Sketch In, Sketch Out: Accelerating both Learning and Inference for Structured Prediction with Kernels

Journée de Statistique 2024

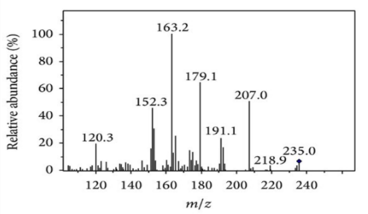
Tamim El Ahmad[⋆], Luc Brogat-Motte[⋆†], Pierre Laforgue[‡], Florence d'Alché-Buc[⋆]
⋆ LTCI, Télécom Paris, Institut Polytechnique de Paris
† L2S, CentraleSupélec
‡ Università degli Studi di Milano

May 28, 2024

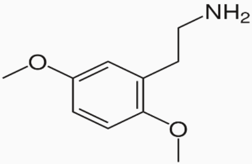
Structured Prediction

Goal: learn a mapping $f : \mathcal{X} \longrightarrow \mathcal{Y}$ with $\mathcal{Y}$ a space of structured objects (graphs, rankings, sequences, binary vectors, etc.).

MS/MS spectra



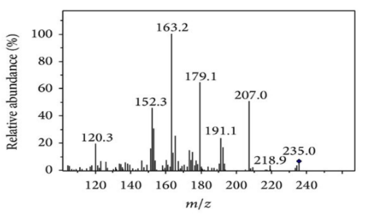
$$\xrightarrow{\quad f \;=\; ? \quad}$$



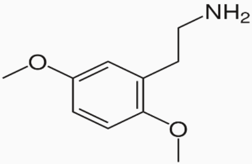
Structured Prediction

Goal: learn a mapping $f : \mathcal{X} \longrightarrow \mathcal{Y}$ with $\mathcal{Y}$ a space of structured objects (graphs, rankings, sequences, binary vectors, etc.).

MS/MS spectra



$$\xrightarrow{\quad f \ = \ ? \quad}$$



Existing works: **Energy-based models** (Lafferty et al., 2001; Taskar et al., 2003; Tsochantaridis et al., 2004; LeCun et al., 2007; Belanger and McCallum, 2016):

$$f(x) = \underset{y \in \mathcal{Y}}{\arg\min} \ E(x, y) \tag{1}$$

Table of contents

Input Output Kernel Regression

Output Kernel Regression

Given a p.d. kernel $k_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ **defining a relevant similarity measure** and $\psi_{\mathcal{Y}} : y \in \mathcal{Y} \mapsto k_{\mathcal{Y}}(\cdot, y) \in \mathcal{H}_{\mathcal{Y}}$,

we define $\Delta(y, y') = \|\psi_{\mathcal{Y}}(y) - \psi_{\mathcal{Y}}(y')\|^2_{\mathcal{H}_{\mathcal{Y}}}$ (Weston et al., 2003; Cortes et al., 2005), and solve

$$\min_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{(X,Y)\sim\rho}[\|\psi_{\mathcal{Y}}(f(X)) - \psi_{\mathcal{Y}}(Y)\|^2_{\mathcal{H}_{\mathcal{Y}}}] \tag{2}$$

Output Kernel Regression

Given a p.d. kernel $k_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ **defining a relevant similarity measure** and $\psi_{\mathcal{Y}} : y \in \mathcal{Y} \mapsto k_{\mathcal{Y}}(\cdot, y) \in \mathcal{H}_{\mathcal{Y}}$,

we define $\Delta(y, y') = \|\psi_{\mathcal{Y}}(y) - \psi_{\mathcal{Y}}(y')\|^2_{\mathcal{H}_{\mathcal{Y}}}$ (Weston et al., 2003; Cortes et al., 2005), and solve

$$\min_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{(X,Y)\sim\rho}[\|\psi_{\mathcal{Y}}(f(X)) - \psi_{\mathcal{Y}}(Y)\|^2_{\mathcal{H}_{\mathcal{Y}}}] \tag{3}$$

How to learn f through $\psi_{\mathcal{Y}}$?

Output Kernel Regression

$$\mathcal{X} \xrightarrow[\displaystyle \hat{f} \approx \operatorname*{arg\,inf}_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{x,y}\Big[\|\psi_{\mathcal{Y}}(f(x)) - \psi_{\mathcal{Y}}(y)\|^2_{\mathcal{H}_{\mathcal{Y}}}\Big]]{\qquad\qquad\qquad\qquad\qquad\qquad} \mathcal{Y}$$

Output Kernel Regression

$$
\begin{tikzcd}
 & \mathcal{H}_{\mathcal{Y}} \arrow[d, bend left=60, "d"] \\
\mathcal{X} \arrow[ur, "\hat{h} \approx \underset{h:\mathcal{X}\to\mathcal{H}_{\mathcal{Y}}}{\arg\inf}\ \mathbb{E}_{x,y}\left[\|h(x) - \psi_{\mathcal{Y}}(y)\|^2_{\mathcal{H}_{\mathcal{Y}}}\right]"] \arrow[r, "\hat{f} = d \circ \hat{h} = \underset{y\in\mathcal{Y}}{\arg\min}\ \underbrace{\|\hat{h}(\,\cdot\,) - \psi_{\mathcal{Y}}(y)\|^2_{\mathcal{H}_{\mathcal{Y}}}}_{E(\cdot,\, y)}"'] & \mathcal{Y} \arrow[u, "\psi_{\mathcal{Y}}"']
\end{tikzcd}
$$

Output Kernel Regression

$$
\begin{tikzcd}[column sep=huge, row sep=huge]
 & \mathcal{H}_{\mathcal{Y}} \arrow[d, bend left, "d"] \\
\mathcal{X} \arrow[ur, "\hat{h} \approx \underset{h:\mathcal{X}\to\mathcal{H}_{\mathcal{Y}}}{\arg\inf}\, \mathbb{E}_{x,y}\left[\|h(x) - \psi_{\mathcal{Y}}(y)\|^2_{\mathcal{H}_{\mathcal{Y}}}\right]"] \arrow[r, "\hat{f} = d \circ \hat{h} = \underset{y \in \mathcal{Y}}{\arg\min}\, \underbrace{\|\hat{h}(\,\cdot\,) - \psi_{\mathcal{Y}}(y)\|^2_{\mathcal{H}_{\mathcal{Y}}}}_{E(\cdot,\, y)}"'] & \mathcal{Y} \arrow[u, "\psi_{\mathcal{Y}}"']
\end{tikzcd}
$$

Which hypothesis space for $\hat{h}$?

How to deal with infinite-dimensional output feature space $\mathcal{H}_{\mathcal{Y}}$?

Input Output Kernel Regression

$$
\begin{tikzcd}
\mathcal{H}_{\mathcal{X}} \arrow[r, "\widehat{H}"] & \mathcal{H}_{\mathcal{Y}} \arrow[d, bend left, "d"] \\
\mathcal{X} \arrow[u, "\psi_{\mathcal{X}}"] \arrow[ur, "\hat{h}(x) = \widehat{H}\psi_{\mathcal{X}}(x) = \sum_{i=1}^{n} \widehat{\alpha}_i(x)\psi_{\mathcal{Y}}(y_i)"] \arrow[r, "\hat{f} = d \circ \hat{h}"'] & \mathcal{Y} \arrow[u, "\psi_{\mathcal{Y}}"']
\end{tikzcd}
$$

Input Output Kernel Regression

$$
\begin{tikzcd}
\mathcal{H}_{\mathcal{X}} \arrow[rr, "\widehat{H}"] & & \mathcal{H}_{\mathcal{Y}} \arrow[d, bend left=60, "d"] \\
\mathcal{X} \arrow[u, "\psi_{\mathcal{X}}"] \arrow[urr, "\hat{h}(x) = \widehat{H}\psi_{\mathcal{X}}(x) = \sum_{i=1}^{n} \widehat{\alpha}_i(x)\psi_{\mathcal{Y}}(y_i)"] \arrow[rr, "\hat{f} = d \circ \hat{h}"'] & & \mathcal{Y} \arrow[u, "\psi_{\mathcal{Y}}"']
\end{tikzcd}
$$

$\hat{\alpha}(x) = (\underbrace{K_X + n\lambda I_n}_{n \times n})^{-1} k_X^x = \widehat{\Omega} k_X^x$ where $n =$ number of training data

Input Output Kernel Regression

$$
\begin{tikzcd}
\mathcal{H}_{\mathcal{X}} \arrow[r, "\widehat{H}"] & \mathcal{H}_{\mathcal{Y}} \arrow[d, bend left, "d"] \\
\mathcal{X} \arrow[u, "\psi_{\mathcal{X}}"] \arrow[ur, "\hat{h}(x) = \widehat{H}\psi_{\mathcal{X}}(x) = \sum_{i=1}^{n} \widehat{\alpha}_i(x)\psi_{\mathcal{Y}}(y_i)"] \arrow[r, "\hat{f} = d \circ \hat{h}"'] & \mathcal{Y} \arrow[u, "\psi_{\mathcal{Y}}"']
\end{tikzcd}
$$

$\hat{\alpha}(x) = (\underbrace{K_X + n\lambda I_n}_{n\times n})^{-1} k_X^x = \widehat{\Omega} k_X^x$ where $n =$ number of training data

Training complexity: $\mathcal{O}\left(n^3\right)$

Input Output Kernel Regression: Inference

$$\hat{f}(x) = d(\hat{h}(x)) = \operatorname*{arg\,min}_{y \in \mathcal{Y}} \left\| \hat{h}(x) - \psi_{\mathcal{Y}}(y) \right\|^2_{\mathcal{H}_{\mathcal{Y}}} =$$

$$\operatorname*{arg\,min}_{y \in \mathcal{Y}} k_{\mathcal{Y}}(y, y) - 2 {k_X^x}^T \widehat{\Omega} k_Y^y$$

Input Output Kernel Regression: Inference

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

$$\underset{y \in \mathcal{Y}}{\arg\min}\ k_{\mathcal{Y}}(y, y) - 2 {k_X^x}^T \widehat{\Omega} k_Y^y$$

- **Test set:** X_{te} of size n_{te}
- **Candidate set:** $\mathcal{Y}_c \subseteq \mathcal{Y}$ of size n_c

$$\underbrace{K_X^{te,tr}}_{n_{te} \times n} \underbrace{\widehat{\Omega}}_{n \times n} \underbrace{K_Y^{tr,c}}_{n \times n_c} \tag{4}$$

Input Output Kernel Regression: Inference

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

$$\underset{y \in \mathcal{Y}}{\arg\min}\ k_{\mathcal{Y}}(y, y) - 2 {k_X^x}^T \widehat{\Omega} k_Y^y$$

- **Test set:** X_{te} of size n_{te}
- **Candidate set:** $\mathcal{Y}_c \subseteq \mathcal{Y}$ of size n_c

$$\underbrace{K_X^{te,tr}}_{n_{te} \times n} \underbrace{\widehat{\Omega}}_{n \times n} \underbrace{K_Y^{tr,c}}_{n \times n_c} \quad (5)$$

Inference complexity: $\mathcal{O}\left(n_{te} n n_c\right)$ if $n_{te} < n \leq n_c$

Fisher consistency and excess risk bound

Lemma 1 and Theorem 3 from Ciliberto et al. (2020). Let $\mathcal{Y}$ be compact, $k_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a p.d. kernel and $\psi_{\mathcal{Y}} : y \mapsto k_{\mathcal{Y}}(\cdot, y)$ s.t. $\|\psi_{\mathcal{Y}}(y)\|_{\mathcal{H}_{\mathcal{Y}}} = 1, \forall y \in \mathcal{Y}$, and

$$f^* = \underset{f:\mathcal{X}\to\mathcal{Y}}{\arg\inf}\ \mathcal{R}(f) = \underset{f:\mathcal{X}\to\mathcal{Y}}{\arg\inf}\ \mathbb{E}_{(x,y)\sim\rho}[\|\psi_{\mathcal{Y}}(f(x)) - \psi_{\mathcal{Y}}(y)\|^2_{\mathcal{H}_{\mathcal{Y}}}]\,.$$

Then,

$$f^*(x) = \underset{y\in\mathcal{Y}}{\arg\min}\ \|h^*(x) - \psi_{\mathcal{Y}}(y)\|^2_{\mathcal{H}_{\mathcal{Y}}} = d \circ h^*(x)\,, \quad h^*(x) = \mathbb{E}_y[\psi_{\mathcal{Y}}(y)|x]\,,$$

almost surely with respect to $\rho_{\mathcal{X}}$.

Fisher consistency and excess risk bound

Lemma 1 and Theorem 3 from Ciliberto et al. (2020). Let $\mathcal{Y}$ be compact, $k_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a p.d. kernel and $\psi_{\mathcal{Y}} : y \mapsto k_{\mathcal{Y}}(\cdot, y)$ s.t. $\|\psi_{\mathcal{Y}}(y)\|_{\mathcal{H}_{\mathcal{Y}}} = 1, \forall y \in \mathcal{Y}$, and

$$f^* = \underset{f:\mathcal{X}\to\mathcal{Y}}{\arg\inf}\ \mathcal{R}(f) = \underset{f:\mathcal{X}\to\mathcal{Y}}{\arg\inf}\ \mathbb{E}_{(x,y)\sim\rho}[\|\psi_{\mathcal{Y}}(f(x)) - \psi_{\mathcal{Y}}(y)\|^2_{\mathcal{H}_{\mathcal{Y}}}] \,.$$

Then,

$$f^*(x) = \underset{y\in\mathcal{Y}}{\arg\min}\ \|h^*(x) - \psi_{\mathcal{Y}}(y)\|^2_{\mathcal{H}_{\mathcal{Y}}} = d \circ h^*(x) \,, \quad h^*(x) = \mathbb{E}_y[\psi_{\mathcal{Y}}(y)|x] \,,$$

almost surely with respect to $\rho_{\mathcal{X}}$.

Moreover, let $h : \mathcal{X} \to \mathcal{H}_{\mathcal{Y}}$ be measurable and $f : \mathcal{X} \to \mathcal{Y}$ such that, for any $x \in \mathcal{X}$,

$$f(x) = \underset{y\in\mathcal{Y}}{\arg\min}\ \|h(x) - \psi_{\mathcal{Y}}(y)\|^2_{\mathcal{H}_{\mathcal{Y}}} = d \circ h(x) \,.$$

Then,

$$\mathcal{R}(f) - \mathcal{R}(f^*) \leq 12\sqrt{\mathcal{E}(h) - \mathcal{E}(h^*)} \,,$$

where $\mathcal{E}(h) = \mathbb{E}_{(x,y)\sim\rho}[\|h(x) - \psi_{\mathcal{Y}}(y)\|^2_{\mathcal{H}_{\mathcal{Y}}}]$.

Advantages of IOKR

1) Strong theoretical grounding: derived from the operator-valued kernel and surrogate methods literature.

Advantages of IOKR

1) Strong theoretical grounding: derived from the operator-valued kernel and surrogate methods literature.

2) Very general algorithm for structured prediction: ability to tackle many different tasks through an appropriate choice of the output kernel

Advantages of IOKR

1) **Strong theoretical grounding:** derived from the operator-valued kernel and surrogate methods literature.

2) **Very general algorithm for structured prediction:** ability to tackle many different tasks through an appropriate choice of the output kernel

3) **Closed-form solution of kernel Ridge regression:** no need for any optimization algorithm to be solved, unlike deep models (Belanger and McCallum, 2016; Belanger et al., 2017; Gygli et al., 2017)

Research question

Can we scale IOKR up to large datasets at both the training and inference phases, especially since they employ not only an input but also an output kernel, while keeping good empirical and theoretical statistical guarantees?

Sketched Input Sketched Output Kernel Regression

Motivation

Motivation: build a **low-rank** approximation $\tilde{h}$ thanks to **input and output** random projectors $\widetilde{P}_X$ and $\widetilde{P}_Y$ to obtain a **scalable** predictor $\tilde{f}$ together with an excess risk bound

Some notations

- For $\mathcal{Z}$ a Polish space, $k_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ a p.d. kernel, $\psi_{\mathcal{Z}}(z) := k_{\mathcal{Z}}(\cdot, z)$, $\mathcal{H}_{\mathcal{Z}}$ its RKHS

Some notations

- For $\mathcal{Z}$ a Polish space, $k_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ a p.d. kernel, $\psi_{\mathcal{Z}}(z) := k_{\mathcal{Z}}(\cdot, z)$, $\mathcal{H}_{\mathcal{Z}}$ its RKHS
- For an i.i.d. sample $(z_i)_{i=1}^n \in \mathcal{Z}^n \sim \rho_z$

Some notations

- For $\mathcal{Z}$ a Polish space, $k_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ a p.d. kernel, $\psi_{\mathcal{Z}}(z) := k_{\mathcal{Z}}(\cdot, z)$, $\mathcal{H}_{\mathcal{Z}}$ its RKHS
- For an i.i.d. sample $(z_i)_{i=1}^n \in \mathcal{Z}^n \sim \rho_z$
- $S_Z : f \in \mathcal{H}_{\mathcal{Z}} \mapsto (1/\sqrt{n})(f(z_1), \ldots, f(z_n))^\top \in \mathbb{R}^n$ sampling operator

Some notations

- For $\mathcal{Z}$ a Polish space, $k_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ a p.d. kernel, $\psi_{\mathcal{Z}}(z) := k_{\mathcal{Z}}(\cdot, z)$, $\mathcal{H}_{\mathcal{Z}}$ its RKHS
- For an i.i.d. sample $(z_i)_{i=1}^n \in \mathcal{Z}^n \sim \rho_z$
- $S_Z : f \in \mathcal{H}_{\mathcal{Z}} \mapsto (1/\sqrt{n})(f(z_1), \ldots, f(z_n))^\top \in \mathbb{R}^n$ sampling operator
- $S_Z^\# : \alpha \in \mathbb{R}^n \mapsto (1/\sqrt{n}) \sum_{i=1}^n \alpha_i \psi_{\mathcal{Z}}(z_i) \in \mathcal{H}_{\mathcal{Z}}$ its adjoint

Some notations

- For $\mathcal{Z}$ a Polish space, $k_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ a p.d. kernel, $\psi_{\mathcal{Z}}(z) := k_{\mathcal{Z}}(\cdot, z)$, $\mathcal{H}_{\mathcal{Z}}$ its RKHS
- For an i.i.d. sample $(z_i)_{i=1}^n \in \mathcal{Z}^n \sim \rho_z$
- $S_Z : f \in \mathcal{H}_{\mathcal{Z}} \mapsto (1/\sqrt{n})(f(z_1), \dots, f(z_n))^\top \in \mathbb{R}^n$ sampling operator
- $S_Z^\# : \alpha \in \mathbb{R}^n \mapsto (1/\sqrt{n}) \sum_{i=1}^n \alpha_i \psi_{\mathcal{Z}}(z_i) \in \mathcal{H}_{\mathcal{Z}}$ its adjoint
- $K_Z = \left(k_{\mathcal{Z}}(z_i, z_j)\right)_{1 \leq i,j \leq n} = n S_Z S_Z^\#$

Some notations

- For $\mathcal{Z}$ a Polish space, $k_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ a p.d. kernel, $\psi_{\mathcal{Z}}(z) := k_{\mathcal{Z}}(\cdot, z)$, $\mathcal{H}_{\mathcal{Z}}$ its RKHS
- For an i.i.d. sample $(z_i)_{i=1}^n \in \mathcal{Z}^n \sim \rho_z$
- $S_Z : f \in \mathcal{H}_{\mathcal{Z}} \mapsto (1/\sqrt{n})(f(z_1), \ldots, f(z_n))^\top \in \mathbb{R}^n$ sampling operator
- $S_Z^\# : \alpha \in \mathbb{R}^n \mapsto (1/\sqrt{n}) \sum_{i=1}^n \alpha_i \psi_{\mathcal{Z}}(z_i) \in \mathcal{H}_{\mathcal{Z}}$ its adjoint
- $K_Z = \left(k_{\mathcal{Z}}(z_i, z_j)\right)_{1 \le i,j \le n} = n S_Z S_Z^\#$
- $C_{\mathcal{Z}} = \mathbb{E}_z[\psi_{\mathcal{Z}}(z) \otimes \psi_{\mathcal{Z}}(z)]$ covariance operator

Some notations

- For $\mathcal{Z}$ a Polish space, $k_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ a p.d. kernel, $\psi_{\mathcal{Z}}(z) := k_{\mathcal{Z}}(\cdot, z)$, $\mathcal{H}_{\mathcal{Z}}$ its RKHS
- For an i.i.d. sample $(z_i)_{i=1}^n \in \mathcal{Z}^n \sim \rho_z$
- $S_Z : f \in \mathcal{H}_{\mathcal{Z}} \mapsto (1/\sqrt{n})(f(z_1), \ldots, f(z_n))^\top \in \mathbb{R}^n$ sampling operator
- $S_Z^\# : \alpha \in \mathbb{R}^n \mapsto (1/\sqrt{n}) \sum_{i=1}^n \alpha_i \psi_{\mathcal{Z}}(z_i) \in \mathcal{H}_{\mathcal{Z}}$ its adjoint
- $K_Z = \left(k_{\mathcal{Z}}(z_i, z_j)\right)_{1 \le i,j \le n} = n S_Z S_Z^\#$
- $C_{\mathcal{Z}} = \mathbb{E}_z[\psi_{\mathcal{Z}}(z) \otimes \psi_{\mathcal{Z}}(z)]$ covariance operator
- $\widehat{C}_Z = (1/n) \sum_{i=1}^n \psi_{\mathcal{Z}}(z_i) \otimes \psi_{\mathcal{Z}}(z_i) = S_Z^\# S_Z$ its empirical counterpart

Low-rank Estimator: from IOKR to SISOKR

$$
\begin{tikzcd}
\mathcal{H}_\mathcal{X} \arrow[rr, "\widehat{H} = S_Y^{\#} S_X \left(S_X^{\#} S_X + \lambda I_{\mathcal{H}_\mathcal{X}}\right)^{-1}"] & & \mathcal{H}_\mathcal{Y} \arrow[d, bend left, "d"] \\
\mathcal{X} \arrow[u, "\psi_\mathcal{X}"] \arrow[urr, "\widehat{h}(x) = \widehat{H}\psi_\mathcal{X}(x) = \sum_{i=1}^{n} \widehat{\alpha}_i(x)\psi_\mathcal{Y}(y_i)"] \arrow[rr, "\widehat{f} = d \circ \widehat{h}"'] & & \mathcal{Y} \arrow[u, "\psi_\mathcal{Y}"']
\end{tikzcd}
$$

Low-rank Estimator: from IOKR to SISOKR

$$
\begin{tikzcd}[column sep=huge, row sep=huge]
\mathcal{H}_{\mathcal{X}} \arrow[r, "{\widetilde{H} = \widetilde{P}_Y S_Y^{\#} S_X \widetilde{P}_X \left(\widetilde{P}_X S_X^{\#} S_X \widetilde{P}_X + \lambda I_{\mathcal{H}_{\mathcal{X}}}\right)^{-1}}"] & \mathcal{H}_{\mathcal{Y}} \arrow[d, bend left, "d"] \\
\mathcal{X} \arrow[u, "{\widetilde{P}_X \psi_{\mathcal{X}}}"] \arrow[ur, "{\widetilde{h}(x) = \widetilde{H}\psi_{\mathcal{X}}(x) = \sum_{i=1}^{n} \widetilde{\alpha}_i(x)\psi_{\mathcal{Y}}(y_i)}"] \arrow[r, "{\widetilde{f} = d \circ \widetilde{h}}"'] & \mathcal{Y} \arrow[u, "{\widetilde{P}_Y \psi_{\mathcal{Y}}}"']
\end{tikzcd}
$$

Sketching

Sketching: random linear projections

Sketching

Sketching: random linear projections

Let $m_{\mathcal{Z}} \ll n$, $R_{\mathcal{Z}} \in \mathbb{R}^{m_{\mathcal{Z}} \times n}$ be a random matrix and n data $(z_i)_{i=1}^n \in \mathcal{Z}$

Sketching

Sketching: random linear projections

Let $m_{\mathcal{Z}} \ll n$, $R_{\mathcal{Z}} \in \mathbb{R}^{m_{\mathcal{Z}} \times n}$ be a random matrix and n data $(z_i)_{i=1}^n \in \mathcal{Z}$

Basic idea: Sketching-based operator $\widetilde{P}_Z$ projects onto the following linear subspace of $\mathcal{H}_{\mathcal{Z}}$

$$\sum_{j=1}^{n} (R_{\mathcal{Z}})_{ij} \psi_{\mathcal{Z}}(z_j) \in \mathcal{H}_{\mathcal{Z}}, \quad i = 1, \ldots, m_{\mathcal{Z}} \tag{6}$$

Construction of the orthogonal projector $\widetilde{P}_Z$

- $\widetilde{C}_Z = S_Z^{\#} R_{\mathcal{Z}}^{\top} R_{\mathcal{Z}} S_Z$

Construction of the orthogonal projector $\widetilde{P}_Z$

- $\widetilde{C}_Z = S_Z^{\#} R_{\mathcal{Z}}^{\top} R_{\mathcal{Z}} S_Z$
- $\widetilde{K}_Z = R_{\mathcal{Z}} K_Z R_{\mathcal{Z}}^{\top}$, and $\left\{ \left(\sigma_i(\widetilde{K}_Z), \tilde{\mathbf{v}}_i^Z \right), i \in [m_{\mathcal{Z}}] \right\}$ its eigenpairs

Construction of the orthogonal projector $\widetilde{P}_Z$

- $\widetilde{C}_Z = S_Z^{\#} R_{\mathcal{Z}}^{\top} R_{\mathcal{Z}} S_Z$
- $\widetilde{K}_Z = R_{\mathcal{Z}} K_Z R_{\mathcal{Z}}^{\top}$, and $\left\{ \left(\sigma_i(\widetilde{K}_Z), \tilde{\mathbf{v}}_i^Z \right), i \in [m_{\mathcal{Z}}] \right\}$ its eigenpairs
- $p_Z = \operatorname{rank}\left(\widetilde{K}_Z\right)$, and for all $1 \leq i \leq p_Z$, $\tilde{e}_i^Z = \sqrt{\frac{n}{\sigma_i(\widetilde{K}_Z)}} S_Z^{\#} R_{\mathcal{Z}}^{\top} \tilde{\mathbf{v}}_i^Z \in \mathcal{H}_{\mathcal{Z}}$

Construction of the orthogonal projector $\widetilde{P}_Z$

- $\widetilde{C}_Z = S_Z^{\#} R_{\mathcal{Z}}^{\top} R_{\mathcal{Z}} S_Z$
- $\widetilde{K}_Z = R_{\mathcal{Z}} K_Z R_{\mathcal{Z}}^{\top}$, and $\left\{ \left(\sigma_i(\widetilde{K}_Z), \tilde{\mathbf{v}}_i^Z \right), i \in [m_{\mathcal{Z}}] \right\}$ its eigenpairs
- $p_Z = \mathrm{rank}\left(\widetilde{K}_Z\right)$, and for all $1 \leq i \leq p_Z$, $\tilde{e}_i^Z = \sqrt{\frac{n}{\sigma_i(\widetilde{K}_Z)}} S_Z^{\#} R_{\mathcal{Z}}^{\top} \tilde{\mathbf{v}}_i^Z \in \mathcal{H}_{\mathcal{Z}}$

Proposition

The $\tilde{e}_i^Z$s are the eigenfunctions, associated to the eigenvalues $\sigma_i(\widetilde{K}_Z)/n$, of $\widetilde{C}_Z$.

- $\widetilde{C}_Z = S_Z^{\#} R_{\mathcal{Z}}^{\top} R_{\mathcal{Z}} S_Z$
- $\widetilde{K}_Z = R_{\mathcal{Z}} K_Z R_{\mathcal{Z}}^{\top}$, and $\left\{ \left(\sigma_i(\widetilde{K}_Z), \tilde{\mathbf{v}}_i^Z \right), i \in [m_{\mathcal{Z}}] \right\}$ its eigenpairs
- $p_Z = \mathrm{rank}\left(\widetilde{K}_Z\right)$, and for all $1 \leq i \leq p_Z$, $\tilde{e}_i^Z = \sqrt{\frac{n}{\sigma_i(\widetilde{K}_Z)}} S_Z^{\#} R_{\mathcal{Z}}^{\top} \tilde{\mathbf{v}}_i^Z \in \mathcal{H}_{\mathcal{Z}}$

Proposition

The $\tilde{e}_i^Z$s are the eigenfunctions, associated to the eigenvalues $\sigma_i(\widetilde{K}_Z)/n$, of $\widetilde{C}_Z$.
Furthermore, let $\widetilde{\mathcal{H}}_{\mathcal{Z}} = \mathrm{span}\left(\tilde{e}_1^Z, \ldots, \tilde{e}_{p_Z}^Z\right)$, the orthogonal projector $\widetilde{P}_Z$ onto $\widetilde{\mathcal{H}}_{\mathcal{Z}}$ writes as

$$\widetilde{P}_Z = (R_{\mathcal{Z}} S_Z)^{\#} \left(R_{\mathcal{Z}} S_Z (R_{\mathcal{Z}} S_Z)^{\#} \right)^{\dagger} R_{\mathcal{Z}} S_Z \,. \tag{7}$$

Construction of the orthogonal projector $\widetilde{P}_Z$

- $\widetilde{C}_Z = S_Z^{\#} R_{\mathcal{Z}}^{\top} R_{\mathcal{Z}} S_Z$
- $\widetilde{K}_Z = R_{\mathcal{Z}} K_Z R_{\mathcal{Z}}^{\top}$, and $\left\{ \left(\sigma_i(\widetilde{K}_Z), \tilde{\mathbf{v}}_i^Z \right), i \in [m_{\mathcal{Z}}] \right\}$ its eigenpairs
- $p_Z = \operatorname{rank}\left(\widetilde{K}_Z\right)$, and for all $1 \leq i \leq p_Z$, $\tilde{e}_i^Z = \sqrt{\frac{n}{\sigma_i(\widetilde{K}_Z)}} S_Z^{\#} R_{\mathcal{Z}}^{\top} \tilde{\mathbf{v}}_i^Z \in \mathcal{H}_{\mathcal{Z}}$

Proposition

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

$$\widetilde{P}_Z = (R_{\mathcal{Z}} S_Z)^{\#} \left(R_{\mathcal{Z}} S_Z (R_{\mathcal{Z}} S_Z)^{\#} \right)^{\dagger} R_{\mathcal{Z}} S_Z \,. \tag{7}$$

Related work on Nyström: Yang et al. (2012); Rudi et al. (2015)

Sketched Input Sketched Output Kernel Regression estimator

Proposition (Expression of SISOKR)

$$\tilde{h}(x) = \sum_{i=1}^{n} \tilde{\alpha}_i(x)\, \psi_{\mathcal{Y}}(y_i)\ , \quad \textit{where} \quad \tilde{\alpha}(x) = R_{\mathcal{Y}}^\top \widetilde{\Omega} R_{\mathcal{X}} k_X^x\,, \tag{8}$$

with

$$\widetilde{\Omega} = \underbrace{(R_{\mathcal{Y}} K_Y R_{\mathcal{Y}}^\top)}_{m_{\mathcal{Y}} \times m_{\mathcal{Y}}}{}^\dagger R_{\mathcal{Y}} K_Y K_X R_{\mathcal{X}}^\top \underbrace{(R_{\mathcal{X}} K_X^2 R_{\mathcal{X}}^\top + n\lambda R_{\mathcal{X}} K_X R_{\mathcal{X}}^\top)}_{m_{\mathcal{X}} \times m_{\mathcal{X}}}{}^\dagger \tag{9}$$

Sketched Input Sketched Output Kernel Regression estimator

Proposition (Expression of SISOKR)

$$\tilde{h}(x) = \sum_{i=1}^{n} \tilde{\alpha}_i(x)\,\psi_{\mathcal{Y}}(y_i)\,, \quad \textit{where} \quad \tilde{\alpha}(x) = R_{\mathcal{Y}}^{\top}\widetilde{\Omega}R_{\mathcal{X}}k_X^x\,, \tag{8}$$

with

$$\widetilde{\Omega} = (\underbrace{R_{\mathcal{Y}}K_Y R_{\mathcal{Y}}^{\top}}_{m_{\mathcal{Y}}\times m_{\mathcal{Y}}})^{\dagger} R_{\mathcal{Y}}K_Y K_X R_{\mathcal{X}}^{\top}(\underbrace{R_{\mathcal{X}}K_X^2 R_{\mathcal{X}}^{\top} + n\lambda R_{\mathcal{X}}K_X R_{\mathcal{X}}^{\top}}_{m_{\mathcal{X}}\times m_{\mathcal{X}}})^{\dagger} \tag{9}$$

Inversion complexity: $\mathcal{O}(n^3) \rightarrow \mathcal{O}(\max(m_{\mathcal{X}}^3, m_{\mathcal{Y}}^3))$

Sketched Input Sketched Output Kernel Regression estimator

Proposition (Expression of SISOKR)

$$\tilde{h}(x) = \sum_{i=1}^{n} \tilde{\alpha}_i(x) \psi_{\mathcal{Y}}(y_i) , \quad \textit{where} \quad \tilde{\alpha}(x) = R_{\mathcal{Y}}^\top \widetilde{\Omega} R_{\mathcal{X}} k_X^x , \tag{8}$$

with

$$\widetilde{\Omega} = (\underbrace{R_{\mathcal{Y}} K_Y R_{\mathcal{Y}}^\top}_{m_{\mathcal{Y}} \times m_{\mathcal{Y}}})^\dagger R_{\mathcal{Y}} K_Y K_X R_{\mathcal{X}}^\top (\underbrace{R_{\mathcal{X}} K_X^2 R_{\mathcal{X}}^\top + n\lambda R_{\mathcal{X}} K_X R_{\mathcal{X}}^\top}_{m_{\mathcal{X}} \times m_{\mathcal{X}}})^\dagger \tag{9}$$

Inversion complexity: $\mathcal{O}(n^3) \to \mathcal{O}(\max(m_{\mathcal{X}}^3, m_{\mathcal{Y}}^3))$

Complexity of $R_{\mathcal{Z}} K_Z$: depends on the sketch matrix, between $\mathcal{O}(n m_{\mathcal{Z}})$ and $\mathcal{O}(n^2 m_{\mathcal{Z}})$

Sketched Input Sketched Output Kernel Regression estimator

Proposition (Expression of SISOKR)

$$\tilde{h}(x) = \sum_{i=1}^{n} \tilde{\alpha}_i(x)\,\psi_{\mathcal{Y}}(y_i)\,, \quad \textit{where} \quad \tilde{\alpha}(x) = R_{\mathcal{Y}}^{\top}\widetilde{\Omega}R_{\mathcal{X}}k_X^x\,, \tag{8}$$

with

$$\widetilde{\Omega} = (\underbrace{R_{\mathcal{Y}}K_YR_{\mathcal{Y}}^{\top}}_{m_{\mathcal{Y}}\times m_{\mathcal{Y}}})^{\dagger}R_{\mathcal{Y}}K_YK_XR_{\mathcal{X}}^{\top}(\underbrace{R_{\mathcal{X}}K_X^2R_{\mathcal{X}}^{\top} + n\lambda R_{\mathcal{X}}K_XR_{\mathcal{X}}^{\top}}_{m_{\mathcal{X}}\times m_{\mathcal{X}}})^{\dagger} \tag{9}$$

Inversion complexity: $\mathcal{O}(n^3) \rightarrow \mathcal{O}(\max(m_{\mathcal{X}}^3, m_{\mathcal{Y}}^3))$

Complexity of $R_{\mathcal{Z}}K_Z$**:** depends on the sketch matrix, between $\mathcal{O}(nm_{\mathcal{Z}})$ and $\mathcal{O}(n^2m_{\mathcal{Z}})$

$\implies$ **Training complexity reduced!**

SISOKR estimator: Inference

- **Test set:** X_{te} of size n_{te}
- **Candidate set:** $\mathcal{Y}_c \subseteq \mathcal{Y}$ of size n_c

$$\underbrace{K_X^{te,tr} R_{\mathcal{X}}^\top}_{n_{te} \times m_{\mathcal{X}}} \underbrace{\widetilde{\Omega}}_{m_{\mathcal{X}} \times m_{\mathcal{Y}}} \underbrace{R_{\mathcal{Y}} K_Y^{tr,c}}_{m_{\mathcal{Y}} \times n_c} \tag{10}$$

SISOKR estimator: Inference

- **Test set:** X_{te} of size n_{te}
- **Candidate set:** $\mathcal{Y}_c \subseteq \mathcal{Y}$ of size n_c

$$\underbrace{K_X^{te,tr} R_{\mathcal{X}}^\top}_{n_{te} \times m_{\mathcal{X}}} \underbrace{\widetilde{\Omega}}_{m_{\mathcal{X}} \times m_{\mathcal{Y}}} \underbrace{R_{\mathcal{Y}} K_Y^{tr,c}}_{m_{\mathcal{Y}} \times n_c} \tag{10}$$

Decoding complexity: $\mathcal{O}(n_{te} n n_c) \rightarrow \mathcal{O}(n_{te} m_{\mathcal{Y}} n_c)$ if $n_{te} \leq m_{\mathcal{X}}, m_{\mathcal{Y}} < n \leq n_c$

SISOKR estimator: Inference

- **Test set:** X_{te} of size n_{te}
- **Candidate set:** $\mathcal{Y}_c \subseteq \mathcal{Y}$ of size n_c

$$\underbrace{K_X^{te,tr} R_{\mathcal{X}}^\top}_{n_{te} \times m_{\mathcal{X}}} \underbrace{\widetilde{\Omega}}_{m_{\mathcal{X}} \times m_{\mathcal{Y}}} \underbrace{R_{\mathcal{Y}} K_Y^{tr,c}}_{m_{\mathcal{Y}} \times n_c} \tag{10}$$

Decoding complexity: $\mathcal{O}(n_{te} n n_c) \to \mathcal{O}(n_{te} m_{\mathcal{Y}} n_c)$ if $n_{te} \leq m_{\mathcal{X}}, m_{\mathcal{Y}} < n \leq n_c$

$\Longrightarrow$ **Inference complexity reduced!**

Theoretical Analysis

Assumptions

Asm. 1 (Attainability): Recall that $h^*(x) := \mathbb{E}_Y[\psi_{\mathcal{Y}}(Y) \mid X = x]$. $h^* \in \mathcal{H}$, i.e. there exists $H : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ with $\|H\|_{\mathsf{HS}} < +\infty$ such that

$$h^*(x) = H\psi_{\mathcal{X}}(x) \quad \forall x \in \mathcal{X}. \tag{11}$$

Assumptions

Asm. 1 (Attainability): Recall that $h^*(x) := \mathbb{E}_Y[\psi_{\mathcal{Y}}(Y) \mid X = x]$. $h^* \in \mathcal{H}$, i.e. there exists $H : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ with $\|H\|_{\mathsf{HS}} < +\infty$ such that

$$h^*(x) = H\psi_{\mathcal{X}}(x) \quad \forall x \in \mathcal{X}. \tag{11}$$

Asm. 2 (Bounded kernel): there exists $\kappa_{\mathcal{Z}} > 0$ such that

$$k_z(z, z) \leq \kappa_{\mathcal{Z}}^2 \quad \forall z \in \mathcal{Z}. \tag{12}$$

Assumptions

Asm. 1 (Attainability): Recall that $h^*(x) := \mathbb{E}_Y[\psi_{\mathcal{Y}}(Y) \mid X = x]$. $h^* \in \mathcal{H}$, i.e. there exists $H : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ with $\|H\|_{\mathsf{HS}} < +\infty$ such that

$$h^*(x) = H\psi_{\mathcal{X}}(x) \quad \forall x \in \mathcal{X}. \tag{11}$$

Asm. 2 (Bounded kernel): there exists $\kappa_{\mathcal{Z}} > 0$ such that

$$k_z(z, z) \leq \kappa_{\mathcal{Z}}^2 \quad \forall z \in \mathcal{Z}. \tag{12}$$

Asm. 3 (Capacity condition): there exists $\gamma_{\mathcal{Z}} \in [0, 1]$ such that

$$Q_{\mathcal{Z}} := \mathrm{Tr}(C_{\mathcal{Z}}^{\gamma_{\mathcal{Z}}}) < +\infty\,. \tag{13}$$

Assumptions

Asm. 1 (Attainability): Recall that $h^*(x) := \mathbb{E}_Y[\psi_{\mathcal{Y}}(Y) \mid X = x]$. $h^* \in \mathcal{H}$, i.e. there exists $H : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ with $\|H\|_{\mathrm{HS}} < +\infty$ such that

$$h^*(x) = H\psi_{\mathcal{X}}(x) \quad \forall x \in \mathcal{X}. \tag{11}$$

Asm. 2 (Bounded kernel): there exists $\kappa_{\mathcal{Z}} > 0$ such that

$$k_z(z, z) \leq \kappa_{\mathcal{Z}}^2 \quad \forall z \in \mathcal{Z}. \tag{12}$$

Asm. 3 (Capacity condition): there exists $\gamma_{\mathcal{Z}} \in [0, 1]$ such that

$$Q_{\mathcal{Z}} := \mathrm{Tr}(C_{\mathcal{Z}}^{\gamma_{\mathcal{Z}}}) < +\infty\,. \tag{13}$$

Asm. 4 (Embedding property): there exists $b_{\mathcal{Z}} > 0$ and $\mu_{\mathcal{Z}} \in [0, 1]$ such that almost surely

$$\psi_{\mathcal{Z}}(z) \otimes \psi_{\mathcal{Z}}(z) \preceq b_{\mathcal{Z}} C_{\mathcal{Z}}^{1-\mu_{\mathcal{Z}}}\,. \tag{14}$$

Assumptions

Asm. 1 (Attainability): Recall that $h^*(x) := \mathbb{E}_Y[\psi_{\mathcal{Y}}(Y) \mid X = x]$. $h^* \in \mathcal{H}$, i.e. there exists $H : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ with $\|H\|_{\mathrm{HS}} < +\infty$ such that

$$h^*(x) = H\psi_{\mathcal{X}}(x) \quad \forall x \in \mathcal{X}. \tag{11}$$

Asm. 2 (Bounded kernel): there exists $\kappa_{\mathcal{Z}} > 0$ such that

$$k_z(z, z) \leq \kappa_{\mathcal{Z}}^2 \quad \forall z \in \mathcal{Z}. \tag{12}$$

Asm. 3 (Capacity condition): there exists $\gamma_{\mathcal{Z}} \in [0, 1]$ such that

$$Q_{\mathcal{Z}} := \mathrm{Tr}(C_{\mathcal{Z}}^{\gamma_{\mathcal{Z}}}) < +\infty\,. \tag{13}$$

Asm. 4 (Embedding property): there exists $b_{\mathcal{Z}} > 0$ and $\mu_{\mathcal{Z}} \in [0, 1]$ such that almost surely

$$\psi_{\mathcal{Z}}(z) \otimes \psi_{\mathcal{Z}}(z) \preceq b_{\mathcal{Z}} C_{\mathcal{Z}}^{1-\mu_{\mathcal{Z}}}. \tag{14}$$

Asm. 5 (Sub-Gaussian sketches): $R_{\mathcal{Z}} \in \mathbb{R}^{m_{\mathcal{Z}} \times n}$ composed with i.i.d. entries s.t. (i) $\mathbb{E}\left[R_{\mathcal{Z}_{ij}}\right] = 0$, (ii) $\mathbb{E}\left[R_{\mathcal{Z}_{ij}}^2\right] = 1/m_{\mathcal{Z}}$ and (iii) $R_{\mathcal{Z}_{ij}} \sim \frac{\nu_{\mathcal{Z}}^2}{m\mathcal{Z}} -$ sub-Gaussian with $\nu_{\mathcal{Z}} \geq 1$.

Corollary (SISOKR learning rates)

Under Asm. 1, 2, 3, 4 and 5, if for all $y \in \mathcal{Y}$, $\|\psi_{\mathcal{Y}}(y)\|_{\mathcal{H}_{\mathcal{Y}}} = \kappa_{\mathcal{Y}}$, *for* $n \in \mathbb{N}$ *sufficiently large such that* $\frac{9}{n}\log(n/\delta) \leq n^{-\frac{1}{1+\gamma_{\mathcal{X}}}} \leq \|C_{\mathcal{X}}\|_{\mathrm{op}}/2$, *and* $\frac{9}{n}\log(n/\delta) \leq n^{-\frac{1}{1+\gamma_{\mathcal{Y}}}} \leq \|C_{\mathcal{Y}}\|_{\mathrm{op}}/2$, *and for sketching size* $m_{\mathcal{X}}, m_{\mathcal{Y}} \in \mathbb{N}$ *such that*

$$m_{\mathcal{X}} \gtrsim \max\left(\nu_{\mathcal{X}}^2 n^{\frac{\gamma_{\mathcal{X}}+\mu_{\mathcal{X}}}{1+\gamma_{\mathcal{X}}}}, \nu_{\mathcal{X}}^4 \log\left(1/\delta\right)\right), \tag{15}$$

$$m_{\mathcal{Y}} \gtrsim \max\left(\nu_{\mathcal{Y}}^2 n^{\frac{\gamma_{\mathcal{Y}}+\mu_{\mathcal{Y}}}{1+\gamma_{\mathcal{Y}}}}, \nu_{\mathcal{Y}}^4 \log\left(1/\delta\right)\right), \tag{16}$$

then with probability $1-\delta$

$$\mathbb{E}[\|\tilde{h}(x) - h^*(x)\|_{\mathcal{H}_{\mathcal{Y}}}^2]^{\frac{1}{2}} \lesssim \log\left(4/\delta\right) n^{-\frac{1-\gamma_{\mathcal{X}} \vee \gamma_{\mathcal{Y}}}{2(1+\gamma_{\mathcal{X}} \vee \gamma_{\mathcal{Y}})}}, \tag{17}$$

and

$$\mathcal{R}(\tilde{f}) - \mathcal{R}(f^*) \lesssim \log\left(4/\delta\right) n^{-\frac{1-\gamma_{\mathcal{X}} \vee \gamma_{\mathcal{Y}}}{2(1+\gamma_{\mathcal{X}} \vee \gamma_{\mathcal{Y}})}}. \tag{18}$$

Experiments

Multi-Label Classification: Statistical Performance

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

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

Multi-Label Classification: Computational Performance

Table 2: Comparison of training/inference computation times (in seconds).

Method	Bibtex	Bookmarks
SISOKR	**1.41 ± 0.03 / 0.46 ± 0.01**	**118 ± 1.5 / 20 ± 0.2**
ISOKR	2.51 ± 0.06 / 0.58 ± 0.01	NA
SIOKR	1.99 ± 0.07 / 1.22 ± 0.03	354 ± 2.1 / 297 ± 2.1
IOKR	2.54 / 1.18	NA

Conclusion

Take-home messages

- Scale up surrogate kernel methods for structured prediction by leveraging random projections, in both input and output feature spaces, to accelerate training and inference phases

Take-home messages

- Scale up surrogate kernel methods for structured prediction by leveraging **random projections**, in **both input and output** feature spaces, to **accelerate training and inference** phases
- Derive **excess risk bounds** for the **sketched estimator**

Take-home messages

- Scale up surrogate kernel methods for structured prediction by leveraging **random projections**, in **both input and output** feature spaces, to **accelerate training and inference** phases
- Derive **excess risk bounds** for the **sketched estimator**
- Show that **sub-Gaussian** sketches are admissible sketches in the sense that they lead to **close to optimal learning rates** with **sketching sizes** $m < n$

Take-home messages

- Scale up surrogate kernel methods for structured prediction by leveraging **random projections**, in **both input and output** feature spaces, to **accelerate training and inference** phases
- Derive **excess risk bounds** for the **sketched estimator**
- Show that **sub-Gaussian** sketches are admissible sketches in the sense that they lead to **close to optimal learning rates** with **sketching sizes** $m < n$
- Provide structured prediction **experiments** on **real-world data sets** showing **similar performances** as **IOKR** while **being faster** in **both training and inference** phases.

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.

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

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

Belanger, D., Yang, B., and McCallum, A. (2017). End-to-end learning for structured prediction energy networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 429–439. PMLR.

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

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

References iii

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

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.

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

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

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

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

Gazagnadou, N., Ibrahim, M., and Gower, R. M. (2021). `RidgeSketch`: A fast sketching based solver for large scale ridge regression.

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

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

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

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

Lacotte, J., Pilanci, M., and Pavone, M. (2019). High-dimensional optimization in adaptive random subspaces. In *Proc. of the 33rd International Conference on Neural Information Processing Systems*, pages 10847–10857.

Lafferty, J. D., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*.

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

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

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

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

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

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

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

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

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

Rudi, A., Carratino, L., and Rosasco, L. (2017). Falkon: an optimal large scale kernel method. In *Proceedings of the 31st International Conference on Advances on Neural Information Processing Systems (NeurIPS)*, pages 3891–3901.

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

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

Sriperumbudur, B. K. and Szabó, Z. (2015). Optimal rates for random fourier features. In *NIPS*.

Taskar, B., Guestrin, C., and Koller, D. (2003). Max-margin markov networks. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press.

Tsochantaridis, I., Hofmann, T., Joachims, T., and Altun, Y. (2004). Support vector machine learning for interdependent and structured output spaces. *Machine Learning*.

Weston, J., Chapelle, O., Vapnik, V., Elisseeff, A., and Schölkopf, B. (2003). Kernel dependency estimation. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 897–904. MIT Press.

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

Woodruff, D. P. (2014). Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1-2):1–157.

Yang, T., Li, Y.-f., Mahdavi, M., Jin, R., and Zhou, Z.-H. (2012). Nyström method vs random fourier features: A theoretical and empirical comparison. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

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

Reminder: positive definite kernels and Reproducing Kernel Hilbert Space

Positive definite kernel: $k_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ such that

- for all $(z, z') \in \mathcal{Z}^2$, $k_{\mathcal{Z}}(z, z') = k_{\mathcal{Z}}(z', z)^\top$
- for all $n \in \mathbb{N}$ and any $(z_i, \alpha_i)_{i=1}^n \in (\mathcal{Z} \times \mathbb{R})^n$, $\sum_{i,j=1}^n \alpha_i \alpha_j k_{\mathcal{Z}}(z_i, z_j) \geqslant 0$

RKHS (Aronszajn, 1950): Hilbert space $\mathcal{H}_{\mathcal{Z}}$ of functions $f : \mathcal{Z} \to \mathbb{R}$ s. t. for all $f \in \mathcal{H}_{\mathcal{Z}}$ and $z \in \mathcal{Z}$

1. $z' \mapsto k_{\mathcal{Z}}(z, z') \in \mathcal{H}_{\mathcal{Z}}$,
2. $\langle f, k_{\mathcal{Z}}(\cdot, z) \rangle_{\mathcal{H}_{\mathcal{Z}}} = f(z)$ **(reproducing property)**.

Vector-Valued Reproducing Kernel Hilbert Space

Operator-valued kernel (Senkene and Tempel'man, 1973; Micchelli and Pontil, 2005; Carmeli et al., 2006, 2010): $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{F})$, where $\mathcal{F}$ is a Hilbert space, such that

- for all $(x, x') \in \mathcal{X}^2$, $\mathcal{K}(x, x') = \mathcal{K}(x', x)^{\#}$
- for all $n \in \mathbb{N}$ and any $(x_i, \varphi_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{F})^n$, $\sum_{i,j=1}^n \left\langle \varphi_i, \mathcal{K}\left(x_i, x_j\right) \varphi_j \right\rangle_{\mathcal{F}} \geqslant 0$

vv-RKHS: Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathcal{F}$ s. t. for all $f \in \mathcal{H}$, $\varphi \in \mathcal{F}$ and $x \in \mathcal{X}$

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

Background: Scalability to large datasets

1) Random Fourier Features (Rahimi and Recht, 2007; Rudi and Rosasco, 2017; Sriperumbudur and Szabó, 2015; Brault et al., 2016; Li et al., 2021)

Background: Scalability to large datasets

1) **Random Fourier Features** (Rahimi and Recht, 2007; Rudi and Rosasco, 2017; Sriperumbudur and Szabó, 2015; Brault et al., 2016; Li et al., 2021)

2) **Sketching (Mahoney et al., 2011; Woodruff, 2014):** dimension reduction approach based on random linear projections

- Nyström approximation ($\iff$ sub-sampling sketch) (Williams and Seeger, 2001; Drineas et al., 2005; Bach, 2013; Rudi et al., 2017; Meanti et al., 2020)
- Gaussian, Randomized Orthogonal Systems, sparse sketches etc. (Yang et al., 2017; Lacotte et al., 2019; Kpotufe and Sriperumbudur, 2020; Lacotte and Pilanci, 2020; Chen and Yang, 2021a; Gazagnadou et al., 2021)

Example: Sketching for scalar Kernel Ridge Regression ($\mathcal{Y} = \mathbb{R}$)

Representer theorem: $\hat{f} = \sum_{i=1}^{n} \hat{\alpha}_i k_x(\cdot, x_i)$, where

$$\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n)^\top = \underset{\alpha \in \mathbb{R}^n}{\arg\min}\ \alpha^\top \left(K_X^2 + n\lambda K_X\right) \alpha - 2Y^\top K_X \alpha$$

$$= \underbrace{(K_X + n\lambda I_n}_{n \times n})^{-1} Y$$

Example: Sketching for scalar Kernel Ridge Regression ($\mathcal{Y} = \mathbb{R}$)

Representer theorem: $\hat{f} = \sum_{i=1}^{n} \hat{\alpha}_i k_x(\cdot, x_i)$, where

$$\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n)^\top = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^n} \; \alpha^\top \left(K_X^2 + n\lambda K_X\right) \alpha - 2Y^\top K_X \alpha$$

$$= \underbrace{(K_X + n\lambda I_n)}_{n \times n}{}^{-1} Y$$

Let $m \ll n$, $R \in \mathbb{R}^{m \times n}$ be a random matrix: $\alpha \leftarrow R^\top \gamma$

Example: Sketching for scalar Kernel Ridge Regression ($\mathcal{Y} = \mathbb{R}$)

Representer theorem: $\hat{f} = \sum_{i=1}^{n} \hat{\alpha}_i k_x(\cdot, x_i)$, where

$$\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n)^\top = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^n} \alpha^\top \left(K_X^2 + n\lambda K_X\right) \alpha - 2Y^\top K_X \alpha$$

$$= \underbrace{(K_X + n\lambda I_n)}_{n \times n}{}^{-1} Y$$

Let $m \ll n$, $R \in \mathbb{R}^{m \times n}$ be a random matrix: $\alpha \leftarrow R^\top \gamma$

$\hat{f} \leftarrow \tilde{f} = \sum_{i=1}^{n} [R^\top \tilde{\gamma}]_i k_x(\cdot, x_i)$, where

$$\tilde{\gamma} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_m)^\top = \operatorname*{arg\,min}_{\gamma \in \mathbb{R}^m} \gamma^\top \left(R K_X^2 R^\top + n\lambda R K_X R^\top\right) \gamma - 2Y^\top K_X R^\top \gamma$$

$$= \underbrace{(R K_X^2 R^\top + n\lambda R K_X R^\top)}_{m \times m}{}^{\dagger} R K_X Y$$

Low-rank estimator

$$\hat{h}(x) = \sum_{i=1}^{n} \hat{\alpha}_i(x) \psi_{\mathcal{Y}}(y_i)\,, \quad \text{where} \quad \hat{\alpha}(x) = (K_X + n\lambda I_n)^{-1} k_X^x$$

Low-rank estimator

$$\begin{aligned}\hat{h}(x) &= \sum_{i=1}^{n} \hat{\alpha}_i(x)\psi_{\mathcal{Y}}(y_i)\,, \quad \text{where} \quad \hat{\alpha}(x) = (K_X + n\lambda I_n)^{-1} k_X^x \\ &= \sqrt{n} S_Y^{\#} \hat{\alpha}(x)\end{aligned}$$

Low-rank estimator

$$
\begin{aligned}
\hat{h}(x) &= \sum_{i=1}^{n} \hat{\alpha}_i(x)\psi_{\mathcal{Y}}(y_i)\,, \quad \text{where} \quad \hat{\alpha}(x) = (K_X + n\lambda I_n)^{-1}k_X^x \\
&= \sqrt{n}S_Y^{\#}\hat{\alpha}(x) \\
&= \sqrt{n}S_Y^{\#}(nS_XS_X^{\#} + n\lambda I_n)^{-1}\sqrt{n}S_X\psi_{\mathcal{X}}(x)
\end{aligned}
$$

Low-rank estimator

$$\begin{aligned}
\hat{h}(x) &= \sum_{i=1}^{n} \hat{\alpha}_i(x)\psi_{\mathcal{Y}}(y_i)\,, \quad \text{where} \quad \hat{\alpha}(x) = (K_X + n\lambda I_n)^{-1}k_X^x \\
&= \sqrt{n} S_Y^{\#} \hat{\alpha}(x) \\
&= \sqrt{n} S_Y^{\#}(n S_X S_X^{\#} + n\lambda I_n)^{-1}\sqrt{n} S_X \psi_{\mathcal{X}}(x) \\
\hat{h}(x) &= S_Y^{\#} S_X \big(S_X^{\#} S_X + \lambda I_{\mathcal{H}_{\mathcal{X}}}\big)^{-1}\psi_{\mathcal{X}}(x)
\end{aligned}$$

Low-rank estimator

$$
\begin{aligned}
\hat{h}(x) &= \sum_{i=1}^{n} \hat{\alpha}_i(x)\psi_{\mathcal{Y}}(y_i)\,, \quad \text{where} \quad \hat{\alpha}(x) = (K_X + n\lambda I_n)^{-1} k_X^x \\
&= \sqrt{n} S_Y^{\#} \hat{\alpha}(x) \\
&= \sqrt{n} S_Y^{\#} (n S_X S_X^{\#} + n\lambda I_n)^{-1} \sqrt{n} S_X \psi_{\mathcal{X}}(x) \\
\hat{h}(x) &= S_Y^{\#} S_X \left(S_X^{\#} S_X + \lambda I_{\mathcal{H}_{\mathcal{X}}}\right)^{-1} \psi_{\mathcal{X}}(x)
\end{aligned}
$$

Goal: Given orthogonal projectors $\widetilde{P}_X$ and $\widetilde{P}_Y$ onto subspaces of $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ resp.

Low-rank estimator

$$\begin{aligned}
\hat{h}(x) &= \sum_{i=1}^{n} \hat{\alpha}_i(x)\psi_{\mathcal{Y}}(y_i)\,, \quad \text{where} \quad \hat{\alpha}(x) = (K_X + n\lambda I_n)^{-1} k_X^x \\
&= \sqrt{n} S_Y^{\#} \hat{\alpha}(x) \\
&= \sqrt{n} S_Y^{\#} (n S_X S_X^{\#} + n\lambda I_n)^{-1} \sqrt{n} S_X \psi_{\mathcal{X}}(x) \\
\hat{h}(x) &= S_Y^{\#} S_X \left(S_X^{\#} S_X + \lambda I_{\mathcal{H}_{\mathcal{X}}}\right)^{-1} \psi_{\mathcal{X}}(x)
\end{aligned}$$

Goal: Given orthogonal projectors $\widetilde{P}_X$ and $\widetilde{P}_Y$ onto subspaces of $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ resp.

$$S_X^{\#} \leftarrow \widetilde{P}_X S_X^{\#} \quad \text{and} \quad S_Y^{\#} \leftarrow \widetilde{P}_Y S_Y^{\#}$$

Low-rank estimator

$$\hat{h}(x) = \sum_{i=1}^{n} \hat{\alpha}_i(x)\psi_{\mathcal{Y}}(y_i)\,, \quad \text{where} \quad \hat{\alpha}(x) = (K_X + n\lambda I_n)^{-1} k_X^x$$
$$= \sqrt{n} S_Y^{\#} \hat{\alpha}(x)$$
$$= \sqrt{n} S_Y^{\#} (n S_X S_X^{\#} + n\lambda I_n)^{-1} \sqrt{n} S_X \psi_{\mathcal{X}}(x)$$
$$\hat{h}(x) = S_Y^{\#} S_X \left(S_X^{\#} S_X + \lambda I_{\mathcal{H}_{\mathcal{X}}}\right)^{-1} \psi_{\mathcal{X}}(x)$$

Goal: Given orthogonal projectors $\widetilde{P}_X$ and $\widetilde{P}_Y$ onto subspaces of $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ resp.

$$S_X^{\#} \leftarrow \widetilde{P}_X S_X^{\#} \quad \text{and} \quad S_Y^{\#} \leftarrow \widetilde{P}_Y S_Y^{\#}$$

$$\tilde{h}(x) = \widetilde{P}_Y S_Y^{\#} S_X \widetilde{P}_X (\widetilde{P}_X S_X^{\#} S_X \widetilde{P}_X + \lambda I_{\mathcal{H}_{\mathcal{X}}})^{-1} \psi_{\mathcal{X}}(x)\,. \tag{19}$$

Complexity of IOKR and SISOKR for various types of sketching

Table 3: Time and space complexities at training and inference for the IOKR and SISOKR algorithms with sub-sampling, p-sparsified ($p \in (0, 1]$) or Gaussian sketching, for a test set of size n_{te} and a candidate set of size n_c, such that $n_{te} \leq m_{\mathcal{X}}, m_{\mathcal{Y}} < n \leq n_c$. For the sake of simplicity, we omit the $\mathcal{O}(\cdot)$ in the following.

	Training		Inference	
Method	Time	Space	Time	Space
IOKR	n^3	n^2	$n_{te} n n_c$	$n n_c$
SISOKR (sub-sampling)	$\max(m_{\mathcal{X}}, m_{\mathcal{Y}}) n$	$\max(m_{\mathcal{X}}, m_{\mathcal{Y}}) n$	$n_{te} m_{\mathcal{Y}} n_c$	$m_{\mathcal{Y}} n_c$
SISOKR (p-sparsified)	$\max(m_{\mathcal{X}}, m_{\mathcal{Y}})^2 pn$	$\max(m_{\mathcal{X}}, m_{\mathcal{Y}}) pn$	$\max(n_{te}, n m_{\mathcal{Y}} p) m_{\mathcal{Y}} n_c$	$n p m_{\mathcal{Y}} n_c$
SISOKR (Gaussian)	$\max(m_{\mathcal{X}}, m_{\mathcal{Y}}) n^2$	n^2	$n m_{\mathcal{Y}} n_c$	$n n_c$

Related works and differences

Rudi et al. (2015):

1. scalar kernel Ridge regression
2. sketching only applied in the input feature space
3. Nyström approximation with uniform or approximate leverage scores sampling

Related works and differences

Rudi et al. (2015):

1. **scalar** kernel Ridge regression
2. sketching **only** applied in the **input** feature space
3. **Nyström** approximation with **uniform** or **approximate leverage scores** sampling

Ciliberto et al. (2020):

1. **vector-valued** kernel Ridge regression, with possibly infinite-dimensional outputs
2. **no approximation** considered

Related works and differences

Rudi et al. (2015):

1. **scalar** kernel Ridge regression
2. sketching **only** applied in the **input** feature space
3. **Nyström** approximation with **uniform** or **approximate leverage scores** sampling

Ciliberto et al. (2020):

1. **vector-valued** kernel Ridge regression, with possibly infinite-dimensional outputs
2. **no approximation** considered

This work:

1. **vector-valued** kernel Ridge regression, with possibly infinite-dimensional outputs
2. sketching applied in **both** the **input and output** feature space
3. generic **sub-Gaussian** sketches

SISOKR Excess-Risk bound

Theorem (SISOKR excess-risk bound)

*Let $\delta \in [0,1]$, $n \in \mathbb{N}$ sufficiently large such that $\lambda = n^{-1/(1+\gamma_{\mathcal{X}})} \geq \frac{9\kappa_{\mathcal{X}}^2}{n}\log(\frac{n}{\delta})$. Under **Asm. 1, 2, 3 and 4**, the following holds with probability at least $1-\delta$*

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

where

$S(n) = c_1 \log(4/\delta) n^{-\frac{1}{2(1+\gamma_{\mathcal{X}})}}$ *(regression error)*

$A_{\rho_z}^{\psi_{\mathcal{Z}}}(\widetilde{P}_Z) = \mathbb{E}_z[\|(\widetilde{P}_Z - I_{\mathcal{H}_{\mathcal{Z}}})\psi_{\mathcal{Z}}(z)\|_{\mathcal{H}_{\mathcal{Z}}}^2]^{\frac{1}{2}}$ *(sketching reconstruction error)*

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

Sub-Gaussian sketch

Definition

A sub-Gaussian sketch $R_{\mathcal{Z}} \in \mathbb{R}^{m_{\mathcal{Z}} \times n}$ is composed with i.i.d. entries such that

$$\mathbb{E}\left[R_{\mathcal{Z}_{ij}}\right] = 0 \tag{20}$$

$$\mathbb{E}\left[R^2_{\mathcal{Z}_{ij}}\right] = 1/m \tag{21}$$

$$R_{\mathcal{Z}_{ij}} \sim \frac{\nu^2_{\mathcal{Z}}}{m} - \text{sub-Gaussian}, \quad \text{with} \quad \nu_{\mathcal{Z}} \geq 1 \tag{22}$$

Sub-Gaussian sketch

Definition

A sub-Gaussian sketch $R_{\mathcal{Z}} \in \mathbb{R}^{m_{\mathcal{Z}} \times n}$ is composed with i.i.d. entries such that

$$\mathbb{E}\left[R_{\mathcal{Z}_{ij}}\right] = 0 \tag{20}$$

$$\mathbb{E}\left[R^2_{\mathcal{Z}_{ij}}\right] = 1/m \tag{21}$$

$$R_{\mathcal{Z}_{ij}} \sim \frac{\nu_{\mathcal{Z}}^2}{m} - \text{sub-Gaussian}, \quad \text{with} \quad \nu_{\mathcal{Z}} \geq 1 \tag{22}$$

Examples:

- matrix composed with i.i.d. Gaussian entries

Sub-Gaussian sketch

Definition

A sub-Gaussian sketch $R_{\mathcal{Z}} \in \mathbb{R}^{m_{\mathcal{Z}} \times n}$ is composed with i.i.d. entries such that

$$\mathbb{E}\left[R_{\mathcal{Z}_{ij}}\right] = 0 \tag{20}$$

$$\mathbb{E}\left[R^2_{\mathcal{Z}_{ij}}\right] = 1/m \tag{21}$$

$$R_{\mathcal{Z}_{ij}} \sim \frac{\nu^2_{\mathcal{Z}}}{m} - \text{sub-Gaussian}, \quad \text{with} \quad \nu_{\mathcal{Z}} \geq 1 \tag{22}$$

Examples:

- matrix composed with i.i.d. Gaussian entries
- matrix composed with i.i.d. bounded random variables

Sub-Gaussian sketch

Definition

A sub-Gaussian sketch $R_{\mathcal{Z}} \in \mathbb{R}^{m_{\mathcal{Z}} \times n}$ is composed with i.i.d. entries such that

$$\mathbb{E}\left[R_{\mathcal{Z}_{ij}}\right] = 0 \tag{20}$$

$$\mathbb{E}\left[R^2_{\mathcal{Z}_{ij}}\right] = 1/m \tag{21}$$

$$R_{\mathcal{Z}_{ij}} \sim \frac{\nu_{\mathcal{Z}}^2}{m} - \text{sub-Gaussian}, \quad \text{with} \quad \nu_{\mathcal{Z}} \geq 1 \tag{22}$$

Examples:

- matrix composed with i.i.d. Gaussian entries
- matrix composed with i.i.d. bounded random variables
- matrix composed with i.i.d. Gaussian/bounded r.v. multiplied with independent Bernoulli r.v. (El Ahmad et al., 2023)

Sub-Gaussian Sketching Reconstruction Error

Theorem (Sub-Gaussian sketching reconstruction error)

Under Asm. 1, 2, 3 and 4, for $\delta \in (0, 1/e]$, $n \in \mathbb{N}$ *sufficiently large such that* $\frac{9}{n}\log(n/\delta) \le n^{-\frac{1}{1+\gamma_{\mathcal{Z}}}} \le \|C_{\mathcal{Z}}\|_{\mathrm{op}}/2$, *then if*

$$m_{\mathcal{Z}} \ge c_4 \max\left(\nu_{\mathcal{Z}}^2 n^{\frac{\gamma_{\mathcal{Z}}+\mu_{\mathcal{Z}}}{1+\gamma_{\mathcal{Z}}}}, \nu_{\mathcal{Z}}^4 \log\left(1/\delta\right)\right), \tag{23}$$

then with probability $1-\delta$

$$\mathbb{E}_z[\|(\widetilde{P}_Z - I_{\mathcal{H}_{\mathcal{Z}}})\psi_{\mathcal{Z}}(z)\|^2_{\mathcal{H}_{\mathcal{Z}}}] \le c_3 n^{-\frac{1-\gamma_{\mathcal{Z}}}{(1+\gamma_{\mathcal{Z}})}} \tag{24}$$

where $c_3, c_4 > 0$ *are constants independents of* $n, m_{\mathcal{Z}}, \delta$ *defined in the proofs.*

Synthetic Least Squares Regression

1) $n = 10,000$, $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, $d = 300$, $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ linear kernels $\implies \mathcal{H}_{\mathcal{X}} = \mathcal{H}_{\mathcal{Y}} = \mathbb{R}^d$

Synthetic Least Squares Regression

1) $n = 10,000$, $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, $d = 300$, $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ linear kernels $\implies \mathcal{H}_{\mathcal{X}} = \mathcal{H}_{\mathcal{Y}} = \mathbb{R}^d$

2) Construct covariance matrices $C_{\mathcal{X}}$ and E such that $\sigma_k(C_{\mathcal{X}}) = k^{-3/2}$ and $\sigma_k(E) = 0.2k^{-1/10}$

Synthetic Least Squares Regression

1) $n = 10,000$, $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, $d = 300$, $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ linear kernels $\implies$ $\mathcal{H}_{\mathcal{X}} = \mathcal{H}_{\mathcal{Y}} = \mathbb{R}^d$

2) Construct covariance matrices $C_{\mathcal{X}}$ and E such that $\sigma_k(C_{\mathcal{X}}) = k^{-3/2}$ and $\sigma_k(E) = 0.2k^{-1/10}$

3) Draw $H_0 \sim \mathcal{N}(0, I_d)$, and for $i \leq n$, $x_i \sim \mathcal{N}(0, C_{\mathcal{X}})$, $\epsilon_i \sim \mathcal{N}(0, E)$,

$$y_i = C_{\mathcal{X}} H_0 x_i + \epsilon_i \tag{25}$$

Synthetic Least Squares Regression

1) $n = 10,000$, $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, $d = 300$, $k_\mathcal{X}$ and $k_\mathcal{Y}$ linear kernels $\implies$ $\mathcal{H}_\mathcal{X} = \mathcal{H}_\mathcal{Y} = \mathbb{R}^d$

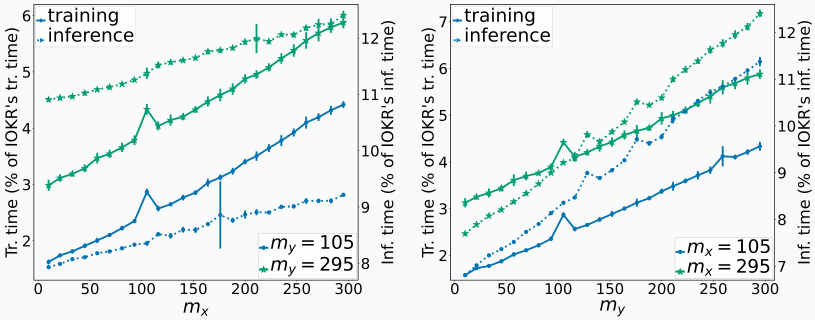
2) Construct covariance matrices $C_\mathcal{X}$ and E such that $\sigma_k(C_\mathcal{X}) = k^{-3/2}$ and $\sigma_k(E) = 0.2k^{-1/10}$

3) Draw $H_0 \sim \mathcal{N}(0, I_d)$, and for $i \leq n$, $x_i \sim \mathcal{N}(0, C_\mathcal{X})$, $\epsilon_i \sim \mathcal{N}(0, E)$,

$$y_i = C_\mathcal{X} H_0 x_i + \epsilon_i \tag{25}$$

4) $(2 \cdot 10^{-2})$-SR input and output sketches

Synthetic Least Squares Regression



(a) Training and inference time w.r.t. $m_{\mathcal{X}}$ for $m_{\mathcal{Y}} \in \{105, 295\}$

(b) Training and inference time w.r.t. $m_{\mathcal{Y}}$ for $m_{\mathcal{X}} \in \{105, 295\}$

Synthetic Least Squares Regression

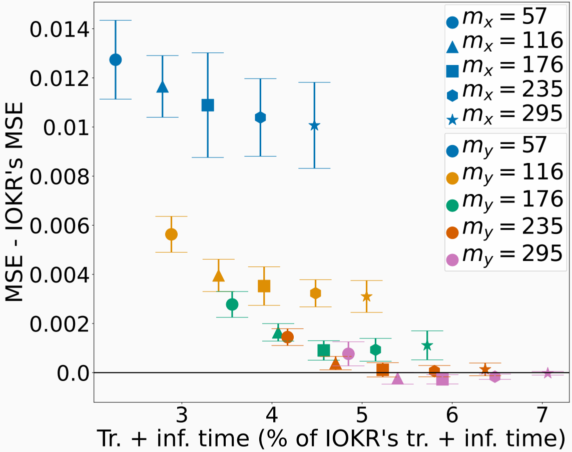
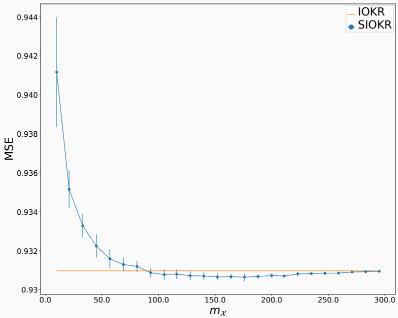
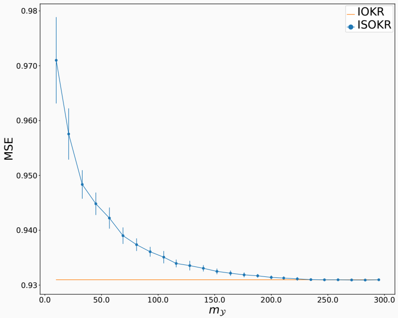


Figure 2: MSE w.r.t. learning time for different values of $m_{\mathcal{X}}$ and $m_{\mathcal{Y}}$

Synthetic Least Squares Regression



(a) SIOKR



(b) ISOKR

Multi-Label Classification

Bibtex and **Bookmarks** (Katakis et al., 2008): tag recommendation problems
Mediamill: detection of semantic concepts in a video

Table 4: Multi-label data sets description.

Data set	n	n_{te}	$n_{features}$	n_{labels}
Bibtex	4880	2515	1836	159
Bookmarks	60000	27856	2150	298
Mediamill	30993	12914	120	101

Multi-Label Classification: Statistical Performance

Table 5: F_1 scores on tag prediction from text data.

Method	Bibtex	Bookmarks	Mediamill
LR	37.2	30.7	NA
SPEN	42.2	34.4	NA
PRLR	44.2	34.9	NA
DVN	44.7	37.1	NA
SISOKR	44.1 ± 0.07	**39.3** ± 0.61	57.26 ± 0.04
ISOKR	44.8 ± 0.01	NA	58.02 ± 0.01
SIOKR	44.7 ± 0.09	39.1 ± 0.04	57.33 ± 0.04
IOKR	**44.9**	NA	**58.17**

Multi-Label Classification: Computational Performance

Table 6: Training/inference times (in seconds).

Method	Bibtex	Bookmarks	Mediamill
SISOKR	$\mathbf{1.41 \pm 0.03}$ / $\mathbf{0.46 \pm 0.01}$	$\mathbf{118 \pm 1.5}$ / $\mathbf{20 \pm 0.2}$	$\mathbf{66 \pm 0.1}$ / $\mathbf{4 \pm 0.01}$
ISOKR	2.51 ± 0.06 / 0.58 ± 0.01	NA	636 ± 3.7 9 ± 0.2
SIOKR	1.99 ± 0.07 / 1.22 ± 0.03	354 ± 2.1 / 297 ± 2.1	199 ± 0.1 / 121 ± 0.02
IOKR	2.54 / 1.18	NA	621 / 204

Metabolite identification

Inputs: tandem mass spectrum of a metabolite (small molecule

Outputs: molecular structure, i.e. fingerprints, encoded by binary vectors of length $d = 7593$

$n = 6974$ and each molecule is associated to a candidate set: median size = 292 and largest = 36,918 fingerprints

Table 7: MSE and standard errors for the metabolite identification problem. SPEN directly predicts outputs in $\mathcal{Y}$, then MSE is not defined.

Method	MSE	Tanimoto-Gaussian loss	Top-1 | 5 | 10 accuracies
SISOKR	0.813 ± 0.002	0.566 ± 0.007	25.1% | 54.2% | 64.7%
ISOKR	0.794 ± 0.003	0.509 ± 0.009	28.0% | 58.9% | 68.9%
SIOKR	0.793 ± 0.002	0.492 ± 0.008	29.5% | 61.3% | 70.9%
IOKR	**0.780** ± 0.002	**0.486** ± 0.008	**29.6%** | **61.6%** | **71.4%**
SPEN	NA	0.537 ± 0.008	25.9% | 54.1% | 64.3%

Metabolite identification

Table 8: Comparison of training/inference computation times (in seconds).

Method	Metabolite
SISOKR	4.05 ± 0.05 / $\mathbf{1112 \pm 29}$
ISOKR	6.25 ± 50.31 / 1133 ± 32
SIOKR	$\mathbf{1.25 \pm 0.02}$ / 1179 ± 37
IOKR	3.54 ± 0.15 / 1191 ± 38

p-Sparsified Sketches: Definition

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

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

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

Computational Property: *Decomposition trick*

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

$$R = R_{\mathrm{SG}}\, R_{\mathrm{SS}}\ ,$$

where

- $R_{\mathrm{SG}} \in \mathbb{R}^{m \times m'} \leftarrow$ deleting the null columns from R
- $R_{\mathrm{SS}} \in \mathbb{R}^{m' \times n} \leftarrow$ sampling the rows of I_n corresponding to the indices of non-zero columns of R.

Example:

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

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

Advantages of sub-sampling sketch

Let $X = \{x_1, \ldots, x_5\}$, $k_X^{x_i} = (k(x_i, x_1), \ldots, k(x_i, x_5))$ and

$$R_{\mathrm{SS}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Advantages of sub-sampling sketch

Let $X = \{x_1, \ldots, x_5\}$, $k_X^{x_i} = (k(x_i, x_1), \ldots, k(x_i, x_5))$ and
$R_{\mathrm{SS}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$

$$R_{\mathrm{SS}} K = \begin{pmatrix} k_X^{x_1} \\ k_X^{x_4} \end{pmatrix} \quad \text{and} \quad R_{\mathrm{SS}} K R_{\mathrm{SS}}^\top = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_4) \\ k(x_4, x_1) & k(x_4, x_4) \end{pmatrix}$$

$\iff$

1. Sample $X' = \{x_1, x_4\}$
2. Directly construct sub-Gram matrices $K_{X',X} \in \mathbb{R}^{2 \times 5}$ and $K_{X',X'} \in \mathbb{R}^{2 \times 2}$

Advantages of sub-sampling sketch

Let $X = \{x_1, \ldots, x_5\}$, $k_X^{x_i} = (k(x_i, x_1), \ldots, k(x_i, x_5))$ and

$$R_{\mathrm{SS}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$R_{\mathrm{SS}} K = \begin{pmatrix} k_X^{x_1} \\ k_X^{x_4} \end{pmatrix} \quad \text{and} \quad R_{\mathrm{SS}} K R_{\mathrm{SS}}^\top = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_4) \\ k(x_4, x_1) & k(x_4, x_4) \end{pmatrix}$$

$\iff$

1. Sample $X' = \{x_1, x_4\}$
2. Directly construct sub-Gram matrices $K_{X',X} \in \mathbb{R}^{2\times 5}$ and $K_{X',X'} \in \mathbb{R}^{2\times 2}$

$\implies$ No need to compute costly matrix multiplications!

$\implies$ No need to compute the whole K and store it in memory!

Time and Space Complexities of $R \cdot K_Z$

Let C_k be the cost of computing $k(x, x')$ for a couple $(x, x') \in \mathcal{X}^2$

- **Standard sketch (e.g. Gaussian):** $\mathcal{O}\left(C_k n^2 + n^2 m\right)$ and $\mathcal{O}\left(n^2\right)$,
- p**-sparsified sketch:** $\mathcal{O}\left(C_k n^2 mp + n^2 m^2 p\right)$ and $\mathcal{O}\left(n^2 mp\right)$.

$\implies$ Complexity reduction if $p < 1/m$

Goal of p-sparsified sketches and related works

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

Goal of p-sparsified sketches and related works

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