Paper Review: Early Stopping and Non-parametric Regression: An Optimal Data-dependent Stopping Rule

Garvesh Raskutti, Martin J. Wainwright and Bin Yu

Group Meeting · Meimei Liu · Jan 22, 2015

Outline

Nonparametric Model Description

- Use a covariate $X \in \mathcal{X}$ to predict a real-valued response $Y \in \mathbb{R}$ by a function $f : \mathcal{X} \to \mathbb{R}$.
- In terms of mean-squared error, the optimal choice is the regression function $f^*(x) = E[Y|x]$, i.e, we observe n samples $\{(x_i, y_i), i = 1, \cdots, n\}$ of the form

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

- Here assume the r.v. w_i are sub-Gaussian with zero-mean and parameter σ, i.e,

$$E[e^{tw_i}] \leq e^{t^2\sigma^2/2} \text{for all } t \in \mathbb{R}$$

- The Goal is to estimate the regression function f^*.

- Problem in Nonparametric setting: Overfitting!
  $\rightarrow$ Solution: Regularization.

- For example, the Kernel Ridge Regression

$$\hat{f}_v := argmin_{f\in\mathcal{H}}\{\frac{1}{2n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \frac{1}{2v}||f||_{\mathcal{H}}^2\}$$

  Stopping rule: e.g, use GCV to choose v.

- An alternative approach is based on early stopping of an iterative algorithm, such as gradient descent applied to the unregularized loss function.

Reproducing Kernel Hilbert Space (RKHS)

- The Hilbert space $\mathcal{H} \subset L^2(\mathbb{P})$ is a RKHS: if there exists a symmetric function $\mathbb{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ s.t:
  (a) for each $x \in \mathcal{X}$, the function $\mathbb{K}(\cdot, x)$ belongs to $\mathcal{H}$.
  (b) reproducing relation $[x]f = f(x) = \langle f, \mathbb{K}(\cdot, x)\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.

- Mercer's Theorem (1909) guarantees that the kernel has an eigen-expansion of the form

$$\mathbb{K}(x, x') = \sum_{k=1}^{\infty} \lambda_k \phi_k(x) \phi_k(x')$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are eigenvalues, and $\{\phi_k\}_{k=0}^{\infty}$ are the associated orthonormal eigenfunctions in $L^2(\mathbb{P})$.

Property in RKHS

- Any function $f \in \mathcal{H}$, $f(x) = \sum_{k=1}^{\infty} \sqrt{\lambda_k} a_k \phi_k(x)$, and the coefficients $a_k = \frac{1}{\sqrt{\lambda_k}} \langle f, \phi_k \rangle_{L^2(\mathbb{P})}$.
- The unit ball for the Hilbert space $\mathcal{H}$ takes the form

$$\mathbb{B}_{\mathcal{H}}(1) = \{f = \sum_{k=1}^{\infty} \sqrt{\lambda_k} b_k \phi_k \text{for some } \sum_{k=1}^{\infty} b_k^2 \leq 1\}$$

- Assume any function f in the unit balls uniformly bounded, i.e $\exists B < \infty$, s.t

$$||f||_{\infty} = \sup_{x \in \mathcal{X}} |f(x)| \leq B \text{for all } f \in \mathbb{B}_{\mathcal{H}}(1)$$

Gradient Update Equation

- Consider minimizing the least squares loss function over some subset of $\mathcal{H}$

$$L(f) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - f(x_i))$$

- It suffices to restrict $f \in$the span of $\{\mathbb{K}(\cdot, x_i), i = 1, \cdots, n\}$
- i.e, we adopt the parameterization $f(\cdot) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i \mathbb{K}(\cdot, x_i)$, for some coefficient vector $w \in \mathbb{R}^n$.
- Using the empirical kernel matrix $K \in \mathbb{R}^{n\times n}$ with entires $[K]_{ij} = \frac{1}{n}\mathbb{K}(x_i, x_j)$, $L(f)$ has the form

$$L(w) = \frac{1}{n}||y_1^n - \sqrt{n}Kw||_2^2.$$

Gradient Update Equation

- We can perform gradient descent in the transformed coordinate system $\theta = \sqrt{K}w$, then

$$L(\theta) = \frac{1}{n}||y_1^n - \sqrt{n}\sqrt{K}\theta||_2^2 = \frac{1}{2n}||y_1^n||_2^2 - \frac{1}{\sqrt{n}}\langle y_1^n, \sqrt{K}\theta\rangle + \frac{1}{2}\theta^T K\theta$$

- Given a sequence of positive step size $\{\alpha_t\}_{t=0}^{\infty}$, the gradient descent algorithm operates via the recursion

$$\theta_{t+1} = \theta_t - \alpha_t \bigtriangledown L(\theta_t) = \theta_t - \alpha_t(K\theta_t - \frac{1}{\sqrt{n}}\sqrt{K}y_1^n)$$

since $\bigtriangledown L(\theta_t) = K\theta - \frac{1}{\sqrt{n}}\sqrt{K}y_1^n$.

Goal in this paper

- At iteration t, we have the estimate θ_t, then compute $w^t = \sqrt{K^{-1}}\theta_t$, then have the estimate $f_t(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^n w_i^t \mathbb{K}(\cdot, x_i)$.
- Goal in this paper:
  (1) measure the error between the sequence $\{f_t\}_{t=0}^{\infty}$ and the true regression function f^* in two ways:
  the $L^2(\mathbb{P}_n)$ norm $||f_t - f^*||_n^2 = \frac{1}{n}\sum_{i=1}^n (f_t(x_i) - f^*(x_i))^2$
  the $L^2(\mathbb{P})$ norm $||f_t - f^*||_2^2 = E[(f_t(X) - f^*(X))^2]$
  (2) Formulate an early stopping strategy to decide precisely how many iteration $\hat{T}$ should be used, in a data-dependent and easily computable manner.

Why do we need early stopping rule?

-To prevent Overfitting. Since too many iterations lead to fitting the noise in the data.

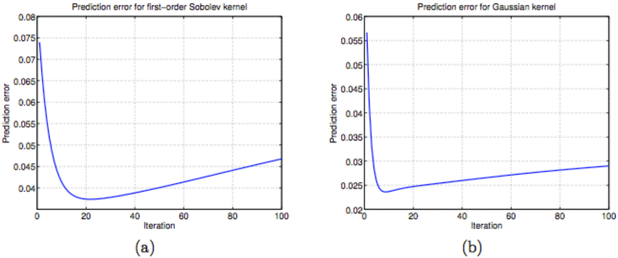

Figure 1: Behavior of gradient descent update (3) with constant step size $\alpha = 0.25$ applied to least-squares loss with $n = 100$ with equi-distant design points $x_i = i/n$ for $i = 1, \ldots, n$, and regression function $f^*(x) = |x-1/2| - 1/2$. Each panel gives plots the $L^2(\mathbb{P}_n)$ error $\|f_t - f^*\|_n^2$ as a function of the iteration number $t = 1, 2, \ldots, 100$. (a) For the first-order Sobolev kernel $\mathbb{K}(x, x') = \min\{x, x'\}$. (b) For the Gaussian kernel $\mathbb{K}(x, x') = \exp(-\frac{1}{2}(x - x')^2)$.

Two important quantities

- The running sum of the step sizes $\eta_t = \sum_{\tau=0}^{t-1} \alpha_\tau$
  The step sizes satisfying the following properties:
  - Boundedness: $0 \leq \alpha_\tau \leq \min\{1, 1/\hat{\lambda}_1\}$ for all $\tau = 0, 1, \cdots$.
  - Non-increasing: $\alpha_{\tau=1} \leq \alpha_\tau$ for all $\tau = 0, 1, \cdots$.
  - Infinite travel: the running sum $\eta_t = \sum_{\tau=0}^{t-1} \alpha_\tau$ diverges as $t \to \infty$.
- A model complexity measure
  $\hat{\mathcal{R}}_K(\varepsilon) = [\frac{1}{n} \sum_{i=1}^{n} \min\{\hat{\lambda}_i, \varepsilon^2\}]^{1/2}$.
  Define the critical empirical radius $\hat{\varepsilon}_n > 0$ as the smallest positive solution to the inequality $\hat{\mathcal{R}}_K(\varepsilon) \leq \varepsilon^2/(2e\sigma)$.

Stopping Rule

$$\hat{T} = argmin\{t \in \mathbb{N} | \hat{\mathcal{R}}_K(1/\sqrt{\eta_t}) > (2e\sigma\eta_t)^{-1}\} - 1$$

- The intuition is that the sum of the step-size η_t acts as a tuning parameter that controls the bias-variance tradeoff.

Theorem 1 for the case of fixed design points $\{x_i\}_{i=1}^n$.

Theorem

Given the stopping time $\hat{T}$ and the critical radius $\hat{\varepsilon}_n$, there are universal positive constants (c_1, c_2) s.t. the following events hold with probability at least $1 - c_1 \exp(-c_2 n \hat{\varepsilon}_n^2)$:
(a) For all iterations $t = 1, 2, \cdots, \hat{T}$: $||f_t - f^*||_n^2 \leq \frac{4}{e\eta_t}$.
(b) At the iteration $\hat{T}$, we have $||f_{\hat{T}} - f^*||_n^2 \leq 12\hat{\varepsilon}_n^2$.
(c) Moreover, for all $t > \hat{T}$, $E[||f_t - f^*||_n^2] \geq \frac{\sigma^2}{4}\eta_t \hat{\mathcal{R}}_K^2(1/\sqrt{\eta_t})$

Remarks for Thm 1

- The bounds (a) and (b) are stated as high probability claims, the expected mean-squared error satisfy

$$E[||f_t - f^*||_n^2] \leq \frac{4}{e\eta_t} \text{for all } \; t \leq \hat{T}$$

- The lower bound (c) shows that for large $t > \hat{T}$, running the iterative algorithm leads to inconsistent estimators for infinite rank kernels.

Thm 2 for the case of random design point $\{x_i\} \sim \mathbb{P}$.

- Define the population version of model complexity measure $\mathcal{R}_{\mathbb{K}}(\varepsilon) = [\frac{1}{n} \sum_{j=1}^{\infty} \min\{\lambda_j, \varepsilon^2\}]^{1/2}$.
  The critical population radius $\varepsilon_n > 0$ as the smallest positive solution to the inequality $40\mathcal{R}_{\mathbb{K}}(\varepsilon) \leq \varepsilon^2/(\sigma)$.

Theorem

With the design variables $\{x_i\}_{i=1}^{n}$ are sampled i.i.d according to $\mathbb{P}$ and the ε_n defined above, there are universal constants $c_j, j = 1, 2, 3$ s.t.

$$||f_{\hat{T}} - f^*||_2^2 \leq c_3 \varepsilon_n^2$$

with probability at least $1 - c_1 \exp(-c_2 n \varepsilon_n^2)$.

Consequences for Kernels with Polynomial Eigendecay

- $\lambda_k \leq C(\frac{1}{k}^{2\beta})$ for some $\beta > 1/2$ and constant C.
  - This type of scaling covers various types of Sobolev spaces, consisting of functions with β derivatives.

Corollary

Suppose that in addition to the assumptions of Thm 2, the kernel class $\hat{H}$ satisfies the polynomial eigenvalue decay for $\beta > 1/2$. Then there is a universal constant c_5 s.t.

$$E[||f_{\hat{T}} - f^*||_2^2] \leq c_5(\frac{\sigma^2}{n})^{\frac{2\beta}{2\beta+1}}$$

  - i.e, the error bound is minimax-optimal.

Consequences for Finite Rank Kernels

- There is some finite integer $m < \infty$ s.t. $\lambda_j = 0$ for all $j \geq m+1$.
- For any integer $d \geq 2$, the kernel $\mathbb{K}(x, x') = (1 + xx')^d$ generates the RKHS of all polynomials with degree at most d. For such kernel, we have

Corollary

If, in addition to the conditions of Thm2, the kernel has finite rank m, then

$$E[||f_{\hat{T}} - f^*||_2^2] \leq c_5 \sigma^2 \frac{m}{n}$$

which achieves the minimax optimal rate in terms of squared $L^2(\mathbb{P})$.

Kernel Ridge Regression (KRR)

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

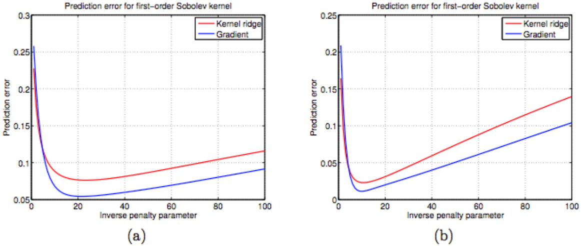

Figure 4: Comparison of the prediction error of the path of kernel ridge regression estimates (17) obtained by varying $\nu \in [1, 100]$ to those of the gradient updates (3) over 100 iterations with constant step size. All simulations were performed with the kernel $\mathbb{K}(x, x') = \min\{|x|, |x'|\}$ based on $n = 100$ samples at the design points $x_i = i/n$ with $f^*(x) = |x - \frac{1}{2}| - \frac{1}{2}$. (a) Noise variance $\sigma^2 = 1$. (b) Noise variance $\sigma^2 = 2$.

Connections of Early Stopping Rule to KRR

- Main idea: when the inverse penalty parameter v is chosen using the same criterion as the stopping rule, i.e $(4\sigma v)^{-1} < \hat{\mathcal{R}}_K(1/\sqrt{v})$, then the prediction error can achieve the same type of bounds.
- Key point: when with penalty term, the continuous parameter v plays the role of discrete parameter $\eta_t = \sum_{\tau=0}^{t-1} \alpha_\tau$.

Corollary

Consider the KRR estimator applied to n i.i.d samples $\{(x_i, y_i)\}$ with σ-sub Gaussian noise. Then there are universal constants (c_1, c_2, c_3) s.t. with probability at least $1 - c_1 \exp(-c_2 n \hat{\varepsilon}_n^2)$:
(a) For all $0 < v < \hat{v}$, $||\hat{f}_v - f^*||_n^2 \leq \frac{2}{v}$
(b) With $\hat{v}$ according to the stopping rule, $||\hat{f}_v - f^*||_n^2 \leq c_3 \hat{\varepsilon}_n^2$.
(c) For all $v > \hat{v}$, $E[||\hat{f}_v - f^*||_n^2] \geq \frac{\sigma^2}{4} v \hat{\mathcal{R}}_K^2(1/\sqrt{v})$.

Compare the corollary with Thm 1, the only difference is that the inverse regularization parameter v replaces the running sum η_t.

Sketch of Proof for Thm 1

- To derive upper bounds on the $L^2(\mathbb{P}_n)$-error in Thm 1, we need to rewrite the gradient update in an alternative form.

$$\theta_{t+1} = \theta_t - \alpha_t(K\theta_t - \frac{1}{\sqrt{n}}\sqrt{K}y_1^n)$$

- Since $f_t(x^n) = \frac{1}{\sqrt{n}}Kw^t = \frac{1}{\sqrt{n}}\sqrt{K}\theta_t$, by multiplying both sides by $\sqrt{K}$, we have $f_{t+1}(x^n) = (I_{n\times n} - \alpha_t K)f_t(x^n) + \alpha_t K y^n$.
- Given the SVD $K = U\Lambda U^T$, and $y^n = f^* + w$, where w is the vector of noise r.v., define the vector $\gamma^t = \frac{1}{\sqrt{n}}U^T f_t(x^n)$, then

$$\gamma^{t+1} = \gamma^t + \alpha_t\Lambda\frac{\tilde{w}}{\sqrt{n}} - \alpha_t\Lambda(\gamma^t - \gamma^*)$$

where $\gamma^* = \frac{1}{\sqrt{n}}U^T f^*(x^n)$, and $\tilde{w} = U^T w$ is a rotated noise vector.

- Since $\gamma^0 = 0$, unwrapping this recursion then yields

$$\gamma^t - \gamma^* = (I - S^t)\frac{\tilde{w}}{\sqrt{n}} - S^t\gamma^*$$

where $S^t = \prod_{\tau=0}^{t-1}(I_{n\times n} - \alpha_\tau \Lambda)$ is called the shrinkage matrix, it indicates the extend of shrinkage towards the origin.

- Properties of Shrinkage Matrices S^t
For all indices $j \in \{1, 2, \cdots, r\}$, S^t satisfy the bounds

$$0 \leq (S^t)^2_{jj} \leq \frac{1}{2e\eta_t\hat{\lambda}_j}$$

$$\frac{1}{2}\min\{1, \eta_t\hat{\lambda}_j\} \leq 1 - S^t_{jj} \leq \min\{1, \eta_t\hat{\lambda}_j\}$$

- $||\gamma^t - \gamma^*||_2^2 \leq \frac{2}{n}||(I - S^t)\tilde{w}||_2^2 + 2||S^t\gamma^*||_2^2$
  $= \frac{2}{n}||(I - S^t)\tilde{w}||_2^2 + 2\sum_{j=1}^{r}[S^t]_{jj}^2(\gamma_{jj}^*)^2 + 2\sum_{j=r+1}^{n}(\gamma_{jj}^*)^2$
- Notice that $||\gamma^t - \gamma^*||_2^2 = \frac{1}{n}||f_t(x^n) - f^*(x^n)||_2^2$, we have the Bias and Variance decomposition as:

$$||f_t - f^*||_n^2 \leq \underbrace{\frac{2}{n}\sum_{j=1}^{r}(S^t)_{jj}^2[U^T f^*(x^n)]_j^2 + \frac{2}{n}\sum_{j=r+1}^{n}[U^T f^*(x^n)]_j^2}_{\text{Squared Bias} B_t^2}$$

$$+ \underbrace{\frac{2}{n}\sum_{j=1}^{r}(1 - S_{jj}^t)^2[U^T w]_j^2}_{\text{Variance} V_t}$$

Bounds on the Bias and Variance

- For all iterations $t = 1, 2, \cdots$, the squared bias is upper bounded as

$$B_t^2 \leq \frac{1}{e\eta_t}$$

Moreover, there is a universal constant $c_1 > 0$ s.t., for any iteration $t = 1, 2, \cdots, \hat{T}$,

$$V_t \leq 5\sigma^2 \eta_t \mathcal{R}_K^2(1/\sqrt{\eta_t})$$

with probability at least $1 - \exp(-c_1 n \hat{\varepsilon}_n^2)$.