Implicit Regularization in Nonconvex Statistical Estimation

Yuxin Chen
Electrical Engineering, Princeton University
Cong Ma  
Princeton ORFE

Kaizheng Wang  
Princeton ORFE

Yuejie Chi  
CMU ECE
Nonconvex estimation problems are everywhere

Empirical risk minimization is usually nonconvex

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

Empirical risk minimization is usually nonconvex

\[
\text{minimize}_x \quad \ell(x; y) \quad \rightarrow \quad \text{may be nonconvex}
\]

subj. to \quad x \in S \quad \rightarrow \quad \text{may be nonconvex}

- 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*
... but is sometimes much nicer than we think

- statistical models
- benign landscape
- exploit geometry
- efficient algorithms
Carefully proceed via iterative optimization procedures without leaving this local basin

- Start from an appropriate initial point

**Optimization-based methods: two-stage approach**

- Initial guess \( x^0 \)
- Basin of attraction

\[
\minimize h \kappa, y \backslash x \quad \text{s.t.} \quad y + b \backslash x = A \backslash x
\]

Find an initial point within a nice local basin surrounding

\[
| \cdot |_2^2 \quad \text{squared magnitude}
\]

with prob.

\[
\begin{align*}
\minimize _{k=1}^{m} X \kappa, k = \cdots \quad \text{s.t.} \quad b \kappa = a
\end{align*}
\]
Optimization-based methods: two-stage approach

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

- Prevents overfitting and improves generalization
  - e.g. $\ell_1$ penalization, SCAD, nuclear norm penalization, ...
Roles of regularization

• Prevents overfitting and improves generalization
  ○ e.g. $\ell_1$ penalization, SCAD, nuclear norm penalization, ...

• Improves computation by stabilizing search directions
  ○ e.g. trimming, projection, regularized loss
Roles of regularization

• Prevents overfitting and improves generalization
  ○ e.g. $\ell_1$ penalization, SCAD, nuclear norm penalization, ...

• Improves computation by stabilizing search directions
  $\implies$ focus of this talk
  ○ e.g. trimming, projection, regularized loss
3 representative nonconvex problems

- phase retrieval
- matrix completion
- blind deconvolution
Are unregularized methods suboptimal for nonconvex estimation?
Are unregularized methods suboptimal for nonconvex estimation?
Are unregularized methods suboptimal for nonconvex estimation?
Missing phase problem

Detectors record **intensities** of diffracted rays

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

*Fig credit: Stanford SLAC*

intensity of electrical field: $|\hat{x}(f_1, f_2)|^2 = \left| \int x(t_1, t_2)e^{-i2\pi(f_1t_1+f_2t_2)} dt_1 dt_2 \right|^2$
Missing phase problem

Detectors record **intensities** of diffracted rays

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

*Fig credit: Stanford SLAC*

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 \( |\hat{x}(f_1, f_2)|^2 \)
Solving quadratic systems of equations

\[ A \times x = Ax \]

\[ y = |Ax|^2 \]

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

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

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

Empirical risk minimization

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

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

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

Two standard conditions that enable geometric convergence of GD

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

Two standard conditions that enable geometric convergence of GD

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

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

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

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

\( \ell_2 \) 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

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

region of local strong convexity + smoothness
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
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
Gradient descent theory revisited

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

region of local strong convexity + smoothness
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 \(\beta/\alpha\) determines rate of convergence
Gradient descent theory revisited

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

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

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

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

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

Population level (infinite samples)

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

_locally_ positive definite and well-conditioned

Consequence: WF converges within $O(\log \frac{1}{\varepsilon})$ iterations if $m \to \infty$
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 $(m \asymp n \log n)$

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

Gaussian designs: $a_k \sim_{\text{i.i.d.}} \mathcal{N}(0, I_n)$, $1 \leq k \leq m$

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

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

condition number $\asymp n$
What does this optimization theory say about WF?

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

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

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

condition number $\approx n$

Consequence (Candès et al ’14): WF attains $\varepsilon$-accuracy within $O(n \log \frac{1}{\varepsilon})$ iterations if $m \approx n \log n$
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 \((m \asymp n \log n)\)

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

Consequence (Candès et al '14): WF attains \(\varepsilon\)-accuracy within \(O(n \log \frac{1}{\varepsilon})\) 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
But wait a minute ...:

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

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

\[ \eta_t = O\left(\frac{1}{n}\right) \]

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

WF converges in $O(n)$ iterations

\[ \uparrow \]

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

\[ \uparrow \]

This choice is suggested by worst-case optimization theory

\[ \uparrow \]

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?

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

Which region enjoys both strong convexity and smoothness?

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

- 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 $x^\triangledown$
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

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

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

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

Theorem 1 (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)
- \( \| x^t - x^\natural \|_2 \lesssim (1 - \frac{\eta}{2})^t \| x^\natural \|_2 \) (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

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

- \( \| x^t - x^\dagger \|_2 \lesssim (1 - \frac{\eta}{2})^t \| x^\dagger \|_2 \) (near-linear convergence)

provided that step size \( \eta \prec \frac{1}{\log n} \) 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

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

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

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

- Computational complexity: \( \frac{n}{\log n} \) times faster than existing theory
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.

\[
\begin{array}{c|c|c|c}
A^{(l)} & x & A^{(l)}x & y^{(l)} = |A^{(l)}x|^2 \\
\hline
\end{array}
\]
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 \textbf{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 \( \{\mathbf{x}^{t,(l)}\} \) are independent of \( \mathbf{a}_l \), and are hence \textbf{incoherent} w.r.t. \( \mathbf{a}_l \) with high prob.
- Leave-one-out iterates \( \mathbf{x}^{t,(l)} \approx \text{true iterates } \mathbf{x}^t \)
- \( |\mathbf{a}_l^\top (\mathbf{x}^t - \mathbf{x}^\dagger)| \leq |\mathbf{a}_l^\top (\mathbf{x}^{t,(l)} - \mathbf{x}^\dagger)| + |\mathbf{a}_l^\top (\mathbf{x}^t - \mathbf{x}^{t,(l)})| \)
This recipe is quite general
Low-rank matrix completion

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

*Fig. credit: Candès*
Prior art

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

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

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

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

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

- projection onto set of incoherent matrices
  - e.g. Chen, Wainwright '15, Zheng, Lafferty '16
Theoretical guarantees

Theorem 2 (Matrix completion)

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

- in $O(\log \frac{1}{\epsilon})$ iterations

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

- in $O\left(\log \frac{1}{\varepsilon}\right)$ iterations w.r.t. $\|\cdot\|_F$, $\|\cdot\|$, and $\|\cdot\|_{2,\infty}$

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

Theorem 2 (Matrix completion)

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

- in $O\left(\log \frac{1}{\varepsilon}\right)$ iterations w.r.t. $\|\cdot\|_F$, $\|\cdot\|$, and $\|\cdot\|_{2,\infty}$ if step size $\eta \lesssim \frac{1}{\sigma_{\text{max}}(M)}$ and sample size $\gtrsim nr^3 \log^3 n$

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

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

\[
\begin{align*}
\text{minimize}_{x, h} & \quad f(x, h) = \sum_{k=1}^{m} \left| b_k^* \left( hx^* - h^\dagger x^\dagger \right) a_k \right|^2 \\
\begin{array}{c}
a_k \sim \mathcal{N}(0, I) \\
\end{array} & \quad \text{and} \quad \{b_k\} : \text{partial Fourier basis}
\end{align*}
\]
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 \\

\text{\(a_k\) i.i.d.} \quad \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
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

\begin{align*}
  a_k & \sim \text{i.i.d. } \mathcal{N}(0, I) \\
  \{b_k\} : & \text{ partial Fourier basis}
\end{align*}

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 $h^\dagger$ is incoherent w.r.t. $\{b_k\}$. Vanilla gradient descent (with spectral initialization) achieves $\varepsilon$ accuracy in $O(\log \frac{1}{\varepsilon})$ iterations, provided that step size $\eta \lesssim 1$ and sample size $m \gtrsim n \text{poly log}(m)$.

- Regularization-free
- Converges in $O(\log \frac{1}{\varepsilon})$ iterations (vs. $O(m \log \frac{1}{\varepsilon})$ iterations in prior theory)
Incoherence region in high dimensions

2-dimensional

high-dimensional (mental representation)

incoherence region is vanishingly small
Complicated dependencies across iterations

- Several prior sample-splitting approaches: require fresh samples at each iteration; not what we actually run in practice
Complicated dependencies across iterations

- Several prior sample-splitting approaches: require fresh samples at each iteration; not what we actually run in practice

- This work: reuses all samples in all iterations
Summary

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

• **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, arXiv:1711.10467