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.

Nonparametric Testing under Early Stopping

• Consider the nonparametric model \( y = f(x) + \epsilon \), and the hypothesis testing problem
  \( H_0 : f = f_0 \) vs. \( H_1 : f \neq f_0 \),
  where \( f_0 \) is a known function.
• A distance-based test statistic is
  \( D_{n,t} = \| f_t - f_0 \|_2^2 \).
• The sequence of iterates \( \{ f_t \} \) is generated as
  \( f_{t+1} = f_t - \alpha_t \nabla L_n(f_t) \),
  where \( \nabla L_n(f) = \frac{1}{n} \sum_{i \in [n]} (f(x_i) - y_i) K(x_i, \cdot) \) is the functional gradient.

Theorem 1: Testing consistency

As long as \( n \to \infty \) and \( t \to \infty \), we have under \( H_0 \),
\[ D_{n,t} = \frac{\mu_{n,t}}{\sigma_{n,t}} \sim N(0,1). \]

Here \( \mu_{n,t} = E_{f \sim D_{n,t}}[D_{n,t}] \) and \( \sigma_{n,t}^2 = \text{Var}_{f \sim D_{n,t}}[D_{n,t}]. \)

• Testing rule is
  \( \phi_{n,t} = I (|D_{n,t} - \mu_{n,t}| \geq z_{1-\alpha} \sigma_{n,t}) \).
  \( \phi_{n,t} = 1 \iff \text{reject } H_0 \).

Theorem 2: Power analysis

For any \( \epsilon > 0 \), there exist positive constants \( C_\epsilon, t_\epsilon \) and \( N_\epsilon \) such that
\[ \inf_{\xi \in \mathcal{C}_{n \geq N_\epsilon}} \inf_{f \in \mathcal{B}} P_{\phi_{n,t}} (\phi_{n,t} = 1|X) \geq 1 - \epsilon, \quad \text{high power} \]

where \( \mathcal{B} = \{ f \in H : \| f \|_1 \leq C \} \) for a constant \( C \) and \( P_{\phi} (\cdot) \) is the probability measure under \( f \).

Early Stopping Rule for Testing

\[ d^2_{n,t} = \frac{1}{\gamma} + \sigma_{n,t}, \quad \text{Bias}^2 \]

\[ d_{n,T^*} = \frac{1}{\gamma} + \sigma_{n,T^*}, \quad \text{Bias}^2 \]

\( \frac{1}{\gamma} = \sum_{i=1}^n \alpha_i \) is the total step size, \( \sigma_{n,T^*} \sim \frac{1}{\sqrt{n}} \sqrt{\sum_{i=1}^n \min(1, \eta_i \tilde{\nu}_i)} \).

• Data-dependent early stopping rule
  \[ T^* := \arg\min \left\{ t \in \mathbb{N} \mid \frac{1}{\gamma} + \sigma_{n,t} \leq \frac{1}{\sqrt{n}} \sqrt{\sum_{i=1}^n \min(1, \eta_i \tilde{\nu}_i)} \right\}. \]

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
\[ d_{n,T^*} = d_n(T^*) \propto \frac{1}{\sqrt{n}}. \]

<table>
<thead>
<tr>
<th>Polynomial kernel (PDK)</th>
<th>Exponential kernel (EDK)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{1}{n} \log \eta )</td>
<td>( \frac{1}{n} )</td>
</tr>
<tr>
<td>( \frac{1}{n} )</td>
<td>( \frac{1}{n} \log n )</td>
</tr>
</tbody>
</table>

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

If \( t < T^* \) or \( t > T^* \), then there exists a positive constant \( C_1 \) such that, with probability approaching 1,
\[ \lim_{n \to \infty} \sup_{f \in \mathcal{B}} \inf_{\xi \in \mathcal{C}_{n \geq N_1}} P_{\phi_{n,t}} (\phi_{n,t} = 1|X) \leq \alpha. \quad \text{low power} \]

Compare with Early Stopping in Estimation

In literature, [1] and [2] proposed the stopping rule to achieve minimax optimal estimation as
\[ \hat{T} := \arg\min \left\{ t \in \mathbb{N} \mid \frac{1}{\gamma} < \frac{1}{\sqrt{n}} \sqrt{\sum_{i=1}^n \min(1, \eta_i \tilde{\nu}_i)} \right\}. \]

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

Numerical Study

Compare with Early Stopping in Estimation

Figure 2 (a), (b) are size and power for PDK; total iteration steps \( T = (n/3)^{1/4} \). (c), (d) are size and power for EDK; total iteration steps \( T = (n/(\log n))^{1/4} \).

References