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 $a_0$ and a sparse signal $x_0$ 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 $a_0$ has shift coherence $\mu$, and $x_0$ follows a random sparsity model with rate $\theta \in \left[\ell_1/p_0, \ell_0/\sqrt{p_0} \sqrt{\log^2 p_0}\right]$. 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 $y = a_0 * x_0$, between a short signal $a_0$ (the motif) and a longer sparse signal $x_0$, whose nonzero entries indicate where in the sample the motif is present. A very similar structure arises in image deblurring (Chan & Wong, 1998), where $y$ is a blurry image, $a_0$ the blur kernel, and $x_0$ 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 $a_0 \in \mathbb{R}^{p_0}$ and a sparse signal $x_0 \in \mathbb{R}^n$ ($n \gg p_0$) from their length-$n$ cyclic convolution $y = a_0 * x_0 \in \mathbb{R}^n$. This SaS model exhibits a basic scaled shift symmetry: for any nonzero scalar $\alpha$ and cyclic shift $s_{\ell}[\cdot]$,

$$\left(\alpha s_{\ell}[a_0]\right) \ast \left(\frac{1}{\alpha} s_{-\ell}[x_0]\right) = y. \quad (1.1)$$

Because of this symmetry, we only expect to recover $a_0$ and $x_0$ 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 $y = a_0 * x_0$ of $a_0 \in \mathbb{R}^{p_0}$ short ($p_0 \ll n$), and $x_0 \in \mathbb{R}^n$ sparse, recover $a_0$ and $x_0$, 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 of a shift $s_{\ell}[a_0]$, rather than guaranteeing recovery of $a_0$ as a whole, and requires certain (complicated) conditions on the convolution matrix associated with $a_0$ (Zhang et al., 2018).

In this paper, describe an algorithm which, under simpler conditions, exactly recovers a scaled shift of the pair $(a_0, x_0)$. 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 $a_0$. Based on

---

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

which corroborates our theoretical claim. Finally, Section 5 whenever a constrained formulation often correspond to shifts (or shift truncations (Zhang et al., 2017)) of the ground truth \(a_0\).

The problem (2.3) is defined in terms of the optimal Lasso cost. This function is challenging to analyze, especially far away from \(a_0\). (Zhang et al., 2017) analyzes the local minima of a simplification of (2.3), obtained by approximating the data fidelity term as

\[
\frac{1}{2} \|a*x - y\|_2^2 = \frac{1}{2} \|a*x\|_2^2 - \langle a*x, y \rangle + \frac{1}{2} \|y\|_2^2,
\]

\[
\approx \frac{1}{2} \|x\|_2^2 - \langle a*x, y \rangle + \frac{1}{2} \|y\|_2^2. \tag{2.4}
\]

This yields a simpler objective function

\[
\varphi_\ell(a) = \min_x \left\{ \frac{1}{2} \|x\|_2^2 - \langle a*x, y \rangle + \frac{1}{2} \|y\|_2^2 + \lambda \|x\|_1 \right\}. \tag{2.5}
\]

We make one further simplification to this problem, replacing the nondifferentiable penalty \(\|\cdot\|_1\) with a smooth approximation \(\varphi_\ell(x)\).

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

\[
\varphi_\ell(a) = \min_x \left\{ \frac{1}{2} \|x\|_2^2 - \langle a*x, y \rangle + \frac{1}{2} \|y\|_2^2 + \lambda \varphi_\ell(x) \right\}. \tag{2.7}
\]

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

\[
\min_a \varphi_\ell(a) \quad \text{s.t.} \quad a \in S^{p-1} \tag{2.8}
\]

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 \(a_0\), rather than a truncated shift. Our approach will leverage the following fact: if we view \(a \in S^{p-1}\) 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 \(s_\ell[a_0]\) 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 \(a_0\) has unit \(\ell^2\) norm, i.e., \(a_0 \in S^{p_0-1}\). Our analysis makes two main assumptions, on the short motif \(a_0\) and the sparse map \(x_0\), respectively:

\[3\text{For a generic} \ a, \text{we have} \langle s_i[a], s_j[a]\rangle \approx 0 \text{ and hence} \|a + x\|_2^2 \approx \|a_0 + x\|_2^2 \approx \|x\|_2^2.\]

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

\[5\text{This surrogate is often named as the pseudo-Huber function.}\]

\[6\text{This is purely a technical convenience. Our theory guarantees recovery of a signed shift} \ (\pm s_\ell[a_0], \pm s_{-\ell}[x_0]) \text{of the truth. If} \|a_0\|_2 \neq 1, \text{identical reasoning implies that our method recovers a scaled shift} \ (\alpha s_\ell[a_0], \alpha^{-1} s_{-\ell}[x_0]) \text{with} \alpha = \pm \frac{1}{\|a_0\|_2}.\]
The first is that distinct shifts \( a_0 \) have small inner product. We define the shift coherence of \( \mu(a_0) \) to be the largest inner product between distinct shifts:

\[
\mu(a_0) = \max_{\ell \neq 0} |\langle a_0, s_\ell a_0 \rangle|.
\]  

(2.9)

\( \mu(a_0) \) is bounded between 0 and 1. Our theory allows any \( \mu \) smaller than some numerical constant. Figure 2 shows three examples of families of \( a_0 \) that satisfy this assumption:

- **Spiky.** When \( a_0 \) is close to the Dirac delta \( \delta_0 \), the shift coherence \( \mu(a_0) \approx 0 \). Here, the observed signal \( y \) consists of a superposition of sharp pulses. This is arguably the easiest instance of SaS deconvolution.

- **Generic.** If \( a_0 \) is chosen uniformly at random from the sphere \( S_{p_0}^{-1} \), its coherence is bounded as \( \mu(a_0) \lesssim \sqrt{1/p_0} \) with high probability.

- **Tapered Generic Lowpass.** Here, \( a_0 \) 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 \( \mu(a_0) \lesssim \beta \). In this model, \( \mu \) does not have to diminish as \( p_0 \) grows — it can be a fixed constant.\(^8\)

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

We assume that \( x_0 \) is a sparse random vector, under Bernoulli-Gaussian distribution, with rate \( \theta \). Concretely speaking, we assume \( x_{0i} = \omega_1 g_i \), where \( \omega_1 \sim \text{Ber}(\theta) \), \( g_i \sim N(0, 1) \) with all random variables are jointly independent. We write this as

\[
x_0 \sim \text{i.i.d. } BG(\theta).
\]  

(2.10)

Here, \( \theta \) is the probability that a given entry \( x_{0i} \) is nonzero. Problems with smaller \( \theta \) are easier to solve. In the extreme case, when \( \theta \ll 1/p_0 \), the observation \( y \) contains many isolated copies of the motif \( a_0 \), and \( a_0 \) 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

\[
\theta \lesssim 1/(p_0 \sqrt{\mu + \sqrt{p_0}}).
\]  

(2.11)

When \( a_0 \) is very shift-incoherent (\( \mu \approx 0 \)), our method succeeds when each length-\( p_0 \) window contains about \( \sqrt{p_0} \) copies of \( a_0 \). When \( \mu \) is larger (as in the generic lowpass model), our method succeeds as long as relatively few copies of \( a_0 \) 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 \( a_0 \) (and \( x_0 \)) up to a signed shift — i.e., we wish to recover some \( \pm s_\ell [a_0] \). The shifts \( \pm s_\ell [a_0] \) 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 \( \pm s_\ell [a_0] \).

To gain intuition into the properties of \( \varphi_\rho \), we first visualize this function in the vicinity of a single shift \( s_\ell [a_0] \) of the ground truth \( a_0 \). In Figure 3-(a), we plot the function value of \( \varphi_\rho \) over \( B_{r_2}^2 (s_\ell [a_0]) \cap S_{p_0}^{-1} \), where \( B_{r_2}^2 (a) \) is a ball of radius \( r \) around \( a \). We make two observations:

\(^8\)The upper right panel of Figure 2 is generated using random DFT components with frequencies smaller than one-third Nyquist.
We make three observations:

- The objective function $\varphi_p$ is strongly convex on this neighborhood of $s_\ell[a_0]$.
- There is a local minimizer very close to $s_\ell[a_0]$.

We next visualize the objective function $\varphi_p$ near the linear span of two different shifts $s_{\ell_1}[a_0]$ and $s_{\ell_2}[a_0]$. More precisely, we plot $\varphi_p$ near the intersection (Figure 3-(b)) of the sphere $\mathbb{S}^{p-1}$ and the linear subspace $S_{\ell_1, \ell_2} = \{ \alpha_1 s_{\ell_1}[a_0] + \alpha_2 s_{\ell_2}[a_0] \mid \alpha \in \mathbb{R}^2 \}$.

We make three observations:

- Again, there is a local minimizer near each shift $s_\ell[a_0]$.
- These are the only local minimizers in the vicinity of $S_{\ell_1, \ell_2}$. In particular, the objective function $\varphi$ exhibits negative curvature along $S_{\ell_1, \ell_2}$ at any superposition $\alpha_1 s_{\ell_1}[a_0] + \alpha_2 s_{\ell_2}[a_0]$ whose weights $\alpha_1$ and $\alpha_2$ are balanced, i.e., $|\alpha_1| \approx |\alpha_2|$.
- Furthermore, the function $\varphi_p$ exhibits positive curvature in directions away from the subspace $S_{\ell_1, \ell_2}$.

Finally, we visualize $\varphi_p$ over the intersection (Figure 3-(c)) of the sphere $\mathbb{S}^{p-1}$ with the linear span of three shifts $s_{\ell_1}[a_0], s_{\ell_2}[a_0], s_{\ell_3}[a_0]$ of the true kernel $a_0$:

$$S_{\ell_1, \ell_2, \ell_3} = \{ \alpha_1 s_{\ell_1}[a_0] + \alpha_2 s_{\ell_2}[a_0] + \alpha_3 s_{\ell_3}[a_0] \mid \alpha \in \mathbb{R}^3 \}$$

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 obtain on every subspace spanned by a few shifts of $a_0$. Indeed, for each subset

$$\mathbf{3.1} \quad \tau \subseteq \{-p_0 + 1, \ldots, p_0 - 1\}$$

define a linear subspace

$$\mathbf{3.2} \quad S_\tau = \{ \sum_{\ell \in \tau} \alpha_\ell s_\ell[a_0] \mid \alpha \in \mathbb{R}^{2p_0-1} \}.$$

The subspace $S_\tau$ is the linear span of the shifts $s_\ell[a_0]$ indexed by $\ell$ in the set $\tau$. Our geometric theory will show that with high probability the function $\varphi_p$ has no spurious local minimizers near any $S_\tau$ for which $|\tau|$ is not too large — say, $|\tau| \leq 40p_0$. Combining all of these subspaces into a single geometric object, define the union of subspaces

$$\mathbf{3.3} \quad \Sigma_{40p_0} = \bigcup_{|\tau| \leq 40p_0} S_\tau.$$

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

- In the neighborhood of $\Sigma_{40p_0}$, all local minimizers are near signed shifts.
- The value of $\varphi_p$ grows in directions away from $\Sigma_{40p_0}$.

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 $y = a_0 * x_0$ with $a_0 \in \mathbb{S}^{p_0-1}$ $\mu$-shift coherent and $x_0 \sim \text{i.i.d.}$ $\text{BG}($ of $\theta$) in $\mathbb{R}^n$ with sparsity rate

$$\mathbf{3.4} \quad \theta \in \left[\frac{c_1}{p_0}, \frac{c_2}{p_0 \sqrt{\mu} + \sqrt{\rho_0}}\right], \frac{1}{\log p_0}.$$

Choose $\rho(x) = \sqrt{x^2 + \delta^2}$ and set $\delta = 0.1/\sqrt{\rho_0}$ in $\varphi_p$. Then there exists $\delta > 0$ and numerical constant $c$ such that if $n \geq \text{poly}(p_0)$, with high probability, every local minimizer $\bar{a}$ of $\varphi_p$ over $\Sigma_{40p_0}$ satisfies $\|\bar{a} - s_{\ell}[a_0]\|_2 \leq c \max \{\mu, p_0^{-1}\}$ for some signed shift $s_\ell[a_0]$ 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 $a_0$ is better conditioned (as a kernel), its coherence $\mu$ is smaller and $x_0$ can be denser.

At a technical level, our proof of Theorem 3.1 shows that (i) $\varphi_\rho(a)$ is strongly convex in the vicinity of each signed shift, and that at every other point $a$ near $\Sigma_{4\ell p_0}$, there is either (ii) a nonzero gradient or (iii) a direction of strict negative curvature; furthermore (iv) the function $\varphi_\rho$ grows away from $\Sigma_{4\ell p_0}$. Points (ii)-(iii) imply that near $\Sigma_{4\ell p_0}$ 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 $\Sigma_{4\ell p_0}$. Moreover, this minimizer is close enough to a shift (here, $\|a - s_\ell(a_0)\|_2 \approx \mu$) for us to exactly recover $s_\ell(a_0)$: we will give a refinement algorithm that produces $(\pm s_\ell(a_0), \pm \sigma_{-\ell}(x_0))$.

3.2. Provable Algorithm for SaS Deconvolution

The objective function $\varphi_\rho$ has good geometric properties on (and near!) the union of subspaces $\Sigma_{4\ell p_0}$. In this section, we show how to use give an efficient method that exactly recovers $a_0$ and $x_0$, up to shift symmetry. Although our geometric analysis only controls $\varphi_\rho$ near $\Sigma_{4\ell p_0}$, we will give a descent method which, with appropriate initialization $a^{(0)}$, produces iterates $a^{(1)}, \ldots, a^{(k)}, \ldots$ that remain close to $\Sigma_{4\ell p_0}$ for all $k$. In short, it is easy to start near $\Sigma_{4\ell p_0}$ and easy to stay near $\Sigma_{4\ell p_0}$. After finding a local minimizer $\bar{a}$, we refine it to produce a signed shift of $(a_0, x_0)$ using alternating minimization.

Our algorithm starts with an initialization scheme which generates $a^{(0)}$ near the union of subspaces $\Sigma_{4\ell p_0}$, which consists of linear combinations of just a few shifts of $a_0$. How can we find a point near this union? Notice that the data $y$ also consists of a linear combination of just a few shifts of $a_0$. Indeed:

$$y = a_0 * x_0 = \sum_{\ell \in \text{supp}(x_0)} a_0 \ell \circ s_\ell(a_0).$$  \hspace{1cm} (3.5)

A length-$p_0$ segment of data $y_{0,...,p_0-1} = [y_0, \ldots, y_{p_0-1}]^T$ captures portions of roughly $2\ell p_0 \ll 4\ell p_0$ shifts $s_\ell(a_0)$.

Many of these copies of $a_0$ are truncated by the restriction to $\{0, \ldots, p_0 - 1\}$. A relatively simple remedy is as follows: first, we zero-pad $y_{0,...,p_0-1}$ to length $p = 3p_0 - 2$, giving

$$[0^{p_0-1}; y_0; \ldots; y_{p_0-1}; 0^{p_0-1}].$$  \hspace{1cm} (3.6)

Zero padding provides enough space to accommodate any shift $s_\ell(a_0)$ with $\ell \in \tau$. We then perform one step of the generalized power method, writing

$$a^{(0)} = -P_{\ell^{p_0-1}} \nabla \varphi_\rho(P_{\ell^{p_0-1}} [0^{p_0-1}; y_0; \ldots; y_{p_0-1}; 0^{p_0-1}]),$$  \hspace{1cm} (3.7)

where $P_{\ell^{p_0-1}}$ 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 $s_\ell(a_0)$ (see appendix). For now, we note that this operation has the effect of (approximately) filling in the missing pieces of the truncated shifts $s_\ell(a_0)$—see Figure 5 for an example. We will prove that with high probability $a^{(0)}$ is indeed close to $\Sigma_{4\ell p_0}$.

The next key observation is that the function $\varphi_\rho$ grows as we move away from the subspace $\Sigma_\tau$, as shown in Figure 3.

\footnote{The power method for minimizing a quadratic form $\xi(a) = \frac{1}{2}a^T Ma$ over the sphere consists of the iteration $a \mapsto -P_{\ell^{p_0-1}} Ma$. Notice that in this mapping, $-Ma = -\nabla \xi(a)$. The generalized power method, for minimizing a function $\varphi$ over the sphere consists of repeatedly projecting $-\nabla \varphi$ onto the sphere, giving the iteration $a \mapsto -P_{\ell^{p_0-1}} \nabla \varphi(a).$ (3.7) can be interpreted as one step of the generalized power method for the objective function $\varphi_\rho$.}
Because of this, a small-stepping descent method will not move far away from \(\Sigma_{4p_0}\). 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 \(-g\) and a negative curvature direction \(-v\). At the \(k\)-th iteration, the algorithm updates \(a^{(k+1)}\) as

\[
a^{(k+1)} \leftarrow P_{g_{p-1}} \left[ a^{(k)} - t g^{(k)} + t^2 v^{(k)} \right] \tag{3.8}
\]

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 \(a^{(0)}\), this method produces a sequence \(a^{(1)}, a^{(2)}, \ldots\) which efficiently converges to a local minimizer \(\bar{a}\) that is near some signed shift \(\pm s\|\bar{a}\|_0\) of the ground truth.

The second step of our algorithm rounds the local minimizer \(\bar{a} \approx \sigma s|\bar{a}|_0\) to produce an exact solution \(\hat{a} = \sigma s|\bar{a}|_0\).

Byproduct, it also exactly recovers the corresponding reweighting that is not in \(x\).

\[
\bar{a}^\ast \approx \sigma s|\bar{a}|_0 \implies x = \sigma s|\bar{a}|_0. \tag{3.10}
\]

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 \(\bar{a} \ast \hat{x} \approx y\).

\[
\bar{a} \ast \hat{x} \approx y. \tag{3.10}
\]

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:

\[
\begin{align*}
&\text{Update } x: \quad x^{(k+1)} \leftarrow \text{argmin}_x \frac{1}{2} \|a^{(k)} \ast x - y\|^2_2 + \lambda^{(k)} \sum_{i \in I^{(k)}} |x_i|, \\
&\text{Update } a: \quad a^{(k+1)} \leftarrow P_{g_{p-1}} \left[ \text{argmin}_a \frac{1}{2} \|a \ast x^{(k+1)} - y\|^2_2 \right], \\
&\text{Update } \lambda \\& I: \quad (k+1) \leftarrow \frac{1}{2} \lambda^{(k)}, \quad I^{(k+1)} \leftarrow \text{supp} (x^{(k+1)}). \tag{3.11}
\end{align*}
\]

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 \(\bar{a}\) 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 \((a_0, x_0)\):

\[
\text{Algorithm 1 Short and Sparse Deconvolution}
\]

**Input** Observation \(y\), motif length \(p_0\), sparsity \(\theta\), shift-coherence \(\mu\), and curvature threshold \(-\eta_0\).

**Minimization:**

Initialize \(a^{(0)} \leftarrow -P_{g_{p-1}} \varphi_{\rho}(P_{g_{p-1}}(0^{p_0-1}; y_{p_0}; \ldots; y_{p_0-1}; 0^{p_0-1})\)), \(\lambda = 0.1/\sqrt{\|y\|^2_0}\) and \(\delta > 0\) in \(\varphi_{\rho}\).

For \(k = 1\) to \(K_1\) do

\[
a^{(k+1)} \leftarrow P_{g_{p-1}} \left[ a^{(k)} - t g^{(k)} + t^2 v^{(k)} \right] \tag{3.8}
\]

Here, \(g^{(k)}\) is the Riemannian gradient; \(v^{(k)}\) is the eigenvector of smallest Riemannian Hessian eigenvalue if less then \(-\eta_0\), with \(\langle v^{(k)}, g^{(k)} \rangle \geq 0\), otherwise let \(v^{(k)} = 0\); and \(t \in (0, 0.1/n\theta)\) satisfies

\[
\varphi_{\rho}(a^{(k+1)} - \varphi_{\rho}(a^{(k)}) - \frac{1}{2} t \|g^{(k)}\|^2_2 - \frac{1}{2} \eta_0 \|v^{(k)}\|^2_2 \tag{3.9}
\]

for Obtaining a near local minimizer \(\bar{a} \leftarrow a^{(K_1)}\).

**Refinement:**

Initialize \(a^{(0)} \leftarrow \bar{a}, \quad \lambda^{(0)} \leftarrow 10(p\theta + \log n)(\mu + 1/p)\)

and \(I^{(0)} \leftarrow \text{supp}(\hat{y} \ast \bar{a})\).

For \(k = 1\) to \(K_2\) do

\[
a^{(k+1)} \leftarrow \text{argmin}_a \frac{1}{2} \|a \ast x - y\|^2_2 + \lambda^{(k)} \sum_{i \in I^{(k)}} |x_i|, \\
\lambda^{(k+1)} \leftarrow \lambda^{(k)} / 2, \quad I^{(k+1)} \leftarrow \text{supp} (x^{(k+1)}) \tag{3.11}
\]

for Output \((\bar{a}, \hat{x}) \leftarrow (a^{(K_2)}, x^{(K_2)})\).

**Theorem 3.2** (Main Algorithmic Theorem). Suppose \(y = a_0 \ast x_0\) where \(a_0 \in \mathbb{S}^{p_0-1}\) is \(\mu\)-truncated shift coherent such that

\[
\max_{i \neq j} \left| \left\langle \mathcal{E}_{p_0} S_i [a_0], \mathcal{E}_{p_0} S_j [a_0] \right\rangle \right| \leq \mu \quad \text{and} \quad x_0 \sim_{\text{i.i.d.}} \text{BG}(\theta) \in \mathbb{R}^n \text{ with } \theta, \mu \text{ satisfying}
\]

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

for some constant \(c_1, c_2, c_3 > 0\). If the signal lengths \(n, p_0\) satisfy \(n > \text{poly}(p_0)\) and \(p_0 > \text{polylog}(n)\), then there exist \(\delta, \eta_0 > 0\) such that with high probability, Algorithm 1
produces $(\hat{a}, \hat{x})$ that are equal to the ground truth up to signed shift symmetry:

$$\| (\hat{a}, \hat{x}) - \sigma(s_\ell[a_0], s_{-\ell}[x_0]) \|_2 \leq \varepsilon \quad (3.13)$$

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

**Proof.** See Appendix C.

### 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* (repeated patterns) in datasets. This demonstrates that some 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 $\varphi_\ell$ 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 $a$ on the presence or absence of spurious local minimizers. In image deblurring, a simplex constraint ($a \geq 0$ and $\|a\|_1 = 1$) 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 $a$ 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 $a$ and observes that this alternative constraint gives more reliable algorithm. (Zhang et al., 2017) studies the geometry of the simplified objective $\varphi_\ell$ over the sphere, and proves that in the dilute limit in which $x_0$ has one nonzero entry, all strict local minima of $\varphi_\ell$ are close to signed shifts truncations of $a_0$. 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 $a_0$, i.e., the restriction of $s_\ell[a_0]$ to the window $\{0, \ldots, p_0 - 1\}$. The analysis of (Zhang et al., 2018) allows the sparse sequence $x_0$ to be denser ($\theta \sim p_0^{2/3}$ for a generic kernel $a_0$, 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 $s_\ell[a_0]$, under complicated conditions on the kernel $a_0$. Our core optimization problem is very similar to (Zhang et al., 2017). However, we obtain exact recovery of both $a_0$ and relatively dense $x_0$, 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 $(a_0, x_0)$ – in particular, for certain support patterns in which $x_0$ 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.

In practice, we suggest setting $\lambda = c_\lambda/\sqrt{p_0}\sigma$ with $c_\lambda \in [0.5, 0.8]$. 
Motivated by a variety of applications, many low-dimensional deconvolution models have been studied in the theoretical literature. In communication applications, the signals $a_0$ and $x_0$ 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 $y_i = a_0 * x_i$ consisting of multiple convolutions $i = 1, \ldots, K$ of a kernel $a_0$ and different sparse vectors $x_i$. Both works develop provable algorithms. There are several key differences with our work. First, both the proposed algorithms and their analysis require $a_0$ 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 $y$ into $K$ pieces; this apparent reduction fails due to boundary effects.

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 $a_0$ and $x_0$. 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 $a_0 \sim \text{Unif}(S^{p_0-1})$ and $x_0 \in \mathbb{R}^n$ with $x_0 \sim_{1,1,d} BG(\theta)$ where $n = 5 \times 10^5$. Note that choosing $a_0$ this way implies $\mu(a_0) \approx \frac{1}{\sqrt{n}}$.

For each instance, we recover $a_0$ and $x_0$ from $y = a_0 * x_0$ 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 $a_{\text{min}}$ is sufficiently close to a solution of SaS deconvolution problem, if

\[
\text{success}(a_{\text{min}}; a_0) := \left\{ \max_t |\langle x_t [a_0], a_{\text{min}} \rangle| > 0.95 \right\}.
\]

5. Discussion

The main drawback of our proposed method is that it does not succeed when the target motif $a_0$ has shift coherence very close to 1. For instance, a common scenario in image blind deconvolution involves deburring 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 $(a_0, x_0)$ consistently— the minimizer of $\varphi_\rho$ is often spurious and is not close to any particular shift of $a_0$. 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 $(a_0, x_0)$ 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 $x_0$ required for success are clearly suboptimal. We also suspect our sparsity-coherence tradeoff between $\mu, \theta$ (roughly, $\theta \approx 1/(\sqrt{p_0})$) is suboptimal, even for
the $\varphi_p$ 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**


