Attribute-Efficient Learning of Monomials Over Highly-Correlated Variables

Alexandr Andoni, Rishabh Dudeja, Daniel Hsu, Kiran Vodrahalli

Columbia University: Computer Science Department, Department of Statistics, and the Data Science Institute

Problem Statement

- Model: Observe n features-response pairs $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$ drawn i.i.d. from the following model:
$$x_i \sim \mathcal{N}(0, \Phi), \quad y_i = f(x_i), \quad f(x) = \Pi_{j \in S} x_j^{\beta_j}.$$
- Goal: Design an algorithm to accurately estimate the unknown function f with small sample complexity (n) and small run-time. Moreover, the unknown function f may depend on only k out of the p features, with $k \ll p$. This models the problem of feature selection in machine learning and statistics.
- Efficiency requirement: Design algorithms that are *attribute-efficient* and require $n = \mathrm{poly}(\log(p), k)$ samples and $\mathrm{poly}(n, p, k)$ run-time.

Contributions

1. We design an attribute-efficient algorithm for learning $f(x)$ using sample size $n = O(k^2 \cdot \mathrm{poly}(\log(p), \log(k)))$ and runtime $\mathrm{poly}(n, p, k)$ time. The algorithm does not have access to Φ. We only assume $\Phi_{i,i} = 1$ for all $i \in [p]$ and $\max_{i \neq j} |\Phi_{i,j}| < 1$.
2. The key algorithmic technique is to apply a **log-transform** to the features and response, and reduce the problem to a sparse linear regression problem.
3. We analyze how the covariance matrix changes after the log-transform, showing that the log-transform eliminates linear dependencies between two or more features.

Algorithm

We use Lasso for concreteness, but any ℓ_1 minimization method works.

Algorithm 1 Learn Sparse Monomial

Require: data matrix $X \in \mathbb{R}^{n \times p}$, responses $y \in \mathbb{R}^n$, regularization parameter $\vartheta > 0$

1: Apply $\log(|\cdot|)$ transformation to data and responses, element-wise: $\hat{X} \leftarrow \log(|X|)$ and $\hat{y} \leftarrow \log(|y|)$.

2: Solve Lasso optimization problem: $\hat{\beta} \leftarrow \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|\hat{X}\beta - y\|_2^2 + \vartheta \|\beta\|_1$.

3: Select variables: $\hat{S} \leftarrow \{j \in [p] : \hat{\beta}_j \neq 0\}$.

4: **return** $\hat{S}$ and $\hat{\beta}$.

The Restricted Eigenvalue Condition (REC)

The following concept is essential to analyzing the performance of the Lasso estimator, and is the main focus of our analysis.

Definition

For $T \subset [p]$ and $q_0 > 0$, define $\mathcal{C}(q_0, T) := \{v \in \mathbb{R}^p : \|v\|_2 = 1, \|v_{T^c}\|_1 \leq q_0 \|v_T\|_1\}$. T is commonly taken to be the non-zero support S of the sparse vector to recover. We say the (q_0, T, A)*-restricted eigenvalue condition (REC)* is satisfied by matrix $A \in \mathbb{R}^{n \times p}$ if $\tilde{\lambda}(q_0, T, A) := \min_{v \in \mathcal{C}(q_0, T)} \frac{1}{n} \|Av\|_2^2 > 0$. When q_0 and T are apparent from context and $|T| = s$, we will simply write $\tilde{\lambda}(s, A)$.

Performance of the Lasso

The following well-known result about the performance of the estimator $\hat{w}_{\text{Lasso}}(\vartheta)$ is due to [2]; the specific form we state is taken from [3].

Theorem

Consider the model $Aw + \eta = b$, and suppose the support S of $w \in \mathbb{R}^p$ has size k, and the measurement matrix $A \in \mathbb{R}^{n \times p}$ satisfies (q_0, S, A)-REC with $q_0 = 3$. For any $\vartheta > 0$ such that $\vartheta \geq (2/n)\|A^T \eta\|_\infty$, the Lasso estimate $\hat{w}_{\text{Lasso}}(\vartheta)$ satisfies
$$\|w - \hat{w}_{\text{Lasso}}(\vartheta)\|_2 \leq \frac{3\vartheta\sqrt{k}}{\tilde{\lambda}(k, 3, S, A)}.$$

Outline of Analysis

We show REC holds with high probability on the log-transformed Gaussian data in two steps:

- Demonstrate the population REC holds.
- Analyze the fluctuation of the empirical REC.

Main Theorem

Theorem

Let $\delta \in (0, 1)$ be an arbitrary confidence parameter. Suppose the covariance matrix Φ satisfies $\Phi_{i,i} = 1, \forall i \in [p]$ and $\max_{i \neq j} |\Phi_{i,j}| < 1 - \epsilon$. Then, the $\log(|\cdot|)$-transformed design matrix $\hat{X} = \log(|X|)$ for X taken from the model with true support $|S| = k$ satisfies
$$\tilde{\lambda}\left(k, \frac{1}{\sqrt{n}}\hat{X}\right) \geq \frac{1}{5}\sqrt{\frac{\epsilon}{\log(16k) + 2}},$$
with probability $1 - \delta$, provided that
$$n \geq C \cdot \frac{k^2 \log(2k)}{\epsilon} \cdot \log^2\left(\frac{2p}{\delta}\right) \cdot \log^2\left(\frac{k \log(k)}{\epsilon} \log\left(\frac{2p}{\delta}\right)\right). \tag{1}$$
In the above display, C is a universal constant.

Proof Sketch

- We first derive a closed form series expression for the population covariance using a Hermite basis expansion.
- We use this expression to derive a closed form series expression for the population restricted eigenvalue.
- We lower-bound the terms of the REC series with a restricted variant of the Gershgorin circle theorem.
- We finish by showing the log-transformed features are sub-exponential and apply a concentration inequality from [5] to the REC quantity.

Characterizing Covariance of Log-Transformed Gaussians

Let $x \sim \mathcal{N}(0, \Phi)$ where $\Phi_{i,i} = 1$ for all $i \in [p]$, and
$$z := \log(|x|), \qquad \Sigma := \mathbb{E}_z[zz^T], \qquad \hat{\Sigma} := \frac{1}{n}\sum_{i=1}^n z^{(i)} z^{(i)T}.$$
Then:

1. $\mathrm{var}(z_i) = \pi^2/8$.
2. The function $a \mapsto \log(|a|)$ admits the following expansion in the Hermite polynomial basis $\{H_l\}_{l \geq 0}$:
$$\log(|a|) = \Sigma_{l=0}^\infty c_{2l} H_{2l}(a), \quad c_{2l} = \frac{(-1)^{l-1} 2^{l-1} (l-1)!}{\sqrt{(2l)!}}.$$
3. $\mathbb{E}[z_i z_j] = \Sigma_{l=0}^\infty c_{2l}^2 \Phi_{i,j}^{2l}$.
4. $\Sigma = c_0^2 \mathbf{1}_{p \times p} + \Sigma_{l=1}^\infty c_{2l}^2 \Phi^{(2l)}$, where $\mathbf{1}_{p \times p}$ is the $p \times p$ matrix of all 1's.

Related Work

- For k-sparse parity functions, there is an attribute-efficient algorithm with run-time $O(p^{k/2})$ due to Dan Spielman [4], and an attribute-inefficient improper learner with sample complexity $n = O(p^{1-1/k})$ and run-time $O(p^4)$ for the noiseless case with an arbitrary distribution over $\{-1, +1\}^p$ due to [4]. There is also an $O(p^{0.8k} \,\mathrm{poly}\,(1/(1-2\eta))$-time (but attribute-inefficient) algorithm of [6].
- [1] considers the problem of learning s-sparse polynomials of degree d with additive noise over real-valued data, but the data must come from a product distribution.

Discussion

We summarize the conceptual contributions of the paper:

1. **Blessing of non-linearity**: The assumptions on the correlation structure needed to learn a class of sparse non-linear functions are less restrictive than those needed to learn sparse linear functions.
2. The minimum eigenvalue of the log-transformed data covariance matrix is strictly positive with high probability, regardless of initial rank. Thus, **nonlinear data transformations can destroy low-rank covariance structure**.

References

[1] A. Andoni, R. Panigrahy, G. Valiant, and L. Zhang.
Learning sparse polynomial functions.
In *Proceedings of the Twenty-fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 500–510, Philadelphia, PA, USA, 2014. Society for Industrial and Applied Mathematics.

[2] P. J. Bickel, Y. Ritov, and A. B. Tsybakov.
Simultaneous analysis of lasso and dantzig selector.
Annals of Statistics, 37:1705–1732, 2009.

[3] T. Hastie, R. Tibshirani, and M. Wainwright.
Statistical Learning with Sparsity: The Lasso and Generalizations.
Chapman & Hall/CRC, 2015.

[4] A. R. Klivans and R. A. Servedio.
Toward attribute efficient learning of decision lists and parities.
Journal of Machine Learning Research, 7(Apr):587–602, 2006.

[5] A. K. Kuchibhotla and A. Chakrabortty.
Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression.
ArXiv e-prints, 2018.

[6] G. Valiant.
Finding correlations in subquadratic time, with applications to learning parities and the closest pair problem.
Journal of the ACM (JACM), 62(2):13, 2015.