Nonparametric Sparsity

John Lafferty
Computer Science Dept.
Machine Learning Dept.
Carnegie Mellon University

Larry Wasserman
Department of Statistics
Machine Learning Dept.
Carnegie Mellon University
Motivation

• “Modern” data are very high dimensional

• In order to be “learnable,” there must be lower-dimensional structure

• Developing practical algorithms with theoretical guarantees for beating the curse of (apparent) dimensionality is a main scientific challenge for our field
Motivation

• Sparsity is emerging as a key concept in statistics and machine learning
• Dramatic progress in recent years on understanding sparsity in parametric settings
• Nonparametric sparsity: Wide open
Outline

• High dimensional learning: Parametric and nonparametric
• Rodeo: Greedy, sparse nonparametric regression
• Extensions of the Rodeo
Parametric Case: Variable Selection in Linear Models

\[ p(x) = Z(\phi) - 1 \exp \left( \sum \text{cliques } \phi C (x C) \right) \]

\[ Z(\phi) = \sum x_{1,\ldots,n} \exp \left( \sum \text{cliques } \phi C (x C) \right) \]

\[ Y = \sum_{j=1}^{d} \beta_j X_j + \epsilon = X^T \beta + \epsilon \]

where \( d \) might be larger than \( n \). Predictive risk

\[ R = \mathbb{E}(Y_{new} - X_{new}^T \beta)^2. \]

Want to choose subset \((X_j : j \in S), S \subset \{1, \ldots, d\}\) to make \( R \) small.

Bias-variance tradeoff:

small \( S \) \( \implies \) Bias ↑ Variance ↓

large \( S \) \( \implies \) Bias ↓ Variance ↑
Lasso/Basis Pursuit
(Chen & Donoho, 1994; Tibshirani, 1996)

\[ \sum_{j=1}^{d} |\beta_j| \leq t \quad \text{Level sets of squared error} \]

For orthogonal designs, solution given by soft thresholding

\[ \hat{\beta}_j = \text{sign}(\beta_j) (|\beta_j| - \lambda)_+ \]
Convex Relaxations for Sparse Signal Recovery

Desired problem:

\[
\begin{align*}
\min & \quad \| \beta \|_0 \\
\text{such that} & \quad \| X \beta - y \|_2 \leq \epsilon
\end{align*}
\]

Requires intractable combinatorial optimization.

Convex optimization surrogate:

\[
\begin{align*}
\min & \quad \| \beta \|_1 \\
\text{such that} & \quad \| X \beta - y \|_2 \leq \epsilon
\end{align*}
\]

Substantial progress recently on theoretical justification

(Candès and Tao, Donoho, Tropp, Meinshausen and Bühlmann, Wainwright, Zhao and Yu, Fan and Peng,...)
Nonparametric Regression

Given \((X_1, Y_1), \ldots, (X_n, Y_n)\) where

\[ Y_i \in \mathbb{R}, \quad X_i = (X_{1i}, \ldots, X_{di})^T \in \mathbb{R}^d, \]

\[ Y_i = m(X_{1i}, \ldots, X_{di}) + \epsilon_i, \quad \mathbb{E}(\epsilon_i) = 0 \]

Risk:

\[ R(m, \hat{m}) = \int \mathbb{E}(\hat{m}(x) - m(x))^2 dx \]

Minimax theorem:

\[ \inf_{\hat{m}} \sup_{m \in \mathcal{F}} R(m, \hat{m}) \asymp \left( \frac{1}{n} \right)^{4/(4+d)} \]

where \(\mathcal{F}\) is class of functions with 2 smooth derivatives. Note the curse of dimensionality.
The Curse of Dimensionality
(Sobolev space of order 2)

d = 20

Risk = 0.01

Risk

sample size

0.0 0.1 0.2 0.3 0.4 0.5
1e+02 1e+04 1e+06 1e+08

dimension

sample size

1e+12
1e+11
4e+11
2e+11
0e+10

10 12 14 16 18 20
Nonparametric Sparsity

- In many applications, reasonable to expect true function depends only on small number of variables
- Assume \( m(x) = m(x_R) \)
  where \( x_R = (x_j)_{j \in R} \) are the relevant variables with \( |R| = r \ll d \)
- Can hope to achieve the better minimax rate \( n^{-4/(4+r)} \)
- Challenge: Variable selection in nonparametric regression
Rodeo: Regularization of derivative expectation operator

- A general strategy for nonparametric estimation: Regularize derivatives of estimator with respect to smoothing parameters
- A simple new algorithm for simultaneous bandwidth and variable selection in nonparametric regression
- Theoretical analysis: Algorithm correctly determines relevant variables, with high probability, and achieves (near) optimal minimax rate of convergence
- Examples showing performance consistent with theory
Key Idea in Rodeo: 
Change of Representation

$$F(h) = F(0) + \int_{0}^{h} F'(x) \, dx$$
Rodeo: The Main Idea

- Use a nonparametric estimator based on a kernel
- Start with large bandwidths in each dimension, for an estimate having small variance but high bias
  - Choosing large bandwidth is like ignoring a variable
- Compute the derivatives of the estimate with respect to bandwidth
- Threshold the derivatives to get a sparse estimate
- **Intuition:** If a variable is irrelevant, then changing the bandwidth in that dimension should only result in a small change in the estimator
Rodeo: The Main Idea

\[ h_1 \]

Start

Rodeo path

Ideal path

Optimal bandwidth

\[ h_2 \]
Using Local Linear Smoothing

The estimator can be written as

\[ \hat{m}_h(x) = \sum_{i=1}^{n} G(X_i, x, h)Y_i \]

Our method is based on the statistic

\[ Z_j = \frac{\partial \hat{m}_h(x)}{\partial h_j} = \sum_{i=1}^{n} G_j(X_i, x, h)Y_i \]

The estimated variance is

\[ s^2_j = \text{Var}(Z_j \mid X_1, \ldots, X_n) = \sigma^2 \sum_{i=1}^{n} G_{j}^2(X_i, x, h) \]
Rodeo: Hard Tresholding Version

1. Select parameter $0 < \beta < 1$ and initial bandwidth $h_0$.

2. Initialize the bandwidths, and activate all covariates:
   (a) $h_j = h_0$, $j = 1, 2, \ldots, d$.
   (b) $A = \{1, 2, \ldots, d\}$

3. While $A$ is nonempty, do for each $j \in A$:
   (a) Compute estimated derivative expectation: $Z_j$ and $s_j$
   (b) Compute threshold $\lambda_j = s_j \sqrt{2 \log n}$.
   (c) If $|Z_j| > \lambda_j$, set $h_j \leftarrow \beta h_j$; otherwise remove $j$ from $A$.

4. Output bandwidths $h^* = (h_1, \ldots, h_d)$ and estimator
   $$\tilde{m}(x) = \hat{m}_{h^*}(x)$$
Example: \( m(x) = 2(x_1 + 1)^3 + 2 \sin(10x_2), \ d = 20 \)
Loss with $r=2$, Increasing Dimension

Leave-one-out cross-validation

Rodeo
Main Result: Near Optimal Rates

Theorem. Suppose that \( d = O(\log n / \log \log n) \), \( h_0 = 1 / \log \log n \), and \( |m_{j,j}(x)| > 0 \). Then the rodeo outputs bandwidths \( h^* \) that satisfy

\[
P(h_j^* = h_0 \text{ for all } j > r) \longrightarrow 1
\]

and for every \( \epsilon > 0 \),

\[
P\left( n^{-1/(4+r) - \epsilon} \leq h_j^* \leq n^{-1/(4+r) + \epsilon} \text{ for all } j \leq r \right) \longrightarrow 1.
\]

Let \( T_n \) be the stopping time of the algorithm. Then

\[
P(t_L \leq T_n \leq t_U) \longrightarrow 1
\]

where

\[
t_L = \frac{1}{(r + 4) \log(1/\beta)} \log \left( \frac{n A_{\min}}{\log n (\log \log n)^d} \right)
\]

\[
t_U = \frac{1}{(r + 4) \log(1/\beta)} \log \left( \frac{n A_{\max}}{\log n (\log \log n)^d} \right)
\]
Greedy Rodeo and LARS

- Rodeo can be viewed as a nonparametric version of least angle regression (LARS), (Efron et al., 2004)
- In forward stagewise, variable selection is incremental. LARS adds the variable most correlated with the residuals of the current fit.
- For the Rodeo, the derivative is essentially the correlation between the output and the derivative of the effective kernel
- Reducing the bandwidth is like adding more of that variable
LARS Regularization Paths

![LARS Regularization Paths Graph]
Greedy Rodeo Bandwidth Paths

Rodeo order: 3 (body mass index), 9 (serum), 7 (serum), 4 (blood pressure), 1 (age), 2 (sex), 8 (serum), 5 (serum), 10 (serum), 6 (serum).

LARS order: 3, 9, 4, 7, 2, 10, 5, 8, 6, 1.
Extensions

- Sparse density estimation
- Local polynomial estimation
- Classification using Rodeo with generalized linear models
- Other nonparametric estimators
- Data-adaptive basis pursuit
Bootstrap Method for the sparse components diagnosis

1. Draw a sample $X^*_1, \ldots, X^*_n$ of size $n_u$ with replacement $B$ times for the following
2. Compute the estimate $\hat{\beta}(i)_k, \cdot u_k = 1, \ldots, d_u$ from data $X^*_1, \ldots, X^*_n$
3. Compute the test statistic $Z_k$ for $k = 1, \ldots, d_u$ as $Z_k = \| 1_B \sum_{i=1}^B \hat{\beta}(i)_k, \cdot \|_2$
4. Output the resulted $Z_k u_k = 1, \ldots, d_x$

Figure 1: The bootstrap method for sparse component identification

Results: Univariate Regression Example

$p$Doppler function $r$: The doppler function is in the following form

$m_{p x r} = \sqrt{x_p 1 - x_r \sin(2.1 \pi x_z z_5 p 17 r)}$

Take $n = 1, \ldots, n_z z u_x = i/n$ and $\epsilon \sim N(p z, \sigma^2 r)$ with $\sigma = z_z 10.0 0.2 0.4 0.6 0.8 1.0
0.0 0.2 0.4 0.6 0.8 1.0
0.0 0.2 0.4 0.6 0.8 1.0

Figure 2: The true regression line of the doppler function and the fitted result from the data-adaptive basis with 36 basis used. From which we see that fitted result is almost perfect. This surprisingly good performance is easily to be explained when we check the 36 data-adaptive basis functions as shown in figure 3 to figure 5.
Data-Adaptive Basis Pursuit

- Recall idea of Rodeo:
  \[ \hat{m}(x) = \hat{m}_1(x) - \int_0^1 \langle \hat{Z}(x, h(s)), \dot{h}(s) \rangle ds \]

- Let \( \Phi(X_i) = \text{vec} \left( Z(X_i, h(s_k)) \cdot dh(s_k) \right) \) over a grid of bandwidths

- Run the Lasso:
  \[
  \min_{\beta} \quad \| Y - \Phi(X)\beta \|_2 \\
  \text{such that} \quad \| \beta \|_1 \leq t
  \]
Data-Adaptive Basis Pursuit

Figure 3: The data-adaptive basis $1, 2, \ldots, 12$ for the data from the doppler function.
Summary

- Sparsity is playing an increasingly important role in statistics and machine learning.
- In order to be “learnable,” there must be lower-dimensional structure.
- Nonparametric sparsity: many open problems.
- Rodeo: conceptually simple and practical, theoretically nice properties.