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
Possible models:

- ordinary linear regression: \( y = \sum_{j=1}^{p} \theta_j x_j + w \)

- general non-parametric model: \( y = f(x_1, x_2, \ldots, x_p) + w \).
High dimensions and sample complexity

Possible models:

- ordinary linear regression: \( y = \sum_{j=1}^{p} \theta_j x_j + w \)
- 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 \( n \asymp \frac{p}{\epsilon^2} \) necessary/sufficient linear in \( p \)
High dimensions and sample complexity

Possible models:

- ordinary linear regression: \( y = \sum_{j=1}^{p} \theta_j x_j + w \)

- 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 \( n \asymp \frac{p}{\epsilon^2} \) necessary/sufficient
  - linear in \( p \)
  - with sparsity \( s \ll p \): sample size \( n \asymp \frac{(s \log p)}{\epsilon^2} \) necessary/sufficient
  - logarithmic in \( p \)
High dimensions and sample complexity

Possible models:

- ordinary linear regression: \( y = \sum_{j=1}^{p} \theta_j x_j + w \)

- 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 \( n \approx \frac{p}{\epsilon^2} \) necessary/sufficient linear in \( p \)
  - with sparsity \( s \ll p \): sample size \( n \approx \frac{(s \log p)/\epsilon^2}{\epsilon^2} \) necessary/sufficient logarithmic in \( p \)

- non-parametric models: \( p \)-dimensional, smoothness \( \alpha \)

  Curse of dimensionality: \( n \approx \frac{(1/\epsilon)^{2+p/\alpha}}{\epsilon^{2+p/\alpha}} \) 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

  $f(x_1, x_2, \ldots, x_p) = \sum_{j \in S} f_j(x_j)$ 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
  \[ f(x_1, x_2, \ldots, x_p) = \sum_{j \in S} f_j(x_j) \]
  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
Noisy samples

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

of unknown function \( f^* \) with:

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

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

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:

\[
\min_{|S| \leq s} \min_{f = \sum_{j \in S} f_j} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \quad \text{subject to} \quad \|y - f\|^2_n
\]
Sparse and smooth

- Disregarding computational cost:

\[
\min_{\| S \| \leq s} \min_{\begin{array}{c}
\sum_{j \in S} f_j \\
\sum_{j \in S} f_j \in \mathcal{H}_j
\end{array}} \frac{1}{n} \sum_{i=1}^{n} \left( y_i - f(x_i) \right)^2
\]

\[
\| y - f \|_n^2
\]

- 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:

\[
\min_{|S| \leq s} \min_{f = \sum_{j \in S} f_j \in \mathcal{H}_j} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2
\]

- 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

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

of unknown function \( f^* = \sum_{j \in S} f_j^* \).
A family of estimators

Noisy samples

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

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

\[ y_i = f^\ast(x_{i1}, x_{i2}, \ldots, x_{ip}) + w_i \quad \text{for } i = 1, 2, \ldots, n \]

of unknown function \( f^\ast = \sum_{j \in S} f_j^\ast \).

**Estimator:**

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

Two kinds of regularization:

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

\[ \|f\|_{1,H} = \sum_{j=1}^{p} \|f_j\|_{H_j}. \]
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 $[K_j]_{i\ell} = K_j(x_{ij}, x_{\ell j})$.

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

(Kimeldorf & Wahba, 1971)

**Original estimator and kernelized form:**

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

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

Polynomial kernel

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

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

\[ K(z, x) = \min\{z, x\} \]

Functions in span of data are Lipschitz:

\[ f(z) = \sum_{i=1}^{n} \alpha_i \min\{z, x\} \]
Empirical results: Unrescaled

MSE versus raw sample size

- $p = 256$
- $p = 128$
- $p = 64$
Empirical results: Appropriately rescaled

MSE versus rescaled sample size

Mean-squared error

Rescaled sample size

- $p = 256$
- $p = 128$
- $p = 64$
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 \( \lambda_j \rightarrow +\infty \) 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

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

**Key intuition:** Decay rate \( \lambda_j \to +\infty \) controls complexity of kernel class.

#### Local Rademacher complexity

(Mendelson, 2002)

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

**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 $\delta_n$ determined by solving

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

Theorem (Raskutti, W. & Yu, 2010)

For appropriate choices of regularization parameters $\rho_n, \mu_n$, we have

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

with high probability.
Consequence: Finite-rank kernels

- a (block) univariate kernel $K$ 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

\[
\| \hat{f} - f^* \|_{L^2(P_n)}^2 \lesssim \frac{s \log p}{n} \quad + \quad \frac{sm}{n}
\]

Cost of subset selection Cost of \( s \)-variate estimation

with high probability.
Consequence: Finite-rank kernels

- A (block) univariate kernel $K$ 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

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

Cost of subset selection Cost of $s$-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 $\alpha$ 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 $\alpha$ 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 $\alpha$, we have

\[ \| \hat{f} - f^* \|_{L^2(P_n)}^2 \lesssim \frac{s \log p}{n} + \frac{s}{n^{2\alpha+1}} \]

Cost of subset selection  Cost of $s$-variate estimation

with high probability.
Consequence: Sobolev kernels

- a univariate Sobolev kernel of smoothness $\alpha$ has eigenvalue decay

\[ \lambda_j \approx \left(\frac{1}{j}\right)^\alpha \]

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

Corollary

For a Sobolev kernel with smoothness $\alpha$, we have

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

Cost of subset selection \hspace{1cm} Cost of s-variate estimation

with high probability.

Note: Either term can dominate, depending on relative scalings of sample size $n$, ambient dimension $p$ and the smoothness $\alpha$. 
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 $\alpha$-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 $\alpha$-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 $s\left(\frac{\log p}{n}\right)^{2\alpha+1}$ for $\alpha$-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 \( \alpha \),

\[ \| \hat{f} - f^* \|_{L^2(P_n)}^2 \lesssim \phi(s, n) \frac{s}{n^{2\alpha} \alpha + 1} + \frac{s \log(p/s)}{n} \]

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:

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

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 $\{(y_i, x_i)\}_{i=1}^{n}$.

Metric entropy classes

Covering number

\[ N(\delta; \mathcal{F}) = \text{smallest } \# \text{ of } \delta\text{-balls needed to cover } \mathcal{F} \]
Metric entropy classes

Covering number

\[ N(\delta; \mathcal{F}) = \text{smallest } \# \text{ of } \delta\text{-balls needed to cover } \mathcal{F} \]

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

\[ N(\delta; \mathcal{F}) = \text{smallest } \# \delta\text{-balls needed to cover } \mathcal{F} \]

1. Polynomial metric entropy:

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

Examples:
- various smoothness classes
- Sobolev classes
**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:

$$
\mathbb{P} \left[ M_n(\mathcal{F}_{s,p,\alpha}) \geq c_0 \left\{ \frac{s \log p}{s n} + s \left( \frac{m}{n} \right) \right\} \right] \geq 1/2.
$$

subset sel.  s-var. est.
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:

   $$\mathbb{P} \left[ M_n(\mathcal{F}_{s,p,\alpha}) \geq c_0 \left\{ \frac{s \log p/s}{n} + s \left( \frac{m}{n} \right) \right\} \right] \geq 1/2.$$ 

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

   $$\mathbb{P} \left[ M_n(\mathcal{F}_{s,p,\alpha}) \geq c_0 \left\{ \frac{s \log p/s}{n} + s \left( \frac{1}{n} \right)^{2\alpha+1} \right\} \right] \geq 1/2.$$
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: $s \delta_n^2$
- 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