Projections Aléatoires Entrée, Sortie : Accélération de l'Apprentissage et de l'Inférence dans la Prédiction Structurée avec Noyaux

Tamim El Ahmad [1] & Luc Brogat-Motte [2] & Pierre Laforgue [3] & Florence d'Alché-Buc [4]

[1] *LTCI, Télécom Paris, IP Paris, France, tamim.elahmad@telecom-paris.fr*
[2] *L2S, CentraleSupélec, France, luc.brogat.motte@l2s.centralesupelec.fr*
[3] *Department of Computer Science, University of Milan, Italy, pierre.laforgue@unimi.it*
[4] *LTCI, Télécom Paris, IP Paris, France, florence.dalche@telecom-paris.fr*

Résumé. Grâce à l'utilisation de l'astuce du noyau dans les espaces d'entrée et de sorties, les méthodes subrogées à noyaux offrent une solution polyvalente avec des fondements théoriques au problème de prédiction structurée. Bien que sur des ensembles de données de tailles modérées elles constituent l'état de l'art, comme dans la chimie-informatique par exemple, elles échouent à passer à l'échelle lorsque le nombre de données d'entraînement devient élevé. Nous proposons d'équiper ces méthodes subrogées à noyaux avec des approximations à l'aide de projections aléatoires, appliquées aux noyaux d'entrée et de sortie. Nous prouvons une borne d'excès de risque sur l'estimateur du problème structuré, atteignant une vitesse de convergence proche de l'estimateur optimal avec des projecteurs de petite dimension en fonction des vitesses de décroissance des valeurs propres des opérateurs de covariance entrée/sortie. D'un point de vue computationnel, nous montrons que les deux approximations ont des impacts distincts mais complémentaires : en entrée on accélère principalement l'apprentissage, tandis qu'en sortie c'est l'inférence qui est accélérée.

Mots-clés. Méthodes à noyaux, Projections aléatoires, Apprentissage machine à grande échelle, Prédiction Structurée, Apprentissage statistique.

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

Keywords. Kernel methods, Sketching, Large-scale Machine Learning, Structured Prediction, Statistical learning.

1 Introduction

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

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

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

However, contrary to deep neural networks, they do not scale neither in memory nor in time without further approximation. The aim of this paper is to equip these methods with kernel approximations to obtain a drastic complexity reduction while maintaining their statistical properties. Several works have highlighted the power of kernel approximations, from Random Fourier Features (Rahimi and Recht, 2007; Brault et al., 2016; Rudi and Rosasco, 2017; Li et al., 2021), to general low-rank approaches (Bach, 2013; Meanti et al., 2020).

In this work we focus on sketching (Mahoney et al., 2011; Woodruff, 2014), a general dimension reduction method based on linear random projections. Applied to kernel approximation, sketching has been widely studied through Nyström's sub-sampling approximation (Williams and Seeger, 2001; Alaoui and Mahoney, 2015; Rudi et al., 2015), and further explored using Gaussian or Randomized Orthogonal Systems (Yang et al., 2017; Lacotte and Pilanci, 2020). Interpreted as a way to provide data-dependent random features (Williams and Seeger, 2001; Yang et al., 2012; Kpotufe and Sriperumbudur, 2020), this approach has allowed to scale up kernel PCA (Sterge and Sriperumbudur, 2022), kernel mean embedding

(Chatalic et al., 2022a,b) or independence tests (Kalinke and Szabó, 2023) while enjoying statistical guarantees. However, sketching has been limited so far to scalar kernel machines. No current approach covers both sides of the coin, i.e., applying approximations to both the input and output kernels. Motivated by surrogate structured prediction, we close this gap and make the following contributions:

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

Notations. We introduce now generic notations for the input (output) space and kernel. If $\mathcal{Z}$ denotes a generic Polish space, $\mathrm{k}_{\mathcal{Z}}$ is a positive definite kernel over $\mathcal{Z}$ and $\psi_{\mathcal{Z}}(z) \coloneqq \mathrm{k}_{\mathcal{Z}}(\cdot, z)$ is the canonical feature map of $\mathrm{k}_{\mathcal{Z}}$. $\mathcal{H}_{\mathcal{Z}}$ denotes the Reproducing Kernel Hilbert Space (RKHS) associated to $\mathrm{k}_{\mathcal{Z}}$. $\mathrm{S}_{\mathrm{Z}} : f \in \mathcal{H}_{\mathcal{Z}} \mapsto (1/\sqrt{n})(f(z_1), \ldots, f(z_n))^\top$ is the sampling operator over $\mathcal{H}_{\mathcal{Z}}$ (Smale and Zhou, 2007). For any operator A, we denote $A^{\#}$ its adjoint. The adjoint of S_{Z} is defined as ${\mathrm{S}_{\mathrm{Z}}}^{\#} : \alpha \in \mathbb{R}^n \mapsto (1/\sqrt{n}) \sum_{i=1}^n \alpha_i \, \psi_{\mathcal{Z}}(z_i)$. If z is a r.v. distributed according to $\rho_{\mathcal{Z}}$, its covariance operator over $\mathcal{H}_Z$ is $\mathrm{C}_{\mathcal{Z}} = \mathbb{E}_z[\psi_{\mathcal{Z}}(z) \otimes \psi_{\mathcal{Z}}(z)]$, and its empirical counterpart $\widehat{\mathrm{C}}_{\mathrm{Z}} = (1/n) \sum_{i=1}^n \psi_{\mathcal{Z}}(z_i) \otimes \psi_{\mathcal{Z}}(z_i) = {\mathrm{S}_{\mathrm{Z}}}^{\#} \mathrm{S}_{\mathrm{Z}}$, where $\{(z_i)_{i=1}^n\}$ is i.i.d. drawn from $\rho_{\mathcal{Z}}$. The Moore-Penrose inverse of M is denoted $M^\dagger$.

2 Background

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

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

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

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

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

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

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

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

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

Input Output ridge Kernel Regression. A common choice to tackle in practice the surrogate regression problem consists in solving a *kernel ridge regression problem*, leveraging kernels in both input and output spaces. The hypothesis space is chosen as a vector-valued Reproducing Kernel Hilbert Space (vv-RKHS) (Senkene and Tempel'man, 1973; Micchelli and Pontil, 2005; Carmeli et al., 2006, 2010). In the same way that RKHS are based on positive symmetric definite kernels, vv-RKHS are based on Operator-Valued Kernels (OVK). In our setting, we define an OVK $\mathcal{K}$, as a mapping $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{H}_{\mathcal{Y}})$, where $\mathcal{L}(\mathcal{H}_{\mathcal{Y}})$ is the set of bounded linear operators on $\mathcal{H}_{\mathcal{Y}}$, and that satisfies the properties recalled in Appendix B. An OVK $\mathcal{K}$ is uniquely associated with a vv-RKHS $\mathcal{H}$, i.e. a Hilbert space of functions from $\mathcal{X}$ to $\mathcal{H}_{\mathcal{Y}}$ that enjoys the reproducing kernel property (see Appendix B).

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

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

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

Pontil, 2005) the following expression:

$$\hat{h}(x) = \sum_{i=1}^{n} \hat{\alpha}_i(x)\, \psi_{\mathcal{Y}}(y_i)\,, \quad \text{with} \quad \hat{\alpha}(x) = \left(\mathrm{K_X} + n\lambda\right)^{-1} \mathrm{k_X^x} \coloneqq \widehat{\Omega}\, \mathrm{k_X^x}\,, \tag{3}$$

where $\mathrm{K_X} = \left(\mathrm{k}_{\mathcal{X}}\left(x_i, x_j\right)\right)_{i,j=1}^{n}$ and $\mathrm{k_X^x} = \left(\mathrm{k}_{\mathcal{X}}(x, x_1), \ldots, \mathrm{k}_{\mathcal{X}}(x, x_n)\right)$. On the other hand, using an operator view one obtains

$$\hat{h}(x) = \widehat{H}\, \psi_{\mathcal{X}}(x)\,, \quad \text{where} \quad \widehat{H} = {\mathrm{S_Y}}^{\#}\, \mathrm{S_X} \left(\widehat{\mathrm{C}}_{\mathrm{X}} + \lambda I\right)^{-1}. \tag{4}$$

The latter expression can be seen as a re-writing of the first (Ciliberto et al., 2016), echoing the KDE equations with finite-dimensional feature maps (Cortes et al., 2005). It can also be related to the conditional kernel empirical mean embedding (Grünewälder et al., 2012).

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

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

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

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

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

Sketching for kernel methods. Applied to kernel methods to reduce their dependency in n, sketching can be seen as linear projections induced by a random matrix R (the sketching matrix) drawn from a probability distribution over $\mathbb{R}^{m \times n}$, where $m \ll n$. Classic examples include Nyström's approximation, where each row of R is randomly drawn from the rows of the identity matrix I_n, and Gaussian sketches, where all entries of R are i.i.d. Gaussian random variables. Nyström's approximation acts as a random training data sub-sampler, but it can be interpreted in many ways. In Drineas et al. (2005); Bach (2013), it is shown to generate a low-rank approximation of the Gram matrix, while in Williams and Seeger (2001); Yang et al. (2012), it is seen as a way to construct data-dependent finite-dimensional random features. In Rudi et al. (2015), instead, it is presented as a projection onto a small subspace of the RKHS. For other sketching schemes such as Gaussian or Randomized Orthogonal Systems, most of the works adopt an optimization viewpoint, where a variable substitution is operated

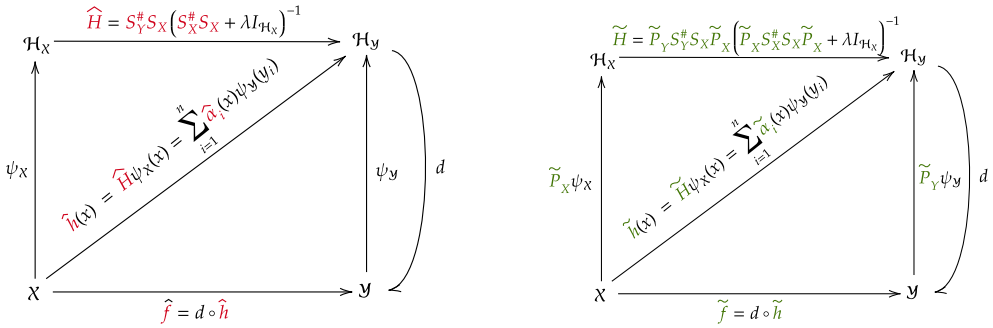

Figure 1: IOKR (left) and SISOKR (right) in the KDE setting. Note that SISOKR consists in IOKR when kernels $\mathrm{k}_{\mathcal{Z}}$ are replaced with their projected versions $\tilde{k}_{\mathcal{Z}}(\cdot,\cdot) = \langle \psi_{\mathcal{Z}}(\cdot), \widetilde{\mathrm{P}}_{\mathrm{Z}}\, \psi_{\mathcal{Z}}(\cdot)\rangle_{\mathcal{H}_{\mathcal{Z}}}$.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

after the application of a Representer theorem (Yang et al., 2017; Lacotte and Pilanci, 2020). An interesting view provided in Kpotufe and Sriperumbudur (2020) explores the construction of random features based on Gaussian sketching. All these works are however limited to sketching the *input* kernel, in scalar regression problems. In this work: (1) we generalize input sketching to vector-valued problems, (2) we sketch the outputs, which is critical to scale-up surrogate methods with kernelized outputs.

3 Sketched Input Sketched Output Kernel Regression

The goal of this section is to construct a low-rank estimator of $\hat{h}$ by using sketching on both the input and output kernels. Note that sketching the feature maps is not desirable here: if we replace the output features $\psi_{\mathcal{Y}}(y_i) \in \mathcal{H}_{\mathcal{Y}}$ with some sketch-dependent approximations $\tilde{\psi}_{\mathcal{Y}}(y_i) \in \mathbb{R}^m$ we become unable to compare the resulting $\tilde{h}$ to the target h^*. Indeed, $\tilde{h}$ is an approximation of $x \mapsto \mathbb{E}_y[\tilde{\psi}_{\mathcal{Y}}(y)|x]$, which is a biased version of h^* due to the sketch realization. Instead, as we show below, seeing sketching as orthogonal projections provides a natural way to solve our problem. Ultimately, this gives rise to an estimator $\tilde{f}$ for structured prediction which is versatile, easy-to-implement, theoretically-based and scalable to large data sets.

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

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

We now show how to design the projection operators using sketching and then derive the novel expression of the low-rank estimator in terms of a weighted combination of the training

outputs: $\tilde{h}(x) = \sum_{i=1}^{n} \tilde{\alpha}_i \, \psi_{\mathcal{Y}}(y_i)$, yielding a reduced computational cost. IOKR and SISOKR approaches are illustrated on Figure 1.

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

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

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

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

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

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

$$\tilde{\alpha}\,(x) = \mathrm{R}_{\mathcal{Y}}{}^{\top} \, \widetilde{\Omega} \, \mathrm{R}_{\mathcal{X}} \, \mathrm{k}_{\mathrm{X}}^{\mathrm{x}} \quad \textit{and} \quad \widetilde{\Omega} = \widetilde{\mathrm{K}}_{\mathrm{Y}}{}^{\dagger} \, \mathrm{R}_{\mathcal{Y}} \, \mathrm{K}_{\mathrm{Y}} \, \mathrm{K}_{\mathrm{X}} \, \mathrm{R}_{\mathcal{X}}{}^{\top} (\mathrm{R}_{\mathcal{X}} \, \mathrm{K}_{\mathrm{X}}{}^{2} \, \mathrm{R}_{\mathcal{X}}{}^{\top} + n\lambda \, \widetilde{\mathrm{K}}_{\mathrm{X}})^{\dagger} \, , \tag{8}$$

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

Note that the matrix quantity that we recover above, $\mathrm{K}_{\mathrm{X}} \, \mathrm{R}_{\mathcal{X}}{}^{\top} \left(\mathrm{R}_{\mathcal{X}} \, \mathrm{K}_{\mathrm{X}}{}^{2} \, \mathrm{R}_{\mathcal{X}}{}^{\top} + n\lambda \, \widetilde{\mathrm{K}}_{\mathrm{X}} \right)^{\dagger} \mathrm{R}_{\mathcal{X}} \, \mathrm{k}_{\mathrm{X}}^{\mathrm{x}}$, is typical to sketched kernel Ridge regression (Rudi et al., 2015; Yang et al., 2017). It allows to reduce the size of the matrix to invert, which is now an $\mathrm{m}_{\mathcal{X}} \times \mathrm{m}_{\mathcal{X}}$ matrix. This is the main reason for the reduction of the learning step's complexity, and is due to the input sketching (still, we need to perform matrix multiplication $\mathrm{R}_{\mathcal{X}} \, \mathrm{K}_{\mathrm{X}}$, whose efficiency depends on the sketch used). Note that output sketching also requires additional operations, but the overall cost of computing $\tilde{\alpha}$ remains negligible compared to $\mathcal{O}(n^3)$. We obtain the corresponding structured prediction estimator $\tilde{f}$ by decoding $\tilde{h}$, i.e., by replacing $\widehat{\Omega}$ by $\widetilde{\Omega}$ in (5). In fact, the main quantity we have to compute for prediction is now

$$\underbrace{\mathrm{K}_{\mathrm{X}}{}^{\mathrm{te,tr}} \, \mathrm{R}_{\mathcal{X}}{}^{\top}}_{n_{\mathrm{te}} \times \mathrm{m}_{\mathcal{X}}} \; \underbrace{\widetilde{\Omega}}_{\mathrm{m}_{\mathcal{X}} \times \mathrm{m}_{\mathcal{Y}}} \; \underbrace{\mathrm{R}_{\mathcal{Y}} \, \mathrm{K}_{\mathrm{Y}}{}^{\mathrm{tr},c}}_{\mathrm{m}_{\mathcal{Y}} \times n_c} \; . \tag{9}$$

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

4 Theoretical Analysis

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

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

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

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

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

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

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

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

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

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

$$\mathbb{E}_x \left[\|\tilde{h}(x) - h^*(x)\|^2_{\mathcal{H}_{\mathcal{Y}}} \right]^{\frac{1}{2}} \leq S(n,\delta) + c_2 A^{\psi_{\mathcal{X}}}_{\rho_{\mathcal{X}}}(\widetilde{P}_X) + A^{\psi_{\mathcal{Y}}}_{\rho_{\mathcal{Y}}}(\widetilde{P}_Y)\,, \tag{10}$$

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

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

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

Proof sketch. The proof relies on a decomposition of the operator $\widetilde{H}$ such that $\tilde{h}(x) = \widetilde{H}\psi_{\mathcal{X}}(x)$, see (44). The first term in (10) corresponds to the non-sketched kernel Ridge regression error, and the second term to the input sketching error. The latter extends both the results of Ciliberto et al. (2020) to sketched estimators, and that of Rudi et al. (2015) to the vector vector-valued case. The third term, i.e., the output sketching error is specific to our framework and derives from the expression of h^* and Jensen's inequality. □

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

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

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

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

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

$$\mathrm{m}_{\mathcal{Z}} \geq c_4 \max\left(\nu_{\mathcal{Z}}^2\, n^{\frac{\gamma_{\mathcal{Z}}+\mu_{\mathcal{Z}}}{1+\gamma_{\mathcal{Z}}}}, \nu_{\mathcal{Z}}^4 \log\left(1/\delta\right)\right), \tag{11}$$

with probability $1-\delta$ we have

$$\mathbb{E}_z\left[\|(\widetilde{\mathrm{P}}_{\mathrm{Z}} - I_{\mathcal{H}_{\mathcal{Z}}})\,\psi_{\mathcal{Z}}(z)\|_{\mathcal{H}_{\mathcal{Z}}}^2\right] \leq c_3\, n^{-\frac{1-\gamma_{\mathcal{Z}}}{1+\gamma_{\mathcal{Z}}}}, \tag{12}$$

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

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

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

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

Corollary 1 (SISOKR learning rates). *Consider the Assumptions of Theorems 1 and 2, that $\|\,\psi_{\mathcal{Y}}(y)\|_{\mathcal{H}_{\mathcal{Y}}} = \kappa_{\mathcal{Y}}$ for all $y \in \mathcal{Y}$, and $n \in \mathbb{N}$ such that $\frac{9}{n}\log(n/\delta) \leq n^{-\frac{1}{1+\gamma_{\mathcal{Z}}}} \leq \|\mathrm{C}_{\mathcal{Z}}\|_{\mathrm{op}}/2$ for $\mathcal{Z} \in \{\mathcal{X}, \mathcal{Y}\}$. Set*

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

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.