Distributed Learning with Minimum Error Entropy Principle

Xin Guo

The Hong Kong Polytechnic University

Joint Statistical Meetings 2018, Jul 27 – Aug 01, 2019

Supported in part by Research Grants Council of Hong Kong
Joint works with Ting Hu and Qiang Wu

Outline

- Definition of the minimum error entropy algorithms
- Distributed MEE
- MEE with semi-supervised data
- Empirical feature-based summary statistics
- Semi-supervised learning with summary statistics

- Renyi's Entropy

$$H(E) = -\log \mathbb{E}(p_E) = -\log \int p_E^2(e)de.$$

- Parzen Window density estimator of prediction error

$$\hat{p}_E(e) = \frac{1}{mh} \sum_{i=1}^{m} G\left(\frac{(e-e_i)^2}{h^2}\right),$$

  - $h > 0$, scaling parameter
  - $D = \{(x_i, y_i)\}_{i=1}^m$, a sample
  - $e_i = g(x_i) - y_i$, error of a predicted function g
  - G: a windowing function, e.g., $G(t) = e^{-t}\mathbf{1}_{t\geq 0}$.

- Regularized Minimum Error Entropy Algorithm

$$f_{D,\lambda} = \arg\min_{f\in\mathcal{H}} \left\{ -\frac{h^2}{m^2} \sum_{i,j=1}^{m} G\left(\frac{[f(x_i,x_j) - y_i + y_j]^2}{h^2} \right) + \lambda\Omega(f) \right\},$$

  - We use $\mathcal{H} = \mathcal{H}_K$, a reproducing kernel Hilbert space generated by a Mercer kernel $K : X^2 \times X^2 \to \mathbb{R}$, and take $\Omega(f) = \|f\|_K^2$.
  - MEE: widely applied in signal processing, regression analysis, feature selection, data clustering, etc., see Erdogmus & Principe, '03; Chen et al., '10; Gokcay & Principe, '02; Shen & Li, '15; Silva et al., '10.
  - Learning theory of MEE, Hu et al., '15; Fan et al., '16; Hu et al., '13.

Distributed MEE

- Distributed learning
  - $D = D_1 \cup D_2 \cup \cdots \cup D_\ell$;
  - For technical reasons, assume $|D_1| = |D_2| = \cdots = |D_\ell|$;
  - $\bar{f}_{D,\lambda} = \sum_{j=1}^{\ell} \frac{|D_j|}{|D|} f_{D_j,\lambda}$.
- Similar setting extensively studied in recent years. Zhang et al., '13&'15; Shamir & Srebro, '14; Blanchard & Mücke, '16; Lin et al., '17; Chang et al., '17; Lin & Zhou, '18; Guo et al., '16.

Convergence of Distributed MEE

Assume

- Effective dimension (Zhang, '02): for some $0 < s \leq 1$,

$$\mathcal{N}(\lambda) := \text{Trace}(L_K(L_K + \lambda I)^{-1}) = O(\lambda^{-s}), \quad \text{as } \lambda \to 0^+,$$

  where $L_K f := \int_{X^2} f(x,u) K_{(x,u)} d\rho_X(x) d\rho_X(u), \quad \forall f \in L^2_{\rho_X^2}$. The case $s \geq 1$ is trivial.
- Regularity of the regression function: $f_\rho \in L_K^r(L^2_{\rho_X^2})$, for some $0 < r \leq 1$.
- Moment condition: for $0 < \sigma, M < \infty$ and ρ_X-almost X,

$$\mathbb{E}\left(|Y|^q | X\right) \leq \frac{1}{2} q! \sigma^2 M^{q-2}, \text{ for any integer } q \geq 2.$$

- $\sup_{0<\tau<\infty} |G'(\tau)| < \infty$, and
  $|G'(a) - G'(0)| \leq c_p a^p$, for any $0 < a < \infty$.

Theorem (G & Hu & Wu, JMLR, under review)

Let $\lambda = m^{-1/(s+\max\{2r,1\})}$,

$$\ell \begin{cases} = 1, & \text{for } 0 < r \leq 1/2, \\ \leq \lambda^{\frac{1}{2}-r} \log^{-4} m, & \text{for } 1/2 < r \leq 1, \end{cases}$$

and $h \geq (\lambda^{-r-1-p} \log^{2p+4} m)^{\frac{1}{2p}}$. *With the above four assumptions, one has with confidence at least* $1-\delta$ *that*

$$\|\bar{f}_{D,\lambda} - f_\rho\|_\rho = O\left(m^{-\frac{r}{s+\max\{2r,1\}}}\right) \log^{2p+4} \frac{16}{\delta}.$$

- When $1/2 \leq r \leq 1$, the above learning rate is minimax optimal for point-wise regularized least squares.

Semi-supervised learning

- $D' = D'_1 \cup \ldots \cup D'_\ell$, a set of unlabeled data. For each $x' \in D'$, fake a label $\tilde{y} = 0$.
- For technical reasons, assume $|D'_1| = |D'_2| = \cdots = |D'_\ell|$;
- Compensation: for $(x, y) \in D$, update the label as $\tilde{y} = \frac{|D'|+|D|}{|D|} y$.
- Write $D^* = D \cup D'$ with $D^*_i = D_i \cup D'_i$ for $1 \le i \le \ell$.
- local output function: $f_{D^*_i,\lambda}$ obtained from semi-supervised data D^*_i with fake labels. Define $\bar{f}_{D^*,\lambda} = \frac{1}{\ell} \sum_{i=1}^{\ell} f_{D^*_i,\lambda}$.
- Unlabeled data are observed
  - helpful for capturing the underlying manifold structures of data distribution. Coifman & Lafon, '06; Bertozzi et al., '18.
  - helpful for relaxing the requirement on single-source minimum sample size in distributed learning. Lin & Zhou, '18; Chang et al., '17; and G & Hu & Wu, manuscript.
  - helpful for improving the convergence under weak regularity assumptions of the regression function. Chang et al., '17; and G & Hu & Wu, manuscript.

Theorem (G & Hu & Wu, JMLR, under review)

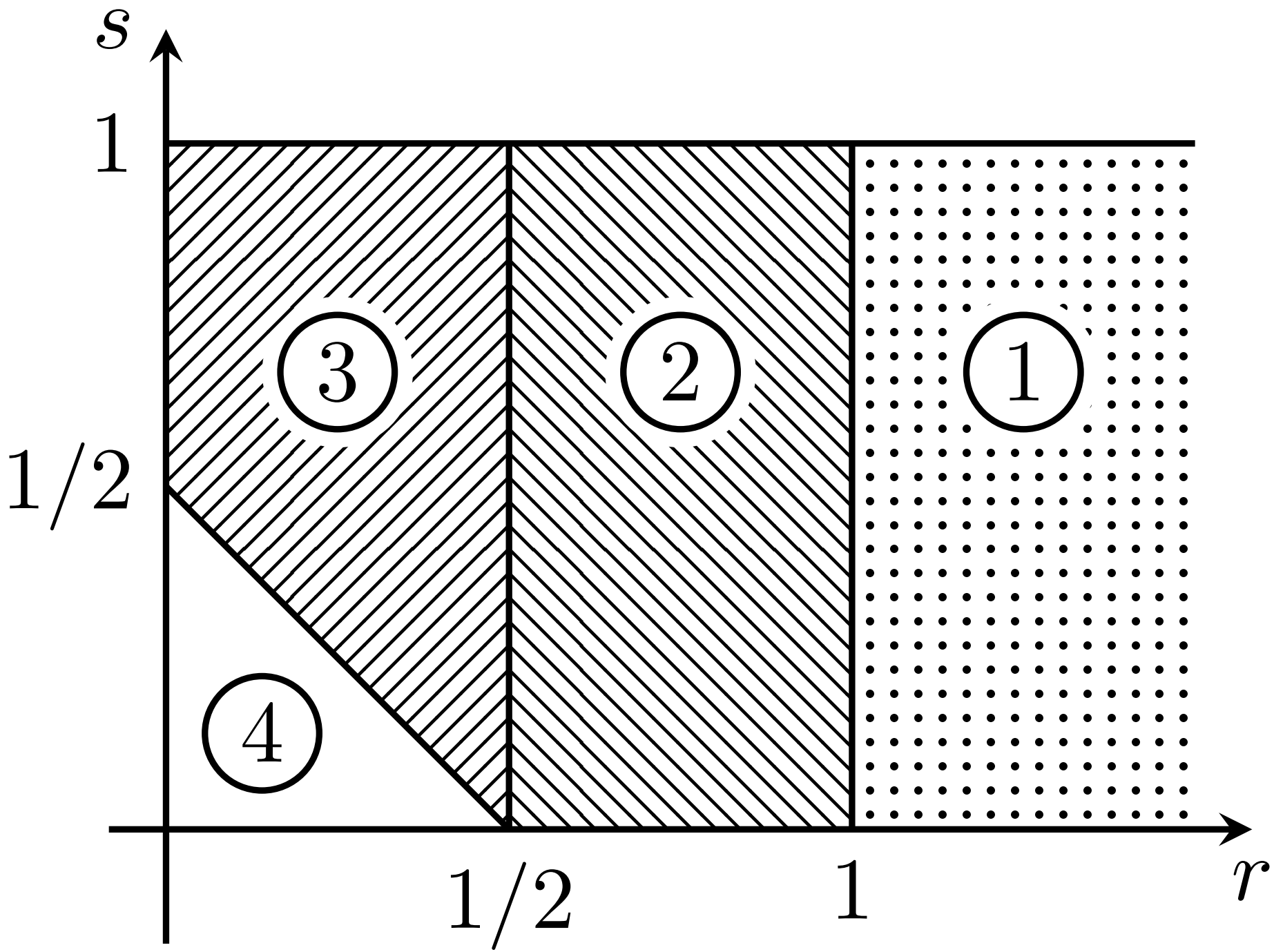
Assume $0 < r \leq 1$, $s + r \geq 1/2$, *and* $|D^*| \geq \max\left\{|D|^{\frac{s+1}{2r+s}}, |D|\right\}$. *Let* $\lambda = |D|^{-\frac{1}{2r+s}}$,

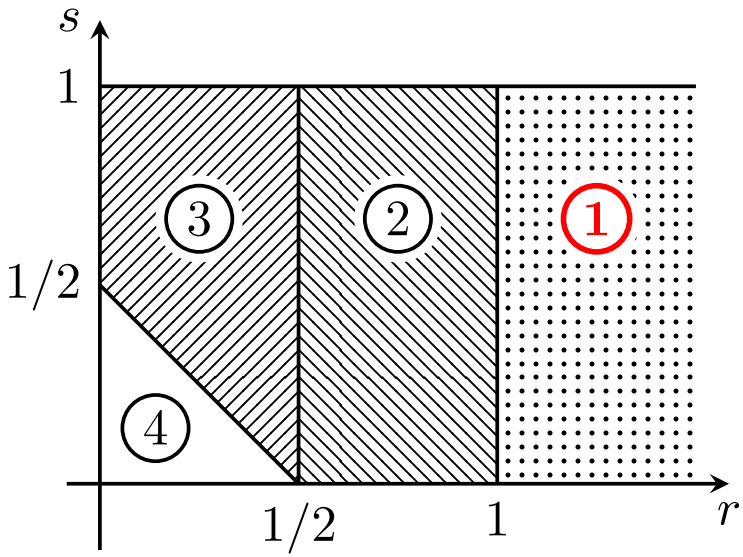
$$\ell \leq \left(\log^{-4}|D|\right) \min\left\{\sqrt{|D^*|\lambda^{1+s}}, \sqrt[3]{|D^*|\lambda^{2-2r-s}}\right\}, \text{ and}$$

$$h \geq \left\{|D|^{\frac{2r+1}{2(2r+s)}} \left[\left(\frac{|D^*|}{\lambda|D|}\right)^{p+\frac{1}{2}} + \left(\frac{|D^*|}{|D|}\right)^{2p+1}\right] \log^{2p+4}|D|\right\}^{\frac{1}{2p}}.$$

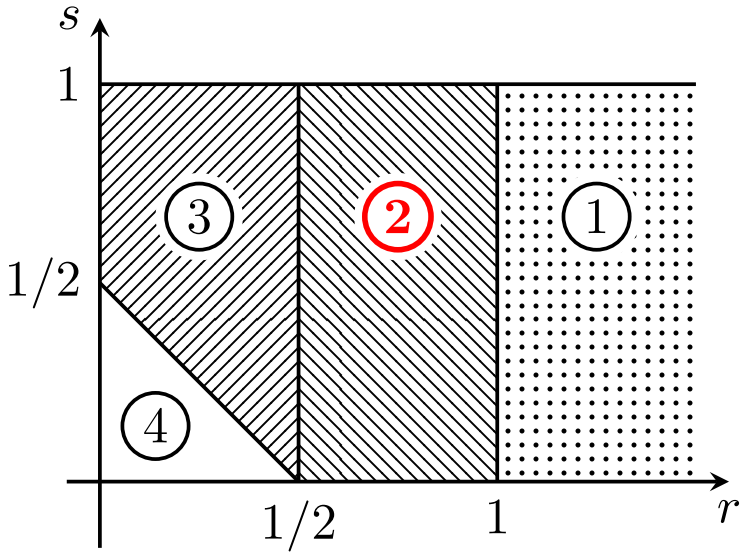
With the above four assumptions, one has with confidence at least $1 - \delta$ *that*

$$\left\|\bar{f}_{D^*,\lambda} - f_\rho\right\|_\rho \leq O\left(|D|^{-\frac{r}{2r+s}}\right) \log^{2p+4}\frac{16}{\delta}.$$

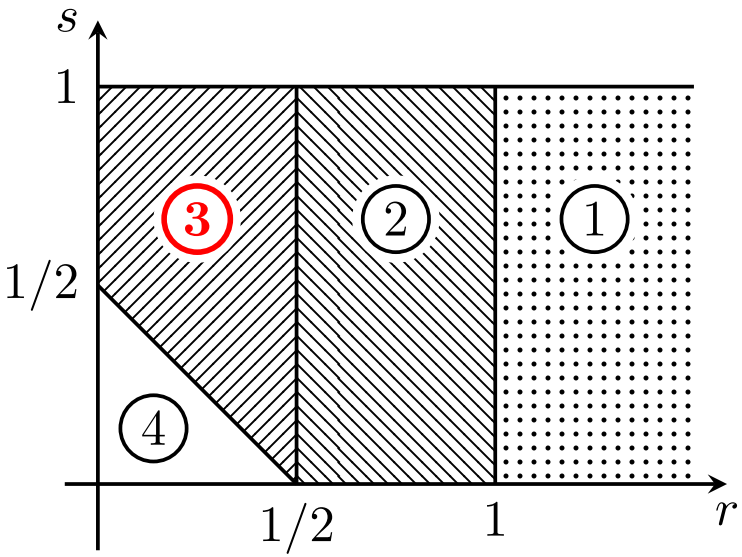

s
1
1/2
3
2
1
4
1/2
1
r



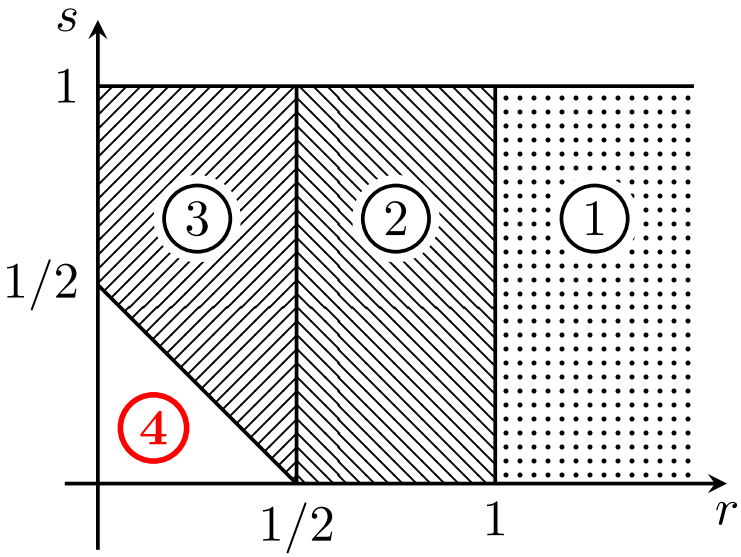
- Saturation is observed for Distributed MEE, similar as regularized least squares. Overcome by spectral algorithms, Lo Gerfo et al., '08; Blanchard & Mücke, '16&'18; Guo et al., '17.



- Distributed MEE achieves the learning rates which equal the minimax optimal rates for point-wise regression learning with methods, Steinwart et al., '09; Caponnetto & De Vito, '07; Bauer et al., '07; Blanchard & Mücke, '18.
- Unlabeled data help to essentially remove the restriction of the maximum number of computing nodes Distributed MEE can be distributed to. Also reported in Chang et al., '17; Lin & Zhou, '18.
- While allowing a "more distributed" computation, unlabeled data may increase the single-node computational complexity and memory requirement.



- Fully supervised Distributed MEE could not indeed be distributed without sacrificing the learning rates.
- The learning rate $O(|D|^{-\frac{r}{2r+s}})$ achieved by semi-supervised distributed MEE matches the optimal learning rate proved in Steinwart et al., '09 under the boundedness assumption of $L_K^r : L^2 \to L^\infty$ for point-wise regularized least squares. While fully supervised Distributed MEE does not achieve this lower bound.
- Semi-supervised Distributed MEE also slacks the upper bound of ℓ.
- Similar phenomenon also reported in Chang et al., '17; Lin & Zhou, '18 for point-wise learning.



- Our analysis of semi-supervised Distributed MEE fails to achieve the rate $O(|D|^{-\frac{r}{2r+s}})$.
- Area 4 seems to be the situation that one should avoid, because of less regularity assumption and small hypothesis space.

Regression with Summary Statistics

- Consider the linear regression model $y = \mathbb{X}\beta + \varepsilon$, and its least squares solution $\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T y$.
- Use $\hat{\beta}' = (\tilde{\mathbb{X}}^T\tilde{\mathbb{X}})^{-1}\mathbb{X}^T y$ instead of $\hat{\beta}$, where $\tilde{\mathbb{X}}$ is the coefficient matrix made by openly accessible and unlabeled data without privacy issues.
- Summary statistics: $\mathbb{X}^T y$.

We thank Prof Jian Huang for introducing to use the following two papers

- Xiang Zhu and Matthew Stephens. Bayesian large-scale multiple regression with sum mary statistics from genome-wide association studies. Ann. Appl. Stat., 11(3):1561-1592, 2017.
- Jin Liu, Can Yang, Yuling Jiao, and Jian Huang. ssLasso: a summary-statistic-based regression using Lasso. Preprint, 2017.

Regression and classification learning

- Given a sample $D = \{(x_i, y_i)\}_{i=1}^m$, find a function f_D that can predict an output $f_D(x) \in Y \subset \mathbb{R}$ for a new instance x. For example, $Y = \mathbb{R}$ for regression, and $Y = \{\pm 1\}$ for binary classification.
- Empirical risk minimization ($\lambda > 0$)

$$f_{D,\lambda} = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{m} \sum_{i=1}^{m} V(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2 \right\}.$$

- $V(t, y)$: the loss function, usually convex on t. For example,
  - $V(t, y) = (t - y)^2$, used in regularized least squares for regression;
  - $V(t, y) = (1 - \tau)(t - y)\mathbf{1}_{t>y} + \tau(y - t)\mathbf{1}_{y \geq t}$, $0 < \tau < 1$, the τ-pinball loss, for quantile regression;
  - $V(t, y) = \max\{0, 1 - ty\}$, the hinge loss, used in SVM for binary classification.
- $\mathcal{H}$: the hypothesis space, for example a reproducing kernel Hilbert space (RKHS), or an artificial neural network.

Reproducing kernel Hilbert space (RKHS)

Let X be a metric space, and $K : X \times X \to \mathbb{R}$. K is called a reproducing kernel on X if it is symmetric ($K(x,u) = K(u,x)$ for any $x, u \in X$), and positive semi-definite (for any $x_1, \ldots, x_m \in X$, the Gram matrix $(K(x_i, x_j))_{1 \le i,j \le m}$ is positive semi-definite). If furthermore, K is continuous, we call it a Mercer kernel.

Examples of Mercer kernels

- linear kernel, $K(x,u) = \langle x, u \rangle$ on $\mathbb{R}^n$;
- Gaussian kernel, $K(x,u) = \exp\left(-\frac{1}{2\sigma^2}\|x-u\|^2\right)$ on $\mathbb{R}^n$, where $\sigma > 0$;
- Polynomial kernel, $K(x,u) = (1 + \langle x, u \rangle)^d$ on $\mathbb{R}^n$, where $d \ge 1$ is an integer;
- Inverse multiquadrics, $K(x,u) = \left(c^2 + \|x-u\|^2\right)^{-\alpha}$ on $\mathbb{R}^n$, for any $c, \alpha > 0$.

Empirical features

- Integral operator $L_K : \mathcal{H}_K \to \mathcal{H}_K$, $L_K f = \int_X K_x f(x) d\rho_X(x)$, where ρ_X is the marginal distribution of $(X \times Y, \rho)$ on X. L_K is symmetric, p.s.d., compact, Hilbert-Schmidt, and of trace class.
- Eigenvalues of L_K: $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$. Normalized eigenvectors $\phi_1, \phi_2, \ldots$, are called features.
- Empirical operator $L_K^{\mathbf{x}} : \mathcal{H}_K \to \mathcal{H}_K$, $L_K^{\mathbf{x}} f := \frac{1}{m} \sum_{i=1}^m f(x_i) K_{x_i}$, where $\mathbf{x} = \{x_i\}_{i=1}^m$. Eigenvalues: $\lambda_1^{\mathbf{x}} \geq \lambda_2^{\mathbf{x}} \geq \ldots \geq 0$. Normalized eigenvectors $\phi_1^{\mathbf{x}}, \phi_2^{\mathbf{x}}, \ldots$ are called empirical features (data dependent features), and are used in kernel principal component analysis (Schölkoph et al., '98; Mika et al., '99; Liu, '04; Hoffmann, '07), kernel ridge regression (Cristianini & Shawe-Taylor, '00; Hastie et al., '01) kernel projection machine (Blanchard et al., '04; Zwald et al., '04), spectral algorithms (Lo Gerfo et al., '08; Caponnetto & Yao '10, Blanchard & Mucke, '16 and '18; Guo et al., '17), and diffusion maps (Coifman & Lafon '06).

About the implementations

- Sampling operator $S_{\mathbf{x}} : \mathcal{H}_K \to \mathbb{R}^m$

$$S_{\mathbf{x}} f = (f(x_1), \ldots, f(x_m))^T .$$

  So for $\mathbf{c} = (c_1, \ldots, c_m)^T \in \mathbb{R}^m$, $S_{\mathbf{x}}^T \mathbf{c} = \sum_{i=1}^m c_i K_{x_i}$.

- Gram matrix $\mathbb{K} = (K(x_i, x_j))_{i,j=1}^m$,

$$\frac{1}{m}\mathbb{K} = \frac{1}{m} S_{\mathbf{x}} S_{\mathbf{x}}^T, \qquad L_K^{\mathbf{x}} = \frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}}.$$

- Let $\frac{1}{m}\mathbb{K} = U \Lambda U^T$ be the eigen decomposition with orthogonal matrix $U = [U_1, \ldots, U_m]$ and eigenvalues $\Lambda = \mathrm{diag}\{\lambda_1^{\mathbf{x}}, \ldots, \lambda_m^{\mathbf{x}}\}$ shared with $L_K^{\mathbf{x}}$. When $\lambda_i^{\mathbf{x}} = 0$, $\phi_i^{\mathbf{x}} \perp \mathrm{span}\{K_{x_i}\}$; when $\lambda_i^{\mathbf{x}} > 0$, one can use

$$\phi_i^{\mathbf{x}} = \frac{1}{\sqrt{m\lambda_i^{\mathbf{x}}}} S_{\mathbf{x}}^T U_i, \qquad U_i = \frac{1}{\sqrt{m\lambda_i^{\mathbf{x}}}} S_{\mathbf{x}} \phi_i^{\mathbf{x}}.$$

LESS

- Let $D = \{(x_i, y_i)\}_{i=1}^m$ be a labeled sample. Write $\mathbf{x} = \{x_i\}$, $\mathbf{y} = (y_1, \ldots, y_m)^T$.
- Let $\mathbf{u} = \{u_i\}_{i=1}^m$ be a separate set of unlabeled data (OK if $\mathbf{x} \subset \mathbf{u}$). As before, we define empirical operator $L_K^{\mathbf{u}}$; eigenvalues and empirical features $\{(\lambda_i^{\mathbf{u}}, \phi_i^{\mathbf{u}})\}_i$; $\|\phi_i^{\mathbf{u}}\|_K \equiv 1$.
- Summary statistics $= (d_1, \ldots, d_N)$ with

$$d_i = \left\langle \phi_i^{\mathbf{u}}, \frac{1}{m} S_{\mathbf{x}}^T \mathbf{y} \right\rangle_K, \quad 1 \le i \le N.$$

- **LESS** (Learning with Empirical feature-based Summary statistics from Semi-supervised data):

$$f_\lambda^{\mathbf{u},D} = (L_K^{\mathbf{u}} + \lambda I)^{-1} \sum_{i=1}^N d_i \phi_i^{\mathbf{u}}.$$

LESS has a natural extension to distributed learning

- define $\mathbf{d}^j = (d_1^j, \ldots, d_N^j)$ by

$$d_i^j = \left\langle \phi_i^{\mathbf{u}}, \frac{1}{|D_j|} \sum_{(x,y) \in D_j} y K_x \right\rangle_K, \qquad 1 \le j \le \ell; \quad 1 \le i \le N.$$

- The average of these statistics is exactly the statistic for LESS with batch learning

$$\mathbf{d} = \sum_{j=1}^{\ell} \frac{|D_j|}{|D|} \mathbf{d}^j.$$

- no upper bound for ℓ.

The convergence of LESS

Assume

- Effective dimension (Zhang, '02): for some $0 < s \leq 1$,

$$\mathcal{N}(\lambda) := \text{Trace}(L_K(L_K + \lambda I)^{-1}) = O(\lambda^{-s}), \quad \text{as } \lambda \to 0^+.$$

- Regularity of the regression function: $f_\rho \in L_K^r(L^2_{\rho_X})$, for some $1/2 \leq r \leq 1$.
- Momentum condition: for $0 < \sigma, M < \infty$ and ρ_X-almost x,

$$\int_Y \left\{ \exp\left\{ \frac{|y - f_\rho(x)|}{M} \right\} - \frac{|y - f_\rho(x)|}{M} - 1 \right\} d\rho(y|x) \leq \frac{\sigma^2}{2M^2}.$$

- N large enough such that $\lambda_{N+1} = O(\lambda)$.

With the above assumptions we have

Theorem (Qin & G, '19)

Assume $|\mathbf{u}| \geq \max\{m, m^{\frac{2}{2r+s}}\}$. *Let* $\lambda = m^{-\frac{1}{2r+s}}$. *One has with confidence at least* $1-\delta$ *that*

$$\left\| f_\lambda^{\mathbf{u},D} - f_\rho \right\|_\rho \leq C m^{-\frac{r}{2r+s}} \log^3 \frac{10}{\delta},$$

which implies that for any $\mu > 0$,

$$\left[\mathbb{E}\left(\left\| f_\lambda^{\mathbf{u},D} - f_\rho \right\|_\rho^\mu \right) \right]^{1/\mu} \leq C \left[10\Gamma(3\mu+1) \right]^{1/\mu} m^{-\frac{r}{2r+s}}.$$

The above convergence rate matches the minimax optimal lower rate for kernel methods.

Theorem (Lin & Zhou, '17)

Assume $|\mathbf{u}_1| = \ldots = |\mathbf{u}_\ell|$ *and* $|D_1| = \ldots = |D_\ell|$. *Let* $|D| = m$ *and* $|D| + |\mathbf{u}| = m^*$. *If*

$$\ell \leq \frac{1}{\log^5 m + 1} \min\left\{ \sqrt{m^*} m^{-\frac{s+1}{4r+2s}}, \sqrt[3]{m^*} m^{\frac{2r+s-2}{6r+3s}} \right\},$$

then with confidence at least $1 - \delta$,

$$\left\| \bar{f}_{D \cup \mathbf{u}, \lambda} - f_\rho \right\|_\rho \leq C m^{-\frac{r}{2r+s}} \log^4 \frac{12}{\delta}.$$

- When $|D_i| = O(1)$, above requirement on ℓ implies $|\mathbf{u}| \gtrsim |D|^{2+\frac{2}{2r+s}}$.
- Meanwhile LESS requires only $|\mathbf{u}| \geq \max\{|D|, |D|^{\frac{2}{2r+s}}\}$.

The privacy benefit of adopting LESS

- Representer theorem [Wahba, '90]:

$$f_{D,\lambda}^{V} = \arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{m}\sum_{i=1}^{m} V(f(x_i),y_i)+\lambda\|f\|_K^2\right\} = \sum_{i=1}^{m} c_i K_{x_i}.$$

- $\{K_{x_i}\}$ (hence $\{x_i\}$) needs to be shipped along with $f_{\mathbf{z},\lambda}^{V}$.
- While LESS ships $f_{\lambda}^{\mathbf{u},D} = (L_K^{\mathbf{u}}+\lambda I)^{-1}\sum_{i=1}^{N} d_i\phi_i^{\mathbf{u}}$, with $\mathbf{u}$ chosen to be free of privacy issues; $d_i = \left\langle \phi_i^{\mathbf{u}}, \frac{1}{m}S_{\mathbf{x}}^T\mathbf{y}\right\rangle_K$, from which it is even difficult to recover m!.
- N can be small when λ_i's decay quickly. For example, λ_i's decay polynomially when K is Sobolev smooth; λ_i's decay exponentially when K is analytic. See (Reade, '84; Little & Reade, '84).

Related Works

- Lin & Zhou, Constr Approx'18; Chang et al., JMLR'17. With the distributed setting $\bar{f}_{D,\lambda} = \sum_{j=1}^{\ell} \frac{|D_j|}{|D|} f_{D_j,\lambda}$, send new instance x_{new} to the sources of data, collect and average $f_{D_j,\lambda}(x_{\text{new}})$. However, this does not protect the private inforamtion in x_{new}.
- Chaudhuri et al., JMLR'11. Random projection based on the Fourier transform of the (translation invariant) kernel. Adding noise during the projection.
- Abadi et al., '16. Adding noise to clipped gradients during the training process.

Papers

- Xin Guo, Ting Hu, and Qiang Wu, Distributed Minimum Error Entropy Algorithms, JMLR, under review
- Huihui Qin and Xin Guo, Semi-supervised learning with summary statistics, Analysis and Applications, to appear

Thank you!