SIAM J. MATH. DATA SCI.
Vol. 2, No. 1, pp. 216–245

Geometry and Symmetry in Short-and-Sparse Deconvolution*

Han-Wen Kuo†, Yuqian Zhang‡, Yenson Lau†, and John Wright†§

Abstract. We study the *short-and-sparse (SaS) deconvolution* problem of recovering a short signal $\boldsymbol{a}_0$ and a sparse signal $\boldsymbol{x}_0$ from their convolution. We propose a method based on nonconvex optimization, which under certain conditions recovers the target short and sparse signals, up to a signed shift symmetry which is intrinsic to this model. This symmetry plays a central role in shaping the optimization landscape for deconvolution. We give a *regional analysis*, which characterizes this landscape geometrically, on a union of subspaces. Our geometric characterization holds when the length-p_0 short signal $\boldsymbol{a}_0$ has shift coherence μ, and $\boldsymbol{x}_0$ follows a random sparsity model with sparsity rate $\theta \in \left[\frac{c_1}{p_0}, \frac{c_2}{p_0\sqrt{\mu}+\sqrt{p_0}}\right] \cdot \frac{1}{\log^2 p_0}$. Based on this geometry, we give a provable method that successfully solves SaS deconvolution with high probability.

Key words. signal reconstruction, blind deconvolution, nonconvex geometry, nonconvex optimization

AMS subject classifications. 94A12, 90C26, 65Y20

DOI. 10.1137/19M1237569

1. Introduction.

Datasets in a wide range of areas, including neuroscience [37], microscopy [15], and astronomy [49], can be modeled as superpositions of translations of a basic motif. Data of this nature can be modeled mathematically as a convolution $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$, between a *short* signal $\boldsymbol{a}_0$ (the motif) and a longer *sparse* signal $\boldsymbol{x}_0$, whose nonzero entries indicate where in the sample the motif is present. A very similar structure arises in image deblurring [14], where $\boldsymbol{y}$ is a blurry image, $\boldsymbol{a}_0$ the blur kernel, and $\boldsymbol{x}_0$ the (edge map) of the target sharp image.

Motivated by these and related problems in imaging and scientific data analysis, we study the *short-and-sparse (SaS) deconvolution* problem of recovering a short signal $\boldsymbol{a}_0 \in \mathbb{R}^{p_0}$ and a sparse signal $\boldsymbol{x}_0 \in \mathbb{R}^n$ ($n \gg p_0$) from their length-n cyclic convolution $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0 \in \mathbb{R}^n$.[1] This SaS model exhibits a basic *scaled shift symmetry*: for any nonzero scalar α and cyclic shift $s_\ell[\cdot]$,

$$\left(\alpha\, s_\ell[\boldsymbol{a}_0]\right) * \left(\tfrac{1}{\alpha}\, s_{-\ell}[\boldsymbol{x}_0]\right) = \boldsymbol{y}. \tag{1.1}$$

Because of this symmetry, we only expect to recover $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$ up to a signed shift (see

*Received by the editors January 8, 2019; accepted for publication (in revised form) December 2, 2019; published electronically February 27, 2020.

https://doi.org/10.1137/19M1237569

Funding: This work was funded by National Science Foundation grants NSF 1343282, NSF CCF 1527809, and NSF IIS 1546411.

†Department of Electronic Engineering and Data Science Institute, Columbia University, New York, NY 10027 USA (hk2673@columbia.edu, yl3027@columbia.edu, jw2966@columbia.edu).

‡Department of Computer Science, Cornell University, Ithaca, NY 14853 USA (yz2409@columbia.edu).

§Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027 USA

[1] In this paper, the cyclic convolution $\boldsymbol{a}_0 * \boldsymbol{x}_0$ assumes $\boldsymbol{a}_0$ to be zero-padded $[\boldsymbol{a}_0, \boldsymbol{0}^{n-p_0}]$ to length n.

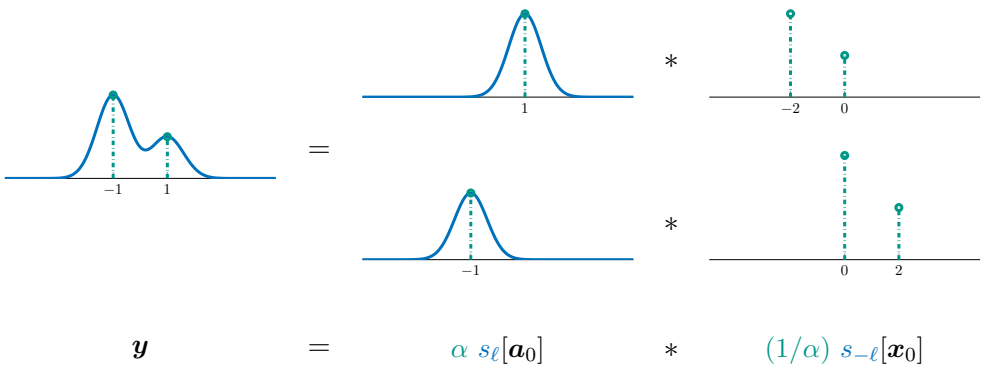

Figure 1. Shift symmetry in short-and-sparse deconvolution. *An observation $\boldsymbol{y}$ (left) which is a convolution of a short signal $\boldsymbol{a}_0$ and a sparse signal $\boldsymbol{x}_0$ (top right) can be equivalently expressed as a convolution of $s_\ell[\boldsymbol{a}_0]$ and $s_{-\ell}[\boldsymbol{x}_0]$, where $s_\ell[\cdot]$ denotes shift ℓ samples. The ground truth signals $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$ can only be identified up to a scaled shift.*

Figure 1). Our problem of interest can be stated more formally as follows.

Problem 1.1 (short-and-sparse deconvolution). Given the cyclic convolution[2] $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0 \in \mathbb{R}^n$ of $\boldsymbol{a}_0 \in \mathbb{R}^{p_0}$ short ($p_0 \ll n$) and $\boldsymbol{x}_0 \in \mathbb{R}^n$ sparse, recover $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$, up to a scaled shift.

Despite a long history and many applications, until recently very little algorithmic theory was available for SaS deconvolution. Much of this difficulty can be attributed to the scale-shift symmetry: natural convex relaxations fail,[3] and nonconvex formulations exhibit a complicated optimization landscape, with many equivalent global minimizers (scaled shifts of the ground truth), additional local minimizers (scaled shift truncations of the ground truth), and a variety of critical points [63, 64]. Currently available theory guarantees approximate recovery of a truncation[4] of a shift $s_\ell[\boldsymbol{a}_0]$, rather than guaranteeing recovery of $\boldsymbol{a}_0$ as a whole, and requires certain (complicated) conditions on the convolution matrix associated with $\boldsymbol{a}_0$ [63].

In this paper, we describe an algorithm which, under simpler conditions, *exactly* recovers a scaled shift of the pair $(\boldsymbol{a}_0, \boldsymbol{x}_0)$. Our algorithm is based on a formulation first introduced in [64], which casts the deconvolution problem as (nonconvex) optimization over the sphere. We characterize the geometry of this objective function and show that near a certain union of subspaces, every local minimizer is very close to a signed shift of $\boldsymbol{a}_0$. Based on this geometric analysis, we give provable methods for SaS deconvolution that exactly recover a scaled shift of $(\boldsymbol{a}_0, \boldsymbol{x}_0)$ whenever $\boldsymbol{a}_0$ is *shift-incoherent* and $\boldsymbol{x}_0$ is a sufficiently sparse random vector. Our geometric analysis highlights the role of symmetry in shaping the objective landscape for SaS deconvolution.

The remainder of this paper is organized as follows. Section 2 introduces our optimization

[2]Our result can be applied to recovering direct convolutions. Let $\boldsymbol{y} \in \mathbb{R}^{p_0+n-1}$ be the direct convolution between $\boldsymbol{a}_0 \in \mathbb{R}^{p_0}$ and $\boldsymbol{x}_0 \in \mathbb{R}^n$; then $\boldsymbol{y}$ can also be expressed as circular convolution between $\boldsymbol{a}_0$ and $[\boldsymbol{x}_0; \boldsymbol{0}^{p_0-1}]$.

[3]Such as matrix lifting relaxation [2, 39], in which $\boldsymbol{a}_0$ or $\boldsymbol{x}_0$ resides in random subspaces without shift symmetry.

[4]That is, the portion of the shifted signal $s_\ell[\boldsymbol{a}_0]$ that falls in the window $\{0, \dots, p_0 - 1\}$.

approach and modeling assumptions. Section 3 introduces our main results—both geometric and algorithmic—and compares them to the literature. Sections 4 and 5 describes the main ideas of our analysis. Section 6 demonstrates the experimental performance of the analyzed algorithm. Finally, section 7 discusses two main limitations of our analysis and describes directions for future work.

2. Formulation and assumptions.

2.1. Nonconvex SaS over the sphere.

Our starting point is the (natural) formulation

$$\min_{\boldsymbol{a},\boldsymbol{x}} \ \underbrace{\tfrac{1}{2}\|\boldsymbol{a}*\boldsymbol{x}-\boldsymbol{y}\|_2^2}_{\text{Data Fidelity}} + \underbrace{\lambda\|\boldsymbol{x}\|_1}_{\text{Sparsity}} \quad \text{s.t.} \quad \|\boldsymbol{a}\|_2 = 1. \tag{2.1}$$

We term this optimization problem the *Bilinear Lasso*, for its resemblance to the Lasso estimator in statistics. Indeed, letting

$$\varphi_{\text{lasso}}(\boldsymbol{a}) \equiv \min_{\boldsymbol{x}} \left\{ \tfrac{1}{2}\|\boldsymbol{a}*\boldsymbol{x}-\boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{x}\|_1 \right\} \tag{2.2}$$

denote the optimal Lasso cost, we see that (2.1) simply optimizes φ_{lasso} with respect to $\boldsymbol{a}$:

$$\min_{\boldsymbol{a}} \ \varphi_{\text{lasso}}(\boldsymbol{a}) \quad \text{s.t.} \quad \|\boldsymbol{a}\|_2 = 1. \tag{2.3}$$

In (2.1)–(2.3), we constrain $\boldsymbol{a}$ to have unit ℓ^2 norm. This constraint breaks the scale ambiguity between $\boldsymbol{a}$ and $\boldsymbol{x}$. Moreover, the choice of constraint manifold has surprisingly strong implications for computation: if $\boldsymbol{a}$ is instead constrained to the simplex, the problem admits trivial global minimizers. In contrast, local minima of the sphere-constrained formulation often correspond to shifts (or shift truncations [64]) of the ground truth $\boldsymbol{a}_0$.

The problem (2.3) is defined in terms of the optimal Lasso cost. This function is challenging to analyze, especially far away from $\boldsymbol{a}_0$. The article [64] analyzes the local minima of a simplification of (2.3), obtained by approximating[5] the data fidelity term as

$$\begin{aligned} \tfrac{1}{2}\|\boldsymbol{a}*\boldsymbol{x}-\boldsymbol{y}\|_2^2 &= \tfrac{1}{2}\|\boldsymbol{a}*\boldsymbol{x}\|_2^2 - \langle \boldsymbol{a}*\boldsymbol{x}, \boldsymbol{y}\rangle + \tfrac{1}{2}\|\boldsymbol{y}\|_2^2 \\ &\approx \tfrac{1}{2}\|\boldsymbol{x}\|_2^2 - \langle \boldsymbol{a}*\boldsymbol{x}, \boldsymbol{y}\rangle + \tfrac{1}{2}\|\boldsymbol{y}\|_2^2 . \end{aligned} \tag{2.4}$$

This yields a simpler objective function,

$$\varphi_{\ell^1}(\boldsymbol{a}) = \min_{\boldsymbol{x}} \left\{ \tfrac{1}{2}\|\boldsymbol{x}\|_2^2 - \langle \boldsymbol{a}*\boldsymbol{x}, \boldsymbol{y}\rangle + \tfrac{1}{2}\|\boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{x}\|_1 \right\}. \tag{2.5}$$

We make one further simplification to this problem, replacing the nondifferentiable penalty $\|\cdot\|_1$ with a smooth approximation $\rho(\boldsymbol{x})$.[6] Our analysis allows for a variety of smooth sparsity surrogates $\rho(\boldsymbol{x})$; for concreteness, we state our main results for the particular penalty[7]

$$\rho(\boldsymbol{x}) = \sum_i \left(\boldsymbol{x}_i^2 + \delta^2\right)^{1/2}. \tag{2.6}$$

[5] For a generic $\boldsymbol{a}$, we have $\langle s_i[\boldsymbol{a}], s_j[\boldsymbol{a}]\rangle \approx 0$ and hence $\|\boldsymbol{a}*\boldsymbol{x}\|_2^2 = \boldsymbol{x}^*\boldsymbol{C}_{\boldsymbol{a}}^*\boldsymbol{C}_{\boldsymbol{a}}\boldsymbol{x} \approx \boldsymbol{x}^*\boldsymbol{I}\boldsymbol{x} = \|\boldsymbol{x}\|_2^2$. The use of φ_ρ performs not as ideal compared to Bilinear Lasso when this approximation is inexact; see section 7.

[6] φ_{ℓ^1} is not twice differentiable everywhere and hence can't be minimized with conventional second order methods.

[7] This particular surrogate is sometimes called the pseudo-Huber function.

For $\delta > 0$, this is a smooth function of $\boldsymbol{x}$; as $\delta \searrow 0$ it approaches $\|\boldsymbol{x}\|_1$. Replacing $\|\cdot\|_1$ with $\rho(\cdot)$, we obtain the objective function which will be our main object of study,

$$\varphi_\rho(\boldsymbol{a}) = \min_{\boldsymbol{x}} \left\{ \tfrac{1}{2} \|\boldsymbol{x}\|_2^2 - \langle \boldsymbol{a} * \boldsymbol{x}, \boldsymbol{y} \rangle + \tfrac{1}{2} \|\boldsymbol{y}\|_2^2 + \lambda \rho(\boldsymbol{x}) \right\}. \tag{2.7}$$

As in [64], we optimize $\varphi_\rho(\boldsymbol{a})$ over the sphere $\mathbb{S}^{p-1}$:

$$\boxed{\min_{\boldsymbol{a}} \ \varphi_\rho(\boldsymbol{a}) \quad \text{s.t.} \quad \boldsymbol{a} \in \mathbb{S}^{p-1}.} \tag{2.8}$$

Here, we set $p = 3p_0 - 2$. As we will see, optimizing over this slightly higher dimensional sphere enables us to recover a (full) shift of $\boldsymbol{a}_0$, rather than a *truncated* shift. Our approach will leverage the following fact: if we view $\boldsymbol{a} \in \mathbb{S}^{p-1}$ as indexed by coordinates $W = \{-p_0 + 1, \ldots, 2p_0 - 1\}$, then for any shifts $\ell \in \{-p_0 + 1, \ldots, p_0 - 1\}$, the support of ℓ-shifted short signal $s_\ell[\boldsymbol{a}_0]$ is entirely contained in interval W. We will give a provable method which recovers a scaled version of one of these canonical shifts.

2.2. Analysis setting and assumptions. For convenience, we assume that $\boldsymbol{a}_0$ has unit ℓ^2 norm, i.e., $\boldsymbol{a}_0 \in \mathbb{S}^{p_0-1}$.[8] Our analysis makes two main assumptions, on the short motif $\boldsymbol{a}_0$ and the sparse map $\boldsymbol{x}_0$, respectively:

The first is that distinct shifts $\boldsymbol{a}_0$ have small inner product. We define the *shift coherence* of $\mu(\boldsymbol{a}_0)$ to be the largest inner product between distinct shifts:

$$\mu(\boldsymbol{a}_0) = \max_{\ell \neq 0} |\langle \boldsymbol{a}_0, s_\ell[\boldsymbol{a}_0] \rangle|. \tag{2.9}$$

The quantity $\mu(\boldsymbol{a}_0)$ is bounded between 0 and 1. Our theory allows any μ smaller than some numerical constant. Figure 2 shows three examples of families of $\boldsymbol{a}_0$ that satisfy this assumption:

- *Spiky.* When $\boldsymbol{a}_0$ is close to the Dirac delta $\boldsymbol{\delta}_0$, the shift coherence $\mu(\boldsymbol{a}_0) \approx 0$.[9] Here, the observed signal $\boldsymbol{y}$ consists of a superposition of sharp pulses. This is arguably the easiest instance of SaS deconvolution.
- *Generic.* If $\boldsymbol{a}_0$ is chosen uniformly at random from the sphere $\mathbb{S}^{p_0-1}$, its coherence is bounded as $\mu(\boldsymbol{a}_0) \lessapprox \sqrt{1/p_0}$ with high probability.
- *Tapered generic low-pass.* Here, $\boldsymbol{a}_0$ is generated by taking a random conjugate symmetric superposition of the first L length-p_0 discrete Fourier transform (DFT) basis signals, windowing (e.g., with a Hamming window) and normalizing to unit ℓ^2 norm. When $L = p_0\sqrt{1-\beta}$, with high probability $\mu(\boldsymbol{a}_0) \lessapprox \beta$. In this model, μ does not have to diminish as p_0 grows—it can be a fixed constant.[10]

[8] This is purely a technical convenience. Our theory guarantees recovery of a signed shift $(\pm s_\ell[\boldsymbol{a}_0], \pm s_{-\ell}[\boldsymbol{x}_0])$ of the truth. If $\boldsymbol{a}_0$ does not have unit norm, identical reasoning implies that our method recovers a scaled shift $\left(\alpha s_\ell[\boldsymbol{a}_0], \alpha^{-1} s_{-\ell}[\boldsymbol{x}_0]\right)$ with $\alpha = \pm \frac{1}{\|\boldsymbol{a}_0\|_2}$.

[9] The use of "$\approx$" here suppresses constant and logarithmic factors.

[10] The upper right panel of Figure 2 is generated using random DFT components with frequencies smaller than one-third Nyquist. Such a kernel is incoherent, with high probability. Many commonly occurring low-pass kernels have $\mu(\boldsymbol{a}_0)$ larger—very close to one. One of the most important limitations of our results is that they do not provide guarantees in this highly coherent situation. See [34].

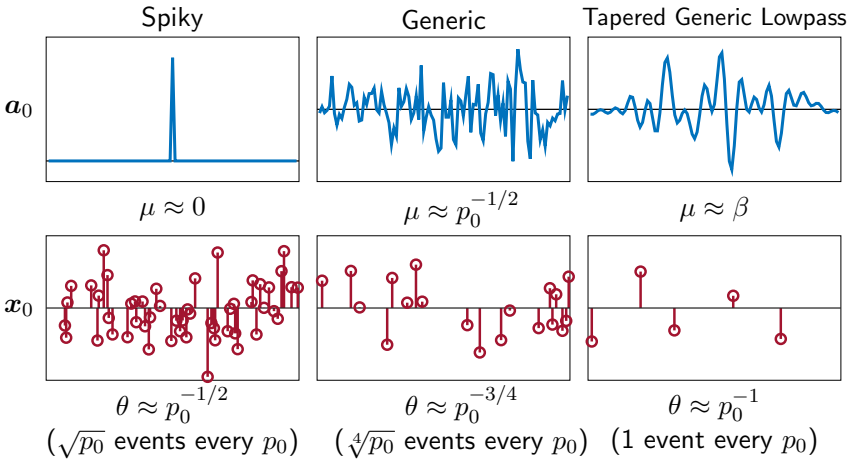

Figure 2. Sparsity-coherence tradeoff. *Top: three families of motifs $\boldsymbol{a}_0$ with varying coherence μ. Bottom: maximum allowable sparsity θ and number of copies θp_0 within each length-p_0 window. Here, we suppress constants and logarithmic factors. When the target motif has smaller shift-coherence μ, our result allows larger θ, and vice versa. This sparsity-coherence tradeoff is made precise in our main result, Theorem 3.1, which, loosely speaking, asserts that when $\theta \lessapprox 1/(p_0\sqrt{\mu} + \sqrt{p_0})$, our method succeeds.*

Intuitively speaking, problems with smaller μ are easier to solve, a claim which will be made precise in our technical results.

We assume that $\boldsymbol{x}_0$ is a sparse random vector. More precisely, we assume that $\boldsymbol{x}_0$ is Bernoulli–Gaussian, with rate θ:

$$\boldsymbol{x}_{0i} = \boldsymbol{\omega}_i \boldsymbol{g}_i, \tag{2.10}$$

where $\boldsymbol{\omega}_i \sim \mathrm{Ber}(\theta)$, $\boldsymbol{g}_i \sim \mathcal{N}(0,1)$, and all random variables are jointly independent. We write this as

$$\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta). \tag{2.11}$$

Here, θ is the probability that a given entry $\boldsymbol{x}_{0i}$ is nonzero. Problems with smaller θ are easier to solve. In the extreme case, when $\theta \ll 1/p_0$, the observation $\boldsymbol{y}$ contains many isolated copies of the motif $\boldsymbol{a}_0$, and $\boldsymbol{a}_0$ can be determined by direct inspection. Our analysis will focus on the nontrivial scenario, when $\theta \gtrapprox 1/p_0$.

Our technical results will articulate *sparsity-coherence* tradeoffs, in which smaller coherence μ enables larger θ, and vice versa. More specifically, in our main theorem, the sparsity-coherence relationship is captured in the form

$$\theta \lessapprox 1/(p_0\sqrt{\mu} + \sqrt{p_0}). \tag{2.12}$$

When the target $\boldsymbol{a}_0$ is very shift-incoherent ($\mu \approx 0$), our method succeeds when each length-p_0 window contains about $\sqrt{p_0}$ copies of $\boldsymbol{a}_0$. When μ is larger (as in the generic low-pass model), our method succeeds as long as relatively few copies of $\boldsymbol{a}_0$ overlap in the observed signal. In Figure 2, we illustrate these tradeoffs for the three models described above.

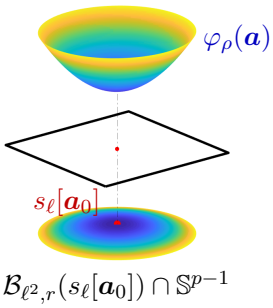

Figure 3. Geometry of φ_ρ near a shift of $\boldsymbol{a}_0$. *Bottom: a portion of the sphere $\mathbb{S}^{p-1}$, colored according to φ_ρ. Top: φ_ρ visualized as height. φ_ρ is strongly convex in this region, and it has a minimizer very close to $s_\ell[\boldsymbol{a}_0]$.*

3. Main results: Geometry and algorithms. In this section, we introduce our main results—on the geometry of φ_ρ (subsection 3.1) and its algorithmic implications (subsection 3.2). Finally, in subsection 3.3, we compare these results with the literature on deconvolution.

3.1. Geometry of the objective φ_ρ. The goal in SaS deconvolution is to recover $\boldsymbol{a}_0$ (and $\boldsymbol{x}_0$) up to a signed shift; i.e., we wish to recover some $\pm s_\ell[\boldsymbol{a}_0]$. The shifts $\pm s_\ell[\boldsymbol{a}_0]$ play a key role in shaping the landscape of φ_ρ. In particular, we will argue that over a certain subset of the sphere, *every local minimum of φ_ρ is close to some $\pm s_\ell[\boldsymbol{a}_0]$*.

To gain intuition into the properties of φ_ρ, we first visualize this function in the vicinity of a single shift $s_\ell[\boldsymbol{a}_0]$ of the ground truth $\boldsymbol{a}_0$. In Figure 3, we plot the function value of φ_ρ over

$$\mathcal{B}_{\ell^2,r}(s_\ell[\boldsymbol{a}_0]) \cap \mathbb{S}^{p-1},$$

where $\mathcal{B}_{\ell^2,r}(\boldsymbol{a})$ is a ball of radius r around $\boldsymbol{a}$. We make two observations:

- The objective function φ_ρ is strongly convex in this neighborhood of $s_\ell[\boldsymbol{a}_0]$.
- There is a local minimizer very close to $s_\ell[\boldsymbol{a}_0]$.

We next visualize the objective function φ_ρ near the linear span of *two* different shifts, $s_{\ell_1}[\boldsymbol{a}_0]$ and $s_{\ell_2}[\boldsymbol{a}_0]$. More precisely, we plot φ_ρ near the intersection (Figure 4, left) of the sphere $\mathbb{S}^{p-1}$ and the linear subspace

$$\mathcal{S}_{\{\ell_1,\ell_2\}} = \{\, \boldsymbol{\alpha}_1 s_{\ell_1}[\boldsymbol{a}_0] + \boldsymbol{\alpha}_2 s_{\ell_2}[\boldsymbol{a}_0] \mid \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathbb{R} \,\}.$$

We make three observations:

- Again, there is a local minimizer near each shift $s_\ell[\boldsymbol{a}_0]$.
- These are the *only* local minimizers in the vicinity of $\mathcal{S}_{\{\ell_1,\ell_2\}}$. In particular, the objective function φ exhibits *negative curvature* along $S_{\{\ell_1,\ell_2\}}$ at any superposition $\boldsymbol{\alpha}_1 s_{\ell_1}[\boldsymbol{a}_0] + \boldsymbol{\alpha}_2 s_{\ell_2}[\boldsymbol{a}_0]$ whose weights $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are balanced, i.e., $|\boldsymbol{\alpha}_1| \approx |\boldsymbol{\alpha}_2|$.
- Furthermore, the function φ_ρ exhibits *positive curvature* in directions away from the subspace $\mathcal{S}_{\ell_1,\ell_2}$.

Finally, we visualize φ_ρ over the intersection (Figure 5, left) of the sphere $\mathbb{S}^{p-1}$ with the linear span of three shifts $s_{\ell_1}[\boldsymbol{a}_0]$, $s_{\ell_2}[\boldsymbol{a}_0]$, $s_{\ell_3}[\boldsymbol{a}_0]$ of the true kernel $\boldsymbol{a}_0$:

$$\mathcal{S}_{\{\ell_1,\ell_2,\ell_3\}} = \{\, \boldsymbol{\alpha}_1 s_{\ell_1}[\boldsymbol{a}_0] + \boldsymbol{\alpha}_2 s_{\ell_2}[\boldsymbol{a}_0] + \boldsymbol{\alpha}_3 s_{\ell_3}[\boldsymbol{a}_0] \mid \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3 \in \mathbb{R} \,\}.$$

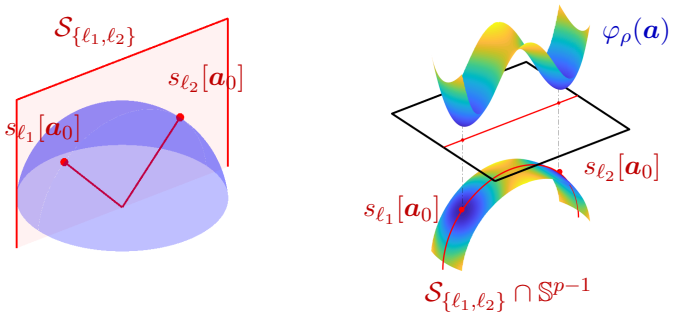

Figure 4. Geometry of φ_ρ near the span $\mathcal{S}_{\{\ell_1,\ell_2\}}$ of two shifts of $\boldsymbol{a}_0$. *Left: each pair of shifts* $s_{\ell_1}[\boldsymbol{a}_0]$, $s_{\ell_2}[\boldsymbol{a}_0]$ *defines a linear subspace* $\mathcal{S}_{\{\ell_1,\ell_2\}}$ *of* $\mathbb{R}^p$. *Center/right: every local minimum of* φ_ρ *near* $\mathcal{S}_{\{\ell_1,\ell_2\}}$ *(red line) is close to either* $s_{\ell_1}[\boldsymbol{a}_0]$ *or* $s_{\ell_2}[\boldsymbol{a}_0]$*; there is a negative curvature in the middle of* $s_{\ell_1}[\boldsymbol{a}_0]$ *and* $s_{\ell_2}[\boldsymbol{a}_0]$*, and* φ_ρ *is convex in direction away from* $\mathcal{S}_{\{\ell_1,\ell_2\}}$.

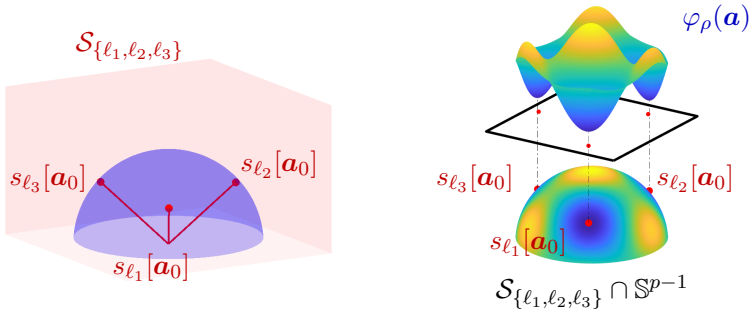

Figure 5. Geometry of φ_ρ over the span $\mathcal{S}_{\{\ell_1,\ell_2,\ell_3\}}$ of three shifts of $\boldsymbol{a}_0$. *The subspace* $\mathcal{S}_{\{\ell_1,\ell_2,\ell_3\}}$ *is three-dimensional; its intersection with the sphere* $\mathbb{S}^{p-1}$ *is isomorphic to a two-dimensional sphere. On this set,* φ_ρ *has local minimizers near each of the* $s_{\ell_i}[\boldsymbol{a}_0]$ *and are the only minimizers near* $\mathcal{S}_{\ell_1,\ell_2,\ell_3}$.

Again, *there is a local minimizer near each signed shift.* At roughly balanced superpositions of shifts, the objective function exhibits negative curvature. As a result, again, the *only* local minimizers are close to signed shifts.

Our main geometric result will show that these properties are obtained from *every* subspace spanned by a few shifts of $\boldsymbol{a}_0$. Indeed, for each subset

$$\boldsymbol{\tau} \subseteq \{-p_0+1, \ldots, p_0-1\}, \tag{3.1}$$

define a linear subspace

$$\mathcal{S}_{\boldsymbol{\tau}} = \left\{ \sum_{\ell \in \boldsymbol{\tau}} \boldsymbol{\alpha}_\ell s_\ell[\boldsymbol{a}_0] \,\middle|\, \boldsymbol{\alpha}_{-p_0+1}, \ldots, \boldsymbol{\alpha}_{p_0-1} \in \mathbb{R} \right\}. \tag{3.2}$$

The subspace $\mathcal{S}_{\boldsymbol{\tau}}$ is the linear span of the shifts $s_\ell[\boldsymbol{a}_0]$ indexed by ℓ in the set $\boldsymbol{\tau}$. Our geometric theory will show that with high probability the function φ_ρ has no spurious local minimizers

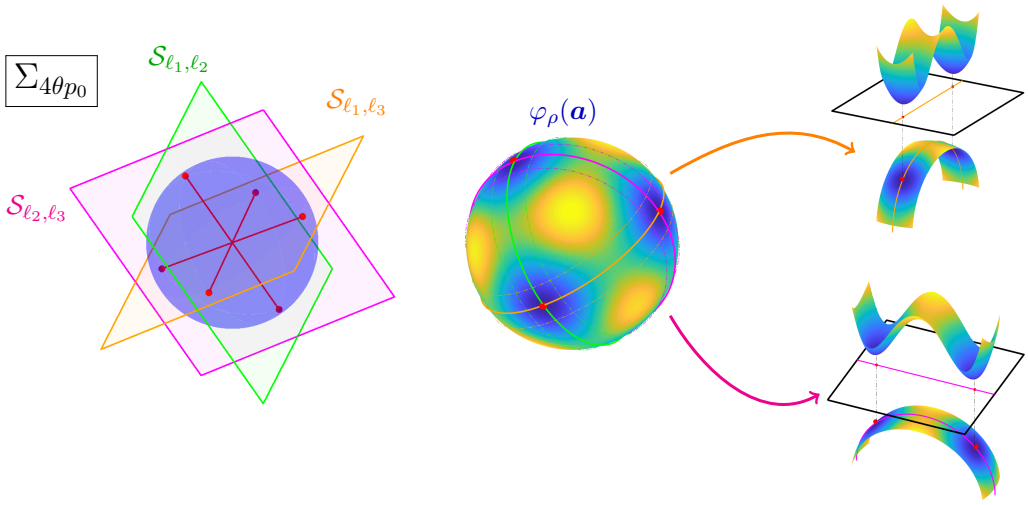

Figure 6. Geometry of φ_ρ over the union of subspaces $\Sigma_{2\theta p_0}$. *Left: schematic representation of the union of subspaces* $\Sigma_{4\theta p_0}$. *For each set* $\boldsymbol{\tau}$ *of at most* $4\theta p_0$ *shifts, we have a subspace* $\mathcal{S}_{\boldsymbol{\tau}}$. *Right:* φ_ρ *has good geometry near this union of subspaces.*

near any $\mathcal{S}_{\boldsymbol{\tau}}$ for which $\boldsymbol{\tau}$ is not too large—say, $|\boldsymbol{\tau}| \le 4\theta p_0$. Combining all of these subspaces into a single geometric object, define the union of subspaces

$$\Sigma_{4\theta p_0} = \bigcup_{|\boldsymbol{\tau}| \le 4\theta p_0} \mathcal{S}_{\boldsymbol{\tau}}. \tag{3.3}$$

Figure 6 (left) gives a schematic representation of this set. We claim the following:

- In the neighborhood of $\Sigma_{4\theta p_0}$, all local minimizers are near signed shifts.
- The value of φ_ρ grows in any direction away from $\Sigma_{4\theta p_0}$.

Our main result formalizes the above observations under two key assumptions: first, that the sparsity rate θ is sufficiently small (relative to the shift coherence μ of p_0), and, second, that the signal length n is sufficiently large.

Theorem 3.1 (main geometric theorem). *Let* $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$ *with* $\boldsymbol{a}_0 \in \mathbb{S}^{p_0-1}$ μ*-shift coherent*[11] *and* $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \text{BG}(\theta) \in \mathbb{R}^n$ *with sparsity rate*

$$\theta \in \left[\frac{c_1}{p_0}, \frac{c_2}{p_0\sqrt{\mu} + \sqrt{p_0}}\right] \cdot \frac{1}{\log^2 p_0}. \tag{3.4}$$

Choose $\rho(x) = \sqrt{x^2 + \delta^2}$ *and set* $\lambda = 0.1/\sqrt{p_0\theta}$ *in* φ_ρ. *Then there exist* $\delta > 0$ *and numerical constant* c *such that if* $n \ge \text{poly}(p_0)$, *with high probability, every local minimizer* $\bar{\boldsymbol{a}}$ *of* φ_ρ *over* $\Sigma_{4\theta p_0}$ *satisfies* $\|\bar{\boldsymbol{a}} - \sigma s_\ell[\boldsymbol{a}_0]\|_2 \le c \max\{\mu, p_0^{-1}\}$ *for some signed shift* $\sigma s_\ell[\boldsymbol{a}_0]$ *of the true kernel. Above,* $c_1, c_2 > 0$ *are positive numerical constants.*

[11] Typically it *is* possible to provide an overestimate $p_0' \ge p_0$. Our theory and algorithm can be applied directly to the overestimate p_0', with the caveat that the sparsity rate θ now scales with p_0' rather than p_0.

Proof. This follows from Theorem 4.1. ∎

The upper bound on θ in (3.4) yields the tradeoff between coherence and sparsity described in Figure 2. Simply put, when $\boldsymbol{a}_0$ is better conditioned (as a kernel), its coherence μ is smaller and $\boldsymbol{x}_0$ can be denser.

At a technical level, our proof of Theorem 3.1 shows that (i) $\varphi_\rho(\boldsymbol{a})$ is strongly convex in the vicinity of each signed shift, and that at every other point $\boldsymbol{a}$ near $\Sigma_{4\theta p_0}$, there is either (ii) a nonzero gradient or (iii) a direction of strict negative curvature; furthermore (iv) the function φ_ρ grows away from $\Sigma_{4\theta p_0}$. Points (ii)–(iii) imply that near $\Sigma_{4\theta p_0}$ there are no "flat" saddles: every saddle point has a direction of strict negative curvature. We will leverage these properties to propose an efficient algorithm for finding a local minimizer near $\Sigma_{4\theta p_0}$. Moreover, this minimizer is close enough to a shift (here, $\|\bar{\boldsymbol{a}} - s_\ell[\boldsymbol{a}_0]\|_2 \lessapprox \mu$) for us to exactly recover $s_\ell[\boldsymbol{a}_0]$: we will give a refinement algorithm that produces $(\pm s_\ell[\boldsymbol{a}_0], \pm s_{-\ell}[\boldsymbol{x}_0])$.

3.2. Provable algorithm for SaS deconvolution. The objective function φ_ρ has good geometric properties on (and near!) the union of subspaces $\Sigma_{4\theta p_0}$. In this section, we show an efficient method that exactly recovers $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$ up to shift symmetry. Although our geometric analysis only controls φ_ρ near $\Sigma_{4\theta p_0}$, we will give a descent method which, with appropriate initialization $\boldsymbol{a}^{(0)}$, produces iterates $\boldsymbol{a}^{(1)}, \ldots, \boldsymbol{a}^{(k)}, \ldots$ that remain close to $\Sigma_{4\theta p_0}$ for all k. In short, it is easy to *start* near $\Sigma_{4\theta p_0}$ and easy to *stay* near $\Sigma_{4\theta p_0}$. After finding a local minimizer $\bar{\boldsymbol{a}}$, we refine it to produce a signed shift of $(\boldsymbol{a}_0, \boldsymbol{x}_0)$ using alternating minimization.

The next two paragraphs give the main ideas behind the principal steps of the algorithm. We then describe its components in more detail (Algorithm 3.1) and state our main algorithmic result (Theorem 3.2), which asserts that under appropriate conditions this method produces a signed shift of $(\boldsymbol{a}_0, \boldsymbol{x}_0)$.

Our algorithm starts with an initialization scheme which generates $\boldsymbol{a}^{(0)}$ near the union of subspaces $\Sigma_{4\theta p_0}$, which consists of linear combinations of just a few shifts of $\boldsymbol{a}_0$. How can we find a point near this union? Notice that *the data* $\boldsymbol{y}$ *also consists of a linear combination of just a few shifts of* $\boldsymbol{a}_0$ Indeed,

$$\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0 \quad = \sum_{\ell \in \mathrm{supp}(\boldsymbol{x}_0)} \boldsymbol{x}_{0\ell} s_\ell[\boldsymbol{a}_0]. \tag{3.5}$$

A length-p_0 segment of data $\boldsymbol{y}_{0,\ldots,p_0-1} = [\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{p_0-1}]^*$ captures portions of roughly $2\theta p_0 \ll 4\theta p_0$ shifts $s_\ell[\boldsymbol{a}_0]$.

Many of these copies of $\boldsymbol{a}_0$ are truncated by the restriction to $\{0, \ldots, p_0 - 1\}$. A relatively simple remedy is as follows: First, we zero-pad $\boldsymbol{y}_{0,\ldots,p_0-1}$ to length $p = 3p_0 - 2$, giving

$$\left[\mathbf{0}^{p_0-1}; \boldsymbol{y}_0; \cdots; \boldsymbol{y}_{p_0-1}; \mathbf{0}^{p_0-1}\right]. \tag{3.6}$$

Zero-padding provides enough space to accommodate any shift $s_\ell[\boldsymbol{a}_0]$ with $\ell \in \boldsymbol{\tau}$. We then

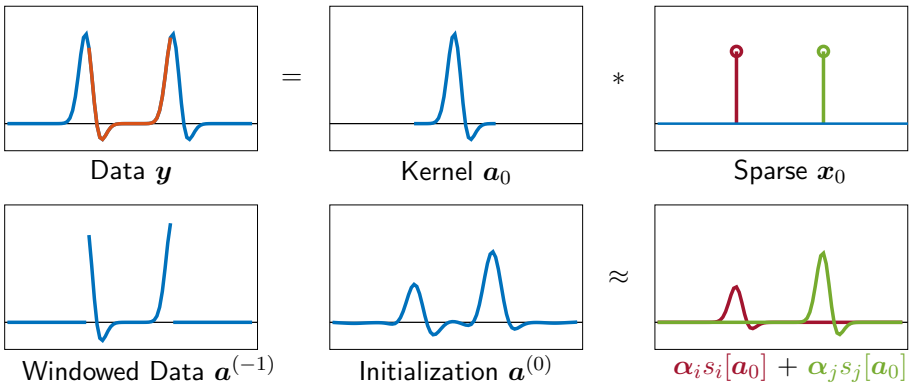

Figure 7. Data-driven initialization. *Using a piece of the observed data $\boldsymbol{y}$ to generate an initial point $\boldsymbol{a}^{(0)}$ that is close to a superposition of shifts $s_\ell[\boldsymbol{a}_0]$ of the ground truth. Top: data $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$ is a superposition of shifts of the true kernel $\boldsymbol{a}_0$. Bottom: a length-p_0 window contains pieces of just a few shifts. Bottom middle: one step of the generalized power method approximately fills in the missing pieces, yielding a near superposition of shifts of $\boldsymbol{a}_0$ (right).*

perform one step of the generalized power method,[12] writing

$$\boldsymbol{a}^{(0)} = -\boldsymbol{P}_{\mathbb{S}^{p-1}} \nabla \varphi_{\ell^1} \left(\boldsymbol{P}_{\mathbb{S}^{p-1}} \left[\boldsymbol{0}^{p_0-1}; \boldsymbol{y}_0; \cdots; \boldsymbol{y}_{p_0-1}; \boldsymbol{0}^{p_0-1} \right] \right), \tag{3.7}$$

where $\boldsymbol{P}_{\mathbb{S}^{p-1}}$ projects onto the sphere. The reasoning behind this construction may seem obscure. We will explain it at a more technical level in section 5 after interpreting the gradient $\nabla \varphi_\rho$ in terms of its action on the shifts $s_\ell[\boldsymbol{a}_0]$ in section 4. For now, we note that this operation has the effect of (approximately) filling in the missing pieces of the truncated shifts $s_\ell[\boldsymbol{a}_0]$; see Figure 7 for an example. We will prove that with high probability $\boldsymbol{a}^{(0)}$ is indeed close to $\Sigma_{4\theta p_0}$.

The next key observation is that the function φ_ρ grows as we move away from the subspace $\mathcal{S}_{\boldsymbol{\tau}}$; see Figure 8. Because of this, a small-stepping descent method will not move far away from $\Sigma_{4\theta p_0}$. For concreteness, we will analyze a variant of the curvilinear search method [23, 24], which moves in a linear combination of the negative gradient direction $-\boldsymbol{g}$ and a negative curvature direction $-\boldsymbol{v}$. At the kth iteration, the algorithm updates $\boldsymbol{a}^{(k+1)}$ as

$$\boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p-1}} \left[\boldsymbol{a}^{(k)} - t\boldsymbol{g}^{(k)} - t^2 \boldsymbol{v}^{(k)} \right] \tag{3.8}$$

with appropriately chosen step size t. The inclusion of a negative curvature direction allows the method to avoid stagnation near saddle points. Indeed, we will prove that, starting from initialization $\boldsymbol{a}^{(0)}$, this method produces a sequence $\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}, \ldots$ which efficiently converges to a local minimizer $\bar{\boldsymbol{a}}$ that is near some signed shift $\pm s_\ell[\boldsymbol{a}_0]$ of the ground truth.

[12] The power method for minimizing a quadratic form $\xi(\boldsymbol{a}) = \frac{1}{2}\boldsymbol{a}^* \boldsymbol{M} \boldsymbol{a}$ over the sphere consists of the iteration $\boldsymbol{a} \mapsto -\boldsymbol{P}_{\mathbb{S}^{p-1}} \boldsymbol{M} \boldsymbol{a}$. Notice that in this mapping, $-\boldsymbol{M}\boldsymbol{a} = -\nabla \xi(\boldsymbol{a})$. The generalized power method, for minimizing a function φ over the sphere, consists of repeatedly projecting $-\nabla \varphi$ onto the sphere, giving the iteration $\boldsymbol{a} \mapsto -\boldsymbol{P}_{\mathbb{S}^{p-1}} \nabla \varphi(\boldsymbol{a})$. Equation (3.7) can be interpreted as one step of the generalized power method for the objective function φ_ρ.

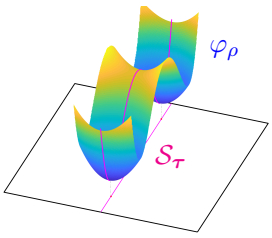

Figure 8. Growth of φ_ρ away from $\mathcal{S}_\tau$. *Because φ_ρ grows away from $\mathcal{S}_\tau$, small-stepping descent methods stay near $\mathcal{S}_\tau$.*

The second step of our algorithm *rounds* the local minimizer $\bar{\boldsymbol{a}} \approx \sigma s_\ell[\boldsymbol{a}_0]$ to produce an exact solution $\widehat{\boldsymbol{a}} = \sigma s_\ell[\boldsymbol{a}_0]$. As a by-product, it also exactly recovers the corresponding signed shift of the true sparse signal, $\widehat{\boldsymbol{x}} = \sigma s_{-\ell}[\boldsymbol{x}_0]$.

Our rounding algorithm is an alternating minimization scheme, which alternates between minimizing the Lasso cost over $\boldsymbol{a}$ with $\boldsymbol{x}$ fixed, and minimizing the Lasso cost over $\boldsymbol{x}$ with $\boldsymbol{a}$ fixed. We make two modifications to this basic idea, both of which are important for obtaining exact recovery. First, unlike the standard Lasso cost, which penalizes all of the entries of $\boldsymbol{x}$, we maintain a running estimate $I^{(k)}$ of the support of $\boldsymbol{x}_0$ and only penalize those entries that are not in $I^{(k)}$:

$$\tfrac{1}{2}\left\|\boldsymbol{a} * \boldsymbol{x} - \boldsymbol{y}\right\|_2^2 + \lambda \sum_{i \notin I^{(k)}} \left|\boldsymbol{x}_i\right|. \tag{3.9}$$

This can be viewed as an extreme form of *reweighting* [11]. Second, our algorithm gradually decreases penalty variable λ to 0, so that eventually

$$\widehat{\boldsymbol{a}} * \widehat{\boldsymbol{x}} \approx \boldsymbol{y}. \tag{3.10}$$

This can be viewed as a *homotopy* or *continuation* method [46, 19]. For concreteness, at the kth iteration the algorithm reads

$$\text{Update } \boldsymbol{x}\text{:} \quad \boldsymbol{x}^{(k+1)} \leftarrow \operatorname*{argmin}_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{a}^{(k)} * \boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda^{(k)} \sum_{i \notin I^{(k)}} \left|\boldsymbol{x}_i\right|; \tag{3.11}$$

$$\text{Update } \boldsymbol{a}\text{:} \quad \boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p-1}}\big[\operatorname*{argmin}_{\boldsymbol{a}} \tfrac{1}{2}\|\boldsymbol{a} * \boldsymbol{x}^{(k+1)} - \boldsymbol{y}\|_2^2\big]; \tag{3.12}$$

$$\text{Update } \lambda \text{ and } I\text{:} \quad \lambda^{(k+1)} \leftarrow \tfrac{1}{2}\lambda^{(k)}, \qquad I^{(k+1)} \leftarrow \operatorname{supp}\big(\boldsymbol{x}^{(k+1)}\big). \tag{3.13}$$

We prove that the iterates produced by this sequence of operations converge to the ground truth at a linear rate, as long as the initializer $\bar{\boldsymbol{a}}$ is sufficiently nearby.

Our overall algorithm is summarized as Algorithm 3.1. Figure 9 illustrates the main steps of this algorithm. Our main algorithmic result states that under essentially the same hypotheses as above, Algorithm 3.1 produces a signed shift of the ground truth $(\boldsymbol{a}_0, \boldsymbol{x}_0)$.

Algorithm 3.1. Short-and-sparse deconvolution.

Input: Observation $\boldsymbol{y}$, motif length p_0, sparsity θ, shift-coherence μ, and curvature threshold $-\eta_v$.

Minimization:

Set $\boldsymbol{a}^{(0)} \leftarrow -\boldsymbol{P}_{\mathbb{S}^{p-1}} \nabla \varphi_\rho \left(\boldsymbol{P}_{\mathbb{S}^{p-1}} \left[\boldsymbol{0}^{p_0-1}; \boldsymbol{y}_0; \cdots; \boldsymbol{y}_{p_0-1}; \boldsymbol{0}^{p_0-1} \right] \right)$.

Set $\lambda = 0.1/\sqrt{p_0\theta}$ [13] and $\delta > 0$ in φ_ρ. For $k = 1, 2, \ldots, K_1$, let

$$\boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p-1}}[\boldsymbol{a}^{(k)} - t\boldsymbol{g}^{(k)} - t^2\boldsymbol{v}^{(k)}], \tag{3.14}$$

where $\boldsymbol{g}^{(k)}$ is the Riemannian gradient; $\boldsymbol{v}^{(k)}$ is the eigenvector of smallest Riemannian Hessian eigenvalue if less than $-\eta_v$ with $\langle \boldsymbol{v}^{(k)}, \boldsymbol{g}^{(k)} \rangle \geq 0$, otherwise let $\boldsymbol{v}^{(k)} = \boldsymbol{0}$; and $t \in (0, 0.1/n\theta]$ satisfies

$$\varphi_\rho(\boldsymbol{a}^{(k+1)}) < \varphi_\rho(\boldsymbol{a}^{(k)}) - \tfrac{1}{2}t\|\boldsymbol{g}^{(k)}\|_2^2 - \tfrac{1}{4}t^4\eta_v\|\boldsymbol{v}^{(k)}\|_2^2 \tag{3.15}$$

to obtain a near local minimizer $\bar{\boldsymbol{a}} \leftarrow \boldsymbol{a}^{(K_1)}$.

Refinement:

Set $\boldsymbol{a}^{(0)} \leftarrow \bar{\boldsymbol{a}}$, $\lambda^{(0)} \leftarrow 10(p\theta + \log n)(\mu + 1/p)$, and $I^{(0)} \leftarrow \mathcal{S}_{\lambda^{(0)}}[\mathrm{supp}(\overleftarrow{\boldsymbol{y}} * \bar{\boldsymbol{a}})]$. For $k = 1, 2, \ldots, K_2$, let

$$\boldsymbol{x}^{(k+1)} \leftarrow \mathrm{argmin}_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{a}^{(k)} * \boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda^{(k)} \textstyle\sum_{i \notin I^{(k)}} |\boldsymbol{x}_i|, \tag{3.16}$$

$$\boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\mathrm{argmin}_{\boldsymbol{a}} \tfrac{1}{2}\|\boldsymbol{a} * \boldsymbol{x}^{(k+1)} - \boldsymbol{y}\|_2^2 \right], \tag{3.17}$$

$$\lambda^{(k+1)} \leftarrow \lambda^{(k)}/2, \qquad I^{(k+1)} \leftarrow \mathrm{supp}(\boldsymbol{x}^{(k+1)}), \tag{3.18}$$

to obtain $(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{x}}) \leftarrow (\boldsymbol{a}^{(K_2)}, \boldsymbol{x}^{(K_2)})$.

Output: Return $(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{x}})$.

Theorem 3.2 (main algorithmic theorem). *Suppose $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$, where $\boldsymbol{a}_0 \in \mathbb{S}^{p_0-1}$ is μ-truncated shift coherent*[14] *such that $\max_{i \neq j} \left| \langle \boldsymbol{\iota}_{p_0}^* s_i[\boldsymbol{a}_0], \boldsymbol{\iota}_{p_0}^* s_j[\boldsymbol{a}_0] \rangle \right| \leq \mu$ and $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta) \in \mathbb{R}^n$ with θ, μ satisfying*

$$\theta \in \left[\frac{c_1}{p_0}, \frac{c_2}{\left(p_0\sqrt{\mu} + \sqrt{p_0}\right)\log^2 p_0} \right], \qquad \mu \leq \frac{c_3}{\log^2 n} \tag{3.19}$$

for some constant $c_1, c_2, c_3 > 0$. If the signal lengths n, p_0 satisfy $n > \mathrm{poly}(p_0)$ and $p_0 > \mathrm{polylog}(n)$, then there exist $\delta, \eta_v > 0$ such that with high probability Algorithm 3.1 produces $(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{x}})$ that are equal to the ground truth up to signed shift symmetry:

$$\left\| (\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{x}}) - \sigma\left(s_\ell[\boldsymbol{a}_0], s_{-\ell}[\boldsymbol{x}_0]\right) \right\|_2 \leq \varepsilon \tag{3.20}$$

for $\sigma \in \{\pm 1\}$ and $\ell \in \{-p_0 + 1, \ldots, p_0 - 1\}$ if $K_1 > \mathrm{poly}(n, p_0)$ and $K_2 > \mathrm{polylog}(n, p_0, \varepsilon^{-1})$.

Proof. See Theorems 5.1 and 5.2. ∎

[13] In practice, we suggest setting $\lambda = c_\lambda/\sqrt{p_0\theta}$ with $c_\lambda \in [0.5, 0.8]$.

[14] The truncated shift coherence is a stronger condition than natural shift coherence. The statement appears mainly due to the limitation of the proof strategy for the algorithm.

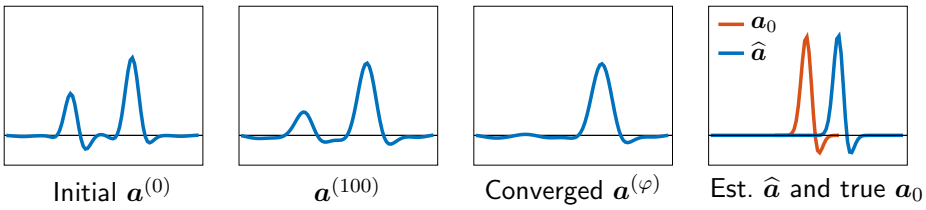

Figure 9. Local minimization and refinement. *Left: data-driven initialization* $\boldsymbol{a}^{(0)}$ *consisting of a near superposition of two shifts. Middle: minimizing* φ_ρ *produces a near shift of* $\boldsymbol{a}_0$. *Right: rounded solution* $\widehat{\boldsymbol{a}}$ *using the Lasso.* $\widehat{\boldsymbol{a}}$ *is very close to a shift of* $\boldsymbol{a}_0$.

When solving SaS deconvolution via minimizing Bilinear Lasso objective (2.2) in practice, the algorithm is analogous to the provable method introduced in Algorithm 3.1, where the curvilinear descent and the refinement step can be realized as alternating gradient descent of both variables $\boldsymbol{a}, \boldsymbol{x}$ in (2.2). Unlike Algorithm 3.1, this alternating gradient method has yet to come with theoretical guarantees, but has shown to be an effective and efficient method for SaS deconvolution problems both in simulation and in reality [34].

3.3. Relationship to the literature. Blind deconvolution is a classical problem in signal processing [54, 12] and has been studied under a variety of hypotheses. In this section, we first discuss the relationship between our results and the existing literature on the SaS version of this problem, and then briefly discuss other deconvolution variants in the theoretical literature.

The SaS model arises in a number of applications. One class of applications involves finding basic motifs (repeated patterns) in datasets. This *motif discovery* problem arises in extracellular spike sorting [37, 20] and calcium imaging [48], where the observed signal exhibits repetitive *short* neuron excitation patterns occurring *sparsely* across time and/or space. Similarly, electron microscopy images [15] arising in study of nanomaterials often exhibit repeated motifs.

Another significant application of SaS deconvolution is *image deblurring*. Typically, the blur kernel is small relative to the image size (*short*) [3, 62, 13, 35, 36]. In natural image deblurring, the target image is often assumed to have relatively few sharp edges [21, 27, 36], and hence have *sparse* derivatives. In scientific image deblurring, e.g., in astronomy [33, 25, 9] and geophysics [28], the target image is often sparse, either in the spatial or wavelet domains, again leading to variants of the SaS model. The literature on blind image deconvolution is large; see, e.g., [31, 10] for surveys.

Variants of the SaS deconvolution problem arise in many other areas of engineering as well. Examples include *blind equalization* in communications [50, 51, 26], *dereverberation* in sound engineering [44, 45], and image *superresolution* [4, 53, 61].

These applications have motivated a great deal of algorithmic work on variants of the SaS problem [32, 8, 6, 31, 43, 10, 56]. In contrast, relatively little theory is available to explain when and why algorithms succeed. Our algorithm minimizes φ_ρ as an approximation to the Lasso cost over the sphere. Our formulation and results have strong precedent in the literature. Lasso-like objective functions have been widely used in image deblurring [62, 14, 21, 35, 52, 60, 18, 30, 36, 59, 47, 64]. A number of insights have been obtained into the

geometry of sparse deconvolution—in particular, into the effect of various constraints on $\boldsymbol{a}$ on the presence or absence of spurious local minimizers. In image deblurring, a simplex constraint ($\boldsymbol{a} \geq \boldsymbol{0}$ and $\|\boldsymbol{a}\|_1 = 1$) arises naturally from the physical structure of the problem [62, 14]. Perhaps surprisingly, simplex-constrained deconvolution admits trivial global minimizers, at which the recovered kernel $\boldsymbol{a}$ is a spike, rather than the target blur kernel [7, 36].

The work [59] imposes the ℓ^2 regularization on $\boldsymbol{a}$ and observes that this alternative constraint gives a more reliable algorithm. In [64], the geometry of the simplified objective φ_{ℓ^1} over the sphere is studied, and it is proved that in the dilute limit in which $\boldsymbol{x}_0$ has one nonzero entry, all strict local minima of φ_{ℓ^1} are close to signed shifts truncations of $\boldsymbol{a}_0$. By adopting a different objective function (based on ℓ^4 maximization) over the sphere, [63] proves that on a certain region of the sphere every local minimum is near a *truncated* signed shift of $\boldsymbol{a}_0$, i.e., the restriction of $s_\ell[\boldsymbol{a}_0]$ to the window $\{0, \ldots, p_0 - 1\}$. The analysis of [63] allows the sparse sequence $\boldsymbol{x}_0$ to be denser ($\theta \sim p_0^{-2/3}$ for a generic kernel $\boldsymbol{a}_0$, as opposed to $\theta \lesssim p_0^{-3/4}$ in our result). Both [64] and [63] guarantee *approximate* recovery of a portion of $s_\ell[\boldsymbol{a}_0]$, under complicated conditions on the kernel $\boldsymbol{a}_0$. Our core optimization problem is very similar to that of [64]. However, we obtain *exact* recovery of both $\boldsymbol{a}_0$ and relatively dense $\boldsymbol{x}_0$ under the much simpler assumption of shift incoherence.

Other aspects of the SaS problem have been studied theoretically. One basic question is under what circumstances the problem is identifiable up to the scaled shift ambiguity. The paper [17] shows that the problem is ill-posed for worst case $(\boldsymbol{a}_0, \boldsymbol{x}_0)$, particularly for certain support patterns in which $\boldsymbol{x}_0$ does not have any isolated nonzero entries. This demonstrates that *some* modeling assumptions on the support of the sparse term are needed. At the same time, this worst-case structure is unlikely to occur, either under the Bernoulli model or in practical deconvolution problems.

Motivated by a variety of applications, much research has focused on low-dimensional deconvolution models in the theoretical literature. In communication applications, the signals $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$ either live in known low-dimensional subspaces or are sparse in some known dictionary [2, 16, 29, 39, 40, 41, 42]. These theoretical works assume that the subspace/dictionary are chosen at random. This low-dimensional deconvolution model does not exhibit the signed shift ambiguity; nonconvex formulations for this model exhibit a different structure from that studied here. In fact, the variant in which both signals belong to known subspaces can be solved by convex relaxation [2]. The SaS model does not appear to be amenable to convexification and exhibits a more complicated nonconvex geometry due to the shift ambiguity. The main motivation for tackling this model lies in the aforementioned applications in imaging and data analysis.

In [38, 57] the related *multi-instance* sparse blind deconvolution problem (MISBD) is studied, where there are K observations $\boldsymbol{y}_i = \boldsymbol{a}_0 * \boldsymbol{x}_i$ consisting of multiple convolutions $i = 1, \ldots, K$ of a kernel $\boldsymbol{a}_0$ and different sparse vectors $\boldsymbol{x}_i$. Both works develop provable algorithms. There are several key differences with our work. First, both the proposed algorithms and their analysis require the kernel to be invertible. Second, despite the apparent similarity between the SaS model and MISBD, these problems are not equivalent. It might seem possible to reduce SaS to MISBD by dividing the single observation $\boldsymbol{y}$ into K pieces; this apparent reduction fails due to boundary effects.

3.4. Notation. All vectors/matrices are written in bold font, $\boldsymbol{a}/\boldsymbol{A}$; indexed values are written as $\boldsymbol{a}_i$, $\boldsymbol{A}_{ij}$. Zero or one vectors are defined as $\mathbf{0}$ or $\mathbf{1}$, and ith canonical basis vector defined as $\boldsymbol{e}_i$. The indices for vectors/matrices all start from 0 and are taken modulo-n, and thus a vector of length n should have its indices labeled as $\{0, 1, \ldots, n-1\}$. We write $[n] = \{0, \ldots, n-1\}$. We often use the capital italic symbols I, J for subsets of $[n]$. We abuse notation slightly and write $[-p] = \{n-p+1, \ldots, n-1, 0\}$ and $[\pm p] = \{n-p+1, \ldots, n-1, 0, 1, \ldots, p-1\}$. Index sets can be labels for vectors; $\boldsymbol{a}_I \in \mathbb{R}^{|I|}$ denotes the restriction of the vector $\boldsymbol{a}$ to coordinates I. Also, we use a check symbol to denote the reversal operator on index set $\check{I} = -I$ and vectors $\check{\boldsymbol{a}}_i = \boldsymbol{a}_{-i}$.

We let $\boldsymbol{P}_C$ denote the projection operator associated with a compact set C. The zero-filling operator $\boldsymbol{\iota}_I : \mathbb{R}^{|I|} \to \mathbb{R}^n$ injects the input vector to higher dimensional Euclidean space, via $(\boldsymbol{\iota}_I \boldsymbol{x})_i = \boldsymbol{x}_{I^{-1}(i)}$ for $i \in I$, and 0 otherwise. Its adjoint operator $\boldsymbol{\iota}_I^*$ can be understood as a subset selection operator which picks up entries of coordinates I. A common zero-filling operator throughout this paper, $\boldsymbol{\iota}$, is an abbreviation of $\boldsymbol{\iota}_{[p]}$, which is often addressed as the zero-padding operator and its adjoint $\boldsymbol{\iota}^*$ as truncation operator.

The convolution operators are all circular with modulo-n: $(\boldsymbol{a} * \boldsymbol{x})_i = \sum_{j\in[n]} \boldsymbol{a}_j \boldsymbol{x}_{i-j}$; also, the convolution operator works on the index set: $I * J = \operatorname{supp}(\mathbf{1}_I * \mathbf{1}_J)$. Similarly, the shift operator $s_\ell[\cdot] : \mathbb{R}^p \to \mathbb{R}^n$ is circular with modulo-n without specification: $(s_\ell[\boldsymbol{a}])_j = (\boldsymbol{\iota}_{[p]}\boldsymbol{a})_{j-\ell}$. Notice that here $\boldsymbol{a}$ can be shorter, $p \le n$. Let $\boldsymbol{C}_{\boldsymbol{a}} \in \mathbb{R}^{n\times n}$ denote a circulant matrix (with modulo-n) for vector $\boldsymbol{a}$, whose jth column is the cyclic shift of $\boldsymbol{a}$ by j: $\boldsymbol{C}_{\boldsymbol{a}}\boldsymbol{e}_j = s_j[\boldsymbol{a}]$. It satisfies for any $b \in \mathbb{R}^n$,

$$\boldsymbol{C}_{\boldsymbol{a}}\boldsymbol{b} = \boldsymbol{a} * \boldsymbol{b}. \tag{3.21}$$

The correlation between $\boldsymbol{a}$ and $\boldsymbol{b}$ can be also written in similar form of convolution operators which reverse one vector before convolution. Define two correlation matrices $\boldsymbol{C}_{\boldsymbol{a}}^*$ and $\check{\boldsymbol{C}}_{\boldsymbol{a}}$ as $\boldsymbol{C}_{\boldsymbol{a}}^*\boldsymbol{e}_j = s_j[\check{\boldsymbol{a}}]$ and $\check{\boldsymbol{C}}_{\boldsymbol{a}}\boldsymbol{e}_j = s_{-j}[\boldsymbol{a}]$. The two operators will satisfy

$$\boldsymbol{C}_{\boldsymbol{a}}^*\boldsymbol{b} = \check{\boldsymbol{a}} * \boldsymbol{b}, \quad \check{\boldsymbol{C}}_{\boldsymbol{a}}\boldsymbol{b} = \boldsymbol{a} * \check{\boldsymbol{b}}. \tag{3.22}$$

4. Geometry of φ_ρ in shift space. Underlying our main geometric and algorithmic results is a relationship between the geometry of the function φ_ρ and the symmetries of the deconvolution problem. In this section, we describe this relationship at a more technical level by interpreting the gradient and Hessian of the function φ_ρ in terms of the shifts $s_\ell[\boldsymbol{a}_0]$ and stating a key lemma which asserts that a certain neighborhood of the union of subspaces $\Sigma_{4\theta p_0}$ can be decomposed into regions of negative curvature, strong gradient, and strong convexity near the target solutions $\pm s_\ell[\boldsymbol{a}_0]$.

4.1. Shifts and correlations. The set $\Sigma_{4\theta p_0}$ is a union of subspaces. Any point $\boldsymbol{a}$ in one of these subspaces $\mathcal{S}_{\boldsymbol{\tau}}$ is a superposition of shifts of $\boldsymbol{a}_0$:

$$\boldsymbol{a} = \sum_{\ell\in\boldsymbol{\tau}} \boldsymbol{\alpha}_\ell s_\ell[\boldsymbol{a}_0]. \tag{4.1}$$

This representation can be extended to a general point $\boldsymbol{a} \in \mathbb{S}^{p-1}$ by writing

$$\boldsymbol{a} = \sum_{\ell\in\boldsymbol{\tau}} \boldsymbol{\alpha}_\ell s_\ell[\boldsymbol{a}_0] + \sum_{\ell\notin\boldsymbol{\tau}} \boldsymbol{\alpha}_\ell s_\ell[\boldsymbol{a}_0]. \tag{4.2}$$

The vector $\boldsymbol{\alpha}$ can be viewed as the coefficients of a decomposition of $\boldsymbol{a}$ into different shifts of $\boldsymbol{a}_0$. This representation is not unique. For $\boldsymbol{a}$ close to $\mathcal{S}_{\boldsymbol{\tau}}$, we can choose a particular $\boldsymbol{\alpha}$ for which $\boldsymbol{\alpha}_{\tau^c}$ is small, a notion that we will formalize below.

For convenience, we introduce a closely related vector $\boldsymbol{\beta} \in \mathbb{R}^n$, whose entries are the inner products between $\boldsymbol{a}$ and the shifts of $\boldsymbol{a}_0$: $\boldsymbol{\beta}_\ell = \langle \boldsymbol{a}, s_\ell[\boldsymbol{a}_0] \rangle$. Since the columns of $\boldsymbol{C}_{\boldsymbol{a}_0}$ are the shifts of $\boldsymbol{a}_0$, we can write

$$\boldsymbol{\beta} = \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{\iota} \boldsymbol{a} \tag{4.3}$$

$$= \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{\iota} \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha} =: \boldsymbol{M} \boldsymbol{\alpha}. \tag{4.4}$$

The matrix $\boldsymbol{M}$ is the Gram matrix of the truncated shifts: $\boldsymbol{M}_{ij} = \langle \boldsymbol{\iota}^* s_i[\boldsymbol{a}_0], \boldsymbol{\iota}^* s_j[\boldsymbol{a}_0] \rangle$. When μ is small, the off-diagonal elements of $\boldsymbol{M}$ are small. In particular, on $\mathcal{S}_\tau$ we may take $\boldsymbol{\alpha}_{\tau^c} = \boldsymbol{0}$, and $\boldsymbol{\beta} \approx \boldsymbol{\alpha}$ in the sense that $\boldsymbol{\beta}_\tau \approx \boldsymbol{\alpha}_\tau$ and the entries of $\boldsymbol{\beta}_{\tau^c}$ are small. For detailed elaboration, see section SM2 in the supplementary material.

4.2. Shifts and the calculus of φ_{ℓ^1}. Our main geometric claims pertain to the function φ_ρ, which is based on a smooth sparsity surrogate $\rho(\cdot) \approx \|\cdot\|_1$. In this section, we sketch the main ideas of the proof as if $\rho(\cdot) = \|\cdot\|_1$ by relating the geometry of the function φ_{ℓ^1} to the vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ introduced above. Working with φ_{ℓ^1} simplifies the exposition; it is also faithful to the structure of our proof, which relates the derivatives of the smooth function φ_ρ to similar quantities associated with the nonsmooth function φ_{ℓ^1}.

The function φ_{ℓ^1} has a relatively simple closed form:

$$\varphi_{\ell^1}(\boldsymbol{a}) = -\tfrac{1}{2} \left\| \mathcal{S}_\lambda \left[\overset{\vee}{\boldsymbol{y}} * \boldsymbol{a} \right] \right\|_2^2. \tag{4.5}$$

Here, $\mathcal{S}_\lambda$ is the *soft thresholding operator*, which is defined for scalars t as

$$\mathcal{S}_\lambda[t] = \operatorname{sign}(t) \max\{|t| - \lambda, 0\} \tag{4.6}$$

and is extended to vectors by applying it elementwise. The operator $\mathcal{S}_\lambda[\boldsymbol{x}]$ shrinks the elements of $\boldsymbol{x}$ toward zero. Small elements become identically zero, resulting in a sparse vector.

Gradient: Sparsifying the correlations β. Our goal is to understand the local minimizers of the function φ_{ℓ^1} over the sphere. The function φ_{ℓ^1} is differentiable. Clearly, any point $\boldsymbol{a}$ at which its gradient (over the sphere) is nonzero cannot be a local minimizer. We first give an expression for the gradient of φ_{ℓ^1} over Euclidean space $\mathbb{R}^p$, and then extend it to the sphere $\mathbb{S}^{p-1}$. Using $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$ and calculus gives

$$\begin{aligned} \nabla \varphi_{\ell^1}(\boldsymbol{a}) &= -\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \overset{\vee}{\boldsymbol{C}}_{\boldsymbol{x}_0} \mathcal{S}_\lambda \left[\overset{\vee}{\boldsymbol{C}}_{\boldsymbol{x}_0} \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{\iota} \boldsymbol{a} \right] \\ &= -\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \overset{\vee}{\boldsymbol{C}}_{\boldsymbol{x}_0} \mathcal{S}_\lambda \left[\overset{\vee}{\boldsymbol{C}}_{\boldsymbol{x}_0} \boldsymbol{\beta} \right] \\ &= -\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\chi}[\boldsymbol{\beta}], \end{aligned} \tag{4.7}$$

where we have simplified the notation by introducing an operator $\boldsymbol{\chi} : \mathbb{R}^n \to \mathbb{R}^n$ as $\boldsymbol{\chi}[\boldsymbol{\beta}] = \overset{\vee}{\boldsymbol{C}}_{\boldsymbol{x}_0} \mathcal{S}_\lambda \left[\overset{\vee}{\boldsymbol{C}}_{\boldsymbol{x}_0} \boldsymbol{\beta} \right]$. This representation exhibits the (negative) gradient as a superposition of shifts of $\boldsymbol{a}_0$ with coefficients given by the entries of $\boldsymbol{\chi}[\boldsymbol{\beta}]$:

$$-\nabla \varphi_{\ell^1}(\boldsymbol{a}) = \sum_\ell \boldsymbol{\chi}[\boldsymbol{\beta}]_\ell \, s_\ell[\boldsymbol{a}_0]. \tag{4.8}$$

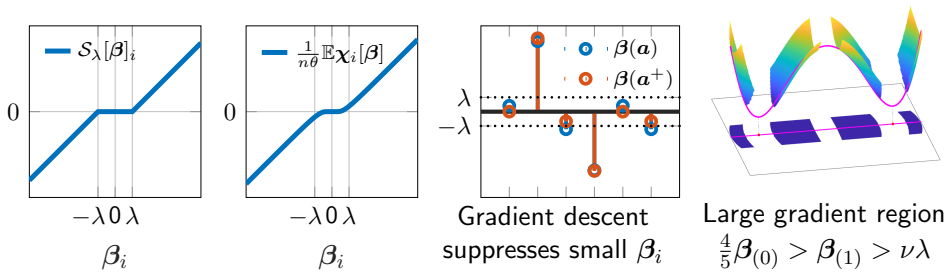

Figure 10. Gradient sparsifies correlations. *Left: the soft thresholding operator $\mathcal{S}_\lambda[\boldsymbol{\beta}]$ shrinks the entries of $\boldsymbol{\beta}$ toward zero, making it sparser. Middle left: the negative gradient $-\nabla\varphi_{\ell^1}$ is a superposition of shifts $s_\ell[\boldsymbol{a}_0]$, with coefficients $\boldsymbol{\chi}_\ell[\boldsymbol{\beta}] \approx \mathcal{S}_\lambda[\boldsymbol{\beta}]_\ell$. Because of this, gradient descent sparsifies $\boldsymbol{\beta}$. Middle right: $\boldsymbol{\beta}(\boldsymbol{a})$ before, and $\boldsymbol{\beta}(\boldsymbol{a}^+)$ after, one projected gradient step $\boldsymbol{a}^+ = \boldsymbol{P}_{\mathbb{S}^{p-1}}[\boldsymbol{a} - t\cdot\operatorname{grad}[\varphi_{\ell^1}](\boldsymbol{a})]$. Notice that the small entries of $\boldsymbol{\beta}$ are shrunk towards zero. Right: the gradient $\operatorname{grad}[\varphi_{\ell^1}](\boldsymbol{a})$ is large whenever it is easy to sparsify $\boldsymbol{\beta}$, particularly when the largest entry $\boldsymbol{\beta}_{(0)} \gg \boldsymbol{\beta}_{(1)} \gg 0$.*

The operator $\boldsymbol{\chi}$ appears complicated. However, its effect is relatively simple: *when $\boldsymbol{x}_0$ is a long random vector, $\boldsymbol{\chi}[\boldsymbol{\beta}]$ acts like a soft thresholding operator on the vector $\boldsymbol{\beta}$.* That is,

$$\frac{1}{n\theta}\cdot\boldsymbol{\chi}[\boldsymbol{\beta}]_\ell \approx \begin{cases} \boldsymbol{\beta}_\ell - \lambda, & \boldsymbol{\beta}_\ell > \lambda, \\ \boldsymbol{\beta}_\ell + \lambda, & \boldsymbol{\beta}_\ell < -\lambda, \\ 0 & \text{otherwise.} \end{cases} \tag{4.9}$$

We show this rigorously below in the proof of our main theorems. Here, we support this claim pictorially by plotting the ℓth entry $\boldsymbol{\chi}[\boldsymbol{\beta}]_\ell$ as $\boldsymbol{\beta}_\ell$ varies; see Figure 10 (middle left) and compare to Figure 10 (left). Because $\boldsymbol{\chi}[\boldsymbol{\beta}]$ suppresses small entries of $\boldsymbol{\beta}$, the strongest contributions to $-\nabla\varphi_{\ell^1}$ in (4.8) will come from shifts $s_\ell[\boldsymbol{a}_0]$ with large $\boldsymbol{\beta}_\ell$. *In particular, the Euclidean gradient is large whenever there is a single preferred shift $s_\ell[\boldsymbol{a}_0]$, i.e., the largest entry of $\boldsymbol{\beta}$ is significantly larger than the second largest entry.*

The (Euclidean) gradient $\nabla\varphi_{\ell^1}$ measures the slope of φ_{ℓ^1} over $\mathbb{R}^n$. We are interested in the slope of φ_{ℓ^1} over the sphere $\mathbb{S}^{p-1}$, which is measured by the Riemannian gradient

$$\begin{aligned} \operatorname{grad}[\varphi_{\ell^1}](\boldsymbol{a}) &= \boldsymbol{P}_{\boldsymbol{a}^\perp}\nabla\varphi_{\ell^1}(\boldsymbol{a}) \\ &= -\boldsymbol{P}_{\boldsymbol{a}^\perp}\sum_\ell \boldsymbol{\chi}_\ell[\boldsymbol{\beta}]\, s_\ell[\boldsymbol{a}_0]. \end{aligned} \tag{4.10}$$

The Riemannian gradient simply projects the Euclidean gradient onto the tangent space $\boldsymbol{a}^\perp$ to $\mathbb{S}^{p-1}$ at $\boldsymbol{a}$. The Riemannian gradient is large whenever

(i) *the negative gradient points to one particular shift*: there is a single preferred shift $s_\ell[\boldsymbol{a}_0]$ so that the Euclidean gradient is large; *and*

(ii) *$\boldsymbol{a}$ is not too close to any shift*: it is possible to move in the tangent space in the direction of this shift.[15] Since the tangent space consists of those vectors orthogonal to $\boldsymbol{a}$, this is possible whenever $s_\ell[\boldsymbol{a}_0]$ is not too aligned with $\boldsymbol{a}$, i.e., $\boldsymbol{a}$ is not too close to $s_\ell[\boldsymbol{a}_0]$.

[15] . . . so the projection of the Euclidean gradient onto the tangent space does not vanish.

Our technical lemma quantifies this situation in terms of the ordered entries of $\boldsymbol{\beta}$. Write $|\boldsymbol{\beta}_{(0)}| \geq |\boldsymbol{\beta}_{(1)}| \geq \cdots$, with corresponding shifts $s_{(0)}[\boldsymbol{a}_0], s_{(1)}[\boldsymbol{a}_0], \ldots$. There is a strong gradient whenever $|\boldsymbol{\beta}_{(0)}|$ is significantly larger than $|\boldsymbol{\beta}_{(1)}|$ and $|\boldsymbol{\beta}_{(1)}|$ is not too small compared to λ: in particular, when $\frac{4}{5}|\boldsymbol{\beta}_{(0)}| > |\boldsymbol{\beta}_{(1)}| > \frac{\lambda}{4\log^2\theta^{-1}}$. In this situation, gradient descent drives $\boldsymbol{a}$ toward $s_{(0)}[\boldsymbol{a}_0]$, reducing $|\boldsymbol{\beta}_{(1)}|, \ldots$, and making the vector $\boldsymbol{\beta}$ sparser. We establish the technical claim that the (Euclidean) gradient of φ_{ℓ^1} sparsifies vectors in shift space in section SM3.

Hessian: Negative curvature breaks symmetry. When there is no single preferred shift, i.e., when $|\boldsymbol{\beta}_{(1)}|$ is close to $|\boldsymbol{\beta}_{(0)}|$, the gradient can be small. Similarly, when $\boldsymbol{a}$ is very close to $\pm s_{(0)}[\boldsymbol{a}_0]$, the gradient can be small. In either of these situations, we need to study the curvature of the function φ to determine whether there are local minimizers.

Strictly speaking, the function φ_{ℓ^1} is not twice differentiable, due to the nonsmoothness of the soft thresholding operator $\mathcal{S}_\lambda[t]$ at $t = \pm\lambda$. Indeed, φ_{ℓ^1} is nonsmooth at any point $\boldsymbol{a}$ for which some entry of $\breve{\boldsymbol{y}} * \boldsymbol{a}$ has magnitude λ. At other points $\boldsymbol{a}$, φ_{ℓ^1} is twice differentiable, and its Hessian is given by

$$\widetilde{\nabla}^2 \varphi_{\ell^1}(\boldsymbol{a}) = -\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \breve{\boldsymbol{C}}_{\boldsymbol{x}_0} \boldsymbol{P}_I \breve{\boldsymbol{C}}_{\boldsymbol{x}_0} \boldsymbol{C}^*_{\boldsymbol{a}_0} \boldsymbol{\iota}, \tag{4.11}$$

with $I = \operatorname{supp}\left(\mathcal{S}_\lambda\left[\breve{\boldsymbol{C}}_{\boldsymbol{y}} \boldsymbol{\iota} \boldsymbol{a}\right]\right)$. We (formally) extend this expression to *every* $\boldsymbol{a} \in \mathbb{R}^n$, terming $\widetilde{\nabla}^2 \varphi_{\ell^1}$ the *pseudo-Hessian* of φ_{ℓ^1}. For appropriately chosen smooth sparsity surrogate ρ, we will see that the (true) Hessian of the smooth function $\nabla^2 \varphi_\rho$ is close to $\widetilde{\nabla}^2 \varphi_{\ell^1}$, and so $\widetilde{\nabla}^2 \varphi_{\ell^1}$ yields useful information about the curvature of φ_ρ.

As with the gradient, the Hessian is complicated, but becomes simpler when the sample size is large. The approximation

$$\widetilde{\nabla}^2 \varphi_{\ell^1}(\boldsymbol{a}) \approx -\sum_\ell s_\ell[\boldsymbol{a}_0] s_\ell[\boldsymbol{a}_0]^* \left(\frac{\partial}{\partial \boldsymbol{\beta}_\ell} \boldsymbol{\chi}_\ell[\boldsymbol{\beta}]\right) \tag{4.12}$$

can be obtained from (4.8) by noting that $\frac{\partial}{\partial \boldsymbol{a}} \boldsymbol{\chi}_\ell[\boldsymbol{\beta}] = \sum_j s_j[\boldsymbol{a}_0] \frac{\partial}{\partial \boldsymbol{\beta}_j} \boldsymbol{\chi}_\ell[\boldsymbol{\beta}]$, that $\frac{\partial}{\partial \boldsymbol{\beta}_j} \boldsymbol{\chi}_\ell[\boldsymbol{\beta}] \approx 0$ for $j \neq \ell$, and that

$$\frac{1}{n\theta} \cdot \frac{\partial \boldsymbol{\chi}_\ell[\boldsymbol{\beta}]}{\partial \boldsymbol{\beta}_\ell} \approx \begin{cases} 0, & |\boldsymbol{\beta}_\ell| \ll \lambda, \\ 1, & |\boldsymbol{\beta}_\ell| \gg \lambda. \end{cases} \tag{4.13}$$

Again, we corroborate this approximation pictorially; see Figure 11.

From this approximation, we can see that the quadratic form $\boldsymbol{v}^* \widetilde{\nabla}^2 \varphi_{\ell^1} \boldsymbol{v}$ takes on a large negative value whenever $\boldsymbol{v}$ is a shift $s_\ell[\boldsymbol{a}_0]$ corresponding to some $|\boldsymbol{\beta}_\ell| \geq \lambda$, or whenever $\boldsymbol{v}$ is a linear combination of such shifts. *In particular, if for some j, $|\boldsymbol{\beta}_{(0)}|, |\boldsymbol{\beta}_{(1)}|, \ldots, |\boldsymbol{\beta}_{(j)}| \gg \lambda$, then φ_{ℓ^1} will exhibit negative curvature in any direction $\boldsymbol{v} \in \operatorname{span}(s_{(0)}[\boldsymbol{a}_0], s_{(1)}[\boldsymbol{a}_0], \ldots, s_{(j)}[\boldsymbol{a}_0])$.*

The (Euclidean) Hessian measures the curvature of the function φ_{ℓ^1} over $\mathbb{R}^n$. The Riemannian Hessian

$$\widetilde{\operatorname{Hess}}[\varphi_{\ell^1}](\boldsymbol{a}) = \boldsymbol{P}_{\boldsymbol{a}^\perp} \left(\underbrace{\widetilde{\nabla}^2 \varphi_{\ell^1}(\boldsymbol{a})}_{\text{Curvature of } \varphi_{\ell^1}} + \underbrace{\langle -\nabla \varphi_{\ell^1}(\boldsymbol{a}), \boldsymbol{a} \rangle \cdot \boldsymbol{I}}_{\text{Curvature of the sphere}} \right) \boldsymbol{P}_{\boldsymbol{a}^\perp} \tag{4.14}$$

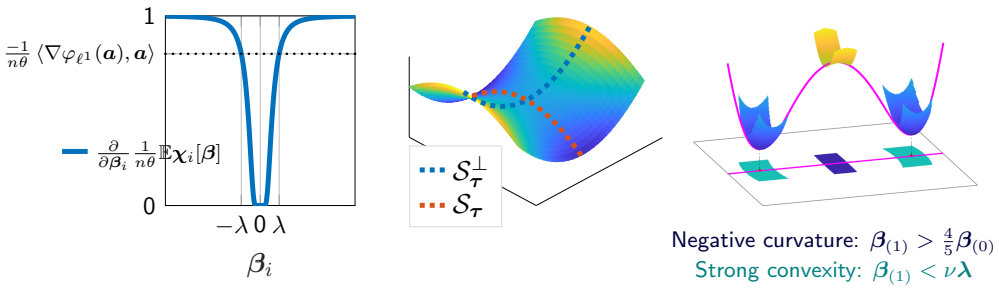

Figure 11. Hessian breaks symmetry. *Left: contribution of* $-s_i[\boldsymbol{a}_0]s_i[\boldsymbol{a}_0]^*$ *to the Euclidean Hessian. If* $|\boldsymbol{\beta}_i| \gg \lambda$*, the Euclidean Hessian exhibits a strong negative component in the* $s_i[\boldsymbol{a}_0]$ *direction. The Riemannian Hessian exhibits negative curvature in directions spanned by* $s_i[\boldsymbol{a}_0]$ *with corresponding* $|\boldsymbol{\beta}_i| \gg \lambda$ *and positive curvature in directions spanned by* $s_i[\boldsymbol{a}_0]$ *with* $|\boldsymbol{\beta}_i| \ll \lambda$*. Middle: this creates negative curvature along the subspace* $\mathcal{S}_{\boldsymbol{\tau}}$ *and positive curvature orthogonal to this subspace. Right: our analysis shows that there is always a direction of negative curvature when* $\boldsymbol{\beta}_{(1)} > \frac{4}{5}\boldsymbol{\beta}_{(0)}$*; conversely, when* $\boldsymbol{\beta}_{(1)} \ll \lambda$ *there is positive curvature in every feasible direction and the function is strongly convex.*

measures the curvature of φ_{ℓ^1} over the sphere. The projection $\boldsymbol{P}_{\boldsymbol{a}^\perp}$ restricts its action to directions $\boldsymbol{v} \perp \boldsymbol{a}$ that are tangent to the sphere. The additional term $\langle -\nabla\varphi_{\ell^1}(\boldsymbol{a}), \boldsymbol{a}\rangle$ accounts for the curvature of the sphere. This term is always positive. The net effect is that directions of strong negative curvature of φ_{ℓ^1} over $\mathbb{R}^n$ become directions of moderate negative curvature over the sphere. Directions of nearly zero curvature over $\mathbb{R}^n$ become directions of positive curvature over the sphere. This has three implications for the geometry of φ_{ℓ^1} over the sphere:

(i) *Negative curvature in symmetry breaking directions*: If $|\boldsymbol{\beta}_{(0)}|, |\boldsymbol{\beta}_{(1)}|, \ldots, |\boldsymbol{\beta}_{(j)}| \gg \lambda$, then φ_{ℓ^1} will exhibit negative curvature in any tangent direction $\boldsymbol{v} \perp \boldsymbol{a}$ which is in the linear span

$$\operatorname{span}(s_{(0)}[\boldsymbol{a}_0], s_{(1)}[\boldsymbol{a}_0], \ldots, s_{(j)}[\boldsymbol{a}_0])$$

of the corresponding shifts of $\boldsymbol{a}_0$.

(ii) *Positive curvature in directions away from* $\mathcal{S}_{\boldsymbol{\tau}}$: The Euclidean Hessian quadratic form $\boldsymbol{v}^*\widetilde{\nabla}^2\varphi_{\ell^1}\boldsymbol{v}$ takes on relatively small values in directions orthogonal to the subspace $\mathcal{S}_{\boldsymbol{\tau}}$. The Riemannian Hessian is positive in these directions, creating positive curvature orthogonal to the subspace $\mathcal{S}_{\boldsymbol{\tau}}$.

(iii) *Strong convexity around minimizers*: Around a minimizer $s_\ell[\boldsymbol{a}_0]$, only a single entry $\boldsymbol{\beta}_\ell$ is large. Any tangent direction $\boldsymbol{v} \perp \boldsymbol{a}$ is nearly orthogonal to the subspace $\operatorname{span}(s_\ell[\boldsymbol{a}_0])$, and hence is a direction of positive (Riemannian) curvature. The objective function φ_ρ is strongly convex around the target solutions $\pm s_\ell[\boldsymbol{a}_0]$.

Figure 11 visualizes these regions of negative and positive curvatures, and the technical claim of positivity/negativity of curvature in shift space is presented in detail in section SM4.

4.3. Any local minimizer is a near shift. We close this section by stating a key theorem, which makes the above discussion precise. We will show that a certain neighborhood of any subspace $\mathcal{S}_{\boldsymbol{\tau}}$ can be covered by regions of *negative curvature*, of *large gradient*, and of *strong convexity* containing target solutions $\pm s_\ell[\boldsymbol{a}_0]$. Furthermore, at the boundary of this neighborhood, the negative gradient points back—*retracts*—toward the subspace $\mathcal{S}_{\boldsymbol{\tau}}$, due to

the (directional) convexity of φ_ρ away from the subspace.

To formally state the result, we need a way of measuring how close $\boldsymbol{a}$ is to the subspace $\mathcal{S}_{\boldsymbol{\tau}}$. For technical reasons, it turns out to be convenient to do this in terms of the coefficients $\boldsymbol{\alpha}$ in the representation

$$\boldsymbol{a} = \sum_{\ell \in \boldsymbol{\tau}} \boldsymbol{\alpha}_\ell s_\ell[\boldsymbol{a}_0] + \sum_{\ell' \in \boldsymbol{\tau}^c} \boldsymbol{\alpha}_{\ell'} s_{\ell'}[\boldsymbol{a}_0]. \tag{4.15}$$

If $\boldsymbol{a} \in \mathcal{S}_{\boldsymbol{\tau}}$, we can take $\boldsymbol{\alpha}$ with $\boldsymbol{\alpha}_{\tau^c} = \mathbf{0}$. We can view the energy $\|\boldsymbol{\alpha}_{\tau^c}\|_2$ as a measure of the distance from $\boldsymbol{a}$ to $\mathcal{S}_{\boldsymbol{\tau}}$. A technical wrinkle arises, because the representation (4.15) is not unique. We resolve this issue by choosing the $\boldsymbol{\alpha}$ that minimizes $\|\boldsymbol{\alpha}_{\tau^c}\|_2$, writing

$$d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) = \inf \left\{ \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \,:\, \sum_\ell \boldsymbol{\alpha}_\ell s_\ell[\boldsymbol{a}_0] = \boldsymbol{a} \right\}. \tag{4.16}$$

The distance $d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}})$ is zero for $\boldsymbol{a} \in \mathcal{S}_{\boldsymbol{\tau}}$. Our analysis controls the geometric properties of φ_ρ over the set of $\boldsymbol{a}$ for which $d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}})$ is not too large. Similar to (3.3), we define an object which contains all points that are close to some $\mathcal{S}_{\boldsymbol{\tau}}$ in the above sense:

$$\Sigma^\gamma_{4\theta p_0} := \bigcup_{|\boldsymbol{\tau}| \le 4\theta p_0} \{\boldsymbol{a} \,:\, d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) \le \gamma\}. \tag{4.17}$$

The aforementioned geometric properties hold over this set.

Theorem 4.1 (geometry of φ_ρ over union of subspaces). *Suppose that $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$, where $\boldsymbol{a}_0 \in \mathbb{S}^{p_0 - 1}$ is μ-shift coherent and $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \text{BG}(\theta) \in \mathbb{R}^n$ satisfying*

$$\theta \in \left[\frac{c'}{p_0}, \frac{c}{p_0 \sqrt{\mu} + \sqrt{p_0}} \right] \cdot \frac{1}{\log^2 p_0} \tag{4.18}$$

for some constants $c', c > 0$. Set $\lambda = 0.1/\sqrt{p_0 \theta}$ in φ_ρ, where $\rho(x) = \sqrt{x^2 + \delta^2}$. There exist numerical constants $C, c'', c''', c_1\text{-}c_4 > 0$ such that if $\delta \le \frac{c'' \lambda \theta^8}{p^2 \log^2 n}$ and $n > C p_0^5 \theta^{-2} \log p_0$, then with probability at least $1 - c'''/n$, for every $\boldsymbol{a} \in \Sigma^\gamma_{4\theta p_0}$, we have the following:

(Negative curvature.) If $|\boldsymbol{\beta}_{(1)}| \ge \nu_1 |\boldsymbol{\beta}_{(0)}|$, then

$$\lambda_{\min}\left(\text{Hess}[\varphi_\rho](\boldsymbol{a})\right) \le -c_1 n \theta \lambda. \tag{4.19}$$

(Large gradient.) If $\nu_1 |\boldsymbol{\beta}_{(0)}| \ge |\boldsymbol{\beta}_{(1)}| \ge \nu_2(\theta)\lambda$, then

$$\|\text{grad}[\varphi_\rho](\boldsymbol{a})\|_2 \ge c_2 n \theta \tfrac{\lambda^2}{\log^2 \theta^{-1}}. \tag{4.20}$$

(Convex near shifts.) If $\nu_2(\theta)\lambda \ge |\boldsymbol{\beta}_{(1)}|$, then

$$\text{Hess}[\varphi_\rho](\boldsymbol{a}) \succ c_3 n \theta \boldsymbol{P}_{\boldsymbol{a}^\perp}. \tag{4.21}$$

(Retraction to subspace.) If $\frac{\gamma}{2} \le d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) \le \gamma$, then for every $\boldsymbol{\alpha}$ satisfying $\boldsymbol{a} = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}$, there exists $\boldsymbol{\zeta}$ satisfying $\text{grad}[\varphi_\rho](\boldsymbol{a}) = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\zeta}$, such that

$$\langle \boldsymbol{\zeta}_{\boldsymbol{\tau}^c}, \boldsymbol{\alpha}_{\boldsymbol{\tau}^c} \rangle \ge c_4 \|\boldsymbol{\zeta}_{\boldsymbol{\tau}^c}\|_2 \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2. \tag{4.22}$$

(Local minimizers.) If $\boldsymbol{a}$ *is a local minimizer,*

$$\min_{\substack{\ell\in[\pm p]\\ \sigma\in\{\pm 1\}}} \|\boldsymbol{a}-\sigma\, s_\ell[\boldsymbol{a}_0]\|_2 \;\le\; \tfrac{1}{2}\max\left\{\mu, p_0^{-1}\right\}, \tag{4.23}$$

where $\nu_1=\frac{4}{5}$, $\nu_2(\theta)=\frac{1}{4\log^2\theta^{-1}}$, *and* $\gamma=\frac{c\cdot\mathrm{poly}(\sqrt{1/\theta},\sqrt{1/\mu})}{\log^2\theta^{-1}}\cdot\frac{1}{\sqrt{p_0}}$.

Proof. See subsection SM6.5. ∎

The retraction property elaborated upon in (4.22) implies that the negative gradient at $\boldsymbol{a}$ points in a direction that decreases $d_\alpha(\boldsymbol{a},\mathcal{S}_{\boldsymbol{\tau}})$. This is a consequence of positive curvature away from $\mathcal{S}_{\boldsymbol{\tau}}$. It essentially implies that the gradient is monotone in $\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}$ space: choose any $\underline{\boldsymbol{a}}\in\mathcal{S}_{\boldsymbol{\tau}}\cap\mathbb{S}^{p-1}$, write $\underline{\boldsymbol{\alpha}}$ to be its coefficient, and let $\underline{\boldsymbol{\zeta}}$ be the coefficient of $\mathrm{grad}[\varphi_\rho](\underline{\boldsymbol{a}})$. Then $\underline{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^c}=\boldsymbol{0}$, $\underline{\boldsymbol{\zeta}}_{\boldsymbol{\tau}^c}\approx\boldsymbol{0}$, and

$$\langle\boldsymbol{\zeta}_{\boldsymbol{\tau}^c}-\underline{\boldsymbol{\zeta}}_{\boldsymbol{\tau}^c},\,\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}-\underline{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^c}\rangle\approx\langle\boldsymbol{\zeta}_{\boldsymbol{\tau}^c}-\boldsymbol{0},\,\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}-\boldsymbol{0}\rangle=\langle\boldsymbol{\zeta}_{\boldsymbol{\tau}^c},\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\rangle>0.$$

Our main geometric claim in Theorem 3.1 is a direct consequence of Theorem 4.1. Moreover, it suggests that as long as we can minimize φ_ρ within the region $\Sigma^\gamma_{4\theta p_0}$, we will solve the SaS deconvolution problem.

5. Provable algorithm.

In light of Theorem 4.1, in this section we introduce a two-part algorithm, Algorithm 3.1, which first applies the curvilinear descent method to find a local minimum of φ_ρ within $\Sigma^\gamma_{4\theta p_0}$, followed by a refinement algorithm that uses alternating minimization to exactly recover the ground truth. This algorithm exactly solves the SaS deconvolution problem.

5.1. Minimization.

There are three major issues in finding a local minimizer within $\Sigma^\gamma_{4\theta p_0}$:

(i) *Initialization.* The initializer $\boldsymbol{a}^{(0)}$ to reside within $\Sigma^\gamma_{4\theta p_0}$.
(ii) *Negative curvature.* The method to avoid stagnating near saddle points of φ_ρ.
(iii) *No exit.* The descent method to remain inside $\Sigma^\gamma_{4\theta p_0}$.

In the following paragraphs, we describe how our proposed algorithm achieves the above desiderata.

Initialization within $\Sigma^\gamma_{4\theta p_0}$. Our data-driven initialization scheme produces $\boldsymbol{a}^{(0)}$, where

$$\begin{aligned}\boldsymbol{a}^{(0)}&=-\boldsymbol{P}_{\mathbb{S}^{p-1}}\nabla\varphi_\rho\left(\boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\boldsymbol{0}^{p_0-1};\boldsymbol{y}_0;\cdots;\boldsymbol{y}_{p_0-1};\boldsymbol{0}^{p_0-1}\right]\right)\\&=-\boldsymbol{P}_{\mathbb{S}^{p-1}}\nabla\varphi_\rho\boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\boldsymbol{P}_{[p_0]}(\boldsymbol{a}_0*\boldsymbol{x}_0)\right]\\&\approx-\boldsymbol{P}_{\mathbb{S}^{p-1}}\nabla\varphi_\rho\left[\boldsymbol{P}_{[p_0]}(\boldsymbol{a}_0*\widetilde{\boldsymbol{x}}_0)\right]\end{aligned}$$

is the normalized gradient vector from a chunk of data $\boldsymbol{a}^{(-1)}:=\boldsymbol{P}_{[p_0]}(\boldsymbol{a}_0*\widetilde{\boldsymbol{x}}_0)$ with $\widetilde{\boldsymbol{x}}_0$ a normalized Bernoulli–Gaussian random vector of length $2p_0-1$. Since $\nabla\varphi_\rho\approx\nabla\varphi_{\ell^1}$, expanding the gradient $\nabla\varphi_{\ell^1}$ and rewriting the gradient $\nabla_{\ell^1}(\boldsymbol{a}^{(-1)})$ in shift space gives us

$$\begin{aligned}-\nabla\varphi_{\rho^1}(\boldsymbol{a}^{(-1)})&\approx\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\widetilde{\boldsymbol{C}}_{\boldsymbol{x}_0}\mathcal{S}_\lambda\left[\widetilde{\boldsymbol{C}}_{\boldsymbol{x}_0}\boldsymbol{C}^*_{\boldsymbol{a}_0}\boldsymbol{P}_{[p_0]}(\boldsymbol{a}_0*\widetilde{\boldsymbol{x}}_0)\right]\\&=\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{\chi}\left[\boldsymbol{C}^*_{\boldsymbol{a}_0}\boldsymbol{P}_{[p_0]}\boldsymbol{C}_{\boldsymbol{a}_0}\widetilde{\boldsymbol{x}}_0\right]\\&\approx\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{\chi}\left[\widetilde{\boldsymbol{x}}_0\right]\\&\approx n\theta\cdot\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\mathcal{S}_\lambda\left[\widetilde{\boldsymbol{x}}_0\right],\end{aligned}$$

where the approximation in the third equation is accurate if the truncated shifts are incoherent:

$$\max_{i\neq j} \left|\left\langle \boldsymbol{\iota}_{p_0}^* s_i[\boldsymbol{a}_0], \boldsymbol{\iota}_{p_0}^* s_j[\boldsymbol{a}_0]\right\rangle\right| \leq \mu \ll 1. \tag{5.1}$$

With this simple approximation, it becomes clear that the coefficients (in shift space) of initializer $\boldsymbol{a}^{(0)}$,

$$\boldsymbol{a}^{(0)} \approx \boldsymbol{P}_{\mathbb{S}^{p-1}} \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \mathcal{S}_\lambda\left[\widetilde{\boldsymbol{x}}_0\right], \tag{5.2}$$

approximate $\mathcal{S}_\lambda\left[\widetilde{\boldsymbol{x}}_0\right]$, which resides near the subspace $\mathcal{S}_{\boldsymbol{\tau}}$, in which $\boldsymbol{\tau}$ contains the nonzero entries of $\widetilde{\boldsymbol{x}}_0$ on $\{-p_0+1,\dots,p_0-1\}$. With high probability, the number of nonzero entries is $|\boldsymbol{\tau}| \lessapprox 4\theta p_0$, and we therefore conclude that our initializer $\boldsymbol{a}^{(0)}$ satisfies

$$\boldsymbol{a}^{(0)} \in \Sigma^{\gamma}_{4\theta p_0}. \tag{5.3}$$

Furthermore, since $\widetilde{\boldsymbol{x}}_0$ is normalized, the largest magnitude for entries of $|\widetilde{\boldsymbol{x}}_0|$ is likely to be around $1/\sqrt{2p_0\theta}$. To ensure that $\mathcal{S}_\lambda\left[\widetilde{\boldsymbol{x}}_0\right]$ does not annihilate all nonzero entries of $\widetilde{\boldsymbol{x}}_0$ (otherwise our initializer $\boldsymbol{a}^{(0)}$ will become $\boldsymbol{0}$), the ideal λ should be slightly less than the largest magnitude of $|\widetilde{\boldsymbol{x}}_0|$. We suggest setting λ in φ_ρ as

$$\lambda = \frac{c}{\sqrt{p_0\theta}} \tag{5.4}$$

for some $c \in (0,1)$.

Many methods have been proposed to optimize functions whose saddle points exhibit strict negative curvature, including the noisy gradient method [22], trust region methods [1, 55], and curvilinear search [58]. Any of the above methods can be adapted to minimize φ_ρ. In this paper, we use the *curvilinear method with restricted stepsize* to demonstrate how to analyze an optimization problem using the geometric properties of φ_ρ over $\Sigma^{\gamma}_{4\theta p_0}$—in particular, negative curvature in symmetry breaking directions and positive curvature away from $\mathcal{S}_{\boldsymbol{\tau}}$.

Curvilinear search uses an update strategy that combines the gradient $\boldsymbol{g}$ and a direction of negative curvature $\boldsymbol{v}$, which here we choose as an eigenvector of the Hessian $\boldsymbol{H}$ with smallest eigenvalue, scaled such that $\boldsymbol{v}^*\boldsymbol{g} \geq 0$. In particular, we set

$$\boldsymbol{a}^+ \leftarrow \boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\boldsymbol{a} - t\boldsymbol{g} - t^2\boldsymbol{v}\right]. \tag{5.5}$$

For small t,

$$\varphi(\boldsymbol{a}^+) \approx \varphi(\boldsymbol{a}) + \langle \boldsymbol{g}, \boldsymbol{\xi}\rangle + \tfrac{1}{2}\boldsymbol{\xi}^*\boldsymbol{H}\boldsymbol{\xi}. \tag{5.6}$$

Since $\boldsymbol{\xi}$ converges to $\boldsymbol{0}$ only if $\boldsymbol{a}$ converges to the local minimizer (otherwise either gradient $\boldsymbol{g}$ is nonzero or there is a negative curvature direction $\boldsymbol{v}$), this iteration produces a local minimizer for φ_ρ, whose saddle points near any $\mathcal{S}_{\boldsymbol{\tau}}$ have negative curvature, we just need to ensure all iterates stay near some such subspace. We prove this by showing the following:

- When $d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) \leq \gamma$, curvilinear steps move a small distance away from the subspace:

$$\left|d_\alpha\left(\boldsymbol{a}^+, \mathcal{S}_{\boldsymbol{\tau}}\right) - d_\alpha\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right)\right| \leq \tfrac{\gamma}{2}. \tag{5.7}$$

- When $d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) \in \left[\frac{\gamma}{2}, \gamma\right]$, curvilinear steps retract toward subspace:

$$d_\alpha\left(\boldsymbol{a}^+, \mathcal{S}_{\boldsymbol{\tau}}\right) \leq d_\alpha\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right). \tag{5.8}$$

Together, we can prove that the iterates $\boldsymbol{a}^{(k)}$ converge to a minimizer, and

$$\forall\, k = 1, 2, \ldots, \quad \boldsymbol{a}^{(k)} \in \Sigma^{\gamma}_{4\theta p_0}. \tag{5.9}$$

We conclude this section with the following theorem.

Theorem 5.1 (convergence of retractive curvilinear search). *Suppose signals $\boldsymbol{a}_0, \boldsymbol{x}_0$ satisfy the conditions of Theorem 4.1, $\theta > 10^3 c/p_0$ $(c > 1)$, and $\boldsymbol{a}_0$ is μ-truncated shift coherent $\max_{i \neq j} \left|\left\langle \boldsymbol{\iota}^*_{p_0} s_i[\boldsymbol{a}_0], \boldsymbol{\iota}^*_{p_0} s_j[\boldsymbol{a}_0] \right\rangle\right| \leq \mu$. Write $\boldsymbol{g} = \operatorname{grad}[\varphi_\rho](\boldsymbol{a})$ and $\boldsymbol{H} = \operatorname{Hess}[\varphi_\rho](\boldsymbol{a})$. When the smallest eigenvalue of $\boldsymbol{H}$ is strictly smaller than $-\eta_v$, let $\boldsymbol{v}$ be the unit eigenvector of smallest eigenvalue, scaled so $\boldsymbol{v}^* \boldsymbol{g} \geq 0$; otherwise let $\boldsymbol{v} = \boldsymbol{0}$. Define a sequence $\left\{\boldsymbol{a}^{(k)}\right\}_{k \in \mathbb{N}}$ where $\boldsymbol{a}^{(0)}$ equals (3.7) and for $k = 1, 2, \ldots, K_1$*

$$\boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\boldsymbol{a}^{(k)} - t\boldsymbol{g}^{(k)} - t^2 \boldsymbol{v}^{(k)}\right], \tag{5.10}$$

with largest $t \in \left(0, \frac{0.1}{n\theta}\right]$ satisfying Armijo steplength

$$\varphi_\rho(\boldsymbol{a}^{(k+1)}) < \varphi_\rho(\boldsymbol{a}^{(k)}) - \tfrac{1}{2}\left(t\|\boldsymbol{g}^{(k)}\|_2^2 + \tfrac{1}{2} t^4 \eta_v \|\boldsymbol{v}^{(k)}\|_2^2\right). \tag{5.11}$$

Then with probability at least $1 - 1/c$, there exists some signed shift $\bar{\boldsymbol{a}} = \pm s_i[\boldsymbol{a}_0]$, where $i \in [\pm p_0]$, such that $\left\|\boldsymbol{a}^{(k)} - \bar{\boldsymbol{a}}\right\|_2 \leq \mu + 1/p$ for all $k \geq K_1 = \operatorname{poly}(n, p)$. Here, $\eta_v = c' n\theta\lambda$ for some $c' < c_1$ in Theorem 4.1.

Proof. See subsection SM7.2. ∎

5.2. Local refinement. In this section, we describe and analyze an algorithm which refines an estimate $\bar{\boldsymbol{a}} \approx \boldsymbol{a}_0$ of the kernel to exactly recover $(\boldsymbol{a}_0, \boldsymbol{x}_0)$. Set

$$\boldsymbol{a}^{(0)} \leftarrow \bar{\boldsymbol{a}}, \qquad \lambda^{(0)} \leftarrow C(p\theta + \log n)(\mu + 1/p), \qquad I^{(0)} \leftarrow \operatorname{supp}(\mathcal{S}_\lambda\left[\boldsymbol{C}^*_{\bar{\boldsymbol{a}}} \boldsymbol{y}\right]). \tag{5.12}$$

We alternatively minimize the Lasso objective with respect to $\boldsymbol{a}$ and $\boldsymbol{x}$:

$$\boldsymbol{x}^{(k+1)} \leftarrow \operatorname*{argmin}_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{a}^{(k)} * \boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda^{(k)} \sum_{i \notin I^{(k)}} |\boldsymbol{x}_i|, \tag{5.13}$$

$$\boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\operatorname*{argmin}_{\boldsymbol{a}} \tfrac{1}{2}\|\boldsymbol{a} * \boldsymbol{x}^{(k+1)} - \boldsymbol{y}\|_2^2\right], \tag{5.14}$$

$$\lambda^{(k+1)} \leftarrow \tfrac{1}{2}\lambda^{(k)}, \qquad I^{(k+1)} \leftarrow \operatorname{supp}\left(\boldsymbol{x}^{(k+1)}\right). \tag{5.15}$$

One departure from standard alternating minimization procedures is our use of a continuation method, which (i) decreases λ, and (ii) maintains a running estimate $I^{(k)}$ of the support set. Our analysis will show that $\boldsymbol{a}^{(k)}$ converges to one of the signed shifts of $\boldsymbol{a}_0$ at a linear rate, in the sense that

$$\min_{\sigma \in \pm 1,\, \ell \in [\pm p_0]} \left\|\boldsymbol{a}^{(k)} - \sigma \cdot s_\ell[\boldsymbol{a}_0]\right\|_2 \leq C' 2^{-k}. \tag{5.16}$$

It should be clear that exact recovery is unlikely if $\boldsymbol{x}_0$ contains many consecutive nonzero entries: in fact in this situation, even *nonblind* deconvolution fails. Therefore to obtain exact recovery it is necessary to put an upper bound on signal dimension n. Here, we introduce the notation κ_I as an upper bound for the number of nonzero entries of $\boldsymbol{x}_0$ in a length-p window:

$$\kappa_I := 6 \max\{\theta p, \log n\}, \tag{5.17}$$

where the indexing and addition should be interpreted modulo n. We will denote the support sets of true sparse vector $\boldsymbol{x}_0$ and recovered $\boldsymbol{x}^{(k)}$ in the intermediate kth steps as

$$I = \operatorname{supp}(\boldsymbol{x}_0), \qquad I^{(k)} = \operatorname{supp}(\boldsymbol{x}^{(k)}). \tag{5.18}$$

Then in the Bernoulli–Gaussian model, with high probability,

$$\max_{\ell} \left| I \cap ([p] + \ell) \right| \le \kappa_I. \tag{5.19}$$

The $\log n$ term reflects the fact that as n becomes enormous (exponential in p), eventually it becomes likely that some length-p window of $\boldsymbol{x}_0$ is densely occupied. In our main theorem statement, we preclude this possibility by putting an upper bound on signal length n with respect to window length p and shift coherence μ. We will assume

$$(\mu + 1/p) \cdot \kappa_I^2 < c \tag{5.20}$$

for some numerical constant $c \in (0, 1)$.

Recall that (4.23) in Theorem 3.1 provides that

$$\|\bar{\boldsymbol{a}} - \boldsymbol{a}_0\|_2 \le (\mu + 1/p), \tag{5.21}$$

which is sufficiently close to $\boldsymbol{a}_0$ as long as (5.19) holds true. Here, we will elaborate upon this by showing that a single iteration of alternating minimization algorithm (5.13)–(5.15) is a contraction mapping for $\boldsymbol{a}$ toward $\boldsymbol{a}_0$.

To this end, at kth iteration, write $T = I^{(k)}$, $J = I^{(k+1)}$, and $\boldsymbol{\sigma}^{(k)} = \operatorname{sign}\left(\boldsymbol{x}^{(k)}\right)$; then first observe that the solution to the reweighted Lasso problem (5.13) can be written as

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{\iota}_J \left(\boldsymbol{\iota}_J^* \boldsymbol{C}_{\boldsymbol{a}^{(k)}}^* \boldsymbol{C}_{\boldsymbol{a}^{(k)}} \boldsymbol{\iota}_J \right)^{-1} \boldsymbol{\iota}_J^* \left(\boldsymbol{C}_{\boldsymbol{a}^{(k)}}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{x}_0 - \lambda^{(k)} \boldsymbol{P}_{J \setminus T} \boldsymbol{\sigma}^{(k+1)} \right), \tag{5.22}$$

and the solution to least squares problem (5.14) will be

$$\boldsymbol{a}^{(k+1)} = \left(\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}^{(k+1)}}^* \boldsymbol{C}_{\boldsymbol{x}^{(k+1)}} \boldsymbol{\iota} \right)^{-1} \left(\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}^{(k+1)}}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota} \boldsymbol{a}_0 \right). \tag{5.23}$$

Here, we are going to illustrate the relationship between $\boldsymbol{a}^{(k+1)} - \boldsymbol{a}_0$ and $\boldsymbol{a}^{(k)} - \boldsymbol{a}_0$ using simple approximations. First, let us assume that $\boldsymbol{a}^{(k)} \approx \boldsymbol{a}_0$, $\boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \approx \boldsymbol{I}$, and $I \approx J \approx T$. Then (5.22) gives

$$\boldsymbol{x}^{(k+1)} \approx \boldsymbol{x}_0, \tag{5.24}$$

$$\begin{aligned} (\boldsymbol{x}^{(k+1)} - \boldsymbol{x}_0) &\approx \boldsymbol{P}_I \left(\boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{x}_0 - \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{a}^{(k)}} \boldsymbol{x}_0 \right) \\ &\approx \boldsymbol{P}_I \left[\boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota} (\boldsymbol{a}_0 - \boldsymbol{a}^{(k)}) \right], \end{aligned} \tag{5.25}$$

which implies, while assuming $\boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{x}_0} \approx n\theta \boldsymbol{I}$, that from (5.23),

$$\begin{aligned}(\boldsymbol{a}^{(k+1)} - \boldsymbol{a}_0) &\approx (n\theta)^{-1} \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}^{(k+1)}}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota} \boldsymbol{a}_0 - \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}^{(k+1)}}^* \boldsymbol{C}_{\boldsymbol{x}^{(k+1)}} \boldsymbol{\iota} \boldsymbol{a}_0 \\ &\approx (n\theta)^{-1} \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} (\boldsymbol{x}_0 - \boldsymbol{x}^{(k+1)}) \\ &\approx (n\theta)^{-1} \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota} \, (\boldsymbol{a}^{(k)} - \boldsymbol{a}_0). \end{aligned} \tag{5.26}$$

Now since $\boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{x}_0} \approx n\theta \, \boldsymbol{e}_0 \boldsymbol{e}_0^*$, this suggests that $(n\theta)^{-1} \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota}$ approximates a contraction mapping with fixed point $\boldsymbol{a}_0$, as follows:

$$\begin{aligned}(n\theta)^{-1} \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota} &\approx \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{e}_0 \boldsymbol{e}_0^* \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{\iota} \\ &\approx \boldsymbol{a}_0 \boldsymbol{a}_0^*. \end{aligned} \tag{5.27}$$

Hence, if we can ensure all of the above approximation is sufficiently and increasingly accurate as the iterate proceeds, the alternating minimization essentially is a power method which finds the leading eigenvector of matrix $\boldsymbol{a}_0 \boldsymbol{a}_0^*$—and the solution to this algorithm is apparently $\boldsymbol{a}_0$. Indeed, we prove that the iterates produced by this sequence of operations converge to the ground truth at a linear rate, as long as it is initialized sufficiently nearby.

Theorem 5.2 (linear rate convergence of alternating minimization). *Suppose $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$, where $\boldsymbol{a}_0$ is μ-shift coherent and $\boldsymbol{x}_0 \sim \mathrm{BG}(\theta)$. Then there exist some constants C, c, c_μ such that if $(\mu + 1/p)\, \kappa_I^2 < c_\mu$ and $n > C\theta^{-2} p^2 \log n$, then with probability at least $1 - c/n$, for any starting point $\boldsymbol{a}^{(0)}$ and $\lambda^{(0)}$, $I^{(0)}$ such that*

$$\left\| \boldsymbol{a}^{(0)} - \boldsymbol{a}_0 \right\|_2 \le \mu + 1/p, \qquad \lambda^{(0)} = 5\kappa_I(\mu + 1/p), \qquad I^{(0)} = \mathrm{supp}\left(\boldsymbol{C}_{\boldsymbol{a}^{(0)}}^* \boldsymbol{y} \right), \tag{5.28}$$

and for $k = 1, 2, \ldots,$

$$\boldsymbol{x}^{(k+1)} \leftarrow \operatorname*{argmin}_{\boldsymbol{x}} \tfrac{1}{2} \|\boldsymbol{a}^{(k)} * \boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda^{(k)} \sum_{i \notin I^{(k)}} |\boldsymbol{x}_i| \,, \tag{5.29}$$

$$\boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p-1}} \big[\operatorname*{argmin}_{\boldsymbol{a}} \tfrac{1}{2} \|\boldsymbol{a} * \boldsymbol{x}^{(k+1)} - \boldsymbol{y}\|_2^2 \big], \tag{5.30}$$

$$\lambda^{(k+1)} \leftarrow \tfrac{1}{2} \lambda^{(k)}, \qquad I^{(k+1)} \leftarrow \mathrm{supp}\left(\boldsymbol{x}^{(k+1)} \right), \tag{5.31}$$

then

$$\left\| \boldsymbol{a}^{(k+1)} - \boldsymbol{a}_0 \right\|_2 \le (\mu + 1/p) 2^{-k} \tag{5.32}$$

for every $k = 0, 1, 2, \ldots$.

Proof. See subsection SM8.3. ∎

Remark 5.3. The estimates $\boldsymbol{x}^{(k)}$ also converges to the ground truth $\boldsymbol{x}_0$ at a linear rate.

6. Experiments. We demonstrate that the tradeoffs between the motif length p_0 and sparsity rate θ produce a transition region for successful SaS deconvolution under generic choices of $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$. For fixed values of $\theta \in [10^{-3}, 10^{-2}]$ and $p_0 \in [10^3, 10^4]$, we draw 50 instances of synthetic data by choosing $\boldsymbol{a}_0 \sim \mathrm{Unif}(\mathbb{S}^{p_0 - 1})$ and $\boldsymbol{x}_0 \in \mathbb{R}^n$ with $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$, where $n = 5 \times 10^5$. Note that choosing $\boldsymbol{a}_0$ this way implies $\mu(\boldsymbol{a}_0) \approx \frac{1}{\sqrt{p_0}}$.

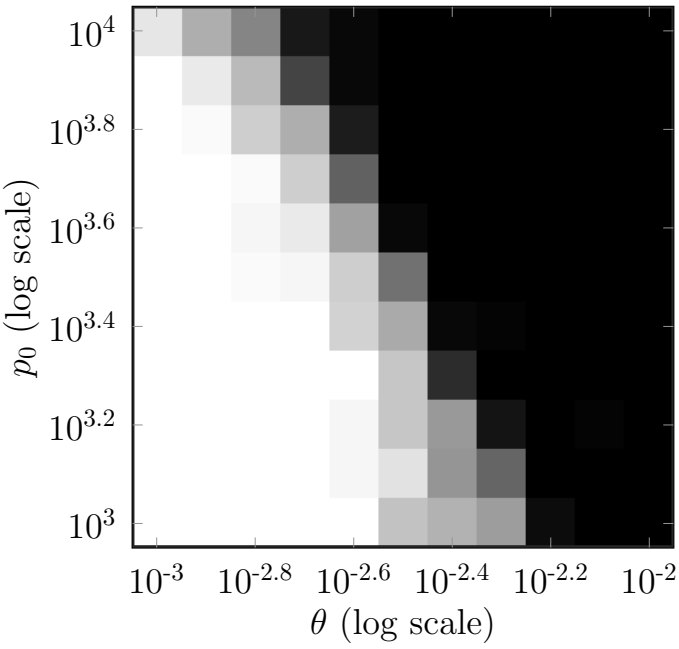

Figure 12. Success probability of SaS deconvolution under generic $\boldsymbol{a}_0$, $\boldsymbol{x}_0$ with varying kernel length p_0 and sparsity rate θ. *When sparsity rate decreases sufficiently with respect to kernel length, successful recovery becomes very likely (brighter), and vice versa (darker). A transition line is shown with slope* $\frac{\log p_0}{\log \theta} \approx -2$, *implying Algorithm 6.1 works with high probability when* $\theta \lessapprox \frac{1}{\sqrt{p_0}}$ *in the generic case.*

For each instance, we recover $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$ from $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$ by minimizing problem (2.5). For ease of computation, we modify Algorithm 3.1 by replacing curvilinear search with the *accelerated Riemannian gradient descent* method (Algorithm 6.1), which is an adaptation of accelerated gradient descent [5] to the sphere. In particular, we apply momentum and increment by the Riemannian gradient via the exponential and logarithmic operators

$$\mathrm{Exp}_{\boldsymbol{a}}(\boldsymbol{u}) \;:=\; \cos(\|\boldsymbol{u}\|_2)\cdot \boldsymbol{a} + \sin(\|\boldsymbol{u}\|_2)\cdot \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2}, \tag{6.1}$$

$$\mathrm{Log}_{\boldsymbol{a}}(\boldsymbol{b}) \;:=\; \arccos(\langle \boldsymbol{a}, \boldsymbol{b}\rangle)\cdot \frac{\boldsymbol{P}_{\boldsymbol{a}^\perp}(\boldsymbol{b}-\boldsymbol{a})}{\|\boldsymbol{P}_{\boldsymbol{a}^\perp}(\boldsymbol{b}-\boldsymbol{a})\|_2}, \tag{6.2}$$

derived from [1]. Here $\mathrm{Exp}_{\boldsymbol{a}} : \boldsymbol{a}^\perp \to \mathbb{S}^{p-1}$ takes a tangent vector of $\boldsymbol{a}$ and produces a new point on the sphere, whereas $\mathrm{Log}_{\boldsymbol{a}} : \mathbb{S}^{p-1} \to \boldsymbol{a}^\perp$ takes a point $\boldsymbol{b} \in \mathbb{S}^{p-1}$ and returns the tangent vector which points from $\boldsymbol{a}$ to $\boldsymbol{b}$.

For each recovery instance, we say the local minimizer $\boldsymbol{a}_{\min}$ generated from Algorithm 6.1 is sufficiently close to a solution of the SaS deconvolution problem if

$$\mathrm{success}(\boldsymbol{a}_{\min}, ; \boldsymbol{a}_0) \;:=\; \{\, \max_\ell |\langle s_\ell[\boldsymbol{a}_0], \boldsymbol{a}_{\min}\rangle| > 0.95 \,\}. \tag{6.3}$$

The result is shown in Figure 12. Our source code can be accessed via the following address:

https://github.com/sbdsphere/sbd_experiments.git

Algorithm 6.1. SaS deconvolution with accelerated Riemannian gradient descent.

Input: Observation $\boldsymbol{y}$, sparsity penalty $\lambda = 0.5/\sqrt{p_0\theta}$, momentum parameter $\eta \in [0, 1)$.
Initialize $\boldsymbol{a}^{(0)} \leftarrow -\boldsymbol{P}_{\mathbb{S}^{p-1}} \nabla \varphi_\rho \left(\boldsymbol{P}_{\mathbb{S}^{p-1}} \left[\boldsymbol{0}^{p_0-1}; [\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{p_0-1}]; \boldsymbol{0}^{p_0-1} \right] \right)$,
for $k = 1, 2, \ldots, K$ **do**
Get momentum: $\boldsymbol{w} \leftarrow \mathrm{Exp}_{\boldsymbol{a}^{(k)}} \left(\eta \cdot \mathrm{Log}_{\boldsymbol{a}^{(k-1)}} (\boldsymbol{a}^{(k)}) \right)$.
Get negative gradient direction: $\boldsymbol{g} \leftarrow -\mathrm{grad}[\varphi_\rho](\boldsymbol{w})$.
Armijo step $\boldsymbol{a}^{(k+1)} \leftarrow \mathrm{Exp}_{\boldsymbol{w}}(t\boldsymbol{g})$, choosing $t \in (0, 1)$ s.t. $\varphi_\rho(\boldsymbol{a}^{(k+1)}) - \varphi_\rho(\boldsymbol{w}) < -t \left\| \boldsymbol{g} \right\|_2^2$.
end for
Output: Return $\boldsymbol{a}^{(K)}$.

7. Discussion.

In this section, we close by discussing the most important limitations of our results when $\boldsymbol{a}_0$ is coherent, regarding scenarios when the signal setting breaches our assumption, especially when $\boldsymbol{x}_0$ is either highly sparse or nonsymmetric, and highlighting corresponding directions for future work.

The main drawback of our proposed method is that it does not succeed when the target motif $\boldsymbol{a}_0$ has shift coherence very close to 1. For instance, a common scenario in image blind deconvolution involves deblurring an image with a smooth, low-pass point spread function (e.g., Gaussian blur). Both our analysis and numerical experiments show that in this situation minimizing φ_ρ does not find the generating signal pairs $(\boldsymbol{a}_0, \boldsymbol{x}_0)$ consistently—the minimizer of φ_ρ is often spurious and is not close to any particular shift of $\boldsymbol{a}_0$. We do not suggest minimizing φ_ρ in this situation. On the other hand, minimizing the Bilinear Lasso objective φ_{lasso} over the sphere often succeeds even if the true signal pair $(\boldsymbol{a}_0, \boldsymbol{x}_0)$ is coherent and dense.

In light of the above observations, we view the analysis of the Bilinear Lasso as the most important direction for future theoretical work on SaS deconvolution. The drop quadratic formulation studied here has commonalities with the Bilinear Lasso: both exhibit local minima at signed shifts, and both exhibit negative curvature in symmetry breaking directions. A major difference (and hence major challenge) is that gradient methods for Bilinear Lasso do not retract to a union of subspaces—they retract to a more complicated, nonlinear set.

Our model assumes $\boldsymbol{x}_0$ to be Bernoulli–Gaussian vectors, which are sparse and symmetric i.i.d. random variables. When $\boldsymbol{x}_0$ is sparse but nonsymmetric, (e.g., Bernoulli), one can apply our result with a simple symmetrization trick, using the concatenated observation vectors $[\boldsymbol{y}, -\boldsymbol{y}]$ as an input to our algorithm.

When $\boldsymbol{x}_0$ is highly sparse and if $\boldsymbol{y}$ is noiseless, it is possible to identify a short copy of $\boldsymbol{a}_0$ via looking for the shortest consecutive nonzero entries within $\boldsymbol{y}$. When $\theta \ll 1/p_0$, these isolated copies are very common. Once θ exceeds $1/p_0$, or when support $\boldsymbol{x}_0$ is not Bernoulli random while being more clustered, they become very uncommon. In particular, the probability of an isolated copy is small unless $n \gtrapprox \exp(p_0\theta)$. Our proposed approach succeeds when $n \geq \mathrm{poly}(p_0)$.

In applications involving noisy data, optimization approaches often outperform direct inspection, even for samples with isolated copies of $\boldsymbol{a}_0$. An intuition for this is that optimization methods aggregate information across the sample. One practical avenue for obtaining the best of both worlds is to try to optimize the choice of data segment used for initialization. This can be a potential improvement for our data-driven initialization scheme, both in theory and in practice.

Finally, there are several directions in which our analysis could be improved. Our lower bounds on the length n of the random vector $\boldsymbol{x}_0$ required for success are clearly suboptimal. We also suspect our sparsity-coherence tradeoff between μ, θ (roughly, $\theta \lessapprox 1/(\sqrt{\mu p_0})$) is suboptimal, even for the φ_ρ objective. Articulating optimal sparsity-coherence tradeoffs is another interesting direction in this line of work. Extending our current result for cases when $\boldsymbol{y}$ is affected by noise can also be a natural next step for future work.

REFERENCES

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, 2009.

[2] A. Ahmed, B. Recht, and J. Romberg, *Blind deconvolution using convex programming*, IEEE Trans. Inform. Theory, 60 (2014), pp. 1711–1732.

[3] G. Ayers and J. C. Dainty, *Iterative blind deconvolution method and its applications*, Opt. Lett., 13 (1988), pp. 547–549.

[4] S. Baker and T. Kanade, *Limits on super-resolution and how to break them*, IEEE Trans. Pattern Anal. Mach. Intell., 24 (2002), pp. 1167–1183.

[5] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202, https://doi.org/10.1137/080716542.

[6] A. J. Bell and T. J. Sejnowski, *An information-maximization approach to blind separation and blind deconvolution*, Neural Comput., 7 (1995), pp. 1129–1159.

[7] A. Benichoux, E. Vincent, and R. Gribonval, *A fundamental pitfall in blind deconvolution with sparse and shift-invariant priors*, in the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2013), IEEE, 2013, pp. 6108–6112.

[8] P. Bones, C. Parker, B. Satherley, and R. Watson, *Deconvolution and phase retrieval with use of zero sheets*, J. Opt. Soc. A, 12 (1995), pp. 1842–1857.

[9] D. Briers, D. D. Duncan, E. R. Hirst, S. J. Kirkpatrick, M. Larsson, W. Steenbergen, T. Stromberg, and O. B. Thompson, *Laser speckle contrast imaging: Theoretical and practical limitations*, J. Biomed. Opt., 18 (2013), 066018.

[10] P. Campisi and K. Egiazarian, *Blind Image Deconvolution: Theory and Applications*, CRC Press, 2016.

[11] E. J. Candes, M. B. Wakin, and S. P. Boyd, *Enhancing sparsity by reweighted ℓ_1 minimization*, J. Fourier Anal. Appl., 14 (2008), pp. 877–905.

[12] M. Cannon, *Blind deconvolution of spatially invariant image blurs with phase*, IEEE Trans. Acoustics Speech Signal Process., 24 (1976), pp. 58–63.

[13] A. S. Carasso, *Direct blind deconvolution*, SIAM J. Appl. Math., 61 (2001), pp. 1980–2007, https://doi.org/10.1137/S0036139999362592.

[14] T. F. Chan and C.-K. Wong, *Total variation blind deconvolution*, IEEE Trans. Image Process., 7 (1998), pp. 370–375.

[15] S. Cheung, Y. Lau, Z. Chen, J. Sun, Y. Zhang, J. Wright, and A. Pasupathy, *Beyond the Fourier transform: A nonconvex optimization approach to microscopy analysis*, submitted.

[16] Y. Chi, *Guaranteed blind sparse spikes deconvolution via lifting and convex optimization*, IEEE J. Selected Topics Signal Process., 10 (2016), pp. 782–794.

[17] S. Choudhary and U. Mitra, *Fundamental Limits of Blind Deconvolution Part* II: *Sparsity-Ambiguity Trade-Offs*, preprint, https://arxiv.org/abs/1503.03184, 2015.

[18] W. Dong, L. Zhang, G. Shi, and X. Wu, *Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization*, IEEE Trans. Image Process., 20 (2011), pp. 1838–1857.

[19] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Ann. Statist., 32 (2004), pp. 407–499.

[20] C. Ekanadham, D. Tranchina, and E. P. Simoncelli, *A blind sparse deconvolution method for neural spike identification*, in Advances in Neural Information Processing Systems 24, 2011, pp. 1440–1448.

[21] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, *Removing camera shake from a single photograph*, ACM Trans. Graphics, 25 (2006), pp. 787–794.

[22] R. Ge, F. Huang, C. Jin, and Y. Yuan, *Escaping from saddle points—online stochastic gradient for tensor decomposition*, in Conference on Learning Theory, 2015, pp. 797–842.
[23] D. Goldfarb, *Curvilinear path steplength algorithms for minimization which use directions of negative curvature*, Math. Programming, 18 (1980), pp. 31–40.
[24] D. Goldfarb, C. Mu, J. Wright, and C. Zhou, *Using negative curvature in solving nonlinear programs*, Comput. Optim. Appl., 68 (2017), pp. 479–502.
[25] S. Harmeling, M. Hirsch, S. Sra, and B. Scholkopf, *Online blind deconvolution for astronomical imaging*, in the 2009 IEEE International Conference on Computational Photography (ICCP 2009), IEEE, 2009, pp. 1–7.
[26] R. Johnson, P. Schniter, T. J. Endres, J. D. Behm, D. R. Brown, and R. A. Casas, *Blind equalization using the constant modulus criterion: A review*, Proc. IEEE, 86 (1998), pp. 1927–1950.
[27] N. Joshi, R. Szeliski, and D. J. Kriegman, *PSF estimation using sharp edge prediction*, in the 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.
[28] K. F. Kaaresen and T. Taxt, *Multichannel blind deconvolution of seismic signals*, Geophys., 63 (1998), pp. 2093–2107.
[29] M. Kech and F. Krahmer, *Optimal injectivity conditions for bilinear inverse problems with applications to identifiability of deconvolution problems*, SIAM J. Appl. Algebra Geom., 1 (2017), pp. 20–37, https://doi.org/10.1137/16M1067469.
[30] D. Krishnan, T. Tay, and R. Fergus, *Blind deconvolution using a normalized sparsity measure*, in the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 233–240.
[31] D. Kundur and D. Hatzinakos, *Blind image deconvolution*, IEEE Signal Process. Mag., 13 (1996), pp. 43–64.
[32] R. Lane and R. Bates, *Automatic multidimensional deconvolution*, J. Opt. Soc. A, 4 (1987), pp. 180–188.
[33] R. G. Lane, *Blind deconvolution of speckle images*, J. Opt. Soc. A, 9 (1992), pp. 1508–1514.
[34] Y. Lau, Q. Qu, H.-W. Kuo, P. Zhou, Y. Zhang, and J. Wright, *Short-and-sparse deconvolution: A geometric approach*, in the 8th International Conference on Learning Representations, 2020; preprint, https://arxiv.org/abs/1908.10959, 2019.
[35] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, *Deconvolution Using Natural Image Priors*, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 2007.
[36] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, *Understanding blind deconvolution algorithms*, IEEE Trans. Pattern Anal. Mach. Intell., 33 (2011), pp. 2354–2367.
[37] M. S. Lewicki, *A review of methods for spike sorting: The detection and classification of neural action potentials*, Network, 9 (1998), pp. R53–R78.
[38] Y. Li and Y. Bresler, *Global geometry of multichannel sparse blind deconvolution on the sphere*, in Advances in Neural Information Processing Systems 31, Curran Associates, 2018, pp. 1132–1143.
[39] Y. Li, K. Lee, and Y. Bresler, *Identifiability in blind deconvolution with subspace or sparsity constraints*, IEEE Trans. Inform. Theory, 62 (2016), pp. 4266–4275.
[40] Y. Li, K. Lee, and Y. Bresler, *Identifiability and stability in blind deconvolution under minimal assumptions*, IEEE Trans. Inform. Theory, 63 (2017), pp. 4619–4633.
[41] S. Ling and T. Strohmer, *Self-calibration and biconvex compressive sensing*, Inverse Problems, 31 (2015), 115002.
[42] S. Ling and T. Strohmer, *Blind deconvolution meets blind demixing: Algorithms and performance bounds*, IEEE Trans. Inform. Theory, 63 (2017), pp. 4497–4520.
[43] J. Markham and J.-A. Conchello, *Parametric blind deconvolution: A robust method for the simultaneous estimation of image and blur*, J. Opt. Soc. A, 16 (1999), pp. 2377–2391.
[44] M. Miyoshi and Y. Kaneda, *Inverse filtering of room acoustics*, IEEE Trans. Acoustics Speech Signal Process., 36 (1988), pp. 145–152.
[45] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, Springer Science & Business Media, 2010.
[46] M. R. Osborne, B. Presnell, and B. A. Turlach, *A new approach to variable selection in least squares problems*, IMA J. Numer. Anal., 20 (2000), pp. 389–403.
[47] D. Perrone and P. Favaro, *Total variation blind deconvolution: The devil is in the details*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2909–2916.
[48] E. A. Pnevmatikakis, D. Soudry, Y. Gao, T. A. Machado, J. Merel, D. Pfau, T. Reardon,

Y. Mu, C. Lacefield, W. Yang et al., *Simultaneous denoising, deconvolution, and demixing of calcium imaging data*, Neuron, 89 (2016), pp. 285–299.
[49] S. K. Saha, *Diffraction-Limited Imaging with Large and Moderate Telescopes*, World Scientific, 2007.
[50] Y. Sato, *A method of self-recovering equalization for multilevel amplitude-modulation systems*, IEEE Trans. Commun., 23 (1975), pp. 679–682.
[51] O. Shalvi and E. Weinstein, *New criteria for blind deconvolution of nonminimum phase systems (channels)*, IEEE Trans. Inform. Theory, 36 (1990), pp. 312–321.
[52] Q. Shan, J. Jia, and A. Agarwala, *High-quality motion deblurring from a single image*, ACM Trans. Graphics, 27 (2008), pp. 1–10.
[53] G. Shtengel, J. A. Galbraith, C. G. Galbraith, J. Lippincott-Schwartz, J. M. Gillette, S. Manley, R. Sougrat, C. M. Waterman, P. Kanchanawong, M. W. Davidson, R. D. Fetter, and H. F. Hess, *Interferometric fluorescent super-resolution microscopy resolves 3D cellular ultrastructure*, Proc. Natl. Acad. Sci. USA, 106 (2009), pp. 3125–3130.
[54] T. G. Stockham, T. M. Cannon, and R. B. Ingebretsen, *Blind deconvolution through digital signal processing*, Proc. IEEE, 63 (1975), pp. 678–692.
[55] J. Sun, Q. Qu, and J. Wright, *Complete dictionary recovery over the sphere* II: *Recovery by Riemannian trust-region method*, IEEE Trans. Inform. Theory, 63 (2017), pp. 885–914.
[56] P. Walk, P. Jung, G. E. Pfander, and B. Hassibi, *Blind Deconvolution with Additional Autocorrelations via Convex Programs*, preprint, https://arxiv.org/abs/1701.04890, 2017.
[57] L. Wang and Y. Chi, *Blind deconvolution from multiple sparse inputs*, IEEE Signal Process. Lett., 23 (2016), pp. 1384–1388.
[58] Z. Wen and W. Yin, *A feasible method for optimization with orthogonality constraints*, Math. Program., 142 (2013), pp. 397–434.
[59] D. Wipf and H. Zhang, *Revisiting Bayesian blind deconvolution*, J. Mach. Learn. Res., 15 (2014), pp. 3595–3634.
[60] L. Xu and J. Jia, *Two-phase kernel estimation for robust motion deblurring*, in the 11th European Conference on Computer Vision, Springer, 2010, pp. 157–170.
[61] J. Yang, J. Wright, T. S. Huang, and Y. Ma, *Image super-resolution via sparse representation*, IEEE Trans. Image Process., 19 (2010), pp. 2861–2873.
[62] Y.-L. You and M. Kaveh, *Anisotropic blind image restoration*, in Proceedings of the 3rd International Conference on Image Processing, Vol. 2, IEEE, 1996, pp. 461–464.
[63] Y. Zhang, H.-W. Kuo, and J. Wright, *Structured local optima in sparse blind deconvolution*, IEEE Trans. Inform. Theory, 66 (2020), pp. 419–452.
[64] Y. Zhang, Y. Lau, H.-W. Kuo, S. Cheung, A. Pasupathy, and J. Wright, *On the global geometry of sphere-constrained sparse blind deconvolution*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 4894–4902.