Implicit Regularization in Nonconvex Statistical Estimation

Yuxin Chen
Electrical Engineering, Princeton University
Nonconvex problems are everywhere

Maximum likelihood is usually nonconvex

$$\max_x \ell(x; y) \rightarrow \text{may be nonconvex}$$
subj. to $$x \in S \rightarrow \text{may be nonconvex}$$
Nonconvex problems are everywhere

Maximum likelihood is usually nonconvex

\[
\begin{align*}
\text{maximize}_x & \quad \ell(x; y) \quad \rightarrow \quad \text{may be nonconvex} \\
\text{subj. to} & \quad x \in S \quad \rightarrow \quad \text{may be nonconvex}
\end{align*}
\]

- low-rank matrix completion
- graph clustering
- dictionary learning
- mixture models
- deep learning
- ...
Nonconvex optimization may be super scary

There may be bumps everywhere and exponentially many local optima

e.g. 1-layer neural net (Auer, Herbster, Warmuth ’96; Vu ’98)
Nonconvex optimization may be super scary

There may be bumps everywhere and exponentially many local optima

e.g. 1-layer neural net (Auer, Herbster, Warmuth ’96; Vu ’98)
... but is sometimes much nicer than we think

Under certain statistical models, we see benign global geometry: **no spurious local optima**

*Fig credit: Sun, Qu & Wright*
nonconvex optimization
statistical models
exploit geometry
nonconvex optimization

exploit geometry

statistical models

nonconvex optimization
Nonconvex algorithms

- Optimization-based methods
  - gradient descent, mirror descent, ADMM, ...
Nonconvex algorithms

- Optimization-based methods
  - gradient descent, mirror descent, ADMM, ...

- Proper regularization
  - better exploit statistical / geometric properties
Optimization-based methods: two-stage approach

• Start from an appropriate initial point
Optimization-based methods: two-stage approach

- Start from an appropriate initial point
- Proceed via some iterative optimization algorithms

initial guess $z^0$

$\mathbf{x}$

basin of attraction

$\mathbf{z}^0$

$\mathbf{z}^1$

$\mathbf{z}^2$

$\mathbf{x}$

basin of attraction
 Roles of regularization

- Prevents overfitting and improves generalization
  - e.g. lasso, SCAD, nuclear norm penalization, ...
Roles of regularization

• Prevents overfitting and improves generalization
  ◦ e.g. lasso, SCAD, nuclear norm penalization, ...

• Improves computation by keeping trajectory in good region
  ◦ e.g. trimming, projection, regularized loss
Roles of regularization

• Prevents overfitting and improves generalization
  ◦ e.g. lasso, SCAD, nuclear norm penalization, ...

• Improves computation by keeping trajectory in good region
  ⇒ focus of this talk
  ◦ e.g. trimming, projection, regularized loss
3 representative nonconvex problems

- Phase retrieval / solving random quadratic systems

find $x \in \mathbb{R}^n$ s.t. $(a_k^\top x)^2 = y_k, \ 1 \leq k \leq m$
3 representative nonconvex problems

- Phase retrieval / solving random quadratic systems

  \[
  \text{find } \mathbf{x} \in \mathbb{R}^n \quad \text{s.t.} \quad (\mathbf{a}_k^\top \mathbf{x})^2 = y_k, \quad 1 \leq k \leq m
  \]

- Low-rank matrix completion

  \[
  \text{find } \mathbf{X} \in \mathbb{R}^{n \times r} \quad \text{s.t.} \quad \mathbf{e}_j^\top \mathbf{X} \mathbf{X}^\top \mathbf{e}_k = M_{j,k}, \quad (j, k) \in \Omega
  \]
3 representative nonconvex problems

- Phase retrieval / solving random quadratic systems
  \[
  \text{find } x \in \mathbb{R}^n \text{ s.t. } (a_k^T x)^2 = y_k, \quad 1 \leq k \leq m
  \]

- Low-rank matrix completion
  \[
  \text{find } X \in \mathbb{R}^{n \times r} \text{ s.t. } e_j^T X X^T e_k = M_{j,k}, \quad (j, k) \in \Omega
  \]

- Blind deconvolution / solving random bilinear systems
  \[
  \text{find } h, x \in \mathbb{C}^n \text{ s.t. } b_k^* h_k x_k^* a_k = y_k, \quad 1 \leq k \leq m
  \]
### 3 representative nonconvex problems

- **Phase retrieval / solving random quadratic systems**
  
  \[
  \text{find } \quad \mathbf{x} \in \mathbb{R}^n \quad \text{s.t.} \quad (\mathbf{a}_k^\top \mathbf{x})^2 = y_k, \quad 1 \leq k \leq m
  \]

- **Low-rank matrix completion**
  
  \[
  \text{find } \quad \mathbf{X} \in \mathbb{R}^{n \times r} \quad \text{s.t.} \quad \mathbf{e}_j^\top \mathbf{X} \mathbf{X}^\top \mathbf{e}_k = M_{j,k}, \quad (j, k) \in \Omega
  \]

- **Blind deconvolution / solving random bilinear systems**
  
  \[
  \text{find } \quad \mathbf{h}, \mathbf{x} \in \mathbb{C}^n \quad \text{s.t.} \quad b_k^* \mathbf{h}_k \mathbf{x}_k^* \mathbf{a}_k = y_k, \quad 1 \leq k \leq m
  \]
Missing phase problem

Detectors record **intensities** of diffracted rays

- electric field \( x(t_1, t_2) \) \( \rightarrow \) Fourier transform \( \hat{x}(f_1, f_2) \)

\[ |\hat{x}(f_1, f_2)|^2 = \left| \int x(t_1, t_2) e^{-i2\pi(f_1t_1+f_2t_2)} dt_1 dt_2 \right|^2 \]

**Fig credit: Stanford SLAC**
Missing phase problem

Detectors record intensities of diffracted rays

- electric field \( x(t_1, t_2) \) \( \rightarrow \) Fourier transform \( \hat{x}(f_1, f_2) \)

\[
\text{intensity of electrical field: } |\hat{x}(f_1, f_2)|^2 = \left| \int x(t_1, t_2) e^{-i2\pi(f_1 t_1 + f_2 t_2)} dt_1 dt_2 \right|^2
\]

Phase retrieval: recover signal \( x(t_1, t_2) \) from intensity \( |\hat{x}(f_1, f_2)|^2 \)
Estimate $x^\dagger \in \mathbb{R}^n$ from $m$ random quadratic samples

\[ y_k = |a_k^\top x^\dagger|^2, \quad k = 1, \ldots, m \]

or \[ y = |Ax^\dagger|^2 \]

where $|z|^2 := [|z_1|^2, \ldots, |z_m|^2]^\top$

*Assume w.l.o.g. $\|x^\dagger\|_2 = 1$*
Wirtinger flow (Candès, Li, Soltanolkotabi ’14)

Empirical loss minimization

\[
\text{minimize}_x \quad f(x) = \frac{1}{m} \sum_{k=1}^{m} \left[ (a_k^\top x)^2 - y_k \right]^2
\]
Wirtinger flow (Candès, Li, Soltanolkotabi ’14)

Empirical loss minimization

\[
\text{minimize}_x \quad f(x) = \frac{1}{m} \sum_{k=1}^{m} \left( (a_k^\top x)^2 - y_k \right)^2
\]

- Initialization by spectral method

- Gradient iterations: for \( t = 0, 1, \ldots \)

\[
x^{t+1} = x^t - \eta_t \nabla f(x^t)
\]
Gradient descent theory revisited

Two standard conditions that enable geometric convergence of GD
Gradient descent theory revisited

Two standard conditions that enable geometric convergence of GD

- (local) regularity condition (or restricted strong convexity)
Gradient descent theory revisited

Two standard conditions that enable geometric convergence of GD

• (local) regularity condition (or restricted strong convexity)
• (local) smoothness

\[ \nabla^2 f(x) \succ 0 \] and is well-conditioned
Gradient descent theory revisited

$f$ is said to be $\alpha$-strongly convex and $\beta$-smooth if

$$0 \preceq \alpha I \preceq \nabla^2 f(x) \preceq \beta I, \quad \forall x$$

$l_2$ error contraction: GD with $\eta = 1/\beta$ obeys

$$\|x^{t+1} - x^\dagger\|_2 \leq \left(1 - \frac{\alpha}{\beta}\right)\|x^t - x^\dagger\|_2$$
Gradient descent theory revisited

$f$ is said to be $\alpha$-strongly convex and $\beta$-smooth if

\[ 0 \leq \alpha I \preceq \nabla^2 f(x) \preceq \beta I, \quad \forall x \]

**$\ell_2$ error contraction:** GD with $\eta = 1/\beta$ obeys

\[ \|x^{t+1} - x^\#\|_2 \leq \left(1 - \frac{\alpha}{\beta}\right) \|x^t - x^\#\|_2 \]
Gradient descent theory revisited

\[ \| x^{t+1} - x^\dagger \|_2 \leq (1 - \alpha/\beta) \| x^t - x^\dagger \|_2 \]

- If this nice region is $\ell_2$ ball and if we start within this region, then GD converges fast
Gradient descent theory revisited

\[ \| x^{t+1} - x^\dag \|_2 \leq (1 - \alpha/\beta) \| x^t - x^\dag \|_2 \]

- region of local strong convexity + smoothness

- If this nice region is $\ell_2$ ball and if we start within this region, then GD converges fast
Gradient descent theory revisited

\[ \|x^{t+1} - x^\#\|_2 \leq (1 - \alpha/\beta) \|x^t - x^\#\|_2 \]

- region of local strong convexity + smoothness

- If this nice region is $\ell_2$ ball and if we start within this region, then GD converges fast
Gradient descent theory revisited

\[ \|x^{t+1} - x^\dagger\|_2 \leq (1 - \alpha/\beta) \|x^t - x^\dagger\|_2 \]

- region of local strong convexity + smoothness

- If this nice region is $\ell_2$ ball and if we start within this region, then GD converges fast
Gradient descent theory revisited

\[ 0 \leq \alpha I \leq \nabla^2 f(x) \leq \beta I, \quad \forall x \]

\textbf{$\ell_2$ error contraction:} GD with $\eta = 1/\beta$ obeys

\[ \|x^{t+1} - x^\#\|_2 \leq \left(1 - \frac{\alpha}{\beta}\right) \|x^t - x^\#\|_2 \]

- Condition number $\beta/\alpha$ determines rate of convergence
Gradient descent theory revisited

\[ 0 \preceq \alpha I \preceq \nabla^2 f(x) \preceq \beta I, \quad \forall x \]

\textbf{\( \ell_2 \) error contraction:} GD with \( \eta = 1/\beta \) obeys

\[ \| x^{t+1} - x^\dagger \|_2 \leq \left( 1 - \frac{\alpha}{\beta} \right) \| x^t - x^\dagger \|_2 \]

- Condition number \( \beta/\alpha \) determines rate of convergence
- Attains \( \varepsilon \)-accuracy within \( O\left( \frac{\beta}{\alpha} \log \frac{1}{\varepsilon} \right) \) iterations
What does this optimization theory say about WF?

Gaussian designs: \( a_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I), \quad 1 \leq k \leq m \)
What does this optimization theory say about WF?

Gaussian designs: $a_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I), \quad 1 \leq k \leq m$

Population level (infinite samples)

$$
\mathbb{E}[\nabla^2 f(x)] = 3 \left( \|x\|_2^2 I + 2xx^\top \right) - \left( \|x^\parallel_2^2 I + 2x^\parallel x^\parallel^\top \right)
$$

locally positive definite and well-conditioned

Consequence: WF converges within logarithmic iterations if $m \to \infty$
What does this optimization theory say about WF?

\[
\text{Gaussian designs: } a_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I), \quad 1 \leq k \leq m
\]

Finite-sample level \((m \asymp n \log n)\)

\[
\nabla^2 f(x) \succ 0
\]
What does this optimization theory say about WF?

**Gaussian designs:** $a_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I), \quad 1 \leq k \leq m$

**Finite-sample level** ($m \asymp n \log n$)

$\nabla^2 f(x) \succ 0$ \hspace{1cm} but ill-conditioned \hspace{1cm} (even locally)

\hspace{2cm} condition number $\asymp n$
What does this optimization theory say about WF?

Gaussian designs: $a_k \sim \mathcal{N}(0, I), \ 1 \leq k \leq m$

Finite-sample level ($m \asymp n \log n$)

$$\nabla^2 f(x) \succ 0 \text{ but ill-conditioned} \quad \text{(even locally)}$$

condition number $\asymp n$

Consequence (Candès et al ’14): WF converges within $O(n)$ iterations if $m \asymp n \log n$
What does this optimization theory say about WF?

Gaussian designs: $a_k \overset{i.i.d.}{\sim} \mathcal{N}(0, I), \quad 1 \leq k \leq m$

Finite-sample level ($m \asymp n \log n$)

$$\nabla^2 f(x) \succ 0 \quad \text{but ill-conditioned} \quad \text{(even locally)}$$

Condition number $\asymp n$

Consequence (Candès et al ’14): WF converges within $O(n)$ iterations if $m \asymp n \log n$

Too slow ... can we accelerate it?
One solution: truncated WF (Chen, Candès ’15)

Regularize / trim gradient components to accelerate convergence
One solution: truncated WF (Chen, Candès ’15)

Regularize / trim gradient components to accelerate convergence

WF: $\eta_t = O(1/n)$
One solution: truncated WF (Chen, Candès ’15)

Regularize / trim gradient components to accelerate convergence

WF: $\eta_t = O(1/n)$ \hspace{1cm} WF with trimming: $\eta_t = O(1)$
One solution: truncated WF (Chen, Candès ’15)

Regularize / trim gradient components to accelerate convergence

WF: \( \eta_t = O(1/n) \)
WF with trimming: \( \eta_t = O(1) \)

With better-controlled search directions, one can proceed much faster
But wait a minute ...

WF converges in $O(n)$ iterations
But wait a minute ...

WF converges in $O(n)$ iterations

Step size taken to be $\eta_t = O(1/n)$
But wait a minute ...

WF converges in \( O(n) \) iterations

Step size taken to be \( \eta_t = O(1/n) \)

This choice is suggested by generic optimization theory
WF converges in $O(n)$ iterations

Step size taken to be $\eta_t = O(1/n)$

This choice is suggested by worst-case optimization theory
But wait a minute ...

WF converges in $O(n)$ iterations

Step size taken to be $\eta_t = O(1/n)$

This choice is suggested by worst-case optimization theory

Does it capture what really happens?
Numerical surprise with $\eta_t = 0.1$

Vanilla GD (WF) can proceed much more aggressively!
A second look at gradient descent theory

Which region enjoys both strong convexity and smoothness?

\[
\nabla^2 f(x) = \frac{1}{m} \sum_{k=1}^{m} \left[ 3(a_k^\top x)^2 - (a_k^\top x^\dagger)^2 \right] a_k a_k^\top
\]
A second look at gradient descent theory

Which region enjoys both strong convexity and smoothness?

\[ \nabla^2 f(x) = \frac{1}{m} \sum_{k=1}^{m} \left[ 3(a_k^\top x)^2 - (a_k^\top x^{\#})^2 \right] a_k a_k^\top \]

- Not smooth if \( x \) and \( a_k \) are too close (coherent)
A second look at gradient descent theory

Which region enjoys both strong convexity and smoothness?

- $x$ is not far away from $x^\dagger$

\[
\begin{align*}
\nabla^2 f(x) &= \frac{1}{m} \sum_{k=1}^{m} [3 (a_k^\top x)^2 - (a_k^\top x^\dagger)^2] a_k a_k^\top \\
\text{Not smooth if } x \text{ and } a_k \text{ are too close (coherent)}
\end{align*}
\]
A second look at gradient descent theory

Which region enjoys both strong convexity and smoothness?

- $x$ is not far away from $x^\dagger$
- $x$ is incoherent w.r.t. sampling vectors (incoherence region)
A second look at gradient descent theory

Which region enjoys both strong convexity and smoothness?

- $x$ is not far away from $x^\dagger$
- $x$ is incoherent w.r.t. sampling vectors (incoherence region)
A second look at gradient descent theory

- region of local strong convexity + smoothness

- Prior theory only ensures that iterates remain in $\ell_2$ ball but not incoherence region
A second look at gradient descent theory

- region of local strong convexity $+$ smoothness

- Prior theory only ensures that iterates remain in $\ell_2$ ball but not incoherence region
A second look at gradient descent theory

- region of local strong convexity + smoothness

- Prior theory only ensures that iterates remain in $\ell_2$ ball but not incoherence region
A second look at gradient descent theory

- region of local strong convexity + smoothness

- Prior theory only ensures that iterates remain in $\ell_2$ ball but not incoherence region
A second look at gradient descent theory

- region of local strong convexity + smoothness

- Prior theory only ensures that iterates remain in $\ell_2$ ball but not incoherence region
A second look at gradient descent theory

- region of local strong convexity + smoothness

- Prior theory only ensures that iterates remain in $\ell_2$ ball but not incoherence region
A second look at gradient descent theory

- region of local strong convexity + smoothness

- Prior theory only ensures that iterates remain in \( \ell_2 \) ball but not incoherence region
A second look at gradient descent theory

- region of local strong convexity + smoothness

- Prior theory only ensures that iterates remain in $\ell_2$ ball but not incoherence region
A second look at gradient descent theory

• Prior theory only ensures that iterates remain in $\ell_2$ ball but not incoherence region

• Prior theory enforces regularization to promote incoherence
Our findings: GD is implicitly regularized

region of local strong convexity + smoothness
Our findings: GD is implicitly regularized

- region of local strong convexity + smoothness
Our findings: GD is implicitly regularized

region of local strong convexity + smoothness
Our findings: GD is implicitly regularized

region of local strong convexity + smoothness
Our findings: GD is implicitly regularized

- region of local strong convexity + smoothness

GD implicitly forces iterates to remain incoherent
Theoretical guarantees

Theorem 1 (Phase retrieval)

Under i.i.d. Gaussian design, WF achieves

- \[ |\mathbf{a}_k^\top (\mathbf{x}^t - \mathbf{x}^\natural)\| \lesssim \sqrt{\log n} \quad \text{(incoherence)} \]
- \[ \|\mathbf{x}^t - \mathbf{x}^\natural\|_2 \lesssim (1 - \eta^2)^t \|\mathbf{x}^\natural\|_2 \quad \text{(near-linear convergence)} \]

provided that step size \( \eta \lesssim \frac{1}{\log n} \) and sample size \( m \gtrsim n \log n \).
Theoretical guarantees

Theorem 1 (Phase retrieval)
Under i.i.d. Gaussian design, WF achieves

- $|\mathbf{a}_k^\top (\mathbf{x}^t - \mathbf{x}^\dagger)| \lesssim \sqrt{\log n}$ (incoherence)
- $\|\mathbf{x}^t - \mathbf{x}^\dagger\|_2 \lesssim (1 - \frac{\eta}{2})^t \|\mathbf{x}^\dagger\|_2$ (near-linear convergence)

provided that step size $\eta \lesssim \frac{1}{\log n}$ and sample size $m \gtrsim n \log n$. 
Theoretical guarantees

Theorem 1 (Phase retrieval)

Under i.i.d. Gaussian design, WF achieves

- $|a_k^\top (x^t - x^\#)| \lesssim \sqrt{\log n}$ (incoherence)
- $\|x^t - x^\#\|_2 \lesssim (1 - \frac{\eta}{2})^t \|x^\#\|_2$ (near-linear convergence)

provided that step size $\eta \lesssim \frac{1}{\log n}$ and sample size $m \gtrsim n \log n$.

- Step size: $\frac{1}{\log n}$ (vs. $\frac{1}{n}$)
Theoretical guarantees

**Theorem 1 (Phase retrieval)**

Under i.i.d. Gaussian design, WF achieves

- \(|a_k^T(x^t - x^h)| \lesssim \sqrt{\log n}\) (incoherence)
- \(\|x^t - x^h\|_2 \lesssim (1 - \eta^2)^t \|x^h\|_2\) (near-linear convergence)

provided that step size \(\eta \lesssim \frac{1}{\log n}\) and sample size \(m \gtrsim n \log n\).

- Step size: \(\frac{1}{\log n}\) (vs. \(\frac{1}{n}\))
- Computational complexity: \(n/\log n\) times faster than existing theory for WF
Key ingredient: leave-one-out analysis

For each $1 \leq l \leq m$, introduce leave-one-out iterates $x^{t,(l)}$ by dropping $l$th measurement

$$A^{(l)} \quad x \quad A^{(l)}x$$

$$y^{(l)} = |A^{(l)}x|^2$$
Key ingredient: leave-one-out analysis

- Leave-one-out iterates $x_{t,(l)}$ are independent of $a_l$, and are hence **incoherent** w.r.t. $a_l$ with high prob.
Key ingredient: leave-one-out analysis

- Leave-one-out iterates $x^{t,(l)}$ are independent of $a_l$, and are hence incoherent w.r.t. $a_l$ with high prob.
- Leave-one-out iterates $x^{t,(l)}$ and true iterates $x^t$ are very close
This recipe is quite general
Low-rank matrix completion

Given partial samples of a *low-rank* matrix $\mathbf{M}$, fill in missing entries.

Fig. credit: Candès
Prior art

\[
\text{minimize}_X \quad f(X) = \sum_{(j,k) \in \Omega} \left( e_j X X^\top e_k - M_{j,k} \right)^2
\]
Prior art

\[ \text{minimize}_X \quad f(X) = \sum_{(j,k) \in \Omega} \left( e_j^\top X X^\top e_k - M_{j,k} \right)^2 \]

Existing theory on gradient descent requires

- regularized loss (solve \( \min_x f(x) + R(x) \) instead)
- projection onto set of incoherent matrices
Prior art

\[
\text{minimize}_X \quad f(X) = \sum_{(j,k) \in \Omega} \left( e_j^\top X X^\top e_k - M_{j,k} \right)^2
\]

Existing theory on gradient descent requires

- regularized loss (solve \( \min_x f(x) + R(x) \) instead)
  - e.g. Keshavan, Montanari, Oh '10, Sun, Luo '14, Ge, Lee, Ma '16
Prior art

\[
\text{minimize}_X \quad f(X) = \sum_{(j,k) \in \Omega} \left( e_j^\top XX^\top e_k - M_{j,k} \right)^2
\]

Existing theory on gradient descent requires

- regularized loss (solve \( \min_x f(x) + R(x) \) instead)
  - e.g. Keshavan, Montanari, Oh ’10, Sun, Luo ’14, Ge, Lee, Ma ’16

- projection onto set of incoherent matrices
  - e.g. Chen, Wainwright ’15
Theoretical guarantees

**Theorem 2 (Matrix completion)**

Suppose $\mathbf{M}$ is rank-$r$, incoherent and well-conditioned. *Vanilla gradient descent* (with spectral initialization) achieves

- geometric convergence

if step size $\eta \lesssim 1/\sigma_{\text{max}}(\mathbf{M})$ and sample size $\gtrsim nr^3 \log^3 n$.
Theoretical guarantees

Theorem 2 (Matrix completion)

Suppose $\mathcal{M}$ is rank-$r$, incoherent and well-conditioned. Vanilla gradient descent (with spectral initialization) achieves

- geometric convergence w.r.t. $\|\cdot\|_F$, $\|\cdot\|$, and $\|\cdot\|_{2,\infty}$

if step size $\eta \lesssim \frac{1}{\sigma_{\text{max}}(\mathcal{M})}$ and sample size $\gtrsim nr^3 \log^3 n$
Theoretical guarantees

**Theorem 2 (Matrix completion)**

Suppose $M$ is rank-$r$, incoherent and well-conditioned. \(\text{Vanilla gradient descent (with spectral initialization) achieves}

- geometric convergence w.r.t. $\| \cdot \|_{F}$, $\| \cdot \|$, and $\| \cdot \|_{2,\infty}$ incoherence

if step size $\eta \lesssim 1/\sigma_{\text{max}}(M)$ and sample size $\gtrsim nr^3 \log^3 n$

- Regularization-free
Theoretical guarantees

**Theorem 2 (Matrix completion)**

Suppose $M$ is rank-$r$, incoherent and well-conditioned. *Vanilla gradient descent (with spectral initialization)* achieves

- **geometric convergence** w.r.t. $\| \cdot \|_F$, $\| \cdot \|$, and $\| \cdot \|_2,\infty$

if step size $\eta \lesssim 1 / \sigma_{\text{max}}(M)$ and sample size $\gtrsim nr^3 \log^3 n$

- Regularization-free
- *Byproduct: vanilla GD controls entrywise error* — errors are spread out across all entries
Reconstruct two signals from their convolution; equivalently,

\[
\text{find } \ h, \ x \in \mathbb{C}^n \ \text{ s.t. } \ b_k^* h_k x_k^* a_k = y_k, \quad 1 \leq k \leq m
\]
Prior art

\[
\text{minimize}_{x, h} \quad f(x, h) = \sum_{k=1}^{m} \left| b_k^* \left( h x^* - h^\dagger x^\dagger \right) a_k \right|^2
\]

\( a_k \sim \text{i.i.d. } \mathcal{N}(0, I) \) \quad and \quad \{b_k\} : \text{partial Fourier basis}
Prior art

\[
\minimize_{x, h} \quad f(x, h) = \sum_{k=1}^{m} \left| b_k^* \left( hx^* - h^{\dagger} x^{\dagger*} \right) a_k \right|^2
\]

\[a_k \sim \mathcal{N}(0, I)\]

and \{b_k\} : partial Fourier basis

Existing theory on gradient descent requires

- regularized loss + projection
  - e.g. Li, Ling, Strohmer, Wei ’16, Huang, Hand ’17, Ling, Strohmer ’17
Prior art

\[
\text{minimize}_{x,h} \quad f(x, h) = \sum_{k=1}^{m} \left| b_k^* \left( hx^* - h^\dagger x^\dagger \right) a_k \right|^2
\]

\[a_k \text{ i.i.d. } \mathcal{N}(0, I) \quad \text{and} \quad \{b_k\} : \text{partial Fourier basis}\]

Existing theory on gradient descent requires

- regularized loss + projection
  - e.g. Li, Ling, Strohmer, Wei '16, Huang, Hand '17, Ling, Strohmer '17
  - requires \(m\) iterations even with regularization
Theoretical guarantees

Theorem 3 (Blind deconvolution)

Suppose $\mathbf{h}^\dagger$ is incoherent w.r.t. $\{\mathbf{b}_k\}$. Vanilla gradient descent (with spectral initialization) achieves

- geometric convergence

provided that step size $\eta \lesssim 1$ and sample size $m \gtrsim n \text{poly} \log(m)$.

- Regularization-free
- Converges in logarithmic iterations (vs. $O(m)$ iterations in prior theory)
Summary

• **Blessings of randomness:** nonconvex statistical optimization may not be that scary

---

Paper: "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution," Cong Ma, Kaizheng Wang, Yuejie Chi, Yuxin Chen.
Summary

• **Blessings of randomness**: nonconvex statistical optimization may not be that scary

• **Implicit regularization**: vanilla gradient descent automatically forces iterates to stay *incoherent*
Summary

- **Blessings of randomness:** nonconvex statistical optimization may not be that scary
- **Implicit regularization:** vanilla gradient descent automatically forces iterates to stay *incoherent*
- Enable error controls in a much stronger sense (e.g. *entrywise error control*)

**Paper:**

“Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution”, Cong Ma, Kaizheng Wang, Yuejie Chi, Yuxin Chen.