How Many Processors Do We Really Need in Parallel Computing?

Guang Cheng and Zuofeng Shang

Purdue University and Binghamton University

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 type of "*Big Data Theory.*"

A Basic Question from Statisticians

How many machines do we really need in parallel computing from a statistical theory perspective?

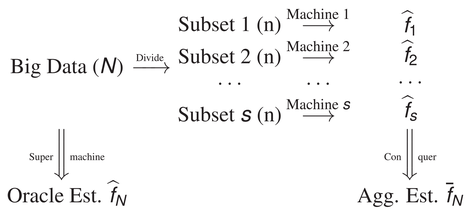
- In this poster, we address this basic, yet fundamentally important, question by carefully analyzing statistical versus computational trade-off in the above D&C framework;
- In particular, an intriguing phase transition phenomenon is discovered for the number of deployed machines that ends up being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

Divide and Conquer Strategy

A Flowchart of D&C

Nonparametric (univariate) regression model:

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



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

A Plot for Computing Time

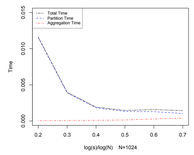


Figure: Computing time for $\bar{f}_N$ when $N = 1024$ as s diverges

Statistical-and-Computational Tradeoff

A Plot for Mean Squared Error

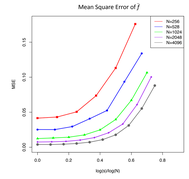


Figure: Mean-square errors of $\bar{f}_N$ under different N as s diverges

Our Major Goal

- We would like to know how fast s is allowed 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 will show 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 the D&C algorithm.

Computational Limit I: Statistical Estimation

- 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 $[0, 1]$;
- At each machine, we obtain a smoothing spline sub-estimate as

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

where $\langle f, g\rangle$ is a roughness penalty and $\lambda > 0$ is a smoothing parameter;
- $\mathrm{MSE}_{f_0}(\bar{f}_N) = E\{\|\bar{f}_N - E\{\bar{f}_N\}\|_2^2\} + \|E\{\bar{f}_N\} - f_0\|_2^2$.

A Graphical Illustration

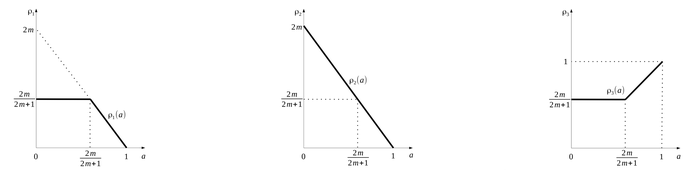


Figure: *Plots of ρ_1, ρ_2, ρ_3 versus a, indicated by thick solid lines. Here, $N^{-\rho_1}$ ($N^{-\rho_2}$) indicates the upper (lower) bound of squared bias of $\bar{f}_N$, and $N^{-\rho_3}$ indicates the upper bound of variance of $\bar{f}_N$. Phase transition occurs at $a = 2m/(2m+1)$. Lower bound result is obtained over $\{f_0 \in S^m(\mathbb{I}) : \|f_0\|_{\mathcal{H}} \le C\}$, i.e., "worst scenario" consideration.*

A *sharp* upper bound $s^* \asymp N^{2m/(2m+1)}$ is established for optimal estimation.

Computational Limit II: Statistical Testing

- 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:

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

- Our testing rule is thus

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

where $\mu_{N,\lambda}$ and $\sigma_{N,\lambda}^2$ are mean and variance of $T_{N,\lambda}$, respectively;
- Testing consistency (Type I error) 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. Specifically, the separation rate achieves its minimal value $N^{-2m/(4m+1)}$ under the following *sharp* upper bound

$$s^{**} \asymp N^{(4m-1)/(4m+1)}.$$

Future Works

Conjecture: is D&C 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.*

A "Theoretical" Suggestion

When applying D&C to massive data, we may allocate machines as:

- 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.

Practical Formulae to be Developed...

- We show that λ should be chosen in the order of N even when each subsample has size n. Hence, the standard generalized cross validation method does not work here;

  How to select λ that is scaled to N in practice?
- So far, we can only give a theoretical upper bound for s.

  How to pick the number of machines in practice using data-depdent method?

Department of Statistics, Purdue University Mail: chengg@stat.purdue.edu WWW: http://www.stat.purdue.edu/∼chengg