Implicit Regularization in Nonconvex Statistical Estimation

Yuejie Chi

Department of Electrical and Computer Engineering

Carnegie Mellon University

May 2018
Acknowledgements

Thanks to my collaborators:

Y. Chen  
Princeton

J. Fan  
Princeton

C. Ma  
Princeton

K. Wang  
Princeton

Y. Li  
CMU

This research is supported by NSF, ONR and AFOSR.
Nonconvex problems are abundant

Empirical risk minimization is usually nonconvex

\[
\text{minimize}_{x} \quad \ell(y; x) \quad \rightarrow \quad \text{nonconvex}
\]
Nonconvex problems are abundant

Empirical risk minimization is usually nonconvex

$$\minimize_{x} \ell(y; x) \rightarrow \text{nonconvex}$$

- low-rank matrix completion
- phase retrieval
- dictionary learning
- blind deconvolution
- mixture models
- deep learning
- ...
Nonconvex optimization is daunting in theory

There may be exponentially many local optima

e.g. a single neuron model (Auer, Herbster, Warmuth ’96)
Exponentially many local minima for perceptron

Given training data \( \{x_i, y_i\}_{i=1}^n \),

\[
\text{minimize}_{\mathbf{w} \in \mathbb{R}^d} \quad \ell_n(\mathbf{w}) := \frac{1}{2n} \sum_{i=1}^n \left( y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i) \right)^2
\]

Theorem (Auer et al., 1995)

Let \( \sigma(\cdot) \) be sigmoid and \( \ell(\cdot) \) be the quadratic loss function. There exists a sequence of training samples \( \{x_i, y_i\}_{i=1}^n \) such that \( \ell_n(\mathbf{w}) \) has \( \lfloor nd \rfloor \) distinct local minima.

No. of local minima grows exponentially with the dimension \( d \).
Exponentially many local minima for perceptron

Given training data \( \{ \mathbf{x}_i, y_i \}_{i=1}^n \),

\[
\text{minimize}_{\mathbf{w} \in \mathbb{R}^d} \quad \ell_n(\mathbf{w}) := \frac{1}{2n} \sum_{i=1}^{n} \left( y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i) \right)^2
\]

Theorem (Auer et al., 1995)

Let \( \sigma(\cdot) \) be sigmoid and \( \ell(\cdot) \) be the quadratic loss function. There exists a sequence of training samples \( \{ \mathbf{x}_i, y_i \}_{i=1}^n \) such that \( \ell_n(\mathbf{w}) \) has \( \left\lfloor \frac{n}{d} \right\rfloor^d \) distinct local minima.

No. of local minima grows exponentially with the dimension \( d \)!
Nonconvex optimization is daunting in theory

There may be exponentially many local optima

e.g. a single neuron model (Auer, Herbster, Warmuth ’96)
Nonconvex optimization is daunting in theory

There may be exponentially many local optima

e.g. a single neuron model (Auer, Herbster, Warmuth ’96)
But they’re solved on a daily basis in practice

Using simple algorithms such as gradient descent, e.g., “back propagation” for training deep neural networks...
Statistical models come to rescue

Data/measurements follow certain **statistical models** and hence are not worst-case instances.

\[
\text{minimize}_{x} \ f(x) = \frac{1}{m} \sum_{i=1}^{m} \ell(y_i; x)
\]
Statistical models come to rescue

Data/measurements follow certain **statistical models** and hence are not worst-case instances.

\[
\text{minimize}_x f(x) = \frac{1}{m} \sum_{i=1}^{m} \ell(y_i; x) \quad \xrightarrow{m \to \infty} \quad \mathbb{E}[\ell(y; x)]
\]
Statistical models come to rescue

Data/measurements follow certain statistical models and hence are not worst-case instances.

\[
\text{minimize}_{\mathbf{x}} \, f(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} \ell(y_i; \mathbf{x}) \quad \xrightarrow{m \to \infty} \quad \mathbb{E}[\ell(y; \mathbf{x})]
\]

empirical risk \approx population risk (often nice!)

\[\theta_0 = [1, 0], \quad \hat{\theta}_n = [0.816, -0.268]\]

Figure credit: Mei, Bai and Montanari
Putting together...

statistical models

benign landscape

global convergence
Computational efficiency?

- statistical models
- benign landscape
- global convergence

But how fast?
What we know in theory

Statistical: efficient

Critical points

Sample complexity

Computational: inefficient

(saddle point, nonsmooth)
What we know in theory

Statistical:
- efficient

Critical points

Computational:
- inefficient
  - (saddle point, nonsmooth)

Smoothness

Sample complexity
What we know in theory

Statistical:
- Efficient
- Inefficient

Critical points
- Sample complexity

Computational:
- Inefficient (saddle point, nonsmooth)
- Efficient

Sample complexity
What we know in theory

Statistical:
- Efficient
- Inefficient
  - Critical points
  - Smoothness

Computational:
- Inefficient
  - (saddle point, nonsmooth)
- Efficient
  - Regularized
  - Unregularized

Can we simultaneously achieve statistical and computational efficiency using unregularized methods?
What we know in theory

**Statistical:**
- Efficient
- Inefficient
- Critical points
- Smoothness

**Computational:**
- Inefficient
  - (saddle point, nonsmooth)
- Efficient
  - Regularized
  - Unregularized
  - Efficiency?

Can we simultaneously achieve statistical and computational efficiency using unregularized methods?
Three problems I care about

\[ \mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t) \]

phase retrieval
matrix completion
blind deconvolution
Regularized gradient descent

\[ \mathbf{x}^{t+1} = \mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t) \]

phase retrieval

matrix completion

blind deconvolution

regularized trimming

regularized cost projection

regularized cost projection

This talk: vanilla gradient descent runs as fast as regularized ones!
Regularized vs. unregularized gradient descent

\[ \mathbf{x}^{t+1} = \mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t) \]

- **Phase retrieval**
  - Regularized
  - Unregularized
  - Suboptimal
  - Computational cost

- **Matrix completion**
  - Regularized
  - Unregularized
  - Regularized cost
  - Projection
  - Temperature

- **Blind deconvolution**
  - Regularized
  - Unregularized
  - Regularized cost
  - Projection
  - Temperature
Regularized vs. unregularized gradient descent

\[ \mathbf{x}^{t+1} = \mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t) \]

This talk: vanilla gradient descent runs as fast as regularized ones!
Set $X^\downarrow = [x_1, x_2, \ldots, x_r]$, then

$$y = \sum_{i=1}^{r} \sigma(a^\top x_i).$$
Shallow neural network with quadratic activation

Set $X^\|$ = $[x_1, x_2, \ldots, x_r]$, then

$$y = \sum_{i=1}^{r} \sigma(a^\top x_i) \sigma(z) = z^2 \sum_{i=1}^{r} (a^\top x_i)^2 = \|a^\top X^\|_2^2.$$
Generalized phase retrieval

\[ A \times X = AX \]

Recover \( X^\dagger \in \mathbb{R}^{n \times r} \) from \( m \) “random” quadratic measurements

\[ y_i = \left\| a_i^\top X^\dagger \right\|_2^2, \quad i = 1, \ldots, m \]
Single neuron with quadratic activation

\[ y = |Ax|^2 \]

Recover \( x^\dagger \in \mathbb{R}^n \) from \( m \) “random” quadratic measurements

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

where \( m \) is about as large as \( n \).

Assume w.l.o.g. \( \|x^\dagger\|_2 = 1 \)
A natural least squares formulation

given: \[ y_k = |a_k^T x|^2, \quad 1 \leq k \leq m \]

\[ \downarrow \]

\[ \text{minimize}_{x \in \mathbb{R}^n} \quad f(x) = \frac{1}{4m} \sum_{k=1}^{m} \left[ |a_k^T x|^2 - y_k \right]^2 \]
A natural least squares formulation

given: \[ y_k = \| a_k^\top x \|^2, \quad 1 \leq k \leq m \]

\[ \downarrow \]

minimize_{x \in \mathbb{R}^n} \quad f(x) = \frac{1}{4m} \sum_{k=1}^{m} \left[ \| a_k^\top x \|^2 - y_k \right]^2

- **pros:** global minimizers are the truth as long as sample size is sufficiently large
A natural least squares formulation

given: \[ y_k = |a_k^\top x|^2, \quad 1 \leq k \leq m \]

\[ \Downarrow \]

minimize \( x \in \mathbb{R}^n \) \[ f(x) = \frac{1}{4m} \sum_{k=1}^{m} \left[ |a_k^\top x|^2 - y_k \right]^2 \]

• **pros:** global minimizers are the truth as long as sample size is sufficiently large

• **cons:** \( f(\cdot) \) is nonconvex
  \[ \longrightarrow \text{computationally challenging!} \]
Two-step nonconvex procedure

\[ \hat{x} = \arg\min_x f(x) := \frac{1}{m} \sum_{i=1}^{m} \ell(y_i; x) \]

- Initialize \( x^0 \) via spectral methods properly;
- Update using simple iterative methods, e.g. gradient descent.
Two-step nonconvex procedure

\[ \hat{x} = \arg\min_x f(x) := \frac{1}{m} \sum_{i=1}^{m} \ell(y_i; x) \]

- Initialize \( x^0 \) via *spectral* methods properly;
- Update using *simple* iterative methods, e.g. gradient descent.
Wirtinger flow (Candès, Li, Soltanolkotabi ’14)

Empirical risk minimization

\[
\min_{\mathbf{x}} \quad f(\mathbf{x}) = \frac{1}{4m} \sum_{k=1}^{m} \left[ (\mathbf{a}_k^\top \mathbf{x})^2 - y_k \right]^2
\]
Wirtinger flow (Candès, Li, Soltanolkotabi ’14)

Empirical risk minimization

$$\text{minimize}_x \quad f(x) = \frac{1}{4m} \sum_{k=1}^{m} \left[ (a_k^T x)^2 - y_k \right]^2$$

- Initialization by spectral method

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

$$x^{t+1} = x^t - \eta \nabla f(x^t)$$
Gradient descent theory revisited

Two standard conditions that enable geometric convergence of GD at least along certain descent directions.
Gradient descent theory revisited

Two standard conditions that enable geometric convergence of GD

• (local) restricted strong convexity

at least along certain descent directions.
Gradient descent theory revisited

Two standard conditions that enable geometric convergence of GD

- (local) restricted strong convexity
- (local) smoothness

at least along certain descent directions.
Gradient descent theory revisited

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

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

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

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

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

region of local strong convexity + smoothness
Gradient descent theory revisited

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

region of local strong convexity + smoothness
Gradient descent theory revisited

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

- region of local strong convexity + smoothness
Gradient descent theory revisited

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

region of local strong convexity + smoothness
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

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

- Condition number \(\frac{\beta}{\alpha}\) determines rate of convergence
Gradient descent theory revisited

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

**\( \ell_2 \) 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 \( \frac{\beta}{\alpha} \) 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: $a_k \sim \mathcal{N}(0, I_n)$, $1 \leq k \leq m$
What does this optimization theory say about WF?

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

Population level (infinite samples)

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

*locally* positive definite and well-conditioned

\[
\mathbf{I}_n \preceq \mathbb{E}[\nabla^2 f(\mathbf{x})] \preceq 10\mathbf{I}_n
\]

**Consequence:** WF converges within \( O(\log \frac{1}{\varepsilon}) \) iterations if

\[
m \to \infty
\]
What does this optimization theory say about WF?

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

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

\[ \nabla^2 f(\mathbf{x}) \quad \text{but ill-conditioned} \quad \text{(even locally)} \]

\[ \text{condition number} \asymp n \]
What does this optimization theory say about WF?

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

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

\[
\nabla^2 f(x) \quad \text{but ill-conditioned} \quad \text{(even locally)}
\]

Condition number \( \asymp n \)

\[
\frac{1}{2} I_n \preceq \nabla^2 f(x) \preceq O(n) I_n
\]
What does this optimization theory say about WF?

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

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

\[
\nabla^2 f(\mathbf{x}) \quad \text{but ill-conditioned} \quad (\text{even locally})
\]

Condition number \( \asymp n \)

\[
\frac{1}{2} I_n \preceq \nabla^2 f(\mathbf{x}) \preceq O(n) I_n
\]

**Consequence** (**Candès et al ’14**): WF attains \( \varepsilon \)-accuracy within \( O(n \log \frac{1}{\varepsilon}) \) iterations with \( \eta \asymp 1/n \) if \( m \asymp n \log n \)
What does this optimization theory say about WF?

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

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

\[
\nabla^2 f(x) \quad \text{but ill-conditioned} \quad \text{(even locally)}
\]

**Condition number** \( \asymp n \)

\[
\frac{1}{2} I_n \preceq \nabla^2 f(x) \preceq O(n)I_n
\]

**Consequence (Candès et al '14):** WF attains \( \varepsilon \)-accuracy within \( O(n \log \frac{1}{\varepsilon}) \) iterations with \( \eta \asymp 1/n \) if \( m \asymp n \log n \)

Too slow ...
Numerical experiment with $\eta_t = 0.1$

Vanilla GD (WF) can proceed much more aggressively!
Numerical experiment with $\eta_t = 0.1$

Generic optimization theory is too pessimistic!
A second look at gradient descent theory

Which region enjoys both strong convexity and smoothness?

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

Which region enjoys both strong convexity and smoothness?

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

- 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 $x^\dagger$

\[ \nabla^2 f(x) = \frac{1}{m} \sum_{k=1}^{m} \left[ (a^\top_k x)^2 - (a^\top_k x^\dagger)^2 \right] a_k a^\top_k \]

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?

- \( \mathbf{x} \) is not far away from \( \mathbf{x}^\dagger \)
- \( \mathbf{x} \) is incoherent w.r.t. sampling vectors (incoherence region)

\[
(1/2) \cdot \mathbf{I}_n \preceq \nabla^2 f(\mathbf{x}) \preceq O(\log n) \cdot \mathbf{I}_n
\]
A second look at gradient descent theory

Which region enjoys both strong convexity and smoothness?

- $\mathbf{x}$ is not far away from $\mathbf{x}^\ddagger$
- $\mathbf{x}$ is incoherent w.r.t. sampling vectors (incoherence region)

$$(1/2) \cdot \mathbf{I}_n \preceq \nabla^2 f(\mathbf{x}) \preceq O(\log n) \cdot \mathbf{I}_n$$
A second look at gradient descent theory

region of local strong convexity + smoothness

- Generic optimization 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

- Generic optimization 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

- Generic optimization 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

- Generic optimization 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

- Generic optimization 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

- Generic optimization 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

- Generic optimization 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

- Generic optimization 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

- Generic optimization theory only ensures that iterates remain in $\ell_2$ ball but not incoherence region

- Existing algorithms enforce regularization, or apply sample splitting 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 even without regularization
Theoretical guarantees

**Theorem (Phase retrieval)**

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

- \( \max_k |a_k^\top (x^t - x^\natural)| \lesssim \sqrt{\log n} \|x^\natural\|_2 \) (incoherence)

- \( \|x^t - x^\natural\|_2 \lesssim (1 - \eta^2) t \|x^\natural\|_2 \) (near-linear convergence)

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

**Big computational saving:** WF attains \( \epsilon \)-accuracy within \( O(\log n \log \frac{1}{\epsilon}) \) iterations with \( \eta \approx \frac{1}{\log n} \) if \( m \approx n \log n \).
Theoretical guarantees

Theorem (Phase retrieval)

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

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

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

Theorem (Phase retrieval)

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

- \[ \max_k | a_k^\top (x^t - \hat{x})| \lesssim \sqrt{\log n} \| \hat{x} \|_2 \text{ (incoherence)} \]
- \[ \| x^t - \hat{x} \|_2 \lesssim (1 - \frac{\eta}{2})^t \| \hat{x} \|_2 \text{ (near-linear convergence)} \]

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

Big computational saving: WF attains \( \varepsilon \)-accuracy within \( O(\log n \log \frac{1}{\varepsilon}) \) iterations with \( \eta \asymp 1/\log n \) if \( m \asymp n \log n \).
Key ingredient: leave-one-out analysis

How to establish $|a_l^\top (x^t - x^\dagger)| \lesssim \sqrt{\log n} \|x^\dagger\|_2$?
Key ingredient: leave-one-out analysis

How to establish $|a_l^\top (x^t - x^\dagger)| \lesssim \sqrt{\log n} \|x^\dagger\|_2$?

Technical difficulty: $x^t$ is statistically dependent with $\{a_l\}$;
Key ingredient: leave-one-out analysis

How to establish $|a_l^\top(x^t - x^\dagger)| \lesssim \sqrt{\log n} \|x^\dagger\|_2$?

Technical difficulty: $x^t$ is statistically dependent with $\{a_l\}$;

Leave-one-out trick: For each $1 \leq l \leq m$, introduce leave-one-out iterates $x^{t,(l)}$ by dropping $l$th sample.
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 $x^{t,(l)} \approx$ true iterates $x^t$
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 \( x^t, (l) \approx \) true iterates \( x^t \)
- Finish by triangle inequality

\[
|a_l^\top (x^t - x^b)| \leq |a_l^\top (x^t, (l) - x^b)| + |a_l^\top (x^t - x^t, (l))|
\]
Incoherence region in high dimensions

2-dimensional

high-dimensional

incoherence region is vanishingly small
No sample splitting

- Several prior works use sample-splitting: require fresh samples at each iteration; not practical but helps analysis.

- This work: reuses all samples in all iterations
This recipe is quite general
Low-rank matrix completion

Given partial samples of a *low-rank* matrix $M$ in an index set $\Omega$, fill in missing entries.

*Applications: recommendation systems, ...*
Incoherence

\[
\begin{bmatrix}
1 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 0
\end{bmatrix}
\quad \text{hard } \mu=n
\]

versus

\[
\begin{bmatrix}
1 & 1 & 1 & \cdots & 1 \\
1 & 1 & 1 & \cdots & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & 1 & \cdots & 1
\end{bmatrix}
\quad \text{easy } \mu=1
\]

Definition (Incoherence for matrix completion)

A rank-\(r\) matrix \(M^\|$ with eigendecomposition \(M^\|$ = \(U^\| \Sigma^\|$ \(U^\|$^T\) is said to be \(\mu\)-incoherent if

\[
\|U^\|^\|_{2,\infty} \leq \sqrt{\frac{\mu}{n}} \|U^\|_F = \sqrt{\frac{\mu r}{n}}.
\]
Matrix completion via vanilla GD

\[
\text{minimize}_{X \in \mathbb{R}^{n \times r}} \quad f(X) = \sum_{(j,k) \in \Omega} (e_j^\top XX^\top e_k - M_{j,k})^2
\]
Prior theory

\[
\minimize_{X \in \mathbb{R}^{n \times r}} f(X) = \sum_{(j,k) \in \Omega} \left( e_j^\top XX^\top e_k - M_{j,k} \right)^2
\]

Existing theory promotes incoherence explicitly:

- 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, Zheng, Lafferty '16
- no theory on vanilla / unregularized gradient descent
Prior theory

\[
\text{minimize } \min_{\mathbf{X} \in \mathbb{R}^{n \times r}} f(\mathbf{X}) = \sum_{(j, k) \in \Omega} \left( e_j^\top \mathbf{X} \mathbf{X}^\top e_k - M_{j,k} \right)^2
\]

Existing theory promotes incoherence explicitly:

- regularized loss (solve \( \min_{\mathbf{X}} f(\mathbf{X}) + R(\mathbf{X}) \) instead)
  - e.g. Keshavan, Montanari, Oh ’10, Sun, Luo ’14, Ge, Lee, Ma ’16
Prior theory

\[
\text{minimize}_{\mathbf{X} \in \mathbb{R}^{n \times r}} \quad f(\mathbf{X}) = \sum_{(j,k) \in \Omega} \left( e_j^\top \mathbf{X} \mathbf{X}^\top e_k - M_{j,k} \right)^2
\]

Existing theory promotes incoherence explicitly:

- regularized loss (solve \( \min_{\mathbf{X}} f(\mathbf{X}) + R(\mathbf{X}) \) 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, Zheng, Lafferty ’16

- no theory on vanilla / unregularized gradient descent
Our theory

**Theorem (Matrix completion)**

Suppose $M = X^b X^b\top$ is rank-$r$, incoherent and well-conditioned. Vanilla GD (with spectral initialization) achieves

- $\max_i \| e_i^\top (X^t - X^b) \|_2 \ll \| X^b \|_{2,\infty}$ (incoherence)
- in $O(\log \frac{1}{\varepsilon})$ iterations

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

**Theorem (Matrix completion)**

Suppose $M = X^b X^b^\top$ is rank-$r$, incoherent and well-conditioned. Vanilla GD (with spectral initialization) achieves

- $\max_i \| e_i^\top (X_t - X^b) \|_2 \ll \| X^b \|_{2,\infty}$ (incoherence)
- in $O\left( \log \frac{1}{\varepsilon} \right)$ iterations w.r.t. $\| \cdot \|_F$, $\| \cdot \|$, and $\| \cdot \|_{2,\infty}$ incoherence

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

- near-optimal entrywise error control $\| X_t X_t^{\top} - M^b^\|_\infty$.
- $O(\log 1/\varepsilon)$ iteration complexity.
- First result on vanilla gradient descent for matrix completion.
Noiseless matrix completion via Vanilla GD

Figure: Relative error of $X^t X^{t\top}$ (measured by $\|\cdot\|_F$, $\|\cdot\|$, $\|\cdot\|_\infty$) vs. iteration count for matrix completion, where $n = 1000$, $r = 10$, $p = 0.1$, and $\eta_t = 0.2$. 
Noisy matrix completion via Vanilla GD

Figure: Squared relative error of the estimate $\hat{X}$ (measured by $\|\cdot\|_F, \|\cdot\|, \|\cdot\|_{2,\infty}$) and $\hat{M} = \hat{X} \hat{X}^\top$ (measured by $\|\cdot\|_\infty$) vs. SNR, where $n = 500$, $r = 10$, $p = 0.1$, and $\eta_t = 0.2$. Here, $\text{SNR} := \frac{\|M^q\|_F^2}{n^2 \sigma^2}$. 
What about random initialization?
Initialization

- spectral initialization gets us reasonably close to truth
• spectral initialization gets us reasonably close to truth

• cannot initialize GD from anywhere, e.g. it might get stuck at local stationary points (e.g. saddle points)
• spectral initialization gets us reasonably close to truth

• cannot initialize GD from anywhere, e.g. it might get stuck at local stationary points (e.g. saddle points)

Can we initialize GD randomly?
What does prior theory say?

- no spurious local mins (Sun et al. ’16)
What does prior theory say?

- no spurious local mins (Sun et al. ’16)
- Vanilla GD with random initialization converges to global min almost surely (Lee et al. ’16)

No convergence rate guarantees for vanilla GD!
Randomly initialized GD for phase retrieval

\[ \eta_t = 0.1, \ a_i \sim \mathcal{N}(0, I_n), \ m = 10n, \ x^0 \sim \mathcal{N}(0, n^{-1} I_n) \]
Randomly initialized GD for phase retrieval

$$\eta_t = 0.1, \ a_i \sim \mathcal{N}(0, I_n), \ m = 10n, \ \mathbf{x}^0 \sim \mathcal{N}(0, n^{-1}I_n)$$

Randomly initialized GD enters local basin within a few iterations
Randomly initialized GD for phase retrieval

\[ \eta_t = 0.1, \ a_i \sim \mathcal{N}(0, I_n), \ m = 10n, \ x^0 \sim \mathcal{N}(0, n^{-1} I_n) \]

Randomly initialized GD enters local basin within a few iterations
These numerical findings can be formalized when $a_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_n)$:

**Theorem (Chen, Chi, Fan, Ma ’18)**

Under i.i.d. Gaussian design, GD with $x^0 \sim \mathcal{N}(0, n^{-1}I_n)$ achieves

$$\text{dist}(x^t, x^\dagger) \leq \gamma (1 - \rho)^{t-T_\gamma} \|x^\dagger\|_2, \quad t \geq T_\gamma$$

for $T_\gamma \preceq \log n$ and some constants $\gamma, \rho > 0$, provided that step size $\eta \asymp 1$ and sample size $m \gtrsim n \text{poly log } m$. 
Theoretical guarantees

$$\text{dist}(\mathbf{x}_t, \mathbf{x}_\text{fl}) \leq \gamma (1 - \rho)^{t - T\gamma} \|\mathbf{x}_\text{fl}\|_2, \quad t \geq T\gamma \asymp \log n$$
Theoretical guarantees

\[ \text{dist}(\bm{x}^t, \bm{x}^\dagger) \leq \gamma (1 - \rho)^{t - T_\gamma} \| \bm{x}^\dagger \|_2, \quad t \geq T_\gamma \lesssim \log n \]

- \textit{Stage 1:} takes \( O(\log n) \) iterations to reach \( \text{dist}(\bm{x}^t, \bm{x}^\dagger) \leq \gamma \)
Theoretical guarantees

\[ \text{dist}(\mathbf{x}^t, \mathbf{x}^\dagger) \leq \gamma (1 - \rho)^{t-T\gamma} \| \mathbf{x}^\dagger \|_2, \quad t \geq T\gamma \asymp \log n \]

- **Stage 1:** takes \( O(\log n) \) iterations to reach \( \text{dist}(\mathbf{x}^t, \mathbf{x}^\dagger) \leq \gamma \)
- **Stage 2:** linear convergence
Theoretical guarantees

\[
\text{dist}(\mathbf{x}^{t}, \mathbf{x}^{\dagger}) \leq \gamma(1 - \rho)^{t-T_{\gamma}}\|\mathbf{x}^{\dagger}\|_{2}, \quad t \geq T_{\gamma} \leq \log n
\]

Randomly initialized WF attains \(\varepsilon\)-accuracy within
\[O\left(\log n + \log \frac{1}{\varepsilon}\right)\] iterations with \(\eta \approx 1\) if \(m \approx n\text{polylog}m\)
Population-level (infinite samples) state evolution

\[ \mathbf{x}^{t+1} = \mathbf{x}^t - \eta \cdot \nabla F(\mathbf{x}^t) \]

population gradient

Let \( \alpha_t := \left| \langle \mathbf{x}^t, \mathbf{x}^\dagger \rangle \right| \),

signal strength

\( \beta_t := \| \mathbf{x}^t - \langle \mathbf{x}^t, \mathbf{x}^\dagger \rangle \mathbf{x}^\dagger \|_2 \)

size of residual component
Population-level (infinite samples) state evolution

\[ \mathbf{x}^{t+1} = \mathbf{x}^t - \eta \cdot \nabla F(\mathbf{x}^t) \]

Let \( \alpha_t := \left| \langle \mathbf{x}^t, \mathbf{x}^\dagger \rangle \right| \),

signal strength

\[ \beta_t := \left\| \mathbf{x}^t - \langle \mathbf{x}^t, \mathbf{x}^\dagger \rangle \mathbf{x}^\dagger \right\|_2 \]

size of residual component

2-parameter dynamics:

\[ \alpha_{t+1} = \left\{ 1 + 3\eta [1 - (\alpha_t^2 + \beta_t^2)] \right\} \alpha_t \]

\[ \beta_{t+1} = \left\{ 1 + \eta [1 - 3(\alpha_t^2 + \beta_t^2)] \right\} \beta_t \]
Back to finite-sample analysis

\[ x^{t+1} = x^t - \eta \nabla f(x^t) \]
Back to finite-sample analysis

\[ x^{t+1} = x^t - \eta \nabla f(x^t) = x^t - \eta \nabla F(x^t) - \eta \left( \nabla f(x^t) - \nabla F(x^t) \right) \]

\[ := r(x^t) \]
Back to finite-sample analysis

\[ x^{t+1} = x^t - \eta \nabla f(x^t) = x^t - \eta \nabla F(x^t) - \eta \left( \nabla f(x^t) - \nabla F(x^t) \right) := r(x^t) \]

- population-level analysis holds \textit{approximately} if \[ r(x^t) \ll x^t - \eta \nabla F(x^t) \]

a region with well-controlled \( r(x) \)
Back to finite-sample analysis

\[
x^{t+1} = x^t - \eta \nabla f(x^t) = x^t - \eta \nabla F(x^t) - \eta \left( \nabla f(x^t) - \nabla F(x^t) \right)
\]

\[
:= r(x^t)
\]

- population-level analysis holds \textit{approximately} if \[ r(x^t) \ll x^t - \eta \nabla F(x^t) \]
- \( r(x^t) \) is well-controlled if \( x^t \) is independent of \( \{a_k\} \)

a region with well-controlled \( r(x) \)
Back to finite-sample analysis

\[ x^{t+1} = x^t - \eta \nabla f(x^t) = x^t - \eta \nabla F(x^t) - \eta \left( \nabla f(x^t) - \nabla F(x^t) \right) \]

\[ := r(x^t) \]

- population-level analysis holds approximately if \( r(x^t) \ll x^t - \eta \nabla F(x^t) \)

- \( r(x^t) \) is well-controlled if \( x^t \) is independent of \( \{a_k\} \)

- **key analysis ingredient:** show \( x^t \) is “nearly-independent” of each \( a_k \) via leave-one-out analysis
Conclusions

**optimization theory + statistical model:** vanilla gradient descent is “implicitly regularized” and runs fast!

**Computational:**
- near dimension-free iteration complexity

**Statistical:**
- near-optimal sample complexity

It will be interesting to study “implicit regularization” via the leave-one-out argument for other algorithms such as alternating minimization, and other problems.


Thank you!