Fast global convergence of gradient methods for high-dimensional statistical recovery

Alekh Agarwal Sahand Negahban Martin Wainwright
UC Berkeley



High-dimensional statistics setup

- **Model Class:** Parameter space $\Omega \subseteq \mathbb{R}^d$, associated family of distributions $\{\mathbb{P}_\theta \mid \theta \in \Omega\}$.
- **Data:** Samples $Z_1^n = (x_i, y_i),\ i = 1, \ldots, n$ *i.i.d.* from $\mathbb{P}_{\theta^*}$, $d \gg n$.
- **Estimation:** Minimize regularized empirical risk:

$$\hat{\theta} \in \arg \min_{\mathcal{R}(\theta) \leq \rho} \left\{ \mathcal{L}(\theta; Z_1^n) \right\}.$$

- **Statistically**, $\widehat{\theta}$ consistent under certain assumptions even when $d \gg n$.

High-dimensional statistics setup

- **Model Class:** Parameter space $\Omega \subseteq \mathbb{R}^d$, associated family of distributions $\{\mathbb{P}_\theta \mid \theta \in \Omega\}$.
- **Data:** Samples $Z_1^n = (x_i, y_i),\ i = 1, \ldots, n$ *i.i.d.* from $\mathbb{P}_{\theta^*}$, $d \gg n$.
- **Estimation:** Minimize regularized empirical risk:

$$\hat{\theta} \in \arg \min_{\mathcal{R}(\theta) \leq \rho} \left\{ \mathcal{L}(\theta; Z_1^n) \right\}.$$

- **Statistically**, $\widehat{\theta}$ consistent under certain assumptions even when $d \gg n$.
- **Computationally**, optimization procedures can be quite slow in high dimensions.

High-dimensional statistics setup

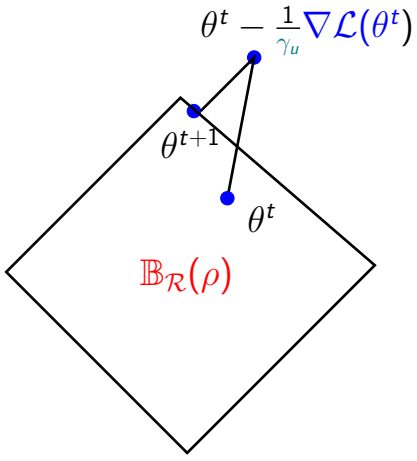
- **Model Class:** Parameter space $\Omega \subseteq \mathbb{R}^d$, associated family of distributions $\{\mathbb{P}_\theta \mid \theta \in \Omega\}$.
- **Data:** Samples $Z_1^n = (x_i, y_i),\ i = 1, \ldots, n$ *i.i.d.* from $\mathbb{P}_{\theta^*}$, $d \gg n$.
- **Estimation:** Minimize regularized empirical risk:

$$\hat{\theta} \in \arg \min_{\mathcal{R}(\theta) \leq \rho} \left\{ \mathcal{L}(\theta; Z_1^n) \right\}.$$

- **Statistically**, $\widehat{\theta}$ consistent under certain assumptions even when $d \gg n$.
- **Computationally**, optimization procedures can be quite slow in high dimensions.

Can optimization for $\widehat{\theta}$ benefit from similar assumptions useful in statistical analysis?

Projected Gradient Descent in high-dimensions

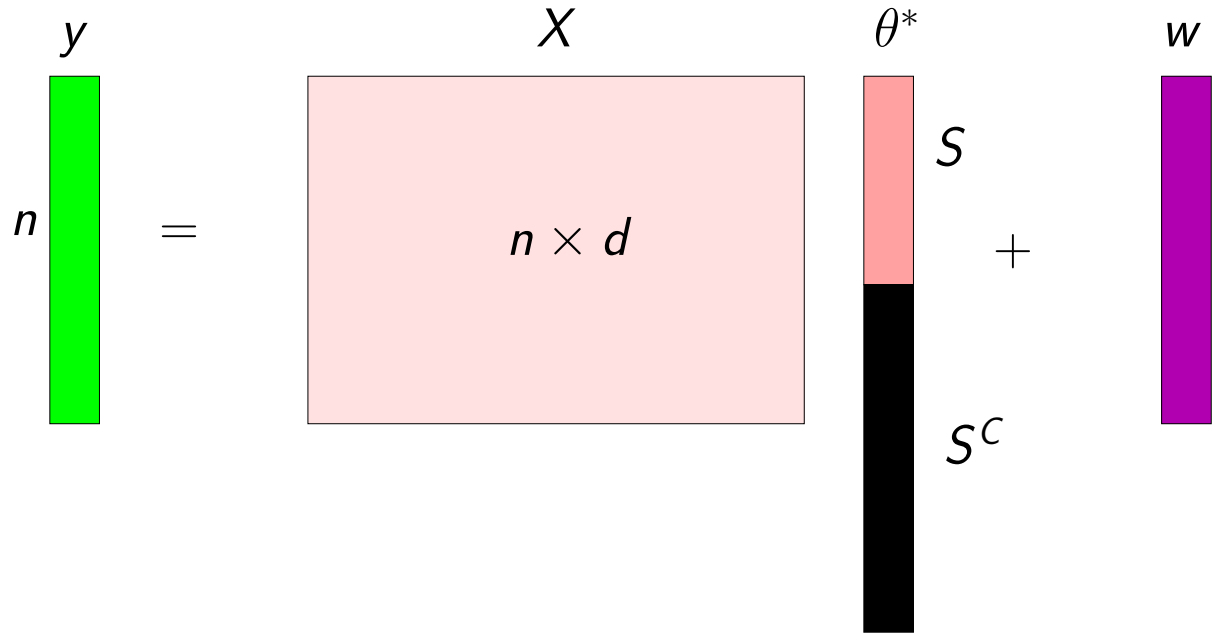


- Iterate:

$$\theta^{t+1} = \Pi_{\mathbb{B}_{\mathcal{R}}(\rho)}\left\{\theta^t - \frac{1}{\gamma_u}\nabla\mathcal{L}(\theta^t)\right\}.$$

- $\mathbb{B}_{\mathcal{R}}(\rho) = \{\theta \mid \mathcal{R}(\theta) \leq \rho\}$.
- Projected gradient descent with constant stepsize.
- γ_u depends on smoothness of $\mathcal{L}$.
- $\mathcal{L}$ **smooth**: sublinear convergence.
- $\mathcal{L}$ **smooth** and **strongly convex**: linear convergence.

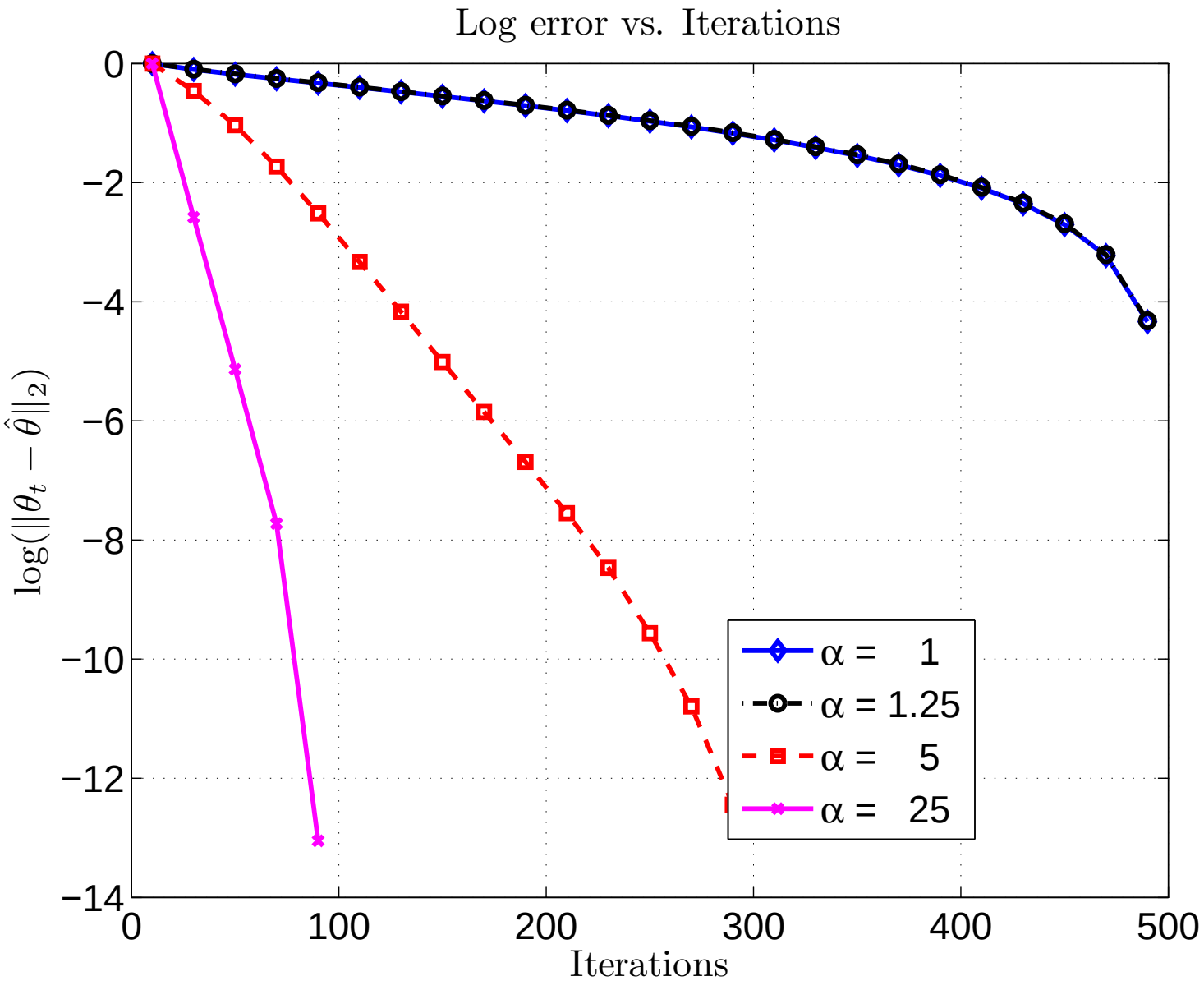
Sparse linear regression example



- n *i.i.d.* samples: $y_i = x_i^T \theta^* + w_i$.
- θ^* has at most s non-zero entries.
- Lasso program:

$$\hat{\theta} \in \arg \min_{\|\theta\|_1 \leq \rho} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 \right\}.$$

Globally linear rates obtained in pratice



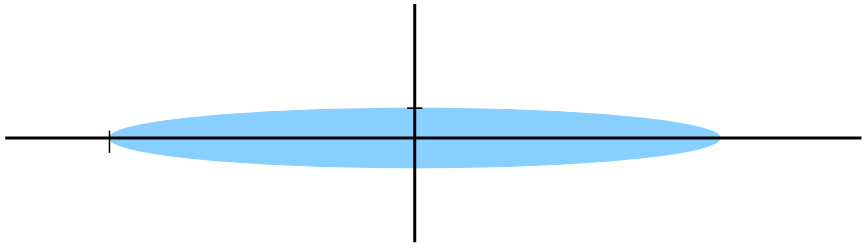
$n = \alpha s \log d.$

Some past theoretical work

- Sublinear convergence of Nesterov's methods under **smoothness**(Beck & Teboulle, 2009; Becker et. al., 2009; Ji & Ye, 2009).
- Local linear convergence rates using **RIP** (Hale et. al., 2008).
- Finite or linear convergence up to *noise variance precision* for compressed sensing (Tropp & Gilbert, 2007; Garg & Khandekar, 2009).
- Global linear convergence using **RIP** and **smoothness** (Bredies & Lorenz, 2008).

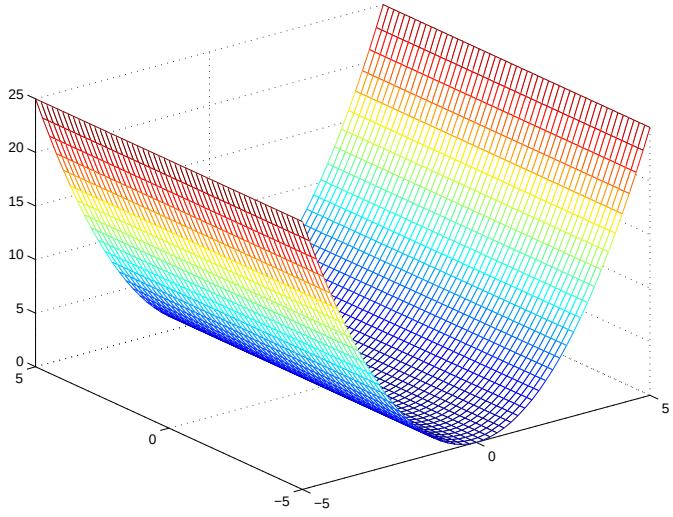
No smoothness or curvature in high dimensions

$$\mathcal{L}(\theta; Z_1^n) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\theta)^2, \quad x_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_{d\times d}).$$



No Smoothness:

$\lambda_{\max}(X^TX/n) = \Theta\left(1 + \sqrt{\frac{d}{n}}\right)$ with high probability.



No Strong Convexity:

$\lambda_{\min}(X^TX/n) = 0$. Hessian rank-deficient when $d > n$.

Restricted Strong Convexity and Smoothness

Definition

$\mathcal{L}$ satisfies RSC condition with $(\gamma_\ell, \kappa_\ell, \delta)$ if for all $\theta, \theta' \in \mathbb{B}_{\mathcal{R}}(\rho)$:

$$\mathcal{L}(\theta') - \underbrace{\left\{\mathcal{L}(\theta) + \langle \nabla \mathcal{L}(\theta),\, \theta' - \theta \rangle\right\}}_{\text{First-order Taylor approx.}} \geq \underbrace{\frac{\gamma_\ell}{2}\|\theta' - \theta\|_2^2}_{\text{Lower curvature}} - \kappa_\ell \underbrace{\delta^2}_{\text{Tolerance}}$$

- Same as strong convexity apart from the $\kappa_\ell \delta^2$ term.
- Lower-bounded curvature when $\|\theta^t - \widehat{\theta}\|_2 \gg \delta$.

Restricted Strong Convexity and Smoothness

Definition

$\mathcal{L}$ satisfies RSC condition with $(\gamma_\ell, \kappa_\ell, \delta)$ if for all $\theta, \theta' \in \mathbb{B}_{\mathcal{R}}(\rho)$:

$$\mathcal{L}(\theta') - \underbrace{\left\{\mathcal{L}(\theta) + \langle \nabla \mathcal{L}(\theta),\, \theta' - \theta \rangle\right\}}_{\text{First-order Taylor approx.}} \geq \underbrace{\frac{\gamma_\ell}{2}\|\theta' - \theta\|_2^2}_{\text{Lower curvature}} - \kappa_\ell \underbrace{\delta^2}_{\text{Tolerance}}$$

Definition

$\mathcal{L}$ satisfies RSM condition with $(\gamma_\ell, \kappa_\ell, \delta)$ if for all $\theta, \theta' \in \mathbb{B}_{\mathcal{R}}(\rho)$:

$$\mathcal{L}(\theta') - \underbrace{\left\{\mathcal{L}(\theta) + \langle \nabla \mathcal{L}(\theta),\, \theta' - \theta \rangle\right\}}_{\text{First-order Taylor approx.}} \leq \underbrace{\frac{\gamma_u}{2}\|\theta' - \theta\|_2^2}_{\text{Upper Curvature}} + \kappa_u \underbrace{\delta^2}_{\text{Tolerance}}$$

Linear convergence of gradient descent

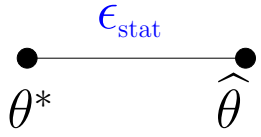
Theorem

Suppose $\mathcal{L}$ satisfies (RSC) and (RSM) with constants $\gamma_\ell, \gamma_u, \kappa_\ell, \kappa_u$. Then

$$\|\theta^t - \widehat{\theta}\|_2^2 \leq c_0 \left(1 - \frac{\gamma_\ell}{4\gamma_u}\right)^t + c_1 \delta^2.$$

- Geometric convergence with contraction factor $1 - \frac{\gamma_\ell}{4\gamma_u}$.
- Convergence to an accuracy δ^2.
- Global convergence, holds for all iterates.

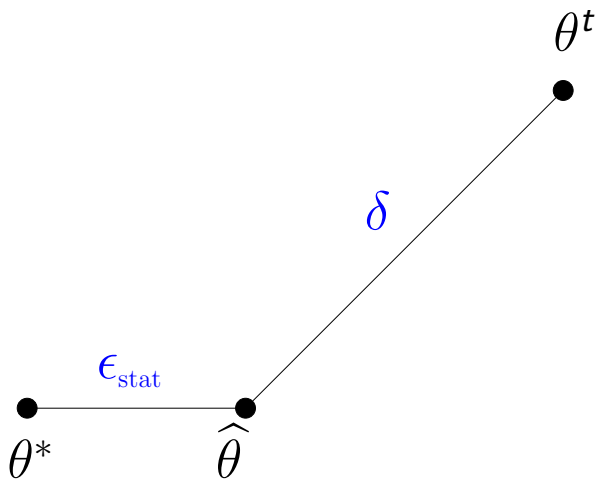
Convergence to statistical precision



$$\epsilon_{\text{stat}} := \mathbb{E}\|\widehat{\theta} - \theta^*\|_2.$$

- Aim to recover true model θ^*.
- Define $\epsilon_{\text{stat}} := \mathbb{E}\|\widehat{\theta} - \theta^*\|_2$.
- Sufficient to guarantee $\delta = \Theta(\epsilon_{\text{stat}})$.
- Need to establish (RSC), (RSM) with $\delta = \epsilon_{\text{stat}}$ by choice of ρ.

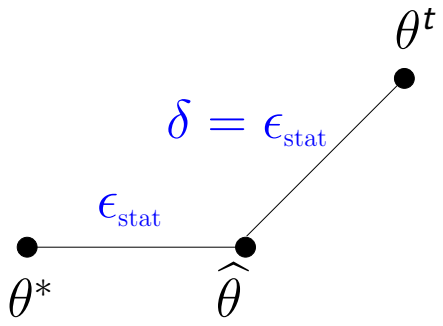
Convergence to statistical precision



θ^t is a bad estimator if $\delta \gg \epsilon_{\text{stat}}$.

- Aim to recover true model θ^*.
- Define $\epsilon_{\text{stat}} := \mathbb{E}\|\widehat{\theta} - \theta^*\|_2$.
- Sufficient to guarantee $\delta = \Theta(\epsilon_{\text{stat}})$.
- Need to establish (RSC), (RSM) with $\delta = \epsilon_{\text{stat}}$ by choice of ρ.

Convergence to statistical precision



θ^t as good as $\widehat{\theta}$ if $\delta = \Theta(\epsilon_{\text{stat}})$.

- Aim to recover true model θ^*.
- Define $\epsilon_{\text{stat}} := \mathbb{E}\|\widehat{\theta} - \theta^*\|_2$.
- Sufficient to guarantee $\delta = \Theta(\epsilon_{\text{stat}})$.
- Need to establish (RSC), (RSM) with $\delta = \epsilon_{\text{stat}}$ by choice of ρ.

Exact sparse linear regression

- $x_i \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$, $y_i = x_i^T \theta^* + w_i$.
- θ^* is s-sparse.
- Set $\rho = \|\theta^*\|_1$.
- (RSC) and (RSM) hold w.h.p. with $\delta^2 = \frac{s \log d}{n}$.

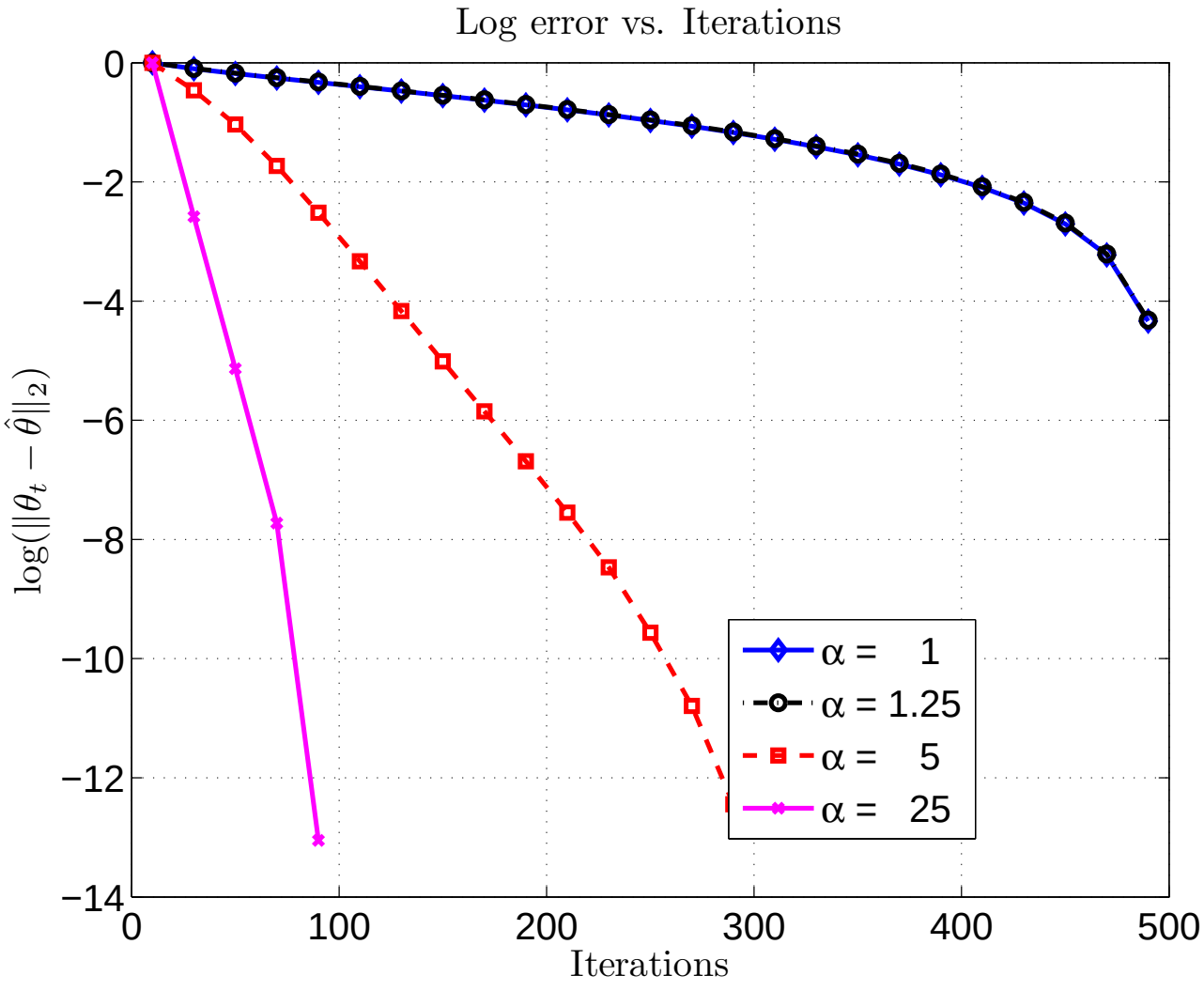
Corollary

There are universal constants c_0, c_1, c_2 such that w.h.p.

$$\|\theta^t - \widehat{\theta}\|_2^2 \leq c_0 \left(1 - \frac{c_2}{\kappa(\Sigma)}\right)^t + c_1 \underbrace{\frac{s \log d}{n}}_{\epsilon^2_{stat}}.$$

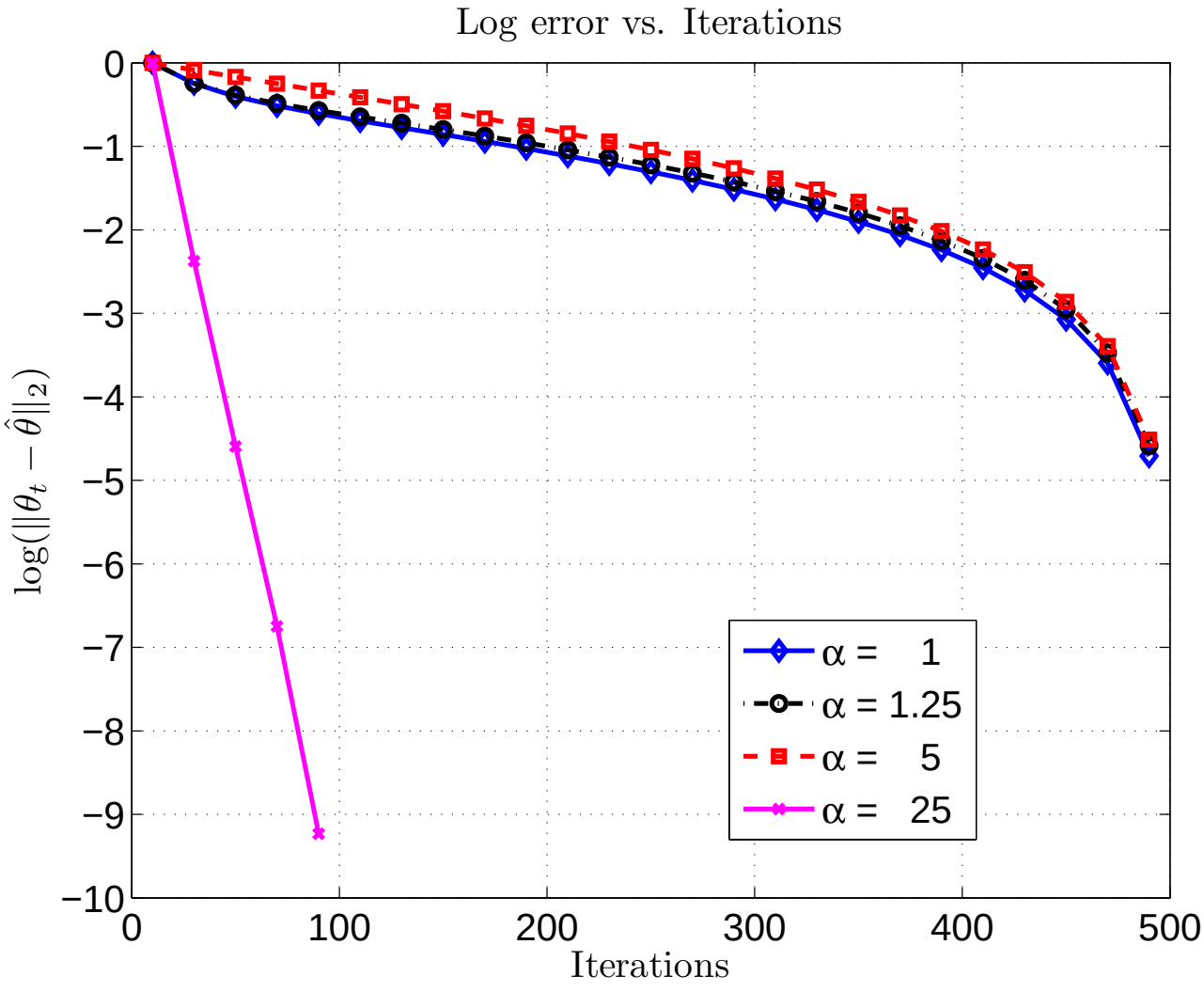
- $\kappa(\Sigma)$ is condition number of Σ.
- Result extends to approximate sparsity.

Experimental results: Sparse linear regression



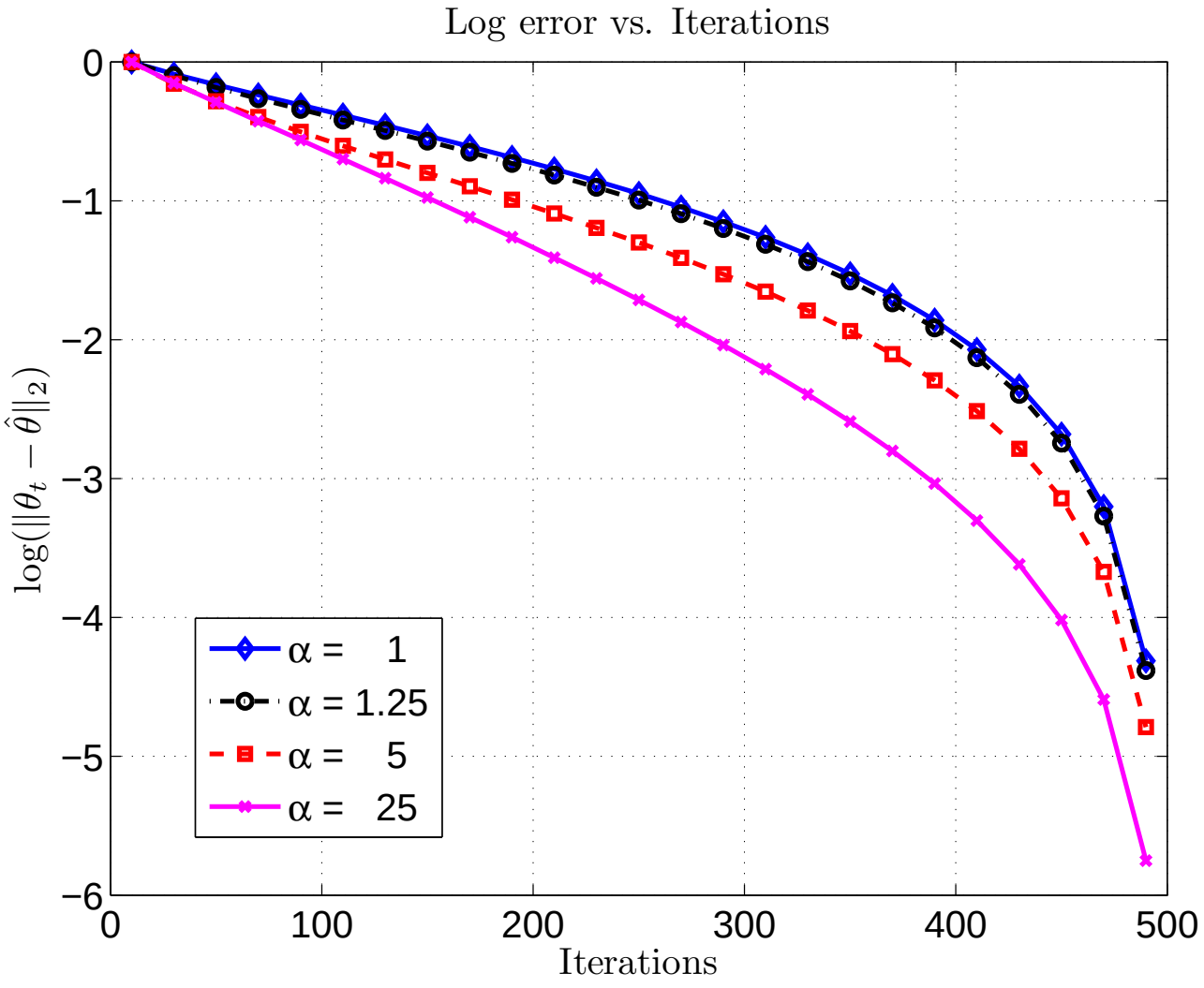
Exact sparsity, RIP ($n = \alpha s \log d$).

Experimental results: Sparse linear regression



Approximate sparsity, RIP ($n = \alpha s \log d$).

Experimental results: Sparse linear regression

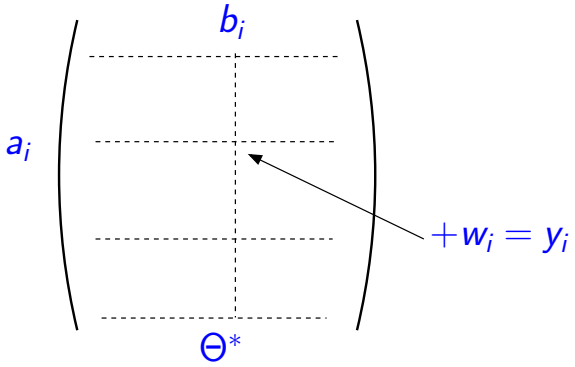


Exact sparsity, Highly correlated design ($n = \alpha s \log d$).

Low-rank matrix completion

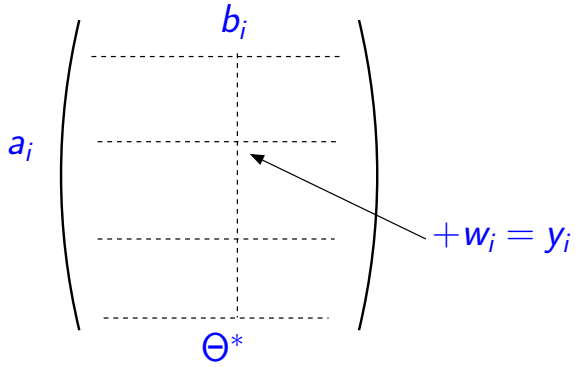
- $y_i = \Theta^*_{a(i),b(i)} + w_i$, $(a(i), b(i))$ picked randomly.

- Assume $\Theta^* \in \mathbb{R}^{m \times m}$ has rank at most r.
- $\mathcal{L}(\Theta; Z_1^n) = \frac{1}{n} \sum_{i=1}^n (y_i - \Theta_{a(i),b(i)})^2$
- $\mathcal{R}(\Theta) = \|\sigma(\Theta)\|_1$, set $\rho = \|\sigma(\Theta^*)\|_1$.



Low-rank matrix completion

- $y_i = \Theta^*_{a(i),b(i)} + w_i$, $(a(i), b(i))$ picked randomly.
- Assume $\Theta^* \in \mathbb{R}^{m \times m}$ has rank at most r.
- $\mathcal{L}(\Theta; Z_1^n) = \frac{1}{n} \sum_{i=1}^n (y_i - \Theta_{a(i),b(i)})^2$
- $\mathcal{R}(\Theta) = \|\sigma(\Theta)\|_1$, set $\rho = \|\sigma(\Theta^*)\|_1$.

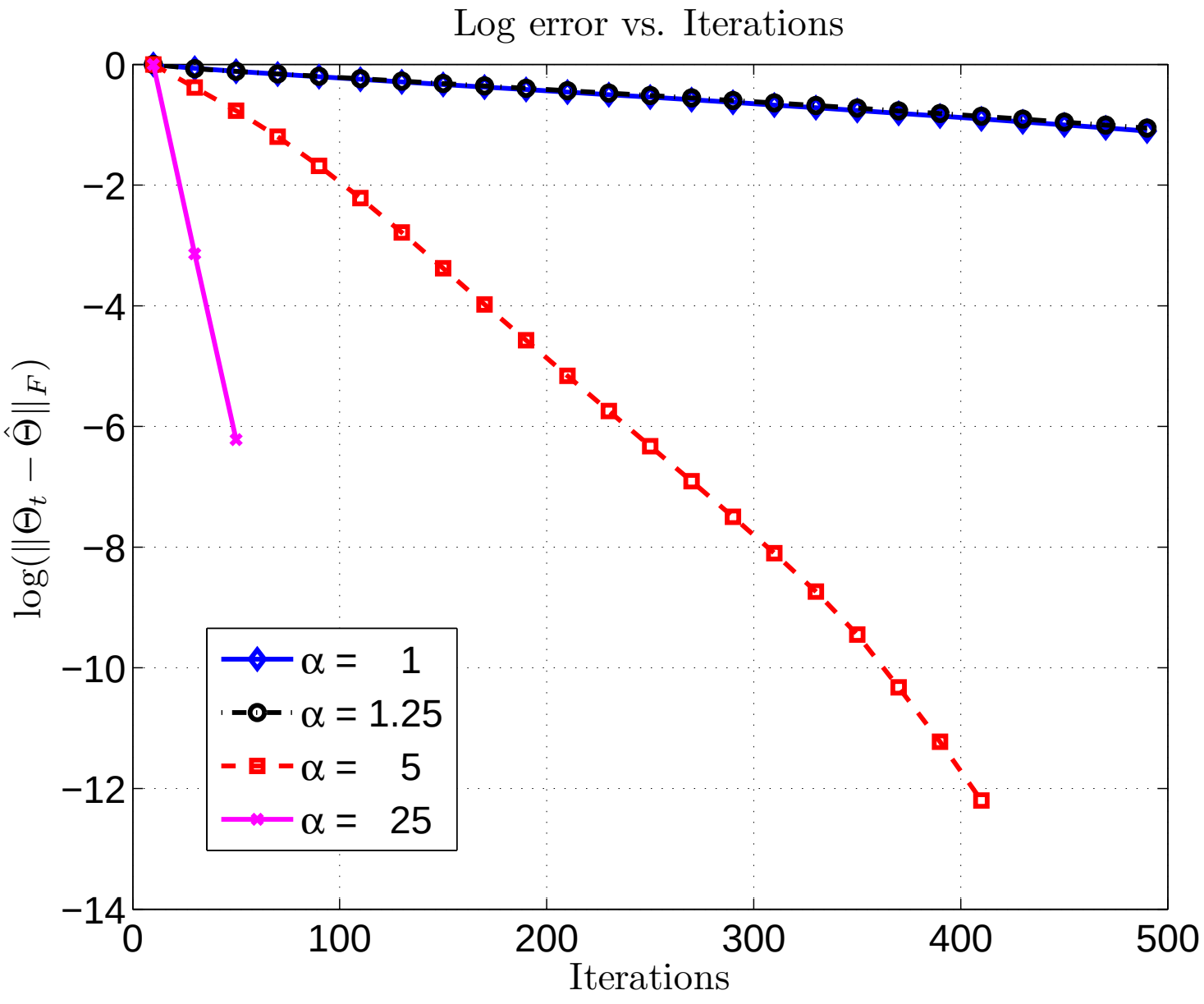


Corollary

There are universal constants c_0, c_1, c_2 and $\nu \in (0, 1)$ such that w.h.p.

$$\|\Theta^t - \widehat{\Theta}\|_F^2 \leq c_0 \nu^t + c_1 \underbrace{\frac{rm \log m}{n}}_{\epsilon^2_{stat}}.$$

Experimental results : Low rank completion



$n = \alpha r m \log m$

Conclusions

- Prove global linear convergence of gradient descent under (RSC), (RSM).
- Show the assumptions hold for several interesting models.
- First global linear convergence result for high dimensional regression, matrix completion etc.
- Convergence only up to statistical precision of the underlying model.
- Exploit good statistical properties for fast optimization.

Extensions

- Results extend to decomposable regularizers (Negahban et al, 2009).
  - Group sparsity with non-overlapping groups.
  - Multi-task learning.
  - Generalized-linear models (e.g. sparse logistic regression).
- Also analyze a Lagrangian version.

Thank You