Proceedings of Machine Learning Research vol 125:1–4, 2020 33rd Annual Conference on Learning Theory

Free Energy Wells and Overlap Gap Property in Sparse PCA (Extended Abstract)

Gérard Ben Arous BENAROUS@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences, New York University

Alexander S. Wein AWEIN@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences, New York University

Ilias Zadik ZADIK@NYU.EDU
Center for Data Science, New York University

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We study a variant of the sparse PCA (principal component analysis) problem in the "hard" regime, where the inference task is possible yet no polynomial-time algorithm is known to exist. Prior work, based on the low-degree likelihood ratio, has conjectured a precise expression for the best possible (sub-exponential) runtime throughout the hard regime. Following instead a statistical physics inspired point of view, we show bounds on the depth of free energy wells for various Gibbs measures naturally associated to the problem. These free energy wells imply hitting time lower bounds that corroborate the low-degree conjecture: we show that a class of natural MCMC (Markov chain Monte Carlo) methods (with worst-case initialization) cannot solve sparse PCA with less than the conjectured runtime. These lower bounds apply to a wide range of values for two tuning parameters: temperature and sparsity misparametrization. Finally, we prove that the Overlap Gap Property (OGP), a structural property that implies failure of certain local search algorithms, holds in a significant part of the hard regime.[1]

1. The Model

We consider the following variant of sparse PCA in the spiked Wigner model (also called *principal submatrix recovery*). Let W be a $\mathrm{GOE}(n)$ matrix, i.e., $n \times n$ symmetric with off-diagonal entries $\mathcal{N}(0, 1/n)$ and diagonal entries $\mathcal{N}(0, 2/n)$, all independent aside from the symmetry $W_{ij} = W_{ji}$. Let x be an unknown k-sparse vector in $\{0,1\}^n$, i.e., exactly k entries are equal to 1. We are interested in recovering x from the observation

$$Y = \frac{\lambda}{k} x x^\top + W$$

where $\lambda > 0$ is the *signal-to-noise ratio*. We study the problem in the limit $n \to \infty$, where the parameters $\lambda = \lambda_n$ and $k = k_n$ may depend on n. We are primarily interested in the *exact recovery* problem: we study algorithms which given Y, output x *with high probability*, i.e., probability tending to 1 as $n \to \infty$. Our regime of interest will be $1 \ll k \ll n$. Throughout, we use the notation $\ll$ to hide factors of $n^{o(1)}$ (although in most cases, $\ll$ will only hide logarithmic factors).

1. This paper is an extended abstract. The full version appears as arXiv preprint *arXiv:2006.10689v1*.

2. Our Contributions

Prior work suggests the existence of a "hard regime" $\sqrt{k/n} \ll \lambda \ll \min\{1, k/\sqrt{n}\}$ where exact recovery is information-theoretically possible but no polynomial-time algorithm is known (see e.g. Baik et al. (2005); Féral and Péché (2007); Amini and Wainwright (2008); Benaych-Georges and Nadakuditi (2009); Johnstone and Lu (2009); Banks et al. (2018); Ding et al. (2019)). More specifically, the work of Ding et al. (2019) suggests the following conjecture regarding a precise expression for the best possible (sub-exponential) runtime throughout the hard regime.

Conjecture 1 *Consider the sparse PCA problem as defined in Section 1. For any λ in the "hard" regime $\sqrt{k/n} \ll \lambda \ll \min\{1, k/\sqrt{n}\}$, any algorithm requires runtime $\exp\left(\tilde{\Omega}\left(\frac{k^2}{\lambda^2 n}\right)\right)$ to achieve exact recovery.*

This prediction is made by Ding et al. (2019) (for a variant of our model where $x_i \in \{0, -1, 1\}$) using the *low-degree likelihood ratio* (Hopkins and Steurer, 2017; Hopkins et al., 2017; Hopkins, 2018), which amounts to studying the power of algorithms based on low-degree polynomials. There are known algorithms which achieve the matching runtime $\exp\left(\tilde{O}\left(\frac{k^2}{\lambda^2 n}\right)\right)$ (Ding et al., 2019; Holtzman et al., 2019). For instance, the following simple algorithm of Ding et al. (2019) proceeds in two steps. The first step is to let $k' \approx \frac{k^2}{\lambda^2 n}$ and solve, by exhaustive search, the optimization problem

$$\underset{v \in S_{k'}}{\operatorname{argmax}}\, v^\top Y v \tag{1}$$

where $S_{k'}$ is the space of k'-sparse vectors

$$S_{k'} = \{v \in \{0,1\}^n \,:\, \|v\|_0 = k'\}, \tag{2}$$

and the final step uses the optimizer v^* to exactly recover x via a simple boosting procedure.

In this work we give evidence in support of Conjecture 1 by showing

- the existence of *free energy wells* (Ben Arous et al., 2018; Gamarnik and Zadik, 2017b, 2019; Gamarnik et al., 2019) in the Gibbs measure (at various temperatures) associated with the optimization problem (1),
- and the existence of the *overlap gap property* (Gamarnik and Zadik, 2017a,b; Gamarnik et al., 2019; Gamarnik and Zadik, 2019; Zadik, 2019) in the space of feasible solutions of the optimization problem (1),

for various choices of the tuning parameter k'. As explained in the full version of the present work, a free energy well of depth D at inverse temperature β implies that a certain class of MCMC methods with parameter β requires time at least $\exp(\Omega(D))$ to solve (1). Our main result can be stated informally as follows.

Theorem (Main result, informal) *Suppose λ is in the "hard" regime $\sqrt{k/n} \ll \lambda \ll \min\{1, k/\sqrt{n}\}$ and that additionally, $\lambda \ll (k/n)^{1/4}$. For any "informative" k' and any $\beta \geq 0$ (possibly depending on n), there exists a free energy well of depth $\tilde{\Omega}\left(\frac{k^2}{\lambda^2 n}\right)$ and the overlap gap property holds, with high probability.*

Here "informative" k' refers to the condition, $\frac{k^2}{\lambda^2 n} \lesssim k' \lesssim \lambda^2 n$, which captures the k' values for which solving the optimization problem (1) is actually useful in the sense that a near-optimal solution can be used to exactly recover x via a simple boosting procedure. Our main result shows that if the condition $\lambda \ll (k/n)^{1/4}$ is satisfied then MCMC cannot improve the runtime of Ding et al. (2019); Holtzman et al. (2019) for *any* choice of inverse temperature β and *any* (informative) choice of misparametrization k'. The main weakness of the result is the condition $\lambda \ll (k/n)^{1/4}$, which is an artifact of the proof. However, in the relatively sparse regime $k \ll n^{1/3}$, the condition $\lambda \ll (k/n)^{1/4}$ holds throughout the entire "hard" regime. Thus we obtain a complete refutation of MCMC methods (across all β and k') throughout a large range of sparsity values (namely $k \ll n^{1/3}$).

Our results are actually somewhat stronger than what we have stated here: even when the condition $\lambda \ll (k/n)^{1/4}$ does not hold, the result still holds for some k' values; in particular, it always holds for all informative $k' \leq k$. One consequence of this is that it is not possible to speed up the algorithm of Ding et al. (2019) by taking their choice of $k' \approx \frac{k^2}{\lambda^2 n}$ (the smallest "informative" k', which in particular is less than k) and solving (1) via MCMC instead of exhaustive search.

Remark 2 *In order for a computational hardness result to be most compelling, the class of algorithms ruled out should capture the best known algorithms. This is indeed the case here in the sense that there exists a choice of parameters, namely $k' \approx \frac{k^2}{\lambda^2 n}$ and $\beta = 0$, for which MCMC (followed by boosting) mimics the algorithm of Ding et al. (2019) and achieves the runtime $\exp\left(\tilde{O}\left(\frac{k^2}{\lambda^2 n}\right)\right)$. For this choice of parameters, MCMC is simply a random walk (ignoring the data Y) on the space of k'-sparse vectors, which will visit all states within time $\exp(\tilde{O}(k')) = \exp\left(\tilde{O}\left(\frac{k^2}{\lambda^2 n}\right)\right)$ with high probability. A consequence is that for this choice of k' and β, any free energy well has depth $\tilde{O}\left(\frac{k^2}{\lambda^2 n}\right)$ and so our lower bound is tight. It is not clear whether MCMC with more natural parameters (e.g. $k' = k$) matches the above runtime; it might in fact be strictly worse. This highlights the importance of allowing $k' \neq k$ in our main result.*

Acknowledgments

A.S.W. is partially supported by NSF grant DMS-1712730 and by the Simons Collaboration on Algorithms and Geometry. I.Z. is supported by a CDS Moore-Sloan postdoctoral fellowship. The authors would like to thank David Gamarnik for helpful comments on an earlier draft of this work. We also thank the anonymous reviewers for their helpful comments.

References

Arash A Amini and Martin J Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. In *2008 IEEE International Symposium on Information Theory*, pages 2454–2458. IEEE, 2008.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

Jess Banks, Cristopher Moore, Roman Vershynin, Nicolas Verzelen, and Jiaming Xu. Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization. *IEEE Transactions on Information Theory*, 64(7):4872–4894, 2018.

Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Algorithmic thresholds for tensor PCA. *arXiv preprint arXiv:1808.00921*, 2018.

Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *arXiv preprint arXiv:0910.2120*, 2009.

Yunzi Ding, Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Subexponential-time algorithms for sparse PCA. *arXiv preprint arXiv:1907.11635*, 2019.

Delphine Féral and Sandrine Péché. The largest eigenvalue of rank one deformation of large wigner matrices. *Communications in mathematical physics*, 272(1):185–228, 2007.

David Gamarnik and Ilias Zadik. High dimensional linear regression with binary coefficients: Mean squared error and a phase transition. *Conference on Learning Theory (COLT)*, 2017a. URL https://arxiv.org/abs/1701.04455.

David Gamarnik and Ilias Zadik. Sparse high dimensional linear regression: Algorithmic barrier and a local search algorithm. *arXiv Preprint*, 2017b. URL https://arxiv.org/abs/1711.04952.

David Gamarnik and Ilias Zadik. The landscape of the planted clique problem: Dense subgraphs and the overlap gap property, 2019.

David Gamarnik, Aukosh Jagannath, and Subhabrata Sen. The overlap gap property in principal submatrix recovery. *arXiv preprint arXiv:1908.09959*, 2019.

Guy Holtzman, Adam Soffer, and Dan Vilenchik. A greedy anytime algorithm for sparse PCA. *arXiv preprint arXiv:1910.06846*, 2019.

Samuel Hopkins. *Statistical Inference and the Sum of Squares Method*. PhD thesis, Cornell University, 2018.

Samuel B Hopkins and David Steurer. Bayesian estimation from few samples: community detection and related problems. *arXiv preprint arXiv:1710.00264*, 2017.

Samuel B Hopkins, Pravesh K Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The power of sum-of-squares for detecting hidden structures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 720–731. IEEE, 2017.

Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

Ilias Zadik. *Computational and statistical challenges in high dimensional statistical models.* PhD thesis, Massachusetts Institute of Technology; Cambridge MA, 2019.