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

T. El Ahmad⋆, P. Laforgue† and F. d'Alché-Buc⋆

⋆ LTCI, Télécom Paris, Institut Polytechnique de Paris † Università degli Studi di Milano

Motivations

Problem: learn $\hat{f} : \mathcal{X} \to \mathbb{R}^d$, where $d \geq 1$, i.e. **multi-output** regression, via **kernel machines** and with a **large number of training data** $\boldsymbol{n}$

Existing works: 4 points are of particular interest, and many works tackle some of them, e.g.:

1. [2] tackles **scalability** to large data sets;
2. [3] goes **beyond least squares**;
3. [4] gives **excess risk bounds**;
4. [5] studies **multi-output** regression.

Goals:

- Provide a general framework to solve large-scale multi-output regression via **decomposable kernels** and **sketching**.
- Derive **excess risk bounds** for such estimator with a **Lipschitz loss** and a $\boldsymbol{K}$**-satisfiable sketch**.
- Provide a new $\boldsymbol{K}$**-satisfiable sketch** sketching distribution adapted to kernel methods, i.e. reducing **time and space complexities**.

Sketched Kernel Machines

Let $\mathcal{K} = kM$, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ p. d. kernel, $M \in \mathbb{R}^{d \times d}$, $\mathcal{H}$ vv-RKHS of $\mathcal{K}$,

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

Non-sketched estimator: $\hat{f} = \sum_{j=1}^{n} k(\cdot, x_j) M \hat{A}_{j:}$, with $\hat{A} \in \mathbb{R}^{n \times d}$ sol. to

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

Sketched estimator: let $s \ll n$ and $S \in \mathbb{R}^{s \times n}$ a random matrix, $\widetilde{K} = SKS^\top$, $\tilde{f} = \sum_{j=1}^{n} k(\cdot, x_j) M [S^\top \tilde{\Gamma}]_{j:}$, with $\tilde{\Gamma} \in \mathbb{R}^{s \times d}$ sol. to

$$\min_{\Gamma \in \mathbb{R}^{s \times d}} \frac{1}{n} \sum_{i=1}^{n} \ell\left([KS^\top \Gamma M]_{i:}, y_i\right) + \frac{\lambda}{2} \operatorname{Tr}\left(\widetilde{K} \Gamma M \Gamma^\top\right)$$

$\Longrightarrow$ from $\boldsymbol{n} \times d$ to $\boldsymbol{s} \times d$ parameters to learn!

K-Satisfiability

Let $K/n = UDU^\top$ (SVD), δ_n^2 the lowest value s. t. $\psi(\delta_n) = (\frac{1}{n}\sum_{i=1}^{n} \min(\delta_n^2, \lambda_i))^{1/2} \leq \delta_n^2$, $d_n = \min \{j \in \{1, \ldots, n\} \colon \lambda_j \leq \delta_n^2\}$, $U_1 \in \mathbb{R}^{n \times d_n}$ and $U_2 \in \mathbb{R}^{n \times (n - d_n)}$ the left and right blocks of U, D_2 the bottom right $(n - d_n)^2$-sub-matrix of D.

Definition 1 (K-satisfiability [1]) *Let $c > 0$ be independent of n. A sketch matrix S is said to be K-satisfiable for c if we have*

$$\left\| (SU_1)^\top SU_1 - I_{d_n} \right\|_{\mathrm{op}} \leq 1/2 \,,$$

$$\left\| SU_2 D_2^{1/2} \right\|_{\mathrm{op}} \leq c\delta_n \,.$$

Intuition: S is K-satisfiable $\Longrightarrow$ isometry on the largest eigenvectors of K/n and small operator norm on the smallest eigenvectors.

Excess Risk Bounds

A. 1: Expected risk is minimized over $\mathcal{H}$ at $f_{\mathcal{H}} = \operatorname{arginf}_{f \in \mathcal{H}} \ \mathbb{E}\left[\ell\left(f\left(X\right), Y\right)\right]$.

A. 2: The hypothesis set considered is the unit ball $\mathcal{B}\left(\mathcal{H}\right)$ of $\mathcal{H}$.

A. 3: $\forall\, y \in \mathbb{R}^d$, $z \mapsto \ell(z, y)$ is L-Lipschitz over $\mathcal{H}(\mathcal{X}) = \{f\left(x\right) : f \in \mathcal{H},\, x \in \mathcal{X}\}$.

A. 4: $\exists\, \kappa > 0$ s. t. $k(x, x) \leq \kappa$, $\forall\, x \in \mathcal{X}$ and M is non-singular.

A. 5: The sketch S is K-satisfiable for a $c > 0$ independent of n.

Theorem 2 *Under **A. 1, 2, 3, 4 and 5**, let $C = 1 + \sqrt{6}c$, for any $\delta \in (0, 1)$, then with probability at least $1 - \delta$,*

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

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

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

p-Sparsified Sketches

Definition 3 *Let $s < n$, $p \in (0, 1]$. A p-sparsified sketch $S \in \mathbb{R}^{s \times n}$ is composed of i.i.d. entries*

$$S_{ij} = \frac{1}{\sqrt{sp}} B_{ij} R_{ij} \,,$$

where $B_{ij} \overset{i.i.d.}{\sim}$ Ber(p) and $R_{ij} \overset{i.i.d.}{\sim}$ Rad($\frac{1}{2}$) (p-SR) or $\mathcal{N}\left(0, 1\right)$ (p-SG).

Theorem 4 *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 $c = \frac{2}{\sqrt{p}}\left(1 + \sqrt{\log(5)}\right) + 1$.*

Decomposition trick: $s' = \sum_{j=1}^{n} \mathbb{I}\{S_{:j} \neq 0_s\} \sim \operatorname{Binom}\left(n, 1 - (1 - p)^s\right) \implies \mathbb{E}\left[s'\right] = n(1 - (1 - p)^s) \underset{p \to 0}{\sim} nsp$,

$$\boldsymbol{S} = \boldsymbol{S_{\mathrm{SG}}}\, \boldsymbol{S_{\mathrm{SS}}}$$

· $S_{\mathrm{SG}} \in \mathbb{R}^{s \times s'}$: **sparse sub-gaussian sketch** obtained by deleting the null columns from S

· $S_{\mathrm{SS}} \in \mathbb{R}^{s' \times n}$: **sub-sampling sketch** obtained by sampling the rows of I_n corresponding to the indices of non-zero columns of S

Let C_k = cost of computing $k(x, x')$, complexities of Gaussian vs p-sparsified sketch:

Time: $\mathcal{O}\left(C_k n^2 + n^2 s\right)$ vs $\mathcal{O}\left(C_k n^2 sp + n^2 s^2 p\right)$

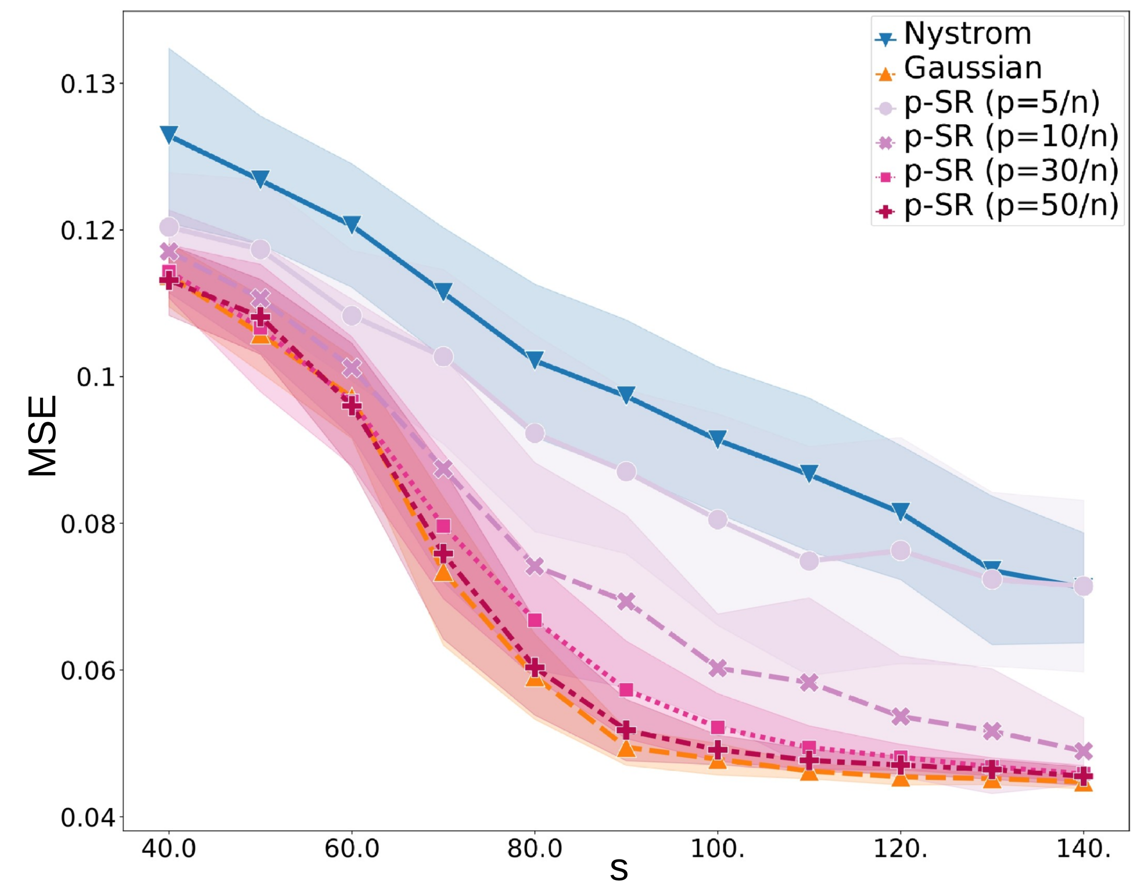
Space: $\mathcal{O}\left(n^2\right)$ vs $\mathcal{O}\left(n^2 sp\right)$

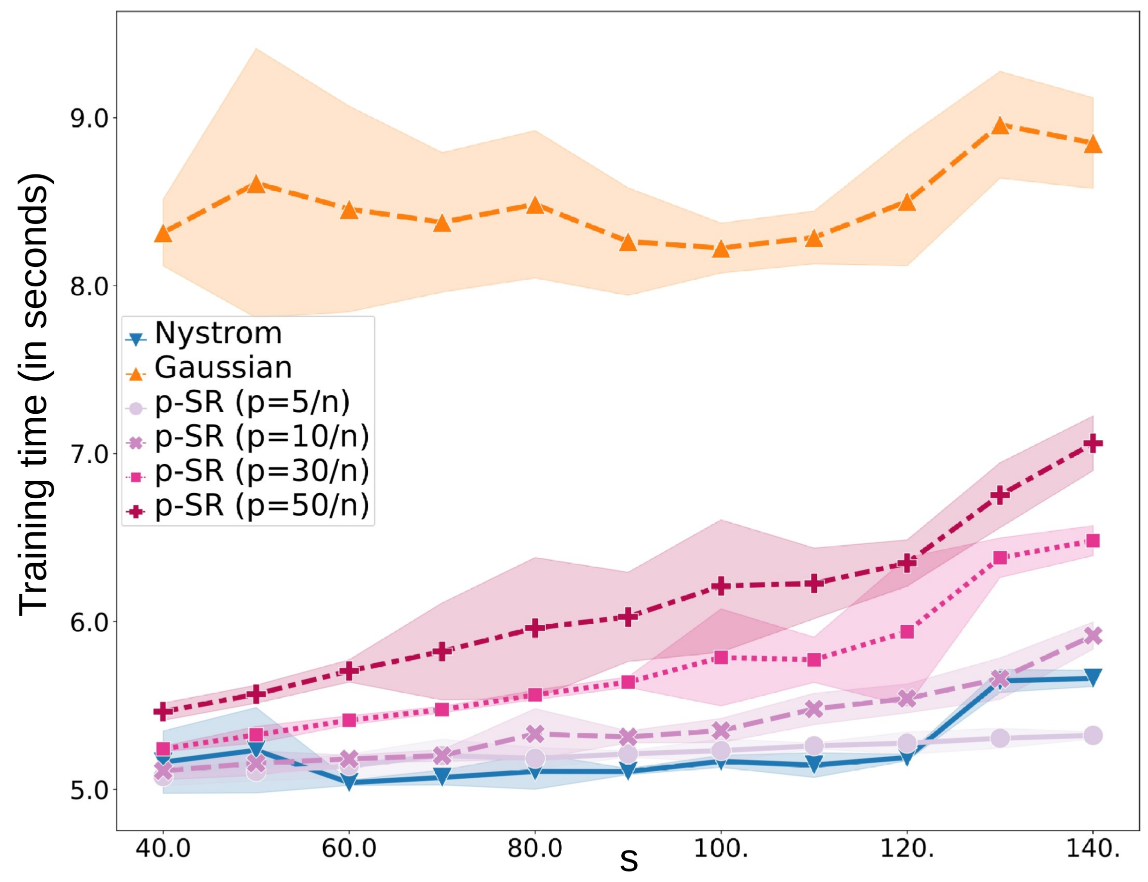
Acknowledgements

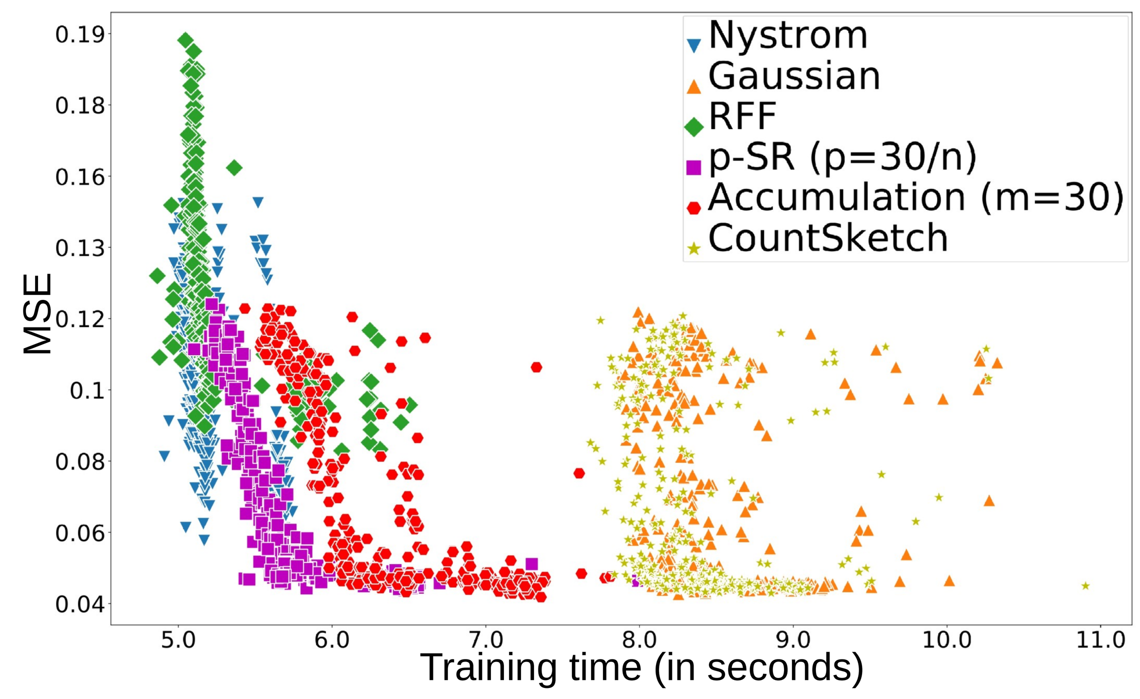
The authors thank Olivier Fercoq for insightful discussions. This work was supported by the Télécom Paris research chair on Data Science and Artificial Intelligence for Digitalized Industry and Services (DSAIDIS) and by the French National Research Agency (ANR) through the ANR-18-CE23-0014 APi project.

Experiments

Synthetic scalar robust regression:







Real-world multi-output joint quantile regression: Acc denotes Accumulation sketch [6]

dataset	Boston		Otoliths	
Sketch	w/o	p-SR	w/o	p-SR
Pinball	**51.28**	54.75	2.78	2.66
Crossing	0.34	0.26	**5.18**	5.46
Time	6.97	1.43	606.8	20.4

dataset	Boston		Otoliths	
Sketch	p-SG	Acc	p-SG	Acc
Pinball	54.78	54.73	**2.64**	2.67
Crossing	**0.11**	0.15	5.43	5.46
Time	**1.38**	1.48	**20.0**	22.1

References

[1] Yang et al. *Randomized sketches for kernels: Fast and optimal nonparametric regression.* The Annals of Statistics, 2017.

[2] Meanti et al. *Kernel methods through the roof: Handling billions of points efficiently.* NeurIPS, 2020.

[3] Lacotte & Pilanci *Adaptive and oblivious randomized subspace methods for high-dimensional optimization: Sharp analysis and lower bounds.*. IEEE, 2021.

[4] Li et al. *Towards a unified analysis of random fourier features.* JMLR, 2021.

[5] Sangnier et al. *Joint quantile regression in vector-valued RKHSs.* NeurIPS, 2016.

[6] Chen & Yang *Accumulations of projections—a unified framework for random sketches in kernel ridge regression.* AISTATS, 2021.