Early Stopping for Nonparametric Testing

Meimei Liu[1] Guang Cheng[2]

Department of Statistical Science, Duke University[1], Department of Statistics, Purdue University[2]

Motivation

- Gradient Descent + Early Stopping [1, 2] can avoid over-fitting and achieve optimal **estimation**.
- We propose a nonparametric **testing** method under early stopping.
- Characterize computational limits, i.e., the optimal stopping time to preserve statistical optimality.

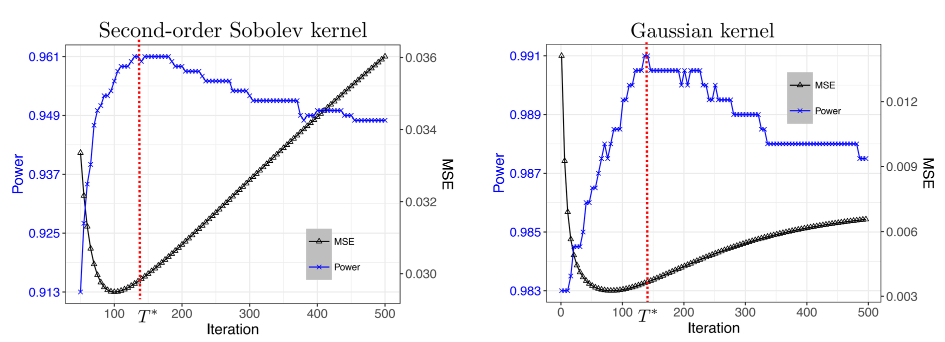


Figure 1:MSE can always be achieved earlier than the maximum power.

Nonparametric Testing under Early Stopping

- Consider the nonparametric model $y = f(x) + \epsilon$, and the hypothesis testing problem

$$H_0 : f = f_0 \text{ vs. } H_1 : f \neq f_0,$$

where f_0 is a known function.

- A distance-based test statistic is

$$D_{n,t} = \|f_t - f_0\|_n^2$$

- The sequence of iterates $\{f_t\}_{t=1}^{\infty}$ is generated as

$$f_{t+1} = f_t - \alpha_t \nabla \mathcal{L}_n(f),$$

where $\nabla \mathcal{L}_n(f) = \frac{1}{n}\Sigma_{i=1}^n (f_t(x_i) - y_i) K(x_i, \cdot)$ is the functional gradient.

Theorem 1: Testing consistency

As long as $n \to \infty$ and $t \to \infty$, we have under H_0,

$$\frac{D_{n,t} - \mu_{n,t}}{\sigma_{n,t}} \xrightarrow{d} N(0,1).$$

Here $\mu_{n,t} = \mathsf{E}_{H_0}[D_{n,t}|X]$ and $\sigma^2_{n,t} = \mathsf{Var}_{H_0}[D_{n,t}|X]$.

- Testing rule is

$$\phi_{n,t} = I(|D_{n,t} - \mu_{n,t}| \geq z_{1-\alpha/2}\sigma_{n,t}),$$

$$\phi_{n,t} = 1 \iff \text{reject } H_0$$

Theorem 2: Power analysis

For any $\varepsilon > 0$, there exist positive constants C_ε, t_ε and N_ε such that

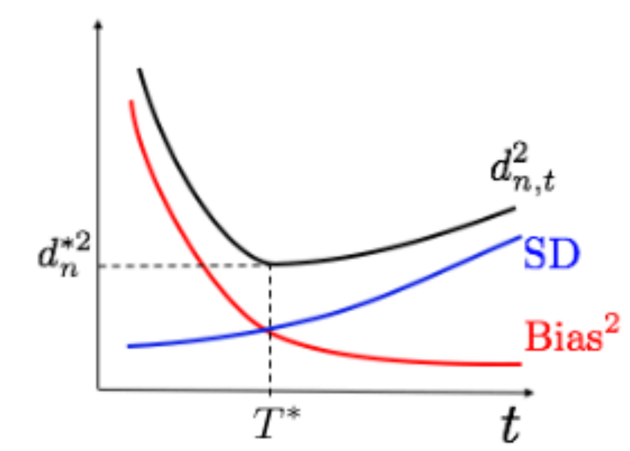
$$\inf_{t \geq t_\varepsilon} \inf_{n \geq N_\varepsilon} \inf_{\substack{f \in \mathcal{B} \\ \|f - f_0\|_n \geq C_\varepsilon d_{n,t}}} P_f(\phi_{n,t} = 1|X) \geq 1 - \varepsilon, \quad \text{(high power)}$$

where $\mathcal{B} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq C\}$ for a constant C and $P_f(\cdot)$ is the probability measure under f.

Early Stopping Rule for Testing

$$d^2_{n,t} = \frac{1}{\eta_t} + \sigma_{n,t}$$

↑ Bias² ↑ SD



$\frac{1}{\eta_t} = \sum_{i=0}^t \alpha_i$ is the total step size, $\sigma_{n,t} \asymp \frac{1}{n}\sqrt{\sum_{i=1}^n \min\{1, \eta_t \widehat{\mu}_i\}}$

- Data-dependent early stopping rule

$$T^* := \arg\min \left\{ t \in \mathbb{N} \mid \frac{1}{\eta_t} < \frac{1}{n}\sqrt{\sum_{i=1}^n \min\{1, \eta_t \widehat{\mu}_i\}} \right\}.$$

This rule involves the empirical eigenvalues of kernel matrix.

Minimax Optimal Testing at T^*

At the iteration T^*, the distance-based test achieves its optimal rate as $d_n^* := d_{n,T^*} \asymp \frac{1}{\sqrt{\eta_{T^*}}}$.

	Polynomial kernel (PDK)	Exponential kernel (EDK)
η_{T^*}	$n^{\frac{4m}{4m+1}}$	$n(\log n)^{-\frac{1}{2p}}$
d_n^*	$n^{-\frac{2m}{4m+1}}$	$n^{-1/2}(\log n)^{\frac{1}{4p}}$

Theorem 3: Sharpness of T^* for PDK and EDK

If $t \ll T^*$ or $t \gg T^*$, then there exists a positive constant C_1 such that, with probability approaching 1,

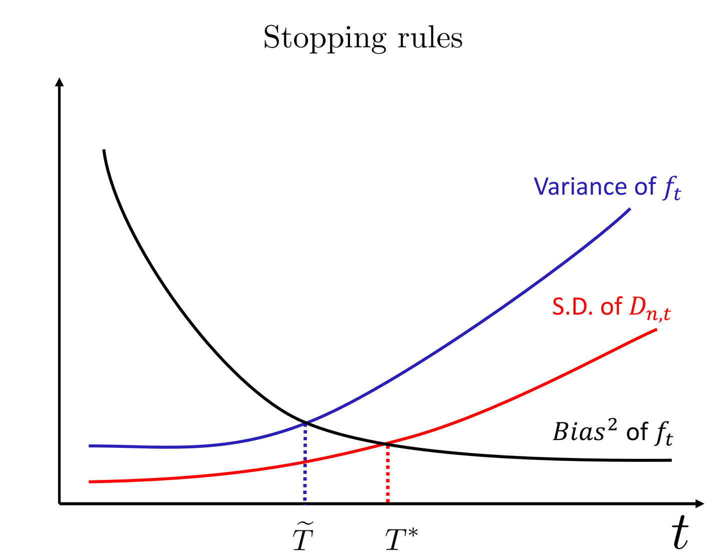
$$\limsup_{n \to \infty} \inf_{\substack{f \in \mathcal{B} \\ \|f - f_0\|_n \geq C_1 d_n^*}} P_f(\phi_{n,t} = 1|X) \leq \alpha. \quad \text{low power}$$

Compare with Early Stopping in Estimation

In literature, [1] and [2] proposed the stopping rule to achieve minimax optimal estimation as

$$\tilde{T} := \arg\min \left\{ t \in \mathbb{N} \mid \frac{1}{\eta_t} < \frac{\sigma}{n} \sum_{i=1}^n \min\{1, \eta_t \widehat{\mu}_i\} \right\}$$

- Estimation: Bias-variance tradeoff
- Testing: Bias-standard deviation tradeoff



Numerical Study

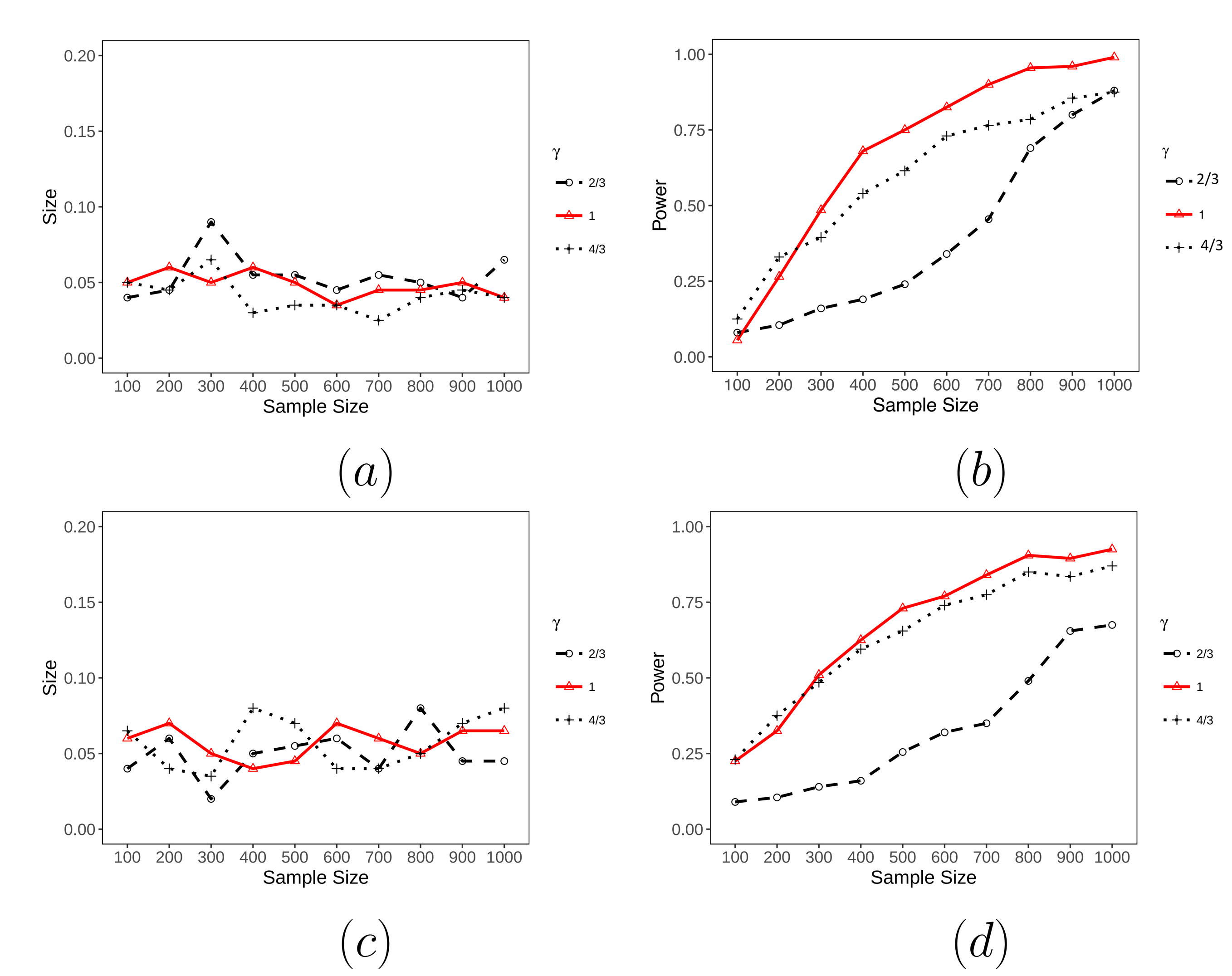


Figure 2:(a), (b) are size and power for PDK; total iteration steps $T = (n^{8/9})^{\gamma}$. (c), (d) are size and power for EDK; total iteration steps $T = (n/(logn)1/4)^{\gamma}$.

References

[1] Raskutti, Garvesh and Wainwright, Martin J and Yu, Bin. *Early stopping and non-parametric regression: an optimal data-dependent stopping rule.* Journal of Machine Learning Research, 15 (1): 335-366, 2014.

[2] Wei, Yuting and Yang, Fanny and Wainwright, Martin J. *Early stopping for kernel boosting algorithms: A general analysis with localized complexities.* NIPS, 6067–6077, 2017.