MANIFOLD GRADIENT DESCENT SOLVES MULTI-CHANNEL SPARSE BLIND DECONVOLUTION PROVABLY AND EFFICIENTLY

Laixi Shi and Yuejie Chi

Dept. of Electrical and Computer Engineering, Carnegie Mellon University
Emails: {laixis,yuejiec}@andrew.cmu.edu

ABSTRACT

Multi-channel sparse blind deconvolution refers to the problem of learning an unknown filter by observing its circulant convolutions with multiple input signals that are sparse. It is challenging to learn the filter efficiently due to the bilinear structure of the observations with respect to the unknown filter and inputs, leading to global ambiguities of identification. We propose a novel approach based on nonconvex optimization over the sphere manifold by minimizing a smooth surrogate of the sparsity-promoting loss function. It is demonstrated that manifold gradient descent with random initializations provably recovers the filter, up to scaling and shift ambiguities, as soon as the number of observations is sufficiently large under a suitable random data model. Numerical experiments are conducted to illustrate the efficiency of the proposed method with comparisons to existing methods.

Index Terms— multi-channel sparse blind deconvolution, nonconvex optimization, manifold gradient descent

1. INTRODUCTION

In various fields of signal processing, computer vision, and inverse problems, a problem of central interest is to simultaneously recover a pair of unknown signals $\boldsymbol{x}$ and $\boldsymbol{g}$ from their convolution. For example, neural or seismic recordings can be modeled as the convolution of a pulse shape (i.e. a filter), corresponding to characteristics of neurons or earth wave propagation, with a spike train modeling time of activations (i.e. a sparse input) [1, 2]. This problem is ill-posed without extra assumptions since the number of unknowns is much larger than the number of observations [3, 4, 5, 6, 7]. Luckily, in many situations, one can make multiple observations sharing the same filter, but with diverse sparse inputs, either spatially or temporally, thanks to the advances of sensing technologies [3, 8]. In this paper, we are thus motivated to identify the filter as well as the sparse inputs leveraging multiple convolutional observations in an efficient manner, a problem termed as multi-channel sparse blind deconvolution (MSBD).

Mathematically, we model each observation $\boldsymbol{y}_i \in \mathbb{R}^n$ as a convolution, between a filter, or an impulse response, $\boldsymbol{g} \in \mathbb{R}^n$, and a sparse input, $\boldsymbol{x}_i \in \mathbb{R}^n$:

$$\boldsymbol{y}_i = \boldsymbol{g} \circledast \boldsymbol{x}_i = \mathcal{C}(\boldsymbol{g})\boldsymbol{x}_i, \quad i = 1, \ldots, p, \tag{1}$$

where the total number of observations is given as p. Here, we consider circulant convolution, denoted as $\circledast$, whose operation is expressed equivalently via pre-multiplying a matrix $\mathcal{C}(\boldsymbol{g})$ to the input $\boldsymbol{x}_i$, where $\mathcal{C}(\boldsymbol{g})$ is a circulant matrix with $\boldsymbol{g}$ as its first column. Our goal is to recover both the filter $\boldsymbol{g}$ and sparse inputs $\{\boldsymbol{x}_i\}_{i=1}^p$ from the observations $\{\boldsymbol{y}_i\}_{i=1}^p$.

Clearly, the problem is challenging due to the bilinear form of the observations with respect to the unknowns, which are not uniquely identifiable. For any circular shift $\mathcal{S}_j(\boldsymbol{g})$ of $\boldsymbol{g}$ by j positions, and any non-zero scalar $\beta \neq 0$, we have $\boldsymbol{y}_i = (\beta\mathcal{S}_j(\boldsymbol{g})) \circledast (\beta^{-1}\mathcal{S}_{-j}(\boldsymbol{x}_i))$, for $j = 1, \ldots, n$. Hence, we can only hope to recover $\boldsymbol{g}$ and $\{\boldsymbol{x_i}\}_{i=1}^p$ accurately up to certain circulant shift and scaling ambiguities.

In this paper, we focus on the case that $\mathcal{C}(\boldsymbol{g})$ is invertible, which is equivalent to requiring all the Fourier coefficients of $\boldsymbol{g}$ are nonzero. This condition plays a critical role in guaranteeing the identifiability of the model as long as p is large enough [9]. Under this assumption, there exists a unique inverse filter (sometimes called an equalizer), $\boldsymbol{g}_{\rm inv} \in \mathbb{R}^n$ such that $\mathcal{C}(\boldsymbol{g}_{\rm inv})\mathcal{C}(\boldsymbol{g}) = \mathcal{C}(\boldsymbol{g})\mathcal{C}(\boldsymbol{g}_{\rm inv}) = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. This allows us to convert the bilinear form (1) into a linear form of the unknowns by multiplying $\mathcal{C}(\boldsymbol{g}_{\rm inv})$ on both sides to obtain:

$$\mathcal{C}(\boldsymbol{g}_{\rm inv})\boldsymbol{y}_i = \boldsymbol{g}_{\rm inv} \circledast \boldsymbol{g} \circledast \boldsymbol{x}_i = \mathcal{C}(\boldsymbol{g}_{\rm inv})\mathcal{C}(\boldsymbol{g})\boldsymbol{x}_i = \boldsymbol{x}_i,$$

for $i = 1, \ldots, p$. A natural strategy is to recover $\boldsymbol{g}_{\rm inv}$ via exploiting the sparsity of the inputs $\{\boldsymbol{x}_i\}_{i=1}^p$, by seeking a vector $\boldsymbol{h}$ that minimizes the cardinality of $\mathcal{C}(\boldsymbol{h})\boldsymbol{y}_i = \mathcal{C}(\boldsymbol{y}_i)\boldsymbol{h}$:

$$\min_{\boldsymbol{h}\in\mathbb{R}^n} \ \frac{1}{p}\sum_{i=1}^{p} \|\mathcal{C}(\boldsymbol{y}_i)\boldsymbol{h}\|_0 \,,$$

where $\|\cdot\|_0$ is the pseudo-ℓ_0 norm that counts the cardinality of the nonzero entries of the input vector. However, this simple formulation is problematic for two obvious reasons: (1) first, due to scaling ambiguity, a trivial solution is $\boldsymbol{h} = \boldsymbol{0}$; (2) second, the cardinality minimization is computationally infeasible. In this paper, we address these issues by reformulating this naive approach into an efficient and provably correct approach.

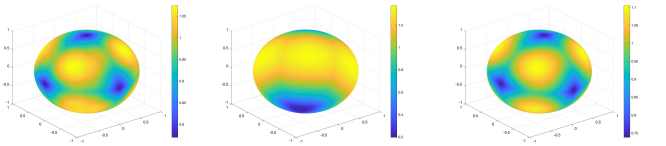
1.1. Our contributions

To address the first issue, a *spherical* constraint $\|\boldsymbol{h}\|_2 = 1$ is added to avoid the scaling ambiguity. For the second issue, the $\|\cdot\|_0$ constraint is relaxed to its convex smooth surrogate $\psi_\mu(z) = \mu \log\cosh(z/\mu)$, where μ controls the smoothness of the surrogate. With slight abuse of notation, we assume $\psi_\mu(\boldsymbol{z}) = \sum_{i=1}^n \psi_\mu(z_i)$ is applied entry-wise, where $\boldsymbol{z} = [z_i]_{1\le i\le n}$. Inspired by [3, 10], we propose the following pre-conditioned optimization problem for MSBD:

$$\min_{\boldsymbol{h}\in\mathbb{R}^n} \ f(\boldsymbol{h}) = \frac{1}{p}\sum_{i=1}^{p} \psi_\mu(\mathcal{C}(\boldsymbol{y}_i)\boldsymbol{R}\boldsymbol{h}) \quad \text{s.t} \quad \|\boldsymbol{h}\|_2 = 1, \tag{2}$$

where $\boldsymbol{R}$ is a pre-conditioning matrix depending only on the observations $\{\boldsymbol{y}_i\}_{i=1}^p$, that we will formally introduce later in Section 2.2.

This work is supported in part by ONR under the grants N00014-18-1-2142 and N00014-19-1-2404, by NSF under the grants CAREER ECCS-1818571, CCF-1806154 and CCF-1901199.



(a) orthogonal filter no pre-conditioning (b) general filter no pre-conditioning (c) general filter with pre-conditioning

Fig. 1. An illustration of the landscape of the empirical loss function $f(\boldsymbol{h})$ with or without the pre-conditioning matrix $\boldsymbol{R}$ in $\mathbb{R}^3$. (a) orthogonal filter $\mathcal{C}(\boldsymbol{g}) = \boldsymbol{I}$, no pre-conditioning is applied; (b) a general filter, no pre-conditioning is applied; (c) same general filter as (b) with pre-conditioning.

Encouragingly, although (2) is a nonconvex optimization problem due to the sphere constraint, under a suitable random model of the sparse inputs, the loss function $f(\boldsymbol{h})$ exhibits benign geometric curvatures as long as the sample size p is sufficiently large. As an illustration, Fig. 1 shows the landscape of $f(\boldsymbol{h})$ when $n = 3$, $p = 30$, and the inputs $\boldsymbol{x}_i$'s are composed of standard Bernoulli-Gaussian (BG) entries [11, Definition 2] with parameter $\theta = 0.3$.[1] When the filter is orthogonal, e.g. $\mathcal{C}(\boldsymbol{g}) = \boldsymbol{I}$, it can be seen from Fig. 1 (a) that the empirical loss function $f(\boldsymbol{h})$ has benign geometry without pre-conditioning (e.g. $\boldsymbol{R} = \boldsymbol{I}$), where the local minimizers are approximately all shift and sign-flipped variants of the ground truth (i.e, $\pm\boldsymbol{e}_i$), and are symmetrically distributed across the sphere. On the other end, for filters that are not orthogonal, the geometry of $f(\boldsymbol{h})$ without pre-conditioning is less well-posed, as illustrated in Fig. 1 (b). For the same non-orthogonal filter, by introducing pre-conditioning, which intuitively stretches the loss surface to mirror the orthogonal case, the pre-conditioned loss function $f(\boldsymbol{h})$ given in (2) has much better geometry (as illustrated in Fig. 1 (c)).

Motivated by the benign geometry of $f(\boldsymbol{h})$ illustrated above, a natural approach to minimize $f(\boldsymbol{h})$ over the sphere is via manifold gradient descent, which is simple and low-complexity. Surprisingly, this simple approach works remarkably well, even with random initializations, for appropriately chosen step sizes. Based on such empirical success, our goal is to address the following question: *can we establish theoretical guarantees of manifold gradient descent to recover the filter?*

In this paper, we prove that despite nonconvexity of (2), with a small number of random initializations, a vanilla manifold gradient descent (MGD) algorithm is capable of accurately recovering both the unknown filter and sparse inputs in polynomial time with high probability, as long as the sample size p is large enough on the order of $p = O(n^{4.5})$ up to logarithmic factors. This result is achieved through an integrated analysis of geometry and optimization, which provides justifications to the empirical success of MGD with random initializations.

1.2. Related work

Our work belongs to the recent line of activities on designing provable nonconvex procedures for high-dimensional statistical estimation, see [12] for an overview. The problem of blind deconvolution / calibration with a single observation (or equivalently, channel) has been studied extensively under different geometric priors such as sparsity and subspace assumptions on both the filter and the input, using convex and nonconvex optimization formulations [4, 6, 7, 13, 14, 15, 16, 17], and in the multi-channel setting [18, 19, 20, 21, 22, 23]. For the same MSBD problem, [18] proposed a convex linear program which has stringent requirements on the condition number of the filter matrix $\mathcal{C}(\boldsymbol{g})$. Our work is most related to [3], which was the first to study a nonconvex formulation for MSBD, by applying an ℓ_4-norm relaxation to the $\|\cdot\|_0$ norm. Our algorithm works at a much lower sample complexity $O(n^{4.5})$ in contract to $O(n^9)$ in [3] and outperforms it in the numerical experiments. Another concurrent work in [24] obtains a sample complexity of $p = O(n^5)$, which is slightly higher than ours, using a different loss function. Other algorithms for multi-channel blind deconvolution include sparse spectral methods [25] and nonconvex regularization [26].

On a different front, MSBD can be regarded as learning a convolutional invertible dictionary, where the proposed algorithm is inspired by approaches for dictionary learning in [10, 27, 28]. However, the approach in [27] is only applicable to orthogonal dictionaries, while we deal with a general invertible convolutional filter. Compared to the sample complexity $O(n^9)$ required for learning complete dictionaries in [10], our result demonstrates the benefit of exploiting convolutional structures, which has a much lower sample complexity of $O(n^{4.5})$.

1.3. Paper organization and notations

The rest of this paper is organized as follows. The theoretical guarantee of the benign geometry and its implications for the convergence of MGD are presented in Section 2. In Section 3, we numerically evaluate the proposed method with comparisons to existing algorithms. Finally, we conclude in Section 4. Due to space limits, the proofs are delayed to the full version [29]. Throughout the paper, we use boldface letters to represent vectors and matrices. For a vector $\boldsymbol{x} \in \mathbb{R}^n$, let x_j denote its jth element, $\boldsymbol{x}_{1:j}$ denote the length-j vector composed of the first j entries of $\boldsymbol{x}$, i.e., the vector $[x_1, x_2, \cdots, x_j]^\top$, $\boldsymbol{x}_{\setminus i}$ denote $\boldsymbol{x}_{1:i-1,i+1:n}$, i.e. the vector removing the ith element of $\boldsymbol{x}$. Let $[n]$ denote the index set $\{1, 2, \cdots, n\}$. If an index $j \notin [n]$ for an n-dimensional vector, then the actual index is computed as in the modulo n sense. Let $\|\cdot\|_2$ represent the ℓ_2 norm of a vector, and $\|\cdot\|$ denote the operator norm of a matrix.

2. MAIN RESULTS

In this section, we provide the theoretical analysis of the benign geometry of the objective function $f(\boldsymbol{h})$, and its algorithmic implications on efficient optimization via MGD.

2.1. Geometry in the Orthogonal Case

We start by considering the simpler case when $\mathcal{C}(\boldsymbol{g})$ is an orthonormal matrix. In this case, we present the following optimization problem without pre-conditioning:

$$\min_{\boldsymbol{h}\in\mathbb{R}^n} \; f_o(\boldsymbol{h}) := \frac{1}{p}\sum_{i=1}^{p} \psi_\mu(\mathcal{C}(\boldsymbol{y}_i)\boldsymbol{h}) \quad \text{s.t} \quad \|\boldsymbol{h}\|_2 = 1. \tag{3}$$

Without loss of generality, we can assume $\mathcal{C}(\boldsymbol{g}) = \boldsymbol{I}$. To see this, denote $\widetilde{\boldsymbol{h}} = \mathcal{C}(\boldsymbol{g})\boldsymbol{h}$, we have $\|\widetilde{\boldsymbol{h}}\|_2 = \|\mathcal{C}(\boldsymbol{g})\boldsymbol{h}\|_2 = 1$ due to the orthonormality of $\mathcal{C}(\boldsymbol{g})$. Therefore, (3) can be equivalently reformulated with respect to $\widetilde{\boldsymbol{h}}$, which corresponds to the case that the ground truth $\boldsymbol{g}_{\text{inv}} = \boldsymbol{e}_1$.

Our main geometric theorem characterizes benign local curvatures of $f_o(\boldsymbol{h})$ around the neighborhoods of $\pm\mathcal{S}_j(\boldsymbol{e}_1), j \in [n]$, which

[1] A random variable $X = \Omega \cdot Z$ is said to be BG with parameter $\theta \in [0, 1]$ if Ω is a Bernoulli random variable with probability θ and Z is a standard Gaussian random variable, where Ω and Z are independent.

are shifted and sign-permuted copies of the ground truth $\boldsymbol{e}_1$. We first introduce $2n$ subsets on the sphere which we will focus on in the following [27, 28]:

$$\mathcal{S}_{\xi}^{(i\pm)} = \left\{ \boldsymbol{h} : h_i \gtrless 0,\ \frac{h_i^2}{\left\|\boldsymbol{h}_{\setminus i}\right\|_\infty^2} \geqslant 1+\xi \right\}, \tag{4}$$

where $\xi \in [0, \infty)$ controls the size of the subsets, $i \in [n]$.

Due to symmetry, we will describe the geometry of $f_o(\boldsymbol{h})$ only for $\mathcal{S}_{\xi}^{(n+)}$, which carries over to the $2n$ subsets in (4). For convenience, we introduce the reparametrization trick in [10] by defining $\boldsymbol{w} = \boldsymbol{h}_{1:n-1} \in \mathbb{B}^{n-1}$ i.e. $\boldsymbol{h}(\boldsymbol{w}) = \left(\boldsymbol{w}, \sqrt{1-\|\boldsymbol{w}\|_2^2}\right)$, where $\mathbb{B}^{n-1} = \{\boldsymbol{w} : \|\boldsymbol{w}\|_2 \leq 1\}$ is the unit ball in $\mathbb{R}^{n-1}$. With this in mind, $f_o(\boldsymbol{h})$ can be equivalently rewritten with respect to $\boldsymbol{w}$ as $\phi_o(\boldsymbol{w}) = f_o(\boldsymbol{h}) = \frac{1}{p}\sum_{i=1}^p \psi_\mu(\mathcal{C}(\boldsymbol{y}_i)\boldsymbol{h}(\boldsymbol{w}))$, whose geometry is characterized in the following theorem.

Theorem 1. *Suppose $\mathcal{C}(\boldsymbol{g}) = \boldsymbol{I}$ and $\boldsymbol{x}_i \sim_{iid} \mathrm{BG}(\theta)$. For any $\xi_0 \in (0,1)$, $\theta \in (0, \frac{1}{3})$, there exist constants $c_1, c_2, c_3, c_4, c_5, C$ such that for $\mu < c_1 \min\{\theta, \xi_0^{1/6} n^{-3/4}\}$, $p \geq \frac{Cn^4}{\theta^2\xi_0^2}\log\left(\frac{n^3\log^{3/2} p \log n}{\mu\theta\xi_0}\right)$, the following holds with probability at least $1 - c_3 p^{-7} - \exp(-c_4 n)$ for $h(\boldsymbol{w}) \in \mathcal{S}_{\xi_0}^{(n+)}$:*

$$\frac{\boldsymbol{w}^\top \nabla\phi_o(\boldsymbol{w})}{\|\boldsymbol{w}\|_2} \geq c_2\xi_0\theta \quad \text{if } \|\boldsymbol{w}\|_2 \geq \frac{\mu}{4\sqrt{2}}, \tag{5a}$$

$$\nabla^2\phi_o(\boldsymbol{w}) \geq \frac{c_2 n\theta}{\mu}\boldsymbol{I} \quad \text{if } \|\boldsymbol{w}\|_2 \leq \frac{\mu}{4\sqrt{2}}, \tag{5b}$$

and the function $\phi_o(\boldsymbol{w})$ has exactly one unique local minimizer $\boldsymbol{w}_*$ near $\boldsymbol{0}$, such that $\|\boldsymbol{w}_* - \boldsymbol{0}\|_2 \leq \frac{c_5\mu}{\theta}\sqrt{\frac{\log^2 p}{p}}$.

Theorem 1 demonstrates the desired benign geometry of the empirical objective function in the subset $\mathcal{S}_{\xi_0}^{(n+)}$ with respect to $\boldsymbol{w}$. Theorem 1 guarantees that with high probability, there exists a unique local minimizer near a shifted ground truth $\boldsymbol{e}_n$, with $\|\boldsymbol{h}_*(\boldsymbol{w}_*) - \boldsymbol{e}_n\|_2 \leq \sqrt{2}\|\boldsymbol{w}_* - \boldsymbol{0}\|_2$. By extending this geometry to the $2n$ subsets $\mathcal{S}_{\xi}^{(i\pm)}$, there exists exactly $2n$ local minimizers which are close to the shifted and sign-permuted ground truths $\{\pm\boldsymbol{e}_i : i \in [n]\}$ respectively. The function $\phi_o(\boldsymbol{w})$ either has a large gradient towards the descent direction when $\|\boldsymbol{w}\|_2$ is large (c.f. (5a)), or is strongly convex when $\|\boldsymbol{w}\|_2$ is small (c.f. (5b)), indicating the geometry is rather benign and suitable for optimization using first-order methods such as MGD.

2.2. Extension to the General Case

We now sketch the extension of the geometry in Theorem 1 to the general case when $\mathcal{C}(\boldsymbol{g})$ is invertible, but not necessarily orthogonal. To preserve the benign geometry, we construct the pre-conditioning matrix $\boldsymbol{R}$ in the following manner according to [3, 10]:

$$\boldsymbol{R} = \left[\frac{1}{\theta np}\sum_{i=1}^p \mathcal{C}(\boldsymbol{y}_i)^\top \mathcal{C}(\boldsymbol{y}_i)\right]^{-1/2}.$$

As the sample size p grows to infinity, the pre-conditioning matrix $\boldsymbol{R}$ asymptotically converges to its expectation value, and the pre-conditioned $f(\boldsymbol{h})$ in (2) also gets closer to the benign geometry of $f_o(\boldsymbol{h})$ in the orthogonal case in Section 2.1. As long as the sample size is large enough, the theorem below suggests that under the same reparameterization $\boldsymbol{h} = \boldsymbol{h}(\boldsymbol{w})$ in Section 2.1, a geometry similar to that of $f_o(\boldsymbol{h})$ in Theorem 1 can be guaranteed for $\phi(\boldsymbol{w}) = f(\boldsymbol{h})$. Let κ be the condition number κ of $\mathcal{C}(\boldsymbol{g})$, which is the ratio of the largest and smallest magnitudes of the DFT coefficients of $\boldsymbol{g}$.

Algorithm 1: Manifold Gradient Descent for MSBD

Input: Observation $\{\boldsymbol{y}_i\}_{i=1}^p$, sparsity θ, step size η, initialization $\boldsymbol{h}^{(0)}$ on the sphere;

for $k \leftarrow 0$ **to** $T-1$ **do**

$$\boldsymbol{h}^{(k+1)} \leftarrow \frac{\boldsymbol{h}^{(k)} - \eta\cdot\partial f(\boldsymbol{h}^{(k)})}{\|\boldsymbol{h}^{(k)} - \eta\cdot\partial f(\boldsymbol{h}^{(k)})\|_2}$$

Output: Return $\boldsymbol{h}^{(T)}$

Theorem 2. *Suppose $\mathcal{C}(\boldsymbol{g})$ is invertible with condition number κ. For any $\xi_0 \in (0,1)$, $\theta \in (0, \frac{1}{3})$, there exist constants c_1, c_2, c_3, c_4, C such that when $\mu < c_1\min\{\theta, \xi_0^{1/6} n^{-3/4}\}$ and*

$$p \geq C\frac{\kappa^8 n^3 \log^4 p \log^2 n}{\theta^4\mu^2\xi_0^2}, \tag{6}$$

the geometry in (5) holds for $\phi(\boldsymbol{w})$ with probability at least $1 - c_3 p^{-7} - \exp(-c_4 n)$ for $\boldsymbol{h}(\boldsymbol{w}) \in \mathcal{S}_{\xi_0}^{(n+)}$. In addition, the function $\phi(\boldsymbol{w})$ has exactly one unique local minimizer $\boldsymbol{w}^\star$ near $\boldsymbol{0}$, such that $\|\boldsymbol{w}^\star - \boldsymbol{0}\|_2 \leq \frac{c_2\kappa^4}{\theta^2}\sqrt{\frac{n\log^3 p\log^2 n}{p}}$.

2.3. Manifold Gradient Descent

Inspired by Theorem 2, a simple MGD algorithm is proposed and summarized in Alg. 1, where $\partial f(\boldsymbol{h}) = (\boldsymbol{I} - \boldsymbol{h}\boldsymbol{h}^\top)\nabla f(\boldsymbol{h})$ is the Riemannian gradient with respect to $\boldsymbol{h}$, and $\nabla f(\boldsymbol{h})$ is the Euclidean gradient of $f(\boldsymbol{h})$. We present the convergence guarantee of Alg. 1 in the following theorem.

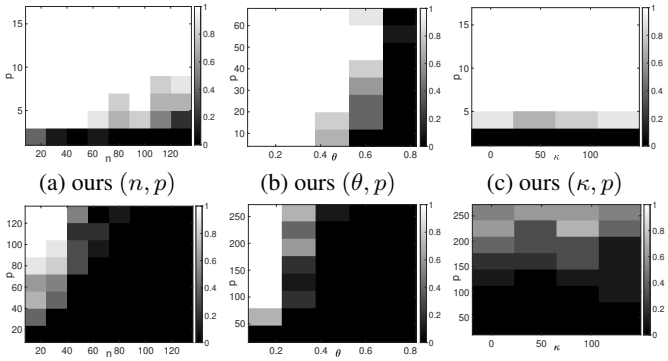
Theorem 3. *Instate the assumptions of Theorem 2. For the MGD algorithm in Alg. 1, if the initialization satisfies $\boldsymbol{h}^{(0)} \in \mathcal{S}_{\xi_0}^{(i\pm)}$, for any $i \in [n]$, then with a step size $\eta \leq \frac{c\mu\xi_0\theta}{n^2\sqrt{\log(np)}}$ for some sufficiently small constant c, the iterates $\boldsymbol{h}^{(k)}, k = 1, 2, \cdots$ stay in $\mathcal{S}_{\xi_0}^{(i\pm)}$ and achieve*

$$\min_{j\in[n]} \|\boldsymbol{h}^{(T)} \pm \mathcal{S}_j(\boldsymbol{g}_{\mathrm{inv}})\|_2 \lesssim \frac{\kappa^4}{\theta^2}\sqrt{\frac{n\log^3 p\log^2 n}{p}} + \epsilon$$

for any $\epsilon > 0$, in $T \lesssim \frac{n}{\mu\eta\xi_0\theta} + \frac{\mu}{n\theta\eta}\log\left(\frac{\mu}{\epsilon}\right)$ iterations.

The above theorem demonstrates that with an initialization in one of the $2n$ subsets $\left\{\mathcal{S}_{\xi_0}^{(i\pm)}, i \in [n]\right\}$, the proposed MGD algorithm, with proper step size, will converge to the unique local minimizer in that subset in a polynomial time. Therefore, the only left ingredient is to make sure a valid initialization can be obtained efficiently. Fortunately, it is known from the following lemma that setting $\xi_0 = 1/(4\log n)$ allows a sufficiently large basin of attraction, such that a random initialization can land into it with a constant probability.

Lemma 1. *[28, Lemma 3] When $\xi_0 = \frac{1}{4\log n}$, the initialization selected uniformly at random on the sphere lies in one of these $2n$ subsets $\left\{\mathcal{S}_{\xi_0}^{(i\pm)}, i \in [n]\right\}$ with probability at least $1/2$.*



(a) ours (n, p) (b) ours (θ, p) (c) ours (κ, p)

(d) [3] (n, p) (e) [3] (θ, p) (f) [3] (κ, p)

Fig. 2. Success rates of the proposed approach (first row) and the approach in [3] (second row) under various parameter settings.

Finally, combining Lemma 1 and Theorem 3, by setting $\xi_0 = 1/(4\log n)$, we can guarantee to recover $\boldsymbol{g}_{\text{inv}}$ accurately up to global ambiguity with high probability, as long as Alg. 1 is initialized uniformly at random over the sphere with $O(\log n)$ times.

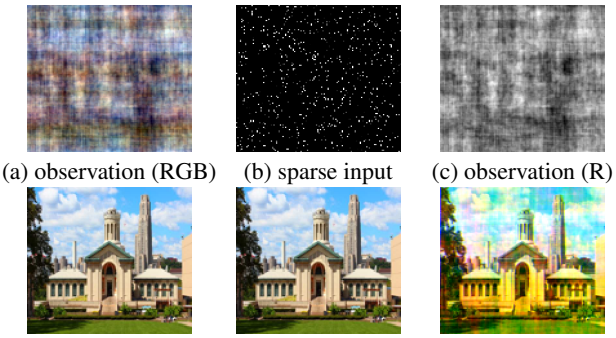
3. NUMERICAL EXPERIMENTS

In this section, we examine the performance of the proposed approach with comparison to [3], which is also based on MGD using a different loss function $L(\boldsymbol{h}) = -\frac{1}{4p}\sum_{i=1}^{p}\|\mathcal{C}(\boldsymbol{y}_i)\boldsymbol{R}\boldsymbol{h}\|_4^4$, on both synthetic and real data.

3.1. Experiments on synthetic data

We first compare the success rate of the proposed approach and that in [3], following a similar simulation setup as in [3]. In each experiment, the sparse inputs are generated following $\text{BG}(\theta)$, and the matrix $\mathcal{C}(\boldsymbol{g})$ with specific κ is synthesized by generating the DFT $\tilde{\boldsymbol{g}}$ of $\boldsymbol{g}$ which is random with the following rules: 1) The DFT $\tilde{\boldsymbol{g}}$ is symmetric to ensure that $\boldsymbol{g}$ is real, i.e., $\tilde{\boldsymbol{g}}_{(j)} = \tilde{\boldsymbol{g}}^*_{(n+2-j)}$, where $*$ denotes the conjugate operation. 2) The gains of $\tilde{\boldsymbol{g}}$ follow a uniform distribution on $[1, \kappa]$, and the phases of $\tilde{\boldsymbol{g}}$ follow a uniform distribution on $[0, 2\pi)$. In all experiments, we run MGD for no more than $T = 200$ iterations with a fixed step size of $\eta = 0.1$ and apply backtracking line search for both methods. For our formulation, we set $\mu = \min(10n^{-5/4}, 0.05)$. For each parameter setting, we conduct 10 Monte Carlo simulations to compute the success rate. Recall that the desired $\boldsymbol{h}$ is a signed shifted version of $\boldsymbol{g}_{\text{inv}}$, i.e., $\mathcal{C}(\boldsymbol{g})\boldsymbol{h} = \pm\boldsymbol{e}_j$ ($j \in [n]$). Therefore, to evaluate the accuracy of the output $\boldsymbol{h}^{(T)}$, we compute $\mathcal{C}(\boldsymbol{g})\boldsymbol{R}\boldsymbol{h}^{(T)}$ with the ground truth $\boldsymbol{g}$, and declare successful recovery if $\|\mathcal{C}(\boldsymbol{g})\boldsymbol{R}\boldsymbol{h}^{(T)}\|_\infty / \|\mathcal{C}(\boldsymbol{g})\boldsymbol{R}\boldsymbol{h}^{(T)}\|_2 > 0.99$.

Fig. 2 (a) and (d) show the success rate of the proposed approach and that in [3] with respect to n and p, where $\theta = 0.3$ and $\kappa = 8$ are fixed. It can be seen that the proposed approach succeeds at a much smaller sample size, where p is smaller than n. This indicates possible room for improvements of our theory. Fig. 2 (b) and (e) shows the success rate of the proposed approach and that in [3] with respect to θ and p, where $n = 64$ and $\kappa = 8$ are fixed. The proposed approach continues to work well even at a relatively high value of θ up to around 0.5. Finally, Fig. 2 (c) and (f) shows the success rate of the proposed approach and that in [3] with respect to κ and p, where $n = 64$ and $\theta = 0.3$ are fixed. Again, the performance of the proposed approach is insensitive to the condition number κ as long as the sample size p is large enough. On the other end, the approach in [3] performs significantly worse than the proposed approach under all the parameter settings.



(a) observation (RGB) (b) sparse input (c) observation (R)

(d) true image (e) recovery via ours (f) recovery via [3]

Fig. 3. Multi-channel sparse blind image deconvolution. Examples of (a) the observation in RGB; (b) the sparse input; (c) the observation for the R-channel alone. (d-f) The RGB image of the truth, the recovery via our method and [3].

3.2. Experiments on 2D image deconvolution

To further evaluate our method, we performance the task of blind image reconstruction and compare with [3]. Suppose multiple circulant convolutions $\{\boldsymbol{y}_i\}_{i=1}^p$ (illustrated in Fig. 3 (a) for the RGB image and Fig. 3 (c) for the R channel only) of an unknown 2D image (illustrated in Fig. 3 (d), the Hamerschlag hall on the campus of CMU) and multiple sparse inputs $\{\boldsymbol{x}_i\}_{i=1}^p$ (illustrated in Fig. 3 (b)) are observed. Here, the size of the observations is $n = 128 \times 128$, $\theta = 0.1$, and the number of observations $p = 1000$, which is significantly smaller than n.

We apply the proposed reconstruction method to each channel of the image, i.e. R, G, B, respectively using the corresponding channel of the observations $\{\boldsymbol{y}_i\}_{i=1}^p$, and obtain the final recovery by summing up the recovered channels. For each channel, the recovered image is computed as $\hat{\boldsymbol{g}} = \mathcal{F}^{-1}\left[\mathcal{F}\left(\boldsymbol{R}\hat{\boldsymbol{h}}\right)^{\odot -1}\right]$, where $\hat{\boldsymbol{h}}$ denotes the output of the algorithm, $\mathcal{F}$ is the 2D DFT operator, and $\boldsymbol{x}^{\odot -1}$ is the entry-wise inverse of a vector $\boldsymbol{x}$. The second row of Fig. 3 shows the true image, final recovered image by our method and [3] (after aligning the shift and sign) in (d), (e) and (f) respectively. It shows that the proposed approach again obtains much better recovery than that in [3].

4. CONCLUSION

This paper proposes a novel nonconvex approach for multi-channel sparse blind deconvolution based on manifold gradient descent with random initializations. Under a Bernoulli-Gaussian model for the sparse inputs, we provide theoretical characterizations for the benign geometric landscape of the loss function, which ensures the global convergence of a properly designed manifold gradient descent with random initializations. We prove that the proposed approach succeeds with high probability as long as the sample complexity satisfies $p = O(n^{4.5})$ up to logarithmic factors, which significantly improves prior art in [3]. Furthermore, our method succeeds in a much larger range of the condition number of $\mathcal{C}(\boldsymbol{g})$ and the sparsity level of inputs. In future work, we plan to improve the sample complexity as well as extend the analysis to the noisy setting.

5. REFERENCES

[1] C. Ekanadham, D. Tranchina, and E. P. Simoncelli, "Recovery of sparse translation-invariant signals with continuous basis pursuit," *IEEE transactions on signal processing*, vol. 59, no. 10, pp. 4735–4744, 2011.

[2] D. Donoho, "On minimum entropy deconvolution," in *Applied time series analysis II*. Elsevier, 1981, pp. 565–608.

[3] Y. Li and Y. Bresler, "Global geometry of multichannel sparse blind deconvolution on the sphere," in *Advances in Neural Information Processing Systems*, 2018, pp. 1132–1143.

[4] Y. Chi, "Guaranteed blind sparse spikes deconvolution via lifting and convex optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 782–794, 2016.

[5] H.-W. Kuo, Y. Lau, Y. Zhang, and J. Wright, "Geometry and symmetry in short-and-sparse deconvolution," in *International Conference on Machine Learning*, 2019, pp. 3570–3580.

[6] S. Ling and T. Strohmer, "Self-calibration and biconvex compressive sensing," *Inverse Problems*, vol. 31, no. 11, p. 115002, 2015.

[7] A. Ahmed, B. Recht, and J. Romberg, "Blind deconvolution using convex programming," *Information Theory, IEEE Transactions on*, vol. 60, no. 3, pp. 1711–1732, 2014.

[8] K. F. Kaaresen and T. Taxt, "Multichannel blind deconvolution of seismic signals," *Geophysics*, vol. 63, no. 6, pp. 2093–2107, 1998.

[9] Y. Li, K. Lee, and Y. Bresler, "A unified framework for identifiability analysis in bilinear inverse problems with applications to subspace and sparsity models," *arXiv preprint arXiv:1501.06120*, 2015.

[10] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere I: Overview and the geometric picture," *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 853–884, 2017.

[11] D. A. Spielman, H. Wang, and J. Wright, "Exact recovery of sparsely-used dictionaries," in *Conference on Learning Theory*, 2012, pp. 1–37.

[12] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-rank matrix factorization: An overview," *IEEE Transactions on Signal Processing*, vol. 67, no. 20, pp. 5239–5269, Oct 2019.

[13] X. Li, S. Ling, T. Strohmer, and K. Wei, "Rapid, robust, and reliable blind deconvolution via nonconvex optimization," *Applied and computational harmonic analysis*, vol. 47, no. 3, pp. 893–934, 2019.

[14] C. Ma, K. Wang, Y. Chi, and Y. Chen, "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution," *Foundations of Computational Mathematics*, pp. 1–182, 2018.

[15] R. Gribonval, G. Chardon, and L. Daudet, "Blind calibration for compressed sensing by convex optimization," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 2713–2716.

[16] Y. Li, K. Lee, and Y. Bresler, "Blind gain and phase calibration via sparse spectral methods," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 3097–3123, 2018.

[17] S. Choudhary and U. Mitra, "Sparse blind deconvolution: What cannot be done," in *2014 IEEE International Symposium on Information Theory*. IEEE, 2014, pp. 3002–3006.

[18] L. Wang and Y. Chi, "Blind deconvolution from multiple sparse inputs," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1384–1388, 2016.

[19] C. Bilen, G. Puy, R. Gribonval, and L. Daudet, "Convex optimization approaches for blind sensor calibration using sparsity," *Signal Processing, IEEE Transactions on*, vol. 62, no. 18, pp. 4847–4856, 2014.

[20] M. F. D. Costa and Y. Chi, "Self-calibrated super resolution," in *Asilomar Conference on Signals, Systems and Computers*. IEEE, 2019.

[21] M. Cho, W. Liao, and Y. Chi, "A non-convex approach to joint sensor calibration and spectrum estimation," in *2018 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, 2018, pp. 398–402.

[22] Y. C. Eldar, W. Liao, and S. Tang, "Sensor calibration for off-the-grid spectral estimation," *Applied and Computational Harmonic Analysis*, 2018.

[23] K. N. Ramamohan, S. P. Chepuri, D. F. Comesana, and G. Leus, "Blind calibration of sparse arrays for DOA estimation with analog and one-bit measurements," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 4185–4189.

[24] Q. Qu, X. Li, and Z. Zhu, "A nonconvex approach for exact and efficient multichannel sparse blind deconvolution," in *Advances in Neural Information Processing Systems*, 2019, pp. 4017–4028.

[25] K. Lee, N. Tian, and J. Romberg, "Fast and guaranteed blind multichannel deconvolution under a bilinear system model," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 4792–4818, 2018.

[26] Y. Xia and S. Li, "Identifiability of multichannel blind deconvolution and nonconvex regularization algorithm," *IEEE Transactions on Signal Processing*, vol. 66, no. 20, pp. 5299–5312, 2018.

[27] Y. Bai, Q. Jiang, and J. Sun, "Subgradient descent learns orthogonal dictionaries," in *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[28] D. Gilboa, S. Buchanan, and J. Wright, "Efficient dictionary learning with gradient descent," in *International Conference on Machine Learning*, 2019, pp. 2252–2259.

[29] L. Shi and Y. Chi, "Manifold gradient descent solves multichannel sparse blind deconvolution provably and efficiently," *arXiv preprint arXiv:1911.11167*, 2019.