Divide and Conquer Kernel Ridge Regression: A Distributed Algorithm with Minimax Optimal Rates

Yuchen Zhang, John C. Duchi, Martin Wainwright (UC Berkeley; http://arxiv.org/pdf/1305.5029; Apr 29, 2014)

Zoltán Szabó

Gatsby Unit, Tea Talk

June 10, 2014

Outline

- Motivation.
- Algorithm.
- Consistency results.

Motivation: non-parametric regression

- Given: $\{(x_i, y_i)\}_{i=1}^N$ training samples ($x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$).
- Assumption: $(x_i, y_i) \overset{i.i.d.}{\sim} \mathbb{P}$.
- Goal: $\hat{f} : \mathcal{X} \to \mathbb{R}$, which predicts "well" on future inputs.
- Objective function: mean square prediction error, i.e.

$$J(f) := \mathbb{E}\left[f(X) - Y\right]^2 \to \min_{f\text{: measurable}} . \tag{1}$$

- Optimal solution (theoretical): regression function

$$f^*(x) = \mathbb{E}[Y|X = x]. \tag{2}$$

Motivation: ridge regressor

- Regularized M-estimators:
  - data-dependent loss + regularization.
  - example: least-squares loss + squared Hilbert norm.
- Our focus:
  - function class = RKHS: $\mathcal{H} = \mathcal{H}(K)$.
  - kernel ridge regression:

$$\hat{f} := \arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} [f(x_i) - y_i]^2 + \lambda \|f\|_{\mathcal{H}}^2 \quad (\lambda > 0). \tag{3}$$

Motivation: analytical solution

- Explicit solution:

$$\hat{f}(\cdot) = \sum_{i=1}^{N} \alpha_i K(\cdot, x_i), \tag{4}$$

where

$$K = [K(x_i, x_j)] \in \mathbb{R}^{N \times N}, \alpha = (K + \lambda N I)^{-1} y \in \mathbb{R}^N. \tag{5}$$

Motivation: analytical solution

- Explicit solution:

$$\hat{f}(\cdot) = \sum_{i=1}^{N} \alpha_i K(\cdot, x_i), \tag{4}$$

where

$$K = [K(x_i, x_j)] \in \mathbb{R}^{N \times N}, \alpha = (K + \lambda N I)^{-1} y \in \mathbb{R}^{N}. \tag{5}$$

- Slight problem:
  - scales terribly,
  - time complexity: $\mathcal{O}\left(N^3\right)$.

Motivation: approximations

- Low-rank methods:
  - Examples: incomplete Cholesky, Nyström approximation.
  - Prediction error guarantees: hardly studied.
- Early stopping methods:
  - Early stopping $\approx$ regularization.
  - Examples: gradient descent, conjugate gradient.
- Time complexity: $\mathcal{O}\left(d^2N\right)$, $\mathcal{O}\left(tN^2\right)$.

Motivation: current approach

- Decomposition-based technique:
  - randomly partition the N samples: m equal sized subsets (S_i).
  - independent ridge regressors: $\hat{f}_i$ ($i = 1, \ldots, m$).
  - average the obtained predictors:

$$\bar{f} = \frac{1}{m}\sum_{i=1}^{m}\hat{f}_i, \quad \hat{f}_i = \arg\min_{f\in\mathcal{H}} \frac{1}{|S_i|}\sum_{(x,y)\in S_i}[f(x)-y]^2 + \lambda\|f\|_{\mathcal{H}}^2.$$

- Time complexity: $\mathcal{O}\left(m\left(\frac{N}{m}\right)^3\right) = O\left(\frac{N^3}{m^2}\right)$.

Algorithm: $\bar{f}$

- Sub-problems: use λ; as if we had N samples.
- Under-regularization: each estimate has
  - small bias, but
  - the variance blows up!
- Average:
  - reduces variance enough,
  - minimax optimality: for certain kernel classes.

Notations

- $(\mathcal{X}, K)$, $(X, Y) \sim \mathbb{P}$, $X \sim \mathbb{P}_X$, $n = \frac{N}{m} = \#$ of blocks.
- $S_K : L^2(\mathbb{P}_X) \to \mathcal{H} = \mathcal{H}(K)$, $id = S_K^* : \mathcal{H} \to L^2(\mathbb{P}_X)$

$$S_K(f)(x) = \int_{\mathcal{X}} K(x, x')\mathrm{d}\mathbb{P}_X(x'), \quad T_K = id \circ S_K. \tag{6}$$

- T_K: compact, positive, self-adjoint operator (if $\mathcal{H}$ is separable, $\|K^{\frac{1}{2}}\|^2_{L^2(\mathbb{P}_X)} := \int_{\mathcal{X}} K(x, x)\mathrm{d}\mathbb{P}_X(x) < \infty$).

Notations

- $(\mathcal{X}, K)$, $(X, Y) \sim \mathbb{P}$, $X \sim \mathbb{P}_X$, $n = \frac{N}{m} = \#$ of blocks.
- $S_K : L^2(\mathbb{P}_X) \to \mathcal{H} = \mathcal{H}(K)$, $id = S_K^* : \mathcal{H} \to L^2(\mathbb{P}_X)$

$$S_K(f)(x) = \int_{\mathcal{X}} K(x, x')\mathrm{d}\mathbb{P}_X(x'), \quad T_K = id \circ S_K. \tag{6}$$

- T_K: compact, positive, self-adjoint operator (if $\mathcal{H}$ is separable, $\|K^{\frac{1}{2}}\|^2_{L^2(\mathbb{P}_X)} := \int_{\mathcal{X}} K(x, x)\mathrm{d}\mathbb{P}_X(x) < \infty$).
- $\xrightarrow{\text{spectral theorem}}$ $\exists$ countable
  - $\{\phi_i\}$ ONS (eigenvectors) $\subseteq L^2(\mathbb{P}_X)$,
  - μ_i eigenvalues (> 0, $\to 0$).
- W.l.o.g.: $\phi_i \in H$.

Mercer theorem: $K \leftarrow \{(\phi_i, \mu_i)\}$

- If $\mathcal{X}$ is compact metric, K is continuous, then

$$K(u,v) = \sum_{j=1}^{\infty} \mu_j \phi_j(u) \phi_j(v). \tag{7}$$

- Note (T_K conditions):
  - $(\mathcal{X}, K)$ conditions $\Rightarrow$ K: bounded.
  - $\mathcal{X}$: compact metric $\Rightarrow$ separable.
  - $\mathcal{X}$: separable, K: continuous$\}$ $\Rightarrow$ $\mathcal{H} = \mathcal{H}(K)$: separable.

Some notations

- $\|h\|_2 := \|h\|_{L_2(\mathbb{P}_X)} = \sqrt{\int_X h^2(x) \mathrm{d}\mathbb{P}(x)}$.
- Our bound on the MSE $\mathbb{E}\left[\left\|\bar{f} - f^*\right\|_2^2\right]$ is formulated in terms of

$$tr(K) = \sum_{j=1}^{\infty} \mu_j, \quad \gamma(\lambda) = \sum_{j=1}^{\infty} \frac{1}{1 + \frac{\lambda}{\mu_j}} \quad, \beta_d = \sum_{j=d+1}^{\infty} \mu_j. \quad (8)$$

- Intuition:
  - $tr(K)$: "size" of the kernel operator (T_K).
  - $\gamma(\lambda)$: "effective dimensionality" of T_K w.r.t. $L^2(\mathbb{P}_X)$.
  - β_d: tail decay of the eigenvalues of T_K ($d \geq 0$ – free parameter). $\beta_0 = tr(K)$.

Assumptions: tail behaviour of ϕ_j-s, bounded variance

- A: $\exists k \geq 2$, $\rho < \infty$ such that $\mathbb{E}\left[\phi_j(X)^{2k}\right] \leq \rho^{2k}$ $(j = 1, 2, \ldots)$.
- A':
  - $\exists \rho < \infty$ such that $\sup_{u \in \mathcal{X}} |\phi_j(u)| \leq \rho$ $(j = 1, 2, \ldots)$.
  - Assumption A' $\Rightarrow$ Assumption A:

$$\mathbb{E}\left[\phi_j(X)^{2k}\right] \leq \mathbb{E}\left[\sup_{u \in \mathcal{X}} |\phi_j(u)|^{2k}\right] \leq \mathbb{E}\left[\rho^{2k}\right] = \rho^{2k}. \qquad (9)$$

- B: $f^* \in \mathcal{H}$. $\exists \sigma > 0$ such that $\forall x \in \mathcal{X}$: $\mathbb{E}[Y - f^*(x)]^2 \leq \sigma^2$.
- Notation+:

$$b(n, d, k) = \max\left[\sqrt{\max(k, \log(d))}, \frac{\max(k, \log(d))}{n^{\frac{1}{2} - \frac{1}{k}}}\right].$$

Main result (C: universal constant)

If $f^* \in \mathcal{H}$, assumptions A and B hold, then

$$\mathbb{E}\left[\left\|\bar{f} - f^*\right\|_2^2\right] \le \left(8 + \frac{12}{m}\right)\lambda \left\|f^*\right\|_{\mathcal{H}}^2 + \frac{12\sigma^2\gamma(\lambda)}{N} + \inf_{d\in\mathbb{N}}\{T_1(d) + T_2(d) + T_3(d)\},$$

$$T_1(d) = \frac{8\rho^4 \left\|f^*\right\|_{\mathcal{H}}^2 tr(K)\beta_d}{\lambda},$$

$$T_2(d) = \frac{4\left\|f^*\right\|_{\mathcal{H}}^2 + 2\sigma^2/\lambda}{m}\left(\mu_{d+1} + \frac{12\rho^4 tr(K)\beta_d}{\lambda}\right),$$

$$T_3(d) = \left[Cb(n,d,k)\frac{\rho^2\gamma(\lambda)}{\sqrt{n}}\right]^k \left\|f^*\right\|_2^2 \left(1 + \frac{2\sigma^2}{m\lambda} + \frac{4\left\|f^*\right\|_{\mathcal{H}}^2}{m}\right).$$

Main result: intuition

- "Simplified" form:

$$\mathbb{E}\left[\left\|\bar{f} - f^*\right\|_2^2\right] = \mathcal{O}\Bigg(\underbrace{\lambda \left\|f^*\right\|_{\mathcal{H}}^2}_{\text{squared bias}} + \underbrace{\frac{\sigma^2\gamma(\lambda)}{N}}_{\text{variance}} \Bigg).$$

- For 3 kernel families, this is "correct" (idea):
  - For large enough d and small enough m: $T_3(d) \leq \frac{\gamma(\lambda)}{N}$.
  - $T_1(d)$, $T_2(d)$: either 0, or smaller then the others.
  - $\lambda = \frac{\gamma(\lambda)}{N}$ fixed point equation $\Rightarrow \lambda^*$. Rate: $\frac{\gamma(\lambda^*)}{N}$.

Consequence-1 (finite rank kernel; example: linear/polynomial)

Assumption: $rank(K) = r$, $\lambda = \frac{r}{N}$, A (or A') and B. If

$$m \leq c \frac{N^{\frac{k-4}{k-2}}}{r^2 \rho^{\frac{4k}{k-2}} \log^{\frac{k}{k-2}}(r)} \quad (A), \qquad m \leq c \frac{N}{r^2 \rho^4 \log(N)} \quad (A'),$$

then

$$\mathbb{E}\left[\left\|\bar{f} - f^*\right\|_2^2\right] = \mathcal{O}\left(\frac{\sigma^2 r}{N}\right). \tag{10}$$

Moreover, (10) is minimax-optimal: $\exists c' > 0$

$$\inf_{f_E} \sup_{f^* \in B_{\mathcal{H}}(1) = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}} \mathbb{E}\left[\left\|f_E - f^*\right\|_2^2\right] \geq c' \frac{r}{N}. \tag{11}$$

Consequence-2 (polynomially decaying eigenvalues; example: Sobolev; C: universal constant)

Assumption: $\mu_j \le Cj^{-2\nu}$ ($j = 1, 2, \ldots$), $\nu > \frac{1}{2}$, $\lambda = \frac{1}{N^{\frac{2\nu}{2\nu+1}}}$, A (or A') and B. If [$c = c(\nu)$]

$$m \le c \left(\frac{N^{\frac{2(k-4)\nu - k}{2\nu+1}}}{\rho^{4k} \log^k(N)} \right)^{\frac{1}{k-2}} \quad (A), \quad m \le c \frac{N^{\frac{2\nu-1}{2\nu+1} \in (0,1)}}{\rho^4 \log(N)} \quad (A'),$$

then

$$\mathbb{E}\left[\left\| \bar{f} - f^* \right\|_2^2 \right] = \mathcal{O}\left(\left(\frac{\sigma^2}{N} \right)^{\frac{2\nu}{2\nu+1} \in \left(\frac{1}{2}, 1\right)} \right). \qquad (12)$$

Moreover, (12) is minimax-optimal.

Consequence-3 (exponentially decaying eigenvalues; example: RBF; $c_i > 0$)

Assumption: $\lambda = \frac{1}{N}$, $\mu_j \le c_1 e^{-c_2 j^2}$, A (or A') and B, $\lambda = \frac{1}{N}$. If

$$m \le c \frac{N^{\frac{k-4}{k-2}}}{\rho^{\frac{4k}{k-2}} \log^{\frac{2k-1}{k-2}}(N)} \quad (A), \qquad m \le c \frac{N}{\rho^4 \log^2(N)} \quad (A'),$$

then

$$\mathbb{E}\left[\left\|\bar{f} - f^*\right\|_2^2\right] = \mathcal{O}\left(\sigma^2 \frac{\sqrt{\log(N)}}{N}\right). \tag{13}$$

Moreover, (13) is minimax-optimal.

Theorem: decomposition trick

$$\mathbb{E}\left\|\bar{f}-f^*\right\|_2^2 = \mathbb{E}\left\|\bar{f}-\mathbb{E}[\bar{f}]+\mathbb{E}[\bar{f}]-f^*\right\|_2^2$$
$$= \mathbb{E}\left[\left\|\bar{f}-\mathbb{E}[\bar{f}]\right\|_2^2\right] + \left\|\mathbb{E}[\bar{f}]-f^*\right\|_2^2 + 2\mathbb{E}\left[\left\langle \bar{f}-\mathbb{E}[\bar{f}], \mathbb{E}[\bar{f}]-f^*\right\rangle_{L^2(\mathbb{P})}\right]$$
$$= \mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m(\hat{f}_i-\mathbb{E}[\hat{f}_i])\right\|_2^2\right] + \left\|\mathbb{E}[\bar{f}]-f^*\right\|_2^2$$
$$\le \frac{1}{m^2} m \sum_{i=1}^m \mathbb{E}\left[\left\|\hat{f}_i-\mathbb{E}[\hat{f}_i]\right\|_2^2\right] + \left\|\mathbb{E}[\hat{f}_1]-f^*\right\|_2^2$$
$$= \frac{1}{m}\mathbb{E}\left[\left\|\hat{f}_1-f^*\right\|_2^2\right] + \left\|\mathbb{E}[\hat{f}_1]-f^*\right\|_2^2 = \frac{variance}{m} + bias$$

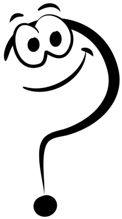
using $f^* \in \mathcal{H}$, $\mathbb{E}[\hat{f}_i] = \arg\min_{f\in\mathcal{H}} \mathbb{E}\left[\left\|\hat{f}_i - f\right\|_2^2\right]$ and (H: Hilbert)

$$\left\|\sum_{i=1}^m h_i\right\|_H^2 \le m\sum_{i=1}^m \|h_i\|_H^2, \mathbb{E}[\bar{f}] = \mathbb{E}[\hat{f}_j], \mathbb{E}\left\langle rnd, const\right\rangle = \left\langle \mathbb{E}[rnd], const\right\rangle.$$

Summary

- Goal: conditional expectation approximation.
- Tool: kernel ridge regression $\leftarrow \mathcal{O}\left(N^3\right)$ time.
- Studied algorithm: simple, parallelizable.
- Result:
  - MSE bound.
  - Explicit rates + minimax optimality for 3 (kernel, $\mathbb{P}$) classes.

Thank you for the attention!



Operator property: definitions

A $T : H \to H$(ilbert) bounded linear operator is

- positive: $\langle Ta, a \rangle_H \geq 0$ ($\forall a \in H$).
- self-adjoint: $T = T^*$.
- compact: $\overline{T(B_E)}$ is compact, $B_H = \{u \in H : \|u\|_H \leq 1\}$.
    - example: finite rank operator.
    - alternative definition: closure of finite rank operators (in operator norm).

Sobolev space

- $\mathcal{X} \subseteq \mathbb{R}^d$: bounded domain. $p \in [1, \infty]$, $|\alpha| = \sum_{i=1}^{d} \alpha_i$.
- Weak derivative of u (extension of the integration by part formula): $D^\alpha u$.
- $W^{m,p}(\mathcal{X}) := \{u \in L^p(\mathcal{X}) : D^\alpha u \in L^p(\mathcal{X}), |\alpha| \leq m\}$.
- Example: $W^{1,\infty}(I)$ = Lipschitz functions on interval I.