Proceedings of Machine Learning Research vol 65:1–6, 2017

High-Dimensional Regression with Binary Coefficients. Estimating Squared Error and a Phase Transition.

David Gamarnik[*] GAMARNIK@MIT.EDU and **Ilias Zadik** IZADIK@MIT.EDU
Massachusetts Institute of Technology

Abstract

We consider a sparse linear regression model $Y = X\beta^* + W$ where X is $n \times p$ matrix Gaussian i.i.d. entries, W is $n \times 1$ noise vector with i.i.d. mean zero Gaussian entries and standard deviation σ, and β^* is $p \times 1$ binary vector with support size (sparsity) k. Using a novel conditional second moment method we obtain a tight up to a multiplicative constant approximation of the optimal squared error $\min_\beta \|Y - X\beta\|_2$, where the minimization is over all k-sparse binary vectors β. The approximation reveals interesting structural properties of the underlying regression problem. In particular,

(a) We establish that $n^* = 2k \log p / \log(2k/\sigma^2 + 1)$ is a phase transition point with the following "all-or-nothing" property. When n exceeds n^*, $(2k)^{-1}\|\beta_2 - \beta^*\|_0 \approx 0$, and when n is below n^*, $(2k)^{-1}\|\beta_2 - \beta^*\|_0 \approx 1$, where β_2 is the optimal solution achieving the smallest squared error. As a corollary n^* is the asymptotic threshold for recovering β^* information theoretically. Note that n^* is asymptotically below the threshold $n_{\text{LASSO/CS}} = (2k + \sigma^2) \log p$, above which the LASSO and Compressive Sensing methods are able to recover β^*.

(b) We compute the squared error for an intermediate problem $\min_\beta \|Y - X\beta\|_2$ where the minimization is restricted to vectors β with $\|\beta - \beta^*\|_0 = 2k\zeta$, for some fixed ratio $\zeta \in [0, 1]$. We show that a lower bound part $\Gamma(\zeta)$ of the estimate, which essentially corresponds to the estimate based on the first moment method, undergoes a phase transition at three different thresholds, namely $n_{\text{inf},1} = \sigma^2 \log p$, which is information theoretic bound for recovering β^* when $k = 1$ and σ is large, then at n^* and finally at $n_{\text{LASSO/CS}}$.

(c) We establish a certain Overlap Gap Property (OGP) on the space of all k-sparse binary vectors β when $n \le ck \log p$ for sufficiently small constant c. By drawing a connection with a similar OGP exhibited by many randomly generated constraint satisfaction problems and statistical physics models, we conjecture that OGP is the source of algorithmic hardness of solving the minimization problem $\min_\beta \|Y - X\beta\|_2$ in the regime $n < n_{\text{LASSO/CS}}$. [1]

Keywords: Linear regression, High-dimensional inference, Second moment method, Phase transitions.

1. Introduction

We consider a high-dimensional linear regression model of the form $Y = X\beta^* + W$ where X is $n \times p$ matrix, W is $n \times 1$ noise vector, and β^* is $p \times 1$ vector of regression coefficients to be recovered from observing X and Y. A great body of literature is devoted to the problem of identifying the underlying regression vector β^*, assuming its support size (the number of coordinates with non-zero

[*] Research supported by the NSF grants CMMI-1335155.

[1] Extended abstract. Full version available at arXiv https://arxiv.org/abs/1701.04455

coefficients) k is sufficiently small. The support recovery problem has attracted a lot of attention in recent years, because it naturally arises in many contexts including signal denoising Chen et al. (2001) and compressive sensing Candes and Tao (2005),Donoho (2006). In this paper we assume that X has i.i.d. standard normal entries, W has i.i.d normal entries with standard deviations σ, and β^* is a binary vector (all entries are either zero or one). The results in the existing literature discussed below are adopted to this setting.

A lot of work has been devoted in particular to finding computationally efficient ways for recovering the support of β^*. Many algorithms have been proven to succeed w.h.p. when both X, W are Gaussian, but all of them require the sample size to satisfy $n \geq (2k + \sigma^2)\log p$, Meinshausen and Bhlmann (2006), Wainwright (2009b), Zhao and Yu (2006). See also the recent book Foucart and Rauhut (2013). In particular, Wainwright (2009b) showed that if $n \geq (1+\epsilon)(2k+\sigma^2)\log p$, for some $\epsilon > 0$ then the optimal solutions of LASSO $\min_{\beta \in \mathbb{R}^p}\{||Y - X\beta||_2^2 + \lambda_p||\beta||_1\}$, for appropriately chosen $\lambda_p > 0$, recovers the support of β^* exactly w.h.p. Furthermore, orthogonal matching pursuit, a simple and popular greedy algorithm, has also been proven to work given again that σ^2 satisfies $\frac{\sigma^2}{k} \to 0$ and $n > (1+\epsilon)\,(2k + \sigma^2)\log p$, Fletcher and Rangan (2009). We note that the impact of σ^2 on this threshold value is asymptotically negligible when $\sigma^2/k \to 0$. It will be convenient for us to maintain it though. Thus we denote $(2k + \sigma^2)\log p$ by $n_{\text{LASSO/CS}}$. At the present time no tractable (polynomial time) algorithms are known for the support recovery when $n \leq n_{\text{LASSO/CS}}$.

On the complimentary direction, an easy corollary of Theorem 2 in Wainwright (2009a), when applied to our context below involving vectors β^* with binary values, shows that if $n < (1-\epsilon)\,\sigma^2 \log p$ then for every support recovery algorithm, a binary vector β^* can be constructed in such a way that the underlying algorithm fails to recover β^* exactly, with probability at least $\frac{\epsilon}{2}$. We let $n_{\text{inf},1} \triangleq \sigma^2 \log p$. The best known information theoretic bound at the moment follows from an improvement of the previous argument in Wang et al. (2010) where it is established that the exact recovery of β^* is information theoretically impossible when n is smaller than $n^* \triangleq 2k\log p/\log(1 + 2k/\sigma^2)$, where n^* is the information theoretic limit of this Gaussian channel for general k, σ^2. Based on the above discussion the regime $n \in [n_{\text{inf},1}, n_{\text{LASSO/CS}}]$ remains largely unexplored from the algorithmic perspective, and the present paper is devoted to studying this regime. Towards this goal, for the regression model $Y = X\beta^* + W$, we consider the corresponding maximum likelihood estimation problem:

$$
\begin{array}{lll}
(\Phi_2) & \min & n^{-\frac{1}{2}}\|Y - X\beta\|_2 \\
 & \text{s.t.} & \beta \in \{0,1\}^p, \\
 & & \|\beta\|_0 = k,
\end{array}
$$

where $\|\beta\|_0$ is the sparsity of β. Namely, it is the cardinality of the set $\{i \in [p] \big| \beta_i \neq 0\}$. We denote by ϕ_2 its optimal value and by β_2 its unique optimal solution. As above, the matrix X is assumed to have i.i.d. standard normal entries, the elements of the noise vector W are assumed to have i.i.d. zero mean normal entries with variance σ^2, and the vector β^* is assumed to be binary k-sparse; $\|\beta^*\|_0 = k$. In particular, we assume that the sparsity k is known to the optimizer.

We address two questions in this paper: (a) What is the value of the squared error estimator $\min_{\beta \in \{0,1\}^p, \|\beta\|_0 = k}\|Y - X\beta\|_2 = \|Y - X\beta_2\|_2$; and (b) how well does the optimal vector β_2 approximate the ground truth vector β^*?

Besides providing a fairly complete answer to these two questions, we also reveal also a geometric property, we call Overlap Gap Property (OGP), in the space of binary k-sparse vectors which holds when $n \in [n_{\inf,1}, cn_{\text{LASSO/CS}}]$ for some small constant $c > 0$ w.h.p. This and similar properties are known to be appearing in many random constraint satisfaction problems Achlioptas et al. (2011), Achlioptas and Coja-Oghlan (2008), Montanari et al. (2011), Coja-Oghlan and Efthymiou (2011), Gamarnik and Sudan (a), Rahman and Virag (2014), Gamarnik and Sudan (b),Gamarnik and Li (2016). It was conjectured that when they appear, they provide fundamental algorithmic barriers for the random constraint satisfaction problem under consideration Achlioptas and Coja-Oghlan (2008). In particular such properties were used in Gamarnik and Sudan (a), Rahman and Virag (2014), Gamarnik and Sudan (b) and Coja-Oghlan et al. (2016) to establish a fundamental barrier on the power of so-called local algorithms for solving certain types of random constraint satisfaction problems. Drawing a connection with these results, we propose the presence of OGP as an evidence of the algorithmic hardness of the problem when $n \in [n_{\inf,1}, cn_{\text{LASSO/CS}}]$ for some $c > 0$.

2. Model and the Main Results

In order to recover β^*, we consider the following constrained optimization problem (Φ_2), defined in Introduction. We denote by $\phi_2 = \phi_2(X, W)$ its optimal value and by β_2 its (unique almost surely) optimal solution.

Consider the following restricted version of the problem Φ_2:

$$\begin{array}{lll} (\Phi_2(\ell)) & \min & n^{-\frac{1}{2}}||Y - X\beta||_2 \\ & \text{s.t.} & \beta \in \{0,1\}^p \\ & & ||\beta||_0 = k, ||\beta - \beta^*||_0 = 2l, \end{array}$$

where $\ell = 0, 1, 2, .., k$. For every fixed ℓ, denote by $\phi_2(\ell)$ the optimal value of $\Phi_2(\ell)$. Clearly $\phi_2 = \min_\ell \phi_2(\ell)$.

Consider for example the extreme cases $\ell = 0$ and $\ell = k$. For $\ell = 0$, the region that defines $\Phi_2(0)$ consists only of the vector β^*. On the other hand, for $\ell = k$, the region that defines $\Phi_2(k)$ consists of all k-sparse binary vectors β, whose common support with β^* is empty.

Our first main result is a structural result for the asymptotic behavor of $\phi_2(\ell)$ for $\ell = 0, 1, 2, \ldots, k$.

Theorem 1 *Suppose $k \log k \leq Cn$ for some constant C for all k, n. Then*

(a) W.h.p. as k increases

$$\phi_2(\ell) \geq e^{-\frac{3}{2}} \sqrt{2l + \sigma^2} \exp\left(-\frac{\ell \log p}{n}\right), \tag{1}$$

for all $0 \leq \ell \leq k$.

(b) Suppose further that $\sigma^2 \leq 2k$. Then for every sufficiently large constant D_0 if it holds also $n \leq k \log p/(3 \log D_0)$, then w.h.p. as k increases, the cardinality of the set

$$\left\{\beta : \|\beta - \beta^*\|_\infty = 2k,\ n^{-\frac{1}{2}}\|Y - X\beta\|_2 \leq D_0\sqrt{2k + \sigma^2} \exp\left(-\frac{k \log p}{n}\right)\right\} \tag{2}$$

is at least $D_0^{\frac{n}{3}}$. In particular, this set is exponentially large in n.

Now we will discuss some implications of Theorem 1. The expression $\left(2\ell+\sigma^2\right)^{\frac{1}{2}}\exp\left(-\frac{\ell\log p}{n}\right)$, appearing in the theorem above, motivates the following notation. Let the function $\Gamma:[0,1]\to\mathbb{R}_+$ be defined by

$$\Gamma\left(\zeta\right)=\left(2\zeta k+\sigma^2\right)^{\frac{1}{2}}\exp\left(-\frac{\zeta k\log p}{n}\right). \tag{3}$$

Then the lower bound (1) can be rewritten as

$$\phi_2\left(\ell\right)\geq e^{-\frac{3}{2}}\Gamma(\ell/k).$$

A similar inequality applies to (2).

An easy monotonicity analysis for the function Γ reveals the following interesting behavior with respect to the thresholds $n_{\mathrm{inf},1}, n^*, n_{\mathrm{LASSO/CS}}$, which are defined in the introduction.

Proposition 2 *The function Γ satisfies the following properties.*

1. *When $n\leq\sigma^2\log p$, Γ is a strictly decreasing function of ζ.*
2. *When $\sigma^2\log p<n<n^*$, Γ is not monotonic and it attains its minimum at $\zeta=1$.*
3. *When $n=n^*$, Γ is not monotonic and it attains its minimum at $\zeta=0$ and $\zeta=1$.*
4. *When $n^*<n<(2k+\sigma^2)\log p$, Γ is not monotonic and it attains its minimum at $\zeta=0$.*
5. *When $n>\left(2k+\sigma^2\right)\log p$, Γ is a strictly increasing function of ζ.*

We use Theorem 1 and the intuition from Proposition 2 to obtain a tight characterization of the performance of Φ_2. Specifically we reveal the following sharp phase transition behavior.

Theorem 3 *Let $\epsilon>0$ be arbitrary. Suppose $\max\{k,\frac{2k}{\sigma^2}+1\}\leq\exp\left(\sqrt{C\log p}\right)$, for some $C>0$ for all k and n. Suppose furthermore that $k\to\infty$ and $\sigma^2/k\to0$ as $k\to\infty$. If $n\geq(1+\epsilon)\,n^*$, then w.h.p. as k increases*

$$\frac{1}{2k}||\beta_2-\beta^*||_0\to0.$$

On the other hand if $\frac{1}{C}k\log k\leq n\leq(1-\epsilon)\,n^*$, then w.h.p. as k increases

$$\frac{1}{2k}||\beta_2-\beta^*||_0\to1.$$

As explained in the introduction, to get an insight into possible reason for the apparent algorithmic hardness of the problem in the regime $n\in[n_{\mathrm{inf},1},n_{\mathrm{LASSO/CS}}]$ we reveal a certain Overlap Gap Property (OGP) on the space of k-sparse binary vectors. We establish in particular, that in this regime the solutions β which are sufficiently "close" to optimality break into two separate clusters – those which are close in $\|\cdot\|_0$ norm to the optimal solution β_2, namely those which have a "large" overlap with β_2, and those which are far from it, namely those which have a "small" overlap with β_2. In the next theorem we establish that the OGP indeed takes place, when the sampling size is bounded away by a constant from $\max\{k\log k, n_{\mathrm{inf},1}\}$ and $n_{\mathrm{LASSO/CS}}$. Given any $r\geq0$, let

$$S_r:=\{\beta\in\{0,1\}^p:||\beta||_0=k,n^{-\frac{1}{2}}||Y-X\beta||_2<r\}.$$

Theorem 4 (The Overlap Gap Property) *Suppose the assumptions of Theorem 1 hold. Suppose in addition $\sigma^2 \to +\infty$. For every sufficiently large constant D_0 there exist sequences $0 < \zeta_{1,k,n} < \zeta_{2,k,n} < 1$ satisfying*

$$\lim_{k\to\infty} k\left(\zeta_{2,k,n} - \zeta_{1,k,n}\right) = +\infty,$$

as $k \to \infty$, and such that if $r_k = D_0 \max\left(\Gamma(0), \Gamma(1)\right)$ and $\max\{\frac{1}{C} k \log k, \left(e^7 D_0^2 + 1\right)\sigma^2 \log p\} \leq n \leq k \log p/(3 \log D_0)$ then w.h.p. as k increases the following holds

(a) For every $\beta \in S_{r_k}$

$$(2k)^{-1}\|\beta - \beta^*\|_0 < \zeta_{1,k,n} \text{ or } (2k)^{-1}\|\beta - \beta^*\|_0 > \zeta_{2,k,n}.$$

(b) $\beta^* \in S_{r_k}$. In particular the set

$$S_{r_k} \cap \{\beta : (2k)^{-1}\|\beta - \beta^*\|_0 < \zeta_{1,k,n}\}$$

is non-empty.

(c) The cardinality of the set

$$|S_{r_k} \cap \{\beta : \|\beta - \beta^*\|_0\} = 2k\}|,$$

is at least $D_0^{\frac{n}{3}}$. In particular the set $S_{r_k} \cap \{\beta : \|\beta - \beta^*\|_0\} = 2k\}$ has exponentially many in n elements.

Acknowledgments

The authors would like to thank Philippe Rigollet for helpful discussions during the preparation of this paper.

References

D. Achlioptas, A. Coja-Oghlan, and F. Ricci-Tersenghi. On the solution space geometry of random formulas. *Random Structures and Algorithms*, 38:251–268, 2011.

Dimitris Achlioptas and Amin Coja-Oghlan. Algorithmic barriers from phase transitions. In *Foundations of Computer Science, 2008. FOCS'08. IEEE 49th Annual IEEE Symposium on*, pages 793–802. IEEE, 2008.

Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, January 2001. ISSN 0036-1445. doi: 10.1137/S003614450037906X. URL http://dx.doi.org/10.1137/S003614450037906X.

A. Coja-Oghlan and C. Efthymiou. On independent sets in random graphs. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 136–144. SIAM, 2011.

Amin Coja-Oghlan, Amir Haqshenas, and Samuel Hetterich. Walksat stalls well below the satisfiability threshold. *arXiv preprint arXiv:1608.00346*, 2016.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Alyson K. Fletcher and Sundeep Rangan. Orthogonal matching pursuit from noisy measurements: A new analysis. In *Proceedings of the 22Nd International Conference on Neural Information Processing Systems*, NIPS'09, pages 540–548, USA, 2009. Curran Associates Inc. ISBN 978-1-61567-911-9. URL http://dl.acm.org/citation.cfm?id=2984093.2984154.

Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Springer, 2013.

David Gamarnik and Quan Li. Finding a large submatrix of a gaussian random matrix. *arXiv preprint arXiv:1602.08529*, 2016.

David Gamarnik and Madhu Sudan. Limits of local algorithms over sparse random graphs. *Annals of Probability*. To appear, a.

David Gamarnik and Madhu Sudan. Performance of sequential local algorithms for the random nae-k-sat problem. *SIAM Journal on Computing*. To appear, b.

Nicolai Meinshausen and Peter Bhlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 06 2006. doi: 10.1214/009053606000000281. URL http://dx.doi.org/10.1214/009053606000000281.

Andrea Montanari, Ricardo Restrepo, and Prasad Tetali. Reconstruction and clustering in random constraint satisfaction problems. *SIAM Journal on Discrete Mathematics*, 25(2):771–808, 2011.

Mustazee Rahman and Balint Virag. Local algorithms for independent sets are half-optimal. *arXiv preprint arXiv:1402.0485*, 2014.

Martin J Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *Information Theory, IEEE Transactions on*, 55(12):5728–5741, 2009a.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5): 2183–2202, 2009b.

Wei Wang, Martin J Wainwright, and Kannan Ramchandran. Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *Information Theory, IEEE Transactions on*, 56(6):2967–2979, 2010.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, December 2006. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1248547.1248637.