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

Yifan Sun UBC Vancouver
Halyun Jeong UBC Vancouver
Julie Nutini UBC Vancouver
Mark Schmidt UBC Vancouver

Abstract

In machine learning, models that generalize better often generate outputs that lie on a low-dimensional manifold. Recently, several works have separately shown finite-time manifold identification by some proximal methods. In this work we provide a unified view by giving a simple condition under which any proximal method using a constant step size can achieve finite-iteration manifold detection. For several key methods (FISTA, DRS, ADMM, SVRG, SAGA, and RDA) we give an iteration bound, characterized in terms of their variable convergence rate and a problem-dependent constant that indicates problem degeneracy. For popular models, this constant is related to certain data assumptions, which gives intuition as to when lower active set complexity may be expected in practice.

1 INTRODUCTION

Consider the classic machine learning problem

$$\min_{x\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(a_i^T x; b_i) + \lambda\mathcal{R}(x), \tag{1}$$

where $\mathcal{L}$ is the loss of a machine learning model (such as squared or logistic loss) and $\mathcal{R}(x)$ is a separable regularizer that promotes x to lie in a low-dimensional manifold. Specifically, the manifold $\mathcal{M}$ is parametrized by an *active set* of indices $\mathcal{Z}$ where

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

For example, when $\mathcal{R}(x)$ is the ℓ_1 norm, then $\mathcal{M}$ is the set of vectors x where $x_i = 0$ whenever $x_i^* = 0$.

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

For simplicity, in this work we restrict our attention to cases where $\mathcal{R}(x)$ is element-wise separable and convex. This includes the ℓ_1 norm and element-wise constraints, which arise in support vector machines (SVMs). However, all our results can be extended to other norms by considering *atomic sparsity* and a generalized definition of the active set,

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

where $B_i x$ is the projection of x onto an orthogonal subspace B_i. This includes the total variation norm for smooth vectors and the group norm for group sparsity.

1.1 Related work

The early works on active set identification (Dunn, 1987; Burke and Moré, 1988; Wright, 1993; Gafni and Bertsekas, 1984) focus on constrained optimization problems, where an inequality constraint is active if satisfied with equality. Extensions to nonsmooth functions have also been explored (Hare and Lewis, 2004; Nutini et al., 2017b). Notable works cover sequential quadratic programming (Burke and Moré, 1988), proximal gradient method (Dunn, 1987), bundle methods (Daniilidis et al., 2009), prox-SVRG (Poon et al., 2018), regularized dual averaging (Lee and Wright, 2012) and variants (Duchi and Ruan, 2016), and block coordinate methods (Tseng and Yun, 2009; Hare, 2011; De Santis et al., 2016; She and Schmidt, 2017). In general, all these works show that finite-time manifold identification is possible if the final solution is non-degenerate. More recently, nonasymptotic active set complexity bounds have become of interest, in particular for the proximal gradient method (Nutini et al., 2017b), ADMM (Liang et al., 2017), and block coordinate methods (Nutini et al., 2017a).

Understanding manifold identification properties has also led to specialized methods for improved performance. For example, in two-stage methods, a first-order proximal method is first used to identify the manifold, and a heavier method (such as Newton's method, or a direct solve) is used to find the true

solution in a much reduced parameter space (Bertsekas, 1974; Ko et al., 1994; Daniilidis et al., 2009). Similarly, a two-metric method is proposed by Gafni and Bertsekas (1984), and uses a mixture of Newton's method on $i \notin \mathcal{Z}$ and the proximal gradient method for $i \in \mathcal{Z}$ in the intermediate iterates, achieving superior convergence results. And, Wright (2012) proposes a linearized proximal coordinate-descent type algorithm that is accelerated by performing Newton steps whenever the function is smooth over the active set currently identified. Active set properties can also be used for parameter selection (Hastie et al., 2004).

Contributions We provide a unified analysis bounding the number of iterations needed for manifold identification, or the *active set complexity*, for all proximal gradient-type methods. The analysis follows 3 main steps:

1. quantify the "amount of nondegeneracy" in the problem,
2. bound the amount of variable error allowed in order for correct manifold identification, which is directly correlated with the "amount of degeneracy", and
3. combine with known variable convergence bounds to compute the active set complexity. (See also Tables 1 and 2 in the Appendix.)

Although the idea of bounding the variable error with the distance to the manifold is not new (Lewis and Wright, 2011) how this distance and variable error are related for various methods is made explicit in this work. In particular, our analysis shows a direct relationship amongst three key quantities:

- $\epsilon(k)$, a monotonic sequence upper bounding variable error;
- $\bar{k}$ a constant such that for all $k > \bar{k}$, $x^{(k)}$ is on the manifold; and
- $\delta_{\min}$ quantifying how "close to degenerate" the problem is[1].

These three key quantities are unified under a single lemma, which we term the "Wiggle Room Lemma"; subsequently, manifold identification rates for new proximal methods can be quickly and easily derived. In contrast, previously they have been often derived from scratch; additionally, in some cases our derived bounds are tighter than those previously derived.

2 ACTIVE SET MANIFOLDS

2.1 Problem statement

Generalizing (1), we consider the problem class

$$\min_{x \in \mathbb{R}^n} \quad f(x) := g(x) + h(x), \tag{2}$$

where g is convex and L-smooth:

$$\|\nabla g(x) - \nabla g(y)\|_2 \leq L\|x - y\|_2, \ \forall \ x, y.$$

The function $h(x)$ is generally a nonsmooth, convex regularizer. We assume the optimal minimizer x^* is unique. To simplify the analysis, we consider only $h(x) = \sum_i h_i(x_i)$ separable. These assumptions are reasonable and often appear in practice.

The first-order optimality condition of (2) is

$$0 \in \nabla g(x^*) + \partial h(x^*), \tag{3}$$

where $\partial h(x)$ is the subdifferential of h at x:(Rockafellar, 2015)

$$\partial h(x) = \{z : h(x) - h(y) \leq z^T(x - y)\}, \ \forall \ y.$$

2.2 Active sets

We characterize the manifold via the *active set*.

Definition 1. (Nutini et al., 2017b) For problems of form (2), we define the *active set* of indices as

$$\mathcal{Z} = \{i : \partial h_i(x_i^*) \text{ is not a singleton}\}, \tag{4}$$

where x^* is the optimum of (2).

Essentially, Def. 1 says that if $i \in \mathcal{Z}$ then $h_i(x_i)$ is nonsmooth at x^*. When $i \in \mathcal{Z}$, the solution to (2) is often trivial. For example, when $h(x) = \|x\|_1$, $x_i^* = 0$ for all $i \in \mathcal{Z}$; and when $h(x)$ is an indicator for an elementwise constraint $x \leq c$, then $x_i^* = c_i$ for all $i \in \mathcal{Z}$. In this case, if we know $\mathcal{Z}$, then the optimization reduces to a smooth unconstrained problem over $\{x_i\}_{i \notin \mathcal{Z}}$.

Definition 2. (Nutini et al., 2017b) For any x, define $\delta_i(x)$ for $i = 1, \ldots, n$ as the maximum scalar d where

$$[-(\nabla g(x))_i - d, -(\nabla g(x))_i + d] \subseteq \partial h_i(x). \tag{5}$$

We can think of $\delta_{\min}$ as the "wiggle room" in the optimality conditions. For many methods, the existance of a strictly positive $\delta_{\min}^*$ exactly corresponds to when finite time manifold identification is possible, since it allows noisy iterates $x^{(k)}$ to partially satisfy optimality conditions, provided $x^{(k)}$ is close enough to x^*.

Definition 3 (Degeneracy)**.** Define the data-dependent constant

$$\delta_{\min} := \min_{i \in \mathcal{Z}} \delta_i(x^*).$$

We say the *problem is degenerate* when $\delta_{\min} = 0$.

In general, finite time manifold identification is *impossible to guarantee* if the problem is degenerate (Burke and Moré, 1988; Lee and Wright, 2012).

[1] $\delta_{\min}$ is commonly written as $\text{dist}(x^*, \text{rbd}(\partial h(x^*)))$, where ∂h is the subdifferential of h and 'rbd' indicates the relative boundary of a set (Rockafellar, 2015).

$\lvert c\rvert$	x^*	$g'(x^*)$	$\partial h(x^*)$	active?
< 1	0	$-c$	$\{\mathbf{sign}(x^*)\}$	no
> 1	$c - \mathbf{sign}(c)$	$-\mathbf{sign}(c)$	$[-1, 1]$	yes
$= 1$	0	$-c$	$[-1, 1]$	degenerate

Table 1: **Small example.** Solution and function properties for (6).

2.3 Example

Consider minimizing the composite scalar function $f(x) = g(x) + h(x)$ where

$$g(x) = \frac{1}{2}(x - c)^2 \quad \text{and} \quad h(x) = |x|. \tag{6}$$

In this case the true solution can be computed analytically (Table 1). Specifically, when $x^* \neq 0$, then $f = g + h$ is smooth at x^* and the active set is empty (Fig. 1 left). When $x^* = 0$, then h is nonsmooth at x^* (and so is f) and $\mathcal{Z} = \{1\}$. In this case, $\delta_{\min} = \delta_1(x^*) = 1 - |g'(x^*)| = 1 - |c|$, and is strictly positive only if $|c| < 1$ (Fig. 1 middle); otherwise, the problem is degenerate (Fig. 1 right). In practice, degeneracy for randomized data is a rare occurrence; in this example, if c is drawn from any smooth distributions, $c = \pm 1$ with probability 0.

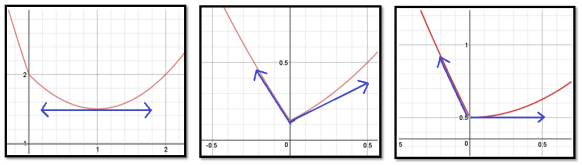

Figure 1: **Simple example.** Left: $c = 2$, and $x^* = 1 > 0$ (inactive). Middle: $c = 1/2$ and $\delta_1(x^*) = 1/2 > 0$ (active). Right: $c = 1$, and both $x^* = \delta_1(x^*) = 0$ (degenerate).

It has been observed that $\delta_{\min}$ plays a key role not only in the iteration complexity for manifold identification, but also for variable convergence. This is often attributed to the two-phase convergence in which many methods seem to accelerate after the manifold has been identified (Wright, 2012; Gafni and Bertsekas, 1984). In particular, when we reduce (2) to $\{x_i : i \in \mathcal{Z}\}$, often the problem becomes better conditioned; that is,

$$\lambda_{\min}(\nabla^2 g(x)) \leq \lambda_{\min}((\nabla^2 g(x))_{\mathcal{Z},\mathcal{Z}}).$$

and this increase often results in noticable speedup in convergence (Liang et al., 2014, 2016).

3 MANIFOLD ID METHODS

Definition 4. For a nondegenerate problem ($\delta_{\min} > 0$) we say a method is *manifold identifying* if it identifies the active set $\mathcal{Z}$ in a finite number of iterations. Specifically, there is a $\bar{k}$ such that for all $k \geq \bar{k}$,

$$x_i^{(k)} = x_i^*, \quad \forall\ i \in \mathcal{Z}.$$

The quantity $\bar{k}$ is the *active set complexity* and depends on $\delta_{\min}$.

There is an inherent one-sided flavor to this definition, in that for indices *outside* the active set, these points may achieve x_i^* at any k, or never at all. For sparse optimization, this means that $x_i^{(k)} = 0$ may occur at any k, even if $i \notin \mathcal{Z}$. Thus, this definition differs from the traditional notion of support identification, where the exact pattern of zeros *and* nonzeros are identified. In general, guaranteeing both sides cannot be done for finite $\bar{k}$, as $\delta_i(x^*) = 0$ whenever $i \notin \mathcal{Z}$; however, in practice both are often identified in finite time.

3.1 Proximal methods

We define the *proximal mapping* of h with respect to a positive definite scaling matrix H as

$$\mathbf{prox}_h^H(z) := \underset{x}{\mathbf{argmin}}\ h(x) + \frac{1}{2}(x - z)^T H(x - z).$$

We also use the unscaled proximal mapping $\mathbf{prox}_h(z) := \mathbf{prox}_h^I(z)$.

There are many important methods that use the proximal mapping. The most common example is the proximal gradient (PG) method, where at each iteration,

$$\begin{aligned} z^{(k)} &= x^{(k)} - t^{(k)} \nabla f(x^{(k)}) \\ x^{(k+1)} &= \mathbf{prox}_{t^{(k)} h}(z^{(k)}) \end{aligned}$$

is applied until $x^{(k)}$ converges to a fixed point x^*. Here, $t^{(k)} > 0$ is some positive step length, and $z^{(k)}$ is the result after a gradient step. Other proximal methods of similar form include the prox-Newton method, the Douglas-Rachford Splitting (DRS) method, and the Alternating Direction Method of Multipliers (ADMM). When using a constant step size, we write $t := t^{(k)}$ for all k.

Lemma 1 (Wiggle Room Lemma)**.** *Consider a generalized proximal algorithm for solving* (2) *of the form*

$$x^{(k+1)} = \mathbf{prox}_{t^{(k)} h}^{H^{(k)}}(z^{(k)}), \tag{7}$$

where $z^{(k)}$ depends on past $x^{(k)}$, $t^{(k)}$, and $H^{(k)}$. Then the active set is identified by $x^{(k+1)}$ when for all $i \in \mathcal{Z}$

$$\left| \left(\frac{1}{t^{(k)}} H^{(k)}(z^{(k)} - x^*) + \nabla g(x^*) \right)_i \right| \leq \delta_{\min}. \tag{8}$$

The proof of Lemma 1 is in Appendix A.

In the case of the PG method, using the triangle inequality, (8) can be satisfied if

$$\left(\frac{1}{t} + L \right) \|x^{(k)} - x^*\|_2 \leq \delta_{\min}. \tag{9}$$

When g is strongly convex, $\|x^{(k)} - x^*\|_2 = O(\exp(-k))$, and by inverting (9) the active set complexity is

$$\bar{k} = O\left(\log\left(\frac{1/t + L}{\delta_{\min}}\right)\right),$$

as reported in Nutini et al. (2017b).

We now use Lemma 1 to prove the manifold identifying property on several well-known proximal methods. In particular, we define $s^{(k)} \in \mathbb{R}^n$ with

$$s_i^{(k)} = \begin{cases} 0, & i \notin \mathcal{Z} \\ \frac{1}{t^{(k)}}\left(z_i^{(k)} - x_i^*\right) + (\nabla g(x^*))_i, & i \in \mathcal{Z} \end{cases} \quad (10)$$

the *active set residual*, and derive $\bar{k}$ where for all $k \geq \bar{k}$, $\|s^{(k)}\|_\infty \leq \delta_{\min}$.

3.2 First-order methods

In this section we derive the active set complexity for several popular first-order deterministic methods. We give $\bar{k}$ as a function of variable error rates, with an explicit rate when g is strongly convex. All proofs can be found in Appendix A, and exact convergence rates (used to compute exact $\bar{k}$) are given in Appendix B. We define $\epsilon_x(k)$ a monotonically decreasing upper bounding sequences ($\epsilon_x(k) \geq \|x^{(k)} - x^*\|_2$) and always assume $t^{(k)} = t$ a constant step size.

We emphasize that strong convexity is not a requirement to show finite manifold identification, but is often used to derive the variable convergence rate $\epsilon_x(k)$. All we require at this point is the uniqueness of the solution x^* to (2).

Accel. Prox. Grad. (aPG) The proximal gradient descent is often accelerated (Nesterov, 2013b) via a simple scheme

$$\begin{aligned} y^{(k+1)} &= x^{(k)} - t\nabla g(x^{(k)}) \\ x^{(k+1)} &= \mathbf{prox}_{th}((1-\gamma^{(k)})y^{(k+1)} + \gamma^{(k)} y^{(k)}) \end{aligned}$$

for a specific sequence of $-1 \leq \gamma^{(k)} \leq 1$. When g is strongly convex, then the iterates $x^{(k)}$ converge to x^* at a linear rate (Nesterov, 2013a).

Theorem 1. *The aPG method identifies the manifold for all $k \geq \bar{k}$ when*

$$\left(\frac{1}{t} + L\right)\epsilon_x(\bar{k} - 1) \leq \delta_{\min}.$$

When g is strongly convex, $\bar{k} = O(\log(\delta_{\min}))$.

Note that manifold identification is with reference to $x^{(k)}$; $y^{(k)}$ may not be in $\mathcal{M}$. More generally, manifold identification is proven only for the output of the proximal mapping. The dependence of aPG on $\epsilon(k)$ (see Appendix B) is the same order as that of PG; however, since the variable convergence rate $\epsilon(k)$ of aPG is faster than PG, the overall active set complexity rate is faster as well. Finally, the manifold identification property does *not* depend on the method itself having monotonic variable convergence, as aPG in general does not. All we need is a monotonic variable *bound*, which may be pessimistic in practice. (See Fig. 2.)

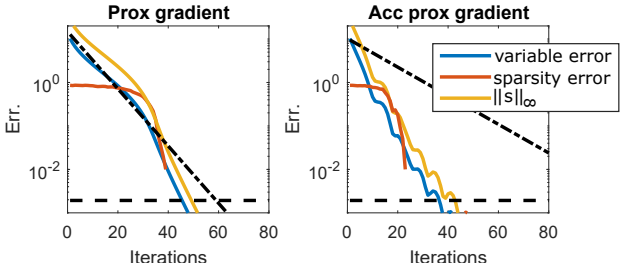

Figure 2: **PG and aPG convergence behavior.** Trajectories of deterministic methods on sparse linear regression. Here, the entries of $A \in \mathbb{R}^{200\times 500}$ are i.i.d. Gaussian and $b = Ax^{\#} + y$. The vector $x^{\#}$ is the sparse ground truth signal, and y is a Gaussian i.i.d. noise vector. The variable error (blue) is bounded above by the dashed decreasing line, which is $\epsilon(k)$ as reported in literature. The ASI error (yellow) reaches 0 when the max absolute value of the residual (red) dips below the horizontal dashed line, which represents $\delta_{\min}$. Before this happens, the manifold has been identified.

Previous work Liang et al. (2017) and Johnstone and Moulin (2015) also discuss manifold identification for aPG. The analysis in Johnstone and Moulin (2015) is similar but is restricted to sparse (ℓ_1 regularized) optimization. In Liang et al. (2017) the analysis is based more on generalized topological analysis and extends to all L-smooth convex functions g, resulting in an active set complexity of $O(\delta_{\min}^2)$ iterations. In comparison, Theorem 1 with strong convexity guarantees $O(\log(\delta_{\min}))$ iterations.

DRS and ADMM Two other (equivalent) non-monotone proximal splitting methods are the DRS method (Douglas and Rachford, 1956; Lions and Mercier, 1979; Eckstein and Bertsekas, 1992), which minimizes (2) using the following scheme

$$\begin{aligned} x^{(k+1)} &= \mathbf{prox}_{th}(z^{(k)}) && (11) \\ y^{(k+1)} &= \mathbf{prox}_{tg}(2x^{(k+1)} - z^{(k)}) && (12) \\ z^{(k+1)} &= z^{(k)} + y^{(k+1)} - x^{(k+1)} && (13) \end{aligned}$$

and ADMM (Gabay and Mercier, 1975; Glowinski and Marroco, 1975) on the reformulation of (2)

$$\min_{x,y} \{g(y) + h(x) : x = y\}. \quad (14)$$

The variable convergence rate is equivalent for both, since in this simplified formulation they are equivalent under the transformation $u^{(k)} = (z^{(k)} - x^{(k)})/t$ where $u^{(k)}$ is the dual ADMM iterate. (See Appendix A.) In

particular, Giselsson and Boyd (2017) showed $x^{(k)} \to x^*$ at a linear rate and He and Yuan (2015) showed $\|x^{(k)} - y^{(k)}\|_2 \to 0$ at a $O(1/\sqrt{k})$ rate.

Theorem 2. *Consider a monotonically decreasing sequence $\epsilon_x(k) \geq \max\{\|x^{(k)} - x^*\|_2, \|x^{(k)} - y^{(k)}\|_2\}$ for all k. Then the active set is identified at $\bar{k}$ when*

1. $(2/t + L)\epsilon_x(\bar{k}) \leq \delta_{\min}$ *for DRS, and*
2. $(2/t + 2L)\epsilon_x(\bar{k}) \leq \delta_{\min}$ *for ADMM on* (14).

For g strongly convex, $\bar{k} = O(\delta_{\min}^2)$ for both methods.

Previous work Liang et al. (2016) Prop. 4.5 gives a rate $\bar{k} = O(\delta_{\min}^2)$ for general convex g. When g is strongly convex, Thm. 2 gives an improved active set complexity of $\bar{k} = O(\delta_{\min})$. Note that if we can show $\|x^{(k)} - y^{(k)}\|_2$ converging linearly, we can improve this to $\bar{k} = O(\log(\delta_{\min}))$. The local linear convergence behavior of DRS-like methods is also discussed in Molinari et al. (2018), which shows finite active set complexity but does not provide a rate for $\bar{k}$.

3.3 Proximal Newton-type methods

When g is twice-differentiable everywhere, a family of proximal Newton-type methods can be described as

$$x^{(k+1)} = \underset{x}{\mathbf{argmin}} \big(\nabla g(x^{(k)})^T x + h(x) + (x - x^{(k)})^T H_{\text{est}}^{(k)} (x - x^{(k)})\big), \tag{15}$$

where $H_{\text{est}}^{(k)}$ is a symmetric positive definite matrix approximating the Hessian at $x^{(k)}$ satisfying the Dennis-More (Dennis and Moré, 1974) criterion

$$\frac{\|(H_{\text{est}}^{(k)} - \nabla^2 g(x^*))(x^{(k+1)} - x^{(k)})\|_2}{\|x^{(k+1)} - x^{(k)}\|_2} \to 0.$$

Specifically, (15) describes the prox-Newton (pN) method when $H_{\text{est}}^{(k)} = \nabla^2 g(x^{(k)})$, and more generally the prox-Quasi-Newton (pQn) method. These methods are well-studied; when g is strongly convex, pN converges q-quadratically (Lee et al., 2014), and pQN converges linearly (Ghanbari and Scheinberg, 2016).

Theorem 3. *When $H_{\text{est}}^{(k)} \preceq L_H I$, both pN and pQN methods identify the manifold for all $k \geq \bar{k}$ when* [2]

$$(L + L_H)\epsilon_x(\bar{k}) \leq \delta_{\min}.$$

For g strongly convex, pQN has active set complexity

$$\bar{k} = O(\log(\delta_{\min})$$

and pN has

$$\bar{k} = O(\log(\log(\delta_{\min})).$$

When $H_{\text{est}}^{(k)} = \nabla^2 g(x^{(k)})$ is exact, $L_H = L$.

An important future extension of this analysis is to include quasi-Newton methods that do not satisfy the Dennis-More condition, such as L-BFGS.

[2] The two-stage convergence behavior of the Newton-type methods (linear to a neighborhood, followed by quadratic convergence) is captured in $\epsilon_x(k)$.

3.4 Stochastic methods

We now extend the analysis to stochastic proximal methods, namely proximal versions of stochastic gradient descent (SGD), the stochastic variance reduced gradient (SVRG) method (Johnson and Zhang, 2013), stochastic average gradient amélioré (SAGA) (Defazio et al., 2014), and the regularized dual averaging (RDA) method (Xiao, 2010). A new challenge in stochastic methods is a potentially diminishing step size. In particular, when $t^{(k)} \to 0$, it is not clear that condition (8) will hold for all $k > \bar{k}$, for any finite $\bar{k}$, even if $x^{(k)} \to x^*$. Condition (8) is a conservative bound; in our experiments, we observe that when $\epsilon_x(k)/t^{(k)} \not\to 0$, manifold identification happens in some problem instances, but not consistently.

We now assume g is a sum of smooth functions

$$g(x) = \frac{1}{m}\sum_{i=1}^{m} g_i(x) \tag{16}$$

and consider methods that sample a single gradient $\nabla g_i(x)$ for $i \in \{1, ..., m\}$ uniformly at each iteration. Exact $\bar{k}$ values can be computed by combining the stated theorems and explicit rates for ϵ_x and ϵ_g. (See Table 2 given in Appendix C.)

SGD, SVRG, SAGA Denote by $i[k]$ the sample picked at iteration k. These methods can be summarized by the iteration scheme

$$x^{(k+1)} = \mathbf{prox}_{t^{(k)}h}(x^{(k)} - t^{(k)} G_{\text{est}}^{(k)}) \tag{17}$$

where $G_{\text{est}}^{(k)}$ is a noisy estimate of the gradient $\nabla g(x^{(k)})$. In particular, we consider

- prox-SGD (pSGD), where $G_{\text{est}}^{(k)} = \nabla g_{i[k]}(x^{(k)})$;
- prox-SVRG (pSVRG) (Johnson and Zhang, 2013), where
$$G_{\text{est}}^{(k)} = \nabla g_{i[k]}(x^{(k)}) - \nabla g_{i[k]}(\tilde{x}) + \nabla g(\tilde{x})$$
and $\tilde{x}$ is the iterate taken at the beginning of the current epoch; and
- prox-SAGA (pSAGA) (Defazio et al., 2014), where
$$G_{\text{est}}^{(k)} = \nabla g_{i[k]}(x^{(k)}) - y_{i[k]}^{(k-1)} + \frac{1}{m}\sum_{i=1}^{m} y_i^{(k-1)}$$
and
$$y_i^{(k)} = \begin{cases} \nabla g_i(x^{(k)}), & i = i[k], \\ y_i^{(k-1)}, & \text{else.} \end{cases}$$

Define $\epsilon_x(k)$ and $\epsilon_g(k)$ two monotonically decaying sequences such that $\epsilon_x(k) \geq \|x^{(k)} - x^*\|_2$ and $\epsilon_g(k) \geq \|\nabla g(x^*) - G_{\text{est}}^{(k)}\|_2$.

Theorem 4. *pSVRG and pSAGA identify the manifold for $k > \bar{k}$ when*

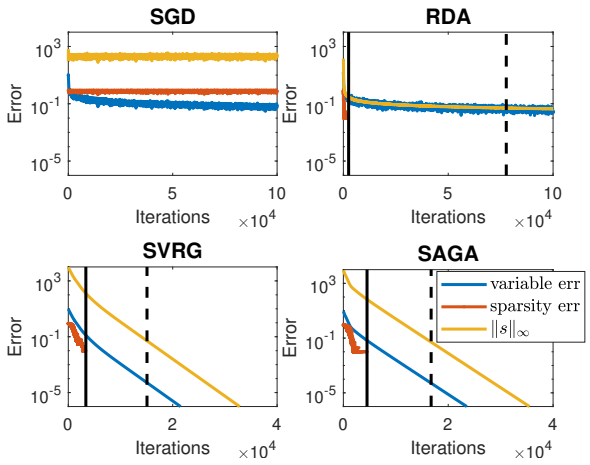

Figure 3: **Trajectories of stochastic methods on sparse linear regression.** Data is generaed as in Fig. 2. For SGD and RDA, the step size rate decays as $t^{(k)} = 0.1/\sqrt{k}$, which is necessary for convergence. For SVRG and SAGA, the step size is constant at $t^{(k)} = 0.1$. The variable error (blue) and the max value of the active set residual (yellow) are plotted. The vertical lines measure when the active set is identified (solid), and when $\|s^{(k)}\|_\infty < \delta_{\min}$ (dashed).

$$\frac{\epsilon_x(\bar{k})}{t^{(\bar{k})}} + \epsilon_g(\bar{k}) \leq \delta_{\min}. \tag{18}$$

When g is strongly convex and $t^{(k)} = t$ a constant step size, $\bar{k} = O(\log(\delta_{\min}))$.

Proof. We invoke (8) for $z^{(k)} = x^{(k)} - t^{(k)} G^{(k)}_{\text{est}}$ and apply the triangle inequality. The rest follows from the linear convergence rate (Appendix C Table 2). □

For pSGD, the condition for finite manifold identification is also (18); however, here $\epsilon_g(k) \not\to 0$ in general. [3] A key advantage of variance reduced methods like pSVRG (Johnson and Zhang, 2013) and pSAGA (Defazio et al., 2014) (as well as other variants like MISO, Finito, and SDCA) is that $\epsilon_g(k) \to 0$. Then, since pSVRG and pSAGA usually employ constant step size, both terms in (18) go to 0, and thus are manifold identifying methods. (See also Poon et al. (2018).)

Previous work Poon et al. (2018) discuss finite-iteration manifold identification of pSVRG and pSAGA, and give some hint to $\bar{k}$ for specific applications, but do not show its relationship with $\delta_{\min}$.

RDA The RDA method (Duchi and Ruan, 2016) and its proximal version (pRDA) (Xiao, 2010) was introduced as a variance-reduced method for stochastic optimization (where we do not need to assume that m is finite in (16)). Unlike pSVRG and pSAGA, in pRDA we require a decaying step size for convergence, and the variable error converges at a sublinear rate.

By some rearrangement, we can rewrite this scheme as

$$\begin{aligned} \bar{g}^{(k)} &= \frac{1}{k}\sum_{i=1}^{k} \nabla g_{i[k]}(x)) \\ x^{(k+1)} &= \mathbf{prox}_{kt^{(k)}h}\left(-kt^{(k)}\bar{g}^{(k)}\right). \end{aligned}$$

Then, using $z^{(k)} = -kt^{(k)}\bar{g}^{(k)}$, condition (8) gives the following result.

Theorem 5. *pRDA identifies the manifold when*

$$\epsilon_g(k) + \frac{B}{kt^{(k)}} \leq \delta_{\min},$$

where $B \geq \|x^{(k)}\|_2$ for all k. Taking $t^{(k)} = 1/\sqrt{k}$ and g strongly convex, $\bar{k} = O\left((\delta_{\min})^4\right)$.

Since most choices of h in our applications are not strongly convex, the only scheme in which RDA is guaranteed to converge is with a *diminishing* step size,[4] such as $t^{(k)} = 1/\sqrt{k}$. Under this choice of step size, Xiao (2010) reports $\mathbb{E}[(\epsilon_x(k))^2] = O(k^{-1/2})$, which gives $\mathbb{E}[\epsilon_g(k)] = O(k^{-1/4})$. Interestingly, Theorem 5 does not depend on $\epsilon_x(k)$, and guarantees active set identification whenever $t^{(k)} > O(1/k)$.

Previous work Lee and Wright (2012) discuss finite-iteration manifold identification of RDA and show that there is a direct relationship between $\epsilon_x(k)$ and $\delta_{\min}$. They also provide sharper probabilistic estimates of $\epsilon_x(k)$, whereas ours is in expectation. Duchi and Ruan (2016) also consider manifold identification of a variantof RDA under a restricted strongly convex assumptions on g.

4 APPLICATIONS

It is now evident that a key factor in the manifold identifying behavior is $\delta_{\min}$. Unfortunately, because $\delta_{\min}$ depends on the optimal value x^* of (2), in practice the value is not available until after the optimization is complete. In this section, we develop some intuition as to when $\delta_{\min}$ may be large, based on data incoherence and training performance. We investigate this for both sparse regularization and bound constraints. All the proofs in this section are in Appendix D.

4.1 Sparse regularization

Consider $h(x) = \lambda\|x\|_1$. Then $\delta_i(x) = \lambda - |\nabla g(x)_i|$. Here, we consider two specific instances of problem (1): sparse linear and logistic regression.

[3] If we assume additional conditions, such as the strong growth condition on f, then $\epsilon_g(k) \to 0$ (Schmidt and Roux, 2013).

[4] Here, we use $1/t^{(k)}$ in place of $\beta^{(k)}$ the step size parameter used in Xiao (2010), so that all methods are easy to compare.

Linear regression Taking

$$\mathcal{L}(a_k^T x; b_k) = (1/2)(a_k^T x - b_k)^2,$$

then

$$\delta_{\min} = \min_{j \in \mathcal{Z}} \lambda - \frac{1}{m}|(A^T(Ax-b))_j|,$$

where $A = [a_1, \ldots, a_m]^T$.

- **Repeated feature entries.** Assume that the data matrix is $A = [a, a, ..., a] = a\mathbf{1}^T \in \mathbb{R}^{m \times n}$. Then for $i \neq j$ and $x_i = 0$ but x_j is nonzero, then x_i is necessarily degenerate. (See Appendix D.)
- **Very incoherent A.** When the data matrix A is almost orthonormal ($A^T A \approx mI$) and fits the data well ($Ax \approx b$) then for fixed λ, one may expect $\delta_{\min}$ to be larger. (Prop. 1.)
- **Incoherent useless entries.** Often, data vectors are *not* totally incoherent. In fact, in classification, we may expect the data matrices restricted to one class to be highly correlated. Note, however, that if $j \in \mathcal{Z}$ then $x_j = 0$, and the jth feature is not used in the final classifier. Such a feature then has little correlation with the sample label, and may be uncorrelated with the other sample features as well.

Assumption 1. *Suppose that $b = F(Ax^\#) + y$ for some smooth map $F : \mathbb{R}^n \to \mathbb{R}^n$ and sparse "truth" vector $x^\#$, and the noise is bounded by $\|y\|_2 \leq \eta\sqrt{m}$. Define $e = x^* - x^\#$. For all columns a_i of A, $\|a_i\|_2 \leq \alpha\sqrt{m}$. For all $i \in \mathcal{Z}$, for all $j \neq i$, $|a_i^T a_j| \leq \rho$.*

Proposition 1. *With Assumption 1 and model $b = Ax^\# + y$, for all $j \in \mathcal{Z}$, in linear regression,*

$$\delta_{\min} \geq \lambda - \rho\|e\|_1 - \alpha^2|e_j| - \alpha\eta.$$

In general, α is some small fixed constant. (If we normalize the data, then $\alpha = 1$.) This proposition suggests that $\delta_{\min}$ can be increased if either λ is larger or ρ, $|e_j|$, and η are smaller.

Figure 4 gives a surface plot of $|x_i^*|$ and $\delta_i(x^*)$ for the MNIST binary classification problem distinguishing 4 and 9, for varying choice of λ. As expected, with sparse regularization, the features selected are those in regions that distinguish between 4 and 9, such as the angular slant of the 4. The $\delta_i(x^*)$ surface plot shows a close inverse, with large values when a feature is not used in classification ($|x_i| \approx 0$). In regions where both plots are blue, the feature is degenerate, suggesting ambiguity in whether it is important.

Sparse logistic regression We now extend our results to sparse logistic regression, where $\sigma(x) := 1/(1+e^{-x})$ the sigmoid function, and

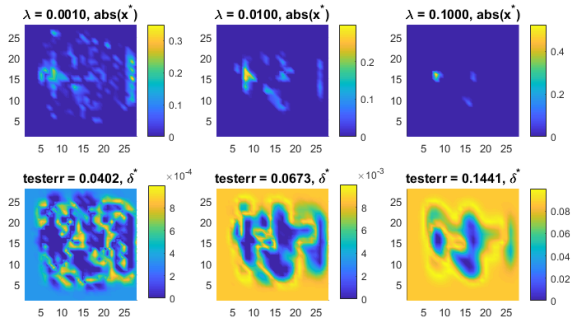

Figure 4: **Feature importance.** MNIST classification of 4 vs 9 digits via sparse linear regression for varying levels of λ. The top row are $|x_i|$ and bottom are $\delta(x_i^*)$. Note the tradeoff between test error and sparsity quality, with the middle column being a desirable "sweet spot".

$$\begin{aligned}\mathcal{L}(a_i^T x; b_i) = &-b_i \log(\sigma(a_i^T x)) \\ &-(1-b_i)\log(1-\sigma(a_i^T x)).\end{aligned} \tag{19}$$

Extending the analysis to logistic regression is more involved, and thus we locally linearize the model.

Proposition 2. *With Assumption 1 and $b = \sigma(Ax^\#) + y$, define*

$$\tau = \max_i \sigma(a_i^T x^*)(1 - \sigma(a_i^T x^*)).$$

Then for all $j \in \mathcal{Z}$, for the sparse logistic problem,

$$\delta_j^* \geq \lambda - \tau\alpha^2\|e\|_2 - \alpha\eta + O(\|e\|_2^2).$$

The constant $\tau < 1/4$ is a measure of the maximum "uncertainty" for a single sample; if all samples are classified easily, then τ is very small. Overall, if τ, η, and $\|e\|_2$ are small, we may expect larger $\delta_{\min}$.

Figure 5 illustrates this result numerically. In almost all cases, the probability of degeneracy ($\delta_{\min} = 0$) is at the tail end of the curve (unless λ is too small). In these experiments, we found varying noise levels and the number of samples m had little effect on the $\delta_{\min}$ distribution. However, the sparsity levels of $x^\#$, the λ value, and the incoherence of A, did have an effect.

4.2 Constrained optimization

Element-wise constraints on x appear after dualization of the hinge-loss function. The most important example of this is the Lagrange dual of SVMs [5]

$$\begin{aligned}\min_{x \in \mathbb{R}^m} \quad & \tfrac{1}{2m}(b \circ x)^T K (b \circ x) - \tfrac{1}{m} x^T \mathbf{1} \\ \text{st} \quad & 0 \leq x_i \leq \lambda, \; i = 1, \ldots, m.\end{aligned} \tag{20}$$

Here, $x \circ y$ represents element-wise multiplication. The matrix $K \in \mathbb{R}^{m \times m}$ is the symmetric positive semidefinite kernel matrix, where $K_{ij} = K_f(a_i, a_j)$ and K_f is

[5] We analyze the version without bias terms in the model.

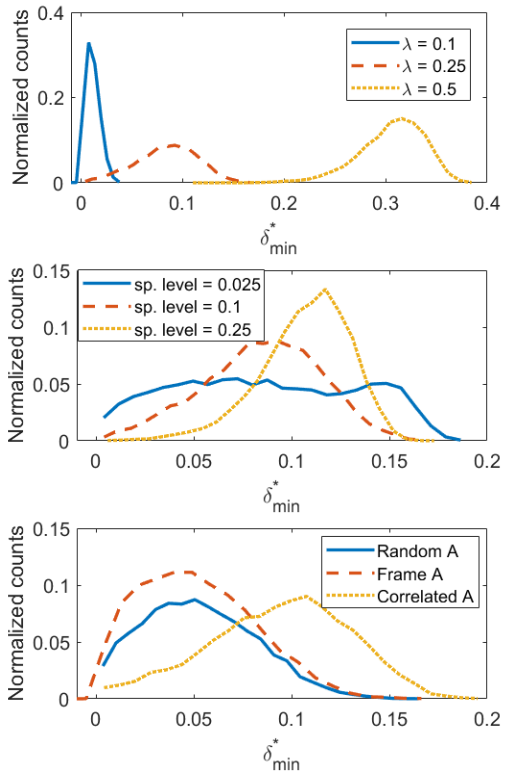

Figure 5: **Histograms of** $\delta_{\min}$. Here, x^* is the solution to 10,000 random instances of logistic regression, where the data is generated as in Fig 2. Unless otherwise stated, A is random Gaussian, $n = 250$, $m = 250$, sparsity level = 10%, noise level = 1/100, $\lambda = 1$. As expected, increasing λ increases $\delta_{\min}$ in expectation. Also, when the sparsity level of $x^{\#}$ is increased, $\delta_{\min}$ concentrates away from 0. Unexpectedly, more correlated data seems to increase $\delta_{\min}$ in practice.

some kernel function. For a linear SVM, $K = AA^T$. After training, the binary classifier for a new data vector $\hat{a}$ is $f(\hat{a}) = \mathbf{sign}(\sum_{i=1}^{m} b_i x_i K_f(a_i, \hat{a}))$.

The function g is the L-smooth objective of (20) (where $L = \|K\|_2$). The constraint can be expressed as a separate $0/\infty$ indicator penalty for each element-wise interval, and $\delta_{\min} = \min_{i \in \mathcal{Z}} |(\nabla g(x^*))_i|$. Additionally, $j \in \mathcal{Z}$ implies $x_j^* \in \{0, \lambda\}$.

Proposition 3. *Let us assume that for all* $j \in \mathcal{Z}$, $|K_{ij}| \leq \rho$ *for* $i \neq j$, *and* $|K_{ii}| \leq \alpha$. *Then for problem* (20), *if* $x_j = \lambda$ *then*

$$\delta_j^* \geq \frac{1}{m} - \frac{\lambda}{m}(\alpha + \rho m).$$

We generally expect α to be a small constant value; for linear SVMs, $\alpha \geq \|a_i\|_2^2$ for all i, and for the radial basis function (RBF) kernel, $\alpha = 1$. When $x_j = 0$, then the data sample j plays no role in the final classifier and we cannot give any guarantees; the value of $\delta_{\min}$ can be arbitrarily close to 0. Now assume that for some $j \in \mathcal{Z}$, $x_j = \lambda$. If we choose λ very small (corresponding to a large hinge loss regularizer in the primal) we will obtain large $\delta_{\min}$; however this can degrade performance as it unnaturally drives x^* to 0. However, limiting $\lambda \leq 1/\alpha$ and driving ρ to 0 can increase $\delta_{\min}$. Active set methods for solving the dual SVM are explored in a number of works, for example Joachims (1998), Usunier et al. (2010). Keerthi and DeCoste (2005), and She and Schmidt (2017).

5 CONCLUSIONS

The ability of certain methods to identify low-dimensional solution manifolds in finite time is known in folklore, but existing proofs are often specific to the algorithm. Here we provide a unified view of proximal methods, with one key lemma that summarizes all the requirements for finite active set complexity. To show the power of this unified analysis, we calculate the step size assumptions and number of iterations needed for manifold identification for seemingly unrelated proximal methods.

The active set complexity is closely reliant on problem degeneracy ($\delta_{\min}$), and both manifold identification and variable convergence is often faster when $\delta_{\min}$ is large. We cannot easily compute $\delta_{\min}$, but analysis and experiments suggest that manipulating data parameters like ρ, τ, and $\|e\|_1$ (for example, through clever feature selection) give favorable properties.

None of our results rely on strong convexity of g when relating $\epsilon(k)$, $\bar{k}$, and $\delta_{\min}$, though we require x^* to be unique. Considering problems with nonunique x^* adds unrealistic complications to the analysis, as the resulting solution manifold may no longer be unique. Of course, the rates $\epsilon_x(k)$ often require g to be strongly convex, and having more general variable error rates is the most natural way to relax this assumption. (Duchi and Ruan, 2016; Yen et al., 2014)

Finally, thus far not much in our analysis has precluded the use of nonconvex functions g and regularizers h. Convexity has several advantages, such as a simple and everywhere interpretable notion of subdifferentials, and simplifies the question of unique stationary points. However, though we focus on convex functions, we believe the intuition in our analysis carries over to more general problems.

Acknowledgements

Yifan Sun is supported by ONR award N00014-17-1-2009. Halyun Jeong is supported by the UBC Data Science Institute and the Pacific Institute of Mathematical Sciences.

References

Bertsekas, D. P. (1974). On the goldstein-levitin-polyak gradient projection method. In *1974 IEEE Conference on Decision and Control including the 13th Symposium on Adaptive Processes*, pages 47–52. IEEE.

Burke, J. V. and Moré, J. J. (1988). On the identification of active constraints. *SIAM Journal on Numerical Analysis*, 25(5):1197–1211.

Daniilidis, A., Sagastizábal, C., and Solodov, M. (2009). Identifying structure of nonsmooth convex functions by the bundle technique. *SIAM Journal on Optimization*, 20(2):820–840.

De Santis, M., Lucidi, S., and Rinaldi, F. (2016). A fast active set block coordinate descent algorithm for \ell_1-regularized least squares. *SIAM Journal on Optimization*, 26(1):781–809.

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654.

Dennis, J. E. and Moré, J. J. (1974). A characterization of superlinear convergence and its application to quasi-newton methods. *Mathematics of computation*, 28(126):549–560.

Douglas, J. and Rachford, H. H. (1956). On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American mathematical Society*, 82(2):421–439.

Duchi, J. and Ruan, F. (2016). Asymptotic optimality in stochastic optimization. *Arxiv Preprint*.

Dunn, J. C. (1987). On the convergence of projected gradient processes to singular critical points. *Journal of Optimization Theory and Applications*, 55(2):203–216.

Eckstein, J. and Bertsekas, D. P. (1992). On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318.

Gabay, D. and Mercier, B. (1975). *A dual algorithm for the solution of non linear variational problems via finite element approximation*. Institut de recherche d'informatique et d'automatique.

Gafni, E. M. and Bertsekas, D. P. (1984). Two-metric projection methods for constrained optimization. *SIAM Journal on Control and Optimization*, 22(6):936–964.

Ghanbari, H. and Scheinberg, K. (2016). Proximal quasi-newton methods for regularized convex optimization with linear and accelerated sublinear convergence rates. *arXiv preprint arXiv:1607.03081*.

Giselsson, P. and Boyd, S. (2017). Linear convergence and metric selection for Douglas-Rachford splitting and ADMM. *IEEE Transactions on Automatic Control*, 62(2):532–544.

Glowinski, R. and Marroco, A. (1975). Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(R2):41–76.

Hare, W. (2011). Identifying active manifolds in regularization problems. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 261–271. Springer.

Hare, W. and Lewis, A. S. (2004). Identifying active constraints via partial smoothness and prox-regularity. *Journal of Convex Analysis*, 11(2):251–266.

Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5(Oct):1391–1415.

He, B. and Yuan, X. (2015). On the convergence rate of Douglas–Rachford operator splitting method. *Mathematical Programming*, 153(2):715–722.

Joachims, T. (1998). Making large-scale svm learning practical. Technical report, Technical report, SFB 475: Komplexitätsreduktion in Multivariaten ....

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323.

Johnstone, P. R. and Moulin, P. (2015). A lyapunov analysis of fista with local linear convergence for sparse opti-mization. *arXiv preprint arXiv:1502.02281*.

Keerthi, S. S. and DeCoste, D. (2005). A modified finite newton method for fast solution of large scale linear svms. *Journal of Machine Learning Research*, 6(Mar):341–361.

Ko, M., Zowe, J., et al. (1994). An iterative two-step algorithm for linear complementarity problems. *Numerische Mathematik*, 68(1):95–106.

Lee, J. D., Sun, Y., and Saunders, M. A. (2014). Proximal newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443.

Lee, S. and Wright, S. J. (2012). Manifold identification in dual averaging for regularized stochastic online learning. *Journal of Machine Learning Research*, 13(Jun):1705–1744.

Lewis, A. S. and Wright, S. J. (2011). Identifying activity. *SIAM Journal on Optimization*, 21(2):597–614.

Liang, J., Fadili, J., and Peyré, G. (2014). Local linear convergence of forward–backward under partial smoothness. In *Advances in Neural Information Processing Systems*, pages 1970–1978.

Liang, J., Fadili, J., and Peyré, G. (2016). Local convergence properties of Douglas–Rachford and ADMM. *arXiv preprint arXiv:1606.02116*.

Liang, J., Fadili, J., and Peyré, G. (2017). Activity identification and local linear convergence of forward–backward-type methods. *SIAM Journal on Optimization*, 27(1):408–437.

Lions, P.-L. and Mercier, B. (1979). Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979.

Molinari, C., Liang, J., and Fadili, J. (2018). Convergence rates of forward–douglas–rachford splitting method. *arXiv preprint arXiv:1801.01088*.

Nesterov, Y. (2013a). Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161.

Nesterov, Y. (2013b). *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.

Nutini, J., Laradji, I., and Schmidt, M. (2017a). Let's make block coordinate descent go fast: Faster greedy rules, message-passing, active-set complexity, and superlinear convergence. *arXiv preprint arXiv:1712.08859*.

Nutini, J., Schmidt, M., and Hare, W. (2017b). "Active-set complexity" of proximal gradient: How long does it take to find the sparsity pattern? *arXiv preprint arXiv:1712.03577*.

Poon, C., Liang, J., and Schönlieb, C.-B. (2018). Local convergence properties of SAGA/prox–SVRG and acceleration. *arXiv preprint arXiv:1802.02554*.

Rockafellar, R. T. (2015). *Convex analysis*. Princeton university press.

Schmidt, M. and Roux, N. L. (2013). Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*.

She, J. and Schmidt, M. (2017). Linear convergence and support vector identification of sequential minimal optimization. In *10th NIPS Workshop on Optimization for Machine Learning*, page 5.

Tseng, P. and Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1-2):387–423.

Usunier, N., Bordes, A., and Bottou, L. (2010). Guarantees for approximate incremental svms. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 884–891.

Wright, S. J. (1993). Identifiable surfaces in constrained optimization. *SIAM Journal on Control and Optimization*, 31(4):1063–1079.

Wright, S. J. (2012). Accelerated block-coordinate relaxation for regularized optimization. *SIAM Journal on Optimization*, 22(1):159–186.

Xiao, L. (2010). Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596.

Yen, I. E.-H., Hsieh, C.-J., Ravikumar, P. K., and Dhillon, I. S. (2014). Constant nullspace strong convexity and fast convergence of proximal methods under high-dimensional settings. In *Advances in Neural Information Processing Systems*, pages 1008–1016.