Machine Learning Theory

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

July, 2017

Outline of the Course

I. Learning tasks, problems, algorithms, and analysis

II. Learning theory of least squares regression

III. Learning theory of classification

IV. Kernel methods in learning theory and data science

V. Sparsity in machine learning

VI. Online learning and stochastic gradient descent

VII. Distributed learning with big data

VIII. Deep learning with deep neural networks

I. Learning Tasks, Problems, Algorithms, and Analysis

I.1. Regression

Given a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$, find a function $f_{\mathbf{z}}$ predicting an output value $f_{\mathbf{z}}(x) \in Y = \mathbb{R}$ for a new input x

If the input space $X \subseteq \mathbb{R}^n$, then each sample point is a vector $x = (x^1, x^2, \ldots, x^n)$

The ambient space dimension n is large, and X is a manifold of much lower dimension

Data matrix $\mathbf{X} = \begin{bmatrix} x_1^1 & x_1^2 & \ldots & x_1^n \\ x_2^1 & x_2^2 & \ldots & x_2^n \\ \vdots & \vdots & \ldots & \vdots \\ x_m^1 & x_m^2 & \ldots & x_m^n \end{bmatrix} \in \mathbb{R}^{m \times n}$

Problems: least squares regression, quantile regression, geometric quantile, inverse regression, ...

Algorithms: empirical risk minimization, regularization schemes, LASSO, ridge regression, elastic net, ...

I.2. Classification

Find a (binary) classifier $f : X \to \{1, -1\} = Y$

It makes a decision for each case $x = (x^1, \ldots, x^n) \in X \subseteq \mathbb{R}^n$:

$$f(x) = 1 \quad \text{(class 1)} \quad \text{or} \quad f(x) = -1 \quad \text{(class 2)}$$

Less demanding if we take f to be $\text{sgn}(f_{\mathbf{z}})$, the sign of a real-valued function $f_{\mathbf{z}} : X \to \mathbb{R}$

Problems: binary classification, multi-class classification, multi-task learning, ...

Algorithms: support vector machines, linear support vector machines, One-vs-Rest, One-vs-One, decision trees, ...

I.3. Clustering

Grouping a set of objects $\mathbf{x} = \{x_i\}_{i=1}^m$ so that objects in the same group (a cluster) are more similar (in some way) to each other than to those in other groups.

Unsupervised learning

Problems: distance-based clustering, density-based clustering, distribution-based clustering, ...

Algorithms: k-means algorithm, Gaussian mixture models using the expectation-maximization (EM) algorithm, spectral clustering by graph Laplacians, ...

I.4. Ranking

Given sample points $\{x_i\}_{i=1}^m \subset X^m$ with ranking labels $\{y_i\}$ or with preference scores $\{r_{i,j}\}_{i,j=1}^m$, find an ordering function $f : X \times X \to \mathbb{R}$ between objects (inputs) so that $f(x, x') < 0$ if x is preferred over x' (the ranking labels satisfy $y < y'$).

Problems: point-wise ranking with loss $V(y-y'-(f(x)-f(x')))$, pairwise ranking with loss $\phi(\text{sgn}(y-y')(f(x,x'))$, listwise ranking, ...

Algorithms: PageRank, SVM Rank, scoring function-based ranking, pairwise learning, ...

Other related learning tasks: metric learning, similarity learning, AUC maximization, gradient learning, ...

I.5. Deep learning

To face scientific challenges arising from big data:
storage bottleneck, algorithmic scalability, ...

Deep learning Applications
Speech Recognition: from hidden Markov models and Gaussian mixture models to restricted Boltzmann machines and end-to-end deep learning speech recognition systems, phoneme error rate brought down from 26% to 17.7%

Computer Vision: Google Street View

Natural Language Processing

Fundamental operations in deep learning:
distributed implementation, stochastic gradient descent

II. Learning Theory of Least Squares Regression

II.1. Classical Least Squares Method

Given samples $\{(x_i, y_i)\}_{i=1}^m$, find a function f that fits the data: $f(x_i) \approx y_i$

Choose $\mathcal{H} = \Pi_{N-1} := \left\{p(x) = \sum_{j=1}^N c_j x^{j-1} : \quad c_j \in \mathbb{R}\right\}$. We can find an approximation $p^* \in \Pi_{N-1}$ of f by solving the least squares problem:

$$p^* = \arg \min_{p \in \Pi_{N-1}} \frac{1}{m} \sum_{i=1}^m (p(x_i) - y_i)^2.$$

Solution: if we denote $\phi_j(x) = x^{j-1}$ for $j = 1, \ldots, N$, then $p^* = \sum_{j=1}^N c_j^* \phi_j$ where $c^* = (c_j^*)_{j=1}^N \in \mathbb{R}^N$ is a solution of the linear system

$$\mathbf{X}^T \mathbf{X} c = \mathbf{X}^T \mathbf{y},$$

where $\mathbf{X} = [\phi_j(x_i)]_{i=1,\ldots,m, j=1,\ldots,N} \in \mathbb{R}^{m \times N}$ and $\mathbf{y} = (y_i)_{i=1}^m \in \mathbb{R}^m$.

II.2. Framework of the Least Squares Regression

Learn $f : X \to Y$ from random samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$

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

marginal distribution ρ_X on X: $\mathbf{x} = \{x_i\}_{i=1}^m$ drawn according to ρ_X, reflecting data structures

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

Learning the **regression function**: $f_\rho(x) = \int_Y y d\rho(y|x)$

$y_i \approx f_\rho(x_i)$

II.3. Hypothesis Space and Target Function

Generalization Error or **risk** of f is $\mathcal{E}(f) := \int_Z (f(x) - y)^2 d\rho$

Since $\int_{\mathbb{R}} \{f_\rho(x) - y\}\, d\rho(y|x) = 0$, we have

$$\begin{aligned}
\mathcal{E}(f) &= \int_X \int_{\mathbb{R}} (f(x) - y)^2 d\rho(y|x) d\rho_X \\
&= \int_X \int_{\mathbb{R}} (f(x) - f_\rho(x) + f_\rho(x) - y)^2 \rho(y|x) d\rho_X \\
&= \int_X \Big\{ \int_{\mathbb{R}} (f(x) - f_\rho(x))^2 + 2(f(x) - f_\rho(x))(f_\rho(x) - y) \\
&\quad + (f_\rho(x) - y)^2 \rho(y|x) \Big\} d\rho_X \\
&= \int_X (f(x) - f_\rho(x))^2 d\rho_X + \int_Z (f_\rho(x) - y)^2 \rho.
\end{aligned}$$

It follows that f_ρ minimizes $\mathcal{E}(f)$ and

$$\|f - f_\rho\|^2_{L^2_{\rho_X}} = \mathcal{E}(f) - \mathcal{E}(f_\rho) \geq 0. \qquad (1)$$

So approximating the regression function f_ρ in the space $L^2_{\rho_X}$ is equivalent to minimizing the **excess generalization error** or excess risk $\mathcal{E}(f) - \mathcal{E}(f_\rho)$.

Hypothesis space $\mathcal{H}$: a bounded subset of $C(X)$

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

$$f_\mathcal{H} = \arg \inf_{f \in \mathcal{H}} \int_Z (f(x) - y)^2 d\rho = \arg \min_{f \in \mathcal{H}} \|f - f_\rho\|^2_{L^2_{\rho_X}}$$

But ρ is unknown. Hence f_ρ and $f_\mathcal{H}$ cannot be found directly. We need to learn from the sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ through approximating $\mathcal{E}(f)$ by its discretized version, the **empirical generalization error** or empirical risk

$$\mathcal{E}_\mathbf{z}(f) = \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2.$$

II.4. Empirical Risk Minimization

The **empirical target function** $f_{\mathbf{z}}$ is defined by the **empirical risk minimization** (ERM, Vapnik) as

$$f_{\mathbf{z}} := \arg\min_{f \in \mathcal{H}} \mathcal{E}_{\mathbf{z}}(f).$$

Main task for ERM: estimate $\|f_{\mathbf{z}} - f_\rho\|^2_{L^2_{\rho_X}} = \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)$ in terms of properties of $(\rho, \mathcal{H})$ as the sample size m increases.

Example 1. Let $X = \{x \in \mathbb{R}^n : |x| \le 1\}$ and ρ_X be the normalized Lebesgue measure on X. If K is the reproducing kernel of the Sobolev space $H^{s+n/2}(X)$ for some index $s > 0$ and $f_\rho \in H^{\theta(s+n/2)}(X)$ for some $\theta > 0$, then we have with confidence $1 - \delta$,

$$\|f_{\mathbf{z}} - f_\rho\|_{L^2_{\rho_X}} = O\left(m^{-\theta/((2+\theta)(1+2n/s))}\right)$$

by taking $\mathcal{H} = \{f \in H^{s+n/2}(X) : \|f\|_{H^{s+n/2}} \le m^{1/((2+\theta)(1+2n/s))}\}$.

In learning theory we often assume that for some $M > 0$, $|y| \leq M$ almost surely. That is, for almost every $x \in X$, the conditional probability distribution $\rho(\cdot|x)$ is supported on $[-M, M]$.

Error decomposition:

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \leq \underbrace{2 \sup_{f \in \mathcal{H}} |\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)|}_{\textbf{sample error}} + \underbrace{\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho)}_{\textbf{approximation error}} \quad (2)$$

Reason:

$$\begin{aligned} \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) &= \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \{\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}})\} \\ &\quad + \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}) + \mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho) \end{aligned}$$

$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) \leq 0$ by definition, and both $f_{\mathbf{z}}$ and $f_{\mathcal{H}}$ lie in $\mathcal{H}$.

II.5. Sample Error Estimates

Sample error $\sup_{f\in\mathcal{H}} |\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)|$

Law of Large Numbers. For a random variable ξ on (Z,ρ) and a random sample $\{z_i\}_{i=1}^m$, $\frac{1}{m}\sum_{i=1}^m \xi(z_i) \to E(\xi) = \mu$ in prob

Central Limit Theorem. $\Pr\left(\frac{\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mu}{\sqrt{\sigma^2/m}} \le x\right) \approx \Phi(x)$,

where σ^2 is the variance of ξ. So $\frac{c_1}{\sqrt{m}} \le \left|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mu\right| \le \frac{c_2}{\sqrt{m}}$ in prob for any $0 < c_1 < c_2 < \infty$

Hoeffding's inequality. If $|\xi| \le \widetilde{M}$, then

$$\text{Prob}\left\{\left|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mu\right| > \epsilon\right\} \le 2\exp\left\{-\frac{m\epsilon^2}{2\widetilde{M}^2}\right\}, \qquad \forall \epsilon > 0.$$

Hence with confidence $1-\delta$, $\left|\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mu\right| \le \frac{\widetilde{M}\sqrt{2\log(2/\delta)}}{\sqrt{m}}$.

Apply to the random variable $\xi = (f(x) - y)^2$ on (Z, ρ). Then $\frac{1}{m}\sum_{i=1}^{m} \xi(z_i) = \mathcal{E}_{\mathbf{z}}(f)$ and $E(\xi) = \mathcal{E}(f)$. We have $|\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)| \leq \frac{\widetilde{M}\sqrt{2\log(2/\delta)}}{\sqrt{m}}$ if $(f(x) - y)^2 \leq \widetilde{M}$ almost surely.

But $f_{\mathbf{z}}$ changes in the hypothesis space $\mathcal{H}$ with the random sample $\mathbf{z}$. So $\xi = (f_{\mathbf{z}}(x) - y)^2$ is not a single random variable on (Z, ρ).

Theory of uniform convergence: bound the sample error $\sup_{f \in \mathcal{H}} |\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)|$ by various capacity measures of the hypothesis space: VC dimension, covering number, entropy integral, ...

Covering Number $\mathcal{N}(\mathcal{H}, \eta)$: the minimal integer $\ell \in \mathbb{N}$ such that there exist ℓ disks in $C(X)$ with radius $\eta > 0$ covering $\mathcal{H}$ in $C(X)$.

Example 2. If $\mathcal{H}$ is the unit ball of $C^s(X)$ on $X \subset \mathbb{R}^n$, then there are constants $0 < C_1 \leq C_2 < \infty$ such that

$$C_1 \, (1/\eta)^{\frac{n}{s}} \leq \log \mathcal{N}(\mathcal{H}, \eta) \leq C_2 \, (1/\eta)^{\frac{n}{s}} \, , \qquad \forall \eta > 0.$$

Theorem 1 *Let $\mathcal{H}$ be a compact subset of $C(X)$ such that for some $\widetilde{M} > 0$ and every $f \in \mathcal{H}$, there holds $|f(x) - y| \leq \widetilde{M}$ almost surely. Then, for any $\epsilon > 0$,*

$$Prob_{\mathbf{z} \in Z^m} \left\{ \sup_{f \in \mathcal{H}} \left| \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f) \right| \leq \epsilon \right\} \geq 1 - 2\mathcal{N} \left(\mathcal{H}, \frac{\epsilon}{8\widetilde{M}} \right) \exp \left\{ - \frac{m\epsilon^2}{8\widetilde{M}^4} \right\}.$$

Cucker-Smale (2002), ...

Proof. Let $\ell = \mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{8\widetilde{M}}\right)$ and $\{f_j\}_{j=1}^{\ell} \subseteq \mathcal{H}$ be such that for each $f \in \mathcal{H}$, there exists some $j \in \{1, \ldots, \ell\}$ such that $\|f - f_j\|_\infty \leq \frac{\epsilon}{8\widetilde{M}}$, i.e., $f \in B_j$ where $B_j = \{f \in \mathcal{H} : \|f - f_j\|_\infty \leq \frac{\epsilon}{8\widetilde{M}}\}$. It follows that $\mathcal{H} = \cup_{j=1}^{\ell} B_j$ and

$$\begin{aligned}
&\sup_{f \in \mathcal{H}} \left|\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)\right| > \epsilon \\
\Longrightarrow\ &\sup_{f \in B_j} \left|\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)\right| > \epsilon \quad \text{for some } j \in \{1, \ldots, \ell\} \\
\Longrightarrow\ &\left|\mathcal{E}_{\mathbf{z}}(f_j) - \mathcal{E}(f_j)\right| > \frac{\epsilon}{2} \quad \text{for some } j \in \{1, \ldots, \ell\}.
\end{aligned}$$

The last inequality follows from the fact that for $f \in B_j$, $|(f(x) - y)^2 - (f_j(x) - y)^2| = |(f(x) - y) + (f_j(x) - y)||f(x) - f_j(x)| \leq 2\widetilde{M}\frac{\epsilon}{8\widetilde{M}} = \frac{\epsilon}{4}$ almost surely and hence

$$\begin{aligned}
\left|\mathcal{E}_{\mathbf{z}}(f_j) - \mathcal{E}(f_j)\right| &\geq \left|\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)\right| - \left|\mathcal{E}_{\mathbf{z}}(f_j) - \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f_j) + \mathcal{E}(f)\right| \\
&\geq \left|\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)\right| - \frac{\epsilon}{2}.
\end{aligned}$$

Thus, the event $\left\{\mathbf{z} \in Z^m : \sup_{f\in\mathcal{H}}\left|\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)\right| > \epsilon\right\}$ is contained in $\cup_{j=1,\ldots,\ell}\left\{\mathbf{z} \in Z^m : \left|\mathcal{E}_{\mathbf{z}}(f_j) - \mathcal{E}(f_j)\right| > \frac{\epsilon}{2}\right\}$ each of which has measure at most $2\exp\left\{-\frac{m\epsilon^2}{8\widetilde{M}^4}\right\}$ by Hoeffding's inequality. Therefore,

$$\text{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathcal{H}}\left|\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)\right| > \epsilon\right\} \leq \ell 2\exp\left\{-\frac{m\epsilon^2}{8\widetilde{M}^4}\right\}.$$

This proves the desired bound.

Corollary 1 *If there exist some constants $p > 0$ and $C_{\mathcal{H}}$ such that $\log\mathcal{N}(\mathcal{H},\epsilon) \leq C_{\mathcal{H}}\epsilon^{-p}$ for any $\epsilon > 0$, then for any $0 < \delta < 1$, we have with confidence at least $1-\delta$,*

$$\sup_{f\in\mathcal{H}}\left|\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f)\right| \leq 4\widetilde{M}^2\left(1 + 4C_{\mathcal{H}}^{\frac{1}{2+p}}\right)\log\frac{2}{\delta}m^{-\frac{1}{2+p}}.$$

II.6. Approximation Error Estimates

$$\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho) = \|f_{\mathcal{H}} - f_\rho\|_2^2 = \inf_{f \in \mathcal{H}} \int (f(x) - f_\rho(x))^2$$

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

Theorem 2 *(Smale-Zhou, 2003) 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, then*

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

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

II.7. Uniform Law of Large Numbers: Theory of Uniform Convergence

to bound $\sup_{f\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m} f(x_i) - \int_X f(x)d\rho\right|$.

Given a bounded set $\mathcal{H}$ of functions on X, when do we have

$$\lim_{\ell\to\infty} \sup_{\rho} \text{Prob}\left\{\sup_{m\geq\ell}\sup_{f\in\mathcal{H}}\left|\frac{1}{m}\sum_{i=1}^{m} f(x_i) - \int_X f(x)d\rho\right| > \epsilon\right\} = 0,\ \forall\epsilon > 0?$$

Characterization. Finiteness of VC dimension for indicator functions: Vapnik-Chervonenkis (1971)

Finiteness of V_γ dimension for general functions: Alon, Ben-David, Cesa-Bianchi, Haussler (1997)

II.8. More Probability Inequalities

Classical Markov inequality: If $\xi \geq 0$ and $t > 0$, then

$$\text{Prob}\{\xi \geq t\} \leq \frac{\mu}{t}.$$

Classical Chebyshev's inequality: Applying the Markov inequality to $(\xi - \mu)^2$ yields

$$\text{Prob}\{|\xi - \mu| \geq t\} = \text{Prob}\{(\xi - \mu)^2 \geq t^2\} \leq \frac{\sigma^2}{t^2}.$$

In particular, we have

$$\text{Prob}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\right| \geq \epsilon\right\} \leq \frac{\sigma^2}{m\epsilon^2}, \qquad \forall \epsilon > 0.$$

Hence with confidence $1 - \delta$, $\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\right| \leq \frac{\sqrt{\sigma^2}}{\sqrt{m\delta}}$. This is weak in terms of the dependence on the confidence δ.

Hoeffding's inequality. If $|\xi| \leq \widetilde{M}$, then

$$\text{Prob}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\right| \geq \epsilon\right\} \leq 2\exp\left\{-\frac{m\epsilon^2}{2\widetilde{M}^2}\right\}, \qquad \forall \epsilon > 0.$$

Hence with confidence $1-\delta$, $\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\right| \leq \frac{\widetilde{M}\sqrt{2\log(2/\delta)}}{\sqrt{m}}$.

Bernstein's inequality. If $|\xi - \mu| \leq \widetilde{M}$, then

$$\text{Prob}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\right| \geq \epsilon\right\} \leq 2\exp\left\{-\frac{m\epsilon^2}{2\left(\sigma^2 + \frac{1}{3}\widetilde{M}\epsilon\right)}\right\}, \qquad \forall \epsilon > 0.$$

Hence with confidence $1-\delta$,

$$\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\right| \leq \frac{\sqrt{2\sigma^2\log(2/\delta)}}{\sqrt{m}} + \frac{\frac{2}{3}\widetilde{M}}{m}.$$

Improvement to optimal rates in the theory of uniform convergence when $\sigma^2 \leq c\mu$ uniformly.

Bernstein's ratio probability inequality. If $\mu \geq 0$, $|\xi - \mu| \leq B$ almost surely, and $\sigma^2 \leq c\mu$ for some $c > 0$, then for every $\epsilon > 0$ and $0 < \alpha \leq 1$,

$$\text{Prob}\left\{\frac{\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\right|}{\sqrt{\mu+\epsilon}} > \alpha\sqrt{\epsilon}\right\} \leq 2\left\{-\frac{\alpha^2 m\epsilon}{2c + \frac{2}{3}B}\right\}.$$

This follows from Bernstein's inequality by taking $\widetilde{\epsilon} = \sqrt{\mu+\epsilon}\alpha\sqrt{\epsilon}$ and noticing $\sigma^2 + \frac{1}{3}\widetilde{M}\widetilde{\epsilon} \leq \left(c + \frac{1}{3}B\right)(\mu+\epsilon)$.

Uniform ratio inequality. Let $\mathcal{G}$ be a set of functions on Z and $c > 0$ such that, for each $g \in \mathcal{G}$, $\mathbf{E}(g) = \int_Z g(z)d\rho \geq 0$, $\mathbf{E}(g^2) \leq c\mathbf{E}(g)$, and $|g - \mathbf{E}(g)| \leq B$ almost surely. Then for every $\epsilon > 0$ and $0 < \alpha \leq 1$, with $\mathbf{E_z}(g) = \frac{1}{m}\sum_{i=1}^{m} g(z_i)$, we have

$$\text{Prob}\left\{\sup_{g\in\mathcal{G}}\frac{\mathbf{E}(g) - \mathbf{E_z}(g)}{\sqrt{\mathbf{E}(g)+\epsilon}} > 4\alpha\sqrt{\epsilon}\right\} \leq \mathcal{N}(\mathcal{G}, \alpha\epsilon)\exp\left\{-\frac{\alpha^2 m\epsilon}{2c + \frac{2}{3}B}\right\}.$$

Example 3. Let $\mathcal{H}$ be a compact and **convex** subset of $C(X)$. If for some $M > 0$ and every $f \in \mathcal{H}$, we have $|f(x) - y| \le M$, then for every ϵ, we have

$$\text{Prob}_{\mathbf{z}\in Z^m}\{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \ge \epsilon\} \le \mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{12M}\right) \exp\left\{-\frac{m\epsilon}{300M^2}\right\}.$$

Proof. Consider the function set

$$\mathcal{G} = \left\{(f(x) - y)^2 - (f_{\mathcal{H}}(x) - y)^2 : f \in \mathcal{H}\right\}.$$

Each function g in $\mathcal{G}$ satisfies $\mathbf{E}(g) = \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}) \ge 0$, $-M^2 \le g(z) \le M^2$, $|g - \mathbf{E}(g)| \le B := 2M^2$, and $|g(z)| \le 2M|f(x) - f_{\mathcal{H}}(x)|$ which implies $\mathbf{E}(g^2) \le 4M^2 \int_X (f - f_{\mathcal{H}})^2$.

A crucial inequality (see Cucker-Smale, or Lemma 3.16 in the book of Cucker-Zhou) caused by the convexity of $\mathcal{H}$:

$$\int_X |f(x) - f_{\mathcal{H}}(x)|^2 \, d\rho_X \le \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}).$$

So $\mathbf{E}(g^2) \leq 4M^2 (\mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}})) = c\mathbf{E}(g)$ with $c = 4M^2$. Thus by the uniform ratio inequality, from the identity $\mathbf{E}_{\mathbf{z}}(g) = \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}})$, for $\epsilon > 0$ and $0 < \alpha \leq 1$, with probability

$$1 - \mathcal{N}(\mathcal{G}, \alpha\epsilon) \exp\left\{-\frac{\alpha^2 m\epsilon}{8M^2 + \frac{2}{3}2M^2}\right\},$$

there holds

$$\sup_{f \in \mathcal{H}} \frac{(\mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}})) - (\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}))}{\sqrt{(\mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}})) + \epsilon}} \leq 4\alpha\sqrt{\epsilon}.$$

Taking $\alpha = \sqrt{2}/8$ and $f = f_{\mathbf{z}}$, and noting that $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) \leq 0$ by definition, we have

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \leq 4\alpha\sqrt{\epsilon}\sqrt{(\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}})) + \epsilon}.$$

Solving the quadratic equation, we have $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}) \leq \epsilon$. The desired inequality now follows by the covering number bound $\mathcal{N}(\mathcal{G}, \alpha\epsilon) \leq \mathcal{N}\left(\mathcal{H}, \frac{\alpha\epsilon}{2M}\right)$ seen from the inequality $\|g_1 - g_2\|_\infty \leq \|(f_1(x) - f_2(x))\,[(f_1(x) - y) + (f_2(x) - y)]\,\|_\infty \leq 2M\|f_1 - f_2\|_\infty$.

Concentration inequality: Let $\mathcal{F}$ be a set of measurable functions on Z, and $B, c > 0, \theta \in [0,1]$ be constants such that each function $f \in \mathcal{F}$ satisfies $\|f\|_\infty \le B$ and $E(f^2) \le c(Ef)^\theta$. If for some $a > 0$ and $p \in (0,2)$,

$$\sup_{m\in\mathbb{N}} \sup_{\mathbf{z}\in Z^m} \log \mathcal{N}_{2,\mathbf{z}}(\mathcal{F},\epsilon) \le a\epsilon^{-p}, \qquad \forall \epsilon > 0,$$

then there exists a constant c_p' depending only on p such that for any $t > 0$, with probability at least $1 - e^{-t}$, there holds

$$E[f] - \frac{1}{m}\sum_{i=1}^m f(z_i) \le \frac{1}{2}\eta^{1-\theta}(Ef)^\theta + c_p'\eta + 2\left(\frac{ct}{m}\right)^{\frac{1}{2-\theta}} + \frac{18Bt}{m}, \quad \forall f \in \mathcal{F},$$

where $\eta := \max\left\{ c^{\frac{2-p}{4-2\theta+p\theta}}\left(\frac{a}{m}\right)^{\frac{2}{4-2\theta+p\theta}},\ B^{\frac{2-p}{2+p}}\left(\frac{a}{m}\right)^{\frac{2}{2+p}} \right\}$.

Here $\mathcal{N}_{2,\mathbf{z}}(\mathcal{F},\epsilon)$ is the empirical covering number defined with the empirical metric $d_{2,\mathbf{z}}(f,g) = \left\{\frac{1}{m}\sum_{i=1}^m |f(z_i) - g(z_i)|^2\right\}^{\frac{1}{2}}$ as

$$\mathcal{N}_{2,\mathbf{z}}(\mathcal{F},\epsilon) = \inf\left\{\ell \in \mathbb{N} : \exists \{f_i\}_{i=1}^\ell \subset \mathcal{F}, \forall f \in \mathcal{F}, \min_i d_{2,\mathbf{z}}(f,f_i) \le \epsilon\right\}.$$

III. Learning Theory of Classification

III.1. Reproducing Kernel Hilbert Spaces

Mercer kernel $K : X \times X \to \mathbb{R}$ is a continuous and symmetric function such that the matrix $(K(x_i, x_j))_{i,j=1}^{\ell}$ is positive semidefinite for any $\ell \in \mathbb{N}$ and $(x_i)_{i=1}^{\ell} \subset X$.

Examples of Mercer kernels on $X \subset \mathbb{R}^n$:

polynomial kernels $K(x, y) = (x \cdot y)^d$ or $(1 + x \cdot y)^d$: Vapnik

Gaussian kernels $K(x, y) = e^{-\frac{|x-y|^2}{\sigma^2}}$ with $\sigma > 0$

spline kernels: Wahba, ...

radial basis functions $K(x, y) = \int_0^{+\infty} e^{-\rho|x-y|^2} d\beta(\rho)$: Schoenberg, Micchelli, Schaback, Wu, Wendland, ...

Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_K$: completion of the span of the set of functions $\{K_t = K(t,\cdot) : t \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ satisfying

$$\langle K_x, K_y \rangle_K = K(x,y), \qquad \forall x, y \in X.$$

Then the norm of the function $f(x) = \sum_{i=1}^{\ell} c_i K_{x_i}$ satisfies

$$\langle \sum_i^{\ell} c_i K_{x_i}, \sum_i^{\ell} c_i K_{x_i} \rangle_K = \| \sum_i^{\ell} c_i K_{x_i} \|_K^2 = \sum_{i,j=1}^{\ell} c_i K(x_i, x_j) c_j \geq 0.$$

$\mathcal{H}_K$ is a subspace of $C(X)$.

Reproducing property of $\mathcal{H}_K$:

$$\langle f, K_x \rangle_K = f(x), \qquad \forall x \in X, f \in \mathcal{H}_K. \tag{3}$$

Embedding (Zhou, 2008): If $X \subset \mathbb{R}^n$ and $K \in C^{2s}(X \times X)$ for some $s > 0$, then the inclusion $J : \mathcal{H}_K \hookrightarrow C^s(X)$ is well-defined and bounded

III.2. Regularization Schemes

Let $V : \mathbb{R}^2 \to \mathbb{R}_+$ be a loss function.

Example 4. Examples of loss functions:

(1) least-square loss $V(f, y) = (f - y)^2$

(2) hinge loss $V(f, y) = (1 - yf)_+ := \max\{0, 1 - yf\}$

(3) ϵ-insensitive loss for support vector regression $V(f, y) = |f - y|_\epsilon := \max\{0, |f - y| - \epsilon\}$ where $\epsilon > 0$. Logistic loss $V(f, y) = \log(1 + e^{-yf})$. Exponential loss $V(f, y) = e^{-yf}$

The **regularization scheme** associated with the loss V and the RKHS $\mathcal{H}_K$ is defined (Evgeniou-Pontil-Poggio, 2000) by

$$f_{\mathbf{z},\lambda} := \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^{m} V(f(x_i), y_i) + \lambda \|f\|_K^2 \right\}, \qquad (4)$$

where $\lambda = \lambda(m) > 0$ is a regularization parameter.

III.3. Representer Theorem

Kimeldorf-Wahba, Schölkopf-Herbrich-Smola, ...

Theorem 3 *If* $\lambda > 0$, *then* $f_{\mathbf{z},\lambda} = \sum_{i=1}^m c_i K(x_i, x)$ *for some* $(c_i)_{i=1}^m \in \mathbb{R}^m$.

Proof. Let $\mathcal{H}_{K,\mathbf{x}} = \text{span}\{K_{x_i}\}_{i=1}^m$ and P be the orthogonal projection from $\mathcal{H}_K$ to $\mathcal{H}_{K,\mathbf{x}}$. Then (3) tells us that for $f \in \mathcal{H}_K$,

$$f(x_i) = \langle f, K_{x_i}\rangle_K = \langle P(f), K_{x_i}\rangle_K = P(f)(x_i), \qquad \forall i = 1, \ldots, m.$$

If we denote $\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^m V(f(x_i), y_i)$, then $\mathcal{E}_{\mathbf{z}}(f) = \mathcal{E}_{\mathbf{z}}(P(f))$. Thus, $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) = \mathcal{E}_{\mathbf{z}}(P(f_{\mathbf{z},\lambda}))$ while $\lambda\|P(f_{\mathbf{z},\lambda})\|_K^2 < \lambda\|f_{\mathbf{z},\lambda}\|_K^2$ unless $f_{\mathbf{z},\lambda} \in \text{span}\{K_{x_i}\}_{i=1}^m$. But $f_{\mathbf{z},\lambda}$ minimizes $\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|_K^2$, so it must lies in $\text{span}\{K_{x_i}\}_{i=1}^m$. This proves the theorem.

Special case: if $V(f, y) = (f - y)^2$ is the least-square loss, then $(c_i)_{i=1}^m$ is given by a linear system

$$\left(\left[K(x_i, x_j)\right]_{i,j=1}^m + m\lambda I\right)(c_j)_{j=1}^m = (y_i)_{i=1}^m.$$

III.4. Binary Classification

Find the sign of f_ρ only!

A (binary) **classifier** is a function $f : X \to \{1, -1\} = Y$

It makes a decision for each case $x = (x^1, \ldots, x^n) \in X \subseteq \mathbb{R}^n$:

$$f(x) = 1 \quad \text{(YES)} \qquad \text{or} \qquad f(x) = -1 \quad \text{(NO)}$$

It ρ is a probability measure on $Z = X \times Y$, then the conditional distribution at $x \in X$ is

$$\begin{cases} P(y = 1|x) \text{ is the probability for the output to be 1 (YES)} \\ P(y = -1|x) \text{ is the probability for the output to be } -1 \text{ (NO).} \end{cases}$$

Best classifier (Bayes rule) f_c:

$$f_c(x) = \begin{cases} 1, & \text{if } \rho(y=1|x) \geq \rho(y=-1|x) \\ -1, & \text{if } \rho(y=1|x) < \rho(y=-1|x). \end{cases}$$

Misclassification error:

$$\mathcal{R}(f) = \text{Prob}\{y \neq f(x)\} = \text{Prob}\{yf(x) = -1\} = \int_Z \chi_{\{yf(x)=-1\}} d\rho$$

We have $\mathcal{R}(f) \geq \mathcal{R}(f_c)$ for any $f : X \to Y$.

The purpose of classification algorithms is to find good approximations $f_{\mathbf{z}}$ of f_c from training data $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \subset Z^m$ drawn according to ρ

Note that f_c is only the sign of the regression function f_ρ: $f_c = \text{sgn}(f_\rho)$

III.5. Support Vector Machine

If we choose the hypothesis space to be signs of functions from $\mathcal{H}_K$, then

$$f^0_{\mathbf{z},\lambda} := \arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{m}\sum_{i=1}^m \phi_0(y_i f(x_i)) + \lambda\|f\|_K^2\right\},$$

where $\phi_0(yf) = \chi_{\{yf<0\}}$ is the $0-1$ loss. By the representer theorem, $f^0_{\mathbf{z},\lambda} = \sum_{i=1}^m c_i^0 K_{x_i}$ where

$$c^0 := \arg\min_{c\in\mathbb{R}^m}\left\{\frac{1}{m}\sum_{i=1}^m \chi_{\{y_i\sum_{j=1}^m c_j K(x_i,x_j)<0\}} + \lambda\sum_{i,j=1}^m c_i K(x_i,x_j)c_j\right\}.$$

Nonconvex optimization problem in $\mathbb{R}^m$!

Support Vector Machine (SVM): $\mathrm{sgn}(f_{\mathbf{z},\lambda})$ with

$$f_{\mathbf{z},\lambda} = \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{m}\sum_{i=1}^{m} \phi(y_i f(x_i)) + \lambda \|f\|_K^2 \right\}. \qquad (5)$$

Requirement for the loss function ϕ: close to ϕ_0 and convex.

Hinge loss: $\phi(yf) = \max\{1-yf, 0\}$.

By the representer theorem, $f_{\mathbf{z},\lambda} = \sum_{i=1}^{m} c_i^{\mathbf{z}} K_{x_i}$ with $c^{\mathbf{z}} \in \mathbb{R}^m$.

ϕ is **convex** $\Longrightarrow$ the optimization problem for $c^{\mathbf{z}}$ is convex!

ϕ is **piecewise linear** $\Longrightarrow$ **Convex quadratic programming**

$$\begin{aligned} \arg\min_{c,\xi\in\mathbb{R}^m} \quad & \tfrac{1}{m}\textstyle\sum_{i=1}^{m} \xi_i + \lambda \sum_{i,j=1}^{m} c_i K(x_i,x_j) c_j, \\ \text{subj. to} \quad & \xi_i \geq 1 - y_i \textstyle\sum_{j=1}^{m} K(x_i,x_j) c_j, \\ & \xi_i \geq 0, \qquad i = 1,\ldots,m. \end{aligned}$$

Rosenblatt (1962): perceptron $K(x,y) = 1 + x \cdot y \quad \mathcal{H}_K = \Pi_1$

Boser-Guyon-Vapnik (1992): $K(x,y) = (1 + x \cdot y)^d \quad \mathcal{H}_K = \Pi_d$

Cortes-Vapnik (1995): general Mercer kernels

Error bound for separable distributions: Vapnik, Cristianini-Shawe-Taylor, ...

Example 5. Assume that ρ is strictly separable by $\mathcal{H}_K$ with margin $\Delta > 0$, meaning that there exists a function $f_{sp} \in \mathcal{H}_K$ such that $\|f_{sp}\|_K = 1$ and $y f_{sp}(x) \geq \Delta$ almost surely. If for some constants $C_0, p > 0$, the covering number of the unit ball B_1 of $\mathcal{H}_K$ satisfies $\log \mathcal{N}(B_1, \eta) \leq C_0 \eta^{-p}$ for every $\eta > 0$, then for any $0 < \delta < 1$, with confidence $1 - \delta$, we have

$$\mathcal{R}(f_{\mathbf{z},\lambda}) \leq \frac{2\lambda}{\Delta} + \widetilde{C}(1 + \frac{1}{\Delta})^2 \log \frac{2}{\delta} m^{-\frac{1}{1+p}}$$

where $\widetilde{C}$ is a positive constant independent of m, Δ or δ.

III.6. Comparison Theorems

T. Zhang, Chen-Wu-Ying-Zhou, Bartlett-Jordan-McAuliffe, ...

Theorem 4 *For any measurable function* $f : X \to \mathbb{R}$,

$$\mathcal{R}(sgn(f)) - \mathcal{R}(f_c) \leq \mathcal{E}(f) - \mathcal{E}(f_c).$$

Theorem 5 *Let* $\phi : \mathbb{R} \to \mathbb{R}_+$ *be convex such that* $\phi'(0) < 0, \phi''(0) > 0$. *Then* $\mathcal{R}(sgn(f)) - \mathcal{R}(f_c) \leq c_\phi \sqrt{\mathcal{E}(f) - \mathcal{E}(f_\rho^\phi)}$, *where* $\mathcal{E}(f) = \int_Z \phi(yf(x))d\rho$ *and* f_ρ^ϕ *is a minimizer of* $\mathcal{E}$*:*

$$f_\rho^\phi(x) = \arg\min_{t\in\mathbb{R}} \{\phi(t)P(y=1|x) + \phi(-t)P(y=-1|x)\}, \quad x \in X.$$

Improvements are possible when the Tsybakov noise condition is satisfies: for some $0 < q \leq \infty$ and $C_q > 0$ there holds

$$\rho_X(\{x \in X : |f_\rho(x)| \leq C_q t\}) \leq t^q, \qquad \forall t > 0.$$

III.7. Error Analysis

Error decomposition (Wu-Ying-Zhou):

$$\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}(f_\rho^\phi) \le \left\{\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})\right\} + \{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)\} + \mathcal{D}(\lambda),$$

where $\mathcal{D}(\lambda)$ is the regularization error defined by

$$\mathcal{D}(\lambda) = \inf_{f \in \mathcal{H}_K} \left\{\mathcal{E}(f) - \mathcal{E}(f_\rho^\phi) + \lambda \|f_\lambda\|_K^2\right\}$$

and f_λ is a minimizer. $\mathcal{D}(\lambda)$ can be estimated from regularity of f_ρ^ϕ with respect to $(\mathcal{H}_K, L^2_{\rho_X})$.

Since both $f_{\mathbf{z},\lambda}$ and f_λ lie in $B_R = \{f \in \mathcal{H}_K : \|f\|_K \le R\}$ with $R = \sqrt{\mathcal{E}(0)/\lambda}$, we know that

$$\left\{\mathcal{E}(f_{\mathbf{z},\lambda}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})\right\} + \{\mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda)\} \le 2 \sup_{f \in B_R} |\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)|.$$

Then the theory of uniform convergence yields error bounds.

IV. Kernel Methods in Learning Theory and Data Science

IV.1. Feature Maps and Dimensionality Reduction

Let X be a subset of $\mathbb{R}^n$ with metric d and probability measure ρ_X. Each $x = (x^1, x^2, \ldots, x^n) \in X$ contains n measurements.

Hypothesis: X is a manifold of dimension J with $J << n$

Purpose of dimensionality reduction: represent points on (X, d, ρ_X) by images in $\mathbb{R}^J$ under a (isometric) map $\mathcal{F} : X \to \mathbb{R}^J$

$$x = (x^1, \ldots, x^n) \quad \to \quad \mathcal{F}(x) = (y^1, \ldots, y^J) = (\psi_1(x), \ldots, \psi_J(x))$$

Let L_K be the integral operator on $L^2_{\rho_X}$ defined by

$$L_K f(x) = \int_X K(x,y) f(y) d\rho_X(y), \qquad x \in X, f \in L^2_{\rho_X}.$$

Let $\{(\lambda_i, \phi_i)\}_{i=1}^{\infty}$ be normalized eigenpairs of L_K with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$ and $\{\phi_i\}$ an orthonormal basis of $L^2_{\rho_X}$.

The feature map $\Phi : X \to \ell^2$ is defined by $\Phi(x) = \begin{bmatrix} \sqrt{\lambda_1}\phi_1(x) \\ \sqrt{\lambda_2}\phi_2(x) \\ \vdots \end{bmatrix}$.

Cristianini-Shawe-Taylor, ...
Understand the dimensionality reduction theoretically:

Take some $\epsilon > 0$ and $J = \#\{i : \lambda_i \geq \epsilon\}$. Then $\mathcal{F} : X \to \ell^2(\mathbb{R}^J)$ is given by $\mathcal{F}(x) = \begin{bmatrix} \sqrt{\lambda_1}\phi_1(x) \\ \sqrt{\lambda_2}\phi_2(x) \\ \vdots \\ \sqrt{\lambda_J}\phi_J(x) \end{bmatrix}$. Then we have $x \in X \subset \mathbb{R}^n \longrightarrow \mathcal{F}(x) \in \mathbb{R}^J$, and $f_\rho(x) \approx \widetilde{f}_\rho(\mathcal{F}(x))$.

Theoretical background

If the metric $d = d_K$ is induced by a Mercer kernel K on X as

$$d_K(x, y) = \|K_x - K_y\|_K = \sqrt{K(x, x) + K(y, y) - 2K(x, y)}$$

By Mercer Theorem,

$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(y) = \langle \Phi(x), \Phi(y) \rangle_{\ell^2}.$$

So the feature map Φ is isometric:

$$\|\Phi(x) - \Phi(y)\|_{\ell^2} = d_K(x, y), \qquad x, y \in X.$$

IV.2. Laplacian Eigenmaps

Let $\mathbf{x} := \{x_i\}_{i=1}^m$ be an i.i.d. sample and $K : X \times X \to \mathbb{R}_+$ (e.g. Gaussian kernel).

Let $K_{\mathbf{x}} = \Big(K(x_i, x_j)\Big)_{i,j=1}^m$ and $D_{\mathbf{x}}$ be the diagonal matrix with main diagonal entries $\sum_{j=1}^m K(x_i, x_j)$. Define a discrete Laplacian matrix $L = L_{\mathbf{x}} = D_{\mathbf{x}} - K_{\mathbf{x}}$. The normalized discrete Laplacian is defined by

$$\widehat{\Delta}_{K,\mathbf{x}} = I - D_{\mathbf{x}}^{-\frac{1}{2}} K_{\mathbf{x}} D_{\mathbf{x}}^{-\frac{1}{2}} =: I - \frac{1}{m}\widehat{K}_{\mathbf{x}}. \tag{6}$$

Laplacian eigenmap algorithm: If $\{\widehat{\lambda}_k, \widehat{v}^{(k)}\}_{k=1}^m$ are eigenpairs of $\widehat{\Delta}_{K,\mathbf{x}}$ with eigenvalues $\widehat{\lambda}_1 = 0 \leq \widehat{\lambda}_2 \leq \ldots \leq \widehat{\lambda}_m$ and $1 \leq J \leq m$, then the data representation is given by

$$x_i \quad \longrightarrow \quad \begin{bmatrix} v_i^{(2)} \\ v_i^{(3)} \\ \vdots \\ v_i^{(J)} \end{bmatrix}. \tag{7}$$

IV.3. Spectral Clustering by Graph Laplacians

Unnormalized Graph Laplacians: $L = L_{\mathbf{x}} = D_{\mathbf{x}} - K_{\mathbf{x}}$

Normalized Graph Laplacians: $\widehat{\Delta}_{K,\mathbf{x}} = I - D_{\mathbf{x}}^{-\frac{1}{2}} K_{\mathbf{x}} D_{\mathbf{x}}^{-\frac{1}{2}}$ or $I - D_{\mathbf{x}}^{-1} K_{\mathbf{x}}$

Spectral clustering algorithm: Let $v \in \mathbb{R}^m$ be an eigenvector of $D_{\mathbf{x}} - K_{\mathbf{x}}$ or $I - D_{\mathbf{x}}^{-1} K_{\mathbf{x}}$ associated with the second smallest eigenvalue. Then the output clusters are given by

$$A = \{j : v_j \geq 0\}, \qquad \widetilde{A} = \{j : v_j < 0\}.$$

Note that

$$\left(I - D_{\mathbf{x}}^{-1} K_{\mathbf{x}}\right) v = \lambda v \quad \Leftrightarrow \quad \left(I - D_{\mathbf{x}}^{-\frac{1}{2}} K_{\mathbf{x}} D_{\mathbf{x}}^{-\frac{1}{2}}\right) \left(D_{\mathbf{x}}^{\frac{1}{2}} v\right) = \lambda \left(D_{\mathbf{x}}^{\frac{1}{2}} v\right).$$

Observe that for $\mathbf{y} \in \mathbb{R}^m$, we have

$$\sum_{i,j=1}^{m} \left(y_i - y_j\right)^2 K(x_i, x_j) = \sum_{i,j=1}^{m} \left(y_i^2 + y_j^2 - 2y_i y_j\right) K(x_i, x_j)$$

$$= \sum_{i=1}^{m} y_i^2 (D_{\mathbf{x}})_{i,i} + \sum_{j=1}^{m} y_j^2 (D_{\mathbf{x}})_{j,j} + 2 \sum_{i,j=1}^{m} y_i y_j K(x_i, x_j)$$

$$= 2\mathbf{y}^T (D_{\mathbf{x}})\, \mathbf{y} - 2\mathbf{y}^T (K_{\mathbf{x}})\, \mathbf{y} = \left(D_{\mathbf{x}}^{\frac{1}{2}}\mathbf{y}\right)^T \widehat{\Delta}_{K,\mathbf{x}} \left(D_{\mathbf{x}}^{\frac{1}{2}}\mathbf{y}\right) \geq 0.$$

So the matrix $\widehat{\Delta}_{K,\mathbf{x}}$ is positive semidefinite. Moreover, by taking $y_i \equiv 1$, we see that 0 is always an eigenvalue with eigenvector $D_{\mathbf{x}}^{\frac{1}{2}}(1, \ldots, 1)^T$. It is not included in the Laplacian eigenmap or graph Laplacians.

Belkin-Niyogi, ...

IV.4. Convergence Analysis

We assume that K is a Mercer kernel and that $p = \int_X K_x d\rho_X \in \mathcal{H}_K$ is positive on X (which is true if K is positive). Let

$$\widetilde{K}(x,y) = \frac{K(x,y)}{\sqrt{p(x)}\sqrt{p(y)}}, \qquad x, y \in X. \tag{8}$$

Theorem 6 *(Smale-Zhou, 2009) Let $k \in \{1, \ldots, m\}$ such that $r := \min\{\lambda_{k-1} - \lambda_k, \lambda_k - \lambda_{k+1}\} > 0$. If $0 < \delta < 1$ and $m \in \mathbb{N}$ satisfy $\frac{4\kappa^2 \log\left(4/\delta\right)}{\sqrt{m} p_0} \le \frac{r}{8}$, then with confidence $1 - \delta$, we have*

$$\left\| \sqrt{\lambda_k}\phi_k - \frac{1}{\sqrt{m\lambda_k}} \sum_{i=1}^{m} v_i^{(k)} \widetilde{K}_{x_i} \right\|_{\widetilde{K}} \le \frac{72\kappa^2 \log\left(4/\delta\right)}{r\sqrt{m} p_0 \sqrt{\lambda_k}}, \tag{9}$$

where $p_0 := \min_{x \in X} p(x)$ and $v^{(k)} \in \ell^2(\mathbf{x})$ is a normalized eigenvector of the matrix $\frac{1}{m}\widehat{K}_{\mathbf{x}}$ with the k-th eigenvalue.

From the theorem, we know that the theoretical truncated feature map can be learned well by the Laplacian eigenmap

$$\mathcal{F}(x) = \begin{bmatrix} \sqrt{\lambda_2}\phi_2(x) \\ \sqrt{\lambda_3}\phi_3(x) \\ \vdots \\ \sqrt{\lambda_J}\phi_J(x) \end{bmatrix} \quad \Longleftrightarrow \quad x_i \quad \longrightarrow \quad \begin{bmatrix} v_i^{(2)} \\ v_i^{(3)} \\ \vdots \\ v_i^{(J)} \end{bmatrix}.$$

Off-sample algorithm: In general, denote $p_{\mathbf{x}} = \frac{1}{m}\sum_{j=1}^{m} K_{x_j}$, we represent points $x \in X \subset \mathbb{R}^n$ with $\widehat{\mathcal{J}}(x) \in \mathbb{R}^J$ given by

$$\begin{bmatrix} \frac{1}{\sqrt{m\widehat{\lambda}_1}}\left(v_1^{(1)}\frac{K(x,x_1)}{\sqrt{p_{\mathbf{x}}(x)}\sqrt{p_{\mathbf{x}}(x_1)}} + \cdots + v_m^{(1)}\frac{K(x,x_m)}{\sqrt{p_{\mathbf{x}}(x)}\sqrt{p_{\mathbf{x}}(x_m)}}\right) \\ \frac{1}{\sqrt{m\widehat{\lambda}_2}}\left(v_1^{(2)}\frac{K(x,x_1)}{\sqrt{p_{\mathbf{x}}(x)}\sqrt{p_{\mathbf{x}}(x_1)}} + \cdots + v_m^{(2)}\frac{K(x,x_m)}{\sqrt{p_{\mathbf{x}}(x)}\sqrt{p_{\mathbf{x}}(x_m)}}\right) \\ \vdots \\ \frac{1}{\sqrt{m\widehat{\lambda}_J}}\left(v_1^{(J)}\frac{K(x,x_1)}{\sqrt{p_{\mathbf{x}}(x)}\sqrt{p_{\mathbf{x}}(x_1)}} + \cdots + v_m^{(J)}\frac{K(x,x_m)}{\sqrt{p_{\mathbf{x}}(x)}\sqrt{p_{\mathbf{x}}(x_m)}}\right) \end{bmatrix}. \quad (10)$$

IV.5. Semi-supervised Learning on Manifolds

If we have unlabelled data $\mathbf{x}^u = \{x_i\}_{i=m+1}^{m+u}$ in addition to the labelled data $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$, a manifold regularization for semi-supervised learning outputs the function

$$f_{\mathbf{z},\mathbf{x}^u,\lambda,\gamma} = \arg\min_{f\in\mathcal{H}_K}\Big\{\frac{1}{m}\sum_{i=1}^m V(y_i, f(x_i)) + \lambda\|f\|_K^2 + \frac{\gamma}{(m+u)^2}\sum_{i,j=1}^{m+u} W_{i,j}\Big(f(x_i) - f(x_j)\Big)^2\Big\},$$

where $W_{i,j}$ are edge weights in the data adjacency graph and $\gamma > 0$ is a semi-supervised learning regularization parameter.

Belkin-Niyogi-Sindhwani, Guo-Lin-Shi, ...

Note that $\frac{1}{(m+u)^2}\sum_{i,j=1}^{m+u} W_{i,j}\Big(f(x_i) - f(x_j)\Big)^2 = \mathbf{f}^T L_{\mathbf{x}}\mathbf{f}$ where $\mathbf{f} = (f(x_1), \ldots, f(x_{m+u}))^T$ and $L_{\mathbf{x}}$ is the unnormalized Graph Laplacian associated with the adjacency matrix $(W_{i,j})$.

V. Sparsity in Machine Learning

V.1. LASSO

Let $\{\phi_i, \ldots, \phi_N\}$ be functions on X. The **least absolute shrinkage and selection operator** (LASSO) outputs the function $f_{\mathbf{lasso}} = \sum_{i=1}^N c_i^{\mathbf{lasso}} \phi_i$ where with $\gamma > 0$,

$$c^{\mathbf{lasso}} = \arg\min_{c \in \mathbb{R}^N} \frac{1}{m} \sum_{i=1}^m \left(y_i - \sum_{j=1}^N c_j \phi_j(x_i) \right)^2 + \gamma \|c\|_1.$$

If we denote $\mathbf{X} = [\phi_j(x_i)]_{i=1,\ldots,m, j=1,\ldots,N} \in \mathbb{R}^{m \times N}$, then it is a convex linear programming optimization problem

$$c^{\mathbf{lasso}} = \arg\min_{c \in \mathbb{R}^N} c^T \left(\frac{1}{m} X^T X \right) c - 2 \left(\frac{1}{m} X^T \mathbf{y} \right)^T c + \gamma \|c\|_1.$$

Tibshirani (1996)

More efficient algorithm: least angle regression (LARS)

Theorem 7 (Orthogonal design) *If* $\frac{1}{m}X^TX = I$, *then by letting* $d = \frac{1}{m}X^T\mathbf{y} \in \mathbb{R}^N$, *for each* j, *we have*

$$c_j^{\mathbf{lasso}} = \left(|d_j| - \frac{\gamma}{2}\right)_+ sgn(d_j) = \begin{cases} d_j - \frac{\gamma}{2}, & \text{if } d_j > \frac{\gamma}{2}, \\ 0, & \text{if } -\frac{\gamma}{2} \le d_j \le \frac{\gamma}{2} \\ d_j + \frac{\gamma}{2}, & \text{if } d_j < -\frac{\gamma}{2}. \end{cases}$$

Proof. Since $X^TX = I$, we find $c^T(\frac{1}{m}X^TX)c - 2(\frac{1}{m}X^T\mathbf{y})^Tc + \gamma\|c\|_1 = \sum_{j=1}^N \left\{c_j^2 - 2d_jc_j + \gamma|c_j|\right\}$. So for each j, $c_j^{\mathbf{lasso}}$ is a minimizer of the univariate function $g_j(x) = x^2 - 2d_jx + \gamma|x|$. Note $g_j'(x) = 2x - 2d_j + \gamma\text{sgn}(x)$.

When $d_j > \frac{\gamma}{2}$, g_j decreases strictly on $(-\infty, d_j - \frac{\gamma}{2})$ and increases strictly on $(d_j - \frac{\gamma}{2}, \infty)$, achieving its only minimizer at $d_j - \frac{\gamma}{2}$. Similarly, when $d_j < -\frac{\gamma}{2}$, g_j achieves its minimizer at $d_j + \frac{\gamma}{2}$. When $-\frac{\gamma}{2} \le d_j \le \frac{\gamma}{2}$, g_j decreases strictly on $(-\infty, 0)$ and increases strictly on $(0, \infty)$, achieving its only minimizer at 0. This proves the theorem.

V.2. Elastic Net

The elastic net outputs the function $f_{\mathbf{en}} = \sum_{i=1}^{N} c_i^{\mathbf{en}} \phi_i$ where with $\gamma_1, \gamma_2 > 0$,

$$c^{\mathbf{en}} = \arg\min_{c \in \mathbb{R}^N} \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{N} c_j \phi_j(x_i) \right)^2 + \gamma_1 \|c\|_1 + \gamma_2 \|c\|_2^2.$$

Observe that $\gamma_2 \|c\|_2^2 = \frac{1}{m} \|0 - \sqrt{m\gamma_2} I c\|_2^2$. If we denote $\mathbf{X}^* = \begin{bmatrix} X \\ \sqrt{m\gamma_2} I_{N \times N} \end{bmatrix} \in \mathbb{R}^{(m+N) \times N}$ and $\mathbf{y}^* = \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} \in \mathbb{R}^{m+N}$, then

$$c^{\mathbf{en}} = c^{\mathbf{lasso}}, \text{ where } (X, \mathbf{y}) \text{ is replaced by } (X^*, \mathbf{y}^*).$$

In the orthogonal design with $\frac{1}{m} X^T X = I$, we have $\frac{1}{m} (X^*)^T X^* = (1 + \gamma_2) I$ and $c^{\mathbf{en}} = \frac{1}{\sqrt{1+\gamma_2}} c^{\mathbf{lasso}}$.

Zou-Hastie (2005)

V.3. Ridge Regression and Ordinary LS

The ridge regression is defined by $f_{\mathbf{rr}} = \sum_{i=1}^{N} c_i^{\mathbf{rr}} \phi_i$ where with $\gamma_2 > 0$,

$$c^{\mathbf{rr}} = \arg\min_{c \in \mathbb{R}^N} \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{N} c_j \phi_j(x_i) \right)^2 + \gamma_2 \|c\|_2^2.$$

The coefficient vector is given by

$$c^{\mathbf{rr}} = \left(\frac{1}{m} X^T X + \gamma_2 I \right)^{-1} \frac{1}{m} X^T \mathbf{y}.$$

The ordinary LS corresponds to the case of $\gamma_2 = 0$ and (when $X^T X$ is invertible) the coefficient vector is given by

$$c^{\mathbf{ls}} = \left(\frac{1}{m} X^T X \right)^{-1} \frac{1}{m} X^T \mathbf{y}.$$

V.4. Sparsity for Orthogonal Design

For the orthogonal design, we have $\frac{1}{m}X^TX = I$ and the vector d is the same as the coefficient vector of the ordinary LS.

$$\begin{aligned}
\textbf{Ordinary least-square} \quad c_j^{\mathbf{ls}} &= d_j, \\
\textbf{Ridge regression} \quad c_j^{\mathbf{rr}} &= \frac{d_j}{1+\gamma_2}, \\
\textbf{Lasso} \quad c_j^{\mathbf{lasso}} &= \left(|d_j| - \frac{\gamma}{2}\right)_+ \mathrm{sgn}(d_j), \\
\textbf{Elastic net} \quad c_j^{\mathbf{en}} &= \frac{1}{\sqrt{1+\gamma_2}}\left(|d_j| - \frac{\gamma}{2}\right)_+ \mathrm{sgn}(d_j).
\end{aligned}$$

Therefore, least-square or ridge regression has hardly any sparsity: the coefficient does not vanish unless d_j is exactly 0.

Lasso and elastic net may have sparse representations: the j-th coefficient vanishes if $|d_j| < \frac{\gamma}{2}$. Note that $d_j = \frac{1}{m}\sum_{i=1}^m y_i\phi_j(x_i) \approx \langle f_\rho, \phi_j\rangle_{L^2_{\rho_X}}$ measures the correlation between f_ρ and ϕ_j.

V.5. Group Effect of Elastic Net for General Design

For general design, we do not have $\frac{1}{m}X^TX = I$ and columns of X may be correlated, making $\sigma^{\mathbf{z}}_{j,k} := \frac{1}{m}\sum_{i=1}^m \phi_j(x_i)\phi_k(x_i)$ to be large for some $j \neq k$.

Group effect of elastic net: if normalized columns j, k of X are highly correlated, making $\sigma^{\mathbf{z}}_{j,k}$ close to 1 (equal to 1 if the columns are identical), then $c_j^{\mathbf{en}} \approx c_k^{\mathbf{en}}$.

Theorem 8 *Assume* $\frac{1}{m}\sum_{i=1}^m (\phi_\ell(x_i))^2 = 1$ *for every* ℓ. *Then*

$$\left|c_j^{\mathbf{en}} - c_k^{\mathbf{en}}\right| \leq \frac{\sqrt{\frac{1}{m}\sum_{i=1}^m y_i^2}}{\gamma_2}\sqrt{2 - 2\sigma^{\mathbf{z}}_{j,k}}.$$

Zou-Hastie, Zhou, ...

Lasso: if $\sigma^{\mathbf{z}}_{j,k} = 1$, we see that for a solution $c^{\mathbf{lasso}}$ of Lasso, any coefficient vector with $(c_j^{\mathbf{lasso}}, c_k^{\mathbf{lasso}})$ replaced by $(s(c_j^{\mathbf{lasso}} + c_k^{\mathbf{lasso}}), (1-s)(c_j^{\mathbf{lasso}} + c_k^{\mathbf{lasso}}))$ is also a solution for any $0 < s < 1$.

V.6. Empirical Feature-Based Learning Algorithms

Let $L_K^{\mathbf{x}}$ be an empirical integral operator on $\mathcal{H}_K$ defined by

$$L_K^{\mathbf{x}} f = \frac{1}{m} \sum_{i=1}^{m} f(x_i) K_{x_i} = \frac{1}{m} \sum_{i=1}^{m} \langle f, K_{x_i} \rangle_K K_{x_i}$$

with normalized eigenpairs $\{(\lambda_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})\}$.

Then

$$\lambda_i^{\mathbf{x}} = \widehat{\lambda}_i^{\mathbf{x}}, \quad \phi_i^{\mathbf{x}} = \sum_{j=1}^{m} (\widehat{\mu}_i)_j K_{x_j} / \sqrt{m \widehat{\lambda}_i^{\mathbf{x}}}, \qquad i = 1, \ldots, d,$$

$$\phi_i^{\mathbf{x}}(x_j) = 0, \qquad i > d$$

where d is the rank of $\mathbb{K} := [K(x_i, x_j)/m]_{i,j=1}^m$ and $\{(\widehat{\lambda}_i^{\mathbf{x}}, \widehat{\mu}_i)\}$ are normalized eigenpairs.

Now we regard the integral operator as one on $\mathcal{H}_K$. Then the normalized eigenfunction $\sqrt{\lambda_k}\phi$ in $\mathcal{H}_K$ is denoted as ϕ_k. By Theorem 9, $\phi_i^{\mathbf{x}} \approx \phi_i$ for each fixed i.

Kernel Projection Machine: Output $\sum_{i=1}^{m} c_{\gamma,i}^{\mathbf{z}} \phi_i^{\mathbf{x}}$ where

$$c_\gamma^{\mathbf{z}} = \arg\min_{c \in \mathbb{R}^m} \left\{ \frac{1}{m} \sum_{i=1}^{m} V\left(\sum_{j=1}^{m} c_j \phi_j^{\mathbf{x}}(x_i), y_i \right) + \gamma \|c\|_0 \right\}.$$

Here $\|c\|_0$ is the number of nonzero entries of the vector $c = (c_1, \cdots, c_m) \in \mathbb{R}^m$.

Zwald, ...

Empirical Feature-Based Learning Algorithms: output $f^{\mathbf{z}} = \sum_{i=1}^{\infty} c_i^{\mathbf{z}} \phi_i^{\mathbf{x}}$ with $0 < q \leq 1$ and $c^{\mathbf{z}} = (c_i^{\mathbf{z}})_{i=1}^{\infty}$ given by

$$c^{\mathbf{z}} = \arg\min_{c \in \ell^2} \left\{ \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{j=1}^{\infty} c_j \phi_j^{\mathbf{x}}(x_i) - y_i \right)^2 + \gamma \|c\|_q^q \right\},$$

where $\|c\|_q^q = \sum_{j=1}^{\infty} |c_j|^q$.

Guo-Zhou (ACHA 2011), Guo-Fan-Zhou (JMLR 2016), ...

VI. Online Learning and Stochastic Gradient Descent

VI.1. Kaczmarz algorithms

Solve a linear system

$$Aw = b,\ A = \begin{bmatrix} a_1^T \\ \vdots \\ a_m^T \end{bmatrix} \in \mathbb{R}^{m\times d}, b \in \mathbb{R}^m \Longleftrightarrow a_i \cdot w = b_i,\ i = 1, \ldots, m.$$

Classical Kaczmarz (1937) algorithm:

$$w_{k+1} = w_k + \frac{b_i - \langle a_i, w_k\rangle}{\|a_i\|^2} a_i, \qquad k = 1, 2, \ldots$$

where $i = k \mod m$, and $w_1 \in \mathbb{R}^d$ is an initial vector.

Randomized Kaczmarz algorithm: Strohmer and Vershynin (2009), Needell (2010), ...

$$w_{k+1} = w_k + \left(\frac{b_{p(k)}}{\|a_{p(k)}\|} - \langle \frac{a_{p(k)}}{\|a_{p(k)}\|}, w_k \rangle \right) \frac{a_{p(k)}}{\|a_{p(k)}\|},$$

where $p(k)$ takes values in $\{1, \ldots, m\}$ with probability proportional to the row norm squares $P(p(k) = j) = \|a_j\|^2 / \sum_\ell \|a_\ell\|^2$.

VI.2. Learning theory of randomized Kaczmarz algorithm

(J. H. Lin-Zhou, JMLR, 2015)

Relaxed randomized Kaczmarz algorithm

$$w_{k+1} = w_k + \eta_k \{y_k - \langle x_k, w_k \rangle\} x_k, \qquad k = 1, 2, \ldots, \qquad (11)$$

where $w_1 = 0$ and $\{(x_k, y_k)\}_k$ is independently drawn according to ρ on $Z = X \times Y$ with $X \subseteq \{x \in \mathbb{R}^d : \|x\| \leq R\}$ and $Y = \mathbb{R}$. Consider homogeneous linear functions $f_w : X \to \mathbb{R}$ with $w \in \mathbb{R}^d$ defined by $f_w(x) := \langle w, x \rangle$.

Theorem 9 *Assume* $\inf_{w \in \mathbb{R}^d} \mathcal{E}(f_w) > 0$. *Then* $\mathbb{E}_{z_1,\ldots,z_T} \|w_{T+1} - w^*\|^2 \to 0$ *for some* $w^* \in \mathbb{R}^d$ *if and only if* $\lim_{k\to\infty} \eta_k = 0$ *and* $\sum_{k=1}^{\infty} \eta_k = \infty$. *In this case,* $\sum_{T=1}^{\infty} \sqrt{\mathbb{E}_{z_1,\ldots,z_T} \|x_{T+1} - x^*\|^2} = \infty$. *If* $\eta_k = \eta_1 k^{-\theta}$ *for some* $\theta \in (0, 1)$ *and* $\eta_1 \in (0, 1)$, *then* $\mathbb{E}_{z_1,\ldots,z_T} \|x_{T+1} - x^*\|^2 = O\left(T^{-\theta}\right)$.

Idea: Fix $w^* \in \mathbb{R}^d$. Then $w_{k+1} - w^* = w_k - w^* + \eta_k \{y_k - \langle x_k, w_k \rangle\} x_k$ and $\|w_{k+1} - w^*\|^2 = \langle w_{k+1} - w^*, w_{k+1} - w^* \rangle$ is

$$\|w_k - w^*\|^2 + 2\eta_k \{y_k - \langle x_k, w_k \rangle\} \langle x_k, w_k - w^* \rangle + \eta_k^2 \{y_k - \langle x_k, w_k \rangle\}^2 \|x_k\|^2.$$

Since w_k is independent of $z_k = (x_k, y_k)$, taking E_{z_k} gives

$$\begin{aligned} E_{z_k}[\|w_{k+1} - w^*\|^2] &= \|w_k - w^*\|^2 + 2\eta_k \int_Z \{y - f_{w_k}(x)\} f_{w_k - w^*}(x) d\rho \\ &\quad + \eta_k^2 \int_Z \{y - f_{w_k}(x)\}^2 \|x\|^2 d\rho. \end{aligned}$$

Take $w^* = \arg\inf_{w \in \mathbb{R}^d} \mathcal{E}(f_w)$ and the hypothesis space $\mathcal{H} = \left\{ f_w \in L^2_{\rho_X} : \ w \in \mathbb{R}^d \right\}$. Then f_{w^*} is the orthogonal projection of f_ρ in $L^2_{\rho_X}$ onto the subspace $\mathcal{H}$. So $f_\rho - f_{w^*} \perp \mathcal{H}$ and $\langle f_\rho, f_{w_k - w^*} \rangle_{L^2_{\rho_X}} = \langle f_{w^*}, f_{w_k - w^*} \rangle_{L^2_{\rho_X}}$. This together with $E[y|x] = f_\rho(x)$ tells us that the above middle term equals

$$\int_X \{f_\rho(x) - f_{w_k}(x)\} f_{w_k - w^*}(x) d\rho_X = \langle f_{w^*} - f_{w_k}, f_{w_k - w^*} \rangle_{L^2_{\rho_X}}.$$

It is $-\|f_{w_k - w^*}\|^2_{L^2_{\rho_X}} = -\int_X \langle w_k - w^*, x \rangle \langle x, w_k - w^* \rangle d\rho_X$.

We introduce the covariance matrix $C_{\rho_X} = E_{\rho_X}[xx^\top]$ associated with the marginal distribution ρ_X. Denote its eigenspace associated with the eigenvalue 0 as V_0 and the orthogonal projection onto V_0 as P_0. Then x_t is orthogonal to V_0 almost surely for each t. It follows that $P_0(x_t) = P_0(x_1) = 0$ for each t. Take the vector w^* to be the minimizer of $\mathcal{E}(f_w)$ in $\mathbb{R}^d$ such that $P_0(w^*) = P_0(x_1) = 0$. With this choice, $w_k - w^*$ is orthogonal to V_0 for each t, and belongs to the orthogonal complement $V_0^\perp$. The eigenvalues of C_{ρ_X} restricted to the subspace $V_0^\perp$ is positive $\lambda_r > 0$. So we have

$$\begin{aligned}&\int_X \langle w_k - w^*, x\rangle\langle x, w_k - w^*\rangle d\rho_X = (w_k - w^*)^\top E[xx^\top](w_k - w^*)\\ &= (w_k - w^*)^\top C_{\rho_X}(w_k - w^*) \geq \lambda_r \|w_k - w^*\|^2.\end{aligned}$$

Also, $\int_Z \{y - f_{w_k}(x)\}^2 \|x\|^2 d\rho$ is bounded by $R^2\mathcal{E}(f_{w_k}) = R^2\mathcal{E}(f_{w^*}) + R^2\|f_{w_k - w^*}\|^2_{L^2_{\rho_X}} = R^2\mathcal{E}(f_{w^*}) + R^2(w_k - w^*)^\top C_{\rho_X}(w_k - w^*)$. Thus when $\eta_k \leq \lambda_r/(R^2\lambda_1)$,

$$E_{z_k}[\|w_{k+1} - w^*\|^2] \;\leq\; (1 - \lambda_r\eta_k)\|w_k - w^*\|^2 + \eta_k^2 R^2 \mathcal{E}(f_{w^*}).$$

VI.3. Online Learning or Stochastic Gradient Descent for Classification

(Ying-Zhou, 2006)

Towards a goal of finding a minimizer of the regularized generalization error $\mathcal{E}^\phi(f)+\frac{\lambda}{2}\|f\|_K^2$ in $\mathcal{H}_K$ with a convex classification loss $\phi : \mathbb{R} \to \mathbb{R}_+$, $Y = \{1, -1\}$ and $\mathcal{E}^\phi(f) = \int_Z \phi(yf(x))d\rho$, we take a functional derivative in $\mathcal{H}_K$:

$$\partial\left(\mathcal{E}^\phi(f) + \frac{\lambda}{2}\|f\|_K^2\right) = \int_Z \phi'_-(yf(x))yK_x d\rho + \lambda f,$$

where $f(x) = \langle f, K_x\rangle_K$ has functional derivative K_x in $\mathcal{H}_K$.

For online learning, the sample pairs $\{z_k = (x_k, y_k)\}$ arrive sequentially. One can process one sample pair at each time. Replace the integral $\int_Z \phi'_-(yf(x))yK_x d\rho$ by one sample value $\phi'_-(y_k f(x_k))y_k K_{x_k}$ at $f = f_k \in \mathcal{H}_K$, we get a stochastic gradient descent algorithm for classification:

$$f_{k+1} = f_k - \eta_k\left\{\phi'_-(y_k f_k(x_k))y_k K_{x_k} + \lambda f_k\right\}, \qquad k = 1, 2, \ldots$$

Fully online learning: replace λ by λ_k

Example 6. For the hinge loss with $\phi(u) = (1-u)_+$ and $f_1 = 0$, the online algorithm is

$$f_{k+1} = \begin{cases} (1-\eta_k\lambda_k)f_k, & \text{if } y_k f_k(x_k) > 1, \\ (1-\eta_k\lambda_k)f_k + \eta_k y_k K_{x_k}, & \text{if } y_k f_t(x_k) \le 1. \end{cases}$$

Assume $K \in C^{2s}(X \times X)$ for some $0 < s \le \frac{1}{2}$, $f_\rho \in C^s(X)$ and the triple (ρ_X, f_c, K) satisfies

$$\inf_{f \in \mathcal{H}_K} \{\|f - f_c\|_{L^1_{\rho_X}} + \frac{\lambda}{2}\|f\|_K^2\} \le \mathcal{D}_0 \lambda^\beta \qquad \forall \lambda > 0$$

for some $0 < \beta \le 1$ and $\mathcal{D}_0 > 0$. Take

$$\lambda_k = \lambda_1 k^{-\frac{1}{4}}, \eta_k = \eta_1 k^{-(\frac{1}{2}+\frac{\beta}{12})}$$

where $\lambda_1 > 0$ and $0 < \eta_1 \le \frac{1}{2\kappa^2+\lambda_1}$. Then

$$E_{z_1,\dots,z_T}\left(\mathcal{R}(\text{sgn}(f_{T+1})) - \mathcal{R}(f_c)\right) = O\left(T^{-\min\{\frac{\beta}{4}, \frac{1}{8}+\frac{\beta}{24}\}}\right).$$

Idea of the analysis: Take the minimizer $f_\lambda \in \mathcal{H}_K$ of $\mathcal{E}^\phi(f) + \frac{\lambda}{2}\|f\|_K^2$. Taking inner products gives

$$\|f_{k+1} - f_\lambda\|_K^2 = \|f_k - f_\lambda\|_K^2 + \eta_k^2 \left\|\phi'_-(y_k f_k(x_k)) y_k K_{x_k} + \lambda f_k\right\|_K^2$$
$$-2\eta_k \left\{\phi'_-(y_k f_k(x_k)) y_k \left(f_k(x_k) - f_\lambda(x_k)\right) + \lambda \langle f_k, f_k - f_\lambda\rangle_K\right\}.$$

By **convexity**, $\phi'_-(a)(b-a) \le \phi(b) - \phi(a)$, the middle term is

$$\phi'_-(y_k f_k(x_k)) \left(y_k f_\lambda(x_k) - y_k f_k(x_k)\right) \le \phi(y_k f_\lambda(x_k)) - \phi(y_k f_k(x_k)).$$

Also, $-\langle f_k, f_k - f_\lambda\rangle_K \le -\|f_k\|_K^2 + \frac{1}{2}\left(\|f_k\|_K^2 + \|f_\lambda\|_K^2\right)$. Hence

$$E_{z_k}[\|f_{k+1} - f_\lambda\|_K^2] \le \|f_k - f_\lambda\|_K^2 + \eta_k^2 \{\ldots\}$$
$$+2\eta_k \left\{\left[\mathcal{E}^\phi(f_\lambda) + \frac{\lambda}{2}\|f_\lambda\|_K^2\right] - \left[\mathcal{E}^\phi(f_k) + \frac{\lambda}{2}\|f_k\|_K^2\right]\right\}.$$

Key lemma: $\left[\mathcal{E}^\phi(f) + \frac{\lambda}{2}\|f\|_K^2\right] - \left[\mathcal{E}^\phi(f_\lambda) + \frac{\lambda}{2}\|f_\lambda\|_K^2\right] \ge \frac{\lambda}{2}\|f - f_\lambda\|_K^2$.
So we have

$$E_{z_k}[\|f_{k+1} - f_\lambda\|_K^2] \le (1 - \lambda\eta_k)\|f_k - f_\lambda\|_K^2 + \eta_k^2 \{\ldots\}.$$

VI.4. Online Linearized Bregman Iteration

(Y. W. Lei-Zhou, 2017)

Linearized Bregman iteration with a threshold parameter $\lambda \geq 0$ produces $\{(w_t, v_t)\}_{t\in\mathbb{N}}$ in $\mathbb{R}^d$ with $w_1 = v_1 = 0 \in \mathbb{R}^d$ as

$$\begin{cases} v_{t+1} = v_t - \eta_t A^\top (Aw_t - y), \\ w_{t+1} = S_\lambda(v_{t+1}), \end{cases} \tag{12}$$

where $S_\lambda : \mathbb{R}^d \to \mathbb{R}^d$ is the soft-thresholding operator defined component-wisely in terms of the soft-thresholding function $S_\lambda : \mathbb{R} \to \mathbb{R}$ given by $S_\lambda(v) := \mathrm{sgn}(v)(|v| - \lambda)_+$.

Convergence analysis: Cai-Osher-Shen, ...

Related algorithms in image processing: Goldstein-Osher, Yin-Osher-Goldfarb-Darbon, Zhang-Burger-Osher, ...

Online Linearized Bregman Iteration:

$$\begin{cases} v_{t+1} = v_t - \eta_t\{\langle w_t, x_t\rangle - y_t\}x_t, \\ w_{t+1} = S_\lambda(v_{t+1}). \end{cases} \tag{13}$$

Let $\Psi(w) = \lambda\|w\|_1 + \frac{1}{2}\|w\|_2^2$ and

$$w^* = \arg\min_{w\in W^*} \Psi(w) \quad \text{with } W^* := \{w \in \mathbb{R}^d : C_{\rho_X} w = E_\rho[xy]\}.$$

Let I be the support of w^* and we denote by $w(I) = (w(i))_{i\in I}$ the restriction of $w \in \mathbb{R}^d$ onto the index set I.

Theorem 10 *Assume* $w^* \neq 0$ *and* $\inf_{w\in\mathbb{R}^d} E_\rho[\|(\langle w, x\rangle - y)x(I)\|_2] > 0$. *Then* $\lim_{T\to\infty} E_{z_1,\ldots,z_{T-1}}[\|w_T - w^*\|_2^2] = 0$ *if and only if the step size sequence satisfies* $\lim_{t\to\infty}\eta_t = 0$ *and* $\sum_{t=1}^\infty \eta_t = \infty$. *In this case, we have*

$$\sum_{T=1}^\infty \sqrt{E_{z_1,\ldots,z_{T-1}}[\|w_T - w^*\|_2^2]} = \infty.$$

Idea: Use the Bregman distance between w and $\tilde{w}$ by

$$D_\Psi^{\tilde{v}}(w, \tilde{w}) = \Psi(w) - \Psi(\tilde{w}) - \langle w - \tilde{w}, \tilde{v} \rangle,$$

where $\tilde{v} \in \partial\Psi(w)$, the subdifferential of Ψ at w. It satisfies

$$D_\Psi^{\tilde{v}}(w, \tilde{w}) \geq \frac{1}{2}\|w - \tilde{w}\|_2^2, \qquad \forall w, \tilde{w} \in \mathbb{R}^d, \tilde{v} \in \partial\Psi(\tilde{w}).$$

and $v_{t+1} \in \partial\Psi(w_{t+1})$.

Error decomposition:

$$D_\Psi^{v_{t+1}}(w, w_{t+1}) - D_\Psi^{v_t}(w, w_t) = \langle w - w_{t+1}, v_t - v_{t+1} \rangle - D_\Psi^{v_t}(w_{t+1}, w_t).$$

Separate $w - w_{t+1}$ into $w - w_t + w_t - w_{t+1}$ and bound $\langle w_t - w_{t+1}, v_t - v_{t+1} \rangle$ by $\frac{1}{2}\|w_t - w_{t+1}\|_2^2 + \frac{1}{2}\|v_t - v_{t+1}\|_2^2$. Then

$$\begin{aligned} D_\Psi^{v_{t+1}}(w, w_{t+1}) - D_\Psi^{v_t}(w, w_t) &\leq \langle w - w_t, v_t - v_{t+1} \rangle + \frac{1}{2}\|v_t - v_{t+1}\|_2^2 \\ &= \eta_t \langle w - w_t, (\langle w_t, x_t \rangle - y_t) x_t \rangle + \frac{1}{2}\eta_t^2 \|(\langle w_t, x_t \rangle - y_t) x_t\|_2^2. \end{aligned}$$

But the function $f(w) = 2^{-1}(\langle w, x_t\rangle - y_t)^2$ is convex. So

$$D_\Psi^{v_t+1}(w, w_{t+1}) - D_\Psi^{v_t}(w, w_t) \leq \frac{\eta_t}{2}\Big[(\langle w, x_t\rangle - y_t)^2 - (\langle w_t, x_t\rangle - y_t)^2\Big] + \frac{\eta_t^2}{2} R^2 (\langle w_t, x_t\rangle - y_t)^2.$$

Taking $w = w^*$ and taking conditional expectations, we have

$$E_{z_t}[D_\Psi^{v_t+1}(w^*, w_{t+1})] - D_\Psi^{v_t}(w^*, w_t) \leq \eta_t[\mathcal{E}(f_{w^*}) - \mathcal{E}(f_{w_t})] + \eta_t^2 R^2 \mathcal{E}(f_{w_t}).$$

Key lemma: for some constant $\overline{C}$, there holds

$$D_\Psi^{v_t}(w^*, w_t) \leq \overline{C}\,[\mathcal{E}(f_{w_t}) - \mathcal{E}(f_{w^*})]\,.$$

Then with two positive constants $\tilde{a}$, $\tilde{b}$, we have

$$E_{z_1,\ldots,z_t}[D_\Psi^{v_t+1}(w^*, w_{t+1})] \leq (1 - \tilde{a}\eta_t) E_{z_1,\ldots,z_{t-1}}[D_\Psi^{v_t}(w^*, w_t)] + \tilde{b}\eta_t^2.$$

This yields $E_{z_1,\ldots,z_{T-1}}[D_\Psi^{v_T}(w^*, w_T)] \to 0$ as $T \to \infty$.

VII. Distributed Learning with Big Data

VII.1. Distributed learning: based on a divide-and-conquer approach

A distributed learning algorithm consisting of three steps:
(1) partitioning the data into disjoint subsets OR the original data are stored distributively
(2) applying a learning algorithm implemented in an individual machine or local 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, privacy-preserving, ...

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: Zhang-Duchi-Wainwright, Shamir-Srebro, Meister-Steinwart, ...

VII.2. 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, Steinwart-Christmann, Cucker-Zhou, ...
Too many papers to mention

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 $\sigma_\rho^2(x)=\int_{\mathcal{Y}}(y-f_\rho(x))^2\,d\rho(y|x)$, ...

VII.3. Previous work on optimal learning rates

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

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

Steinwart-Hush-Scovel (2009): If $f_\rho \in \mathcal{H}_K$, $\lambda_i = O\left(i^{-2\alpha}\right)$ 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 $[-M, M]$.

Smale-Zhou (2007): If $f_\rho = L_K(g_\rho)$ for some $g \in L^2_{\rho_X}$ and $|y| \le M$ almost surely, then for any $0 < \delta < 1$, with confidence $1 - \delta$ there holds

$$\|f_{\mathbf{z},\lambda} - f_\rho\|_K \le 7\Big(\kappa M \log(2/\delta)\Big)^{1/3} \|g\|^{2/3}_{L^2_{\rho_X}} \Big(\frac{1}{m}\Big)^{\frac{1}{6}}$$

where $\lambda = \Big(32\kappa M \log(2/\delta)/\|g\|_\rho\Big)^{2/3} m^{-1/3}$ and $\kappa = \sup_{x\in X} \sqrt{K(x,x)}$.

VII.4. Distributed Learning with Regularization Schemes

(S. B. Lin-X. Guo-Zhou, JMLR, 2017)

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

$$\lambda_i = O(i^{-2\alpha}) \quad \Longrightarrow \quad \mathcal{N}(\lambda) = O(\lambda^{-\frac{1}{2\alpha}})$$

Theorem 11 *Assume* $|y| \le M$ *and* $f_\rho = L_K^r(g_\rho)$ *for some* $\frac{1}{2} \le r \le 1$ *and* $g_\rho \in L^2_{\rho_X}$. *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*

$$m \le N^{\min\left\{\frac{6\alpha(2r-1)+1}{5(4\alpha r+1)}, \frac{2\alpha(2r-1)}{4\alpha r+1}\right\}}, \tag{14}$$

then by taking $\lambda = N^{-\frac{2\alpha}{4\alpha r+1}}$, *we have*

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

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

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

$$\|\varphi_i\|_{L^\infty_{\rho_X}} \le A \text{ or } \|\varphi_i\|_{L^{2k}_{\rho_X}} = \left\{E\left[|\varphi_i(x)|^{2k}\right]\right\}^{\frac{1}{2k}} \le A, \quad \forall i \qquad (15)$$

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

$$m = O(N^{\frac{2\alpha-1}{2\alpha+1}}/(A^4 \log N)) \qquad (16)$$

or $m = O((N^{\frac{2(k-4)\alpha-k}{2\alpha+1}}/(A^{4k}\log^k N))^{\frac{1}{k-2}})$, and $\lambda = N^{\frac{2\alpha}{2\alpha+1}}$.

But checking the eigenfunction assumption (15) for a general marginal distribution is difficult.

VII.5. Optimal rates for regularization: by-product

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

Theorem 12 *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*

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

VII.6. Novelty: Second order decomposition

Empirical integral operator $L_{K,D}$ with data D on $\mathcal{H}_K$ is

$$L_{K,D}(f) = \frac{1}{N}\sum_{i=1}^{N} f(x_i)K_{x_i} = \frac{1}{N}\sum_{i=1}^{N} \langle f, K_{x_i}\rangle_K K_{x_i},$$

where $K_x = K(\cdot, x)$. Then

$$f_{D,\lambda} = \left(L_{K,D} + \lambda I\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} y_i K_{x_i}\right).$$

Note that $\frac{1}{N}\sum_{i=1}^{N} y_i K_{x_i} \approx \int_Z y K_x d\rho = \int_X f_\rho(x) K_x d\rho_X = L_K(f_\rho)$.
For the operators $A = L_{K,D} + \lambda I$ and $B = L_K + \lambda I$, we have

$$A^{-1} - B^{-1} = B^{-1}\{B - A\}A^{-1} \qquad \text{(first order decomposition)}$$

and

$$A^{-1} - B^{-1} = B^{-1}\{B - A\}B^{-1} + B^{-1}\{B - A\}A^{-1}\{B - A\}B^{-1}$$

our second order decomposition.

Second order decomposition used to solve two open problems on kernel partial least squares and kernel gradient descent: S. B. Lin-Zhou (J. Fourier Anal. Appl. 2017), S. B. Lin-Zhou (Constr. Approx. 2017)

It has been applied to other problems: Z. C. Guo-L. Shi (ACHA 2017), Z. C. Guo-S. B. Lin-L. Shi (ACHA 2017), ...

Advantages of our analysis: general results without any eigenfunction assumption and error estimates in the $\mathcal{H}_K$ metric (Smale-Zhou 2007)

VII.7. Distributed learning with spectral algorithms

The spectral algorithm associated with K and a filter function $g_\lambda : [0, \kappa^2] \to \mathbb{R}$ with parameter $\lambda > 0$ is

$$f_{g_\lambda, D, \lambda} = g_\lambda(L_{K,D}) \frac{1}{N} \sum_{i=1}^{N} y_i K_{x_i}.$$

Here $g_\lambda(L_{K,D})$ is a linear operator on $\mathcal{H}_K$ defined by spectral calculus: for a set $\{(\sigma_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})_i\}$ of normalized eigenpairs of $L_{K,D}$, $g_\lambda(L_{K,D}) = \sum_i g_\lambda(\sigma_i^{\mathbf{x}}) \phi_i^{\mathbf{x}} \otimes \phi_i^{\mathbf{x}} = \sum_i g_\lambda(\sigma_i^{\mathbf{x}}) \langle \cdot, \phi_i^{\mathbf{x}} \rangle_K \phi_i^{\mathbf{x}}$.

It is an approximate inversion of L_K or $L_{K,D}$ so that $f_{g_\lambda, D, \lambda} \approx L_K^{-1} L_K(f_\rho) \approx f_\rho$ since $\frac{1}{N} \sum_{i=1}^{N} y_i K_{x_i} \approx L_K(f_\rho)$

Least squares regularization: $g_\lambda(\sigma) = \frac{1}{\lambda + \sigma}$ and $g_\lambda(L_{K,D}) = (\lambda I + L_{K,D})^{-1} \approx L_K^{-1}$

Analysis of the classical spectral algorithms

Engl-Hanke-Neubauer, Lo Gerfo-Rosasco-Odone-De Vito-Verri, Bauer-Pereverzev-Rosasco, Blanchard-Krämer, ...

The filter function $g_\lambda : [0, \kappa^2] \to \mathbb{R}$ with $0 < \lambda \leq \kappa^2$ has *qualification* $\nu_g \geq \frac{1}{2}$ if there exists a positive constant b independent of λ such that

$$\sup_{0<\sigma\leq\kappa^2} |g_\lambda(\sigma)| \leq \frac{b}{\lambda}, \qquad \sup_{0<\sigma\leq\kappa^2} |g_\lambda(\sigma)\sigma| \leq b,$$

and

$$\sup_{0<\sigma\leq\kappa^2} |1 - g_\lambda(\sigma)\sigma|\sigma^\nu \leq \gamma_\nu \lambda^\nu, \qquad \forall\ 0 < \nu \leq \nu_g,$$

where $\gamma_\nu > 0$ is a constant depending only on $\nu \in (0, \nu_g]$.

Examples of spectral algorithms with different filter functions

Tikhonov Regularization: $g_\lambda(\sigma) = \frac{1}{\lambda+\sigma}$ with $\lambda > 0$: $\nu_g = 1$.

Landweber Iteration: $g_\lambda(\sigma) = \sum_{i=0}^{t-1}(1-\sigma)^i$ with $\lambda = \frac{1}{t}$ for some $t \in \mathbb{N}$, then $\nu_g = \infty$.

Spectral Cut-off: $g_\lambda(\sigma) = \begin{cases} \frac{1}{\sigma}, & \text{if } \sigma \geq \lambda, \\ 0, & \text{if } \sigma < \lambda. \end{cases}$ Then $\nu_g = \infty$.

Accelerated Landweber Iteration: $g_\lambda(\sigma) = p_t(\sigma)$ with $\lambda = t^{-2}$, $t \in \mathbb{N}$, and p_t a polynomial of degree $t-1$. A special case of the accelerated Landweber iteration is the ν method.

Distributed learning with spectral algorithms

Z. C. Guo-S. B. Lin-Zhou (Inverse Problems, 2017)

Represent $L_{K,D}(f) = \frac{1}{N}\sum_{i=1}^{N} c_i K_{x_i}$ by a singular value decomposition of the matrix $K[\mathbf{x}] := \frac{1}{N}(K(x_i, x_j))_{i,j=1}^{N}$:

$$(c_i)_{i=1}^{N} = g_\lambda(K[\mathbf{x}])(y_i)_{i=1}^{N}$$

Distributed learning with spectral algorithms:

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

to reduce the memory and computing costs to handle big data

Theorem 13 *Assume* $|y| \leq M$ *and the filter function* $g_\lambda : [0, \kappa^2] \to \mathbb{R}$ *with* $0 < \lambda \leq \kappa^2$ *has a qualification* $\nu_g \geq \frac{1}{2}$. *If* $f_\rho = L_K^r(g_\rho)$ *for some* $\frac{1}{2} \leq r \leq \nu_g$ *and* $g_\rho \in L^2_{\rho_X}$, $\mathcal{N}(\lambda) = O(\lambda^{-\beta})$ *for some* $\beta > 0$, $|D_j| = \frac{N}{m}$ *for* $j = 1, \ldots, m$, $\lambda = N^{-\frac{1}{2r+\beta}}$, *and*

$$m \leq N^{\min\left\{\frac{2}{2r+\beta}, \frac{2r-1}{2r+\beta}\right\}},$$

then

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

Other work:

for spectral algorithms: L. H. Dicker-D. P. Foster-D. Hsu, arXiv, May 2016

for spectral algorithms and distributed spectral algorithms: G. Blanchard-N. Mücke, arXiv, October 2016

VII.8. Distributed semi-supervised learning

X. Y. Chang-S. B. Lin-D. X. Zhou (JMLR, 2017)

When $f_\rho \in \mathcal{H}_K$ with $r = \frac{1}{2}$, the restriction on # of local processors in equation (14) of Theorem 11 means $m = O(1)$.

Idea: use additional unlabeled data to remove the restriction.

Let $\tilde{D}_j(x) = \{x_1^j, \ldots, x_{|\tilde{D}_j|}^j\}$ be the unlabeled data drawn independently from ρ_X stored on the j-th local processor. Construct $D^* = \bigcup_{j=1}^m D_j^*$ with $D_j^* = \left\{(x_i^*, y_i^*)\right\}_{i=1}^{|D_j|+|\tilde{D}_j(x)|}$ given by

$$x_i^* = \begin{cases} x_i, & \text{if } (x_i, y_i) \in D_j, \\ \tilde{x}_i, & \text{if } \tilde{x}_i \in \tilde{D}_j(x), \end{cases} \quad \text{and} \quad y_i^* = \begin{cases} \frac{|D_j^*|}{|D_j|} y_i, & \text{if } (x_i, y_i) \in D_j, \\ 0, & \text{otherwise.} \end{cases}$$

Distributed semi-supervised learning with regularization schemes:

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

Theorem 14 *Assume* $|y| \leq M$, $f_\rho = L_K^r(g_\rho)$ *for some* $\frac{1}{2} \leq r \leq 1$ *and* $g_\rho \in L^2_{\rho_X}$, *and* $\mathcal{N}(\lambda) = O(\lambda^{-\frac{1}{2\alpha}})$ *for some* $\alpha > 0$. *If* $\lambda = N^{-\frac{2\alpha}{4\alpha r+1}}$, $|D_1| = \ldots = |D_m|$, $|D_1^*| = \ldots = |D_m^*|$ *and*

$$m \leq \min\left\{N^{\frac{4\alpha r+2-2\alpha}{4\alpha r+1}},\ |D^*|N^{-\frac{2\alpha+1}{4\alpha r+1}}\right\},$$

then we have

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

In particular, if $|D^*| = N^{1+\frac{2\alpha}{4\alpha r+1}}$ *with additional unlabel data of size* $N^{1+\frac{2\alpha}{4\alpha r+1}} - N$, *and* $s \geq 1/2$, $m \leq N^{\frac{4\alpha r}{4\alpha r+1}}$, *the distributed semi-supervised learning algorithm achieves the optimal learning rate.*

Our analysis even works for $0 < r < \frac{1}{2}$. Simulation with Million Song data is carried out.

VIII. Deep Learning with Deep Neural Networks

VIII.1. Shallow Nets and Classical Results

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous activation function.

Cybenko (1989), Hornik (1991), Barron (1993): 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 the Fourier transform $\hat{f}$ of $f \in L^2(\mathbb{R}^d)$ satisfies $\int_{\mathbb{R}^d} |w||\hat{f}(w)|dw < \infty$, 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 $\|f_N - f\|^2_{L^2(\mathbb{R}^d)} \leq \frac{c_f}{N}$, where c_f is a constant independent of N.

Mhaskar (1996): If for some $b \in \mathbb{R}$, $\phi^{(k)}(b) \neq 0$ for all $k \in \mathbb{Z}_+$, then for any f in the Sobolev space $W_2^r(\mathbb{R}^d)$ with $r \in \mathbb{N}$, there holds

$$\|f_N - f\|_{L^2(\mathbb{R}^d)} \leq c\|f\|_{W_2^r(\mathbb{R}^d)} N^{-\frac{r}{d}}.$$

Micchelli, Pinkus, Chui-Li, Shaham-Cloningen-Coifman, ...

Universal Approximation Property (Lin-Pinkus-Schocken 1994): Consider the function set on $\mathbb{R}^d$ with $d \in \mathbb{N}$ defined by

$$\mathcal{M}(\sigma) = \text{span}\left\{\sigma(w \cdot x + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\right\}.$$

Then $\mathcal{M}(\sigma)$ is dense in $C(\mathbb{R}^d)$ in the topology of uniform convergence on all compact subsets of $\mathbb{R}^d$ if and only if σ is not a polynomial.

Lack of localized approximation: A neural networks with the activation function σ is said to provide **localized approximation** if for every compact subset K of $\mathbb{R}^d$,

$$\inf_{g \in \mathcal{M}(\sigma)} \left\| g - \chi_{[-1,1]^d} \right\|_{L^1(K)} = 0.$$

Chui-Li-Mhaskar (1994): the neural network with the activation function $\sigma = \chi_{[0,\infty)}$ does not provide localized approximation.

VIII.2. 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 $c_i \in \mathbb{R}, w_{i,j} \in \mathbb{R}^d, a_{i,j}, b_{i,j} \in \mathbb{R}$.

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

VIII.3. Deep Neural Network Architectures

Convolutional neural networks mainly for image and speech applications

Recursive neural networks mainly for natural language processing

Deep belief networks: no connection within layers: deep Boltzmann machines

A large literature in comparing representational complexity

Little is known on approximation or learning ability of structured deep neural networks

VIII.4. Compositional structures of functions

Minimax rates of estimations (Stone 1982): if a regression function f is Lipschitz on $\mathbb{R}^d$ α with $0 < \alpha < 1$, then the optimal minimax rate of statistical regression estimators with N samples is $N^{-\frac{2\alpha}{2\alpha+d}}$.

Additive models (Stone 1985): $f(x_1, \ldots, x_d) = f_1(x_1) + \ldots + f_d(x_d)$ with minimax rate $N^{-\frac{2\alpha}{2\alpha+1}}$.
Raskutti-Wainwright-Yu (IEEE TIT, 2011), Yuan-Zhou (AoS, 2016), Christmann-Zhou (AA, 2016), ...

Interaction models (Stone 1994): $f = \sum_{I \subseteq \{1,\ldots,d\}, |I| = d^*} f_I(x_I)$ with minimax rate $N^{-\frac{2\alpha}{2\alpha+d^*}}$. Here $d^* \in \{1, \ldots, d\}$ 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 (Härdle and Stoker 1989): $f = g(a \cdot x)$ for some $a \in \mathbb{R}^d$ and $g : \mathbb{R} \to \mathbb{R}$

Projection pursuit (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)

Simple case: $f = g(f_1(x_{I_1}), f_2(x_{I_2}), \ldots, f_{d^*}(x_{I_{d^*}}))$

Generalized hierarchical model: $f = g(a_1 \cdot x, \ldots, a_{d^*} \cdot x)$

Generalized hierarchical interaction model: $f = \sum_k g_k(f_{1,k}, \ldots, f_{d^*,k})$ with $f_{i,k}(x)$ generalized hierarchical model

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

Extended List of References

[1] F. Cucker and D. X. Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, 2007.

[2] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, 2000.

[3] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, 1998.

[4] B. Schölkopf and A. J. Smola, Learning with Kernels, MIT Press, Cambridge, 2002.

[5] M. Belkin and P. Niyogi, Semisupervised learning on Riemannian manifolds, Machine Learning 56 (2004), 209–239.

[6] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Royal. Stat. Soc. 58 (1996), 267–288.

[7] H. Zou and T. Hastie, Regularization and variable selection via the elastic net, J. Royal. Stat. Soc. 67 (2005), 301–320.

[8] Y. Ying and D. X. Zhou, Online regularized classification algorithms, IEEE Trans. Inform. Theory 52 (2006), 4775–4788.

[9] J. H. Lin and D. X. Zhou, Learning theory of randomized Kaczmarz algorithm, J. Machine Learning Research 16 (2015), 3341–3365.

[10] S. B. Lin, X. Guo, and D. X. Zhou, Distributed learning with regularized least squares, J. Machine Learning Research, 2017, to appear.

[11] Z. C. Guo, S. B. Lin, and D. X. Zhou, Learning theory of distributed spectral algorithms, Inverse problems 33 (2017).