

# The nonsmooth landscape of blind deconvolution

Mateo Díaz

MD825@CORNELL.EDU

Center for Applied Mathematics, Cornell University, Ithaca, NY 14853

## Abstract

The blind deconvolution problem aims to recover a rank-one matrix from a set of rank-one linear measurements. Recently, Charisopoulos et al. [9] introduced a nonconvex nonsmooth formulation that can be used, in combination with an initialization procedure, to provably solve this problem under standard statistical assumptions. In practice, however, initialization is unnecessary. As we demonstrate numerically, a randomly initialized subgradient method consistently solves the problem. In pursuit of a better understanding of this phenomenon, we study the random landscape of this formulation. We characterize in closed form the landscape of the population objective and describe the approximate location of the stationary points of the sample objective. In particular, we show that the set of spurious critical points lies close to a codimension two subspace. In doing this, we develop tools for studying the landscape of a broader family of singular value functions, these results may be of independent interest.

## 1. Introduction

An increasing amount of research has shown how matrix recovery problems, which in the worst case are hard, become tractable under appropriate statistical assumptions. Examples include phase retrieval [7, 17, 42], blind deconvolution [9, 31], matrix sensing [36, 45], matrix completion [19, 41], and robust PCA [6, 43], among others [5, 8, 30, 40]. Convex relaxations have proven to be a great tool to tackle these problems, but they often require lifting the problem to a higher dimensional space and consequently end up being computationally expensive. Thus, focus has shifted back to iterative methods for nonconvex formulations that operate in the natural parameter space. One of the difficulties of nonconvex optimization is that, in general, it is hard to find global minimizers. To overcome this issue, recent works have suggested two stage methods: One starts by running an *initialization procedure* – usually based on spectral techniques – and then refines the solution by warm-starting a *local search method* that minimizes a nonconvex formulation. This thread of ideas has proven very successful, and we refer the reader to [10] for a survey.

Initialization procedures are nontrivial to develop and can sometimes be more expensive than the refinement stage. Thus, it is important to understand when initialization methods are superfluous. There are iterative methods, for specific problems, that provably converge to minimizers [5, 12, 19, 20, 26]. Analysis of these methods are of two types: those based on studying the iterate sequence [16, 25, 46], and those based on characterizing the landscape of smooth loss functions [20, 33, 39].

In this work, we study the landscape of a nonsmooth nonconvex formulation (2) for the (real) *blind deconvolution* problem. Unlike the aforementioned works, we consider a nonsmooth loss, which presents fundamentally different technical challenges. We show that, as the number of measurements grow, the set of spurious stationary points converges to a codimension two subspace. This suggests that there is an extensive region with friendly geometry.

The blind deconvolution problem aims to recover a pair of real vectors  $(\bar{w}, \bar{x}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$  from a set of  $m$  observations given by

$$y_i = \langle a_i, \bar{w} \rangle \langle \bar{x}, b_i \rangle \quad \text{for all } i \in 1, \dots, m \quad (1)$$

where  $a_i$  and  $b_i$  are known vectors for all indices. This problem has important applications in a variety of different fields, we describe two below.

**Signal processing.** The complex analogue of this problem is intimately linked to the problem of recovering a pair of vectors  $(u, v)$  from the convolution  $(Au) * (Bv)$ , where  $A$  and  $B$  are tall-skinny matrices. In fact, when passed to the frequency domain, this problem becomes equivalent to the one mentioned above. This problem has applications in image deblurring and channel protection with random codes [1, 39].

**Shallow neural networks.** Solving this problem is equivalent to learning the weights of a shallow neural network with bilinear activation functions. Taking  $\{(a_i, b_i), y_i\}$  as training data, writing the output of the network as  $y = \sigma(a^\top w, b^\top x)$ , with  $(w, x) \in \mathbb{R}^{d_1 \times d_2}$ , and the setting the activation function to  $\sigma(z_1, z_2) = z_1 z_2$ .

To tackle the blind deconvolution problem, [9] proposed the following nonconvex nonsmooth formulation

$$\operatorname{argmin}_{w, x} f_S(w, x) \triangleq \frac{1}{m} \sum_{i=1}^m |\langle a_i, w \rangle \langle x, b_i \rangle - y_i|. \quad (2)$$

The authors of [9] designed a two-stage method based on this formulation and showed that if the measuring vectors,  $a_i$  and  $b_i$ , are i.i.d standard Gaussian, then their algorithm converges rapidly to a solution whenever  $m \gtrsim (d_1 + d_2)$ .<sup>1</sup> Nonetheless, experimentally the initialization stage seems to be superfluous. Indeed, a simple randomly-initialized subgradient algorithm is successful at solving the problem most of the time provided that  $m/(d_1 + d_2)$  is big enough. See Section 3 for experimental evidence of this claim.

Aiming to get a better understanding of the high-dimensional geometry of  $f_S$ , we study the landscape of  $f_S$  when  $A$  and  $B$  are standard Gaussian random matrices. Following the line of ideas in [13], we think of  $f_S$  as the empirical average approximation of the *population objective*

$$f_P(w, x) \triangleq \mathbb{E} f_S(w, x) = \mathbb{E} \left( |a^\top (wx - \bar{w}\bar{x}) b^\top| \right),$$

where  $a \in \mathbb{R}^{d_1}$  and  $b \in \mathbb{R}^{d_2}$  are standard Gaussian vectors. From now on, we will refer to  $f_S$  as the *sample objective*. The rationale is simple: we will describe the stationary points of  $f_P$ , then we will prove that the graph of the subdifferential  $\partial f_S$  concentrates around the graph of  $\partial f_P$  and combine these to describe the landscape of  $f_S$ .<sup>2</sup> This strategy allows us to show that the set of spurious stationary points converges to a codimension two subspace at a controlled rate. We remark that these results are geometrical and not computational.

**Related work.** There is a vast recent literature on blind deconvolution. A variety of algorithmic solutions have been proposed, including convex relaxations [1, 2], Riemannian optimization methods [21], gradient descent algorithms [31, 34], and nonsmooth procedures [9]. Related to this work, the authors of [24, 44] studied variations of the blind deconvolution problem via landscape analysis;

1. This is information-theoretically optimal up to constants [11, 23].

2. We will give a formal definition of  $\partial f$  in Section A.

their approach is based on smooth formulations and therefore their tools are of a different nature. Besides algorithms, researchers have also been interested in information-theoretical limits of the problem under different assumptions [11, 23, 32].

On the other hand, the study of the high-dimensional landscape of nonconvex formulations is an emergent area of research. Examples for smooth formulations include the analysis for phase retrieval [39], matrix completion [19], robust PCA [20], and synchronization networks [33]. The majority of these results focus on using second order information to show that under reasonable assumptions the formulations do not exhibit spurious stationary points. The machinery developed for nonsmooth formulations is based on different ideas and is more case-oriented. Despite this, there are remarkable examples [3, 13, 18, 22]. Our work is closer to the paper [13]; the authors of this article studied a similar nonsmooth formulation for the phase retrieval problem, which can be regarded as a symmetric analogue of blind deconvolution.

## 2. Main Contributions

In this section we state our contributions, all the proofs are defer to appendices. Before starting, let us observe that one can only wish to recover the pair  $(\bar{w}, \bar{x})$  up to scaling. In fact, the measurements (1) are invariant under the mapping  $(w, x) \mapsto (\alpha w, x/\alpha)$  for any  $\alpha \neq 0$ . Hence the set of solutions of the problems is defined as  $\mathcal{S} \triangleq \{(\alpha \bar{w}, \bar{x}/\alpha) \mid \alpha \in \mathbb{R} \setminus \{0\}\}$ .

### 2.1. Population objective

We start by observing that  $f_P$  is a singular value function, see Appendix A for a formal definition. Indeed, if we set  $X = wx^\top - \bar{w}\bar{x}^\top$  then due to the orthogonal invariance of the Gaussian distribution we get

$$f_P(w, x) = \mathbb{E}|a^\top U \text{diag}(\sigma(X)) V^\top b| = \mathbb{E}|\sigma_1(X)a_1b_1 + \sigma_2(X)a_2b_2|, \quad (3)$$

where of course  $U \text{diag}(\sigma(X)) V^\top$  is the singular value decomposition of  $X$ . This simple observation leads to our first result, a closed form characterization of this function in terms  $\sigma(X)$ . We defer the proof to Appendix C.

**Proposition 1 (Population objective)** *The population objective can be written as*

$$f_P(w, x) = \sigma_{\max}(X) \sum_{n=0}^{\infty} \left( \frac{(2n)!}{2^{2n}(n!)^2} \right)^2 \frac{(1 - \kappa^{-2}(X))^n}{1 - 2n} \quad (4)$$

where  $X = wx^\top - \bar{w}\bar{x}^\top$  and  $\kappa(X) = \sigma_{\max}(X)/\sigma_{\min}(X)$  is the condition number of  $X$ .

The next result characterizes in closed form the set of critical points of the population objective. To derive this result, we develop tools to study the critical points of a family of singular value functions. These arguments and the proof of this theorem can be found in Appendix B.

**Theorem 2** *The set of critical points of the population objective  $f_P$  is exactly*

$$\mathcal{S} \cup \{0\} \cup \{(w, x) \mid \langle w, \bar{w} \rangle = 0, \langle x, \bar{x} \rangle = 0, \text{ and } wx^\top = 0\}.$$

Revealing that the set of extraneous critical points of  $f_P$  is the subspace  $(\bar{w}, 0)^\perp \cap (0, \bar{x})^\perp$ . When the signal  $(\bar{w}, \bar{x})$  lives in  $\mathbb{R}^2$  the landscape of the population objective is rather simple, the only critical points are the solutions and zero, see Figure 2 in the appendix. This is not the case in higher dimensions where an entire subspace of critical points appear.

## 2.2. Sample objective

In this section we describe the approximate locations of the critical points of the sample objective. Unlike in the smooth case, nonsmooth losses do not exhibit point-wise concentration of the subgradients, or in other words,  $\partial f_S(w, x)$  doesn't converges to  $\partial f_P(w, x)$ . To overcome this obstacle, we show that the graph of  $\partial f_S$  approaches that of  $\partial f_P$  at a quantifiable rate. Intuitively, this means that if  $(w, x)$  is a critical point of  $f_S$ , then nearby there exists a point  $(\hat{w}, \hat{x})$  such that  $\text{dist}(0, \partial f_P(\hat{w}, \hat{x}))$  small.

The following result can be regarded as an analogous version of Theorem 2 for the sample objective. However, the proof of this result is more involved and will require us to study the location of epsilon critical points of the population. We defer the development of these arguments to Appendices E and F.

**Theorem 3** *Consider the sample objective (2) generated with two Gaussian matrices  $A$  and  $B$ . For any fixed  $\nu > 1$  there exist numerical constants  $c_1, c_2, c_3 > 0$  such that if  $m \geq c_1(d_1 + d_2 + 1)$ , then with probability at least  $1 - c_2 \exp(-c_3(d_1 + d_2 + 1))$ , every stationary point  $(w, x)$  of  $f_S$  for which  $\|(w, x)\| \leq \nu \|(\bar{w}, \bar{x})\|$  satisfies at least one of the following conditions*

1. *Near zero.*

$$\|(w, x)\| \leq \|(\bar{w}, \bar{x})\| \Delta,$$

2. *Near a solution.*

$$\|wx^\top - \bar{w}\bar{x}^\top\| \lesssim (\nu^2 + 1) \|\bar{w}\bar{x}^\top\| \Delta,$$

3. *Near orthogonal.*

$$\begin{cases} |\langle w, \bar{w} \rangle| \lesssim (\nu^2 + 1) \|(w, x)\| \|\bar{w}\| \Delta, \\ |\langle x, \bar{x} \rangle| \lesssim (\nu^2 + 1) \|(w, x)\| \|\bar{x}\| \Delta. \end{cases}$$

where  $\Delta = \left( \frac{d_1 + d_2 + 1}{m} \log \left( \frac{m}{d_1 + d_2 + 1} \right) \right)^{\frac{1}{8}}$ .

We remark that one can further prove that with high probability, there exists a neighborhood around the solutions set  $\mathcal{S}$  in which the only critical points are the solutions [9]. Hence at the cost of potentially increasing  $c_1$ , the second condition can be strengthened to  $(\bar{w}, \bar{x}) \in \mathcal{S}$ .

## 3. Experiments

In this section we empirically investigate the behavior of a randomly-initialized subgradient algorithm applied to (2). It is known that well-tuned subgradient algorithms converge to critical points for any locally Lipschitz function [15]. Further, this type of iterative procedures are computationally cheap, easy to implement, and widely used in practice. This makes them a great proxy for our purposes. For the experiments, we use Polyak's subgradient method, a classical algorithm known to exhibit rapid convergence near solutions for *sharp weakly-convex functions* [14].<sup>3</sup> Polyak's method is an iterative algorithm given by

$$x_{k+1} \leftarrow x_k - \left( \frac{f(x_k) - \inf f}{\|g_k\|^2} \right) g_k \quad \text{with } g_k \in \partial f(x_k). \quad (5)$$

3. It was shown in [9] that  $f_S$  satisfies these assumptions with high probability provided  $m \gtrsim (d_1 + d_2)$ .

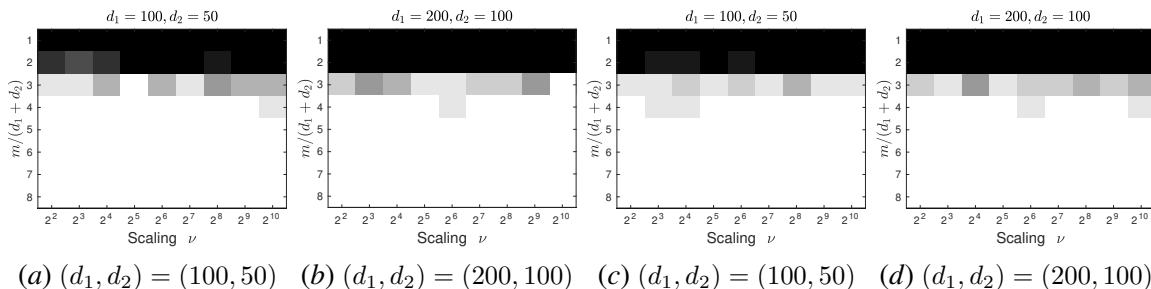


Figure 1: Empirical probability of recovery with random Gaussian initialization (two images on the left) and random initialization on a cube (two images on the right) for different dimensions. White denotes probability one and black denotes probability zero.

Notice that the step size requires us to know the minimum value, which in this case is exactly zero. Polyak’s algorithm was used in [8] as one of the procedures in the two-stage method for blind deconvolution.

In all the experiments the goal is to recover a pair of canonical vectors  $(e_1, e_1) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ . Observe that this instance is a good representative of the average performance of the method due to the rotational invariance of the measurements. We evaluate the frequency of successful recovery of (5) using two different random initialization strategies:

1. **(Uniform over a cube)** We set  $(w_0, x_0)$  to be an uniform vector on the cube  $[-\nu, \nu]^{d_1+d_2}$ .
2. **(Random Gaussian)** We set  $w_0$  and  $x_0$  to be distributed  $\mathcal{N}(0, \frac{\nu^2}{d_1} I_{d_1})$  and  $\mathcal{N}(0, \frac{\nu^2}{d_2} I_{d_2})$ , respectively. This ensures that with high probability, both  $\|w_0\|$  and  $\|x_0\|$  are close to  $\nu$ .

We generate phase transition plots for both initialization strategies by varying the value of  $\nu$  and  $C \triangleq m/(d_1 + d_2)$  between  $\{2^4, 2^5, \dots, 2^{10}\}$  and  $\{1, 2, \dots, 8\}$ , respectively. For each choice of parameters we generate ten random instances  $(w_0, x_0, A, B)$  and record in how many instances Polyak’s method achieves a relative error smaller than  $10^{-6}$ . The method stops whenever it reaches 100 000 iterations or the function value is less than  $10^{-10}$ . We repeat these experiments for two different pairs of dimensions,  $(d_1, d_2) \in \{(100, 50), (200, 100)\}$ , see the results in Figure 1.

A first immediate observation is that the random initialization, the dimension, and the scaling parameter  $\nu$  do not seem to be affecting the recovery frequency of the algorithm. The only parameter that controls the recovery frequency is  $C$ . This is intuitively consistent with Theorem 3, since this parameter determines the concentration of spurious critical points around a subspace. Nonetheless, the effect of this parameter seems to be stronger in practice. Indeed, the probability of recovery exhibits a sharp phase transition at  $C \sim 3$ .

**Reproducible research.** The experiments were run in a 2013 MacBook Pro with 2.4 GHz Intel Core i5 Processor and 8 GB of RAM. All the results and code concerning these experiments are publicly available in <https://github.com/mateodd25/BlindDeconvolutionLandscape>.

**Acknowledgments.** I thank Jose Bastidas, Damek Davis, Dmitriy Drusvyatskiy, Robert Kleinberg, and Mauricio Velasco for insightful and encouraging conversations. Finally, I would like to thank my advisor Damek Davis for research funding during the completion of this work.

## References

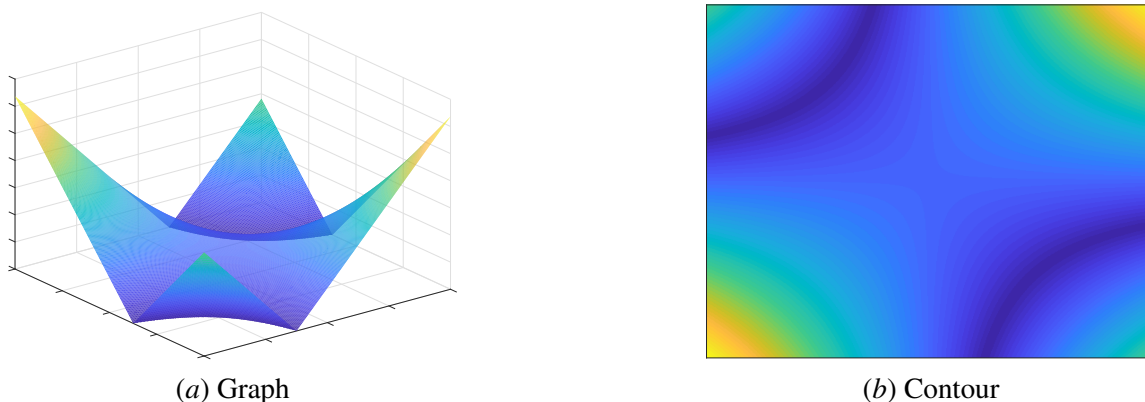
- [1] Ali Ahmed, Benjamin Recht, and Justin Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014.
- [2] Ali Ahmed, Alireza Aghasi, and Paul Hand. Blind deconvolutional phase retrieval via convex programming. *arXiv preprint arXiv:1806.08091*, 2018.
- [3] Yu Bai, Qijia Jiang, and Ju Sun. Subgradient descent learns orthogonal dictionaries. *arXiv preprint arXiv:1810.10702*, 2018.
- [4] J. M. Borwein and Q. J. Zhu. *Techniques of Variational Analysis*. Springer Verlag, New York, 2005. ISBN 0387242988.
- [5] Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pages 2757–2765, 2016.
- [6] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):Art. 11, 37, 2011. ISSN 0004-5411. doi: 10.1145/1970392.1970395. URL <http://dx.doi.org/10.1145/1970392.1970395>.
- [7] E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: theory and algorithms. *IEEE Trans. Inform. Theory*, 61(4):1985–2007, 2015. ISSN 0018-9448. URL <https://doi.org/10.1109/TIT.2015.2399924>.
- [8] Vasileios Charisopoulos, Yudong Chen, Damek Davis, Mateo Díaz, Lijun Ding, and Dmitriy Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *arXiv preprint arXiv:1904.10020*, 2019.
- [9] Vasileios Charisopoulos, Damek Davis, Mateo Díaz, and Dmitriy Drusvyatskiy. Composite optimization for robust blind deconvolution. *arXiv preprint arXiv:1901.01624*, 2019.
- [10] Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *arXiv preprint arXiv:1809.09573*, 2018.
- [11] Sunav Choudhary and Urbashi Mitra. Sparse blind deconvolution: What cannot be done. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 3002–3006. IEEE, 2014.
- [12] Diego Cifuentes. Burer-monteiro guarantees for general semidefinite programs. *arXiv preprint arXiv:1904.07147*, 2019.
- [13] Damek Davis, Dmitriy Drusvyatskiy, and Courtney Paquette. The nonsmooth landscape of phase retrieval. *arXiv preprint arXiv:1711.03247*, 2017.
- [14] Damek Davis, Dmitriy Drusvyatskiy, Kellie J. MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. *arXiv preprint arXiv:1803.02461*, 2018.

- [15] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, Jan 2019. ISSN 1615-3383. doi: 10.1007/s10208-018-09409-5. URL <https://doi.org/10.1007/s10208-018-09409-5>.
- [16] Lijun Ding and Yudong Chen. The leave-one-out approach for matrix completion: Primal and dual analysis. *arXiv preprint arXiv:1803.07554*, 2018.
- [17] Y. C. Eldar and S. Mendelson. Phase retrieval: stability and recovery guarantees. *Appl. Comput. Harmon. Anal.*, 36(3):473–494, 2014. ISSN 1063-5203. doi: 10.1016/j.acha.2013.08.003. URL <http://dx.doi.org/10.1016/j.acha.2013.08.003>.
- [18] Salar Fattahi and Somayeh Sojoudi. Exact guarantees on the absence of spurious local minima for non-negative robust principal component analysis. *arXiv preprint arXiv:1812.11466*, 2018.
- [19] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2973–2981. Curran Associates, Inc., 2016. URL <http://papers.nips.cc/paper/6048-matrix-completion-has-no-spurious-local-minimum.pdf>.
- [20] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.
- [21] Wen Huang and Paul Hand. Blind deconvolution by steepest descent algorithm on a quotient manifold. *arXiv preprint arXiv:1710.03309v2*, 2018.
- [22] Cedric Jozs, Yi Ouyang, Richard Zhang, Javad Lavaei, and Somayeh Sojoudi. A theory on the absence of spurious optimality. *arXiv preprint arXiv:1805.08204*, 2018.
- [23] Michael Kech and Felix Krahmer. Optimal injectivity conditions for bilinear inverse problems with applications to identifiability of deconvolution problems. *SIAM Journal on Applied Algebra and Geometry*, 1(1):20–37, 2017.
- [24] Han-Wen Kuo, Yenson Lau, Yuqian Zhang, and John Wright. Geometry and symmetry in short-and-sparse deconvolution. *arXiv preprint arXiv:1901.00256*, 2019.
- [25] Jeongyeol Kwon and Constantine Caramanis. Global convergence of em algorithm for mixtures of two component linear regression. *arXiv preprint arXiv:1810.05752*, 2018.
- [26] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257, 2016.
- [27] A. S. Lewis. Derivatives of spectral functions. *Math. Oper. Res.*, 21(3):576–588, 1996. ISSN 0364-765X. doi: 10.1287/moor.21.3.576. URL <http://dx.doi.org/10.1287/moor.21.3.576>.
- [28] A. S. Lewis. Nonsmooth analysis of eigenvalues. *Math. Program.*, 84(1, Ser. A):1–24, 1999. ISSN 0025-5610.

- [29] Adrian S Lewis and Hristo S Sendov. Nonsmooth analysis of singular values. part i: Theory. *Set-Valued Analysis*, 13(3):213–241, 2005.
- [30] Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and Rene Vidal. Nonconvex robust low-rank matrix recovery. *arXiv preprint arXiv:1809.09237*, 2018.
- [31] Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *arXiv preprint arXiv:1606.04933*, 2016.
- [32] Yanjun Li, Kiryung Lee, and Yoram Bresler. Identifiability in blind deconvolution with subspace or sparsity constraints. *IEEE Transactions on Information Theory*, 62(7):4266–4275, 2016.
- [33] Shuyang Ling, Ruitu Xu, and Afonso S Bandeira. On the landscape of synchronization networks: A perspective from nonconvex optimization. *arXiv preprint arXiv:1809.11083*, 2018.
- [34] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *arXiv preprint arXiv:1711.10467*, 2017.
- [35] B. S. Mordukhovich. *Variational analysis and generalized differentiation. I*, volume 330 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2006. ISBN 978-3-540-25437-9; 3-540-25437-4. Basic theory.
- [36] Dohyung Park, Anastasios Kyriillidis, Constantine Caramanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. *arXiv preprint arXiv:1609.03240*, 2016.
- [37] Inder K. Rana. *An introduction to measure and integration*, volume 45 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2002. ISBN 0-8218-2974-2. doi: 10.1090/gsm/045. URL <https://doi.org/10.1090/gsm/045>.
- [38] R. T. Rockafellar and R. J-B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften, Vol 317, Springer, Berlin, 1998.
- [39] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *To appear in Found. Comp. Math.*, *arXiv:1602.06664*, 2017.
- [40] Ju sun. Provable nonconvex methods/algorithms. URL <https://sunju.org/research/nonconvex/>. Accessed on 2019-05-13.
- [41] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Trans. Inform. Theory*, 62(11):6535–6579, 2016. ISSN 0018-9448. doi: 10.1109/TIT.2016.2598574. URL <https://doi-org.offcampus.lib.washington.edu/10.1109/TIT.2016.2598574>.
- [42] Irène Waldspurger. Phase retrieval with random gaussian sensing vectors by alternating projections. *IEEE Transactions on Information Theory*, 64(5):3301–3312, 2018.



- [43] Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust PCA via gradient descent. In *Neural Information Processing Systems Conference (NIPS)*, 2016.
- [44] Yuqian Zhang, Yenson Lau, Han-wen Kuo, Sky Cheung, Abhay Pasupathy, and John Wright. On the global geometry of sphere-constrained sparse blind deconvolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4894–4902, 2017.
- [45] Kai Zhong, Prateek Jain, and Inderjit S. Dhillon. Efficient matrix sensing using rank-1 Gaussian measurements. In *Algorithmic learning theory*, volume 9355 of *Lecture Notes in Comput. Sci.*, pages 3–18. Springer, Cham, 2015. doi: 10.1007/978-3-319-24486-0\_1. URL [https://doi.org/10.1007/978-3-319-24486-0\\_1](https://doi.org/10.1007/978-3-319-24486-0_1).
- [46] Yiqiao Zhong and Nicolas Boumal. Near-optimal bounds for phase synchronization. *SIAM Journal on Optimization*, 28(2):989–1016, 2018.


 Figure 2: Population objective  $d_1 = d_2 = 1$ .

## Appendix A. Preliminaries

We will follow standard notation. The symbols  $\mathbb{R}$  and  $\mathbb{R}_+$  denote the real line and the nonnegative reals, respectively. The set of extended reals  $\mathbb{R} \cup \{+\infty\}$  is written as  $\overline{\mathbb{R}}$ . We always endow  $\mathbb{R}^d$  with its standard inner product,  $\langle x, y \rangle = x^\top y$ , and its induced norm  $\|x\| = \sqrt{\langle x, x \rangle}$ . We also use  $\|x\|_1 = \sum |x_i|$  to denote the  $\ell_1$ -norm. For a set  $S \subseteq \mathbb{R}^d$ , we denote the distance from a point  $x$  to the set by  $\text{dist}(x, S) = \inf_{y \in S} \|x - y\|$ . For any pair of real-valued functions  $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ , we say that  $f \lesssim g$  if there exists a constant  $C$  such that  $f \leq Cg$ . Moreover, we write  $f \asymp g$  if both  $f \lesssim g$  and  $g \lesssim f$ .

The adjoint of a linear operator  $A : \mathbb{R}^d \rightarrow \mathbb{R}^n$  is indicated by  $A^\top : \mathbb{R}^n \rightarrow \mathbb{R}^d$ . Assuming  $d \leq n$ , the map  $\sigma : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}_+^d$  returns the vector of ordered singular values of a matrix with  $\sigma_1(A) \geq \sigma_2(A) \geq \dots \geq \sigma_d(A)$ . We will use the symbols  $\|A\|_{\text{op}} = \sigma_1(A)$  and  $\|A\|_F = \|\sigma(A)\|_2$  to indicate the operator and Frobenius norm, respectively. When not specified it is understood that  $\|A\| := \|A\|_{\text{op}}$ . We will use the symbol  $O(d)$  to denote the set of  $d \times d$  orthogonal matrix.

**Variational analysis.** Since we will handle nonsmooth functions, we need a definition of generalized derivatives. We refer the interested reader to some excellent references on the subject [4, 35, 38]. Let  $f : \mathbb{R}^d \rightarrow \overline{\mathbb{R}}$  be a lower semicontinuous proper function and  $\bar{x}$  be a point. The *Fréchet subdifferential*  $\widehat{\partial}f(x)$  is the set of all vectors  $\xi$  for which

$$f(x) \geq f(\bar{x}) + \langle \xi, x - \bar{x} \rangle + o(\|x - \bar{x}\|) \quad \text{as } x \rightarrow \bar{x}.$$

Intuitively,  $\xi \in \widehat{\partial}f(\bar{x})$  if the function  $x \mapsto f(\bar{x}) + \langle \xi, x - \bar{x} \rangle$  locally minorizes  $f$  up to first order information. Unfortunately, the set-valued mapping  $x \mapsto \widehat{\partial}f(x)$  lacks some desirable topological properties. For this reason it is useful to consider an extension. The *limiting subdifferential*  $\partial f(\bar{x})$  is the set of all  $\xi$  such that there are sequences  $(x_n)$  and  $(\xi_n)$  with  $\xi_n \in \widehat{\partial}f(x_n)$  satisfying  $(x_n, f(x_n), \xi_n) \rightarrow (\bar{x}, f(\bar{x}), \xi)$ . It is well-known that  $\partial f(\bar{x})$  reduces to the classical derivative when  $f$  is Fréchet differentiable and that for  $f$  convex,  $\partial f(\bar{x})$  is equal to the usual convex subdifferential

$$\xi \in \partial f(\bar{x}) \quad \iff \quad f(x) \geq f(\bar{x}) + \langle \xi, x - \bar{x} \rangle \quad \forall x.$$

We say that a point  $\bar{x}$  is *stationary* if  $0 \in \partial f(\bar{x})$ . The *graph* of  $\partial f$  is given by  $\text{gph } \partial f = \{(x, \zeta) \mid \zeta \in \partial f(x)\}$ .

For  $\rho > 0$ , we say that  $f$  is  $\rho$ -weakly convex if the regularized function  $f + \frac{\rho}{2} \|\cdot\|_2^2$  is convex. This encompasses a broad class of functions: Any function that can be decomposed as  $f = h \circ g$ , where  $h : \mathbb{R}^m \rightarrow \mathbb{R}$  is a Lipschitz convex function and  $g : \mathbb{R}^d \rightarrow \mathbb{R}^m$  is smooth map, is weakly convex. It is worth noting that for functions that can be decomposed in this fashion, the chain rule [38, Theorem 10.6] yields  $\partial f(x) = \nabla g(x)^\top \partial h(g(x))$  for all  $x$ .

**Singular value functions.** For a pair of dimensions  $d_1, d_2$  we will denote  $d = \min\{d_1, d_2\}$ . A function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is *symmetric* if  $f(\pi x) = f(x)$  for any permutation matrix  $\pi \in \{0, 1\}^{d \times d}$ . Additionally, a function  $f$  is *sign invariant* if  $f(sx) = f(x)$  for any diagonal matrix  $s \in \{-1, 0, 1\}^{d \times d}$  with diagonal entries in  $\{\pm 1\}$ . We say that  $f_\sigma : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}$  is a *singular value function* if it can be decomposed as  $f_\sigma = (f \circ \sigma)$  for a symmetric sign invariant function  $f$ . A simple and illuminating example is the Frobenius norm, since  $\|A\|_F = \|\sigma(A)\|_2$ . This type of function has been deeply studied in variational analysis [27–29].

A pair of matrices  $X$  and  $Y$  in  $\mathbb{R}^{d_1 \times d_2}$  have a *simultaneous ordered singular value decomposition* if there exist matrices  $U \in O(d_1)$  and  $V \in O(d_2)$  such that  $X = U \text{diag}(\sigma(X)) V^\top$  and  $Y = U \text{diag}(\sigma(Y)) V^\top$ . We will make great use of the following remarkable theorem.

**Theorem 4 (Theorem 7.1 in [29])** *The limiting subdifferential of a singular value function  $f_\sigma = f \circ \sigma$  at a matrix  $M \in \mathbb{R}^{d_1 \times d_2}$  is given by*

$$\partial f_\sigma(M) = \{U \text{diag}(\zeta) V^\top \mid \zeta \in \partial f(\sigma(M)) \text{ and } U \text{diag}(\sigma(M)) V^\top = M\}. \quad (6)$$

Hence  $M$  and any of its subgradients have simultaneous ordered singular value decomposition.

## Appendix B. Population objective

### B.1. Landscape analysis for a class of singular value functions

From now on we consider an arbitrary function  $g : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \rightarrow \mathbb{R}$  for which there exists a rank one matrix  $\bar{w}\bar{x}^\top$  and a singular value function  $f_\sigma$  satisfying

$$g(w, x) = f_\sigma(w x^\top - \bar{w}\bar{x}^\top) = f \circ \sigma \left( w x^\top - \bar{w}\bar{x}^\top \right).$$

This gives us two useful characterizations of  $g$  that we will use throughout. In the following section we will see a way of recasting  $f_P$  in this form.

A simple application of the chain rule yields

$$\partial g(w, x) = \left\{ \begin{bmatrix} Yx \\ Y^\top w \end{bmatrix} \mid Y \in \partial f_\sigma(X) \right\}. \quad (7)$$

Notice that we already have a description of  $\partial f_\sigma(X)$  given by Theorem 4, that is  $Y \in \partial f_\sigma(X)$  if and only if there exists matrices  $U \in O(d_1)$  and  $V \in O(d_2)$  satisfying

$$\sigma(Y) \in \partial f(\sigma(X)), \quad Y = U \text{diag}(\sigma(Y)) V^\top, \quad \text{and} \quad X = U \text{diag}(\sigma(X)) V^\top. \quad (8)$$

Equipped with these tools we derive the following result regarding the critical points of  $g$ . We defer a proof to Appendix D.

**Theorem 5** Suppose that  $(w, x)$  is a stationary point for  $g$ , i.e.  $Yx = 0, Y^\top w = 0$ . Then at least one of the following conditions hold:

1. **Small objective.**  $g(w, x) \leq g(\bar{w}, \bar{x})$ ,
2. **Zero.**  $(w, x) = 0$ ,
3. **One zero component.**  $\langle w, \bar{w} \rangle = \langle x, \bar{x} \rangle = 0$ ,  $wx^\top = 0$ , and (assuming that  $x$  is not zero)  $Yx = 0$  (similarly for  $w$ ).
4. **Small product norm.**  $\langle w, \bar{w} \rangle = \langle x, \bar{x} \rangle = 0$ ,  $\text{rank}(Y) = 1$ , and  $0 < \|wx^\top\| < \|\bar{w}\bar{x}^\top\|$ .

Moreover, if  $(\bar{w}, \bar{x})$  minimizes  $g$ , then  $(w, x)$  is a critical point if, and only if, it satisfies 1, 2, 3, or 4 for some  $Y \in \partial f_\sigma(X)$ .

## B.2. Landscape of the population objective

Our goal now is to apply Theorem 5 to describe the landscape of  $f_P$ . In order to do it we need to write  $f_P(w, x) = f \circ \sigma(X)$  with  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  a symmetric sign-invariant convex function. An easy way to do this is to define

$$f(s_1, \dots, s_d) = \mathbb{E} \left( \left| \sum_{i=1}^d a_i b_i s_i \right| \right).$$

To use Theorem 5, we need to study  $\partial f$ . The next lemma shows that the function is actually differentiable at every point but zero. We defer the proof of this result to Appendix D.1.

**Lemma 6** For any nonzero vector  $s \in \mathbb{R}_+^d \setminus \{0\}$ , the partial derivatives of  $f$  satisfy

$$\frac{\partial f}{\partial s_j}(s) = \sqrt{\frac{2}{\pi}} s_j \mathbb{E} \left[ a_j^2 \left( \sum_i (a_i s_i)^2 \right)^{-\frac{1}{2}} \right]. \quad (9)$$

This lemma gives us the final tool to derive the main theorem regarding the landscape of  $f_P$ .

**Proof** [Proof of Theorem 2] Notice that  $(\bar{w}, \bar{x})$  minimizes the population objective  $f_P$ , therefore Theorem 5 gives a complete description of the critical points. Let us examine each one of the conditions in this theorem.

The points in  $\{(w, x) \mid wx^\top = \bar{w}\bar{x}^\top\}$  and  $\{0\}$  are contained in the set of stationary points because they satisfy the first and second condition, respectively.

Now, let  $(w, x) \in \{\bar{w}\}^\perp \times \{\bar{x}\}^\perp$  such that  $wx^\top = 0$ . Thus, the matrix  $X$  is rank 1, and consequently (9) reveals that any  $Y \in \partial f_\sigma(X)$  satisfies  $\sigma(Y) = \nabla f(\sigma(X)) = (2/\pi, 0, \dots, 0)$ . Therefore, due to (8), we get  $Y = \frac{2}{\pi} \bar{w}\bar{x}^\top / \|\bar{w}\| \|\bar{x}\|$ . Without loss of generality, assume  $x$  is not zero. Then,  $\|Yx\| = \frac{2}{\pi \|\bar{x}\|} |\langle x, \bar{x} \rangle| = 0$  and, consequently,  $(w, x)$  is stationary.

On the other hand, let  $(w, x) \in \{\bar{w}\}^\perp \times \{\bar{x}\}^\perp$  such that  $0 < \|wx^\top\| < \|\bar{w}\bar{x}^\top\|$ . Therefore, the matrix  $X$  is rank 2 and so (9) gives that  $\sigma_2(Y) > 0$  for all  $Y \in \partial f_\sigma(X)$ . Hence,  $(w, x)$  is not a stationary point, giving the result.  $\blacksquare$

### Appendix C. Proof of Proposition 1

Recall that we defined the functions  $f : \mathbb{R}_+^d \rightarrow \mathbb{R}$  and  $f_\sigma : \mathbb{R}^{d \times n} \rightarrow \mathbb{R}$  to be such that  $f_P(w, x) = f_\sigma(X) = f(\sigma(X))$ . It is known that for constants  $c_1, c_2 \in \mathbb{R}_+$  we have that  $c_1 b_1 + c_2 b_2 \stackrel{(d)}{=} \sqrt{c_1^2 + c_2^2} b_1$ , where  $\stackrel{(d)}{=}$  denotes equality in distribution. Then

$$\begin{aligned}
 f(s_1, s_2, 0, \dots, 0) &= \mathbb{E}(|s_1 a_1 b_1 + s_2 a_2 b_2|) \\
 &= \mathbb{E}(\mathbb{E}(|s_1 a_1 b_1 + s_2 a_2 b_2| \mid a_1, a_2)) \\
 &= \mathbb{E}\left(\mathbb{E}(\sqrt{(s_1 a_1)^2 + (s_2 a_2)^2} |b_1| \mid a_1, a_2)\right) \\
 &= \sqrt{\frac{2}{\pi}} \mathbb{E} \sqrt{(s_1 a_1)^2 + (s_2 a_2)^2} \\
 &= \frac{\sqrt{\pi}}{\sqrt{2\pi^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \sqrt{(a_1 s_1)^2 + (a_2 s_2)^2} \exp\left(-\frac{a_1^2 + a_2^2}{2}\right) da_1 da_2 \\
 &= 4 \frac{\sqrt{\pi}}{\sqrt{2\pi^2}} \int_0^{\infty} \int_0^{\infty} \sqrt{(a_1 s_1)^2 + (a_2 s_2)^2} \exp\left(-\frac{a_1^2 + a_2^2}{2}\right) da_1 da_2 \\
 &= 2 \frac{\sqrt{2\pi}}{\pi^2} \int_0^{\infty} \int_0^{\pi/2} r^2 \sqrt{s_1^2 \cos^2 \theta + s_2^2 \sin^2 \theta} \exp\left(-\frac{r^2}{2}\right) d\theta dr \\
 &= \frac{2}{\pi} \int_0^{\pi/2} \sqrt{s_1^2 \cos^2 \theta + s_2^2 \sin^2 \theta} d\theta \\
 &= \frac{2s_1}{\pi} \int_0^{\pi/2} \sqrt{\cos^2 \theta + \frac{s_2^2}{s_1^2} \sin^2 \theta} d\theta \\
 &= \frac{2s_1}{\pi} \int_0^{\pi/2} \sqrt{1 - \left(1 - \frac{s_2^2}{s_1^2}\right) \sin^2 \theta} d\theta \\
 &= \frac{2s_1}{\pi} E\left(1 - \frac{s_2^2}{s_1^2}\right)
 \end{aligned}$$

where  $E(\cdot)$  is the complete elliptic integral of the second kind (with parameter  $m = k^2$ ). Thus altogether we obtain

$$f_\sigma(X) = \sigma_{\max}(X) \sum_{n=0}^{\infty} \left( \frac{(2n)!}{2^{2n} (n!)^2} \right)^2 \frac{(1 - \kappa^{-2}(X))^n}{1 - 2n}$$

where  $\kappa(X) = \sigma_{\max}(X)/\sigma_{\min}(X)$  is the condition number of  $X$ .

### Appendix D. Proof of Theorem 5

The proof of this result builds upon the next three lemmas. We will prove these lemmas and before we dive into the proof. Recall that  $U \in O(d_1)$  and  $V \in O(d_2)$  are any pair of matrices for which  $X = U\sigma(X)V = \sum_i \sigma_i(X) U_i V_i^\top$ .

**Lemma 7** *The following are true.*

1. **Anticorrelation.** The next equalities hold

$$\langle U_1, w \rangle \langle x, V_2 \rangle = \langle U_1, \bar{w} \rangle \langle \bar{x}, V_2 \rangle \quad \text{and} \quad \langle U_2, w \rangle \langle x, V_1 \rangle = \langle U_2, \bar{w} \rangle \langle \bar{x}, V_1 \rangle.$$

2. **Singular values.** The singular values of  $X$  satisfy

$$\sigma_1(X) = \langle U_1, w \rangle \langle x, V_1 \rangle - \langle U_1, \bar{w} \rangle \langle \bar{x}, V_1 \rangle \geq 0,$$

$$\sigma_2(X) = \langle U_2, w \rangle \langle x, V_2 \rangle - \langle U_2, \bar{w} \rangle \langle \bar{x}, V_2 \rangle \geq 0.$$

3. **Correlation.** Assume that  $\sigma_2(wx^\top - \bar{w}\bar{x}^\top) > 0$ , then  $\text{span}\{x, \bar{x}\} = \text{span}\{V_1, V_2\}$ ,  $\text{span}\{w, \bar{w}\} = \text{span}\{U_1, U_2\}$ , and consequently,

$$\begin{aligned} \langle w, \bar{w} \rangle &= \langle U_1, w \rangle \langle U_1, \bar{w} \rangle + \langle U_2, w \rangle \langle U_2, \bar{w} \rangle, \\ \langle x, \bar{x} \rangle &= \langle V_1, x \rangle \langle V_1, \bar{x} \rangle + \langle V_2, x \rangle \langle V_2, \bar{x} \rangle. \end{aligned}$$

**Proof** The first equality in item one follows by observing that  $U_1^\top X V_2 = 0$ , expanding the expression on the left-hand-side gives the result. The same argument starting from  $U_2^\top X V_1 = 0$  gives the other equality. The second item follows by definition.

To prove the last item note that

$$\langle U_i, w \rangle x - \langle U_i, \bar{w} \rangle \bar{x} = X^\top U_i = \sigma_i(X) V_i \quad \forall i \in \{1, 2\}.$$

Dividing through by  $\sigma_i(X)$  in the previous inequality shows that  $\text{span}\{x, \bar{x}\} = \text{span}\{V_1, V_2\}$ . Therefore, we can write  $x = \langle x, V_1 \rangle V_1 + \langle x, V_2 \rangle V_2$  and  $\bar{x} = \langle \bar{x}, V_1 \rangle V_1 + \langle \bar{x}, V_2 \rangle V_2$ . Hence,

$$\langle x, \bar{x} \rangle = \langle \langle x, V_1 \rangle V_1 + \langle x, V_2 \rangle V_2, \langle \bar{x}, V_1 \rangle V_1 + \langle \bar{x}, V_2 \rangle V_2 \rangle = \langle V_1, x \rangle \langle V_1, \bar{x} \rangle + \langle V_2, x \rangle \langle V_2, \bar{x} \rangle$$

An analogous argument shows the statement for  $w$  and  $\bar{w}$ . ■

**Lemma 8** *The following hold true.*

1. **Maximum correlation.**

$$\begin{aligned} \max\{|\sigma_1(Y) \langle v_1, x \rangle|, |\sigma_2(Y) \langle v_2, x \rangle|\} &\leq \|Yx\|, \\ \max\{|\sigma_1(Y) \langle u_1, w \rangle|, |\sigma_2(Y) \langle u_2, w \rangle|\} &\leq \|Y^\top w\|. \end{aligned} \tag{10}$$

2. **Objective gap.**

$$g(w, x) - g(\bar{w}, \bar{x}) \leq \sigma_1(Y) \sigma_1(X) + \sigma_2(Y) \sigma_2(X). \tag{11}$$

**Proof** Note that  $\|Yx\| \geq \langle z, Yx \rangle$  for all  $z \in \mathbb{S}^{d-1}$ , then the very first claim follows by testing with  $z \in \{\pm U_1, \pm U_2\}$ . An analogous argument gives the statement for  $w$ . Recall that  $f$  is convex, consequently  $f_\sigma$  is convex and the subgradient inequality gives

$$g(w, x) - g(\bar{w}, \bar{x}) = f_\sigma(X) - f_\sigma(0) \leq \langle Y, X \rangle = \sigma_1(Y) \sigma_1(X) + \sigma_2(Y) \sigma_1(X). \quad \blacksquare$$

**Lemma 9** Assume  $\bar{w} \in \mathbb{R}^{d_1}$  and  $\bar{x} \in \mathbb{R}^{d_2}$  are nonzero vectors. Set  $X = wx^\top + \bar{w}\bar{x}^\top$ , then  $X$  is a rank 1 matrix if, and only if,  $w = \lambda\bar{w}$  or  $x = \lambda\bar{x}$  for some  $\lambda \in \mathbb{R}$ .

**Proof** It is trivial to see that if the later holds then  $X$  is rank 1. Let us prove the other direction. Notice that if any of the vectors is zero we are done, so assume that none of them is. Recall that all the columns of  $X$  are span from one vector. Consider the case where  $x$  and  $\bar{x}$  have different support (i.e. set of nonzero entries), then it is immediate that  $w$  and  $\bar{w}$  have to be multiples of each other.

Now assume that this is not the case, without loss of generality assume that  $w \notin \text{span}\{\bar{w}\}$  and  $x$  and  $\bar{x}$  are nonzero and their first component is equal to one. Then the first column of  $X$  is equal to  $w + \bar{w}$ , furthermore the second column is equal to  $x_2w + \bar{x}_2\bar{w}$  has to be a multiple of the first one. By assumption  $w, \bar{w}$  are linearly independent therefore  $x_2 = \bar{x}_2$ . Using the same procedure for the rest of the entries we obtain  $x = \bar{x}$ .  $\blacksquare$

We are now in good shape to describe the landscape of the function  $g$ .

**Proof** [Proof of Theorem 5] To prove that at least one of the conditions hold we will show that if the first two don't hold then at least one of the other two have to hold. Assume that that the first two conditions are not satisfied, therefore  $g(w, x) > g(\bar{w}, \bar{x})$  and  $(w, x) \neq (0, 0)$ . Let us furnished some facts before we prove this is the case. Notice that from (11) we can derive

$$0 < \sigma_1(Y)\sigma_1(X) + \sigma_2(Y)\sigma_1(X) \leq 2\sigma_1(Y)\sigma_1(X),$$

thus  $\sigma_1(Y), \sigma_1(X) > 0$ . On the other hand, since  $(w, x)$  is critical inequalities (10) immediately give

$$\sigma_1(Y)\langle V_1, x \rangle = \sigma_2(Y)\langle V_2, x \rangle = 0, \quad \text{and} \quad \sigma_1(Y)\langle U_1, w \rangle = \sigma_2(Y)\langle U_2, w \rangle = 0. \quad (12)$$

So  $\langle V_1, x \rangle = 0$  and  $\langle U_1, w \rangle = 0$ , then the first claim in Lemma 7 gives. Additionally, this and the second claim in Lemma 7 imply that

$$\langle U_1, \bar{w} \rangle \langle \bar{x}, V_2 \rangle = \langle U_2, \bar{w} \rangle \langle \bar{x}, V_1 \rangle = 0, \quad \text{and} \quad -\langle U_1, \bar{w} \rangle \langle \bar{x}, V_1 \rangle = \sigma_1(X) > 0.$$

Combining these two gives  $\langle U_2, \bar{w} \rangle = \langle \bar{x}, V_2 \rangle = 0$ . Then by applying the second claim in Lemma 7 we get  $\sigma_2(X) = \langle U_2, w \rangle \langle x, V_2 \rangle$ . Using Equations (12) we conclude that  $\sigma_2(Y)\sigma_2(X) = 0$ .

Now we will show that at least one of the conditions holds, depending on the value of  $\sigma_2(X)$ , let us consider two cases:

**Case 1.** Assume  $\sigma_2(X) = 0$ . This means that  $X = wx^\top - \bar{w}\bar{x}^\top$  is a rank 1 matrix. By Lemma 9 we have that  $w = \lambda\bar{w}$  or  $x = \lambda\bar{x}$  for some  $\lambda \in \mathbb{R}$ . Note that if  $w = \lambda\bar{w}$  then  $U_1 = \pm\bar{w}/\|\bar{w}\|$ , then using Equation 12 we get that  $\lambda\|\bar{w}\| = 0$ . Which implies that  $\lambda = 0$ , and consequently  $w = 0$ . An analogous argument applies when  $x = \lambda\bar{x}$ . By assumption we have that  $Yx = 0$  and  $Y^\top w = 0$ . Additionally, since  $X = -\bar{w}\bar{x}^\top$  we get that that  $U_1 = \pm\bar{w}/\|\bar{w}\|$  and  $V_1 = \pm\bar{x}/\|\bar{x}\|$ . Recall that  $Y = U\text{diag}(\sigma(Y))V^\top$ , then using the fact that  $(w, x)$  is critical we conclude  $\langle w, \bar{w} \rangle = \langle x, \bar{x} \rangle = 0$ . Implying that property three holds.

**Case 2.** Assume  $\sigma_2(X) \neq 0$ . This immediately implies that  $\sigma_2(Y) = 0$ . By the third part of Lemma 7 we get that

$$\langle x, \bar{x} \rangle = \langle V_1, x \rangle \langle V_1, \bar{x} \rangle + \langle V_2, x \rangle \langle V_2, \bar{x} \rangle = 0$$

and analogously  $\langle w, \bar{w} \rangle = 0$ . Moreover, since  $w \perp \bar{w}$  and  $x \perp \bar{x}$  (and none of them are zero by assumption) we get that  $(w/\|w\|, x/\|x\|)$  and  $(\bar{w}/\|\bar{w}\|, \bar{x}/\|\bar{x}\|)$  are pairs of left and right singular

vectors, with associated singular values  $w^\top X x = \|w x^\top\|$  and  $\bar{w}^\top X \bar{x} = \|\bar{w} \bar{x}^\top\|$ , respectively. Assume that  $\|w x^\top\| \geq \|\bar{w} \bar{x}^\top\|$ , thus  $0 = w^\top Y x = \|w x^\top\| \sigma_1(Y) > 0$ , yielding a contradiction. Hence the condition four holds true.

Finally, we will prove the reverse statement. Assume that  $(\bar{w}, \bar{x})$  minimize  $g$ . In this case, the set of points that satisfies the first conditions is the collection of minimizers so they are critical. Clearly  $(w, x) = 0$  is always a stationary point, since  $\|Y^\top w\| = \|Y x\| = 0$ . Now let's construct a certificate  $Y \in \partial f_\sigma(X)$  that ensures criticality for the remaining cases.

Assume that  $(w, x)$  that  $w x^\top = 0$ , without loss of generality let's assume that  $w = 0$ . Further, assume that there exists  $Y \in \partial f_\sigma(w, x)$  such that  $Y x = 0$  and  $\langle x, \bar{x} \rangle = 0$ . It is immediate that  $(w, x)$  is a stationary point.

Assume that  $(w, x)$  is such that  $0 < \|w x^\top\| < \|\bar{w} \bar{x}^\top\|$ ,  $\langle w, \bar{w} \rangle = \langle x, \bar{x} \rangle = 0$  and there exists  $Y \in \partial f_\sigma(X)$  with  $\sigma_2(Y) = 0$ . By our argument above since  $\|w x^\top\| < \|\bar{w} \bar{x}^\top\|$ , any pair of admissible matrices  $U, V$  satisfy  $U_1 = \pm \bar{w} / \|\bar{w}\|$  and  $V_1 = \pm \bar{x} / \|\bar{x}\|$ . Therefore

$$Y x = (\sigma_1(Y) U_1 V_1^\top) x = \pm \frac{\sigma_1(Y)}{\|\bar{x}\|} \langle \bar{x}, x \rangle U_1 = 0,$$

analogously  $Y^\top w = 0$ . ■

### D.1. Proof of Lemma 6

It is well-known that if  $(a_1, a_2, \dots, a_d)$  is fixed (i.e. if we conditioned on it), then

$$\sum_{i=1}^d a_i b_i s_i \stackrel{(d)}{=} \left( \sum_{i=1}^d (a_i s_i)^2 \right)^{\frac{1}{2}} b$$

and  $b$  is a standard normal random variable independent of the rest of the data. Therefore

$$\begin{aligned} f(s_1, \dots, s_d) &= \mathbb{E} \left( \left| \sum_{i=1}^d a_i b_i s_i \right| \right) = \mathbb{E} \left( \mathbb{E} \left( \left| \sum_{i=1}^d a_i b_i s_i \right| \middle| a_1, \dots, a_d \right) \right) \\ &= \mathbb{E} \left( \left( \sum_{i=1}^d (a_i s_i)^2 \right)^{\frac{1}{2}} \mathbb{E}(b \mid a_1, \dots, a_d) \right) = \sqrt{\frac{2}{\pi}} \mathbb{E} \left( \sum_{i=1}^d (a_i s_i)^2 \right)^{\frac{1}{2}}. \end{aligned}$$

Now, we need a technical tool in order to procede.

**Theorem 10 (Leibniz Integral Rule, Theorem 5.4.12 in [37])** *Let  $U$  be an open subset of  $\mathbb{R}^d$  and  $\Omega$  be a measure space. Suppose that the function  $h : U \times \Omega \rightarrow \mathbb{R}$  satisfies the following:*

1. *For all  $x \in U$ , the function  $h(x, \cdot)$  is Lebesgue integrable.*
2. *For almost all  $w \in \Omega$ , if we define  $h^\omega(\cdot) = f(\cdot, w)$  the partial derivatives  $\frac{\partial h^\omega}{\partial x_i}(x)$  exists for all  $x \in U$ .*
3. *There is an integrable function  $\Phi : \Omega \rightarrow \mathbb{R}$  such that  $|\frac{\partial h^\omega}{\partial x_i}(x)| \leq \Phi(\omega)$  for all  $x \in U$  and almost every  $w \in \Omega$ .*



Then, we have that for all  $x \in U$

$$\frac{\partial}{\partial x_i} \int_{\Omega} h(x, \omega) d\omega = \int_{\Omega} \frac{\partial h^\omega}{\partial x_i}(x) d\omega.$$

This theorem tell us that we can swap partial derivatives and integrals provided that the function satisfies all the conditions above. Consider  $\Omega$  to be the set  $\mathbb{R}^d$  endowed with the Borel  $\sigma$ -algebra and the multivariate Gaussian measure. Define  $h : \mathbb{R}^d \times \Omega \rightarrow \mathbb{R}$  to be given by

$$(s, a) \mapsto \left( \sum_{i=1}^d (a_i s_i)^2 \right)^{\frac{1}{2}}.$$

Take  $s \in \mathbb{R}^d \setminus \{0\}$  to be an arbitrary element, set  $S = \{u \in \mathbb{R}^d \mid \text{supp}(s) \subseteq \text{supp}(u)\}$ , and define  $U = B_\epsilon(s)$  with  $\epsilon$  small enough such that  $U \subseteq S$  and  $\inf_{u \in U} \min_{i \in \text{supp}(s)} |u_i| > 0$ . Then it is easy to see that the first two conditions hold, in particular the second condition hold for all  $a \neq 0$ . Further, for any  $x \in U$

$$\begin{aligned} \left| \frac{\partial h^a}{\partial s_j}(x) \right| &= \left| \frac{a_j^2 x_j}{\left( \sum_i^d (a_i x_i)^2 \right)^{\frac{1}{2}}} \right| \leq \frac{\sup_{u \in U} \|u\|_\infty}{\inf_{u \in U} \min_{i \in \text{supp}(s)} |u_i|} \frac{\sum_{i \in \text{supp}(s)} a_j^2}{\left( \sum_{i \in \text{supp}(s)} a_i^2 \right)^{\frac{1}{2}}} \\ &\leq \frac{\sup_{u \in U} \|u\|_\infty}{\inf_{u \in U} \min_i |u_i|} \left( \sum_{i \in \text{supp}(s)} a_i^2 \right)^{\frac{1}{2}}, \end{aligned}$$

where the last function is integrable with respect to the Gaussian measure. Thus, Theorem 10 ensures that the function  $f$  is differentiable at every nonzero point. Consequently, for all  $s \in \mathbb{R}^d \setminus \{0\}$

$$\frac{\partial f}{\partial s_j}(s) = \sqrt{\frac{2}{\pi}} s_j \mathbb{E} \frac{a_j^2}{\left( \sum_i^d (a_i s_i)^2 \right)^{\frac{1}{2}}}.$$

## Appendix E. Approximate critical points of a spectral function family

In Section B we characterize the points for which  $0 \in \partial f_P(w, x)$ . In order to derive similar results for  $f_S$  we will need to understand  $\epsilon$ -critical points of  $f_P$ , i.e. points  $(w, x)$  for which  $\text{dist}(0, \partial f(w, x)) \leq \epsilon$ . Just as before we adopt a more general viewpoint and consider spectral functions of the form  $g(w, x) = f \circ \sigma(wx^\top - \bar{w}\bar{x}^\top)$ .

The main result in this section is Theorem 13. Given the fact that we don't have second order information in the form of a Hessian, we need to appeal to a different kind of growth condition. Turns out that the natural condition for this problem is

$$g(w, x) - g(\bar{w}, \bar{x}) \geq \kappa \left\| wx^\top - \bar{w}\bar{x}^\top \right\|_F \quad \forall (w, x) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}, \quad (13)$$

for some  $\kappa > 0$ . Intuitively this means that the function grows sharply away from minimizers.

Before we dive into the main theorem, let us provide some technical lemmas.

**Lemma 11** *Suppose there exists a constant  $\kappa > 0$  such that (13) holds. Then, for any point  $(w, x)$  such that  $wx^\top \neq \bar{w}\bar{x}^\top$  we have  $\sigma_1(Y) + \sigma_2(Y) \geq \kappa$ .*

**Proof** By definition  $\sigma_2(X) \leq \sigma_1(X) \leq \|wx^\top - \bar{w}\bar{x}^\top\|_F$ . Then, applying (11) gives

$$\begin{aligned} \kappa \left\| wx^\top - \bar{w}\bar{x}^\top \right\|_F &\leq g(w, x) - g(\bar{w}, \bar{x}) \leq \sigma_1(Y)\sigma_1(X) + \sigma_2(Y)\sigma_2(X) \\ &\leq (\sigma_1(Y) + \sigma_2(Y)) \left\| wx^\top - \bar{w}\bar{x}^\top \right\|_F. \end{aligned}$$

■

**Lemma 12** *Suppose there exists a constant  $\kappa > 0$  such that (13) holds. Then any pair  $(w, x) \in \mathbb{R}^{d_1+d_2} \setminus \{0\}$  satisfies*

$$\frac{1}{\min\{\|w\|, \|x\|\}} \left( \kappa \left\| wx^\top - \bar{w}\bar{x}^\top \right\| - (\sigma_1(Y) + \sigma_2(Y)) \left\| \bar{w}\bar{x}^\top \right\| \right) \leq \text{dist}(0; \partial g(w, x)).$$

**Proof** Notice that the result holds trivially for any pair such that  $wx^\top = \bar{w}\bar{x}^\top$ . Let's assume that this is not the case. Recall that  $\partial g(w, x) = \partial f_\sigma(X)x \times (\partial f_\sigma(X))^\top w$ . Pick  $Y \in \partial g(w, x)$  such that  $\text{dist}(0, \partial g(w, x)) = \sqrt{\|Yx\|^2 + \|Y^\top w\|^2}$ . Using the convexity of  $f_\sigma$  we get

$$\begin{aligned} \kappa \|wx^\top - \bar{w}\bar{x}^\top\|_F &\leq g(w, x) - g(\bar{w}, \bar{x}) = f_\sigma(X) - f_\sigma(0) \\ &\leq \langle Y, wx^\top - \bar{w}\bar{x}^\top \rangle \\ &\leq \|x\| \|Y^\top w\| + |w^\top Yx| \leq \|x\| \text{dist}(0, \partial g(w, x)) + |w^\top Yx|, \end{aligned}$$

where the last inequality follows by Cauchy-Schwartz. Applying the same argument using  $w^\top Yx \leq \|w\| \|Yx\|$  gives

$$g(w, x) - g(\bar{w}, \bar{x}) \leq \min\{\|w\|, \|x\|\} \text{dist}(0, \partial g(w, x)) + |w^\top Yx|.$$

Now, let's bound the second term on the right-hand-side. Note that

$$|w^\top Yx| = |\langle Y, wx^\top \rangle| \leq \|Y\| \|wx^\top\| \leq (\sigma_1(Y) + \sigma_2(Y)) \left\| \bar{w}\bar{x}^\top \right\|.$$

The result follows immediately. ■

We can now prove the main result of this section, a detailed location description of  $\varepsilon$ -critical points. This can be thought of as a quantitative version of Corollary 5. Its proof is however more involved due to the inexactness of the assumptions.

**Theorem 13** *Assume that  $\|\bar{w}\| = \|\bar{x}\|$  and that there exists a constant  $\kappa > 0$  such that (13) holds. Further assume that  $\sigma_1(Y)$  is bounded by some numerical constant.<sup>4</sup> Let  $\zeta = (Yx, Y^\top w) \in \partial g(w, x)$ , and set  $\varepsilon = \|\zeta\|$ . Then if  $wx^\top = 0$  we have that*

$$\max\{\|Yx\|, \|Y^\top w\|\} \leq \varepsilon, \quad \text{and} \quad \begin{cases} |\langle w, \bar{w} \rangle| & \lesssim \varepsilon \|\bar{w}\| \\ |\langle x, \bar{x} \rangle| & \lesssim \varepsilon \|\bar{x}\| \end{cases}.$$

4. This is implied for example when  $f$  is Lipschitz.

On the other hand, if  $wx^\top \neq 0$  and  $\|(w, x)\| \leq \nu \|(\bar{w}, \bar{x})\|$  for some fixed  $\nu > 1$ . There exists a constant<sup>5</sup>  $\gamma > 0$  such that if  $\varepsilon \leq \gamma \max\{\|w\|, \|x\|\}$  then  $\|wx^\top\| \lesssim \|\bar{w}\bar{x}^\top\|$  and at least one of the following holds

1.

$$\max\{\|w\|, \|x\|\} \|wx^\top - \bar{w}\bar{x}^\top\| \lesssim \varepsilon \|\bar{w}\bar{x}^\top\|$$

2.

$$\min\{\|w\|, \|x\|\} \lesssim \varepsilon \quad \text{and} \quad \begin{cases} |\langle w, \bar{w} \rangle| \lesssim \nu^2 \varepsilon \|\bar{w}\| \\ |\langle x, \bar{x} \rangle| \lesssim \nu^2 \varepsilon \|\bar{x}\| \end{cases}.$$

3.

$$\sigma_2(Y) \lesssim \frac{\varepsilon}{\max\{\|w\|, \|x\|\}} \quad \text{and} \quad \begin{cases} |\langle w, \bar{w} \rangle| \lesssim \nu^2 \varepsilon \|\bar{w}\| \\ |\langle x, \bar{x} \rangle| \lesssim \nu^2 \varepsilon \|\bar{x}\| \end{cases}.$$

**Proof** First assume that  $wx^\top = 0$ , then it is clear that  $\max\{\|Yx\|, \|Y^\top w\|\} = \|\zeta\| \leq \varepsilon$ . Without loss of generality assume that  $x = 0$ , let  $Y = U\sigma(Y)V^\top$  be the singular value decomposition. Since  $X = -\bar{w}\bar{x}^\top$  then  $U_1 = \pm\bar{w}/\|\bar{w}\|$  and  $V_1 = \pm\bar{x}/\|\bar{x}\|$  and so

$$\varepsilon \geq \|Y^\top w\| = \left\| \frac{\sigma_1(Y)}{\|\bar{w}\|} \langle \bar{w}, w \rangle V_1 + z \right\| \geq \frac{\sigma_1(Y)}{\|\bar{w}\|} |\langle \bar{w}, w \rangle| \geq \frac{\kappa}{2\|\bar{w}\|} \langle \bar{w}, w \rangle \quad (14)$$

where  $z$  is orthogonal to  $V_1$  and the second inequality follows by Lemma 11. This proves the first statement in the theorem.

We know move to the ‘‘On the other hand’’ statement, assume  $wx^\top \neq 0$  and  $\|(w, x)\| \leq \nu \|(\bar{w}, \bar{x})\|$ . Notice that the result holds immediately if  $(w, x) \in \{(\alpha\bar{w}, \bar{x}/\alpha) \mid \alpha \in \mathbb{R}\}$ . Further, due to Theorem 5 it also holds when  $\varepsilon = 0$ . Let us assume that none of these two conditions are satisfied.

We will start by showing that  $\|wx^\top\| \lesssim \|\bar{w}\bar{x}^\top\|$ . Set

$$\delta = \frac{\sqrt{2}}{\kappa} (\sigma_1(Y) + \sigma_2(Y)) + 1. \quad (15)$$

We showed in Lemma 11 that  $(\sigma_1(Y) + \sigma_2(Y)) \geq \kappa$  and thus  $\delta > 1$ .

**Claim 1** The inequality  $\|wx^\top\| \leq \delta \|\bar{w}\bar{x}^\top\|$  holds true.

**Proof** Seeking contradiction assume that this is not the case. By the previous lemma

$$\frac{\sqrt{2}}{\|wx^\top\|} \kappa \|wx^\top - \bar{w}\bar{x}^\top\| - \frac{\varepsilon}{\max\{\|w\|, \|x\|\}} \leq (\sigma_1(Y) + \sigma_2(Y)) \frac{\|\bar{w}\bar{x}^\top\|}{\|wx^\top\|}. \quad (16)$$

Notice that

$$\frac{\sqrt{2}}{\|wx^\top\|} \kappa \|wx^\top - \bar{w}\bar{x}^\top\| = \sqrt{2} \kappa \left\| \frac{wx^\top}{\|wx^\top\|} - \frac{\bar{w}\bar{x}^\top}{\|wx^\top\|} \right\| \geq \sqrt{2} \kappa \left| 1 - \frac{1}{\delta} \right|.$$

---

5. Independent of  $\nu$ .

If we set  $\gamma < \frac{\sqrt{2}\kappa}{2} |1 - \frac{1}{\delta}|$  then we ensure  $\frac{\varepsilon}{\max\{\|w\|, \|x\|\}} < \frac{\sqrt{2}\kappa}{2} |1 - \frac{1}{\delta}|$ , thus

$$\begin{aligned} \frac{\sqrt{2}\kappa |1 - \frac{1}{\delta}|}{2(\sigma_1(Y) + \sigma_2(Y))} &\leq \frac{1}{(\sigma_1(Y) + \sigma_2(Y))} \left( \frac{\sqrt{2}}{\|wx^\top\|} \kappa \|wx^\top - \bar{w}\bar{x}^\top\| - \frac{\varepsilon}{\max\{\|w\|, \|x\|\}} \right) \\ &\leq \frac{\|\bar{w}\bar{x}^\top\|}{\|wx^\top\|} < \frac{1}{\delta}. \end{aligned}$$

Rearranging we get

$$|\delta - 1| < \frac{\sqrt{2}}{\kappa} (\sigma_1(Y) + \sigma_2(Y)),$$

leading to a contradiction. ■

We now move on to proving that at least one of the three conditions has to hold. To this end, define

$$\rho_1 = \frac{\max\{\|w\|, \|x\|\}}{\sqrt{2}} \quad \text{and} \quad \rho_2 = \frac{1}{\kappa} \max\{2\sqrt{2}(1 + \delta), 4\sigma_1(Y)\} \frac{\|\bar{w}\bar{x}^\top\|}{\max\{\|w\|, \|x\|\}}.$$

Observe that if assume that if  $\varepsilon\rho_2 \geq \|wx^\top - \bar{w}\bar{x}^\top\|$  then the result holds immediately. Assume from now on that  $\varepsilon\rho_2 < \|wx^\top - \bar{w}\bar{x}^\top\|$ . Our road map is as follows, we will start by assuming  $\min\{\|w\|, \|x\|\} \leq 2\varepsilon/\kappa$  and we will show that this implies the second condition in item two. Then we will move to assume that  $\min\{\|w\|, \|x\|\} > 2\varepsilon/\kappa$  and show that item three has to hold.

Before we continue let us list some important facts. By Lemma 8

$$\max\{\sigma_1(Y)|\langle V_1, x \rangle|, \sigma_2(Y)|\langle V_2, x \rangle|, \sigma_1(Y)|\langle U_1, w \rangle|, \sigma_2(Y)|\langle U_2, w \rangle|\} \leq \varepsilon \quad (17)$$

which together with  $\sigma_1(Y) > \kappa/2$  implies that

$$\max\{|\langle U_1, w \rangle|, |\langle V_1, x \rangle|\} \leq \frac{\varepsilon}{\sigma_1(Y)} \leq \frac{2\varepsilon}{\kappa}. \quad (18)$$

Notice that this implies by Lemma 7

$$|\langle U_1, \bar{w} \rangle \langle \bar{x}, V_2 \rangle| = |\langle U_1, w \rangle \langle x, V_2 \rangle| \leq \frac{2\|x\|\varepsilon}{\kappa} \quad \text{and} \quad |\langle U_2, \bar{w} \rangle \langle \bar{x}, V_1 \rangle| \leq \frac{2\|w\|\varepsilon}{\kappa}. \quad (19)$$

Observe that

$$\max\{\|w\|, \|x\|\} \leq \|(w, x)\| \leq \nu \|(\bar{w}, \bar{x})\| = \sqrt{2}\nu \min\{\|\bar{w}\|, \|\bar{x}\|\}. \quad (20)$$

We can now continue with the proof. We will now assume that  $\min\{\|w\|, \|x\|\} \leq 4\delta\varepsilon/\kappa$  and prove that item two holds.

**Claim 2** *Assume that  $\min\{\|w\|, \|x\|\} \leq 4\delta\varepsilon/\kappa$ . Then*

$$|\langle w, \bar{w} \rangle| \lesssim \nu^2 \varepsilon \|\bar{w}\| \quad \text{and} \quad |\langle x, \bar{x} \rangle| \lesssim \nu^2 \varepsilon \|\bar{x}\|.$$

**Proof** Notice

$$\left| \left\langle w, \frac{\bar{w}}{\|\bar{w}\|} \right\rangle \right| \leq \left| \left\langle w, \frac{\bar{w}}{\|\bar{w}\|} - U_1 \right\rangle \right| + |\langle w, U_1 \rangle| \leq \|w\| \left\| \frac{\bar{w}}{\|\bar{w}\|} - U_1 \right\| + \frac{2\varepsilon}{\kappa}$$

where the last inequality follows by Cauchy-Schwartz and (18). A similar argument gives the same bound with  $\|\bar{w}/\|\bar{w}\| + U_1\|$  instead. Now we need to make use of the Davis-Kahan Theorem.

**Theorem 14** *Let  $A, \hat{A} \in \mathbb{R}^{d_1 \times d_2}$  with  $\text{rank}(A) = 1$ . Let  $A = U\sigma(A)V^\top$  and  $\hat{A} = \hat{U}\sigma(\hat{A})\hat{V}^\top$  be their singular value decompositions. Then the*

$$\sin \theta(V_1, \hat{V}_1) \leq \frac{2(2\sigma_1(A) + \|A - \hat{A}\|_{\text{op}})}{\|A\|^2} \|A - \hat{A}\|,$$

the same bound holds for  $U_1, \hat{U}_1$ .

By letting  $A = -\bar{w}\bar{x}^\top$  and  $\hat{A} = wx^\top - \bar{w}\bar{x}^\top$  in the previous theorem we get

$$\begin{aligned} \min \left\{ \left\| \frac{\bar{w}}{\|\bar{w}\|} + U_1 \right\|, \left\| \frac{\bar{w}}{\|\bar{w}\|} - U_1 \right\| \right\} &\leq \sqrt{2} \sin(\theta(\bar{w}/\|\bar{w}\|, U_1)) \\ &\leq 2\sqrt{2} \frac{(2\|\bar{w}\bar{x}^\top\| + \|wx^\top\|)}{\|\bar{w}\bar{x}^\top\|^2} \|wx^\top\| \\ &\leq 2\sqrt{2}(2 + \delta) \frac{\|wx^\top\|}{\|\bar{w}\bar{x}^\top\|} \\ &\leq 2\sqrt{2}(2 + \delta)\nu \frac{\varepsilon}{\|\bar{w}\|} \end{aligned}$$

where the last inequality follows since  $\|wx^\top\| \leq \varepsilon \max\{\|w\|, \|x\|\}$  and (20). Hence from the previous inequalities we derive

$$\left| \left\langle w, \frac{\bar{w}}{\|\bar{w}\|} \right\rangle \right| \leq \|w\| \left\| \frac{\bar{w}}{\|\bar{w}\|} - U_1 \right\| + \frac{2\varepsilon}{\kappa} \leq 2\sqrt{2}(2 + \delta)\nu \frac{\|w\|}{\|\bar{w}\|} \varepsilon + \frac{2\varepsilon}{\kappa} = \left( 2\sqrt{2}(2 + \delta)\nu^2 + \frac{2}{\kappa} \right) \varepsilon.$$

A completely analogous result holds for  $|\langle x, \bar{x} \rangle|$ . ■

Suppose now that  $\min\{\|w\|, \|x\|\} > 4\delta\varepsilon/\kappa$ . In the remainder of the proof we will show that in this case, item three has to hold.

**Claim 3** *The rank of  $X = wx^\top - \bar{w}\bar{x}^\top$  is two.*

**Proof** Assume  $w = \lambda\bar{w}$ , then  $U_1 = \pm w/\|\bar{w}\|$  and using the same computation as in Equation (14) we get  $\lambda\|\bar{w}\| \leq 2\varepsilon/\kappa \leq 4\delta\varepsilon/\kappa$  which implies  $\min\{\|w\|, \|x\|\} \leq 4\delta\varepsilon/\kappa$ , yielding a contradiction. An analogous argument holds for  $x = \lambda\bar{x}$ . Thus, Lemma 9 implies that  $\sigma_2(wx^\top - \bar{w}\bar{x}^\top) > 0$ . ■

**Claim 4**  $\sigma_2(Y) < \frac{\varepsilon}{\rho_1}$ .

**Proof** Without loss of generality suppose  $\|w\| = \max\{\|w\|, \|x\|\}$ . Assume seeking contradiction that this isn't true, thus  $\sigma_2(Y) \geq \varepsilon/\rho_1$  then Inequality (17) gives  $|\langle U_2, w \rangle| \leq \rho_1$ . Furthermore, notice that due to Lemma 7 we have that  $\|w\|^2 = \langle U_1, w \rangle^2 + \langle U_2, w \rangle^2$  and consequently  $|\langle U_1, w \rangle| \geq \sqrt{\|w\|^2 - \rho_1^2}$ . Again, due to (17)

$$\sigma_1(Y) \leq \frac{\varepsilon}{|\langle U_1, w \rangle|} \leq \frac{\varepsilon}{\sqrt{\|w\|^2 - \rho_1^2}}.$$

In turn this implies

$$\begin{aligned} \kappa\varepsilon\rho_2 < \kappa \left\| wx^\top - \bar{w}\bar{x}^\top \right\| &\leq g(w, x) - g(\bar{w}, \bar{x}) \leq \sigma_1(Y)\sigma_1(X) + \sigma_2(Y)\sigma_2(X) \\ &\leq 2\sigma_1(Y)\sigma_1(X) \\ &\leq 2\frac{\varepsilon}{\sqrt{\|w\|^2 - \rho_1^2}} |\langle U_1, w \rangle \langle x, V_1 \rangle - \langle U_1, \bar{w} \rangle \langle \bar{x}, V_1 \rangle| \\ &\leq 2\frac{\varepsilon}{\sqrt{\|w\|^2 - \rho_1^2}} \left( \|wx^\top\| + \|\bar{w}\bar{x}^\top\| \right) \\ &\leq \frac{2\sqrt{2}\varepsilon}{\|w\|} (1 + \delta) \|\bar{w}\bar{x}^\top\| \end{aligned}$$

Rearranging we get

$$\rho < \frac{2\sqrt{2}(1 + \delta)}{\kappa} \frac{\|\bar{w}\bar{x}^\top\|}{\max\{\|w\|, \|x\|\}},$$

yielding a contradiction. ■

We now need to prove an additional claim.

**Claim 5**  $|\langle U_2, \bar{w} \rangle| \leq |\langle U_1, \bar{w} \rangle|$  and  $|\langle V_2, \bar{x} \rangle| \leq |\langle V_1, \bar{x} \rangle|$

**Proof** Seeking contradiction we assume the possible contrary cases.

*Case 1.* Assume  $|\langle U_2, \bar{w} \rangle| > |\langle U_1, \bar{w} \rangle|$  and  $|\langle V_2, \bar{x} \rangle| > |\langle V_1, \bar{x} \rangle|$ , then (18) and (19) imply

$$\max\{|\langle U_1, w \rangle \langle V_1, x \rangle|, |\langle U_1, \bar{w} \rangle \langle V_1, \bar{x} \rangle|\} \leq \frac{2 \min\{\|w\|, \|x\|\} \varepsilon}{\kappa}.$$

From which we derive

$$\kappa\varepsilon\rho_2 < g(w, x) - g(\bar{w}, \bar{x}) \leq 2\sigma_1(Y)\sigma_1(X) \leq 4\sigma_1(Y)\delta \frac{\|\bar{w}\bar{x}^\top\|}{\max\{\|w\|, \|x\|\}} \varepsilon.$$

contradicting the definition of  $\rho_2$ .

*Case 2.* Assume that  $|\langle U_2, \bar{w} \rangle| \leq |\langle U_1, \bar{w} \rangle|$  and  $|\langle V_2, \bar{x} \rangle| > |\langle V_1, \bar{x} \rangle|$ . Notice that  $\|\bar{w}\|^2 = \langle U_1, \bar{w} \rangle^2 + \langle U_2, \bar{w} \rangle^2$ , hence  $|\langle U_1, \bar{w} \rangle| \geq \|\bar{w}\|/\sqrt{2}$  and similarly  $|\langle V_2, \bar{x} \rangle| > \|\bar{x}\|/\sqrt{2}$ . Thus,

$$\frac{\|\bar{w}\|}{\sqrt{2}} \leq |\langle U_1, \bar{w} \rangle| \leq \frac{2\|x\|\varepsilon}{\kappa|\langle \bar{x}, V_2 \rangle|} < \frac{2\sqrt{2}\|x\|\varepsilon}{\kappa\|\bar{x}\|}.$$

This implies that

$$\min\{\|w\|, \|x\|\} \leq \|w\| \leq \delta \frac{\|\bar{w}\bar{x}^\top\|}{\|x\|} < \frac{4\delta\varepsilon}{\kappa},$$

yielding a contradiction. ■

Without loss of generality let us assume  $\|w\| \leq \|x\|$ .

**Claim 6**

$$|\langle w, \bar{w} \rangle| \lesssim \varepsilon \|\bar{w}\| \quad \text{and} \quad |\langle x, \bar{x} \rangle| \lesssim \nu^2 \varepsilon \|\bar{x}\|.$$

**Proof** By the previous claim and the fact that  $\|\bar{w}\|^2 = \langle U_1, \bar{w} \rangle^2 + \langle U_2, \bar{w} \rangle^2$  we get that  $|\langle U_1, \bar{w} \rangle| \geq \|\bar{w}\|/\sqrt{2}$ , combining this with (19) gives

$$|\langle \bar{x}, V_2 \rangle| \leq \frac{2\sqrt{2}\|x\|\varepsilon}{\kappa\|\bar{w}\|} \leq \frac{4\delta}{\kappa}\nu\varepsilon$$

Then by Lemma 7

$$\begin{aligned} |\langle x, \bar{x} \rangle| &= |\langle V_1, x \rangle \langle V_1, \bar{x} \rangle + \langle V_2, x \rangle \langle V_2, \bar{x} \rangle| \leq |\langle V_1, x \rangle \langle V_1, \bar{x} \rangle| + |\langle V_2, x \rangle \langle V_2, \bar{x} \rangle| \\ &\leq \frac{2\varepsilon}{\kappa} \|\bar{x}\| + \|x\| |\langle V_2, \bar{x} \rangle| \\ &\leq \left( \frac{2}{\kappa} + \frac{4\delta}{\kappa} \nu^2 \right) \varepsilon \|\bar{x}\| \leq \left( \frac{2}{\kappa} + \frac{4\delta}{\kappa} \nu^2 \right) \varepsilon \|\bar{x}\|. \end{aligned}$$

where we used (20). Notice that the same analysis gives

$$|\langle w, \bar{w} \rangle| \leq \left( \frac{2}{\kappa} + \frac{2\sqrt{2}\delta}{\kappa} \frac{\|w\|}{\|x\|} \right) \varepsilon \|\bar{w}\| \leq \left( \frac{2}{\kappa} + \frac{2\sqrt{2}\delta}{\kappa} \right) \varepsilon \|\bar{w}\|.$$

■

This last claim finishes the proof of the theorem. ■

**Appendix F. Proof of Theorem 3**

In order to prove the theorem we will apply three steps: we will show that the graphs of  $\partial f_S$  and  $\partial f_P$  are close, then use Theorem 13 to study the  $\varepsilon$ -critical points of  $f_P$  and finally conclude about the landscape of  $f_S$  by combining the previous two steps. The following two propositions handle the first part.

**Proposition 15** *Fix two functions  $f, g : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \rightarrow \mathbb{R}$  such that  $g$  is  $\rho$ -weakly convex. Suppose that there exists a point  $(\bar{w}, \bar{x})$  and a real  $\delta > 0$  such that the inequality*

$$|f(w, x) - g(w, x)| \leq \delta \left\| wx^\top - \bar{w}\bar{x}^\top \right\|_F \quad \text{holds for all } (w, x) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}.$$

*Then for any stationary point  $(w, x)$  of  $g$ , there exists a point  $(\hat{w}, \hat{x})$  satisfying*

$$\begin{cases} \|(w, x) - (\hat{w}, \hat{x})\| &\leq 2\sqrt{\frac{\delta \|wx^\top - \bar{w}\bar{x}^\top\|}{\rho + \delta}} \\ \|\text{dist}(0, \partial f(\hat{w}, \hat{x}))\| &\leq \left( \delta + \sqrt{2\delta(\rho + \delta)} \right) (\|(w, x)\| + \|(\bar{w}, \bar{x})\|). \end{cases}$$

**Proof** The proposition is a corollary of Theorem 6.1 of [13]. Recall that for a function  $l : \mathbb{R}^d \rightarrow \overline{\mathbb{R}}$  the Lipschitz constant at  $\bar{y} \in \mathbb{R}^d$  is given by

$$\text{lip}(l, \bar{y}) := \limsup_{y \rightarrow \bar{y}} \frac{|l(y) - l(\bar{y})|}{|y - \bar{y}|}.$$

Set  $u(x) = \delta \|wx^\top - \bar{w}\bar{x}^\top\|$ , and  $l(x) = -\delta \|wx^\top - \bar{w}\bar{x}^\top\|$ . It is easy to see that at differentiable points the gradient of  $l(\cdot)$  is equal to

$$\nabla l(w, x) = -\frac{\delta}{\|wx^\top - \bar{w}\bar{x}^\top\|_F} \begin{bmatrix} (wx^\top - \bar{w}\bar{x}^\top)x \\ (wx^\top - \bar{w}\bar{x}^\top)^\top w \end{bmatrix} \implies \|\nabla l(w, x)\| \leq \delta \|(w, x)\|$$

Then, since  $\text{lip}(l; w, x) = \limsup_{(w', x') \rightarrow (w, x)} \|\nabla l(w', x')\|$ , we can over estimate

$$\text{lip}(l; w, x) \leq \delta (\|(w, x)\| + \|(\bar{w}, \bar{x})\|).$$

Thus applying Theorem 6.1 of [13] we get that for all  $\gamma > 0$  there exists  $(\hat{w}, \hat{x})$  such that  $\|(w, x) - (\hat{w}, \hat{x})\| \leq 2\gamma$  and

$$\text{dist}(0, \partial f(\hat{w}, \hat{x})) \leq 2\rho\gamma + 2\delta \frac{\|wx^\top - \bar{w}\bar{x}^\top\|}{\gamma} + \delta (\|(\hat{w}, \hat{x})\| + \|(\bar{w}, \bar{x})\|)$$

By the triangular inequality we get  $\|(\hat{w}, \hat{x})\| \leq 2\gamma + \|(w, x)\|$  and therefore

$$\text{dist}(0, \partial f(\hat{w}, \hat{x})) \leq 2(\rho + \delta)\gamma + 2\delta \frac{\|wx^\top - \bar{w}\bar{x}^\top\|}{\gamma} + \delta \|(w, x)\|.$$

Hence setting  $\gamma = \sqrt{\frac{\delta \|wx^\top - \bar{w}\bar{x}^\top\|}{\rho + \delta}}$ , gives

$$\begin{aligned} \text{dist}(0, \partial f(\hat{w}, \hat{x})) &\leq 2\sqrt{\delta(\rho + \delta)} \|wx^\top - \bar{w}\bar{x}^\top\| + \delta (\|(\hat{w}, \hat{x})\| + \|(\bar{w}, \bar{x})\|) \\ &\leq 2\sqrt{\delta(\rho + \delta)} (\|wx^\top\| + \|\bar{w}\bar{x}^\top\|) + \delta (\|(\hat{w}, \hat{x})\| + \|(\bar{w}, \bar{x})\|) \\ &\leq 2\sqrt{\delta(\rho + \delta)} \left( \sqrt{\|wx^\top\|} + \sqrt{\|\bar{w}\bar{x}^\top\|} \right) + \delta (\|(\hat{w}, \hat{x})\| + \|(\bar{w}, \bar{x})\|) \\ &\leq \left( \delta + \sqrt{2\delta(\rho + \delta)} \right) (\|(w, x)\| + \|(\bar{w}, \bar{x})\|) \end{aligned}$$

where we used that  $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$  and  $ab \leq (a^2 + b^2)/2$ . ■

**Proposition 16** *There exist numerical constants  $c_1, c_2 > 0$  such that for all  $(w, x) \in \mathbb{R}^{d_1 \times d_2}$  we have*

$$\left| f_S(w, x) - f_P(w, x) \right| \lesssim \left( \frac{d_1 + d_2 + 1}{m} \log \left( \frac{m}{d_1 + d_2 + 1} \right) \right)^{\frac{1}{2}} \|wx^\top - \bar{w}\bar{x}^\top\| \quad (21)$$

with probability at least  $1 - 2\exp(-c_1(d_1 + d_2 + 1))$  provided  $m \geq c_2(d_1 + d_2 + 1)$ .



**Proof** The proof of this proposition is almost entirely analogous to the proof of Theorem 4.6 in [9] (noting that Gaussian matrices satisfy the hypothesis of this result). The proof follows exactly the same up to Equation (4.19) in the aforementioned paper. Where the authors proved that there exists constants  $c_1, c_2, c_3, c_4 > 0$  such that for any  $t \in (0, c_4)$  the following uniform concentration bound holds

$$\left| f_S(w, x) - f_P(w, x) \right| \leq \frac{3}{2} t \|wx^\top - \bar{w}\bar{x}^\top\|_F \quad \text{for all } (w, x) \in \mathbb{R}^{d_1 \times d_2}$$

with probability at least  $1 - 2 \exp(c_1(d_1 + d_2 + 1 + 1) \log(c_2/t) - c_3 t^2 m)$ . This probability is at least  $1 - 2 \exp(-c_3 t^2 m/2)$  provided that

$$\frac{d_1 + d_2 + 1}{m} \leq \frac{c_3 t^2}{2c_1 \log(c_2/t)}. \quad (22)$$

Set  $t = \max\left(\sqrt{\frac{2c_1}{c_3}}, c_2\right) \left(\frac{d_1 + d_2 + 1}{m} \log\left(\frac{m}{d_1 + d_2 + 1}\right)\right)^{\frac{1}{2}}$ . This choice ensures that (22) holds, since

$$\frac{d_1 + d_2 + 1}{m} \leq \frac{(d_1 + d_2 + 1) \log(m/(d_1 + d_2 + 1))}{m \log\left(\frac{m}{d_1 + d_2 + 1} \log^{-1}\left(\frac{m}{d_1 + d_2 + 1}\right)\right)} \leq \frac{c_3 t^2}{2c_1 \log(c_2/t)}.$$

We can ensure that  $t \in (0, c_4)$  by setting  $m \geq C(d_1 + d_2 + 1)$  with  $C$  sufficiently large. This proves the result (after relabeling the constants).  $\blacksquare$

We are finally in position to proof the theorem.

**Proof** [Proof of Theorem 3] Fix  $v \geq 1$  and a fix point  $(w, x)$  satisfying  $\|(w, x)\| \leq \nu \|(\bar{w}, \bar{x})\|$ . Proposition 16 shows that there exist constants  $c_1, c_2 > 0$  such that with probability at least  $1 - 2 \exp(-c_1(d_1 + d_2 + 1))$  we have

$$|f_S(w, x) - f_P(w, x)| \leq \tilde{O}\left(\left(\frac{d_1 + d_2 + 1}{m}\right)^{\frac{1}{2}}\right) \|wx^\top - \bar{w}\bar{x}^\top\|_F \quad \forall (w, x) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$$

provided that  $m \geq c_2(d_1 + d_2 + 1)$ . To ease the notation let us denote  $\bar{\Delta} := \tilde{O}\left(\left(\frac{d_1 + d_2 + 1}{m}\right)^{\frac{1}{2}}\right)$ .

Assume that we are in the event in which this holds, it is known that  $f_S$  is  $\rho$ -weakly with high probability provided that  $m \geq C(d_1 + d_2 + 1)$ , see Section 3 and Theorem 4.6 in [9]. Now, assume that  $m$  is big enough and we are in the intersection of this two events. This holds with probability  $1 - c_3 \exp(c_1(d_1 + d_2 + 1))$  (for some possibly different constants  $c_1, c_3$ ). Hence by Proposition 15 there exists a point  $(\hat{w}, \hat{x})$  such that

$$\|(w, x) - (\hat{w}, \hat{x})\| \leq \frac{2}{\sqrt{\rho}} \sqrt{\bar{\Delta}} D_{wx} \quad \text{and} \quad \text{dist}(0, \partial f(\bar{w}, \bar{x})) \leq C \sqrt{\bar{\Delta}} D_{wx}$$

where  $D_{wx} = \|(w, x)\| + \|(\bar{w}, \bar{x})\|$ .

Notice that if  $\|(w, x)\| \leq \bar{\Delta}^{\frac{1}{4}} \|(\bar{w}, \bar{x})\|$  holds then the result holds immediately. So assume that this inequality is not satisfied. So we can lower bound

$$\|(\hat{w}, \hat{x})\| \geq \|(w, x)\| - \|(\hat{w}, \hat{x}) - (w, x)\| \geq \left(1 - 2\left(\frac{\bar{\Delta}}{\rho}\right)^{\frac{1}{2}} \left(1 + \bar{\Delta}^{-\frac{1}{4}}\right)\right) \|(w, x)\| \geq \frac{1}{2} \|(w, x)\|$$

where the first inequality follow by applying the triangle inequality and the last inequality follows for  $m$  sufficiently large, since we can ensure that for such  $m$  the term in the parenthesis is bigger than  $1/2$ . Therefore,

$$\begin{aligned} \text{dist}(0, \partial f(\widehat{w}, \widehat{x})) &\leq C\Delta^{\frac{1}{2}} (\|(w, x)\| + \|(\bar{w}, \bar{x})\|) \\ &\leq C\Delta^{\frac{1}{2}} \left(1 + \Delta^{-\frac{1}{4}}\right) \|(w, x)\| \\ &\leq 2C\Delta^{\frac{1}{2}} \left(1 + \Delta^{-\frac{1}{4}}\right) \|(\widehat{w}, \widehat{x})\| \\ &\leq 4C\Delta^{\frac{1}{4}} \|(\widehat{w}, \widehat{x})\|. \end{aligned}$$

Hence, by reducing  $\bar{\Delta}$  if necessary we can guarantee that  $\text{dist}(0, \partial f(\widehat{w}, \widehat{x})) \leq \gamma \|(\widehat{w}, \widehat{x})\|$  and consequently Theorem 13 gives that at least one of the following two holds

$$\max\{\|\widehat{w}\|, \|\widehat{x}\|\} \|\widehat{w}\widehat{x}^\top - \bar{w}\bar{x}^\top\| \lesssim \Delta^{\frac{1}{2}} D_{wx} \|\bar{w}\bar{x}^\top\| \quad \text{and} \quad \begin{cases} |\langle w, \bar{w} \rangle| \lesssim \nu^2 \Delta^{\frac{1}{2}} D_{wx} \|\bar{w}\| \\ |\langle x, \bar{x} \rangle| \lesssim \nu^2 \Delta^{\frac{1}{2}} D_{wx} \|\bar{x}\| \end{cases} \quad (23)$$

Let us prove that this implies the statement of the theorem.

**Case 1.** Assume that the second condition in (23) holds. Then

$$|\langle w, \bar{w} \rangle| \leq |\langle \widehat{w}, \bar{w} \rangle| + \|\bar{w}\| \|\widehat{w} - w\| \lesssim (\nu^2 + 1) \Delta^{\frac{1}{2}} D_{wx} \|\bar{w}\| \lesssim (\nu^2 + 1) \Delta^{\frac{1}{4}} \|(w, x)\| \|\bar{w}\|$$

where we used that  $\Delta^{\frac{1}{2}} D_{wx} \lesssim \Delta^{\frac{1}{4}} \|(w, x)\|$  for  $m$  big enough. A similar argument yields the result for  $|\langle w, x \rangle|$ .

**Case 2.** On the other hand, if the first condition holds, there exist  $e_w \in \mathbb{R}^{d_1}$ ,  $e_x \in \mathbb{R}^{d_2}$  such that  $\widehat{w} = w + e_w$  and  $\widehat{x} = x + e_x$  with  $\|e_w\|, \|e_x\| \leq \Delta^{\frac{1}{2}} D_{wx}$ . Then

$$\begin{aligned} \|(w, x)\| \|wx^\top - \bar{w}\bar{x}^\top\| &\leq \|(w, x)\| \|wx^\top - \widehat{w}\widehat{x}^\top\| + \|(w, x)\| \|\widehat{w}\widehat{x}^\top - \bar{w}\bar{x}^\top\| \\ &\leq \|(w, x)\| \|wx^\top - \widehat{w}\widehat{x}^\top\| + 2\|(\widehat{w}, \widehat{x})\| \|\widehat{w}\widehat{x}^\top - \bar{w}\bar{x}^\top\| \\ &\lesssim \|(w, x)\| \|wx^\top - (w + e_w)(x + e_x)^\top\| + \Delta^{\frac{1}{2}} D_{wx} \|\bar{w}\bar{x}^\top\| \\ &\leq \|(w, x)\| \left( \|we_x^\top\| + \|e_w x^\top\| + \|e_w e_x^\top\| \right) + \Delta^{\frac{1}{2}} D_{wx} \|\bar{w}\bar{x}^\top\| \\ &\leq \|(w, x)\| \Delta^{\frac{1}{2}} D_{wx} \left( \|w\| + \|x^\top\| + \Delta^{\frac{1}{2}} D_{wx} \right) + \Delta^{\frac{1}{2}} D_{wx} \|\bar{w}\bar{x}^\top\| \\ &\lesssim \|(w, x)\| \Delta^{\frac{1}{2}} D_{wx} \left( \|(w, x)\| + \Delta^{\frac{1}{4}} \|(w, x)\| \right) + \Delta^{\frac{1}{2}} D_{wx} \|\bar{w}\bar{x}^\top\| \\ &\lesssim \|(w, x)\|^2 \Delta^{\frac{1}{2}} D_{wx} + \Delta^{\frac{1}{2}} D_{wx} \|\bar{w}\bar{x}^\top\| \\ &\lesssim (\nu^2 + 1) \|\bar{w}\bar{x}^\top\| \Delta^{\frac{1}{2}} D_{wx} \\ &\lesssim (\nu^2 + 1) \Delta^{\frac{1}{4}} \|(w, x)\| \|\bar{w}\bar{x}^\top\|. \end{aligned}$$

Proving the desired result. ■