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?

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.

Mathematical Setup

**Problem class**

\[ \min f(x) := g(x) + h(x) \]

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

**Manifolds and active sets**

**Active set**

\[ Z = \{ i : h(x_i^*) \text{ is not a singleton} \} \]

- \( h(x) = \|x\|_1 \rightarrow Z = \{ i : x_i = 0 \} \)
- \( l \leq u \leq Z \rightarrow \{ i : x_i = u_i \text{ or } x_i = l_i \} \)

**Solution manifold**

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

A method \( x^{(k)} \rightarrow x^* \) identifies the manifold at \( k \) if

\[ \forall k > k, x^{(k)} \in \mathcal{M} \]

**Proximal methods**

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

\[ x^{(k+1)} = \text{prox}_{\phi_{H^{(k)}}}^{(k)}(x^{(k)}) \]

- \( x^{(k)} \) depends on past: \( x^{(1)}, \ldots, x^{(k)} \)
- \( \text{prox}_{\phi_{H^{(k)}}}^{(k)}(z) := \arg\min h(z) + \|z - x^{(k)}\|^2 H(z - x^{(k)}) \)
- Examples: Prox. grad. descent, FISTA, DRS, ADMM

Wiggle room lemma

Define

\[ \delta_i = \max \{ \delta : -\nabla g(x^*) + \partial h_i(x^*) \cap \{ |\delta| \leq \delta \} \} \]

If, for all \( i \in Z \),

\[ \left\| H^{(k)}(x^{(k)} - x^*) + \nabla g(x^{(k)}) \right\|_i \leq \delta_i \]

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

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**
  \[ \max |\omega^{(k)}| \leq \frac{1}{2} \left( \|x^{(k)} - x^*\|^2 + \|\nabla g(x^*) - \nabla g(x^{(k)})\|^2 \right) \]

How much wiggle room

- **Manifold ID rates depend on** \( \delta_{\min} \)
  \[ \rightarrow \text{but need } x^* \text{ to compute } \delta_{\min} \]

- **We can empirically connect it to problem parameters**
  \[ \rightarrow \text{e.g., regularization weight, ground truth sparsity} \]

- **Open question:** Can we infer it from knowledge of the data distribution of our problem?

References

- **Prox DRS / ADMM**: Liang, Fadili, and Peyré (2016)
- **Prox SAGA / SVRG**: Poon, Liang, and Schönherr (2018)
- **Prox RDA**: Lee and Wright (2012), Duchi and Ruan (2016)