Paper Review: Randomized sketches for kernels: Fast and optimal non-parametric regression

Yun Yang, Mert Pilanci and Martin J. Wainwright

Group Meeting · Meimei Liu · Nov 03, 2015

Outline

Model Description

- Consider the model

$$y_i = f^*(x_i) + \sigma w_i, \text{for } i = 1, 2, \cdots, n$$

where the regression function $f^*(x) = E[Y|x]$, the sequence $\{w_i\}_{i=1}^n$ consists of i.i.d. standard Gaussian variates.

- Suppose f^* belong to a RKHS with kernel function $\mathcal{K}(\cdot, \cdot)$, define the empirical kernel matrix K with entries $K_{ij} = n^{-1}\mathcal{K}(x_i, x_j)$. The KRR estimator

$$\hat{f}_{KRR} := argmin_{f \in \mathcal{H}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n ||f||_{\mathcal{H}}^2 \right\} \quad (1)$$

- By Representer Theorem, $\hat{f}_{KRR}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i^{\dagger} \mathcal{K}(\cdot, x_i)$, with w_i obtained by solving the quadratic program

$$w^{\dagger} = argmin_{w \in \mathbb{R}^n} \left\{ \frac{1}{2} w^T K^2 w - w^T \frac{Ky}{\sqrt{n}} + \lambda_n w^T K w \right\} \quad (2)$$

- KRR estimator could achieve the minimax prediction error for various classes of kernels.

- By Representer Theorem, $\hat{f}_{KRR}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i^{\dagger} \mathcal{K}(\cdot, x_i)$, with w_i obtained by solving the quadratic program

$$w^{\dagger} = argmin_{w \in \mathbb{R}^n} \left\{ \frac{1}{2} w^T K^2 w - w^T \frac{Ky}{\sqrt{n}} + \lambda_n w^T K w \right\} \quad (2)$$

- KRR estimator could achieve the minimax prediction error for various classes of kernels.
- But consider the computational complexity:
  - The n dimensional quadratic program in (2) requires $\mathcal{O}(n^3)$ via QR decomposition.
  - The n dimensional matrix K is dense in general, so requires storage of order n^2 numbers.

Randomized Sketch Approach

- This paper considers approximations to KRR based on random projections of the data. Define a sketch matrix as $S \in \mathbb{R}^{m \times n}$, where $m \ll n$ is the projection dimension.
- $K \to SK$: approximate K by projecting its row and column subspaces to a randomly chosen $m-$dimensional subspace.
- The sketched KRR estimate $\hat{f}(\cdot) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (S^T \hat{\alpha})_i \mathcal{K}(\cdot, x_i)$ is given by first solving

$$\hat{\alpha} = argmin_{\alpha \in \mathbb{R}^m} \left\{ \frac{1}{2} \alpha^T (SK)(KS^T) \alpha - \alpha^T S \frac{Ky}{\sqrt{n}} + \lambda_n \alpha^T SKS^T \alpha \right\} \quad (3)$$

Purpose of this Paper

- i.e., change the n dimensional quadratic program to m dimensional. The computational complexity becomes $\mathcal{O}(m^3)$.
- The purpose of this paper:
  - what classes of random projection matrices could be used?
  - how small can the projection dimension m be chosen while still retaining minimax optimality of the original KRR estimate?

several classes of random sketch matrices S

- Sub-Gaussian sketches: the row s_i is zero-mean 1-sub-Gaussian as for any fixed unit vector $u \in \mathbb{R}^n$, we have

$$P(|\langle u, s_i \rangle \geq t|) \leq 2 \exp^{-\frac{n\delta^2}{2}} \textit{ for all } \delta \geq 0$$

For example, matrices with i.i.d. Gaussian entries, i.i.d. Bernoulli entries

- Randomized orthogonal system (ROS) sketches: i.i.d. rows s_i has the form

$$s_i = \sqrt{\frac{n}{m}} R H^T p_i, for\ i = 1, \cdots, m,$$

where R is a random diagonal matrix whose entries are i.i.d. Rademacher variables, H is a fixed orthonormal matrix $H \in \mathbb{R}^{n \times n}$, e.g., the Hadamard matrix. $\{p_i\}$ is a random subset of m rows sampled uniformly from $I_{n \times n}$.

- Sub-sampling sketches: $s_i = \sqrt{\frac{n}{m}} p_i$, where the $\{p_1, \cdots, p_m\}$ are drawn uniformly at random without replacement from $I_{n \times n}$.
- this is equivalent to the Nystrom approximation.

Kernel Complexity measures

The eigendecomposition of kernel matrix $K = UDU^T$, where U is an orthonormal matrix, D is diagonal with $\hat{\mu}_1 \geq \hat{\mu}_2 \geq \cdots \geq \hat{\mu}_n \geq 0$.

- Kernel complexity function:

$$\hat{\mathcal{R}}(\delta) = \sqrt{\frac{1}{n}\sum_{j=1}^{n} \min\{\delta^2, \hat{\mu}_j\}}$$

  i.e., a rescaled sum of the eigenvalues, truncated at level δ^2.

- critical radius is the smallest positive solution $\delta_n > 0$ to the inequality

$$\frac{\hat{\mathcal{R}}(\delta)}{\delta} \leq \frac{\delta}{\sigma}$$

Statistical dimension

Define the statistical dimension of the kernel as

$$d_n := argmin_{j=1,\cdots,n}\{\hat{\mu}_j \leq \delta_n^2\}$$

i.e., $\hat{\mu}_j > \delta_n^2$ for all $j \in \{1, 2, \cdots, d_n\}$. And

$$\hat{\mathcal{R}}(\delta_n) = [\frac{d_n}{n}\delta_n^2 + \frac{1}{n}\sum_{j=d_n+1}^{n} \hat{\mu}_j]^{1/2}$$

d_n controls a type of bias-variance tradeoff.
when the tail sum $\sum_{j=d_n+1}^{n} \hat{\mu}_j \preceq d_n\delta_n^2$, $\delta_n^2 \asymp \frac{\sigma^2 d_n}{n}$.

Theorem 1 (Critical radius and minimax risk)

Theorem

Given n i.i.d. samples $\{(y_i, x_i)\}_{i=1}^n$ from the standard non-parametric regression model over any regular kernel class, any estimator $\tilde{f}$ has prediction error lower bounded as

$$\sup_{\|f^*\|_{\mathcal{H}} \leq 1} \mathbb{E}\|\tilde{f} - f^*\|_n^2 \geq c_l \delta_n^2$$

where $c_l > 0$ is a numerical constant, and δ_n is the critical radius.

Remark: The critical radius is a fundamental lower bound on the performance of any estimator.

K-satisfiable Condition for sketched kernel optimality

For $K = UDU^T$, let $U_1 \in \mathbb{R}^{n \times d_n}$ be the left bolck of U, $U_2 \in \mathbb{R}^{n \times (n-d_n)}$ be the right block.

- Say S is K-satisfiable if there is a universal constant c such that
$$|||(SU_1)^T SU_1 - I_{d_n}|||_{op} \leq 1/2$$
$$|||SU_2 D_2^{1/2}|||_{op} \leq c\delta_n$$
where $D_2 = diag\{\hat{\mu}_{d_n+1}, \cdots, \hat{\mu}_n\}$.
- Intuitively, a sketch matrix S is "good" if the sub-matrix $SU_1 \in \mathbb{R}^{m \times d_n}$ is relatively close to an isometry, whereas $SU_2 \in \mathbb{R}^{m \times (n-d_n)}$ has a relatively small operator norm.

Theorem 2 (Upper Bound for the sketched KRR)

Theorem

Given n i.i.d. samples $\{(y_i, x_i)\}_{i=1}^n$ *from the standard nonparametric regression model, consider the sketched KRR problem (3) based on a K-satisfiable sketch matrix S. Then for any* $\lambda_n \geq 2\delta_n^2$, *the sketched regression estimate* $\hat{f}$ *satisfies the bound*

$$||\hat{f} - f^*||_n^2 \leq c_u\{\lambda_n + \delta_n^2\}$$

with probability greater than $1 - c_1 \exp^{-c_2 n\delta_n^2}$.

Significance of the critical radius δ_n

δ_n is used to specify bounds on the prediction error in KRR estimator. See Theorem 2.

- Example 1 (Polynomial kernel) For some integer $D \geq 1$, the kernel function given by $\mathcal{K}_{poly}(u, v) = (1 + < u, v >)^D$ generates $f(x) = \sum_{j=0}^{D} a_j x^j$. So K always has at most $\min\{D+1, n\}$ non-zero eigenvalues. Consequently,

$$\hat{\mathcal{R}}(\delta) \leq c\sqrt{\frac{D+1}{n}}\delta$$

Then $\delta_n^2 \preceq \sigma^2 \frac{D+1}{n}$, i.e., $||\hat{f} - f^*||_n^2 \preceq \sigma^2 \frac{D+1}{n}$ with high probability.

- Example 2 (First-order Sobolev space) Consider the kernel $\mathcal{K}_{sob}(u, v) = \min\{u, v\}$, it generates the function class

$$\mathcal{H}^1[0,1] = \{f : [0,1] \to \mathbb{R} | f(0) = 0, \text{and } f \text{ is abs. cts. with} \int_0^1 [f'(x)]^2 dx < \infty\}$$

The population level eigenvalues are given by $\mu_j = (\frac{2}{(2j-1)^2\pi})^2$ for $j = 1, 2, \cdots$.
Then it can be calculated that $\delta_n^2 \asymp (\sigma^2/n)^{2/3}$.

- Question: How small can the projection dimension m be chosen while still retaining such minimax optimality?
- A natural conjecture is that the projection dimension m proportional to the statistical dimension d_n.

Let the sketch dimension satisfies a lower bound of the form

$$m \geq \begin{cases} cd_n & \text{for Gaussian sketches,} \\ cd_n log^4(n) & \text{for ROS sketches.} \end{cases} \tag{4}$$

Also define the function

$$\phi(m, d_n, n) := \begin{cases} c_1 \exp^{-c_2 m} & \text{for Gaussian,} \\ c_1[\exp^{-c_2 \frac{m}{d_n log^2(n)}} + \exp^{-c_2 d_n log^2(n)}] & \text{for ROS sketches.} \end{cases}$$

Guarantees for Gaussian and ROS sketches

Corollary

Given n i.i.d. samples $\{(y_i, x_i)\}_{i=1}^n$ from the standard nonparametric regression model, consider the sketched KRR problem (3) based on a sketch dimension m satisfying the lower bound (4). Then there is a universal constant c'_u s.t. for any $\lambda_n \geq 2\delta_n^2$, the sketched regression estimate $\hat{f}$ satisfies the bound

$$||\hat{f} - f^*||_n^2 \leq c'_u\{\lambda_n + \delta_n^2\}$$

with probability greater than $1 - \phi(m, d_n, n) - c_3 \exp^{-c_4 n \delta_n^2}$.

Remark

- For the D^{th} order polynomial kernel from Example 1, d_n is at most $D + 1$, so that a sketch size of order $D + 1$ is sufficient. This is special, since it has no dependence on the sample size.
- For the first-order Sobolev kernel from Example 2, $d_n \asymp n^{1/3}$, so $m \asymp n^{1/3}$ is required.

Simulation

- Consider the Sobolev kernel $\mathcal{K}_{sob}(u, v) = \min\{u, v\}$ in Example 2.
- n i.i.d. samples are generated with $\sigma = 1$, and regression function

$$f^*(x) = |x + 0.5| - 0.5$$

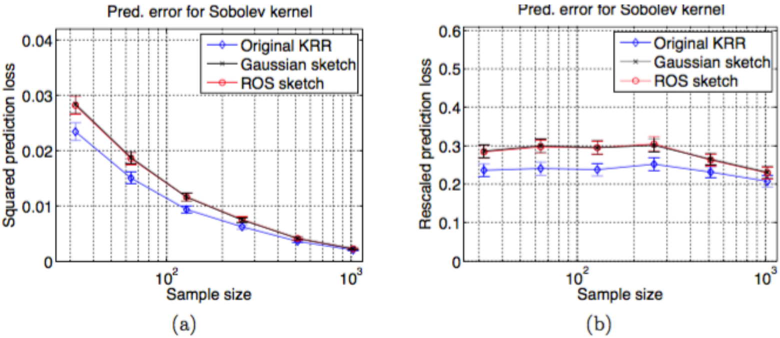


Figure 1. Prediction error versus sample size for original KRR, Gaussian sketch, and ROS sketches for the Sobolev one kernel for the function $f^*(x) = |x + 0.5| - 0.5$. In all cases, each point corresponds to the average of 100 trials, with standard errors also shown. (a) Squared prediction error $\|\widehat{f} - f^*\|_n^2$ versus the sample size $n \in \{32, 64, 128, 256, 1024\}$ for projection dimension $m = \lceil n^{1/3} \rceil$. (b) Rescaled prediction error $n^{2/3}\|\widehat{f} - f^*\|_n^2$ versus the sample size.

Proof intuition of Theorem 2

$$\frac{1}{2}||\hat{f} - f^*||_n^2 \leq \underbrace{||f^\dagger - f^*||_n^2}_{\text{Approximation error}} + \underbrace{||f^\dagger - \hat{f}||_n^2}_{\text{Estimation error}}$$

where $f^\dagger = \frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha^\dagger \mathcal{K}(\cdot, x_i)$ is a zero-noise version of the KRR estimator within the range space of S^T. $\alpha^\dagger$ is achieved by

$$\alpha^\dagger = argmin_{\alpha \in \mathbb{R}^m} \left\{ \frac{1}{2n}||z^* - \sqrt{n}KS^T\alpha||_2^2 + \lambda_n||\sqrt{K}S^T\alpha||_2^2 \right\}$$

with $z^* := (f^*(x_1), \cdots, f^*(x_n))$.

- Lemma 1 (Control of estimation error)
  Under the condition of Theorem 2, we have

  $$||f^\dagger - \hat{f}||_n^2 \leq c\delta_n^2$$

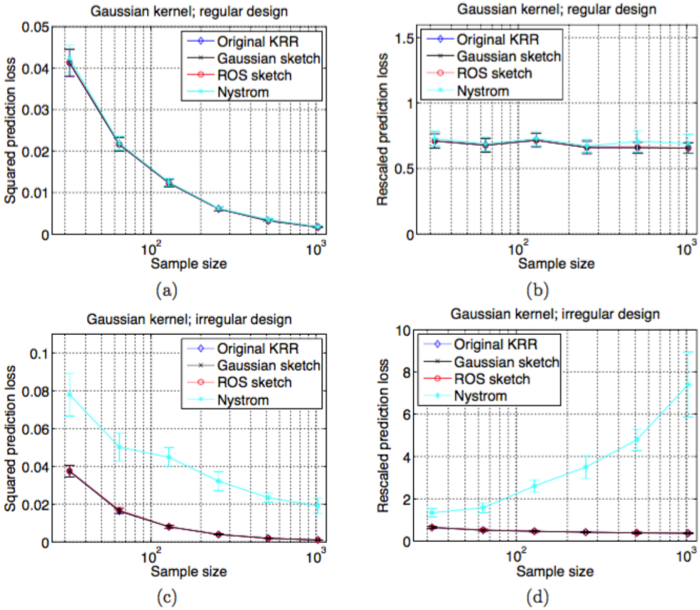
  with probability at least $1 - c_1 \exp -c_2 n\delta_n^2$
- Lemma 2 (Control of approximation error)
  For any K-satisfiable sketch matrix S, we have

  $$||f^\dagger - f^*||_n^2 \leq c\{\lambda_n + \delta_n^2\}\text{and} \;\; ||f^\dagger||_{\mathcal{H}} \leq c\{1 + \frac{\delta_n^2}{\lambda_n}\}$$

Comparison with Nystrom-based approaches

Nystrom approximation: uniformly sampling a subset of $p-$ columns of the kernel matrix K.

- In Bach's paper, $p \succeq n||diag(K(K+\lambda_n I)^{-1})||_\infty logn$
- There are many classes of kernel matrices for which the Nystrom approximation will be poor.


Gaussian kernel; regular design
Squared prediction loss
Sample size
Original KRR
Gaussian sketch
ROS sketch
Nystrom
(a)
Gaussian kernel; regular design
Rescaled prediction loss
Sample size
Original KRR
Gaussian sketch
ROS sketch
Nystrom
(b)
Gaussian kernel; irregular design
Squared prediction loss
Sample size
Original KRR
Gaussian sketch
ROS sketch
Nystrom
(c)
Gaussian kernel; irregular design
Rescaled prediction loss
Sample size
Original KRR
Gaussian sketch
ROS sketch
Nystrom
(d)