Sharp convergence guarantees for iterative algorithms in random (nonconvex) optimization

or

Can you predict *if* and *how fast*
your model-fitting algorithm will converge?

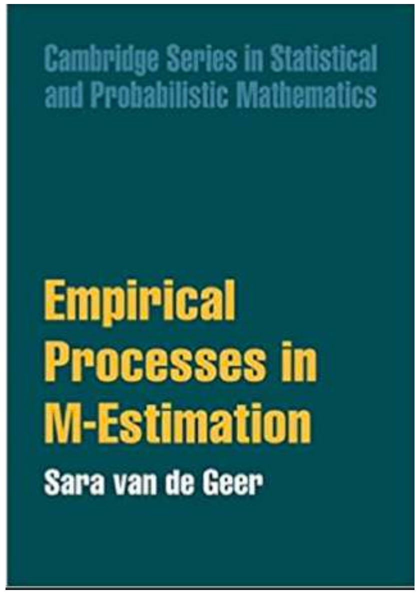
Ashwin Pananjady

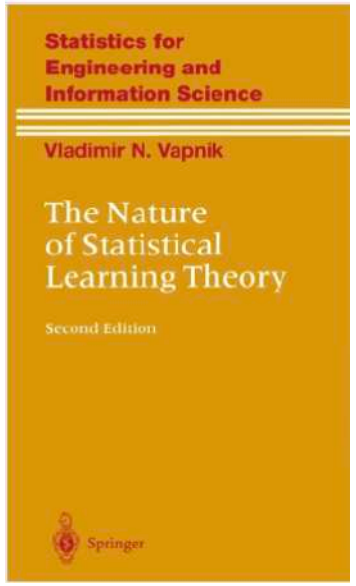
Schools of Industrial and Systems Engineering and
Electrical and Computer Engineering, Georgia Tech

Kabir
Chandrasekher
(Stanford)

Christos
Thrampoulidis
(UBC)

Mengqi
Lou
(Georgia Tech)



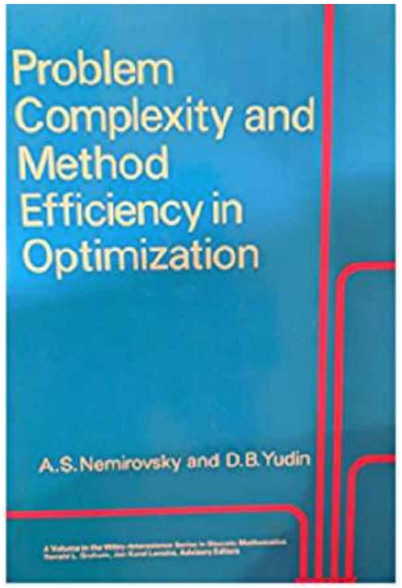


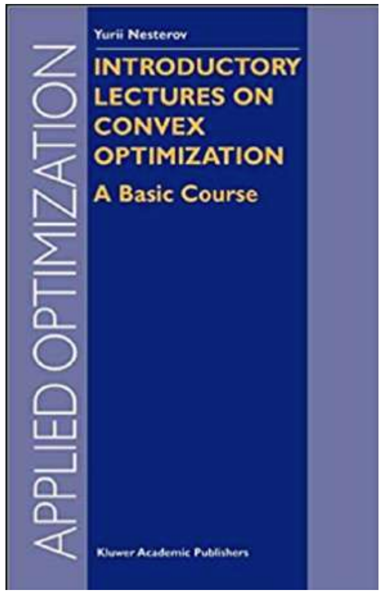
M-estimation in supervised learning

$$\mathbb{R}^d \quad \mathbb{R}$$

$$(\boldsymbol{x}_i, y_i)_{i=1}^n$$

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^n \ell_i(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i)$$





- Optimal solution $\widehat{\boldsymbol{\theta}}$ has nice properties under noise model, e.g.

$$\mathbb{E}[y_i|\boldsymbol{x}_i] = f_{\boldsymbol{\theta}^*}(\boldsymbol{x}_i)$$

- General purpose: Rate of estimation measured in terms of sample size and model geometry around $\boldsymbol{\theta}^*$.
- Optimality guarantees: Minimax lower bounds and information theory

- Iterative algorithms to converge to $\widehat{\boldsymbol{\theta}}$
- *Efficiency* of method: #iterations to get ε-close (typically upper bounds under convexity/smoothness notions)
- Optimality guarantees: Oracle complexity lower bounds over worst-case family of loss functions

Largely successful in convex optimization approaches to M-estimation:
Decoupling statistics from optimization

$$(\boldsymbol{x}_i, y_i)_{i=1}^n \qquad \widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^n \ell_i(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i)$$

Q1 Do worst case efficiency estimates reflect performance in model-fitting problems with random data?

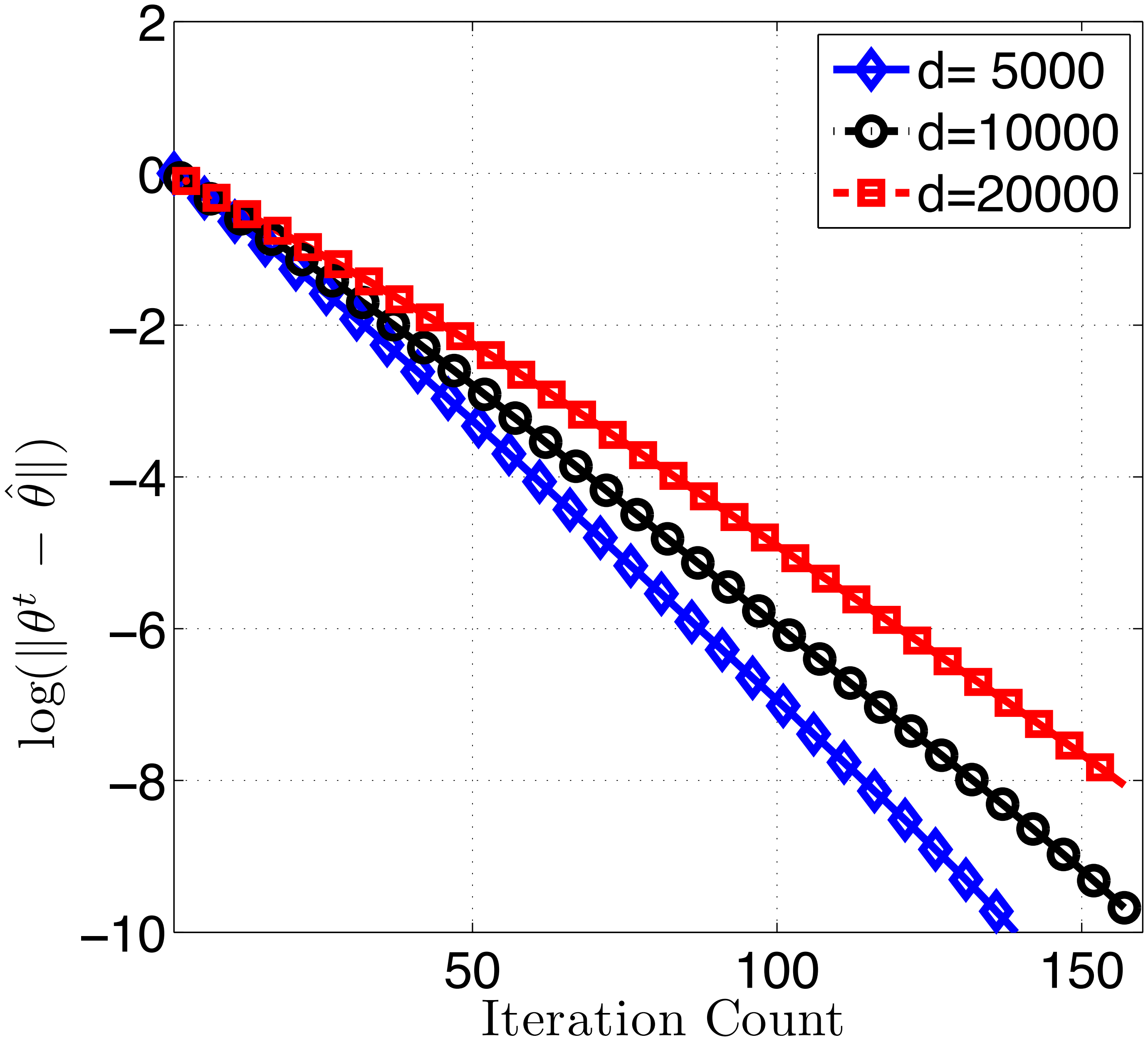
Q2 Can we provide exact efficiency estimates in random ensembles of optimization problems?

Q3 Is there hope of developing a parallel understanding for nonconvex problems?

Hoare '62 Smale '83 Spielman & Tang '04

Illustration 1: Projected gradient descent on L1 constrained quadratic programming with standard Gaussian covariates

$n = 2500$ Sparsity $k = \sqrt{d}$



$$y_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta}^* \rangle + \epsilon_i$$

$$\|\boldsymbol{\theta}^*\|_0 = k, \quad \|\boldsymbol{\theta}^*\|_2 = 1$$

- Worst-case efficiency guarantees pessimistic
- Different convergence rates depending on problem size: “Larger problems are harder” (but theory does not capture this phenomenon)

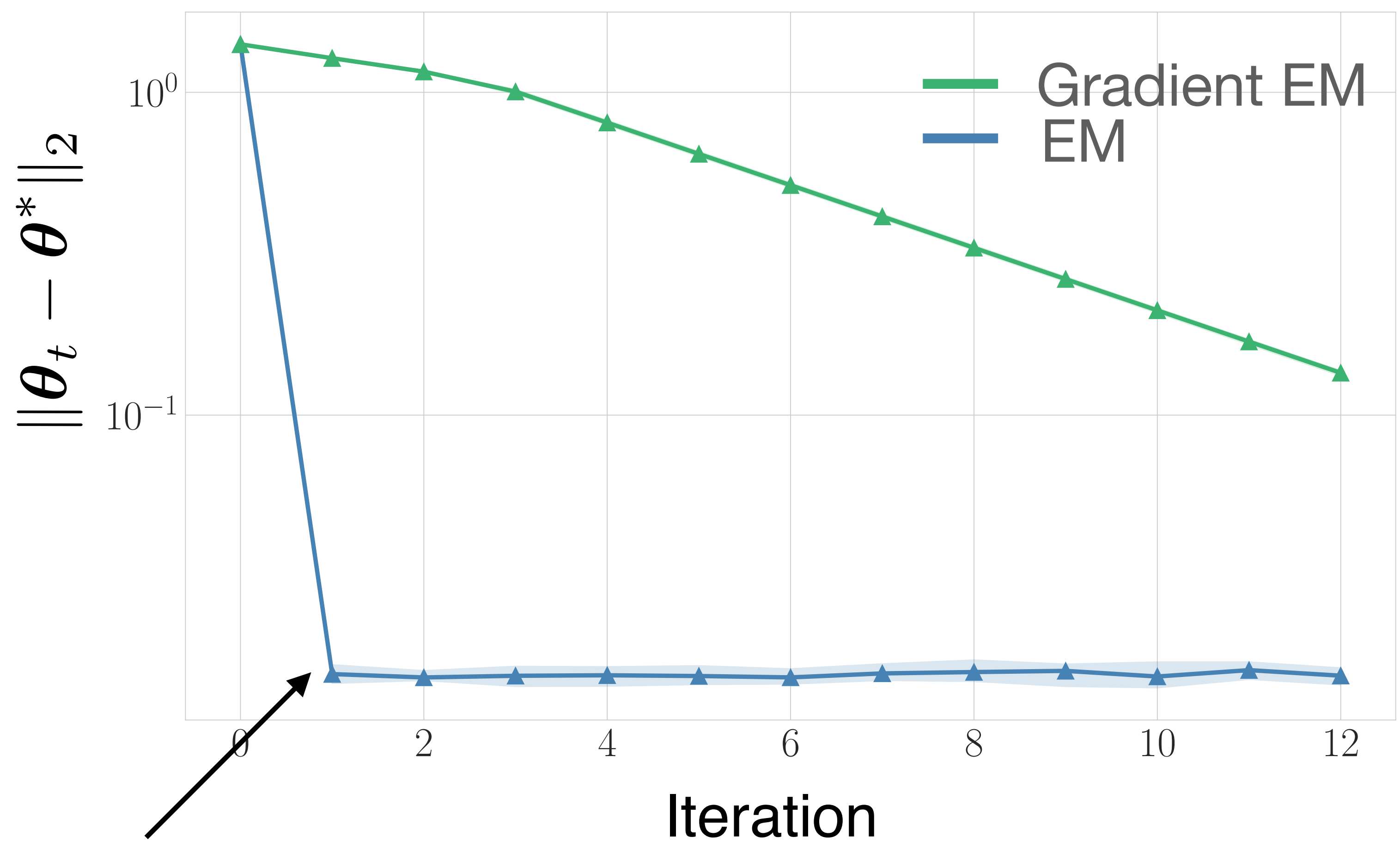
Agarwal, Negahban, Wainwright ’12

Illustration 2: Expectation maximization algorithms on symmetric Gaussian mixture

$$\boldsymbol{x}_i \sim \frac{1}{2}\mathsf{N}(\boldsymbol{\theta}^*, \boldsymbol{I}) + \frac{1}{2}\mathsf{N}(-\boldsymbol{\theta}^*, \boldsymbol{I})$$

$$\|\boldsymbol{\theta}^*\|_2 = 1$$

$d = 250, \quad n = 10{,}000$



Noise floor $\asymp \sigma\sqrt{d/n}$

- Non-convex optimization problem but parameter estimation possible
- Gradient EM and EM exhibit different convergence behavior but common analysis tool “cannot capture distinction”

Daskalakis, Tzamos, Zampetakis ’17 Balakrishnan, Wainwright, Yu ’17

- Rigorous comparisons between iterative model-fitting methods
- Not just by comparing upper bounds on efficiency!
- Want to answer the following questions in nonconvex problems:
  - Does the algorithm converge (to a statistically useful neighborhood) from a given initialization?
  - Does the algorithm converge *globally,* from a random initialization?
  - How fast does the algorithm converge?

Desiderata

- General-purpose, iterate-by-iterate predictions of solution quality if:
  - Each iteration of algorithm is solution to convex program*
  - Data in optimization problem is suitably Gaussian (i.e. Gaussian conditioned on past iterations)
- Distinguishes convergent behavior from otherwise
- Upper *and* lower bounds on convergence rates and exact error floor

This talk

AMP and first-order algorithms

Message-passing algorithms for compressed sensing

David L. Donoho[a,1], Arian Maleki[b], and Andrea Montanari[a,b,1]

The dynamics of message passing on dense graphs, with applications to compressed sensing

Mohsen Bayati
Department of Electrical Engineering
Stanford University

Andrea Montanari
Departments of Electrical Engineering and Statistics
Stanford University

An Iterative Construction of Solutions of the TAP Equations for the Sherrington–Kirkpatrick Model

Erwin Bolthausen*

Institute of Mathematics, Universität Zürich, Zürich, Switzerland. E-mail: eb@math.uzh.ch

The estimation error of general first order methods

Michael Celentano* Andrea Montanari*† Yuchen Wu*

Other sharp predictions

Sharp Time–Data Tradeoffs for Linear Inverse Problems

Samet Oymak*† Benjamin Recht*‡ Mahdi Soltanolkotabi§

Halting Time is Predictable for Large Models: A Universality Property and Average-case Analysis

Courtney Paquette*† Bart van Merriënboer* Elliot Paquette† Fabian Pedregosa*

SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality

Courtney Paquette*† Kiwon Lee† Fabian Pedregosa* Elliot Paquette†

- Applies under weaker randomness conditions
- Specific settings, typically asymptotic

General-purpose analysis tool: Population-based

STATISTICAL GUARANTEES FOR THE EM ALGORITHM: FROM POPULATION TO SAMPLE-BASED ANALYSIS[1]

BY SIVARAMAN BALAKRISHNAN*,†, MARTIN J. WAINWRIGHT† AND BIN YU†

University of California, Berkeley and Carnegie Mellon University†*

Global Convergence of the EM Algorithm for Mixtures of Two Component Linear Regression

Jeongyeol Kwon* KWONCHUNGLI@UTEXAS.EDU
The University of Texas at Austin
Wei Qian* WQ34@CORNELL.EDU
Cornell University
Constantine Caramanis CONSTANTINE@UTEXAS.EDU
The University of Texas at Austin
Yudong Chen YUDONG.CHEN@CORNELL.EDU
Cornell University
Damek Davis DSD95@CORNELL.EDU
Cornell University

- Complementary views: Landscape analysis, properties of loss function verified w.h.p.

Running example: Phase retrieval with a real signal

Model

$$y_i = |\langle \boldsymbol{x}_i, \boldsymbol{\theta}^* \rangle| + \epsilon_i, \; i = 1, 2, \ldots$$

$\boldsymbol{x}_i \sim \mathsf{N}(0, \boldsymbol{I}_d)$ $\quad \epsilon_i = \mathsf{N}(0, \sigma^2)$ $\quad \|\boldsymbol{\theta}^*\|_2 = 1$

MLE: Minimizer of nonconvex loss

$$\mathfrak{R}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - |\langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle|)^2$$

Algorithms

"Subgradient" method

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \eta \nabla \mathfrak{R}_n(\boldsymbol{\theta}_t) \\ &= \boldsymbol{\theta}_t - \frac{2\eta}{n} \sum_{i=1}^{n} (|\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle| - y_i) \cdot \mathsf{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle) \cdot \boldsymbol{x}_i \end{aligned}$$

Alternating projections

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \mathsf{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle) \cdot \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle)^2 \\ &= \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathsf{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle) \cdot \boldsymbol{x}_i y_i \right) \end{aligned}$$

Sample-splitting: Fresh data $(\boldsymbol{X}, \boldsymbol{y}) \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$ in each iteration

$$\Lambda = n/d > 1$$

Gershberg/Saxton/Fienup '70s-'80s $\quad$ Candes et al. '15 $\quad$ Zhang et al. '17 $\quad$ Netrapalli, Kwon et al. '12-'17

Running example: Phase retrieval with a real signal

Model

$$y_i = |\langle \boldsymbol{x}_i, \boldsymbol{\theta}^* \rangle| + \epsilon_i, \; i = 1, 2, \ldots$$

$\boldsymbol{x}_i \sim \mathsf{N}(0, \boldsymbol{I}_d)$ $\quad \epsilon_i = \mathsf{N}(0, \sigma^2)$ $\quad \|\boldsymbol{\theta}^*\|_2 = 1$

MLE: Minimizer of nonconvex loss

$$\mathfrak{R}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - |\langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle|)^2$$

Algorithms

"Subgradient" method

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla \mathfrak{R}_n(\boldsymbol{\theta}_t)$$
$$= \boldsymbol{\theta}_t - \frac{2\eta}{n} \sum_{i=1}^{n} (|\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle| - y_i) \cdot \mathsf{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle) \cdot \boldsymbol{x}_i$$

Alternating projections

$$\boldsymbol{\theta}_{t+1} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \mathsf{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle) \cdot \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle)^2$$
$$= \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathsf{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle) \cdot \boldsymbol{x}_i y_i \right)$$

Sample-splitting: Fresh data $(\boldsymbol{X}, \boldsymbol{y}) \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$ in each iteration

$$\Lambda = n/d > 1$$

Chen and Candes '17 Mondelli and Montanari '18 Mondelli and Venkataramanan '20 Maillard et al. '20

Running example: Phase retrieval with a real signal

Model

$$y_i = |\langle \boldsymbol{x}_i, \boldsymbol{\theta}^* \rangle| + \epsilon_i,\ i = 1, 2, \ldots$$

$\boldsymbol{x}_i \sim \mathsf{N}(0, \boldsymbol{I}_d)$ $\quad \epsilon_i = \mathsf{N}(0, \sigma^2)$ $\quad \|\boldsymbol{\theta}^*\|_2 = 1$

MLE: Minimizer of nonconvex loss

$$\mathfrak{R}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - |\langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle|)^2$$

Algorithms

"Subgradient" method

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \eta \nabla \mathfrak{R}_n(\boldsymbol{\theta}_t) \\ &= \boldsymbol{\theta}_t - \frac{2\eta}{n} \sum_{i=1}^{n} (|\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle| - y_i) \cdot \mathsf{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle) \cdot \boldsymbol{x}_i \end{aligned}$$

Alternating projections

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \mathsf{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle) \cdot \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle)^2 \\ &= \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathsf{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle) \cdot \boldsymbol{x}_i y_i \right) \end{aligned}$$

Sample-splitting: Fresh data $(\boldsymbol{X}, \boldsymbol{y}) \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$ in each iteration

$$\Lambda = n/d > 1$$

Waldspurger '17 Zhang '18-'20 Duchi, Ruan '18 Ghosh et al. '19-'21 Lee et al. '21

Background on population-based analysis

- View each iteration of the algorithm as a random operator

$$\mathcal{T}_n : \boldsymbol{\theta}_t \mapsto \boldsymbol{\theta}_{t+1}$$

- Use infinite sample limit $\mathcal{T}_\infty$ to guide analysis of the empirical iterates

Alternating projections

$$\mathcal{T}_n(\boldsymbol{\theta}_t)$$

$$= \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathsf{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle) \cdot \boldsymbol{x}_i y_i \right)$$

$$\mathcal{T}_\infty(\boldsymbol{\theta}_t) = \mathbb{E}[\mathsf{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle) \cdot \boldsymbol{x}_i y_i]$$

Subgradient method

$$\mathcal{T}_n(\boldsymbol{\theta}_t)$$

$$= \boldsymbol{\theta}_t - \frac{2\eta}{n} \sum_{i=1}^{n} (|\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle| - y_i) \cdot \mathsf{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle) \cdot \boldsymbol{x}_i$$

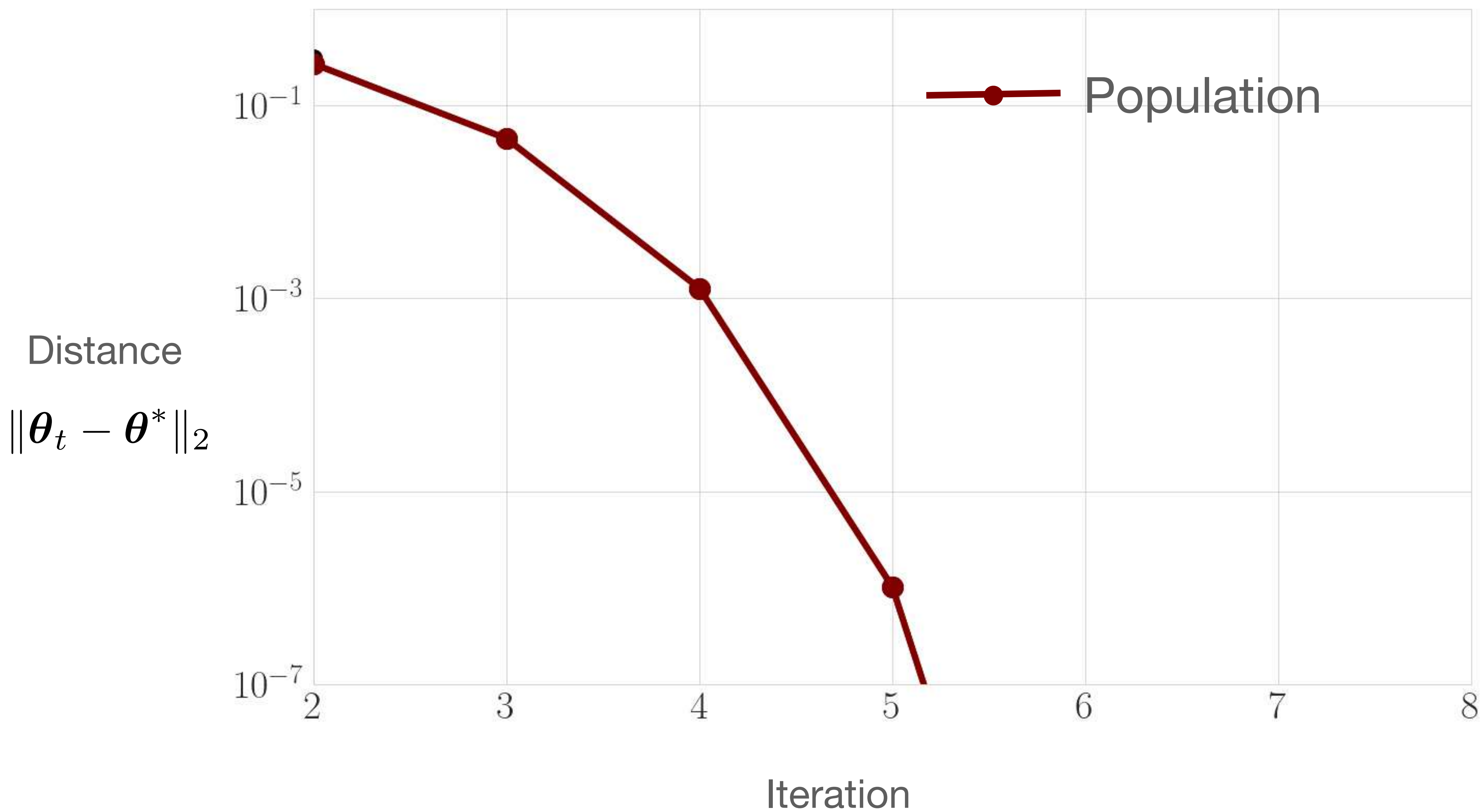
$$\mathcal{T}_\infty(\boldsymbol{\theta}_t) = (1 - 2\eta) \cdot \boldsymbol{\theta}_t + 2\eta \cdot \mathbb{E}[\mathsf{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle) \cdot \boldsymbol{x}_i y_i]$$

Note	For $\eta = 1/2$, both population updates coincide!

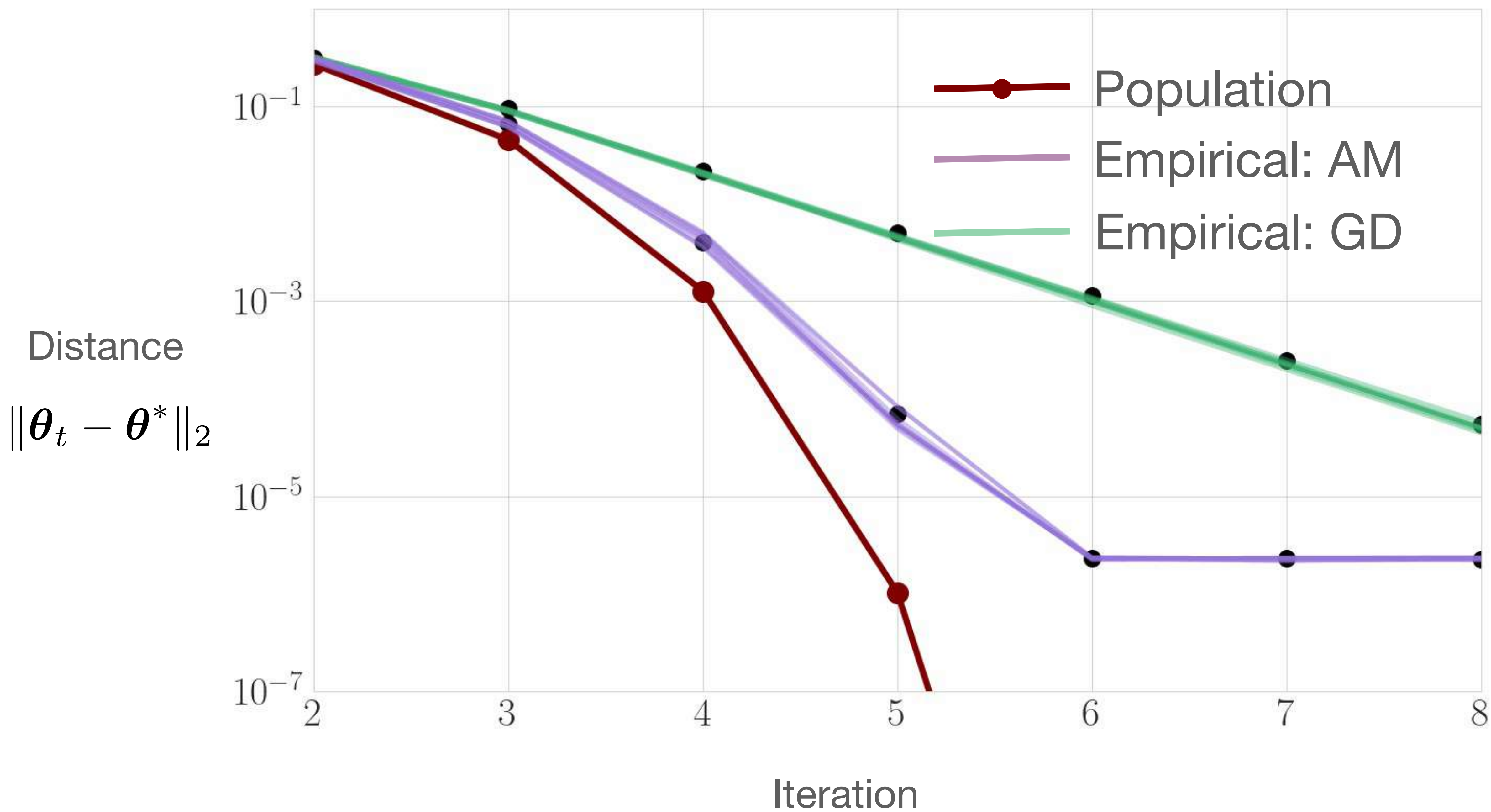
Balakrishnan, Wainwright, Yu '17 Population-based analysis of SAA Similar ideas in ODE literature

$d = 800, \quad n = 16{,}000, \quad \sigma = 10^{-6}$

$\eta = 1/2$



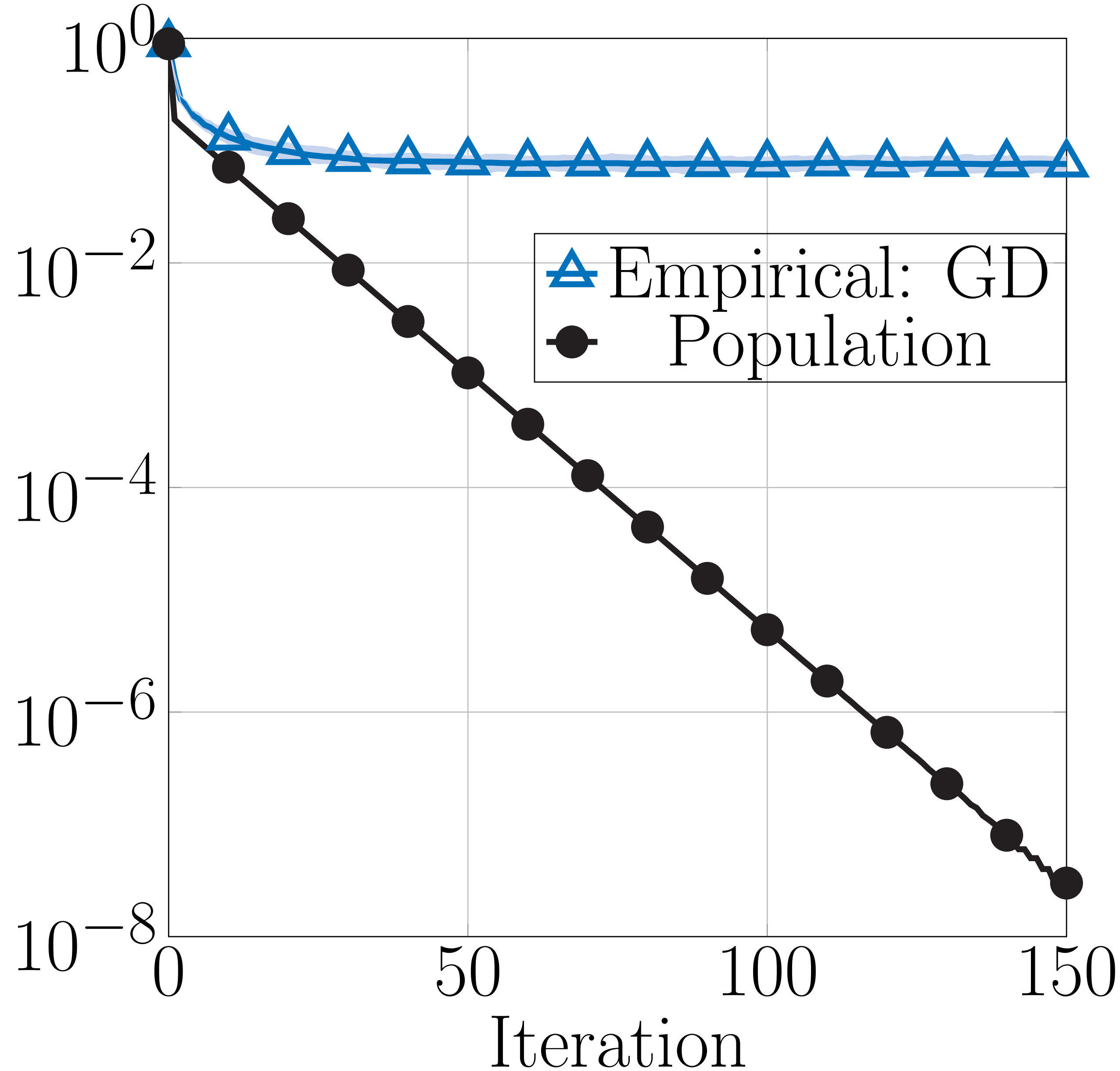
$d = 800, \quad n = 16{,}000, \quad \sigma = 10^{-6}$ $\eta = 1/2$



- Same population update, but different convergence behavior
- Population update significantly optimistic in both cases

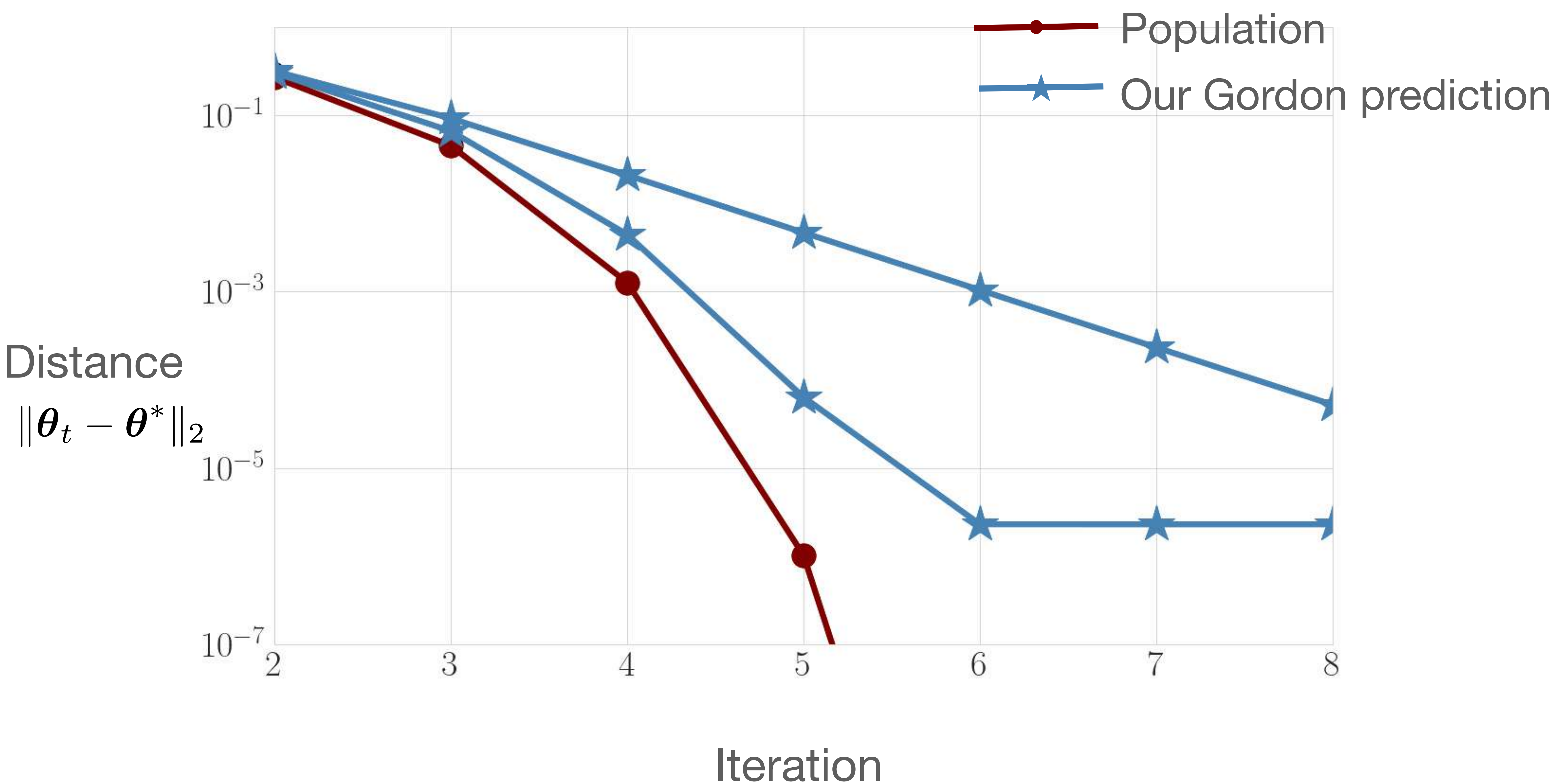
Distance

$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|_2$

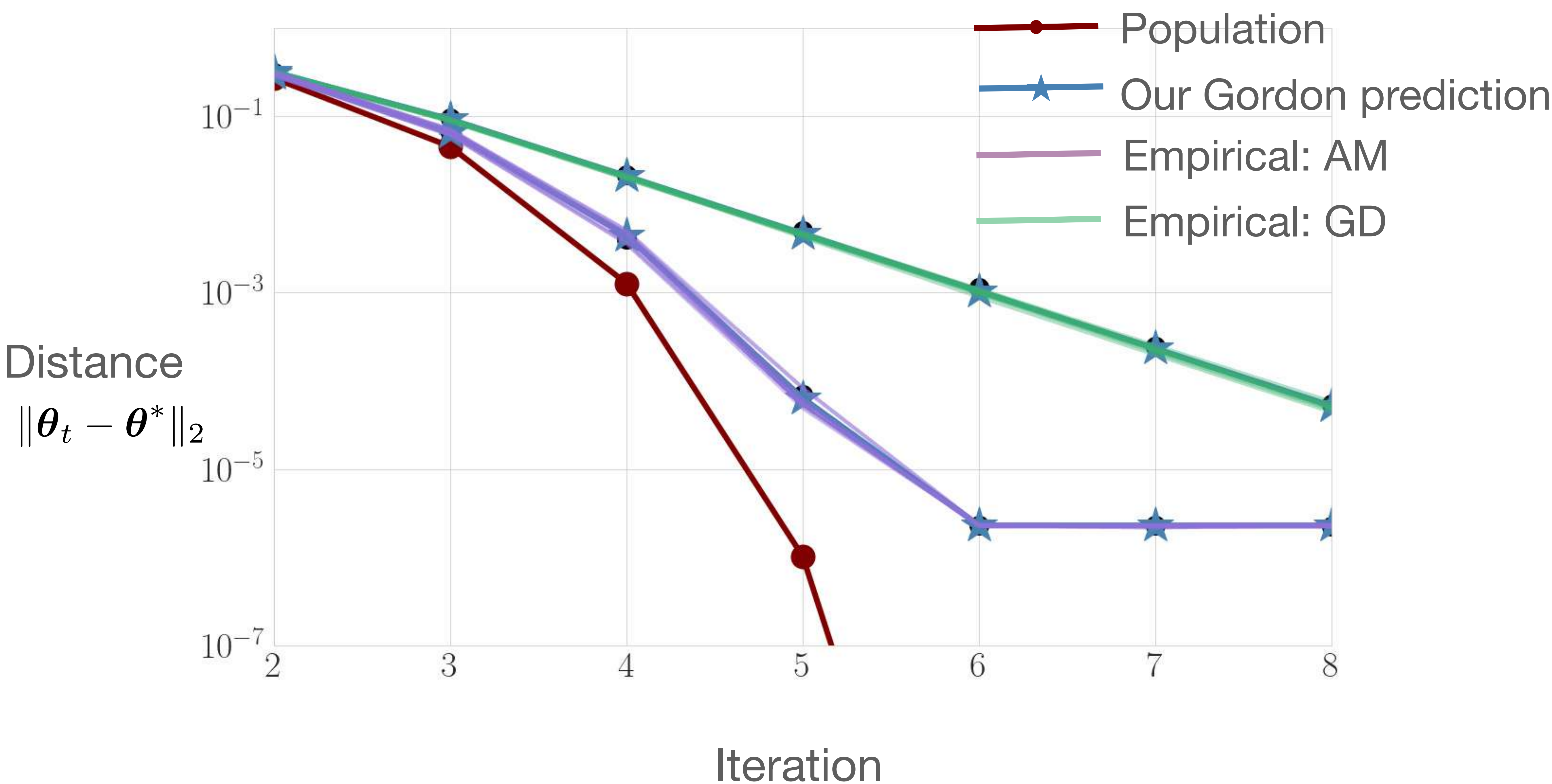


$d = 600, \quad n = 6{,}000, \quad \sigma = 0 \qquad \eta = 0.95$

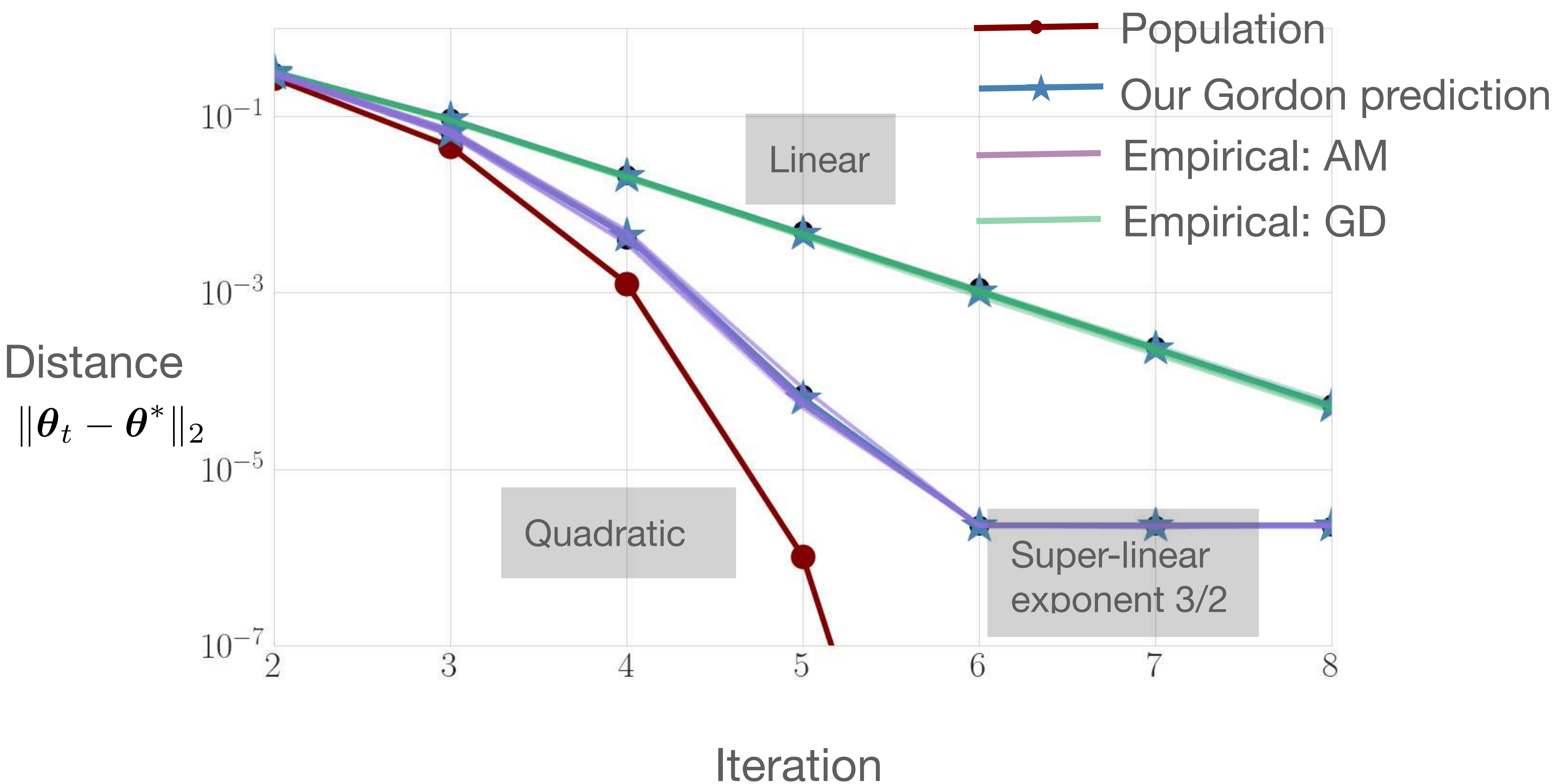
$d = 800, \quad n = 16{,}000, \quad \sigma = 10^{-6}$ $\eta = 1/2$

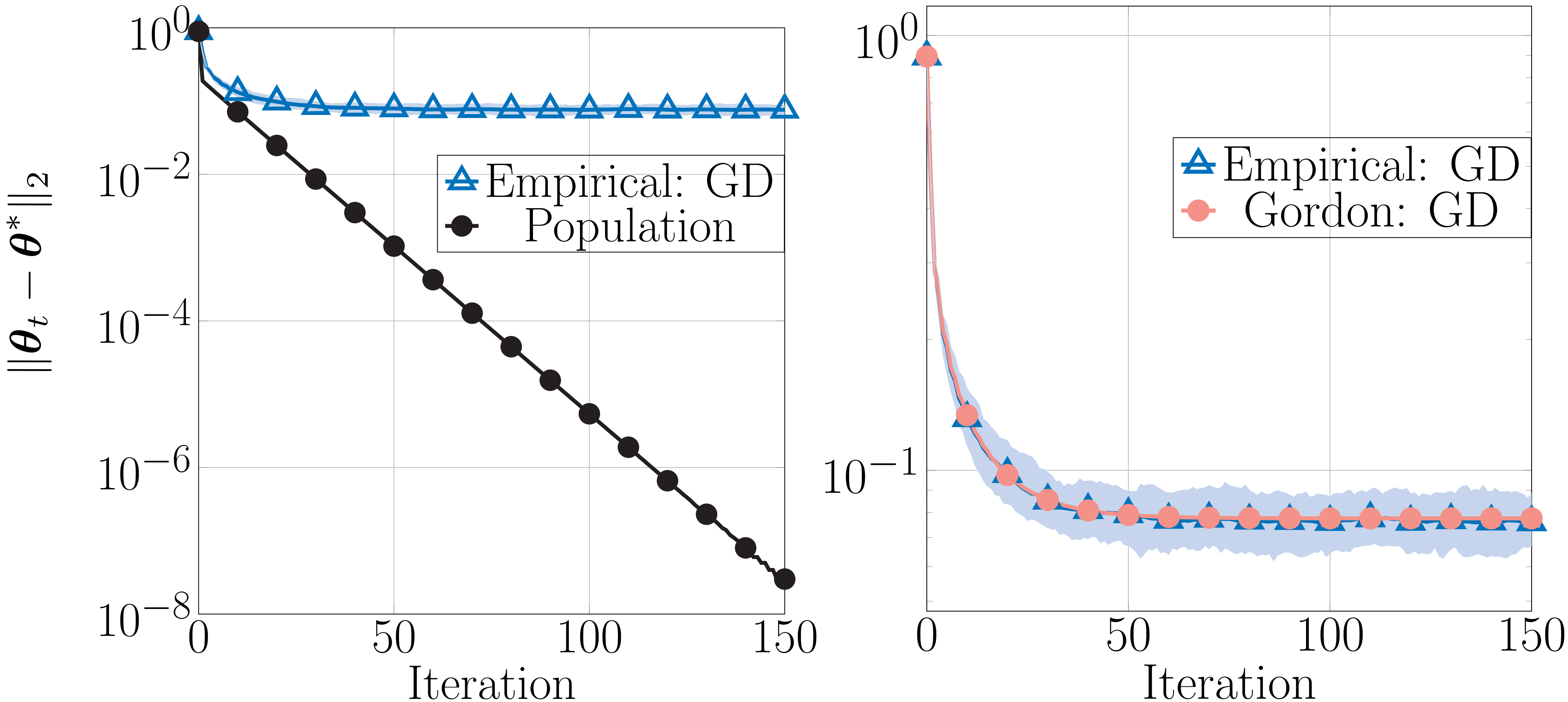


$d = 800, \quad n = 16{,}000, \quad \sigma = 10^{-6}$ $\eta = 1/2$



$d = 800, \quad n = 16{,}000, \quad \sigma = 10^{-6}$ $\qquad \eta = 1/2$





$d = 600, \quad n = 6{,}000, \quad \sigma = 0 \qquad \eta = 0.95$

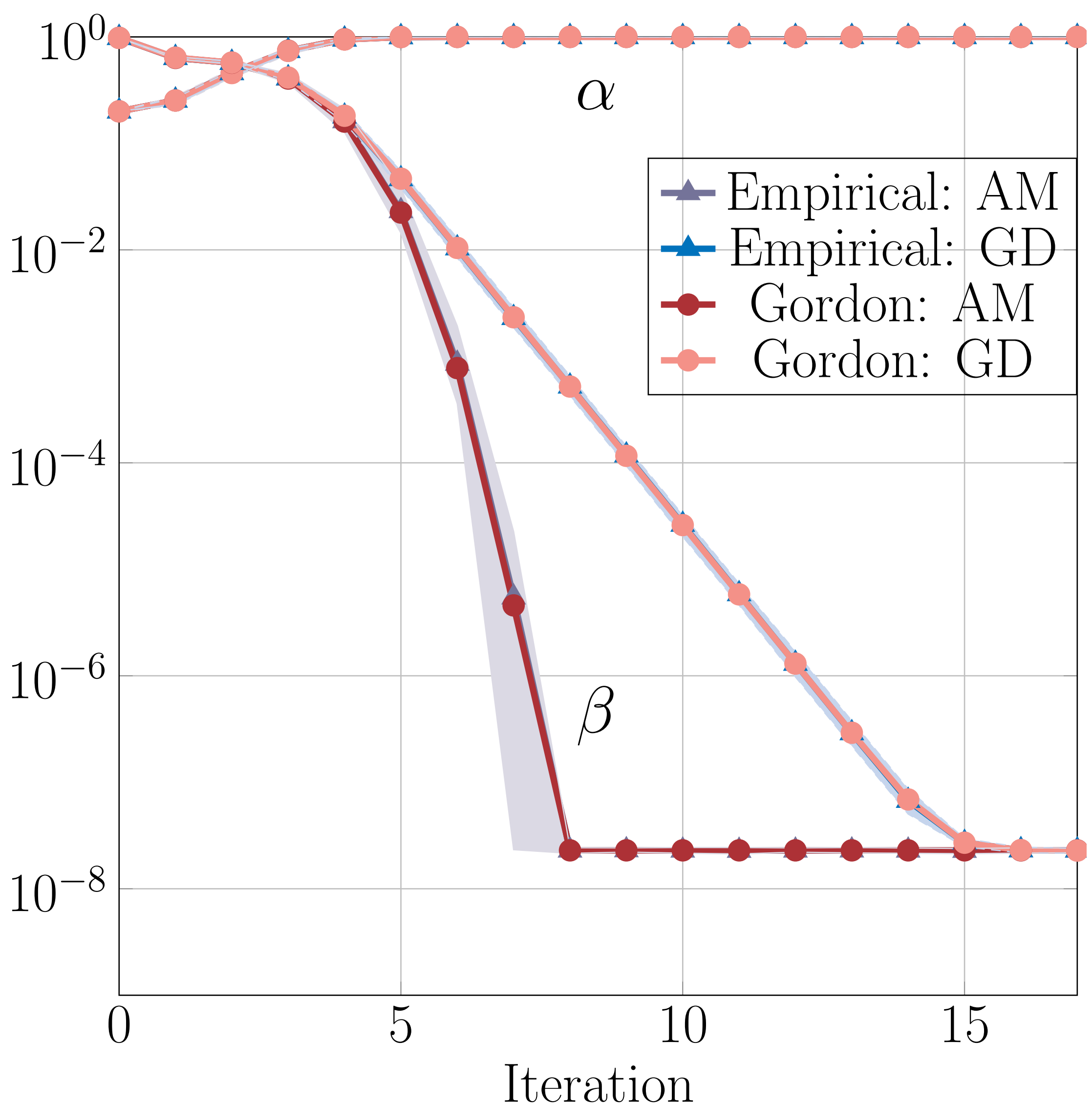
n : per-iteration sample size

d : dimension

$\Lambda = n/d$

$$\alpha(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \boldsymbol{\theta}^* \rangle$$

$$\beta(\boldsymbol{\theta}) = \| \boldsymbol{P}^{\perp}_{\boldsymbol{\theta}^*} \boldsymbol{\theta} \|_2$$



- Gordon prediction is state evolution update

$$(\alpha(\boldsymbol{\theta}_t), \beta(\boldsymbol{\theta}_t)) \mapsto (\alpha^{\text{gor}}_{t+1}, \beta^{\text{gor}}_{t+1})$$
$$\approx (\alpha(\boldsymbol{\theta}_{t+1}), \beta(\boldsymbol{\theta}_{t+1}))$$

- Finite sample "correction" to population prediction:

$$\left(\beta^{\text{gor}}_{t+1}\right)^2 = \left(\beta^{\text{pop}}_{t+1}\right)^2 + \mathcal{O}(\Lambda^{-1}) \cdot \Delta(\alpha_t, \beta_t; \sigma)$$

Chen et al. '19 Tan & Vershynin '19 Wu & Zhou '20

Part I:

General iterate-by-iterate recipe
if each iteration convex, Gaussianity

Part II:

Explicit one-step prediction for general class
of models and methods

Part III:

Consequences for nonconvex model-fitting:
Global convergence prediction

Workhorse: The Convex Gaussian Minimax theorem

$$\boldsymbol{G} \in \mathbb{R}^{n\times d} \quad \boldsymbol{G}_{i,j} \overset{i.i.d.}{\sim} \mathsf{N}(0,1)$$

Primary

$$\mathcal{P}(\boldsymbol{G}) := \min_{\boldsymbol{u}\in\mathcal{U}} \max_{\boldsymbol{v}\in\mathcal{V}} \; \langle \boldsymbol{v}, \boldsymbol{G}\boldsymbol{u} \rangle + Q(\boldsymbol{u},\boldsymbol{v})$$

Auxiliary

$$\mathcal{A}(\boldsymbol{\gamma}_d, \boldsymbol{\gamma}_n) := \min_{\boldsymbol{u}\in\mathcal{U}} \max_{\boldsymbol{v}\in\mathcal{V}} \; \|\boldsymbol{v}\|_2 \cdot \langle \boldsymbol{\gamma}_d, \boldsymbol{u} \rangle + \|\boldsymbol{u}\|_2 \cdot \langle \boldsymbol{\gamma}_n, \boldsymbol{v} \rangle + Q(\boldsymbol{u},\boldsymbol{v})$$

$$\boldsymbol{\gamma}_d \sim \mathsf{N}(0, \boldsymbol{I}_d) \qquad \boldsymbol{\gamma}_n \sim \mathsf{N}(0, \boldsymbol{I}_n)$$

Theorem

Suppose Q is continuous. Then the following hold for all scalars t :

(a) We have

$$\mathbb{P}(\mathcal{P}(\boldsymbol{G}) \le t) \le 2\mathbb{P}(\mathcal{A}(\boldsymbol{\gamma}_d, \boldsymbol{\gamma}_n) \le t)$$

Gordon '85, '88 Stojnic '13 Thrampoulidis, Oymak, Hassibi '15

$$\boldsymbol{\zeta} = (\alpha, \beta)$$

I: The recipe

Step 1: Write an iteration as solution to convex program; then write objective in bilinear form

$$\boldsymbol{\theta}_{t+1} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_t, \boldsymbol{X}, \boldsymbol{y})$$

Variational form/Fenchel conjugate

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_t, \boldsymbol{X}, \boldsymbol{y}) \overset{(d)}{=} \min_{\boldsymbol{u}\in\mathcal{U}} \max_{\boldsymbol{v}\in\mathcal{V}} \ \langle \boldsymbol{v}, \boldsymbol{G}\boldsymbol{u}\rangle + Q(\boldsymbol{u}, \boldsymbol{v})$$

Step 2: Invoke CGMT to replace matrix of Gaussian variables with two vectors

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_t, \boldsymbol{X}, \boldsymbol{y})$$

$$\approx \min_{\boldsymbol{u}\in\mathcal{U}} \max_{\boldsymbol{v}\in\mathcal{V}} \ \|\boldsymbol{v}\|_2 \cdot \langle \boldsymbol{\gamma}_d, \boldsymbol{u}\rangle + \|\boldsymbol{u}\|_2 \cdot \langle \boldsymbol{\gamma}_n, \boldsymbol{v}\rangle \ + Q(\boldsymbol{u}, \boldsymbol{v})$$

$$=: \min_{\boldsymbol{\theta}\in\mathbb{R}^d} \ \mathfrak{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_t, \boldsymbol{\gamma}_d, \boldsymbol{\gamma}_n)$$

Step 3: Scalarize: Obtain equiv. low-dimensional problem, solve

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \ \mathfrak{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_t, \boldsymbol{\gamma}_d, \boldsymbol{\gamma}_n)$$

$$\approx \quad \min_{\boldsymbol{\zeta}\in\mathbb{R}^2} \ \overline{L}(\boldsymbol{\zeta}; \boldsymbol{\zeta}_t)$$

Obtain deterministic Gordon state evolution prediction

Step 4: Use growth conditions on the losses to make statements about minimizers. Empirical process theory

II: Prediction for a general class of problems

Model

i.i.d. observations: $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$

$$y_i = f(\langle \boldsymbol{x}_i, \boldsymbol{\theta}^* \rangle; q_i) + \epsilon_i$$

$$\boldsymbol{x}_i \sim \mathsf{N}(0, \boldsymbol{I}_d)$$

$$\epsilon_i = \mathsf{N}(0, \sigma^2)$$

$$q_i \sim \mathcal{Q}$$

$$\|\boldsymbol{\theta}^*\|_2 = 1$$

- Phase retrieval
- Mixtures of regressions
- Mixtures of single-index models

Algorithms

Higher-order methods

$$\boldsymbol{\theta}_{t+1} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{n} \|\omega(\boldsymbol{X}\boldsymbol{\theta}_t, \boldsymbol{y}) - \boldsymbol{X}\boldsymbol{\theta}\|_2^2$$

$$= \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \omega(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle, y_i) \cdot \boldsymbol{x}_i \right)$$

First-order methods

$$\boldsymbol{\theta}_{t+1} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \|\boldsymbol{\theta} - \boldsymbol{\theta}_t + \frac{2\eta}{n} \sum_{i=1}^n \omega(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle, y_i) \cdot \boldsymbol{x}_i\|_2^2$$

$$= \boldsymbol{\theta}_t - \frac{2\eta}{n} \sum_{i=1}^n \omega(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle, y_i) \cdot \boldsymbol{x}_i$$

- Alternating projections, Newton methods
- Expectation maximization (EM), Newton EM
- (Sub-)gradient descent, gradient EM

One-step Gordon update and deviation bound

$$\alpha = \alpha(\boldsymbol{\theta}_t)$$
$$\beta = \beta(\boldsymbol{\theta}_t)$$

$$Z_1, Z_2, Z_3 \overset{i.i.d.}{\sim} \mathsf{N}(0,1) \qquad Q \sim \mathcal{Q} \qquad \Omega(\alpha,\beta) := \omega(\alpha Z_1 + \beta Z_2, f(Z_1; Q) + \sigma Z_3)$$

	First-order	Higher-order
α^{gor}	$\alpha - 2\eta \cdot \mathbb{E}[Z_1\Omega]$	$\mathbb{E}[Z_1\Omega]$
β^{gor}	$\sqrt{(\beta - 2\eta \cdot \mathbb{E}[Z_2\Omega])^2 + 4\eta^2\mathbb{E}[\Omega^2]/\Lambda}$	$\sqrt{\mathbb{E}[Z_2\Omega]^2 + (\Lambda-1)^{-1} \cdot (\mathbb{E}[\Omega^2] - \mathbb{E}[Z_1\Omega]^2 - \mathbb{E}[Z_2\Omega]^2)}$

Theorem

If $\Lambda \geq C$ and $n \gtrsim \log(1/\delta)$, then with probability greater than $1-\delta$:

$$|\beta^{\mathsf{gor}} - \beta(\boldsymbol{\theta}_{t+1})| \lesssim \left(\frac{\log(1/\delta)}{n}\right)^{1/4} \quad \text{and} \quad |\alpha^{\mathsf{gor}} - \alpha(\boldsymbol{\theta}_{t+1})| \lesssim \left(\frac{\log^7(1/\delta)}{n}\right)^{1/2}.$$

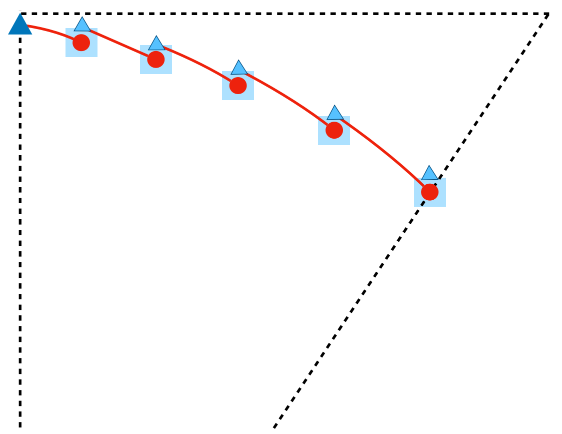
- Fully non-asymptotic result, parallel component concentration requires additional argument

Part II | Explicit one-step prediction for general class of models and methods

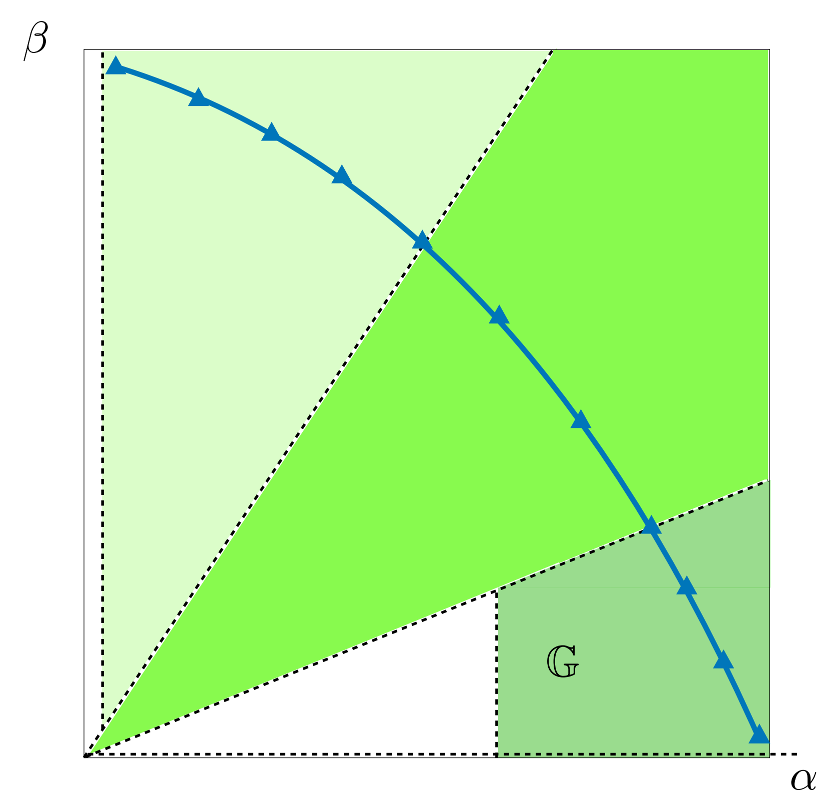
III: Convergence guarantees

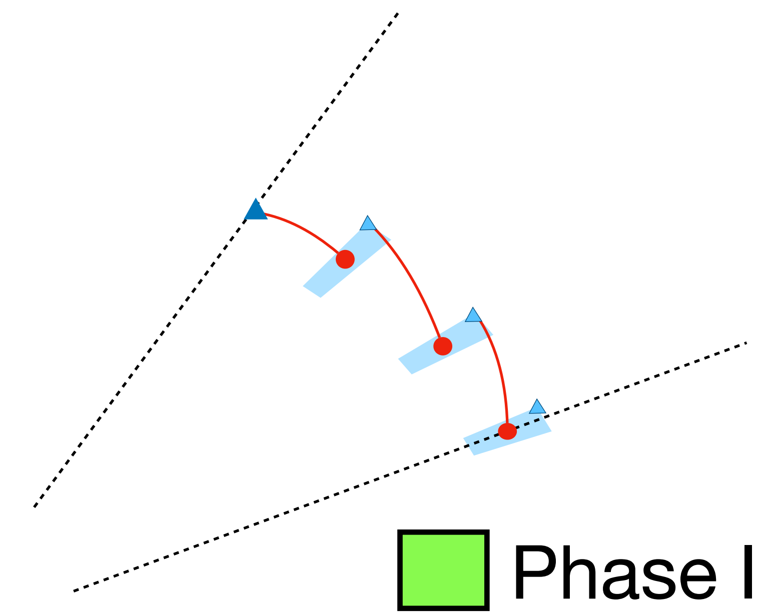
$$\alpha(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \boldsymbol{\theta}^* \rangle$$

$$\beta(\boldsymbol{\theta}) = \|\boldsymbol{P}_{\boldsymbol{\theta}^*}^{\perp} \boldsymbol{\theta}\|_2$$



Phase I

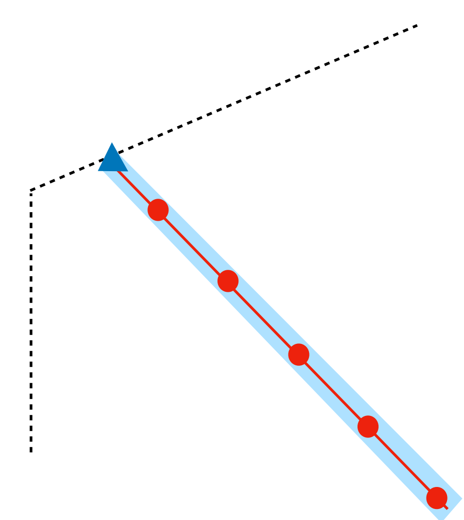




Phase II

- $\mathcal{O}(n^{-1/2})$—deviations on parallel component crucial!

Phase III



$$\zeta = (\alpha, \beta)$$

III: Convergence guarantees

- Natural parameter estimation metrics can be expressed in terms of state only:

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 = \sqrt{(1-\alpha)^2 + \beta^2} \qquad\qquad \angle(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \tan^{-1}(\beta/\alpha)$$

- Gordon state evolution operator: $\mathcal{S}_{\mathsf{gor}} : (\alpha, \beta) \mapsto (\alpha^{\mathsf{gor}}, \beta^{\mathsf{gor}})$

Linear: $$c \cdot \mathsf{d}(\zeta) + \varepsilon/2 \le \mathsf{d}(\mathcal{S}_{\mathsf{gor}}(\zeta)) \le C \cdot \mathsf{d}(\zeta) + \varepsilon$$

Superlinear: $$c \cdot [\mathsf{d}(\zeta)]^{\xi} + \varepsilon/2 \le \mathsf{d}(\mathcal{S}_{\mathsf{gor}}(\zeta)) \le C \cdot [\mathsf{d}(\zeta)]^{\xi} + \varepsilon$$

Example: Global convergence of AM for phase retrieval

Theorem

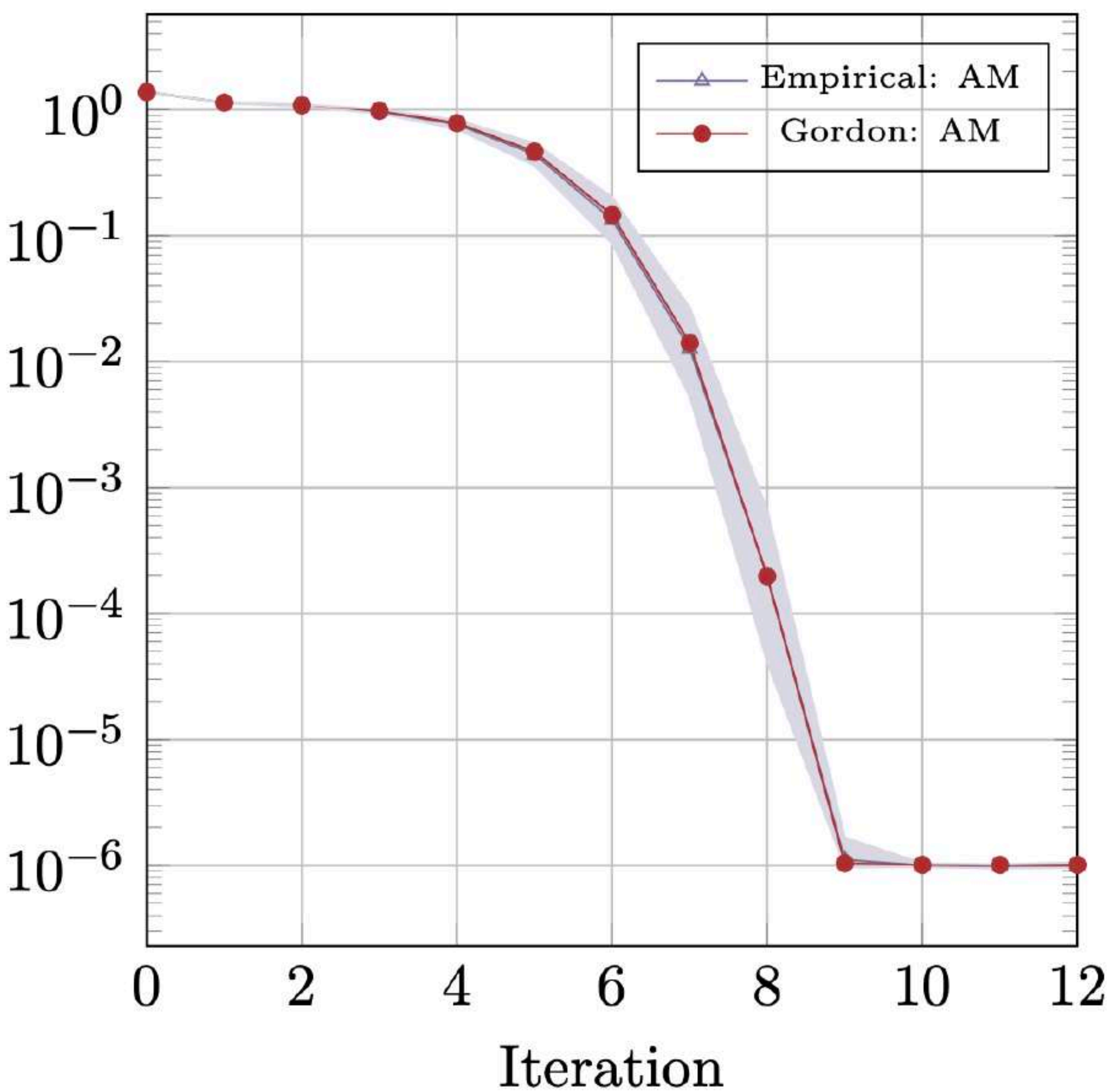
(a) The Gordon state evolution update converges in L2 superlinearly with exponent $3/2$ within the local region $\mathbb{G}$ to level $\varepsilon = \sigma\sqrt{d/n}$.

(b) If $\boldsymbol{\theta} \in \mathbb{G}$, then with probability at least $1 - 2Tn^{-10}$:

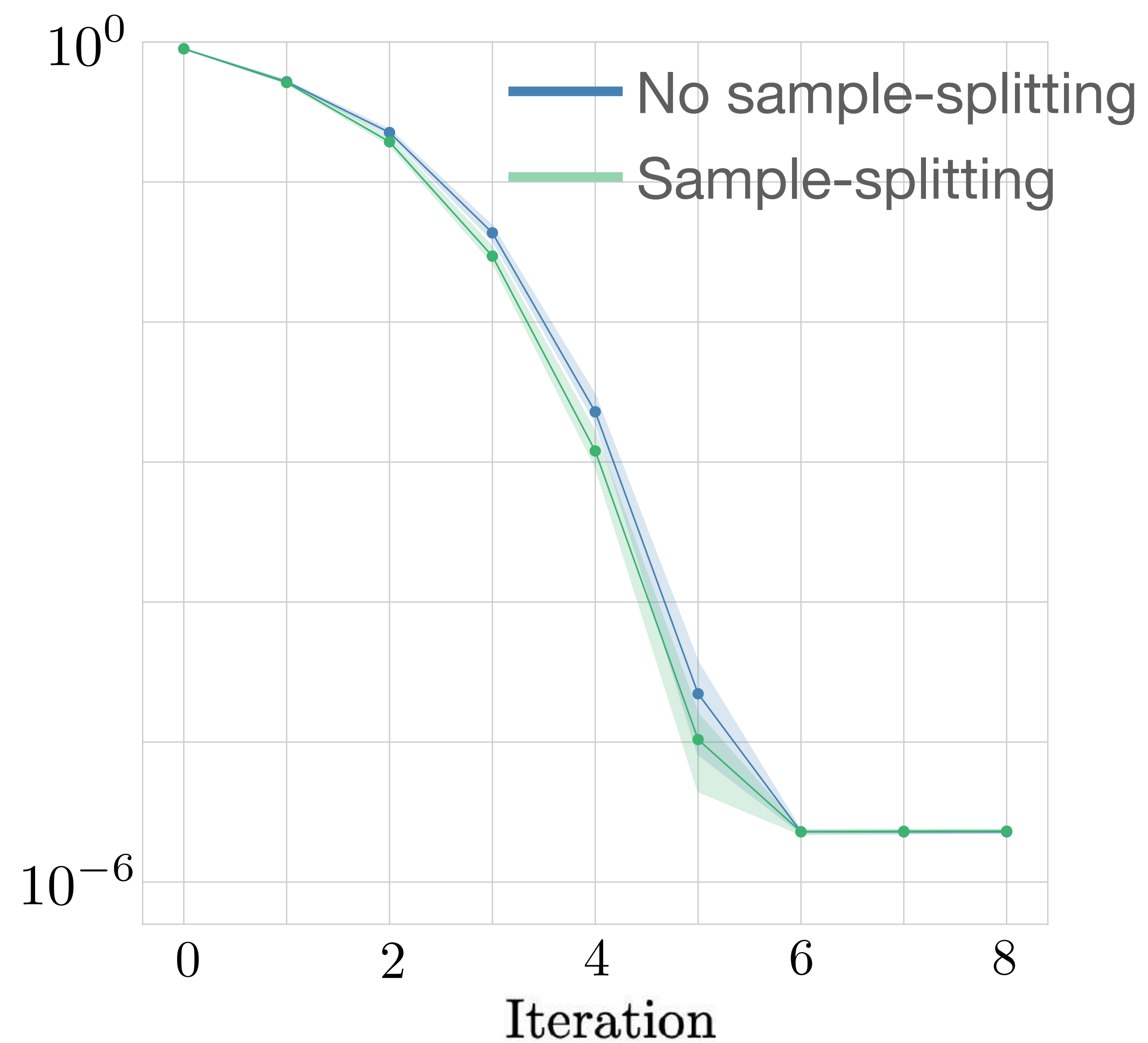
$$\max_{1 \le t \le T} |\mathsf{d}_{\ell_2}(\mathcal{S}^t_{\mathsf{gor}}(\boldsymbol{\zeta})) - \|\mathcal{T}^t_n(\boldsymbol{\theta}) - \boldsymbol{\theta}^*\|_2| \lesssim \left(\frac{\log n}{n}\right)^{1/4}.$$

- Parallel results for mixtures of regressions (angular not L2, AM converges linearly)
- Sample-splitting results in logarithmic blowup in total sample size

Global convergence prediction



Comparison: No sample-splitting



Zooming out: A vignette

- Key "meta" observation: Nonconvex optimization can be reduced to iterative convex M-estimation in high dimensions
- Drawback of Gordon approach: Suboptimal non asymptotic guarantees that necessitate additional work to prove global convergence

Question

Are there other ways to arrive at deterministic predictions with optimal concentration rates?

$$y_i = \langle \boldsymbol{x}_i, \boldsymbol{\mu}^\star \rangle \cdot \langle \boldsymbol{z}_i, \boldsymbol{\nu}^\star \rangle + \epsilon_i$$

$$\boldsymbol{x}_i, \boldsymbol{z}_i \sim \mathcal{N}(0, \boldsymbol{I})$$

$$\mathfrak{R}_n(\boldsymbol{\mu}, \boldsymbol{\nu}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \boldsymbol{x}_i, \boldsymbol{\mu} \rangle \cdot \langle \boldsymbol{z}_i, \boldsymbol{\nu} \rangle)^2$$

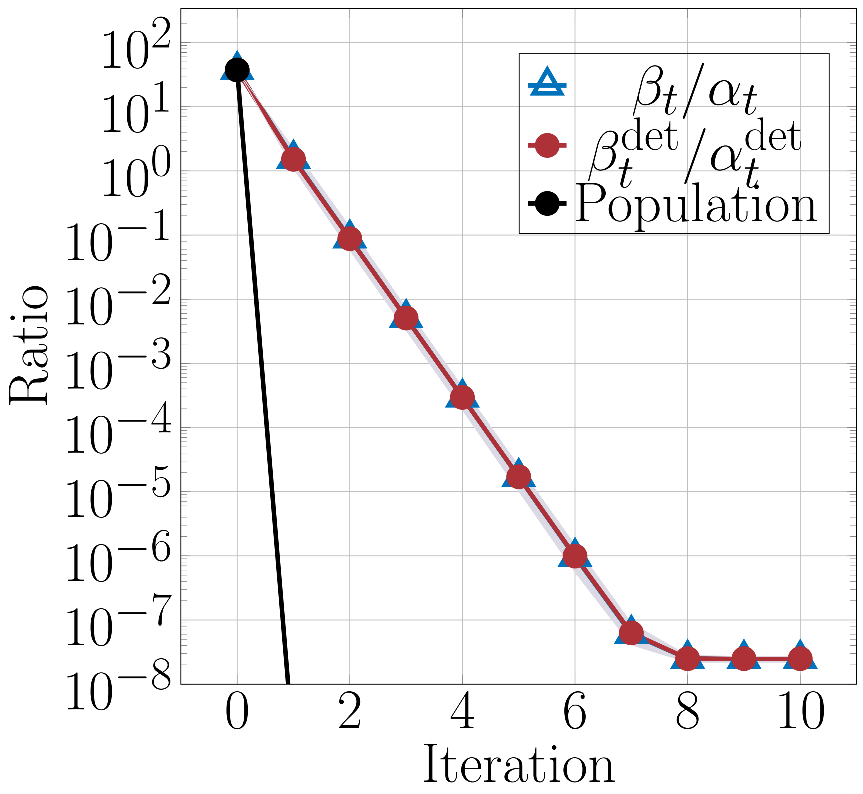
$$\boldsymbol{\nu}_{t+1} = \arg\min_{\nu \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \boldsymbol{x}_i, \boldsymbol{\mu}_t \rangle \cdot \langle \boldsymbol{z}_i, \boldsymbol{\nu} \rangle)^2$$

$$\boldsymbol{\mu}_{t+1} = \arg\min_{\mu \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle \boldsymbol{x}_i, \boldsymbol{\mu} \rangle \cdot \langle \boldsymbol{z}_i, \boldsymbol{\nu}_{t+1} \rangle)^2$$

- Also other popular higher-order methods, e.g., composite optimization

RMT-based prediction for AM in rank one bilinear sensing

- Uninformative population update
- Ratio (tangent of angle) converges linearly from random init. to noise floor
- Analysis enabled by “direct”, optimal, non-asymptotic concentration bound



Theorem

If $\Lambda \geq C$ and $n \gtrsim \log(1/\delta)$, then with probability greater than $1-\delta$:

$$|\alpha^{\mathsf{det}} - \alpha(\boldsymbol{\mu}_{t+1})| \lesssim \left(\frac{\mathsf{polylog}(n/\delta)}{n}\right)^{1/2} \quad \text{and} \quad |\beta^{\mathsf{det}} - \beta(\boldsymbol{\mu}_{t+1})| \lesssim \left(\frac{\mathsf{polylog}(n/\delta)}{n}\right)^{1/2}.$$

El Karoui et al. ’12-’18

- The population method can mis-predict efficiency in model-fitting
- Sharp characterizations of **convergence behavior** for iterative algorithms as well as **statistical accuracy** post-convergence.
- *Key properties:*
  - Each iteration is solution to convex optimization problem
  - Data is Gaussian conditioned on the past

Takeaways

- Removing the sample-splitting assumption
- Weakening the Gaussianity assumption
- Using sharp upper and lower bounds for "algorithmic" model-selection and hyperparameter tuning
- Broadly applicable machinery of reducing to iterative convex M-estimation: Can this say anything about your favorite model-fitting algorithm?

Open questions

Sharp global convergence guarantees for iterative nonconvex optimization with random data, with Chandrasekher and Thrampoulidis (under revision in Annals of Statistics)

Higher order methods for rank one bilinear sensing: Random initialization and sharp predictions, with Chandrasekher and Lou (coming soon)

Backup

$S_t = \mathsf{span}(\boldsymbol{\theta}^*, \boldsymbol{\theta}_t)$

Example derivation: AM for phase retrieval

Step 1: Write an iteration as solution to convex program; then write objective in bilinear form

$$\boldsymbol{\theta}_{t+1} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_t, \boldsymbol{X}, \boldsymbol{y})$$

Variational form/Fenchel conjugate

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_t, \boldsymbol{X}, \boldsymbol{y}) \overset{(d)}{=} \min_{\boldsymbol{u} \in \mathcal{U}} \max_{\boldsymbol{v} \in \mathcal{V}} \; \langle \boldsymbol{v}, \boldsymbol{G}\boldsymbol{u} \rangle + Q(\boldsymbol{u}, \boldsymbol{v})$$

Step 2: Invoke **CGMT** to replace matrix of Gaussians with two vectors

$$\begin{aligned}
&\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_t, \boldsymbol{X}, \boldsymbol{y}) \\
&\approx \min_{\boldsymbol{u} \in \mathcal{U}} \max_{\boldsymbol{v} \in \mathcal{V}} \; \|\boldsymbol{v}\|_2 \cdot \langle \boldsymbol{\gamma}_d, \boldsymbol{u} \rangle + \|\boldsymbol{u}\|_2 \cdot \langle \boldsymbol{\gamma}_n, \boldsymbol{v} \rangle \\
&\qquad + Q(\boldsymbol{u}, \boldsymbol{v}) \\
&=: \min_{\boldsymbol{\theta} \in \mathbb{R}^d} \; \mathfrak{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_t, \boldsymbol{\gamma}_d, \boldsymbol{\gamma}_n)
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\theta}_{t+1} &= \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \mathsf{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}_t \rangle) \cdot \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle)^2 \\
&= \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{1}{\sqrt{n}} \|\boldsymbol{X}\boldsymbol{\theta} - \mathsf{diag}(\mathsf{sign}(\boldsymbol{X}\boldsymbol{\theta}_t)) \cdot \boldsymbol{y}\|_2
\end{aligned}$$

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \max_{\|\boldsymbol{v}\|_2 \le 1} \langle \boldsymbol{v}, \boldsymbol{X}\boldsymbol{P}_{S_t}^{\perp}\boldsymbol{\theta} \rangle + \langle \boldsymbol{v}, \boldsymbol{X}\boldsymbol{P}_{S_t}\boldsymbol{\theta} \rangle - \langle \boldsymbol{v}, \mathsf{diag}(\mathsf{sign}(\boldsymbol{X}\boldsymbol{\theta}_t)) \cdot \boldsymbol{y} \rangle$$

$$\begin{aligned}
\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \max_{\|\boldsymbol{v}\|_2 \le 1} \; &\|\boldsymbol{v}\|_2 \cdot \langle \boldsymbol{\gamma}_d, \boldsymbol{P}_{S_t}^{\perp}\boldsymbol{\theta} \rangle + \|\boldsymbol{P}_{S_t}^{\perp}\boldsymbol{\theta}\|_2 \cdot \langle \boldsymbol{v}, \boldsymbol{\gamma}_n \rangle \\
&+ \langle \boldsymbol{v}, \boldsymbol{X}\boldsymbol{P}_{S_t}\boldsymbol{\theta} \rangle - \langle \boldsymbol{v}, \mathsf{diag}(\mathsf{sign}(\boldsymbol{X}\boldsymbol{\theta}_t)) \cdot \boldsymbol{y} \rangle
\end{aligned}$$

Gordon '85, '88 Stojnic '13 Thrampoulidis, Oymak, Hassibi '15

$\alpha = \langle \boldsymbol{\theta}, \boldsymbol{\theta}^* \rangle \quad \mu = \frac{\langle \boldsymbol{\theta}, \boldsymbol{P}^{\perp}_{\boldsymbol{\theta}^*} \boldsymbol{\theta}_t \rangle}{\|\boldsymbol{P}^{\perp}_{\boldsymbol{\theta}^*} \boldsymbol{\theta}_t\|_2} \quad \nu = \|\boldsymbol{P}^{\perp}_{S_t} \boldsymbol{\theta}\|_2 \quad \beta^2 = \mu^2 + \nu^2$

$S_t = \mathsf{span}(\boldsymbol{\theta}^*, \boldsymbol{\theta}_t)$

Step 3: Scalarize: Obtain equiv. low-dimensional problem, solve

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathfrak{L}(\boldsymbol{\theta}; \boldsymbol{\theta}_t, \boldsymbol{\gamma}_d, \boldsymbol{\gamma}_n)$$

$$\approx \min_{\boldsymbol{\zeta} \in \mathbb{R}^3} \overline{L}(\boldsymbol{\zeta}; \boldsymbol{\zeta}_t)$$

Step 4: Use growth conditions on the losses to make statements about minimizers.

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \left(-\frac{\|\boldsymbol{P}^{\perp}_{S_t} \boldsymbol{\theta}\|_2 \|\boldsymbol{P}^{\perp}_{S_t} \boldsymbol{\gamma}_d\|_2}{\sqrt{n}} + \|\mathsf{diag}(\mathsf{sign}(\boldsymbol{X}\boldsymbol{\theta}_t)) \cdot \boldsymbol{y} - \boldsymbol{X}\boldsymbol{P}_{S_t}\boldsymbol{\theta} - \|\boldsymbol{P}^{\perp}_{S_t}\boldsymbol{\theta}\|_2 \cdot \boldsymbol{\gamma}_n\|_2 \right)_+$$

$$\approx \min_{\substack{\alpha, \mu \\ \nu \geq 0}} \left(-\frac{\nu}{\sqrt{\Lambda}} + \sqrt{\mathbb{E}(\Omega_t - \alpha Z_1 - \mu Z_2 - \nu Z')^2} \right)_+$$

$\mathsf{sign}(\alpha_t Z_1 + \mu_t Z_2) \cdot |Z_1|$

$$\alpha^{\mathsf{gor}}_{t+1} = 1 - \frac{1}{\pi}(2\phi_t - \sin(2\phi_t)) \qquad \phi_t = \tan^{-1}(\beta_t / \alpha_t)$$

$$\beta^{\mathsf{gor}}_{t+1} = \sqrt{\frac{4}{\pi^2}\sin^4(\phi_t) + \frac{1 - (1 - \frac{1}{\pi}(2\phi_t - \sin(2\phi_t)))^2 - \frac{4}{\pi^2}\sin^4(\phi_t)}{\Lambda - 1}}$$

$\mathcal{O}(\beta_t^4)$

$\mathcal{O}(\beta_t^3)$