Fast Kernel Methods for Generic Lipschitz Losses via p-Sparsified Sketches

Anonymous authors
Paper under double-blind review

Abstract

Kernel methods are learning algorithms that enjoy solid theoretical foundations while suffering from important computational limitations. Sketching, which consists in looking for solutions among a subspace of reduced dimension, is a well-studied approach to alleviate these computational burdens. However, statistically-accurate sketches, such as the Gaussian one, usually contain few null entries, such that their application to kernel methods and their non-sparse Gram matrices remains slow in practice. In this paper, we show that sparsified Gaussian (and Rademacher) sketches still produce theoretically-valid approximations while allowing for important time and space savings thanks to an efficient *decomposition trick*. To support our method, we derive excess risk bounds for both single and multiple output kernel problems, with generic Lipschitz losses, hereby providing new guarantees for a wide range of applications, from robust regression to multiple quantile regression. Our theoretical results are complemented with experiments showing the empirical superiority of our approach over state-of-the-art sketching methods.

1 Introduction

Kernel methods hold a privileged position in machine learning, as they allow to tackle a large variety of learning tasks in a unique and generic framework, that of Reproducing Kernel Hilbert Spaces (RKHSs), while enjoying solid theoretical foundations (Steinwart & Christmann, 2008b; Scholkopf & Smola, 2018). From scalar-valued to multiple output regression (Micchelli & Pontil, 2005; Carmeli et al., 2006; 2010), these approaches play a central role in nonparametric learning, showing a great flexibility. However, when implemented naively, kernel methods raise major issues in terms of time and memory complexity, and are often thought of as limited to "fat data", i.e., datasets of reduced size but with a large number of input features. One way to scale up kernel methods are the Random Fourier Features (Rahimi & Recht, 2007; Rudi & Rosasco, 2017; Sriperumbudur & Szabó, 2015; Li et al., 2021), but they mainly apply to shift-invariant kernels. Another popular approach is to use sketching methods, first exemplified with Nyström approximations (Williams & Seeger, 2001; Drineas et al., 2005; Bach, 2013; Rudi et al., 2015). Indeed, sketching has recently gained a lot of interest in the kernel community due to its wide applicability (Yang et al., 2017; Lacotte et al., 2019; Kpotufe & Sriperumbudur, 2020; Lacotte & Pilanci, 2020; Gazagnadou et al., 2022) and its spectacular successes when combined to preconditioners and GPUs (Meanti et al., 2020).

Sketching as a random projection method (Mahoney et al., 2011; Woodruff, 2014) is rooted in the Johnson-Lindenstrauss lemma (Johnson & Lindenstrauss, 1984), and consists in working in reduced dimension subspaces while benefiting from theoretical guarantees. Learning with sketched kernels has mostly been studied in the case of scalar-valued regression, in particular in the emblematic case of Kernel Ridge Regression (Alaoui & Mahoney, 2015; Avron et al., 2017; Yang et al., 2017; Chen & Yang, 2021a). For several identified sketching types (e.g., Gaussian, Randomized Orthogonal Systems, adaptive sub-sampling), the resulting estimators come with theoretical guarantees under the form of the minimax optimality of the empirical approximation error. However, an important blind spot of the above works is their limitation to the square loss. Few papers go beyond Ridge Regression, and usually exclusively with sub-sampling schemes (Zhang et al., 2012; Li et al., 2016; Della Vecchia et al., 2021). In this work, we derive excess risk bounds for sketched kernel machines with generic Lipschitz-continuous losses, under standard assumption on the sketch matrix, solving an open problem from Yang et al. (2017). Doing so, we provide theoretical guarantees for a wide range of applications, from robust regression, based either on the Huber loss (Huber, 1964) or ϵ-insensitive losses (Steinwart & Christmann, 2008a), to quantile regression, tackled through the minimization of the pinball loss (Koenker, 2005). Further, we address this question in the general context of single and multiple output regression. Learning vector-valued functions using matrix-valued kernels (Micchelli & Pontil, 2005) have been primarily motivated by multi-task learning. Although equivalent in functional terms to scalar-valued kernels on pairs of input and tasks (Hein & Bousquet, 2004, Proposition 5), matrix-valued kernels (Álvarez et al., 2012) provide a way to define a larger variety of statistical learning problems by distinguishing the role of the inputs from that of the tasks. The computational and memory burden is naturally heavier in multi-task/multi-output regression, as the dimension of the output space plays an inevitable role, making approximation methods for matrix-valued kernel machines a crucial issue. To our knowledge, this work is the first to address this problem under the angle of sketching. It is however worth mentioning Baldassarre et al. (2012), who explored spectral filtering approaches for multiple output regression, and the generalization of Random Fourier Features to operator-valued kernels by Brault et al. (2016).

An important challenge when sketching kernel machines is that the sketched items, e.g., the Gram matrix, are usually dense. Plain sketchin matrices, such as the Gaussian one, then induce significantly more calculations than sub-sampling methods, which can be computed by applying a mask over the Gram matrix. Sparse sketching matrices (Clarkson & Woodruff, 2017; Nelson & Nguyên, 2013; Cohen, 2016; Derezinski et al., 2021) constitute an important line of research to reduce complexity while keeping good statistical properties when applied to sparse matrices (e.g., matrices induced by graphs), which is not the case of a Gram matrix. Motivated by these considerations, we analyze a family of sketches, unified under the name of p-sparsified sketches, that achieve interesting trade-offs between statistical accuracy (Gaussian sketches can be recovered as a particular case of p-sparsified sketches) and computational efficiency. The p-sparsified sketches are also memory-efficient, as they do not require computing and storing the full Gram matrix upfront. Besides theoretical analysis, we provide extensive experiments showing the superiority of p-sparsified sketches over SOTA approaches such as accumulation sketches (Chen & Yang, 2021a).

Contributions. Our goal is to provide a framework to speed-up both scalar and matrix-valued kernel methods which is as general as possible while maintaining good theoretical guarantees. For that purpose, we present three contributions, which may be of independent interest.

- We derive excess risk bounds for sketched kernel machines with generic Lipschitz-continuous losses, both in the scalar and multiple output cases. We hereby solve an open problem from Yang et al. (2017), and provide a first analysis to the sketching of vector-valued kernel methods.
- We show that sparsified Gaussian and Rademacher sketches provide valid approximations when applied to kernels methods. They maintain theoretical guarantees while inducing important space and computation savings, as opposed to plain sketches.
- We discuss how to learn these new sketched kernel machines, by means of an approximated feature map. We finally present experiments using Lipschitz-continuous losses, such as robust and quantile regression, on both synthetic and real-world datasets, supporting the relevance of our approach.

Notation. For any matrix $A \in \mathbb{R}^{m \times p}$, $A^\dagger$ is its pseudo-inverse, $\|A\|_{\mathrm{op}}$ its operator norm, $A_{i:} \in \mathbb{R}^p$ its i-th row, and $A_{:j} \in \mathbb{R}^m$ its j-th column. The identity matrix of dimension d is I_d. For a couple of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with distribution P, P_X is the marginal distribution of X. For $f : \mathcal{X} \longrightarrow \mathcal{Y}$, we use $\mathbb{E}\left[f\right] = \mathbb{E}_{P_X}\left[f(X)\right]$, $\mathbb{E}\left[\ell_f\right] = \mathbb{E}_P\left[\ell\left(f(X), Y\right)\right]$ and $\mathbb{E}_n[\ell_f] = \frac{1}{n}\sum_{i=1}^n \ell(f(x_i), y_i)$ for any function $\ell : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}$.

2 Sketching Kernels Machines with Lipschitz-Continuous Losses

In this section, we derive excess risk bounds for sketched kernel machines with generic Lipschitz losses, for both scalar and multiple output regression.

2.1 Scalar Kernel Machines

We consider a general regression framework, from an input space $\mathcal{X}$ to some scalar output space $\mathcal{Y} \subseteq \mathbb{R}$. Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ such that $z \mapsto \ell(z, y)$ is proper, lower semi-continuous and convex for every y, our goal is to estimate $f^* = \arg\inf_{f \in \mathcal{H}} \mathbb{E}_{(X,Y)\sim P} \left[\ell\left(f(X), Y\right)\right]$, where $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ is a hypothesis set, and P is a joint distribution over $\mathcal{X} \times \mathcal{Y}$. Since P is usually unknown, we assume that we have access to a training dataset $\{(x_i, y_i)\}_{i=1}^n$ composed of i.i.d. realisations drawn from P. We recall the definitions of a scalar-valued kernel and its RKHS (Aronszajn, 1950).

Definition 1 (Scalar-valued kernel)**.** *A scalar-valued kernel is a symmetric function* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *such that for all* $n \in \mathbb{N}$*, and any* $(x_i)_{i=1}^n \in \mathcal{X}^n$*,* $(\alpha_i)_{i=1}^n \in \mathbb{R}^n$*, we have* $\sum_{i,j=1}^n \alpha_i \, k\left(x_i, x_j\right) \alpha_j \geq 0$*.*

Theorem 1 (RKHS)**.** *Let* k *be a kernel on* $\mathcal{X}$*. Then, there exists a unique Hilbert space of functions* $\mathcal{H}_k \subset \mathcal{F}\left(\mathcal{X}, \mathbb{R}\right)$*, where* $\mathcal{F}\left(\mathcal{X}, \mathbb{R}\right)$ *denotes the set of functions from* $\mathcal{X}$ *to* $\mathbb{R}$*, such that* $k\left(\cdot, x\right) \in \mathcal{H}_k$ *for all* $x \in \mathcal{X}$*, and such that we have* $h\left(x\right) = \langle h, k\left(\cdot, x\right)\rangle_{\mathcal{H}_k}$ *for any* $(h, x) \in \mathcal{H}_k \times \mathcal{X}$*.*

A kernel machine computes a proxy for f^* by solving

$$\min_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^n \ell(f(x_i), y_i) + \frac{\lambda_n}{2} \|f\|_{\mathcal{H}_k}^2 \,, \tag{1}$$

where $\lambda_n > 0$ is a regularization parameter. By the representer theorem (Kimeldorf & Wahba, 1971; Schölkopf et al., 2001), the solution to Problem (1) is given by $\hat{f}_n = \sum_{i=1}^n \hat{\alpha}_i \, k(\cdot, x_i)$, with $\hat{\alpha} \in \mathbb{R}^n$ the solution to

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \ell([K\alpha]_i, y_i) + \frac{\lambda_n}{2} \alpha^\top K \alpha \,, \tag{2}$$

where $K \in \mathbb{R}^{n \times n}$ is the kernel Gram matrix such that $K_{ij} = k(x_i, x_j)$.

Definition 2 (Regularized Kernel-based Sketched Estimator)**.** *Given a random matrix* $S \in \mathbb{R}^{s \times n}$*, with* $s << n$*, sketching consists in imposing the substitution* $\alpha = S^\top \gamma$ *in the empirical risk minimization problem stated in Equation* (2)*. We then obtain an optimisation problem of reduced size on* γ*, that yields the sketched estimator* $\tilde{f}_s = \sum_{i=1}^n [S^\top \tilde{\gamma}]_i \, k(\cdot, x_i)$*, where* $\tilde{\gamma} \in \mathbb{R}^s$ *is a solution to*

$$\min_{\gamma \in \mathbb{R}^s} \frac{1}{n} \sum_{i=1}^n \ell([KS^\top \gamma]_i, y_i) + \frac{\lambda_n}{2} \gamma^\top S K S^\top \gamma \,. \tag{3}$$

The literature is rich in examples of distributions that can be used to generate the sketching matrix S. For instance, the sub-sampling matrices, where each line of S is sampled from the n^2-identity matrix, have been widely studied in the context of kernel methods. They are computationally efficient from both time and space perspectives, and yield the so-called Nyström approach (Williams & Seeger, 2001; Rudi et al., 2015). More complex distributions, such as Randomized Orthogonal System (ROS) sketching or Gaussian sketch matrices, have also been considered (Yang et al., 2017). In this work, we first give a general theoretical analysis of regularized kernel-based sketched estimators for any K-satisfiable sketch matrix (Definition 3). Then, we introduce the p-sparsified sketches and prove their K-satisfiablity, as well as their relevance for kernel methods in terms of statistical and computational trade-off.

Works about sketched kernel machines usually assess the performance of $\tilde{f}_s$ by upper bounding its squared $L^2(\mathbb{P}_N)$ error, i.e., $(1/n) \sum_{i=1}^n (\tilde{f}_s(x_i) - f_{\mathcal{H}_k}(x_i))^2$, where $f_{\mathcal{H}_k}$ is the minimizer of the true risk over $\mathcal{H}_k$, supposed to be attained (Yang et al., 2017, Equation 2), or through its (relative) recovery error $\|\tilde{f}_s - \hat{f}_n\|_{\mathcal{H}_k} / \|\hat{f}_n\|_{\mathcal{H}_k}$, see Lacotte & Pilanci (2020, Theorem 3). In contrast, we focus on the excess risk of $\tilde{f}_s$, the original quantity of interest. As revealed by the proof of Theorem 2, the approximation error of the excess of risk can be controlled in terms of the $L^2(\mathbb{P}_N)$ error, and we actually recover the results from Yang et al. (2017) when we particularize to the square loss with bounded outputs (second bound in Theorem 2). Furthermore, studying the excess risk allows to better position the performances of $\tilde{f}_s$ among the known off-the-shelf kernel-based estimators available for the targeted problem. To achieve this study, we rely on the key notion of K-satisfiability for a sketch matrix (Yang et al., 2017; Liu et al., 2019; Chen & Yang, 2021a).

Let $K/n = UDU^\top$ be the eigendecomposition of the Gram matrix, where $D = \operatorname{diag}(\mu_1, \ldots, \mu_n)$ stores the eigenvalues of K/n in decreasing order. Let δ_n^2 be the critical radius of K/n, i.e., the lowest value such that $\psi(\delta_n) = (\frac{1}{n}\sum_{i=1}^n \min(\delta_n^2, \mu_i))^{1/2} \leq \delta_n^2$. The existence and uniqueness of δ_n^2 is guaranteed for any unit ball in the RKHS associated with a positive definite kernel (Bartlett et al., 2006; Yang et al., 2017). Note that δ_n^2 is similar to the parameter $\tilde{\varepsilon}^2$ used in Yang et al. (2012) to analyze Nyström approximation for kernel methods. We define the statistical dimension of K as $d_n = \min\ \left\{j \in \{1, \ldots, n\} \colon \mu_j \leq \delta_n^2\right\}$, with $d_n = n$ if no such index j exists.

Definition 3 (K-satisfiability of a sketching matrix, Yang et al. 2017)**.** *Let $c > 0$ be independent of n, $U_1 \in \mathbb{R}^{n \times d_n}$ and $U_2 \in \mathbb{R}^{n \times (n-d_n)}$ be the the left and right blocks of the matrix U previously defined, and $D_2 = \operatorname{diag}(\mu_{d_n+1}, \ldots, \mu_n)$. A sketch matrix S is said to be K-satisfiable for c if we have*

$$\left\| (SU_1)^\top SU_1 - I_{d_n} \right\|_{\mathrm{op}} \leq 1/2\,, \quad \text{and} \quad \left\| SU_2 D_2^{1/2} \right\|_{\mathrm{op}} \leq c\delta_n\,. \tag{4}$$

Roughly speaking, a sketching matrix is K-satisfiable if it defines an isometry on the largest eigenvectors of K, and has a small operator norm on the smallest eigenvectors. For random sketching matrices, it is common to show K-satisfiability with high probability under some condition on the sketch size s, see e.g., Yang et al. (2017, Lemma 5) for Gaussian sketches, Chen & Yang (2021a, Theorem 8) for Accumulation sketches. In Section 3, we show similar results for p-sparsified sketches.

To derive our excess risk bounds, we place ourselves in the framework of Li et al. (2021), see Sections 2.1 and 3 therein. Namely, we assume that the true risk is minimized over $\mathcal{H}_k$ at $f_{\mathcal{H}_k} := \arg\inf_{f \in \mathcal{H}_k}\ \mathbb{E}\left[\ell\left(f\left(X\right), Y\right)\right]$. The existence of $f_{\mathcal{H}_k}$ is standard in the literature (Caponnetto & De Vito, 2007; Rudi & Rosasco, 2017; Yang et al., 2017), and implies that $f_{\mathcal{H}_k}$ has bounded norm, see e.g., Rudi & Rosasco (2017, Remark 2). Similarly to Li et al. (2021), we also assume that estimators returned by Empirical Risk Minimization have bounded norm. Hence, all estimators considered in the present paper belong to some ball of finite radius R. However, we highlight that our results do not require prior knowledge on R, and hold uniformly for all finite R. As a consequence, we consider without loss of generality as hypothesis set the unit ball $\mathcal{B}\left(\mathcal{H}_k\right)$ in $\mathcal{H}_k$, up to an *a posteriori* rescaling of the bounds by R to recover the general case.

Assumption 1. *The true risk is minimized at $f_{\mathcal{H}_k}$.*

Assumption 2. *The hypothesis set considered is $\mathcal{B}\left(\mathcal{H}_k\right)$.*

Assumption 3. *For all $y \in \mathcal{Y}$, $z \mapsto \ell(z, y)$ is L-Lipschitz.*

Assumption 4. *For all $x, x' \in \mathcal{X}$, we have $k(x, x') \leq \kappa$.*

Assumption 5. *The sketch S is K-satisfiable.*

Note that we discuss some directions to relax Assumption 2 in Appendix B. Many loss functions satisfy Assumption 3, such as the hinge loss ($L = 1$), used in SVMs (Cortes & Vapnik, 1995), the ϵ-insensitive ℓ_1 (Drucker et al., 1997), the κ-Huber loss, known for robust regression (Huber, 1964), the pinball loss, used in quantile regression (Steinwart & Christmann, 2011), or the square loss with bounded outputs. Assumption 4 is standard (e.g., $\kappa = 1$ for the Gaussian kernel). Under Assumptions 1 to 5 we have the following result.

Theorem 2. *Suppose that Assumptions 1 to 5 hold, and let $C = 1 + \sqrt{6}c$, with c the constant from Assumption 5. Then, for any $\delta \in (0, 1)$ with probability at least $1 - \delta$ we have*

$$\mathbb{E}\left[\ell_{\tilde{f}_s}\right] \leq \mathbb{E}\left[\ell_{f_{\mathcal{H}_k}}\right] + LC\sqrt{\lambda_n + \delta_n^2} + \frac{\lambda_n}{2} + 8L\sqrt{\frac{\kappa}{n}} + 2\sqrt{\frac{8\log\left(4/\delta\right)}{n}}\,. \tag{5}$$

Furthermore, if $\ell\left(z, y\right) = \left(z - y\right)^2/2$ and $\mathcal{Y} \subset [0, 1]$, with probability at least $1 - \delta$ we have

$$\mathbb{E}\left[\ell_{\tilde{f}_s}\right] \leq \mathbb{E}\left[\ell_{f_{\mathcal{H}_k}}\right] + \left(C^2 + \frac{1}{2}\right)\lambda_n + C^2\delta_n^2 + 8\frac{\kappa + \sqrt{\kappa}}{\sqrt{n}} + 2\sqrt{\frac{8\log\left(4/\delta\right)}{n}}\,. \tag{6}$$

Proof sketch. The proof relies on the decomposition of the excess risk into two generalization error terms and an approximation error term, i.e.,

$$\mathbb{E}[\ell_{\tilde{f}_s}] - \mathbb{E}[\ell_{f_{\mathcal{H}_k}}] = \mathbb{E}[\ell_{\tilde{f}_s}] - \mathbb{E}_n[\ell_{\tilde{f}_s}] + \mathbb{E}_n[\ell_{\tilde{f}_s}] - \mathbb{E}_n[\ell_{f_{\mathcal{H}_k}}] + \mathbb{E}_n[\ell_{f_{\mathcal{H}_k}}] - \mathbb{E}[\ell_{f_{\mathcal{H}_k}}]. \tag{7}$$

The two generalization errors (of $\tilde{f}_s$ and $f_{\mathcal{H}_k}$) can be bounded using Bartlett & Mendelson (2003, Theorem 8) together with Assumptions 1 to 4. For the last term, we can use Jensen's inequality and the Lipschitz continuity of the loss to upper bound this approximation error by the square root of the sum of the square residuals of the Kernel Ridge Regression with targets the $f_{\mathcal{H}_k}(x_i)$. The latter can in turn be upper bounded using Assumptions 1 and 5 and Lemma 2 from Yang et al. (2017). When considering the square loss, Jensen's inequality is not necessary anymore, leading to the improved second term in the right-hand side of the last inequality in Theorem 2. □

Recall that the rates in Theorem 2 are incomparable as is to that of Yang et al. (2017, Theorem 2), since we focus on the excess risk while the authors study the squared $L^2(\mathbb{P}_N)$ error. Precisely, we recover their results as a particular case with the square loss and bounded outputs, up to the generalization errors. Instead, note that we do recover the rates of Li et al. (2021, Theorem 1), based on a similar framework. Our bounds feature two different terms: a quantity related to the generalization errors, and a quantity governed by δ_n, deriving from the K-satisfiability analysis. The behaviour of the critical radius δ_n crucially depends on the choice of the kernel. In Yang et al. (2017), the authors compute its decay rate for different kernels. For instance, we have $\delta_n^2 = \mathcal{O}(\sqrt{\log(n)}/n)$ for the Gaussian kernel, $\delta_n^2 = \mathcal{O}(1/n)$ for polynomial kernels, or $\delta_n^2 = \mathcal{O}(n^{-2/3})$ for first-order Sobolev kernels. Note finally that by setting $\lambda_n \propto 1/\sqrt{n}$ we attain a rate of $\mathcal{O}(1/\sqrt{n})$, that is minimax for the kernel ridge regression, see Caponnetto & De Vito (2007).

Remark 1. *Note that a standard additional assumption on the second order moments of the functions in* $\mathcal{H}_k$ *(Bartlett et al., 2005) allows to derive refined learning rates for the generalization errors. These refined rates are expressed in terms of* $\hat{r}^{\star}_{\mathcal{H}_k}$*, the fixed point of a new sub-root function* $\hat{\psi}_n$*. In order to make the approximation error of the same order, it is then necessary to prove the* K*-satisfiability of* S *with respect to* $\hat{r}^{\star 2}_{\mathcal{H}_k}$ *instead of* δ_n^2*. Whether it is possible to prove such a* K*-satisfiability for standard sketches is however a nontrivial question, left as future work.*

2.2 Matrix-valued Kernel Machines

In this section, we extend our results to multiple output regression, tackled in vector-valued RKHSs. Note that the output space $\mathcal{Y}$ is now a subset of $\mathbb{R}^d$, with $d \geq 2$. We start by recalling important notions about Matrix-Valued Kernels (MVKs) and vector-valued RKHSs (vv-RKHSs).

Definition 4 (Matrix-valued kernel)**.** *An MVK is an application* $\mathcal{K}\colon \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathbb{R}^d)$*, where* $\mathcal{L}(\mathbb{R}^d)$ *is the set of bounded linear operators on* $\mathbb{R}^d$*, such that* $\mathcal{K}(x, x') = \mathcal{K}(x', x)^\top$ *for all* $(x, x') \in \mathcal{X}^2$*, and such that for all* $n \in \mathbb{N}$ *and any* $(x_i, y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ *we have* $\sum_{i,j=1}^n y_i^\top \mathcal{K}(x_i, x_j)\, y_j \geqslant 0$*.*

Theorem 3 (Vector-valued RKHS)**.** *Let* $\mathcal{K}$ *be an MVK. There is a unique Hilbert space* $\mathcal{H}_{\mathcal{K}} \subset \mathcal{F}(\mathcal{X}, \mathbb{R}^d)$*, the vv-RKHS of* $\mathcal{K}$*, such that for all* $x \in \mathcal{X}$*,* $y \in \mathbb{R}^d$ *and* $f \in \mathcal{H}_{\mathcal{K}}$ *we have* $x' \mapsto \mathcal{K}(x, x')\, y \in \mathcal{H}_{\mathcal{K}}$*, and* $\langle f, \mathcal{K}(\cdot, x)\, y \rangle_{\mathcal{H}} = f(x)^\top y$*.*

Note that we focus in this paper on the finite-dimensional case, i.e., $\mathcal{Y} \subset \mathbb{R}^d$, such that for all $x, x' \in \mathcal{X}$, we have $\mathcal{K}(x, x') \in \mathbb{R}^{d \times d}$. For a training sample $\{x_1, \ldots, x_n\}$, we define the Gram matrix as $\mathbf{K} = (\mathcal{K}(x_i, x_j))_{1 \leq i,j \leq n} \in \mathbb{R}^{nd \times nd}$. A common assumption consists in considering decomposable kernels: we assume that there exist a scalar kernel k and a positive semidefinite matrix $M \in \mathbb{R}^{d \times d}$ such that for all $x, x' \in \mathcal{X}$ we have $\mathcal{K}(x, x') = k(x, x')M$. The Gram matrix can then be written $\mathbf{K} = K \otimes M$, where $K \in \mathbb{R}^{n \times n}$ is the scalar Gram matrix, and $\otimes$ denotes the Kronecker product. Decomposable kernels are widely spread in the literature as they provide a good compromise between computational simplicity and expressivity —note that in particular they encapsulate independent learning, achieved with $M = I_d$. We now discuss two examples of relevant output matrices.

Example 1. *In joint quantile regression, one is interested in predicting* d *different conditional quantiles of an output* y *given the input* x*. If* $(\tau_i)_{i \leq d} \in (0, 1)$ *denote the* d *different quantile levels, it has been shown in*

Sangnier et al. (2016) that choosing $M_{ij} = \exp(-\gamma(\tau_i - \tau_j)^2)$ *favors close predictions for close quantile levels, while limiting crossing effects.*

Example 2. *In multiple output regression, it is possible to leverage prior knowledge on the task relationships to design a relevant output matrix* M*. For instance, let* P *be the* $d \times d$ *adjacency matrix of a graph in which the vertices are the tasks and an edge exists between two tasks if and only if they are (thought to be) related. Denoting by* L_P *the graph Laplacian associated to* P*, Evgeniou et al. (2005) and Sheldon (2008) have proposed to use* $M = (\mu L_P + (1-\mu) I_d)^{-1}$*, with* $\mu \in [0,1]$*. When* $\mu = 0$*, we have* $M = I_d$ *and all tasks are considered independent. When* $\mu = 1$*, we only rely on the prior knowledge encoded in* P*.*

Given a sample $(x_i, y_i)_{i=1}^n \in \left(\mathcal{X}, \mathbb{R}^d\right)^n$ and a decomposable kernel $\mathcal{K} = kM$ (its associated vv-RKHS is $\mathcal{H}_{\mathcal{K}}$), the penalized empirical risk minimisation problem is

$$\min_{f \in \mathcal{H}_{\mathcal{K}}} \frac{1}{n} \sum_{i=1}^n \ell(f(x_i), y_i) + \frac{\lambda_n}{2} \|f\|_{\mathcal{H}_{\mathcal{K}}}^2 , \tag{8}$$

where $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a loss such that $z \mapsto \ell(z, y)$ is proper, lower semi-continuous and convex for all $y \in \mathbb{R}^d$. By the vector-valued representer theorem (Micchelli & Pontil, 2005), we have that the solution to Problem (8) writes $\hat{f}_n = \sum_{j=1}^n \mathcal{K}(\cdot, x_j)\hat{\alpha}_j = \sum_{j=1}^n k(\cdot, x_j) M \hat{\alpha}_j$, where $\hat{A} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n)^\top \in \mathbb{R}^{n \times d}$ is the solution to the problem

$$\min_{A \in \mathbb{R}^{n \times d}} \frac{1}{n} \sum_{i=1}^n \ell\left([KAM]_{i:}^\top, y_i\right) + \frac{\lambda_n}{2} \operatorname{Tr}\left(KAMA^\top\right) .$$

In this context, sketching consists in making the substitution $A = S^\top \Gamma$, where $S \in \mathbb{R}^{s \times n}$ is a sketch matrix and $\Gamma \in \mathbb{R}^{s \times d}$ is the parameter of reduced dimension to be learned. The solution to the sketched problem is then $\tilde{f}_s = \sum_{j=1}^n k(\cdot, x_j) M \left[S^\top \tilde{\Gamma}\right]_{j:}$, with $\tilde{\Gamma} \in \mathbb{R}^{s \times d}$ minimizing

$$\frac{1}{n} \sum_{i=1}^n \ell\left(\left[KS^\top \Gamma M\right]_{i:}, y_i\right) + \frac{\lambda_n}{2} \operatorname{Tr}\left(SKS^\top \Gamma M \Gamma^\top\right) .$$

Theorem 4. *Suppose that Assumptions 1 to 5 hold, that* $\mathcal{K} = kM$ *is a decomposable kernel with* M *invertible, and let* C *as in Theorem 2. Then for any* $\delta \in (0,1)$ *with probability at least* $1 - \delta$ *we have*

$$\mathbb{E}\left[\ell_{\tilde{f}_s}\right] \leq \mathbb{E}\left[\ell_{f_{\mathcal{H}_{\mathcal{K}}}}\right] + LC\sqrt{\lambda_n + \|M\|_{\mathrm{op}}\, \delta_n^2} + \frac{\lambda_n}{2} + 8L\sqrt{\frac{\kappa \operatorname{Tr}(M)}{n}} + 2\sqrt{\frac{8 \log(4/\delta)}{n}} . \tag{9}$$

Furthermore, if $\ell(z, y) = \|z - y\|_2^2 / 2$ *and* $\mathcal{Y} \subset \mathcal{B}\left(\mathbb{R}^d\right)$*, with probability at least* $1 - \delta$ *we have that*

$$\mathbb{E}\left[\ell_{\tilde{f}_s}\right] \leq \mathbb{E}\left[\ell_{f_{\mathcal{H}_k}}\right] + \left(C^2 + \frac{1}{2}\right)\lambda_n + C^2 \|M\|_{\mathrm{op}}\, \delta_n^2 + 8 \operatorname{Tr}(M)^{1/2} \frac{\kappa \|M\|_{\mathrm{op}}^{1/2} + \kappa^{1/2}}{\sqrt{n}} + 2\sqrt{\frac{8 \log(4/\delta)}{n}} . \tag{10}$$

Proof sketch. The proof follows that of Theorem 2. The main challenge is to adapt Yang et al. (2017, Lemma 2) to the multiple output setting. To do so, we leverage that $\mathcal{K}$ is decomposable, such that the K-satisfiability of S is sufficient, where K the scalar Gram matrix. □

Note that for $M = I_d$ (independent prior), the third term of the right-hand side of both inequalities become of order $\sqrt{d/n}$, that is typical of multiple output problems. If moreover we instantiate the bound for $d = 1$, we recover exactly Theorem 2. To the best of our knowledge, Theorem 4 is the first theoretical result about sketched vector-valued kernel machines. We highlight that it applies to generic Lipschitz losses and provides a bound directly on the excess risk.

2.3 Algorithmic details

We now discuss how to solve single and multiple output optimization problems. Let $\{(\tilde{\mu}_i, \tilde{\mathbf{v}}_i), i \in [s]\}$ be the eigenpairs of $SKS^\top$ in descending order, $\tilde{U} = [\tilde{U}_{ij}]_{s\times s} = (\tilde{\mathbf{v}}_1, \dots, \tilde{\mathbf{v}}_s)$, $r = \operatorname{rank}(SKS^\top)$, and $\tilde{K}_r = \tilde{U}_r \tilde{D}_r^{-1/2}$, where $\tilde{D}_r = \operatorname{diag}(\tilde{\mu}_1, \dots, \tilde{\mu}_r)$, and $\tilde{U}_r = (\tilde{\mathbf{v}}_1, \dots, \tilde{\mathbf{v}}_r)$.

Proposition 1. *Solving Problem (3) is equivalent to solving*

$$\min_{\omega \in \mathbb{R}^r} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\omega^\top \mathbf{z}_S\left(x_i\right), y_i\right) + \frac{\lambda_n}{2} \|\omega\|_2^2 \,, \tag{11}$$

where $\mathbf{z}_S(x) = \tilde{K}_r^\top S \left(k\left(x, x_1\right), \dots, k\left(x, x_n\right)\right)^\top \in \mathbb{R}^r$.

Problem (11) thus writes as a linear problem with respect to the feature maps induced by the sketch, generalizing the results established in Yang et al. (2012) for sub-sampling sketches. When considering multiple outputs, it is also possible to derive a linear feature map version when the kernel is decomposable. These feature maps are of the form $\mathbf{z}_S \otimes M^{1/2}$, yielding matrices of size $nd \times rd$ that are prohibitive in terms of space, see Appendix E. Note that an alternative way is to see sketching as a projection of the $k(\cdot, x_i)$ into $\mathbb{R}^r$ (Chatalic et al., 2021). Instead, we directly learn Γ. For both single and multiple output problems, we consider losses not differentiable everywhere in Section 4 and apply ADAM Stochastic Subgradient Descent (Kingma & Ba, 2015) for its ability to handle large datasets.

Remark 2. *In the previous sections, sketching is always leveraged in primal problems. However, for some of the loss functions we consider, dual problems are usually more attractive (Cortes & Vapnik, 1995; Laforgue et al., 2020). This naturally raises the question of investigating the interplay between sketching and duality on the algorithmic level. More details can be found in Appendix F.*

3 p-Sparsified Sketches

We now introduce the p-sparsified sketches, and establish their K-satisfiability. The p-sparsified sketching matrices are composed of i.i.d. Rademacher or centered Gaussian entries, multiplied by independent Bernoulli variables of parameter p (the non-zero entries are scaled to ensure that S defines an isometry in expectation). The sketch sparsity is controlled by p, and when the latter becomes small enough, S contains many columns full of zeros. It is then possible to rewrite S as the product of a sub-Gaussian and a sub-sampling sketch of reduced size, which greatly accelerates the computations.

Definition 5. *Let $s < n$, and $p \in (0, 1]$. A p-Sparsified Rademacher (p-SR) sketching matrix is a random matrix $S \in \mathbb{R}^{s\times n}$ whose entries S_{ij} are independent and identically distributed (i.i.d.) as follows*

$$S_{ij} = \begin{cases} \frac{1}{\sqrt{sp}} & \text{with probability} \quad \frac{p}{2} \\ 0 & \text{with probability} \quad 1-p \\ \frac{-1}{\sqrt{sp}} & \text{with probability} \quad \frac{p}{2} \end{cases} \tag{12}$$

A p-Sparsified Gaussian (p-SG) sketching matrix is a random matrix $S \in \mathbb{R}^{s\times n}$ whose entries S_{ij} are i.i.d. as follows

$$S_{ij} = \begin{cases} \frac{1}{\sqrt{sp}} G_{ij} & \text{with probability} \quad p \\ 0 & \text{with probability} \quad 1-p \end{cases} \tag{13}$$

where the G_{ij} are i.i.d. standard normal random variables. Note that standard Gaussian sketches are a special case of p-SG sketches, corresponding to $p = 1$.

Several works partially addressed p-SR sketches in the past literature. For instance, Baraniuk et al. (2008) establish that p-SR sketches satisfy the Restricted Isometry Property (based on concentration results from Achlioptas (2001)), but only for $p = 1$ and $p = 1/3$. In Li et al. (2006), the authors consider generic p-SR sketches, but do not provide any theoretical result outside of a moment analysis. The *i.i.d. sparse embedding matrices* from Cohen (2016) are basically m/s-SR sketches, where $m \geq 1$, leading each column to have exactly m nonzero elements in expectation. However, we were not able to reproduce the proof of the Johnson-Linderstrauss property proposed by the author for his sketch (Theorem 4.2 in the paper, equivalent to the first claim of K-satisfiability, left-hand side of (4)). More precisely, we think that the assumptions considering "each entry is independently nonzero with probability m/s" and "each column has a fixed number of nonzero entries" (m here) are conflicting. As far as we know, this is the first time p-SG sketches are introduced in the literature. Note that both (12) and (13) can be rewritten as $S_{ij} = (1/\sqrt{sp})B_{ij}R_{ij}$, where the B_{ij} are i.i.d. Bernouilli random variables of parameter p, and the R_{ij} are i.i.d. random variables, independent from the B_{ij}, such that $\mathbb{E}[R_{ij}] = 0$ and $\mathbb{E}[R_{ij}R_{i'j'}] = 1$ if $i = i'$ and $j = j'$, and 0 otherwise. Namely, for p-SG sketches $R_{ij} = G_{ij}$ is a standard Gaussian variable while for p-SR sketches it is a Rademacher random variable. It is then easy to check that p-SR and p-SG sketches define isometries in expectation. In the next theorem, we show that p-sparsified sketches are K-satisfiable with high probability.

Theorem 5. *Let S be a p-sparsified sketching matrix. Then, there are some universal constants $C_0, C_1 > 0$ and a constant $c(p)$, increasing with p, such that for $s \geq \max\left(C_0 d_n/p^2, \delta_n^2 n\right)$ and with a probability at least $1 - C_1 e^{-sc(p)}$, the sketch S is K-satisfiable for $c = \frac{2}{\sqrt{p}}\left(1 + \sqrt{\log(5)}\right) + 1$.*

Proof sketch. To prove the left-hand side of (4), we use Boucheron et al. (2013, Theorem 2.13), which shows that any i.i.d. sub-Gaussian sketch matrix satisfies the Johnson-Lindenstrauss lemma with high probability. To prove the right-hand side of (4), we work conditionally on a realization of the B_{ij}, and use concentration results of Lipschitz functions of Rademacher or Gaussian random variables (Tao, 2012). We highlight that such concentration results do not hold for sub-Gaussian random variables in general, preventing from showing K-satisfiability of generic sparsified sub-Gaussian sketches. Note that having $S_{ij} \propto B_{ij}R_{ij}$ is key, and that sub-sampling uniformly at random non-zero entries instead of using i.i.d. Bernoulli variables would make the proof significantly more complex. We highlight that Theorem 5 strictly generalizes Yang et al. (2017, Lemma 5), recovered for $p = 1$, and extends the results to Rademacher sketches. ☐

Computational property of p-sparsified sketches. In addition to be statistically accurate, p-sparsified sketches are computationally efficient. Indeed, recall that the main quantity one has to compute when sketching a kernel machine is the matrix $SKS^\top$. With standard Gaussian sketches, that are known to be theoretically accurate, this computation takes $\mathcal{O}(sn^2)$ operations. Sub-sampling sketches are notoriously less precise, but since they act as masks over the Gram matrix K, computing $SKS^\top$ can be done in $\mathcal{O}(s^2)$ operations only, without having to store the entire Gram matrix upfront. Now, let $S \in \mathbb{R}^{s\times n}$ be a p-sparsified sketch, and $s' = \sum_{j=1}^n \mathbb{I}\{S_{:j} \neq 0_s\}$ be the number of columns of S with at least one nonzero element. The crucial observation that makes S computationally efficient is that we have

$$S = S_{\mathrm{SG}}\, S_{\mathrm{SS}}\,, \tag{14}$$

where $S_{\mathrm{SG}} \in \mathbb{R}^{s\times s'}$ is obtained by deleting the null columns from S, and $S_{\mathrm{SS}} \in \mathbb{R}^{s'\times n}$ is a sub-Sampling sketch whose sampling indices correspond to the indices of the columns in S with at least one non-zero entry[1]. We refer to (14) as the *decomposition trick*. This decomposition is key, as we can apply first a fast sub-sampling sketch, and then a sub-Gaussian sketch on the sub-sampled Gram matrix of reduced size. Note that s' is a random variable. By independence of the entries, each column is null with probability $(1-p)^s$. Then, by the independence of the columns we have that s' follows a Binomial distribution with parameters n and $1-(1-p)^s$, such that $\mathbb{E}[s'] = n(1-(1-p)^s)$.

Hence, the sparsity of the p-sparsified sketches, controlled by parameter p, is an interesting degree of freedom to add: it preserves statistical guarantees (Theorem 5) while speeding-up calculations (14). Of course, there is no free lunch and one looses on one side what is gained on the other: when p decreases (sparser sketches), the lower bound to get guarantees $s \gtrsim d_n/p^2$ increases, but the expected number of non-null columns s' decreases, thus accelerating computations (note that for $p = 1$ we exactly recover the lower bound and number of non-null columns for Gaussian sketches). By substituting $s = C_0 d_n/p^2$ into $\mathbb{E}[s']$, one can show that it is optimal to set $p \approx 0.7$, independently from C_0 and d_n. This value minimizes computations while maintaining the guarantees. However, the lower bound in Theorem 5 is a sufficient condition, that might be conservative.

---

[1] Precisely, S_{SS} is the identity matrix $I_{s'}$, augmented with $n - s'$ null columns inserted at the indices of the null columns of S.

Looking at the problem of setting s and p from the practitioner point of view, we also provide more aggressive empirical guidelines. Indeed, although this regime is not covered by Theorem 5, experiments show that setting s as for the Gaussian sketch and p smaller than $1/s$ yield very interesting results, see Figure 1(c). Overall, p-sparsified sketches (*i*) generalize Gaussian sketches by introducing sparsity as a new degree of freedom, (*ii*) enjoy a regime in which theoretical guarantees are preserved and computations (slightly) accelerated, (*iii*) empirically yield competitive results also in aggressive regimes not covered by theory, thus achieving a wide range of intesting accuracy/computations tradeoffs.

Related works. Sparse sketches have been widely studied in the literature, see Clarkson & Woodruff (2017); Nelson & Nguyên (2013); Derezinski et al. (2021). However these sketches are well-suited when applied to sparse matrices (e.g., matrices induced by graphs). In fact, given a matrix A, computing SA with these types of sketching has a time complexity of the order of $\text{nnz}\,(A)$, the number of nonzero elements of A. Besides, these sketches usually are constructed such that each column has at least one nonzero element (e.g. CountSketch, OSNAP), hence no *decomposition trick* is possible. Regarding kernel methods, since a Gram matrix is typically dense (e.g., with the Gaussian kernel, $\text{nnz}\,(K) = n^2$), and since no decomposition trick can be applied, one has to compute the whole matrix K and store it, such that time and space complexity implied by such sketches are of the order of n^2. In practice, we show that we can set p small enough to computationally outperform classical sparse sketches and still obtain similar statistical performance. Note that an important line of research is devoted to improve the statistical performance of Nyström's approximation, either by adaptive sampling (Kumar et al., 2012; Wang & Zhang, 2013; Gittens & Mahoney, 2013), or leverage scores (Alaoui & Mahoney, 2015; Musco & Musco, 2017; Rudi et al., 2018; Chen & Yang, 2021b). We took the opposite route, as p-SG sketches are accelerated but statistically degraded versions of the Gaussian sketch.

4 Experiments

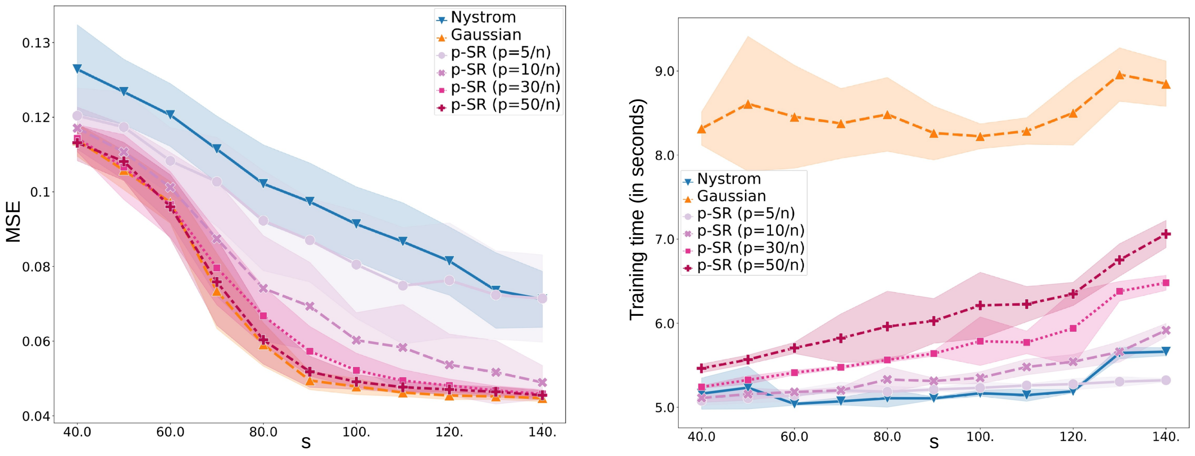
We now empirically compare the performance of p-sparsified sketches against state-of-he-art approaches, namely Nyström approximation (Williams & Seeger, 2001), Gaussian sketch (Yang et al., 2017), Accumulation sketch (Chen & Yang, 2021a), CountSketch (Clarkson & Woodruff, 2017) and Random Fourier Features (Rahimi & Recht, 2007). We chose not to benchmark ROS sketches as CountSketch has equivalent statistical accuracy while being faster to compute. Results reported are averaged over 30 replicates.

4.1 Scalar regression

Robust regression. We generate a dataset composed of $n = 10,000$ training datapoints: $9,900$ input points drawn i.i.d. from $\mathcal{U}\left([0_{10}, \mathbb{1}_{10}]\right)$ and 100 other drawn i.i.d. from $\mathcal{N}\left(1.5\mathbb{1}_{10}, 0.25 I_{10}\right)$. The outputs are generated as $y = f^{\star}(x) + \epsilon$, where $\epsilon \sim \mathcal{N}\left(0, 1\right)$ and

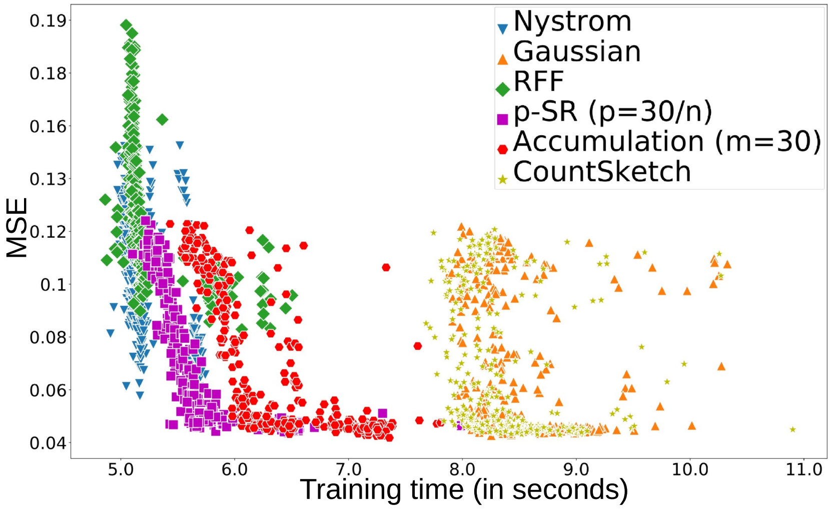
$$f^{\star}(x) = 0.1 e^{4x_1} + \frac{4}{1 + e^{-20(x_2 - 0.5)}} + 3x_3 + 2x_4 + x_5\,,$$

as introduced in Friedman (1991). We generate a test set of $n_{te} = 10,000$ points in the same way. We use the Gaussian kernel and select its bandwidth —as well as parameters λ_n and κ (and ϵ for ϵ-SVR)— via 5-folds cross-validation. We solve this 1D regression problem using the κ-Huber loss, described in Appendix G. We learn the sketched kernel machines for different values of s (from 40 to 140) and several values of p, the probability of being non-null in a p-SR sketch. Figure 1(a) presents the test error as a function of the sketch size s. Figure 1(b) shows the corresponding computational training time. All methods reduce their test error, measured in terms of the relative Mean Squared Error (MSE) when s increases. Note that increasing p increases both the precision and the training time, as expected. This behaviour recalls the Accumulation sketches, since we observe a form of interpolation between the Nyström and Gaussian approximations. The behaviour of all the different sketched kernel machines is shown in Figure 1(c), where each of them appears as a point (training time, test MSE). We observe that p-SR sketches attain the smallest possible error ($MSE \leq 0.05$) at the lowest training time budget (mostly around $5.6 < time < 6.6$). Moreover, p-SR sketches obtain a similar precision range as the Accumulation sketches, but for smaller training times (both approaches improve upon CountSketch and Gaussian sketch in that respect). Nyström sketching, which similarly to our approach does not need computing the entire Gram matrix, is fast to compute. The method is however



(a) Test relative MSE w.r.t. s with κ-Huber.

(b) Training time (seconds) w.r.t. s with κ-Huber.



(c) Test relative MSE w.r.t. training times with κ-Huber.

Figure 1: Trade-off between Accuracy and Efficiency for p-SR sketches with κ-Huber loss on synthetic dataset.

known to be sensitive to the non-homogeneity of the marginal distribution of the input data (Yang et al., 2017, Section 3.3). In contrast, the sub-Gaussian mixing matrix S_{SG} in (14) makes p-sparsified sketches more robust, as empirically shown in Figure 1(c). See Appendix H.1 for results on p-SG sketches.

4.2 Vector-valued regression

Joint quantile regression. We choose the quantile levels as follows $\tau = (0.1, 0.3, 0.5, 0.7, 0.9)$. We apply a subgradient algorithm to minimize the pinball loss described in Appendix G with ridge regularization and a kernel $\mathcal{K} = kM$ with M discussed in Example 1, and k a Gaussian kernel. We select regularisation parameter λ_n and bandwidth of kernel σ^2 via a 5-fold cross-validation. We showcase the behaviour of the proposed algorithm for Joint Sketched Quantile Regression on two datasets: the Boston Housing dataset (Harrison Jr & Rubinfeld, 1978), composed of 506 data points devoted to house price prediction, and the Fish Otoliths dataset (Moen et al., 2018; Ordoñez et al., 2020), dedicated to fish age prediction from images of otoliths (calcium carbonate structures), composed of a train and test sets of size 3780 and 165 respectively. The results are averages over 10 random $70\% - 30\%$ train-test splits for Boston dataset. For the Otoliths dataset we kept the initial given train-test split. The results are reported in Table 1. Sketching allows for a massive reduction of the training times while preserving the statistical performances. As a comparison, according to the results of Sangnier et al. (2016), the best benchmark result for the Boston dataset in terms of test pinball loss is 47.4, while best test crossing loss is 0.48, which shows that our implementation does not compete in terms of quantile prediction but preserves the non-crossing property.

Multi-output regression. We finally conducted experiments on multi-output kernel ridge regression. We used decomposable kernels, and took the largest datasets introduced in Spyromitros-Xioufis et al. (2016). They consist in four datasets, divided in two groups: River Flow (rf1 and rf2) both composed of 4108 training data, and Supply Chain Management (scm1d and scm20d) composed of 8145 and 7463 training data respectively (more details and additional results can be found in Appendix H.2). We compare our non-sketched decomposable matrix-valued kernel machine with the sketched version. For the sake of conciseness, we only report here the Average Relative Root Mean Squared Error (ARRMSE), see Table 2 and Appendix H.2. For all datasets, sketching shows strong computational improvements while maintaining the accuracy of non-sketched approaches.

Note that for both joint quantile regression and multi-output regression the results obtained after sketching (no matter the sketch chosen) are almost the same as that attained without sketching. It might be explained by two factors. First, the datasets studied have relatively small training sizes (from 354 training data for Boston to 8145 for scm1d). Second, predicting jointly multiple outputs is a complex task, so that it appears more natural to obtain less differences and variance using various types of sketches (or no sketch). However, in all cases sketching induces a huge time saver.

Table 1: Test pinball and crossing loss and training times (in seconds) with and without sketching ($s = 50$).

Dataset	Metrics	w/o Sketch	$20/n_{tr}$-SR	$20/n_{tr}$-SG	Acc. $m = 20$	CountSketch
Boston	Pinball loss	**51.28 ± 0.67**	54.75 ± 0.74	54.78 ± 0.72	54.73 ± 0.75	54.60 ± 0.72
	Crossing loss	0.34 ± 0.13	0.26 ± 0.08	0.11 ± 0.07	0.15 ± 0.07	**0.10 ± 0.05**
	Training time	6.97 ± 0.25	1.43 ± 0.07	1.38 ± 0.08	1.48 ± 0.05	**1.23 ± 0.07**
otoliths	Pinball loss	2.78	2.66 ± 0.02	**2.64 ± 0.02**	2.67 ± 0.03	2.65 ± 0.02
	Crossing loss	**5.18**	5.46 ± 0.06	5.43 ± 0.05	5.46 ± 0.06	5.44 ± 0.05
	Training time	606.8	20.4 ± 0.5	**20.0 ± 0.3**	22.1 ± 0.4	20.9 ± 0.3

Table 2: ARRMSE and training times (in sec) with square loss and $s = 100$ when using Sketching.

Dataset	Metrics	w/o Sketch	$20/n_{tr}$-SR	$20/n_{tr}$-SG	Acc. $m = 20$	CountSketch
rf1	ARRMSE	**0.575**	0.584 ± 0.003	0.583 ± 0.003	0.592 ± 0.001	**0.575 ± 0.0005**
	Training time	1.73	**0.22 ± 0.025**	0.25 ± 0.005	0.60 ± 0.0004	0.66 ± 0.013
rf2	ARRMSE	**0.578**	0.671 ± 0.009	0.656 ± 0.006	0.796 ± 0.006	0.715 ± 0.011
	Training time	1.77	0.28 ± 0.003	**0.27 ± 0.003**	0.82 ± 0.003	0.62 ± 0.001
scm1d	ARRMSE	**0.418**	0.422 ± 0.002	0.423 ± 0.001	0.423 ± 0.001	0.420 ± 0.001
	Training time	9.36	**0.45 ± 0.022**	**0.45 ± 0.019**	0.86 ± 0.006	2.49 ± 0.035
scm20d	ARRMSE	0.755	0.754 ± 0.003	0.754 ± 0.003	**0.753 ± 0.001**	0.754 ± 0.002
	Training time	6.16	**0.38 ± 0.016**	**0.38 ± 0.017**	0.70 ± 0.032	1.91 ± 0.047

5 Conclusion

We proposed excess-risk bounds for sketched kernel machines in the context of Lipschitz-continuous losses, with results valid for both scalar and matrix-valued kernels. We introduced a novel sketching scheme that leverages the good empirical statistical guarantees of the Gaussian Sketching while combining them with the low cost of Nyström sketching. Numerical experiments show that this novel scheme opens the door to many applications beyond the squared loss. Improvements on multi-output regression can certainly be obtained by applying low-rank considerations in the output space as well.

References

Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 274–281, 2001.

Ahmet Alacaoglu, Olivier Fercoq, and Volkan Cevher. Random extrapolation for primal-dual coordinate descent. In *Proc of the International Conference on Machine Learning (ICML)*, pp. 191–201. PMLR, 2020.

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

Mauricio A. Álvarez, Lorenzo Rosasco, and Neil D. Lawrence. Kernels for vector-valued functions: a review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, pp. 337–404, 1950.

Haim Avron, Kenneth L Clarkson, and David P Woodruff. Faster kernel ridge regression using sketching and preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1116–1138, 2017.

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

Luca Baldassarre, Lorenzo Rosasco, Annalisa Barla, and Alessandro Verri. Multi-output learning via spectral filtering. *Machine Learning*, 87(3):259–301, 2012.

Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2003.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Ann. Statist.*, 33 (4):1497–1537, 2005.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Romain Brault, Markus Heinonen, and Florence Buc. Random fourier features for operator-valued kernels. In *Asian Conference on Machine Learning*, pp. 110–125. PMLR, 2016.

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

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

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

Antonin Chambolle, Matthias J Ehrhardt, Peter Richtárik, and Carola-Bibiane Schonlieb. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2018.

Antoine Chatalic, Luigi Carratino, Ernesto De Vito, and Lorenzo Rosasco. Mean nyström embeddings for adaptive compressive learning, 2021.

Yifan Chen and Yun Yang. Accumulations of projections—a unified framework for random sketches in kernel ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 2953–2961. PMLR, 2021a.

Yifan Chen and Yun Yang. Fast statistical leverage score approximation in kernel ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 2935–2943. PMLR, 2021b.

Kenneth L. Clarkson and David P. Woodruff. Low-rank approximation and regression in input sparsity time. *J. ACM*, 63(6), jan 2017. ISSN 0004-5411. doi: 10.1145/3019134. URL `https://doi.org/10.1145/3019134`.

Michael B. Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the 2016 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 278–287, 2016.

Laurent Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of optimization theory and applications*, 158(2):460–479, 2013.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Andrea Della Vecchia, Jaouad Mourtada, Ernesto De Vito, and Lorenzo Rosasco. Regularized erm on random subspaces. In *International Conference on Artificial Intelligence and Statistics*, pp. 4006–4014. PMLR, 2021.

Michal Derezinski, Zhenyu Liao, E. Dobriban, and Michael W. Mahoney. Sparse sketches with small inversion bias. In *COLT*, 2021.

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

Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pp. 155–161, 1997.

Theodoros Evgeniou, Charles A. Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6(21):615–637, 2005.

Olivier Fercoq and Pascal Bianchi. A coordinate-descent primal-dual algorithm with large step size and possibly nonseparable functions. *SIAM Journal on Optimization*, 29(1):100–134, Jan 2019.

Jerome H Friedman. Multivariate adaptive regression splines. *The annals of statistics*, pp. 1–67, 1991.

Nidham Gazagnadou, Mark Ibrahim, and Robert M. Gower. Ridgesketch: A fast sketching based solver for large scale ridge regression. *SIAM Journal on Matrix Analysis and Applications*, 43(3):1440–1468, 2022. doi: 10.1137/21M1422963. URL `https://doi.org/10.1137/21M1422963`.

Alex Gittens and Michael Mahoney. Revisiting the nystrom method for improved large-scale machine learning. *Proceedings of the 30th International Conference on Machine Learning*, 28(3):567–575, 17–19 Jun 2013.

William Groves and Maria Gini. On optimizing airline ticket purchase timing. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(1):1–28, 2015.

David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.

Matthias Hein and Olivier Bousquet. Kernels, associated structures and generalizations. Technical Report 127, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, July 2004.

Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pp. 73–101, 1964.

William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space 26. *Contemporary mathematics*, 26:28, 1984.

George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Roger Koenker. *Quantile regression.* Cambridge university press, 2005.

Samory Kpotufe and Bharath K. Sriperumbudur. Gaussian sketching yields a J-L lemma in RKHS. In Silvia Chiappa and Roberto Calandra (eds.), *AISTATS 2020*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3928–3937. PMLR, 2020.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling methods for the nyström method. *J. Mach. Learn. Res.*, 13:981–1006, 2012.

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

Jonathan Lacotte, Mert Pilanci, and Marco Pavone. High-dimensional optimization in adaptive random subspaces. In *Proc. of the 33rd International Conference on Neural Information Processing Systems*, pp. 10847–10857, 2019.

Pierre Laforgue, Alex Lambert, Luc Brogat-Motte, and Florence d'Alché Buc. Duality in rkhss with infinite dimensional outputs: Application to robust losses. In *International Conference on Machine Learning*, pp. 5598–5607. PMLR, 2020.

Ping Li, Trevor J Hastie, and Kenneth W Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 287–296, 2006.

Zhe Li, Tianbao Yang, Lijun Zhang, and Rong Jin. Fast and accurate refined nyström-based kernel svm. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a unified analysis of random fourier features. *Journal of Machine Learning Research*, 22(108):1–51, 2021.

Meimei Liu, Zuofeng Shang, and Guang Cheng. Sharp theoretical analysis for nonparametric testing under random projection. In Alina Beygelzimer and Daniel Hsu (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2175–2209. PMLR, 25–28 Jun 2019.

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

Jiří Matoušek. *Lectures on Discrete Geometry.* Graduate Texts in Mathematics. Springer, 2013.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pp. 3–17. Springer, 2016.

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

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

Endre Moen, Nils Olav Handegard, Vaneeda Allken, Ole Thomas Albert, Alf Harbitz, and Ketil Malde. Automatic interpretation of otoliths using deep learning. *PLoS One*, 13(12):e0204713, 2018.

Cameron Musco and Christopher Musco. Recursive sampling for the nyström method. *Advances in Neural Information Processing Systems*, 2017:3834–3846, 2017.

Jelani Nelson and Huy L. Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 117–126, 2013. doi: 10.1109/FOCS.2013.21.

Alba Ordoñez, Line Eikvil, Arnt-Børre Salberg, Alf Harbitz, Sean Meling Murray, and Michael C Kampffmeyer. Explaining decisions of deep neural networks used for fish age prediction. *PloS one*, 15(6):e0235013, 2020.

Ali Rahimi and B. Recht. Random features for large scale kernel machines. *NIPS*, 20:1177–1184, 01 2007.

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

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, 28, 2015.

Alessandro Rudi, Daniele Calandriello, Luigi Carratino, and Lorenzo Rosasco. On fast leverage score sampling and optimal learning. In *NeurIPS*, 2018.

Maxime Sangnier, Olivier Fercoq, and Florence d'Alché Buc. Joint quantile regression in vector-valued RKHSs. In *Advances in Neural Information Processing Systems (NeurIPS)*, Barcelona, France, December 2016.

Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* Adaptive Computation and Machine Learning Series, 2018.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *Computational Learning Theory: 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001 Amsterdam, The Netherlands, July 16–19, 2001 Proceedings 14*, pp. 416–426. Springer, 2001.

Daniel Sheldon. Graphical multi-task learning, 2008.

Eleftherios Spyromitros-Xioufis, Grigorios Tsoumakas, William Groves, and Ioannis Vlahavas. Multi-target regression via input space expansion: treating targets as inputs. *Machine Learning*, 104(1):55–98, 2016.

Bharath K Sriperumbudur and Zoltán Szabó. Optimal rates for random fourier features. In *NIPS*, 2015.

Ingo Steinwart and Andreas Christmann. Sparsity of svms that use the epsilon-insensitive loss. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou (eds.), *Advances in Neural Information Processing Systems 21 (NeurIPS)*, pp. 1569–1576. Curran Associates, Inc., 2008a.

Ingo Steinwart and Andreas Christmann. *Support vector machines.* Springer Science & Business Media, 2008b.

Ingo Steinwart and Andreas Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 2011.

Michel Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques*, 81(1):73–205, 1995.

Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

Bang Cong Vu. A splitting algorithm for dual monotone inclusions involving cocoercive operators, 2011.

Shusen Wang and Zhihua Zhang. Improving cur matrix decomposition and the nyström approximation via adaptive sampling. *J. Mach. Learn. Res.*, 14(1):2729–2769, jan 2013.

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

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10 (1-2):1–157, 2014. doi: 10.1561/0400000060. URL `https://doi.org/10.1561/0400000060`.

Tianbao Yang, Yu-feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

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

Kai Zhang, Liang Lan, Zhuang Wang, and Fabian Moerchen. Scaling up kernel svm on limited resources: A low-rank linearization approach. In Neil D. Lawrence and Mark Girolami (eds.), *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22, pp. 1425–1434, 2012.