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

CAp 2023

Tamim El Ahmad*, Pierre Laforgue†, Florence d'Alché-Buc*
⋆ LTCI, Télécom Paris, Institut Polytechnique de Paris
† Università degli Studi di Milano

July 5, 2023

Supervised Scalar Regression

We have:

- i.i.d. training sample $(x_i, y_i)_{i=1}^n \in (\mathcal{X}, \mathbb{R})^n \sim P$
- loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0, \infty)$

<u>Goal:</u> Approach $f^\star = \underset{f:\mathcal{X}\to\mathbb{R}}{\arg\inf}\ \mathbb{E}_{(X,Y)\sim P}\left[\ell\left(f(X), Y\right)\right] (ERM)$.

Supervised Scalar Regression

We have:

- i.i.d. training sample $(x_i, y_i)_{i=1}^n \in (\mathcal{X}, \mathbb{R})^n \sim P$
- loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0, \infty)$

<u>**Goal:**</u> Approach $f^\star = \underset{f:\mathcal{X}\to\mathbb{R}}{\arg\inf}\ \mathbb{E}_{(X,Y)\sim P}\left[\ell\left(f(X), Y\right)\right]$ (ERM).

$\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$ is too large: which hypothesis space?

Reminder: positive definite kernels and Reproducing Kernel Hilbert Space

Positive definite kernel: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

- for all $(x, x') \in \mathcal{X}^2$, $k(x, x') = k(x', x)$
- for all $n \in \mathbb{N}$ and any $(x_i, \alpha_i)_{i=1}^n \in (\mathcal{X} \times \mathbb{R})^n$, $\sum_{i,j=1}^n \alpha_i \alpha_j k\left(x_i, x_j\right) \geqslant 0$

RKHS (Aronszajn, 1950): k is uniquely associated to a Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$ s. t. for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$

1. $x' \mapsto k(x, x') \in \mathcal{H}$,
2. $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ **(reproducing property)**.

Kernel-Based Regression

Given k and its associated RKHS $\mathcal{H}$, $\lambda_n > 0$

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

Kernel-Based Regression

Given k and its associated RKHS $\mathcal{H}$, $\lambda_n > 0$

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

Representer Theorem: $\hat{f} = \sum_{j=1}^{n} k(\cdot, x_j) \hat{\alpha}_j$, where

$$(\hat{\alpha}_1, \dots, \hat{\alpha}_n)^\top = \hat{\alpha} = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell\left([K\alpha]_{i:}^\top, y_i \right) + \frac{\lambda_n}{2} \alpha^\top K \alpha \,.$$

Optimisation problem on n parameters: can we reduce n?

Table of contents

1. Sketched Kernel Machines

2. *p*-Sparsified Sketches

3. Experiments

4. Conclusion

Sketched Kernel Machines

First Idea: Sub-Sampling, i.e. Nyström Approximation

$\tilde{f} = \sum_{j=1}^{s} k(\cdot, x_{i_j})\tilde{\gamma}_j$, where

$$(\tilde{\gamma}_1, \ldots, \tilde{\gamma}_s)^\top = \tilde{\gamma} = \operatorname*{arg\,min}_{\gamma \in \mathbb{R}^s} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\left[\underbrace{K_{ns}}_{n \times s} \gamma\right]_{i:}^\top, y_i\right) + \frac{\lambda_n}{2} \gamma^\top \underbrace{K_{ss}}_{s \times s} \gamma \,.$$

First Idea: Sub-Sampling, i.e. Nyström Approximation

$\tilde{f} = \sum_{j=1}^{s} k(\cdot, x_{i_j})\tilde{\gamma}_j$, where

$$(\tilde{\gamma}_1, \ldots, \tilde{\gamma}_s)^\top = \tilde{\gamma} = \underset{\gamma \in \mathbb{R}^s}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell \left(\left[\underbrace{K_{ns}}_{n \times s} \gamma \right]_{i:}^\top, y_i \right) + \frac{\lambda_n}{2} \gamma^\top \underbrace{K_{ss}}_{s \times s} \gamma \, .$$

Sampling the wrong data can lead to poor results $\implies$ data-dependent sampling schemes (e.g. leverage scores) (Alaoui and Mahoney, 2015; Musco and Musco, 2017; Rudi et al., 2018; Chen and Yang, 2021b)

Sub-Sampling is Random Projection

Let $n = 5$, $X = \{x_1, \ldots, x_5\}$, $k_X^x = (k(x, x_1), \ldots, k(x, x_5))$, $s = 2$ and

$$S = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Sub-Sampling is Random Projection

Let $n = 5$, $X = \{x_1, \dots, x_5\}$, $k_X^x = (k(x, x_1), \dots, k(x, x_5))$, $s = 2$ and
$S = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$

$$K_{sn} = \begin{pmatrix} k_X^{x_1} \\ k_X^{x_4} \end{pmatrix} = SK \quad \text{and} \quad K_{ss} = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_4) \\ k(x_4, x_1) & k(x_4, x_4) \end{pmatrix} = SKS^\top$$

Sub-Sampling is Random Projection

Let $n = 5$, $X = \{x_1, \ldots, x_5\}$, $k_X^x = (k(x, x_1), \ldots, k(x, x_5))$, $s = 2$ and
$S = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$

$$K_{sn} = \begin{pmatrix} k_X^{x_1} \\ k_X^{x_4} \end{pmatrix} = SK \quad \text{and} \quad K_{ss} = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_4) \\ k(x_4, x_1) & k(x_4, x_4) \end{pmatrix} = SKS^\top$$

$\tilde{f} = \sum_{j=1}^{s} k(\cdot, x_{i_j})\tilde{\gamma}_j = \sum_{j=1}^{n} k(\cdot, x_{i_j})[S^\top \tilde{\gamma}]_j$, where

$$(\tilde{\gamma}_1, \ldots, \tilde{\gamma}_s)^\top = \tilde{\gamma} = \operatorname*{arg\,min}_{\gamma \in \mathbb{R}^s} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\left[KS^\top \gamma\right]_{i:}^\top, y_i\right) + \frac{\lambda_n}{2} \gamma^\top SKS^\top \gamma \, .$$

Sub-Sampling is Random Projection

Let $n = 5$, $X = \{x_1, \dots, x_5\}$, $k_X^x = (k(x, x_1), \dots, k(x, x_5))$, $s = 2$ and
$S = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$

$$K_{sn} = \begin{pmatrix} k_X^{x_1} \\ k_X^{x_4} \end{pmatrix} = SK \quad \text{and} \quad K_{ss} = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_4) \\ k(x_4, x_1) & k(x_4, x_4) \end{pmatrix} = SKS^\top$$

$\tilde{f} = \sum_{j=1}^{s} k(\cdot, x_{i_j}) \tilde{\gamma}_j = \sum_{j=1}^{n} k(\cdot, x_{i_j}) [S^\top \tilde{\gamma}]_j$, where

$$(\tilde{\gamma}_1, \dots, \tilde{\gamma}_s)^\top = \tilde{\gamma} = \underset{\gamma \in \mathbb{R}^s}{\arg\min} \, \frac{1}{n} \sum_{i=1}^{n} \ell \left(\left[KS^\top \gamma \right]_{i:}^\top, y_i \right) + \frac{\lambda_n}{2} \gamma^\top SKS^\top \gamma \,.$$

Could we use other random matrix distributions?

Johnson-Linderstrauss Lemma

Lemma
Given $0 < \varepsilon < 1$*, a set* $\mathcal{S}$ *of* n *points in* $\mathbb{R}^D$*, and an integer* $d > 8(\log n)/\varepsilon^2$*, there is a linear map* $h : \mathbb{R}^D \to \mathbb{R}^d$ *such that*

$$(1-\varepsilon)\|u-v\|^2 \leq \|h(u)-h(v)\|^2 \leq (1+\varepsilon)\|u-v\|^2 \ ,$$

for all $u, v \in \mathcal{S}$*.*

Lemma
Given $0 < \varepsilon < 1$, a set $\mathcal{S}$ of n points in $\mathbb{R}^D$, and an integer $d > 8(\log n)/\varepsilon^2$, there is a linear map $h : \mathbb{R}^D \to \mathbb{R}^d$ such that

$$(1-\varepsilon)\left\|u-v\right\|^2 \leq \left\|h(u)-h(v)\right\|^2 \leq (1+\varepsilon)\left\|u-v\right\|^2 ,$$

for all $u, v \in \mathcal{S}$.

Most famous proof:

1. take $h = \frac{1}{\sqrt{d}} S \in \mathbb{R}^{d \times D}$, where $S_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \implies$ **Gaussian sketching**
2. prove the above equation with high probability

Gaussian sketching then?

$$(\tilde{\gamma}_1, \ldots, \tilde{\gamma}_s)^\top = \hat{\gamma} = \operatorname*{arg\,min}_{\gamma \in \mathbb{R}^s} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\left[KS^\top \gamma\right]_i, y_i\right) + \frac{\lambda_n}{2} \gamma^\top SKS^\top \gamma\,.$$

Gaussian sketching then?

$$(\tilde{\gamma}_1, \ldots, \tilde{\gamma}_s)^\top = \hat{\gamma} = \operatorname*{arg\,min}_{\gamma \in \mathbb{R}^s} \frac{1}{n} \sum_{i=1}^{n} \ell\left(\left[KS^\top \gamma\right]_i, y_i\right) + \frac{\lambda_n}{2} \gamma^\top SKS^\top \gamma \,.$$

Problems:

1. computing SK: $\mathcal{O}\left(n^2 s\right)$ time complexity $\rightarrow$ **still high complexity**
2. storing K: $\mathcal{O}\left(n^2\right)$ space complexity $\rightarrow$ **space complexity does not change**

Which property should sketching distributions satisfy?

- $K/n = UDU^\top$
- $D = \mathrm{diag}\left(\mu_1, \ldots, \mu_n\right)$ where $\mu_1 \geq \ldots \geq \mu_n$
- δ_n^2 the lowest value s. t. $\psi(\delta_n) = (\frac{1}{n}\sum_{i=1}^{n} \min(\delta_n^2, \mu_i))^{1/2} \leq \delta_n^2$ (Bartlett et al., 2005)
- $d_n = \min \ \left\{j \in \{1, \ldots, n\} \colon \mu_j \leq \delta_n^2\right\}$

Which property should sketching distributions satisfy?

- $K/n = UDU^\top$
- $D = \mathrm{diag}\,(\mu_1, \ldots, \mu_n)$ where $\mu_1 \geq \ldots \geq \mu_n$
- δ_n^2 the lowest value s. t. $\psi(\delta_n) = (\frac{1}{n}\sum_{i=1}^n \min(\delta_n^2, \mu_i))^{1/2} \leq \delta_n^2$ (Bartlett et al., 2005)
- $d_n = \min\ \left\{ j \in \{1, \ldots, n\} \colon \mu_j \leq \delta_n^2 \right\}$

Definition (K-satisfiability (Yang et al., 2017))

Let $c > 0$ independent of n. Let $U_1 \in \mathbb{R}^{n \times d_n}$ and $U_2 \in \mathbb{R}^{n \times (n-d_n)}$ be the left and right blocks of matrix U previously defined, and $D_2 = \mathrm{diag}\,(\mu_{d_n+1}, \ldots, \mu_n)$. A sketch matrix S is said to be K-satisfiable for c if S is such that

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

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

p-Sparsified Sketches

Definition

Let $s < n$, and $p \in (0, 1]$. 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 $B_{ij} \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$ and $R_{ij} \overset{\text{i.i.d.}}{\sim} \text{Rad}(\frac{1}{2})$ (p-SR) or $\mathcal{N}(0, 1)$ (p-SG).

K-Satisfiability of p-Sparsified Sketches

Theorem

Let S be a p-sparsified sketch. 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 $c = \frac{2}{\sqrt{p}}\left(1 + \sqrt{\log(5)}\right) + 1$.

Intuitive behavior of p:

- $p = 1$: we recover Yang et al. (2017)'s result for Gaussian sketching
- the larger it is, the denser S is, and the more likely S is K-satisfiable
- the smaller it is, the larger s is needed

Computational Property: *Decomposition trick*

Let $s' = \sum_{j=1}^{n} \mathbb{I}\{S_{:j} \neq 0_s\}$,

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

where

- $S_{\mathrm{SG}} \in \mathbb{R}^{s \times s'}$: **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

Example:

$$\begin{pmatrix} 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$s' \sim \mathrm{Binom}\left(n, 1-(1-p)^s\right) \implies \mathbb{E}\left[s'\right] = n(1-(1-p)^s) \underset{p\to 0}{\sim} nsp$$

Time and Space Complexities

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

Time and Space Complexities

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

p-sparsified sketch's goal $\rightarrow$ best of both worlds:

1. computational efficiency of sub-sampling sketch
2. statistical accuracy of Rademacher or Gaussian sketch

Time and Space Complexities

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

p-sparsified sketch's goal $\rightarrow$ best of both worlds:

1. computational efficiency of sub-sampling sketch
2. statistical accuracy of Rademacher or Gaussian sketch

Related works:

1. sub-sampling sketch with data-dependent sampling schemes (e.g. leverage scores) (Alaoui and Mahoney, 2015; Musco and Musco, 2017; Rudi et al., 2018; Chen and Yang, 2021b)
2. accumulation sketch (Chen and Yang, 2021a): sum of sub-sampling sketches

Experiments

Scalar regression with synthetic dataset: settings

1) $n = 10,000$, $(x_i, y_i) \in \mathbb{R}^{10} \times \mathbb{R}$

2) Inhomogeneous input data distribution

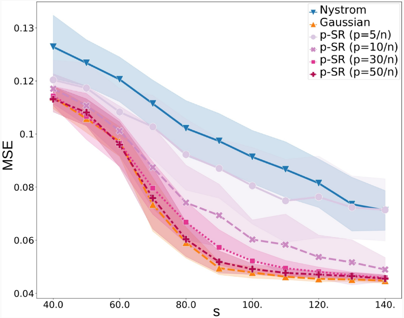
$$x_i \sim \begin{cases} \mathcal{U}\left([0_{10}, \mathbb{1}_{10}]\right), & \text{if } i = 1, \ldots, 9,900\,, \\ \mathcal{N}\left(1.5\mathbb{1}_{10}, 0.25 I_{10}\right), & \text{if } i = 9,901, \ldots, 10,000\,, \end{cases}$$

3) $y = f^\star(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0,1)$ and

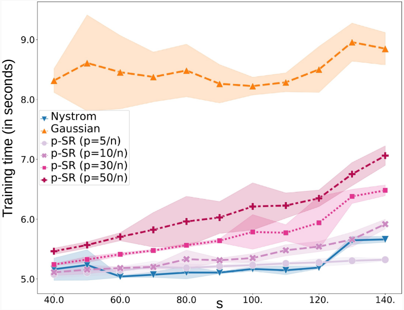
$$f^\star(x) = 0.1 \exp\left(4x_1\right) + \frac{4}{1 + \exp\left(-20\left(x_2 - 0.5\right)\right)} + 3x_3 + 2x_4 + x_5\,.$$

4) loss: κ-Huber

Scalar regression with synthetic dataset



(a) Test relative MSE w.r.t. sketch size s



(b) Training time (sec) w.r.t. sketch size s

Scalar regression with synthetic dataset

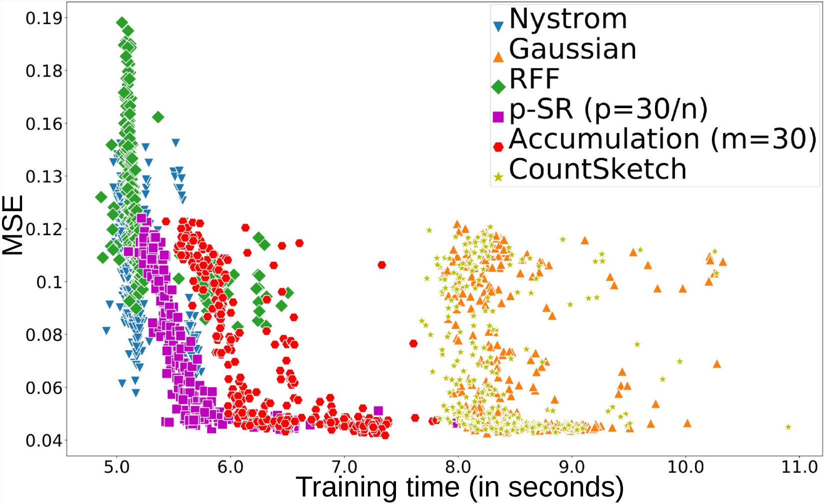


Figure 2: Test relative MSE w.r.t. training times

Conclusion

Conclusion

- Extend scalar regression framework to multi-output regression thanks to decomposable matrix-valued kernels

Conclusion

- Extend scalar regression framework to multi-output regression thanks to decomposable matrix-valued kernels
- Extend previous results and provide excess risk bounds for the multiple output setting and with any generic Lipschitz loss thanks to decomposable matrix-valued kernels

Conclusion

- Extend scalar regression framework to multi-output regression thanks to decomposable matrix-valued kernels
- Extend previous results and provide excess risk bounds for the multiple output setting and with any generic Lipschitz loss thanks to decomposable matrix-valued kernels
- Provide new K-satisfiable sketching distribution – p-sparsified – well-suited to kernel methods thanks to the *decomposition trick*

Conclusion

- Extend scalar regression framework to multi-output regression thanks to decomposable matrix-valued kernels
- Extend previous results and provide excess risk bounds for the multiple output setting and with any generic Lipschitz loss thanks to decomposable matrix-valued kernels
- Provide new K-satisfiable sketching distribution – p-sparsified – well-suited to kernel methods thanks to the *decomposition trick*
- When the input data distribution shows some inhomogeneity, p-sparsified sketches
    1. outperform Nyström approximation and RFFs
    2. compete with statistically accurate sketches (Gaussian, CountSketch, Accumulation) while being faster

Conclusion

- Extend scalar regression framework to multi-output regression thanks to decomposable matrix-valued kernels
- Extend previous results and provide excess risk bounds for the multiple output setting and with any generic Lipschitz loss thanks to decomposable matrix-valued kernels
- Provide new K-satisfiable sketching distribution – p-sparsified – well-suited to kernel methods thanks to the *decomposition trick*
- When the input data distribution shows some inhomogeneity, p-sparsified sketches
    1. outperform Nyström approximation and RFFs
    2. compete with statistically accurate sketches (Gaussian, CountSketch, Accumulation) while being faster
- Sketched kernel algorithms show similar performances – and even outperform in some cases – non-sketched kernel algorithms, while being significantly faster

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.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, pages 337–404.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local rademacher complexities. *Ann. Statist.*, 33(4):1497–1537.

References ii

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

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

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

Koenker, R. (2005). *Quantile regression*. Cambridge university press.

Moen, E., Handegard, N. O., Allken, V., Albert, O. T., Harbitz, A., and Malde, K. (2018). Automatic interpretation of otoliths using deep learning. *PLoS One*, 13(12):e0204713.

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

Ordoñez, A., Eikvil, L., Salberg, A.-B., Harbitz, A., Murray, S. M., and Kampffmeyer, M. C. (2020). Explaining decisions of deep neural networks used for fish age prediction. *PloS one*, 15(6):e0235013.

Rudi, A., Calandriello, D., Carratino, L., and Rosasco, L. (2018). On fast leverage score sampling and optimal learning. In *NeurIPS*.

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

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

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.

Lipschitz Losses

$\ell(y, y') = g(y - y')$, where g is:

- **For κ-Huber**: For $\kappa > 0$:

$$\forall y \in \mathcal{Y}, g(y) = \begin{cases} \frac{1}{2}\|y\|_{\mathcal{Y}}^2 & \text{if } \|y\|_{\mathcal{Y}} \leq \kappa \\ \kappa\left(\|y\|_{\mathcal{Y}} - \frac{\kappa}{2}\right) & \text{otherwise} \end{cases}.$$

- **The pinball loss (Koenker, 2005) for joint quantile regression**:
  For d quantile levels, $\tau_1 < \tau_2 < \ldots < \tau_d$ with $\tau_i \in (0, 1)$, we define:

$$\ell_\tau(f(x), y) = L_\tau(f(x) - y\mathbb{1}_d),$$

  with the following definition for L_τ the extension of pinball loss to $\mathbb{R}^d$ (Sangnier et al., 2016):
  For $r \in \mathbb{R}^d$:

$$L_\tau(r) = \sum_{j=1}^{d} \begin{cases} \tau_j r_j & \text{if } r_j \geq 0, \\ (\tau_j - 1) r_j & \text{if } r_j < 0. \end{cases}$$

Example: Kernel Ridge Multi-Output Regression

With $\mathcal{K} = kI_d$

- **Without sketching:** $\hat{A} = (K + n\lambda I_n)^{-1} Y \implies$ inversion of $n \times n$ matrix
- **With sketching:** $\tilde{\Gamma} = \left(SK^2S^\top + n\lambda SKS^\top\right)^{-1} SKY \implies$ inversion of $s \times s$ matrix

Previous work

Settings in Yang et al. (2017):

- $d = 1 \implies$ scalar regression only
- $\ell(y, y') = (y - y')^2 \implies$ KRR only
- $y_i = f^\star(x_i) + \sigma\omega_i$, where ω_is i.i.d. standard Gaussian variates
- Focus on the squared $L^2(\mathbb{P}_n)$ error, i.e., $\left\|\tilde{f}_s - f^\star\right\|_n^2 = \frac{1}{n}\sum_{i=1}^n \left(\tilde{f}_s(x_i) - f^\star(x_i)\right)^2 \implies$ not excess risk in expectation

Previous work

Settings in Yang et al. (2017):

- $d = 1 \implies$ scalar regression only
- $\ell(y, y') = (y - y')^2 \implies$ KRR only
- $y_i = f^\star(x_i) + \sigma\omega_i$, where ω_is i.i.d. standard Gaussian variates
- Focus on the squared $L^2(\mathbb{P}_n)$ error, i.e., $\left\|\tilde{f}_s - f^\star\right\|_n^2 = \frac{1}{n}\sum_{i=1}^n \left(\tilde{f}_s(x_i) - f^\star(x_i)\right)^2 \implies$ not excess risk in expectation

Yang et al. (2017, Theorem 2): If $f^* \in \mathcal{H}$, then for any $\lambda \geq 2\delta_n^2$, with a probability greater than $1 - c_1 e^{-c_2 n \delta_n^2}$

$$\left\|\tilde{f}_s - f^*\right\|_n^2 \leq c_u\left(\lambda + \delta_n^2\right), \tag{1}$$

where c_u only depends on $\|f^*\|_{\mathcal{H}}$.

Theoretical Guarantees

Assumptions

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

Assumptions

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

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

Assumptions

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

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
$\mathcal{H}(\mathcal{X}) = \{f(x) : f \in \mathcal{H},\, x \in \mathcal{X}\}$.

Assumptions

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
$\mathcal{H}(\mathcal{X}) = \{f(x) : f \in \mathcal{H},\, x \in \mathcal{X}\}$.

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

Assumptions

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

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

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

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.

Excess Risk Bound

Theorem

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 $1 - \delta$,

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

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

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

Sketch of proof: Decomposition Error

$$
\begin{aligned}
\mathbb{E}[\ell_{\tilde{f}_S}] - \mathbb{E}[\ell_{f_{\mathcal{H}_k}}] &= \mathbb{E}_{(X,Y)\sim P}[\ell(\tilde{f}_S(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(\tilde{f}_S(x_i), y_i) \leftarrow \text{gen. error} \\
&+ \frac{1}{n}\sum_{i=1}^{n} \ell(\tilde{f}_S(x_i), y_i) - \frac{1}{n}\sum_{i=1}^{n} \ell(f_{\mathcal{H}_k}(x_i), y_i) \leftarrow \text{approx. error} \\
&+ \frac{1}{n}\sum_{i=1}^{n} \ell(f_{\mathcal{H}_k}(x_i), y_i) - \mathbb{E}_{(X,Y)\sim P}[\ell(f_{\mathcal{H}_k}(X), Y)] \leftarrow \text{gen. error}
\end{aligned}
$$

Sketch of proof: Approximation Error

Let $\mathcal{H}_S = \left\{ f = \sum_{i=1}^{n} k(\cdot, x_i) M \left[S^\top \tilde{\Gamma} \right]_i \mid \gamma \in \mathbb{R}^{s \times d} \right\}$

$$\frac{1}{n} \sum_{i=1}^{n} \ell(\tilde{f}_S(x_i), y_i) - \frac{1}{n} \sum_{i=1}^{n} \ell(f_{\mathcal{H}_k}(x_i), y_i)$$

$$\leq \inf_{\substack{f \in \mathcal{H}_S \\ \|f\|_{\mathcal{H}_\mathcal{K}} \leq 1}} \frac{L}{n} \sum_{i=1}^{n} \| f(x_i) - f_{\mathcal{H}_\mathcal{K}}(x_i) \|_2 \leftarrow \text{A. 2}$$

$$\leq L \inf_{\substack{f \in \mathcal{H}_S \\ \|f\|_{\mathcal{H}_\mathcal{K}} \leq 1}} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \| f(x_i) - f_{\mathcal{H}_\mathcal{K}}(x_i) \|_2^2} \leftarrow \text{Jensen}$$

Which sketching distribution to use for kernels?

With $\mathcal{K} = kI_d$

- **Without sketching:** $\hat{A} = (K + n\lambda I_n)^{-1} Y \implies$ inversion of $n \times n$ matrix
- **With sketching:** $\hat{\Gamma} = \left(SK^2S^\top + n\lambda SKS^\top\right)^{-1} SKY \implies$ inversion of $s \times s$ matrix

Which sketching distribution to use for kernels?

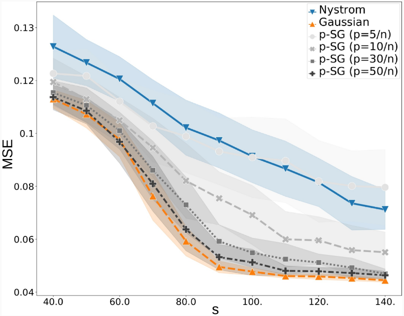
With $\mathcal{K} = kI_d$

- **Without sketching:** $\hat{A} = (K + n\lambda I_n)^{-1} Y \implies$ inversion of $n \times n$ matrix
- **With sketching:** $\hat{\Gamma} = \left(SK^2S^\top + n\lambda SKS^\top\right)^{-1} SKY \implies$ inversion of $s \times s$ matrix

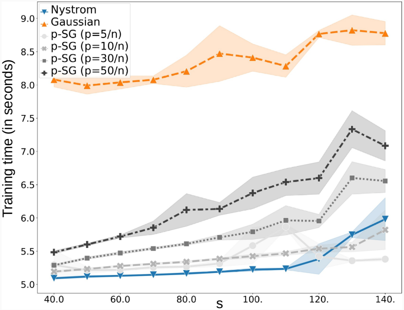
Problems:

1. computing SK: $\mathcal{O}\left(n^2 s\right)$ time complexity $\rightarrow$ still high complexity
2. storing K: $\mathcal{O}\left(n^2\right)$ space complexity $\rightarrow$ space complexity does not change

Scalar regression with synthetic dataset



(a) Test relative MSE w.r.t. sketch size s



(b) Training time (sec) w.r.t. sketch size s

Scalar regression with synthetic dataset

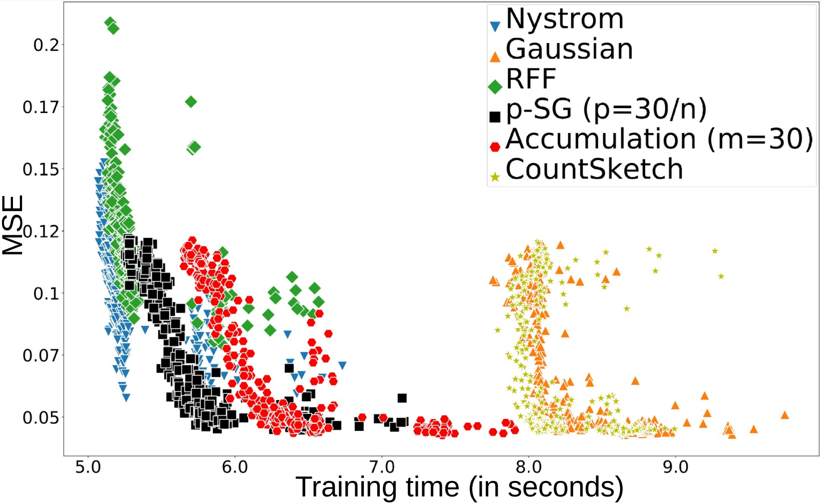


Figure 4: Test relative MSE w.r.t. training times with κ-Huber

Joint Quantile Regression on real data

- Boston dataset (Harrison Jr and Rubinfeld, 1978): house price prediction, $n = 506$
- Otoliths dataset (Moen et al., 2018; Ordoñez et al., 2020): fish age prediction, $n = 3780$

Quantile levels to predict: (0.1, 0.3, 0.5, 0.7, 0.9)

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

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

Multi-target Regression on real data

- rf1 and rf2 datasets (Spyromitros-Xioufis et al., 2016): river network flows prediction, $n = 4108, 4108$
- scm1d and scm20d datasets (Spyromitros-Xioufis et al., 2016): products price prediction, $n = 8145, 7463$

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. $m = 20$
rf1	ARRMSE	**0.575**	0.584 ± 0.003	0.583 ± 0.003	0.592 ± 0.001
	Training time	1.73	$\mathbf{0.22 \pm 0.025}$	0.25 ± 0.005	0.60 ± 0.0004
rf2	ARRMSE	**0.578**	0.671 ± 0.009	0.656 ± 0.006	0.796 ± 0.006
	Training time	1.77	0.28 ± 0.003	$\mathbf{0.27 \pm 0.003}$	0.82 ± 0.003
scm1d	ARRMSE	**0.418**	0.422 ± 0.002	0.423 ± 0.001	0.423 ± 0.001
	Training time	9.36	$\mathbf{0.45 \pm 0.022}$	$\mathbf{0.45 \pm 0.019}$	0.86 ± 0.006
scm20d	ARRMSE	0.755	0.754 ± 0.003	0.754 ± 0.003	$\mathbf{0.753 \pm 0.001}$
	Training time	6.16	$\mathbf{0.38 \pm 0.016}$	$\mathbf{0.38 \pm 0.017}$	0.70 ± 0.032