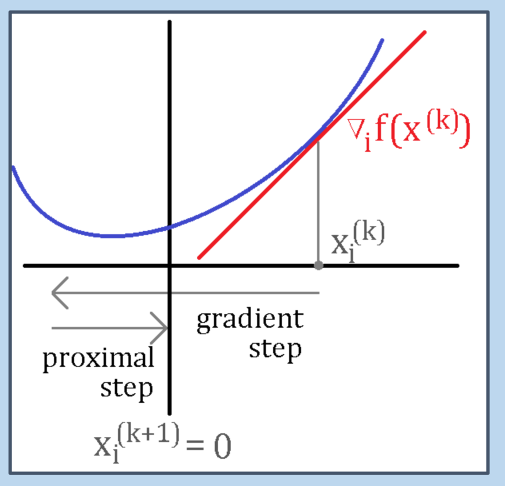
Are we there yet?
Manifold ID of gradient-related proximal methods

Yifan Sun, Halyun Jeong, Julie Nutini, Mark Schmidt
University of British Columbia, Vancouver

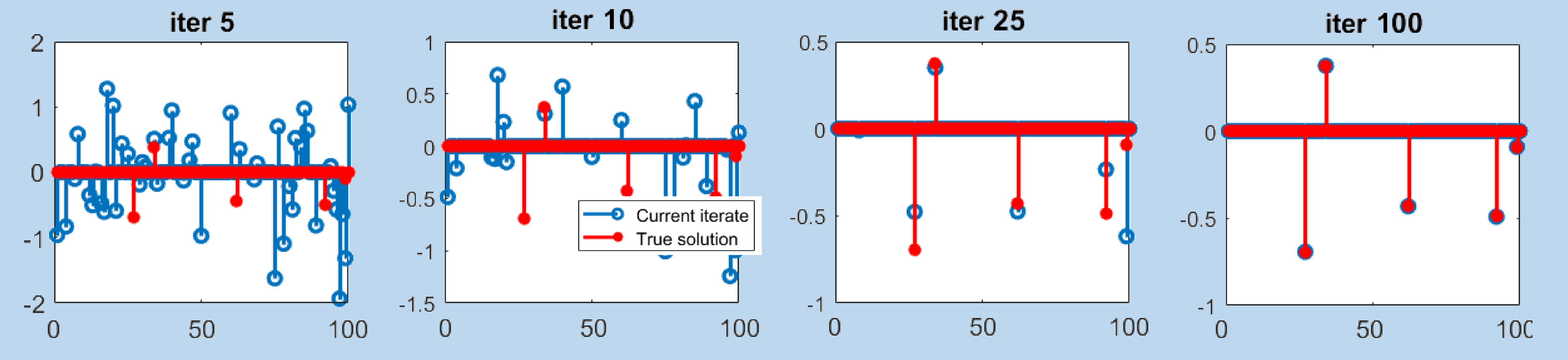
Observation

Proximal methods often **snap to** the solution manifold quickly.

Can we **predict** when this happens?

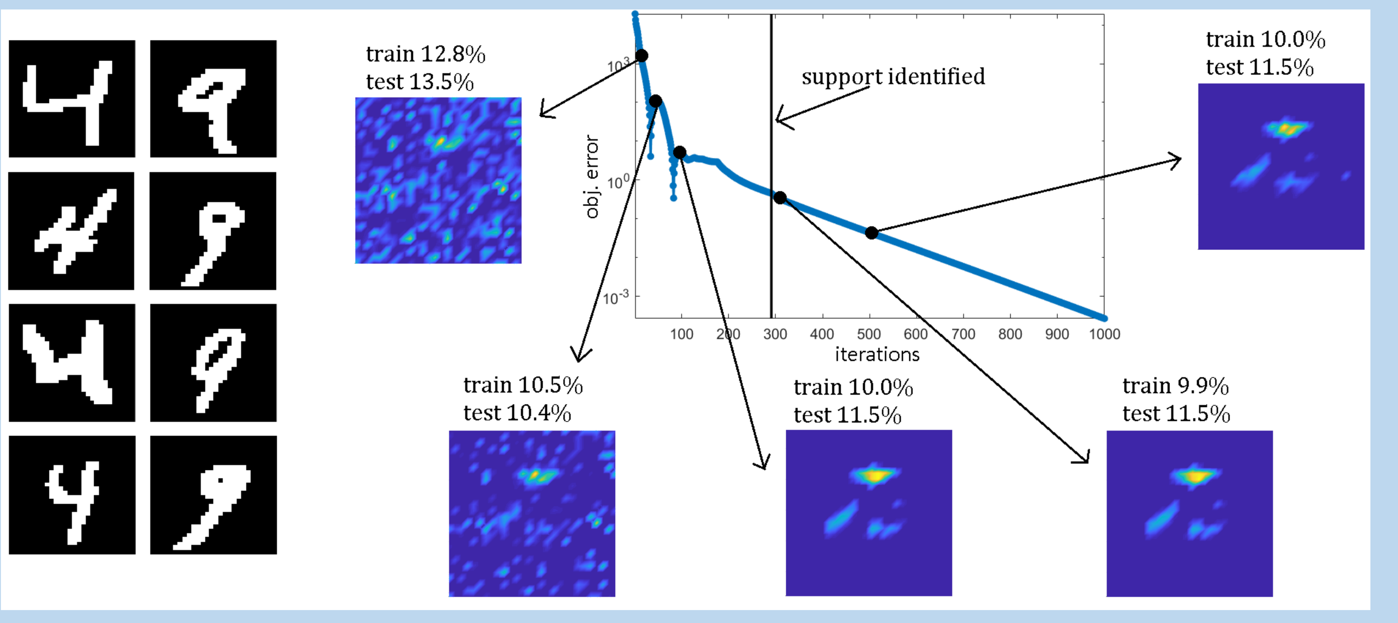


One-step "snapping" action of the prox. grad method.



Sparse LASSO
In 25 iterations, true support identified.
However, the iterate values are still converging.

Sparse Log. Reg. for 4/9 disambiguation
Sparsity, train, and test error converges in very few iterations, but objective error keeps decreasing.



Motivation

Reason 1: Learning sparsity pattern often enough
- Feature selection
- Identifying correlations between variables
- Identifying support vectors

Reason 2: Solving over reduced support may be easy
- Smaller problem → can use more powerful solver (e.g. Newton's)
- Better conditioned Hessian → faster convergence

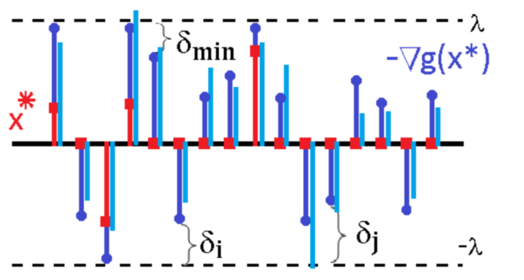
Contribution

We provide a **simple** and **geometrically intuitive** framework to **easily** compute the **manifold ID rates** for **proximal methods**.

Wiggle room lemma

Define

$$\delta_i = \max \{\delta : -\nabla g(x^*)_i + d \in \partial h_i(x^*),\ \forall |d| \le \delta\}$$



If, for all $i \in \mathcal{Z}$,

$$\underbrace{\left|\left(\frac{1}{t^{(k)}} H^{(k)}(z^{(k)} - x^*) + \nabla g(x^*)\right)_i\right|}_{\omega^{(k)}} \le \delta_i$$

Then $x^{(k+1)} \in \mathcal{M}$.

method	rate	rate if strongly convex
Prox grad	$(1/t + L)\epsilon_x \le \delta_{\min}$	$O(\log(1/\delta_{\min}))$
Acc prox grad	$(1/t + L)\epsilon_x \le \delta_{\min}$	$O(\log(1/\delta_{\min}))$
Prox DRS/ADMM	$(2/t + 2L)\epsilon_x \le \delta_{\min}$	$O(1/\delta_{\min}^2)$
Prox Newton	$2L\epsilon_x \le \delta_{\min}$	$O(\log\log(\delta_{\min}))$
Prox Quasi Newton	$(L + L_H)\epsilon_x \le \delta_{\min}$	$O(\log(1/\delta_{\min}))$
Prox SGD	None	None
Prox SAGA / SVRG*	$\epsilon_x/t + \epsilon_g \le \delta_{\min}$	$O(\log(1/\delta_{\min}))$
Prox RDA*	$\epsilon_g + B/(kt) \le \delta_{\min}$	$O(1/\delta_{\min}^4)$

Table: Rates for manifold identification.
$\epsilon_x = \|x - x^*\|_2,\ \epsilon_g = \|\nabla g(x) - \nabla g(x^*)\|_2,\ \delta_{\min} = \min_{i\in\mathcal{Z}} \delta_i$.

The optimality condition for a nonsmooth problem has "built in" **wiggle room.**

Proximal methods ensure that, near optimality, the error **snaps within this wiggle room**.

This gives a **framework** to **quickly** compute **many** manifold ID rates.

How to derive rates

- Prox gradient descent

$$\max_i |\omega_i^{(k)}| \le \frac{1}{t}\underbrace{\|x^{(k)} - x^*\|_2}_{\text{error in var.}\to 0} + \underbrace{\|\nabla g(x^*) - \nabla g(x^{(k)})\|_2}_{\text{error in grad}\to 0}$$

Rate if $g(x)$ is strongly convex:

$$\bar{k} = O\left(\log\left(\frac{1/t+L}{\delta_{\min}}\right)\right) \text{ [NSH '17]}$$

- Prox stochastic gradient descent

$$\omega^{(k)} = \underbrace{\frac{1}{t^{(k)}}(x^{(k)} - x^*)}_{\text{scaled error in var.}\overset{?}{\to} 0} + \underbrace{\nabla g(x^*) - \nabla g(x^{(k)})}_{\text{error in grad}\not\to 0}$$

Does not identify manifold!

Mathematical Setup

Problem class

$$\min_x f(x) := \underbrace{g(x)}_{\substack{L\text{-smooth}\\ \text{loss}\\ \text{function}}} + \underbrace{h(x)}_{\substack{\text{nonsmooth}\\ \text{separable}\\ \text{regularizer}}}$$

- x^* is a unique minimizer
- $h(x)$: $\|x\|_1$, elementwise constraints, hinge loss

Manifolds and active sets

Active set

$\mathcal{Z} = \{i : \partial h(x_i^*)$ is not a singleton $\}$

- $h(x) = \|x\|_1 \to \mathcal{Z} = \{i : x_i^* = 0\}$
- $l \le x \le u \to \mathcal{Z} = \{i : x_i = u_i$ or $x_i = l_i\}$

Solution manifold

$$\mathcal{M} = \{x : x_i = x_i^*,\ \forall i \in \mathcal{Z}\}$$

A method $x^{(k)} \to x^*$ *identifies the manifold* at $\bar{k}$ if $\forall k > \bar{k},\ x^{(k)} \in \mathcal{M}$.

Proximal methods

Consider methods with iteration updates $x^{(k)} \to x^*$ via

$$x^{(k+1)} = \mathbf{prox}_{t^{(k)}h}^{H^{(k)}}(z^{(k)})$$

- $z^{(k)}$ depends on past $x^{(1)}, ..., x^{(k)}$
- $\mathbf{prox}_h^H(z) := \underset{x}{\mathbf{argmin}}\ h(x) + \frac{1}{2}(x - z)^T H(x - z)$
- Examples: Prox. grad. descent, FISTA, DRS, ADMM

How much wiggle room

- Manifold ID rates depend on $\delta_{\min}$.
  → but need x^* to compute $\delta_{\min}$!
- We **can** empirically connect it to problem parameters
  → e.g. regularization weight, ground truth sparsity
- **Open question**: Can we infer it from knowledge of the data distribution of our problem?

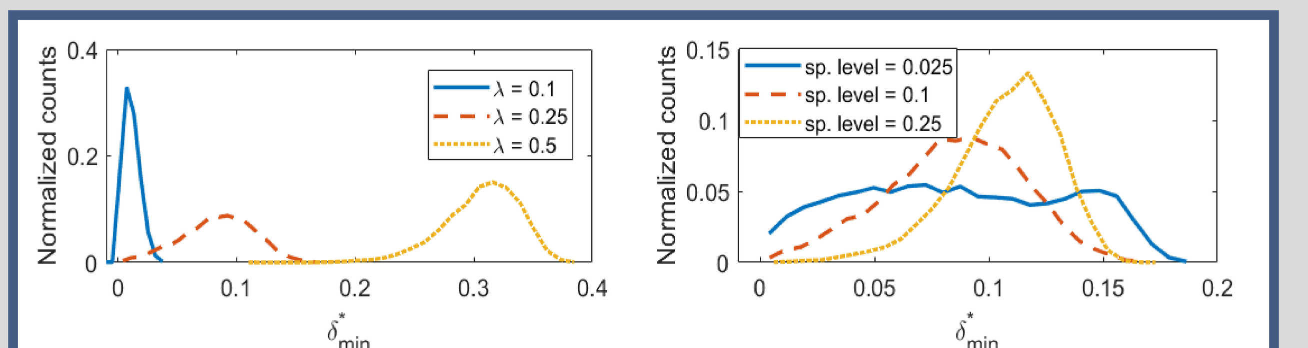


Figure: Histogram of $\delta_{\min} = \min_{i\in\mathcal{Z}} \delta_i$ for randomly generated logistic regression problems.

References

- **Manifold ID:** Bertsekas (1974), Dennis and Moré (1974), Gafni and Bertsekas (1984), Dunn (1987), Burke and Moré (1988), Wright (1993), Ko et al. (1994), Hare and Lewis (2004), Daniilidis, Sagastizábal, and Solodov (2009)
- **Prox grad:** Johnstone and Moulin (2015), Liang, Fadili, and Peyré (2017), Nutini, Schmidt, and Hare, (2017)
- **Prox DRS / ADMM:** Liang, Fadili, and Peyré (2016)
- **Prox SAGA / SVRG:** Poon, Liang, and Schönlieb (2018)
- **Prox RDA:** Lee and Wright (2012), Duchi and Ruan (2016)