How Many Processors Do We Really Need in Parallel Computing?

Guang Cheng[1]

Department of Statistics
Purdue University

BaT Group Meeting
Feb 10, 2016

[1] Acknowledge NSF, ONR and Simons Foundation.

Big Data Era

At the 2010 Google Atmosphere Convention, Googles CEO Eric Schmidt pointed out that,

"There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days."

No wonder that the era of Big Data has arrived. The best examples of Big Data are based on everyday life, medical records of all patients in a large healthcare network; world climate; a wireless sensor network; etc.

Challenges of Big Data

The massive sample size of Big Data introduces unique computational and statistical challenges summarized as *4Ds*:

- Distributed: computation and storage bottleneck;
- Dirty: unstructured data cursed by heterogeneity;
- Dimensionality: scale with sample size;
- Dynamic: varying and unknown underlying distribution.

Challenges of Big Data

The massive sample size of Big Data introduces unique computational and statistical challenges summarized as *4Ds*:

- Distributed: computation and storage bottleneck;
- Dirty: unstructured data cursed by heterogeneity;
- Dimensionality: scale with sample size;
- Dynamic: varying and unknown underlying distribution.

Challenges of Big Data

The massive sample size of Big Data introduces unique computational and statistical challenges summarized as *4Ds*:

- Distributed: computation and storage bottleneck;
- Dirty: unstructured data cursed by heterogeneity;
- Dimensionality: scale with sample size;
- Dynamic: varying and unknown underlying distribution.

Challenges of Big Data

The massive sample size of Big Data introduces unique computational and statistical challenges summarized as *4Ds*:

- Distributed: computation and storage bottleneck;
- Dirty: unstructured data cursed by heterogeneity;
- Dimensionality: scale with sample size;
- Dynamic: varying and unknown underlying distribution.

Parallel Computing on Big Data

- In the parallel computing environment, a common practice is to distribute a massive dataset to multiple processors, and then aggregate local results obtained from separate machines into global counterparts;
- The above Divide-and-Conquer (D&C) strategy often requires a *growing* number of machines to deal with an increasingly large dataset;
- This computational consideration leads to the emergence of the so-called "*Splitotics Theory*," a statistical foundation of "*Big Data Theory.*"

Parallel Computing on Big Data

- In the parallel computing environment, a common practice is to distribute a massive dataset to multiple processors, and then aggregate local results obtained from separate machines into global counterparts;
- The above Divide-and-Conquer (D&C) strategy often requires a *growing* number of machines to deal with an increasingly large dataset;
- This computational consideration leads to the emergence of the so-called "*Splitotics Theory*," a statistical foundation of "*Big Data Theory.*"

Parallel Computing on Big Data

- In the parallel computing environment, a common practice is to distribute a massive dataset to multiple processors, and then aggregate local results obtained from separate machines into global counterparts;
- The above Divide-and-Conquer (D&C) strategy often requires a *growing* number of machines to deal with an increasingly large dataset;
- This computational consideration leads to the emergence of the so-called "*Splitotics Theory*," a statistical foundation of "*Big Data Theory.*"

A Basic Question from Statisticians

What is the computational limit for parallel processing from a purely statistical theory perspective?
(or shall we allocate *all* our computational resources to analyze massive data?)

- In this talk, we address this basic, yet fundamentally important, question by carefully analyzing statistical versus computational trade-off in the D&C framework;
- In particular, an intriguing phase transition phenomenon is discovered for the number of deployed machines that ends up being a simple proxy for computing cost, for both statistical estimation and testing.

A Basic Question from Statisticians

What is the computational limit for parallel processing from a purely statistical theory perspective?
(or shall we allocate *all* our computational resources to analyze massive data?)

- In this talk, we address this basic, yet fundamentally important, question by carefully analyzing statistical versus computational trade-off in the D&C framework;
- In particular, an intriguing phase transition phenomenon is discovered for the number of deployed machines that ends up being a simple proxy for computing cost, for both statistical estimation and testing.

A Basic Question from Statisticians

What is the computational limit for parallel processing from a purely statistical theory perspective?
(or shall we allocate *all* our computational resources to analyze massive data?)

- In this talk, we address this basic, yet fundamentally important, question by carefully analyzing statistical versus computational trade-off in the D&C framework;
- In particular, an intriguing phase transition phenomenon is discovered for the number of deployed machines that ends up being a simple proxy for computing cost, for both statistical estimation and testing.

Outline

A Flowchart of D&C

Consider a univariate nonparametric regression model:

$$Y = f_0(Z) + \epsilon.$$

$$
\begin{array}{cccccc}
 & & \text{Subset } 1 \text{ (n)} & \xrightarrow{\text{Machine } 1} & \widehat{f}_1 \\
\text{Big Data } (N) & \xrightarrow{\text{Divide}}{}^{2} & \text{Subset } 2 \text{ (n)} & \xrightarrow{\text{Machine } 2} & \widehat{f}_2 \\
 & & \cdots & \cdots & \cdots \\
 & & \text{Subset } s \text{ (n)} & \xrightarrow{\text{Machine } s} & \widehat{f}_s \\
\text{Super} \Bigg\Downarrow \text{machine} & & & & \text{Con} \Bigg\Downarrow \text{quer} \\
\text{Oracle Est. } \widehat{f}_N & & & & \text{Agg. Est. } \bar{f}_N
\end{array}
$$

$$\bar{f}_N = (1/s) \sum_{j=1}^{s} \widehat{f}_n^{(j)}.$$

[2]For simplicity, we assume equal-sized splitting, i.e., $N = s \times n$.

Statistical-and-Computational Tradeoff

- We use the number of deployed machines s as a simple proxy for computing time.
- How fast can we allow s to diverge (w.r.t. N), say $s = N^a$, such that the aggregated estimate $\bar{f}_N$ is minimax optimal or nonparametric testing based on $\bar{f}_N$ is minimax optimal?
- We prove that there indeed exists a *sharp* upper bound for s, below which statistical optimality is achievable and above which statistical optimality is impossible.
- The sharpness is important in that it captures the *intrinsic* computational limit of D&C algorithm (existing literature is only concerned with (non-sharp) upper bound).

Statistical-and-Computational Tradeoff

- We use the number of deployed machines s as a simple proxy for computing time.
- How fast can we allow s to diverge (w.r.t. N), say $s = N^{a}$, such that the aggregated estimate $\bar{f}_N$ is minimax optimal or nonparametric testing based on $\bar{f}_N$ is minimax optimal?
- We prove that there indeed exists a *sharp* upper bound for s, below which statistical optimality is achievable and above which statistical optimality is impossible.
- The sharpness is important in that it captures the *intrinsic* computational limit of D&C algorithm (existing literature is only concerned with (non-sharp) upper bound).

Statistical-and-Computational Tradeoff

- We use the number of deployed machines s as a simple proxy for computing time.
- How fast can we allow s to diverge (w.r.t. N), say $s = N^a$, such that the aggregated estimate $\bar{f}_N$ is minimax optimal or nonparametric testing based on $\bar{f}_N$ is minimax optimal?
- We prove that there indeed exists a *sharp* upper bound for s, below which statistical optimality is achievable and above which statistical optimality is impossible.
- The sharpness is important in that it captures the *intrinsic* computational limit of D&C algorithm (existing literature is only concerned with (non-sharp) upper bound).

Statistical-and-Computational Tradeoff

- We use the number of deployed machines s as a simple proxy for computing time.
- How fast can we allow s to diverge (w.r.t. N), say $s = N^{a}$, such that the aggregated estimate $\bar{f}_N$ is minimax optimal or nonparametric testing based on $\bar{f}_N$ is minimax optimal?
- We prove that there indeed exists a *sharp* upper bound for s, below which statistical optimality is achievable and above which statistical optimality is impossible.
- The sharpness is important in that it captures the *intrinsic* computational limit of D&C algorithm (existing literature is only concerned with (non-sharp) upper bound).

A Plot for Computing Time

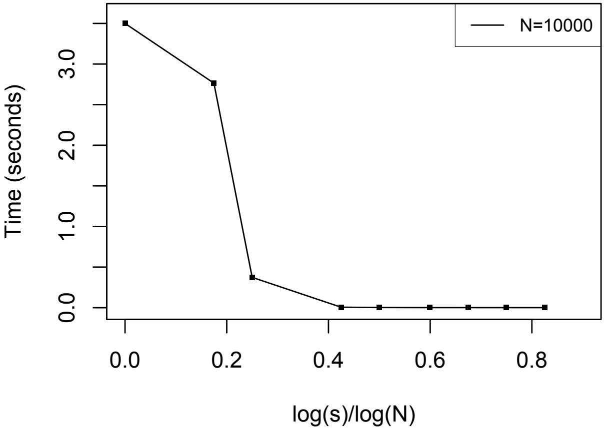


Figure: *Computing time of $\bar{f}_N$ based on a single replication under different choices of s when $N = 10,000$. The larger the s, the smaller the computing time.*

A Plot for Mean Squared Error

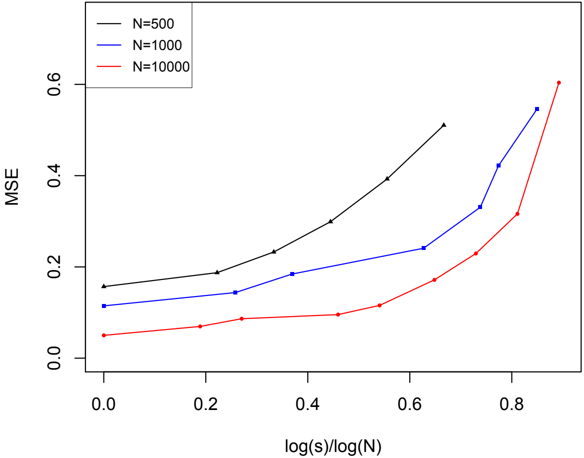
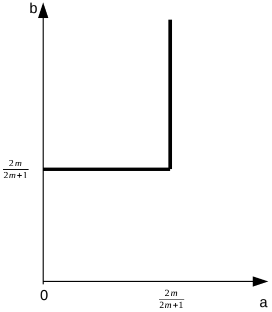


Figure: *MSE of* $\bar{f}_N$ *based on 500 independent replications with* $f_0(z) = 0.6b_{30,17}(z) + 0.4b_{3,11}$. *MSE stays at low levels as* $\log s/\log N \leq 0.2$.

Phase Transition Diagram

Results are based on smoothing spline regression with a smoothing parameter λ and smoothness $m \geq 1$.



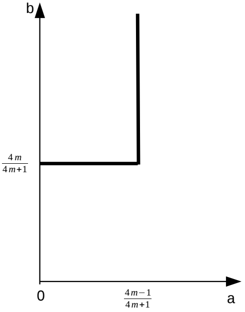


Figure: *Two lines indicate the choices of $s \asymp N^a$ and $\lambda \asymp N^{-b}$, leading to minimax optimal estimation (left) and minimax optimal testing (right). Whereas (a, b)'s outside these two lines lead to suboptimal rates.*

Smoothing Spline Model with *Fixed* Design

- Observe samples from the following model

$$y_l = f(l/N) + \epsilon_l,\ l = 0, 1, \ldots, N-1,$$

  where ϵ_l's are *iid* zero-mean r.v.s with unit variances;
- The N samples $\{y_l, l/N\}_{l=0}^{N-1}$ are distributed to s machines with each machine being assigned n samples;
- We want the N covariates $t_l = l/N$ to appear in s machines as "evenly" as possible over the entire interval $[0, 1]$.

Smoothing Spline Model with *Fixed* Design

- Observe samples from the following model

$$y_l = f(l/N) + \epsilon_l,\ l = 0, 1, \ldots, N-1,$$

  where ϵ_l's are *iid* zero-mean r.v.s with unit variances;
- The N samples $\{y_l, l/N\}_{l=0}^{N-1}$ are distributed to s machines with each machine being assigned n samples;
- We want the N covariates $t_l = l/N$ to appear in s machines as "evenly" as possible over the entire interval $[0, 1]$.

Smoothing Spline Model with *Fixed* Design

- Observe samples from the following model

$$y_l = f(l/N) + \epsilon_l,\ l = 0, 1, \ldots, N-1,$$

  where ϵ_l's are *iid* zero-mean r.v.s with unit variances;
- The N samples $\{y_l, l/N\}_{l=0}^{N-1}$ are distributed to s machines with each machine being assigned n samples;
- We want the N covariates $t_l = l/N$ to appear in s machines as "evenly" as possible over the entire interval $[0, 1]$.

- Let $t_{1,j}, \ldots, t_{n,j}$ be evenly spaced points across $[0, 1]$ (with a gap $1/n$), i.e.,

$$t_{i,j} = \frac{is - s + j - 1}{N};$$

- For $1 \leq j \leq s$, the sub-model at the jth machine is

$$Y_{i,j} = f(t_{i,j}) + \epsilon_{i,j}, \; i = 1, \ldots, n,$$

where $\epsilon_{i,j} = \epsilon_{is-s+j-1}$ and $Y_{i,j} = y_{is-s+j-1}$.

- Let $t_{1,j}, \ldots, t_{n,j}$ be evenly spaced points across $[0,1]$ (with a gap $1/n$), i.e.,

$$t_{i,j} = \frac{is - s + j - 1}{N};$$

- For $1 \leq j \leq s$, the sub-model at the jth machine is

$$Y_{i,j} = f(t_{i,j}) + \epsilon_{i,j},\ i = 1, \ldots, n,$$

where $\epsilon_{i,j} = \epsilon_{is-s+j-1}$ and $Y_{i,j} = y_{is-s+j-1}$.

Smoothing Spline Estimate

At each machine, we obtain a smoothing spline sub-estimate as

$$\widehat{f}_j = \arg\min_{f \in S^m(\mathbb{I})} \frac{1}{n} \sum_{i=1}^{n} (Y_{i,j} - f(t_{i,j}))^2 + \lambda \|f\|_{\mathcal{H}}^2, \tag{1}$$

where $\langle f, g \rangle_{\mathcal{H}} = \int_0^1 f^{(m)}(t) g^{(m)}(t) dt$ is a roughness penalty and $\lambda > 0$ is a smoothing parameter;

Trigonometric Eigen-System

We consider an m-order ($m > 1/2$) periodic Sobolev space

$$S^m(\mathbb{I}) = \{\sum_{\nu=1}^{\infty} f_\nu \phi_\nu(\cdot) : \sum_{\nu=1}^{\infty} f_\nu^2 \gamma_\nu < \infty\},$$

where for $k = 1, 2, \ldots$,

$$\phi_{2k-1}(t) = \sqrt{2}\cos(2\pi k t), \quad \phi_{2k}(t) = \sqrt{2}\sin(2\pi k t),$$

$$\gamma_{2k-1} = \gamma_{2k} = (2\pi k)^{2m}.$$

Main Theorem I: Upper Bound of MSE

Theorem

Suppose $h > 0$ and N is divisible by n. Then there exist absolute positive constants $b_m, c_m \geq 1$ (depending on m only) such that for any fixed $1 \leq s \leq N$,

$$E\{\|\bar{f}_N - E\{\bar{f}_N\}\|_2^2\} \leq b_m \left(N^{-1} + (N\lambda^{1/(2m)})^{-1} A_n(m,\lambda) \right),$$

$$\|E\{\bar{f}_N\} - f_0\|_2 \leq c_m \sqrt{\|f_0\|_{\mathcal{H}}(\lambda + n^{-2m} + N^{-1})},$$

where $A_n(m,\lambda) = \int_0^{\pi n \lambda^{1/(2m)}} (1 + x^{2m})^{-1} dx$.

An Upper Bound s^*

- From the above theorem and the well known fact that

$$\mathrm{MSE}_{f_0}(\bar{f}_N) = E\{\|\bar{f}_N - E\{\bar{f}_N\}\|_2^2\} + \|E\{\bar{f}_N\} - f_0\|_2^2,$$

  we conclude that $\bar{f}_N$ is minimax optimal[3] if one of the following two conditions hold:
  - $s = O(N^{2m/(2m+1)})$ and $\lambda \asymp N^{-2m/(2m+1)}$;
  - $s \asymp N^{2m/(2m+1)}$ and $\lambda = o(N^{-2m/(2m+1)})$;
- Denote $s^* = N^{2m/(2m+1)}$ and $\lambda^* = N^{-2m/(2m+1)}$;
- Improve the upper bound $s = O(N^{(2m-1)/(2m+1)}/\log N)$ derived in Zhang, Duchi and Wainwright (COLT'13) under $\lambda = \lambda^*$. For example, when $m = 2$ (cubic spline),

$$N^{0.6}/\log N \Longrightarrow N^{0.8}$$

[3] in the sense that $\|\bar{f}_N - f_0\|_2 = O_P(N^{-m/(2m+1)})$

An Upper Bound s^*

- From the above theorem and the well known fact that

$$\mathrm{MSE}_{f_0}(\bar{f}_N) = E\{\|\bar{f}_N - E\{\bar{f}_N\}\|_2^2\} + \|E\{\bar{f}_N\} - f_0\|_2^2,$$

  we conclude that $\bar{f}_N$ is minimax optimal[3] if one of the following two conditions hold:
  - $s = O(N^{2m/(2m+1)})$ and $\lambda \asymp N^{-2m/(2m+1)}$;
  - $s \asymp N^{2m/(2m+1)}$ and $\lambda = o(N^{-2m/(2m+1)})$;
- Denote $s^* = N^{2m/(2m+1)}$ and $\lambda^* = N^{-2m/(2m+1)}$;
- Improve the upper bound $s = O(N^{(2m-1)/(2m+1)}/\log N)$ derived in Zhang, Duchi and Wainwright (COLT'13) under $\lambda = \lambda^*$. For example, when $m = 2$ (cubic spline),

$$N^{0.6}/\log N \Longrightarrow N^{0.8}$$

[3] in the sense that $\|\bar{f}_N - f_0\|_2 = O_P(N^{-m/(2m+1)})$

An Upper Bound s^*

- From the above theorem and the well known fact that

$$\mathrm{MSE}_{f_0}(\bar{f}_N) = E\{\|\bar{f}_N - E\{\bar{f}_N\}\|_2^2\} + \|E\{\bar{f}_N\} - f_0\|_2^2,$$

  we conclude that $\bar{f}_N$ is minimax optimal[3] if one of the following two conditions hold:
  - $s = O(N^{2m/(2m+1)})$ and $\lambda \asymp N^{-2m/(2m+1)}$;
  - $s \asymp N^{2m/(2m+1)}$ and $\lambda = o(N^{-2m/(2m+1)})$;
- Denote $s^* = N^{2m/(2m+1)}$ and $\lambda^* = N^{-2m/(2m+1)}$;
- Improve the upper bound $s = O(N^{(2m-1)/(2m+1)}/\log N)$ derived in Zhang, Duchi and Wainwright (COLT'13) under $\lambda = \lambda^*$. For example, when $m = 2$ (cubic spline),

$$N^{0.6}/\log N \Longrightarrow N^{0.8}$$

[3] in the sense that $\|\bar{f}_N - f_0\|_2 = O_P(N^{-m/(2m+1)})$

An Upper Bound s^*

- From the above theorem and the well known fact that

$$\mathrm{MSE}_{f_0}(\bar{f}_N) = E\{\|\bar{f}_N - E\{\bar{f}_N\}\|_2^2\} + \|E\{\bar{f}_N\} - f_0\|_2^2,$$

  we conclude that $\bar{f}_N$ is minimax optimal[3] if one of the following two conditions hold:
  - $s = O(N^{2m/(2m+1)})$ and $\lambda \asymp N^{-2m/(2m+1)}$;
  - $s \asymp N^{2m/(2m+1)}$ and $\lambda = o(N^{-2m/(2m+1)})$;
- Denote $s^* = N^{2m/(2m+1)}$ and $\lambda^* = N^{-2m/(2m+1)}$;
- Improve the upper bound $s = O(N^{(2m-1)/(2m+1)}/\log N)$ derived in Zhang, Duchi and Wainwright (COLT'13) under $\lambda = \lambda^*$. For example, when $m = 2$ (cubic spline),

$$N^{0.6}/\log N \Longrightarrow N^{0.8}$$

[3] in the sense that $\|\bar{f}_N - f_0\|_2 = O_P(N^{-m/(2m+1)})$

An Upper Bound s^*

- From the above theorem and the well known fact that

$$\mathrm{MSE}_{f_0}(\bar{f}_N) = E\{\|\bar{f}_N - E\{\bar{f}_N\}\|_2^2\} + \|E\{\bar{f}_N\} - f_0\|_2^2,$$

  we conclude that $\bar{f}_N$ is minimax optimal[3] if one of the following two conditions hold:
  - $s = O(N^{2m/(2m+1)})$ and $\lambda \asymp N^{-2m/(2m+1)}$;
  - $s \asymp N^{2m/(2m+1)}$ and $\lambda = o(N^{-2m/(2m+1)})$;
- Denote $s^* = N^{2m/(2m+1)}$ and $\lambda^* = N^{-2m/(2m+1)}$;
- Improve the upper bound $s = O(N^{(2m-1)/(2m+1)}/\log N)$ derived in Zhang, Duchi and Wainwright (COLT'13) under $\lambda = \lambda^*$. For example, when $m = 2$ (cubic spline),

$$N^{0.6}/\log N \Longrightarrow N^{0.8}$$

.

[3] in the sense that $\|\bar{f}_N - f_0\|_2 = O_P(N^{-m/(2m+1)})$

Is there any room to improve s^*?

Main Theorem II: Lower Bound of MSE

Theorem

Suppose $h > 0$ and N is divisible by n. Then for any constant $C > 0$, it holds that for any fixed $1 < s < N$,

$$\sup_{\substack{f_0 \in S^m(\mathbb{I}) \\ \|f_0\|_{\mathcal{H}} \leq C}} \|E\{\bar{f}_N\} - f_0\|_2^2 \geq C(a_m n^{-2m} - 8N^{-1}),$$

where $a_m \in (0,1)$ is an absolute constant depending on m only.

Remark: This is a "worst scenario" result. It implies that once s is beyond some threshold, the minimax rate optimality will break down for some true f_0.

Un-Improvable s^*

- Our second theorem implies that

$$\sup_{\substack{f_0 \in S^m(\mathbb{I}) \\ \|f_0\|_{\mathcal{H}} \le C}} \mathrm{MSE}_{f_0}(\bar{f}) \ge \sup_{\substack{f_0 \in S^m(\mathbb{I}) \\ \|f_0\|_{\mathcal{H}} \le C}} \|E\{\bar{f}\} - f_0\|_2^2 \ge C(a_m n^{-2m} - 8N^{-1});$$

- It is easy to see that the above lower bound is strictly slower than the optimal rate $N^{-2m/(2m+1)}$ when $s \gg s^* = N^{2m/(2m+1)}$ (no matter how λ is chosen). Therefore, s^* cannot be further improved;
- From the above *sharpness* result, we claim that D&C approach prefers more smooth function (larger m) since we can save more computational efforts (larger s).

Un-Improvable s^*

- Our second theorem implies that

$$\sup_{\substack{f_0 \in S^m(\mathbb{I}) \\ \|f_0\|_{\mathcal{H}} \le C}} \mathrm{MSE}_{f_0}(\bar{f}) \ge \sup_{\substack{f_0 \in S^m(\mathbb{I}) \\ \|f_0\|_{\mathcal{H}} \le C}} \|E\{\bar{f}\} - f_0\|_2^2 \ge C(a_m n^{-2m} - 8N^{-1});$$

- It is easy to see that the above lower bound is strictly slower than the optimal rate $N^{-2m/(2m+1)}$ when $s \gg s^* = N^{2m/(2m+1)}$ (no matter how λ is chosen). Therefore, s^* cannot be further improved;
- From the above *sharpness* result, we claim that D&C approach prefers more smooth function (larger m) since we can save more computational efforts (larger s).

Un-Improvable s^*

- Our second theorem implies that

$$\sup_{\substack{f_0 \in S^m(\mathbb{I}) \\ \|f_0\|_{\mathcal{H}} \le C}} \mathrm{MSE}_{f_0}(\bar{f}) \ge \sup_{\substack{f_0 \in S^m(\mathbb{I}) \\ \|f_0\|_{\mathcal{H}} \le C}} \|E\{\bar{f}\} - f_0\|_2^2 \ge C(a_m n^{-2m} - 8N^{-1});$$

- It is easy to see that the above lower bound is strictly slower than the optimal rate $N^{-2m/(2m+1)}$ when $s \gg s^* = N^{2m/(2m+1)}$ (no matter how λ is chosen). Therefore, s^* cannot be further improved;
- From the above *sharpness* result, we claim that D&C approach prefers more smooth function (larger m) since we can save more computational efforts (larger s).

A Graphical Illustration

Set $\lambda \asymp N^{-2m/(2m+1)}$ and $s = N^a$ for $0 \leq a \leq 1$.

$$\text{Upper bound of squared bias: } N^{-\rho_1(a)}$$
$$\text{Lower bound of squared bias: } N^{-\rho_2(a)}$$
$$\text{Upper bound of variance: } N^{-\rho_3(a)}.$$

A Graphical Illustration

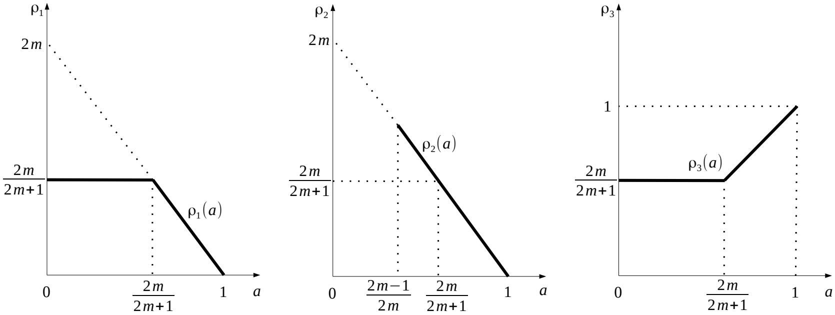


Figure: *Plots of $\rho_1(a), \rho_2(a), \rho_3(a)$ versus a, indicated by thick solid lines. $\rho_1(a)$, $\rho_2(a)$ and $\rho_3(a)$ indicate upper bound of squared bias, lower bound of squared bias and upper bound of variance, respectively. Note that $\rho_2(a)$ is plotted only for $(2m-1)/(2m) < a \leq 1$; when $0 \leq a \leq (2m-1)/(2m)$, $\rho_2(a) = \infty$, which is omitted.*

What is the corresponding s^* for testing?[4]

[4] The minimax optimality of nonparametric testing is in the sense of Ingster (1993).

Wald-Type Test

- Consider the same smoothing spline model:

$$Y_{i,j} = f(t_{i,j}) + \epsilon_{i,j}, \ i = 1, \ldots, n,$$

where $\epsilon_{i,j}$ are *iid* $N(0,1)$[5] and $f \in S^m(\mathbb{I})$.

- Test the following simple hypothesis:

$$H_0 : f = 0 \quad \text{v.s.} \quad H_1 : f \in S^m(\mathbb{I})\backslash\{0\};$$

- Define a Wald-type test statistic[6]:

$$T_{N,\lambda} = \|\bar{f}_N\|_2^2.$$

[5] Gaussian error assumption is not essential.

[6] Aggregated likelihood ratio test also works with more technicalities.

Wald-Type Test

- Consider the same smoothing spline model:

$$Y_{i,j} = f(t_{i,j}) + \epsilon_{i,j}, \ i = 1, \ldots, n,$$

  where $\epsilon_{i,j}$ are *iid* $N(0,1)$[5] and $f \in S^m(\mathbb{I})$.
- Test the following simple hypothesis:

$$H_0 : f = 0 \ \text{ v.s. } \ H_1 : f \in S^m(\mathbb{I}) \backslash \{0\};$$

- Define a Wald-type test statistic[6]:

$$T_{N,\lambda} = \|\bar{f}_N\|_2^2.$$

[5] Gaussian error assumption is not essential.

[6] Aggregated likelihood ratio test also works with more technicalities.

Wald-Type Test

- Consider the same smoothing spline model:

$$Y_{i,j} = f(t_{i,j}) + \epsilon_{i,j}, \ i = 1, \ldots, n,$$

  where $\epsilon_{i,j}$ are *iid* $N(0,1)$[5] and $f \in S^m(\mathbb{I})$.
- Test the following simple hypothesis:

$$H_0 : f = 0 \quad \text{v.s.} \quad H_1 : f \in S^m(\mathbb{I}) \backslash \{0\};$$

- Define a Wald-type test statistic[6]:

$$T_{N,\lambda} = \|\bar{f}_N\|_2^2.$$

---

[5] Gaussian error assumption is not essential.

[6] Aggregated likelihood ratio test also works with more technicalities.

Main Theorem III: Testing Consistency

Theorem

Suppose that $\lambda \to 0$, $n \to \infty$ when $N \to \infty$, and $\lim_{N\to\infty} n\lambda^{1/2m}$ exists (which could be infinity). Then, we have under H_0,

$$\frac{T_{N,\lambda} - \mu_{N,\lambda}}{\sigma_{N,\lambda}} \xrightarrow{d} N(0,1), \quad \textit{as } N \to \infty,$$

where $\mu_{N,\lambda} := E_{H_0}\{T_{N,\lambda}\}$ and $\sigma^2_{N,\lambda} := Var_{H_0}\{T_{N,\lambda}\}$.

Remark: Our testing rule is thus

$$\phi_{N,\lambda} = I(|T_{N,\lambda} - \mu_{N,\lambda}| \geq z_{1-\alpha/2}\sigma_{N,\lambda}).$$

Comments

- It is a bit surprising that testing consistency essentially requires no condition on s as long as $N \to \infty$. In other words, s can be either fixed or diverge at any rate;
- However, the (non-asymptotic) power of our proposed test depends on s in a very subtle manner.

Comments

- It is a bit surprising that testing consistency essentially requires no condition on s as long as $N \to \infty$. In other words, s can be either fixed or diverge at any rate;
- However, the (non-asymptotic) power of our proposed test depends on s in a very subtle manner.

Minimal Separation Rate

- Separation rate of a testing method is defined as a rate measuring the deviation of local alternative sequences $H_{1n} : f = f_n$ to null hypothesis $H_0 : f = 0$ such that a correct rejection of H_{1n} can be triggered;
- Minimal separation rate refects an intrinsic power of a testing method (Ingster, 1994);
- We next examine the impact of s on the separation rate of our proposed test. We are particularly interested in the choice of s leading to the minimal separation rate.

Minimal Separation Rate

- Separation rate of a testing method is defined as a rate measuring the deviation of local alternative sequences $H_{1n}: f = f_n$ to null hypothesis $H_0: f = 0$ such that a correct rejection of H_{1n} can be triggered;
- Minimal separation rate refects an intrinsic power of a testing method (Ingster, 1994);
- We next examine the impact of s on the separation rate of our proposed test. We are particularly interested in the choice of s leading to the minimal separation rate.

Minimal Separation Rate

- Separation rate of a testing method is defined as a rate measuring the deviation of local alternative sequences $H_{1n} : f = f_n$ to null hypothesis $H_0 : f = 0$ such that a correct rejection of H_{1n} can be triggered;
- Minimal separation rate refects an intrinsic power of a testing method (Ingster, 1994);
- We next examine the impact of s on the separation rate of our proposed test. We are particularly interested in the choice of s leading to the minimal separation rate.

Main Theorem IV: Type II Error

The following theorem says that our test can correctly reject $H_{1n} : f = f_n$ with a dominating probability once its alternative values f_n deviates from the null value 0 by an amount

$$d_{N,\lambda} = \sqrt{\lambda + n^{-2m} + \sigma_{N,\lambda}}.$$

Theorem

Suppose that $\lambda \to 0$, $n \to \infty$ when $N \to \infty$, and $\lim_{N\to\infty} n\lambda^{1/2m}$ exists (which could be infinity). Then for any $\varepsilon > 0$, there exist $C_\varepsilon, N_\varepsilon > 0$ s.t. for any $N \geq N_\varepsilon$,

$$\inf_{\substack{f \in \mathcal{B} \\ \|f\|_2 \geq C_\varepsilon d_{N,\lambda}}} P_f\left(\phi_{N,\lambda} = 1\right) \geq 1 - \varepsilon, \tag{2}$$

where $\mathcal{B} = \{f \in S^m(\mathbb{I}) : \|f\|_{\mathcal{H}} \leq C\}$ for a positive constant C.

An Upper Bound s^{**}

- The above theorem implies that the separation rate $d_{N,\lambda}$ achieves its minimal possible rate $d^*_{N,\lambda} := N^{-2m/(4m+1)}$ if one of the following two conditions hold:
  - $s = O(N^{(4m-1)/(4m+1)})$ and $\lambda \asymp N^{-4m/(4m+1)}$;
  - $s \asymp N^{(4m-1)/(4m+1)}$ and $\lambda = o(N^{-4m/(4m+1)})$;
- Denote $s^{**} = N^{(4m-1)/(4m+1)}$ and $\lambda^{**} = N^{-4m/(4m+1)}$.

An Upper Bound s^{**}

- The above theorem implies that the separation rate $d_{N,\lambda}$ achieves its minimal possible rate $d^*_{N,\lambda} := N^{-2m/(4m+1)}$ if one of the following two conditions hold:
  - $s = O(N^{(4m-1)/(4m+1)})$ and $\lambda \asymp N^{-4m/(4m+1)}$;
  - $s \asymp N^{(4m-1)/(4m+1)}$ and $\lambda = o(N^{-4m/(4m+1)})$;
- Denote $s^{**} = N^{(4m-1)/(4m+1)}$ and $\lambda^{**} = N^{-4m/(4m+1)}$.

An Upper Bound s^{**}

- The above theorem implies that the separation rate $d_{N,\lambda}$ achieves its minimal possible rate $d^*_{N,\lambda} := N^{-2m/(4m+1)}$ if one of the following two conditions hold:
  - $s = O(N^{(4m-1)/(4m+1)})$ and $\lambda \asymp N^{-4m/(4m+1)}$;
  - $s \asymp N^{(4m-1)/(4m+1)}$ and $\lambda = o(N^{-4m/(4m+1)})$;
- Denote $s^{**} = N^{(4m-1)/(4m+1)}$ and $\lambda^{**} = N^{-4m/(4m+1)}$.

An Upper Bound s^{**}

- The above theorem implies that the separation rate $d_{N,\lambda}$ achieves its minimal possible rate $d^*_{N,\lambda} := N^{-2m/(4m+1)}$ if one of the following two conditions hold:
  - $s = O(N^{(4m-1)/(4m+1)})$ and $\lambda \asymp N^{-4m/(4m+1)}$;
  - $s \asymp N^{(4m-1)/(4m+1)}$ and $\lambda = o(N^{-4m/(4m+1)})$;
- Denote $s^{**} = N^{(4m-1)/(4m+1)}$ and $\lambda^{**} = N^{-4m/(4m+1)}$.

Main Theorem V: Un-Improvable s^{**}

Theorem

Suppose that $s \gg s^{**}$, $\lambda \to 0$, $n \to \infty$ *when* $N \to \infty$, *and* $\lim_{N\to\infty} n\lambda^{1/2m}$ *exists (which could be infinity). Then there exists a positive sequence* $\beta_{N,\lambda}$ *with* $\lim_{N\to\infty} \beta_{N,\lambda} = \infty$ *s.t.*

$$\limsup_{N\to\infty} \inf_{\substack{f\in\mathcal{B} \\ \|f\|_2 \geq \beta_{N,\lambda} d^*_{N,\lambda}}} P_f\left(\phi_{N,\lambda} = 1\right) \leq \alpha. \tag{3}$$

Recall that $(1-\alpha)$ *is the pre-specified significance level.*

Remark: The above theorem says that when $s \gg s^{**}$, our proposed test is no longer powerful even when $\|f\|_2 \gg d^*_{N,\lambda}$. In other words, our test method fails to be optimal. Therefore, $s^{**} = N^{(4m-1)/(4m+1)}$ is a *sharp* upper bound.

A "Theoretical" Suggestion

When applying divide-and-conquer method to process massive data, we may want to allocate our data as follows:

- Distribute to
$$s \asymp N^{2m/(2m+1)}$$
machines for obtaining an optimal estimate;
- Distribute to
$$s \asymp N^{4m/(4m+1)}$$
machines for performing an optimal test.

A “Theoretical” Suggestion

When applying divide-and-conquer method to process massive data, we may want to allocate our data as follows:

- Distribute to
$$s \asymp N^{2m/(2m+1)}$$
machines for obtaining an optimal estimate;
- Distribute to
$$s \asymp N^{4m/(4m+1)}$$
machines for performing an optimal test.

A "Theoretical" Suggestion

When applying divide-and-conquer method to process massive data, we may want to allocate our data as follows:

- Distribute to
$$s \asymp N^{2m/(2m+1)}$$
machines for obtaining an optimal estimate;
- Distribute to
$$s \asymp N^{4m/(4m+1)}$$
machines for performing an optimal test.

Conjecture: D&C is a new form of tuning?

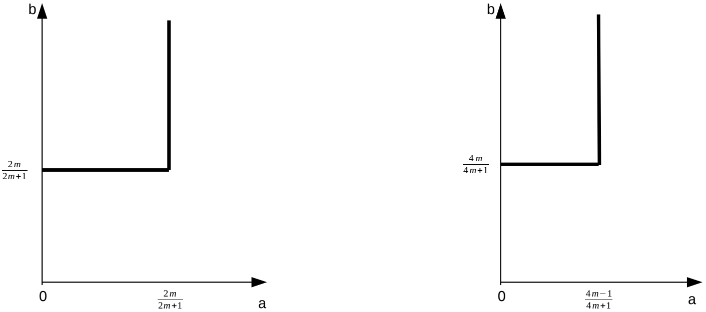


Figure: *Two lines indicate the choices of $s \asymp N^a$ and $\lambda \asymp N^{-b}$, leading to minimax optimal estimation (left) and minimax optimal testing (right). Whereas (a, b)'s outside these two lines lead to suboptimal rates.*