Supervised Kernel Thinning


Albert Gong


Kyuseong Choi


Raaz Dwivedi

Cornell University

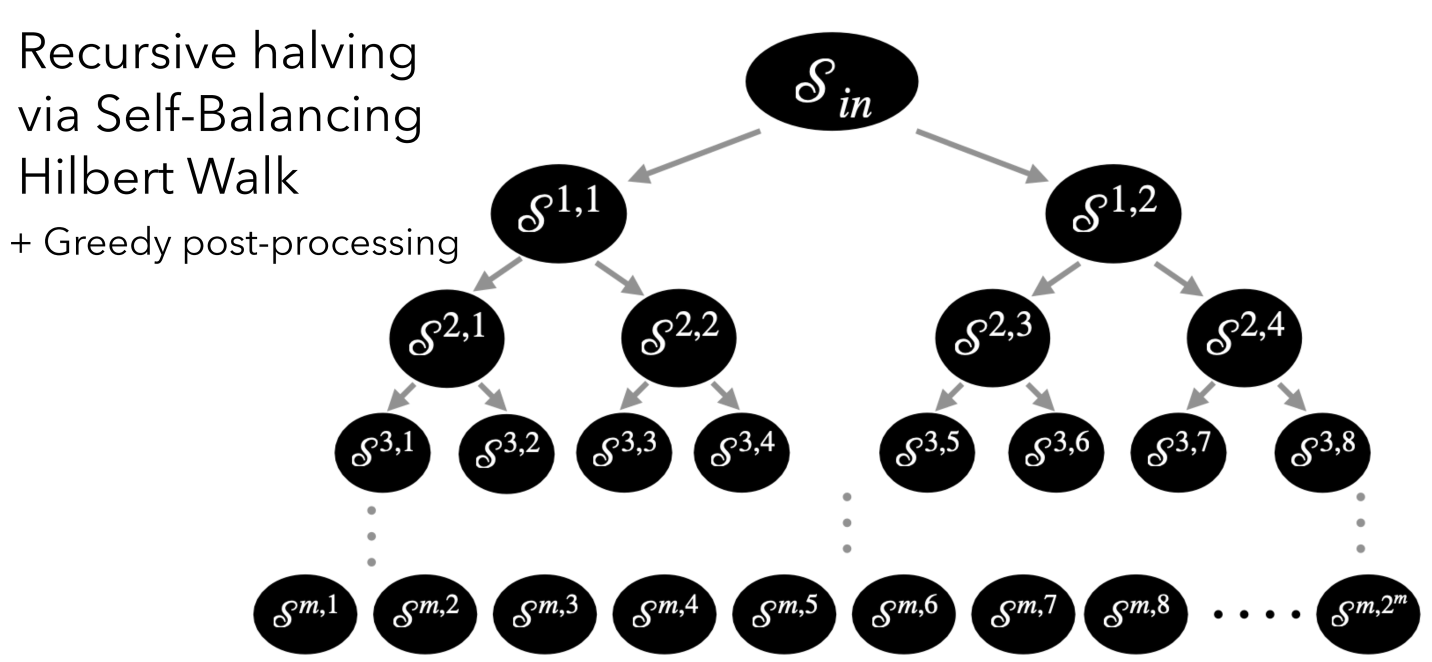
CORNELL TECH — HOME OF THE JACOBS TECHNION-CORNELL INSTITUTE

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation: Kernel methods are powerful ways of fitting regression models.
Problem: Computationally slow when sample size is large. E.g., n^3 training and n inference time with sample size n for kernel ridge regression
Goal: Speed-up without loss of statistical accuracy.
Idea: Use distribution compression algorithms, in particular kernel thinning.

Unsupervised Kernel Thinning

$x_1, \dots, x_n$



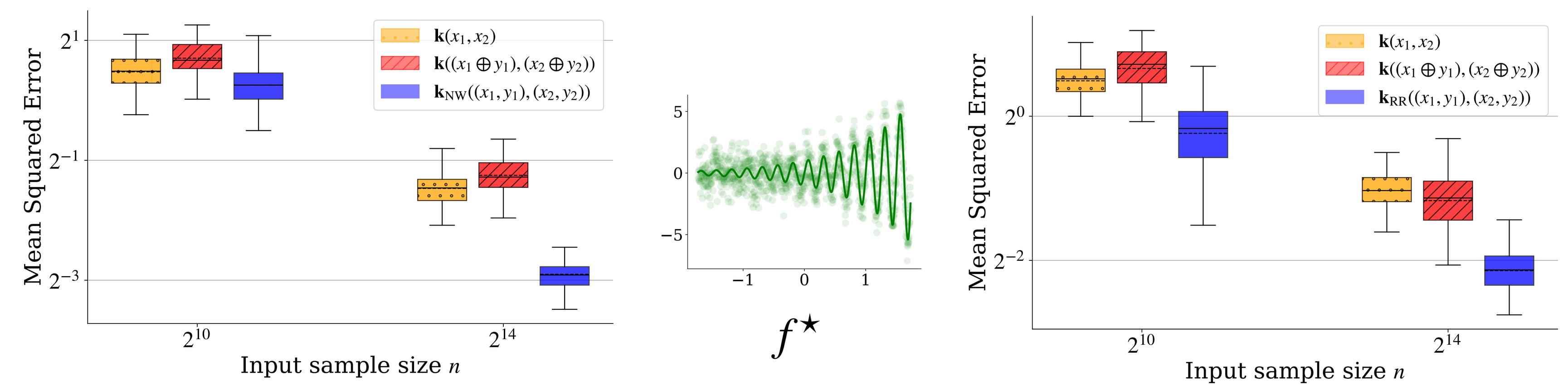
$x'_1, \dots, x'_{n_{\text{out}}}$

$$\left|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \frac{1}{n_{\text{out}}}\sum_{i=1}^{n_{\text{out}}} f(x'_i)\right| \lesssim \frac{\|f\|_{\mathbf{k}}\sqrt{\log(n_{\text{out}})}}{n_{\text{out}}}$$

1. Valid for f lying in the RKHS of $\mathbf{k}$
2. Minimax even with $n_{\text{out}} = \sqrt{n}$ for Gaussian $\mathbf{k}$ and various set of input points
3. Near-linear runtime when $n_{\text{out}} = \sqrt{n}$

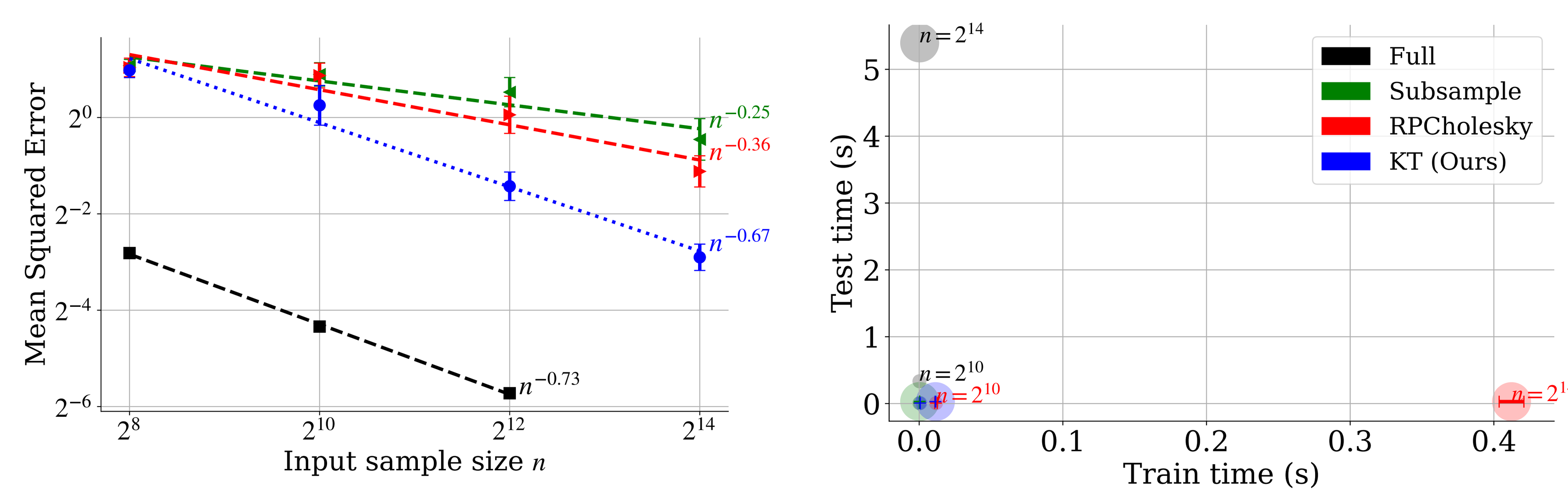
Dwivedi & Mackey '21, '22, '24
Shetty-Dwivedi-Mackey '22
Domingo-Enrich-Dwivedi-Mackey '23
Li-Dwivedi-Mackey '24

What happens if we directly apply unsupervised kernel thinning? Speed-up but with poor accuracy.

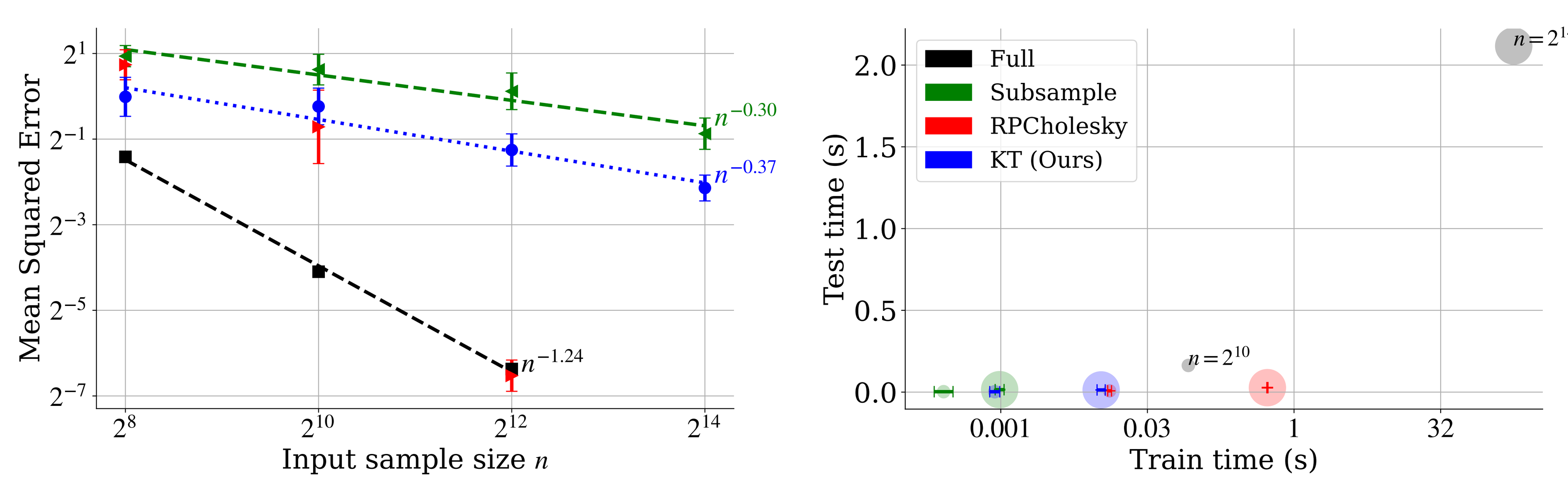


Key Innovation

1. Express the problem or solution as an average over functions
2. Identify kernel $\mathbf{k}'$ whose RKHS contains these functions
3. Apply Kernel thinning with $(x_i, y_i)_{i=1}^n$ with $\mathbf{k}'$
4. Enjoy 10^2 to 10^5x speed-up and better-than-i.i.d. error rates!



Kernel smoothing with $\mathbf{k}$ = Wendland



Kernel ridge regression with $\mathbf{k}$ = Gaussian

Kernel Smoothing (Nadaraya-Watson)

$$\hat{f}_{\text{NW}}(x) = \frac{\frac{1}{n}\sum_{i=1}^n y_i \mathbf{k}(x, x_i)}{\frac{1}{n}\sum_{i=1}^n \mathbf{k}(x, x_i)}$$

i. $\mathbf{k}(x, \cdot)$ lies in the RKHS of $\mathbf{k}$
ii. $(x', y') \mapsto y' \cdot \mathbf{k}(x, x')$ lies in the RKHS of $y_1 y_2 \cdot \mathbf{k}(x_1, x_2)$!

$\Downarrow$

Both denominator and numerator functions lie in the RKHS of
$\mathbf{k}(x_1, x_2) + y_1 y_2 \cdot \mathbf{k}(x_1, x_2)$

Kernel Ridge Regression (KRR)

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

i. f^2 lies in RKHS of $\mathbf{k}^2(x_1, x_2)$
ii. $(x, y) \mapsto y \cdot f(x)$ lies in RKHS of $y_1 y_2 \cdot \mathbf{k}(x_1, x_2)$
iii. y^2 lies in RKHS of $(y_1 y_2)^2$

$\Downarrow$

KRR loss lies in the RKHS of
$\mathbf{k}^2(x_1, x_2) + y_1 y_2 \cdot \mathbf{k}(x_1, x_2) + (y_1 y_2)^2$

	Nadaraya-Watson			KRR		
	Full	Sub-sample	Ours*	Full	Sub-sample	Ours**
MSE	$n^{-\frac{2\beta}{2\beta+d}}$	$n^{-\frac{\beta}{2\beta+d}}$	$n^{-\frac{\beta}{\beta+d}}$	$\sigma^2 \frac{m}{n}$	$\sigma^2 \frac{m}{\sqrt{n}}$	$\frac{m}{n}\|f^\star\|_{\mathbf{k}}^2$
Training	n	$\sqrt{n}$	$n \log^3 n$	n^3	$n^{1.5}$	$n^{1.5}$
Inference	n	$\sqrt{n}$	$\sqrt{n}$	n	$\sqrt{n}$	$\sqrt{n}$

Assumptions:
* $f^\star$ is β Holder for $\beta \in (0,2]$, $\mathbf{k}$ has compact support, and $n_{\text{out}} = \sqrt{n}$
** $f^\star$ is in the RKHS of $\mathbf{k}$, $\mathbf{k}$ has rank m, and $n_{\text{out}} = \sqrt{n}$

$$\frac{\left|\frac{1}{n}\sum_{i=1}^n f^2(x_i) - \frac{1}{n_{\text{out}}}\sum_{i=1}^{n_{\text{out}}} f^2(x'_i)\right|}{\frac{1}{n}\sum_{i=1}^n f^2(x_i)} \lesssim \frac{\sqrt{m \log(n_{\text{out}})}}{n_{\text{out}}}$$

when compressing with $\mathbf{k}^2$ for finite rank $\mathbf{k}$