# Geometry and Symmetry in Short-and-Sparse Deconvolution

**Han-Wen Kuo** [1 2] **Yuqian Zhang** [3] **Yenson Lau** [1 2] **John Wright** [1 2 4]

## 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 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 $\mu$, and $\boldsymbol{x}_0$ follows a random sparsity model with 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.

## 1. Introduction

Datasets in a wide range of areas, including neuroscience (Lewicki, 1998), microscopy (Cheung et al., 2017) and astronomy (Saha, 2007), 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 (Chan & Wong, 1998), 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.

[1]Data Science Institute, Columbia University, New York City, NY, USA [2]Department of Electrical Engineering, Columbia University, New York City, NY, USA [3]Department of Computer Science, Cornell University, Ithaca, NY, USA [4]Department of Applied Math and Applied Physics, Columbia University, New York City, NY, USA. Correspondence to: Han-Wen Kuo <hk2673@columbia.edu>.

*Proceedings of the $36^{th}$ International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. 

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[1] $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0 \in \mathbb{R}^n$. This SaS model exhibits a basic *scaled shift symmetry*: for any nonzero scalar $\alpha$ 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 Figure 1). Our problem of interest can be stated more formally as:

**Problem 1.1** (Short-and-Sparse Deconvolution)**.** *Given the cyclic convolution $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0 \in \mathbb{R}^n$ 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, and nonconvex formulations exhibit a complicated optimization landscape, with many equivalent global minimizers (scaled shifts of the ground truth) and additional local minimizers (scaled shift truncations of the ground truth), and a variety of critical points (Zhang et al., 2017; 2018). Currently available theory guarantees approximate recovery of a truncation[2] 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 $\boldsymbol{a}_0$ (Zhang et al., 2018).

In this paper, 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 (Zhang et al., 2017), 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

[1]Our result applies to direct convolution by zero padding both $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$.

[2]I.e., the portion of the shifted signal $s_\ell[\boldsymbol{a}_0]$ that falls in the window $\{0, \ldots, p_0 - 1\}$.

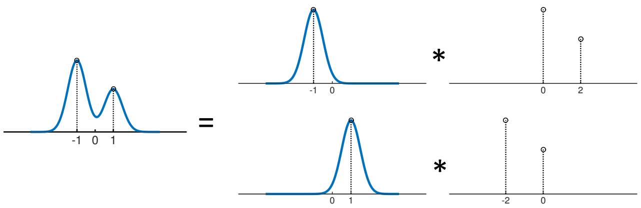


*Figure 1. Shift symmetry in Short-and-Sparse deconvolution.* An observation $\boldsymbol{y}$ (left) is a convolution of a short signal $\boldsymbol{a}_0$ and a sparse signal $\boldsymbol{x}_0$ (top right) can be equivalently expressed as a convolution of $s_\ell[\boldsymbol{a}_0]$ and $s_{-\ell}[\boldsymbol{x}_0]$,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

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 approach and modeling assumptions. Section 3 introduces our main results — both geometric and algorithmic — and compares them to the literature. In Section 4 we present a experimental result which corroborates our theoretical claim. Finally, Section 5 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

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

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

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

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

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

In (2.1)-(2.3), we constrain $\boldsymbol{a}$ to have unit $\ell^2$ 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 (Zhang et al., 2017)) 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 $\boldsymbol{a}_0$. (Zhang et al., 2017) analyzes the local minima of a simplification of (2.3), obtained by approximating[3] the data fidelity term as

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

This yields a simpler objective function

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

We make one further simplification to this problem, replacing the nondifferentiable penalty $\left\|\cdot\right\|_1$ with a smooth approximation $\rho(\boldsymbol{x})$.[4] Our analysis allows for a variety of smooth sparsity surrogates $\rho(\boldsymbol{x})$; for concreteness, we state our main results for the particular penalty[5]

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

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

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

As in (Zhang et al., 2017), we optimize $\varphi_\rho(\boldsymbol{a})$ over the sphere $\mathbb{S}^{p-1}$:

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

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 $\ell$-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 $\ell^2$ norm, i.e., $\boldsymbol{a}_0 \in \mathbb{S}^{p_0-1}$[6]. Our analysis makes two main assumptions, on the short motif $\boldsymbol{a}_0$ and the sparse map $\boldsymbol{x}_0$, respectively:

[3] For a generic $\boldsymbol{a}$, we have $\langle s_i[\boldsymbol{a}], s_j[\boldsymbol{a}]\rangle \approx 0$ and hence $\left\|\boldsymbol{a}*\boldsymbol{x}\right\|_2^2 \approx \left\|\boldsymbol{e}_0*\boldsymbol{x}\right\|_2^2 = \left\|\boldsymbol{x}\right\|_2^2$.

[4] Objective $\varphi_{\ell^1}$ is not twice differentiable everywhere, hence cannot be minimized with conventional second order methods.

[5] This surrogate is often named as the pseudo-Huber function.

[6] 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 $\left\|\boldsymbol{a}_0\right\|_2 \neq 1$, 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 $\alpha = \pm\frac{1}{\left\|\boldsymbol{a}_0\right\|_2}$.

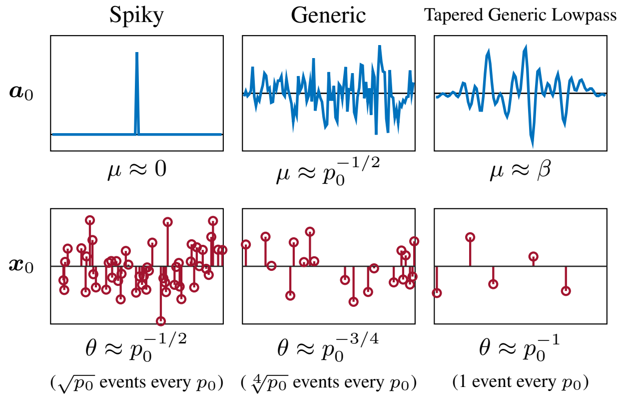


*Figure 2. Sparsity-coherence tradeoff:* We show three families of motifs $\boldsymbol{a}_0$ with varying coherence $\mu$ (top), with their maximum allowable sparsity $\theta$ and number of copies $\theta p_0$ within each length-$p_0$ window respectively (bot). When the target motif has smaller shift-coherence $\mu$, our result allows larger $\theta$, and vise 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.

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:

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

$\mu(\boldsymbol{a}_0)$ is bounded between 0 and 1. Our theory allows any $\mu$ 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 $\mu(\boldsymbol{a}_0) \approx 0$.[7] 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 Lowpass.* 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 Hamming window) and normalizing to unit $\ell^2$ norm. When $L = p_0\sqrt{1-\beta}$, with high probability $\mu(\boldsymbol{a}_0) \lessapprox \beta$. In this model, $\mu$ does not have to diminish as $p_0$ grows – it can be a fixed constant.[8]

[7]The use of "$\approx$" here suppresses constant and log factors.

[8]The upper right panel of Figure 2 is generated using random DFT components with frequencies smaller then 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.

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

We assume that $\boldsymbol{x}_0$ is a sparse random vector, under Bernoulli-Gaussian distribution, with rate $\theta$. Concretely speaking, we assume $\boldsymbol{x}_{0i} = \boldsymbol{\omega}_i \boldsymbol{g}_i$, where $\boldsymbol{\omega}_i \sim \mathrm{Ber}(\theta)$, $\boldsymbol{g}_i \sim \mathcal{N}(0,1)$ with all random variables are jointly independent. We write this as

$$\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta). \tag{2.10}$$

Here, $\theta$ is the probability that a given entry $\boldsymbol{x}_{0i}$ is nonzero. Problems with smaller $\theta$ 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* trade-offs, in which smaller coherence $\mu$ enables larger $\theta$, and vice-versa. More specifically, in our main theorem, the sparsity-coherence relationship is captured in the form

$$\theta \lessapprox 1/(p_0\sqrt{\mu} + \sqrt{p_0}). \tag{2.11}$$

When $\boldsymbol{a}_0$ 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 $\mu$ 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.

## 3. Main Results: Geometry and Algorithms

### 3.1. Geometry of the Objective $\varphi_\rho$

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 $\varphi_\rho$. In particular, we will argue that over a certain subset of the sphere, *every local minimum of $\varphi_\rho$ is close to some* $\pm s_\ell[\boldsymbol{a}_0]$.

To gain intuition into the properties of $\varphi_\rho$, 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-(a), we plot the function value of $\varphi_\rho$ 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 $\varphi_\rho$ is strongly convex on 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 $\varphi_\rho$ 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 $\varphi_\rho$ near the intersection (Figure 3-(b)) of the sphere $\mathbb{S}^{p-1}$ and the linear subspace

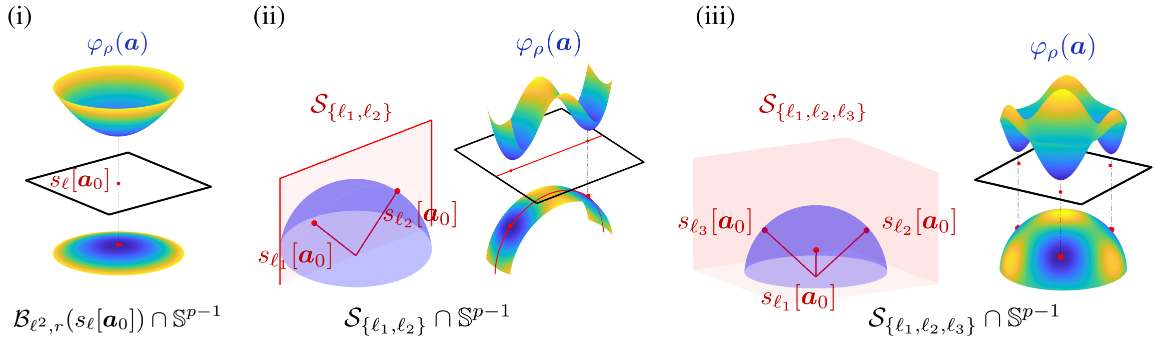


*Figure 3. Geometry of $\varphi_\rho$ near span of shift(s) of $\boldsymbol{a}_0$.* (i) A portion of the sphere $\mathbb{S}^{p-1}$ near $s_\ell[\boldsymbol{a}_0]$ (bot) colored according to height of $\varphi_\rho$ (top); $\varphi_\rho$ is strongly convex in this region, and it has a minimizer very close to $s_\ell[\boldsymbol{a}_0]$. (ii) 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 $\mathbb{R}^p$ (left), in which every local minimum of $\varphi_\rho$ near $\mathcal{S}_{\{\ell_1,\ell_2\}}$ is close to either $s_{\ell_1}[\boldsymbol{a}_0]$ or $s_{\ell_2}[\boldsymbol{a}_0]$ (right); there is a negative curvature in the middle of $s_{\ell_1}[\boldsymbol{a}_0]$, $s_{\ell_2}[\boldsymbol{a}_0]$, and $\varphi_\rho$ is convex in direction away from $\mathcal{S}_{\ell_1,\ell_2}$. (iii) 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 (left). On this set, $\varphi_\rho$ has local minimizers near each of the $s_{\ell_i}[\boldsymbol{a}_0]$, and are the only minimizers near $\mathcal{S}_{\ell_1,\ell_2,\ell_3}$ (right).

$$\mathcal{S}_{\{\ell_1,\ell_2\}} = \left\{ \boldsymbol{\alpha}_1 s_{\ell_1}[\boldsymbol{a}_0] + \boldsymbol{\alpha}_2 s_{\ell_2}[\boldsymbol{a}_0] \,\middle|\, \boldsymbol{\alpha} \in \mathbb{R}^2 \right\}.$$

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 $\varphi$ 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 $\varphi_\rho$ exhibits *positive curvature* in directions away from the subspace $\mathcal{S}_{\ell_1,\ell_2}$.

Finally, we visualize $\varphi_\rho$ over the intersection (Figure 3-(c)) 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 $\boldsymbol{a}_0$:

$$\mathcal{S}_{\{\ell_1,\ell_2,\ell_3\}} = \left\{ \boldsymbol{\alpha}_1 s_{\ell_1}[\boldsymbol{a}_0] + \boldsymbol{\alpha}_2 s_{\ell_2}[\boldsymbol{a}_0] + \boldsymbol{\alpha}_3 s_{\ell_3}[\boldsymbol{a}_0] \,\middle|\, \boldsymbol{\alpha} \in \mathbb{R}^3 \right\}.$$

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 obtain on *every* subspace spanned by a few shifts of $\boldsymbol{a}_0$. Indeed, for each subset

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

define a linear subspace

$$\mathcal{S}_{\boldsymbol{\tau}} = \left\{ \textstyle\sum_{\ell \in \boldsymbol{\tau}} \boldsymbol{\alpha}_\ell s_\ell[\boldsymbol{a}_0] \,\middle|\, \boldsymbol{\alpha} \in \mathbb{R}^{2p_0 - 1} \right\}. \tag{3.2}$$

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

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

Figure 4 gives a schematic portrait of this set. We claim:

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

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

**Theorem 3.1** (Main Geometric Theorem)**.** *Let $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$ with $\boldsymbol{a}_0 \in \mathbb{S}^{p_0-1}$ $\mu$-shift coherent and $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta) \in \mathbb{R}^n$ 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 $\varphi_\rho$. Then there exists $\delta > 0$ and numerical constant $c$ such that if $n \ge \mathrm{poly}(p_0)$, with high probability, every local minimizer $\bar{\boldsymbol{a}}$ of $\varphi_\rho$ over $\Sigma_{4\theta p_0}$ satisfies $\|\bar{\boldsymbol{a}} - \sigma s_\ell[\boldsymbol{a}_0]\|_2 \le c \max\left\{\mu, p_0^{-1}\right\}$ for some signed shift $\sigma s_\ell[\boldsymbol{a}_0]$ of the true kernel. Above, $c_1, c_2 > 0$ are positive numerical constants.*

**Proof.** See Appendix B. ■

The upper bound on $\theta$ 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 $\mu$ 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,

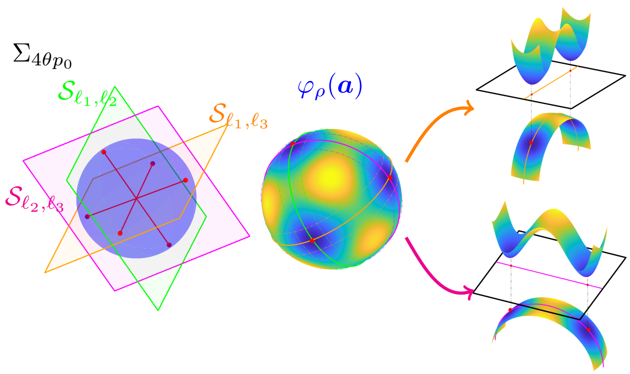


*Figure 4. Geometry of $\varphi_\rho$ over the union of subspaces $\Sigma_{4\theta p_0}$.* We show schematic representation of the union of subspaces $\Sigma_{4\theta p_0}$ (left). For each set $\boldsymbol{\tau}$ of at most $4\theta p_0$ shifts, we have a subspace $\mathcal{S}_{\boldsymbol{\tau}}$, by which $\varphi_\rho$ has good geometry near (right).

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 $\varphi_\rho$ 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 $\varphi_\rho$ has good geometric properties on (and near!) the union of subspaces $\Sigma_{4\theta p_0}$. In this section, we show how to use give an efficient method that exactly recovers $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$, up to shift symmetry. Although our geometric analysis only controls $\varphi_\rho$ near $\Sigma_{4\theta p_0}$, we will give a descent method which, with appropriate initialization $\boldsymbol{a}^{(0)}$, produces iterates $\boldsymbol{a}^{(1)}, \dots, \boldsymbol{a}^{(k)}, \dots$ 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.

Our algorithm starts with a 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:

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

A length-$p_0$ segment of data $\boldsymbol{y}_{0,\dots,p_0-1} = [\boldsymbol{y}_0, \dots, \boldsymbol{y}_{p_0-1}]^T$ 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 $\{0, \dots, p_0 - 1\}$. A relatively simple remedy is as follows: first, we zero-pad $\boldsymbol{y}_{0,\dots,p_0-1}$ to length $p = 3p_0 - 2$, giving

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

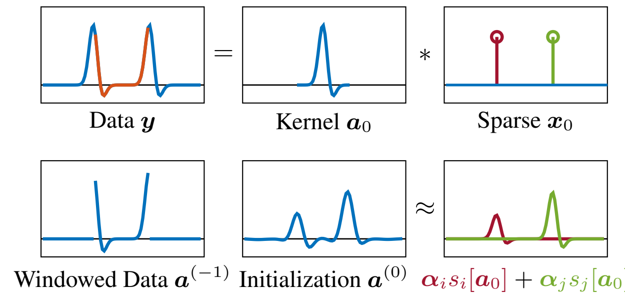


*Figure 5. Data-driven initialization.* Using a piece of the observed data $\boldsymbol{y}$ to generate an initial point that is close to a superposition of shifts $s_\ell[\boldsymbol{a}_0]$ of the ground truth. Data $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$ is a superposition of shifts of the true kernel $\boldsymbol{a}_0$ (top). A length-$p_0$ windowed $\boldsymbol{y}$ contains pieces of just a few shifts as $\boldsymbol{a}^{(-1)}$, one step of the generalized power method approximately fills in its missing pieces, yielding $\boldsymbol{a}^{(0)}$ as a near superposition of shifts of $\boldsymbol{a}_0$ (bot).

Zero padding provides enough space to accommodate any shift $s_\ell[\boldsymbol{a}_0]$ with $\ell \in \boldsymbol{\tau}$. We then perform one step of the generalized power method[9], 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, but can be clarified after interpreting the gradient $\nabla\varphi_\rho$ in terms of its action on the shifts $s_\ell[\boldsymbol{a}_0]$ (see appendix). 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 5 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 $\varphi_\rho$ grows as we move away from the subspace $\mathcal{S}_{\boldsymbol{\tau}}$, as shown in Figure 3. 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 (Goldfarb, 1980; Goldfarb et al., 2017) , which moves in a linear combination of the negative gradient direction $-\boldsymbol{g}$ and a negative curvature direction $-\boldsymbol{v}$. At the $k$-th iteration, the algorithm updates $\boldsymbol{a}^{(k+1)}$ as

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

[9] 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 $\varphi$ over the sphere consists of repeatedly projecting $-\nabla\varphi$ onto the sphere, giving the iteration $\boldsymbol{a} \mapsto -\boldsymbol{P}_{\mathbb{S}^{p-1}}\nabla\varphi(\boldsymbol{a})$. (3.7) can be interpreted as one step of the generalized power method for the objective function $\varphi_\rho$.

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.

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 byproduct, 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 $I^{(k)}$:

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

This can be viewed as an extreme form of *reweighting* (Candes et al., 2008). Second, our algorithm gradually decreases penalty variable $\lambda$ 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 (Osborne et al., 2000; Efron et al., 2004). For concreteness, at $k$-th iteration the algorithm reads:

Update $\boldsymbol{x}$:

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

Update $\boldsymbol{a}$:

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

Update $\lambda$ and $I$:

$$\lambda^{(k+1)} \leftarrow \tfrac{1}{2}\lambda^{(k)}, \quad I^{(k+1)} \leftarrow \operatorname{supp}\left(\boldsymbol{x}^{(k+1)}\right). \tag{3.11}$$

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 1. Figure 6 illustrates the main steps of this algorithm. Our main algorithmic result states that under closely related hypotheses as above, Algorithm 1 produces a signed shift of the ground truth $(\boldsymbol{a}_0, \boldsymbol{x}_0)$:

**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* $\mu$*-truncated shift coherent such that* $\max_{i \neq j}\left|\left\langle \boldsymbol{\iota}_{p_0}^* s_i[\boldsymbol{a}_0], \boldsymbol{\iota}_{p_0}^* s_j[\boldsymbol{a}_0]\right\rangle\right| \leq \mu$ *and* $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta) \in \mathbb{R}^n$ *with* $\theta, \mu$ *satisfying*

$$\theta \in \left[\frac{c_1}{p_0}, \frac{c_2}{(p_0\sqrt{\mu}+\sqrt{p_0})\log^2 p_0}\right], \quad \mu \leq \frac{c_3}{\log^2 n} \tag{3.12}$$

*for some constant* $c_1, c_2, c_3 > 0$*. If the signal lengths* $n, p_0$ *satisfy* $n > \operatorname{poly}(p_0)$ *and* $p_0 > \operatorname{polylog}(n)$*, then there exist* $\delta, \eta_v > 0$ *such that with high probability, Algorithm 1 produces* $(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{x}})$ *that are equal to the ground truth up to signed shift symmetry:*

$$\left\|\left(\widehat{\boldsymbol{a}},\, \widehat{\boldsymbol{x}}\right) - \sigma\left(s_\ell[\boldsymbol{a}_0],\, s_{-\ell}[\boldsymbol{x}_0]\right)\right\|_2 \leq \varepsilon \tag{3.13}$$

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

**Proof.** See Appendix C. ■

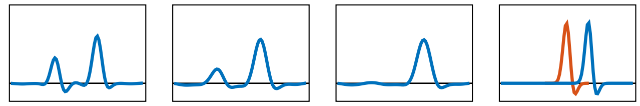


*Figure 6. Local minimization and refinement.* Data-driven initialization $\boldsymbol{a}^{(0)}$ consists of a near-superposition of two shifts (left), and minimizing $\varphi_\rho$ produces a near shift of $\boldsymbol{a}_0$ as $\bar{\boldsymbol{a}}$ (mid). Finally the rounded solution $\widehat{\boldsymbol{a}}$ using the Lasso is almost identical to a shift of $\boldsymbol{a}_0$ (right).

**Algorithm 1** Short and Sparse Deconvolution

**input** Observation $\boldsymbol{y}$, motif length $p_0$, sparsity $\theta$, shift-coherence $\mu$, and curvature threshold $-\eta_v$.

**Minimization**:
Initialize $\boldsymbol{a}^{(0)} \leftarrow -\boldsymbol{P}_{\mathbb{S}^{p-1}} \nabla\varphi_\rho\big(\boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\boldsymbol{0}^{p_0-1}; \boldsymbol{y}_0; \cdots; \boldsymbol{y}_{p_0-1}; \boldsymbol{0}^{p_0-1}\right]\big)$, $\lambda = 0.1/\sqrt{p_0\theta}$[10] and $\delta > 0$ in $\varphi_\rho$.
**for** $k = 1$ **to** $K_1$ **do**
$$\boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p-1}}[\boldsymbol{a}^{(k)} - t\boldsymbol{g}^{(k)} - t^2\boldsymbol{v}^{(k)}]$$
Here, $\boldsymbol{g}^{(k)}$ is the Riemannian gradient; $\boldsymbol{v}^{(k)}$ is the eigenvector of smallest Riemannian Hessian eigenvalue if less then $-\eta_v$ with $\left\langle \boldsymbol{v}^{(k)}, \boldsymbol{g}^{(k)}\right\rangle \geq 0$, otherwise let $\boldsymbol{v}^{(k)} = \boldsymbol{0}$; and $t \in (0, 0.1/n\theta]$ satisfies
$$\varphi_\rho(\boldsymbol{a}^{(k+1)}) < \varphi_\rho(\boldsymbol{a}^{(k)}) - \tfrac{1}{2}t\|\boldsymbol{g}^{(k)}\|_2^2 - \tfrac{1}{4}t^4\eta_v\|\boldsymbol{v}^{(k)}\|_2^2$$
**end for**
Obtain a near local minimizer $\bar{\boldsymbol{a}} \leftarrow \boldsymbol{a}^{(K_1)}$.

**Refinement**:
Initialize $\boldsymbol{a}^{(0)} \leftarrow \bar{\boldsymbol{a}}$, $\lambda^{(0)} \leftarrow 10(p\theta + \log n)(\mu + 1/p)$ and $I^{(0)} \leftarrow \mathcal{S}_{\lambda^{(0)}}\left[\operatorname{supp}(\check{\boldsymbol{y}} * \bar{\boldsymbol{a}})\right]$.
**for** $k = 1$ **to** $K_2$ **do**
$$\boldsymbol{x}^{(k+1)} \leftarrow \operatorname{argmin}_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{a}^{(k)} * \boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda^{(k)} \textstyle\sum_{i \notin I^{(k)}} |\boldsymbol{x}_i|$$
$$\boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\operatorname{argmin}_{\boldsymbol{a}} \tfrac{1}{2}\|\boldsymbol{a} * \boldsymbol{x}^{(k+1)} - \boldsymbol{y}\|_2^2\right]$$
$$\lambda^{(k+1)} \leftarrow \lambda^{(k)}/2, \qquad I^{(k+1)} \leftarrow \operatorname{supp}(\boldsymbol{x}^{(k+1)})$$
**end for**
**output** $(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{x}}) \leftarrow (\boldsymbol{a}^{(K_2)}, \boldsymbol{x}^{(K_2)})$

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

### 3.3. Relationship to the Literature

Blind deconvolution is a classical problem in signal processing (Stockham et al., 1975; Cannon, 1976), 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 short-and-sparse version of this problem, and then briefly discuss other deconvolution variants in the theoretical literature.

The short-and-sparse 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 (Lewicki, 1998; Ekanadham et al., 2011) and calcium imaging (Pnevmatikakis et al., 2016), where the observed signal exhibits repetitive *short* neuron excitation patterns occurring *sparsely* across time and/or space. Similarly, electron microscopy images (Cheung et al., 2017) 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*) (Ayers & Dainty, 1988; You & Kaveh, 1996; Carasso, 2001; Levin et al., 2007; 2011). In natural image deblurring, the target image is often assumed to have relatively few sharp edges (Fergus et al., 2006; Joshi et al., 2008; Levin et al., 2011), and hence have *sparse* derivatives. In scientific image deblurring, e.g., in astronomy (Lane, 1992; Harmeling et al., 2009; Briers et al., 2013) and geophysics (Kaaresen & Taxt, 1998), 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., (Kundur & Hatzinakos, 1996; Campisi & Egiazarian, 2016) for surveys.

Variants of the SaS deconvolution problem arise in many other areas of engineering as well. Examples include *blind equalization* in communications (Sato, 1975; Shalvi & Weinstein, 1990; Johnson et al., 1998), *dereverberation* in sound engineering (Miyoshi & Kaneda, 1988; Naylor & Gaubitch, 2010) and image *super-resolution* (Baker & Kanade, 2002; Shtengel et al., 2009; Yang et al., 2010).

These applications have motivated a great deal of algorithmic work on variants of the SaS problem (Lane & Bates, 1987; Bones et al., 1995; Bell & Sejnowski, 1995; Kundur & Hatzinakos, 1996; Markham & Conchello, 1999; Campisi & Egiazarian, 2016; Walk et al., 2017). In contrast, relatively little theory is available to explain when and why algorithms succeed. Our algorithm minimizes $\varphi_\rho$ 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 (You & Kaveh, 1996; Chan & Wong, 1998; Fergus et al., 2006; Levin et al., 2007; Shan et al., 2008; Xu & Jia, 2010; Dong et al., 2011; Krishnan et al., 2011; Levin et al., 2011; Wipf & Zhang, 2014; Perrone & Favaro, 2014; Zhang et al., 2017). 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 ($\boldsymbol{a} \geq \mathbf{0}$ and $\|\boldsymbol{a}\|_1 = 1$) arises naturally from the physical structure of the problem (You & Kaveh, 1996; Chan & Wong, 1998). 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 (Levin et al., 2011; Benichoux et al., 2013).

(Wipf & Zhang, 2014) imposes the $\ell^2$ regularization on $\boldsymbol{a}$ and observes that this alternative constraint gives more reliable algorithm. (Zhang et al., 2017) studies the geometry of the simplified objective $\varphi_{\ell^1}$ over the sphere, and proves that in the dilute limit in which $\boldsymbol{x}_0$ has one nonzero entry, all strict local minima of $\varphi_{\ell^1}$ are close to signed shifts truncations of $\boldsymbol{a}_0$. By adopting a different objective function (based on $\ell^4$ maximization) over the sphere, (Zhang et al., 2018) 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 $\{0, \dots, p_0 - 1\}$. The analysis of (Zhang et al., 2018) 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 (Zhang et al., 2017) and (Zhang et al., 2018) 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 (Zhang et al., 2017). However, we obtains *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. (Choudhary & Mitra, 2015) shows that the problem ill-posed for worst case $(\boldsymbol{a}_0, \boldsymbol{x}_0)$ – in particular, 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. Nevertheless, this worst case structure is unlikely to occur, either under the Bernoulli model, or in practical deconvolution problems.

Motivated by a variety of applications, many low-dimensional deconvolution models have been studied 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 (Ahmed et al., 2014; Li et al., 2016; Chi, 2016; Ling & Strohmer, 2015; Li et al., 2017; Ling & Strohmer, 2017; Kech & Krahmer, 2017). 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 (Ahmed et al., 2014). 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.

(Wang & Chi, 2016; Li & Bresler, 2018) study the related *multi-instance* sparse blind deconvolution problem (MISBD), 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 $\boldsymbol{a}_0$ to be invertible. Second, SaS model and MISBD are not equivalent despite the apparent similarity between them. 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.

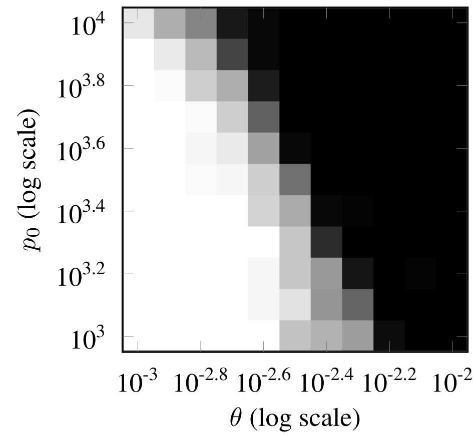


*Figure 7. Success probability of SaS deconvolution under generic $\boldsymbol{a}_0$, $\boldsymbol{x}_0$ with varying kernel length $p_0$, and sparsity rate $\theta$.* When sparsity rate decreases sufficiently with respect to kernel length, successful recovery becomes very likely (brighter), and vice versa (darker). A transition line is shown with slope $\frac{\log p_0}{\log \theta} \approx -2$, implying our method works with high probability when $\theta \lesssim \frac{1}{\sqrt{p_0}}$ in generic case.

## 4. Experiments

We demonstrate that the tradeoffs between the motif length $p_0$ and sparsity rate $\theta$ 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 $\mu(\boldsymbol{a}_0) \approx \frac{1}{\sqrt{p_0}}$.

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 1 by replacing curvilinear search with accelerated Riemannian gradient descent method (See appendix M). In Figure 7, we say the local minimizer $\boldsymbol{a}_{\min}$ is sufficiently close to a solution of SaS deconvolution problem, if

$$\mathrm{success}(\boldsymbol{a}_{\min}, ; \boldsymbol{a}_0) \;:=\; \left\{ \max_\ell |\langle s_\ell[\boldsymbol{a}_0], \boldsymbol{a}_{\min}\rangle| > 0.95 \right\}. \tag{4.1}$$

## 5. Discussion

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 $\varphi_\rho$ does not find the generating signal pairs $(\boldsymbol{a}_0, \boldsymbol{x}_0)$ consistently—the minimizer of $\varphi_\rho$ is often spurious and is not close to any particular shift of $\boldsymbol{a}_0$. We do not suggest minimizing $\varphi_\rho$ in this situation. On the other hand, minimizing the bilinear lasso objective $\varphi_{\text{lasso}}$ 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.

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 $\mu, \theta$ (roughly, $\theta \lesssim 1/(\sqrt{\mu} p_0)$) is suboptimal, even for the $\varphi_\rho$ objective. Articulating optimal sparsity-coherence tradeoffs for is another interesting direction for future work.


## Acknowledgements

The authors gratefully acknowledge support from NSF 1343282, NSF CCF 1527809, and NSF IIS 1546411.

## References


Ahmed, A., Recht, B., and Romberg, J. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014.

Ayers, G. and Dainty, J. C. Iterative blind deconvolution method and its applications. *Optics letters*, 13(7):547–549, 1988.

Baker, S. and Kanade, T. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, 2002.

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

Benichoux, A., Vincent, E., and Gribonval, R. A fundamental pitfall in blind deconvolution with sparse and shift-invariant priors. In *ICASSP-38th International Conference on Acoustics, Speech, and Signal Processing-2013*, 2013.

Bones, P., Parker, C., Satherley, B., and Watson, R. Deconvolution and phase retrieval with use of zero sheets. *JOSA A*, 12(9):1842–1857, 1995.

Briers, D., Duncan, D. D., Hirst, E. R., Kirkpatrick, S. J., Larsson, M., Steenbergen, W., Stromberg, T., and Thompson, O. B. Laser speckle contrast imaging: theoretical and practical limitations. *Journal of biomedical optics*, 18(6):066018, 2013.

Campisi, P. and Egiazarian, K. *Blind image deconvolution: theory and applications*. CRC press, 2016.

Candes, E. J., Wakin, M. B., and Boyd, S. P. Enhancing sparsity by reweighted âĎŞ 1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008.

Cannon, M. Blind deconvolution of spatially invariant image blurs with phase. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(1):58–63, 1976.

Carasso, A. S. Direct blind deconvolution. *SIAM Journal on Applied Mathematics*, 61(6):1980–2007, 2001.

Chan, T. F. and Wong, C.-K. Total variation blind deconvolution. *IEEE transactions on Image Processing*, 7(3): 370–375, 1998.

Cheung, S., Lau, Y., Chen, Z., Sun, J., Zhang, Y., Wright, J., and Pasupathy, A. Beyond the fourier transform: A nonconvex optimization approach to microscopy analysis. *Submitted*, 2017.

Chi, Y. Guaranteed blind sparse spikes deconvolution via lifting and convex optimization. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):782–794, June 2016.

Choudhary, S. and Mitra, U. Fundamental limits of blind deconvolution part ii: Sparsity-ambiguity trade-offs. *arXiv preprint arXiv:1503.03184*, 2015.

Dong, W., Zhang, L., Shi, G., and Wu, X. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857, 2011.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. Least angle regression. *The Annals of statistics*, 32(2): 407–499, 2004.

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

Fergus, R., Singh, B., Hertzmann, A., Roweis, S. T., and Freeman, W. T. Removing camera shake from a single photograph. In *ACM transactions on graphics (TOG)*, volume 25, pp. 787–794. ACM, 2006.

Goldfarb, D. Curvilinear path steplength algorithms for minimization which use directions of negative curvature. *Mathematical programming*, 18(1):31–40, 1980.

Goldfarb, D., Mu, C., Wright, J., and Zhou, C. Using negative curvature in solving nonlinear programs. *Computational Optimization and Applications*, 68(3):479–502, 2017.

Harmeling, S., Hirsch, M., Sra, S., and Scholkopf, B. Online blind deconvolution for astronomical imaging. In *2009 IEEE International Conference on Computational Photography (ICCP 2009)*, pp. 1–7. IEEE, 2009.

Johnson, R., Schniter, P., Endres, T. J., Behm, J. D., Brown, D. R., and Casas, R. A. Blind equalization using the constant modulus criterion: A review. *Proceedings of the IEEE*, 86(10):1927–1950, 1998.

Joshi, N., Szeliski, R., and Kriegman, D. J. Psf estimation using sharp edge prediction. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8. IEEE, 2008.

Kaaresen, K. F. and Taxt, T. Multichannel blind deconvolution of seismic signals. *Geophysics*, 63(6):2093–2107, 1998.

Kech, M. and Krahmer, F. Optimal injectivity conditions for bilinear inverse problems with applications to identifiability of deconvolution problems. *SIAM Journal on Applied Algebra and Geometry*, 1(1):20–37, 2017.

Krishnan, D., Tay, T., and Fergus, R. Blind deconvolution using a normalized sparsity measure. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 233–240. IEEE, 2011.

Kundur, D. and Hatzinakos, D. Blind image deconvolution. *IEEE signal processing magazine*, 13(3):43–64, 1996.

Lane, R. and Bates, R. Automatic multidimensional deconvolution. *JOSA A*, 4(1):180–188, 1987.

Lane, R. G. Blind deconvolution of speckle images. *JOSA A*, 9(9):1508–1514, 1992.

Levin, A., Fergus, R., Durand, F., and Freeman, W. T. Deconvolution using natural image priors. *Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory*, 3, 2007.

Levin, A., Weiss, Y., Durand, F., and Freeman, W. T. Understanding blind deconvolution algorithms. *IEEE transactions on pattern analysis and machine intelligence*, 33 (12):2354–2367, 2011.

Lewicki, M. S. A review of methods for spike sorting: the detection and classification of neural action potentials. *Network: Computation in Neural Systems*, 9(4):R53–R78, 1998.

Li, Y. and Bresler, Y. Global geometry of multichannel sparse blind deconvolution on the sphere. *arXiv preprint arXiv:1404.4104*, 2018.

Li, Y., Lee, K., and Bresler, Y. Identifiability in blind deconvolution with subspace or sparsity constraints. *IEEE Transactions on Information Theory*, 62(7):4266–4275, 2016.

Li, Y., Lee, K., and Bresler, Y. Identifiability and stability in blind deconvolution under minimal assumptions. *IEEE Transaction of Information Theory*, 2017.

Ling, S. and Strohmer, T. Self-calibration and biconvex compressive sensing. *Inverse Problems*, 31(11):115002, 2015.

Ling, S. and Strohmer, T. Blind deconvolution meets blind demixing: Algorithms and performance bounds. *IEEE Transactions on Information Theory*, 63(7):4497–4520, 2017.

Markham, J. and Conchello, J.-A. Parametric blind deconvolution: a robust method for the simultaneous estimation of image and blur. *JOSA A*, 16(10):2377–2391, 1999.

Miyoshi, M. and Kaneda, Y. Inverse filtering of room acoustics. *IEEE Transactions on acoustics, speech, and signal processing*, 36(2):145–152, 1988.

Naylor, P. A. and Gaubitch, N. D. *Speech dereverberation*. Springer Science & Business Media, 2010.

Osborne, M. R., Presnell, B., and Turlach, B. A. A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3):389–403, 2000.

Perrone, D. and Favaro, P. Total variation blind deconvolution: The devil is in the details. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2909–2916, 2014.

Pnevmatikakis, E. A., Soudry, D., Gao, Y., Machado, T. A., Merel, J., Pfau, D., Reardon, T., Mu, Y., Lacefield, C., Yang, W., et al. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron*, 89(2): 285–299, 2016.

Saha, S. K. *Diffraction-limited imaging with large and moderate telescopes*. World Scientific, 2007.

Sato, Y. A method of self-recovering equalization for multilevel amplitude-modulation systems. *IEEE Transactions on communications*, 23(6):679–682, 1975.

Shalvi, O. and Weinstein, E. New criteria for blind deconvolution of nonminimum phase systems (channels). *IEEE Transactions on information theory*, 36(2):312–321, 1990.

Shan, Q., Jia, J., and Agarwala, A. High-quality motion deblurring from a single image. In *Acm transactions on graphics (tog)*, volume 27, pp. 73. ACM, 2008.

Shtengel, G., Galbraith, J. A., Galbraith, C. G., Lippincott-Schwartz, J., Gillette, J. M., Manley, S., Sougrat, R., Waterman, C. M., Kanchanawong, P., Davidson, M. W., et al. Interferometric fluorescent super-resolution microscopy resolves 3d cellular ultrastructure. *Proceedings of the National Academy of Sciences*, 106(9):3125–3130, 2009.

Stockham, T. G., Cannon, T. M., and Ingebretsen, R. B. Blind deconvolution through digital signal processing. *Proceedings of the IEEE*, 63(4):678–692, 1975.

Walk, P., Jung, P., Pfander, G. E., and Hassibi, B. Blind deconvolution with additional autocorrelations via convex programs. *arXiv preprint arXiv:1701.04890*, 2017.

Wang, L. and Chi, Y. Blind deconvolution from multiple sparse inputs. *IEEE Signal Processing Letters*, 23(10): 1384–1388, 2016.

Wipf, D. and Zhang, H. Revisiting bayesian blind deconvolution. *The Journal of Machine Learning Research*, 15 (1):3595–3634, 2014.

Xu, L. and Jia, J. Two-phase kernel estimation for robust motion deblurring. In *European conference on computer vision*, pp. 157–170. Springer, 2010.

Yang, J., Wright, J., Huang, T. S., and Ma, Y. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

You, Y.-L. and Kaveh, M. Anisotropic blind image restoration. In *Image Processing, 1996. Proceedings., International Conference on*, volume 2, pp. 461–464. IEEE, 1996.

Zhang, Y., Lau, Y., Kuo, H.-w., Cheung, S., Pasupathy, A., and Wright, J. On the global geometry of sphere-constrained sparse blind deconvolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4894–4902, 2017.

Zhang, Y., Kuo, H.-W., and Wright, J. Structured local optima in sparse blind deconvolution. *arXiv preprint arXiv:1806.00338*, 2018.