Sparse and Smooth: An optimal convex relaxation for high-dimensional regression

Martin Wainwright

UC Berkeley
Departments of Statistics, and EECS
July 2011

Joint work with Garvesh Raskutti and Bin Yu, UC Berkeley

Non-parametric regression

Goal: How to predict output from covariates?

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

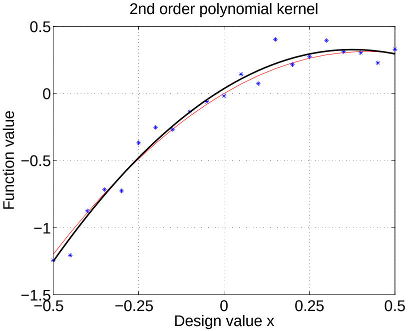
Examples: Medical diagnosis; Geostatistics; Astronomy; Video denoising ...

Non-parametric regression

Goal: How to predict output from covariates?

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

Examples: Medical diagnosis; Geostatistics; Astronomy; Video denoising ...



(a) Second-order poly.

(b) First-order Sobolev

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: $y = f(x_1, x_2, \ldots, x_p) + w$.

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: $y = f(x_1, x_2, \ldots, x_p) + w$.

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: $y = f(x_1, x_2, \ldots, x_p) + w$.

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

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: $y = f(x_1, x_2, \ldots, x_p) + w$.

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

Sparse additive models

- additive models $f(x_1, x_2, \ldots, x_p) = \sum_{j=1}^{p} f_j(x_j)$

  (Stone, 1985; Hastie & Tibshirani, 1990)

- additivity with sparsity

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

Sparse additive models

- additive models $f(x_1, x_2, \ldots, x_p) = \sum_{j=1}^{p} f_j(x_j)$

  (Stone, 1985; Hastie & Tibshirani, 1990)

- additivity with sparsity

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

- studied by previous authors:
  - Lin & Zhang, 2006: COSSO relaxation
  - Ravikumar et al., 2007: SPAM back-fitting procedure
  - Meier et al., 2007
  - Koltchinski & Yuan, 2008, 2010.

Sparse and smooth

Noisy samples

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

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

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

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:

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

Sparse and smooth

- Disregarding computational cost:

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

- 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:

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

- 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

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

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

A family of estimators

Noisy samples

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

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

Estimator:

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

A family of estimators

Noisy samples

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

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

Estimator:

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

Two kinds of regularization:

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

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

Efficient implementation by kernelization

Representer theorem: Reduces to convex program involving:

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

(Kimeldorf & Wahba, 1971)

Original estimator and kernelized form:

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

Efficient implementation by kernelization

Representer theorem: Reduces to convex program involving:

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

(Kimeldorf & Wahba, 1971)

Original estimator and kernelized form:

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

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

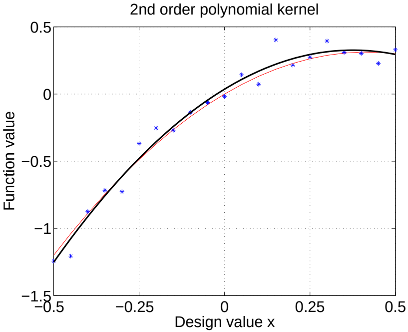
Example: Polynomial kernels

Polynomial kernel

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

Functions in span of data:

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



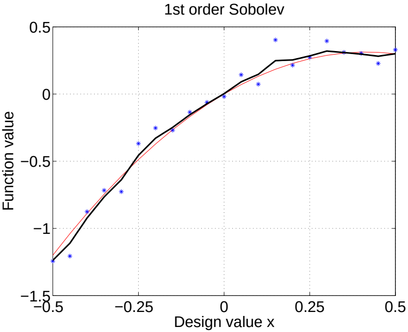
Example: First-order Sobolev kernel

First-order Sobolev kernel

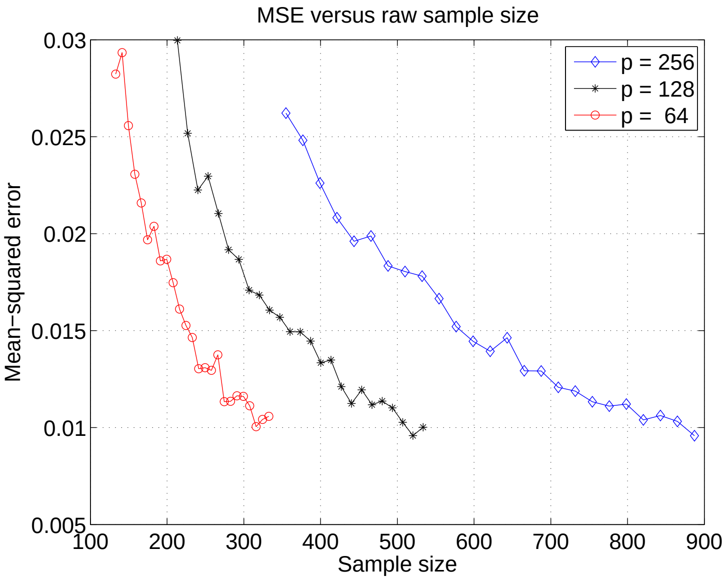
$$\mathbb{K}(z, x) = \min\{z, x\}$$

Functions in span of data are Lipschitz:

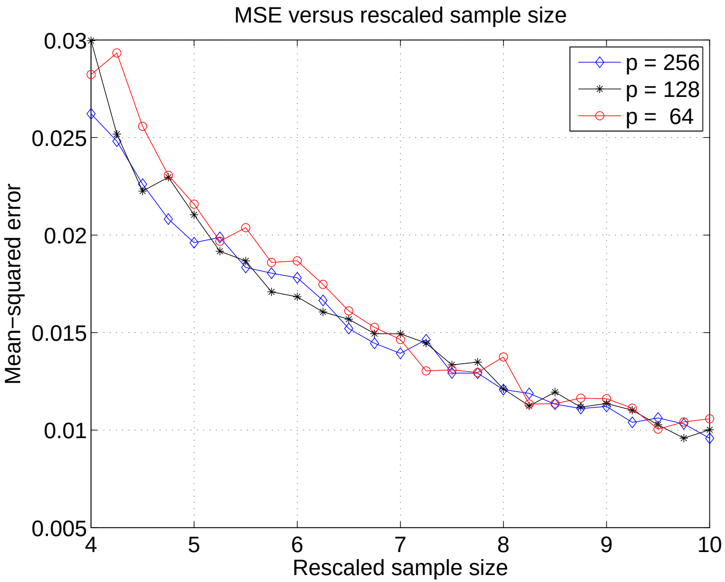
$$f(z) = \sum_{i=1}^{n} \alpha_i \min\{z, x\}$$



Empirical results: Unrescaled



Empirical results: Apppropriately 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 $\lambda_j \to +\infty$ 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 $\lambda_j \to +\infty$ controls complexity of kernel class.

Local Rademacher complexity

(Mendelson, 2002)

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

Example: For Sobolev kernels:

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

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

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

Theorem (Raskutti, W. & Yu, 2010)

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

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

with high probability.

Consequence: 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 we have

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

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 we have

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

with high probability.

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

Consequence: Sobolev kernels

- a univariate Sobolev kernel of smoothness α has eigenvalue decay

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

- examples:
  - $\alpha = 1$: Lipschitz functions
  - $\alpha = 2$: twice differentiable functions

Consequence: Sobolev kernels

- a univariate Sobolev kernel of smoothness α has eigenvalue decay

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

- examples:
  - $\alpha = 1$: Lipschitz functions
  - $\alpha = 2$: twice differentiable functions

Corollary

For a Sobolev kernel with smoothness α, we have

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

with high probability.

Consequence: Sobolev kernels

- a univariate Sobolev kernel of smoothness α has eigenvalue decay

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

- examples:
  - $\alpha = 1$: Lipschitz functions
  - $\alpha = 2$: twice differentiable functions

Corollary

For a Sobolev kernel with smoothness α, we have

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

with high probability.

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

Rates from past work

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

Rates 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

Rates 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 $s^3 \frac{\log p}{n}$

Rates 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 $s\big(\frac{\log p}{n}\big)^{\frac{2\alpha}{2\alpha+1}}$ 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:

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

- similar rates claimed to be optimal

Rates with global boundedness

Koltchinski & Yuan, 2010:

- analyzed same estimator but under global boundedness:

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

- 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,

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

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

Information-theoretic lower bounds

Thus far:

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

Question:

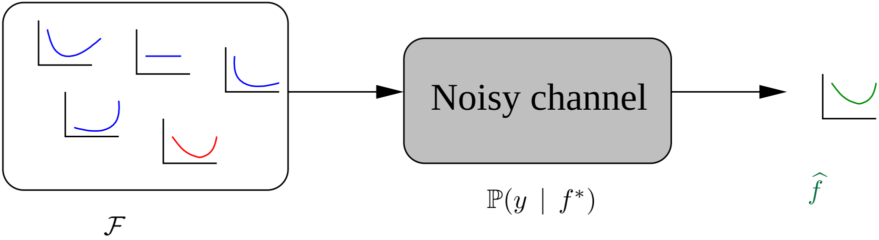
But are these "good" results?

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

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

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

Function estimation as channel coding



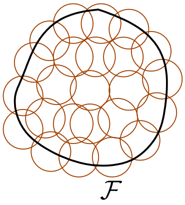
1. Nature chooses a function f^* from a class $\mathcal{F}$.

2. User makes n observations of f^* from a noisy channel.

3. Function estimation as decoding: return estimate $\widehat{f}$ based on samples $\{(y_i, x_i)\}_{i=1}^n$.

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

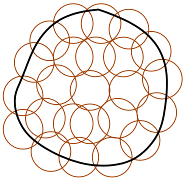
Metric entropy classes



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 $\mathcal{F}$

---

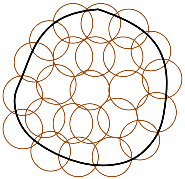
1. Logarithmic metric entropy

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

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 $\mathcal{F}$

---

1. 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, 2009)

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

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

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

Lower bounds on minimax risk

Theorem (Raskutti, W. & Yu, 2009)

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

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

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

2. *For function class $\mathcal{F}$ with α-polynomial metric entropy:*

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

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: $s\delta_n^2$
- many open questions:
  - allowing groupings of variables (doublets, triplets etc.)
  - extension to other structured non-parametric models
  - trade-offs between computational and statistical efficiency

Pre-print:

Raskutti, Wainwright & Yu, 2010
Minimax-optimal rates for sparse additive models over kernel classes
Available at http://arxiv.org/abs/1008.3654.