Rapid, Robust, and Reliable Blind Deconvolution via Nonconvex Optimization

Shuyang Ling

Department of Mathematics, UC Davis

Oct.18th, 2016
Acknowledgements

Research in collaboration with:

- Prof. Xiaodong Li (UC Davis)
- Prof. Thomas Strohmer (UC Davis)
- Dr. Ke Wei (UC Davis)

This work is sponsored by NSF-DMS and DARPA.
Outline

- Applications in image deblurring and wireless communication
- Mathematical models and convex approach
- A nonconvex optimization approach towards blind deconvolution
What is blind deconvolution?

Suppose we observe a function $y$ which consists of the convolution of two unknown functions, the blurring function $f$ and the signal of interest $g$, plus noise $w$. How to reconstruct $f$ and $g$ from $y$?

$$y = f * g + w.$$ 

It is obviously a highly ill-posed *bilinear inverse* problem...

- Much more difficult than ordinary deconvolution...but has important applications in various fields.
- Solvability? What conditions on $f$ and $g$ make this problem solvable?
- How? What algorithms shall we use to recover $f$ and $g$?
Why do we care about blind deconvolution?

**Image deblurring**

Let $f$ be the blurring kernel and $g$ be the original image, then $y = f * g$ is the blurred image.

**Question:** how to reconstruct $f$ and $g$ from $y$?

\[
\begin{align*}
\mathbf{y} &= f * g + \mathbf{w} \\
\text{blurred image} &= \text{blurring kernel} * \text{original image} + \text{noise}
\end{align*}
\]
Why do we care about blind deconvolution?

Joint channel and signal estimation in wireless communication

Suppose that a signal $x$, encoded by $A$, is transmitted through an unknown channel $f$. How to reconstruct $f$ and $x$ from $y$?

$$y = f * Ax + w.$$
Subspace assumptions

We start from the original model

\[ y = f \ast g + w. \]

As mentioned before, it is an ill-posed problem. Hence, this problem is unsolvable without further assumptions...

**Subspace assumption**

Both \( f \) and \( g \) belong to known subspaces: there exist known tall matrices \( B \in \mathbb{C}^{L \times K} \) and \( A \in \mathbb{C}^{L \times N} \) such that

\[ f = Bh_0, \quad g = Ax_0, \]

for some unknown vectors \( h_0 \in \mathbb{C}^K \) and \( x_0 \in \mathbb{C}^N \).
Model under subspace assumption

In the frequency domain,

\[
\hat{y} = \hat{f} \odot \hat{g} + w = \text{diag}(\hat{f})\hat{g} + w,
\]

where “\( \odot \)” denotes entry-wise multiplication. We assume \( y \) and \( \hat{y} \) are both of length \( L \).

Subspace assumption

Both \( \hat{f} \) and \( \hat{g} \) belong to known subspaces: there exist known tall matrices \( \hat{B} \in \mathbb{C}^{L \times K} \) and \( \hat{A} \in \mathbb{C}^{L \times N} \) such that

\[
\hat{f} = \hat{B}h_0, \quad \hat{g} = \hat{A}x_0,
\]

for some unknown vectors \( h_0 \in \mathbb{C}^K \) and \( x_0 \in \mathbb{C}^N \). Here \( \hat{B} = FB \) and \( \hat{A} = FA \).

The degree of freedom for unknowns: \( K + N \); number of constraint: \( L \). To make the solution identifiable, we require \( L \geq K + N \) at least.
Subspace assumption is flexible and useful in applications.

- In imaging deblurring, $B$ can be the support of the blurring kernel; $A$ is a wavelet basis.
- In wireless communication, $B$ corresponds to time-limitation of the channel and $A$ is an encoding matrix.
Mathematical model

\[ y = \text{diag}(Bh_0)Ax_0 + w, \]

where \( \frac{w}{d_0} \sim \frac{1}{\sqrt{2}}N(0, \sigma^2 I_L) + \frac{i}{\sqrt{2}}N(0, \sigma^2 I_L) \) and \( d_0 = \|h_0\|\|x_0\| \).
Mathematical model

\[ y = \text{diag}(Bh_0)Ax_0 + w, \]

where \( \frac{w}{\|d_0\|} \sim \frac{1}{\sqrt{2}} \mathcal{N}(0, \sigma^2 I_L) + \frac{i}{\sqrt{2}} \mathcal{N}(0, \sigma^2 I_L) \) and \( d_0 = \|h_0\|\|x_0\| \).

One might want to solve the following nonlinear least squares problem,

\[ \min F(h, x) := \| \text{diag}(Bh)Ax - y \|^2. \]

Difficulties:

1. **Nonconvexity**: \( F \) is a nonconvex function; algorithms (such as gradient descent) are likely to get trapped at local minima.

2. **No performance guarantees.**
Convex approach and lifting

Two-step convex approach
(a) Lifting: convert bilinear to linear constraints
(b) Solving a SDP relaxation to recover $hx^*$. 

Let $a_i$ be the $i$-th column of $A^*$ and $b_i$ be the $i$-th column of $B^*$. 

$y_i = (Bh_0)^i x_0 a_i + w_i = b^*_i h_0 x_0 a_i + w_i$.

Let $X_0 := h_0 x_0$ and define the linear operator $A: C^{K \times N} \rightarrow C^L$ as,

$A(Z) := \{b^*_i Z a_i\}_{i=1}^L = \{\langle Z, b^*_i a_i \rangle\}_{i=1}^L$.

Then, there holds $y = A(X_0) + w$. 

In this way, $A^*(z) = \sum_{i=1}^L z_i b_i a_i^*: C^L \rightarrow C^{K \times N}$. 

Shuyang Ling (UC Davis)
Convex approach and lifting

Two-step convex approach

(a) Lifting: convert bilinear to linear constraints
(b) Solving a SDP relaxation to recover $hx^*$.

Step 1: lifting

Let $a_i$ be the $i$-th column of $A^*$ and $b_i$ be the $i$-th column of $B^*$.

$$y_i = (Bh_0)x_0^*a_i + w_i = b_i^*h_0x_0^*a_i + w_i,$$

Let $X_0 := h_0x_0^*$ and define the linear operator $A : \mathbb{C}^{K \times N} \rightarrow \mathbb{C}^L$ as,

$$A(Z) := \{b_i^*Za_i\}_{i=1}^L = \{\langle Z, b_i a_i^* \rangle\}_{i=1}^L.$$

Then, there holds

$$y = A(X_0) + w.$$

In this way, $A^*(z) = \sum_{i=1}^L z_i b_i a_i^* : \mathbb{C}^L \rightarrow \mathbb{C}^{K \times N}$. 
Convex relaxation and state of the art

**Step 2: nuclear norm minimization**

Consider the convex envelop of rank($Z$): nuclear norm $\|Z\|_* = \sum \sigma_i(Z)$.

$$\min \|Z\|_* \quad \text{s.t.} \quad \mathcal{A}(Z) = \mathcal{A}(X_0).$$

Convex optimization can be solved within polynomial time.
Convex relaxation and state of the art

Step 2: nuclear norm minimization

Consider the convex envelop of $\text{rank}(Z)$: nuclear norm $\|Z\|_* = \sum \sigma_i(Z)$.

$$\min \|Z\|_* \quad \text{s.t.} \quad A(Z) = A(X_0).$$

Convex optimization can be solved within polynomial time.

Theorem [Ahmed-Recht-Romberg 11]

Assume $y = \text{diag}(Bh_0)Ax_0$, $A : L \times N$ is a complex Gaussian random matrix,

$$B^*B = I_K, \quad \|b_i\|^2 \leq \frac{\mu_{\text{max}}^2 K}{L}, \quad L\|Bh_0\|_\infty^2 \leq \mu_h^2,$$

the above convex relaxation recovers $X = h_0x_0^*$ exactly with high probability if

$$C_0 \max(\mu_{\text{max}}^2 K, \mu_h^2 N) \leq \frac{L}{\log^3 L}.$$
Pros and Cons of Convex Approach

Pros and Cons

- **Pros**: Simple and comes with theoretic guarantees
- **Cons**: Computationally too expensive to solve SDP

Our Goal: **rapid, robust, reliable nonconvex approach**

- **Rapid**: linear convergence
- **Robust**: stable to noise
- **Reliable**: provable and comes with theoretical guarantees; number of measurements close to information-theoretic limits.
A nonconvex optimization approach?

An increasing list of nonconvex approach to various problems:
- Phase retrieval: by Candés, Li, Soltanolkotabi, Chen, Wright, etc...
- Matrix completion: by Sun, Luo, Montanari, etc...
- Various problems: by Recht, Wainwright, Constantine, etc...

Two-step philosophy for provable nonconvex optimization

(a) Use spectral initialization to construct a starting point inside “the basin of attraction”;  
(b) Simple gradient descent method.

The key is to build up “the basin of attraction”.

The basin of attraction relies on the following three observations.

**Observation 1: Unboundedness of solution**
- If the pair \((h_0, x_0)\) is a solution to \(y = \text{diag}(Bh_0)Ax_0\), then so is the pair \((\alpha h_0, \alpha^{-1}x_0)\) for any \(\alpha \neq 0\).

- Thus the blind deconvolution problem always has infinitely many solutions of this type. We can recover \((h_0, x_0)\) only up to a scalar.

- It is possible that \(\|h\| \gg \|x\|\) (vice versa) while \(\|h\| \cdot \|x\| = d_0\). Hence we define \(\mathcal{N}_{d_0}\) to balance \(\|h\|\) and \(\|x\|\):

\[
\mathcal{N}_{d_0} := \{(h, x) : \|h\| \leq 2\sqrt{d_0}, \|x\| \leq 2\sqrt{d_0}\}.
\]
Observation 2: Incoherence

Our numerical experiments have shown that the algorithm’s performance depends on how much $b_i$ and $h_0$ are correlated.

$$\mu_h^2 := \frac{L \| Bh_0 \|_\infty^2}{\| h_0 \|_2^2} = L \frac{\max_i |b_i^* h_0|^2}{\| h_0 \|_2^2},$$

the smaller $\mu_h$, the better.

Therefore, we introduce the $\mathcal{N}_\mu$ to control the incoherence:

$$\mathcal{N}_\mu := \{ h : \sqrt{L} \| Bh \|_\infty \leq 4\mu \sqrt{d_0} \}.$$  

“Incoherence” is not a new idea. In matrix completion, we also require the left and right singular vectors of the ground truth cannot be too “aligned” with those of measurement matrices $\{ b_i a_i^* \}_{1 \leq i \leq L}$. The same philosophy applies here.
Observation 3: “Close” to the ground truth

We define $\mathcal{N}_\varepsilon$ to quantify closeness of $(h, x)$ to true solution, i.e.,

$$\mathcal{N}_\varepsilon := \{(h, x) : \|hx^* - h_0x_0^*\|_F \leq \varepsilon d_0\}.$$ 

We want to find an initial guess close to $(h_0, x_0)$.
Based on the three observations above, we define the three neighborhoods (denoting \( d_0 = \| h_0 \| \| x_0 \| \) and \( 0 < \varepsilon \leq \frac{1}{15} \):

\[
\begin{align*}
\mathcal{N}_{d_0} &:= \{(h, x) : \|h\| \leq 2\sqrt{d_0}, \|x\| \leq 2\sqrt{d_0}\} \\
\mathcal{N}_{\mu} &:= \{h : \sqrt{L}\|Bh\|_{\infty} \leq 4\mu\sqrt{d_0}\} \\
\mathcal{N}_\varepsilon &:= \{(h, x) : \|hx^* - h_0x_0^*\|_{F} \leq \varepsilon d_0\}.
\end{align*}
\]

We first obtain a good initial guess \((u_0, v_0) \in \mathcal{N}_{d_0} \cap \mathcal{N}_\mu \cap \mathcal{N}_\varepsilon\), which is followed by regularized gradient descent.
Objective function: a variant of projected gradient descent

The objective function \( \tilde{F} \) consists of two parts: \( F \) and \( G \):

\[
\min_{(h,x)} \tilde{F}(h,x) := F(h,x) + G(h,x)
\]

\[ \text{least squares term} \quad \text{regularization term} \]

where \( F(h,x) := \|A(hx^*) - y\|^2 = \|\text{diag}(Bh)Ax - y\|^2 \) and

\[
G(h,x) := \rho \left[ G_0 \left( \frac{\|h\|^2}{2d} \right) + G_0 \left( \frac{\|x\|^2}{2d} \right) + \sum_{l=1}^{L} G_0 \left( \frac{L|b_l^*h|^2}{8d\mu^2} \right) \right].
\]

Here \( G_0(z) = \max\{z - 1, 0\}^2 \), \( \rho \approx d^2 \), \( d \approx d_0 \) and \( \mu \geq \mu_h \).

Regularization forces iterates \((u_t, v_t)\) inside \( \mathcal{N}_{d_0} \cap \mathcal{N}_{\mu} \cap \mathcal{N}_{\varepsilon} \).
Algorithm: Wirtinger Gradient Descent

Step 1: Initialization via spectral method and projection:
1. Compute $\mathcal{A}^*(y)$, (since $\mathbb{E}(\mathcal{A}^*(y)) = h_0 x_0^*$);
2. Find the leading singular value, left and right singular vectors of $\mathcal{A}^*(y)$, denoted by $(d, \hat{h}_0, \hat{x}_0)$ respectively;
3. $u_0 := P_{\mathcal{N}_\mu}(\sqrt{d}\hat{h}_0)$ and $v_0 := \sqrt{d}\hat{x}_0$;
4. Output: $(u_0, v_0)$.

Step 2: Gradient descent with constant stepsize $\eta$:
1. Initialization: obtain $(u_0, v_0)$ via Algorithm 1.
2. for $t = 1, 2, \ldots, \text{do}$
3. $u_t = u_{t-1} - \eta \nabla \tilde{F}_h(u_{t-1}, v_{t-1})$
4. $v_t = v_{t-1} - \eta \nabla \tilde{F}_x(u_{t-1}, v_{t-1})$
5. end for
Main theorem

**Theorem:** [Li-Ling-Strohmer-Wei, 2016]

Let $B$ be a tall partial DFT matrix and $A$ be a complex Gaussian random matrix. If the number of measurements satisfies

$$L \geq C(\mu_h^2 + \sigma^2)(K + N) \log^2(L)/\varepsilon^2,$$

(i) then the initialization $(u_0, v_0) \in \frac{1}{\sqrt{3}} N_{d_0} \cap \frac{1}{\sqrt{3}} N_{\mu} \cap N_{\frac{2}{5}\varepsilon}$;

(ii) the regularized gradient descent algorithm creates a sequence $(u_t, v_t)$ in $N_{d_0} \cap N_{\mu} \cap N_{\varepsilon}$ satisfying

$$\|u_t v_t^* - h_0 x_0^*\|_F \leq (1 - \alpha)^t \varepsilon d_0 + c_0 \|A^*(w)\|$$

with high probability where $\alpha = O\left(\frac{1}{(1+\sigma^2)(K+N)\log^2 L}\right)$.
Remarks

(a) If $w = 0$, $(u_t, v_t)$ converges to $(h_0, x_0)$ linearly.

$$\|u_t v_t^* - h_0 x_0^*\|_F \leq (1 - \alpha)^t \varepsilon d_0 \to 0, \text{ as } t \to \infty$$

(b) If $w \neq 0$, $(u_t, v_t)$ converges to a small neighborhood of $(h_0, x_0)$ linearly.

$$\|u_t v_t^* - h_0 x_0^*\|_F \to c_0 \|A^*(w)\|, \text{ as } t \to \infty$$

where

$$\|A^*(w)\| = \mathcal{O} \left( \sigma d_0 \sqrt{\frac{(K + N) \log L}{L}} \right) \to 0, \text{ if } L \to \infty.$$ 

As $L$ is becoming larger and larger, the effect of noise diminishes. (Recall linear least squares.)
Numerical experiments

Nonconvex approach v.s. convex approach:

\[
\min_{(h,x)} \tilde{F}(h,x) \quad \text{v.s.} \quad \min \|Z\|_* \quad \text{s.t.} \|A(Z) - y\| \leq \eta.
\]

Nonconvex method requires fewer measurements to achieve exact recovery than convex method. Moreover, if \(A\) is a partial Hadamard matrix, our algorithm still gives satisfactory performance.

\(K = N = 50, \ B\) is a low-frequency DFT matrix.
The number of measurements $L$ does depend linearly on $\mu_h^2$.

Our algorithm yields stable recovery if the observation is noisy.

Here $K = N = 100$. 

\[ L \text{ v.s. Incoherence } \mu_h^2 \text{ and stability} \]
MRI Image deblurring:

Here $B$ is a partial DFT matrix and $A$ is a partial wavelet matrix.

When the subspace $B$, $(K = 65)$ or support of blurring kernel is known: $g \approx Ax$ : image of $512 \times 512$; $A$ : wavelet subspace corresponding to the $N = 20000$ largest Haar wavelet coefficients of $g$. 
When the subspace $B$ or support of blurring kernel is unknown: we assume the support of blurring kernel is contained in a small box; $N = 35000$. 
Important ingredients of proof

The first three conditions hold over “the basin of attraction” \( \mathcal{N}_{d_0} \cap \mathcal{N}_\mu \cap \mathcal{N}_\varepsilon \).

**Condition 1: Local Regularity Condition**

Guarantee sufficient decrease in each iterate and linear convergence of \( \tilde{F} \):

\[
\| \nabla \tilde{F}(h, x) \|^2 \geq \omega \tilde{F}(h, x)
\]

where \( \omega > 0 \) and \((h, x) \in \mathcal{N}_{d_0} \cap \mathcal{N}_\mu \cap \mathcal{N}_\varepsilon \).

**Condition 2: Local Smoothness Condition**

Governs rate of convergence. Let \( z = (h, x) \). There exists a constant \( C_L \) (Lipschitz constant of gradient) such that

\[
\| \nabla \tilde{F}(z + t\Delta z) - \nabla \tilde{F}(z) \| \leq C_L t\| \Delta z \|, \quad \forall 0 \leq t \leq 1,
\]

for all \( \{(z, \Delta z) : z + t\Delta z \in \mathcal{N}_{d_0} \cap \mathcal{N}_\mu \cap \mathcal{N}_\varepsilon, \forall 0 \leq t \leq 1\} \).
**Condition 3: Local Restricted Isometry Property**

Transfer convergence of objective function to convergence of iterates.

\[
\frac{3}{4} \| h x^* - h_0 x_0^* \|^2_F \leq \| A(h x^* - h_0 x_0^*) \|^2 \leq \frac{5}{4} \| h x^* - h_0 x_0^* \|^2_F
\]

holds uniformly for all \((h, x) \in \mathcal{N}_{d_0} \cap \mathcal{N}_\mu \cap \mathcal{N}_\varepsilon\).

**Condition 4: Robustness Condition**

Provide stability against noise.

\[
\| A^*(w) \| \leq \frac{\varepsilon d_0}{10 \sqrt{2}}.
\]

where \(A^*(w) = \sum_{l=1}^L w_l b_l a_l^*\) is a sum of \(L\) rank-1 random matrices. It concentrates around \(0\).
Two-page proof

Condition 1 + 2 $\implies$ Linear convergence of $\tilde{F}$

Proof.

Let $z_{t+1} = z_t - \eta \nabla \tilde{F}(z_t)$ with $\eta \leq \frac{1}{C_L}$. By using modified descent lemma,

$$\tilde{F}(z_t + \eta \nabla \tilde{F}(z_t)) \leq \tilde{F}(z_t) - (2\eta + C_L \eta^2) \| \nabla \tilde{F}(z_t) \|^2 \leq \tilde{F}(z_t) - \eta \omega \tilde{F}(z_t)$$

which gives $\tilde{F}(z_{t+1}) \leq (1 - \eta \omega)^t \tilde{F}(z_0)$. \qed
Condition 3 $\iff$ Linear convergence of $\|u_t v_t^* - h_0 x_0^*\|_F$.

It follows from $\tilde{F}(z_t) \geq F(z_t) \geq \frac{3}{4} \|u_t v_t^* - h_0 x_0^*\|_F^2$. Hence, linear convergence of objective function also implies linear convergence of iterates.

Condition 4 $\iff$ Proof of stability theory

If $L$ is sufficiently large, $A^*(w)$ is small since $\|A^*(w)\| \to 0$. There holds

$$\|A(hx^* - h_0 x_0^*) - w\|^2 \approx \|A(hx^* - h_0 x_0^*)\|^2 + \sigma^2 d_0^2.$$

Hence, the objective function behaves “almost like” $\|A(hx^* - h_0 x_0^*)\|^2$, the noiseless version of $F$ if the sample size is sufficiently large.
Conclusion: The proposed algorithm is the first blind deconvolution algorithm that is numerically efficient, robust against noise and comes with rigorous recovery guarantees under subspace conditions.
Conclusion: The proposed algorithm is the first blind deconvolution algorithm that is numerically efficient, robust against noise and comes with rigorous recovery guarantees under subspace conditions.

- Can we remove the regularizers $G(h, x)$ in the blind deconvolution?
- Can we generalize it to blind-deconvolution-blind-demixing problem, i.e., $y = \sum_{i=1}^{r} \text{diag}(B_i h_i) A_i x_i$?
- Can we show if similar result holds for other types of $A$?
- What if $x$ or $h$ is sparse/both of them are sparse?
- Better choice of $B$ in image deblurring?