Lecture 3

Approximate Kernel Methods

(Computational vs. Statistical Trade-off)

Bharath K. Sriperumbudur* & Dougal J. Sutherland†

*Department of Statistics, Pennsylvania State University

†Gatsby Unit, University College London

Data Science Summer School
École Polytechnique
June 2019

Outline

- Motivating examples
  - Kernel ridge regression and kernel PCA
- Approximation methods
- Computational vs. statistical trade off

Kernel Ridge regression: Feature Map and Kernel Trick

- **Given:** $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$
- **Task:** Find a regressor $f \in \mathcal{H}$ (some feature space) s.t. $f(x_i) \approx y_i$.
- **Idea:** Map x_i to $\Phi(x_i)$ and do linear regression,

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (\langle f, \Phi(x_i) \rangle_{\mathcal{H}} - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \quad (\lambda > 0)$$

- **Solution:** For $\Phi(\mathbf{X}) := (\Phi(x_1), \ldots, \Phi(x_n)) \in \mathbb{R}^{\dim(\mathcal{H}) \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$f = \underbrace{\frac{1}{n} \left(\frac{1}{n} \Phi(\mathbf{X}) \Phi(\mathbf{X})^\top + \lambda I_{\dim(\mathcal{H})} \right)^{-1} \Phi(\mathbf{X}) \mathbf{y}}_{primal}$$

$$= \underbrace{\frac{1}{n} \Phi(\mathbf{X}) \left(\frac{1}{n} \Phi(\mathbf{X})^\top \Phi(\mathbf{X}) + \lambda I_n \right)^{-1} \mathbf{y}}_{dual}$$

Kernel Ridge regression: Feature Map and Kernel Trick

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$
- Task: Find a regressor $f \in \mathcal{H}$ (some feature space) s.t. $f(x_i) \approx y_i$.
- Idea: Map x_i to $\Phi(x_i)$ and do linear regression,

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (\langle f, \Phi(x_i) \rangle_{\mathcal{H}} - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \quad (\lambda > 0)$$

- Solution: For $\Phi(\mathbf{X}) := (\Phi(x_1), \ldots, \Phi(x_n)) \in \mathbb{R}^{\dim(\mathcal{H}) \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$f = \underbrace{\frac{1}{n} \left(\frac{1}{n} \Phi(\mathbf{X}) \Phi(\mathbf{X})^\top + \lambda I_{\dim(\mathcal{H})} \right)^{-1} \Phi(\mathbf{X}) \mathbf{y}}_{\text{primal}}$$

$$= \underbrace{\frac{1}{n} \Phi(\mathbf{X}) \left(\frac{1}{n} \Phi(\mathbf{X})^\top \Phi(\mathbf{X}) + \lambda I_n \right)^{-1} \mathbf{y}}_{\text{dual}}$$

Kernel Ridge regression: Feature Map and Kernel Trick

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$
- Task: Find a regressor $f \in \mathcal{H}$ (some feature space) s.t. $f(x_i) \approx y_i$.
- Idea: Map x_i to $\Phi(x_i)$ and do linear regression,

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (\langle f, \Phi(x_i) \rangle_{\mathcal{H}} - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \quad (\lambda > 0)$$

- Solution: For $\Phi(\mathbf{X}) := (\Phi(x_1), \ldots, \Phi(x_n)) \in \mathbb{R}^{\dim(\mathcal{H}) \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$f = \underbrace{\frac{1}{n} \left(\frac{1}{n} \Phi(\mathbf{X}) \Phi(\mathbf{X})^\top + \lambda I_{\dim(\mathcal{H})} \right)^{-1} \Phi(\mathbf{X}) \mathbf{y}}_{primal}$$

$$= \underbrace{\frac{1}{n} \Phi(\mathbf{X}) \left(\frac{1}{n} \Phi(\mathbf{X})^\top \Phi(\mathbf{X}) + \lambda I_n \right)^{-1} \mathbf{y}}_{dual}$$

Kernel Ridge regression: Feature Map and Kernel Trick

- Prediction: Given $t \in \mathcal{X}$

$$f(t) = \langle f, \Phi(t) \rangle_{\mathcal{H}} = \frac{1}{n}\mathbf{y}^\top \Phi(\mathbf{X})^\top \left(\frac{1}{n}\Phi(\mathbf{X})\Phi(\mathbf{X})^\top + \lambda I_{\dim(\mathcal{H})} \right)^{-1} \Phi(t)$$

$$= \frac{1}{n}\mathbf{y}^\top \left(\frac{1}{n}\Phi(\mathbf{X})^\top \Phi(\mathbf{X}) + \lambda I_n \right)^{-1} \Phi(\mathbf{X})^\top \Phi(t)$$

As before

$$\Phi(\mathbf{X})^\top \Phi(\mathbf{X}) = \underbrace{\begin{bmatrix} \langle \Phi(x_1), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_1), \Phi(x_n) \rangle_{\mathcal{H}} \\ \langle \Phi(x_2), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_2), \Phi(x_n) \rangle_{\mathcal{H}} \\ \vdots & \ddots & \vdots \\ \langle \Phi(x_n), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_n), \Phi(x_n) \rangle_{\mathcal{H}} \end{bmatrix}}_{k(x_i,x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}}}$$

and

$$\Phi(\mathbf{X})^\top \Phi(t) = \left[\langle \Phi(x_1), \Phi(t) \rangle_{\mathcal{H}}, \ldots, \langle \Phi(x_n), \Phi(t) \rangle_{\mathcal{H}} \right]^\top$$

Kernel Ridge regression: Feature Map and Kernel Trick

- Prediction: Given $t \in \mathcal{X}$

$$f(t) = \langle f, \Phi(t) \rangle_{\mathcal{H}} = \frac{1}{n}\mathbf{y}^\top \Phi(\mathbf{X})^\top \left(\frac{1}{n}\Phi(\mathbf{X})\Phi(\mathbf{X})^\top + \lambda I_{\dim(\mathcal{H})} \right)^{-1} \Phi(t)$$

$$= \frac{1}{n}\mathbf{y}^\top \left(\frac{1}{n}\Phi(\mathbf{X})^\top\Phi(\mathbf{X}) + \lambda I_n \right)^{-1} \Phi(\mathbf{X})^\top \Phi(t)$$

As before

$$\Phi(\mathbf{X})^\top\Phi(\mathbf{X}) = \underbrace{\begin{bmatrix} \langle \Phi(x_1), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_1), \Phi(x_n) \rangle_{\mathcal{H}} \\ \langle \Phi(x_2), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_2), \Phi(x_n) \rangle_{\mathcal{H}} \\ \vdots & \ddots & \vdots \\ \langle \Phi(x_n), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_n), \Phi(x_n) \rangle_{\mathcal{H}} \end{bmatrix}}_{k(x_i,x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}}}$$

and

$$\Phi(\mathbf{X})^\top\Phi(t) = [\langle \Phi(x_1), \Phi(t) \rangle_{\mathcal{H}}, \ldots, \langle \Phi(x_n), \Phi(t) \rangle_{\mathcal{H}}]^\top$$

Remarks

- The primal formulation requires the knowledge of feature map Φ (and of course $\mathcal{H}$) and these could be infinite dimensional.
- The dual formulation is entirely determined by kernel evaluations, Gram matrix and $(k(x_i, t))_i$. But poor scalability: $O(n^3)$.

Kernel Principal Component Analysis

- Dimensionality reduction
- Given: $\{(x_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$
- Task: Find a low-dimensional representation for (x_i).

$$\max_{\|f\|_{\mathcal{H}}=1} \mathrm{Var}\left(\langle f, \Phi(x_1)\rangle_{\mathcal{H}}, \langle f, \Phi(x_2)\rangle_{\mathcal{H}}, \ldots, \langle f, \Phi(x_n)\rangle_{\mathcal{H}}\right)$$

$$\equiv \max_{\|f\|_{\mathcal{H}}=1} \frac{1}{n}\sum_{i=1}^n \langle f, \Phi(x_i)\rangle_{\mathcal{H}}^2 - \left(\frac{1}{n}\sum_{i=1}^n \langle f, \Phi(x_i)\rangle_{\mathcal{H}}\right)^2$$

$$\equiv \max_{\|f\|_{\mathcal{H}}=1} \langle f, \hat{\boldsymbol{\Sigma}} f\rangle_{\mathcal{H}}$$

where

$$\hat{\boldsymbol{\Sigma}} := \frac{1}{n}\sum_{i=1}^n \Phi(x_i)\otimes\Phi(x_i) - \left(\frac{1}{n}\sum_{i=1}^n \Phi(x_i)\right)\otimes\left(\frac{1}{n}\sum_{i=1}^n \Phi(x_i)\right)$$

$$= \frac{1}{n}\Phi(\mathbf{X})\left(I_d - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)\Phi(\mathbf{X})^\top =: \Phi(\mathbf{X})\mathbf{H}\Phi(\mathbf{X})^\top.$$

Kernel Principal Component Analysis

- Dimensionality reduction
- Given: $\{(x_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$
- Task: Find a low-dimensional representation for (x_i).

$$\max_{\|f\|_{\mathcal{H}}=1} \mathrm{Var}\left(\langle f, \Phi(x_1)\rangle_{\mathcal{H}}, \langle f, \Phi(x_2)\rangle_{\mathcal{H}}, \ldots, \langle f, \Phi(x_n)\rangle_{\mathcal{H}}\right)$$

$$\equiv \max_{\|f\|_{\mathcal{H}}=1} \frac{1}{n}\sum_{i=1}^n \langle f, \Phi(x_i)\rangle_{\mathcal{H}}^2 - \left(\frac{1}{n}\sum_{i=1}^n \langle f, \Phi(x_i)\rangle_{\mathcal{H}}\right)^2$$

$$\equiv \max_{\|f\|_{\mathcal{H}}=1} \langle f, \hat{\boldsymbol{\Sigma}} f\rangle_{\mathcal{H}}$$

where

$$\hat{\boldsymbol{\Sigma}} := \frac{1}{n}\sum_{i=1}^n \Phi(x_i)\otimes\Phi(x_i) - \left(\frac{1}{n}\sum_{i=1}^n \Phi(x_i)\right)\otimes\left(\frac{1}{n}\sum_{i=1}^n \Phi(x_i)\right)$$

$$= \frac{1}{n}\Phi(\mathbf{X})\left(I_d - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)\Phi(\mathbf{X})^\top =: \Phi(\mathbf{X})\mathbf{H}\Phi(\mathbf{X})^\top.$$

Kernel Principal Component Analysis

- Dimensionality reduction
- Given: $\{(x_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$
- Task: Find a low-dimensional representation for (x_i).

$$
\begin{aligned}
&\max_{\|f\|_{\mathcal{H}}=1} \operatorname{Var}\left(\langle f, \Phi(x_1)\rangle_{\mathcal{H}}, \langle f, \Phi(x_2)\rangle_{\mathcal{H}}, \dots, \langle f, \Phi(x_n)\rangle_{\mathcal{H}}\right) \\
&\equiv \max_{\|f\|_{\mathcal{H}}=1} \frac{1}{n}\sum_{i=1}^n \langle f, \Phi(x_i)\rangle_{\mathcal{H}}^2 - \left(\frac{1}{n}\sum_{i=1}^n \langle f, \Phi(x_i)\rangle_{\mathcal{H}}\right)^2 \\
&\equiv \max_{\|f\|_{\mathcal{H}}=1} \langle f, \hat{\mathbf{\Sigma}} f\rangle_{\mathcal{H}}
\end{aligned}
$$

where

$$
\begin{aligned}
\hat{\mathbf{\Sigma}} &:= \frac{1}{n}\sum_{i=1}^n \Phi(x_i)\otimes\Phi(x_i) - \left(\frac{1}{n}\sum_{i=1}^n \Phi(x_i)\right)\otimes\left(\frac{1}{n}\sum_{i=1}^n \Phi(x_i)\right) \\
&= \frac{1}{n}\Phi(\mathbf{X})\left(I_d - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)\Phi(\mathbf{X})^\top =: \Phi(\mathbf{X})\mathbf{H}\Phi(\mathbf{X})^\top.
\end{aligned}
$$

Remarks

- The primal formulation requires the knowledge of feature map Φ (and of course $\mathcal{H}$) and these could be infinite dimensional.
- The dual formulation is entirely determined by kernel evaluations, Gram matrix and $(k(x_i, t))_i$. But poor scalability: $O(n^3)$.

Approximation Schemes

- Incomplete Cholesky factorization (Fine and Scheinberg, JMLR 2001)
- Sketching (Yang et al., 2015)
- Sparse greedy approximation (Smola and Schölkopf, NIPS 2000)
- Nyström method (Williams and Seeger, NIPS 2001)
- Random Fourier features (Rahimi and Recht, NIPS 2008), ...

Key Ideas

- Approach 1: Finite dimensional approximation to $\Phi(x)$
  - Perform linear method on $\Phi_m(x) \in \mathbb{R}^m$.
  - Involves $\Phi_m(\mathbf{X})\Phi_m(\mathbf{X})^\top \in \mathbb{R}^{m \times m}$.
- Approach 2: Approximate the representer
  - The representer theorem yields that the solution lies in

$$\left\{ f \in \mathcal{H} \middle| f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) : (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n \right\}$$

  - Instead, restrict the solution to

$$\left\{ f \in \mathcal{H} \middle| f = \sum_{i=1}^{m} \beta_i k(\cdot, \tilde{x}_i) : (\beta_1, \ldots, \beta_m) \in \mathbb{R}^m \right\}$$

Key Ideas

- Approach 1: Finite dimensional approximation to $\Phi(x)$
  - Perform linear method on $\Phi_m(x) \in \mathbb{R}^m$.
  - Involves $\Phi_m(\mathbf{X})\Phi_m(\mathbf{X})^\top \in \mathbb{R}^{m \times m}$.
- Approach 2: Approximate the representer
  - The representer theorem yields that the solution lies in

$$\left\{ f \in \mathcal{H} \middle| f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) : (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n \right\}$$

  - Instead, restrict the solution to

$$\left\{ f \in \mathcal{H} \middle| f = \sum_{i=1}^{m} \beta_i k(\cdot, \tilde{x}_i) : (\beta_1, \ldots, \beta_m) \in \mathbb{R}^m \right\}$$

Approach 1: Finite dimensional approximation to $\Phi(x)$

Random Fourier Approximation

- $\mathcal{X} = \mathbb{R}^d$; k be continuous and translation-invariant, i.e., $k(x, y) = \psi(x - y)$.
- Bochner's theorem: ψ is positive definite if and only if

$$k(x, y) = \int_{\mathbb{R}^d} e^{\sqrt{-1}\langle\omega, x-y\rangle_2} \, d\Lambda(\omega),$$

  where Λ is a finite non-negative Borel measure on $\mathbb{R}^d$.
  - k is symmetric and therefore Λ is a "symmetric" measure on $\mathbb{R}^d$.
- Therefore

$$k(x, y) = \int_{\mathbb{R}^d} \cos(\langle\omega, x - y\rangle_2) \, d\Lambda(\omega).$$

Random Fourier Approximation

- $\mathcal{X} = \mathbb{R}^d$; k be continuous and translation-invariant, i.e., $k(x, y) = \psi(x - y)$.
- Bochner's theorem: ψ is positive definite if and only if

$$k(x, y) = \int_{\mathbb{R}^d} e^{\sqrt{-1}\langle \omega, x-y \rangle_2} \, d\Lambda(\omega),$$

  where Λ is a finite non-negative Borel measure on $\mathbb{R}^d$.
  - k is symmetric and therefore Λ is a "symmetric" measure on $\mathbb{R}^d$.
- Therefore

$$k(x, y) = \int_{\mathbb{R}^d} \cos(\langle \omega, x - y \rangle_2) \, d\Lambda(\omega).$$

Random Feature Approximation

(Rahimi and Recht, 2008a): Draw $(\omega_j)_{j=1}^m \overset{i.i.d.}{\sim} \Lambda$.

$$k_m(x,y) = \frac{1}{m}\sum_{j=1}^m \cos\left(\langle \omega_j, x-y\rangle_2\right) = \langle \Phi_m(x), \Phi_m(y)\rangle_{\mathbb{R}^{2m}},$$

$$\approx k(x,y) = \underbrace{\langle k(\cdot,x), k(\cdot,y)\rangle_{\mathcal{H}}}_{\langle \Phi(x),\Phi(y)\rangle_{\mathcal{H}}}$$

where

$$\Phi_m(x) = \frac{1}{\sqrt{m}}(\overbrace{\cos(\langle\omega_1,x\rangle_2)}^{\varphi_1(x)},\ldots,\cos(\langle\omega_m,x\rangle_2),\sin(\langle\omega_1,x\rangle_2),\ldots,\sin(\langle\omega_m,x\rangle_2)).$$

How good is the approximation?

(S and Szabó, NIPS 2016):

$$\sup_{x,y\in\mathscr{S}} |k_m(x,y) - k(x,y)| = O_{a.s.}\left(\sqrt{\frac{\log|\mathscr{S}|}{m}}\right)$$

Optimal convergence rate

- Other results are known but they are non-optimal (Rahimi and Recht, NIPS 2008; Sutherland and Schneider, UAI 2015).

Ridge Regression: Random Feature Approximation

- **Given:** $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$
- **Task:** Find a regressor f s.t. $f(x_i) \approx y_i$.
- **Idea:** Map x_i to $\Phi_m(x_i)$ and do linear regression,

$$\min_{w \in \mathbb{R}^{2m}} \frac{1}{n} \sum_{i=1}^{n} (\langle w, \Phi_m(x_i) \rangle_{\mathbb{R}^{2m}} - y_i)^2 + \lambda \|w\|^2_{\mathbb{R}^{2m}} \quad (\lambda > 0)$$

- **Solution:** For $\Phi_m(\mathbf{X}) := (\Phi_m(x_1), \ldots, \Phi_m(x_n)) \in \mathbb{R}^{2m \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$f = \underbrace{\frac{1}{n} \left(\frac{1}{n} \Phi_m(\mathbf{X}) \Phi_m(\mathbf{X})^\top + \lambda I_{2m} \right)^{-1} \Phi_m(\mathbf{X}) \mathbf{y}}_{primal}$$

$$= \underbrace{\frac{1}{n} \Phi_m(\mathbf{X}) \left(\frac{1}{n} \Phi_m(\mathbf{X})^\top \Phi_m(\mathbf{X}) + \lambda I_n \right)^{-1} \mathbf{y}}_{dual}$$

Computation: $O(m^2 n)$.

Ridge Regression: Random Feature Approximation

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$
- Task: Find a regressor f s.t. $f(x_i) \approx y_i$.
- Idea: Map x_i to $\Phi_m(x_i)$ and do linear regression,

$$\min_{w \in \mathbb{R}^{2m}} \frac{1}{n} \sum_{i=1}^{n} (\langle w, \Phi_m(x_i) \rangle_{\mathbb{R}^{2m}} - y_i)^2 + \lambda \|w\|_{\mathbb{R}^{2m}}^2 \quad (\lambda > 0)$$

- Solution: For $\Phi_m(\mathbf{X}) := (\Phi_m(x_1), \ldots, \Phi_m(x_n)) \in \mathbb{R}^{2m \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$f = \underbrace{\frac{1}{n} \left(\frac{1}{n} \Phi_m(\mathbf{X}) \Phi_m(\mathbf{X})^\top + \lambda I_{2m} \right)^{-1} \Phi_m(\mathbf{X}) \mathbf{y}}_{primal}$$

$$= \underbrace{\frac{1}{n} \Phi_m(\mathbf{X}) \left(\frac{1}{n} \Phi_m(\mathbf{X})^\top \Phi_m(\mathbf{X}) + \lambda I_n \right)^{-1} \mathbf{y}}_{dual}$$

Computation: $O(m^2 n)$.

Ridge Regression: Random Feature Approximation

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$
- Task: Find a regressor f s.t. $f(x_i) \approx y_i$.
- Idea: Map x_i to $\Phi_m(x_i)$ and do linear regression,

$$\min_{w \in \mathbb{R}^{2m}} \frac{1}{n} \sum_{i=1}^{n} (\langle w, \Phi_m(x_i) \rangle_{\mathbb{R}^{2m}} - y_i)^2 + \lambda \|w\|_{\mathbb{R}^{2m}}^2 \quad (\lambda > 0)$$

- Solution: For $\Phi_m(\mathbf{X}) := (\Phi_m(x_1), \ldots, \Phi_m(x_n)) \in \mathbb{R}^{2m \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$f = \underbrace{\frac{1}{n} \left(\frac{1}{n} \Phi_m(\mathbf{X}) \Phi_m(\mathbf{X})^\top + \lambda I_{2m} \right)^{-1} \Phi_m(\mathbf{X}) \mathbf{y}}_{primal}$$

$$= \underbrace{\frac{1}{n} \Phi_m(\mathbf{X}) \left(\frac{1}{n} \Phi_m(\mathbf{X})^\top \Phi_m(\mathbf{X}) + \lambda I_n \right)^{-1} \mathbf{y}}_{dual}$$

Computation: $O(m^2 n)$.

What happens statistically?

Kernel ridge regression: $(X_i, Y_i)_{i=1}^n \stackrel{iid}{\sim} \rho_{XY}$.

- $\mathcal{R}^*_{\mathbf{P}} = \inf_{f \in L^2(\rho_X)} \mathbb{E}|f(X) - Y|^2 = \mathbb{E}|f^*(X) - Y|^2$.
- Penalized risk minimization: $O(n^3)$

$$f_n = \arg \inf_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(X_i)|_2^2 + \lambda \|f\|_{\mathcal{H}}^2$$

- Penalized risk minimization (approximate): $O(m^2 n)$

$$f_{m,n} = \arg \inf_{f \in \mathcal{H}_m} \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(X_i)|_2^2 + \lambda \|f\|_{\mathcal{H}_m}^2$$

What happens statistically?

$$\underbrace{\mathcal{R}_{\mathbf{P}}(f_{m,n})}_{\mathbb{E}|f_{m,n}(X)-Y|^2} - \mathcal{R}^*$$

$$= \underbrace{(\mathcal{R}_{\mathbf{P}}(f_{m,n}) - \mathcal{R}_{\mathbf{P}}(f_n))}_{\text{error due to approximation}} + (\mathcal{R}_{\mathbf{P}}(f_n) - \mathcal{R}^*_{\mathbf{P}})$$

- (Rahimi and Recht, 2008b): $(m \wedge n)^{-\frac{1}{2}}$
- (Rudi and Rosasco, 2016): If $m \geq n^{\alpha}$ where $\frac{1}{2} \leq \alpha < 1$ with α depending on the properties of f^*, then $f_{m,n}$ achieves the minimax optimal rate as obtained in the case with no approximation.

Computational gain with no statistical loss!!

What happens statistically?

$$\underbrace{\mathcal{R}_{\mathbf{P}}(f_{m,n})}_{\mathbb{E}|f_{m,n}(X)-Y|^2} - \mathcal{R}^*$$

$$= \underbrace{(\mathcal{R}_{\mathbf{P}}(f_{m,n}) - \mathcal{R}_{\mathbf{P}}(f_n))}_{\text{error due to approximation}} + (\mathcal{R}_{\mathbf{P}}(f_n) - \mathcal{R}^*_{\mathbf{P}})$$

- (Rahimi and Recht, 2008b): $(m \wedge n)^{-\frac{1}{2}}$
- (Rudi and Rosasco, 2016): If $m \geq n^{\alpha}$ where $\frac{1}{2} \leq \alpha < 1$ with α depending on the properties of f^*, then $f_{m,n}$ achieves the minimax optimal rate as obtained in the case with no approximation.

Computational gain with no statistical loss!!

What happens statistically?

$$\underbrace{\mathcal{R}_{\mathbf{P}}(f_{m,n})}_{\mathbb{E}|f_{m,n}(X)-Y|^2} - \mathcal{R}^*$$

$$= \underbrace{(\mathcal{R}_{\mathbf{P}}(f_{m,n}) - \mathcal{R}_{\mathbf{P}}(f_n))}_{\text{error due to approximation}} + (\mathcal{R}_{\mathbf{P}}(f_n) - \mathcal{R}^*_{\mathbf{P}})$$

- (Rahimi and Recht, 2008b): $(m \wedge n)^{-\frac{1}{2}}$
- (Rudi and Rosasco, 2016): If $m \geq n^{\alpha}$ where $\frac{1}{2} \leq \alpha < 1$ with α depending on the properties of f^*, then $f_{m,n}$ achieves the minimax optimal rate as obtained in the case with no approximation.

Computational gain with no statistical loss!!

Kernel PCA: Random Feature Approximation

- Perform linear PCA on $\{\Phi_m(x_i)\}_{i=1}^n$.
- Approximate KPCA finds $\beta \in \mathbb{R}^m$ that solves

$$\sup_{\|\beta\|_2=1} \mathrm{Var}[\{\langle \beta, \Phi_m(x_i)\rangle_2\}_{i=1}^n] = \sup_{\|\beta\|_2=1} \left\langle \beta, \hat{\boldsymbol{\Sigma}}_m \beta \right\rangle_2$$

where

$$\hat{\boldsymbol{\Sigma}}_m := \frac{1}{n}\sum_{i=1}^n \Phi_m(x_i) \otimes \Phi_m(x_i) - \left(\frac{1}{n}\sum_{i=1}^n \Phi_m(x_i)\right) \otimes \left(\frac{1}{n}\sum_{i=1}^n \Phi_m(x_i)\right).$$

- Same as doing kernel PCA in $\mathcal{H}_m$ where

$$\mathcal{H}_m = \left\{ f = \sum_{i=1}^m \beta_i \varphi_i \; : \; \beta \in \mathbb{R}^m \right\}$$

is an RKHS induced by the reproducing kernel k_m.

- Computation: $O(m^2 n)$

Kernel PCA: Random Feature Approximation

- Perform linear PCA on $\{\Phi_m(x_i)\}_{i=1}^n$.
- Approximate KPCA finds $\beta \in \mathbb{R}^m$ that solves

$$\sup_{\|\beta\|_2=1} \mathrm{Var}[\{\langle \beta, \Phi_m(x_i)\rangle_2\}_{i=1}^n] = \sup_{\|\beta\|_2=1} \left\langle \beta, \hat{\boldsymbol{\Sigma}}_m \beta \right\rangle_2$$

where

$$\hat{\boldsymbol{\Sigma}}_m := \frac{1}{n}\sum_{i=1}^n \Phi_m(x_i) \otimes \Phi_m(x_i) - \left(\frac{1}{n}\sum_{i=1}^n \Phi_m(x_i)\right) \otimes \left(\frac{1}{n}\sum_{i=1}^n \Phi_m(x_i)\right).$$

- Same as doing kernel PCA in $\mathcal{H}_m$ where

$$\mathcal{H}_m = \left\{ f = \sum_{i=1}^m \beta_i \varphi_i \;:\; \beta \in \mathbb{R}^m \right\}$$

is an RKHS induced by the reproducing kernel k_m.

- Computation: $O(m^2 n)$

Kernel PCA: Random Feature Approximation

- Perform linear PCA on $\{\Phi_m(x_i)\}_{i=1}^n$.
- Approximate KPCA finds $\beta \in \mathbb{R}^m$ that solves

$$\sup_{\|\beta\|_2=1} \mathrm{Var}[\{\langle \beta, \Phi_m(x_i)\rangle_2\}_{i=1}^n] = \sup_{\|\beta\|_2=1} \left\langle \beta, \hat{\mathbf{\Sigma}}_m \beta \right\rangle_2$$

  where

$$\hat{\mathbf{\Sigma}}_m := \frac{1}{n}\sum_{i=1}^n \Phi_m(x_i) \otimes \Phi_m(x_i) - \left(\frac{1}{n}\sum_{i=1}^n \Phi_m(x_i)\right) \otimes \left(\frac{1}{n}\sum_{i=1}^n \Phi_m(x_i)\right).$$

- Same as doing kernel PCA in $\mathcal{H}_m$ where

$$\mathcal{H}_m = \left\{ f = \sum_{i=1}^m \beta_i \varphi_i \,:\, \beta \in \mathbb{R}^m \right\}$$

  is an RKHS induced by the reproducing kernel k_m.
- Computation: $O(m^2 n)$

Reconstruction Error

- Linear PCA

$$\mathbb{E}_{X\sim\mathbb{P}}\left\|(X-\mu)-\sum_{i=1}^{\ell}\langle (X-\mu),\phi_i\rangle_2\phi_i\right\|_2^2$$

- Kernel PCA

$$\mathbb{E}_{X\sim\mathbb{P}}\left\|\tilde{k}(\cdot,X)-\sum_{i=1}^{\ell}\langle \tilde{k}(\cdot,X),\phi_i\rangle_{\mathcal{H}}\phi_i\right\|_{\mathcal{H}}^2$$

where $\tilde{k}(\cdot,x)=k(\cdot,x)-\int k(\cdot,x)\,d\mathbb{P}(x)$.

- However, the eigenfunctions of approximate empirical KPCA lie in $\mathcal{H}_m$, which is finite dimensional and not contained in $\mathcal{H}$.

Reconstruction Error

- Linear PCA

$$\mathbb{E}_{X\sim\mathbb{P}}\left\|(X-\mu)-\sum_{i=1}^{\ell}\langle (X-\mu),\phi_i\rangle_2\phi_i\right\|_2^2$$

- Kernel PCA

$$\mathbb{E}_{X\sim\mathbb{P}}\left\|\tilde{k}(\cdot,X)-\sum_{i=1}^{\ell}\langle \tilde{k}(\cdot,X),\phi_i\rangle_{\mathcal{H}}\phi_i\right\|_{\mathcal{H}}^2$$

where $\tilde{k}(\cdot,x)=k(\cdot,x)-\int k(\cdot,x)\,d\mathbb{P}(x)$.

- However, the eigenfunctions of approximate empirical KPCA lie in $\mathcal{H}_m$, which is finite dimensional and not contained in $\mathcal{H}$.

Reconstruction Error

- Linear PCA

$$\mathbb{E}_{X\sim\mathbb{P}}\left\|(X-\mu)-\sum_{i=1}^{\ell}\langle(X-\mu),\phi_i\rangle_2\phi_i\right\|_2^2$$

- Kernel PCA

$$\mathbb{E}_{X\sim\mathbb{P}}\left\|\tilde{k}(\cdot,X)-\sum_{i=1}^{\ell}\langle\tilde{k}(\cdot,X),\phi_i\rangle_{\mathcal{H}}\phi_i\right\|_{\mathcal{H}}^2$$

where $\tilde{k}(\cdot,x)=k(\cdot,x)-\int k(\cdot,x)\,d\mathbb{P}(x)$.

- However, the eigenfunctions of approximate empirical KPCA lie in $\mathcal{H}_m$, which is finite dimensional and not contained in $\mathcal{H}$.

Embedding to $L^2(\mathbb{P})$ (S and Sterge, 2017)

What we have?

- Population eigenfunctions $(\phi_i)_{i\in I}$ of Σ: these form a subspace in $\mathcal{H}$.
- Empirical eigenfunctions $(\hat{\phi}_i)_{i=1}^n$ of $\hat{\Sigma}$: these form a subspace in $\mathcal{H}$.
- Eigenvectors after approximation, $(\hat{\phi}_{i,m})_{i=1}^m$ of $\hat{\Sigma}_m$: these form a subspace in $\mathbb{R}^m$
- We embed them in a common space before comparing. The common space is $L^2(\mathbb{P})$.
- (Inclusion operator) $\mathcal{I} : \mathcal{H} \to L^2(\mathbb{P})$, $f \mapsto f - \int_{\mathcal{X}} f(x)\, d\mathbb{P}(x)$
- (Approximation operator) $\mathcal{U} : \mathbb{R}^m \to L^2(\mathbb{P})$,

$$\boldsymbol{\alpha} \mapsto \sum_{i=1}^{m} \alpha_i \left(\varphi_i - \int_{\mathcal{X}} \varphi_i(x)\, d\mathbb{P}(x) \right)$$

Embedding to $L^2(\mathbb{P})$ (S and Sterge, 2017)

What we have?

- Population eigenfunctions $(\phi_i)_{i\in I}$ of Σ: these form a subspace in $\mathcal{H}$.
- Empirical eigenfunctions $(\hat{\phi}_i)_{i=1}^n$ of $\hat{\Sigma}$: these form a subspace in $\mathcal{H}$.
- Eigenvectors after approximation, $(\hat{\phi}_{i,m})_{i=1}^m$ of $\hat{\Sigma}_m$: these form a subspace in $\mathbb{R}^m$
- We embed them in a common space before comparing. The common space is $L^2(\mathbb{P})$.

- (Inclusion operator) $\mathcal{I} : \mathcal{H} \to L^2(\mathbb{P})$, $f \mapsto f - \int_{\mathcal{X}} f(x)\, d\mathbb{P}(x)$
- (Approximation operator) $\mathcal{U} : \mathbb{R}^m \to L^2(\mathbb{P})$,

$$\boldsymbol{\alpha} \mapsto \sum_{i=1}^{m} \alpha_i \left(\varphi_i - \int_{\mathcal{X}} \varphi_i(x)\, d\mathbb{P}(x) \right)$$

Reconstruction Error (S and Sterge, 2017)

$\left(\frac{\mathcal{I}\phi_i}{\sqrt{\lambda_i}}\right)_{i=1}^{\infty}$ form an ONS for $L^2(\mathbb{P})$. Define $\tilde{k}(\cdot,x) = k(\cdot,x) - \mu_{\mathbb{P}}$ and $\tau > 0$.

- Population KPCA:

$$R_\ell = \mathbb{E}\left\|\mathcal{I}\tilde{k}(\cdot,X) - \sum_{i=1}^{\ell}\left\langle \frac{\mathcal{I}\phi_i}{\sqrt{\lambda_i}}, \mathcal{I}\tilde{k}(\cdot,X)\right\rangle_{L^2(\mathbb{P})} \frac{\mathcal{I}\phi_i}{\sqrt{\lambda_i}}\right\|_{L^2(\mathbb{P})}^2$$

- Empirical KPCA:

$$R_{n,\ell} = \mathbb{E}\left\|\mathcal{I}\tilde{k}(\cdot,X) - \sum_{i=1}^{\ell}\left\langle \frac{\mathcal{I}\hat{\phi}_i}{\sqrt{\hat{\lambda}_i}}, \mathcal{I}\tilde{k}(\cdot,X)\right\rangle_{L^2(\mathbb{P})} \frac{\mathcal{I}\hat{\phi}_i}{\sqrt{\hat{\lambda}_i}}\right\|_{L^2(\mathbb{P})}^2$$

- Approximate Empirical KPCA:

$$R_{m,n,\ell} = \mathbb{E}\left\|\mathcal{I}\tilde{k}(\cdot,X) - \sum_{i=1}^{\ell}\left\langle \frac{\mathcal{U}\hat{\phi}_{i,m}}{\sqrt{\hat{\lambda}_{i,m}}}, \mathcal{I}\tilde{k}(\cdot,X)\right\rangle_{L^2(\mathbb{P})} \frac{\mathcal{U}\hat{\phi}_{i,m}}{\sqrt{\hat{\lambda}_{i,m}}}\right\|_{L^2(\mathbb{P})}^2$$

Reconstruction Error (S and Sterge, 2017)

$\left(\frac{\mathcal{I}\phi_i}{\sqrt{\lambda_i}}\right)_{i=1}^{\infty}$ form an ONS for $L^2(\mathbb{P})$. Define $\tilde{k}(\cdot,x) = k(\cdot,x) - \mu_{\mathbb{P}}$ and $\tau > 0$.

- Population KPCA:

$$R_\ell = \mathbb{E}\left\|\mathcal{I}\tilde{k}(\cdot,X) - \sum_{i=1}^{\ell}\left\langle \frac{\mathcal{I}\phi_i}{\sqrt{\lambda_i}}, \mathcal{I}\tilde{k}(\cdot,X)\right\rangle_{L^2(\mathbb{P})} \frac{\mathcal{I}\phi_i}{\sqrt{\lambda_i}}\right\|_{L^2(\mathbb{P})}^2$$

- Empirical KPCA:

$$R_{n,\ell} = \mathbb{E}\left\|\mathcal{I}\tilde{k}(\cdot,X) - \sum_{i=1}^{\ell}\left\langle \frac{\mathcal{I}\hat{\phi}_i}{\sqrt{\hat{\lambda}_i}}, \mathcal{I}\tilde{k}(\cdot,X)\right\rangle_{L^2(\mathbb{P})} \frac{\mathcal{I}\hat{\phi}_i}{\sqrt{\hat{\lambda}_i}}\right\|_{L^2(\mathbb{P})}^2$$

- Approximate Empirical KPCA:

$$R_{m,n,\ell} = \mathbb{E}\left\|\mathcal{I}\tilde{k}(\cdot,X) - \sum_{i=1}^{\ell}\left\langle \frac{\mathcal{U}\hat{\phi}_{i,m}}{\sqrt{\hat{\lambda}_{i,m}}}, \mathcal{I}\tilde{k}(\cdot,X)\right\rangle_{L^2(\mathbb{P})} \frac{\mathcal{U}\hat{\phi}_{i,m}}{\sqrt{\hat{\lambda}_{i,m}}}\right\|_{L^2(\mathbb{P})}^2$$

Reconstruction Error (S and Sterge, 2017)

$\left(\frac{\mathcal{I}\phi_i}{\sqrt{\lambda_i}}\right)_{i=1}^{\infty}$ form an ONS for $L^2(\mathbb{P})$. Define $\tilde{k}(\cdot,x) = k(\cdot,x) - \mu_{\mathbb{P}}$ and $\tau > 0$.

- Population KPCA:

$$R_\ell = \mathbb{E}\left\|\mathcal{I}\tilde{k}(\cdot,X) - \sum_{i=1}^{\ell}\left\langle \frac{\mathcal{I}\phi_i}{\sqrt{\lambda_i}}, \mathcal{I}\tilde{k}(\cdot,X)\right\rangle_{L^2(\mathbb{P})} \frac{\mathcal{I}\phi_i}{\sqrt{\lambda_i}}\right\|_{L^2(\mathbb{P})}^2$$

- Empirical KPCA:

$$R_{n,\ell} = \mathbb{E}\left\|\mathcal{I}\tilde{k}(\cdot,X) - \sum_{i=1}^{\ell}\left\langle \frac{\mathcal{I}\hat{\phi}_i}{\sqrt{\hat{\lambda}_i}}, \mathcal{I}\tilde{k}(\cdot,X)\right\rangle_{L^2(\mathbb{P})} \frac{\mathcal{I}\hat{\phi}_i}{\sqrt{\hat{\lambda}_i}}\right\|_{L^2(\mathbb{P})}^2$$

- Approximate Empirical KPCA:

$$R_{m,n,\ell} = \mathbb{E}\left\|\mathcal{I}\tilde{k}(\cdot,X) - \sum_{i=1}^{\ell}\left\langle \frac{\mathcal{U}\hat{\phi}_{i,m}}{\sqrt{\hat{\lambda}_{i,m}}}, \mathcal{I}\tilde{k}(\cdot,X)\right\rangle_{L^2(\mathbb{P})} \frac{\mathcal{U}\hat{\phi}_{i,m}}{\sqrt{\hat{\lambda}_{i,m}}}\right\|_{L^2(\mathbb{P})}^2$$

Result (S and Sterge, 2017)

Clearly $R_\ell \to 0$ as $\ell \to \infty$. The goal is to study the convergence rates for R_ℓ, $R_{n,\ell}$ and $R_{m,n,\ell}$ as $\ell, m, n \to \infty$.

Suppose $\lambda_i \asymp i^{-\alpha}$, $\alpha > \frac{1}{2}$, $\ell = n^{\frac{\theta}{\alpha}}$ and $m = n^\gamma$ where $\theta > 0$ and $0 < \gamma < 1$.

- $n^{-2\theta(1-\frac{1}{2\alpha})} \lesssim R_\ell \lesssim n^{-2\theta(1-\frac{1}{2\alpha})}$
- $$R_{n,\ell} \lesssim \begin{cases} n^{-2\theta(1-\frac{1}{2\alpha})}, & 0 < \theta \le \frac{\alpha}{2(3\alpha-1)} \quad (\textit{bias dominates}) \\ n^{-(\frac{1}{2}-\theta)}, & \frac{\alpha}{2(3\alpha-1)} \le \theta < \frac{1}{2} \quad (\textit{variance dominates}) \end{cases}$$
- $$R_{m,n,\ell} \lesssim \begin{cases} n^{-2\theta(1-\frac{1}{2\alpha})}, & 0 < \theta \le \frac{\alpha}{2(3\alpha-1)} \\ n^{-(\frac{1}{2}-\theta)}, & \frac{\alpha}{2(3\alpha-1)} \le \theta < \frac{1}{2} \end{cases}$$

  for $\gamma > 2\theta$.

No statistical loss

Result (S and Sterge, 2017)

Clearly $R_\ell \to 0$ as $\ell \to \infty$. The goal is to study the convergence rates for R_ℓ, $R_{n,\ell}$ and $R_{m,n,\ell}$ as $\ell, m, n \to \infty$.

Suppose $\lambda_i \asymp i^{-\alpha}$, $\alpha > \frac{1}{2}$, $\ell = n^{\frac{\theta}{\alpha}}$ and $m = n^\gamma$ where $\theta > 0$ and $0 < \gamma < 1$.

- $n^{-2\theta(1-\frac{1}{2\alpha})} \lesssim R_\ell \lesssim n^{-2\theta(1-\frac{1}{2\alpha})}$
- $$R_{n,\ell} \lesssim \begin{cases} n^{-2\theta(1-\frac{1}{2\alpha})}, & 0 < \theta \leq \frac{\alpha}{2(3\alpha-1)} \quad \textit{(bias dominates)} \\ n^{-(\frac{1}{2}-\theta)}, & \frac{\alpha}{2(3\alpha-1)} \leq \theta < \frac{1}{2} \quad \textit{(variance dominates)} \end{cases}$$
- $$R_{m,n,\ell} \lesssim \begin{cases} n^{-2\theta(1-\frac{1}{2\alpha})}, & 0 < \theta \leq \frac{\alpha}{2(3\alpha-1)} \\ n^{-(\frac{1}{2}-\theta)}, & \frac{\alpha}{2(3\alpha-1)} \leq \theta < \frac{1}{2} \end{cases}$$

  for $\gamma > 2\theta$.

No statistical loss

Result (S and Sterge, 2017)

Clearly $R_\ell \to 0$ as $\ell \to \infty$. The goal is to study the convergence rates for R_ℓ, $R_{n,\ell}$ and $R_{m,n,\ell}$ as $\ell, m, n \to \infty$.

Suppose $\lambda_i \asymp i^{-\alpha}$, $\alpha > \frac{1}{2}$, $\ell = n^{\frac{\theta}{\alpha}}$ and $m = n^\gamma$ where $\theta > 0$ and $0 < \gamma < 1$.

- $n^{-2\theta\left(1-\frac{1}{2\alpha}\right)} \lesssim R_\ell \lesssim n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}$
-
$$R_{n,\ell} \lesssim \begin{cases} n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}, & 0 < \theta \le \frac{\alpha}{2(3\alpha-1)} \quad \textit{(bias dominates)} \\ n^{-\left(\frac{1}{2}-\theta\right)}, & \frac{\alpha}{2(3\alpha-1)} \le \theta < \frac{1}{2} \quad \textit{(variance dominates)} \end{cases}$$
-
$$R_{m,n,\ell} \lesssim \begin{cases} n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}, & 0 < \theta \le \frac{\alpha}{2(3\alpha-1)} \\ n^{-\left(\frac{1}{2}-\theta\right)}, & \frac{\alpha}{2(3\alpha-1)} \le \theta < \frac{1}{2} \end{cases}$$
  for $\gamma > 2\theta$.

No statistical loss

Result (S and Sterge, 2017)

Clearly $R_\ell \to 0$ as $\ell \to \infty$. The goal is to study the convergence rates for R_ℓ, $R_{n,\ell}$ and $R_{m,n,\ell}$ as $\ell, m, n \to \infty$.

Suppose $\lambda_i \asymp i^{-\alpha}$, $\alpha > \frac{1}{2}$, $\ell = n^{\frac{\theta}{\alpha}}$ and $m = n^\gamma$ where $\theta > 0$ and $0 < \gamma < 1$.

- $n^{-2\theta\left(1-\frac{1}{2\alpha}\right)} \lesssim R_\ell \lesssim n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}$
- $$R_{n,\ell} \lesssim \begin{cases} n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}, & 0 < \theta \leq \frac{\alpha}{2(3\alpha-1)} \quad (\textit{bias dominates}) \\ n^{-\left(\frac{1}{2}-\theta\right)}, & \frac{\alpha}{2(3\alpha-1)} \leq \theta < \frac{1}{2} \quad (\textit{variance dominates}) \end{cases}$$
- $$R_{m,n,\ell} \lesssim \begin{cases} n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}, & 0 < \theta \leq \frac{\alpha}{2(3\alpha-1)} \\ n^{-\left(\frac{1}{2}-\theta\right)}, & \frac{\alpha}{2(3\alpha-1)} \leq \theta < \frac{1}{2} \end{cases}$$

  for $\gamma > 2\theta$.

No statistical loss

Result (S and Sterge, 2017)

Clearly $R_\ell \to 0$ as $\ell \to \infty$. The goal is to study the convergence rates for R_ℓ, $R_{n,\ell}$ and $R_{m,n,\ell}$ as $\ell, m, n \to \infty$.

Suppose $\lambda_i \asymp i^{-\alpha}$, $\alpha > \frac{1}{2}$, $\ell = n^{\frac{\theta}{\alpha}}$ and $m = n^\gamma$ where $\theta > 0$ and $0 < \gamma < 1$.

- $n^{-2\theta\left(1-\frac{1}{2\alpha}\right)} \lesssim R_\ell \lesssim n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}$
-
$$R_{n,\ell} \lesssim \begin{cases} n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}, & 0 < \theta \leq \frac{\alpha}{2(3\alpha-1)} \quad \textit{(bias dominates)} \\ n^{-\left(\frac{1}{2}-\theta\right)}, & \frac{\alpha}{2(3\alpha-1)} \leq \theta < \frac{1}{2} \quad \textit{(variance dominates)} \end{cases}$$
-
$$R_{m,n,\ell} \lesssim \begin{cases} n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}, & 0 < \theta \leq \frac{\alpha}{2(3\alpha-1)} \\ n^{-\left(\frac{1}{2}-\theta\right)}, & \frac{\alpha}{2(3\alpha-1)} \leq \theta < \frac{1}{2} \end{cases}$$

  for $\gamma > 2\theta$.

No statistical loss

Approach 2: Approximate the representer

Ridge Regression: Nyström Approximation

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$
- Task: Find a regressor f s.t. $f(x_i) \approx y_i$.
- Idea: Restrict f to $\mathcal{H}_m = \{f \in \mathcal{H} : f = \sum_{i=1}^m \alpha_i k(\cdot, \tilde{x}_i)\}$ and do kernel ridge regression,

$$\min_{f \in \mathcal{H}_m} \frac{1}{n} \sum_{i=1}^n (\langle f, k(\cdot, x_i) \rangle_{\mathcal{H}} - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \quad (\lambda > 0)$$

- Solution: Since $f \in \mathcal{H}_m$,

$$\min_{\alpha \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^n ([\tilde{\mathbf{K}}_{nm}\alpha]_i - y_i)^2 + \lambda \alpha^\top \tilde{\mathbf{K}}_{mm}\alpha \quad (\lambda > 0)$$

where $[\tilde{\mathbf{K}}_{mm}]_{ij} = k(\tilde{x}_i, \tilde{x}_j)$ and $[\tilde{\mathbf{K}}_{mn}]_{ij} = k(\tilde{x}_i, x_j)$. Therefore

$$\alpha = (\tilde{\mathbf{K}}_{nm}^\top \tilde{\mathbf{K}}_{nm} + n\lambda \tilde{\mathbf{K}}_{mm})^{-1} \tilde{\mathbf{K}}_{nm}^\top \mathbf{y}.$$

Computation: $O(m^2 n)$.

Ridge Regression: Nyström Approximation

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$
- Task: Find a regressor f s.t. $f(x_i) \approx y_i$.
- Idea: Restrict f to $\mathcal{H}_m = \{f \in \mathcal{H} : f = \sum_{i=1}^m \alpha_i k(\cdot, \tilde{x}_i)\}$ and do kernel ridge regression,

$$\min_{f \in \mathcal{H}_m} \frac{1}{n} \sum_{i=1}^n (\langle f, k(\cdot, x_i) \rangle_{\mathcal{H}} - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \quad (\lambda > 0)$$

- Solution: Since $f \in \mathcal{H}_m$,

$$\min_{\alpha \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^n ([\tilde{\mathbf{K}}_{nm}\alpha]_i - y_i)^2 + \lambda \alpha^\top \tilde{\mathbf{K}}_{mm} \alpha \quad (\lambda > 0)$$

where $[\tilde{\mathbf{K}}_{mm}]_{ij} = k(\tilde{x}_i, \tilde{x}_j)$ and $[\tilde{\mathbf{K}}_{mn}]_{ij} = k(\tilde{x}_i, x_j)$. Therefore

$$\alpha = (\tilde{\mathbf{K}}_{nm}^\top \tilde{\mathbf{K}}_{nm} + n\lambda \tilde{\mathbf{K}}_{mm})^{-1} \tilde{\mathbf{K}}_{nm}^\top \mathbf{y}.$$

Computation: $O(m^2 n)$.

Ridge Regression: Nyström Approximation

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$
- Task: Find a regressor f s.t. $f(x_i) \approx y_i$.
- Idea: Restrict f to $\mathcal{H}_m = \{f \in \mathcal{H} : f = \sum_{i=1}^m \alpha_i k(\cdot, \tilde{x}_i)\}$ and do kernel ridge regression,

$$\min_{f \in \mathcal{H}_m} \frac{1}{n} \sum_{i=1}^n (\langle f, k(\cdot, x_i)\rangle_{\mathcal{H}} - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \quad (\lambda > 0)$$

- Solution: Since $f \in \mathcal{H}_m$,

$$\min_{\alpha \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^n ([\tilde{\mathbf{K}}_{nm}\alpha]_i - y_i)^2 + \lambda \alpha^\top \tilde{\mathbf{K}}_{mm}\alpha \quad (\lambda > 0)$$

where $[\tilde{\mathbf{K}}_{mm}]_{ij} = k(\tilde{x}_i, \tilde{x}_j)$ and $[\tilde{\mathbf{K}}_{mn}]_{ij} = k(\tilde{x}_i, x_j)$. Therefore

$$\alpha = (\tilde{\mathbf{K}}_{nm}^\top \tilde{\mathbf{K}}_{nm} + n\lambda \tilde{\mathbf{K}}_{mm})^{-1} \tilde{\mathbf{K}}_{nm}^\top \mathbf{y}.$$

Computation: $O(m^2 n)$.

Kernel PCA: Nyström Approximation

- $$\arg\sup\left\{\left\langle f, \hat{\boldsymbol{\Sigma}} f\right\rangle_{\mathcal{H}} : f \in \mathcal{H}_m,\ \|f\|_{\mathcal{H}} = 1\right\},$$

  where

  $$\mathcal{H}_m := \left\{f \in \mathcal{H} : f = \sum_{i=1}^{m} \beta_i k(\cdot, \tilde{x}_i) : (\beta_1, \ldots, \beta_m) \in \mathbb{R}^m\right\}.$$

- $$\boldsymbol{\beta} = \tilde{\mathbf{K}}_{mm}^{-1/2}\boldsymbol{u}$$

  where $\boldsymbol{u}$ is an eigenvector of $\frac{1}{n}\tilde{\mathbf{K}}_{mm}^{-1/2}\tilde{\mathbf{K}}_{nm}^{\top}\tilde{\mathbf{K}}_{nm}\tilde{\mathbf{K}}_{mm}^{-1/2}$.

  Computation: $O(m^2 n)$.

Computational vs. Statistical Trade-off

- Results similar to Approach 1 are derived for KRR with Nyström approximation (Bach 2013, Alaoui and Mahoney, 2015, Rudi et al., 2015).
- Kernel PCA with Nyström approximation

Many directions and open questions . . .

Questions

Thank You

References I

Fine, S. and Scheinberg, K. (2001).
Efficient SVM training using low-rank kernel representations.
Journal of Machine Learning Research, 2:243–264.

Rahimi, A. and Recht, B. (2008a).
Random features for large-scale kernel machines.
In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc.

Rahimi, A. and Recht, B. (2008b).
Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning.
In *NIPS*, pages 1313–1320.

Rudi, A. and Rosasco, L. (2016).
Generalization properties of learning with random features.
https://arxiv.org/pdf/1602.04474.pdf.

Smola, A. J. and Schölkopf, B. (2000).
Sparse greedy matrix approximation for machine learning.
In *Proc. 17th International Conference on Machine Learning*, pages 911–918. Morgan Kaufmann, San Francisco, CA.

Sriperumbudur, B. K. and Sterge, N. (2017).
Statistical consistency of kernel PCA with random features.
arXiv:1706.06296.

Sriperumbudur, B. K. and Szabó, Z. (2015).
Optimal rates for random Fourier features.
In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 1144–1152. Curran Associates, Inc.

Sutherland, D. and Schneider, J. (2015).
On the error of random fourier features.
In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 862–871.

Williams, C. and Seeger, M. (2001).
Using the Nyström method to speed up kernel machines.
In T. K. Leen, T. G. Diettrich, V. T., editor, *Advances in Neural Information Processing Systems 13*, pages 682–688, Cambridge, MA. MIT Press.

References II

Yang, Y., Pilanci, M., and Wainwright, M. J. (2015).
Randomized sketches for kernels: Fast and optimal non-parametric regression.
Technical report.
https://arxiv.org/pdf/1501.06195.pdf.