

High-dimensional statistics: Some progress and challenges ahead

Martin Wainwright

UC Berkeley
Departments of Statistics, and EECS

University College, London Master Class: Lecture 3

Joint work with: Alekh Agarwal, Arash Amini, Po-Ling Loh,
Sahand Negahban, Garvesh Raskutti, Pradeep Ravikumar, Bin Yu.

Non-parametric regression

Goal: How to predict output from covariates?

- given covariates $(x_1, x_2, x_3, \dots, x_p)$
- output variable y
- want to predict y based on (x_1, \dots, x_p)

Non-parametric regression

Goal: How to predict output from covariates?

- given covariates $(x_1, x_2, x_3, \dots, x_p)$
- output variable y
- want to predict y based on (x_1, \dots, x_p)

Different models: Ordered in terms of complexity/richness:

- linear
- non-linear but still parametric
- semi-parametric
- non-parametric

Non-parametric regression

Goal: How to predict output from covariates?

- given covariates $(x_1, x_2, x_3, \dots, x_p)$
- output variable y
- want to predict y based on (x_1, \dots, x_p)

Different models: Ordered in terms of complexity/richness:

- linear
- non-linear but still parametric
- semi-parametric
- non-parametric

Challenge:

How to control statistical and computational complexity for large number of predictors p ?

High dimensions and sample complexity

Possible models:

- ordinary linear regression: $y = \underbrace{\sum_{j=1}^p \theta_j x_j}_{\langle \theta, x \rangle} + w$
- general non-parametric model: $y = f(x_1, x_2, \dots, x_p) + w$.

High dimensions and sample complexity

Possible models:

- ordinary linear regression: $y = \underbrace{\sum_{j=1}^p \theta_j x_j}_{\langle \theta, x \rangle} + w$

- general non-parametric model: $y = f(x_1, x_2, \dots, x_p) + w.$

Sample complexity: How many samples n for reliable prediction?

- linear models

- ▶ without any structure: sample size $n \asymp \underbrace{p/\epsilon^2}_{\text{linear in } p}$ necessary/sufficient

High dimensions and sample complexity

Possible models:

- ordinary linear regression: $y = \underbrace{\sum_{j=1}^p \theta_j x_j}_{\langle \theta, x \rangle} + w$

- general non-parametric model: $y = f(x_1, x_2, \dots, x_p) + w.$

Sample complexity: How many samples n for reliable prediction?

- linear models

- ▶ without any structure: sample size $n \asymp \underbrace{p/\epsilon^2}$ necessary/sufficient

- ▶ with sparsity $s \ll p$: sample size $n \asymp \underbrace{(s \log p)/\epsilon^2}_{\text{logarithmic in } p}$ necessary/sufficient

High dimensions and sample complexity

Possible models:

- ordinary linear regression: $y = \underbrace{\sum_{j=1}^p \theta_j x_j}_{\langle \theta, x \rangle} + w$

- general non-parametric model: $y = f(x_1, x_2, \dots, x_p) + w.$

Sample complexity: How many samples n for reliable prediction?

- linear models

- ▶ without any structure: sample size $n \asymp \underbrace{p/\epsilon^2}_{\text{linear in } p}$ necessary/sufficient

- ▶ with sparsity $s \ll p$: sample size $n \asymp \underbrace{(s \log p)/\epsilon^2}_{\text{logarithmic in } p}$ necessary/sufficient

- non-parametric models: p -dimensional, smoothness α

Curse of dimensionality: $n \asymp \underbrace{(1/\epsilon)^{2+p/\alpha}}_{\text{Exponential in } p}$

Structure in non-parametric regression

Upshot: **Essential** to impose structural constraints for high-dimensional non-parametric models.

Reduced dimension models:

- dimension-reducing function: $\varphi : \mathbb{R}^p \rightarrow \mathbb{R}^k$, where $k \ll p$
- lower-dimensional function: $g : \mathbb{R}^k \rightarrow \mathbb{R}$
- composite function: $f : \mathbb{R}^p \rightarrow \mathbb{R}$

$$f(x_1, x_2, \dots, x_p) = g(\varphi(x_1, x_2, \dots, x_p))$$

Structure in non-parametric regression

Reduced dimension models:

- dimension-reducing function: $\varphi : \mathbb{R}^p \rightarrow \mathbb{R}^k$, where $k \ll p$
- lower-dimensional function: $g : \mathbb{R}^k \rightarrow \mathbb{R}$
- composite function: $f : \mathbb{R}^p \rightarrow \mathbb{R}$

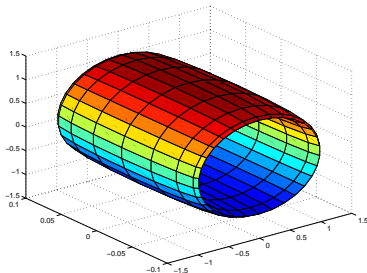
$$f(x_1, x_2, \dots, x_p) = g(\varphi(x_1, x_2, \dots, x_p))$$

Example: Regression on k -dimensional manifold:

Form of model

$$f(x_1, x_2, \dots, x_p) = g(\varphi(x_1, x_2, \dots, x_p))$$

φ is co-ordinate mapping.



Structure in non-parametric regression

Reduced dimension models:

- dimension-reducing function: $\varphi : \mathbb{R}^p \rightarrow \mathbb{R}^k$, where $k \ll p$
- lower-dimensional function: $g : \mathbb{R}^k \rightarrow \mathbb{R}$
- composite function: $f : \mathbb{R}^p \rightarrow \mathbb{R}$

$$f(x_1, x_2, \dots, x_p) = g(\varphi(x_1, x_2, \dots, x_p))$$

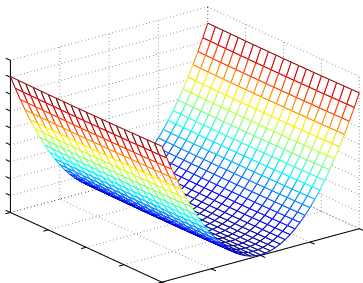
Example: Ridge functions

Form of model:

$$f(x_1, x_2, \dots, x_p) = \sum_{j=1}^k g_j(\langle a_j, x \rangle)$$

Dimension-reducing mapping

$$\varphi(x_1, \dots, x_p) = Ax \quad \text{for some } A \in \mathbb{R}^{k \times p}.$$



Remainder of lecture

1 Sparse additive models

- ▶ formulation, applications
- ▶ families of estimators
- ▶ efficient implementation as SOCP

2 Statistical rates

- ▶ Kernel complexity
- ▶ Subset selection plus univariate function estimation

3 Minimax lower bounds

- ▶ Statistics as channel coding
- ▶ Metric entropy and lower bounds

Sparse additive models

- additive models $f(x_1, x_2, \dots, x_p) = \sum_{j=1}^p f_j(x_j)$
(Stone, 1985; Hastie & Tibshirani, 1990)

Sparse additive models

- additive models $f(x_1, x_2, \dots, x_p) = \sum_{j=1}^p f_j(x_j)$
(Stone, 1985; Hastie & Tibshirani, 1990)
- additivity with sparsity

$$f(x_1, x_2, \dots, x_p) = \sum_{j \in S} f_j(x_j) \quad \text{for unknown subset of cardinality } |S| = s$$

Sparse additive models

- additive models $f(x_1, x_2, \dots, x_p) = \sum_{j=1}^p f_j(x_j)$
(Stone, 1985; Hastie & Tibshirani, 1990)
- additivity with sparsity

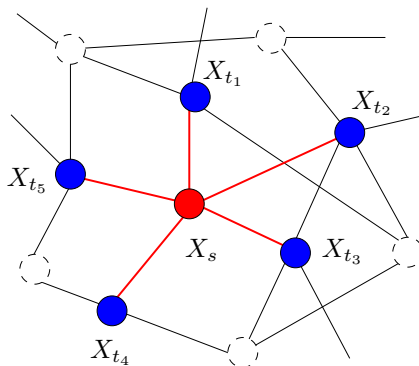
$$f(x_1, x_2, \dots, x_p) = \sum_{j \in S} f_j(x_j) \quad \text{for unknown subset of cardinality } |S| = s$$

- studied by previous authors:
 - ▶ Lin & Zhang, 2006: COSSO relaxation,
 - ▶ Ravikumar et al., 2007: SPAM back-fitting procedure, consistency
 - ▶ Bach et al., 2008: multiple kernel learning (MLK), consistency in classical setting
 - ▶ Meier et al., 2007, $L^2(\mathbb{P}_n)$ regularization
 - ▶ Koltchinski & Yuan, 2008, 2010.
 - ▶ Raskutti, W. & Yu, 2009: minimax lower bounds

Application: Copula methods and graphical models

- transform $X_j \mapsto Z_j = f_j(X_j)$
- model (Z_1, \dots, Z_p) as jointly Gaussian Markov random field

$$\mathbb{P}(z_1, z_2, \dots, z_p) \propto \exp \left\{ \sum_{(s,t) \in E} \theta_{st} z_s z_t \right\}.$$



- exploit Markov properties: neighborhood-based selection for learning graphs (Besag, 1974; Meinshausen & Buhlmann, 2006)
- combined with copula method: semi-parametric approach to graphical model learning (Liu, Lafferty & Wasserman, 2009)

Sparse and smooth

Noisy samples

$$y_i = f^*(x_{i1}, x_{i2}, \dots, x_{ip}) + w_i \quad \text{for } i = 1, 2, \dots, n$$

of unknown function f^* with:

- sparse representation: $f^* = \sum_{j \in S} f_j^*$
- univariate functions are smooth: $f_j \in \mathcal{H}_j$

Sparse and smooth

Noisy samples

$$y_i = f^*(x_{i1}, x_{i2}, \dots, x_{ip}) + w_i \quad \text{for } i = 1, 2, \dots, n$$

of unknown function f^* with:

- sparse representation: $f^* = \sum_{j \in S} f_j^*$
- univariate functions are smooth: $f_j \in \mathcal{H}_j$

- Disregarding computational cost:

$$\min_{|S| \leq s} \min_{\substack{f = \sum_{j \in S} f_j \\ f_j \in \mathcal{H}_j}} \underbrace{\frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2}_{\|y - f\|_n^2}$$

Sparse and smooth

- Disregarding computational cost:

$$\min_{|S| \leq s} \min_{\substack{f = \sum_{j \in S} f_j \\ f_j \in \mathcal{H}_j}} \underbrace{\frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2}_{\|y - f\|_n^2}$$

- 1-Hilbert-norm as convex surrogate:

$$\|f\|_{1, \mathcal{H}} := \sum_{j=1}^p \|f_j\|_{\mathcal{H}_j}$$

Sparse and smooth

- Disregarding computational cost:

$$\min_{|S| \leq s} \min_{\substack{f = \sum_{j \in S} f_j \\ f_j \in \mathcal{H}_j}} \underbrace{\frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2}_{\|y - f\|_n^2}$$

- 1-Hilbert-norm as convex surrogate:

$$\|f\|_{1, \mathcal{H}} := \sum_{j=1}^p \|f_j\|_{\mathcal{H}_j}$$

- 1- $L_2(\mathbb{P}_n)$ -norm as convex surrogate:

$$\|f\|_{1, n} := \sum_{j=1}^p \|f_j\|_{L^2(\mathbb{P}_n)}$$

where $\|f_j\|_{L^2(\mathbb{P}_n)}^2 := \frac{1}{n} \sum_{i=1}^n f_j^2(x_{ij})$.

A family of estimators

Noisy samples

$$y_i = f^*(x_{i1}, x_{i2}, \dots, x_{ip}) + w_i \quad \text{for } i = 1, 2, \dots, n$$

of unknown function $f^* = \sum_{j \in S} f_j^*$.

A family of estimators

Noisy samples

$$y_i = f^*(x_{i1}, x_{i2}, \dots, x_{ip}) + w_i \quad \text{for } i = 1, 2, \dots, n$$

of unknown function $f^* = \sum_{j \in S} f_j^*$.

Estimator:

$$\hat{f} \in \arg \min_{f = \sum_{j=1}^p f_j} \left\{ \frac{1}{n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p f_j(x_{ij}) \right)^2 + \rho_n \|f\|_{1, \mathcal{H}} + \mu_n \|f\|_{1, n} \right\}.$$

A family of estimators

Noisy samples

$$y_i = f^*(x_{i1}, x_{i2}, \dots, x_{ip}) + w_i \quad \text{for } i = 1, 2, \dots, n$$

of unknown function $f^* = \sum_{j \in S} f_j^*$.

Estimator:

$$\hat{f} \in \arg \min_{f = \sum_{j=1}^p f_j} \left\{ \frac{1}{n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p f_j(x_{ij}) \right)^2 + \rho_n \|f\|_{1, \mathcal{H}} + \mu_n \|f\|_{1, n} \right\}.$$

Two kinds of regularization:

$$\|f\|_{1, n} = \sum_{j=1}^p \|f_j\|_{L^2(\mathbb{P}_n)} = \sum_{j=1}^p \sqrt{\frac{1}{n} \sum_{i=1}^n f_j^2(x_{ij})}, \quad \text{and}$$

$$\|f\|_{1, \mathcal{H}} = \sum_{j=1}^p \|f_j\|_{\mathcal{H}_j}.$$

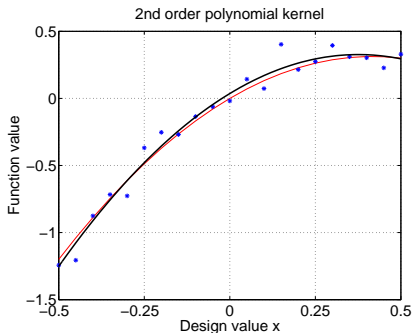
Example: Polynomial kernels

Polynomial kernel

$$\mathbb{K}(z, x) = (1 + \langle z, x \rangle)^d$$

Functions in span of data:

$$f(z) = \sum_{i=1}^n \alpha_i (1 + \langle z, x_i \rangle)^d$$



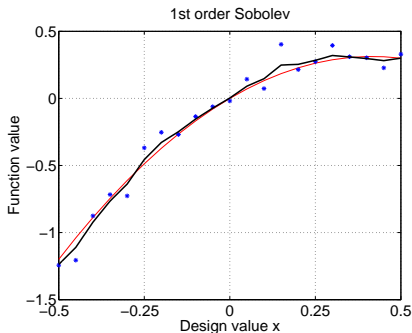
Example: First-order Sobolev kernel

First-order Sobolev kernel

$$\mathbb{K}(z, x) = 1 + \min\{z, x\}$$

Functions in span of data are Lipschitz:

$$f(z) = \sum_{i=1}^n \alpha_i (1 + \min\{z, x_i\})$$



Efficient implementation by kernelization

Representer theorem: Reduces to convex program involving:

- matrix $A = (\alpha_1, \alpha_2, \dots, \alpha_p) \in \mathbb{R}^{n \times p}$.
- empirical kernel matrices $[K_j]_{il} = \mathbb{K}_j(x_{ij}, x_{lj})$.

(Kimeldorf & Wahba, 1971)

Original estimator and kernelized form:

$$\hat{f} \in \arg \min_{f = \sum_{j=1}^p f_j} \left\{ \frac{1}{n} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p f_j(x_{ij}) \right)^2 + \rho_n \sum_{j=1}^p \|f_j\|_{\mathcal{H}_j} + \mu_n \sum_{j=1}^p \|f_j\|_{L^2(\mathbb{P}_n)} \right\}$$

Efficient implementation by kernelization

Representer theorem: Reduces to convex program involving:

- matrix $A = (\alpha_1, \alpha_2, \dots, \alpha_p) \in \mathbb{R}^{n \times p}$.
- empirical kernel matrices $[K_j]_{il} = \mathbb{K}_j(x_{ij}, x_{lj})$.

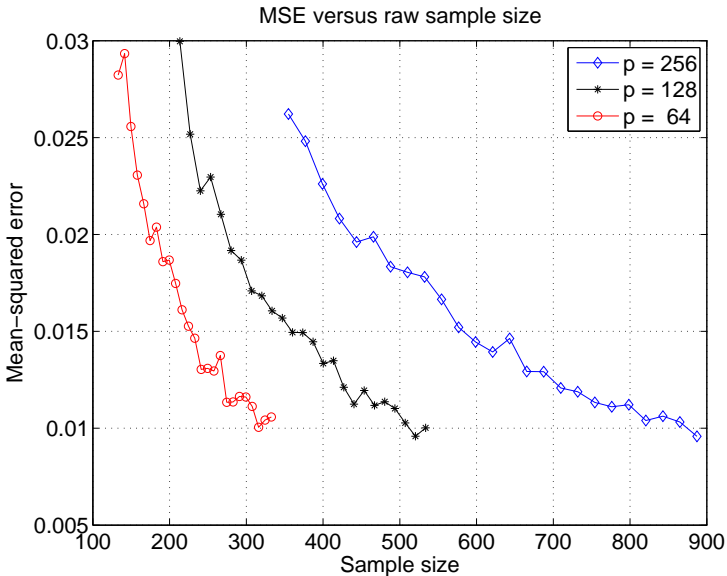
(Kimeldorf & Wahba, 1971)

Original estimator and kernelized form:

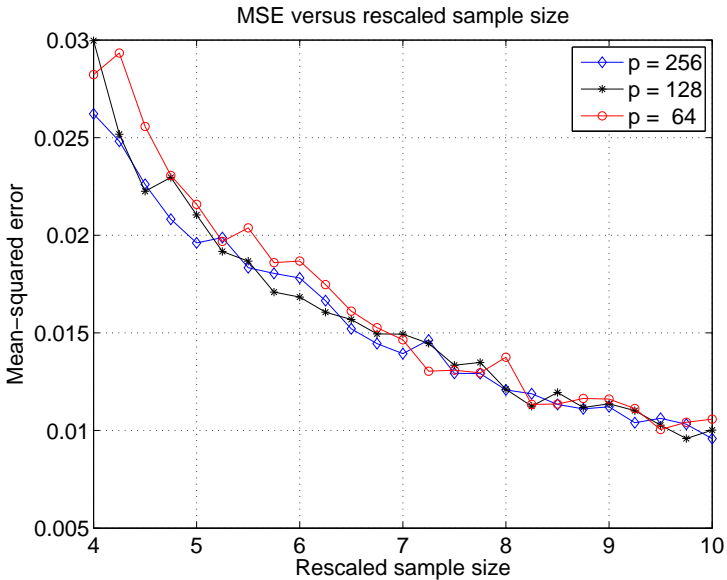
$$\hat{f} \in \arg \min_{f = \sum_{j=1}^p f_j} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - \sum_{j=1}^p f_j(x_{ij}))^2 + \rho_n \sum_{j=1}^p \|f_j\|_{\mathcal{H}_j} + \mu_n \sum_{j=1}^p \|f_j\|_{L^2(\mathbb{P}_n)} \right\}$$

$$\hat{A} \in \arg \min_{A \in \mathbb{R}^{n \times p}} \left\{ \frac{1}{n} \|y - \sum_{j=1}^p K_j \alpha_j\|_2^2 + \rho_n \sum_{j=1}^p \sqrt{\alpha_j^T K_j \alpha_j} + \mu_n \sum_{j=1}^p \sqrt{\alpha_j^T K_j^2 \alpha_j} \right\}.$$

Empirical results: Unrescaled



Empirical results: Appropriately rescaled



Decay rate of kernel eigenvalues

Mercer's theorem: orthonormal basis $\{\phi_j\}$ and non-negative eigenvalues $\{\lambda_j\}$ such that

$$\mathbb{K}(z, x) = \sum_{j=1}^{\infty} \lambda_j \phi_j(z) \phi_j(x).$$

Key intuition: Decay rate $\lambda_j \rightarrow 0$ controls complexity of kernel class.

Decay rate of kernel eigenvalues

Mercer's theorem: orthonormal basis $\{\phi_j\}$ and non-negative eigenvalues $\{\lambda_j\}$ such that

$$\mathbb{K}(z, x) = \sum_{j=1}^{\infty} \lambda_j \phi_j(z) \phi_j(x).$$

Key intuition: Decay rate $\lambda_j \rightarrow 0$ controls complexity of kernel class.

Local Rademacher complexity

(Mendelson, 2002)

$$\mathcal{R}_{\mathbb{K}}(\delta) := \frac{1}{\sqrt{n}} \left[\sum_{j=1}^{\infty} \min \{ \lambda_j, \delta^2 \} \right]^{1/2}.$$

Decay rate of kernel eigenvalues

Mercer's theorem: orthonormal basis $\{\phi_j\}$ and non-negative eigenvalues $\{\lambda_j\}$ such that

$$\mathbb{K}(z, x) = \sum_{j=1}^{\infty} \lambda_j \phi_j(z) \phi_j(x).$$

Key intuition: Decay rate $\lambda_j \rightarrow 0$ controls complexity of kernel class.

Local Rademacher complexity

(Mendelson, 2002)

$$\mathcal{R}_{\mathbb{K}}(\delta) := \frac{1}{\sqrt{n}} \left[\sum_{j=1}^{\infty} \min \{ \lambda_j, \delta^2 \} \right]^{1/2}.$$

Example: Gaussian kernel $\lambda_j \asymp \exp(-j^2)$

Decay rate of kernel eigenvalues

Mercer's theorem: orthonormal basis $\{\phi_j\}$ and non-negative eigenvalues $\{\lambda_j\}$ such that

$$\mathbb{K}(z, x) = \sum_{j=1}^{\infty} \lambda_j \phi_j(z) \phi_j(x).$$

Key intuition: Decay rate $\lambda_j \rightarrow 0$ controls complexity of kernel class.

Local Rademacher complexity

(Mendelson, 2002)

$$\mathcal{R}_{\mathbb{K}}(\delta) := \frac{1}{\sqrt{n}} \left[\sum_{j=1}^{\infty} \min \{ \lambda_j, \delta^2 \} \right]^{1/2}.$$

Example: For Sobolev smoothness kernels:

- First-order (Lipschitz): $\lambda_j \asymp (1/j)^2$
- Second-order (Twice diff'ble): $\lambda_j \asymp (1/j)^4$

Achievable results

Model:

- f^* has $s \ll p$ non-zero components
- each univariate component f_j^* in same univariate Hilbert space \mathcal{H} with eigenvalues $\{\lambda_j\}$
- critical univariate rate δ_n determined by solving

$$\delta^2 \asymp \mathcal{R}_{\mathbb{K}}(\delta_n) = \frac{1}{\sqrt{n}} \left[\sum_{j=1}^{\infty} \min\{\lambda_j, \delta^2\} \right]^{1/2}$$

Achievable results

Model:

- f^* has $s \ll p$ non-zero components
- each univariate component f_j^* in same univariate Hilbert space \mathcal{H} with eigenvalues $\{\lambda_j\}$
- critical univariate rate δ_n determined by solving

$$\delta^2 \asymp \mathcal{R}_{\mathbb{K}}(\delta_n) = \frac{1}{\sqrt{n}} \left[\sum_{j=1}^{\infty} \min\{\lambda_j, \delta^2\} \right]^{1/2}$$

Theorem (Raskutti, W. & Yu, 2010, 2012)

For appropriate choices of regularization parameters ρ_n, μ_n , we have

$$\|\hat{f} - f^*\|_{L^2(\mathbb{P}_n)}^2 \lesssim \underbrace{\frac{s \log p}{n}}_{\text{Cost of subset selection}} + \underbrace{s \delta_n^2}_{\text{Cost of } s\text{-variate estimation}}$$

with high probability.

Consequence: Kernels with exponential decay

- univariate kernel with α -exponential eigendecay eigenvalue decay

$$\lambda_j \asymp \exp(-j^\alpha)$$

- example: Gaussian kernel ($\alpha = 2$)

Consequence: Kernels with exponential decay

- univariate kernel with α -exponential eigenvalue decay

$$\lambda_j \asymp \exp(-j^\alpha)$$

- example: Gaussian kernel ($\alpha = 2$)

Corollary

For kernel with α -exponential decay,

$$\|\hat{f} - f^*\|_{L^2(\mathbb{P}_n)}^2 \lesssim \underbrace{\frac{s \log p}{n}}_{\text{Cost of subset selection}} + \underbrace{\frac{s(\log n)^{2/\alpha}}{n}}_{\text{Cost of } s\text{-variate estimation}}$$

with high probability.

Consequence: Kernels with exponential decay

- univariate kernel with α -exponential eigenvalue decay

$$\lambda_j \asymp \exp(-j^\alpha)$$

- example: Gaussian kernel ($\alpha = 2$)

Corollary

For kernel with α -exponential decay,

$$\|\hat{f} - f^*\|_{L^2(\mathbb{P}_n)}^2 \lesssim \underbrace{\frac{s \log p}{n}}_{\text{Cost of subset selection}} + \underbrace{\frac{s(\log n)^{2/\alpha}}{n}}_{\text{Cost of } s\text{-variate estimation}}$$

with high probability.

Note: Either term can dominate, depending on relative scalings of sample size n , ambient dimension p and decay rate α .

Consequence: Kernels with polynomial decay

- univariate kernel with α -polynomial eigendecay eigenvalue decay

$$\lambda_j \asymp (1/j)^\alpha$$

- examples: Sobolev smoothness classes
 - ▶ $\alpha = 2$: Lipschitz functions
 - ▶ $\alpha = 4$: twice differentiable functions

Consequence: Kernels with polynomial decay

- univariate kernel with α -polynomial eigendecay eigenvalue decay

$$\lambda_j \asymp (1/j)^\alpha$$

- examples: Sobolev smoothness classes
 - ▶ $\alpha = 2$: Lipschitz functions
 - ▶ $\alpha = 4$: twice differentiable functions

Corollary

For kernel with α -polynomial decay:

$$\|\hat{f} - f^*\|_{L^2(\mathbb{P}_n)}^2 \lesssim \underbrace{\frac{s \log p}{n}}_{\text{Cost of subset selection}} + \underbrace{\frac{s}{n^{\frac{2\alpha}{2\alpha+1}}}}_{\text{Cost of } s\text{-variate estimation}}$$

with high probability.

Consequence: Kernels with polynomial decay

- univariate kernel with α -polynomial eigendecay eigenvalue decay

$$\lambda_j \asymp (1/j)^\alpha$$

- examples: Sobolev smoothness classes
 - ▶ $\alpha = 2$: Lipschitz functions
 - ▶ $\alpha = 4$: twice differentiable functions

Corollary

For kernel with α -polynomial decay:

$$\|\widehat{f} - f^*\|_{L^2(\mathbb{P}_n)}^2 \lesssim \underbrace{\frac{s \log p}{n}}_{\text{Cost of subset selection}} + \underbrace{\frac{s}{n^{\frac{2\alpha}{2\alpha+1}}}}_{\text{Cost of } s\text{-variate estimation}}$$

with high probability.

Note: Either term can dominate, depending on relative scalings of sample size n , ambient dimension p and the smoothness α .

Consequence: Finite-rank kernels

- a (block) univariate kernel \mathbb{K} has rank m if $\lambda_j = 0$ for all $j > m$.
- many examples:
 - ▶ linear function classes in \mathbb{R}^m
 - ▶ polynomials of degree $d = m - 1$ in \mathbb{R}

Consequence: Finite-rank kernels

- a (block) univariate kernel \mathbb{K} has rank m if $\lambda_j = 0$ for all $j > m$.
- many examples:
 - ▶ linear function classes in \mathbb{R}^m
 - ▶ polynomials of degree $d = m - 1$ in \mathbb{R}

Corollary

For any kernel with rank m , we have

$$\|\widehat{f} - f^*\|_{L^2(\mathbb{P}_n)}^2 \lesssim \underbrace{\frac{s \log p}{n}}_{\text{Cost of subset selection}} + \underbrace{\frac{sm}{n}}_{\text{Cost of } s\text{-variate estimation}}$$

with high probability.

Consequence: Finite-rank kernels

- a (block) univariate kernel \mathbb{K} has rank m if $\lambda_j = 0$ for all $j > m$.
- many examples:
 - ▶ linear function classes in \mathbb{R}^m
 - ▶ polynomials of degree $d = m - 1$ in \mathbb{R}

Corollary

For any kernel with rank m , we have

$$\|\hat{f} - f^*\|_{L^2(\mathbb{P}_n)}^2 \lesssim \underbrace{\frac{s \log p}{n}}_{\text{Cost of subset selection}} + \underbrace{\frac{sm}{n}}_{\text{Cost of } s\text{-variate estimation}}$$

with high probability.

Note: Either term can dominate, depending on relative scalings of ambient dimension p and kernel rank m .

High-dimensional results from past work

- Ravikumar et al, 2008:
 - ▶ “back-fitting” method for sparse additive models
 - ▶ establish consistency but do not track sparsity s

High-dimensional results from past work

- Ravikumar et al, 2008:
 - ▶ “back-fitting” method for sparse additive models
 - ▶ establish consistency but do not track sparsity s

- Meier et al., 2008:
 - ▶ regularize with $\|f\|_{n,1}$:
 - ▶ establish a rate of the order $s\left(\frac{\log p}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$ for α -smooth Sobolev spaces

High-dimensional results from past work

- Ravikumar et al, 2008:
 - ▶ “back-fitting” method for sparse additive models
 - ▶ establish consistency but do not track sparsity s

- Meier et al., 2008:
 - ▶ regularize with $\|f\|_{n,1}$:
 - ▶ establish a rate of the order $s\left(\frac{\log p}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$ for α -smooth Sobolev spaces

- Koltchinski & Yuan, 2008:
 - ▶ regularize with $\|f\|_{\mathcal{H},1}$
 - ▶ establish rates involving terms at least $s^3 \frac{\log p}{n}$

High-dimensional results from past work

- Ravikumar et al, 2008:
 - ▶ “back-fitting” method for sparse additive models
 - ▶ establish consistency but do not track sparsity s

- Meier et al., 2008:
 - ▶ regularize with $\|f\|_{n,1}$:
 - ▶ establish a rate of the order $s\left(\frac{\log p}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$ for α -smooth Sobolev spaces

- Koltchinski & Yuan, 2008:
 - ▶ regularize with $\|f\|_{\mathcal{H},1}$
 - ▶ establish rates involving terms at least $s^3 \frac{\log p}{n}$

- Concurrent work: Koltchinski & Yuan, 2010:
 - ▶ analyze same estimator but under a global boundedness condition
 - ▶ rates are not minimax-optimal

Rates with global boundedness

Koltchinski & Yuan, 2010:

- analyzed same estimator but under global boundedness:

$$\|f^*\|_\infty = \left\| \sum_{j \in S} f_j^* \right\|_\infty = \sum_{j \in S} \|f_j^*\|_\infty \leq B.$$

- similar rates claimed to be optimal

Rates with global boundedness

Koltchinski & Yuan, 2010:

- analyzed same estimator but under global boundedness:

$$\|f^*\|_\infty = \left\| \sum_{j \in \mathcal{S}} f_j^* \right\|_\infty = \sum_{j \in \mathcal{S}} \|f_j^*\|_\infty \leq B.$$

- similar rates claimed to be optimal

Proposition (Raskutti, W. & Yu, 2010)

Faster rates are possible under global boundedness conditions. For any Sobolev kernel with smoothness α ,

$$\|\hat{f} - f^*\|_{L^2(\mathbb{P}_n)}^2 \lesssim \phi(s, n) \frac{s}{n^{\frac{2\alpha}{2\alpha+1}}} + \frac{s \log(p/s)}{n}$$

for a function such that $\phi(s, n) = o(1)$ if $s \gtrsim \sqrt{n}$.

Information-theoretic lower bounds

Thus far:

- polynomial-time algorithm based on solving SOCP
- upper bounds on error that hold w.h.p.

Question:

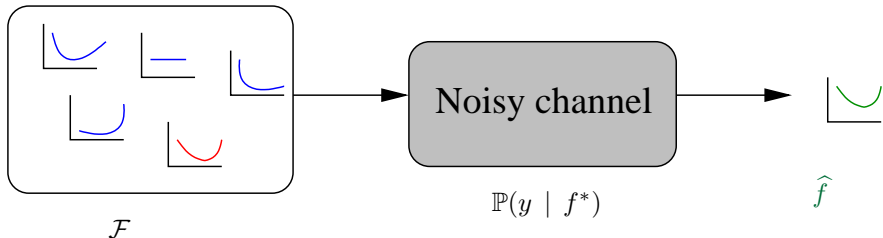
But are these “good” results?

Statistical minimax: For a function class \mathcal{F} , define the minimax error:

$$\mathfrak{M}_n(\mathcal{F}_{s,p,\alpha}) := \inf_{\hat{f}} \sup_{f^* \in \mathcal{F}_{s,p,\alpha}} \|\hat{f} - f^*\|_2^2.$$

Lower bounds behavior of any algorithm over class \mathcal{F} .

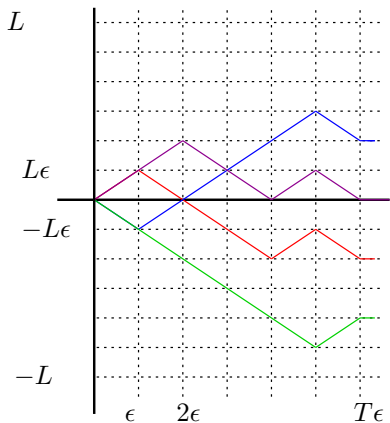
Function estimation as channel coding



- 1 Nature chooses a **function** f^* from a class \mathcal{F} .
- 2 User makes n observations of f^* from a noisy channel.
- 3 Function estimation as decoding: return estimate \hat{f} based on samples $\{(y_i, x_i)\}_{i=1}^n$.

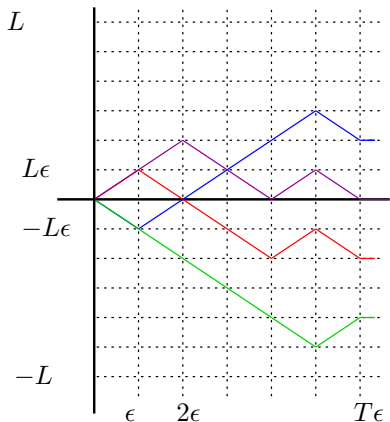
(Hasminskii, 1978, Birge, 1981, Yang & Barron, 1999)

Complexity of function classes



- complexity measured by packing number (Kolmogorov & Tikhomirov, 1960)
- ϵ -packing set: functions $\{f^1, f^2, \dots, f^M\}$ such that $\|f^j - f^k\| > \epsilon$ for all $j \neq k$

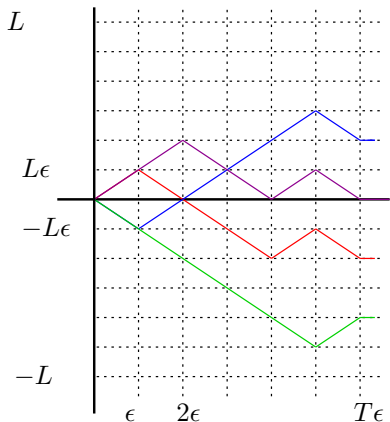
Complexity of function classes



- complexity measured by packing number (Kolmogorov & Tikhomirov, 1960)
- ϵ -packing set: functions $\{f^1, f^2, \dots, f^M\}$ such that $\|f^j - f^k\| > \epsilon$ for all $j \neq k$
- for Lipschitz functions in 1-dimension:

$$M(\epsilon) \asymp 2^{(L/\epsilon)}.$$

Complexity of function classes



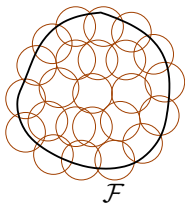
- complexity measured by packing number (Kolmogorov & Tikhomirov, 1960)
- ϵ -packing set: functions $\{f^1, f^2, \dots, f^M\}$ such that $\|f^j - f^k\| > \epsilon$ for all $j \neq k$
- for Lipschitz functions in p -dimensions

$$M(\epsilon) \asymp 2^{(L/\epsilon)^p}.$$

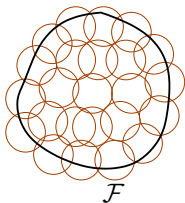
Metric entropy classes

Covering number

$N(\delta; \mathcal{F}) =$ smallest # δ -balls needed to cover \mathcal{F}



Metric entropy classes



Covering number

$N(\delta; \mathcal{F}) =$ smallest # δ -balls needed to cover \mathcal{F}

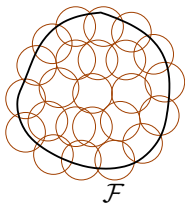
1 Logarithmic metric entropy

$$\log N(\delta; \mathcal{F}) \asymp m \log(1/\delta)$$

Examples:

- ▶ parametric classes
- ▶ finite-rank kernels
- ▶ any function class with finite VC dimension

Metric entropy classes



Covering number

$N(\delta; \mathcal{F}) =$ smallest # δ -balls needed to cover \mathcal{F}

❶ Polynomial metric entropy:

$$\log N(\delta; \mathcal{F}) \asymp \left(\frac{1}{\delta}\right)^{\frac{1}{\alpha}}$$

Examples:

- ▶ various smoothness classes
- ▶ Sobolev classes

Lower bounds on minimax risk

Theorem (Raskutti, W. & Yu, 2010)

Under the same conditions, there is a constant $c_0 > 0$ such that:

① For function class \mathcal{F} with m -logarithmic metric entropy:

$$\mathbb{P} \left[\mathfrak{M}_n(\mathcal{F}_{s,p,\alpha}) \geq c_0 \left\{ \underbrace{\frac{s \log p/s}{n}}_{\text{subset sel.}} + \underbrace{s \left(\frac{m}{n} \right)}_{\text{s-var. est.}} \right\} \right] \geq 1/2.$$

Lower bounds on minimax risk

Theorem (Raskutti, W. & Yu, 2010)

Under the same conditions, there is a constant $c_0 > 0$ such that:

- ❶ For function class \mathcal{F} with m -logarithmic metric entropy:

$$\mathbb{P} \left[\mathfrak{M}_n(\mathcal{F}_{s,p,\alpha}) \geq c_0 \left\{ \underbrace{\frac{s \log p/s}{n}}_{\text{subset sel.}} + \underbrace{s \left(\frac{m}{n} \right)}_{\text{s-var. est.}} \right\} \right] \geq 1/2.$$

- ❷ For function class \mathcal{F} with α -polynomial metric entropy:

$$\mathbb{P} \left[\mathfrak{M}_n(\mathcal{F}_{s,p,\alpha}) \geq c_0 \left\{ \underbrace{\frac{s \log p/s}{n}}_{\text{subset sel.}} + \underbrace{s \left(\frac{1}{n} \right)^{\frac{2\alpha}{2\alpha+1}}}_{\text{s-var. est.}} \right\} \right] \geq 1/2.$$

Summary

- structure is essential for high-dimensional non-parametric models
- sparse and smooth additive models:
 - ▶ convex relaxation based on a composite regularizer
 - ▶ attains minimax-optimal rates for kernel classes:
 - ★ cost of subset selection: $s \frac{\log p/s}{n}$
 - ★ cost of s -variate function estimation: $s\delta_n^2$

Summary

- structure is essential for high-dimensional non-parametric models
- sparse and smooth additive models:
 - ▶ convex relaxation based on a composite regularizer
 - ▶ attains minimax-optimal rates for kernel classes:
 - ★ cost of subset selection: $s \frac{\log p/s}{n}$
 - ★ cost of s -variate function estimation: $s\delta_n^2$
- many open questions:
 - ▶ functional ANOVA decompositions: allowing groupings of variables (doublets, triplets etc.)
 - ▶ other types of high-dimensional non-parametric models:
 - ★ kernel PCA/CCA with structural constraints
 - ★ structured density estimation
 - ▶ trade-offs between computational and statistical efficiency

Summary

- structure is essential for high-dimensional non-parametric models
- sparse and smooth additive models:
 - ▶ convex relaxation based on a composite regularizer
 - ▶ attains minimax-optimal rates for kernel classes:
 - ★ cost of subset selection: $s \frac{\log p/s}{n}$
 - ★ cost of s -variate function estimation: $s\delta_n^2$
- many open questions:
 - ▶ functional ANOVA decompositions: allowing groupings of variables (doublets, triplets etc.)
 - ▶ other types of high-dimensional non-parametric models:
 - ★ kernel PCA/CCA with structural constraints
 - ★ structured density estimation
 - ▶ trade-offs between computational and statistical efficiency

Related paper:

Raskutti, W., & Yu (2012). Minimax-optimal rates for sparse additive models over kernel classes. *Journal of Machine Learning Research*, March 2012.