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 [1] **Luc Brogat-Motte** [1] **Pierre Laforgue** [2] **Florence d'Alché-Buc** [1]

[1] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2] Department of Computer Science, University of Milan, Italy
Correspondance to: tamim.elahmad@telecom-paris.fr

Abstract

Surrogate kernel-based methods offer a flexible solution to structured output prediction by leveraging the kernel trick in both input and output spaces. In contrast to energy-based models, they avoid to pay the cost of inference during training, while enjoying statistical guarantees. However, without approximation, these approaches are condemned to be used only on a limited amount of training data. In this paper, we propose to equip surrogate kernel methods with approximations based on sketching, seen as low rank projections of feature maps both on input and output feature maps. We showcase the approach on Input Output Kernel ridge Regression (or Kernel Dependency Estimation) and provide excess risk bounds that can be in turn directly plugged on the final predictive model. An analysis of the complexity in time and memory show that sketching the input kernel mostly reduces training time while sketching the output kernel allows to reduce the inference time. Furthermore, we show that Gaussian and sub-Gaussian sketches are admissible sketches in the sense that they induce projection operators ensuring a small excess risk. Experiments on different tasks consolidate our findings.

1 Introduction

Ubiquitous in real-world applications, objects composed of different elements that interact with each others, a.k.a. 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 in a supervised learning task, i.e., either as input or output variables, structured objects raise distinct challenges when learning a predictive model. Classification and regression from structured input require a continuous representation that can be learned through a deep neural network (see for instance, (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 output variables impacts directly the definition of the loss function (Ciliberto et al., 2020; Nowak et al., 2019; Cabannes et al., 2021), and therefore the learning problem itself. To handle this general task, an abundant literature has been developed within different directions, each reflecting a way to relax somehow the combinatorial nature of the problem which appears both in training and in prediction. Energy-based approaches relax the structured prediction problem into the learning of a scalar score function LeCun et al. (2007); Tsochantaridis et al. (2005); Belanger and McCallum (2016); Deshwal et al. (2019). End-to-end learning typically exploits a differentiable model, as well as a differentiable loss, to make a structured prediction, enabling gradient descent (Long et al., 2015; Niculae et al., 2018; Berthet et al., 2020). Surrogate methods rely on the implicit embedding of the output variable into a Hilbert Space and solve a surrogate regression problem into this new output space. If they avoid the burden of the inference step (decoding) during learning contrary to energy-based methods, at testing time they also pay the complexity price of the inference. Rare are the methods that simultaneously combine scalability in time and memory

Preprint. Under review.

at learning and inference time, a wide scope of applicability on different structures while offering statistical guarantees of the provided estimator Osokin et al. (2017); Cabannes et al. (2021). In this work, we focus on surrogate methods (Ciliberto et al., 2020) and their implementation as kernel methods, namely the input output kernel regression approaches (Cortes et al., 2005; Brouard et al., 2016b). Recent works have shown that they enjoy consistency and their excess risk is governed by those of the surrogate regression. Moreover, they are well appropriate to make prediction from one (structured) modality to another (structured) one since kernels can be leveraged in the input space as well as the output space. 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 equip these methods with kernel approximations to obtain a drastic reduction of computation and memory at training and testing time while keeping their statistical properties. Several works have successfully highlighted the power of kernel approximation methods such as Random Fourier Features (Rahimi and Recht, 2007; Brault et al., 2016; Rudi and Rosasco, 2017; Li et al., 2021), more generally low-rank approaches (Bach, 2013) and even reached spectacular results as shown by Meanti et al. (2020). Sketched scalar kernel machines have been widely studied first with a particular type of sketching, namely Nyström approximation (Williams and Seeger, 2001; Alaoui and Mahoney, 2015; Rudi et al., 2015), and some works exist handling other sketching distribution, such as Gaussian or Randomized Orthogonal Systems (Yang et al., 2017; Lacotte and Pilanci, 2020). Furthermore, a line of research consists in interpreting such an approximation as data-dependent random features (see e.g. Williams and Seeger (2001); Yang et al. (2012) for Nyström case and Kpotufe and Sriperumbudur (2020) for Gaussian case), leading to interesting applications for kernel PCA (Sterge and Sriperumbudur, 2022) or kernel mean embedding (Chatalic et al., 2022a,b). Current approaches to kernel approximation apply on scalar or vector-valued kernels defined on the input space but no method covers both sides of the coin, aka approximation of both input and output kernels.

Contributions. By extending the random projection interpretation of Nyström's method for scalar-valued kernels to arbitrary sketching matrices and vector-valued RKHSs, we present four contributions

- We apply sketching to both the inputs and the outputs of the Ridge vector-valued regression problems solved as subroutines when performing structured prediction with kernels. This is shown to accelerate respectively the learning and the inference steps of the approach.
- We derive excess risk bounds for the sketched estimator thus obtained, which are controlled by the properties of the sketched projection operator.
- We show that Gaussian and sub-Gaussian sketches are admissible sketches in the sense that they lead to close to optimal learning rates with sketching sizes $m \ll n$.
- We provide structured prediction experiments on real-world datasets which confirm our method's relevance.

2 Structured Prediction with input and output kernels

In this section, we introduce Structured Prediction when leveraging a kernel-induced loss. Surrogate kernel methods are recalled in this context with an emphasis on Input Output Kernel ridge Regression.

Notation. If $\mathcal{Z}$ denotes a generic Polish space, k_z is a positive definite kernel over $\mathcal{Z}$ and $\chi(z) := k_z(\cdot, z)$ is the canonical feature map of k_z. $\mathcal{H}_z$ denotes the Reproducing Kernel Hilbert Space (RKHS) associated to k_z. $S_Z : f \in \mathcal{H}_z \mapsto (1/\sqrt{n})(f(z_1), \dots, f(z_n))^\top$ is the sampling operator over $\mathcal{H}_z$ (Smale and Zhou, 2007). For any operator A, we denote $A^\#$ its adjoint. The adjoint of the sampling operator is defined as $S_Z^\# : \alpha \in \mathbb{R}^n \mapsto (1/\sqrt{n}) \sum_{i=1}^n \alpha_i \chi(z_i)$. If z is a random variable distributed according the probability distribution ρ_z, the covariance operator over $\mathcal{H}_Z$ is denoted $C_Z = \mathbb{E}_z[\chi(z) \otimes \chi(z)]$ and the empirical covariance operator over $\mathcal{H}_Z$ writes as $\widehat{C}_Z = (1/n) \sum_{i=1}^n \chi(z_i) \otimes \chi(z_i) = S_Z^\# S_Z$, where $\{(z_i)_{i=1}^n\}$ is an i.i.d. sample drawn from the probability distribution ρ_z. For any matrix M, we denote $M^\dagger$ its Moore-Penrose inverse.

Structured prediction with surrogate kernel methods. Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ the output space. In general, $\mathcal{Y}$ is finite and extremely large, which makes the task of approximating the relationship between an input variable X with values in $\mathcal{X}$ and an output random variable Y with values in $\mathcal{Y}$ intractable in a direct way Nowozin and Lampert (2011). In this work, we rely on a

kernel-induced loss to fix the goal of structured prediction and leverage the kernel trick to define a tractable surrogate regression problem.

Suppose a positive definite kernel $k_y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that measures how close are two objects from $\mathcal{Y}$. Denote $\mathcal{H}_y$ the Reproducing Kernel Hilbert Space associated to k_y and $\psi : \mathcal{Y} \to \mathcal{H}_y$, the canonical feature map such as $\psi(y) := k_y(\cdot, y), \forall y \in \mathcal{Y}$. We consider the loss function induced by the kernel k_y defined by $\ell : (y, y') \to \|\psi(y) - \psi(y')\|^2_{\mathcal{H}_y}$, for any pair $(y, y') \in \mathcal{Y} \times \mathcal{Y}$. Its relevance for the structured prediction problem at hand directly depends on the kernel chosen (see examples in Section 5), and the wide variety of kernels defined over structured objects (Gärtner, 2008; Borgwardt et al., 2020) ensures its great flexibility.

Given a fixed but unknown joint probability distribution ρ defined on $\mathcal{X} \times \mathcal{Y}$, the goal of *Structured Prediction* is to find an estimator $\hat{f}$ of the target function

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

where $\mathcal{R}(f) = \mathbb{E}_{(x,y)\sim\rho}\left[\|\psi(y) - \psi(f(x))\|^2_{\mathcal{H}_y}\right]$, using a training i.i.d. sample $\{(x_1, y_1), ..., (x_n, y_n)\}$ drawn from ρ.
To tackle this problem, various works Cortes et al. (2005); Geurts et al. (2006); Brouard et al. (2011); Ciliberto et al. (2016) have proposed to proceed in a two-step approach, referred to as Output Kernel Regression (OKR):

1. **Surrogate Regression:** Find $\hat{h}$ an estimator of the surrogate target $h^* : x \mapsto \mathbb{E}[\psi(y)|x]$ the minimizer of $\arg\min_{h:\mathcal{X}\to\mathcal{H}_y} \mathbb{E}\left[\|h(x) - \psi(y)\|^2_{\mathcal{H}_y}\right]$, exploiting the *implicit* knowledge of the training sample $\{(x_1, \psi(y_1)), ..., (x_n, \psi(y_n))\}$.
2. **Pre-image or decoding**: Define $\hat{f}$ by decoding $\hat{h}$, i.e.,

$$\hat{f}(x) := \arg\min_{y\in\mathcal{Y}} \|\hat{h}(x) - \psi(y)\|^2_{\mathcal{H}_y}\,.$$

The surrogate regression problem in Step 1 is much easier to handle than the initial Structured Prediction problem described in (1): it avoids to learn f through the composition with the implicit feature map ψ, and relegates the difficulty of manipulating structured objects to Step 2, i.e., at testing time. Besides, the issue of vector-valued regression into an infinite dimensional space can be overcome by leveraging the kernel trick in the output space.

Moreover, these OKR approaches belong to the general framework of SELF (Ciliberto et al., 2016) and ILE (Ciliberto et al., 2020) and enjoy valuable theoretical guarantees. They are Fisher consistent, meaning that h^* gives exactly rise to the target f^* after decoding, and that the excess risk of $\hat{f}$ is controlled by that of $\hat{h}$, as shown in Ciliberto et al. (2016, Theorem 2).

Input Output Kernel Regression. A natural choice to tackle the surrogate regression problem consists in solving a kernel ridge regression problem, assuming that the hypothesis space is a vector-valued RKHS (Senkene and Tempel'man, 1973; Micchelli and Pontil, 2005; Carmeli et al., 2006, 2010). In a nutshell, if $\mathcal{F}$ denotes a Hilbert space, a mapping $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{F})$, where $\mathcal{L}(\mathcal{F})$ is the set of bounded linear operators on $\mathcal{F}$, is an operator-valued kernel (OVK) if it satisfies the following properties: $\mathcal{K}(x, x') = \mathcal{K}(x', x)^{\#}$ for all $(x, x') \in \mathcal{X}^2$ and such that for all $n \in \mathbb{N}$ and any $(x_i, \varphi_i))_{i=1}^n \in (\mathcal{X} \times \mathcal{F})^n$ we have $\sum_{i,j=1}^n \langle \varphi_i, \mathcal{K}(x_i, x_j)\,\varphi_j)\rangle_{\mathcal{F}} \geqslant 0$.

Similarly to the scalar case, given an OVK $\mathcal{K}$, one can define a unique Hilbert space $\mathcal{H}_\mathcal{K}$ of functions with values in $\mathcal{F}$ associated to $\mathcal{K}$, that enjoys the reproducing kernel property, i.e., such that for all $x \in \mathcal{X}$, $\varphi \in \mathcal{F}$ and $f \in \mathcal{H}_K$ we have $x' \mapsto \mathcal{K}(x, x')\,\varphi \in \mathcal{F}$, and $\langle f, \mathcal{K}(\cdot, x)\,\varphi\rangle_{\mathcal{H}_K} = \langle f(x), \varphi\rangle_{\mathcal{F}}$. The reader can refer to (Carmeli et al., 2010) for further details on vv-RKHSs.

In what follows, we are interested in functions with values in the RKHS $\mathcal{H}_y$ and we opt for the decomposable identity operator-valued kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{H}_y)$, defined as: $\mathcal{K}(x, x') = k_x(x, x')\, I_{\mathcal{H}_y}$, where $k_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite scalar-valued kernel on $\mathcal{X}$. By symmetry with the output space, we denote $\mathcal{H}_x$, the RKHS associated to k_x and for sake of simplicity $\mathcal{H} := \mathcal{H}_\mathcal{K}$, the RKHS associated to $\mathcal{K}$.

We are now well equipped to describe Input Output Kernel ridge Regression (IOKR for short) (Brouard et al., 2011, 2016b; Ciliberto et al., 2020) also introduced as Kernel Dependency Estimation by Weston et al. (2003); Cortes et al. (2005). In IOKR, the estimator of the surrogate regression is obtained by solving the following ridge regression problem within $\mathcal{H}$, given a regularisation penalty $\lambda > 0$,

$$\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \|\psi(y_i) - h(x_i)\|^2_{\mathcal{H}_y} + \lambda \|h\|^2_{\mathcal{H}} \,. \tag{2}$$

Interestingly, as noticed by several authors, the unique solution of the above kernel ridge regression problem can be expressed in different ways. We describe the two of them which are exploited in this work. First, $\hat{h}(x)$ computes a weighted combination of the training outputs

$$\hat{h}(x) = \sum_{i=1}^{n} \hat{\alpha}_i(x) \psi(y_i), \tag{3}$$

with

$$\hat{\alpha}(x) = (K_X + n\lambda)^{-1} \kappa^x_X = \widehat{\Omega} \kappa^x_X \,, \tag{4}$$

where $K_X = (k_x(x_i, x_j))^n_{i,j=1} \in \mathbb{R}^{n \times n}$, and $\kappa^x_X = (k_x(x, x_1), \ldots, k_x(x, x_n))^\top$.

Second, $\hat{h}(x)$ also results from the application of an operator $\widehat{H}$ on $\phi(x)$, i.e.,

$$\hat{h}(x) = \widehat{H} \phi(x), \tag{5}$$

where

$$\widehat{H} = S^{\#}_Y S_X \left(\widehat{C}_X + \lambda I\right)^{-1}. \tag{6}$$

The first expression can be obtained by applying a representer theorem in vv-RKHS (Micchelli and Pontil, 2005) and solving the resulting quadratic program. The second expression can be seen as a re-writing of the first one as highlighted in Ciliberto et al. (2016, Lemma 17) and may also be related to the conditional kernel empirical mean embedding Grünewälder et al. (2012).

The final estimator $\hat{f}$ is computed using the weighted expression in Equation (4) to benefit from the kernel trick:

$$\hat{f}(x) = \operatorname*{arg\,min}_{y \in \mathcal{Y}} k_y(y, y) - 2 {\kappa^x_X}^T \widehat{\Omega} \kappa^y_Y \tag{7}$$

where $\kappa^y_Y = (k_y(y, y_1), \ldots, k_y(y, y_n))^\top$.
The training phase involves thus the inversion of a $n \times n$ matrix which cost without any approximation is $\mathcal{O}(n^3)$.
In practice, the decoding step is performed by searching in a candidate set $\mathcal{Y}_c \subseteq \mathcal{Y}$ of size n_c. Hence, given that we want to perform predictions on a test set X_{te} of size n_{te}, the main quantity to compute is

$$\underbrace{K_{te,tr}}_{n_{te} \times n} \underbrace{\widehat{\Omega}}_{n \times n} \underbrace{K^y_{tr,c}}_{n \times n_c}, \tag{8}$$

where $K_{te,tr} = (k_x(x^{te}_i, x_j))_{1 \le i \le n_{te}, 1 \le j \le n} \in \mathbb{R}^{n_{te} \times n}$ and $K^y_{tr,c} = \left(k_y(y_i, y^c_j)\right)_{1 \le i \le n, 1 \le j \le n_c} \in \mathbb{R}^{n \times n_c}$. The complexity of decoding part is $\mathcal{O}\left(n^2 n_c\right)$, considering $n_{te} < n \le n_c$. IOKR thus suffers, like most kernel methods applied in a naive fashion, from a heavy computational cost. In order to alleviate this limitation and widen the applicability of this method, we develop a general sketching approach that applies both in the input and output feature spaces to accelerate the training and the decoding steps.

3 Sketched Input Sketched Output Kernel Regression

In this section, we describe how to construct $\tilde{h}$, a low-rank approximate estimator of $\hat{h}$, thanks to orthogonal projection operators $\widetilde{P}_X$ and $\widetilde{P}_Y$ onto subspaces of $\mathcal{H}_x$ and $\mathcal{H}_y$ respectively. A general sketching approach is proposed to define such projectors. Ultimately this gives rise to a novel estimator $\tilde{f}$ for structured prediction.

Low-rank estimator. Starting from the expression of $\hat{h}$ in (5), we replace the sampling operators on both sides, S_X and S_Y, by their projected counterparts, $\widetilde{P}_X S_X^{\#}$ and $\widetilde{P}_Y S_Y^{\#}$, in (6), so as to encode dimension reduction. The proposed low-rank estimator expresses as follows:

$$\tilde{h}(x) = \widetilde{P}_Y S_Y^{\#} S_X \widetilde{P}_X (\widetilde{P}_X S_X^{\#} S_X \widetilde{P}_X + \lambda I_{\mathcal{H}_x})^{-1} \phi(x) \,. \tag{9}$$

In the following, we show how to design the projection operators using sketching and then derive the novel expression of the low-rank estimator in terms of weighted combination of the training outputs: $\tilde{h}(x) = \sum_{i=1}^{n} \tilde{\alpha}_i \psi(y_i)$, yielding a reduced computational cost.

Sketching. In this work, we chose to leverage sketching (Mahoney et al., 2011; Woodruff, 2014) to obtain random projectors within the input and output feature spaces. Indeed, sketching consists in approximating a feature map $\chi : \mathcal{Z} \to \mathcal{H}_z$ by projecting it thanks to a random projection operator $\widetilde{P}_Z \in \mathcal{H}_z \otimes \mathcal{H}_z$ defined as follows. Given a random matrix $R_z \in \mathbb{R}^{m_z \times n}$, n data $(z_i)_{i=1}^n \in \mathcal{Z}$ and $m_z \ll n$, the linear subspace defining $\widetilde{P}_Z$ is constructed as the linear subspace generated by the span of the following m_z random vectors

$$\sum_{j=1}^{n} (R_z)_{ij} \chi(z_j) \in \mathcal{H}_z, \quad i = 1, \ldots, m_z \,.$$

This random matrix is drawn from a distribution on $\mathbb{R}^{m_z \times n}$, including classical examples such as sub-sampling sketches where each row is randomly drawn from the rows of the identity matrix I_n (sub-sampling sketches correspond to Nyström approximation (Rudi et al., 2015)) or Gaussian sketches (Kpotufe and Sriperumbudur, 2020) where every entries are i.i.d. Gaussian random variables. One can compute the closed form of such a projector $\widetilde{P}_Z$ as showed in Proposition 2, in Appendix B.

Sketched Input Sketched Output Kernel Regression estimator (SISOKR) We now give the expression of the SISOKR estimator $\tilde{h}$ amenable to practical implementations (see Appendix B for the proof).

Proposition 1 (Expression of SISOKR). *For all* $x \in \mathcal{X}$, $\tilde{h}(x) = \sum_{i=1}^{n} \tilde{\alpha}_i(x) \psi(y_i)$ *where*

$$\tilde{\alpha}(x) = R_y^\top \widetilde{\Omega} R_x \kappa_X^x \,, \tag{10}$$

with $\widetilde{\Omega} = \widetilde{K}_Y^{\dagger} R_y K_Y K_X R_x^\top (R_x K_X^2 R_x^\top + n\lambda \widetilde{K}_X)^{\dagger}$.

We here recover a classical quantity when leveraging sketching for Kernel Ridge Regression, i.e. $K_X R_x^\top (R_x K_X^2 R_x^\top + n\lambda R_x K_X R_x^\top)^{\dagger} R_x \kappa_X^x$ (Rudi et al., 2015; Yang et al., 2017). This quantity allows to reduce the size of the matrix to invert, which is now an $m_x \times m_x$ matrix. This is the main reduction that allows to lower fitting step's complexity, thanks to the input sketching. We still need to perform matrix multiplications $R_x K_X$, which can be very efficient both in time and space complexity for sketches such as sub-sampling or p-sparsified (El Ahmad et al., 2022) sketches, or more costly for Gaussian sketches for instance. Furthermore, output sketching gives additional operations to perform, but the overall cost of computing $\tilde{\alpha}$ complexity can be negligible compared to $\mathcal{O}(n^3)$ according to the sketching scheme used for both input and output kernels. We obtain the corresponding structured prediction estimator $\tilde{f}$ by decoding $\tilde{h}$, i.e. by replacing $\widehat{\Omega}$ by $\widetilde{\Omega}$ in (7). Finally, the purpose of output sketching is to accelerate inference. In fact, the main quantity to compute now when performing predictions is

$$\underbrace{K_{te,tr} R_x^\top}_{n_{te} \times m_x} \underbrace{\widetilde{\Omega}}_{m_x \times m_y} \underbrace{R_y K_{tr,c}^y}_{m_y \times n_c} \,. \tag{11}$$

Time complexity of such an operation is $\mathcal{O}(n_c n_{te} \min(m_x, m_y))$ if $n_{te} > m_y$, and $\mathcal{O}(n_c m_y \min(m_x, n_{te}))$ otherwise. We then obtain a significant complexity reduction, as we do not have any dependency in $n^2 n_c$ anymore.

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 problem that we consider, we upper bound in Theorem 1 the excess-risk of the sketched kernel ridge estimator, exhibiting some approximation errors due to sketching.

Then, in Section 4.2, we provide bounds for these approximation error terms. Finally, this allows us to study in Section 4.3 under which setting the proposed estimators $\tilde{h}$ and $\tilde{f}$ obtain substantial computational gains, while still benefiting from a close to optimal learning rate.

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) quantifying the hardness of the learning problem.

Assumption 1 (Attainability)**.** We assume that $h^* \in \mathcal{H}$, i.e. there exists a linear operator $H : \mathcal{H}_x \to \mathcal{H}_y$ with $\|H\|_{\text{HS}} < +\infty$ such that

$$h^*(x) = H\phi(x) \quad \forall\, x \in \mathcal{X}. \tag{12}$$

This is a standard assumption in the context of least-squares regression (Caponnetto and De Vito, 2007), making the target h^* belonging to the hypothesis space.
We now describe a set of generic assumptions that have to be satisfied by both input and output kernels k_x and k_y.

Assumption 2 (Bounded kernel)**.** We consider positive definite bounded kernels $k_z : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, i.e. we assume that there exists $\kappa_z > 0$ such that

$$k_z(z, z) \leq \kappa_z^2 \quad \forall\, z \in \mathcal{Z}. \tag{13}$$

We note $\kappa_x, \kappa_y > 0$ for the input and output kernels k_x, k_y, respectively.

Assumption 3 (Capacity condition)**.** We assume that there exists $\gamma_z \in [0, 1]$ such that

$$Q_z := \text{Tr}(C_Z^{\gamma_z}) < +\infty\,. \tag{14}$$

This assumption is always verified for $\gamma_z = 1$ (as $\text{Tr}(C_Z) = \mathbb{E}[\|\chi(z)\|^2_{\mathcal{H}_z}] < +\infty$ from Assumption 2), and the smaller the γ_z the faster is the eigenvalue decay of C_Z. It measures the regularity of the features $\chi(z) \in \mathcal{H}_z$ for $z \sim \rho_z$. As a limiting case, when C_Z is finite rank, $\gamma_z = 0$.

Assumption 4 (Embedding property)**.** We assume that there exists $b_z > 0$ and $\mu_z \in [0, 1]$ such that almost surely

$$\chi(z) \otimes \chi(z) \preceq b_z C_Z^{1-\mu_z}. \tag{15}$$

This assumption is always verified for $\mu_z = 1$ (as $\chi(z) \otimes \chi(z) \preceq \kappa_z^2 I_{\mathcal{H}_z}$ from Assumption 2), and the smaller the μ_z the stronger is the assumption. It allows to control the regularity of the functions in $\mathcal{H}_z$ with respect to the L^∞-norm, by giving $\|h\|_{\mathcal{H}_z} \leq b_z^{1/2} \|h\|^{\mu}_{\mathcal{H}_z} \mathbb{E}[h(z)^2]^{(1-\mu)/2}$ (See Pillaud-Vivien et al., 2018). As a limiting case, when C_Z is finite rank, $\mu_z = 0$.

4.1 SISOKR Excess-Risk bound

In this section, we provide 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 $\delta \in [0, 1]$, $n \in \mathbb{N}$ sufficiently large such that $\lambda = n^{-1/(1+\gamma_x)} \geq \frac{9\kappa_x^2}{n} \log(\frac{n}{\delta})$. Under our set of assumptions, the following holds with probability at least $1 - \delta$*

$$\mathbb{E}[\|\tilde{h}(x) - h^*(x)\|^2_{\mathcal{H}_y}]^{\frac{1}{2}} \leq S(n) + c_2 A^{\phi}_{\rho_x}(\widetilde{P}_X) + A^{\psi}_{\rho_y}(\widetilde{P}_Y)$$

where

$$S(n) = c_1 \log(4/\delta) n^{-\frac{1}{2(1+\gamma_x)}} \quad \textit{(regression error)}$$

$$A^{\chi}_{\rho_z}(\widetilde{P}_Z) = \mathbb{E}_z[\|(\widetilde{P}_Z - I_{\mathcal{H}_z})\chi(z)\|^2_{\mathcal{H}_z}]^{\frac{1}{2}} \textit{(sketch. error)}$$

and $c_1, c_2 > 0$ are constants independent of n and δ defined in the proofs.

This bound is a sum of three terms. The first one is a regression error whose dependency in n is the same as the kernel ridge regression (without sketching). In particular, this learning rate has been shown to be optimal under our set of assumptions in a minimax sense (see 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 $\widetilde{P}_X, \widetilde{P}_Y$ of the feature maps ϕ, ψ through the input and output marginal distributions, respectively.

4.2 Sketching Reconstruction Error

In this section, we provide bounds on the sketching reconstruction error for the family of subgaussian sketches.

Subgaussian sketches are defined as follows.

Definition 1. A subgaussian sketch $R \in \mathbb{R}^{m\times n}$ is composed with i.i.d. elements such that $\mathbb{E}[R_{ij}] = 0$, $\mathbb{E}\left[R_{ij}^2\right] = 1/m$ and R_{ij} is $\frac{\nu^2}{m}$-subgaussian, for all $1 \leq i \leq m$ and $1 \leq j \leq n$, where $\nu \geq 1$.

First, a standard normal r. v. is 1-subgaussian. Moreover, thanks to Hoeffding's lemma, any r.v. taking values in a bounded interval $[a, b]$ is $(b-a)^2/4$-subgaussian. Hence, a sketch matrix composed with i.i.d. Gaussian or bounded r.v. is a subgaussian sketch. Finally, the p-sparsified sketches (El Ahmad et al., 2022) are subgaussian with $\nu^2 = 1/p$, $p \in]0, 1]$.

Theorem 2 (Subgaussian sketching reconstruction error)**.** *For $\delta \in (0, 1/e]$, $n \in \mathbb{N}$ sufficiently large such that $\frac{9}{n}\log(n/\delta) \leq n^{-\frac{1}{1+\gamma_z}} \leq \|C_Z\|_{\text{op}}/2$, then if*

$$m \geq c_4 \max\left(\nu_z^2 n^{\frac{\gamma_z+\mu_z}{1+\gamma_z}}, \nu_z^4 \log(1/\delta)\right), \tag{16}$$

then with probability $1-\delta$

$$\mathbb{E}_z[\|(\widetilde{P}_Z - I_{\mathcal{H}_z})\chi(z)\|_{\mathcal{H}_z}^2] \leq c_3 n^{-\frac{1-\gamma_z}{(1+\gamma_z)}} \tag{17}$$

where $c_3, c_4 > 0$ are constants independents of n, m, δ defined in the proofs.

This theorem shows that using a sketching size $m = \mathcal{O}(n^{\frac{\gamma_z+\mu_z}{1+\gamma_z}})$ allows to obtain a reconstruction error of order $\mathcal{O}(n^{-\frac{1-\gamma_z}{(1+\gamma_z)}})$. Hence, depending on the regularity of the distribution (defined through our set of assumptions), one can obtain a small reconstruction error when using a small sketching size. For instance, if $\mu_z = \gamma_z = 1/3$, one obtain a reconstruction error of order $\mathcal{O}(n^{-1/2})$ by using a sketching size of order $\mathcal{O}(n^{1/2}) \ll \mathcal{O}(n)$. As a limiting case, when $\mu_z = \gamma_z = 0$, one obtain a reconstruction error of order $\mathcal{O}(n^{-1})$ by using a constant sketching size.

4.3 SISOKR Learning Rates

For the sake of presentation, we use $\lesssim$ to keep only the dependencies in $n, \delta, \nu, \gamma, \mu$. We note $a \vee b := \max(a, b)$.

Corollary 1 (SISOKR learning rates)**.** *Under the Assumptions of Theorems 1 and 2, if for all $y \in \mathcal{Y}$, $\|\psi(y)\|_{\mathcal{H}_y} = \kappa_y$, for $n \in \mathbb{N}$ sufficiently large such that $\frac{9}{n}\log(n/\delta) \leq n^{-\frac{1}{1+\gamma_x}} \leq \|C_X\|_{\text{op}}/2$, and $\frac{9}{n}\log(n/\delta) \leq n^{-\frac{1}{1+\gamma_y}} \leq \|C_Y\|_{\text{op}}/2$, and for sketching size $m_x, m_y \in \mathbb{N}$ such that*

$$m_x \gtrsim \max\left(\nu_x^2 n^{\frac{\gamma_x+\mu_x}{1+\gamma_x}}, \nu_x^4 \log(1/\delta)\right), \tag{18}$$

$$m_y \gtrsim \max\left(\nu_y^2 n^{\frac{\gamma_y+\mu_y}{1+\gamma_y}}, \nu_y^4 \log(1/\delta)\right), \tag{19}$$

then with probability $1-\delta$

$$\mathbb{E}[\|\tilde{h}(x) - h^*(x)\|_{\mathcal{H}_y}^2]^{\frac{1}{2}} \lesssim \log(4/\delta)\, n^{-\frac{1-\gamma_x\vee\gamma_y}{2(1+\gamma_x\vee\gamma_y)}}, \tag{20}$$

and

$$\mathcal{R}(\tilde{f}) - \mathcal{R}(f^*) \lesssim \log(4/\delta)\, n^{-\frac{1-\gamma_x\vee\gamma_y}{2(1+\gamma_x\vee\gamma_y)}}. \tag{21}$$

Proof. Applying Theorems 1 and 2 to bound $A_{\rho_x}^{\phi}(\widetilde{P}_X)$ and $A_{\rho_y}^{\psi}(\widetilde{P}_Y)$ gives (20). Applying the comparison inequality from Ciliberto et al. (2020) to the loss $\Delta(y, y') = \|\psi(y) - \psi(y')\|_{\mathcal{H}_y}^2$ allows to conclude the proof for Eq. (21). □

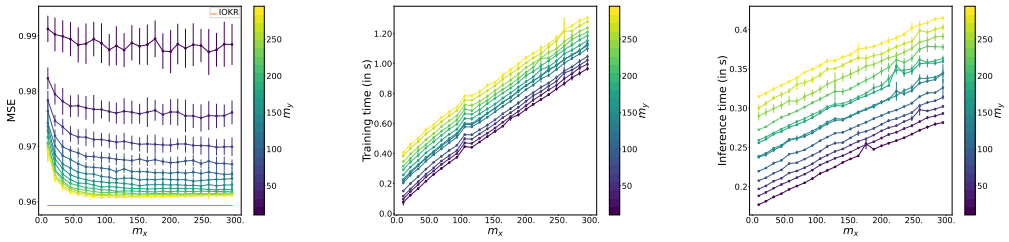

Figure 1: Trade-off between Accuracy and Efficiency for a SISOKR model with $(2 \cdot 10^{-3})$-SR input and output sketches.

This corollary shows that under strong enough regularity assumptions, the proposed estimators benefit from a close to optimal learning rate but only requires a small input and output sketching sizes, leading to significant computational gain for the training and the decoding steps. For instance, if $\mu_x = \mu_y = \gamma_x = \gamma_y = 1/3$, 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 requires sketching sizes m_x, m_y of order $\mathcal{O}(n^{1/2}) \ll \mathcal{O}(n)$. As a limiting case, when $\mu_x = \mu_y = \gamma_x = \gamma_y = 0$, one obtains a, optimal learning rate of $\mathcal{O}(n^{-1/2})$ by using constant sketching sizes.

Remark 1 (Related Work). 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 will consist in similar derivations than in Rudi et al. (2015). Nevertheless, we cannot apply directly their results in our setting because of the three following differences: the definition of R_z, we consider vector-valued ridge regression, we perform sketching also on the output features. We provided more details about this in Appendix G.

Remark 2 (Other Sketches). Although we focused on subgaussian sketches, any sketching distribution admitting concentration bounds for operators on separable Hilbert spaces allows to bound the quantity $A^{\chi}_{\rho_z}(\widetilde{P}_Z)$ 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.

Remark 3 (Application to Least Squares Regression). This model and theoretical framework applies to any least squares regression problem with identity separable input kernel and separable Hilbert output space $\mathcal{Y}$. It corresponds to having the linear output kernel $k_y(\cdot, \cdot) = \langle \cdot, \cdot \rangle_{\mathcal{Y}}$, and then $\psi = I_{\mathcal{Y}}$.

Remark 4 (Comparison to K-satisfiability). Note that our framework significantly departs from that of K-satisfiability (Yang et al., 2017; Chen and Yang, 2021), a popular approach to analyze sketching in kernel methods. First, we highlight that K-satisfiability provides Gram matrix-specific bounds (through the critical radius), while ours are expressed in terms of quantities characteristics to the kernel. It then allows to upper bound the squared $L^2(\mathbb{P}_n)$ error between $\tilde{h}$ and h^*, while our projection point of view provides a direct control on the excess risk. Finally, it is worth noting that our approach shows that sub-Gaussian sketches are admissible, which cannot be proven through K-satisfiability.

5 Experiments

In this section, we present experiments on synthetic and real-world datasets. In the following, SIOKR and ISOKR denote the models with sketching leveraged only on the inputs (respectively outputs). We focus on uniform sub-sampling (Rudi et al., 2015) and p-SR/SG sketches (El Ahmad et al., 2022), which are covered by our theoretical analysis. Results reported are averaged over 30 replicates, unless for the metabolite's experiments where 5 replicates are averaged.

5.1 Synthetic Least Squares Regression

We generate a synthetic dataset of least-squares regression, with $n = 10,000$ training data points, $\mathcal{X} = \mathcal{Y} = \mathcal{H}_y = \mathbb{R}^d$ and $d = 300$. We construct covariance matrices C and E by drawing randomly their eigenvectors and such that their eigenvalues are $\sigma_k(C) = k^{-3/2}$ and $\sigma_k(E) = 0.2k^{-1/10}$. We draw $H_0 \in \mathbb{R}^{d \times d}$ with i.i.d. coefficients from the standard normal distribution and set $H = CH_0$.

Table 1: F_1 score on tag prediction from text data.

Method	Bibtex	Bookmarks
SISOKR	44.1 ± 0.07	**39.3** ± 0.61
ISOKR	44.8 ± 0.01	NA
SIOKR	44.7 ± 0.09	39.1 ± 0.04
IOKR	**44.9**	NA
LR	37.2	30.7
NN	38.9	33.8
SPEN	42.2	34.4
PRLR	44.2	34.9
DVN	44.7	37.1

Table 2: MSE and standard errors for the metabolite identification problem. SPEN directly predicts outputs in $\mathcal{Y}$, then MSE is not defined.

Method	MSE	Tanimoto-Gaussian loss	Top-1 | 5 | 10 accuracies
SISOKR	0.832 ± 0.002	0.597 ± 0.009	22.7% | 50.6% | 61.4%
ISOKR	0.825 ± 0.002	0.566 ± 0.009	24.2% | 53.1% | 63.5%
SIOKR	0.793 ± 0.002	0.507 ± 0.010	28.5% | 59.9% | 69.6%
IOKR	**0.780** ± 0.002	**0.486** ± 0.008	**29.6**% | **61.6**% | **71.4**%
SPEN	NA	0.537 ± 0.008	25.9% | 54.1% | 64.3%

For $i \leq n$, we generate inputs $x_i \sim \mathcal{N}(0, C)$, noise $\epsilon_i \sim \mathcal{N}(0, E)$ and outputs $y_i = Hx_i + \epsilon_i$. We generate validation and test sets of $n_{val} = n_{te} = 1000$ points in the same way.

Such a choice of matrices C with a polynomial eigenvalue decay, E with very low eigenvalues and eigenvalue decay, and $H = CH_0$ allows to enforce a high eigenvalue decay for C_Y, since it will have a similar eigenvalue decay as C, and favorable settings to deploy sketching, as the true regression function H is low rank and sketching induces low-rank estimators by construction, encoded by the sketch sizes m_x and m_y and the projection operators $\widetilde{P}_X$ and $\widetilde{P}_Y$, see (9) and Proposition 2. As stated by Corollary 1, the higher are the eigenvalue decays of C_X and C_Y, the lower we can set m_x and m_y.

We used the Gaussian kernel and selected its bandwidth —as well as the regularisation penalty λ— via 1-fold cross-validation. We learn the SISOKR model for different values of m_x and m_y (from 10 to 295) and $(2 \cdot 10^{-2})$-SR input and output sketches. Figure 1(a) presents test errors, measured in terms of Mean Squared Error (MSE), for many choices of m_y, as a function of the sketch size m_x, and the test error of IOKR, i.e. non-sketched model. Figure 1(b) and Figure 1(c) show the corresponding computational training and inference time. Note that for such a problem where $\mathcal{Y} = \mathcal{H}_y$, there is no decoding step during inference. However, we perform an artificial pre-image problem to illustrate the computational benefit from sketching during this phase in general structured prediction problems. We observe that the MSE decreases as the sketch sizes m_x and m_y increase, and more precisely, it decreases faster with respect to m_x than m_y. This might be due to the fact that we used a linear kernel on the output space, which is only 300-dimensional, hence we benefit less from sketching the output kernel than sketching the input Gaussian kernel, whose RKHS $\mathcal{H}_x$ is infinite-dimensional. This can also explain why we obtain a degraded MSE performance compared to IOKR, while when performing sketching solely on the input kernel, we closely reach the performance of IOKR (see Figure 2 in Appendix H.1). Turning to training and inference times, we observe a reduction compared to IOKR, where training takes around 0.06 to 1.3 seconds for sketch models and 16 seconds for IOKR, and inference takes around 0.07 to 0.32 seconds for sketch models and 3.2 seconds for IOKR Furthermore, training time decreases when m_x and m_y decrease but is more sensitive to input sketch size m_x, while inference time decreases too when m_x and m_y decrease but is more sensitive to output sketch size m_y as expected.

5.2 Multi-Label Classification

We compare our sketched models with SOTA multi-label and structured prediction methods, namely IOKR (Brouard et al., 2016b), logistic regression (LR) trained independently for each label (Lin et al., 2014), a two-layer neural network with cross entropy loss (NN) (Belanger and McCallum, 2016), the

Table 3: Comparison of training/inference computation times (in seconds).

Data set	IOKR	SIOKR	ISOKR	SISOKR
Bibtex	2.54 / 1.18	1.99 ± 0.07 / 1.22 ± 0.03	2.51 ± 0.06 / 0.58 ± 0.01	**1.41** ± 0.03 / **0.46** ± 0.01
Bookmarks	NA	354 ± 2.1 / 297 ± 2.1	NA	**118** ± 1.5 / **20** ± 0.2
Metabolite	1.96 ± 0.40 / 957 ± 28	**1.56** ± 0.02 / 940 ± 28	3.46 ± 0.22 / 878 ± 23	2.85 ± 0.10 / **770** ± 25

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.

Bibtex and Bookmarks (Katakis et al., 2008) 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 (webpage) or Bibtex entry to the social bookmarking system Bibsonomy. Bibtex contains $n = 4880$ training points, while Bookmarks contains $n = 60,000$ training points (see Table 4 for details).

For all multi-label experiments, we used Gaussian input and output kernels with widths σ^2_{input} and σ^2_{output}. We used p-SG input sketches for SIOKR models, p-SG output sketches for ISOKR models, and uniform sub-sampling input sketches and p-SG output sketches for SISOKR models. For Bibtex experiments, we chose $p = 4 \cdot 10^{-3}$, $m_x = 2,250$ and $m_y = 200$, and for Bookmarks experiments, $p = 3 \cdot 10^{-4}$, $m_x = 13,000$ and $m_y = 750$. All the training data were used as candidate sets. Performance of the algorithms are measured by example-based F1 score. We selected the hyper-parameters λ, σ^2_{input} and σ^2_{output} in logarithmic grids by 5-fold cross-validation.

The results in Table 1 show than surrogate methods (first four lines) can compete with SOTA methods. In the case of Bibtex, sketched models preserve good performance compared to IOKR while being faster in training phase for SIOKR and SISOKR, and significantly faster in inference phase for ISOKR and SISOKR, see Table 3. Since Bookmarks data set is too large, storing the whole n^2-Gram matrix K_X exceeds CPU's space limitations. Hence, we only tested SIOKR and SISOKR models on this data set, which outperform other methods. Note that, with the same sketch matrix R_x, SIOKR's training phase is faster than SISOKR's one because there is not additional computations on the output Gram matrix K_Y. In Table 3, SISOKR is faster during training for multi-label data set since the input sketching used is more efficient (sub-sampling vs. p-SG).

5.3 Metabolite Identification

Identifying small molecules, called metabolites, in a biological sample is a problem of high interest in metabolomics. Mass spectrometry is a widespread method to extract distinctive features from a biological sample in the form of a tandem mass (MS/MS) spectrum. Given the tandem mass spectrum of a metabolite, the objective is to predict its molecular structure. These molecular structures are represented by binary vectors of length $d = 7593$, called fingerprints. Each value of this binary vector 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 in $n = 6974$ training mass spectrums, the median size of the candidate sets is 292 and the largest candidate set contains $36,918$ fingerprints. Therefore, metabolite identification is a problem with high-dimensional complex outputs, making the choice of the output kernel crucial, a small train set and a large number of candidates, making the inference step long.

Our numerical experimental protocol is similar to Brouard et al. (2016a) (5-CV Outer / 4-CV Inner loops), we used probability product input kernel for mass spectra and Gaussian-Tanimoto output kernel – with width σ^2 – for the molecular fingerprints. We selected the hyper-parameters λ and σ^2 in logarithmic grids. For the sketched model, we used p-SR input sketches for SIOKR models, p-SR output sketches for ISOKR models, and uniform sub-sampling input sketches and p-SR output sketches for SISOKR models with $p = 3 \cdot 10^{-3}$, $m_x = 2,500$ and $m_y = 300$.

We compare our sketched models with IOKR and SPEN, see Table 2. Results for SPEN were taken from Brogat-Motte et al. (2022)). Although SIOKR competes with IOKR, and ISOKR and SISOKR compete with SPEN, we observe than IOKR outperform all methods. However, sketched models allow training and inference time reduction, see Table 3.

6 Conclusion

In this paper, we scale up surrogate kernel methods for structured prediction by leveraging random projections, in both input and output feature spaces, to accelerate training and inference phases. We develop a theoretical study of the novel estimator and highlight the impact of input and output sketching in the risk decomposition. We extend the existing theory on Nyström approximation and derive the error induced by generic subgaussian sketches. Experiments on structured prediction problems confirm the advantages of the approach. If this paper focuses on the kernel-induced square loss and the ridge estimator, note that output sketching can be applied to other kernelized non-parametric estimators.

Acknowledgements

This work was supported by the Télécom Paris research chair on Data Science and Artificial Intelligence for Digitalized Industry and Services (DSAIDIS).

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.

Chen, Y. and Yang, Y. (2021). Accumulations of projections—a unified framework for random sketches in kernel ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pages 2953–2961. 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)*.

El Ahmad, T., Laforgue, P., and d'Alché-Buc, F. (2022). p-sparsified sketches for fast multiple output kernel methods.

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.

Katakis, I., Tsoumakas, G., and Vlahavas, I. (2008). Multilabel text classification for automated tag suggestion. In *Proceedings of the ECML/PKDD*, volume 18, page 5. Citeseer.

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.

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.

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.

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.

A Notation 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 ρ_X and ρ_Y the marginal distributions of the inputs and outputs, respectively.

Linear algebra notations. For an operator A, $A^{\#}$ denotes 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. In order 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 notations. For any function $h : \mathcal{X} \to \mathcal{H}_y$, we define its least-squares expected risk as

$$\mathcal{E}(h) = \mathbb{E}_{\rho}\left[\|h(x) - \psi(y)\|^2_{\mathcal{H}_y}\right]. \tag{22}$$

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

RKHS notations. We denote by $\mathcal{H}_x$ and $\mathcal{H}_y$ the RKHSs associated to the input $k_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and output $k_y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ kernels, respectively. We denote by $\phi : \mathcal{X} \to \mathcal{H}_x$ and $\psi : \mathcal{Y} \to \mathcal{H}_y$ the canonical feature maps $\phi(x) = k_x(x, .)$ and $\psi(y) = k_y(y, .)$, respectively. We denote by $\mathcal{H}$ the vv-RKHS associated to the operator-valued kernel $\mathcal{K} = kI_{\mathcal{H}_y}$. 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.

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

Sketching operators.

- We denote $R_x \in \mathbb{R}^{m_x \times n}$ and $R_y \in \mathbb{R}^{m_y \times n}$ the input and output sketch matrices with $m_x < n$ and $m_y < n$,
- $\widetilde{C}_X = S_X^{\#} R_x^{\top} R_x S_X$ and $\widetilde{C}_Y = S_Y^{\#} R_y^{\top} R_y S_Y$,
- $\widetilde{K}_X = R_x K_X R_x^{\top}$ and $\widetilde{K}_Y = R_y K_Y R_y^{\top}$.

B 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.

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

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

$\widetilde{P}_Z$ expression. Let $\left\{\left(\sigma_i(\widetilde{K}_Z), \tilde{\mathbf{v}}_i^z\right), i\in[m_z]\right\}$ be the eigenpairs of $\widetilde{K}_Z$ ranked in the descending order of eigenvalues, $p_z = \operatorname{rank}\left(\widetilde{K}_Z\right)$, and for all $1\le i\le p_z$, $\tilde{e}_i^z = \sqrt{\frac{n}{\sigma_i(\widetilde{K}_Z)}}S_Z^{\#}R_z^{\top}\tilde{\mathbf{v}}_i^z$.

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

$$\widetilde{P}_Z = (R_zS_Z)^{\#}\left(R_zS_Z(R_zS_Z)^{\#}\right)^{\dagger}R_zS_Z\,. \tag{23}$$

Proof. For $1\le i\le p_z$

$$\widetilde{C}_Z\,\tilde{e}_i^z = S_Z^{\#}R_z^{\top}R_zS_Z\left(\sqrt{\frac{n}{\sigma_i(\widetilde{K}_Z)}}S_Z^{\#}R_z^{\top}\tilde{\mathbf{v}}_i^z\right) \tag{24}$$

$$= \sqrt{\frac{n}{\sigma_i(\widetilde{K}_Z)}}S_Z^{\#}R_z^{\top}\left(\frac{1}{n}\widetilde{K}_Z\right)\tilde{\mathbf{v}}_i^z \tag{25}$$

$$= \frac{1}{\sqrt{n\sigma_i(\widetilde{K}_Z)}}S_Z^{\#}R_z^{\top}\sigma_i(\widetilde{K}_Z)\tilde{\mathbf{v}}_i^z \tag{26}$$

$$= \frac{\sigma_i(\widetilde{K}_Z)}{n}\tilde{e}_i^z\,. \tag{27}$$

Moreover, we verify that $\operatorname{span}\left(\tilde{e}_1^z,\dots,\tilde{e}_{p_z}^z\right)$ forms an orthonormal basis. Let $1\le i,j\le p_z$,

$$\left\langle\tilde{e}_i^z,\tilde{e}_j^z\right\rangle_{\mathcal{H}_x} = \left\langle\sqrt{\frac{n}{\sigma_i(\widetilde{K}_Z)}}S_Z^{\#}R_z^{\top}\tilde{\mathbf{v}}_i^z, \sqrt{\frac{n}{\sigma_j(\widetilde{K}_Z)}}S_Z^{\#}R_z^{\top}\tilde{\mathbf{v}}_j^z\right\rangle_{\mathcal{H}_z} \tag{28}$$

$$= \frac{n}{\sqrt{\sigma_i(\widetilde{K}_Z)\sigma_j(\widetilde{K}_Z)}}{\tilde{\mathbf{v}}_i^z}^{\top}R_zS_ZS_Z^{\#}R_z^{\top}\tilde{\mathbf{v}}_j^z \tag{29}$$

$$= \frac{n}{\sqrt{\sigma_i(\widetilde{K}_Z)\sigma_j(\widetilde{K}_Z)}}{\tilde{\mathbf{v}}_i^z}^{\top}\left(\frac{1}{n}\widetilde{K}_Z\right)\tilde{\mathbf{v}}_j^z \tag{30}$$

$$= \frac{\sigma_j(\widetilde{K}_Z)}{\sqrt{\sigma_i(\widetilde{K}_Z)\sigma_j(\widetilde{K}_Z)}}{\tilde{\mathbf{v}}_i^z}^{\top}\tilde{\mathbf{v}}_j^z \tag{31}$$

$$= \delta_{ij}\,, \tag{32}$$

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

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

$$\widetilde{P}_Z = nS_Z^{\#}R_z^{\top}\widetilde{K}_Z^{\dagger}R_zS_Z = (R_zS_Z)^{\#}\left(R_zS_Z(R_zS_Z)^{\#}\right)^{\dagger}R_zS_Z\,. \tag{33}$$

□

Remark 5. With R_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 $\psi(y_i)$s.

Proof. Recall that $\tilde{h}(x) = \widetilde{P}_Y S_Y^{\#} S_X \widetilde{P}_X (\widetilde{P}_X S_X^{\#} S_X \widetilde{P}_X + \lambda I_{\mathcal{H}_x})^{-1} \phi(x)$. By Lemma 1 and especially (37), we obtain that

$$\tilde{h}(x) = \sqrt{n} \widetilde{P}_Y S_Y^{\#} K_X R_x^\top \left(R_x K_X^2 R_x^\top + n\lambda R_x K_X R_x^\top \right)^\dagger R_x S_X \phi(x) \,. \tag{34}$$

Finally, by Lemma 2 and with $\alpha(x) = K_X R_x^\top \left(R_x K_X^2 R_x^\top + n\lambda R_x K_X R_x^\top \right)^\dagger R_x S_X \phi(x)$, we have that $\tilde{h}(x) = \sum_{i=1}^n \tilde{\alpha}_i(x) \psi(y_i)$ where

$$\tilde{\alpha}(x) = R_y^\top \widetilde{K}_Y^\dagger R_y K_Y K_X R_x^\top (R_x K_X^2 R_x^\top + n\lambda \widetilde{K}_X)^\dagger R_x \kappa_X^x \,. \tag{35}$$

□

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

Lemma 1. *Let $\widetilde{H} = \widetilde{P}_Y S_Y^{\#} S_X \widetilde{P}_X (\widetilde{P}_X S_X^{\#} S_X \widetilde{P}_X + \lambda I_{\mathcal{H}_x})^{-1}$. The following two expansions hold true*

$$\widetilde{H} = \widetilde{P}_Y S_Y^{\#} S_X \tilde{\eta}(\widehat{C}_X) \,, \tag{36}$$

where $\tilde{\eta}(\widehat{C}_X) = V(V^{\#} \hat{C}_X V + \lambda I_{\mathcal{H}_x})^{-1} V^{\#}$ and for algorithmic purposes

$$\widetilde{H} = \sqrt{n} \widetilde{P}_Y S_Y^{\#} K_X R_x^\top \left(R_x K_X^2 R_x^\top + n\lambda R_x K_X R_x^\top \right)^\dagger R_x S_X \,. \tag{37}$$

Proof. Let us prove (36) first.

$$\begin{aligned}
\tilde{H} &= \widetilde{P}_Y S_Y^{\#} S_X \widetilde{P}_X (\widetilde{P}_X S_X^{\#} S_X \widetilde{P}_X + \lambda I_{\mathcal{H}_x})^{-1} && (38)\\
&= \widetilde{P}_Y S_Y^{\#} S_X V V^{\#} (V V^{\#} S_X^{\#} S_X V V^{\#} + \lambda I_{\mathcal{H}_x})^{-1} && (39)\\
&= \widetilde{P}_Y S_Y^{\#} S_X V (V^{\#} \hat{C}_X V + \lambda I_{\mathcal{H}_x})^{-1} V^{\#} && (40)\\
&= \widetilde{P}_Y S_Y^{\#} S_X \tilde{\eta}(\widehat{C}_X) \,, && (41)
\end{aligned}$$

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

Now, we focus on proving (37). First, we have that

$$\tilde{H} = \widetilde{P}_Y S_Y^{\#} S_X V (V^{\#} \hat{C}_{X\lambda} V)^\dagger V^{\#} \,. \tag{42}$$

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 $UU^\top = I_{p_x}$ and $V^{\#} V = I_{p_x}$, we have that,

$$\tilde{H} = \widetilde{P}_Y S_Y^{\#} S_X V \Sigma U^\top \left(U \Sigma V^{\#} \widehat{C}_{X\lambda} V \Sigma U^\top \right)^\dagger U \Sigma V^{\#} \,. \tag{43}$$

Then, since $R_x S_X = U \Sigma V^{\#}$,

$$\tilde{H} = \widetilde{P}_Y S_Y^{\#} S_X (R_x S_X)^{\#} \left(R_x S_X \left(\widehat{C}_X + \lambda I_{\mathcal{H}_x} \right) (R_x S_X)^{\#} \right)^\dagger R_x S_X \,. \tag{44}$$

Finally, using the fact that $\widehat{C}_X = S_X^{\#} S_X$ and $K_X = n S_X S_X^{\#}$, we obtain that

$$\widetilde{H} = \sqrt{n} \widetilde{P}_Y S_Y^{\#} K_X R_x^\top \left(R_x K_X^2 R_x^\top + n\lambda R_x K_X R_x^\top \right)^\dagger R_x S_X \,. \tag{45}$$

□

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 $h(x) = \sqrt{n}\widetilde{P}_Y S_Y^{\#}\alpha(x)$ with $\alpha : \mathcal{X} \to \mathbb{R}^n$, then $h(x) = \sum_{i=1}^n R_y^\top \widetilde{K}_Y^\dagger R_y K_Y \alpha(x)\psi(y_i)$.*

Proof.

$$h(x) = \sqrt{n}\widetilde{P}_Y S_Y^{\#}\alpha(x) \tag{46}$$
$$= \sqrt{n} S_Y^{\#} R_y^\top \widetilde{K}_Y^\dagger R_y \left(n S_Y S_Y^{\#}\right)\alpha(x) \tag{47}$$
$$= \sqrt{n} S_Y^{\#} R_y^\top \widetilde{K}_Y^\dagger R_y K_Y \alpha(x) \tag{48}$$
$$= \sum_{i=1}^n R_y^\top \widetilde{K}_Y^\dagger R_y K_Y \alpha(x)\psi(y_i)\,. \tag{49}$$

□

C 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 $\delta \in [0,1]$, $n \in \mathbb{N}$ sufficiently large such that $\lambda = n^{-1/(1+\gamma_x)} \geq \frac{9\kappa_x^2}{n}\log(\frac{n}{\delta})$. Under our set of assumptions, the following holds with probability at least $1-\delta$*

$$\mathbb{E}[\|\tilde{h}(x) - h^*(x)\|_{\mathcal{H}_y}^2]^{\frac{1}{2}} \leq S(n) + c_2 A_{\rho_x}^{\phi}(\widetilde{P}_X) + A_{\rho_y}^{\psi}(\widetilde{P}_Y)$$

where

$$S(n) = c_1 \log(4/\delta) n^{-\frac{1}{2(1+\gamma_x)}} \quad \textit{(regression error)}$$
$$A_{\rho_z}^{\chi}(\widetilde{P}_Z) = \mathbb{E}_z[\|(\widetilde{P}_Z - I_{\mathcal{H}_z})\chi(z)\|_{\mathcal{H}_z}^2]^{\frac{1}{2}} \textit{(sketch. error)}$$

and $c_1, c_2 > 0$ are constants independent of n and δ defined in the proofs.

Proof. Our proofs consists in decompositions and then applying the probabilistic bounds given in Section E.

We have

$$\mathbb{E}[\|\tilde{h}(x) - h^*(x)\|^2]^{1/2} = \|(\tilde{H} - H)S^{\#}\|_{\mathrm{HS}} \tag{50}$$

with $\tilde{H} = \widetilde{P}_Y S_Y^{\#} S_X \tilde{\eta}(\widehat{C}_X)$.

Then, defining $H_\lambda = HC_X(C_X + \lambda I)^{-1}$, we decompose

$$\tilde{H} - H = \widetilde{P}_Y\left(S_Y^{\#}S_X - H_\lambda \widehat{C}_X\right)\tilde{\eta}(\widehat{C}_X) + \widetilde{P}_Y H_\lambda\left(\widehat{C}_X\tilde{\eta}(\widehat{C}_X) - I_{\mathcal{H}_x}\right) + \left(\widetilde{P}_Y H_\lambda - H\right) \tag{51}$$

such that

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

with

$$(A) = \left\|\left(S_Y^{\#}S_X - H_\lambda\widehat{C}_X\right)\tilde{\eta}(\widehat{C}_X)C_X^{1/2}\right\|_{\mathrm{HS}} \tag{52}$$
$$(B) = \left\|H_\lambda\left(\widehat{C}_X\tilde{\eta}(\widehat{C}_X) - I_{\mathcal{H}_x}\right)C_X^{1/2}\right\|_{\mathrm{HS}} \tag{53}$$
$$(C) = \left\|(\widetilde{P}_Y H_\lambda - H)C_X^{1/2}\right\|_{\mathrm{HS}} \tag{54}$$

Then, from Lemmas 3 to 5, we obtain

$$\|(\tilde{H}-H)S^{\#}\|_{\mathrm{HS}} \leq 2\sqrt{3}M\log(4/\delta)n^{-\frac{1}{2(1+\gamma_x)}} + 2\sqrt{3}\|H\|_{\mathrm{HS}}\|(I-\widetilde{P}_X)C_X^{1/2}\|_{\mathrm{op}} + \mathbb{E}_y\left[\left\|\left(\widetilde{P}_Y - I_{\mathcal{H}_y}\right)\psi(y)\right\|_{\mathcal{H}_y}^2\right]^{1/2}. \tag{55}$$

Then, notice that

$$\|(I-\widetilde{P}_X)C_X^{1/2}\|_{\mathrm{op}} \leq \|(I-\widetilde{P}_X)C_X^{1/2}\|_{\mathrm{HS}} \tag{56}$$

$$= \mathbb{E}_x\left[\left\|\left(\widetilde{P}_X - I_{\mathcal{H}_x}\right)\phi(x)\right\|_{\mathcal{H}_x}^2\right]^{1/2}. \tag{57}$$

We conclude by defining

$$c_1 = 2\sqrt{3}M, \tag{58}$$

$$c_2 = 2\sqrt{3}\|H\|_{\mathrm{HS}}. \tag{59}$$

□

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

$$(A) \leq 2M\log(4/\delta)n^{-\frac{1}{2(1+\gamma_x)}}. \tag{60}$$

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

Proof. We have

$$(A) \leq \underbrace{\left\|\left(S_Y^{\#}S_X - H_\lambda\widehat{C}_X\right)C_{X\lambda}^{-1/2}\right\|_{\mathrm{HS}}}_{(A.1)} \times \underbrace{\|C_{X\lambda}^{1/2}\tilde{\eta}(\widehat{C}_X)C_X^{1/2}\|_{\mathrm{op}}}_{(A.2)} \tag{61}$$

Moreover, we have

$$(A.2) \leq \|\widehat{C}_{X\lambda}^{1/2}\tilde{\eta}(\widehat{C}_X)\widehat{C}_{X\lambda}^{1/2}\|_{\mathrm{op}}\|\widehat{C}_{X\lambda}^{-1/2}C_{X\lambda}^{1/2}\|_{\mathrm{op}}^2\|C_{X\lambda}^{-1/2}C_X^{1/2}\|_{\mathrm{op}} \tag{62}$$

$$\leq \|\widehat{C}_{X\lambda}^{1/2}\tilde{\eta}(\widehat{C}_X)\widehat{C}_{X\lambda}^{1/2}\|_{\mathrm{op}}\|\widehat{C}_{X\lambda}^{-1/2}C_{X\lambda}^{1/2}\|_{\mathrm{op}}^2 \tag{63}$$

because $\|C_{X\lambda}^{-1/2}C_X^{1/2}\|_{\mathrm{op}} \leq 1$.

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

$$(A) \leq 2M\log(4/\delta)n^{-\frac{1}{2(1+\gamma_x)}}. \tag{64}$$

□

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

$$(B) \leq 2\sqrt{3}\|H\|_{\mathrm{HS}}(\lambda^{1/2} + \|(I-\widetilde{P}_X)C_X^{1/2}\|_{\mathrm{op}}) \tag{65}$$

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

$$\widehat{C}_X\tilde{\eta}(\widehat{C}_X) - I_{\mathcal{H}_x} = \widehat{C}_{X\lambda}\tilde{\eta}(\widehat{C}_X) - \lambda\tilde{\eta}(\widehat{C}_X) - I_{\mathcal{H}_x} \tag{66}$$

$$= (I-\widetilde{P}_X)\widehat{C}_{X\lambda}\tilde{\eta}(\widehat{C}_X) + \widetilde{P}_X\widehat{C}_{X\lambda}\tilde{\eta}(\widehat{C}_X) - \lambda\tilde{\eta}(\widehat{C}_X) - I_{\mathcal{H}_x} \tag{67}$$

$$= (I-\widetilde{P}_X)\widehat{C}_{X\lambda}\tilde{\eta}(\widehat{C}_X) - \lambda\tilde{\eta}(\widehat{C}_X) - (\widetilde{P}_X - I_{\mathcal{H}_x}), \tag{68}$$

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

Then, we have

$$(B) \leq \|H_\lambda\|_{\mathrm{HS}} \left\| \left(\widehat{C}_X \tilde{\eta}(\widehat{C}_X) - I_{\mathcal{H}_x} \right) C_X^{1/2} \right\|_{\mathrm{op}} \tag{69}$$

$$\leq \|H_\lambda\|_{\mathrm{HS}} \left(\|(I - \widetilde{P}_X)\widehat{C}_{X\lambda}\tilde{\eta}(\widehat{C}_X)C_X^{1/2}\|_{\mathrm{op}} + \lambda \|\tilde{\eta}(\widehat{C}_X)C_X^{1/2}\|_{\mathrm{op}} + \|(\widetilde{P}_X - I_{\mathcal{H}_x})C_X^{1/2}\|_{\mathrm{op}} \right) \tag{70}$$

But,

$$\|H_\lambda\|_{\mathrm{HS}} \leq \left\| H \left(C_X C_{X\lambda}^{-1} - I_{\mathcal{H}_x} \right) \right\|_{\mathrm{HS}} + \|H\|_{\mathrm{HS}} \tag{71}$$

$$= \left\| H \left(C_X - C_{X\lambda} \right) C_{X\lambda}^{-1} \right\|_{\mathrm{HS}} + \|H\|_{\mathrm{HS}} \tag{72}$$

$$= \lambda \left\| H C_{X\lambda}^{-1} \right\|_{\mathrm{HS}} + \|H\|_{\mathrm{HS}} \tag{73}$$

$$\leq 2\|H\|_{\mathrm{HS}}. \tag{74}$$

And,

$$\|(I - \widetilde{P}_X)\widehat{C}_{X\lambda}\tilde{\eta}(\widehat{C}_X)C_X^{1/2}\|_{\mathrm{op}} \leq \|(I - \widetilde{P}_X)\widehat{C}_{X\lambda}^{1/2}\|_{\mathrm{op}} \|\widehat{C}_{X\lambda}^{1/2}\tilde{\eta}(\widehat{C}_X)\widehat{C}_{X\lambda}^{1/2}\|_{\mathrm{op}} \|\widehat{C}_{X\lambda}^{-1/2}C_X^{1/2}\|_{\mathrm{op}}. \tag{75}$$

And,

$$\|(I - \widetilde{P}_X)\widehat{C}_{X\lambda}^{1/2}\|_{\mathrm{op}} \leq \|(I - \widetilde{P}_X)C_{X\lambda}^{1/2}\|_{\mathrm{op}} \|C_{X\lambda}^{-1/2}\widehat{C}_{X\lambda}^{1/2}\|_{\mathrm{op}}. \tag{76}$$

And,

$$\|(I - \widetilde{P}_X)C_{X\lambda}^{1/2}\|_{\mathrm{op}} \leq \|(I - \widetilde{P}_X)C_X^{1/2}\|_{\mathrm{op}} + \lambda^{1/2}. \tag{77}$$

Moreover,

$$\begin{aligned} \left\| \lambda\tilde{\eta}\left(\widehat{C}_X\right) C_X^{1/2} \right\|_{\mathrm{op}} &\leq \lambda \left\| \widehat{C}_{X\lambda}^{-1/2} \right\|_{\mathrm{op}} \left\| \widehat{C}_{X\lambda}^{1/2}\tilde{\eta}\left(\widehat{C}_X\right)\widehat{C}_{X\lambda}^{1/2} \right\|_{\mathrm{op}} \left\| \widehat{C}_{X\lambda}^{-1/2}C_{X\lambda}^{1/2} \right\|_{\mathrm{op}} \left\| C_{X\lambda}^{-1/2}C_X^{1/2} \right\|_{\mathrm{op}} \\ &\leq \lambda^{1/2} \left\| \widehat{C}_{X\lambda}^{1/2}\tilde{\eta}\left(\widehat{C}_X\right)\widehat{C}_{X\lambda}^{1/2} \right\|_{\mathrm{op}} \left\| \widehat{C}_{X\lambda}^{-1/2}C_{X\lambda}^{1/2} \right\|_{\mathrm{op}}. \end{aligned}$$

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

$$(B) \leq 4\sqrt{3}\|H\|_{\mathrm{HS}}(\lambda^{1/2} + \|(I - \widetilde{P}_X)C_X^{1/2}\|_{\mathrm{op}}) \tag{78}$$

□

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

$$(C) \leq \mathbb{E}_y \left[\left\| \left(\widetilde{P}_Y - I_{\mathcal{H}_y} \right) \psi(y) \right\|_{\mathcal{H}_y}^2 \right]^{1/2} + \lambda^{1/2} \|H\|_{\mathrm{HS}}. \tag{79}$$

Proof. We have

$$(C) = \left\| (\widetilde{P}_Y H(I_{\mathcal{H}_x} - \lambda C_{X\lambda}^{-1}) - H) C_X^{1/2} \right\|_{\mathrm{HS}} \tag{80}$$

$$\leq \left\| (\widetilde{P}_Y - I_{\mathcal{H}_y}) H C_X^{1/2} \right\|_{\mathrm{HS}} + \lambda^{1/2} \|H\|_{\mathrm{HS}} \tag{81}$$

$$= \mathbb{E}[\|(\widetilde{P}_Y - I_{\mathcal{H}_y})h^*(x)\|_{\mathcal{H}_y}^2]^{1/2} + \lambda^{1/2} \|H\|_{\mathrm{HS}}. \tag{82}$$

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

$$\begin{aligned} \mathbb{E}_x \left[\left\| \left(\widetilde{P}_Y - I_{\mathcal{H}_y} \right) h^*(x) \right\|_{\mathcal{H}_y}^2 \right] &= \mathbb{E}_x \left[\left\| \left(\widetilde{P}_Y - I_{\mathcal{H}_y} \right) \mathbb{E}_{\rho(y|x)}[\psi(y)] \right\|_{\mathcal{H}_y}^2 \right] \\ &= \mathbb{E}_x \left[\left\| \mathbb{E}_{\rho(y|x)} \left[\left(\widetilde{P}_Y - I_{\mathcal{H}_y} \right) \psi(y) \right] \right\|_{\mathcal{H}_y}^2 \right] \\ &\leq \mathbb{E}_x \left[\mathbb{E}_{\rho(y|x)} \left[\left\| \left(\widetilde{P}_Y - I_{\mathcal{H}_y} \right) \psi(y) \right\|_{\mathcal{H}_y}^2 \right] \right] \\ &= \mathbb{E}_y \left[\left\| \left(\widetilde{P}_Y - I_{\mathcal{H}_y} \right) \psi(y) \right\|_{\mathcal{H}_y}^2 \right]. \end{aligned}$$

□

D Sketching reconstruction error

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

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

$$m \geq c_4 \max\left(\nu_z^2 n^{\frac{\gamma_z+\mu_z}{1+\gamma_z}}, \nu_z^4 \log(1/\delta)\right), \tag{16}$$

then with probability $1-\delta$

$$\mathbb{E}_z[\|(\widetilde{P}_Z - I_{\mathcal{H}_z})\chi(z)\|_{\mathcal{H}_z}^2] \leq c_3 n^{-\frac{1-\gamma_z}{(1+\gamma_z)}} \tag{17}$$

where $c_3, c_4 > 0$ are constants independents of n, m, δ defined in the proofs.

Proof. For $t > 0$, we have

$$\begin{aligned}
\mathbb{E}_z\left[\left\|\left(\widetilde{P}_Z - I_{\mathcal{H}_z}\right)\chi(z)\right\|_{\mathcal{H}_z}^2\right] &= \mathrm{Tr}\left(\left(\widetilde{P}_Z - I_{\mathcal{H}_z}\right)\mathbb{E}_z\left[\chi(z)\otimes\chi(z)\right]\right) \\
&= \left\|\left(\widetilde{P}_Z - I_{\mathcal{H}_z}\right)C_Z^{1/2}\right\|_{\mathrm{HS}}^2 \\
&\leq \left\|\left(\widetilde{P}_Z - I_{\mathcal{H}_z}\right)\widehat{C}_{Zt}^{1/2}\right\|_{\mathrm{op}}^2 \left\|\widehat{C}_{Zt}^{-1/2}C_{Zt}^{1/2}\right\|_{\mathrm{op}}^2 \\
&\quad \times \left\|C_{Zt}^{-1/2}C_Z^{1/2}\right\|_{\mathrm{HS}}^2 .
\end{aligned}$$

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

$$\left\|\widehat{C}_{Zt}^{-1/2}C_{Zt}^{1/2}\right\|_{\mathrm{op}}^2 \leq 2\,. \tag{83}$$

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

$$\left\|{C_{Zt}}^{-1/2}C_Z^{1/2}\right\|_{\mathrm{HS}}^2 \leq Q_z t^{-\gamma_z}\,. \tag{84}$$

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

$$\mathbb{E}_y\left[\left\|\left(\widetilde{P}_Z - I_{\mathcal{H}_y}\right)\chi(z)\right\|_{\mathcal{H}_z}^2\right] \leq 6Q_z t^{1-\gamma_z}. \tag{85}$$

Finally, choosing $t = n^{-\frac{1}{1+\gamma}}$, defining $c_3 = 6Q_z$, $c_4 = 576\mathfrak{C}^2 b_z Q_z$, and noticing $\mathcal{N}_z^\infty(t) \leq b_z Q_z t^{-(\gamma_z+\mu_z)}$ (from Lemmas 11 and 12), allows to conclude the proof.

□

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

$$\left\|\left(\widetilde{P}_Z - I_{\mathcal{H}_z}\right)\widehat{C}_{Zt}^{1/2}\right\|_{\mathrm{op}}^2 \leq 3t\,. \tag{86}$$

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

$$\left\|\left(\widetilde{P}_Z - I_{\mathcal{H}_z}\right)\widehat{C}_{Zt}^{1/2}\right\|_{\mathrm{op}}^2 \leq \frac{t}{1-\beta_z(t)}\,, \tag{87}$$

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

Now, applying Lemma 7, with the condition

$$m_z \geq \max\left(432\mathfrak{C}^2\nu^2\mathcal{N}_z^\infty(t), 576\mathfrak{C}^2\nu^4\log\left(1/\delta\right)\right), \tag{88}$$

we obtain $\beta_z(t) \leq 2/3$, which gives

$$\left\|\left(\widetilde{P}_Z - I_{\mathcal{H}_y}\right)\widehat{C}_{Zt}^{1/2}\right\|_{\mathrm{op}}^2 \leq 3t\,. \tag{89}$$

□

Lemma 7. *Let R_z be as in Definition 1 and $\mathcal{N}_z^\infty(t)$ as in Definition 2. For all $\delta \in (0, 1/e]$, $\frac{9}{n}\log\left(\frac{n}{\delta}\right) \leq t \leq \|C_Z\|_{\mathrm{op}} - \frac{9}{n}\log\left(\frac{n}{\delta}\right)$ and $m_z \geq \max\left(6\mathcal{N}_z^\infty(t), \log\left(1/\delta\right)\right)$, with probability at least $1-\delta$,*

$$\left\|\widehat{C}_{Zt}^{-1/2}\left(\widehat{C}_Z - \widetilde{C}_Z\right)\widehat{C}_{Zt}^{-1/2}\right\|_{\mathrm{op}} \leq \mathfrak{C}\frac{2\sqrt{2}\nu\sqrt{6\mathcal{N}_z^\infty(t)} + 8\nu^2\sqrt{\log\left(1/\delta\right)}}{\sqrt{m_z}}\,, \tag{90}$$

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

Proof. We define the following random variables

$$W_i = \sqrt{\frac{m_z}{n}}\sum_{j=1}^n (R_z)_{ij}\widehat{C}_{Zt}^{-1/2}\chi(z_j) \in \mathcal{H}_z \quad \text{for } i = 1, \dots m_z. \tag{91}$$

In order to use the concentration bound given in Theorem 3, 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 $u \in \mathcal{H}_z$ and $v = \widehat{C}_{Zt}^{-1/2}u$, we have that $\langle W_i, u\rangle_{\mathcal{H}_z} = \sqrt{\frac{m_z}{n}}\sum_{j=1}^n (R_z)_{ij}v(z_j)$. Hence, using the definition of a sub-Gaussian sketch, we have

$$\begin{aligned}
\|\langle W_i, u\rangle_{\mathcal{H}_z}\|_{L_2(\mathbb{P})}^2 &= \mathbb{E}_R\left[|\langle W_i, u\rangle_{\mathcal{H}_z}|^2\right] && (92)\\
&= \frac{1}{n}\sum_{j=1}^n v(z_j)^2 && (93)\\
&< +\infty\,. && (94)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}_{R_z}\left[\exp\left(c\langle W_i, u\rangle_{\mathcal{H}_z}\right)\right] &= \mathbb{E}_{R_z}\left[\exp\left(\sum_{j=1}^n c\sqrt{\frac{m_z}{n}}R_{z_{ij}}v(z_j)\right)\right]\\
&= \prod_{j=1}^n \mathbb{E}_{R_z}\left[\exp\left(c\sqrt{\frac{m_z}{n}}R_{z_{ij}}v(z_j)\right)\right]\\
&\leq \prod_{j=1}^n \exp\left(\frac{c^2 m_z v(z_j)^2}{2n}\frac{\nu^2}{m_z}\right)\\
&= \exp\left(\frac{c^2\nu^2}{2n}\sum_{j=1}^n v(z_j)^2\right)\\
&= \exp\left(\frac{c^2\nu^2}{2}\|\langle W_i, u\rangle_{\mathcal{H}_z}\|_{L_2(\mathbb{P})}^2\right)\,.
\end{aligned}$$

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

$$\mathbb{E}_R\left[\exp\left(\frac{\langle W_i, u\rangle^2_{\mathcal{H}_z}}{8\nu^2 \left\|\langle W_i, u\rangle_{\mathcal{H}_z}\right\|^2_{L_2(\mathbb{P})}}\right) - 1\right] \leq 1\,. \tag{95}$$

We deduce that

$$\left\|\langle W_i, u\rangle_{\mathcal{H}_z}\right\|_{\varphi_2} \leq 2\sqrt{2}\nu \left\|\langle W_i, u\rangle_{\mathcal{H}_z}\right\|_{L_2(\mathbb{P})}\,. \tag{96}$$

We conclude that the W_is are subgaussian with $B = 2\sqrt{2}\nu$.

The W_is are pre-gaussian. We define $Z = \sqrt{\frac{m_z}{n}}\sum_{j=1}^n G_j \widehat{C}_{Zt}^{-1/2}\chi(z_j)$, with $G_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/m_z)$. 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 3, and by using also Lemma 14, and the condition $m_z \geq \max\left(6\mathcal{N}_z^\infty(t), \log\left(1/\delta\right)\right)$, we obtain

$$\left\|\widehat{C}_{Zt}^{-1/2}\left(\widehat{C}_Z - \widetilde{C}_Z\right)\widehat{C}_{Zt}^{-1/2}\right\|_{\text{op}} \leq \mathfrak{C}\frac{2\sqrt{2}\nu\sqrt{6\mathcal{N}_z^\infty(t)} + 8\nu^2\sqrt{\log\left(1/\delta\right)}}{\sqrt{m_z}}\,. \tag{97}$$

□

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

$$(A.1) \leq M\log(4/\delta)n^{-\frac{1}{2(1+\gamma_x)}} \tag{98}$$

where the constant M depends on $\kappa_y, \|H\|_{\text{HS}}, \delta$.

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 $\|\widehat{C}_{X\lambda}^{-1/2}C_{X\lambda}^{1/2}\|_{\text{op}}$ (Rudi et al., 2013, Lemma 3.6))**.** *If $\frac{9}{n}\log\frac{n}{\delta} \leq \lambda \leq \|C\|_{\text{op}}$, then with probability $1-\delta$*

$$\|\widehat{C}_{X\lambda}^{-1/2}C_{X\lambda}^{1/2}\|_{\text{op}} \leq \sqrt{2}. \tag{99}$$

□

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

$$\|C_{X\lambda}^{-1/2}\widehat{C}_{X\lambda}^{1/2}\|_{\text{op}} \leq \sqrt{\frac{3}{2}}. \tag{100}$$

Proof. We have

$$\|C_{X\lambda}^{-1/2}\widehat{C}_{X\lambda}^{1/2}\|_{\mathrm{op}} = \|C_{X\lambda}^{-1/2}\widehat{C}_{X\lambda}C_{X\lambda}^{-1/2}\|_{\mathrm{op}}^{1/2} \tag{101}$$

$$= \|I + C_{X\lambda}^{-1/2}(\widehat{C}_X - C_X)C_{X\lambda}^{-1/2}\|_{\mathrm{op}}^{1/2} \tag{102}$$

$$\leq \left(1 + \|C_{X\lambda}^{-1/2}(\widehat{C}_X - C_X)C_{X\lambda}^{-1/2}\|_{\mathrm{op}}\right)^{1/2} \tag{103}$$

$$\leq \sqrt{\frac{3}{2}} \tag{104}$$

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

Theorem 3 (Subgaussian 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 $\mathcal{H}_z$ 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 $1 - e^{-\tau}$,*

$$\|\hat{\Sigma} - \Sigma\| \leq \mathfrak{C}\|\Sigma\| \left(B\sqrt{\frac{\mathbf{r}(\Sigma)}{m}} \vee \frac{\mathbf{r}(\Sigma)}{m} \vee B^2\sqrt{\frac{\tau}{m}} \vee B^2\frac{\tau}{m} \right), \tag{105}$$

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

□

F Auxiliary results and definitions

Definition 2. For $t > 0$, we define the random variable

$$\mathcal{N}_z(t) = \langle \chi(z), C_{Zt}^{-1}\chi(z)\rangle_{\mathcal{H}_z} \tag{106}$$

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

$$d_{\mathrm{eff}}^Z(t) = \mathbb{E}\left[\mathcal{N}_z(t)\right] = \mathrm{Tr}\left(C_Z C_{Zt}^{-1}\right), \quad \mathcal{N}_z^\infty(t) = \sup_{z\in\mathcal{Z}} \mathcal{N}_z(t). \tag{107}$$

We note $\mathcal{N}_x^\infty, d_{\mathrm{eff}}^X(t), \gamma_x, Q_y, \mathcal{N}_y^\infty, d_{\mathrm{eff}}^Y(t), \gamma_y, Q_y$ for the input and output kernels k_x, k_y, respectively.

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

$$d_{\mathrm{eff}}^Z(t) \leq Q_z t^{-\gamma_z}. \tag{108}$$

Proof. We have

$$d_{\mathrm{eff}}^Z(t) = \mathrm{Tr}\left(C_Z C_{Zt}^{-1}\right) \tag{109}$$

$$\leq \mathrm{Tr}\left(C_Z^{\gamma_z}\right)\|C_Z^{1-\gamma_z}C_{Zt}^{-1}\|_{\mathrm{op}} \tag{110}$$

$$\leq Q_z t^{-\gamma_z}. \tag{111}$$

□

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

$$\mathcal{N}_z^\infty(t) \leq b_z d_{\mathrm{eff}}^Z(t) t^{-\mu_z}. \tag{112}$$

Proof. We have

$$\mathcal{N}_z^\infty(t) = \sup_{z\in\mathcal{Z}} \langle \chi(z), C_{Zt}^{-1}\chi(z)\rangle_{\mathcal{H}_z} \tag{113}$$

$$\leq b_z \,\mathrm{Tr}(C_{Zt}^{-1}C_Z^{1-\mu_z}) \tag{114}$$

$$\leq b_z \,\mathrm{Tr}(C_{Zt}^{-1}C_Z)\|C_{Zt}^{-\mu_z}\|_{\mathrm{op}} \tag{115}$$

$$\leq b_z d_{\mathrm{eff}}^Z(t) t^{-\mu_z}. \tag{116}$$

□

We recall the following deterministic bound.

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

$$\|\widehat{C}_{X\lambda}^{1/2}\tilde{\eta}(\widehat{C}_X)\widehat{C}_{X\lambda}^{1/2}\|_{\mathrm{op}} \leq 1. \tag{117}$$

We introduce here some notations and definitions from Koltchinskii and Lounici (2017). Let W be a centered random variable in $\mathcal{H}_z$, W is weakly square integrable iff $\|\langle W, u\rangle_{\mathcal{H}_z}\|_{L_2(\mathbb{P})}^2 := \mathbb{E}\left[|\langle W, u\rangle_{\mathcal{H}_z}|^2\right] < +\infty$, for any $u \in \mathcal{H}_z$. 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

$$\|\eta\|_\varphi = \inf\left\{C > 0 : \mathbb{E}\left[\varphi\left(|\eta|/C\right)\right] \leq 1\right\}. \tag{118}$$

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 $E = \mathcal{H}_z$.

Definition 3. A centered random variable X in $\mathcal{H}_z$ will be called sub-Gaussian iff, for all $u \in \mathcal{H}_z$, there exists $B > 0$ such that

$$\|\langle X, u\rangle_{\mathcal{H}_z}\|_{\phi_2} \leq B\,\|\langle X, u\rangle_{\mathcal{H}_z}\|_{L_2(\mathbb{P})}. \tag{119}$$

Definition 4. A weakly square integrable centered random variable X in $\mathcal{H}_z$ with covariance operator Σ is called pre-Gaussian iff there exists a centered Gaussian random variable Y in $\mathcal{H}_z$ with the same covariance operator Σ.

Lemma 14 (Expectancy, covariance, and intrinsic dimension of the W_is)**.** *Defining $W_i = \sqrt{\frac{m_z}{n}}\sum_{j=1}^n (R_z)_{ij}\widehat{C}_{Zt}^{-1/2}\chi(z_j) \in \mathcal{H}_z$ for $i = 1, \dots m_z$ where R_z is a subgaussian sketch, the following hold true*

$$\mathbb{E}_{R_z}[W_i] = 0 \tag{120}$$

$$\Sigma = \mathbb{E}_{R_z}[W_i \otimes W_i] = \widehat{C}_{Zt}^{-1/2}\widehat{C}_Z\widehat{C}_{Zt}^{-1/2} \tag{121}$$

$$\widehat{\Sigma} = \frac{1}{m_z}\sum_{i=1}^{m_z}\langle f, W_i\rangle_{\mathcal{H}_z}W_i = \widehat{C}_{Zt}^{-1/2}\widetilde{C}_Z\widehat{C}_{Zt}^{-1/2} \tag{122}$$

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

$$r(\Sigma) = \frac{\mathbb{E}_{R_z}\left[\|X_i\|_{\mathcal{H}_z}\right]^2}{\|\Sigma\|_{\mathrm{op}}} \leq 6\mathcal{N}_z^\infty(t). \tag{123}$$

Proof. First, it is straightforward to check that

$$\frac{1}{m_z}\sum_{i=1}^{m_z}\langle f, W_i\rangle_{\mathcal{H}_z}W_i = \widehat{C}_{Zt}^{-1/2}\widetilde{C}_Z\widehat{C}_{Zt}^{-1/2}. \tag{124}$$

Then, since $\mathbb{E}_{R_z}[(R_z)_{i:}] = 0$,

$$\mathbb{E}_{R_z}[W_i] = \sqrt{\frac{m_z}{n}}\widehat{C}_{Zt}^{-1/2}S_Z^\#\,\mathbb{E}_{R_z}[(R_z)_{i:}] = 0. \tag{125}$$

Then,

$$(W_i \otimes W_i)\,f = \langle f, W_i\rangle_{\mathcal{H}_z}W_i \tag{126}$$

$$= \langle f, \sqrt{m_z}\widehat{C}_{Zt}^{-1/2}S_Z^\#(R_z)_{i:}\rangle_{\mathcal{H}_z}\sqrt{m_z}\widehat{C}_{Zt}^{-1/2}S_Z^\#(R_z)_{i:} \tag{127}$$

$$= m_z\left((R_z)_{i:}^\top S_Z\widehat{C}_{Zt}^{-1/2}f\right)\widehat{C}_{Zt}^{-1/2}S_Z^\#(R_z)_{i:} \tag{128}$$

$$= \widehat{C}_{Zt}^{-1/2}S_Z^\#\left(m_z R_{z_{i:}}(R_z)_{i:}^\top\right)S_Z\widehat{C}_{Zt}^{-1/2}f, \tag{129}$$

and since $\mathbb{E}_{R_z}\left[m_z(R_z)_{i:}R_{z_{i:}}^\top\right] = I_n$,

$$\Sigma = \mathbb{E}_{R_z}\left[W_i \otimes W_i\right] \tag{130}$$
$$= \widehat{C}_{Zt}^{-1/2} S_Z^{\#}\, \mathbb{E}_{R_z}\left[m_z R_{z_{i:}} R_{z_{i:}}^\top\right] S_Z \widehat{C}_{Zt}^{-1/2} \tag{131}$$
$$= \widehat{C}_{Zt}^{-1/2} \widehat{C}_Z \widehat{C}_{Zt}^{-1/2}\,. \tag{132}$$

Then,

$$\mathbb{E}_{R_z}\left[\|X_i\|_{\mathcal{H}_z}\right]^2 \leq \mathbb{E}_{R_z}\left[\|X_i\|_{\mathcal{H}_z}^2\right] \quad \text{(by Jensen's inequality)} \tag{133}$$
$$= m_z\, \mathbb{E}_{R_z}\left[\langle \widehat{C}_{Zt}^{-1/2} S_Z^{\#} R_{z_{i:}}, \widehat{C}_{Zt}^{-1/2} S_Z^{\#} R_{z_{i:}} \rangle_{\mathcal{H}_z}\right] \tag{134}$$
$$= \frac{m_z}{n}\, \mathbb{E}_{R_z}\left[\langle \sum_{j=1}^n R_{z_{ij}} \chi(z_j), \sum_{l=1}^n R_{z_{il}} \widehat{C}_{Zt}^{-1} \chi(z_l) \rangle_{\mathcal{H}_z}\right] \tag{135}$$
$$= \frac{m_z}{n}\, \mathbb{E}_{R_z}\left[\sum_{j,l=1}^n R_{z_{ij}} R_{z_{il}} \langle \chi(z_j), \widehat{C}_{Zt}^{-1} \chi(z_l) \rangle_{\mathcal{H}_y}\right] \tag{136}$$
$$= \frac{m_z}{n} \sum_{j=1}^n \frac{1}{m_z} \langle \chi(z_j), \widehat{C}_{Zt}^{-1} \chi(z_j) \rangle_{\mathcal{H}_z} \tag{137}$$
$$= \mathrm{Tr}\left(\widehat{C}_{Zt}^{-1} \widehat{C}_Z\right) \tag{138}$$
$$= \left\| \widehat{C}_{Zt}^{-1/2} \widehat{C}_Z^{1/2} \right\|_{\mathrm{HS}}^2 \tag{139}$$
$$\leq \left\| \widehat{C}_{Zt}^{-1/2} C_{Zt}^{1/2} \right\|_{\mathrm{op}}^2 \left\| C_{Zt}^{-1/2} \widehat{C}_Z^{1/2} \right\|_{\mathrm{HS}}^2\,. \tag{140}$$

But,

$$\left\| C_{Zt}^{-1/2} \widehat{C}_Z^{1/2} \right\|_{\mathrm{HS}}^2 = \mathrm{Tr}\left(C_{Zt}^{-1} \widehat{C}_Z\right) \tag{141}$$
$$= \mathrm{Tr}\left(C_{Zt}^{-1} \left(\frac{1}{n} \sum_{i=1}^n \chi(z_i) \otimes \chi(z_i)\right)\right) \tag{142}$$
$$= \frac{1}{n} \sum_{i=1}^n \mathrm{Tr}\left(C_{Zt}^{-1} \left(\chi(z_i) \otimes \chi(z_i)\right)\right) \tag{143}$$
$$= \frac{1}{n} \sum_{i=1}^n \left\langle \chi(z_i), C_{Zt}^{-1} \chi(z_i) \right\rangle_{\mathcal{H}_y} \tag{144}$$
$$= \frac{1}{n} \sum_{i=1}^n \mathcal{N}_{z_i}(t) \tag{145}$$
$$\leq \mathcal{N}_z^\infty(t)\,. \tag{146}$$

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

$$\mathbb{E}_{R_z}\left[\|X_i\|_{\mathcal{H}_z}\right]^2 \leq 2\mathcal{N}_z^\infty(t). \tag{147}$$

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

We conclude that

$$\frac{\mathbb{E}_{R_z}\left[\|W_i\|_{\mathcal{H}_z}\right]^2}{\|\Sigma\|_{\mathrm{op}}} \leq 6\mathcal{N}_z^\infty(t). \tag{148}$$

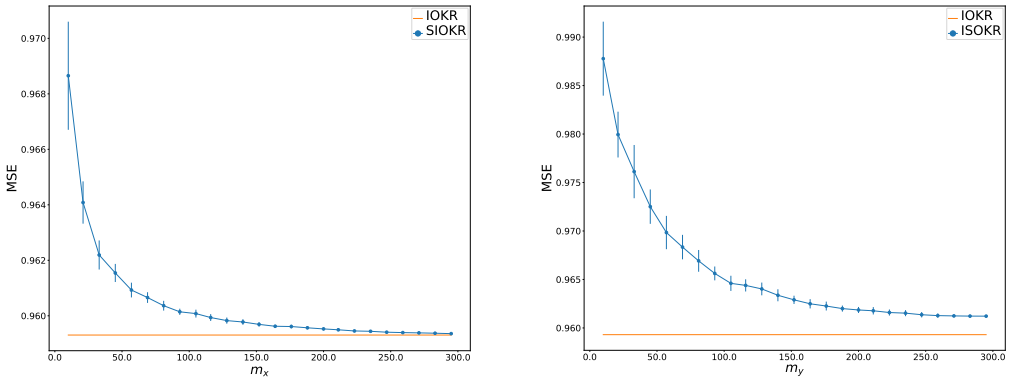

Figure 2: Test MSE with respect to m_x and m_y for a SIOKR and ISOKR model respectively with $(2 \cdot 10^{-3})$-SR input and output sketches.

Table 4: Multi-label data sets description.

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

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 $\frac{9}{n}\log\left(\frac{n}{\delta}\right) \leq t' \leq \|C_Z\|_{\text{op}}$, then $C_{Zt'} \preceq 2\widehat{C}_{Zt'}$, which implies $2\left\|\widehat{C}_Z\right\|_{\text{op}} \geq \|C_Z\|_{\text{op}} - t'$. Now, taking $t' = \frac{9}{n}\log\left(\frac{n}{\delta}\right)$, we obtain $\|C_Z\|_{\text{op}} - \frac{9}{n}\log\left(\frac{n}{\delta}\right) \leq 2\left\|\widehat{C}_Z\right\|_{\text{op}}$. □

G Contributions and Previous Work

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 will 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 subgaussian sketching instead of Nyström sketching as in Rudi et al. (2015).

H Additional Experiments

H.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 for SIOKR and ISOKR models.

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

In this section, you can find more details about training and testing sizes, number of features of the inputs and number of labels to predict of Bibtex and Bookmarks data sets in Table 4.