mathbb{E}_x[\psi_{\mathcal{X}}(x) \otimes \psi_{\mathcal{X}}(x)]$ and $\mathrm{C}_{\mathcal{Y}} = \mathbb{E}_y[\psi_{\mathcal{Y}}(y) \otimes \psi_{\mathcal{Y}}(y)]$,
- $\mathrm{S_X} : f \in \mathcal{H}_{\mathcal{X}} \mapsto \frac{1}{\sqrt{n}}(f(x_1), \ldots, f(x_n))^{\top} \in \mathbb{R}^n$,
- ${\mathrm{S_X}}^{\#} : \alpha \in \mathbb{R}^n \mapsto \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i \psi_{\mathcal{X}}(x_i) \in \mathcal{H}_{\mathcal{X}}$,
- $\mathrm{S_Y} : f \in \mathcal{H}_{\mathcal{Y}} \mapsto \frac{1}{\sqrt{n}}(f(y_1), \ldots, f(y_n))^{\top} \in \mathbb{R}^n$,
- ${\mathrm{S_Y}}^{\#} : \alpha \in \mathbb{R}^n \mapsto \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i \psi_{\mathcal{Y}}(y_i) \in \mathcal{H}_{\mathcal{X}}$,

Sketching operators.

- We denote $\mathrm{R}_{\mathcal{X}} \in \mathbb{R}^{\mathrm{m}_{\mathcal{X}} \times n}$ and $\mathrm{R}_{\mathcal{Y}} \in \mathbb{R}^{\mathrm{m}_{\mathcal{Y}} \times n}$ the input and output sketch matrices with $\mathrm{m}_{\mathcal{X}} < n$ and $\mathrm{m}_{\mathcal{Y}} < n$,
- $\widetilde{\mathrm{C}}_{\mathrm{X}} = {\mathrm{S_X}}^{\#} {\mathrm{R}_{\mathcal{X}}}^{\top} \mathrm{R}_{\mathcal{X}} \mathrm{S_X}$ and $\widetilde{\mathrm{C}}_{\mathrm{Y}} = {\mathrm{S_Y}}^{\#} {\mathrm{R}_{\mathcal{Y}}}^{\top} \mathrm{R}_{\mathcal{Y}} \mathrm{S_Y}$,
- $\widetilde{\mathrm{K}}_{\mathrm{X}} = \mathrm{R}_{\mathcal{X}} \mathrm{K_X} {\mathrm{R}_{\mathcal{X}}}^{\top}$ and $\widetilde{\mathrm{K}}_{\mathrm{Y}} = \mathrm{R}_{\mathcal{Y}} \mathrm{K_Y} {\mathrm{R}_{\mathcal{Y}}}^{\top}$.

B REMINDERS ABOUT VECTOR-VALUED REPRODUCING KERNEL HILBERT SPACES AND OPERATOR-VALUED KERNELS

We recall the definitions of an OVK and its vv-RKHS. Let $\mathcal{F}$ be a Hilbert space and $\mathcal{L}(\mathcal{F})$ the set of bounded linear operators on $\mathcal{F}$.

Definition 2 (Operator-valued kernel)**.** *An OVK is a mapping $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{F})$ such that*

- $\mathcal{K}(x, x') = \mathcal{K}(x', x)^{\#}$ *for all* $(x, x') \in \mathcal{X}^2$*;*
- $\sum_{i,j=1}^{n} \langle \varphi_i, \mathcal{K}(x_i, x_j)\, \varphi_j) \rangle_{\mathcal{F}} \geqslant 0$ *for all* $n \in \mathbb{N}$ *and* $(x_i, \varphi_i))_{i=1}^{n} \in (\mathcal{X} \times \mathcal{F})^n$*.*

Similarly to the scalar case, an OVK is uniquely associated to a vv-RKHS $\mathcal{H}$.

Theorem 3 (vector-valued RKHS)**.** *Let $\mathcal{K}$ be an OVK. There is a unique Hilbert space $\mathcal{H}$ of functions from $\mathcal{X}$ to $\mathcal{F}$, the vv-RKHS of $\mathcal{K}$, such that for all $x \in \mathcal{X}$, $\varphi \in \mathcal{F}$ and $f \in \mathcal{H}$*

- $x' \mapsto \mathcal{K}(x, x')\, \varphi \in \mathcal{F}$*;*
- $\langle f, \mathcal{K}(\cdot, x)\, \varphi \rangle_{\mathcal{H}} = \langle f(x), \varphi \rangle_{\mathcal{F}}$ *(reproducing property).*

C PRELIMINARY RESULTS

In this section, we present useful preliminary results about kernel ridge operators and sketching properties, as well as the proof Proposition 1 that give the expressions of the SISOKR estimator.

Useful kernel ridge operators properties. The following results hold true.

- $\widehat{\mathrm{C}}_{\mathrm{X}} = \frac{1}{n} \sum_{i=1}^{n} \psi_{\mathcal{X}}(x_i) \otimes \psi_{\mathcal{X}}(x_i) = \mathrm{S_X}^{\#}\, \mathrm{S_X}$ and $\widehat{\mathrm{C}}_{\mathrm{Y}} = \frac{1}{n} \sum_{i=1}^{n} \psi_{\mathcal{Y}}(y_i) \otimes \psi_{\mathcal{Y}}(y_i) = \mathrm{S_Y}^{\#}\, \mathrm{S_Y}$,
- $\mathrm{K_X} = n\, \mathrm{S_X}\, \mathrm{S_X}^{\#}$ and $\mathrm{K_Y} = n\, \mathrm{S_Y}\, \mathrm{S_Y}^{\#}$,
- Under the attainability assumption (Ciliberto et al., 2020, Lemma B.2, B.4, B.9) show that:
  - For all $x \in \mathcal{X}$, $\hat{h}(x) = \widehat{H}\, \psi_{\mathcal{X}}(x)$, where $\widehat{H} = \mathrm{S_Y}^{\#}\, \mathrm{S_X}\, \widehat{\mathrm{C}}_{\mathrm{X}\lambda}^{\ -1}$.
  - $\mathbb{E}[\|\hat{h}(x) - h^*(x)\|^2]^{1/2} = \|(\widehat{H} - H) S^{\#}\|_{\mathrm{HS}}$.

Useful sketching properties. We remind some useful notations and provide the expression of $\widetilde{\mathrm{P}}_{\mathrm{Z}}$, leading to the expression of the SISOKR estimator.

Expression of $\widetilde{\mathrm{P}}_{\mathrm{Z}}$. Let $\left\{ (\sigma_i(\widetilde{\mathrm{K}}_{\mathrm{Z}}), \tilde{\mathbf{v}}_i^{Z}), i \in [\mathrm{m}_{\mathcal{Z}}] \right\}$ be the eigenpairs of $\widetilde{\mathrm{K}}_{\mathrm{Z}}$ ranked in descending order of eigenvalues, $\mathrm{p_Z} = \mathrm{rank}\left(\widetilde{\mathrm{K}}_{\mathrm{Z}}\right)$, and for all $1 \leq i \leq \mathrm{p_Z}$, $\tilde{e}_i^{Z} = \sqrt{\frac{n}{\sigma_i(\widetilde{\mathrm{K}}_{\mathrm{Z}})}}\, \mathrm{S_Z}^{\#}\, \mathrm{R_{\mathcal{Z}}}^{\top}\, \tilde{\mathbf{v}}_i^{Z}$.

Proposition 2. *The $\tilde{e}_i^{Z}$s are the eigenfunctions, associated to the eigenvalues $\sigma_i(\widetilde{\mathrm{K}}_{\mathrm{Z}})/n$ of $\widetilde{\mathrm{C}}_{\mathrm{Z}}$. Furthermore, let $\widetilde{\mathcal{H}}_{\mathcal{Z}} = \mathrm{span}\left(\tilde{e}_1^{z}, \ldots, \tilde{e}_{\mathrm{p_Z}}^{z}\right)$, the orthogonal projector $\widetilde{\mathrm{P}}_{\mathrm{Z}}$ onto $\widetilde{\mathcal{H}}_{\mathcal{Z}}$ writes as*

$$\widetilde{\mathrm{P}}_{\mathrm{Z}} = (\mathrm{R_{\mathcal{Z}}}\, \mathrm{S_Z})^{\#} \left(\mathrm{R_{\mathcal{Z}}}\, \mathrm{S_Z} (\mathrm{R_{\mathcal{Z}}}\, \mathrm{S_Z})^{\#}\right)^{\dagger} \mathrm{R_{\mathcal{Z}}}\, \mathrm{S_Z}\ . \tag{16}$$

Proof. For $1 \leq i \leq \mathrm{p_Z}$

$$\widetilde{\mathrm{C}}_{\mathrm{Z}}\ \tilde{e}_i^Z = {\mathrm{S_Z}}^{\#}\,{\mathrm{R}_{\mathcal{Z}}}^{\top}\,\mathrm{R}_{\mathcal{Z}}\,\mathrm{S_Z}\left(\sqrt{\frac{n}{\sigma_i(\widetilde{\mathrm{K}}_{\mathrm{Z}})}}\,{\mathrm{S_Z}}^{\#}\,{\mathrm{R}_{\mathcal{Z}}}^{\top}\,\tilde{\mathbf{v}}_i^Z\right) \tag{17}$$

$$= \sqrt{\frac{n}{\sigma_i(\widetilde{\mathrm{K}}_{\mathrm{Z}})}}\,{\mathrm{S_Z}}^{\#}\,{\mathrm{R}_{\mathcal{Z}}}^{\top}\left(\frac{1}{n}\,\widetilde{\mathrm{K}}_{\mathrm{Z}}\right)\tilde{\mathbf{v}}_i^Z \tag{18}$$

$$= \frac{1}{\sqrt{n\sigma_i(\widetilde{\mathrm{K}}_{\mathrm{Z}})}}\,{\mathrm{S_Z}}^{\#}\,{\mathrm{R}_{\mathcal{Z}}}^{\top}\,\sigma_i(\widetilde{\mathrm{K}}_{\mathrm{Z}})\tilde{\mathbf{v}}_i^Z \tag{19}$$

$$= \frac{\sigma_i(\widetilde{\mathrm{K}}_{\mathrm{Z}})}{n}\tilde{e}_i^Z\,. \tag{20}$$

Moreover, we verify that span $\left(\tilde{e}_1^Z, \ldots, \tilde{e}_{\mathrm{p_Z}}^Z\right)$ forms an orthonormal basis. Let $1 \leq i, j \leq \mathrm{p_Z}$,

$$\left\langle \tilde{e}_i^Z, \tilde{e}_j^Z \right\rangle_{\mathcal{H}_{\mathcal{X}}} = \left\langle \sqrt{\frac{n}{\sigma_i(\widetilde{\mathrm{K}}_{\mathrm{Z}})}}\,{\mathrm{S_Z}}^{\#}\,{\mathrm{R}_{\mathcal{Z}}}^{\top}\,\tilde{\mathbf{v}}_i^Z, \sqrt{\frac{n}{\sigma_j(\widetilde{\mathrm{K}}_{\mathrm{Z}})}}\,{\mathrm{S_Z}}^{\#}\,{\mathrm{R}_{\mathcal{Z}}}^{\top}\,\tilde{\mathbf{v}}_j^Z \right\rangle_{\mathcal{H}_{\mathcal{Z}}} \tag{21}$$

$$= \frac{n}{\sqrt{\sigma_i(\widetilde{\mathrm{K}}_{\mathrm{Z}})\sigma_j(\widetilde{\mathrm{K}}_{\mathrm{Z}})}}{\tilde{\mathbf{v}}_i^Z}^{\top}\,\mathrm{R}_{\mathcal{Z}}\,\mathrm{S_Z}\,{\mathrm{S_Z}}^{\#}\,{\mathrm{R}_{\mathcal{Z}}}^{\top}\,\tilde{\mathbf{v}}_j^Z \tag{22}$$

$$= \frac{n}{\sqrt{\sigma_i(\widetilde{\mathrm{K}}_{\mathrm{Z}})\sigma_j(\widetilde{\mathrm{K}}_{\mathrm{Z}})}}{\tilde{\mathbf{v}}_i^Z}^{\top}\left(\frac{1}{n}\,\widetilde{\mathrm{K}}_{\mathrm{Z}}\right)\tilde{\mathbf{v}}_j^Z \tag{23}$$

$$= \frac{\sigma_j(\widetilde{\mathrm{K}}_{\mathrm{Z}})}{\sqrt{\sigma_i(\widetilde{\mathrm{K}}_{\mathrm{Z}})\sigma_j(\widetilde{\mathrm{K}}_{\mathrm{Z}})}}{\tilde{\mathbf{v}}_i^Z}^{\top}\tilde{\mathbf{v}}_j^Z \tag{24}$$

$$= \delta_{ij}\,, \tag{25}$$

where $\delta_{ij} = 0$ if $i \neq j$, and 1 otherwise.

Finally, it is easy to check that the orthogonal projector onto span $\left(\tilde{e}_1^Z, \ldots, \tilde{e}_{\mathrm{p_Z}}^Z\right)$, i.e. $\widetilde{\mathrm{P}}_{\mathrm{Z}} : f \in \mathcal{H}_{\mathcal{Z}} \mapsto \sum_{i=1}^{\mathrm{p_Z}} \left\langle f, \tilde{e}_i^Z \right\rangle_{\mathcal{H}_{\mathcal{Z}}} \tilde{e}_i^Z$ rewrites as

$$\widetilde{\mathrm{P}}_{\mathrm{Z}} = n\,{\mathrm{S_Z}}^{\#}\,{\mathrm{R}_{\mathcal{Z}}}^{\top}\,{\widetilde{\mathrm{K}}_{\mathrm{Z}}}^{\dagger}\,\mathrm{R}_{\mathcal{Z}}\,\mathrm{S_Z} = {(\mathrm{R}_{\mathcal{Z}}\,\mathrm{S_Z})}^{\#}\left(\mathrm{R}_{\mathcal{Z}}\,\mathrm{S_Z}{(\mathrm{R}_{\mathcal{Z}}\,\mathrm{S_Z})}^{\#}\right)^{\dagger}\mathrm{R}_{\mathcal{Z}}\,\mathrm{S_Z}\,. \tag{26}$$

□

Remark 4. *With* $\mathrm{R}_{\mathcal{X}}$ *a sub-sampling matrix, we recover the linear operator* L_m *introduced in Yang et al. (2012) for the study of Nyström approximation and its eigendecomposition. Moreover, we also recover the projection operator* P_m *from Rudi et al. (2015) and follow the footsteps of the proposed extension "Nyström with sketching matrices".*

Algorithm. We here give the proof of Proposition 1 that provides an expression of the SISOKR estimator $\tilde{h}$ as a linear combination of the $\psi_{\mathcal{Y}}(y_i)$s.

Proposition 1 (Expression of SISOKR)**.** $\forall\, x \in \mathcal{X}$,

$$\tilde{h}\,(x) = \sum_{i=1}^{n} \tilde{\alpha}_i\,(x)\,\psi_{\mathcal{Y}}\,(y_i)\,,$$

where $\tilde{\alpha}\,(x) = {\mathrm{R}_{\mathcal{Y}}}^{\top}\,\widetilde{\Omega}\,\mathrm{R}_{\mathcal{X}}\,\mathrm{k}_{\mathrm{X}}^{\mathrm{x}}$ *and*

$$\widetilde{\Omega} = {\widetilde{\mathrm{K}}_{\mathrm{Y}}}^{\dagger}\,\mathrm{R}_{\mathcal{Y}}\,\mathrm{K_Y}\,\mathrm{K_X}\,\mathrm{R}_{\mathcal{X}}{}^{\top}(\mathrm{R}_{\mathcal{X}}\,{\mathrm{K_X}}^2\,{\mathrm{R}_{\mathcal{X}}}^{\top} + n\lambda\,\widetilde{\mathrm{K}}_{\mathrm{X}})^{\dagger}\,,$$

with $\widetilde{\mathrm{K}}_{\mathrm{X}} = \mathrm{R}_{\mathcal{X}}\,\mathrm{K_X}\,{\mathrm{R}_{\mathcal{X}}}^{\top}$ *and* $\widetilde{\mathrm{K}}_{\mathrm{Y}} = \mathrm{R}_{\mathcal{Y}}\,\mathrm{K_Y}\,{\mathrm{R}_{\mathcal{Y}}}^{\top}$.

Proof. Recall that $\tilde{h}(x) = \widetilde{\mathrm{P}}_{\mathrm{Y}}\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\,\mathrm{S}_{\mathrm{X}}\,\widetilde{\mathrm{P}}_{\mathrm{X}}(\widetilde{\mathrm{P}}_{\mathrm{X}}\,{\mathrm{S}_{\mathrm{X}}}^{\#}\,\mathrm{S}_{\mathrm{X}}\,\widetilde{\mathrm{P}}_{\mathrm{X}} + \lambda I_{\mathcal{H}_{\mathcal{X}}})^{-1}\,\psi_{\mathcal{X}}(x)$. By Lemma 1 and especially (30), we obtain that

$$\tilde{h}(x) = \sqrt{n}\,\widetilde{\mathrm{P}}_{\mathrm{Y}}\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\,\mathrm{K}_{\mathrm{X}}\,{\mathrm{R}_{\mathcal{X}}}^{\top}\left(\mathrm{R}_{\mathcal{X}}\,{\mathrm{K}_{\mathrm{X}}}^{2}\,{\mathrm{R}_{\mathcal{X}}}^{\top} + n\lambda\,\mathrm{R}_{\mathcal{X}}\,\mathrm{K}_{\mathrm{X}}\,{\mathrm{R}_{\mathcal{X}}}^{\top}\right)^{\dagger}\mathrm{R}_{\mathcal{X}}\,\mathrm{S}_{\mathrm{X}}\,\psi_{\mathcal{X}}(x)\,. \tag{27}$$

Finally, by Lemma 2 and with $\alpha(x) = \mathrm{K}_{\mathrm{X}}\,{\mathrm{R}_{\mathcal{X}}}^{\top}\left(\mathrm{R}_{\mathcal{X}}\,{\mathrm{K}_{\mathrm{X}}}^{2}\,{\mathrm{R}_{\mathcal{X}}}^{\top} + n\lambda\,\mathrm{R}_{\mathcal{X}}\,\mathrm{K}_{\mathrm{X}}\,{\mathrm{R}_{\mathcal{X}}}^{\top}\right)^{\dagger}\mathrm{R}_{\mathcal{X}}\,\mathrm{S}_{\mathrm{X}}\,\psi_{\mathcal{X}}(x)$, we have that $\tilde{h}\,(x) = \sum_{i=1}^{n}\tilde{\alpha}_i\,(x)\,\psi_{\mathcal{Y}}\,(y_i)$ where

$$\tilde{\alpha}\,(x) = {\mathrm{R}_{\mathcal{Y}}}^{\top}\,{\widetilde{\mathrm{K}}_{\mathrm{Y}}}^{\dagger}\,\mathrm{R}_{\mathcal{Y}}\,\mathrm{K}_{\mathrm{Y}}\,\mathrm{K}_{\mathrm{X}}\,{\mathrm{R}_{\mathcal{X}}}^{\top}({\mathrm{R}_{\mathcal{X}}}\,{\mathrm{K}_{\mathrm{X}}}^{2}\,{\mathrm{R}_{\mathcal{X}}}^{\top} + n\lambda\,\widetilde{\mathrm{K}}_{\mathrm{X}})^{\dagger}\,\mathrm{R}_{\mathcal{X}}\,\mathrm{k}_{\mathrm{X}}^{\mathrm{x}}\,. \tag{28}$$

□

Before stating and proving Lemmas 1 and 2, and similarly to Rudi et al. (2015), let $\mathrm{R}_{\mathcal{X}}\,\mathrm{S}_{\mathrm{X}} = U\Sigma V^{\#}$ be the SVD of $\mathrm{R}_{\mathcal{X}}\,\mathrm{S}_{\mathrm{X}}$ where $U : \mathbb{R}^{\mathrm{px}} \to \mathbb{R}^{\mathrm{m}_{\mathcal{X}}}$, $\Sigma : \mathbb{R}^{\mathrm{px}} \to \mathbb{R}^{\mathrm{px}}$, $V : \mathbb{R}^{\mathrm{px}} \to \mathcal{H}_{\mathcal{X}}$, and $\Sigma = \mathrm{diag}(\sigma_1(\mathrm{R}_{\mathcal{X}}\,\mathrm{S}_{\mathrm{X}}), \ldots, \sigma_{\mathrm{px}}(\mathrm{R}_{\mathcal{X}}\,\mathrm{S}_{\mathrm{X}}))$ with $\sigma_1(\mathrm{R}_{\mathcal{X}}\,\mathrm{S}_{\mathrm{X}}) \geq \ldots \geq \sigma_{\mathrm{px}}(\mathrm{R}_{\mathcal{X}}\,\mathrm{S}_{\mathrm{X}}) > 0$, $UU^{\top} = I_{\mathrm{px}}$ and $V^{\#}V = I_{\mathrm{px}}$. We are now ready to prove the following lemma for the expansion induced by input sketching.

Lemma 1. *Let $\widetilde{H} = \widetilde{\mathrm{P}}_{\mathrm{Y}}\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\,\mathrm{S}_{\mathrm{X}}\,\widetilde{\mathrm{P}}_{\mathrm{X}}(\widetilde{\mathrm{P}}_{\mathrm{X}}\,{\mathrm{S}_{\mathrm{X}}}^{\#}\,\mathrm{S}_{\mathrm{X}}\,\widetilde{\mathrm{P}}_{\mathrm{X}} + \lambda I_{\mathcal{H}_{\mathcal{X}}})^{-1}$. The following two expansions hold true*

$$\widetilde{H} = \widetilde{\mathrm{P}}_{\mathrm{Y}}\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\,\mathrm{S}_{\mathrm{X}}\,\tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}})\,, \tag{29}$$

where $\tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}}) = V(V^{\#}\,\widehat{\mathrm{C}}_{\mathrm{X}}\,V + \lambda I_{\mathcal{H}_{\mathcal{X}}})^{-1}V^{\#}$ and for algorithmic purposes

$$\widetilde{H} = \sqrt{n}\,\widetilde{\mathrm{P}}_{\mathrm{Y}}\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\,\mathrm{K}_{\mathrm{X}}\,{\mathrm{R}_{\mathcal{X}}}^{\top}\left(\mathrm{R}_{\mathcal{X}}\,{\mathrm{K}_{\mathrm{X}}}^{2}\,{\mathrm{R}_{\mathcal{X}}}^{\top} + n\lambda\,\mathrm{R}_{\mathcal{X}}\,\mathrm{K}_{\mathrm{X}}\,{\mathrm{R}_{\mathcal{X}}}^{\top}\right)^{\dagger}\mathrm{R}_{\mathcal{X}}\,\mathrm{S}_{\mathrm{X}}\,. \tag{30}$$

Proof. Let us prove (29) first.

$$\begin{aligned}
\tilde{H} &= \widetilde{\mathrm{P}}_{\mathrm{Y}}\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\,\mathrm{S}_{\mathrm{X}}\,\widetilde{\mathrm{P}}_{\mathrm{X}}(\widetilde{\mathrm{P}}_{\mathrm{X}}\,{\mathrm{S}_{\mathrm{X}}}^{\#}\,\mathrm{S}_{\mathrm{X}}\,\widetilde{\mathrm{P}}_{\mathrm{X}} + \lambda I_{\mathcal{H}_{\mathcal{X}}})^{-1} && (31)\\
&= \widetilde{\mathrm{P}}_{\mathrm{Y}}\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\,\mathrm{S}_{\mathrm{X}}\,VV^{\#}(VV^{\#}\,{\mathrm{S}_{\mathrm{X}}}^{\#}\,\mathrm{S}_{\mathrm{X}}\,VV^{\#} + \lambda I_{\mathcal{H}_{\mathcal{X}}})^{-1} && (32)\\
&= \widetilde{\mathrm{P}}_{\mathrm{Y}}\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\,\mathrm{S}_{\mathrm{X}}\,V(V^{\#}\hat{\mathrm{C}_{\mathcal{X}}}V + \lambda I_{\mathcal{H}_{\mathcal{X}}})^{-1}V^{\#} && (33)\\
&= \widetilde{\mathrm{P}}_{\mathrm{Y}}\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\,\mathrm{S}_{\mathrm{X}}\,\tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}})\,, && (34)
\end{aligned}$$

using the so-called push-through identity $(I + UV)^{-1}U = U(I + VU)^{-1}$.

Now, we focus on proving (30). First, we have that

$$\tilde{H} = \widetilde{\mathrm{P}}_{\mathrm{Y}}\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\,\mathrm{S}_{\mathrm{X}}\,V(V^{\#}\,\widehat{\mathrm{C}}_{\mathrm{X}\lambda}\,V)^{\dagger}V^{\#}\,. \tag{35}$$

Then, using the fact that U has orthonormal columns, $U^{\top}$ has orthonormal rows and Σ is a full-rank matrix, together with the fact that $UU^{\top} = I_{\mathrm{px}}$ and $V^{\#}V = I_{\mathrm{px}}$, we have that,

$$\tilde{H} = \widetilde{\mathrm{P}}_{\mathrm{Y}}\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\,\mathrm{S}_{\mathrm{X}}\,V\Sigma U^{\top}\left(U\Sigma V^{\#}\,\widehat{\mathrm{C}}_{\mathrm{X}\lambda}\,V\Sigma U^{\top}\right)^{\dagger}U\Sigma V^{\#}\,. \tag{36}$$

Then, since $\mathrm{R}_{\mathcal{X}}\,\mathrm{S}_{\mathrm{X}} = U\Sigma V^{\#}$,

$$\tilde{H} = \widetilde{\mathrm{P}}_{\mathrm{Y}}\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\,\mathrm{S}_{\mathrm{X}}(\mathrm{R}_{\mathcal{X}}\,\mathrm{S}_{\mathrm{X}})^{\#}\left(\mathrm{R}_{\mathcal{X}}\,\mathrm{S}_{\mathrm{X}}\left(\widehat{\mathrm{C}}_{\mathrm{X}} + \lambda I_{\mathcal{H}_{\mathcal{X}}}\right)(\mathrm{R}_{\mathcal{X}}\,\mathrm{S}_{\mathrm{X}})^{\#}\right)^{\dagger}\mathrm{R}_{\mathcal{X}}\,\mathrm{S}_{\mathrm{X}}\,. \tag{37}$$

Finally, using the fact that $\widehat{\mathrm{C}}_{\mathrm{X}} = {\mathrm{S}_{\mathrm{X}}}^{\#}\,\mathrm{S}_{\mathrm{X}}$ and $\mathrm{K}_{\mathrm{X}} = n\,\mathrm{S}_{\mathrm{X}}\,{\mathrm{S}_{\mathrm{X}}}^{\#}$, we obtain that

$$\widetilde{H} = \sqrt{n}\,\widetilde{\mathrm{P}}_{\mathrm{Y}}\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\,\mathrm{K}_{\mathrm{X}}\,{\mathrm{R}_{\mathcal{X}}}^{\top}\left(\mathrm{R}_{\mathcal{X}}\,{\mathrm{K}_{\mathrm{X}}}^{2}\,{\mathrm{R}_{\mathcal{X}}}^{\top} + n\lambda\,\mathrm{R}_{\mathcal{X}}\,\mathrm{K}_{\mathrm{X}}\,{\mathrm{R}_{\mathcal{X}}}^{\top}\right)^{\dagger}\mathrm{R}_{\mathcal{X}}\,\mathrm{S}_{\mathrm{X}}\,. \tag{38}$$

□

Now we state and prove the lemma for the expansion induced by output sketching.

Lemma 2. *For all $x \in \mathcal{X}$, for any $h \in \mathcal{H}$ that writes as $h(x) = \sqrt{n}\,\widetilde{\mathrm{P}}_{\mathrm{Y}}\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\,\alpha(x)$ with $\alpha : \mathcal{X} \to \mathbb{R}^n$, then $h(x) = \sum_{i=1}^{n} {\mathrm{R}_{\mathcal{Y}}}^{\top}\,{\widetilde{\mathrm{K}}_{\mathrm{Y}}}^{\dagger}\,\mathrm{R}_{\mathcal{Y}}\,\mathrm{K}_{\mathrm{Y}}\,\alpha(x)\,\psi_{\mathcal{Y}}(y_i)$.*

Proof.

$$h(x) = \sqrt{n}\,\widetilde{\mathrm{P}}_{\mathrm{Y}}\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\,\alpha(x) \tag{39}$$

$$= \sqrt{n}\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\,{\mathrm{R}_{\mathcal{Y}}}^{\top}\,{\widetilde{\mathrm{K}}_{\mathrm{Y}}}^{\dagger}\,\mathrm{R}_{\mathcal{Y}}\left(n\,\mathrm{S}_{\mathrm{Y}}\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\right)\alpha(x) \tag{40}$$

$$= \sqrt{n}\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\,{\mathrm{R}_{\mathcal{Y}}}^{\top}\,{\widetilde{\mathrm{K}}_{\mathrm{Y}}}^{\dagger}\,\mathrm{R}_{\mathcal{Y}}\,\mathrm{K}_{\mathrm{Y}}\,\alpha(x) \tag{41}$$

$$= \sum_{i=1}^{n} {\mathrm{R}_{\mathcal{Y}}}^{\top}\,{\widetilde{\mathrm{K}}_{\mathrm{Y}}}^{\dagger}\,\mathrm{R}_{\mathcal{Y}}\,\mathrm{K}_{\mathrm{Y}}\,\alpha(x)\,\psi_{\mathcal{Y}}(y_i)\,. \tag{42}$$

□

D SISOKR EXCESS-RISK BOUND

In this section, we provide the proof of Theorem 1 which gives a bound on the excess-risk of the proposed approximated regression estimator with both input and output sketching (SISOKR).

Theorem 1 (SISOKR excess-risk bound)**.** *Let $\delta \in (0, 1]$, $n \in \mathbb{N}$ such that $\lambda = n^{-1/(1+\gamma_{\mathcal{X}})} \geq \frac{9\,{\kappa_{\mathcal{X}}}^2}{n} \log(\frac{n}{\delta})$. Under Assumptions 1 to 4, with probability $1-\delta$ we have*

$$\mathbb{E}_x\left[\|\tilde{h}(x) - h^*(x)\|^2_{\mathcal{H}_{\mathcal{Y}}}\right]^{\frac{1}{2}} \leq S(n,\delta) + c_2 A^{\psi_{\mathcal{X}}}_{\rho_{\mathcal{X}}}(\widetilde{P}_X) + A^{\psi_{\mathcal{Y}}}_{\rho_{\mathcal{Y}}}(\widetilde{P}_Y)\,, \tag{9}$$

where $S(n,\delta) = c_1 \log(4/\delta)\, n^{-\frac{1}{2(1+\gamma_{\mathcal{X}})}}$ and

$$A^{\psi_{\mathcal{Z}}}_{\rho_{\mathcal{Z}}}(\widetilde{P}_Z) = \mathbb{E}_z\left[\|(\widetilde{\mathrm{P}}_{\mathrm{Z}} - I_{\mathcal{H}_{\mathcal{Z}}})\,\psi_{\mathcal{Z}}(z)\|^2_{\mathcal{H}_{\mathcal{Z}}}\right]^{\frac{1}{2}},$$

with $c_1, c_2 > 0$ constants independent of n and δ.

Proof. Our proofs consist of decompositions and then applying the probabilistic bounds given in Section F.

We have

$$\mathbb{E}[\|\tilde{h}(x) - h^*(x)\|^2]^{1/2} = \|(\tilde{H} - H)S^{\#}\|_{\mathrm{HS}} \tag{43}$$

with $\tilde{H} = \widetilde{P}_Y\,{\mathrm{S}_{\mathrm{Y}}}^{\#}\,\mathrm{S}_{\mathrm{X}}\,\tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}})$.

Then, defining $H_\lambda = H\,\mathrm{C}_{\mathcal{X}}(\mathrm{C}_{\mathcal{X}} + \lambda I)^{-1}$, we decompose

$$\tilde{H} - H = \widetilde{P}_Y\left({\mathrm{S}_{\mathrm{Y}}}^{\#}\,\mathrm{S}_{\mathrm{X}} - H_\lambda\,\widehat{\mathrm{C}}_{\mathrm{X}}\right)\tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}}) + \widetilde{P}_Y H_\lambda\left(\widehat{\mathrm{C}}_{\mathrm{X}}\,\tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}}) - I_{\mathcal{H}_{\mathcal{X}}}\right) + \left(\widetilde{P}_Y H_\lambda - H\right) \tag{44}$$

such that

$$\|(\tilde{H} - H)S^{\#}\|_{\mathrm{HS}} \leq (A) + (B) + (C) \tag{45}$$

with

$$(A) = \left\|\left({\mathrm{S}_{\mathrm{Y}}}^{\#}\,\mathrm{S}_{\mathrm{X}} - H_\lambda\,\widehat{\mathrm{C}}_{\mathrm{X}}\right)\tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}})\,{\mathrm{C}_{\mathcal{X}}}^{1/2}\right\|_{\mathrm{HS}} \tag{46}$$

$$(B) = \left\|H_\lambda\left(\widehat{\mathrm{C}}_{\mathrm{X}}\,\tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}}) - I_{\mathcal{H}_{\mathcal{X}}}\right){\mathrm{C}_{\mathcal{X}}}^{1/2}\right\|_{\mathrm{HS}} \tag{47}$$

$$(C) = \left\|(\widetilde{P}_Y H_\lambda - H)\,{\mathrm{C}_{\mathcal{X}}}^{1/2}\right\|_{\mathrm{HS}} \tag{48}$$

Then, from Lemmas 3 to 5, we obtain

$$\|(\tilde{H} - H)S^{\#}\|_{\mathrm{HS}} \leq 2\sqrt{3}M \log(4/\delta) n^{-\frac{1}{2(1+\gamma_{\mathcal{X}})}} + 2\sqrt{3}\|H\|_{\mathrm{HS}} \|(I - \widetilde{P}_X)\, \mathrm{C}_{\mathcal{X}}{}^{1/2}\|_{\mathrm{op}} \tag{49}$$

$$+ \mathbb{E}_y \left[\left\| \left(\widetilde{\mathrm{P}}_{\mathrm{Y}} - I_{\mathcal{H}_{\mathcal{Y}}} \right) \psi_{\mathcal{Y}}(y) \right\|_{\mathcal{H}_{\mathcal{Y}}}^2 \right]^{1/2}. \tag{50}$$

Then, notice that

$$\left\|(I - \widetilde{P}_X)\, \mathrm{C}_{\mathcal{X}}{}^{1/2}\right\|_{\mathrm{op}} \leq \left\|(I - \widetilde{P}_X)\, \mathrm{C}_{\mathcal{X}}{}^{1/2}\right\|_{\mathrm{HS}} \tag{51}$$

$$= \mathbb{E}_x \left[\left\| \left(\widetilde{\mathrm{P}}_{\mathrm{X}} - I_{\mathcal{H}_{\mathcal{X}}} \right) \psi_{\mathcal{X}}(x) \right\|_{\mathcal{H}_{\mathcal{X}}}^2 \right]^{1/2}. \tag{52}$$

We conclude by defining

$$c_1 = 2\sqrt{3}M, \tag{53}$$

$$c_2 = 2\sqrt{3}\|H\|_{\mathrm{HS}}. \tag{54}$$

□

Lemma 3 (Bound (A)). *Let $\delta \in [0,1]$, $n \in \mathbb{N}$ sufficiently large such that $\lambda = n^{-1/(1+\gamma)} \geq \frac{9\kappa_{\mathcal{X}}{}^2}{n} \log(\frac{n}{x})$ Under our set of assumptions, the following holds with probability at least $1-\delta$*

$$(A) \leq 2M \log(4/\delta) n^{-\frac{1}{2(1+\gamma_{\mathcal{X}})}}. \tag{55}$$

where the constant M depends on $\kappa_{\mathcal{Y}}, \|H\|_{\mathrm{HS}}, \delta$.

Proof. We have

$$(A) \leq \underbrace{\left\| \left(\mathrm{S_Y}^{\#}\, \mathrm{S_X} - H_\lambda\, \widehat{\mathrm{C}}_{\mathrm{X}} \right) \mathrm{C}_{\mathcal{X}\lambda}{}^{-1/2} \right\|_{\mathrm{HS}}}_{(A.1)} \times \underbrace{\|\, \mathrm{C}_{\mathcal{X}\lambda}{}^{1/2}\, \tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}})\, \mathrm{C}_{\mathcal{X}}{}^{1/2}\, \|_{\mathrm{op}}}_{(A.2)} \tag{56}$$

Moreover, we have

$$(A.2) \leq \|\, \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{1/2}\, \tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}})\, \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{1/2}\, \|_{\mathrm{op}} \|\, \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{-1/2}\, \mathrm{C}_{\mathcal{X}\lambda}{}^{1/2}\, \|_{\mathrm{op}}^2 \|\, \mathrm{C}_{\mathcal{X}\lambda}{}^{-1/2}\, \mathrm{C}_{\mathcal{X}}{}^{1/2}\, \|_{\mathrm{op}} \tag{57}$$

$$\leq \|\, \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{1/2}\, \tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}})\, \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{1/2}\, \|_{\mathrm{op}} \|\, \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{-1/2}\, \mathrm{C}_{\mathcal{X}\lambda}{}^{1/2}\, \|_{\mathrm{op}}^2 \tag{58}$$

because $\|\, \mathrm{C}_{\mathcal{X}\lambda}{}^{-1/2}\, \mathrm{C}_{\mathcal{X}}{}^{1/2}\, \|_{\mathrm{op}} \leq 1$.

Finally, by using the probabilistic bounds given in Lemmas 8 and 9, and Lemma 13, we obtain

$$(A) \leq 2M \log(4/\delta) n^{-\frac{1}{2(1+\gamma_{\mathcal{X}})}}. \tag{59}$$

□

Lemma 4 (Bound (B)). *If $\frac{9}{n} \log \frac{n}{\delta} \leq \lambda \leq \|\, \mathrm{C}_{\mathcal{X}}\, \|_{\mathrm{op}}$, then with probability $1-\delta$*

$$(B) \leq 2\sqrt{3}\|H\|_{\mathrm{HS}} (\lambda^{1/2} + \|(I - \widetilde{P}_X)\, \mathrm{C}_{\mathcal{X}}{}^{1/2}\|_{\mathrm{op}}) \tag{60}$$

Proof. We do a similar decomposition than in Rudi et al. (2015, Theorem 2):

$$\widehat{\mathrm{C}}_{\mathrm{X}}\, \tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}}) - I_{\mathcal{H}_{\mathcal{X}}} = \widehat{\mathrm{C}}_{\mathrm{X}\lambda}\, \tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}}) - \lambda \tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}}) - I_{\mathcal{H}_{\mathcal{X}}} \tag{61}$$

$$= (I - \widetilde{P}_X)\, \widehat{\mathrm{C}}_{\mathrm{X}\lambda}\, \tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}}) + \widetilde{P}_X\, \widehat{\mathrm{C}}_{\mathrm{X}\lambda}\, \tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}}) - \lambda \tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}}) - I_{\mathcal{H}_{\mathcal{X}}} \tag{62}$$

$$= (I - \widetilde{P}_X)\, \widehat{\mathrm{C}}_{\mathrm{X}\lambda}\, \tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}}) - \lambda \tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}}) - (\widetilde{P}_X - I_{\mathcal{H}_{\mathcal{X}}}), \tag{63}$$

as $\widetilde{P}_X \, \widehat{\mathrm{C}}_{\mathrm{X}\lambda} \, \tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}}) = \widetilde{P}_X$.

Then, we have

$$(B) \leq \|H_\lambda\|_{\mathrm{HS}} \left\| \left(\widehat{\mathrm{C}}_{\mathrm{X}} \, \tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}}) - I_{\mathcal{H}_\mathcal{X}} \right) \mathrm{C}_{\mathcal{X}}{}^{1/2} \right\|_{\mathrm{op}} \tag{64}$$

$$\leq \|H_\lambda\|_{\mathrm{HS}} \left(\left\| (I - \widetilde{P}_X) \, \widehat{\mathrm{C}}_{\mathrm{X}\lambda} \, \tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}}) \, \mathrm{C}_{\mathcal{X}}{}^{1/2} \right\|_{\mathrm{op}} + \lambda \left\| \tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}}) \, \mathrm{C}_{\mathcal{X}}{}^{1/2} \right\|_{\mathrm{op}} + \left\| (\widetilde{P}_X - I_{\mathcal{H}_\mathcal{X}}) \, \mathrm{C}_{\mathcal{X}}{}^{1/2} \right\|_{\mathrm{op}} \right) \tag{65}$$

But,

$$\|H_\lambda\|_{\mathrm{HS}} \leq \left\| H \left(\mathrm{C}_{\mathcal{X}} \, \mathrm{C}_{\mathcal{X}\lambda}{}^{-1} - I_{\mathcal{H}_\mathcal{X}} \right) \right\|_{\mathrm{HS}} + \|H\|_{\mathrm{HS}} \tag{66}$$

$$= \left\| H \left(\mathrm{C}_{\mathcal{X}} - \mathrm{C}_{\mathcal{X}\lambda} \right) \mathrm{C}_{\mathcal{X}\lambda}{}^{-1} \right\|_{\mathrm{HS}} + \|H\|_{\mathrm{HS}} \tag{67}$$

$$= \lambda \left\| H \, \mathrm{C}_{\mathcal{X}\lambda}{}^{-1} \right\|_{\mathrm{HS}} + \|H\|_{\mathrm{HS}} \tag{68}$$

$$\leq 2\|H\|_{\mathrm{HS}}. \tag{69}$$

And,

$$\left\| (I - \widetilde{P}_X) \, \widehat{\mathrm{C}}_{\mathrm{X}\lambda} \, \tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}}) \, \mathrm{C}_{\mathcal{X}}{}^{1/2} \right\|_{\mathrm{op}} \leq \left\| (I - \widetilde{P}_X) \, \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{1/2} \right\|_{\mathrm{op}} \left\| \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{1/2} \, \tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}}) \, \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{1/2} \right\|_{\mathrm{op}} \left\| \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{-1/2} \, \mathrm{C}_{\mathcal{X}}{}^{1/2} \right\|_{\mathrm{op}}. \tag{70}$$

And,

$$\left\| (I - \widetilde{P}_X) \, \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{1/2} \right\|_{\mathrm{op}} \leq \left\| (I - \widetilde{P}_X) C_{X\lambda}^{1/2} \right\|_{\mathrm{op}} \left\| \mathrm{C}_{\mathcal{X}\lambda}{}^{-1/2} \, \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{1/2} \right\|_{\mathrm{op}}. \tag{71}$$

And,

$$\left\| (I - \widetilde{P}_X) \, \mathrm{C}_{\mathcal{X}\lambda}{}^{1/2} \right\|_{\mathrm{op}} \leq \left\| (I - \widetilde{P}_X) \, \mathrm{C}_{\mathcal{X}}{}^{1/2} \right\|_{\mathrm{op}} + \lambda^{1/2}. \tag{72}$$

Moreover,

$$\left\| \lambda \tilde{\eta} \left(\widehat{\mathrm{C}}_{\mathrm{X}} \right) \mathrm{C}_{\mathcal{X}}{}^{1/2} \right\|_{\mathrm{op}} \leq \lambda \left\| \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{-1/2} \right\|_{\mathrm{op}} \left\| \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{1/2} \, \tilde{\eta} \left(\widehat{\mathrm{C}}_{\mathrm{X}} \right) \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{1/2} \right\|_{\mathrm{op}} \left\| \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{-1/2} \, \mathrm{C}_{\mathcal{X}\lambda}{}^{1/2} \right\|_{\mathrm{op}} \left\| \mathrm{C}_{\mathcal{X}\lambda}{}^{-1/2} \, \mathrm{C}_{\mathcal{X}}{}^{1/2} \right\|_{\mathrm{op}}$$

$$\leq \lambda^{1/2} \left\| \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{1/2} \, \tilde{\eta} \left(\widehat{\mathrm{C}}_{\mathrm{X}} \right) \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{1/2} \right\|_{\mathrm{op}} \left\| \widehat{\mathrm{C}}_{\mathrm{X}\lambda}{}^{-1/2} \, \mathrm{C}_{\mathcal{X}\lambda}{}^{1/2} \right\|_{\mathrm{op}}.$$

Conclusion. Using the probabilistic bounds given in Lemmas 9, 10, and Lemma 13, we obtain

$$(B) \leq 4\sqrt{3}\|H\|_{\mathrm{HS}} \left(\lambda^{1/2} + \left\| (I - \widetilde{P}_X) \, \mathrm{C}_{\mathcal{X}}{}^{1/2} \right\|_{\mathrm{op}} \right) \tag{73}$$

□

Lemma 5 (Bound (C))**.** *We have*

$$(C) \leq \mathbb{E}_y \left[\left\| \left(\widetilde{\mathrm{P}}_{\mathrm{Y}} - I_{\mathcal{H}_\mathcal{Y}} \right) \psi_{\mathcal{Y}}(y) \right\|_{\mathcal{H}_\mathcal{Y}}^2 \right]^{1/2} + \lambda^{1/2} \, \|H\|_{\mathrm{HS}} \, . \tag{74}$$

Proof. We have

$$(C) = \left\| (\widetilde{P}_Y H (I_{\mathcal{H}_\mathcal{X}} - \lambda \, \mathrm{C}_{\mathcal{X}\lambda}{}^{-1}) - H) \, \mathrm{C}_{\mathcal{X}}{}^{1/2} \right\|_{\mathrm{HS}} \tag{75}$$

$$\leq \left\| (\widetilde{P}_Y - I_{\mathcal{H}_\mathcal{Y}}) H \, \mathrm{C}_{\mathcal{X}}{}^{1/2} \right\|_{\mathrm{HS}} + \lambda^{1/2} \, \|H\|_{\mathrm{HS}} \tag{76}$$

$$= \mathbb{E}[\|(\widetilde{P}_Y - I_{\mathcal{H}_\mathcal{Y}}) h^*(x)\|_{\mathcal{H}_\mathcal{Y}}^2]^{1/2} + \lambda^{1/2} \, \|H\|_{\mathrm{HS}} \, . \tag{77}$$

We conclude the proof as follows. Using the fact that $h^*(x) = \mathbb{E}_{\rho(y|x)}[\psi_{\mathcal{Y}}(y)]$, the linearity of $\widetilde{\mathrm{P}}_{\mathrm{Y}} - I_{\mathcal{H}_{\mathcal{Y}}}$ and the convexity of $\|\cdot\|^2_{\mathcal{H}_{\mathcal{Y}}}$, by the Jensen's inequality we obtain that

$$\mathbb{E}_x\left[\left\|\left(\widetilde{\mathrm{P}}_{\mathrm{Y}} - I_{\mathcal{H}_{\mathcal{Y}}}\right) h^*(x)\right\|^2_{\mathcal{H}_{\mathcal{Y}}}\right] = \mathbb{E}_x\left[\left\|\left(\widetilde{\mathrm{P}}_{\mathrm{Y}} - I_{\mathcal{H}_{\mathcal{Y}}}\right) \mathbb{E}_{\rho(y|x)}[\psi_{\mathcal{Y}}(y)]\right\|^2_{\mathcal{H}_{\mathcal{Y}}}\right] \tag{78}$$

$$= \mathbb{E}_x\left[\left\|\mathbb{E}_{\rho(y|x)}\left[\left(\widetilde{\mathrm{P}}_{\mathrm{Y}} - I_{\mathcal{H}_{\mathcal{Y}}}\right)\psi_{\mathcal{Y}}(y)\right]\right\|^2_{\mathcal{H}_{\mathcal{Y}}}\right] \tag{79}$$

$$\leq \mathbb{E}_x\left[\mathbb{E}_{\rho(y|x)}\left[\left\|\left(\widetilde{\mathrm{P}}_{\mathrm{Y}} - I_{\mathcal{H}_{\mathcal{Y}}}\right)\psi_{\mathcal{Y}}(y)\right\|^2_{\mathcal{H}_{\mathcal{Y}}}\right]\right] \tag{80}$$

$$= \mathbb{E}_y\left[\left\|\left(\widetilde{\mathrm{P}}_{\mathrm{Y}} - I_{\mathcal{H}_{\mathcal{Y}}}\right)\psi_{\mathcal{Y}}(y)\right\|^2_{\mathcal{H}_{\mathcal{Y}}}\right]. \tag{81}$$

□

E SKETCHING RECONSTRUCTION ERROR

We provide here a bound on the reconstruction error of a sketching approximation.

Theorem 2 (sub-Gaussian sketching reconstruction error)**.** *For $\delta \in (0, 1/e]$, $n \in \mathbb{N}$ sufficiently large such that $\frac{9}{n}\log(n/\delta) \leq n^{-\frac{1}{1+\gamma_{\mathcal{Z}}}} \leq \|\mathrm{C}_{\mathcal{Z}}\|_{\mathrm{op}}/2$, then if*

$$\mathrm{m}_{\mathcal{Z}} \geq c_4 \max\left(\nu^2_{\mathcal{Z}}\, n^{\frac{\gamma_{\mathcal{Z}}+\mu_{\mathcal{Z}}}{1+\gamma_{\mathcal{Z}}}}, \nu^4_{\mathcal{Z}}\log(1/\delta)\right), \tag{10}$$

with probability $1-\delta$ we have

$$\mathbb{E}_z\left[\|(\widetilde{\mathrm{P}}_{\mathrm{Z}} - I_{\mathcal{H}_{\mathcal{Z}}})\,\psi_{\mathcal{Z}}(z)\|^2_{\mathcal{H}_{\mathcal{Z}}}\right] \leq c_3\, n^{-\frac{1-\gamma_{\mathcal{Z}}}{1+\gamma_{\mathcal{Z}}}}, \tag{11}$$

where $c_3, c_4 > 0$ are constants independents of $n, \mathrm{m}_{\mathcal{Z}}, \delta$.

Proof. For $t > 0$, we have

$$\mathbb{E}_z\left[\left\|\left(\widetilde{\mathrm{P}}_{\mathrm{Z}} - I_{\mathcal{H}_{\mathcal{Z}}}\right)\psi_{\mathcal{Z}}(z)\right\|^2_{\mathcal{H}_{\mathcal{Z}}}\right] = \mathrm{Tr}\left(\left(\widetilde{\mathrm{P}}_{\mathrm{Z}} - I_{\mathcal{H}_{\mathcal{Z}}}\right)\mathbb{E}_z\left[\psi_{\mathcal{Z}}(z)\otimes\psi_{\mathcal{Z}}(z)\right]\right) \tag{82}$$

$$= \left\|\left(\widetilde{\mathrm{P}}_{\mathrm{Z}} - I_{\mathcal{H}_{\mathcal{Z}}}\right){\mathrm{C}_{\mathcal{Z}}}^{1/2}\right\|^2_{\mathrm{HS}} \tag{83}$$

$$\leq \left\|\left(\widetilde{\mathrm{P}}_{\mathrm{Z}} - I_{\mathcal{H}_{\mathcal{Z}}}\right){\widehat{\mathrm{C}}_{\mathrm{Zt}}}^{1/2}\right\|^2_{\mathrm{op}} \left\|{\widehat{\mathrm{C}}_{\mathrm{Zt}}}^{-1/2}\,{\mathrm{C}_{\mathcal{Z}\mathrm{t}}}^{1/2}\right\|^2_{\mathrm{op}} \left\|{\mathrm{C}_{\mathcal{Z}\mathrm{t}}}^{-1/2}\,{\mathrm{C}_{\mathcal{Z}}}^{1/2}\right\|^2_{\mathrm{HS}}. \tag{84}$$

Lemma 9 gives that, for $\delta \in (0,1)$, if $\frac{9}{n}\log\left(\frac{n}{\delta}\right) \leq t \leq \|\mathrm{C}_{\mathcal{Z}}\|_{\mathrm{op}}$, then with probability $1-\delta$

$$\left\|{\widehat{\mathrm{C}}_{\mathrm{Zt}}}^{-1/2}\,{\mathrm{C}_{\mathcal{Z}\mathrm{t}}}^{1/2}\right\|^2_{\mathrm{op}} \leq 2\,. \tag{85}$$

Moreover, since $\left\|{\mathrm{C}_{\mathcal{Z}\mathrm{t}}}^{-1/2}\,{\mathrm{C}_{\mathcal{Z}}}^{1/2}\right\|^2_{\mathrm{HS}} = \mathrm{Tr}\left({\mathrm{C}_{\mathcal{Z}\mathrm{t}}}^{-1}\mathrm{C}_{\mathcal{Z}}\right) = \mathrm{d}^{\mathcal{Z}}_{\mathrm{eff}}(t)$, Lemma 11 gives that

$$\left\|{\mathrm{C}_{\mathcal{Z}\mathrm{t}}}^{-1/2}\,{\mathrm{C}_{\mathcal{Z}}}^{1/2}\right\|^2_{\mathrm{HS}} \leq \mathrm{Q}_{\mathcal{Z}}\, t^{-\gamma_{\mathcal{Z}}}. \tag{86}$$

Then, using the Lemma 6, and multiplying the bounds, gives

$$\mathbb{E}_y\left[\left\|\left(\widetilde{\mathrm{P}}_{\mathrm{Z}} - I_{\mathcal{H}_{\mathcal{Y}}}\right)\psi_{\mathcal{Z}}(z)\right\|^2_{\mathcal{H}_{\mathcal{Z}}}\right] \leq 6\,\mathrm{Q}_{\mathcal{Z}}\, t^{1-\gamma_{\mathcal{Z}}}. \tag{87}$$

Finally, choosing $t = n^{-\frac{1}{1+\gamma_{\mathcal{Z}}}}$, defining $c_3 = 6\,\mathrm{Q}_{\mathcal{Z}}$, $c_4 = 576\mathfrak{C}^2\,\mathrm{b}_{\mathcal{Z}}\,\mathrm{Q}_{\mathcal{Z}}$, and noticing $\mathcal{N}^{\infty}_{\mathcal{Z}}(t) \leq \mathrm{b}_{\mathcal{Z}}\,\mathrm{Q}_{\mathcal{Z}}\, t^{-(\gamma_{\mathcal{Z}}+\mu_{\mathcal{Z}})}$ (from Lemmas 11 and 12), allows to conclude the proof.

□

Lemma 6. *Let $\mathcal{N}_{\mathcal{Z}}^{\infty}(t)$ be as in Definition 3. For all $\delta \in (0, 1/e]$, $\frac{9}{n}\log\left(\frac{n}{\delta}\right) \leq t \leq \|\mathrm{C}_{\mathcal{Z}}\|_{\mathrm{op}} - \frac{9}{n}\log\left(\frac{n}{\delta}\right)$ and $\mathrm{m}_{\mathcal{Z}} \geq \max\left(432\mathfrak{C}^2\,{\nu_{\mathcal{Z}}}^2\,\mathcal{N}_{\mathcal{Z}}^{\infty}(t), 576\mathfrak{C}^2\,{\nu_{\mathcal{Z}}}^4\log\left(1/\delta\right)\right)$, with probability at least $1-\delta$,*

$$\left\|\left(\widetilde{\mathrm{P}}_{\mathrm{Z}} - I_{\mathcal{H}_{\mathcal{Z}}}\right)\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{1/2}\right\|_{\mathrm{op}}^2 \leq 3t\,. \tag{88}$$

Proof. Using Propositions 3 and 7 from Rudi et al. (2015), we have, for $t > 0$,

$$\left\|\left(\widetilde{\mathrm{P}}_{\mathrm{Z}} - I_{\mathcal{H}_{\mathcal{Z}}}\right)\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{1/2}\right\|_{\mathrm{op}}^2 \leq \frac{t}{1-\beta_{\mathcal{Z}}(t)}\,, \tag{89}$$

with $\beta_{\mathcal{Z}}(t) = \sigma_{\max}\left(\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\left(\widehat{\mathrm{C}}_{\mathrm{Z}} - \widetilde{\mathrm{C}}_{\mathrm{Z}}\right)\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\right)$.

Now, applying Lemma 7, with the condition

$$\mathrm{m}_{\mathcal{Z}} \geq \max\left(432\mathfrak{C}^2\,{\nu_{\mathcal{Z}}}^2\,\mathcal{N}_{\mathcal{Z}}^{\infty}(t), 576\mathfrak{C}^2\,{\nu_{\mathcal{Z}}}^4\log\left(1/\delta\right)\right), \tag{90}$$

we obtain $\beta_{\mathcal{Z}}(t) \leq 2/3$, which gives

$$\left\|\left(\widetilde{\mathrm{P}}_{\mathrm{Z}} - I_{\mathcal{H}_{\mathcal{Y}}}\right)\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{1/2}\right\|_{\mathrm{op}}^2 \leq 3t\,. \tag{91}$$

□

Lemma 7. *Let $\mathrm{R}_{\mathcal{Z}}$ be as in Definition 1 and $\mathcal{N}_{\mathcal{Z}}^{\infty}(t)$ as in Definition 3. For all $\delta \in (0, 1/e]$, $\frac{9}{n}\log\left(\frac{n}{\delta}\right) \leq t \leq \|\mathrm{C}_{\mathcal{Z}}\|_{\mathrm{op}} - \frac{9}{n}\log\left(\frac{n}{\delta}\right)$ and $\mathrm{m}_{\mathcal{Z}} \geq \max\left(6\mathcal{N}_{\mathcal{Z}}^{\infty}(t), \log\left(1/\delta\right)\right)$, with probability at least $1-\delta$,*

$$\left\|\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\left(\widehat{\mathrm{C}}_{\mathrm{Z}} - \widetilde{\mathrm{C}}_{\mathrm{Z}}\right)\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\right\|_{\mathrm{op}} \leq \mathfrak{C}\frac{2\sqrt{2}\,\nu_{\mathcal{Z}}\sqrt{6\mathcal{N}_{\mathcal{Z}}^{\infty}(t)} + 8\,{\nu_{\mathcal{Z}}}^2\sqrt{\log\left(1/\delta\right)}}{\sqrt{\mathrm{m}_{\mathcal{Z}}}}\,, \tag{92}$$

where $\mathfrak{C}$ is a universal constant independent of $\mathcal{N}_{\mathcal{Z}}^{\infty}(t)$, δ and $\mathrm{m}_{\mathcal{Z}}$.

Proof. We define the following random variables

$$W_i = \sqrt{\frac{\mathrm{m}_{\mathcal{Z}}}{n}}\sum_{j=1}^{n}(\mathrm{R}_{\mathcal{Z}})_{ij}\,\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\,\psi_{\mathcal{Z}}(z_j) \in \mathcal{H}_{\mathcal{Z}} \qquad \text{for } i = 1, \dots \mathrm{m}_{\mathcal{Z}}\,. \tag{93}$$

In order to use the concentration bound given in Theorem 4, we show that the W_is are i.i.d. weakly square integrable centered random vectors with covariance operator Σ, sub-Gaussian, and pre-Gaussian.

The W_is are weakly square integrable. Let $u \in \mathcal{H}_{\mathcal{Z}}$ and $v = \widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\,u$, we have that $\langle W_i, u\rangle_{\mathcal{H}_{\mathcal{Z}}} = \sqrt{\frac{\mathrm{m}_{\mathcal{Z}}}{n}}\sum_{j=1}^{n}(\mathrm{R}_{\mathcal{Z}})_{ij}v(z_j)$. Hence, using the definition of a sub-Gaussian sketch, we have

$$\|\langle W_i, u\rangle_{\mathcal{H}_{\mathcal{Z}}}\|_{L_2(\mathbb{P})}^2 = \mathbb{E}_{\mathrm{R}_{\mathcal{Z}}}\left[|\langle W_i, u\rangle_{\mathcal{H}_{\mathcal{Z}}}|^2\right] \tag{94}$$

$$= \frac{1}{n}\sum_{j=1}^{n} v(z_j)^2 \tag{95}$$

$$< +\infty\,. \tag{96}$$

The W_is are sub-Gaussian. Let $c \in \mathbb{R}$, using the independence and sub-Gaussianity of the $R_{z_{ij}}$, we have

$$\mathbb{E}_{\mathrm{R}_\mathcal{Z}}\left[\exp\left(c\langle W_i, u\rangle_{\mathcal{H}_\mathcal{Z}}\right)\right] = \mathbb{E}_{\mathrm{R}_\mathcal{Z}}\left[\exp\left(\sum_{j=1}^{n} c\sqrt{\frac{\mathrm{m}_\mathcal{Z}}{n}} R_{z_{ij}} v(z_j)\right)\right] \tag{97}$$

$$= \prod_{j=1}^{n} \mathbb{E}_{\mathrm{R}_\mathcal{Z}}\left[\exp\left(c\sqrt{\frac{\mathrm{m}_\mathcal{Z}}{n}} R_{z_{ij}} v(z_j)\right)\right] \tag{98}$$

$$\leq \prod_{j=1}^{n} \exp\left(\frac{c^2\,\mathrm{m}_\mathcal{Z}\, v(z_j)^2}{2n} \frac{{\nu_\mathcal{Z}}^2}{\mathrm{m}_\mathcal{Z}}\right) \tag{99}$$

$$= \exp\left(\frac{c^2\,{\nu_\mathcal{Z}}^2}{2n} \sum_{j=1}^{n} v(z_j)^2\right) \tag{100}$$

$$= \exp\left(\frac{c^2\,{\nu_\mathcal{Z}}^2}{2} \left\|\langle W_i, u\rangle_{\mathcal{H}_\mathcal{Z}}\right\|^2_{L_2(\mathbb{P})}\right). \tag{101}$$

Hence, $\langle W_i, u\rangle_{\mathcal{H}_\mathcal{Z}}$ is a $\frac{1}{2}\,{\nu_\mathcal{Z}}^2 \left\|\langle W_i, u\rangle_{\mathcal{H}_\mathcal{Z}}\right\|^2_{L_2(\mathbb{P})}$-sub-Gaussian random variable. Then, the Orlicz condition of sub-Gaussian random variables gives

$$\mathbb{E}_{\mathrm{R}_\mathcal{Z}}\left[\exp\left(\frac{\langle W_i, u\rangle^2_{\mathcal{H}_\mathcal{Z}}}{8\,{\nu_\mathcal{Z}}^2 \left\|\langle W_i, u\rangle_{\mathcal{H}_\mathcal{Z}}\right\|^2_{L_2(\mathbb{P})}}\right) - 1\right] \leq 1\,. \tag{102}$$

We deduce that

$$\left\|\langle W_i, u\rangle_{\mathcal{H}_\mathcal{Z}}\right\|_{\varphi_2} \leq 2\sqrt{2}\,\nu_\mathcal{Z} \left\|\langle W_i, u\rangle_{\mathcal{H}_\mathcal{Z}}\right\|_{L_2(\mathbb{P})}\,. \tag{103}$$

We conclude that the W_is are sub-Gaussian with $B = 2\sqrt{2}\,\nu_\mathcal{Z}$.

The W_is are pre-gaussian. We define $Z = \sqrt{\frac{\mathrm{m}_\mathcal{Z}}{n}} \sum_{j=1}^{n} G_j\, {\widehat{\mathrm{C}}_{\mathrm{Zt}}}^{-1/2}\, \psi_\mathcal{Z}(z_j)$, with $G_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/\mathrm{m}_\mathcal{Z})$. Z is a Gaussian random variable that admits the same covariance operator as the W_is. So, the W_i are pre-Gaussian.

Applying concentration bound. Because the W_is are i.i.d. weakly square integrable centered random variables, we can apply Theorem 4, and by using also Lemma 14, and the condition $\mathrm{m}_\mathcal{Z} \geq \max\left(6\mathcal{N}^\infty_\mathcal{Z}(t), \log(1/\delta)\right)$, we obtain

$$\left\|{\widehat{\mathrm{C}}_{\mathrm{Zt}}}^{-1/2}\left(\widehat{\mathrm{C}}_{\mathrm{Z}} - \widetilde{\mathrm{C}}_{\mathrm{Z}}\right){\widehat{\mathrm{C}}_{\mathrm{Zt}}}^{-1/2}\right\|_{\mathrm{op}} \leq \mathfrak{C}\frac{2\sqrt{2}\,\nu_\mathcal{Z}\sqrt{6\mathcal{N}^\infty_\mathcal{Z}(t)} + 8\,{\nu_\mathcal{Z}}^2\sqrt{\log(1/\delta)}}{\sqrt{\mathrm{m}_\mathcal{Z}}}\,. \tag{104}$$

□

F PROBABILISTIC BOUNDS

In this section, we provide all the probabilistic bounds used in our proofs. In particular, we restate bounds from other works for the sake of providing a self-contained work. We order them in the same in order of appearance in our proofs.

Lemma 8 (Bound $(A.1) = \left\|\left(\mathrm{S_Y}^{\#}\,\mathrm{S_X} - H_\lambda\,\widehat{\mathrm{C}}_{\mathrm{X}}\right)\mathrm{C}_{\mathcal{X}\lambda}{}^{-1/2}\right\|_{\mathrm{HS}}$ (Ciliberto et al., 2020, Theorem B.10)). *Let $\delta \in [0,1]$, $n \in \mathbb{N}$ sufficiently large such that $\lambda = n^{-1/(1+\gamma_\mathcal{X})} \geq \frac{9\,{\kappa_\mathcal{X}}^2}{n}\log(\frac{n}{x})$ Under our set of assumptions, the following holds with probability at least $1-\delta$*

$$(A.1) \leq M\log(4/\delta)n^{-\frac{1}{2(1+\gamma_\mathcal{X})}} \tag{105}$$

where the constant M depends on $\kappa_\mathcal{Y}, \|H\|_{\mathrm{HS}}, \delta$.

Proof. This lemma can be obtained from Ciliberto et al. (2020, Theorem B.10), by noticing that the bound of Theorem B.10 is obtained by upper bounding the sum of $(A.1)$ and a positive term, such that the bound of Ciliberto et al. (2020, Theorem B.10) is an upper bound of $(A.1)$.

Lemma 9 (Bound $\left\| \widehat{\mathrm{C}}_{\mathrm{Z}\lambda}{}^{-1/2} \, \mathrm{C}_{\mathcal{Z}\lambda}{}^{1/2} \right\|_{\mathrm{op}}$ (Rudi et al., 2013, Lemma 3.6)). *If* $\frac{9}{n} \log \frac{n}{\delta} \leq \lambda \leq \| \mathrm{C}_{\mathcal{Z}} \|_{\mathrm{op}}$, *then we have with probability* $1 - \delta$

$$\| \widehat{\mathrm{C}}_{\mathrm{Z}\lambda}{}^{-1/2} \, \mathrm{C}_{\mathcal{Z}\lambda}{}^{1/2} \|_{\mathrm{op}} \leq \sqrt{2}. \tag{106}$$

□

Lemma 10 (Bound $\left\| \mathrm{C}_{\mathcal{Z}\lambda}{}^{-1/2} \, \widehat{\mathrm{C}}_{\mathrm{Z}\lambda}{}^{1/2} \right\|_{\mathrm{op}}$). *If* $\frac{9}{n} \log \frac{n}{\delta} \leq \lambda \leq \| \mathrm{C}_{\mathcal{Z}} \|_{\mathrm{op}}$, *then with probability* $1 - \delta$

$$\left\| \mathrm{C}_{\mathcal{Z}\lambda}{}^{-1/2} \, \widehat{\mathrm{C}}_{\mathrm{Z}\lambda}{}^{1/2} \right\|_{\mathrm{op}} \leq \sqrt{\frac{3}{2}}. \tag{107}$$

Proof. We have

$$\begin{aligned}
\left\| \mathrm{C}_{\mathcal{Z}\lambda}{}^{-1/2} \, \widehat{\mathrm{C}}_{\mathrm{Z}\lambda}{}^{1/2} \right\|_{\mathrm{op}} &= \left\| \mathrm{C}_{\mathcal{Z}\lambda}{}^{-1/2} \, \widehat{\mathrm{C}}_{\mathrm{Z}\lambda} \, \mathrm{C}_{\mathcal{Z}\lambda}{}^{-1/2} \right\|_{\mathrm{op}}^{1/2} && (108) \\
&= \left\| I + \mathrm{C}_{\mathcal{Z}\lambda}{}^{-1/2} (\widehat{\mathrm{C}}_{\mathrm{Z}} - \mathrm{C}_{\mathcal{Z}}) \, \mathrm{C}_{\mathcal{Z}\lambda}{}^{-1/2} \right\|_{\mathrm{op}}^{1/2} && (109) \\
&\leq \left(1 + \left\| \mathrm{C}_{\mathcal{Z}\lambda}{}^{-1/2} (\widehat{\mathrm{C}}_{\mathrm{Z}} - \mathrm{C}_{\mathcal{Z}}) \, \mathrm{C}_{\mathcal{Z}\lambda}{}^{-1/2} \right\|_{\mathrm{op}} \right)^{1/2} && (110) \\
&\leq \sqrt{\frac{3}{2}} && (111)
\end{aligned}$$

with probability at least $1 - \delta$, where the last inequality is from Rudi et al. (2013, Lemma 3.6).

Theorem 4 (sub-Gaussian concentration bound (Koltchinskii and Lounici, 2017, Theorem 9)). *Let* $W, W_1, \ldots, W_m$ *be i.i.d. weakly square integrable centered random vectors in a separable Hilbert space* $\mathcal{H}_{\mathcal{Z}}$ *with covariance operator* Σ. *If* W *is sub-Gaussian and pre-Gaussian, then there exists a constant* $\mathfrak{C} > 0$ *such that, for all* $\tau \geq 1$, *with probability at least* $1 - e^{-\tau}$,

$$\|\hat{\Sigma} - \Sigma\| \leq \mathfrak{C} \|\Sigma\| \left(B \sqrt{\frac{\mathbf{r}(\Sigma)}{m}} \vee \frac{\mathbf{r}(\Sigma)}{m} \vee B^2 \sqrt{\frac{\tau}{m}} \vee B^2 \frac{\tau}{m} \right), \tag{112}$$

where $B > 0$ *is the constant such that* $\| \langle W, u \rangle_{\mathcal{H}_{\mathcal{Y}}} \|_{\varphi_2} \leq B \, \| \langle W, u \rangle_{\mathcal{H}_{\mathcal{Y}}} \|_{L_2(\mathbb{P})}$ *for all* $u \in \mathcal{H}_{\mathcal{Z}}$.

□

G AUXILIARY RESULTS AND DEFINITIONS

Definition 3. *For* $t > 0$, *we define the random variable*

$$\mathcal{N}(z, t) = \langle \psi_{\mathcal{Z}}(z), \mathrm{C}_{\mathcal{Z}\mathrm{t}}{}^{-1} \, \psi_{\mathcal{Z}}(z) \rangle_{\mathcal{H}_{\mathcal{Z}}} \tag{113}$$

with $z \in \mathcal{Z}$ *distributed according to* $\rho_{\mathcal{Z}}$ *and let*

$$\mathrm{d}_{\mathrm{eff}}^{\mathcal{Z}}(t) = \mathbb{E}_z \left[\mathcal{N}(z, t) \right] = \mathrm{Tr} \left(\mathrm{C}_{\mathcal{Z}} \, \mathrm{C}_{\mathcal{Z}\mathrm{t}}{}^{-1} \right), \quad \mathcal{N}_{\mathcal{Z}}^{\infty}(t) = \sup_{z \in \mathcal{Z}} \mathcal{N}(z, t). \tag{114}$$

We note $\mathcal{N}_{\mathcal{X}}^{\infty}, \mathrm{d}_{\mathrm{eff}}^{\mathcal{X}}(t), \gamma_{\mathcal{X}}, \mathrm{Q}_{\mathcal{Y}}, \mathcal{N}_{\mathcal{Y}}^{\infty}, \mathrm{d}_{\mathrm{eff}}^{\mathcal{Y}}(t), \gamma_{\mathcal{Y}}, \mathrm{Q}_{\mathcal{Y}}$ *for the input and output kernels* $\mathrm{k}_{\mathcal{X}}, k_y$, *respectively.*

Lemma 11. *When Assumption 3 holds then we have*

$$\mathrm{d}_{\mathrm{eff}}^{\mathcal{Z}}(t) \leq \mathrm{Q}_{\mathcal{Z}} \, t^{-\gamma_{\mathcal{Z}}}. \tag{115}$$

Proof. We have

$$\mathrm{d}_{\mathrm{eff}}^{\mathcal{Z}}(t) = \mathrm{Tr}\left(\mathrm{C}_{\mathcal{Z}}\,{\mathrm{C}_{\mathcal{Z}\mathrm{t}}}^{-1}\right) \tag{116}$$
$$\leq \mathrm{Tr}\left({\mathrm{C}_{\mathcal{Z}}}^{\gamma_{\mathcal{Z}}}\right) \| \, {\mathrm{C}_{\mathcal{Z}}}^{1-\gamma_{\mathcal{Z}}}\,{\mathrm{C}_{\mathcal{Z}\mathrm{t}}}^{-1} \, \|_{\mathrm{op}} \tag{117}$$
$$\leq \mathrm{Q}_{\mathcal{Z}}\, t^{-\gamma_{\mathcal{Z}}}. \tag{118}$$

□

Lemma 12. *When Assumption 4 holds then we have*

$$\mathcal{N}_{\mathcal{Z}}^{\infty}(t) \leq \mathrm{b}_{\mathcal{Z}}\, \mathrm{d}_{\mathrm{eff}}^{\mathcal{Z}}(t) t^{-\mu_{\mathcal{Z}}}. \tag{119}$$

Proof. We have

$$\mathcal{N}_{\mathcal{Z}}^{\infty}(t) = \sup_{z\in\mathcal{Z}} \langle \psi_{\mathcal{Z}}(z), {\mathrm{C}_{\mathcal{Z}\mathrm{t}}}^{-1}\,\psi_{\mathcal{Z}}(z)\rangle_{\mathcal{H}_{\mathcal{Z}}} \tag{120}$$
$$\leq \mathrm{b}_{\mathcal{Z}}\,\mathrm{Tr}({\mathrm{C}_{\mathcal{Z}\mathrm{t}}}^{-1}\,{\mathrm{C}_{\mathcal{Z}}}^{1-\mu_{\mathcal{Z}}}) \tag{121}$$
$$\leq \mathrm{b}_{\mathcal{Z}}\,\mathrm{Tr}({\mathrm{C}_{\mathcal{Z}\mathrm{t}}}^{-1}\,\mathrm{C}_{\mathcal{Z}})\| \, {\mathrm{C}_{\mathcal{Z}\mathrm{t}}}^{-\mu_{\mathcal{Z}}} \, \|_{\mathrm{op}} \tag{122}$$
$$\leq \mathrm{b}_{\mathcal{Z}}\,\mathrm{d}_{\mathrm{eff}}^{\mathcal{Z}}(t) t^{-\mu_{\mathcal{Z}}}. \tag{123}$$

□

We recall the following deterministic bound.

Lemma 13 (Bound $\| \, {\widehat{\mathrm{C}}_{\mathrm{X}\lambda}}^{1/2}\,\tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}})\,{\widehat{\mathrm{C}}_{\mathrm{X}\lambda}}^{1/2} \, \|_{\mathrm{op}}$ (Rudi et al., 2015, Lemma 8))**.** *For any $\lambda > 0$,*

$$\| \, {\widehat{\mathrm{C}}_{\mathrm{X}\lambda}}^{1/2}\,\tilde{\eta}(\widehat{\mathrm{C}}_{\mathrm{X}})\,{\widehat{\mathrm{C}}_{\mathrm{X}\lambda}}^{1/2} \, \|_{\mathrm{op}} \leq 1. \tag{124}$$

We introduce here some notations and definitions from Koltchinskii and Lounici (2017). Let W be a centered random variable in $\mathcal{H}_{\mathcal{Z}}$, W is weakly square integrable iff $\|\langle W, u\rangle_{\mathcal{H}_{\mathcal{Z}}}\|_{L_2(\mathbb{P})}^2 := \mathbb{E}\left[|\langle W, u\rangle_{\mathcal{H}_{\mathcal{Z}}}|^2\right] < +\infty$, for any $u \in \mathcal{H}_{\mathcal{Z}}$. Moreover, we define the Orlicz norms. For a convex nondecreasing function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\varphi(0) = 0$ and a random variable η on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, the φ-norm of η is defined as

$$\|\eta\|_{\varphi} = \inf\left\{C > 0 : \mathbb{E}\left[\varphi\left(|\eta|/C\right)\right] \leq 1\right\}. \tag{125}$$

The Orlicz φ_1- and φ_2-norms coincide to the functions $\varphi_1(u) = e^u - 1, u \geq 0$ and $\varphi_2(u) = e^{u^2} - 1, u \geq 0$. Finally, Koltchinskii and Lounici (2017) introduces the definitions of sub-Gaussian and pre-Gaussian random variables in a separable Banach space E. We focus on the case where $E = \mathcal{H}_{\mathcal{Z}}$.

Definition 4. *A centered random variable X in $\mathcal{H}_{\mathcal{Z}}$ will be called sub-Gaussian iff, for all $u \in \mathcal{H}_{\mathcal{Z}}$, there exists $B > 0$ such that*

$$\|\langle X, u\rangle_{\mathcal{H}_{\mathcal{Z}}}\|_{\psi_{\mathcal{X}_2}} \leq B\,\|\langle X, u\rangle_{\mathcal{H}_{\mathcal{Z}}}\|_{L_2(\mathbb{P})}. \tag{126}$$

Definition 5. *A weakly square integrable centered random variable X in $\mathcal{H}_{\mathcal{Z}}$ with covariance operator Σ is called pre-Gaussian iff there exists a centered Gaussian random variable Y in $\mathcal{H}_{\mathcal{Z}}$ with the same covariance operator Σ.*

Lemma 14 (Expectancy, covariance, and intrinsic dimension of the W_is)**.** *Defining $W_i = \sqrt{\frac{\mathrm{m}_{\mathcal{Z}}}{n}}\sum_{j=1}^{n}(\mathrm{R}_{\mathcal{Z}})_{ij}\,{\widehat{\mathrm{C}}_{\mathrm{Z}\mathrm{t}}}^{-1/2}\,\psi_{\mathcal{Z}}(z_j) \in \mathcal{H}_{\mathcal{Z}}$ for $i = 1, \dots \mathrm{m}_{\mathcal{Z}}$ where $\mathrm{R}_{\mathcal{Z}}$ is a sub-Gaussian sketch, the following hold true*

$$\mathbb{E}_{\mathrm{R}_{\mathcal{Z}}}\left[W_i\right] = 0 \tag{127}$$
$$\Sigma = \mathbb{E}_{\mathrm{R}_{\mathcal{Z}}}\left[W_i \otimes W_i\right] = {\widehat{\mathrm{C}}_{\mathrm{Z}\mathrm{t}}}^{-1/2}\,\widehat{\mathrm{C}}_{\mathrm{Z}}\,{\widehat{\mathrm{C}}_{\mathrm{Z}\mathrm{t}}}^{-1/2} \tag{128}$$
$$\widehat{\Sigma} = \frac{1}{\mathrm{m}_{\mathcal{Z}}}\sum_{i=1}^{\mathrm{m}_{\mathcal{Z}}}\langle f, W_i\rangle_{\mathcal{H}_{\mathcal{Z}}} W_i = {\widehat{\mathrm{C}}_{\mathrm{Z}\mathrm{t}}}^{-1/2}\,\widetilde{\mathrm{C}}_{\mathrm{Z}}\,{\widehat{\mathrm{C}}_{\mathrm{Z}\mathrm{t}}}^{-1/2} \tag{129}$$

and for $\delta \in (0,1)$*, if* $\frac{9}{n}\log\left(\frac{n}{\delta}\right) \leq t \leq \|\mathrm{C}_{\mathcal{Z}}\|_{\mathrm{op}} - \frac{9}{n}\log\left(\frac{n}{\delta}\right)$*, then with probability* $1-\delta$

$$r\left(\Sigma\right) = \frac{\mathbb{E}_{\mathrm{R}_{\mathcal{Z}}}\left[\|X_i\|_{\mathcal{H}_{\mathcal{Z}}}\right]^2}{\|\Sigma\|_{\mathrm{op}}} \leq 6\mathcal{N}_{\mathcal{Z}}^{\infty}(t)\,. \tag{130}$$

Proof. First, it is straightforward to check that

$$\frac{1}{\mathrm{m}_{\mathcal{Z}}}\sum_{i=1}^{\mathrm{m}_{\mathcal{Z}}}\langle f, W_i\rangle_{\mathcal{H}_{\mathcal{Z}}} W_i = \widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\,\widetilde{\mathrm{C}}_{\mathrm{Z}}\,\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\,. \tag{131}$$

Then, since $\mathbb{E}_{\mathrm{R}_{\mathcal{Z}}}\left[(\mathrm{R}_{\mathcal{Z}})_{i:}\right] = 0$,

$$\mathbb{E}_{\mathrm{R}_{\mathcal{Z}}}\left[W_i\right] = \sqrt{\frac{\mathrm{m}_{\mathcal{Z}}}{n}}\,\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\,\mathrm{S}_{\mathrm{Z}}{}^{\#}\,\mathbb{E}_{\mathrm{R}_{\mathcal{Z}}}\left[(\mathrm{R}_{\mathcal{Z}})_{i:}\right] = 0\,. \tag{132}$$

Then,

$$\begin{aligned}
\left(W_i \otimes W_i\right) f &= \langle f, W_i\rangle_{\mathcal{H}_{\mathcal{Z}}} W_i && (133)\\
&= \langle f, \sqrt{\mathrm{m}_{\mathcal{Z}}}\,\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\,\mathrm{S}_{\mathrm{Z}}{}^{\#}(\mathrm{R}_{\mathcal{Z}})_{i:}\rangle_{\mathcal{H}_{\mathcal{Z}}}\sqrt{\mathrm{m}_{\mathcal{Z}}}\,\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\,\mathrm{S}_{\mathrm{Z}}{}^{\#}(\mathrm{R}_{\mathcal{Z}})_{i:} && (134)\\
&= \mathrm{m}_{\mathcal{Z}}\left((\mathrm{R}_{\mathcal{Z}})_{i:}^{\top}\,\mathrm{S}_{\mathrm{Z}}\,\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\,f\right)\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\,\mathrm{S}_{\mathrm{Z}}{}^{\#}(\mathrm{R}_{\mathcal{Z}})_{i:} && (135)\\
&= \widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\,\mathrm{S}_{\mathrm{Z}}{}^{\#}\left(\mathrm{m}_{\mathcal{Z}}(\mathrm{R}_{\mathcal{Z}})_{i:}(\mathrm{R}_{\mathcal{Z}})_{i:}^{\top}\right)\mathrm{S}_{\mathrm{Z}}\,\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\,f\,, && (136)
\end{aligned}$$

and since $\mathbb{E}_{\mathrm{R}_{\mathcal{Z}}}\left[\mathrm{m}_{\mathcal{Z}}(\mathrm{R}_{\mathcal{Z}})_{i:}(\mathrm{R}_{\mathcal{Z}})_{i:}^{\top}\right] = I_n$,

$$\begin{aligned}
\Sigma &= \mathbb{E}_{\mathrm{R}_{\mathcal{Z}}}\left[W_i \otimes W_i\right] && (137)\\
&= \widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\,\mathrm{S}_{\mathrm{Z}}{}^{\#}\,\mathbb{E}_{\mathrm{R}_{\mathcal{Z}}}\left[\mathrm{m}_{\mathcal{Z}}(\mathrm{R}_{\mathcal{Z}})_{i:}(\mathrm{R}_{\mathcal{Z}})_{i:}^{\top}\right]\mathrm{S}_{\mathrm{Z}}\,\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2} && (138)\\
&= \widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\,\widehat{\mathrm{C}}_{\mathrm{Z}}\,\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\,. && (139)
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbb{E}_{\mathrm{R}_{\mathcal{Z}}}\left[\|X_i\|_{\mathcal{H}_{\mathcal{Z}}}\right]^2 &\leq \mathbb{E}_{\mathrm{R}_{\mathcal{Z}}}\left[\|X_i\|_{\mathcal{H}_{\mathcal{Z}}}^2\right] \quad \text{(by Jensen's inequality)} && (140)\\
&= \mathrm{m}_{\mathcal{Z}}\,\mathbb{E}_{\mathrm{R}_{\mathcal{Z}}}\left[\langle\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\,\mathrm{S}_{\mathrm{Z}}{}^{\#}(\mathrm{R}_{\mathcal{Z}})_{i:}, \widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\,\mathrm{S}_{\mathrm{Z}}{}^{\#}(\mathrm{R}_{\mathcal{Z}})_{i:}\rangle_{\mathcal{H}_{\mathcal{Z}}}\right] && (141)\\
&= \frac{\mathrm{m}_{\mathcal{Z}}}{n}\,\mathbb{E}_{\mathrm{R}_{\mathcal{Z}}}\left[\langle\sum_{j=1}^{n}\mathrm{R}_{\mathcal{Z}_{\mathrm{ij}}}\,\psi_{\mathcal{Z}}(z_j), \sum_{l=1}^{n}\mathrm{R}_{\mathcal{Z}_{\mathrm{il}}}\,\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1}\,\psi_{\mathcal{Z}}(z_l)\rangle_{\mathcal{H}_{\mathcal{Z}}}\right] && (142)\\
&= \frac{\mathrm{m}_{\mathcal{Z}}}{n}\,\mathbb{E}_{\mathrm{R}_{\mathcal{Z}}}\left[\sum_{j,l=1}^{n}\mathrm{R}_{\mathcal{Z}_{\mathrm{ij}}}\,\mathrm{R}_{\mathcal{Z}_{\mathrm{il}}}\langle\psi_{\mathcal{Z}}(z_j), \widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1}\,\psi_{\mathcal{Z}}(z_l)\rangle_{\mathcal{H}_{\mathcal{Y}}}\right] && (143)\\
&= \frac{\mathrm{m}_{\mathcal{Z}}}{n}\sum_{j=1}^{n}\frac{1}{\mathrm{m}_{\mathcal{Z}}}\langle\psi_{\mathcal{Z}}(z_j), \widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1}\,\psi_{\mathcal{Z}}(z_j)\rangle_{\mathcal{H}_{\mathcal{Z}}} && (144)\\
&= \mathrm{Tr}\left(\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1}\,\widehat{\mathrm{C}}_{\mathrm{Z}}\right) && (145)\\
&= \left\|\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\,\widehat{\mathrm{C}}_{\mathrm{Z}}{}^{1/2}\right\|_{\mathrm{HS}}^2 && (146)\\
&\leq \left\|\widehat{\mathrm{C}}_{\mathrm{Zt}}{}^{-1/2}\,\mathrm{C}_{\mathcal{Z}\mathrm{t}}{}^{1/2}\right\|_{\mathrm{op}}^2\left\|\mathrm{C}_{\mathcal{Z}\mathrm{t}}{}^{-1/2}\,\widehat{\mathrm{C}}_{\mathrm{Z}}{}^{1/2}\right\|_{\mathrm{HS}}^2\,. && (147)
\end{aligned}$$

But,

$$\left\| {\mathrm{C}_{\mathcal{Z}\mathrm{t}}}^{-1/2} \, \widehat{\mathrm{C}}_{\mathrm{Z}} \, {}^{1/2} \right\|_{\mathrm{HS}}^2 = \mathrm{Tr}\left({\mathrm{C}_{\mathcal{Z}\mathrm{t}}}^{-1} \, \widehat{\mathrm{C}}_{\mathrm{Z}} \right) \tag{148}$$

$$= \mathrm{Tr}\left({\mathrm{C}_{\mathcal{Z}\mathrm{t}}}^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \psi_{\mathcal{Z}}(z_i) \otimes \psi_{\mathcal{Z}}(z_i) \right) \right) \tag{149}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathrm{Tr}\left({\mathrm{C}_{\mathcal{Z}\mathrm{t}}}^{-1} \left(\psi_{\mathcal{Z}}(z_i) \otimes \psi_{\mathcal{Z}}(z_i) \right) \right) \tag{150}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left\langle \psi_{\mathcal{Z}}(z_i), {\mathrm{C}_{\mathcal{Z}\mathrm{t}}}^{-1} \, \psi_{\mathcal{Z}}(z_i) \right\rangle_{\mathcal{H}_{\mathcal{Y}}} \tag{151}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathcal{N}(z_i, t) \tag{152}$$

$$\leq \mathcal{N}_{\mathcal{Z}}^{\infty}(t) \, . \tag{153}$$

Then, from Lemma 9, for $\delta \in (0,1)$, and $\frac{9}{n} \log\left(\frac{n}{\delta}\right) \leq t \leq \|\mathrm{C}_{\mathcal{Z}}\|_{\mathrm{op}}$, then with probability $1-\delta$,

$$\mathbb{E}_{\mathrm{R}_{\mathcal{Z}}}\left[\|X_i\|_{\mathcal{H}_{\mathcal{Z}}} \right]^2 \leq 2\mathcal{N}_{\mathcal{Z}}^{\infty}(t). \tag{154}$$

Then, $\|\Sigma\|_{\mathrm{op}} = \left\| {\widehat{\mathrm{C}}_{\mathrm{Zt}}}^{-1/2} \, \widehat{\mathrm{C}}_{\mathrm{Z}} \, {}^{1/2} \right\|_{\mathrm{op}}^2 \geq 1/3$ for $t \leq 2 \left\| \widehat{\mathrm{C}}_{\mathrm{Z}} \right\|_{\mathrm{op}}$.

We conclude that

$$\frac{\mathbb{E}_{\mathrm{R}_{\mathcal{Z}}}\left[\|W_i\|_{\mathcal{H}_{\mathcal{Z}}} \right]^2}{\|\Sigma\|_{\mathrm{op}}} \leq 6\mathcal{N}_{\mathcal{Z}}^{\infty}(t). \tag{155}$$

Finally, in order to obtain a condition on t that does not depend on empirical quantities, we use Lemma 9 which gives that, for any $\frac{9}{n} \log\left(\frac{n}{\delta}\right) \leq t' \leq \|\mathrm{C}_{\mathcal{Z}}\|_{\mathrm{op}}$, then $\mathrm{C}_{\mathcal{Z}\mathrm{t}'} \preceq 2\, \widehat{\mathrm{C}}_{\mathrm{Zt}'}$, which implies $2 \left\| \widehat{\mathrm{C}}_{\mathrm{Z}} \right\|_{\mathrm{op}} \geq \|\mathrm{C}_{\mathcal{Z}}\|_{\mathrm{op}} - t'$. Now, taking $t' = \frac{9}{n} \log\left(\frac{n}{\delta}\right)$, we obtain $\|\mathrm{C}_{\mathcal{Z}}\|_{\mathrm{op}} - \frac{9}{n} \log\left(\frac{n}{\delta}\right) \leq 2 \left\| \widehat{\mathrm{C}}_{\mathrm{Z}} \right\|_{\mathrm{op}}$.

□

H CONTRIBUTIONS AND PREVIOUS WORKS

Excess-risk bounds for sketched kernel ridge regression have been provided in Rudi et al. (2015) in the case of Nyström subsampling, and scalar-valued ridge regression. Our proofs consist in similar derivations than in Rudi et al. (2015). Nevertheless, we cannot apply directly their results in our setting. More precisely, we do the following additional derivations.

1. Additional decompositions to deal with:
   (a) vector-valued regression instead of scalar-valued regression as in Rudi et al. (2015)
   (b) input and output approximated feature maps
2. Novel probabilistic bounds to deal with gaussian and sub-Gaussian sketching instead of Nyström sketching as in Rudi et al. (2015).

I ADDITIONAL EXPERIMENTS

I.1 Simulated Data Set for Least Squares Regression

We report here some results about statistical performance on the synthetic data set described in Section 5. First, we give an additional figure showing the MSE with respect to $\mathrm{m}_{\mathcal{X}}$ and $\mathrm{m}_{\mathcal{Y}}$ of the SISOKR model, see Figure 3.

As reported in Figure 4, SIOKR outperforms IOKR from $\mathrm{m}_{\mathcal{X}} = 100$, and ISOKR obtains very similar result to IOKR from $\mathrm{m}_{\mathcal{Y}} = 250$.

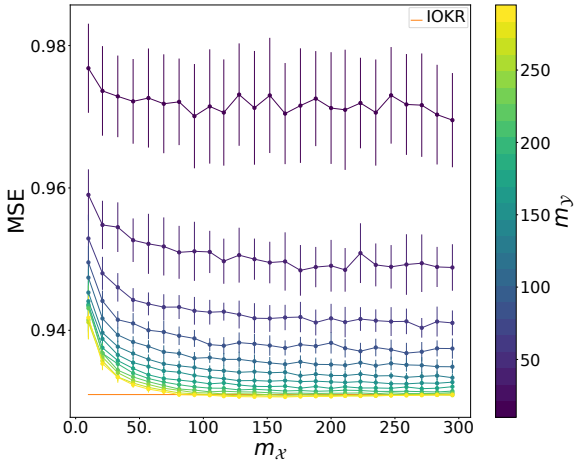

Figure 3: Test MSE with respect to $\mathrm{m}_{\mathcal{X}}$ and $\mathrm{m}_{\mathcal{Y}}$ for the SISOKR model with $(2 \cdot 10^{-3})$-SR input and output sketches.

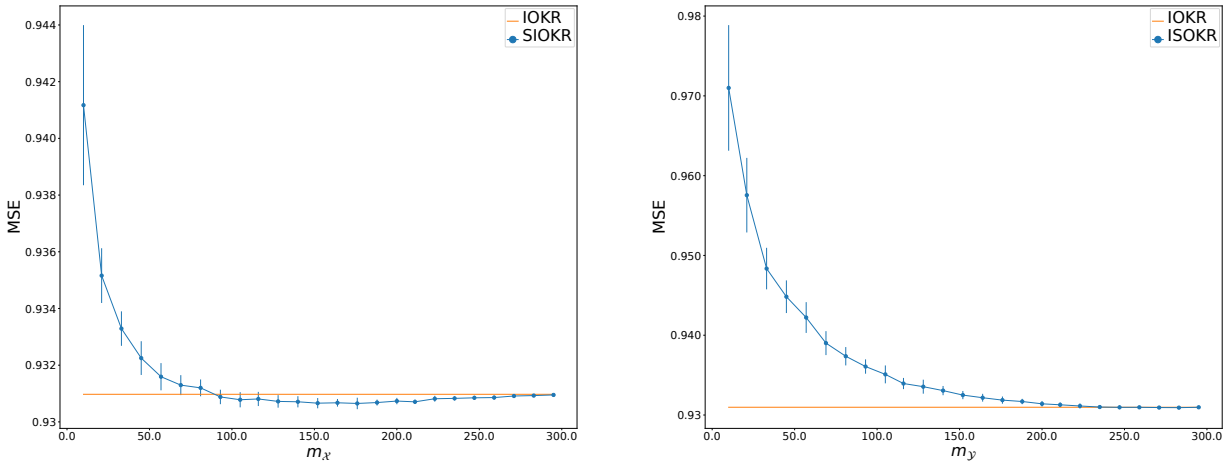

Figure 4: Test MSE with respect to $\mathrm{m}_{\mathcal{X}}$ and $\mathrm{m}_{\mathcal{Y}}$ for a SIOKR and ISOKR model respectively with $(2 \cdot 10^{-3})$-SR input and output sketches.

I.2 More Details about Multi-Label Classification Data Set

In this section, you can find more details about training and testing sizes, the number of features of the inputs, and the number of labels to predict of Bibtex, Bookmarks, and Mediamill data sets in Table 5.

Table 5: Multi-label data sets description.

Data set	n	n_{te}	$n_{features}$	n_{labels}
Bibtex	4880	2515	1836	159
Bookmarks	60000	27856	2150	298
Mediamill	30993	12914	120	101