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

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

A Manifold identification complexity for deterministic methods

In this section, we give the manifold identification complexity of deterministic methods, as a function of $\delta_{\min}$ and $\epsilon_x(k)$. Auxillary results are also presented when needed.

A.1 Proof of Lemma 1

Proof. First, by the optimality conditions of (2), we have

$$-\nabla g(x^*) \in \partial h(x^*).$$

By definition of $\delta_{\min}$, if $h(x) = \sum_i h_i(x_i)$, then additionally for all $i \in \mathcal{Z}$, any vector u where $|u_i + \nabla g(x^*)_i| \leq \delta_{\min}$ also satisfies

$$u_i \in \partial h_i(x_i^*).$$

Using (8) we see that $u = \frac{1}{t^{(k)}} H^{(k)} \left(z^{(k)} - x^*\right)$ satisfies this property, and therefore

$$\frac{1}{t^{(k)}} H^{(k)} \left(z_i^{(k)} - x_i^*\right) \in \partial h_i(x_i^*)$$

which is true if and only if

$$x_i^* = \mathbf{prox}_{t^{(k)} h_i}^{H^{(k)}}(z_i^{(k)}).$$

Since the solution to the prox is unique, then this implies that for all $i \in \mathcal{Z}$,

$$x_i^{(k+1)} = x_i^*.$$

□

A.2 Accelerated proximal gradient descent

We give more details to the accelerated proximal gradient method and derive theorem 1. The proximal gradient descent is often accelerated (Nesterov, 2013b) via a simple scheme

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

When g is convex, $\lambda^{(0)} = 0$ and

$$\lambda^{(k)} = \frac{1 + \sqrt{1 + 4(\lambda^{(k-1)})^2}}{2}, \quad \gamma^{(k)} = \frac{1 - \lambda^{(k)}}{\lambda^{(k+1)}}.$$

When g is strongly convex, $\gamma^{(k)} = \gamma$ is constant, as

$$\gamma = \frac{1 - \sqrt{\kappa}}{1 + \sqrt{\kappa}}, \qquad \kappa = \frac{L}{\mu}.$$

In either case, we prove finite time manifold identification.

Lemma 1. *In the non-strongly convex case, for all k, $-1 \leq \gamma^{(k)} \leq 1$.*

Proof. Note that $\lambda^{(0)} = 0$ and $\lambda^{(k)}$ is monotonically increasing for all k, since

$$\lambda^{(k)} = \frac{1 + \sqrt{1 + 4(\lambda^{(k-1)})^2}}{2} \geq \frac{\sqrt{4(\lambda^{(k-1)})^2}}{2} = \lambda^{(k-1)}.$$

Now, note that

$$-1 \leq \gamma^{(k)} \leq 1 \iff 1 - \lambda^{(k+1)} \leq \lambda^{(k)} \leq 1 + \lambda^{(k+1)}.$$

This is true if for all scalar $x > 0$,

$$1 - \frac{1 + \sqrt{1 + 4x^2}}{2} \leq x \leq 1 + \frac{1 + \sqrt{1 + 4x^2}}{2}$$

which is true if and only if for all $x \geq 0$,

$$\frac{1 - \sqrt{1 + 4x^2}}{2} - x \leq 0, \quad \frac{3 + \sqrt{1 + 4x^2}}{2} - x \geq 0.$$

To see this is satisfied, note that the term $\frac{1-\sqrt{1+4x^2}}{2} - x$ is monotonically decreasing, so it achieves its maximum value of 0 when $x = 0$. Similarly, the term $\frac{3+\sqrt{1+4x^2}}{2} - x$ is also monotonically decreasing (which can be seen by checking the derivative) so it achieves its minimum value of $3/2 > 0$ at the limit when $x \to +\infty$. □

Note that in actuality, with $\lambda^{(0)} = 0$, $-1 < \gamma^{(k)} < 0$ for all $k > 2$, corresponding to extrapolation steps.

Proof of Thm 1

Proof. First we simplify the term in Lemma 1:

$$\begin{aligned}
&\frac{1}{t}(z^{(k)} - x^*) + \nabla g(x^*) \\
= \;& \frac{1}{t}((1-\gamma^{(k)})y^{(k+1)} + \gamma^{(k)} y^{(k)}) - x^*) \\
&+\nabla g(x^*) \\
= \;& \frac{1}{t}((1-\gamma^{(k)})(x^{(k)} - t\nabla g(x^{(k)}) \\
&+\gamma^{(k)}(x^{(k-1)} - t\nabla g(x^{(k-1)})) - x^*) \\
&+\nabla g(x^*) \\
= \;& \frac{1}{t}((1-\gamma^{(k)})x^{(k)} + \gamma^{(k)} x^{(k-1)} - x^*) \\
&-(1-\gamma^{(k)})\nabla g(x^{(k)}) - \gamma^{(k)}\nabla g(x^{(k-1)}) \\
&+\nabla g(x^*)
\end{aligned}$$

Bounding the norms

$$\begin{aligned}
&\left\| \frac{1}{t}(z^{(k)} - x^*) + \nabla g(x^*) \right\|_2 \\
\overset{(a)}{\leq} \;& \frac{1-\gamma^{(k)}}{t}\|x^{(k)} - x^*\|_2 + \frac{\gamma^{(k)}}{t}\|x^{(k-1)} - x^*\|_2 \\
&+(1-\gamma^{(k)})\|\nabla g(x^{(k)}) - \nabla g(x^*)\|_2 \\
&+\gamma^{(k)}\|\nabla g(x^{(k-1)}) - \nabla g(x^*)\|_2 \\
\overset{(b)}{\leq} \;& \frac{1-\gamma^{(k)}}{t}\epsilon(k) + \frac{\gamma^{(k)}}{t}\epsilon(k-1) \\
&+L(1-\gamma^{(k)})\epsilon(k) + L\gamma^{(k)}\epsilon(k-1) \\
\overset{(c)}{\leq} \;& \left(\frac{1}{t} + L\right)\epsilon(k-1)
\end{aligned}$$

where (a) comes from triangle inequality, (b) from L-smoothness, and (c) from the fact that $\epsilon(k)$ is a monotonically decreasing sequence.

The explicit $\bar{k}$ rate for strongly convex $g(x)$ comes from combining this with Table 1. □

A.3 Douglas-Rachford Splitting (DRS)

We give a proof of theorem 2, which gives the manifold identification rate for DRS. In the next section, we show the equivalence of ADMM with DRS, thus extending the complexity result from DRS to ADMM.

Proof. The update $y^{(k+1)}$ is exactly that which solves

$$2x^{(k+1)} - z^{(k)} - y^{(k+1)} = t\nabla g(y^{(k+1)}).$$

Therefore

$$\begin{aligned}
&\|\frac{1}{t}(z^{(k)} - x^*) + \nabla g(x^*)\|_2 \\
= \;& \|\frac{1}{t}(2x^{(k+1)} - y^{(k+1)} - x^*) \\
&-(\nabla g(y^{(k+1)}) - \nabla g(x^*))\|_2 \\
\leq \;& (2/t + L)\epsilon(k).
\end{aligned}$$

The rest follows from Lemma 1, combined with Table 1. □

A.4 Alternating direction method of multipliers (ADMM)

In this section, we briefly elaborate on the details of ADMM. We then show its equivalence of DRS, and extend the manifold identification complexity rate from DRS to ADMM.

First, we show that ADMM on (14) is equivalent to the DRS splitting. This equivalence is well-known, dating back to Gabay (1983). Here, we give a simplified proof for our particular formulation, largely based off the class notes in [1] Writing the augmented Lagrangian of (14) as

$$\mathcal{L}_t(x, y; u) = g(y) + h(x) + u^T(x - y) + \frac{1}{2t}\|x - y\|_2^2,$$

the ADMM algorithm can be summarized via the update scheme

$$\begin{aligned}
x^{(k+1)} &= \mathbf{prox}_{th}(y^{(k)} - tu^{(k)}) &(1)\\
y^{(k+1)} &= \mathbf{prox}_{tg}(x^{(k+1)} + tu^{(k)}) &(2)\\
u^{(k+1)} &= u^{(k)} + \frac{1}{t}(x^{(k+1)} - y^{(k+1)}). &(3)
\end{aligned}$$

Lemma 2. *The DRS method expressed in* (11)-(13) *is equivalent to the ADMM method expressed as* (1)-(3) *given the change of variables* $u^{(k)} = (z^{(k)} - x^{(k)})/t$.

Proof. Recall the DRS scheme for $\rho = 1$ is

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

where we flip h and g. (This does not affect convergence rates.)

We first swap the order of x and z and reindex k

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

Now we swap x and z and don't reindex.

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

[1] `www.seas.ucla.edu/~vandenbe/236C/lectures/dr.pdf`

Now we replace $z^{(k)}$ with $u^{(k)} = (z^{(k)} - x^{(k)})/t$.

$$\begin{aligned} y^{(k+1)} &= \mathbf{prox}_{th}(x^{(k)} - tu^{(k)}) \\ x^{(k+1)} &= \mathbf{prox}_{tg}(y^{(k+1)} + tu^{(k)}) \\ u^{(k+1)} &= u^{(k)} + \frac{1}{t}(y^{(k+1)} - x^{(k+1)}) \end{aligned}$$

This is now exactly the iteration scheme (1)-(3) for ADMM, with y and x flipped. □

We now prove theorem 2 for ADMM.

Proof. The optimality condition for the update of $y^{(k+1)}$ is

$$x^{(k)} + tu^{(k-1)} - y^{(k)} = t\nabla g(y^{(k)}).$$

Therefore,

$$\begin{aligned} & \frac{1}{t}(z^{(k)} - x^*) + \nabla g(x^*) \\ = \; & \frac{1}{t}(y^{(k)} - tu^{(k)} - x^*) + \nabla g(x^*) \\ = \; & \frac{1}{t}(x^{(k)} - x^*) + (u^{(k-1)} - u^{(k)}) \\ & +(\nabla g(x^*) - \nabla g(y^{(k)})) \\ = \; & \frac{1}{t}(x^{(k)} - x^*) + \frac{1}{t}(y^{(k)} - x^{(k)}) \\ & +(\nabla g(x^*) - \nabla g(x^{(k)})) \\ & +(\nabla g(x^{(k)}) - \nabla g(y^{(k)})) \end{aligned}$$

and therefore

$$\left\| \frac{1}{t}(z^{(k)} - x^*) + \nabla g(x^*) \right\|_2 \leq (2/t + 2L)\epsilon(k).$$

The rest follows from Lemma 1 combined with Table 1. □

A.5 Proximal Newton type

We now prove theorem 3, which gives the manifold identification complexity rate for the proximal Newton method.

Proof. Invoking Lemma 1, we just need to find the conditions such that

$$|(H_{\text{est}}(x^{(k)} - x^*) + \nabla g(x^{(k)}) - \nabla g(x^*))_i| \leq \delta_{\min}$$

for all $i \in \mathcal{Z}$.

We see that

$$\begin{aligned} & |(H_{\text{est}}(x^{(k)} - x^*) + \nabla g(x^{(k)}) - \nabla g(x^*))_i| \\ \leq \; & \|H_{\text{est}}(x^{(k)} - x^*)\|_2 + \|\nabla g(x^{(k)}) - \nabla g(x^*)\|_2 \\ \leq \; & (L_H + L)\epsilon(k). \end{aligned}$$

To get a rate of $\bar{k}$ when $g(x)$ is strongly convex, we combine this result with Table 1. □

B Deterministic methods error bounds

In this section, we give detailed rates for variable convergence in deterministic methods (see Table 1).

For DRS and ADMM, two important convergence rates are required here to calculate $\bar{k}$. Taking $\bar{z}$ as the fixed point of the DRS iteration scheme, from Giselsson and Boyd (2017) we have a linear variable convergence rate for the variable error $\|x^{(k+1)} - x^*\|_2 = O(\exp(k))$, and from He and Yuan (2015) we also have another rate for the stationarity of $z^{(k)}$ as $\|x^{(k)} - y^{(k)}\|^2 = O(1/k)$. Both rates are used to calculate $\epsilon_x(k)$.

C Stochastic methods error bounds

C.1 Table of rates

In this section, we give detailed rates for variable convergence rates ($\epsilon_x(k) \geq \|x^{(k)} - x^*\|_2$) and gradient convergence rates ($\epsilon_g(k) \geq \|G_{\text{est}}^{(k)} - \nabla g(x^*)\|_2$) in proximal stochastic methods (see Table 2). In some cases the exact rates were hard to find, so we include our own derivations when necessary.

C.2 Proximal stochastic gradient method

Consider the update scheme

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

We now give a simple proof of the $O(1/\sqrt{k})$ convergence rate of prox-SGD for μ-strongly convex g. The convergence proof for SGD is well-known, and is not new for prox-SGD, but we include it here for completeness.

Lemma 3. *Assume that $\mathbb{E}[G_{\text{est}}^{(k)}] = \nabla g(x^{(k)})$ and there is a V where $\|\nabla g(x^*)\| \leq V$ and $\|G_{\text{est}}^{(k)}\| \leq V$ for all k. Assume that g is strongly convex with modulus μ. Then with a step length sequence $t^{(k)} = 1/(\mu k)$, we have*

$$\mathbb{E}[\|x^{(k)} - x^*\|_2^2] \leq \frac{1}{k}\max\{\|x^{(1)} - x^*\|^2, 4V^2/\mu^2\}. \quad (4)$$

Proof. For ease of notation, we use $t = t^{(k)}$, $G_{\text{est}} = G_{\text{est}}^{(k)}$, $x = x^{(k)}$, and $x^+ = x^{(k+1)}$.

Since $g(x)$ is μ-strongly convex, we have

$$\begin{aligned} g(x^*) - g(x) &\geq \langle \nabla g(x), x^* - x\rangle + \frac{\mu}{2}\|x - x^*\|^2 \\ g(x) - g(x^*) &\geq \langle \nabla g(x^*), x - x^*\rangle + \frac{\mu}{2}\|x - x^*\|^2 \end{aligned}$$

method	Bound on $\|x^{(k)} - x^*\|_2$	Bound on $\|z^{(k)} - z^{(k-1)}\|_2$	stepsize
PG	$(1-\mu/L)^{k/2}B_0$	n/a	$2/L$
aPG	$\frac{1}{\sqrt{\mu}}(1-\sqrt{\mu/L})^{k/2}\sqrt{d_0}$ (Nesterov, 2013a)	n/a	$2/L$
DRS/ADMM	$C_2 C_3^{k+1}$ (Giselsson and Boyd, 2017)	$\sqrt{\frac{C_1}{k+1}}$ (He and Yuan, 2015)	$t > 0$
pN	$(\frac{L_2}{2\mu})^{2^k-1}B_0^{2^k}$ (Thm 3.4 (Lee et al., 2014))	n/a	1
pQN	$2d_0 C_5^k/\mu$ (Ghanbari and Scheinberg, 2016)	n/a	1

Table 1: **Variable convergence rates of deterministic methods.** We assume $g(x)$ is μ-strongly convex, to obtain variable convergence rates more readily. L_2 is the smoothness of the Hessian $\|\nabla^2 g(x) - \nabla^2 g(y)\|_2 \leq L_2\|x-y\|_2$. $B_0 = \|x_0 - x^*\|$. We assume $B_0 < 1$. Otherwise, we can run the scheme until some $\tilde{k}$ where $B_{\tilde{k}} < 1$ and shift all the iterate indices. $d_0 = f(x_0) - f(x^*)$. For DRS and ADMM, $C_1 = \frac{z_0^2 + \|th(z^{(0)}) - th(z^*)\|^2}{4}$, $C_2 = \frac{z_0}{1+t\mu}$, $C_3 = \frac{1+\rho}{2}$ with $\rho = \max\{\frac{tL-1}{tL+1}, \frac{1-t\mu}{t\mu+1}\}$ and $z_0 = \|z^{(0)} - z^*\|$ where $z^{(k)} \to z^*$. $C_4 = 1 - \mu/(\mu+L)$. For pN and pQN, $C_5 = 1 - \mu/(\mu+M)$, and $mI \preceq H_{\text{est}}^{(k)} \preceq MI$.

method	$\mathbb{E}[(\epsilon_x(k))^2]$	Grad. Var.	stepsize
pSGD	$\frac{\max\{B_0^2, 4V^2/\mu^2\}}{k}$	n/a	$1/(\mu k)$
pSVRG	$\rho_{\text{SVRG}}^k(B_0^2 + 2td_0)$ (Poon et al., 2018)	$\mathbb{E}[(\epsilon_g(k))^2] = 4L_Q(d_{k-1} + \rho^k d_0)$	$1/(4L_Q)$ (Xiao and Zhang, 2014)
pSAGA (Defazio et al., 2014)	$\left(1 - \frac{\mu}{2(\mu n + L)}\right)^k B_0^{\text{S}}$	$\epsilon_g(k) = L_{\max}\left(\epsilon_x(k) + 2\epsilon_x(k-m)\right)$	$\frac{1}{2(\mu n+L)}$
pRDA(Xiao, 2010)	$\frac{C_{\text{RDA}}}{\sqrt{k}}$	$\mathbb{E}[\epsilon_g(k)] = \frac{\sigma}{\sqrt{k}} + \frac{C_{\text{RDA}}^g}{k^{1/4}}$	$1/\sqrt{k}$

Table 2: **Variable convergence rates for stochastic methods.** We assume $g(x)$ is μ-strongly convex, to obtain variable convergence rates more readily. $B_0 = \|x_0 - x^*\|_2$. m is number of terms in composite sum and n is the length of x. $d_k = f(x^{(k)}) - f(x^*)$. $L_{\max} = \max_{i=1,\ldots,m} L_i$ where each g_i is L_i-smooth. For prox-SAGA, $B_0^{\text{S}} = (B_0^2 + \frac{n}{\mu n + L}(d_0 - \langle \nabla g(x^*), x^{(0)} - x^* \rangle))$. The gradient variance bound is $\mathbb{E}[\|g_k\|^2] \leq V^2$. For prox-SVRG, we also require $\rho_{\text{SVRG}} = \max\{1 - t\mu, 4Lt(n+1)\} < 1$. For prox-RDA, $C_{\text{RDA}} = 2(h_0 + V^2/\mu^2)$, $C_{\text{RDA}}^g = \frac{4}{3}L_{\max}\sqrt{h_0 + 3V^2/(2\mu^2)}$. and σ^2 is the gradient norm $\|\nabla g_i\|_2$ variance at x^*. For many of these rates, more detailed derivations are given in Appendix C.

and therefore

$$\langle \nabla g(x) - \nabla g(x^*), x - x^* \rangle \geq \mu \|x - x^*\|^2. \quad (5)$$

Now,

$$\begin{aligned} & \|x^+ - x^*\|^2 \\ \overset{(a)}{=} & \|\mathbf{prox}_{th}(x - tG_{\text{est}}) - \mathbf{prox}_{th}(x^* - t\nabla g(x^*))\|^2 \\ \overset{(b)}{\leq} & \|(x - x^*) - t(G_{\text{est}} - \nabla g(x^*))\|^2 \\ = & \|x - x^*\|_2^2 + t^2 \|G_{\text{est}} - \nabla g(x^*)\|^2 \\ & -2t\langle x - x^*, G_{\text{est}} - \nabla g(x^*)\rangle \\ \overset{(c)}{\leq} & \|x - x^*\|_2^2 + 4t^2V^2 - 2t\langle x - x^*, G_{\text{est}} - \nabla g(x^*)\rangle \end{aligned}$$

where (a) comes from the fixed point property $x^* = \mathbf{prox}_{th}(x^* - t\nabla g(x^*))$, (b) from nonexpansiveness of prox, and (c) from the relation

$$\|G_{\text{est}} - \nabla g(x^*)\|_2^2 \leq (\|G_{\text{est}}\|_2 + \|\nabla g(x^*)\|_2)^2 \leq (2V)^2.$$

Taking expectations gives

$$\begin{aligned} & \mathbb{E}[\|x^+ - x^*\|^2 | x] \\ \leq & \|x - x^*\|_2^2 - 2t\langle x - x^*, \mathbb{E}[G_{\text{est}}] - \nabla g(x^*)\rangle \\ & +4t^2V^2 \\ = & \|x - x^*\|_2^2 - 2t\langle x - x^*, \nabla g(x) - \nabla g(x^*)\rangle \\ & +4t^2V^2 \\ \overset{(a)}{\leq} & (1 - 2t\mu)\|x - x^*\|_2^2 + 4t^2V^2. \end{aligned}$$

where (a) is from invoking (5).

Using the nested expectations property, we have that

$$\begin{aligned} \mathbb{E}[\|x^+ - x^*\|_2] &= \mathbb{E}[\mathbb{E}[\|x^+ - x^*\|_2 | x]] \\ &\leq (1 - 2t\mu)\mathbb{E}[\|x - x^*\|_2^2] + 4t^2V^2. \end{aligned}$$

Now the rest follows from induction. Clearly, (4) is satisfied for $k = 1$. Now assume it is satisfied for some k, and consider the $k+1$ term. Take $C = \max\{\|x^{(1)} - x^*\|^2, 4V^2/\mu^2\}$ and $t^{(k)} = 1/(\mu k)$. Then

$$\begin{aligned} & \mathbb{E}[\|x^{(k+1)} - x^*\|_2^2] \\ \leq & (1 - 2/k)\mathbb{E}[\|x - x^*\|_2^2] + 4\mu^2V^2/k^2 \\ \leq & (1 - 2/k)C/k + C/k^2 \\ \leq & \frac{C}{k+1}. \end{aligned}$$

□

C.3 SAGA Gradient error bound

The SAGA method is discussed in Defazio et al. (2014), and $\mathbb{E}[\epsilon_x(k)^2] = O(c^k)$ variable convergence rates for strongly convex g are derived. Here, we give the gradient convergence rates $\epsilon_g(k)$ as a function of $\epsilon_x(k)$, which are needed in the overall manifold identification complexity rate (theorem 4).

Lemma 4. *Suppose that $\epsilon_x(k) \geq \|x^{(k)} - x^*\|_2$, and $L_{\max}$ is the maximum Lipschitz smooth parameter in g_i, e.g.*

$$\|\nabla g_i(x) - \nabla g_i(y)\|_2 \leq L_{\max}\|x - y\|_2, \ \forall i = 1, ..., m.$$

Then in prox-SAGA,

$$\|G_{\text{est}} - \nabla g(x^*)\|_2 \leq L_{\max}\left(\epsilon_x(k) + 2\epsilon_x(k - m)\right).$$

Proof. Define $\hat{x}^{(i)}$ such that $\nabla g_i(\hat{x}^{(i)}) = y_i^{(k-1)}$. Then

$$\begin{aligned} & G_{\text{est}} - \nabla g(x^*) \\ = & \nabla g_{i[k]}(x^{(k)}) - \nabla g_{i[k]}(x^*) \\ & +\nabla g_{i[k]}(x^*) - \nabla g_{i[k]}(\hat{x}^{(i[k])}) \\ & +\frac{1}{m}\sum_{i=1}^m (\nabla g_i(\hat{x}^{(i)}) - \nabla g_i(\hat{x}^*)). \end{aligned}$$

Therefore

$$\begin{aligned} & \|G_{\text{est}} - \nabla g(x^*)\|_2 \\ \overset{(a)}{\leq} & \|\nabla g_{i[k]}(x^{(k)}) - \nabla g_{i[k]}(x^*)\|_2 \\ & +\|\nabla g_{i[k]}(x^*) - \nabla g_{i[k]}(\hat{x}^{(i[k])})\|_2 \\ & +\frac{1}{m}\sum_{i=1}^m \|\nabla g_i(\hat{x}^{(i)}) - \nabla g_i(x^*)\|_2 \\ \overset{(b)}{\leq} & L_{\max}\left(\|x^{(k)} - x^*\|_2 + \|x^* - \hat{x}^{(i[k])})\|_2\right) \\ & +\frac{L_{\max}}{m}\sum_{i=1}^m \|\hat{x}^{(i)} - x^*\|_2 \\ \overset{(c)}{\leq} & L_{\max}(\epsilon_x(k) + 2\epsilon_x(k - m)). \end{aligned}$$

where (a) is from triangle inequality applied to each summed term, (b) uses $L_{\max}$ smoothness of each component term g_i, and (c) is since $\epsilon_x(k)$ is a monotonically decreasing sequence and $\hat{x}^{(i)}$ is at most m iterations stale. □

where (a) comes from triangle inequality

C.4 RDA Gradient error bound

In this section, we list and rework known variable convergence rates of pRDA in Xiao (2010) in order to derive a variable and gradient convergence rate.

Lemma 5. *For any random variable Z, $\mathbb{E}[Z^2] \geq \mathbb{E}[Z]^2$.*

Proof. The variance $\mathbb{E}[(Z-\mathbb{E}[Z])^2] = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 \geq 0$, so $\mathbb{E}[Z^2] \geq \mathbb{E}[Z]^2$. □

Lemma 6. *Assume g is μ-strongly convex. Using $t^{(k)} = \mu/\sqrt{k}$, for general convex h, the expected variable convergence rate is*

$$\mathbb{E}[\|x^{(t+1)} - x^*\|^2] \leq \frac{C_{RDA}}{\sqrt{k}}$$

where $C_{\text{RDA}} = 2(h_0 + V^2/\mu^2)$, $V^2 = \max_{i,x} \mathbb{E}[\|\nabla g_i(x)\|^2]$, and $h_0 = \max_{x \in D} h(x)$ for some reasonable domain D.

This result was first presented and proven in Xiao (2010). We list it here to show the simplification steps needed in our gradient bound.

Proof. Using $t^{(k)} = \mu/\sqrt{k}$, we have

$$\begin{aligned}\mathbb{E}[\|x^{(t+1)} - x^*\|^2] &\overset{(a)}{\leq} \frac{2(\mu h_0 + V^2/\mu)}{\mu(k + 1/\sqrt{k})}\sqrt{k} \\ &\leq 2(h_0 + V^2/\mu^2)\frac{1}{\sqrt{k}}\end{aligned}$$

where (a) is the exact rate reported in Xiao (2010). □

Lemma 7. *Assume that $\mathbb{E}[\|\nabla g_i(x^*) - \nabla g(x^*)\|^2] \leq \sigma^2$ (e.g. the gradients have bounded variance) and bounded norm $\mathbb{E}[\|\nabla g_i(x)\|_2^2] \leq V^2$ for all i, x. Take $d_0 = \max_{x \in \mathcal{D}} h(x)$ where $\mathcal{D}$ is a bounded region where all iterates reside. Assume that there is a maximal Lipschitz constant for each component, e.g.*

$$\|\nabla g_i(x) - \nabla g_i(y)\| \leq L_{\max}\|x - y\|_2, \ \forall i = 1, ..., m.$$

Then

$$\mathbb{E}[\|G_{\text{est}} - \nabla g(x^*)\|_2] \leq \frac{\sigma}{\sqrt{k}} + \frac{C^g_{\text{RDA}}}{k^{1/4}}.$$

where

$$C^g_{\text{RDA}} = \frac{4}{3} L_{\max}\sqrt{h_0 + 3V^2/(2\mu^2)}.$$

Proof.

$$\begin{aligned}& G_{\text{est}} - \nabla g(x^*) \\ = \ & \frac{1}{t^{(k)}k}x^{(k)} + \frac{1}{k}\sum_{j=1}^{k} \nabla g_{i[j]}(x^{(j)}) - \nabla g(x^*) \\ = \ & \frac{1}{k}\sum_{j=1}^{k} \nabla g_{i[j]}(x^{(j)}) - \frac{1}{k}\sum_{j=1}^{k} \nabla g_{i[j]}(x^*) \\ & + \frac{1}{k}\sum_{j=1}^{k} \nabla g_{i[j]}(x^*) - \nabla g(x^*) + \frac{1}{t^{(k)}k}x^{(k)} + \\ = \ & A + B + C\end{aligned}$$

where

$$\begin{aligned}A &= \frac{1}{k}\sum_{j=1}^{k} \nabla g_{i[j]}(x^{(j)}) - \frac{1}{k}\sum_{j=1}^{k} \nabla g_{i[j]}(x^*), \\ B &= \frac{1}{k}\sum_{j=1}^{k} \nabla g_{i[j]}(x^*) - \nabla g(x^*), \\ C &= \frac{1}{t^{(k)}k}x^{(k)}.\end{aligned}$$

Since by triangle in equality $\|G_{\text{est}} - \nabla g(x^*)\|_2 \leq \|A\|_2 + \|B\|_2 + \|C\|_2$, we have

$$\mathbb{E}[\|G_{\text{est}} - \nabla g(x^*)\|_2] \leq \mathbb{E}[\|A\|_2] + \mathbb{E}[\|B\|_2] + \mathbb{E}[\|C\|_2]$$

by linearity.

We now bound each term. If i is sampled uniformly, taking $X = \nabla g_i(x^*)$ as a random vector in $\mathbb{R}^n$, then

$$\mathbb{E}[X] = \nabla g(x^*), \quad \mathbb{E}[\|X - \mathbb{E}[X]\|_2^2] = \sigma^2$$

then

$$\mathbb{E}[\|B\|_2^2] = \mathbb{E}[\|\bar{X}^{(k)} - \mathbb{E}[X]\|_2^2] = \frac{\sigma^2}{k},$$

where $\bar{X}^{(k)} = \frac{1}{k}\sum_{j=1}^{k} \nabla g_{i[j]}(x^*)$ the sample mean of X with k samples.

Now taking the random variable $Z = \|X^{(k)} - \mathbb{E}[X]\|_2$, we also use Lemma 5 and bound

$$\mathbb{E}[Z]^2 \leq \mathbb{E}[Z^2] = \frac{\sigma^2}{k}$$

and thus

$$\mathbb{E}[\|B\|_2] = \mathbb{E}[Z] \leq \frac{\sigma}{\sqrt{k}}.$$

Also,

$$\begin{aligned}\|A\|_2 &\leq \frac{1}{k}\sum_{j=1}^{k} \|\nabla g_{i[j]}(x^{(j)}) - \nabla g_{i[j]}(x^*)\|_2 \\ &\leq \frac{L_{\max}}{k}\sum_{j=1}^{k} \|x^{(j)} - x^*\|_2\end{aligned}$$

Taking $Z = \|x^{(j)} - x^*\|_2$ and using Lemma 5, we have $\mathbb{E}[\|x^{(j)} - x^*\|_2] \leq \sqrt{\mathbb{E}[\|x^{(j)} - x^*\|_2^2]}$, which is the square root of the reported quantities in literature, given in table (2). Now

$$\begin{aligned}\mathbb{E}[\|A\|_2] &\leq \frac{L_{\max}}{k}\sum_{j=1}^{k} \mathbb{E}[\|x^{(j)} - x^*\|_2] \\ &\leq \frac{L_{\max}}{k}\sum_{j=1}^{k} \sqrt{\mathbb{E}[\|x^{(j)} - x^*\|_2^2]} \\ &\overset{(a)}{\leq} \frac{L_{\max}\sqrt{C_{\text{RDA}}}}{k}\sum_{j=1}^{k} \frac{1}{j^{1/4}}\end{aligned}$$

where we use Lemma 6 in (a). Now, using a comparison test, we have

$$\begin{aligned}\sum_{j=1}^{k}\frac{1}{j^{1/4}} &\leq 1+\int_1^{k+1}\frac{1}{t^{1/4}}dt\\ &= \frac{4}{3}(k+1)^{3/4}-\frac{1}{3}\\ &\leq \frac{4}{3}k^{3/4}\end{aligned}$$

and thus

$$\mathbb{E}[\|A\|_2]\leq\frac{4}{3}L_{\max}\sqrt{C_{\text{RDA}}}k^{-1/4}.$$

This gives the result.

□

D $\delta_{\min}$ propositions

In this section, we first show that when A contains repeated column entries i and j, then the solution is degenerate if $x_i = 0 \neq x_j$. Then we discuss and give proofs to the propositions in Section 4 involving $\delta_{\min}$'s statistical properties. We first prove a bound for sparse linear regression, then sparse logistic regression, then the dual of support vector machines.

Lemma 8. *Denote a_j the jth column of A. If $a_i = a_j$, then for $g(x) = \mathcal{L}(Ax; b)$ any loss function, $\delta_i^* = \delta_j^*$.*

Proof. Without loss of generality, assume $i = 1$ and $j = 2$. Then $A = [a, a, \tilde{A}]$ and

$$Ax = \tilde{A}\tilde{x} + \bar{x}a, \qquad \bar{x} = x_i + x_j.$$

Then for $d(z) = \mathcal{L}(\tilde{A}z + \bar{x}a; b)$

$$(\nabla g(x))_j = a^T\nabla_z d(z) = (\nabla g(x))_i.$$

Then $\delta_i^* = \delta_j^*$. □

Corollary 1. *If $a_i = a_j$ and $x_i = 0$ but x_j is nonzero, then x_i is necessarily degenerate.*

Proof of Proposition 1.

Proof. This readily comes from the definition, triangle inequalities, and Cauchy Schwartz. For all $j \in \mathcal{Z}$,

$$\begin{aligned}\delta_j^* &= \lambda-\frac{1}{m}|a_j^T(Ax-b)|\\ &= \lambda-\frac{1}{m}|a_j^T(Ax-(Ax^{\#}-y))|\\ &\geq \lambda-\frac{1}{m}|a_j^TAe|-\frac{1}{m}|a_j^Ty|\\ &\geq \lambda-\rho\|e\|_1-\frac{1}{m}\|a_j\|_2^2|e_j|-\frac{1}{m}\|a_j\|_2\|y\|_2\\ &\geq \lambda-\rho\|e\|_1-\alpha^2|e_j|-\alpha\eta.\end{aligned}$$

□

Proof of Proposition 2.

Proof. Taking $z_i = \sigma(a_i^Tx)$, we can write the gradient of g succinctly as

$$\nabla g(x) = A^T(z-b).$$

(As an abuse of notation, define the vector mapping $\sigma:\mathbb{R}^n\to\mathbb{R}^n$ as $\sigma(x)_i = \sigma(x_i)$.) Now for any $j \in \mathcal{Z}$,

$$\begin{aligned}\delta_j^* &= \lambda-|\frac{1}{m}a_j^T(\sigma(Ax^*)-b)|\\ &= \lambda-|\frac{1}{m}a_j^T(\sigma(Ax^*)-\sigma(Ax^{\#})-y)|.\end{aligned}$$

We can approximate the right hand side using a first-order Tayler series on σ. Taking D an $m\times m$ diagional matrix with $D_{ii} = \sigma(a_i^Tx^*)(1-\sigma(a_i^Tx^*))$ then the first order linearization of $\sigma(Ax^{\#})$ from a reference point of x^* is

$$\sigma(Ax^{\#}) = \sigma(Ax^*) + A^TDe + O(\|e\|^2).$$

Let us consider the regime in which $\|e\|$ is very small, e.g. $x^* \approx x^{\#}$. Then the lower bound on δ_j^* can be approximated, so that

$$\begin{aligned}\delta_j^* &\gtrapprox \lambda-|-\frac{1}{m}a_j^T(DAe)+a_j^Ty|\\ &\geq \lambda-\frac{1}{m}|a_j^T(DAe)|-\frac{1}{m}|a_j^Ty|.\end{aligned}$$

Note that for all i, $0 \leq D_{ii} \leq \tau$. For any vector a and b, where $(a\circ b)_i = a_ib_i$, we have

$$|a^TDb|\overset{(a)}{\leq}\tau\|a\circ b\|_1\overset{(b)}{\leq}\tau\|a\|_2\|b\|_2$$

where (a) is by Holder's inequality (since if $D = \mathbf{diag}(d)$ then $|a^TDb| = d^T(a\circ b)$, and (b) is by Cauchy-Schwartz. Therefore

$$|a_j^T(DAe)|\leq\tau\alpha^2\|e\|_2m$$

and

$$\delta_j^*\gtrapprox\lambda-\tau\alpha^2\|e\|_2-\alpha\eta.$$

□

Proof of Proposition 3.

Proof. Define $\tilde{K} = \mathbf{diag}(b)K\mathbf{diag}(b)$. Then the objective of (20) is

$$g(x) = \frac{1}{2m}x^T\tilde{K}x-\frac{1}{m}x^T\mathbf{1}$$

with gradient

$$\nabla g(x) = \frac{1}{m}(\tilde{K}x-\mathbf{1}).$$

Note first that for all $j \in \mathcal{Z}$, $x_j^* \in \{0, \lambda\}$, and at $x = x^*$, each descent direction must necessarily try to leave the feasible region. Therefore,

$$-\nabla g(x^*)_j \leq 0 \quad \text{if} \quad x_j^* = 0,$$

and

$$-\nabla g(x^*)_j \geq 0 \quad \text{if} \quad x_j^* = \lambda.$$

Therefore

$$\begin{aligned}
\delta_j^* &= |\nabla g(x^*)_j| \\
&= \begin{cases} (\nabla g(x^*))_j, & x_j^* = 0 \\ -(\nabla g(x^*))_j, & x_j^* = \lambda. \end{cases} \\
&= \begin{cases} \dfrac{1}{m}\left(\tilde{K}_{jj}x_j^* + \displaystyle\sum_{i\neq j}\tilde{K}_{ij}x_i^* - 1\right), & x_j^* = 0 \\ \dfrac{1}{m}\left(1 - \tilde{K}_{jj}x_j^* - \displaystyle\sum_{i\neq j}\tilde{K}_{ij}x_i^*\right), & x_j^* = \lambda. \end{cases} \\
&= \begin{cases} \dfrac{1}{m}\left(\displaystyle\sum_{i\neq j}\tilde{K}_{ij}x_i^* - 1\right), & x_j^* = 0 \\ \dfrac{1}{m}\left(1 - \tilde{K}_{jj}\lambda - \displaystyle\sum_{i\neq j}\tilde{K}_{ij}x_i^*\right), & x_j^* = \lambda. \end{cases}
\end{aligned}$$

When $x_j^* = 0$, the term $\sum_{i\neq j}\tilde{K}_{ij}x_i^*$ can be arbitrarily close to 0, and thus no interesting lower bound for δ_j^* can be made. When $x_j^* = \lambda$, note that

$$|\tilde{K}_{jj}x_j + \sum_{i\neq j}\tilde{K}_{ij}x_j| \leq \alpha\lambda + \rho\lambda m$$

and thus

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

□

References

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.

Gabay, D. (1983). Chapter ix applications of the method of multipliers to variational inequalities. In *Studies in mathematics and its applications*, volume 15, pages 299–331. Elsevier.

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.

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

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.

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.

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

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

Xiao, L. and Zhang, T. (2014). A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075.