Implicit Regularization Properties of Early-Stopped Gradient-Based Algorithms

Varun Kanade

(based on joint work with Tomas Vaškevičius and Patrick Rebeschini)

DEPARTMENT OF
COMPUTER
SCIENCE

LMS-Bath Symposium: Mathematics of Machine Learning

August 3 2020

Types of Regularization

- Goal of machine learning: compute $h^\star \in \operatorname{argmin}_{h \in \mathcal{H}} \mathbb{E}[\ell(h, Z)]$

Types of Regularization

- Goal of machine learning: compute $h^\star \in \text{argmin}_{h \in \mathcal{H}} \mathbb{E}[\ell(h, Z)]$
- Unknown data distribution leads to minimize **ERM** $= \frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_i)$

Types of Regularization

- Goal of machine learning: compute $h^\star \in \operatorname{argmin}_{h \in \mathcal{H}} \mathbb{E}[\ell(h, Z)]$
- Unknown data distribution leads to minimize **ERM** $= \frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_i)$
- Finite amount of data ($n < \infty$) leads to **necessity of regularization:**

Types of Regularization

- Goal of machine learning: compute $h^\star \in \text{argmin}_{h \in \mathcal{H}} \mathbb{E}[\ell(h, Z)]$
- Unknown data distribution leads to minimize **ERM** $= \frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_i)$
- Finite amount of data ($n < \infty$) leads to **necessity of regularization:**

Explicit

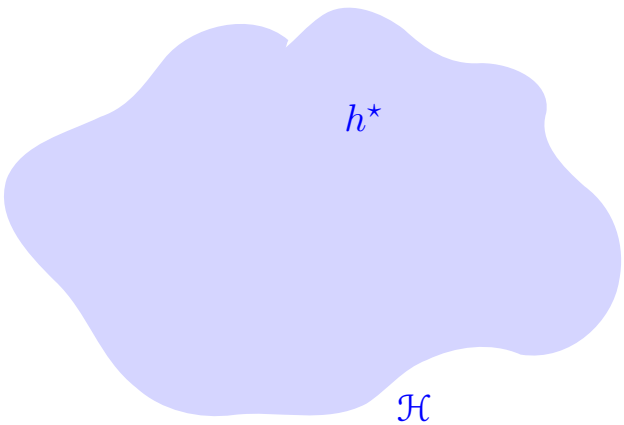
- Consider **ERM** $+ \lambda$**Pen**(h)
- **Statistics:** Choose penalty λ, **Pen**(h)
- **Optimization:** (try to) solve it

Types of Regularization

- Goal of machine learning: compute $h^\star \in \text{argmin}_{h \in \mathcal{H}} \mathbb{E}[\ell(h, Z)]$
- Unknown data distribution leads to minimize **ERM** $= \frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_i)$
- Finite amount of data ($n < \infty$) leads to **necessity of regularization:**

Explicit

- Consider **ERM** $+ \lambda$**Pen**(h)
- **Statistics:** Choose penalty λ, **Pen**(h)
- **Optimization:** (try to) solve it

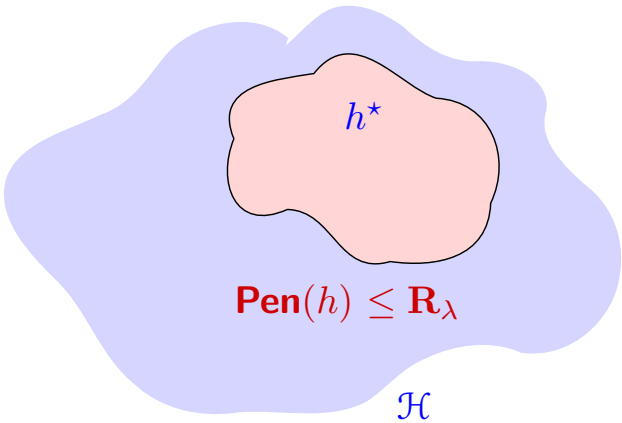


Types of Regularization

- Goal of machine learning: compute $h^\star \in \text{argmin}_{h \in \mathcal{H}} \mathbb{E}[\ell(h, Z)]$
- Unknown data distribution leads to minimize **ERM** $= \frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_i)$
- Finite amount of data ($n < \infty$) leads to **necessity of regularization:**

Explicit

– Consider **ERM** $+ \lambda$**Pen**(h)
– **Statistics:** Choose penalty λ, **Pen**(h)
– **Optimization:** (try to) solve it



Types of Regularization

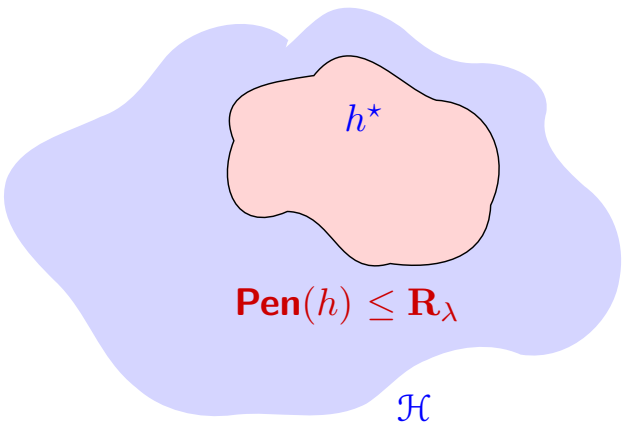
- Goal of machine learning: compute $h^\star \in \text{argmin}_{h \in \mathcal{H}} \mathbb{E}[\ell(h, Z)]$
- Unknown data distribution leads to minimize **ERM** $= \frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_i)$
- Finite amount of data ($n < \infty$) leads to **necessity of regularization:**

Explicit

- Consider **ERM** $+ \lambda$**Pen**(h)
- **Statistics:** Choose penalty λ, **Pen**(h)
- **Optimization:** (try to) solve it

Implicit

- Consider **unpenalized ERM**
- **Statistics + Optimization:** Choose:
  - **Parametrization**
  - **Solver**
  - **(Hyper)Parameters**

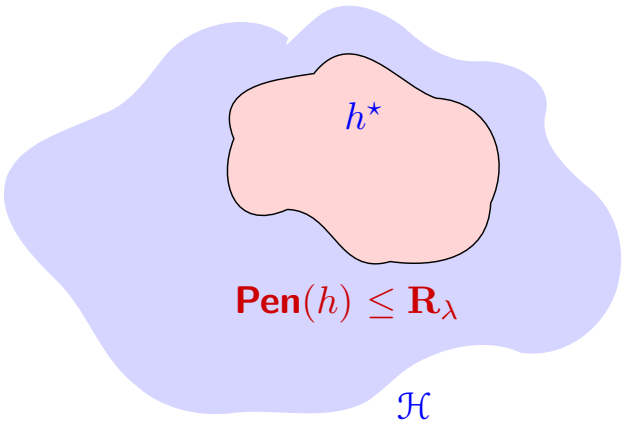


Types of Regularization

- Goal of machine learning: compute $h^\star \in \text{argmin}_{h \in \mathcal{H}} \mathbb{E}[\ell(h, Z)]$
- Unknown data distribution leads to minimize **ERM** $= \frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_i)$
- Finite amount of data ($n < \infty$) leads to **necessity of regularization:**

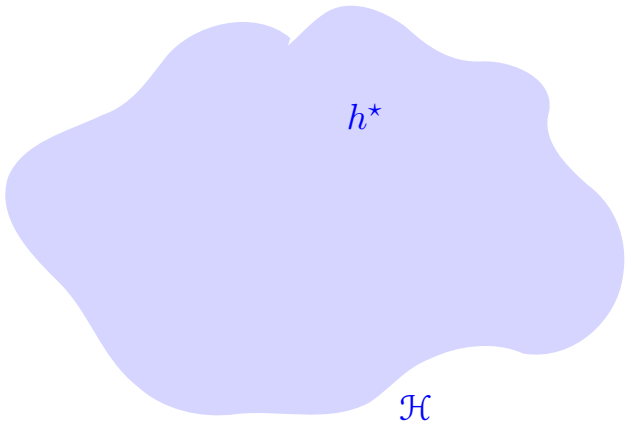
Explicit

- Consider **ERM** $+ \lambda$**Pen**(h)
- **Statistics:** Choose penalty λ, **Pen**(h)
- **Optimization:** (try to) solve it



Implicit

- Consider **unpenalized ERM**
- **Statistics + Optimization:** Choose:
  - **Parametrization**
  - **Solver**
  - **(Hyper)Parameters**

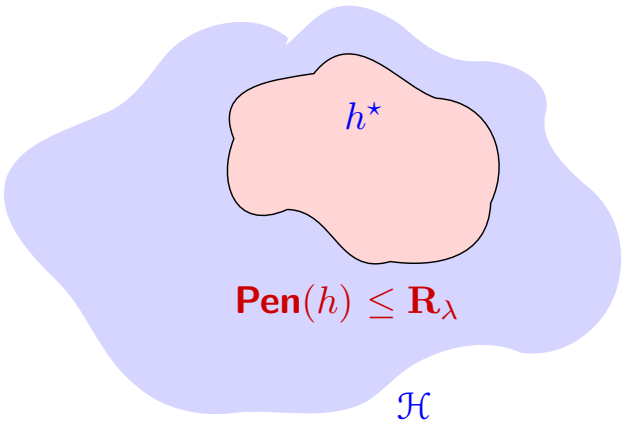


Types of Regularization

- Goal of machine learning: compute $h^\star \in \text{argmin}_{h \in \mathcal{H}} \mathbb{E}[\ell(h, Z)]$
- Unknown data distribution leads to minimize **ERM** $= \frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_i)$
- Finite amount of data ($n < \infty$) leads to **necessity of regularization:**

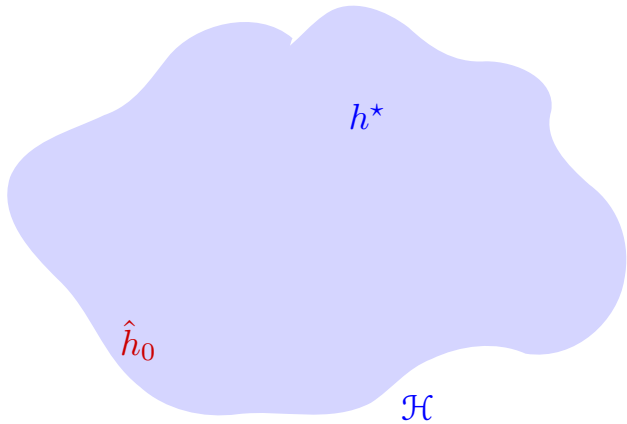
Explicit

- Consider **ERM** $+ \lambda$**Pen**(h)
- **Statistics:** Choose penalty λ, **Pen**(h)
- **Optimization:** (try to) solve it



Implicit

- Consider **unpenalized ERM**
- **Statistics + Optimization:** Choose:
  - **Parametrization**
  - **Solver**
  - **(Hyper)Parameters**

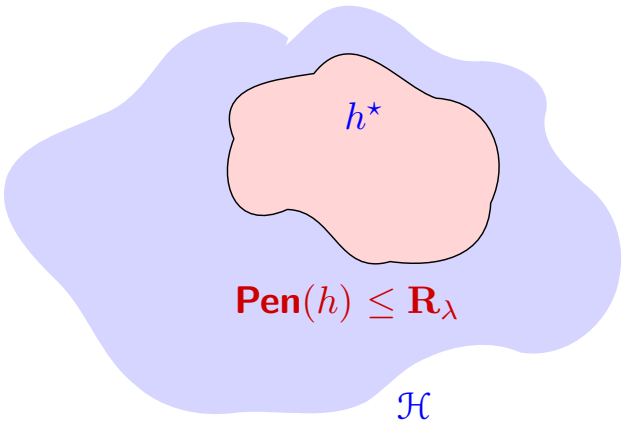


Types of Regularization

- Goal of machine learning: compute $h^\star \in \text{argmin}_{h \in \mathcal{H}} \mathbb{E}[\ell(h, Z)]$
- Unknown data distribution leads to minimize **ERM** $= \frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_i)$
- Finite amount of data ($n < \infty$) leads to **necessity of regularization:**

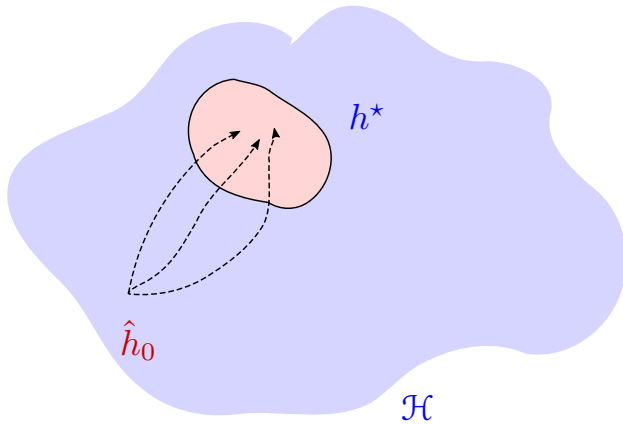
Explicit

- Consider **ERM** $+ \lambda$**Pen**(h)
- **Statistics:** Choose penalty λ, **Pen**(h)
- **Optimization:** (try to) solve it



Implicit

- Consider **unpenalized ERM**
- **Statistics + Optimization:** Choose:
  - **Parametrization**
  - **Solver**
  - **(Hyper)Parameters**

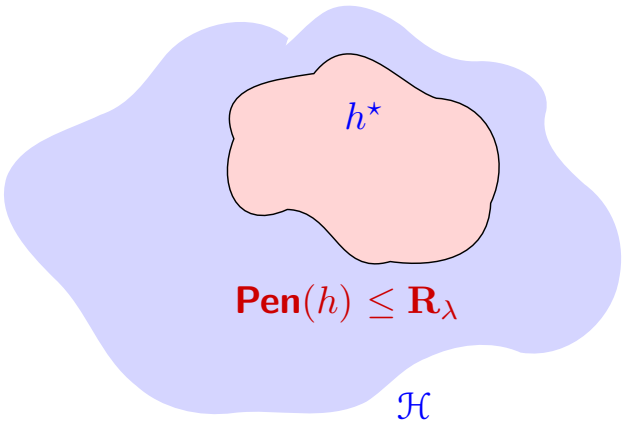


Types of Regularization

- Goal of machine learning: compute $h^\star \in \text{argmin}_{h \in \mathcal{H}} \mathbb{E}[\ell(h, Z)]$
- Unknown data distribution leads to minimize **ERM** $= \frac{1}{n} \sum_{i=1}^{n} \ell(h, Z_i)$
- Finite amount of data ($n < \infty$) leads to **necessity of regularization:**

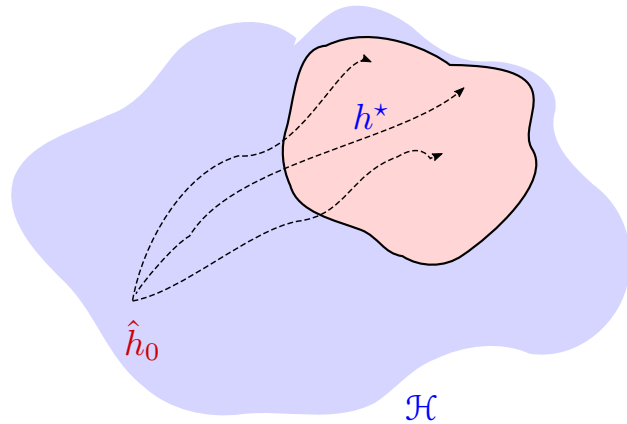
Explicit

- Consider **ERM** $+ \lambda$**Pen**(h)
- **Statistics:** Choose penalty λ, **Pen**(h)
- **Optimization:** (try to) solve it



Implicit

- Consider **unpenalized ERM**
- **Statistics + Optimization:** Choose:
  - **Parametrization**
  - **Solver**
  - **(Hyper)Parameters**



Implicit Regularization for Ridge Regression

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$$

Implicit Regularization for Ridge Regression

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$$

For example:

- **Statistics:**
  $\lambda^\star \sim \frac{1}{\sqrt{n}} \Rightarrow$ **error** $\lesssim \frac{1}{\sqrt{n}}$
  (minimax optimal rates)
  [Goldenshluger and Tsybakov, 2001, Caponnetto and De Vito, 2007], ...

Implicit Regularization for Ridge Regression

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$$

For example:

- **Statistics:**
  $\lambda^\star \sim \frac{1}{\sqrt{n}} \Rightarrow$ **error** $\lesssim \frac{1}{\sqrt{n}}$
  (minimax optimal rates)
  [Goldenshluger and Tsybakov, 2001, Caponnetto and De Vito, 2007], ...
- **Optimization:** strongly convex but $\lambda^\star \sim \frac{1}{\sqrt{n}} \Rightarrow$ high iteration complexity for GD $\Rightarrow$ Newton?

Implicit Regularization for Ridge Regression

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

Implicit

Statistics + Optimization:

For example:

- **Statistics:**
  $\lambda^\star \sim \frac{1}{\sqrt{n}} \Rightarrow$ **error** $\lesssim \frac{1}{\sqrt{n}}$
  (minimax optimal rates)
  [Goldenshluger and Tsybakov, 2001, Caponnetto and De Vito, 2007], ...
- **Optimization:** strongly convex but $\lambda^\star \sim \frac{1}{\sqrt{n}} \Rightarrow$ high iteration complexity for GD $\Rightarrow$ Newton?

Implicit Regularization for Ridge Regression

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

Implicit

Statistics + Optimization:

- **Parametrization**

$$\mathcal{L}(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2$$

For example:

– **Statistics:**
$\lambda^\star \sim \frac{1}{\sqrt{n}} \Rightarrow$ **error** $\lesssim \frac{1}{\sqrt{n}}$
(minimax optimal rates)
[Goldenshluger and Tsybakov, 2001, Caponnetto and De Vito, 2007], ...

– **Optimization:** strongly convex but $\lambda^\star \sim \frac{1}{\sqrt{n}} \Rightarrow$ high iteration complexity for GD $\Rightarrow$ Newton?

Implicit Regularization for Ridge Regression

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

For example:

- **Statistics:** $\lambda^\star \sim \frac{1}{\sqrt{n}} \Rightarrow$ **error** $\lesssim \frac{1}{\sqrt{n}}$ **(minimax optimal rates)** [Goldenshluger and Tsybakov, 2001, Caponnetto and De Vito, 2007], ...
- **Optimization:** strongly convex but $\lambda^\star \sim \frac{1}{\sqrt{n}} \Rightarrow$ high iteration complexity for GD $\Rightarrow$ Newton?

Implicit

Statistics + Optimization:

- **Parametrization**

$$\mathcal{L}(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2$$

- **Solver**

$$\mathbf{w}_0 = \mathbf{0}$$
$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\nabla\mathcal{L}(\mathbf{w}_t)$$

Implicit Regularization for Ridge Regression

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

For example:

- **Statistics:** $\lambda^\star \sim \frac{1}{\sqrt{n}} \Rightarrow$ **error** $\lesssim \frac{1}{\sqrt{n}}$ **(minimax optimal rates)** [Goldenshluger and Tsybakov, 2001, Caponnetto and De Vito, 2007], ...
- **Optimization:** strongly convex but $\lambda^\star \sim \frac{1}{\sqrt{n}} \Rightarrow$ high iteration complexity for GD $\Rightarrow$ Newton?

Implicit

Statistics + Optimization:

- **Parametrization**

$$\mathcal{L}(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2$$

- **Solver**

$$\mathbf{w}_0 = \mathbf{0}$$
$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\nabla\mathcal{L}(\mathbf{w}_t)$$

- **Parameters** most results establish connection at optimality:

$$\boxed{\eta t^\star \sim \frac{1}{\lambda^\star}}$$

[Bühlmann and Yu, 2003, Yao et al., 2007, Bauer et al., 2007, Raskutti et al., 2014],...

Implicit Regularization for Ridge Regression

But even stronger results for the **optimization path:**
(connections already established in prior literature)

- **Empirically:** [Friedman and Popescu, 2004]
- **Theory:** for Gradient Flow ($\eta \to 0$), with **no assumptions on $\mathbf{X}$**, we have: [Suggala et al., 2018, Ali et al., 2019]

$$\boxed{t \sim \frac{1}{\lambda}} \qquad \textbf{for all } t \text{ and } \lambda$$

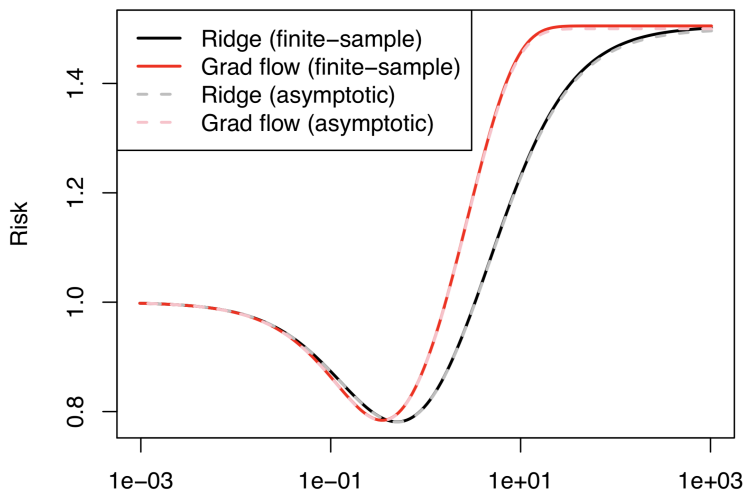


Figure: Risk versus t or $\frac{1}{\lambda}$. Figure taken from [Ali et al., 2019]

Implicit Regularization for Ridge Regression

Strong Connection between GD and Ridge Regression

This has motivated a lot of research on **computationally efficient methods:**

- Acceleration
- Stochastic methods
- Mini-batching
- Averaging
- Sketching
- Sub-sampling
- Preconditioning
- Parallel and distributed architectures
- ...

Success story for Ridge Regression. What about sparse recovery?

Implicit Regularization for Sparse Recovery?

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1$$

Implicit Regularization for Sparse Recovery?

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1$$

Related to our setting:

- **Statistics:** $\lambda^\star \sim \sigma\sqrt{\log d}/\sqrt{n}$
  $\Rightarrow$ **error** $\lesssim \sigma\sqrt{k \log d}/\sqrt{n}$
  (minimax optimal rates)
  [Wainwright, 2019]—book

Implicit Regularization for Sparse Recovery?

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1$$

Related to our setting:

- **Statistics:** $\lambda^\star \sim \sigma\sqrt{\log d}/\sqrt{n}$
  $\Rightarrow$ **error** $\lesssim \sigma\sqrt{k \log d}/\sqrt{n}$
  (minimax optimal rates)
  [Wainwright, 2019]—book
- **Opt:** prox. methods (ISTA, etc.)
  [Bach et al., 2012]—monograph
  $\widetilde{O}(1)$ iteration complexity
  [Agarwal et al., 2010],...
  $\Rightarrow \widetilde{O}(nd)$ comp. complexity
  (computational optimality)

Implicit Regularization for Sparse Recovery?

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1$$

Related to our setting:

- **Statistics:** $\lambda^\star \sim \sigma\sqrt{\log d}/\sqrt{n}$
  $\Rightarrow$ **error** $\lesssim \sigma\sqrt{k\log d}/\sqrt{n}$
  (minimax optimal rates)
  [Wainwright, 2019]—book
- **Opt:** prox. methods (ISTA, etc.)
  [Bach et al., 2012]—monograph
  $\widetilde{O}(1)$ iteration complexity
  [Agarwal et al., 2010],...
  $\Rightarrow \widetilde{O}(nd)$ comp. complexity
  (computational optimality)

Implicit

Implicit Regularization for Sparse Recovery?

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1$$

Most literature on implicit reg. for sparse recovery deals with:

- **Limit statements:**
  - At convergence: $t \to \infty$
  - Inifinitesimal step size: $\eta \to 0$
  - Infinitesimal initial. size: $\alpha \to 0$
- **No noise (or limited noise)**
- **No computational efficiency**

Implicit

Implicit Regularization for Sparse Recovery?

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1$$

Most literature on implicit reg. for sparse recovery deals with:

- **Limit statements:**
  - At convergence: $t \to \infty$
  - Inifinitesimal step size: $\eta \to 0$
  - Infinitesimal initial. size: $\alpha \to 0$
- **No noise (or limited noise)**
- **No computational efficiency**

Implicit

Literature connected to ℓ_1-norm:

- **Coordinate-Descent / AdaBoost:** [Hastie et al., 2001, Efron et al., 2004, Rosset et al., 2004, Zhang and Yu, 2005],...

Implicit Regularization for Sparse Recovery?

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1$$

Most literature on implicit reg. for sparse recovery deals with:

- **Limit statements:**
  - At convergence: $t \to \infty$
  - Inifinitesimal step size: $\eta \to 0$
  - Infinitesimal initial. size: $\alpha \to 0$
- **No noise (or limited noise)**
- **No computational efficiency**

Implicit

Literature connected to ℓ_1-norm:

- **Coordinate-Descent / AdaBoost:** [Hastie et al., 2001, Efron et al., 2004, Rosset et al., 2004, Zhang and Yu, 2005],...
- **Steepest Descent:** [Gunasekar et al., 2018]

Implicit Regularization for Sparse Recovery?

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1$$

Most literature on implicit reg. for sparse recovery deals with:

- **Limit statements:**
  - At convergence: $t \to \infty$
  - Inifinitesimal step size: $\eta \to 0$
  - Infinitesimal initial. size: $\alpha \to 0$
- **No noise (or limited noise)**
- **No computational efficiency**

Implicit

Literature connected to ℓ_1-norm:

- **Coordinate-Descent / AdaBoost:** [Hastie et al., 2001, Efron et al., 2004, Rosset et al., 2004, Zhang and Yu, 2005],...
- **Steepest Descent:** [Gunasekar et al., 2018]
- **Gradient Flow:** Low-rank matrix recovery with param. $\mathbf{W} = \mathbf{U}\mathbf{U}^\mathsf{T}$ [Gunasekar et al., 2017, Li et al., 2018] **RIP**

  conjecture: for $\mathbf{w}^\star \overset{\Downarrow}{\succcurlyeq} 0$ and $\mathbf{w}_t = \mathbf{u}_t \odot \mathbf{u}_t$ GD on $\mathbf{u}_t$ yields min. ℓ_1-**norm** solution

Implicit Regularization for Sparse Recovery?

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1$$

Most literature on implicit reg. for sparse recovery deals with:

- **Limit statements:**
  - At convergence: $t \to \infty$
  - Inifinitesimal step size: $\eta \to 0$
  - Infinitesimal initial. size: $\alpha \to 0$
- **No noise (or limited noise)**
- **No computational efficiency**

Implicit

Literature connected to ℓ_1-norm:

- **Coordinate-Descent / AdaBoost:** [Hastie et al., 2001, Efron et al., 2004, Rosset et al., 2004, Zhang and Yu, 2005],...
- **Steepest Descent:** [Gunasekar et al., 2018]
- **Gradient Flow:** Low-rank matrix recovery with param. $\mathbf{W} = \mathbf{U}\mathbf{U}^\mathsf{T}$ [Gunasekar et al., 2017, Li et al., 2018] **RIP**

  conjecture: for $\mathbf{w}^\star \overset{\Downarrow}{\succcurlyeq} 0$ and $\mathbf{w}_t = \mathbf{u}_t \odot \mathbf{u}_t$ GD on $\mathbf{u}_t$ yields min. ℓ_1-**norm** solution
- **Gradient Descent:** **Zhao et al. [2019]** **(concurrent work, more on this later)**

Implicit Regularization for Sparse Recovery?

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1$$

Most literature on implicit reg. for sparse recovery deals with:

- **Limit statements:**
  - At convergence: $t \to \infty$
  - Inifinitesimal step size: $\eta \to 0$
  - Infinitesimal initial. size: $\alpha \to 0$
- **No noise (or limited noise)**
- **No computational efficiency**

Implicit

Literature connected to ℓ_1-norm:

- **Coordinate-Descent / AdaBoost:** [Hastie et al., 2001, Efron et al., 2004, Rosset et al., 2004, Zhang and Yu, 2005],...
- **Steepest Descent:** [Gunasekar et al., 2018]
- **Gradient Flow:** Low-rank matrix recovery with param. $\mathbf{W} = \mathbf{U}\mathbf{U}^\mathsf{T}$ [Gunasekar et al., 2017, Li et al., 2018] **RIP**

  conjecture: for $\mathbf{w}^\star \overset{\Downarrow}{\succcurlyeq} 0$ and $\mathbf{w}_t = \mathbf{u}_t \odot \mathbf{u}_t$ GD on $\mathbf{u}_t$ yields min. ℓ_1-**norm** solution
- **Gradient Descent:** **Zhao et al. [2019] (concurrent work, more on this later)**

Q. Can build a theory of early stopping for optimal noisy sparse recovery?

Implicit Regularization for Sparse Recovery

Explicit **Implicit**

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1$$

Implicit Regularization for Sparse Recovery

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1$$

Implicit

- **Parametrization**

Implicit Regularization for Sparse Recovery

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1$$

Implicit

- **Parametrization**

$$\mathcal{L}(\mathbf{u}, \mathbf{v}) = \frac{1}{n} \|\mathbf{X}(\mathbf{u} \odot \mathbf{u} - \mathbf{v} \odot \mathbf{v}) - \mathbf{y}\|_2^2$$

Implicit Regularization for Sparse Recovery

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1$$

Implicit

- **Parametrization**

  $$\mathcal{L}(\mathbf{u}, \mathbf{v}) = \frac{1}{n} \|\mathbf{X}(\mathbf{u} \odot \mathbf{u} - \mathbf{v} \odot \mathbf{v}) - \mathbf{y}\|_2^2$$

- **Solver**

Implicit Regularization for Sparse Recovery

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1$$

Implicit

- **Parametrization**

$$\mathcal{L}(\mathbf{u},\mathbf{v}) = \frac{1}{n}\|\mathbf{X}(\mathbf{u}\odot\mathbf{u} - \mathbf{v}\odot\mathbf{v}) - \mathbf{y}\|_2^2$$

- **Solver** GD on $\mathbf{u}$ and $\mathbf{v}$:

$$\mathbf{u}_0 = \mathbf{v}_0 = \boldsymbol{\alpha}\mathbf{1}$$

$$\mathbf{u}_{t+1} = \mathbf{u}_t - \eta\frac{\partial\mathcal{L}(\mathbf{u}_t,\mathbf{v}_t)}{\partial\mathbf{u}_t}$$

$$\mathbf{v}_{t+1} = \mathbf{v}_t - \eta\frac{\partial\mathcal{L}(\mathbf{u}_t,\mathbf{v}_t)}{\partial\mathbf{v}_t}$$

$$\mathbf{w}_{t+1} = \mathbf{u}_{t+1}\odot\mathbf{u}_{t+1} - \mathbf{v}_{t+1}\odot\mathbf{v}_{t+1}$$

Implicit Regularization for Sparse Recovery

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1$$

Intuition:

- Let $\mathbf{w}_t = \mathbf{u}_t \odot \mathbf{u}_t$
- GD is tied to ℓ_2 geometry
- GD on $\mathbf{u}$ should be tied to ℓ_1 for $\mathbf{w}$:

$$\|\mathbf{u}_t\|_2^2 = \|\mathbf{w}_t\|_1$$

Implicit

- **Parametrization**

$$\mathcal{L}(\mathbf{u},\mathbf{v}) = \frac{1}{n}\|\mathbf{X}(\mathbf{u}\odot\mathbf{u} - \mathbf{v}\odot\mathbf{v}) - \mathbf{y}\|_2^2$$

- **Solver** GD on $\mathbf{u}$ and $\mathbf{v}$:

$$\mathbf{u}_0 = \mathbf{v}_0 = \boldsymbol{\alpha}\mathbf{1}$$

$$\mathbf{u}_{t+1} = \mathbf{u}_t - \eta\frac{\partial\mathcal{L}(\mathbf{u}_t,\mathbf{v}_t)}{\partial\mathbf{u}_t}$$

$$\mathbf{v}_{t+1} = \mathbf{v}_t - \eta\frac{\partial\mathcal{L}(\mathbf{u}_t,\mathbf{v}_t)}{\partial\mathbf{v}_t}$$

$$\mathbf{w}_{t+1} = \mathbf{u}_{t+1}\odot\mathbf{u}_{t+1} - \mathbf{v}_{t+1}\odot\mathbf{v}_{t+1}$$

Implicit Regularization for Sparse Recovery

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1$$

Intuition:

- Let $\mathbf{w}_t = \mathbf{u}_t \odot \mathbf{u}_t$
- GD is tied to ℓ_2 geometry
- GD on $\mathbf{u}$ should be tied to ℓ_1 for $\mathbf{w}$:

$$\|\mathbf{u}_t\|_2^2 = \|\mathbf{w}_t\|_1$$

Implicit

- **Parametrization**

$$\mathcal{L}(\mathbf{u},\mathbf{v}) = \frac{1}{n}\|\mathbf{X}(\mathbf{u}\odot\mathbf{u} - \mathbf{v}\odot\mathbf{v}) - \mathbf{y}\|_2^2$$

- **Solver** GD on $\mathbf{u}$ and $\mathbf{v}$:

$$\mathbf{u}_0 = \mathbf{v}_0 = \alpha\mathbf{1}$$

$$\mathbf{u}_{t+1} = \mathbf{u}_t - \eta\frac{\partial\mathcal{L}(\mathbf{u}_t,\mathbf{v}_t)}{\partial\mathbf{u}_t}$$

$$\mathbf{v}_{t+1} = \mathbf{v}_t - \eta\frac{\partial\mathcal{L}(\mathbf{u}_t,\mathbf{v}_t)}{\partial\mathbf{v}_t}$$

$$\mathbf{w}_{t+1} = \mathbf{u}_{t+1}\odot\mathbf{u}_{t+1} - \mathbf{v}_{t+1}\odot\mathbf{v}_{t+1}$$

- **Parameters** for minimax results:

Implicit Regularization for Sparse Recovery

Explicit

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1$$

Intuition:

- Let $\mathbf{w}_t = \mathbf{u}_t \odot \mathbf{u}_t$
- GD is tied to ℓ_2 geometry
- GD on $\mathbf{u}$ should be tied to ℓ_1 for $\mathbf{w}$:

$$\|\mathbf{u}_t\|_2^2 = \|\mathbf{w}_t\|_1$$

Implicit

- **Parametrization**

$$\mathcal{L}(\mathbf{u},\mathbf{v}) = \frac{1}{n}\|\mathbf{X}(\mathbf{u}\odot\mathbf{u} - \mathbf{v}\odot\mathbf{v}) - \mathbf{y}\|_2^2$$

- **Solver** GD on $\mathbf{u}$ and $\mathbf{v}$:

$$\mathbf{u}_0 = \mathbf{v}_0 = \boldsymbol{\alpha}\mathbf{1}$$

$$\mathbf{u}_{t+1} = \mathbf{u}_t - \eta\frac{\partial\mathcal{L}(\mathbf{u}_t,\mathbf{v}_t)}{\partial\mathbf{u}_t}$$

$$\mathbf{v}_{t+1} = \mathbf{v}_t - \eta\frac{\partial\mathcal{L}(\mathbf{u}_t,\mathbf{v}_t)}{\partial\mathbf{v}_t}$$

$$\mathbf{w}_{t+1} = \mathbf{u}_{t+1}\odot\mathbf{u}_{t+1} - \mathbf{v}_{t+1}\odot\mathbf{v}_{t+1}$$

- **Parameters** for minimax results:

$$\boxed{\frac{\eta t^\star}{\log\frac{1}{\alpha}} \sim \frac{1}{\lambda^\star}}$$

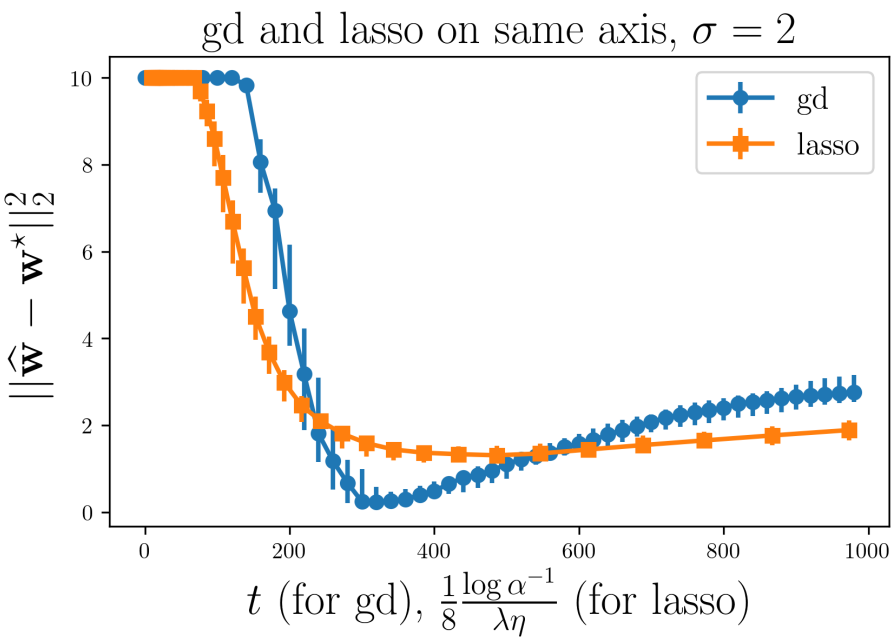
but optimization path is different..

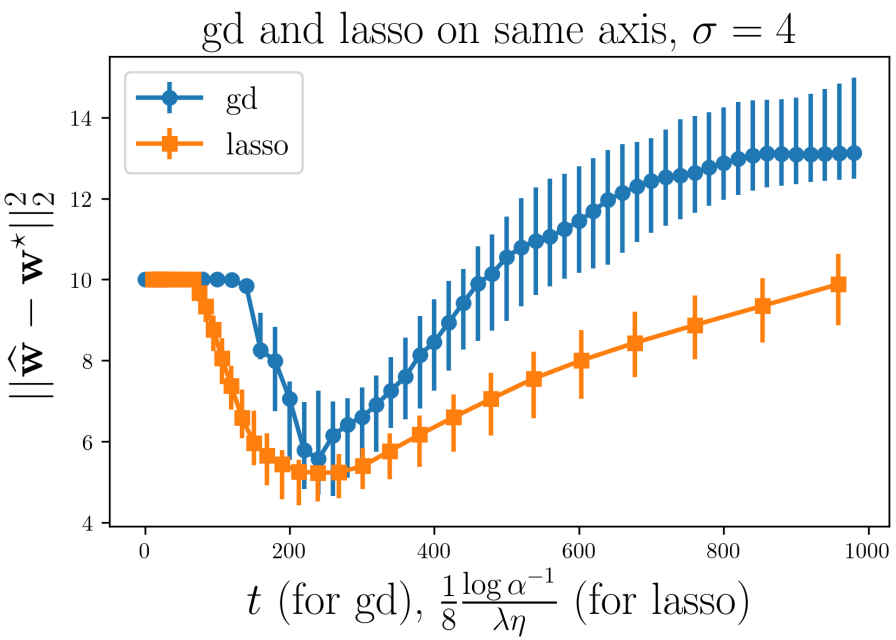
Implicit Regularization for Sparse Recovery

- For minimax rates, we can prove:

$$\frac{\eta t^\star}{\log \frac{1}{\alpha}} \sim \frac{1}{\lambda^\star}$$

- But **opt. paths and properties of estimators (GD vs. Lasso) are different**





On Parametrization and Multiplicative Updates

- Parametrization previously used in:
  - Hoff [2017]: to turn (convex) non-smooth program into (non-convex) smooth
  - Gunasekar et al. [2017]: to address matrix sensing

On Parametrization and Multiplicative Updates

- Parametrization previously used in:
  - Hoff [2017]: to turn (convex) non-smooth program into (non-convex) smooth
  - Gunasekar et al. [2017]: to address matrix sensing

- Parameterization turns **additive** updates into **multiplicative** updates:

$$\mathbf{u}_{t+1} = \mathbf{u}_t \odot \left(\mathbb{1} - 4\eta \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} \left(\mathbf{w}_t - \mathbf{w}^\star \right) - \frac{1}{n} \mathbf{X}^\top \xi \right) \right)$$

$$\mathbf{v}_{t+1} = \mathbf{v}_t \odot \left(\mathbb{1} + 4\eta \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} \left(\mathbf{w}_t - \mathbf{w}^\star \right) - \frac{1}{n} \mathbf{X}^\top \xi \right) \right)$$

On Parametrization and Multiplicative Updates

- Parametrization previously used in:
  - Hoff [2017]: to turn (convex) non-smooth program into (non-convex) smooth
  - Gunasekar et al. [2017]: to address matrix sensing

- Parameterization turns **additive** updates into **multiplicative** updates:

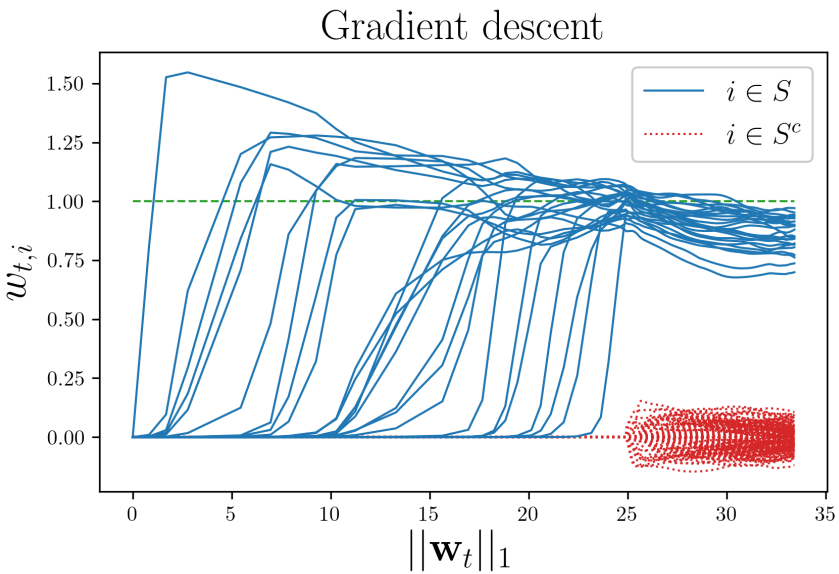
$$\mathbf{u}_{t+1} = \mathbf{u}_t \odot \left(\mathbb{1} - 4\eta \left(\frac{1}{n} \mathbf{X}^\mathsf{T} \mathbf{X} \left(\mathbf{w}_t - \mathbf{w}^\star \right) - \frac{1}{n} \mathbf{X}^\mathsf{T} \xi \right) \right)$$

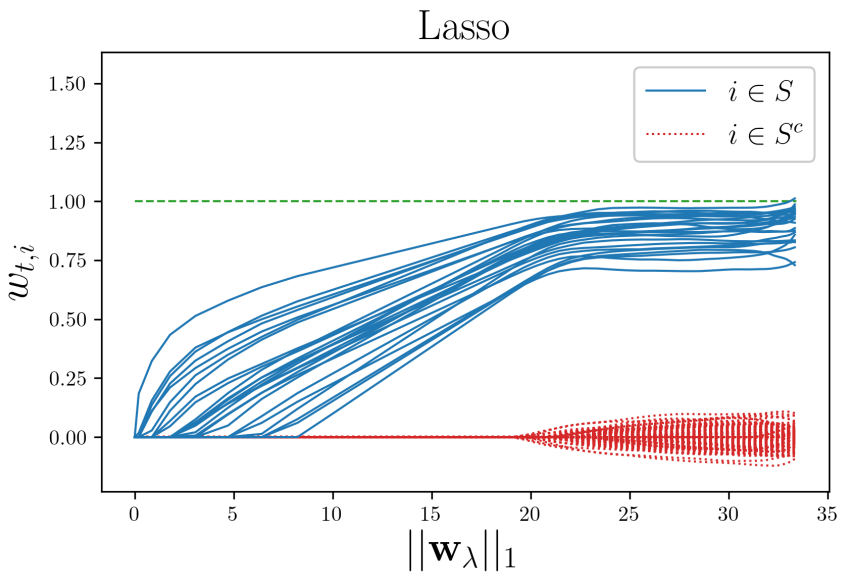
$$\mathbf{v}_{t+1} = \mathbf{v}_t \odot \left(\mathbb{1} + 4\eta \left(\frac{1}{n} \mathbf{X}^\mathsf{T} \mathbf{X} \left(\mathbf{w}_t - \mathbf{w}^\star \right) - \frac{1}{n} \mathbf{X}^\mathsf{T} \xi \right) \right)$$

- Compare to updates on canonical parametrization $\mathcal{L}(\mathbf{w})$ (for Ridge):

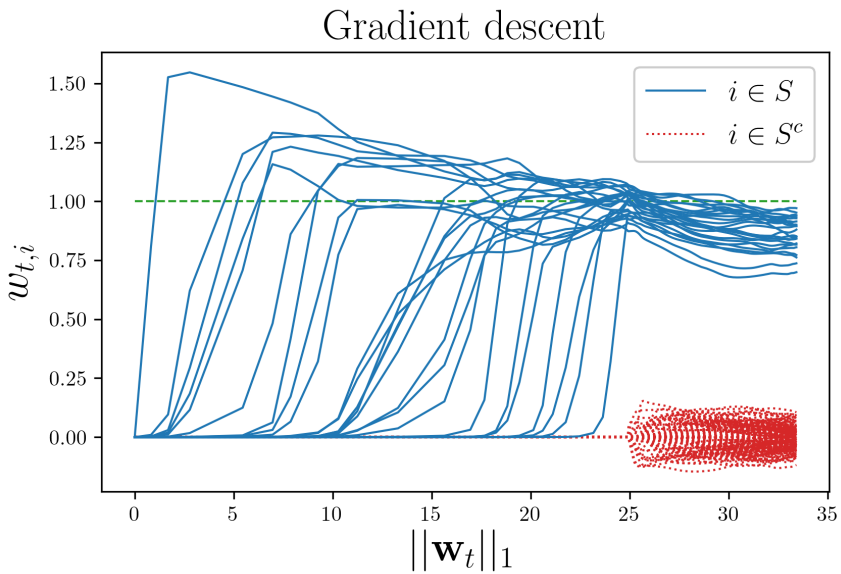
$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla \mathcal{L}(\mathbf{w}_t) = \mathbf{w}_t - \frac{2\eta}{n} \left(\mathbf{X}^\mathsf{T} \mathbf{X} (\mathbf{w}_t - \mathbf{w}^\star) - \mathbf{X}^\mathsf{T} \xi \right)$$

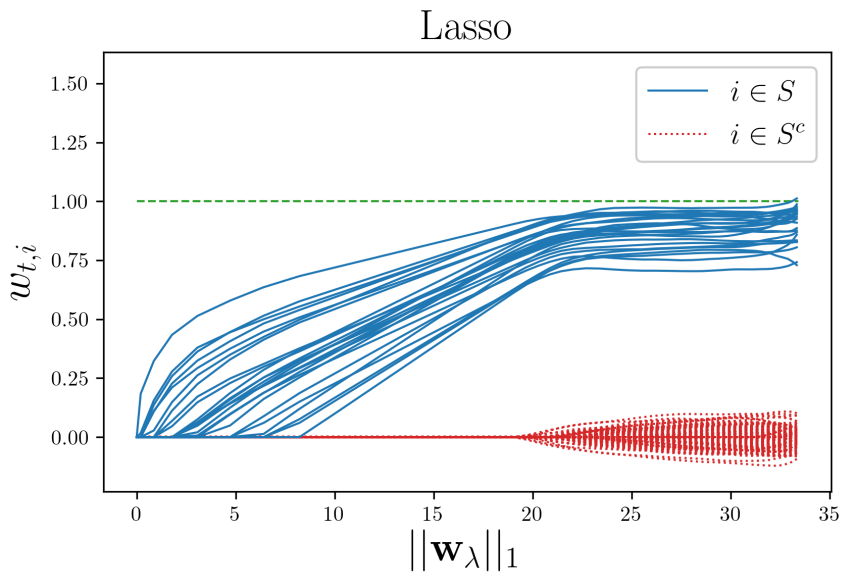
Comparison with Lasso





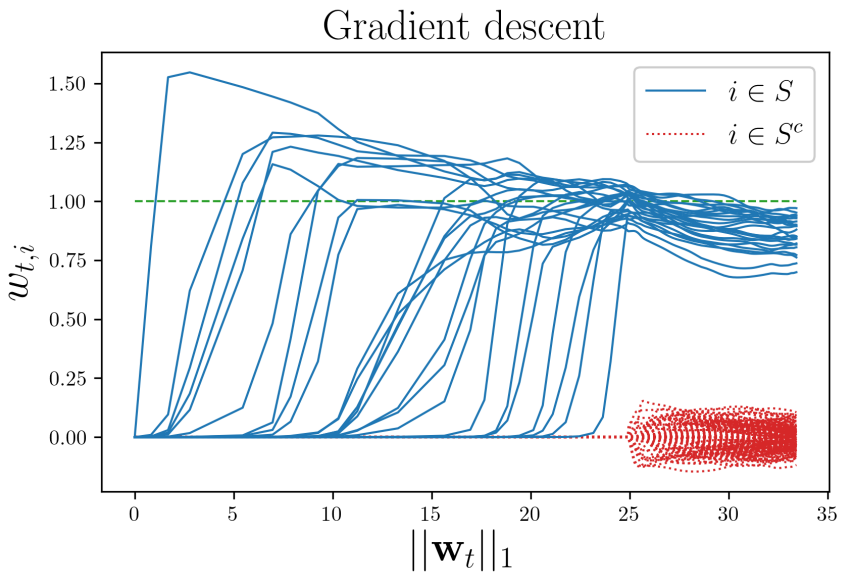
Comparison with Lasso

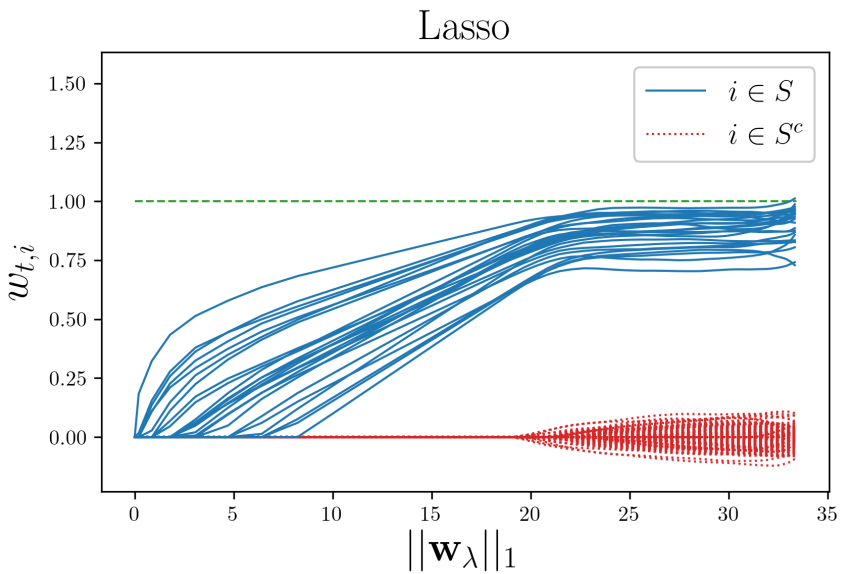




Similar to Lasso: Sparse iterates/solutions, minimax rates

Comparison with Lasso

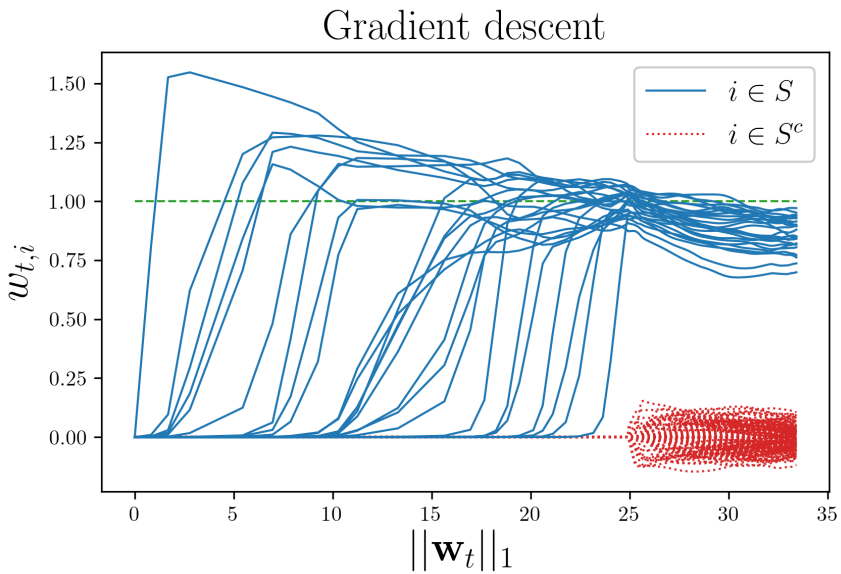


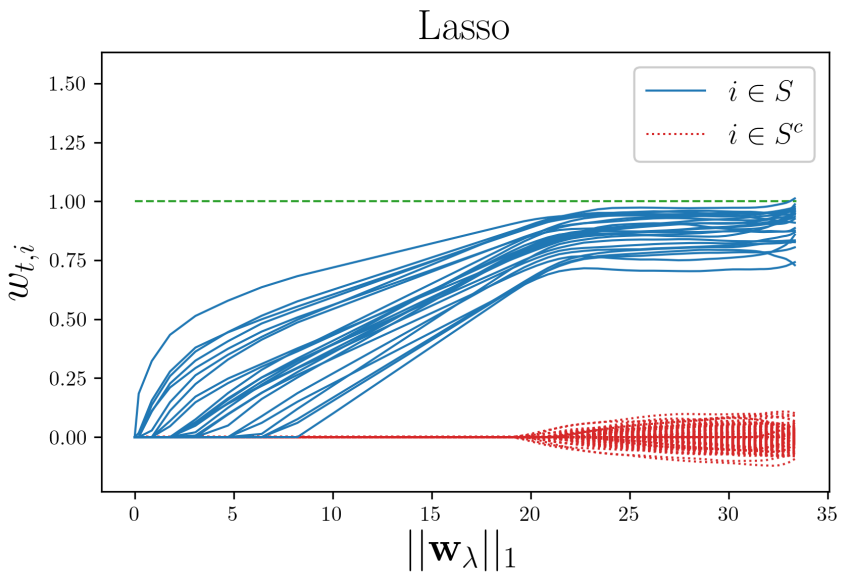


Similar to Lasso: Sparse iterates/solutions, minimax rates

Different than Lasso:

Comparison with Lasso



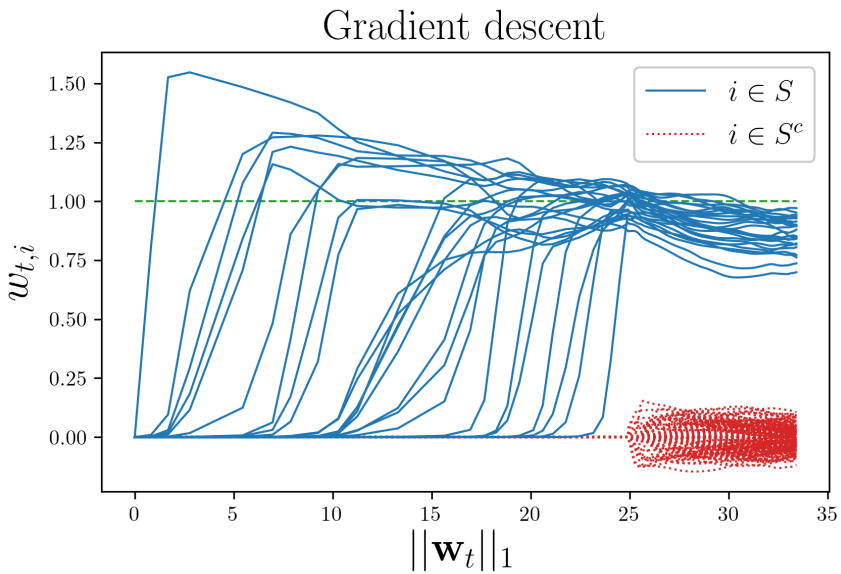


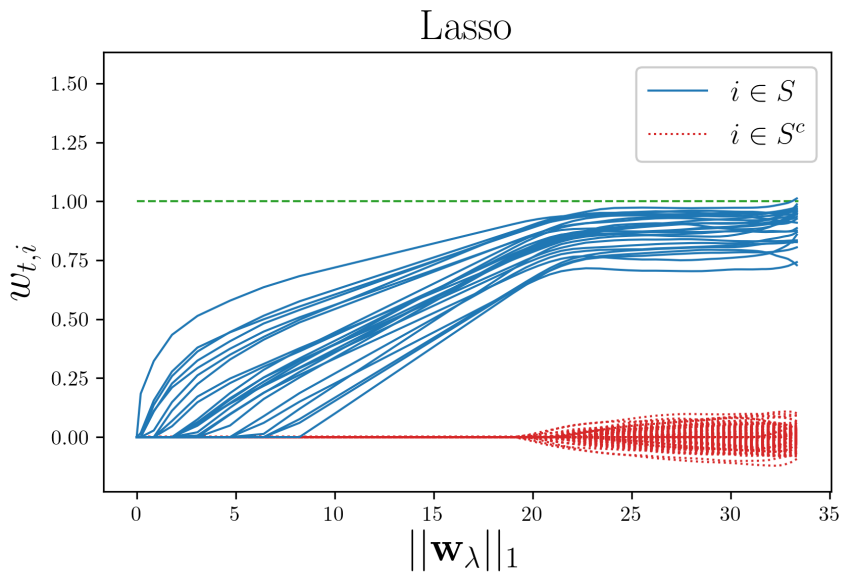
Similar to Lasso: Sparse iterates/solutions, minimax rates

Different than Lasso:

- Coordinates fitted one-by-one

Comparison with Lasso



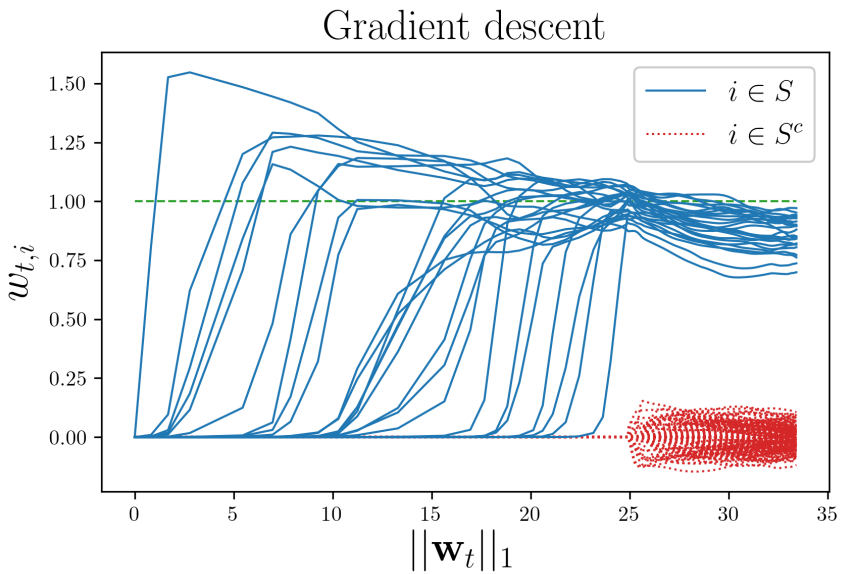


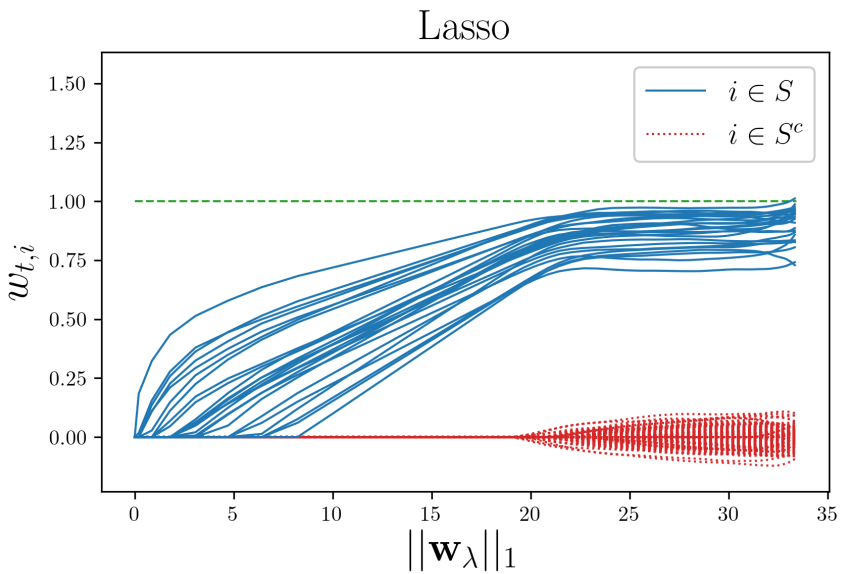
Similar to Lasso: Sparse iterates/solutions, minimax rates

Different than Lasso:

- Coordinates fitted one-by-one
- **Instance adaptivity** for high signal-to-noise (beyond minimax; no $\log d$ bias)

Comparison with Lasso





Similar to Lasso: Sparse iterates/solutions, minimax rates

Different than Lasso:

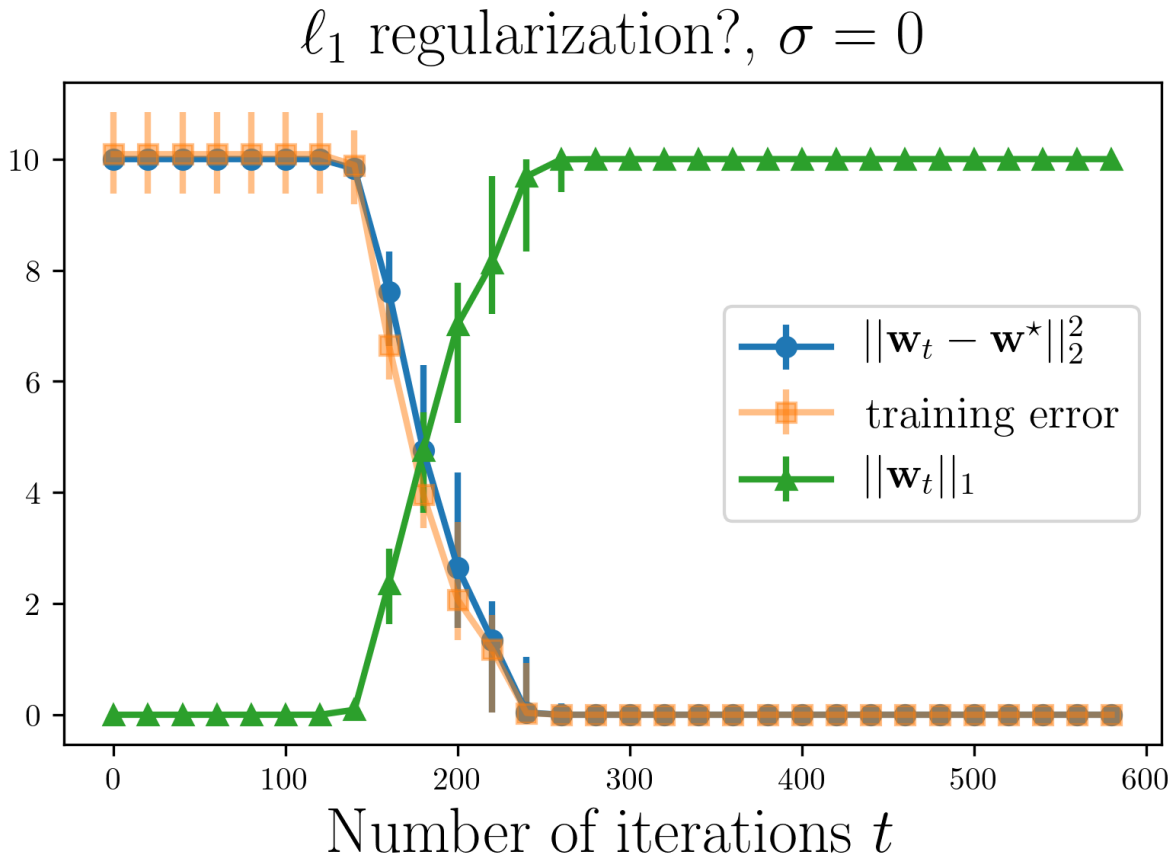
- Coordinates fitted one-by-one
- **Instance adaptivity** for high signal-to-noise (beyond minimax; no $\log d$ bias)
- **Comput. optimality via early stopping** (model selection via GD iterates)

Noiseless Setting

Empirical evidence that:

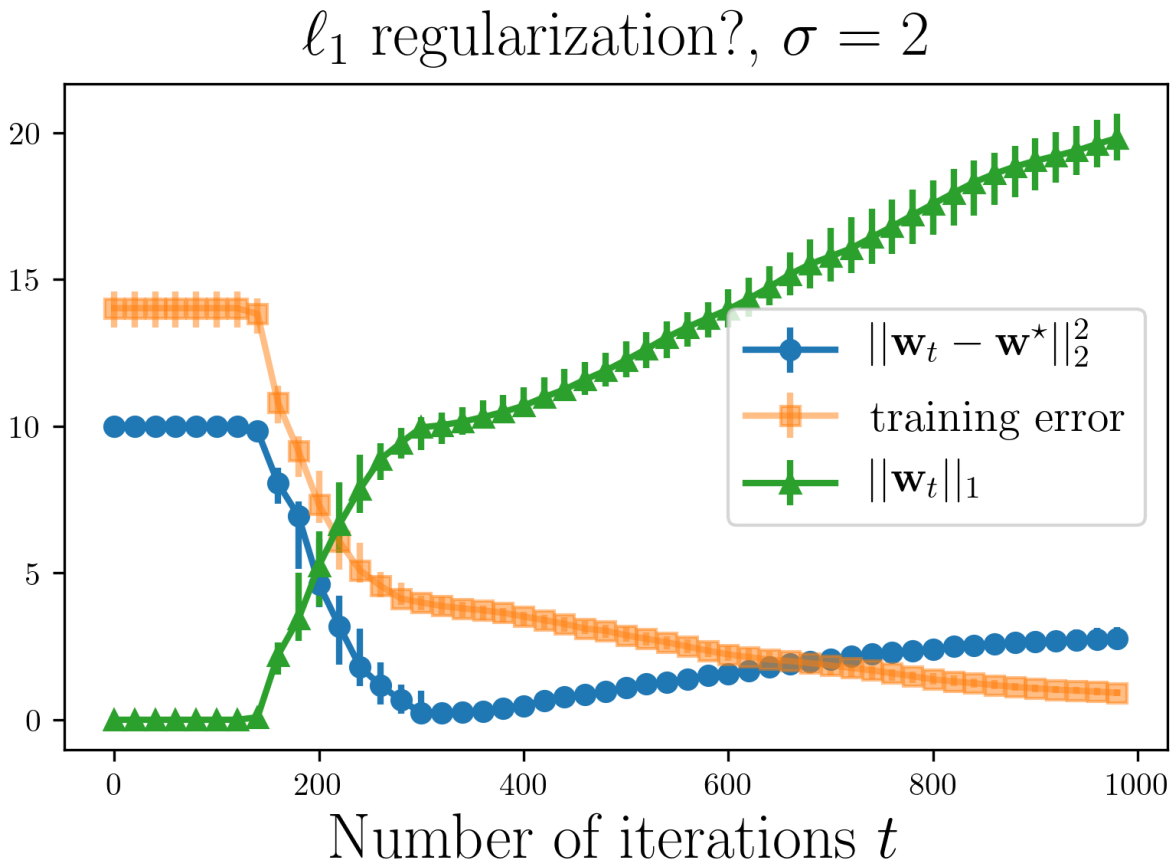
- **Monotonicity:** Training time controls complexity of solution (ℓ_1-**norm**)
- **At convergence** GD yields min. ℓ_1-**norm** solution

  (consistent with conjecture of Gradient Flow in [Gunasekar et al., 2017])



Noisy Setting

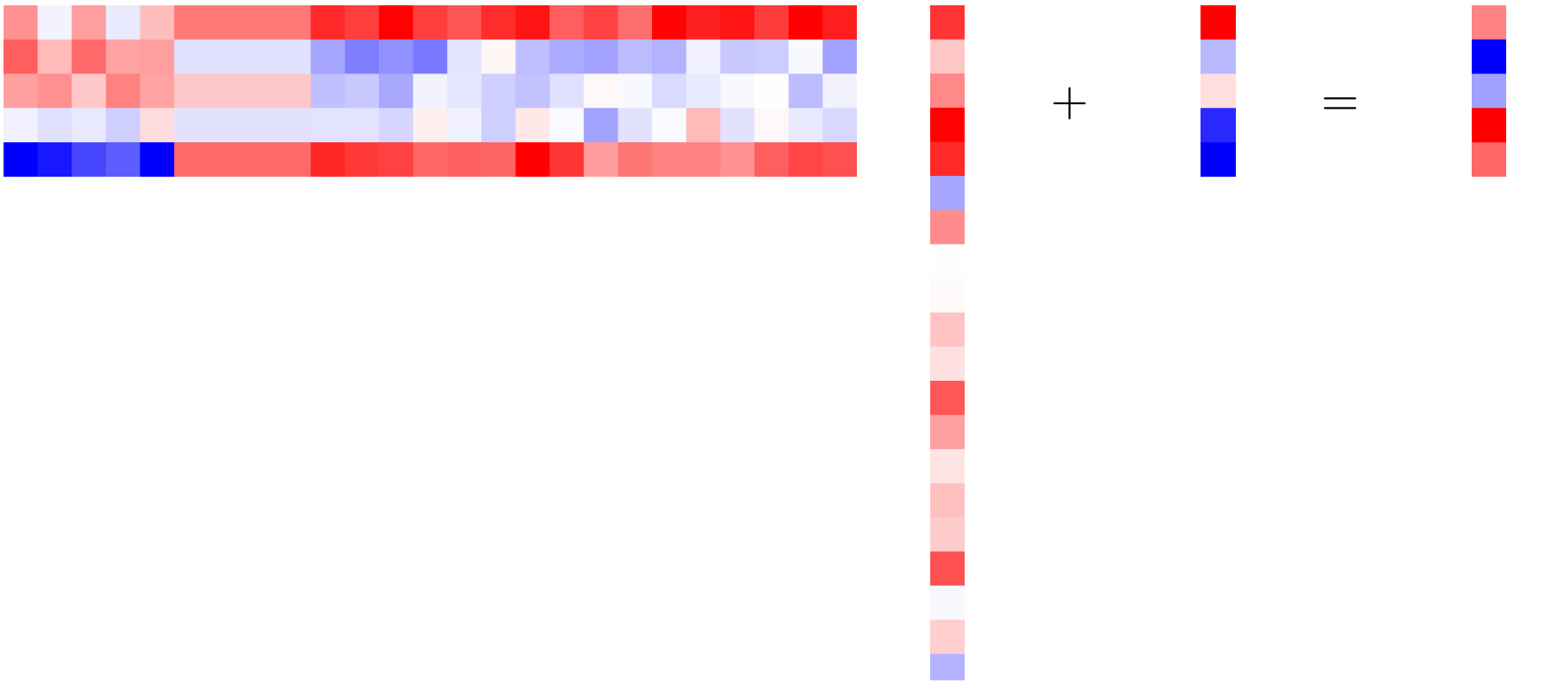
Noisy setting is fundamentally different: early stopping is needed



Training Error: $\frac{1}{n}\|\mathbf{X}w_t - y\|_2^2$

Problem Setting

$\mathbf{X} \in \mathbb{R}^{n \times d}$ $\mathbf{w}^\star \in \mathbb{R}^d$ $\xi \in \mathbb{R}^n$ $\mathbf{y} \in \mathbb{R}^n$



Problem Setting

$\mathbf{X} \in \mathbb{R}^{n \times d}$ $\mathbf{w}^\star \in \mathbb{R}^d$ $\xi \in \mathbb{R}^n$ $\mathbf{y} \in \mathbb{R}^n$

Objective

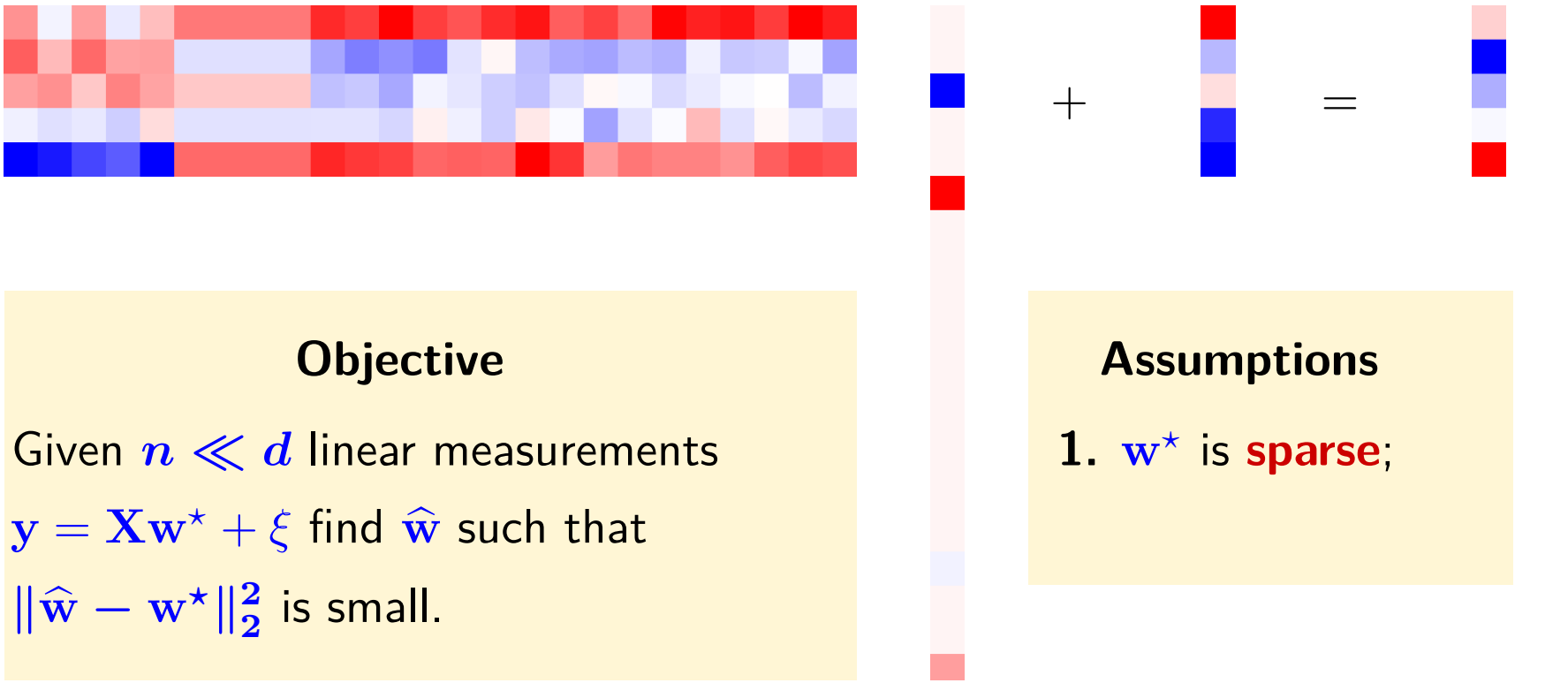
Given $n \ll d$ linear measurements

$\mathbf{y} = \mathbf{X}\mathbf{w}^\star + \xi$ find $\widehat{\mathbf{w}}$ such that

$\|\widehat{\mathbf{w}} - \mathbf{w}^\star\|_2^2$ is small.

Problem Setting

$\mathbf{X} \in \mathbb{R}^{n \times d}$ $\mathbf{w}^\star \in \mathbb{R}^d$ $\xi \in \mathbb{R}^n$ $\mathbf{y} \in \mathbb{R}^n$



Objective

Given $n \ll d$ linear measurements

$\mathbf{y} = \mathbf{X}\mathbf{w}^\star + \xi$ find $\widehat{\mathbf{w}}$ such that

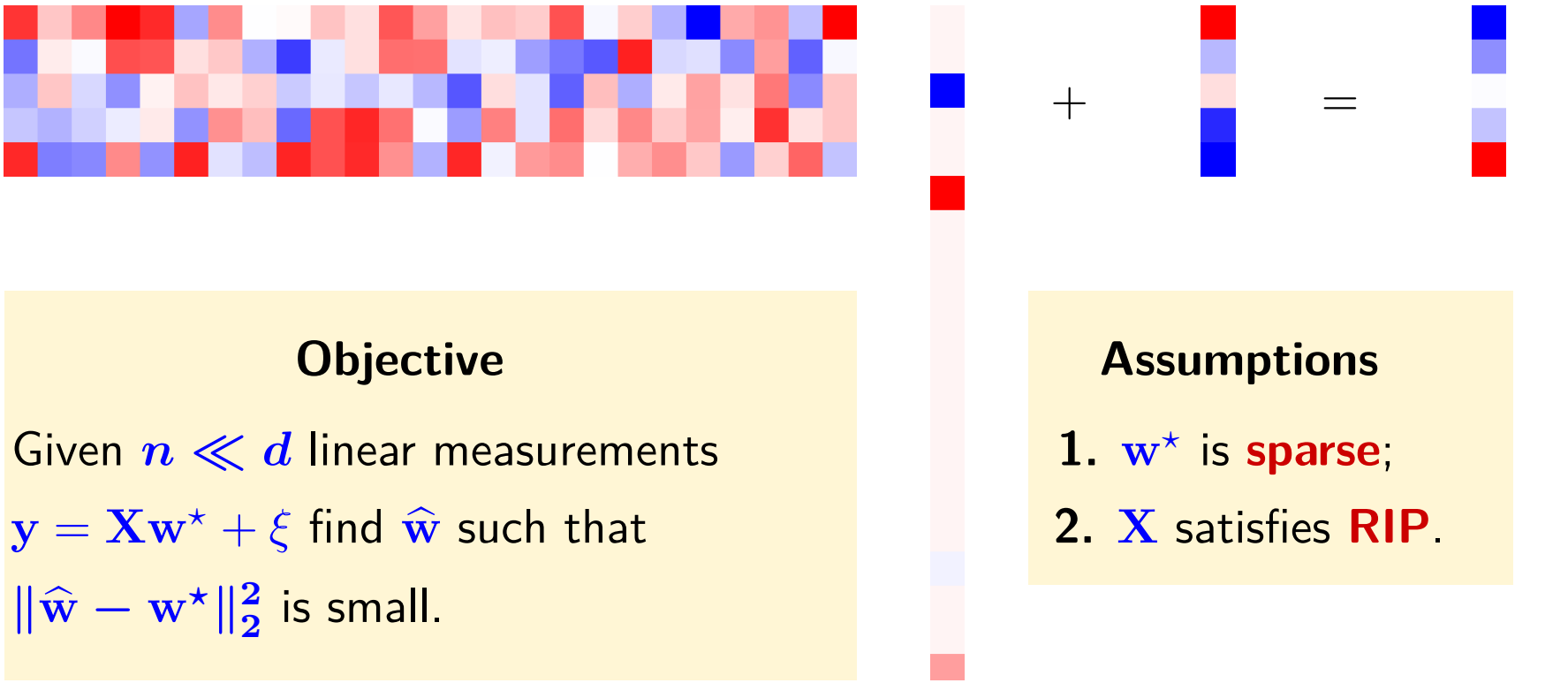
$\|\widehat{\mathbf{w}} - \mathbf{w}^\star\|_2^2$ is small.

Assumptions

1. $\mathbf{w}^\star$ is **sparse**;

Problem Setting

$\mathbf{X} \in \mathbb{R}^{n \times d}$ $\mathbf{w}^\star \in \mathbb{R}^d$ $\xi \in \mathbb{R}^n$ $\mathbf{y} \in \mathbb{R}^n$



Objective

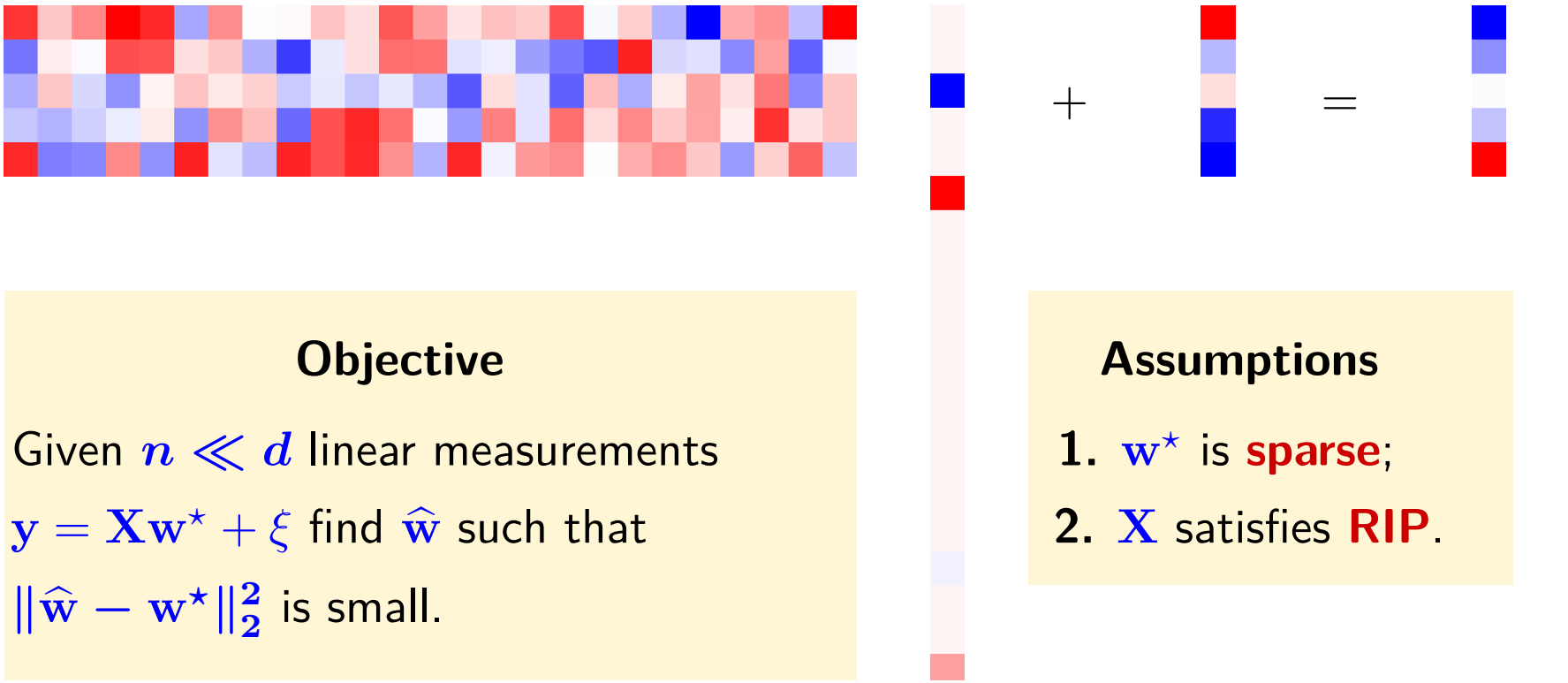
Given $n \ll d$ linear measurements $\mathbf{y} = \mathbf{X}\mathbf{w}^\star + \xi$ find $\widehat{\mathbf{w}}$ such that $\|\widehat{\mathbf{w}} - \mathbf{w}^\star\|_2^2$ is small.

Assumptions

1. $\mathbf{w}^\star$ is **sparse**;
2. $\mathbf{X}$ satisfies **RIP**.

Problem Setting

$\mathbf{X} \in \mathbb{R}^{n \times d}$ $\mathbf{w}^\star \in \mathbb{R}^d$ $\xi \in \mathbb{R}^n$ $\mathbf{y} \in \mathbb{R}^n$



Objective

Given $n \ll d$ linear measurements $\mathbf{y} = \mathbf{X}\mathbf{w}^\star + \xi$ find $\widehat{\mathbf{w}}$ such that $\|\widehat{\mathbf{w}} - \mathbf{w}^\star\|_2^2$ is small.

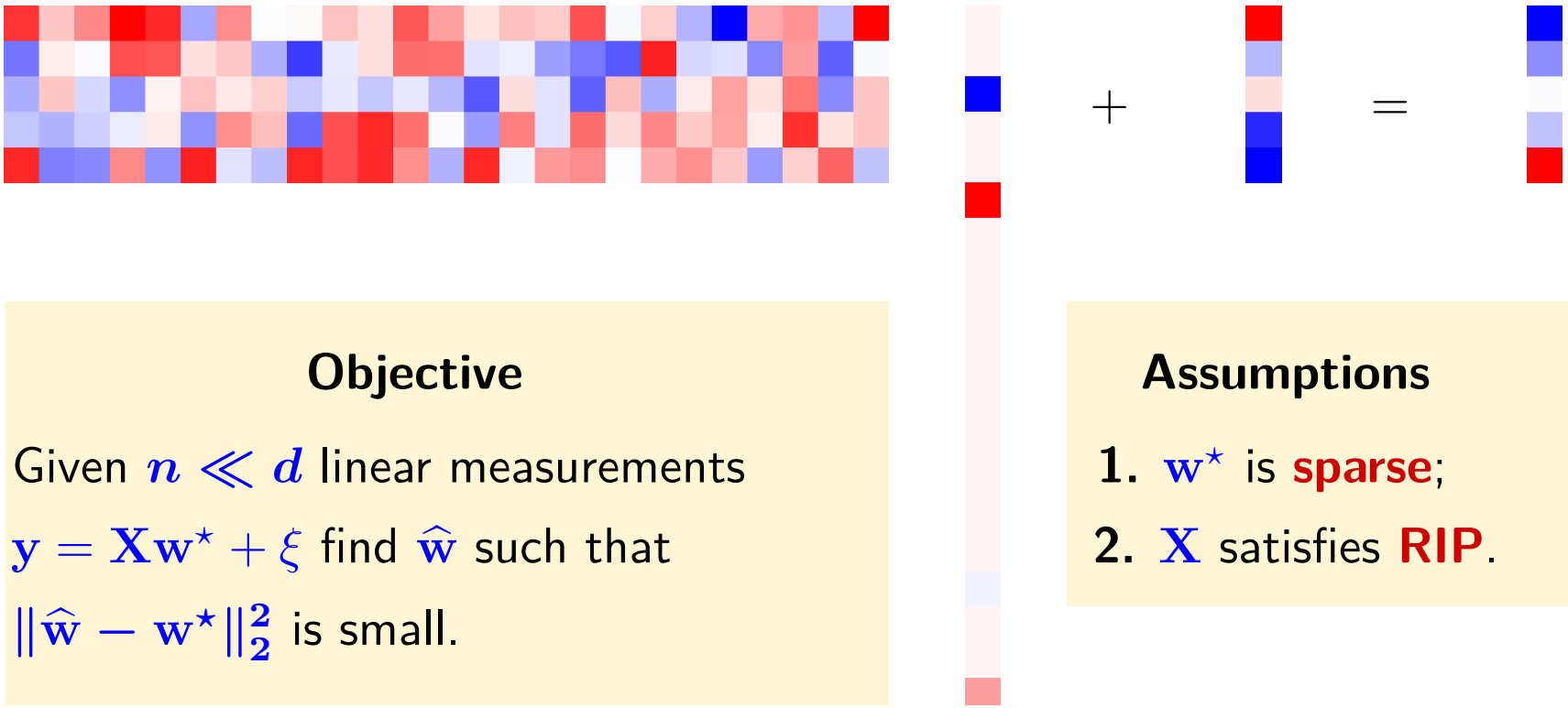
Assumptions

1. $\mathbf{w}^\star$ is **sparse**;
2. $\mathbf{X}$ satisfies **RIP**.

1. Assume $\|w^\star\|_0 = k$

Problem Setting

$\mathbf{X} \in \mathbb{R}^{n \times d}$ $\mathbf{w}^\star \in \mathbb{R}^d$ $\xi \in \mathbb{R}^n$ $\mathbf{y} \in \mathbb{R}^n$



Objective

Given $n \ll d$ linear measurements

$\mathbf{y} = \mathbf{X}\mathbf{w}^\star + \xi$ find $\widehat{\mathbf{w}}$ such that

$\|\widehat{\mathbf{w}} - \mathbf{w}^\star\|_2^2$ is small.

Assumptions

1. $\mathbf{w}^\star$ is **sparse**;
2. $\mathbf{X}$ satisfies **RIP**.

1. Assume $\|w^\star\|_0 = k$
2. Assume $\mathbf{X}/\sqrt{n}$ satisfy **RIP** with $\delta = \widetilde{O}(1/\sqrt{k})$, namely,

$$(1-\delta)\,\|\mathbf{w}\|_2^2 \le \left\|\mathbf{X}\mathbf{w}/\sqrt{n}\right\|_2^2 \le (1+\delta)\,\|\mathbf{w}\|_2^2 \qquad \text{for any } (k+1)\text{-sparse } \mathbf{w} \in \mathbb{R}^d$$

Main Theorem

- Define $w^\star_{\max} = \max_i |w^\star_i|$ and $w^\star_{\min} = \min_{i:w^\star_i \neq 0} |w^\star_i|$

Main Theorem

- Define $w^\star_{\max} = \max_i |w^\star_i|$ and $w^\star_{\min} = \min_{i:w^\star_i \neq 0} |w^\star_i|$
- Pick any $\varepsilon \in (0, 1)$

Main Theorem

- Define $w^\star_{\max} = \max_i |w^\star_i|$ and $w^\star_{\min} = \min_{i:w^\star_i \neq 0} |w^\star_i|$
- Pick any $\varepsilon \in (0, 1)$
- Set initialization size $0 < \alpha < \frac{\varepsilon^2}{(2d+1)^2 \vee (\boldsymbol{w}^\star_{\max})^2}$ **(poly. in param.)**

Main Theorem

- Define $w^\star_{\max} = \max_i |w^\star_i|$ and $w^\star_{\min} = \min_{i:w^\star_i \neq 0} |w^\star_i|$
- Pick any $\varepsilon \in (0, 1)$
- Set initialization size $0 < \alpha < \frac{\varepsilon^2}{(2d+1)^2 \vee (\boldsymbol{w^\star_{\max}})^2}$ **(poly. in param.)**
- Set the learning rate $\eta \leq \frac{1}{20 \boldsymbol{w^\star_{\max}}}$ **(to prevent explosion)**

Main Theorem

- Define $w^\star_{\max} = \max_i |w^\star_i|$ and $w^\star_{\min} = \min_{i:w^\star_i \neq 0} |w^\star_i|$
- Pick any $\varepsilon \in (0, 1)$
- Set initialization size $0 < \alpha < \frac{\varepsilon^2}{(2d+1)^2 \vee (\boldsymbol{w^\star_{\max}})^2}$ **(poly. in param.)**
- Set the learning rate $\eta \leq \frac{1}{20 \boldsymbol{w^\star_{\max}}}$ **(to prevent explosion)**

Lemma: $\boldsymbol{w^\star_{\max}}$ can be estimated up to factor $\boldsymbol{2}$ with cost $\boldsymbol{nd}$

Main Theorem

- Define $w^\star_{\max} = \max_i |w^\star_i|$ and $w^\star_{\min} = \min_{i:w^\star_i \neq 0} |w^\star_i|$
- Pick any $\varepsilon \in (0,1)$
- Set initialization size $0 < \alpha < \frac{\varepsilon^2}{(2d+1)^2 \vee (\boldsymbol{w^\star_{\max}})^2}$ **(poly. in param.)**
- Set the learning rate $\eta \leq \frac{1}{20 \boldsymbol{w^\star_{\max}}}$ **(to prevent explosion)**

Lemma: $\boldsymbol{w^\star_{\max}}$ can be estimated up to factor 2 with cost $\boldsymbol{nd}$

Theorem (Vaskevicius, Kanade, Rebeschini 2019)

After

$$t^\star = O\left(\frac{w^\star_{\max}}{w^\star_{\min} \vee \left\| \frac{1}{n} \mathbf{X}^\mathsf{T} \xi \right\|_\infty \vee \varepsilon} \cdot \frac{1}{\eta w^\star_{\max}} \cdot \log \frac{1}{\alpha} \right)$$

iterations,

Main Theorem

- Define $w^\star_{\max} = \max_i |w^\star_i|$ and $w^\star_{\min} = \min_{i:w^\star_i \neq 0} |w^\star_i|$
- Pick any $\varepsilon \in (0,1)$
- Set initialization size $0 < \alpha < \frac{\varepsilon^2}{(2d+1)^2 \vee (\boldsymbol{w^\star_{\max}})^2}$ **(poly. in param.)**
- Set the learning rate $\eta \leq \frac{1}{20 \boldsymbol{w^\star_{\max}}}$ **(to prevent explosion)**

Lemma: $\boldsymbol{w^\star_{\max}}$ can be estimated up to factor 2 with cost $\boldsymbol{nd}$

Theorem (Vaskevicius, Kanade, Rebeschini 2019)

After

$$t^\star = O\left(\frac{w^\star_{\max}}{w^\star_{\min} \vee \left\| \frac{1}{n} \mathbf{X}^\mathsf{T} \xi \right\|_\infty \vee \varepsilon} \cdot \frac{1}{\eta w^\star_{\max}} \cdot \log \frac{1}{\alpha} \right)$$

iterations, the GD iterate $\mathbf{w}_{t^\star}$ *satisfies*

$$\|\mathbf{w}_{t^\star} \odot \mathbf{1}_{S^c}\|_\infty \lesssim \sqrt{\alpha} < \frac{\varepsilon}{d}$$

Main Theorem

- Define $w^\star_{\max} = \max_i |w^\star_i|$ and $w^\star_{\min} = \min_{i:w^\star_i \neq 0} |w^\star_i|$
- Pick any $\varepsilon \in (0,1)$
- Set initialization size $0 < \alpha < \frac{\varepsilon^2}{(2d+1)^2 \vee (\boldsymbol{w^\star_{\max}})^2}$ **(poly. in param.)**
- Set the learning rate $\eta \leq \frac{1}{20 \boldsymbol{w^\star_{\max}}}$ **(to prevent explosion)**

Lemma: $\boldsymbol{w^\star_{\max}}$ can be estimated up to factor 2 with cost $\boldsymbol{nd}$

Theorem (Vaskevicius, Kanade, Rebeschini 2019)

After

$$t^\star = O\left(\frac{w^\star_{\max}}{w^\star_{\min} \vee \left\| \frac{1}{n} \mathbf{X}^\mathsf{T} \xi \right\|_\infty \vee \varepsilon} \cdot \frac{1}{\eta w^\star_{\max}} \cdot \log \frac{1}{\alpha} \right)$$

iterations, the GD iterate $\mathbf{w}_{t^\star}$ *satisfies*

$$\|\mathbf{w}_{t^\star} \odot \mathbf{1}_{S^c}\|_\infty \lesssim \sqrt{\alpha} < \frac{\varepsilon}{d}$$

$$\|\mathbf{w}_{t^\star} \odot \mathbf{1}_S - \mathbf{w}^\star\|_\infty \lesssim \begin{cases} \left\| \frac{1}{n} \mathbf{X}^\mathsf{T} \xi \right\|_\infty \vee \varepsilon & \text{always holds} \\ \end{cases}$$

Main Theorem

- Define $w^\star_{\max} = \max_i |w^\star_i|$ and $w^\star_{\min} = \min_{i:w^\star_i \neq 0} |w^\star_i|$
- Pick any $\varepsilon \in (0,1)$
- Set initialization size $0 < \alpha < \frac{\varepsilon^2}{(2d+1)^2 \vee (\boldsymbol{w^\star_{\max}})^2}$ **(poly. in param.)**
- Set the learning rate $\eta \leq \frac{1}{20 \boldsymbol{w^\star_{\max}}}$ **(to prevent explosion)**

Lemma: $\boldsymbol{w^\star_{\max}}$ can be estimated up to factor 2 with cost $\boldsymbol{nd}$

Theorem (Vaskevicius, Kanade, Rebeschini 2019)

After

$$t^\star = O\left(\frac{w^\star_{\max}}{w^\star_{\min} \vee \left\| \frac{1}{n} \mathbf{X}^\mathsf{T} \xi \right\|_\infty \vee \varepsilon} \cdot \frac{1}{\eta w^\star_{\max}} \cdot \log \frac{1}{\alpha} \right)$$

iterations, the GD iterate $\mathbf{w}_{t^\star}$ *satisfies*

$$\left\| \mathbf{w}_{t^\star} \odot \mathbf{1}_{S^c} \right\|_\infty \lesssim \sqrt{\alpha} < \frac{\varepsilon}{d}$$

$$\left\| \mathbf{w}_{t^\star} \odot \mathbf{1}_S - \mathbf{w}^\star \right\|_\infty \lesssim \begin{cases} \left\| \frac{1}{n} \mathbf{X}^\mathsf{T} \xi \right\|_\infty \vee \varepsilon & \textit{always holds} \\ \left\| \frac{1}{n} \mathbf{X}^\mathsf{T} \xi \odot \mathbf{1}_S \right\|_\infty \vee \varepsilon & \textit{if } w^\star_{\min} \gtrsim \left\| \frac{1}{n} \mathbf{X}^\mathsf{T} \xi \right\|_\infty \end{cases}$$

Statistical Rates

Corollary (Noiseless Recovery)

Let $\xi = 0$. *Then GD yields* $\boxed{\|\mathbf{w}_{t^\star} - \mathbf{w}^\star\|_2^2 \lesssim k\varepsilon^2}$

Statistical Rates

Corollary (Noiseless Recovery)

Let $\xi = 0$. *Then GD yields* $\boxed{\|\mathbf{w}_{t^\star} - \mathbf{w}^\star\|_2^2 \lesssim k\varepsilon^2}$

Corollary (Minimax Rates in the Noisy Setting)

Let ξ *have i.i.d.* σ^2*-sub-Gaussian entries.*

Statistical Rates

Corollary (Noiseless Recovery)

Let $\xi = 0$. *Then GD yields* $\boxed{\|\mathbf{w}_{t^\star} - \mathbf{w}^\star\|_2^2 \lesssim k\varepsilon^2}$

Corollary (Minimax Rates in the Noisy Setting)

Let ξ *have i.i.d.* σ^2*-sub-Gaussian entries. Let* $\varepsilon = 4\sqrt{\frac{\sigma^2 \log(2d)}{n}}$.

Statistical Rates

Corollary (Noiseless Recovery)

Let $\xi = 0$. *Then GD yields* $\boxed{\|\mathbf{w}_{t^\star} - \mathbf{w}^\star\|_2^2 \lesssim k\varepsilon^2}$

Corollary (Minimax Rates in the Noisy Setting)

Let ξ *have i.i.d.* σ^2*-sub-Gaussian entries. Let* $\varepsilon = 4\sqrt{\frac{\sigma^2 \log(2d)}{n}}$. *Then,*

$$t^\star = O\left(\frac{w^\star_{\max}\sqrt{n}}{\sigma\sqrt{\log d}} \cdot \log\frac{1}{\alpha}\right) = \widetilde{\boldsymbol{O}}\left(\frac{\boldsymbol{w^\star_{\max}\sqrt{n}}}{\boldsymbol{\sigma}}\right)$$

Statistical Rates

Corollary (Noiseless Recovery)

Let $\xi = 0$. *Then GD yields* $\boxed{\|\mathbf{w}_{t^\star} - \mathbf{w}^\star\|_2^2 \lesssim k\varepsilon^2}$

Corollary (Minimax Rates in the Noisy Setting)

Let ξ *have i.i.d.* σ^2*-sub-Gaussian entries. Let* $\varepsilon = 4\sqrt{\frac{\sigma^2 \log(2d)}{n}}$. *Then,*

$$t^\star = O\left(\frac{w^\star_{\max}\sqrt{n}}{\sigma\sqrt{\log d}} \cdot \log\frac{1}{\alpha}\right) = \widetilde{\boldsymbol{O}}\left(\frac{\boldsymbol{w^\star_{\max}\sqrt{n}}}{\boldsymbol{\sigma}}\right)$$

and, with probability at least $1 - 1/(8d^3)$, *GD yields*

$$\boxed{\|\mathbf{w}_{t^\star} - \mathbf{w}^\star\|_2^2 \lesssim \frac{k\sigma^2 \log d}{n}}$$

Statistical Rates

Corollary (Noiseless Recovery)

Let $\xi = 0$. *Then GD yields* $\boxed{\|\mathbf{w}_{t^\star} - \mathbf{w}^\star\|_2^2 \lesssim k\varepsilon^2}$

Corollary (Minimax Rates in the Noisy Setting)

Let ξ *have i.i.d.* σ^2*-sub-Gaussian entries. Let* $\varepsilon = 4\sqrt{\frac{\sigma^2 \log(2d)}{n}}$. *Then,*

$$t^\star = O\left(\frac{w^\star_{\max}\sqrt{n}}{\sigma\sqrt{\log d}} \cdot \log\frac{1}{\alpha}\right) = \widetilde{\boldsymbol{O}}\left(\frac{\boldsymbol{w^\star_{\max}\sqrt{n}}}{\boldsymbol{\sigma}}\right)$$

and, with probability at least $1 - 1/(8d^3)$, *GD yields*

$$\boxed{\|\mathbf{w}_{t^\star} - \mathbf{w}^\star\|_2^2 \lesssim \frac{k\sigma^2 \log d}{n}}$$

- ε controls the size of the smallest coordinates of $\mathbf{w}^\star$ that GD can recover
- To achieve minimax rates, GD has to recover everything as big as $\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty$

Proof Idea

- Let $\mathbf{w}_t = \mathbf{u}_t \odot \mathbf{u}_t$

Proof Idea

- Let $\mathbf{w}_t = \mathbf{u}_t \odot \mathbf{u}_t$
- By **RIP** $\mathbf{X}^\mathsf{T}\mathbf{X}/n \approx \mathbf{I}$ for sparse vectors. Assume $\mathbf{X}^\mathsf{T}\mathbf{X}/n = \mathbf{I}$

Proof Idea

- Let $\mathbf{w}_t = \mathbf{u}_t \odot \mathbf{u}_t$
- By **RIP** $\mathbf{X}^\mathsf{T}\mathbf{X}/n \approx \mathbf{I}$ for sparse vectors. Assume $\mathbf{X}^\mathsf{T}\mathbf{X}/n = \mathbf{I}$
- Each coordinate evolves independently of the others as:

$$\mathbf{w}_{t+1} = \mathbf{w}_t \odot (\mathbf{1} - 4\eta(\mathbf{w}_t - \mathbf{w}^\star + \mathbf{X}^\mathsf{T}\xi/n))^2$$

Proof Idea

- Let $\mathbf{w}_t = \mathbf{u}_t \odot \mathbf{u}_t$
- By **RIP** $\mathbf{X}^\mathsf{T}\mathbf{X}/n \approx \mathbf{I}$ for sparse vectors. Assume $\mathbf{X}^\mathsf{T}\mathbf{X}/n = \mathbf{I}$
- Each coordinate evolves independently of the others as:

$$\mathbf{w}_{t+1} = \mathbf{w}_t \odot (\mathbf{1} - 4\eta(\mathbf{w}_t - \mathbf{w}^\star + \mathbf{X}^\mathsf{T}\xi/n))^2$$

- Hence we only need to understand one-dimensional sequences

$$x_{t+1} = x_t(1 - 4\eta(x_t - x^\star))^2 \qquad \text{with } x_0 = \alpha^2$$

Proof Idea

- Let $\mathbf{w}_t = \mathbf{u}_t \odot \mathbf{u}_t$
- By **RIP** $\mathbf{X}^\mathsf{T}\mathbf{X}/n \approx \mathbf{I}$ for sparse vectors. Assume $\mathbf{X}^\mathsf{T}\mathbf{X}/n = \mathbf{I}$
- Each coordinate evolves independently of the others as:

$$\mathbf{w}_{t+1} = \mathbf{w}_t \odot (\mathbf{1} - 4\eta(\mathbf{w}_t - \mathbf{w}^\star + \mathbf{X}^\mathsf{T}\xi/n))^2$$

- Hence we only need to understand one-dimensional sequences

$$x_{t+1} = x_t(1 - 4\eta(x_t - x^\star))^2 \qquad \text{with } x_0 = \alpha^2$$

- **Prop.** Let $0 < \alpha^2 \le \frac{x^\star}{2}$, $\eta \lesssim 1/x^\star$. Given $\varepsilon > 0$ and $t \gtrsim \frac{1}{\eta x^\star} \log \frac{(x^\star)^2}{\alpha^2 \varepsilon}$:

$$x^\star - \varepsilon \le x_t \le x^\star$$

Proof Idea

- Let $\mathbf{w}_t = \mathbf{u}_t \odot \mathbf{u}_t$
- By **RIP** $\mathbf{X}^\mathsf{T}\mathbf{X}/n \approx \mathbf{I}$ for sparse vectors. Assume $\mathbf{X}^\mathsf{T}\mathbf{X}/n = \mathbf{I}$
- Each coordinate evolves independently of the others as:

$$\mathbf{w}_{t+1} = \mathbf{w}_t \odot (\mathbf{1} - 4\eta(\mathbf{w}_t - \mathbf{w}^\star + \mathbf{X}^\mathsf{T}\xi/n))^2$$

- Hence we only need to understand one-dimensional sequences

$$x_{t+1} = x_t(1 - 4\eta(x_t - x^\star))^2 \qquad \text{with } x_0 = \alpha^2$$

- **Prop.** Let $0 < \alpha^2 \le \frac{x^\star}{2}$, $\eta \lesssim 1/x^\star$. Given $\varepsilon > 0$ and $t \gtrsim \frac{1}{\eta x^\star} \log \frac{(x^\star)^2}{\alpha^2 \varepsilon}$:

$$x^\star - \varepsilon \le x_t \le x^\star$$

- The i-th coord. converges in $O(\frac{1}{\eta\left|w_i^\star + (\mathbf{X}^\mathsf{T}\xi)_i/n\right|} \log \frac{\left|w_i^\star + (\mathbf{X}^\mathsf{T}\xi)_i/n\right|^2}{\alpha^2\varepsilon})$ iterations

Proof Idea

- Let $\mathbf{w}_t = \mathbf{u}_t \odot \mathbf{u}_t$
- By **RIP** $\mathbf{X}^\mathsf{T}\mathbf{X}/n \approx \mathbf{I}$ for sparse vectors. Assume $\mathbf{X}^\mathsf{T}\mathbf{X}/n = \mathbf{I}$
- Each coordinate evolves independently of the others as:

$$\mathbf{w}_{t+1} = \mathbf{w}_t \odot (\mathbf{1} - 4\eta(\mathbf{w}_t - \mathbf{w}^\star + \mathbf{X}^\mathsf{T}\xi/n))^2$$

- Hence we only need to understand one-dimensional sequences

$$x_{t+1} = x_t(1 - 4\eta(x_t - x^\star))^2 \qquad \text{with } x_0 = \alpha^2$$

- **Prop.** Let $0 < \alpha^2 \le \frac{x^\star}{2}$, $\eta \lesssim 1/x^\star$. Given $\varepsilon > 0$ and $t \gtrsim \frac{1}{\eta x^\star} \log \frac{(x^\star)^2}{\alpha^2 \varepsilon}$:

$$x^\star - \varepsilon \le x_t \le x^\star$$

- The i-th coord. converges in $O(\frac{1}{\eta|w_i^\star + (\mathbf{X}^\mathsf{T}\xi)_i/n|} \log \frac{|w_i^\star + (\mathbf{X}^\mathsf{T}\xi)_i/n|^2}{\alpha^2 \varepsilon})$ iterations
- Hence, all coordinates converge **exponentially fast at different rates**

Proof Idea

- Sequence fitting **signal**: $x_{t+1} = x_t(1 - 4\eta(x_t - x^\star))^2$

Proof Idea

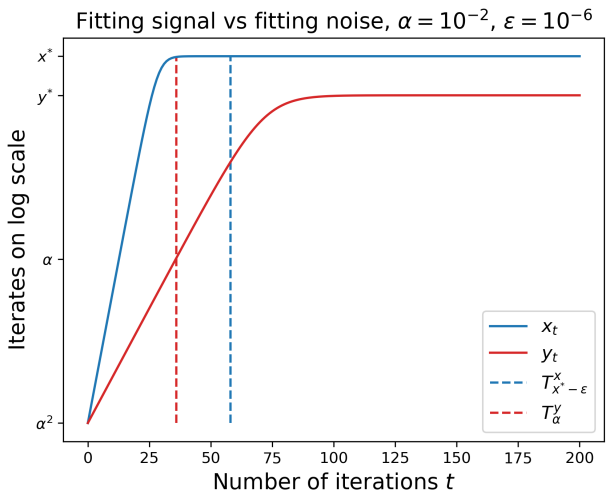
- Sequence fitting **signal**: $x_{t+1} = x_t(1 - 4\eta(x_t - x^\star))^2$
- Sequence fitting **noise**: $y_{t+1} = y_t(1 - 4\eta(y_t - y^\star))^2$ with $y^\star = \left\| \frac{1}{n}\mathbf{X}^\mathsf{T}\xi \right\|_\infty$

Proof Idea

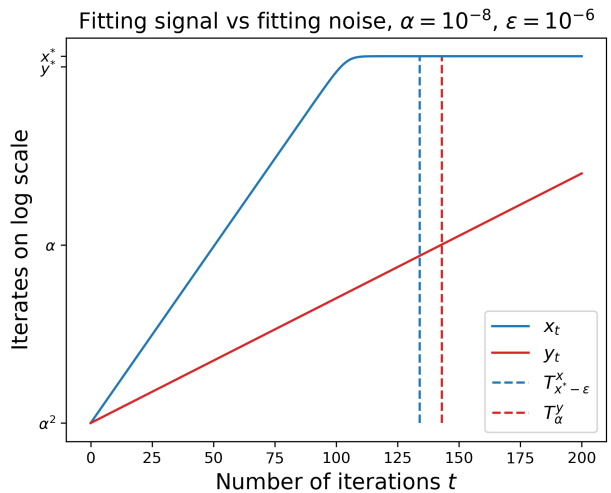
- Sequence fitting **signal**: $x_{t+1} = x_t(1 - 4\eta(x_t - x^\star))^2$
- Sequence fitting **noise**: $y_{t+1} = y_t(1 - 4\eta(y_t - y^\star))^2$ with $y^\star = \left\| \frac{1}{n}\mathbf{X}^\mathsf{T}\xi \right\|_\infty$
- **Goal:** fit the sequence $(x_t)_{t\geq 0}$ to $x^\star$ within ε error **before** $(y_t)_{t\geq 0}$ exceeds α

Proof Idea

- Sequence fitting **signal**: $x_{t+1} = x_t(1 - 4\eta(x_t - x^\star))^2$
- Sequence fitting **noise**: $y_{t+1} = y_t(1 - 4\eta(y_t - y^\star))^2$ with $y^\star = \left\| \frac{1}{n}\mathbf{X}^\mathsf{T}\xi \right\|_\infty$
- **Goal:** fit the sequence $(x_t)_{t\geq 0}$ to $x^\star$ within ε error **before** $(y_t)_{t\geq 0}$ exceeds α
- If $x^\star \gtrsim y^\star$, then for any $\varepsilon > 0$ there is α **small enough** so that $T^x_{x^\star - \varepsilon} \leq T^y_\alpha$:



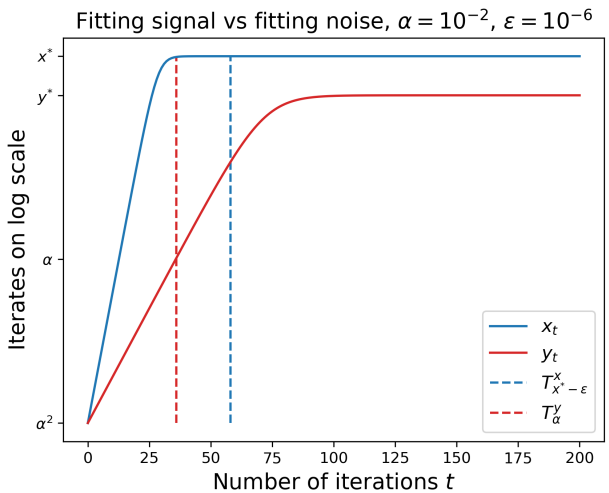
(a) α **too large**



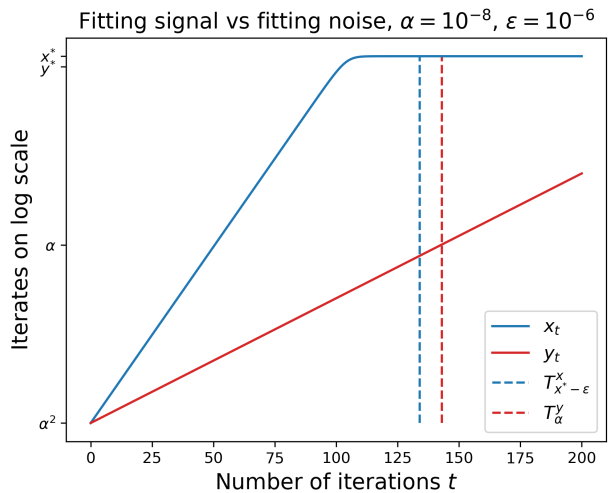
(b) α **small enough:** signal fitted before noise goes above α

Proof Idea

- Sequence fitting **signal**: $x_{t+1} = x_t(1 - 4\eta(x_t - x^\star))^2$
- Sequence fitting **noise**: $y_{t+1} = y_t(1 - 4\eta(y_t - y^\star))^2$ with $y^\star = \left\| \frac{1}{n}\mathbf{X}^\mathsf{T}\xi \right\|_\infty$
- **Goal:** fit the sequence $(x_t)_{t\geq 0}$ to $x^\star$ within ε error **before** $(y_t)_{t\geq 0}$ exceeds α
- If $x^\star \gtrsim y^\star$, then for any $\varepsilon > 0$ there is α **small enough** so that $T^x_{x^\star - \varepsilon} \leq T^y_\alpha$:



(a) α **too large**



(b) α **small enough:** signal fitted before noise goes above α

BY SAME IDEA: GD fits coordinates one by one!

Constant Step Size yields $O(\sqrt{n})$ Iteration Complexity

Our theorem prescribes

$$t^\star = O\left(\frac{w^\star_{\max}\sqrt{n}}{\sigma\sqrt{\log d}} \cdot \log\frac{1}{\alpha}\right) = \widetilde{O}\left(\frac{w^\star_{\max}\sqrt{n}}{\sigma}\right)$$

which yields a total cost $\widetilde{O}(n^{3/2}d)$, **not optimal:** cost of reading data is $O(nd)$

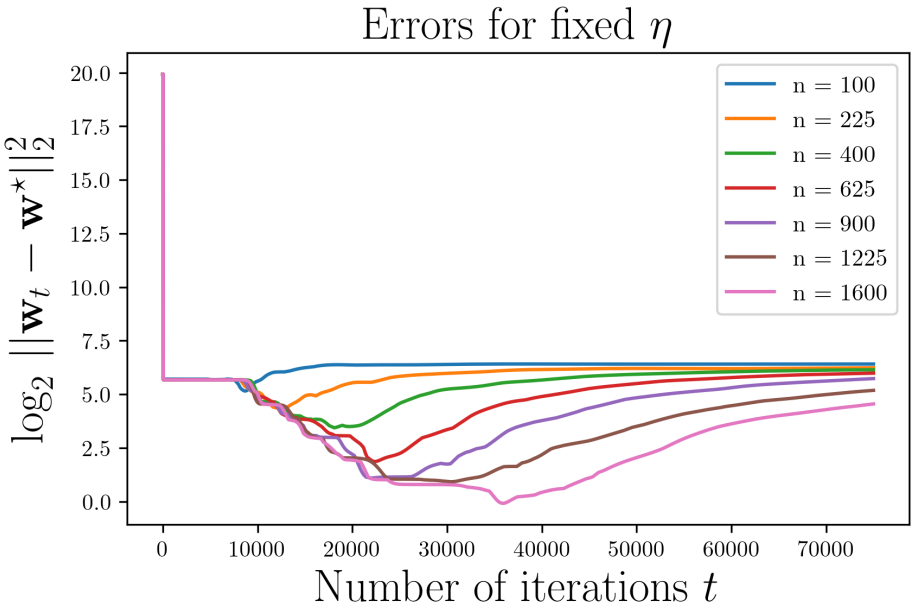


Figure: $n = 100k^2$, for $k = 1, 1.5, 2, 2.5, 3, 3.5, 4$

Q: Can speed up convergence and get computational optimality (mod $\log$ terms)?

Small Step Size Hurts Fitting Small Coordinates

Different coordinates are fitted at different rates: the smaller the later are fitted.

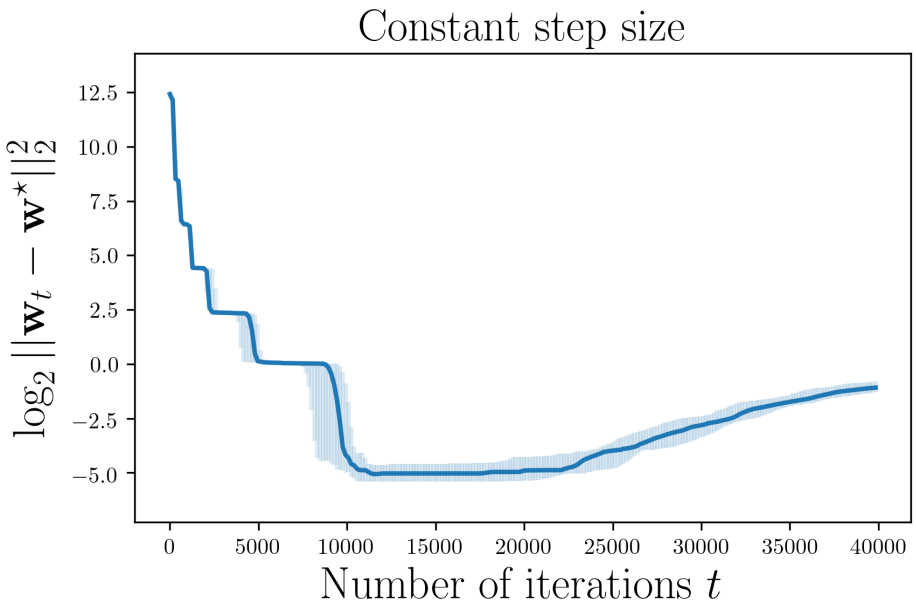


Figure: $\mathbf{w}^\star = (64, 32, 16, 8, 4, 2, 1, 0, \ldots, 0)$. Algorithm with constant step size spends approximately twice the time to fit each coordinate that the previous one

Small Step Size Hurts Fitting Small Coordinates

Different coordinates are fitted at different rates: the smaller the later are fitted.

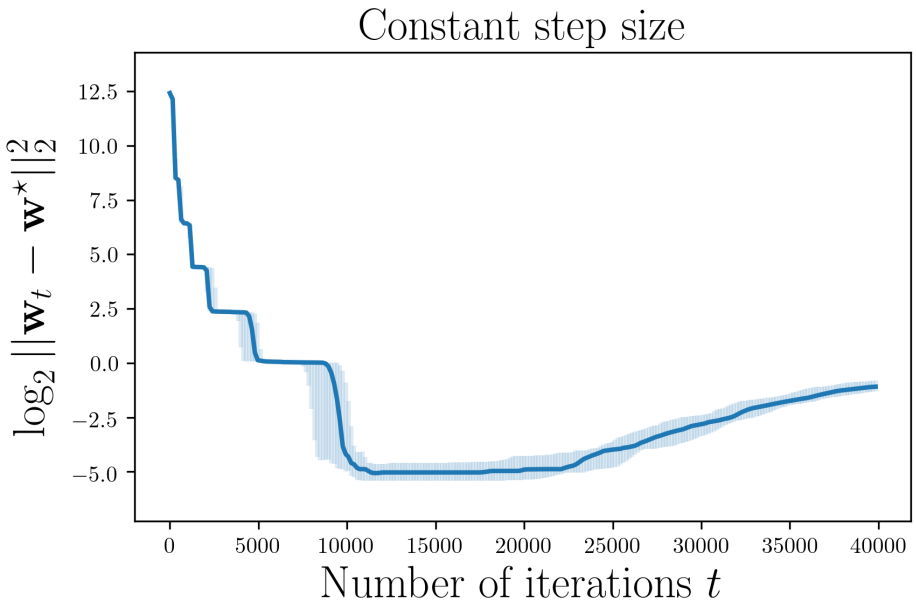


Figure: $\mathbf{w}^\star = (64, 32, 16, 8, 4, 2, 1, 0, \ldots, 0)$. Algorithm with constant step size spends approximately twice the time to fit each coordinate that the previous one

IDEA: Use different learning rates for different coordinates

- If RIP exact and no noise, then $\eta_i \sim \frac{1}{w_i^\star}$ would yield convergence in $O(\log \frac{w_i^\star}{\alpha})$
- We need refined estimates of $w_i^\star$ for each coordinate i

Increasing Step Sizes Scheme yields Comp. Optimality

Increasing Step Sizes + Early Stopping $\Rightarrow$ Computational Optimality

Increasing Step Sizes Scheme yields Comp. Optimality

Increasing Step Sizes + Early Stopping $\Rightarrow$ Computational Optimality

- Estimate $w^{\star}_{\max}$ up to factor 2 in time $O(nd)$ **(Lemma)**

Increasing Step Sizes Scheme yields Comp. Optimality

Increasing Step Sizes + Early Stopping $\Rightarrow$ Computational Optimality

- Estimate $w^\star_{\max}$ up to factor 2 in time $O(nd)$ **(Lemma)**
- For $i = 1, \ldots, d$ set $\eta_i \asymp 1/w^\star_{\max}$ and $C = 1/8$

Increasing Step Sizes Scheme yields Comp. Optimality

Increasing Step Sizes + Early Stopping $\Rightarrow$ Computational Optimality

- Estimate $w^\star_{\max}$ up to factor 2 in time $O(nd)$ **(Lemma)**
- For $i = 1, \ldots, d$ set $\eta_i \asymp 1/w^\star_{\max}$ and $C = 1/8$
- Repeat:
  1. Run gradient descent for $\Omega(\log \alpha^{-1})$ iterations

Increasing Step Sizes Scheme yields Comp. Optimality

Increasing Step Sizes + Early Stopping $\Rightarrow$ Computational Optimality

- Estimate $w^\star_{\max}$ up to factor 2 in time $O(nd)$ **(Lemma)**
- For $i = 1, \ldots, d$ set $\eta_i \asymp 1/w^\star_{\max}$ and $C = 1/8$
- Repeat:
  1. Run gradient descent for $\Omega(\log \alpha^{-1})$ iterations
  2. By this time for all i such that $|w^\star_i| > w^\star_{\max}/2$ we have $|w_{t,i}| > C w^\star_{\max}$

Increasing Step Sizes Scheme yields Comp. Optimality

Increasing Step Sizes + Early Stopping $\Rightarrow$ Computational Optimality

- Estimate $w^\star_{\max}$ up to factor 2 in time $O(nd)$ **(Lemma)**
- For $i = 1, \ldots, d$ set $\eta_i \asymp 1/w^\star_{\max}$ and $C = 1/8$
- Repeat:
  1. Run gradient descent for $\Omega(\log \alpha^{-1})$ iterations
  2. By this time for all i such that $|w^\star_i| > w^\star_{\max}/2$ we have $|w_{t,i}| > C w^\star_{\max}$
  3. For all i such that $|w_{t,i}| \leq C w^\star_{\max}$ **double the step size η_i**

Increasing Step Sizes Scheme yields Comp. Optimality

Increasing Step Sizes + Early Stopping $\Rightarrow$ Computational Optimality

- Estimate $w^\star_{\max}$ up to factor 2 in time $O(nd)$ **(Lemma)**
- For $i = 1, \ldots, d$ set $\eta_i \asymp 1/w^\star_{\max}$ and $C = 1/8$
- Repeat:
  1. Run gradient descent for $\Omega(\log \alpha^{-1})$ iterations
  2. By this time for all i such that $|w^\star_i| > w^\star_{\max}/2$ we have $|w_{t,i}| > C w^\star_{\max}$
  3. For all i such that $|w_{t,i}| \leq C w^\star_{\max}$ **double the step size η_i**
  4. Divide C by 2 and go back to step 1

Increasing Step Sizes Scheme yields Comp. Optimality

Increasing Step Sizes + Early Stopping $\Rightarrow$ Computational Optimality

- Estimate $w^\star_{\max}$ up to factor 2 in time $O(nd)$ **(Lemma)**
- For $i = 1, \dots, d$ set $\eta_i \asymp 1/w^\star_{\max}$ and $C = 1/8$
- Repeat:
  1. Run gradient descent for $\Omega(\log \alpha^{-1})$ iterations
  2. By this time for all i such that $|w^\star_i| > w^\star_{\max}/2$ we have $|w_{t,i}| > C w^\star_{\max}$
  3. For all i such that $|w_{t,i}| \leq C w^\star_{\max}$ **double the step size η_i**
  4. Divide C by 2 and go back to step 1

Theorem

Using the increasing step sizes scheme, ***all previous results hold with***

$$t^\star = O\left(\log\left(\frac{w^\star_{\max}\sqrt{n}}{\sigma\sqrt{\log d}}\right)\log\frac{1}{\alpha}\right)$$

Iteration complexity $\widetilde{O}(1)$ $\Rightarrow$ total computational complexity $\widetilde{O}(nd)$

Computational Optimality

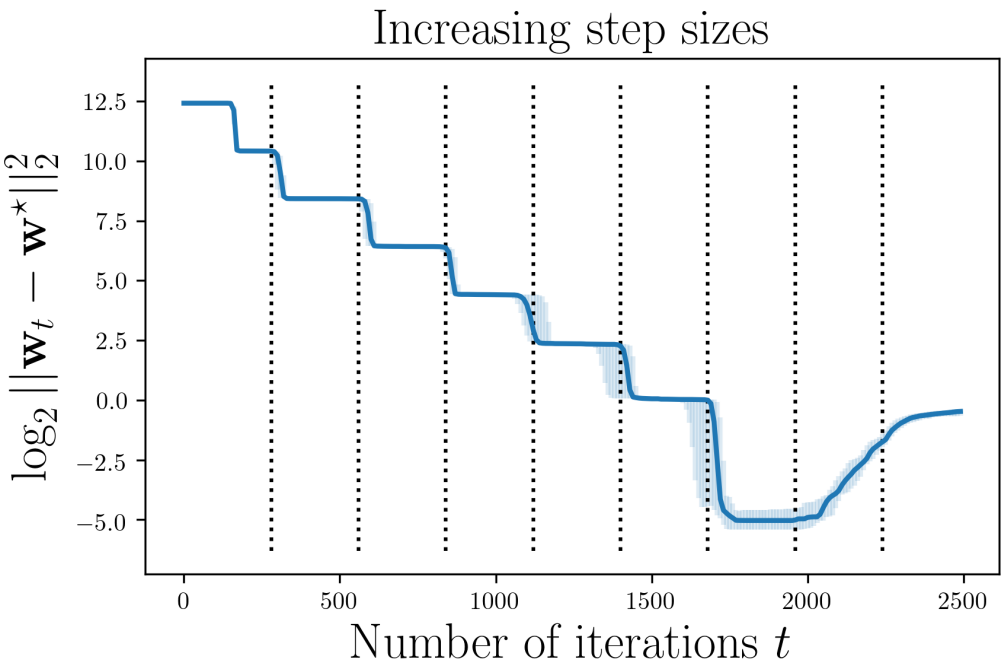
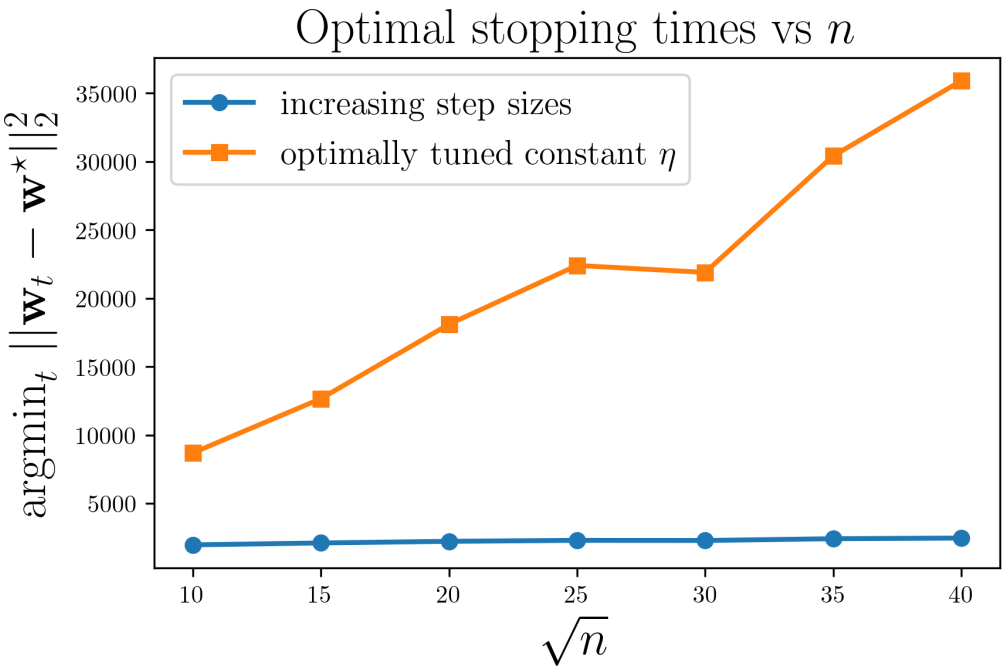


Figure: $\mathbf{w}^\star = (64, 32, 16, 8, 4, 2, 1, 0, \ldots, 0)$. Algorithm with increasing step sizes fits each coordinate at approximately the same number of iterations

Computational Optimality



Is there a more general picture?

Is there a more general picture?

- Gradient updates using Hadamard parametrization:

$$\mathbf{u}_{t+1} = \mathbf{u}_t \odot \left(\mathbb{1} - 4\eta \left(\underbrace{\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{w}_t - \mathbf{y})}_{=\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w})} \right) \right)$$

$$\mathbf{v}_{t+1} = \mathbf{v}_t \odot \left(\mathbb{1} + 4\eta \left(\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{w}_t - \mathbf{y}) \right) \right)$$

Is there a more general picture?

- Gradient updates using Hadamard parametrization:

$$\mathbf{u}_{t+1} = \mathbf{u}_t \odot \left(\mathbb{1} - 4\eta \left(\underbrace{\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{w}_t - \mathbf{y})}_{=\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w})} \right) \right)$$

$$\mathbf{v}_{t+1} = \mathbf{v}_t \odot \left(\mathbb{1} + 4\eta \left(\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{w}_t - \mathbf{y}) \right) \right)$$

- For small η these updates can be written as:

$$\mathbf{w}_{t+1} = \mathbf{w}_t^+ \odot \exp\left(-\eta \nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}_t)\right) - \mathbf{w}_t^- \odot \exp\left(\eta \nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}_t)\right)$$

Is there a more general picture?

- Gradient updates using Hadamard parametrization:

$$\mathbf{u}_{t+1} = \mathbf{u}_t \odot \left(\mathbb{1} - 4\eta \left(\underbrace{\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{w}_t - \mathbf{y})}_{=\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w})} \right) \right)$$

$$\mathbf{v}_{t+1} = \mathbf{v}_t \odot \left(\mathbb{1} + 4\eta \left(\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\mathbf{w}_t - \mathbf{y}) \right) \right)$$

- For small η these updates can be written as:

$$\mathbf{w}_{t+1} = \mathbf{w}_t^+ \odot \exp\left(-\eta \nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}_t)\right) - \mathbf{w}_t^- \odot \exp\left(\eta \nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}_t)\right)$$

- This is the EG$\pm$ algorithm of Kivinen and Warmuth [1997] and was shown by Ghai et al. [2019] to be unconstrained mirror descent initialized at 0 with the mirror map given by the hyperbolic entropy:

$$\psi_\gamma(\mathbf{w}) = \sum_{i=1}^{d} \left(w_i \cdot \operatorname{arcsinh}(w_i/\gamma) - \sqrt{w_i^2 + \gamma^2} \right)$$

Unconstrained Mirror Descent with Squared Error

- The optimization objective is the constrained squared error of a linear model (not necessarily well-specified)

$$\mathcal{L}(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

Unconstrained Mirror Descent with Squared Error

- The optimization objective is the constrained squared error of a linear model (not necessarily well-specified)

$$\mathcal{L}(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

- Update rule in continuous time:

$$\frac{d}{dt}\mathbf{w}_t = -\left(\nabla^2\psi(\mathbf{w}_t)\right)^{-1}\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}_t),$$

where $\psi : \mathbb{R}^d \to \mathbb{R}$ is a strictly convex differentiable function whose gradient is surjective, called a *mirror map*.

Unconstrained Mirror Descent with Squared Error

- The optimization objective is the constrained squared error of a linear model (not necessarily well-specified)

$$\mathcal{L}(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

- Update rule in continuous time:

$$\frac{d}{dt}\mathbf{w}_t = -\left(\nabla^2\psi(\mathbf{w}_t)\right)^{-1}\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}_t),$$

  where $\psi : \mathbb{R}^d \to \mathbb{R}$ is a strictly convex differentiable function whose gradient is surjective, called a *mirror map*.
- Setting $\psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$ gives vanilla gradient descent.

Unconstrained Mirror Descent with Squared Error

- The optimization objective is the constrained squared error of a linear model (not necessarily well-specified)

$$\mathcal{L}(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

- Update rule in continuous time:

$$\frac{d}{dt}\mathbf{w}_t = -\left(\nabla^2\psi(\mathbf{w}_t)\right)^{-1}\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}_t),$$

where $\psi : \mathbb{R}^d \to \mathbb{R}$ is a strictly convex differentiable function whose gradient is surjective, called a *mirror map*.

- Setting $\psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$ gives vanilla gradient descent.
- Discrete-time updates given by:

$$\nabla\psi(\mathbf{w}_{t+1}) - \nabla\psi(\mathbf{w}_t) = -\eta\nabla_w\mathcal{L}(\mathbf{w}_t)$$

Analysis of Mirror Descent (Optimization)

- A key quantity in the analysis is the Bregman divergence

$$D_\psi(\mathbf{w}, \mathbf{w}_0) = \psi(\mathbf{w}) - \psi(\mathbf{w}_0) - \langle \nabla\psi(\mathbf{w}_0), \mathbf{w} - \mathbf{w}_0 \rangle.$$

Analysis of Mirror Descent (Optimization)

- A key quantity in the analysis is the Bregman divergence

$$D_\psi(\mathbf{w}, \mathbf{w}_0) = \psi(\mathbf{w}) - \psi(\mathbf{w}_0) - \langle \nabla\psi(\mathbf{w}_0), \mathbf{w} - \mathbf{w}_0 \rangle.$$

- For a reference point $\mathbf{w}^\star$ (not necessarily optimal) we have

$$-\frac{d}{dt} D_\psi(\mathbf{w}^\star, \mathbf{w}_t) = \langle -\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}_t), \mathbf{w}^\star - \mathbf{w}_t \rangle \underbrace{\geq \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}^\star)}_{\text{by convexity}}$$

Analysis of Mirror Descent (Optimization)

- A key quantity in the analysis is the Bregman divergence

$$D_\psi(\mathbf{w}, \mathbf{w}_0) = \psi(\mathbf{w}) - \psi(\mathbf{w}_0) - \langle \nabla\psi(\mathbf{w}_0), \mathbf{w} - \mathbf{w}_0 \rangle.$$

- For a reference point $\mathbf{w}^\star$ (not necessarily optimal) we have

$$-\frac{d}{dt} D_\psi(\mathbf{w}^\star, \mathbf{w}_t) = \langle -\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}_t), \mathbf{w}^\star - \mathbf{w}_t \rangle \underbrace{\geq \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}^\star)}_{\text{by convexity}}$$

- Thus, we have:

$$\frac{1}{T} D_\psi(\mathbf{w}^\star, \mathbf{w}_0) = \frac{1}{T} \int_0^T -\frac{d}{dt} D_\psi(\mathbf{w}^\star, \mathbf{w}) dt \geq \frac{1}{T} \int_0^T \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}^\star) dt$$

Analysis of Mirror Descent (Optimization)

- A key quantity in the analysis is the Bregman divergence

$$D_\psi(\mathbf{w}, \mathbf{w}_0) = \psi(\mathbf{w}) - \psi(\mathbf{w}_0) - \langle \nabla\psi(\mathbf{w}_0), \mathbf{w} - \mathbf{w}_0 \rangle.$$

- For a reference point $\mathbf{w}^\star$ (not necessarily optimal) we have

$$-\frac{d}{dt} D_\psi(\mathbf{w}^\star, \mathbf{w}_t) = \langle -\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}_t), \mathbf{w}^\star - \mathbf{w}_t \rangle \underbrace{\geq \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}^\star)}_{\text{by convexity}}$$

- Thus, we have:

$$\frac{1}{T} D_\psi(\mathbf{w}^\star, \mathbf{w}_0) = \frac{1}{T} \int_0^T -\frac{d}{dt} D_\psi(\mathbf{w}^\star, \mathbf{w}) dt \geq \frac{1}{T} \int_0^T \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}^\star) dt$$

- This suggests picking the following average as the solution:

$$\bar{\mathbf{w}} = \int_0^T \mathbf{w}_t dt$$

Analysis of Mirror Descent (Statistics)

- How do we get a handle on the statistical properties of mirror descent?

Analysis of Mirror Descent (Statistics)

- How do we get a handle on the statistical properties of mirror descent?
- For optimization, we simply used

$$-\frac{d}{dt} D_\psi(\mathbf{w}^\star, \mathbf{w}_t) = \langle -\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}_t), \mathbf{w}^\star - \mathbf{w}_t \rangle \underbrace{\geq \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}^\star)}_{\text{by convexity}}$$

Analysis of Mirror Descent (Statistics)

- How do we get a handle on the statistical properties of mirror descent?
- For optimization, we simply used

$$-\frac{d}{dt} D_\psi(\mathbf{w}^\star, \mathbf{w}_t) = \langle -\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}_t), \mathbf{w}^\star - \mathbf{w}_t \rangle \underbrace{\geq \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}^\star)}_{\text{by convexity}}$$

- Instead, when $\mathcal{L}(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$, we have the following *equality*:

$$-\frac{d}{dt} D_\psi(\mathbf{w}^\star, \mathbf{w}_t) = \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}^\star) + \frac{1}{n}\|\mathbf{X}\mathbf{w}_t - \mathbf{X}\mathbf{w}^\star\|^2$$

Analysis of Mirror Descent (Statistics)

- How do we get a handle on the statistical properties of mirror descent?
- For optimization, we simply used

$$-\frac{d}{dt} D_\psi(\mathbf{w}^\star, \mathbf{w}_t) = \langle -\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}_t), \mathbf{w}^\star - \mathbf{w}_t \rangle \underbrace{\geq \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}^\star)}_{\text{by convexity}}$$

- Instead, when $\mathcal{L}(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$, we have the following *equality*:

$$-\frac{d}{dt} D_\psi(\mathbf{w}^\star, \mathbf{w}_t) = \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}^\star) + \frac{1}{n}\|\mathbf{X}\mathbf{w}_t - \mathbf{X}\mathbf{w}^\star\|^2$$

- The same analysis becomes

$$\frac{1}{T} D_\psi(\mathbf{w}^\star, \mathbf{w}_0) = \frac{1}{T} \int_0^T \left(\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}^\star) + \frac{1}{n}\|\mathbf{X}\mathbf{w}_t - \mathbf{X}\mathbf{w}^\star\|^2 \right) dt$$

Analysis of Mirror Descent (Statistics)

- How do we get a handle on the statistical properties of mirror descent?
- For optimization, we simply used

$$-\frac{d}{dt} D_\psi(\mathbf{w}^\star, \mathbf{w}_t) = \langle -\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}_t), \mathbf{w}^\star - \mathbf{w}_t \rangle \underbrace{\geq \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}^\star)}_{\text{by convexity}}$$

- Instead, when $\mathcal{L}(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$, we have the following *equality*:

$$-\frac{d}{dt} D_\psi(\mathbf{w}^\star, \mathbf{w}_t) = \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}^\star) + \frac{1}{n}\|\mathbf{X}\mathbf{w}_t - \mathbf{X}\mathbf{w}^\star\|^2$$

- The same analysis becomes

$$\frac{1}{T} D_\psi(\mathbf{w}^\star, \mathbf{w}_0) = \frac{1}{T} \int_0^T \left(\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}^\star) + \frac{1}{n}\|\mathbf{X}\mathbf{w}_t - \mathbf{X}\mathbf{w}^\star\|^2 \right) dt$$

- Stop at a time $T^\star$, such that the *offset* condition holds:

$$\mathcal{L}(\mathbf{w}_{T^\star}) - \mathcal{L}(\mathbf{w}^\star) + \frac{1}{n}\|\mathbf{X}\mathbf{w}_{T^\star} - \mathbf{X}\mathbf{w}^\star\|^2 \leq \varepsilon$$

Offset Rademacher Complexity

Offset Rademacher Complexity

- Slightly informally, an estimator $\hat{g}$ and a class $\mathcal{F}$ satisfy the offset condition with parameters $\varepsilon \geq 0$, $c > 0$, if

$$\mathcal{L}(\hat{g}) - \mathcal{L}(g_{\mathcal{F}}) + c\|\hat{g} - g_{\mathcal{F}}\|_n^2 \leq \varepsilon$$

- Above, $\hat{g}$ *need not* be in $\mathcal{F}$, $g_{\mathcal{F}} \in \mathcal{F}$ is the minimizer of the *true* risk, and the last term is the ℓ_2 distance between $\hat{g}$ and $g_{\mathcal{F}}$ on the (training) sample.

Offset Rademacher Complexity

- Slightly informally, an estimator $\hat{g}$ and a class $\mathcal{F}$ satisfy the offset condition with parameters $\varepsilon \geq 0$, $c > 0$, if

$$\mathcal{L}(\hat{g}) - \mathcal{L}(g_{\mathcal{F}}) + c\|\hat{g} - g_{\mathcal{F}}\|_n^2 \leq \varepsilon$$

- Above, $\hat{g}$ *need not* be in $\mathcal{F}$, $g_{\mathcal{F}} \in \mathcal{F}$ is the minimizer of the *true* risk, and the last term is the ℓ_2 distance between $\hat{g}$ and $g_{\mathcal{F}}$ on the (training) sample.
- Offset Rademacher Complexity [Liang et al. [2015]]

$$\mathrm{RAD}_n(\mathcal{F}, c) = \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\left(2\sigma_i f(x_i) - c f(x_i)^2\right)\right\}\right]$$

Results I

For a class of functions $\mathcal{F}$ and an estimator $\hat{g}$, let $\mathcal{E}(\hat{g}, \mathcal{F})$ denote the excess risk of $\hat{g}$ with respect to the class $\mathcal{F}$.

Theorem (Vaškevičius, Kanade, Rebeschini 2020)

Fix any $\mathbf{w}_0$, $R > 0$, let ψ be a mirror map, and let $\mathcal{F}(\mathbf{w}_0, R) = \{g_{\mathbf{w}} : D_\psi(\mathbf{w}, \mathbf{w}_0) \leq R\}$. For any $\varepsilon > 0$, there exists a data-dependent stopping time $t^\star \leq 2R/\varepsilon$ and constants c_1, c_2 that depend on boundedness constants of the data, we have:

$$\mathbb{E}[\mathcal{E}(g_{\mathbf{w}_{t^\star}}, \mathcal{F}(\mathbf{w}_0, R))] \leq c_1 \mathbb{E}[\mathrm{RAD}_n(\mathcal{F}(\alpha_0, R) - g_{\mathcal{F}(\alpha_0, R)}, c_2)] + \varepsilon.$$

Results II

Application to in-sample predictions under ℓ_1-constraints.

Theorem (Vaškevičius, Kanade, Rebeschini 2020)

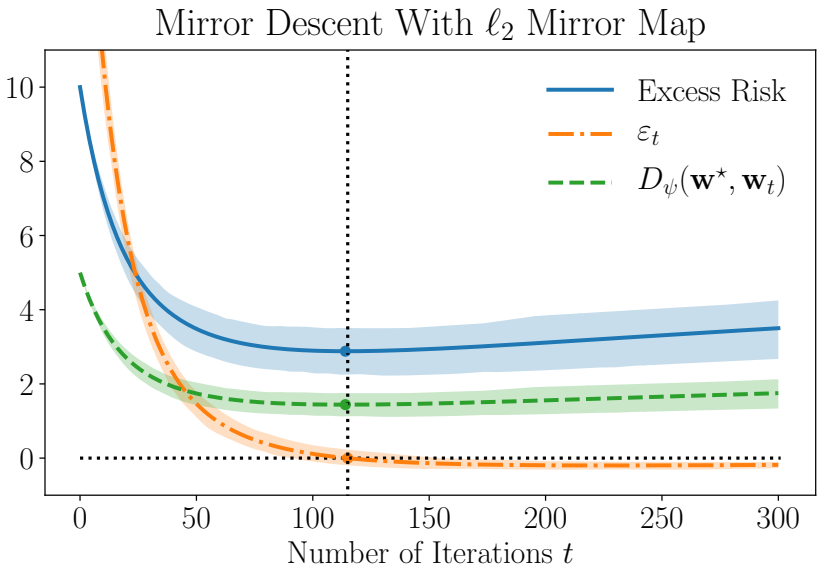
Suppose that $\mathbf{X}$ is a fixed-design matrix with columns bounded in ℓ_2 norm and that $\mathbf{y} = \mathbf{X}\mathbf{w}^\star + \xi$, where ξ is a vector with i.i.d. zero-mean σ^2-sub-Gaussian noise. When using mirror descent with hyperbolic entropy as a mirror map,

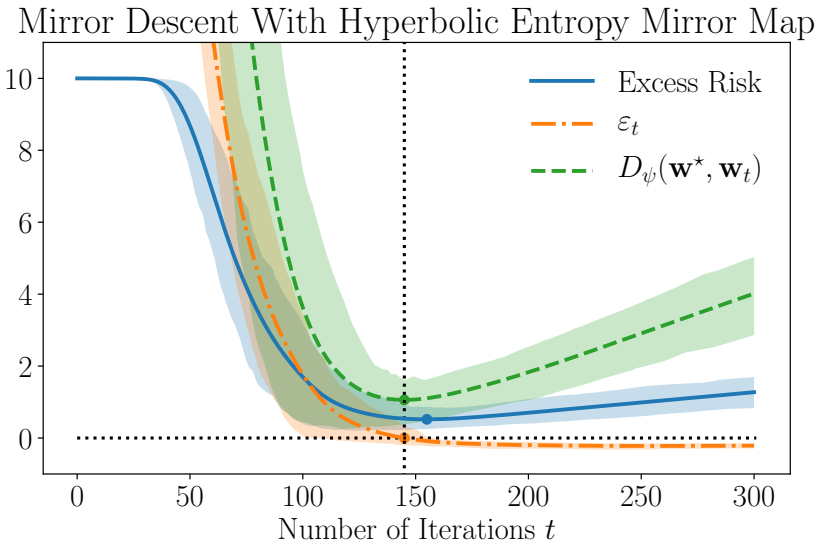
$$\psi_\gamma(\mathbf{w}) = \sum_{i=1}^{d} \left(w_i \cdot \operatorname{arcsinh}(w_i/\gamma) - \sqrt{w_i^2 + \gamma^2} \right),$$

there exists a data-dependent stopping time $t^\star \lesssim \sqrt{n}/(\eta \cdot \sigma\sqrt{\log d})$, such that with high probability:

$$\frac{1}{n}\|\mathbf{X}\mathbf{w}^\star - \mathbf{X}\mathbf{w}_{t^\star}\|_2^2 \lesssim \frac{\|\mathbf{w}^\star\|_1 \cdot \sigma \cdot \sqrt{\log d}}{\sqrt{n}} \cdot \log(1/\gamma).$$

Comparison between ℓ_2 and Hyperbolic Entropy Mirror Maps





Here $\varepsilon_t = \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}^\star) + \|\mathbf{X}\mathbf{w}_t - \mathbf{X}\mathbf{w}^\star\|^2$.

Summary and Research Directions

Main contribution:

Under the RIP, implicitly-reg. GD (parametriz. + initializ. + early stopping) yields:

- Optimal statistical rates (minimax)
- Instance adaptivity (beyond minimax, dim.-free rates for high signal-to-noise)
- Optimal computational cost (modulo log terms, same cost of reading data)

Summary and Research Directions

Main contribution:

Under the RIP, implicitly-reg. GD (parametriz. + initializ. + early stopping) yields:

- Optimal statistical rates (minimax)
- Instance adaptivity (beyond minimax, dim.-free rates for high signal-to-noise)
- Optimal computational cost (modulo log terms, same cost of reading data)

Mirror Descent

- Implicit Regularization Properties of Early-Stopped Mirror Descent
- Analysis of excess risk using offset Rademacher complexities

Summary and Research Directions

Main contribution:

Under the RIP, implicitly-reg. GD (parametriz. + initializ. + early stopping) yields:

- Optimal statistical rates (minimax)
- Instance adaptivity (beyond minimax, dim.-free rates for high signal-to-noise)
- Optimal computational cost (modulo log terms, same cost of reading data)

Mirror Descent

- Implicit Regularization Properties of Early-Stopped Mirror Descent
- Analysis of excess risk using offset Rademacher complexities

Future Research Directions:

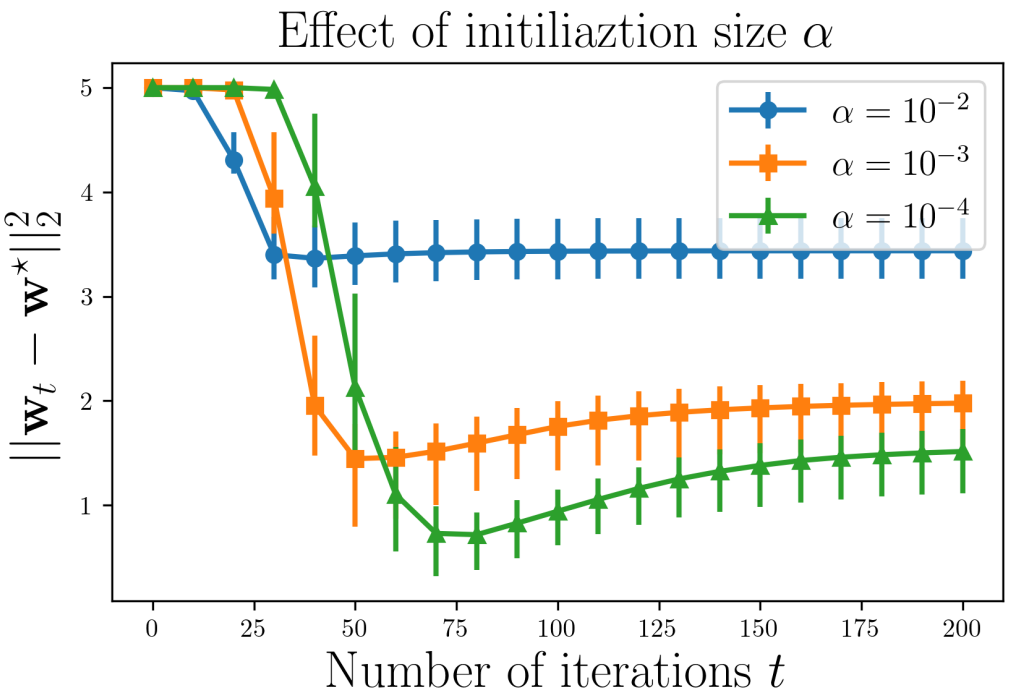
- Analysis of fast rates for sparse recovery using mirror descent framework?
- Understanding loss functions beyond squared loss
- Mirror descent to optimize over non-convex "balls"?

Extra Slides

Effects of Initialization Size: Error Size and Stopping Time

Trade-off: Smaller initialization size α yields:

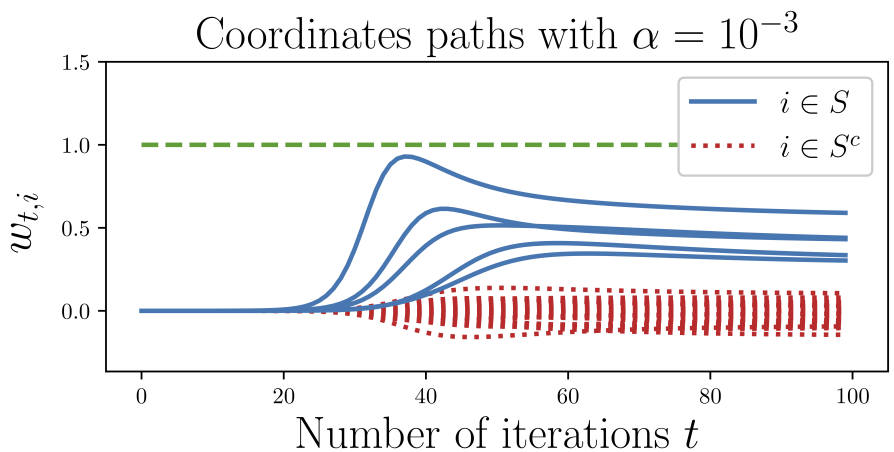
- Smaller error ($\|\mathbf{w}_{t^\star} \odot \mathbf{1}_{S^c}\|_\infty \lesssim \sqrt{\alpha}$)
- Longer stopping time ($t^\star \sim \log 1/\alpha$)



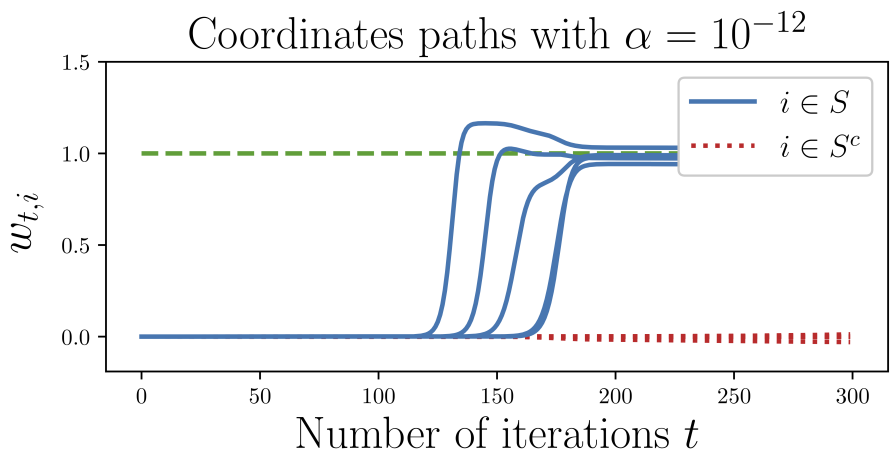
Effects of Initialization Size: Coordinates Path

If initialization size is small enough, Thm yields $\|\mathbf{w}_{t^\star} \odot \mathbf{1}_{S^c}\|_\infty \lesssim \sqrt{\alpha}$:

- Error outside of true support decreases with α
- GD stops before fitting coordinates outside true support S



(a) Initialization **not** small enough



(b) Initialization **small enough**

Instance Adaptivity: Dim-Free Rates (beyond minimax)

- Lasso suffers from a dimension-dependent bias ($\log d$)

Instance Adaptivity: Dim-Free Rates (beyond minimax)

- Lasso suffers from a dimension-dependent bias ($\log d$)
  - $\mathbf{X}^\mathsf{T}\mathbf{X}/n = \mathbf{I} \Rightarrow$ Lasso $w_i^\lambda = \mathrm{sign}(w_i^{\mathrm{LS}})(|w_i^{\mathrm{LS}}| - \lambda)_+$, with $\mathbf{w}^{\mathrm{LS}}$ least squares sol.

Instance Adaptivity: Dim-Free Rates (beyond minimax)

- Lasso suffers from a dimension-dependent bias ($\log d$)
  - $\mathbf{X}^\mathsf{T}\mathbf{X}/n = \mathbf{I} \Rightarrow$ Lasso $w_i^\lambda = \mathrm{sign}(w_i^{\mathrm{LS}})(|w_i^{\mathrm{LS}}| - \lambda)_+$, with $\mathbf{w}^{\mathrm{LS}}$ least squares sol.
  - For sub-Gaussian noise, minimax rates achieved by $\lambda = \Theta(\sqrt{\sigma^2 \mathbf{log(}\boldsymbol{d}\mathbf{)}/n})$

Instance Adaptivity: Dim-Free Rates (beyond minimax)

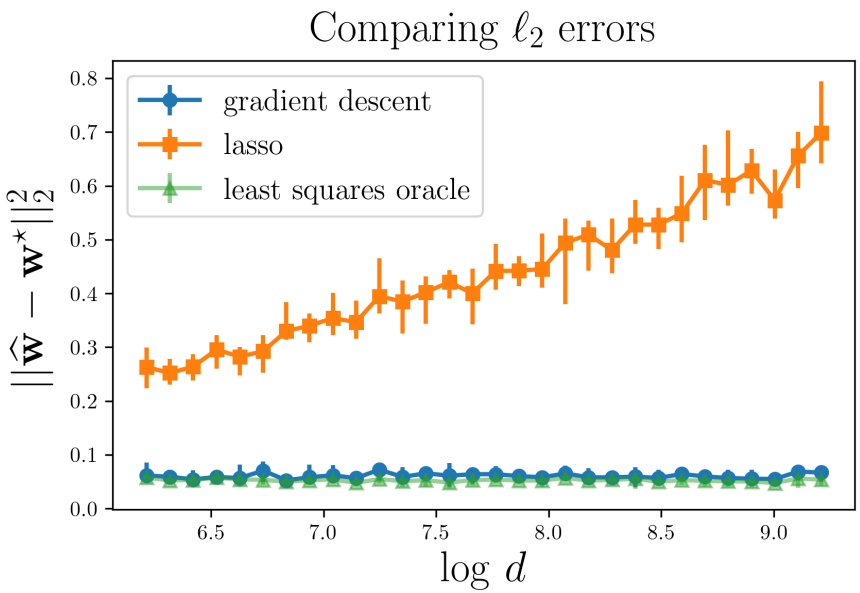
- Lasso suffers from a dimension-dependent bias ($\log d$)
  - $\mathbf{X}^\top \mathbf{X}/n = \mathbf{I} \Rightarrow$ Lasso $w_i^\lambda = \mathrm{sign}(w_i^{\mathrm{LS}})(|w_i^{\mathrm{LS}}| - \lambda)_+$, with $\mathbf{w}^{\mathrm{LS}}$ least squares sol.
  - For sub-Gaussian noise, minimax rates achieved by $\lambda = \Theta(\sqrt{\sigma^2 \log(d)/n})$
- In contrast, **in a high signal-to-noise ratio setting, GD has no bias and achieves better rates than minimax:**

Instance Adaptivity: Dim-Free Rates (beyond minimax)

- Lasso suffers from a dimension-dependent bias ($\log d$)
  - $\mathbf{X}^\mathsf{T}\mathbf{X}/n = \mathbf{I} \Rightarrow$ Lasso $w_i^\lambda = \mathrm{sign}(w_i^{\mathrm{LS}})(|w_i^{\mathrm{LS}}| - \lambda)_+$, with $\mathbf{w}^{\mathrm{LS}}$ least squares sol.
  - For sub-Gaussian noise, minimax rates achieved by $\lambda = \Theta(\sqrt{\sigma^2 \log(d)/n})$
- In contrast, **in a high signal-to-noise ratio setting, GD has no bias and achieves better rates than minimax:**

Corollary

If $w^\star_{\min} \gtrsim \left\| \frac{1}{n}\mathbf{X}^\mathsf{T}\xi \right\|_\infty$, *sub-Gaussian noise, then* $\|\mathbf{w}_{t^\star} - \mathbf{w}^\star\|_2^2 \lesssim k\frac{\sigma^2 \log k}{n}$ *w.h.p.*



Instance Adaptivity: Statistical Phase Transitions

- GD only recovers coord.'s on S growing faster than on S^C: $|w_i^\star| \gtrsim \left\| \frac{1}{n} \mathbf{X}^\mathsf{T} \xi \right\|_\infty$

Instance Adaptivity: Statistical Phase Transitions

- GD only recovers coord.'s on S growing faster than on S^C: $|w_i^\star| \gtrsim \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty$
  - For other coordinates on S, even if GD does not recover them, the error is proportional to $\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty$ per coordinate (the minimax rate is $k\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty^2$)

Instance Adaptivity: Statistical Phase Transitions

- GD only recovers coord.'s on S growing faster than on S^C: $|w_i^\star| \gtrsim \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty$
  - For other coordinates on S, even if GD does not recover them, the error is proportional to $\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty$ per coordinate (the minimax rate is $k\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty^2$)
- If $w_{\min}^\star - \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty > \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty$ **all** coordinates on the true support S grow exponentially at a faster rate than **all** the coordinates on S^C

Instance Adaptivity: Statistical Phase Transitions

- GD only recovers coord.'s on S growing faster than on S^C: $|w_i^\star| \gtrsim \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty$
  - For other coordinates on S, even if GD does not recover them, the error is proportional to $\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty$ per coordinate (the minimax rate is $k\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty^2$)
- If $w_{\min}^\star - \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty > \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty$ **all** coordinates on the true support S grow exponentially at a faster rate than **all** the coordinates on S^C
- At $w_{\min}^\star = 2\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty$, phase transitions **to dim.-independent error**

Instance Adaptivity: Statistical Phase Transitions

- GD only recovers coord.'s on S growing faster than on S^C: $|w_i^\star| \gtrsim \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty$
  - For other coordinates on S, even if GD does not recover them, the error is proportional to $\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty$ per coordinate (the minimax rate is $k\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty^2$)
- If $w_{\min}^\star - \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty > \left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty$ **all** coordinates on the true support S grow exponentially at a faster rate than **all** the coordinates on S^C
- At $w_{\min}^\star = 2\left\|\frac{1}{n}\mathbf{X}^\mathsf{T}\xi\right\|_\infty$, phase transitions **to dim.-independent error**

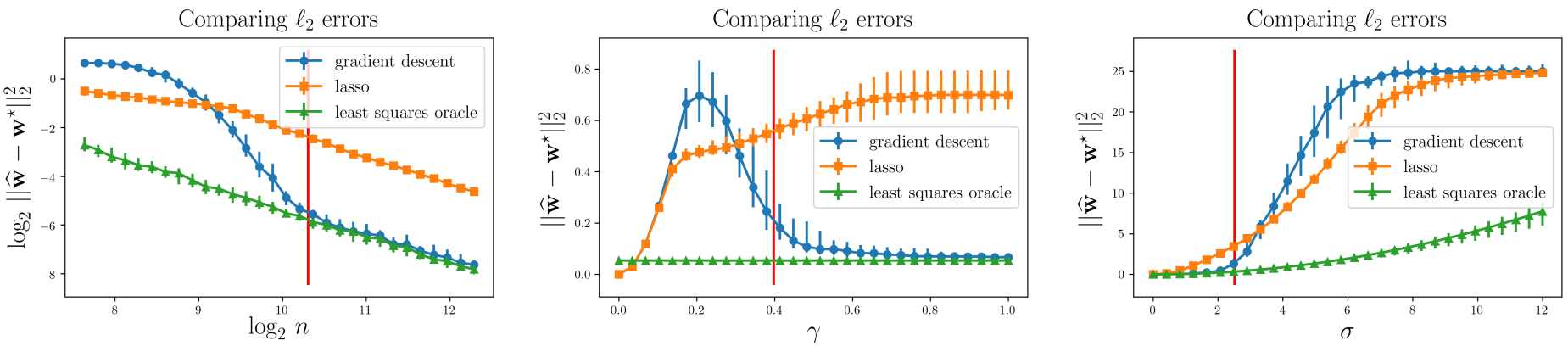
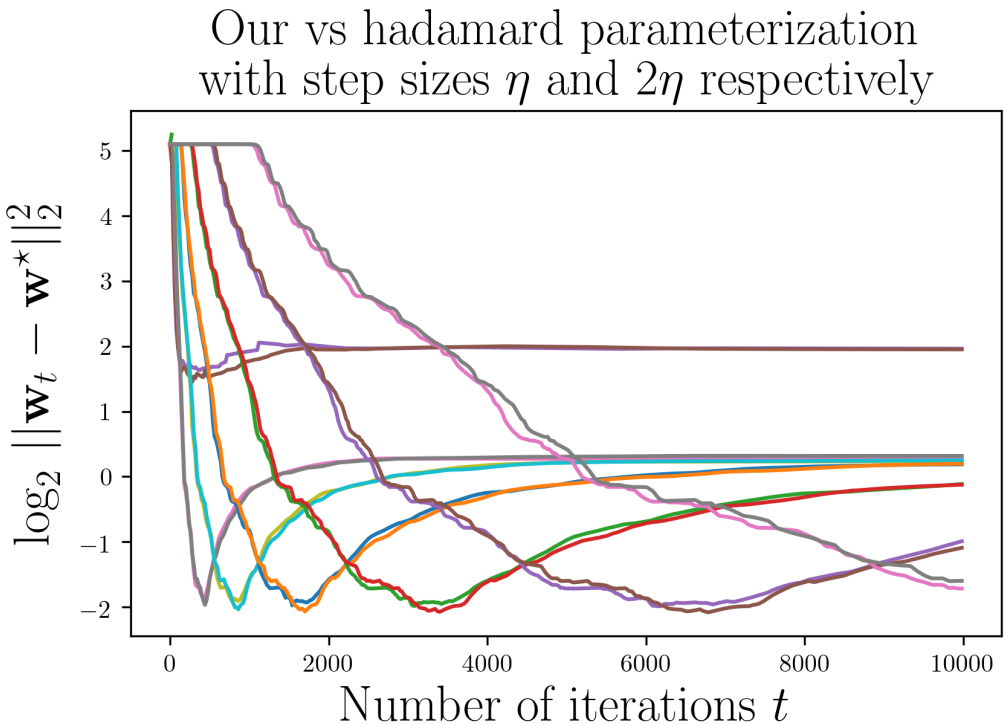


Figure: Let $\mathbf{w}^\star = \gamma \mathbf{1}_S$. Red lines are solutions to $\gamma = 2 \cdot \frac{\sigma\sqrt{2\log(2d)}}{\sqrt{n}}$ for sub-Gauss. noise

Concurrent work: Zhao et al. [2019]

Zhao et al. [2019] studies a closely related Hadamard product reparameterization $\mathbf{w}_t = \mathbf{u}_t \odot \mathbf{v}_t$ and uses GD to implicitly induce sparsity

(our parametrization: $\mathbf{w}_t = \mathbf{u}_t \odot \mathbf{u}_t - \mathbf{v}_t \odot \mathbf{v}_t$)



Parametrization is very similar, but algorithms, analysis and results are not!

Concurrent work: Zhao et al. [2019]

Similarities: RIP condition, minimax rates, instance adaptivity

Concurrent work: Zhao et al. [2019]

Similarities: RIP condition, minimax rates, instance adaptivity

Differences:

Concurrent work: Zhao et al. [2019]

Similarities: RIP condition, minimax rates, instance adaptivity

Differences:

- They have **worse conditions on step size**, depending on **unknown** quantities

Concurrent work: Zhao et al. [2019]

Similarities: RIP condition, minimax rates, instance adaptivity

Differences:

- They have **worse conditions on step size**, depending on **unknown** quantities
  - They require $\eta \lesssim \frac{\boldsymbol{w^\star_{\min}}}{w^\star_{\max}} (\log \frac{d}{\alpha})^{-1}$ while we require $\eta \lesssim 1/w^\star_{\max}$

Concurrent work: Zhao et al. [2019]

Similarities: RIP condition, minimax rates, instance adaptivity

Differences:

- They have **worse conditions on step size**, depending on **unknown** quantities
  - They require $\eta \lesssim \frac{w^\star_{\min}}{w^\star_{\max}} (\log \frac{d}{\alpha})^{-1}$ while we require $\eta \lesssim 1/w^\star_{\max}$
  - Our theory show how $w^\star_{\max}$ can be computed from the data, while in their case η is additional hyperparameter to be tuned

Concurrent work: Zhao et al. [2019]

Similarities: RIP condition, minimax rates, instance adaptivity

Differences:

- They have **worse conditions on step size**, depending on **unknown** quantities
  - They require $\eta \lesssim \frac{w^\star_{\min}}{w^\star_{\max}} (\log \frac{d}{\alpha})^{-1}$ while we require $\eta \lesssim 1/w^\star_{\max}$
  - Our theory show how $w^\star_{\max}$ can be computed from the data, while in their case η is additional hyperparameter to be tuned
- Their theory **does not properly handle noisy settings** and **cannot recover smallest possible signals**

Concurrent work: Zhao et al. [2019]

Similarities: RIP condition, minimax rates, instance adaptivity

Differences:

- They have **worse conditions on step size**, depending on **unknown** quantities
  - They require $\eta \lesssim \frac{w^\star_{\min}}{w^\star_{\max}} (\log \frac{d}{\alpha})^{-1}$ while we require $\eta \lesssim 1/w^\star_{\max}$
  - Our theory show how $w^\star_{\max}$ can be computed from the data, while in their case η is additional hyperparameter to be tuned
- Their theory **does not properly handle noisy settings** and **cannot recover smallest possible signals**
  - Let $\kappa := \frac{w^\star_{\max}}{w^\star_{\min}}$ They require RIP $\delta \lesssim \frac{1}{\kappa \sqrt{k} \log(d/\alpha)}$, while we have $\delta \lesssim \frac{1}{\sqrt{k} \log \kappa}$

Concurrent work: Zhao et al. [2019]

Similarities: RIP condition, minimax rates, instance adaptivity

Differences:

- They have **worse conditions on step size**, depending on **unknown** quantities
  - They require $\eta \lesssim \frac{w^\star_{\min}}{w^\star_{\max}} (\log \frac{d}{\alpha})^{-1}$ while we require $\eta \lesssim 1/w^\star_{\max}$
  - Our theory show how $w^\star_{\max}$ can be computed from the data, while in their case η is additional hyperparameter to be tuned
- Their theory **does not properly handle noisy settings** and **cannot recover smallest possible signals**
  - Let $\kappa := \frac{w^\star_{\max}}{w^\star_{\min}}$ They require RIP $\delta \lesssim \frac{1}{\kappa \sqrt{k} \log(d/\alpha)}$, while we have $\delta \lesssim \frac{1}{\sqrt{k} \log \kappa}$
  - If $w^\star_{\min} \asymp \sigma \sqrt{\log d}/\sqrt{n}$, they have $\delta = O(1/(\sqrt{k}\sqrt{n}))$, which is in general impossible to satisfy with random design matrices (e.g. $\mathbf{X}$ i.i.d. Gaussian)

Concurrent work: Zhao et al. [2019]

Similarities: RIP condition, minimax rates, instance adaptivity

Differences:

- They have **worse conditions on step size**, depending on **unknown** quantities
  - They require $\eta \lesssim \frac{w^\star_{\min}}{w^\star_{\max}} (\log \frac{d}{\alpha})^{-1}$ while we require $\eta \lesssim 1/w^\star_{\max}$
  - Our theory show how $w^\star_{\max}$ can be computed from the data, while in their case η is additional hyperparameter to be tuned
- Their theory **does not properly handle noisy settings** and **cannot recover smallest possible signals**
  - Let $\kappa := \frac{w^\star_{\max}}{w^\star_{\min}}$ They require RIP $\delta \lesssim \frac{1}{\kappa \sqrt{k} \log(d/\alpha)}$, while we have $\delta \lesssim \frac{1}{\sqrt{k} \log \kappa}$
  - If $w^\star_{\min} \asymp \sigma \sqrt{\log d}/\sqrt{n}$, they have $\delta = O(1/(\sqrt{k}\sqrt{n}))$, which is in general impossible to satisfy with random design matrices (e.g. $\mathbf{X}$ i.i.d. Gaussian)
- They only consider constant step size $\Rightarrow$ **do not achieve comput. optimality**

Concurrent work: Zhao et al. [2019]

Similarities: RIP condition, minimax rates, instance adaptivity

Differences:

- They have **worse conditions on step size**, depending on **unknown** quantities
  - They require $\eta \lesssim \frac{w^\star_{\min}}{w^\star_{\max}} (\log \frac{d}{\alpha})^{-1}$ while we require $\eta \lesssim 1/w^\star_{\max}$
  - Our theory show how $w^\star_{\max}$ can be computed from the data, while in their case η is additional hyperparameter to be tuned
- Their theory **does not properly handle noisy settings** and **cannot recover smallest possible signals**
  - Let $\kappa := \frac{w^\star_{\max}}{w^\star_{\min}}$ They require RIP $\delta \lesssim \frac{1}{\kappa \sqrt{k} \log(d/\alpha)}$, while we have $\delta \lesssim \frac{1}{\sqrt{k} \log \kappa}$
  - If $w^\star_{\min} \asymp \sigma \sqrt{\log d}/\sqrt{n}$, they have $\delta = O(1/(\sqrt{k}\sqrt{n}))$, which is in general impossible to satisfy with random design matrices (e.g. $\mathbf{X}$ i.i.d. Gaussian)
- They only consider constant step size $\Rightarrow$ **do not achieve comput. optimality**
  - Due to different constraints on step sizes, even in the case of constant step size our algorithm is can be faster by a factor $\sqrt{n}$

Summary and Research Directions

Main contribution:

Under the RIP, implicitly-reg. GD (parametriz. + initializ. + early stopping) yields:

- Optimal statistical rates (minimax)
- Instance adaptivity (beyond minimax, dim.-free rates for high signal-to-noise)
- Optimal computational cost (modulo log terms, same cost of reading data)

Summary and Research Directions

Main contribution:

Under the RIP, implicitly-reg. GD (parametriz. + initializ. + early stopping) yields:

- Optimal statistical rates (minimax)
- Instance adaptivity (beyond minimax, dim.-free rates for high signal-to-noise)
- Optimal computational cost (modulo log terms, same cost of reading data)

Further improvements: (we have empirical evidence)

- Optimal sample rates
- Restricted Eigenvalue (RE) condition, to allow for correlated design

Summary and Research Directions

Main contribution:

Under the RIP, implicitly-reg. GD (parametriz. + initializ. + early stopping) yields:

- Optimal statistical rates (minimax)
- Instance adaptivity (beyond minimax, dim.-free rates for high signal-to-noise)
- Optimal computational cost (modulo log terms, same cost of reading data)

Further improvements: (we have empirical evidence)

- Optimal sample rates
- Restricted Eigenvalue (RE) condition, to allow for correlated design

General Research Directions:

- Establish **general math. framework** for implicit reg. and sparse recovery (cf. bias-variance for ridge regression, *basic inequality* for M estimators, connection to localized complexity measures)
- Establish a complete theory of early-stopping for sparse estimation (see above), **prediction**, **var. selection**, **oracle ineq.**, with focus on **comput. efficiency**

Summary and Research Directions

Main contribution:

Under the RIP, implicitly-reg. GD (parametriz. + initializ. + early stopping) yields:

- Optimal statistical rates (minimax)
- Instance adaptivity (beyond minimax, dim.-free rates for high signal-to-noise)
- Optimal computational cost (modulo log terms, same cost of reading data)

Further improvements: (we have empirical evidence)

- Optimal sample rates
- Restricted Eigenvalue (RE) condition, to allow for correlated design

General Research Directions:

- Establish **general math. framework** for implicit reg. and sparse recovery (cf. bias-variance for ridge regression, *basic inequality* for M estimators, connection to localized complexity measures)
- Establish a complete theory of early-stopping for sparse estimation (see above), **prediction**, **var. selection**, **oracle ineq.**, with focus on **comput. efficiency**
  - Explicit link with known penalty terms related to sparse recovery?

Summary and Research Directions

Main contribution:

Under the RIP, implicitly-reg. GD (parametriz. + initializ. + early stopping) yields:

- Optimal statistical rates (minimax)
- Instance adaptivity (beyond minimax, dim.-free rates for high signal-to-noise)
- Optimal computational cost (modulo log terms, same cost of reading data)

Further improvements: (we have empirical evidence)

- Optimal sample rates
- Restricted Eigenvalue (RE) condition, to allow for correlated design

General Research Directions:

- Establish **general math. framework** for implicit reg. and sparse recovery (cf. bias-variance for ridge regression, *basic inequality* for M estimators, connection to localized complexity measures)
- Establish a complete theory of early-stopping for sparse estimation (see above), **prediction**, **var. selection**, **oracle ineq.**, with focus on **comput. efficiency**
  - Explicit link with known penalty terms related to sparse recovery?
  - Can we apply some of the **techniques for ridge regression** (cf. slide 4)?

Further Improvements - Optimal* Sample Complexity

- Recall that we require the RIP constant δ to satisfy $\delta = \widetilde{O}(1/\sqrt{k})$

Further Improvements - Optimal* Sample Complexity

- Recall that we require the RIP constant δ to satisfy $\delta = \widetilde{O}(1/\sqrt{k})$
- Satisfying such an assumption requires $n \gtrsim \boldsymbol{k^2} \log(ed/k)$

Further Improvements - Optimal* Sample Complexity

- Recall that we require the RIP constant δ to satisfy $\delta = \widetilde{O}(1/\sqrt{k})$
- Satisfying such an assumption requires $n \gtrsim \boldsymbol{k^2} \log(ed/k)$
  - By random-matrix theory, $\|\mathbf{X}^\mathsf{T}\mathbf{X}/n - \mathbf{I}\| \lesssim \sqrt{k/n} + k/n$, $\|\cdot\|$ operator norm

Further Improvements - Optimal* Sample Complexity

- Recall that we require the RIP constant δ to satisfy $\delta = \widetilde{O}(1/\sqrt{k})$
- Satisfying such an assumption requires $n \gtrsim \boldsymbol{k^2} \log(ed/k)$
  - By random-matrix theory, $\|\mathbf{X}^\mathsf{T}\mathbf{X}/n - \mathbf{I}\| \lesssim \sqrt{k/n} + k/n$, $\|\cdot\|$ operator norm
  - Hence, we need $n \gtrsim k^2$ to satisfty $\|\mathbf{X}^\mathsf{T}\mathbf{X}/n - \mathbf{I}\| \lesssim 1/\sqrt{k}$

Further Improvements - Optimal* Sample Complexity

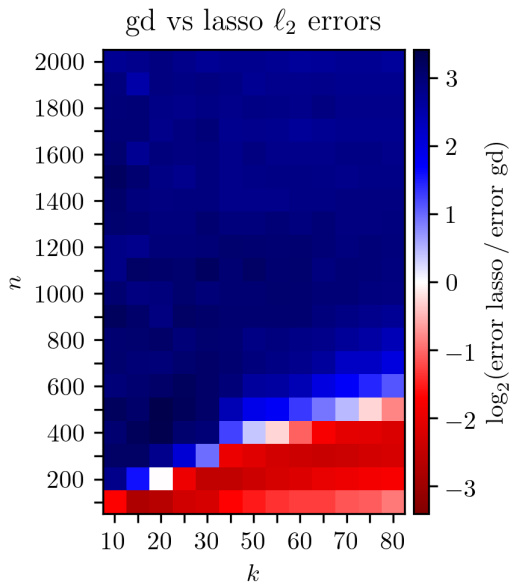
- Recall that we require the RIP constant δ to satisfy $\delta = \widetilde{O}(1/\sqrt{k})$
- Satisfying such an assumption requires $n \gtrsim \boldsymbol{k^2} \log(ed/k)$
  - By random-matrix theory, $\|\mathbf{X}^\mathsf{T}\mathbf{X}/n - \mathbf{I}\| \lesssim \sqrt{k/n} + k/n$, $\|\cdot\|$ operator norm
  - Hence, we need $n \gtrsim k^2$ to satisfty $\|\mathbf{X}^\mathsf{T}\mathbf{X}/n - \mathbf{I}\| \lesssim 1/\sqrt{k}$

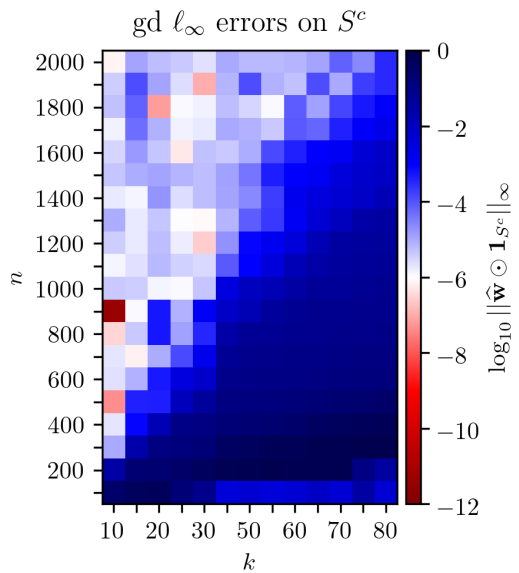
Sub-optimal sample complexity due to our analysis, not to algorithm:

Further Improvements - Optimal* Sample Complexity

- Recall that we require the RIP constant δ to satisfy $\delta = \widetilde{O}(1/\sqrt{k})$
- Satisfying such an assumption requires $n \gtrsim \boldsymbol{k^2} \log(ed/k)$
  - By random-matrix theory, $\|\mathbf{X}^\mathsf{T}\mathbf{X}/n - \mathbf{I}\| \lesssim \sqrt{k/n} + k/n$, $\|\cdot\|$ operator norm
  - Hence, we need $n \gtrsim k^2$ to satisfty $\|\mathbf{X}^\mathsf{T}\mathbf{X}/n - \mathbf{I}\| \lesssim 1/\sqrt{k}$

Sub-optimal sample complexity due to our analysis, not to algorithm:





(a) Sample complexity linear in k is enough for GD to match and eventually exceed **ℓ_2-error** performance of the Lasso

(b) Sample complexity linear in k is enough for GD to achieve the **ℓ_∞-error** in our main theorem: $\|\mathbf{w}_{t^\star} \odot \mathbf{1}_{S^c}\|_\infty \lesssim \sqrt{\alpha} < \frac{\varepsilon}{d}$

Further Improvements - Relaxing RIP Assumption

- Lasso attains minimax rates under a **Restricted Eigenvalue condition:**

Further Improvements - Relaxing RIP Assumption

- Lasso attains minimax rates under a **Restricted Eigenvalue condition:**
  - **RE(γ):** $\|\mathbf{X}\mathbf{w}\|_2^2/n \geq \gamma\|\mathbf{w}\|_2^2$ for vectors $\mathbf{w}$ satisfying the cone condition $\|\mathbf{w}_{S^c}\|_1 \leq c\|\mathbf{w}_S\|_1$ for a suitable choice of constant $c \geq 1$

Further Improvements - Relaxing RIP Assumption

- Lasso attains minimax rates under a **Restricted Eigenvalue condition:**
  - **RE(γ):** $\|\mathbf{X}\mathbf{w}\|_2^2/n \geq \gamma\|\mathbf{w}\|_2^2$ for vectors $\mathbf{w}$ satisfying the cone condition $\|\mathbf{w}_{S^c}\|_1 \leq c\|\mathbf{w}_S\|_1$ for a suitable choice of constant $c \geq 1$
  - Only imposes constraints on *lower* eigenvalues of $\mathbf{X}^\mathsf{T}\mathbf{X}/n$

Further Improvements - Relaxing RIP Assumption

- Lasso attains minimax rates under a **Restricted Eigenvalue condition:**
  - **RE(γ):** $\|\mathbf{X}\mathbf{w}\|_2^2/n \geq \gamma\|\mathbf{w}\|_2^2$ for vectors $\mathbf{w}$ satisfying the cone condition $\|\mathbf{w}_{S^c}\|_1 \leq c\|\mathbf{w}_S\|_1$ for a suitable choice of constant $c \geq 1$
  - Only imposes constraints on *lower* eigenvalues of $\mathbf{X}^\mathsf{T}\mathbf{X}/n$
  - The RE can be satisfied by random *correlated* designs

Further Improvements - Relaxing RIP Assumption

- Lasso attains minimax rates under a **Restricted Eigenvalue condition:**
  - **RE(γ):** $\|\mathbf{X}\mathbf{w}\|_2^2/n \geq \gamma\|\mathbf{w}\|_2^2$ for vectors $\mathbf{w}$ satisfying the cone condition $\|\mathbf{w}_{S^c}\|_1 \leq c\|\mathbf{w}_S\|_1$ for a suitable choice of constant $c \geq 1$
  - Only imposes constraints on *lower* eigenvalues of $\mathbf{X}^\mathsf{T}\mathbf{X}/n$
  - The RE can be satisfied by random *correlated* designs
- RE is **necessary** for fast rates for any poly. time algorithm [Zhang et al., 2014]

Further Improvements - Relaxing RIP Assumption

- Lasso attains minimax rates under a **Restricted Eigenvalue condition:**
  - **RE(γ):** $\|\mathbf{X}\mathbf{w}\|_2^2/n \geq \gamma\|\mathbf{w}\|_2^2$ for vectors $\mathbf{w}$ satisfying the cone condition $\|\mathbf{w}_{S^c}\|_1 \leq c\|\mathbf{w}_S\|_1$ for a suitable choice of constant $c \geq 1$
  - Only imposes constraints on *lower* eigenvalues of $\mathbf{X}^\mathsf{T}\mathbf{X}/n$
  - The RE can be satisfied by random *correlated* designs
- RE is **necessary** for fast rates for any poly. time algorithm [Zhang et al., 2014]

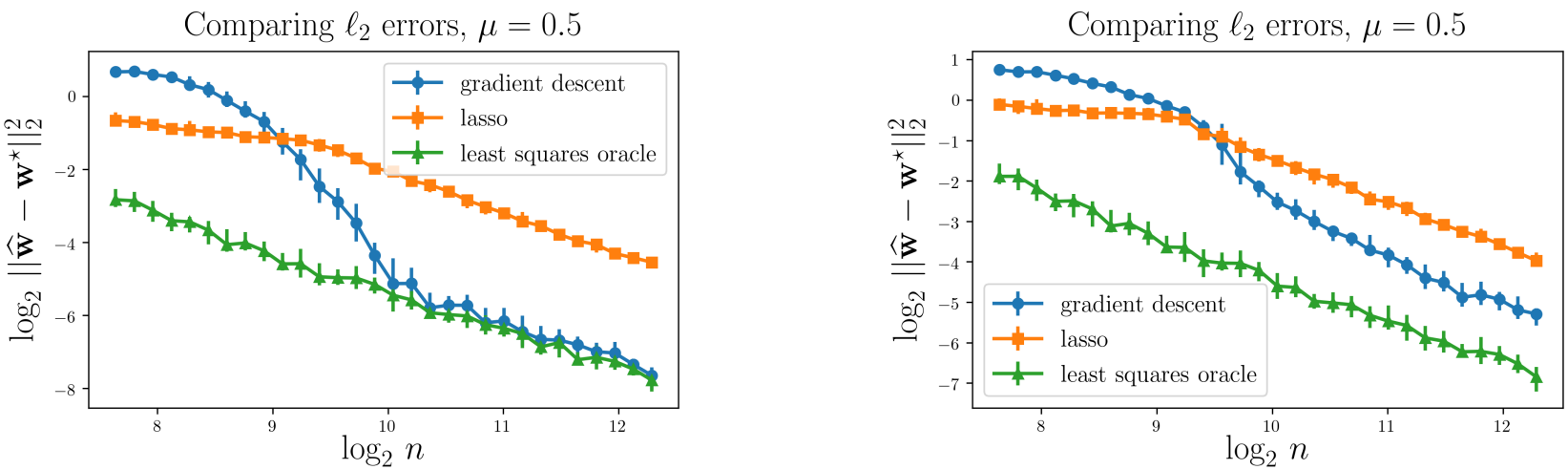


Figure: i.i.d. Gaussian ensembles, covariance matrices $(1-\mu)\mathbf{I} + \mu\mathbf{1}\mathbf{1}^\mathsf{T}$ for $\mu = 0$ and 0.5. For $\mu = 0.5$ the RIP fails but RE condition holds w.h.p. Our method achieves the fast rates and eventually outperforms the lasso even when we violate the RIP assumption

References I

A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010.

A. Ali, J. Z. Kolter, and R. J. Tibshirani. A continuous-time view of early stopping for least squares regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1370–1378, 2019.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

P. Bühlmann and B. Yu. Boosting with the l2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

J. Friedman and B. E. Popescu. Gradient directed regularization. *Technical report*, 2004.

References II

U. Ghai, E. Hazan, and Y. Singer. Exponentiated gradient meets gradient descent. *arXiv preprint arXiv:1902.01903*, 2019.

A. Goldenshluger and A. Tsybakov. Adaptive prediction and estimation in linear regression with infinitely many parameters. *The Annals of Statistics*, 29(6):1601–1619, 2001.

S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.

S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1827–1836, 2018.

T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27 (2):83–85, 2001.

P. D. Hoff. Lasso, fractional norm and structured sparse estimation using a hadamard product parametrization. *Computational Statistics & Data Analysis*, 115:186–198, 2017.

References III

J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *information and computation*, 132(1):1–63, 1997.

Y. Li, T. Ma, and H. Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47, 2018.

T. Liang, A. Rakhlin, and K. Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285, 2015.

G. Raskutti, M. J. Wainwright, and B. Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

S. Rosset, J. Zhu, and T. Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5(Aug):941–973, 2004.

A. Suggala, A. Prasad, and P. K. Ravikumar. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems*, pages 10608–10619, 2018.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

References IV

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

T. Zhang and B. Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, 2005.

Y. Zhang, M. J. Wainwright, and M. I. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948, 2014.

P. Zhao, Y. Yang, and Q.-C. He. Implicit regularization via hadamard product over-parametrization in high-dimensional linear regression. *arXiv preprint arXiv:1903.09367*, 2019.