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

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

⋆ LTCI, Télécom Paris, † L2S, CentraleSupélec, ‡ Università degli Studi di Milano

Motivation

Problem. Learn a decision function $\hat{f} : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ is a structured space.

Existing works. $\hat{f} = d \circ \hat{h}$: 2-step surrogate method based on input/output kernels [1, 2, 3].

1. **generic** (i.e., able to handle different tasks);
2. **grounded theoretically**;
3. **simple algorithmically**;
4. **not scalable** (both in training and inference).

We want to 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 $\tilde{f}$.

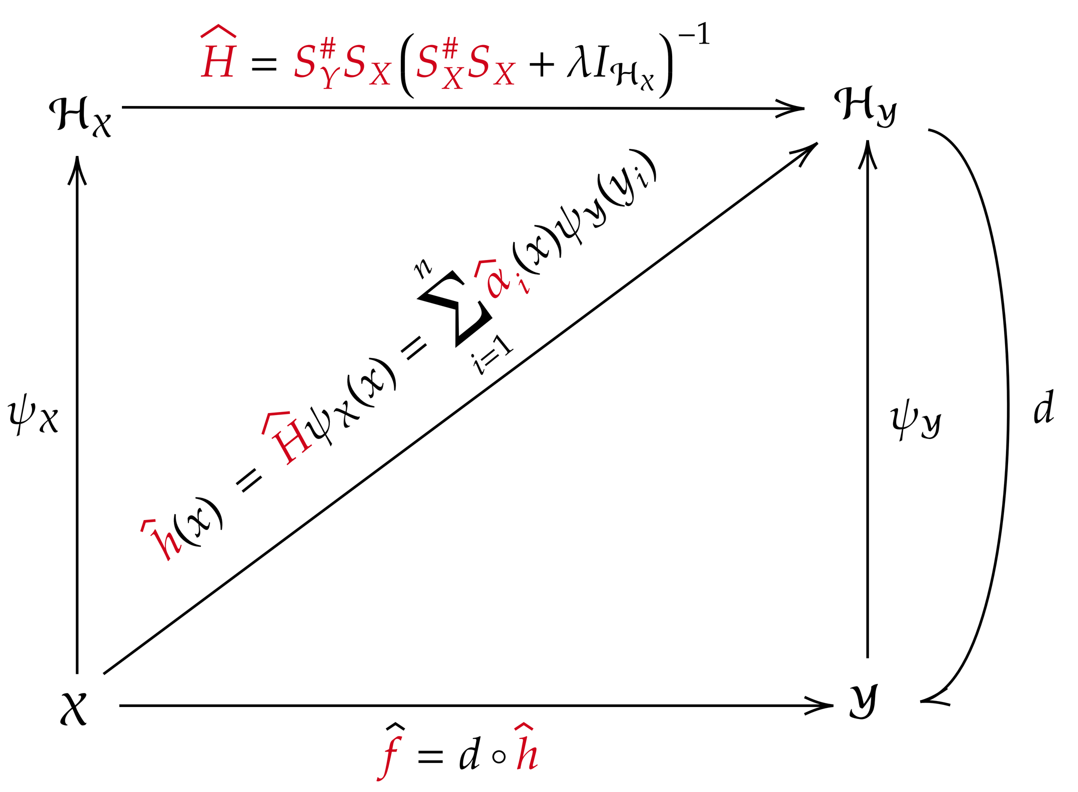
Some Notations

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

Given an i.i.d. sample $(z_i)_{i=1}^n \in \mathcal{Z}^n \sim \rho_\mathcal{Z}$, let

- $S_Z : f \in \mathcal{H}_\mathcal{Z} \mapsto \frac{1}{\sqrt{n}}(f(z_1), \dots, f(z_n)) \in \mathbb{R}^n$
- $S_Z^\# : \alpha \in \mathbb{R}^n \mapsto \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha_i \psi_\mathcal{Z}(z_i) \in \mathcal{H}_\mathcal{Z}$
- $K_Z = (k_\mathcal{Z}(z_i, z_j))_{1 \le i,j \le n} = n S_Z S_Z^\#$
- $C_Z = \mathbb{E}_z[\psi_\mathcal{Z}(z) \otimes \psi_\mathcal{Z}(z)]$
- $\widehat{C}_Z = (1/n) \sum_{i=1}^n \psi_\mathcal{Z}(z_i) \otimes \psi_\mathcal{Z}(z_i) = S_Z^\# S_Z$

Input Output Kernel Regression



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

Complexity: $\mathcal{O}(n^3)$

Inference: for a candidate set $\mathcal{Y}_c \subseteq \mathcal{Y}$ of size n_c

$$d(\psi_y(y)) = \underset{y' \in \mathcal{Y}_c}{\operatorname{argmin}} \ \|\psi_y(y) - \psi_y(y')\|^2_{\mathcal{H}_y}$$

For a test set X_{te} of size n_{te}

$$\underbrace{K_X^{te,tr}}_{n_{te} \times n} \underbrace{(K_X + n\lambda I_n)}_{n \times n}^{-1} \underbrace{K_Y^{tr,c}}_{n \times n_c}$$

Complexity: $\mathcal{O}(n^2 n_c)$ if $n_{te} < n \le n_c$

References

[1] Weston et al. *Kernel dependency estimation.* NeurIPS '03

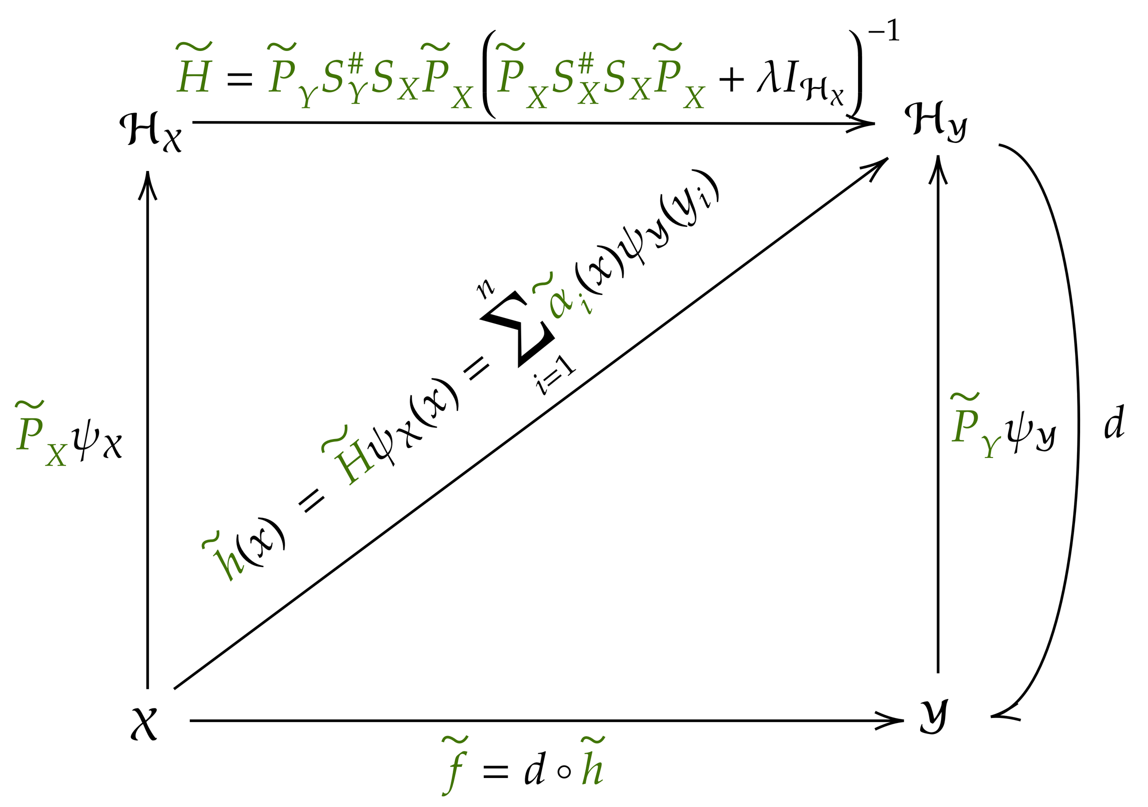
[2] Brouard et al. *Input output kernel regression: supervised and semi-supervised structured output prediction with operator-valued kernels.* JMLR '16.

[3] Ciliberto et al. *A general framework for consistent structured prediction with implicit loss embeddings.* JMLR '20.

[4] Rudi et al. *Less is more: Nyström computational regularization.* NeurIPS '15.

SISOKR: low-rank estimator

Contribution: build a **low-rank** approximation $\tilde{h}$ of $\hat{h}$ thanks to orthogonal projectors $\widetilde{P}_X$ and $\widetilde{P}_Y$.



How to build $\widetilde{P}_X$ and $\widetilde{P}_Y$? By sketching [4], i.e., **random linear projections**: let $m_\mathcal{Z} \ll n$ and $R_\mathcal{Z} \in \mathbb{R}^{m_\mathcal{Z} \times n}$ be a random matrix,

$$\widetilde{P}_\mathcal{Z} = (R_\mathcal{Z} S_\mathcal{Z})^\# \left(R_\mathcal{Z} S_\mathcal{Z} (R_\mathcal{Z} S_\mathcal{Z})^\#\right)^\dagger R_\mathcal{Z} S_\mathcal{Z}$$

Training: $\tilde{\alpha}(x) = R_\mathcal{Y}^\top \widetilde{\Omega}_Y \widetilde{\Omega}_X R_\mathcal{X} k_X^x$ with

$$\widetilde{\Omega}_Y = \underbrace{(R_\mathcal{Y} K_Y R_\mathcal{Y}^\top)}_{m_\mathcal{Y} \times m_\mathcal{Y}}^\dagger R_\mathcal{Y} K_Y ,$$

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

Complexity: $\mathcal{O}(m_\mathcal{X}^3 + m_\mathcal{Y}^3)$

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

Complexity: $\mathcal{O}(n_{te} m_\mathcal{Y} n_c)$ if $n_{te} \le m_\mathcal{X}, m_\mathcal{Y}$

Theoretical Guarantees

A 1 (Attainability) $\exists H : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ s.t. $\|H\| < \infty$ and $h^*(x) := \mathbb{E}_y[\psi_\mathcal{Y}(y) \mid x] = H\psi_\mathcal{X}(x)$.

A 2 (Bounded kernel) $k_\mathcal{Z}(z, z) \le \kappa_\mathcal{Z}^2, \forall z \in \mathcal{Z}$.

A 3 (Capacity) $Q_\mathcal{Z} := \operatorname{Tr}(C_\mathcal{Z}^{\gamma_\mathcal{Z}}) < +\infty$.

A 4 (Embedding) $\psi_\mathcal{Z}(z) \otimes \psi_\mathcal{Z}(z) \preceq b_\mathcal{Z} C_\mathcal{Z}^{1-\mu_\mathcal{Z}}$ a.s.

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

Theorem (SISOKR learning rate). *Assume that* ***A 1-5*** *hold, and that* $\|\psi_\mathcal{Y}(y)\|_{\mathcal{H}_\mathcal{Y}} = \kappa_\mathcal{Y}$. *For* $n \in \mathbb{N}$ *s.t.* $\frac{9}{n}\log(n/\delta) \le n^{-\frac{1}{1+\gamma_\mathcal{Z}}} \le \|C_\mathcal{Z}\|_{\text{op}}/2$, *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 $1 - \delta$ *we have*

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

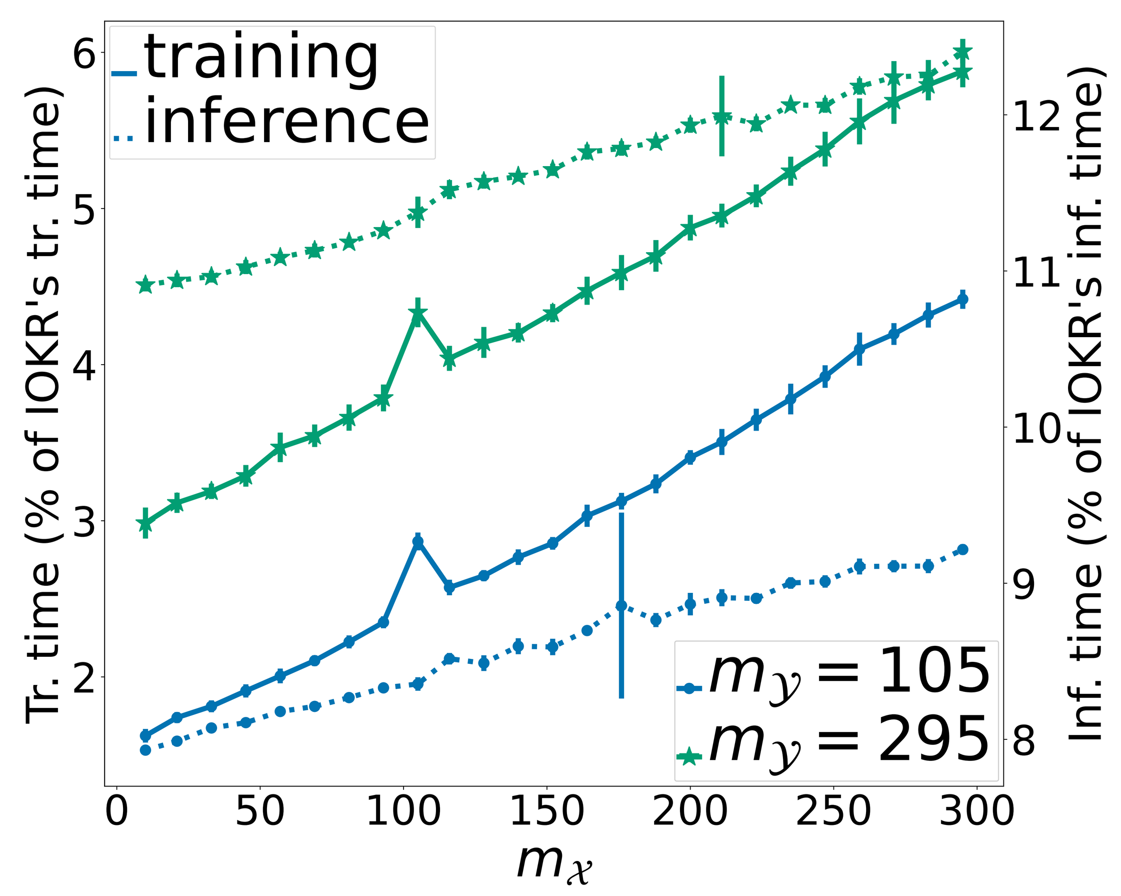
where $\mathcal{R}(f) = \mathbb{E}_{(x,y)\sim\rho}\left[\|\psi_\mathcal{Y}(y) - \psi_\mathcal{Y}(f(x))\|^2_{\mathcal{H}_\mathcal{Y}}\right]$.

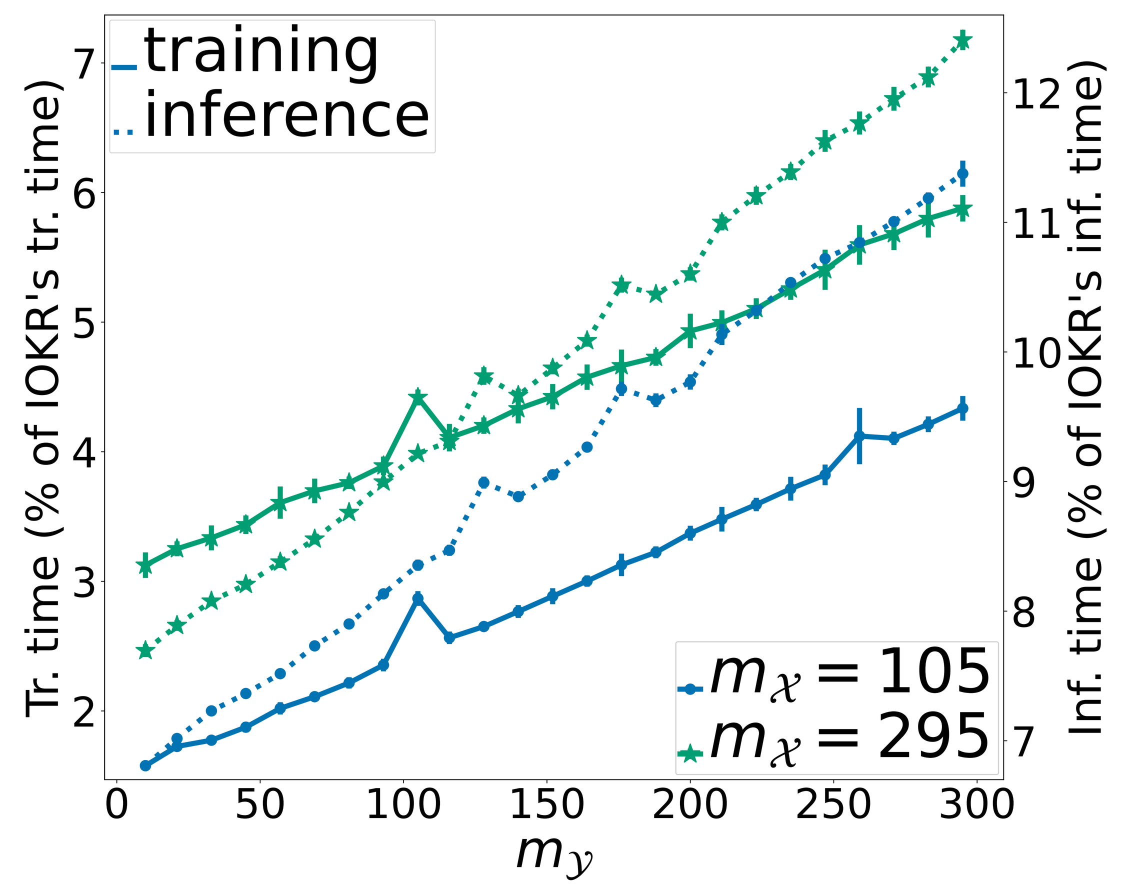
Acknowledgements

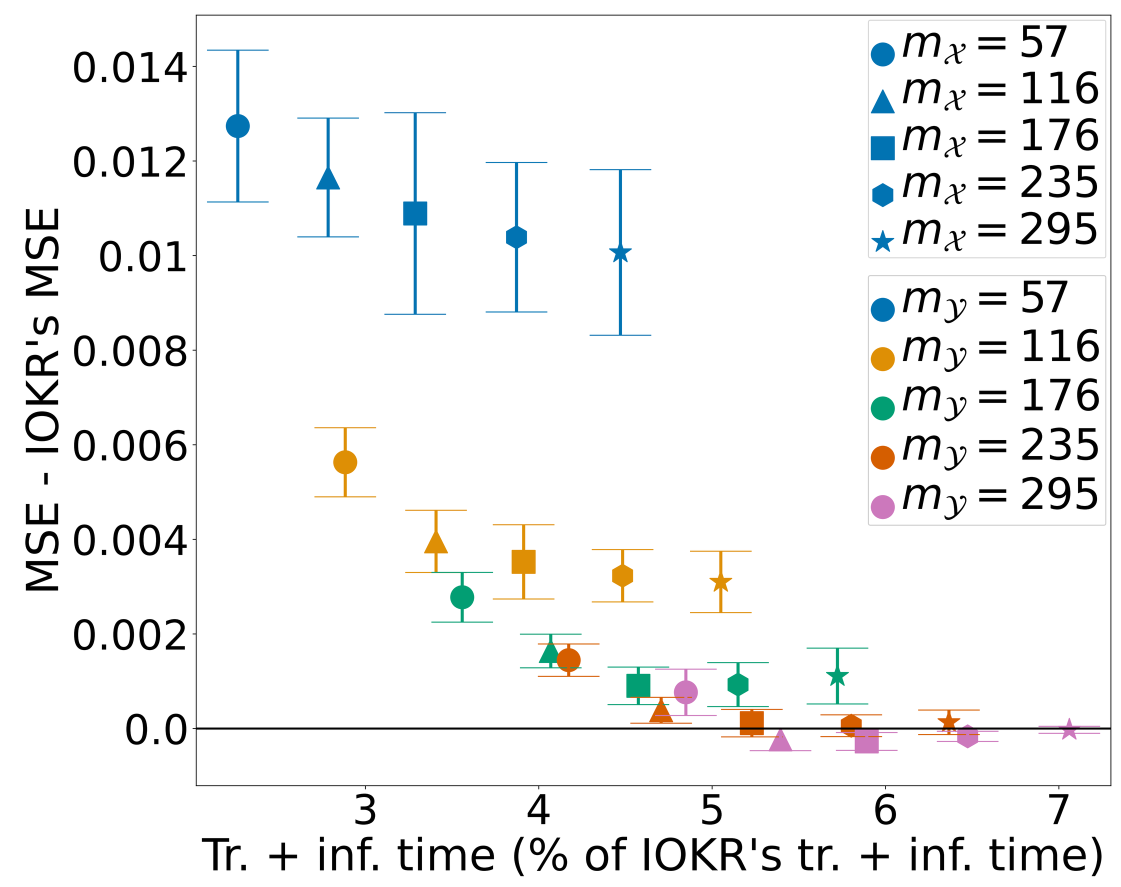
The authors thank the Télécom Paris research chair on Data Science and Artificial Intelligence for Digitalized Industry and Services (DSAIDIS).

Experiments

Synthetic least squares regression:







Real-world multi-label classification:

F1 scores

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

Training/inference times (in sec)

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