Analysis of Distributed Learning Algorithms

Ding-Xuan Zhou

City University of Hong Kong

E-mail: mazhou@cityu.edu.hk

Supported in part by Research Grants Council of Hong Kong

Start

November 5, 2016

Outline of the Talk

I. Distributed learning with big data

II. Least squares regression and and regularization

III. Distributed learning with regularization schemes

IV. Optimal rates for regularization

V. Other distributed learning algorithms

VI. Further topics

I. Distributed learning with big data

Big data leads to scientific challenges:
storage bottleneck, algorithmic scalability, ...

Distributed learning: based on a divide-and-conquer approach

A distributed learning algorithm consisting of three steps:
(1) partitioning the data into disjoint subsets
(2) applying a learning algorithm implemented in an individual machine or processor to each data subset to produce an individual output
(3) synthesizing a global output by utilizing some average of the individual outputs

Advantages: reducing the memory and computing costs to handle big data

If we divide a sample $D = \{(x_i, y_i)\}_{i=1}^N$ of input-output pairs into disjoint subsets $\{D_j\}_{j=1}^m$, applying a learning algorithm to the much smaller data subset D_j gives an output f_{D_j}, and the global output might be $\overline{f}_D = \frac{1}{m}\sum_{j=1}^m f_{D_j}$.

The distributed learning method has been observed to be very successful in many practical applications. There a challenging theoretical question is raised:

If we had a "big machine" which could implement the same learning algorithm to the whole data set D to produce an output f_D, could $\overline{f}_D$ be as efficient as f_D?

Recent work: Zhou-Chawla-Jin-Williams, Zhang-Duchi-Wainwright, Shamir-Srebro, ...

II. Least squares regression and and regularization

II.1. Model for the least squares regression. Learn $f : \mathcal{X} \to \mathcal{Y}$ from a random sample $D = \{(x_i, y_i)\}_{i=1}^N$

Take $\mathcal{X}$ to be a compact metric space and $\mathcal{Y} = \mathbf{R}$. $y \approx f(x)$ Due to noises or other uncertainty, we assume a (unknown) probability measure ρ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ governs the sampling.

marginal distribution ρ_X on $\mathcal{X}$: $\mathbf{x} = \{x_i\}_{i=1}^N$ drawn according to ρ_X

conditional distribution $\rho(\cdot|x)$ at $x \in \mathcal{X}$

Learning the **regression function**: $f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x)$

$y_i \approx f_\rho(x_i)$

II.2. Error decomposition and ERM

$\mathcal{E}^{ls}(f) = \int_{\mathcal{Z}}(f(x)-y)^2 d\rho$ minimized by f_ρ:

$$\mathcal{E}^{ls}(f) - \mathcal{E}^{ls}(f_\rho) = \|f - f_\rho\|^2_{L^2_{\rho_X}} =: \|f - f_\rho\|^2_\rho \geq 0.$$

Classical Approach of Empirical Risk Minimization (ERM)
Let $\mathcal{H}$ be a compact subset of $C(\mathcal{X})$ called hypothesis space (model selection). The ERM algorithm is given by

$$f_D = \arg\min_{f\in\mathcal{H}} \mathcal{E}^{ls}_D(f), \qquad \mathcal{E}^{ls}_D(f) = \frac{1}{N}\sum_{i=1}^{N}(f(x_i)-y_i)^2.$$

Target function $f_{\mathcal{H}}$: best approximation of f_ρ in $\mathcal{H}$

$$f_{\mathcal{H}} = \arg\min_{f\in\mathcal{H}} \mathcal{E}^{ls}(f) = \arg\inf_{f\in\mathcal{H}} \int_{\mathcal{Z}}(f(x)-y)^2 d\rho$$

II.3. Approximation error

Analysis. $\|f_D - f_\rho\|^2_{L^2_{\rho_X}} = \int_{\mathcal{X}} (f_D(x) - f_\rho(x))^2 d\rho_X$ is bounded by $2 \sup_{f \in \mathcal{H}} \left| \mathcal{E}^{ls}_D(f) - \mathcal{E}^{ls}(f) \right| + \left\{ \mathcal{E}^{ls}(f_{\mathcal{H}}) - \mathcal{E}^{ls}(f_\rho) \right\}$.

Approximation Error. Smale-Zhou (Anal. Appl. 2003)

$$\mathcal{E}^{ls}(f_{\mathcal{H}}) - \mathcal{E}^{ls}(f_\rho) = \|f_{\mathcal{H}} - f_\rho\|^2_{L^2_{\rho_X}} = \inf_{f \in \mathcal{H}} \int (f(x) - f_\rho(x))^2 d\rho_X$$

$f_{\mathcal{H}} \approx f_\rho$ when $\mathcal{H}$ is rich

Theorem 1 *Let B be a Hilbert space (such as a Sobolev space or a reproducing kernel Hilbert space). If $B \subset L^2_{\rho_X}$ is dense and $\theta > 0$, then*

$$\inf_{\|f\|_B \le R} \|f - f_\rho\|_{L^2_{\rho_X}} = O(R^{-\theta})$$

if and only if f_ρ lies in the interpolation space $(B, L^2_{\rho_X})_{\frac{\theta}{1+\theta}, \infty}$.

II.4. Examples of hypothesis spaces

Sobolv spaces: if $\mathcal{X} \subset \mathbb{R}^n$, ρ_X is the normalized Lebesgue measure, and B is the Sobolev space H^s with $s > n/2$, then $(H^s, L^2_{\rho_X})_{\frac{\theta}{1+\theta},\infty}$ is the Besov space $B^{\frac{\theta}{1+\theta}s}_{2,\infty}$ and $H^{\frac{\theta}{1+\theta}s} \subset B^{\frac{\theta}{1+\theta}s}_{2,\infty} \subset H^{\frac{\theta}{1+\theta}s-\epsilon}$ for any $\epsilon > 0$.

Range of power of integral operator: if $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a Mercer kernel (continuous, symmetric and positive semidefinite), then the integral operator L_K on $L^2_{\rho_X}$ is defined by

$$L_K(f)(x) = \int_{\mathcal{X}} K(x,y) f(y) d\rho_X(y), \qquad x \in \mathcal{X}.$$

The r-th power L^r_K is well defined for any $r \geq 0$. Its range $L^r_K(L^2_{\rho_X})$ gives the RKHS $\mathcal{H}_K = L^{1/2}_K(L^2_{\rho_X})$ and for $0 < r \leq 1/2$, $L^r_K(L^2_{\rho_X}) \subset (\mathcal{H}_K, L^2_{\rho_X})_{2r,\infty}$ and $(\mathcal{H}_K, L^2_{\rho_X})_{2r,\infty} \subset L^{r-\epsilon}_K(L^2_{\rho_X})$ for any $\epsilon > 0$ when the support of ρ_X is $\mathcal{X}$. So we may assume

$$f_\rho = L^r_K(g_\rho) \quad \text{for some } r > 0, g_\rho \in L^2_{\rho_X}.$$

II.5. Least squares regularization

$$f_{D,\lambda} := \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{N}\sum_{i=1}^{N} (f(x_i)-y_i)^2 + \lambda\|f\|_K^2 \right\}, \quad \lambda > 0.$$

A large literature in learning theory: books by Vapnik, Schölkopf-Smola, Wahba, Anthony-Bartlett, Shawe-Taylor-Cristianini, Steinwa Christmann, Cucker-Zhou, ...
many papers: Cucker-Smale, Zhang, De Vito-Caponnetto-Rosasco, Smale-Zhou, Lin-Zeng-Fang-Xu, Yao, Chen-Xu, Shi-Feng-Zhou, Wu-Ying-Zhou, ...

regularity of f_ρ

complexity of $\mathcal{H}_K$: covering numbers, decay of eigenvalues $\{\lambda_i\}$ of L_K, effective dimension, ...

decay of y: $|y| \le M$, exponential decay, moment decaying condition, $\mathbb{E}[|y|^q] < \infty$ for some $q > 2$, $\sigma_\rho^2 \in L^p_{\rho_X}$ for the conditional variance $\sigma_\rho^2(x) = \int_{\mathcal{Y}} (y - f_\rho(x))^2 \, d\rho(y|x)$, ...

III. Distributed learning with regularization schemes

Join work with S. B. Lin and X. Guo (under major revision for JMLR)

Distributed learning with the data disjoint union $D = \cup_{j=1}^m D_j$:

$$\overline{f}_{D,\lambda} = \sum_{j=1}^m \frac{|D_j|}{|D|} f_{D_j,\lambda}$$

Define the effective dimension to measure the complexity of $\mathcal{H}_K$ with respect to ρ_X as

$$\mathcal{N}(\lambda) = \mathsf{Tr}\left((L_K + \lambda I)^{-1} L_K\right) = \sum_i \frac{\lambda_i}{\lambda_i + \lambda}, \qquad \lambda > 0.$$

Note that $\lambda_i = O(i^{-2\alpha})$ implies $\mathcal{N}(\lambda) = O(\lambda^{-\frac{1}{2\alpha}})$

III.1. Error analysis for distributed learning

Theorem 2 *Assume* $|y| \le M$ *and* $f_\rho = L_K^r(g_\rho)$ *for some* $0 \le r \le \frac{1}{2}$ *and* $g_\rho \in \mathcal{H}_K$. *If* $\mathcal{N}(\lambda) = O(\lambda^{-\frac{1}{2\alpha}})$ *for some* $\alpha > 0$, $|D_j| = \frac{N}{m}$ *for* $j = 1, \ldots, m$, *and* $m \le N^{\min\left\{\frac{12\alpha r+1}{5(4\alpha r+2\alpha+1)}, \frac{4\alpha r}{4\alpha r+2\alpha+1}\right\}}$, *then by taking* $\lambda = N^{-\frac{2\alpha}{4\alpha r+1}}$, *we have*

$$E\left[\left\|\overline{f}_{D,\lambda} - f_\rho\right\|_\rho\right] = O\left(N^{-\frac{\alpha+2\alpha r}{2\alpha+4\alpha r+1}}\right).$$

If $f_\rho \in \mathcal{H}_K$ *and* $m \le N^{\frac{1}{4+6\alpha}}$, *the choice* $\lambda = \left(\frac{m}{N}\right)^{\frac{2\alpha}{2\alpha+1}}$ *yields*

$$E\left[\left\|\overline{f}_{D,\lambda} - f_{D,\lambda}\right\|_\rho\right] = O\left(N^{-\frac{\alpha}{2\alpha+1}} m^{-\frac{1}{4\alpha+2}}\right)$$

and

$$E\left[\left\|\overline{f}_{D,\lambda} - f_{D,\lambda}\right\|_K\right] = O\left(\frac{1}{\sqrt{m}}\right).$$

III.2. Previous work: Zhang-Duchi-Wainwright (2015):
If the normalized eigenfunctions $\{\varphi_i\}_i$ of L_K on $L^2_{\rho_X}$ satisfy

$$\|\varphi_i\|^{2k}_{L^{2k}_{\rho_X}} = E\left[|\varphi_i(x)|^{2k}\right] \le A^{2k}, \qquad i = 1, 2, \ldots,$$

for some constants $k > 2$ and $A < \infty$, $f_\rho \in \mathcal{H}_K$ and $\lambda_i = O(i^{-2\alpha})$ for some $\alpha > 1/2$, then $E\left[\left\|\overline{f}_{D,\lambda} - f_\rho\right\|^2_\rho\right] = O\left(N^{-\frac{2\alpha}{2\alpha+1}}\right)$ when $\lambda = N^{\frac{2\alpha}{2\alpha+1}}$ and $m = O((N^{\frac{2(k-4)\alpha-k}{2\alpha+1}}/(A^{4k}\log^k N))^{\frac{1}{k-2}})$.

An example of a C^∞ Mercer kernel without uniform boundedness of the eigenfunctions: Zhou (2002)

Advantages of our analysis:
(1) General results without any eigenfunction assumption
(2) Error estimates in the $\mathcal{H}_K$ metric (Smale-Zhou 2007)
(3) A novel second order decomposition applicable to other algorithms

IV. Optimal rates for regularization: by-product

Caponnetto-DeVito (2007): If $\lambda_i \approx i^{-2\alpha}$ with some $\alpha > 1/2$, then with $\lambda = \left(\frac{\log N}{N}\right)^{\frac{2\alpha}{2\alpha+1}}$,

$$\lim_{\tau\to\infty} \limsup_{N\to\infty} \sup_{\rho} \text{prob}\left[\left\|f_{D,\lambda_N} - f_\rho\right\|_\rho^2 \le \tau\left(\frac{\log N}{N}\right)^{\frac{2\alpha}{2\alpha+1}}\right] = 1.$$

Steinwart-Hush-Scovel (2009): If $\lambda_i = O\left(i^{-2\alpha}\right)$ with some $\alpha > 1/2$, and for some constant $C > 0$, the pair (K, ρ_X) satisfies

$$\|f\|_\infty \le C\|f\|_K^{\frac{1}{2\alpha}}\|f\|_\rho^{1-\frac{1}{2\alpha}}, \qquad \forall f \in \mathcal{H}_K,$$

then with $\lambda = N^{-\frac{2\alpha}{2\alpha+1}}$,

$$E\left[\left\|\pi_M\left(f_{D,\lambda}\right) - f_\rho\right\|_\rho^2\right] = O\left(N^{-\frac{2\alpha}{2\alpha+1}}\right).$$

Here π_M is the projection onto the interval $[-M, M]$.

Our result: $$E\left[\left\|f_{D,\lambda}-f_\rho\right\|_\rho\right]=O\left(N^{-\frac{\alpha}{2\alpha+1}}\right).$$

Theorem 3 *Assume* $E[y^2]<\infty$ *and* $\sigma_\rho^2\in L^p_{\rho_X}$ *for some* $1\le p\le\infty$. *If* $f_\rho=L_K^r(g_\rho)$ *for some* $g_\rho\in L^2_{\rho_X}$ *and* $0<r\le 1$, *and* $\mathcal{N}(\lambda)=O(\lambda^{-\frac{1}{2\alpha}})$ *for some* $\alpha>0$, *then by taking* $\lambda=N^{-\frac{2\alpha}{2\alpha\max\{2r,1\}+1}}$ *we have*

$$E\left[\left\|f_{D,\lambda}-f_\rho\right\|_\rho\right]=O\left(N^{-\frac{2r\alpha}{2\alpha\max\{2r,1\}+1}+\frac{1}{2p}\frac{2\alpha-1}{2\alpha\max\{2r,1\}+1}}\right).$$

In particular, when $p=\infty$ *(the conditional variances are uniformly bounded), we have*

$$E\left[\left\|f_{D,\lambda}-f_\rho\right\|_\rho\right]=O\left(N^{-\frac{2r\alpha}{2\alpha\max\{2r,1\}+1}}\right).$$

Second order decomposition used to solve two conjecture on kernel partial least squares: S. B. Lin-Zhou

V. Other distributed learning algorithms

Distributed learning with spectral algorithms based on SVD of Gramian matrices $\Big(K(x_i,x_j)\Big)_{i,j=1}^N$: Z. C. Guo-S. B. Lin-Zhou

Distributed learning with stochastic gradient descent: S. B. Lin-Zhou

Distributed learning with additional unlabeled data: X. Y. Chang-S. B. Lin-Zhou

VI. Further topics with distributed learning and deep nets

VI.1. Approximation theory of deep nets

Classical results on shallow nets (Cybenko 1989, Hornik 1991, Barron 1993, Mhaskar 1996): if σ is C^∞ strictly increasing function satisfying $\lim_{x\to-\infty}\sigma(x)=0$ and $\lim_{x\to\infty}\sigma(x)=1$ (sigmoidal function), and if f is in the Sobolev space $W_2^r(\mathbb{R}^d)$, then for every $N\in\mathbb{N}$, there exists a function $f_N(x)=\sum_{i=1}^N c_i\sigma(w_i\cdot x+b_i)$ with $c_i\in\mathbb{R}, w_i\in\mathbb{R}^d, b_i\in\mathbb{R}$ such that

$$\|f_N-f\|^2_{L^2(\mathbb{R}^d)}=O(N^{-2r/d}).$$

Lack of localized approximation (Chui-Li-Mhaskar 1994): the neural network with the activation function $\sigma=\chi_{[0,\infty)}$ does not provide localized approximation meaning that for every compact subset K of $\mathbb{R}^d$,

$$\inf_{N\in\mathbb{N},c_i,w_i,b_i}\left\|\sum_{i=1}^N c_i\sigma(w_i\cdot x+b_i)-\chi_{[-1,1]^d}\right\|_{L^1(K)}=0.$$

Approximation by deep nets

Neural network with 2 hidden layers:

$$f(x) = \sum_{i=1}^{n_2} c_i \sigma \left(\sum_{j=1}^{n_1} a_{i,j} \sigma \left(w_{i,j} \cdot x + b_{i,j} \right) \right) + c_0$$

with $c_i \in \mathbb{R}, a_{i,j} \in \mathbb{R}^d, b_{i,j} \in \mathbb{R}$.

Chui-Li-Mhaskar (1994): the neural network with with 2 hidden layers and an activation measurable function σ satisfying $\lim_{x \to -\infty} \sigma(x) = 0$, $\lim_{x \to \infty} \sigma(x) = 1$ and $\|\sigma\|_\infty < \frac{2d}{2d-1}$ provides localized approximation.

Eldan-Shamir (2016): an example of a function expressible by a 3-layer feedforward neural network cannot be approximated by any 2-layer neural network to certain accuracy unless the width is exponential in the dimension.

Telgarsky (2016): more examples

Neural network with 4 hidden layers: Shaham-Cloningen-Coifman (2016) For the rectify linear function $\sigma(x) = \max\{x, 0\}$, a depth-4 neural networks with N units can achieve the approximation order of $O(N^{-2/d})$ if f is C^2 on a smooth d-dimensional Riemannian manifold without boundary.

Robust and distributed learning with deep nets: Chui-Lin-Zhou (in progress)

VI.2. Stochastic gradient descent and mirror descent: Y. W. Lei-Zhou (Neural Computation 2016), Y. M. Ying-Zhou (ACHA 2016)

Learning with a mirror map $\Psi : \mathbb{R}^d \to \mathbb{R}$, a loss ϕ, and a convex regularizer r:

$$w_{t+1} = \arg\min_{w \in \mathbb{R}^d} \eta_t \langle w - w_t, \phi'_-(y_t, \langle w_t, x_t \rangle x_t \rangle + \eta_t r(w) + D_\Psi(w, w_t),$$

where η_t is a step size and $D_\Psi(w, \tilde{w})$ is the Bregman distance between w and $\tilde{w}$.

Motivation: capture the geometry involving ℓ_p norms with $p \geq 1$ where $\Psi_p(x) = \frac{1}{2}\|x\|_p^2$.

VI.3. Compositional models for deep nets

Additive models (Stone 1985): $f(x_1, \ldots, x_d) = f_1(x_1) + \ldots + f_d(x_d)$
M. Yuan-Zhou (Ann. Stat. 2016), Christmann-Zhou (Anal. Appl. 2016)

Interaction models (Stone 1994):

$$f(x_1, \ldots, x_d) = \sum_{I \subseteq \{1,\ldots,d\}, |I| = d^*} f_I(x_I)$$

with $d^* \in \{1, \ldots, d\}$ and for $I = \{i_1, \ldots, i_{d^*}\} \subseteq \{1, \ldots, d\}$ with $|I| = d^*$, $x_I = (x_{i_1}, \ldots, x_{i_{d^*}})$.

Single index models and Projection pursuit (Härdle and Stoker 1989, Friedman and Stuetzle 1981): $f(x_1, \ldots, x_d) = \sum_{k=1}^{K} g_k(a_k \cdot x)$ with $K \in \mathbb{N}$, $a_k \in \mathbb{R}^d$ and univariate functions g_k

Hierarchical interaction models (Kohler1 and Krzyzak 2016): $f(x_1, \ldots, x_d) = g(f_1(x_{I_1}), f_2(x_{I_2}), \ldots, f_{d^*}(x_{I_{d^*}})$ with $d^* \in \{1, \ldots, d\}$ and $I_i \subseteq \{1, \ldots, d\}$ with $|I_i| = d^*$

Compositional functions: Mhaskar-Liao-Poggio, Mhaskar-Poggio (2016)

THANK YOU!