Sparsity and Error Analysis of Empirical Feature-Based Regularization Schemes

Xin Guo

The Hong Kong Polytechnic University

June 13, ICSA2016, Atlanta

Joint work with Jun Fan and Ding-Xuan Zhou

Nonlinear Regression Problems and Kernel Methods

- Consider the nonlinear regression problem with a sample $\{(x_i, y_i)\}_{i=1}^m \overset{\text{iid}}{\sim} \rho$.
- Input space: $X \supset \{x_i\}_{i=1}^m$. For example $X \subset \mathbb{R}^d$, or X is some discrete set: text, amino acid sequences, etc..
- Output space: $Y \subset \mathbb{R}$.
- There is already a large literature in statistics. Learning theory focuses on high-dimensional problems, especially those without a known mechanism. E.g., MHC-peptide binding affinity prediction (Shen et al., FoCM 2015).
- Kernel method (radial basis functions, polynomial/Sobolev/Gaussian kernels, etc.): $K : X \times X \to \mathbb{R}$ symmetric, positive semi-definite (the kernel matrix $(K(x_i, x_j))_{m\times m}$ p.s.d., for any $x_i, \ldots, x_m \in X$). $K_x(u) := K(x, u)$. One uses $\sum_{i=1}^m c_i K_{x_i}$ to approximate f_ρ.

Regularized Least Squares

$$f_\gamma^{\mathbf{z}} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \gamma \|f\|_{\mathcal{H}_K}^2 \right\}$$

- $\mathcal{H}_K$: the reproducing kernel Hilbert space, which is the completion of $\operatorname{spam}\{K_x : x \in X\}$, under the inner product induced by $\langle K_x, K_u \rangle_{\mathcal{H}_K} := K(x,u)$. The parameter $\gamma > 0$ is in practice determined by cross-validation (Cucker&Smale, 2002; Caponnetto, 2006; Caponnetto&Yao, 2010).
- Representer theorem (Wahba, 1990): $f_\gamma^{\mathbf{z}} = \sum_{i=1}^{m} c_i K_{x_i}$.
- The convergence to the regression function f_ρ $(= \int_Y y d\rho(y|\cdot) = \mathbb{E}[Y|\cdot])$: Caponnetto&De Vito, 2007; Bauer et al., 2007; Smale&Zhou,m 2007). Minimax lower bound of learning rate: Yang&Barron, 1999; Bauer et al., 2007; Caponnetto&De Vito, 2007; DeVore et al., 2004; Suzuki et al., 2012; Steinwart et al., 2009.

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: $\lambda_1 \geq \lambda_2 \geq ... \geq 0$. Normalized eigenvectors $\phi_i, \phi_2, ...$, 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: $\lambda_1^{\mathbf{x}} \geq \lambda_2^{\mathbf{x}} \geq ... \geq 0$. Normalized eigenvectors $\phi_1^{\mathbf{x}}, \phi_2^{\mathbf{x}}, ...$ are called empirical features (data dependent features), and are used in kernel principal component analysis (Schölkoph et al., 1998), kernel ridge regression (Cristianini&Shawe-Taylor, 2000; Hastie et al., 2001) kernel projection machine (Blanchard et al., 2004), spectral algorithms (Lo Gerfo et al., 2008; Caponnetto&Yao 2010), and diffusion maps (Coifman&Lafon 2006).

The Regularized Kernel Principal Component Analysis Scheme (RKPCA)

- Output function: $f_\gamma^{\mathbf{z}} = \sum_{i=1}^m c_i^{\mathbf{z}} \phi_i^{\mathbf{x}}$, where $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ and

$$c^{\mathbf{z}} = (c_1^{\mathbf{z}}, \dots, c_m^{\mathbf{z}})$$

$$= \arg\min_{c \in \mathbb{R}^m} \left\{ \frac{1}{m} \sum_{i=1}^m \left(\sum_{j=1}^m c_j \phi_j^{\mathbf{x}}(x_i) - y_i \right)^2 + \gamma \sum_{j=1}^m \Omega(|c_j|) \right\}$$

- The penalty function Ω: $[0, \infty) \to [0, \infty)$ is nonzero, continuous, concave, and $\Omega(0) = 0$.
- The case $\Omega(|c|) = |c|^2$ corresponds to the kernel ridge regression (Cristianini&Shawe-Taylor, 2000; Hastie et al., 2001)
- In general, we say Ω has a concave exponent $q \in [0, 1]$, if there exists some constant C such that $\Omega(c) \leq C c^q$, for any $c \in (0, 1]$.
- The SCAD penalty (Fan&Li, 2001) corresponds to $q = 1$.

RKPCA Decomposed to 1-dim Problems

- Representation of empirical features

$$\phi_i^{\mathbf{x}} = \sum_{j=1}^{m} \frac{(\hat{\mu}_i)_j}{\sqrt{\hat{\lambda}_i^{\mathbf{x}}}} K_{x_i},$$

where $\hat{\lambda}_i^{\mathbf{x}}$ and $\hat{\mu}_i$ are the i'th eigenvalue and the corresponding normalized eigenvector of the kernel Gram matrix $\mathbb{K} = (K(x_r, x_s))_{m \times m}$.

-

$$\frac{1}{m} \sum_{i=1}^{m} \phi_r^{\mathbf{x}}(x_i) \phi_s^{\mathbf{x}}(x_i) = \delta_{rs} \lambda_r^{\mathbf{x}}.$$

RKPCA Decomposed to 1-dim Problems

Theorem (G-Fan-Zhou 2016)

The vector $c^{\mathbf{z}} = (c_1^{\mathbf{z}}, \ldots, c_m^{\mathbf{z}})$ *is a solution of the RKPCA problem if and only if for each* i, $c_i^{\mathbf{z}}$ *is a minimizer of the univariate function*

$$h_i(c) = \lambda_i^{\mathbf{z}}(c - S_i^{\mathbf{z}})^2 + \gamma\Omega(|c|), \quad c \in \mathbb{R},$$

where

$$S_i^{\mathbf{z}} = \begin{cases} \frac{1}{m\lambda_i^{\mathbf{x}}} \sum_{j=1}^m y_j \phi_i^{\mathbf{x}}(x_j), & \text{if } \lambda_i^{\mathbf{x}} > 0, \\ 0 & \text{if } \lambda_i^{\mathbf{x}} = 0. \end{cases}$$

Sparsity and Learning Rates of RKPCA

Theorem (G-Fan-Zhou 2016)

Assume $f_\rho \in L_K^r(\mathcal{H}_K)$ *with* $r > \frac{1}{2}$, *and that* Ω *has a concave exponent* $q \in [0,1]$. *Suppose that for some positive constants* D_1, D_2 *and* α, *the eigenvalues* $\{\lambda_i\}$ *of* L_K *decay polynomially as* $D_1 i^{-\alpha} \leq \lambda_i \leq D_2 i^{-\alpha}$, $\forall i \in \mathbb{N}$ *with* $2\alpha \max\{r,1\} > 1$. *Let* $0 < \delta < 1$. *If we choose*

$$\gamma = C_1 (D_2/\lambda_1)^{r+1} \left(\log \frac{4m}{\delta}\right)^{1+2r} m^{-\frac{1+r}{1+2r}}, \tag{1}$$

then with confidence $1-\delta$ *we have*

$$c_i^{\mathbf{z}} = 0, \qquad \forall\ m^{\theta_{\text{sp}}} + 1 \leq i \leq m \quad \text{with } \theta_{\text{sp}} = \frac{1}{\alpha(1+2r)} < 1 \tag{2}$$

and

$$\|f^{\mathbf{z}} - f_\rho\|_K \leq C_2 \left(\log \frac{4m}{\delta}\right)^{1+2r} m^{-\theta}, \quad \theta = \frac{4\alpha r - 2(2-q)}{4(2r+1)(2-q)\alpha},$$

where C_1 *and* C_2 *are constants independent of* m *or* δ.

- The eigenvalue decay condition above is typical for Sobolev smooth kernels on domains in Euclidean spaces, with α depending on the smoothness of the kernel (Reade, 1984; Sacks&Ylvisaker, 1966, 1968, 1970).
- The above regularity condition requres

$$q > \frac{2}{2\alpha r + 1},$$

and a larger regularity r leads to a wider range of the concave exponent q.

- For $r \gg 1$, the above learning rate $\theta \uparrow \frac{1}{2(2-q)}$.

Theorem (G-Fan-Zhou 2016)

Assume $f_\rho \in L_K^r(\mathcal{H}_K)$ with $r > \frac{1}{2}$, and that Ω has a concave exponent $q \in [0,1]$. Suppose that for some positive constants D_1, D_2 and β, the eigenvalues $\{\lambda_i\}$ of L_K decay exponentially as $D_1\beta^{-i} \le \lambda_i \le D_2\beta^{-i}$, $\forall i \in \mathbb{N}$. Let $0 < \delta < 1$. If we choose γ as (1), then with confidence $1-\delta$ we have

$$c_i^{\mathbf{z}} = 0, \qquad \forall\ \frac{\log(m+1)}{(1+2r)\log\beta} + 1 \le i \le m \tag{3}$$

and

$$\|f^{\mathbf{z}} - f_\rho\|_K \le C_2 \left(\log\frac{4m}{\delta}\right)^{2r+1} m^{-\theta}, \qquad \theta = \frac{r}{(2-q)(1+2r)},$$

where C_2 is a constant independent of m or δ (to be specified in the proof).

Conclusions

- By applying concave penalty to the coefficients of empirical features, sparsity is achieved without sacrificing learning rate very much.
- Our analysis suggests that as the concave exponent q increases to 1, the learning ability of RKPCA improves.
- A more regular regression function leads to a wider range of the concave exponent q.
- Xin Guo, `x.guo@polyu.edu.hk`

Thank You!