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

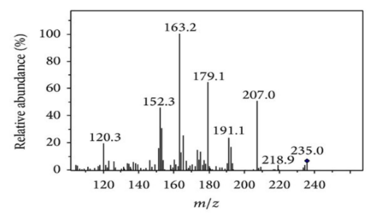
AISTATS 2024

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

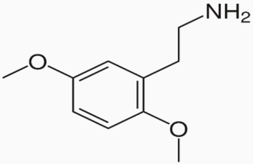
July 16, 2024

Structured prediction

Emblematic example of metabolite identification (Brouard et al., 2016a; Schymanski et al., 2017):



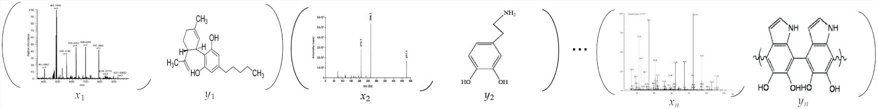
$\xrightarrow{f\ =\ ?}$



x y

Structured prediction in supervised settings

Supervised settings: n i.i.d. training sample $(x_i, y_i)_{i=1}^n \in (\mathcal{X}, \mathcal{Y})^n \sim \rho$

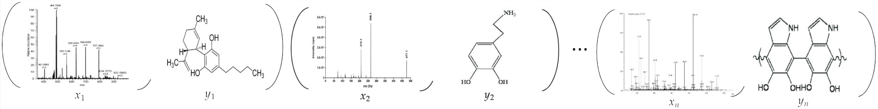


Given a loss function $\Delta : \mathcal{Y}^2 \to \mathbb{R}$

$$f^* = \operatorname*{arg\,inf}_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{(x,y)\sim\rho}[\Delta(f(x), y)] \approx \operatorname*{arg\,inf}_{f:\mathcal{X}\to\mathcal{Y}} \frac{1}{n}\sum_{i=1}^{n} \Delta(f(x_i), y_i) = \hat{f}$$

Structured prediction in supervised settings

Supervised settings: n i.i.d. training sample $(x_i, y_i)_{i=1}^n \in (\mathcal{X}, \mathcal{Y})^n \sim \rho$



Given a loss function $\boldsymbol{\Delta} : \mathcal{Y}^2 \to \mathbb{R}$

$$f^* = \underset{f:\mathcal{X}\to\mathcal{Y}}{\arg\inf}\ \mathbb{E}_{(x,y)\sim\rho}[\boldsymbol{\Delta}(f(x), y)] \approx \underset{f:\mathcal{X}\to\mathcal{Y}}{\arg\inf}\ \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Delta}(f(x_i), y_i) = \hat{f}$$

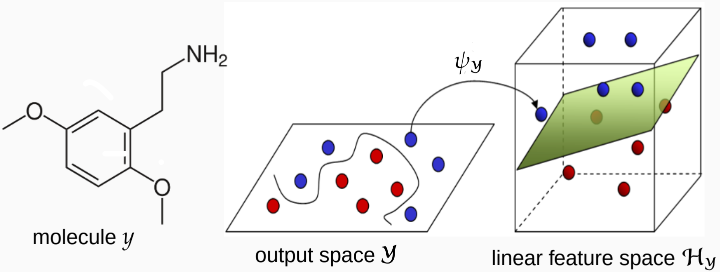
How to design a loss $\boldsymbol{\Delta}$ taking into account the structure of $\mathcal{Y}$?

Table of contents

Input Output Kernel Regression

Kernel methods: output representation

Linear method after embedding through feature map $\psi_{\mathcal{Y}} : \mathcal{Y} \to \mathcal{H}_{\mathcal{Y}}$:
choice of kernel $\iff$ choice of representation



$\langle \psi_{\mathcal{Y}}(y), \psi_{\mathcal{Y}}(y') \rangle_{\mathcal{H}_{\mathcal{Y}}} = k_{\mathcal{Y}}(y, y')$: relevant similarity measure over $\mathcal{Y}$

$$\Longrightarrow \; \boldsymbol{\Delta}(y, y') = \|\boldsymbol{\psi_{\mathcal{Y}}}(y) - \boldsymbol{\psi_{\mathcal{Y}}}(y')\|^2_{\boldsymbol{\mathcal{H}_{\mathcal{Y}}}} = 2 - 2\boldsymbol{k_{\mathcal{Y}}}(y, y')$$

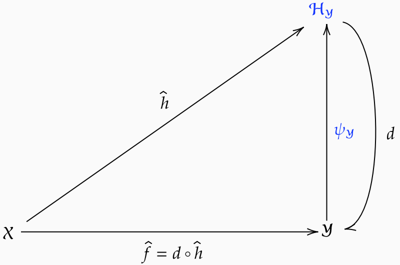
($\forall\, y \in \mathcal{Y}$, $\|\boldsymbol{\psi_{\mathcal{Y}}}\|_{\boldsymbol{\mathcal{H}_{\mathcal{Y}}}} = 1$ without loss of generality)

Versatility: tackle various tasks through an appropriate choice of $\boldsymbol{\psi_{\mathcal{Y}}}$ (e.g. SOTA performance on metabolite identification (Brouard et al., 2016a) and label ranking (Korba et al., 2018) datasets)

Output Kernel Regression: a surrogate approach

Surrogate (2-step) method (Weston et al., 2003; Cortes et al., 2005; Brouard et al., 2011; Kadri et al., 2013):

1. $\hat{h} = \underset{h:\mathcal{X}\to\mathcal{H}_\mathcal{Y}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \|h(x_i) - \psi_\mathcal{Y}(y_i)\|^2_{\mathcal{H}_\mathcal{Y}}$ (training step)
2. $\hat{f}(x) = d \circ \hat{h}(x) = \underset{y\in\mathcal{Y}}{\arg\min} \|\hat{h}(x) - \psi_\mathcal{Y}(y)\|^2_{\mathcal{H}_\mathcal{Y}}$ (inference step)



Theoretical guarantees: for measurable $h : \mathcal{X} \to \mathcal{H}_\mathcal{Y}$ and $f = d \circ h$, **f's excess risk is bounded by h's excess risk** (Ciliberto et al., 2020)

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

$$\widehat{f}(x) = d \circ \widehat{h}(x) = \arg\max_{y \in \mathcal{Y}} \sum_{i=1}^{n} \widehat{\alpha}(x)_i k_{\mathcal{Y}}(y_i, y)$$

IOKR: Weston et al. (2003); Cortes et al. (2005); Brouard et al. (2011); Kadri et al. (2013); Brouard et al. (2016b); Korba et al. (2018)

IOKR: training and inference complexities

1. **Training:** $\hat{h}(x) = \sum_{i=1}^{n} \hat{\boldsymbol{\alpha}}(x)_i \boldsymbol{\psi}_{\boldsymbol{\mathcal{Y}}}(y_i)$ where

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

$\implies \mathcal{O}\left(n^3\right)$ time complexity

IOKR: training and inference complexities

1. **Training:** $\hat{h}(x) = \sum_{i=1}^{n} \hat{\boldsymbol{\alpha}}(x)_i \boldsymbol{\psi_{\mathcal{Y}}}(y_i)$ where

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

$\implies \mathcal{O}\left(n^3\right)$ time complexity

2. **Inference:** $\hat{f}(x) = \operatorname*{arg\,max}_{y \in \mathcal{Y}} \sum_{i=1}^{n} \hat{\boldsymbol{\alpha}}(x)_i \boldsymbol{k_{\mathcal{Y}}}(y_i, y) = {k_X^x}^T \widehat{\boldsymbol{\Omega}} \boldsymbol{k_Y}^y$

- **Test set:** $X^{te} = \{x_1^{\mathrm{te}}, \ldots, x_{n_{\mathrm{te}}}^{\mathrm{te}}\}$ of size n_{te}
- **Candidate set:** $Y^{\mathrm{c}} = \{y_1^{\mathrm{c}}, \ldots, y_{n_{\mathrm{c}}}^{\mathrm{c}}\}$ of size n_{c}

$$\underbrace{K_X^{\mathbf{te,tr}}}_{n_{\mathrm{te}} \times n} \underbrace{\widehat{\boldsymbol{\Omega}}}_{n \times n} \underbrace{K_Y^{\mathbf{tr,c}}}_{n \times n_{\mathrm{c}}}$$

$$\hat{f}(x_i^{\mathrm{te}}) = y_j^{\mathrm{c}} \quad \text{where} \quad j = \operatorname*{arg\,max}_{1 \le j \le n_{\mathrm{c}}} [K_X^{\mathbf{te,tr}} \widehat{\boldsymbol{\Omega}} K_Y^{\mathbf{tr,c}}]_{ij}$$

$\implies \mathcal{O}\left(n_{\mathrm{te}} n n_{\mathrm{c}}\right)$ time complexity if $n_{\mathrm{te}} < n \le n_{\mathrm{c}}$

Challenges raised by IOKR

1. Scalability: obtain $\tilde{f} = d \circ \tilde{h}$, computationally efficient version of $\hat{f} = d \circ \hat{h}$, when learning from big data, i.e. large n

2. Theory: obtain excess risk bound of $\tilde{f} = d \circ \tilde{h}$

Key tool for scalability: Random Fourier Features vs Sketching

a) **Random Fourier Features** (Rahimi and Recht, 2007; Sriperumbudur and Szabó, 2015): for $m_{\mathcal{Y}} \ll n$,

$$\langle \psi_{\mathcal{Y}}(y), \psi_{\mathcal{Y}}(y') \rangle_{\mathcal{H}_{\mathcal{Y}}} \approx \langle \tilde{\psi}_{\mathcal{Y}}(y), \tilde{\psi}_{\mathcal{Y}}(y') \rangle_{\mathbb{R}^{m_{\mathcal{Y}}}}$$

$$\implies \boldsymbol{\Delta}(y, y') = \|\boldsymbol{\psi}_{\mathcal{Y}}(y) - \boldsymbol{\psi}_{\mathcal{Y}}(y')\|^2_{\mathcal{H}_{\mathcal{Y}}} \approx \|\tilde{\boldsymbol{\psi}}_{\mathcal{Y}}(y) - \tilde{\boldsymbol{\psi}}_{\mathcal{Y}}(y')\|^2_{\mathbb{R}^{m_{\mathcal{Y}}}} = \widetilde{\boldsymbol{\Delta}}(y, y')$$

$$\implies \widetilde{\boldsymbol{\Delta}} \text{ approximated loss}$$

Key tool for scalability: Random Fourier Features vs Sketching

a) **Random Fourier Features** (Rahimi and Recht, 2007; Sriperumbudur and Szabó, 2015): for $m_{\mathcal{Y}} \ll n$,

$$\langle \psi_{\mathcal{Y}}(y), \psi_{\mathcal{Y}}(y') \rangle_{\mathcal{H}_{\mathcal{Y}}} \approx \langle \tilde{\psi}_{\mathcal{Y}}(y), \tilde{\psi}_{\mathcal{Y}}(y') \rangle_{\mathbb{R}^{m_{\mathcal{Y}}}}$$

$$\implies \boldsymbol{\Delta}(y, y') = \|\boldsymbol{\psi}_{\mathcal{Y}}(y) - \boldsymbol{\psi}_{\mathcal{Y}}(y')\|^2_{\mathcal{H}_{\mathcal{Y}}} \approx \|\tilde{\boldsymbol{\psi}}_{\mathcal{Y}}(y) - \tilde{\boldsymbol{\psi}}_{\mathcal{Y}}(y')\|^2_{\mathbb{R}^{m_{\mathcal{Y}}}} = \widetilde{\boldsymbol{\Delta}}(y, y')$$

$$\implies \widetilde{\boldsymbol{\Delta}} \text{ approximated loss}$$

b) **Sketching** (Williams and Seeger, 2001; Rudi et al., 2015; Yang et al., 2017): for $m_{\mathcal{Y}} \ll n$, $R_{\mathcal{Y}} \in \mathbb{R}^{m_{\mathcal{Y}} \times n}$

$$\text{span}\left(\left(\psi_{\mathcal{Y}}(y_i)\right)_{i=1}^{n}\right) \leftarrow \text{span}\left(\left(\sum_{j=1}^{n} [R_{\mathcal{Y}}]_{ij} \psi_{\mathcal{Y}}(y_j)\right)_{i=1}^{m_{\mathcal{Y}}}\right)$$

$$\implies \boldsymbol{\Delta} \text{ remains unchanged!}$$

Sketched Input Sketched Output Kernel Regression

Motivation

Motivation: build a **low-rank** approximation $\tilde{h}$ of $\hat{h}$ 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 an i.i.d. sample $(z_i)_{i=1}^n \in \mathcal{Z}^n \sim \rho_Z$:

- $S_Z : f \in \mathcal{H}_{\mathcal{Z}} \mapsto (1/\sqrt{n})(\langle f, \psi_{\mathcal{Z}}(z_1)\rangle_{\mathcal{H}_{\mathcal{Z}}}, \ldots, \langle f, \psi_{\mathcal{Z}}(z_n)\rangle_{\mathcal{H}_{\mathcal{Z}}})^\top \in \mathbb{R}^n$ sampling operator

- $S_Z^\# : \alpha \in \mathbb{R}^n \mapsto (1/\sqrt{n}) \sum_{i=1}^n \alpha_i \psi_{\mathcal{Z}}(z_i) \in$ **span** $\left((\boldsymbol{\psi_{\mathcal{Z}}(z_i)})_{i=1}^n\right)$ its adjoint

- $C_{\mathcal{Z}} = \mathbb{E}_z[\psi_{\mathcal{Z}}(z) \otimes \psi_{\mathcal{Z}}(z)]$ 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:

  $\widehat{C}_Z : \mathcal{H}_{\mathcal{Z}} \to$ **span** $\left((\boldsymbol{\psi_{\mathcal{Z}}(z_i)})_{i=1}^n\right)$

$$
\begin{tikzcd}[column sep=huge, row sep=huge]
\mathcal{H}_{\mathcal{X}} \arrow[rr, "\widehat{H} = S_Y^{\#} S_X \left(S_X^{\#} S_X + \lambda I_{\mathcal{H}_{\mathcal{X}}}\right)^{-1}"] & & \mathcal{H}_{\mathcal{Y}} \arrow[dd, bend left, "d"] \\
 & & \\
\mathcal{X} \arrow[uu, "\psi_{\mathcal{X}}"] \arrow[uurr, "\widehat{h}(x) = \sum_{i=1}^{n} \widehat{\alpha}(x)_i \psi_{\mathcal{Y}}(y_i) = \widehat{H} \psi_{\mathcal{X}}(x)"] \arrow[rr, "\widehat{f}(x) = d \circ \widehat{h}(x) = \arg\max_{y \in \mathcal{Y}} \sum_{i=1}^{n} \widehat{\alpha}(x)_i k_{\mathcal{Y}}(y_i, y)"'] & & \mathcal{Y} \arrow[uu, "\psi_{\mathcal{Y}}"']
\end{tikzcd}
$$

Low-rank estimator: from IOKR to SISOKR

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

$$\widetilde{P}_Z : \mathcal{H}_\mathcal{Z} \to \widetilde{\mathcal{H}}_\mathcal{Z} \text{ where } \widetilde{\mathcal{H}}_\mathcal{Z} := \operatorname{span}\left(\left(\sum_{j=1}^{n} [R_\mathcal{Z}]_{ij} \psi_\mathcal{Z}(z_j)\right)_{i=1}^{m_\mathcal{Z}}\right)$$

How to build these projectors?

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

- $\widehat{C}_Z = S_Z^{\#} S_Z = (1/n) \sum_{i=1}^{n} \psi_{\mathcal{Z}}(z_i) \otimes \psi_{\mathcal{Z}}(z_i)$
- $\widetilde{C}_Z = S_Z^{\#} R_{\mathcal{Z}}^{\top} R_{\mathcal{Z}} S_Z = \frac{1}{n} \sum_{l=1}^{m_{\mathcal{Z}}} \left(\sum_{i=1}^{n} R_{\mathcal{Z}_{li}} \psi_{\mathcal{Z}}(z_i) \right) \otimes \left(\sum_{j=1}^{n} R_{\mathcal{Z}_{lj}} \psi_{\mathcal{Z}}(z_j) \right)$
- $\widetilde{K}_Z = R_{\mathcal{Z}} K_Z R_{\mathcal{Z}}{}^{\top}$, and $\left\{ \left(\sigma_i(\widetilde{K}_Z), \tilde{\mathbf{u}}_i^Z \right), i \in [m_{\mathcal{Z}}] \right\}$ its eigenpairs
- $p_Z = \operatorname{rank}\left(\widetilde{K}_Z\right)$, and for all $1 \leq i \leq p_Z$, $\tilde{\boldsymbol{e}}_i^Z = \sqrt{\frac{n}{\sigma_i(\widetilde{K}_Z)}} S_Z^{\#} R_{\mathcal{Z}}^{\top} \tilde{\mathbf{u}}_i^Z \in \mathcal{H}_{\mathcal{Z}}$

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

- $\widehat{C}_Z = S_Z^{\#} S_Z = (1/n) \sum_{i=1}^{n} \psi_{\mathcal{Z}}(z_i) \otimes \psi_{\mathcal{Z}}(z_i)$
- $\widetilde{C}_Z = S_Z^{\#} R_{\mathcal{Z}}^{\top} R_{\mathcal{Z}} S_Z = \frac{1}{n} \sum_{l=1}^{m_{\mathcal{Z}}} \left(\sum_{i=1}^{n} R_{\mathcal{Z}_{li}} \psi_{\mathcal{Z}}(z_i)\right) \otimes \left(\sum_{j=1}^{n} R_{\mathcal{Z}_{lj}} \psi_{\mathcal{Z}}(z_j)\right)$
- $\widetilde{K}_Z = R_{\mathcal{Z}} K_Z R_{\mathcal{Z}}{}^{\top}$, and $\left\{\left(\sigma_i(\widetilde{K}_Z), \tilde{\mathbf{u}}_i^Z\right), i \in [m_{\mathcal{Z}}]\right\}$ its eigenpairs
- $p_Z = \operatorname{rank}\left(\widetilde{K}_Z\right)$, and for all $1 \leq i \leq p_Z$, $\tilde{\mathbf{e}}_i^Z = \sqrt{\frac{n}{\sigma_i(\widetilde{K}_Z)}} S_Z^{\#} R_{\mathcal{Z}}^{\top} \tilde{\mathbf{u}}_i^Z \in \mathcal{H}_{\mathcal{Z}}$

Proposition (El Ahmad et al., 2024)

The $\tilde{\mathbf{e}}_i^Z$s are the **eigenfunctions**, associated to the eigenvalues $\sigma_i(\widetilde{K}_Z)/n$, of $\widetilde{C}_Z$, whose range is $\mathbf{span}\left(\left(\sum_{j=1}^{n} R_{\mathcal{Z}_{ij}} \psi_{\mathcal{Z}}(z_j)\right)_{i=1}^{m_{\mathcal{Z}}}\right)$. Then, $\widetilde{E}^Z = (\tilde{\mathbf{e}}_1^Z, \ldots, \tilde{\mathbf{e}}_{p_Z}^Z)$ is an **orthonormal basis** of $\mathbf{span}\left(\left(\sum_{j=1}^{n} R_{\mathcal{Z}_{ij}} \psi_{\mathcal{Z}}(z_j)\right)_{i=1}^{m_{\mathcal{Z}}}\right)$, and $\widetilde{P}_Z$ writes as

$$\widetilde{P}_Z = \sum_{i=1}^{p_Z} \langle \cdot, \tilde{\mathbf{e}}_i^Z \rangle_{\mathcal{H}_{\mathcal{Z}}} \tilde{\mathbf{e}}_i^Z = (R_{\mathcal{Z}} S_Z)^{\#} \left(R_{\mathcal{Z}} S_Z (R_{\mathcal{Z}} S_Z)^{\#}\right)^{\dagger} R_{\mathcal{Z}} S_Z \,.$$

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

Sketched Input Sketched Output Kernel Regression estimator

Proposition (El Ahmad et al., 2024)

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

with

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

Sketched Input Sketched Output Kernel Regression estimator

Proposition (El Ahmad et al., 2024)

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

with

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

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 sketching matrix, between $\mathcal{O}(nm_{\mathcal{Z}})$ and $\mathcal{O}(n^2 m_{\mathcal{Z}})$

$\Longrightarrow$ **Training complexity reduced thanks to input sketching!**

SISOKR estimator: Inference

$$\tilde{f}(x) = \underset{y \in \mathcal{Y}}{\arg\max} \; \sum_{i=1}^{n} \tilde{\alpha}(x)_i k_{\mathcal{Y}}(y_i, y) = \underset{y \in \mathcal{Y}}{\arg\max} \; {k_X^x}^T R_{\mathcal{X}}^\top \widetilde{\Omega} R_{\mathcal{Y}} k_Y{}^y$$

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

$$\tilde{f}(x_i^{\text{te}}) = y_j^{\text{c}} \quad \text{where} \quad j = \underset{1 \le j \le n_{\text{c}}}{\arg\max} \, [K_X^{\mathbf{te},\mathbf{tr}} R_{\mathcal{X}}^\top \widetilde{\Omega} R_{\mathcal{Y}} K_Y^{\mathbf{tr},\mathbf{c}}]_{ij}$$

SISOKR estimator: Inference

$$\tilde{f}(x) = \underset{y \in \mathcal{Y}}{\arg\max} \sum_{i=1}^{n} \tilde{\boldsymbol{\alpha}}(x)_i \boldsymbol{k_{\mathcal{Y}}}(y_i, y) = \underset{y \in \mathcal{Y}}{\arg\max} \, {k_X^x}^T R_{\mathcal{X}}^\top \widetilde{\Omega} R_{\mathcal{Y}} \boldsymbol{k_Y}^y$$

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

$$\tilde{f}(x_i^{\mathrm{te}}) = y_j^{\mathrm{c}} \quad \text{where} \quad j = \underset{1 \leq j \leq n_{\mathrm{c}}}{\arg\max} \, [K_X^{\mathbf{te,tr}} R_{\mathcal{X}}^\top \widetilde{\Omega} R_{\mathcal{Y}} K_Y^{\mathbf{tr,c}}]_{ij}$$

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

$\Longrightarrow$ Inference complexity reduced thanks to output sketching!

Scalability ✓!

Theoretical guarantees

Theoretical guarantees of SISOKR

Let

$$\mathcal{R}(f) = \mathbb{E}_{(x,y)\sim\rho}[\boldsymbol{\Delta}(f(x), y)]\,,$$

and

$$f^* = \underset{f:\mathcal{X}\to\mathcal{Y}}{\arg\inf}\ \mathbb{E}_{(x,y)\sim\rho}[\boldsymbol{\Delta}(f(x), y)]\,,$$

we want to control

$$\mathcal{R}(\tilde{f}) - \mathcal{R}(f^*) \leq \quad \textbf{?}$$

Assumptions

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

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

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

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

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

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

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

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

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 $\nu_{\mathcal{Z}} \geq 1$.

Theorem: SISOKR learning rates (El Ahmad et al., 2024)

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 $\mathcal{Z} \in \{\mathcal{X}, \mathcal{Y}\}$ and for $n \in \mathbb{N}$ sufficiently large such that $\frac{9}{n} \log(n/\delta) \leq n^{-\frac{1}{1+\gamma_{\mathcal{Z}}}} \leq \|C_{\mathcal{Z}}\|_{\mathrm{op}}/2$, and for sketching sizes $m_{\mathcal{Z}}, \in \mathbb{N}$ such that

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

then with probability $1-\delta$

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

and

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

Theory ✓!

Experiments

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 $\implies$ $\mathcal{H}_{\mathcal{X}} = \mathcal{H}_{\mathcal{Y}} = \mathbb{R}^d$

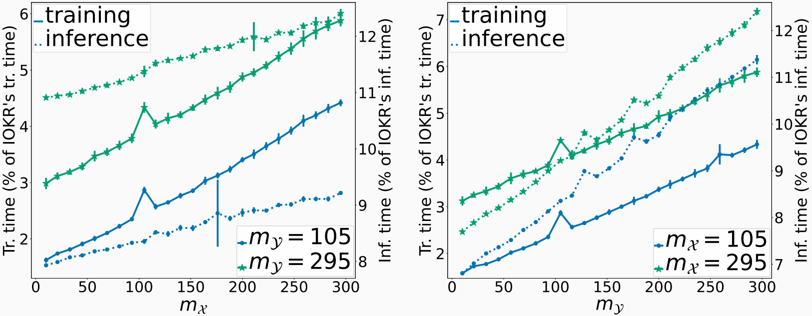
2) Construct covariance matrices $C_{\mathcal{X}}$ and E such that $\sigma_k(C_{\mathcal{X}}) = k^{-3/2}$ and $\sigma_k(E) = 0.2k^{-1/10}$

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

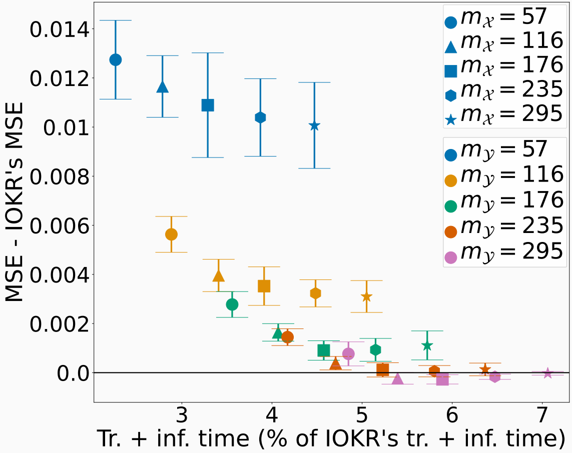
$$y_i = C_{\mathcal{X}} H_0 x_i + \epsilon_i$$

4) $20/n$-SR input and output sketches (sub-Gaussian)

Synthetic least squares regression



Synthetic least squares regression



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	$\mathbf{1.41 \pm 0.03}$ / $\mathbf{0.46 \pm 0.01}$	$\mathbf{118 \pm 1.5}$ / $\mathbf{20 \pm 0.2}$
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

Synthetic and real-world experiments: take-home messages

1) a) Input sketching: mainly accelerates the **training** phase

1) b) Output sketching: accelerates the **inference** phase

Synthetic and real-world experiments: take-home messages

1) a) **Input sketching:** mainly accelerates the **training** phase

1) b) **Output sketching:** accelerates the **inference** phase

2) **Optimal computational/statistical trade-off:** statistical performance **converges** when $m_{\mathcal{X}}/m_{\mathcal{Y}}$ **increases** $\implies$ **no need to set them too high!**

Synthetic and real-world experiments: take-home messages

1) a) Input sketching: mainly accelerates the **training** phase

1) b) Output sketching: accelerates the **inference** phase

2) Optimal computational/statistical trade-off: statistical performance **converges** when $m_{\mathcal{X}} / m_{\mathcal{Y}}$ **increases** $\implies$ **no need to set them too high!**

3) Benefits from sketching w.r.t. the number of training data n:

small	**intermediate**	**large**
No benefit from sketching	SISOKR **accelerates** IOKR while being **as accurate**	SISOKR is **tractable unlike** IOKR

n →

Conclusion

Conclusion

- SISOKR: sketching on both input/output kernels to accelerate both training/inference steps
- Sketching as a way to build orthogonal projectors onto low-dimensional subspace of the feature space
- Excess risk bound leading to a consistent theoretical analysis of SISOKR
- Experiments: SISOKR accelerates IOKR or make it tractable

References

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

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

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

Caldarelli, E., Chatalic, A., Colomé, A., Molinari, C., Ocampo-Martinez, C., Torras, C., and Rosasco, L. (2024). Linear quadratic control of nonlinear systems with koopman operator learning and the nyström method.

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.

El Ahmad, T., Brogat-Motte, L., Laforgue, P., and d'Alché Buc, F. (2024). Sketch in, sketch out: Accelerating both learning and inference for structured prediction with kernels. In Dasgupta, S., Mandt, S., and Li, Y., editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 109–117. PMLR.

Kadri, H., Ghavamzadeh, M., and Preux, P. (2013). A generalized kernel approach to structured output learning. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 471–479, Atlanta, Georgia, USA. PMLR.

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.

Korba, A., Garcia, A., and d'Alché-Buc, F. (2018). A structured prediction approach for label ranking. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Meanti, G., Chatalic, A., Kostic, V. R., Novelli, P., massimiliano pontil, and Rosasco, L. (2023). Estimating koopman operators with sketching to provably learn large scale dynamical systems. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

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

Schymanski, E., Ruttkies, C., and Krauss, M. e. a. (2017). Critical assessment of small molecule identification 2016: automated methods. *J Cheminform*, 9:22.

Sriperumbudur, B. K. and Szabó, Z. (2015). Optimal rates for random fourier features. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1144–1152, Cambridge, MA, USA. MIT Press.

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.

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.

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	$n m_{\mathcal{Y}} n_c$	$n n_c$

Sketching sizes selection strategy

Goal: set the minimal value of $m_{\mathcal{Z}}$ s.t. it captures the information contained in the empirical covariance operator $\widehat{C}_Z = \frac{1}{n}\sum_{i=1}^n \psi_{\mathcal{Z}}(z_i) \otimes \psi_{\mathcal{Z}}(z_i)$

However: computing the SVD of $\widehat{C}_Z$ is costing, i.e. $\mathcal{O}(n^3)$ in time.

1. Approximate leverage scores of $\widehat{C}_X$ and $\widehat{C}_Y$

2. Empirical approach: given training/inference budgets of time $T_{\mathrm{tr}}/T_{\mathrm{inf}}$, set low $m_{\mathcal{X}}$ and $m_{\mathcal{Y}}$ and evaluate the performance of $\tilde{f}$ until reaching one of the following condition:

- convergence of the performance of $\tilde{f}$
- training time attains T_{tr} or inference time attains T_{te}

Selection of $m_{\mathcal{X}}$

$\tilde{h}^{\text{SIOKR}}(x) = \sum_{i=1}^{n} \tilde{\alpha}_i^{\text{SIOKR}}(x)\, \psi_{\mathcal{Y}}(y_i)$ where

$$\tilde{\alpha}^{\text{SIOKR}}(x) = K_X R_{\mathcal{X}}^\top (R_{\mathcal{X}} K_X^2 R_{\mathcal{X}}^\top + n\lambda R_{\mathcal{X}} K_X R_{\mathcal{X}}^\top)^\dagger$$

Set the optimal $m_{\mathcal{X}}$ according to a training budget of time $T_{\mathbf{tr}}$ and the performance of $\tilde{h}^{\text{SIOKR}}$ in terms of surrogate regression error on the validation set, i.e. minimizing

$$\sum_{i=1}^{n_{\text{val}}} \left\| \tilde{h}^{\text{SIOKR}}(x_i^{\text{val}}) - \psi_{\mathcal{Y}}(y_i^{\text{val}}) \right\|_{\mathcal{H}_{\mathcal{Y}}}^2$$

$$= \sum_{i=1}^{n_{\text{val}}} \tilde{\alpha}^{\text{SIOKR}}\left(x_i^{\text{val}}\right)^\top K_Y \tilde{\alpha}^{\text{SIOKR}}\left(x_i^{\text{val}}\right) - 2\tilde{\alpha}^{\text{SIOKR}}\left(x_i^{\text{val}}\right)^\top k_Y^{y_i^{\text{val}}} + k_{\mathcal{Y}}(y_i^{\text{val}}, y_i^{\text{val}})$$

$\Longrightarrow$ allows to cope with the inference phase

Selection of $m_{\mathcal{Y}}$

Set the optimal $m_{\mathcal{Y}}$ according to an inference budget of time T_{inf} and the performance of the *perfect* h estimator on the validation set, i.e.

$$h : (x, y) \mapsto \widetilde{P}_Y \psi_{\mathcal{Y}}(y)$$

$$f(x_i^{\mathsf{val}}) = y_j^{\mathsf{c}} \quad \text{where} \quad j = \operatorname*{arg\,max}_{1 \leq j \leq n_{\mathsf{c}}} [K_Y^{\mathsf{val},\mathsf{tr}} R_{\mathcal{Y}}^\top \widetilde{K}_Y^\dagger R_{\mathcal{Y}} K_Y^{\mathsf{tr},\mathsf{c}}]_{ij}$$

$\implies$ allows to cope with the training phase

Theory: previous 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 (El Ahmad et al., 2024):

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

Related recent works on Koopman operators: (Meanti et al., 2023; Caldarelli et al., 2024)

SISOKR excess risk bound

Theorem (El Ahmad et al., 2024)

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

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

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

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

Sub-Gaussian sketching reconstruction error

Theorem (El Ahmad et al., 2024)

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

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

then with probability $1-\delta$

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

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

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	4 880	2 515	1 836	159
Bookmarks	60 000	27 856	2 150	298
Mediamill	30 993	12 914	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	$\mathbf{39.3} \pm 0.61$	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	**1.41 $\pm$ 0.03 / 0.46 $\pm$ 0.01**	**118 $\pm$ 1.5 / 20 $\pm$ 0.2**	**66 $\pm$ 0.1 / 4 $\pm$ 0.01**
ISOKR	2.51 $\pm$ 0.06 / 0.58 $\pm$ 0.01	NA	636 $\pm$ 3.7 9 $\pm$ 0.2
SIOKR	1.99 $\pm$ 0.07 / 1.22 $\pm$ 0.03	354 $\pm$ 2.1 / 297 $\pm$ 2.1	199 $\pm$ 0.1 / 121 $\pm$ 0.02
IOKR	2.54 / 1.18	NA	621 / 204

Metabolite identification

Inputs: tandem mass spectra of metabolites

Outputs: molecular structures, i.e. fingerprints, encoded by binary vectors of length $d = 7593 \rightarrow$ **probability product kernel**

$n = 5\,579$ and **each molecule is associated with a specific candidate set**: **median size = 292** and **largest = 36 918** fingerprints $\rightarrow$ **Gaussian-Tanimoto kernel**

Method	kernel loss	Top-1 | 5 | 10 accuracies	training	inference
SPEN	0.537 ± 0.008	25.9% | 54.1% | 64.3%	NA	NA
SISOKR	0.566 ± 0.007	25.1% | 54.2% | 64.7%	4.05 ± 0.05	$\mathbf{1112 \pm 29}$
ISOKR	0.509 ± 0.009	28.0% | 58.9% | 68.9%	6.25 ± 50.31	1133 ± 32
SIOKR	0.492 ± 0.008	29.5% | 61.3% | 70.9%	$\mathbf{1.25 \pm 0.02}$	1179 ± 37
IOKR	$\mathbf{0.486 \pm 0.008}$	**29.6%** | **61.6%** | **71.4%**	3.54 ± 0.15	1191 ± 38