Supplementary Material for "Local Convergence Properties of SAGA/Prox-SVRG and Acceleration"

Clarice Poon [1] **Jingwei Liang** [1] **Carola-Bibiane Schönlieb** [1]

Abstract

In this supplementary material, we provide not only the proofs of the main theorems of the submitted manuscript, and moreover numerical examples to demonstrate: proximal stochastic gradient descent has no manifold identification property, the consequence of the non-degeneracy condition being not satisfied, and rate estimations comparison between deterministic Forward–Backward splitting algorithm and SAGA/Prox-SVRG.

1. Prox-SGD has no manifold identification properties

We present a simple example to illustrate the fact the Prox-SGD cannot have manifold identification properties in general. Consider the following minimisation problem

$$\min_{x\in\mathbb{R}^3} \frac{1}{3}\|x\|_1 + \frac{1}{3}\sum_{i=1}^3 \frac{1}{2}\|\mathcal{K}_i x - b_i\|^2,$$

where

$$\mathcal{K} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & \sqrt{3} \end{bmatrix} \quad \text{and} \quad b = \begin{pmatrix} 2 \\ \sqrt{2}/3 \\ \sqrt{3}/4 \end{pmatrix}.$$

The optimal solution is $x^\star = (1,0,0)^T$ and writing $F(x) \stackrel{\text{def}}{=} \frac{1}{6}\|\mathcal{K}x - b\|^2$, we have that the non-degeneracy condition

$$-\nabla F(x^\star) = \frac{1}{3}\begin{pmatrix} 1 \\ \sqrt{2}/3 \\ \sqrt{3}/4 \end{pmatrix} \in \mathrm{ri}\big(\tfrac{1}{3}\partial\|x^\star\|_1\big), \quad \text{where} \quad (\partial\|x^\star\|)_i = \mathrm{sign}(x_i) = \begin{cases} +1 & : x_i > 0, \\ [-1,+1] & : x_i = 0, \\ -1 & : x_i < 0, \end{cases}$$

It is straightforward to verify that $\|\nabla f_i(x) - \nabla F(x)\| \geq \|\nabla F(x)\|$ for all $i = 1,2,3$. Moreover, if Prox-SGD is starting with $x_0 = (\mu,0,0)^T$ with $\mu \in \mathbb{R}$, then with probability $2/3$ the first iterate of the algorithm satisfies $x_1 \notin \mathcal{M}_{x^\star} = \big\{(x,0,0) : x \in \mathbb{R}\big\}$. In fact, x_1 will have 2 non-zero entries if $|\mu| > \gamma_1$ and $i_1 \in \{2,3\}$. Figure 1.1 shows the support sizes of the Prox-SGD iterates over 10^6 iterations.

2. Global convergence of SAGA/Prox-SVRG

To prove Theorem 2.1 and 2.2, the lemma below is needed which is classical result from stochastic analysis (Neveu, 1975).

Lemma 2.1 (Supermartingale convergence). *Let Y_k, Z_k and W_k, $k = 0,1,\ldots$, be three sequences of random variables and let $\mathcal{F}_k$, $k = 0,1,\ldots$, be sets of random variables such that $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k. Suppose that:*

(i) *The random variables Y_k, Z_k and W_k are non-negative, and are functions of the random variables in $\mathcal{F}_k$.*

(ii) *For each k, we have $\mathbb{E}(Y_{k+1}|\mathcal{F}_k) \leq Y_k - Z_k + W_k$.*

[1]DAMTP, University of Cambridge, Cambridge, United Kingdom. Correspondence to: Jingwei Liang <jl993@cam.ac.uk>, Clarice Poon <C.M.H.S.Poon@maths.cam.ac.uk>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

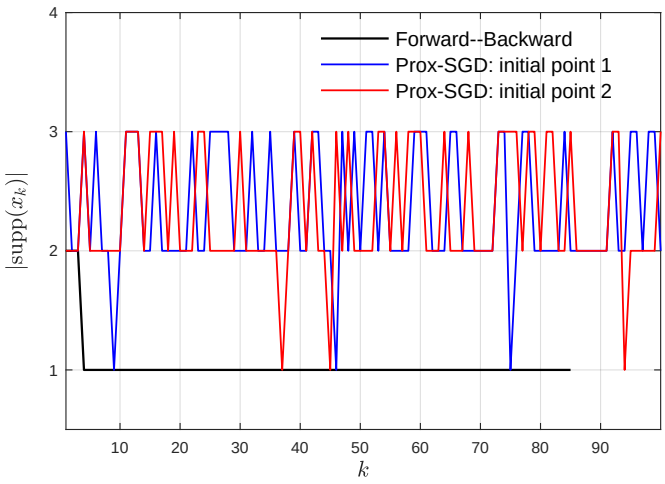


Figure 1.1: Support identification comparison between FB and Prox-SGD. For Prox-SGD, "initial point 1" starts with an arbitrary point with all three elements non-zero; "initial point 2" starts with the point $10x^\star$. The maximum number of iteration for Prox-SGD is 10^6, the blue and red lines are sub-sampled, one out of every 10^4 points.

(iii) *With probability 1,* $\sum_k W_k < \infty$.

Then we have $\sum_k Z_k < \infty$ *and the sequence* Y_k *converges to a non-negative random variable* Y *with probability 1.*

Proof of Theorem 2.1. The convergence of the objective function value for $\gamma_k \equiv \frac{1}{3L}$ is already studied in (Defazio et al., 2014), here for the completeness of the proof, we shall keep the convergence proof of the objective function.

The proof of the theorem consists of several steps. First is the convergence of the objective function value. Let $\phi_{k,i}$ be the point such that $g_{k,i} = \nabla f_i(\phi_{k,i})$, then following the proof in the original SAGA paper (Defazio et al., 2014), define the following Lyapunov function $\mathcal{L}$,

$$\mathcal{L}_k \overset{\text{def}}{=} \mathcal{L}(x_k, \{\phi_{k,i}\}_{i=1}^m) \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^m f_i(\phi_{k,i}) - F(x^\star) - \frac{1}{m}\sum_{i=1}^m \langle \nabla f_i(x^\star),\, \phi_{i,k} - x^\star\rangle + c\|x_k - x^\star\|^2$$

for some appropriate $c > 0$. Denote $\mathbb{E}_k[\cdot]$ the conditional expectation on step k. Then following the Appendix C of the supplementary material of (Defazio et al., 2014), one can show that

$$\mathbb{E}_k[\mathcal{L}_{k+1}] \leq \mathcal{L}_k - \frac{1}{4m}\mathbb{E}_k[\Phi(x_{k+1}) - \Phi(x^\star)]. \tag{2.1}$$

Since $\mathbb{E}_k[\Phi(x_{k+1}) - \Phi(x^\star)]$ is a non-negative random variable of the k^{th} iteration, it then follows that $\{\mathcal{L}_k\}_{k\in\mathbb{N}}$ is a supermartingale owing to Lemma 2.1. Therefore $\{\mathcal{L}_k\}_{k\in\mathbb{N}}$ converges to a non-negative random variable $\mathcal{L}^\star$ with probability 1. At the same time, with probability 1, $\|x_k - x^\star\|^2 \leq \frac{1}{c}\mathcal{L}_k$, hence $\{x_k\}_{k\in\mathbb{N}}$ is a bounded sequence and every cluster point of $\{x_k\}_{k\in\mathbb{N}}$ is a global minimiser of Φ. Moreover, from Lemma 2.1 and (2.1), we have

$$\sum_{k=0}^\infty \big(\mathbb{E}_k[\Phi(x_{k+1}) - \Phi(x^\star)]\big) \leq \mathcal{L}_0 < +\infty$$

holds almost surely. Define a new random variable $y_j \overset{\text{def}}{=} \sum_{k\geq j}\mathbb{E}_k[\Phi(x_{k+1}) - \Phi(x^\star)]$, clearly we have $\{y_j\}_{k\in\mathbb{N}}$ is non-increasing and converges to 0 as $j \to +\infty$. As a consequence, by the monotone convergence theorem, we have

$$0 = \mathbb{E}\Big[\lim_{j\to+\infty} y_j\Big] = \lim_{j\to+\infty}\mathbb{E}[y_j] = \lim_{j\to+\infty}\sum\nolimits_{k\geq j}\mathbb{E}[\Phi(x_{k+1}) - \Phi(x^\star)] = \lim_{j\to+\infty}\mathbb{E}\Big[\sum\nolimits_{k\geq j}\left(\Phi(x_{k+1}) - \Phi(x^\star)\right)\Big],$$

which implies

$$\mathbb{E}\Big[\sum\nolimits_{k\geq j}\left(\Phi(x_{k+1}) - \Phi(x^\star)\right)\Big] < +\infty \Longrightarrow \sum\nolimits_k\big(\Phi(x_{k+1}) - \Phi(x^\star)\big) < +\infty \text{ almost surely}, \tag{2.2}$$

hence $\Phi(x_k) \to \Phi(x^\star)$ almost surely.

With the boundedness of $\{x_k\}_{k\in\mathbb{N}}$, the second step is to prove that $\{\|x_k - x^\star\|\}_{k\in\mathbb{N}}$ is convergent. Define a new sequence

$$w_k \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^m f_i(\phi_{k,i}) - F(x^\star) - \frac{1}{m}\sum_{i=1}^m \langle \nabla f_i(x^\star),\, \phi_{i,k} - x^\star\rangle.$$

Observe that

$$\mathbb{E}_k[w_{k+1}] = \frac{1}{m}F(x_k) - F(x^\star) - \frac{1}{m}\langle \nabla F(x^\star),\, x_k - x^\star\rangle + \big(1 - \frac{1}{m}\big)w_k.$$

Since $x^\star \in \operatorname{Argmin}(\Phi)$ is a global minimiser, we have $-\nabla F(x^\star) \in \partial R(x^\star)$ and $\langle -\nabla F(x^\star),\, x_k - x^\star\rangle \leq R(x_k) - R(x^\star)$, therefore from above equality we further obtain

$$\mathbb{E}_k[w_{k+1}] \leq \frac{1}{m}\big(\Phi(x_k) - \Phi(x^\star)\big) + \big(1 - \frac{1}{m}\big)w_k.$$

Taking expectations over all previous steps for both sides and summing from $k = 0$ to j yields

$$\mathbb{E}[w_{j+1}] + \frac{1}{m}\sum_{k=1}^j \mathbb{E}[w_k] \leq \frac{1}{m}\sum_{k=0}^j \mathbb{E}\big[\Phi(x_k) - \Phi(x^\star)\big] + \big(1 - \frac{1}{m}\big)\mathbb{E}[w_0].$$

As a result, taking j to $+\infty$ implies that $\mathbb{E}[\sum_{k=1}^j w_k] < +\infty$, hence $\sum_{k=1}^j w_k < +\infty$ almost surely. Moreover, $w_k \to 0$ with probability 1. From the convergence result of $\{\mathcal{L}_k\}_{k\in\mathbb{N}}$ and $\{w_k\}_{k\in\mathbb{N}}$, we have that almost surely $\{\|x_k - x^\star\|\}_{k\in\mathbb{N}}$ is bounded and convergent.

Next we prove the almost sure convergence of the sequence $\{x_k\}_{k\in\mathbb{N}}$. Let $\{x_i^\star\}_i$ be a countable subset of the relative interior $\operatorname{ri}(\operatorname{Argmin}(\Phi))$ that is dense in $\operatorname{Argmin}(\Phi)$. From the almost sure convergence of $\|x_k - x^\star\|, x^\star \in \operatorname{Argmin}(\Phi)$, we have that for each i, the probability $\operatorname{Prob}(\{\|x_k - x_i^\star\|\}_{k\in\mathbb{N}}$ is not convergent$) = 0$. Therefore

$$\begin{aligned}\operatorname{Prob}\big(\forall i, \exists b_i \text{ s.t. } \lim_{k\to+\infty}\|x_k - x_i^\star\|\big) &= 1 - \operatorname{Prob}(\{\|x_k - x_i^\star\|\}_{k\in\mathbb{N}} \text{ is not convergent})\\ &\geq 1 - \textstyle\sum_i \operatorname{Prob}(\{\|x_k - x_i^\star\|\}_{k\in\mathbb{N}} \text{ is not convergent}) = 1,\end{aligned}$$

where the inequality follows from the union bound, *i.e.* for each i, $\{\|x_k - x_i^\star\|\}_{k\in\mathbb{N}}$ is a convergent sequence. For a contradiction, suppose that there are convergent sub-sequences $\{u_{k_j}\}_{k_j}$ and $\{v_{k_j}\}_{k_j}$ of $\{x_k\}_{k\in\mathbb{N}}$ which converge to their limiting points $u^\star$ and $v^\star$ respectively, with $\|u^\star - v^\star\| = r > 0$. Since $\Phi(x_k)$ converges to $\inf\Phi$, these two limiting points are necessarily in $\operatorname{Argmin}(\Phi)$. Since $\{x_i^\star\}_i$ is dense in $\operatorname{Argmin}(\Phi)$, we may assume that for all $\epsilon > 0$, we have $x_{i_1}^\star$ and $x_{i_2}^\star$ are such that $\|x_{i_1}^\star - u^\star\| < \epsilon$ and $\|x_{i_2}^\star - v^\star\| < \epsilon$. Therefore, for all k_j sufficiently large,

$$\|u_{k_j} - x_{i_1}^\star\| \leq \|u_{k_j} - u^\star\| + \|u^\star + x_{i_1}^\star\| < \|u_{k_j} - u^\star\| + \epsilon.$$

On the other hand, for sufficiently large j, we have

$$\|v_{k_j} - x_{i_1}^\star\| \geq \|v^\star - u^\star\| - \|u^\star - x_{i_1}^\star\| - \|v_{k_j} - v^\star\| > r - \epsilon - \|v_{k_j} - v^\star\| > r - 2\epsilon.$$

This contradicts with the fact that $x_k - x_{i_1}^\star$ is convergent. Therefore, we must have $u^\star = v^\star$, hence there exists $\bar{x} \in \operatorname{Argmin}(\Phi)$ such that $x_k \to \bar{x}$.

Finally, to see that $\varepsilon_k^{\text{SAGA}} \to 0$, from Lemma 6 of (Defazio et al., 2014),

$$\frac{1}{m}\sum_{i=1}^m \|\nabla f_i(\phi_{k,i}) - \nabla f_i(x^*)\|^2 \leq 2Lw_k \to 0,$$

therefore, combining this with the fact that ∇f_j is L-Lipschitz and $x_k \to x^*$, it follows that

$$\|\varepsilon_k^{\text{SAGA}}\| \leq \|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(\phi_{k,i})\| + \frac{1}{m}\sum_{j=1}^m \|\nabla f_j(\phi_{k,i}) - \nabla f_j(x_k)\| \to 0,$$

which concludes the proof. □

To prove Theorem 2.2, we require the following lemma, which is a direct consequence of Eq. (16) and Corollary 3 of (Xiao & Zhang, 2014).

Lemma 2.2. *Assume that F is α_F-strongly convex and R is α_R-strongly convex. Let $\{x_{\ell,p}\}_{\ell,p}$ be the sequence generated by Prox-SVRG. Then, conditional on step $k = \ell P + p - 1$, we have*

$$\begin{aligned}&(1+\gamma\alpha_R)\mathbb{E}_k[\|x_{\ell,p}-x^\star\|^2]\\&\leq (1-\gamma\alpha_F)\|x_{\ell,p-1}-x^\star\|^2 - 2\gamma\big(\Phi(x_{\ell,p})-\Phi(x^\star)\big) + 8L\gamma^2\left(\Phi(x_{\ell,p-1})-\Phi(x^\star)+\Phi(\tilde{x}_\ell)-\Phi(x^\star)\right).\end{aligned} \tag{2.3}$$

Proof of Theorem 2.2. We begin with the remark that following the arguments in the proof of Theorem 2.1, to show that $x_{\ell,p} \to x^\star$ almost surely for some $x^\star \in \operatorname{argmin}(\Phi)$, it is sufficient to prove that $\|x_{\ell,p} - x^\star\|$ is convergent. By Lemma 2.2 with $\alpha_R = \alpha_F = 0$, we have that conditional on step $k = \ell P + p - 1$,

$$\mathbb{E}_k[\|x_{\ell,p}-x^\star\|^2] + 2\gamma\mathbb{E}_k[\Phi(x_{\ell,p})-\Phi(x^\star)] \leq \|x_{\ell,p-1}-x^\star\|^2 + 8L\gamma^2\left(\Phi(x_{\ell,p-1})-\Phi(x^\star)+\Phi(\tilde{x}_\ell)-\Phi(x^\star)\right). \tag{2.4}$$

Summing (2.4) over $p = 1, \ldots, P$ and taking expectation on the random variables $i_1, \ldots, i_P$, we obtain that

$$\begin{aligned}&\mathbb{E}[\|x_{\ell,P}-x^\star\|^2] + 2\gamma\mathbb{E}[\Phi(x_{\ell,P})-\Phi(x^\star)] + 2\gamma(1-4L\gamma)\textstyle\sum_{j=1}^{P-1}\mathbb{E}[\Phi(x_{\ell,j})-\Phi(x^\star)]\\&\leq \|\tilde{x}_\ell-x^\star\|^2 + 8L\gamma^2(P+1)\big(\Phi(\tilde{x}_\ell)-\Phi(x^\star)\big).\end{aligned} \tag{2.5}$$

Since $\gamma \leq \frac{1}{4L(P+2)}$, which yields $2\gamma(1-4L\gamma) \geq \gamma^2$, we obtain from (2.5)

$$\begin{aligned}&\mathbb{E}[\|x_{\ell,P}-x^\star\|^2] + (2\gamma-\gamma^2)\mathbb{E}[\Phi(x_{\ell,P})-\Phi(x_*)] + \gamma^2\textstyle\sum_{j=1}^{P}\mathbb{E}[\Phi(x_{\ell,j})-\Phi(x^\star)]\\&\leq \|\tilde{x}_\ell-x^\star\|^2 + 8L\gamma^2(P+1)(\Phi(\tilde{x}_\ell)-\Phi(x^\star)).\end{aligned}$$

Moreover, under "Option I", by defining the non-negative random variables

$$T_\ell \overset{\text{def}}{=} \|\tilde{x}_\ell-x^\star\|^2 + (2\gamma-\gamma^2)(\Phi(\tilde{x}_\ell)-\Phi(x^\star)) \;\text{ and }\; S_{\ell+1} \overset{\text{def}}{=} \textstyle\sum_{j=1}^{P}\big(\Phi(x_{\ell,j})-\Phi(x^\star)\big).$$

It follows from $8L\gamma^2(P+1) \leq 2\gamma - \gamma^2$ that

$$\mathbb{E}[T_{\ell+1}] \leq T_\ell - \gamma^2\mathbb{E}[S_{\ell+1}]. \tag{2.6}$$

So, by the super-martingale convergence theorem, $\{T_\ell\}_{\ell\in\mathbb{N}}$ converges to a non-negative random variable and $\sum_\ell S_\ell < +\infty$ holds almost surely. In particular, we have $S_\ell \to 0$ as $\ell \to \infty$ and hence, $\Phi(\tilde{x}_\ell) \to \Phi(x^\star)$ as $\ell \to \infty$. Therefore, $\|\tilde{x}_\ell - x^\star\|^2$ converges almost surely. Following the proof of Theorem 2.1, we can then show that $\tilde{x}_\ell$ converges to an optimal point $x^\star$ almost surely.

Now we prove that the inner iteration sequence $\{x_{\ell,p}\}_{1\leq p\leq P,\, \ell\in\mathbb{N}}$ also converge to $x^\star$ as $\ell \to \infty$. Consider the inequality (2.4), and define the non-negative random variables

$$V_{\ell,p} \overset{\text{def}}{=} \|x_{\ell,p}-x^\star\|^2 + 2\gamma\big(\Phi(x_{\ell,p})-\Phi(x^\star)\big) \;\text{ and }\; W_{\ell,p} \overset{\text{def}}{=} 8L\gamma^2\big(\Phi(\tilde{x}_\ell)-\Phi(x^\star)\big). \tag{2.7}$$

Equation (2.4) implies that

$$\mathbb{E}[V_{\ell,p}] \leq V_{\ell,p-1} + W_{\ell,p-1},$$

and moreover $\sum_{\ell,p} W_{\ell,p} = \sum_\ell S_\ell < \infty$ holds almost surely. Therefore, the super martingale convergence theorem implies that $\{V_{\ell,p}\}_{p\in\{1,\cdots,P\},\ell\in\mathbb{N}}$ converges to a non-negative random variable. Moreover, since $\Phi(x_{\ell,p}) \to \Phi(x^\star)$, it follows that the sequence $\{\|x_{\ell,p}-x^\star\|\}_{p\in\{1,\cdots,P\},\ell\in\mathbb{N}}$ is convergent.

To prove the ergodic convergence rate of $\Phi(\bar{x}_k) - \Phi(x^\star)$, observe that by convexity of Φ and Jensen's inequality, we have

$$\mathbb{E}[S_{\ell+1}] \geq P\mathbb{E}\left[\Phi\left(\tfrac{1}{P}\textstyle\sum_{j=1}^{P} x_{\ell,j}\right) - \Phi(x^\star)\right],$$

which further implies, owing to (2.6),

$$P\gamma^2\mathbb{E}\left[\Phi\left(\frac{1}{P}\textstyle\sum_{j=1}^{P} x_{\ell,j}\right) - \Phi(x^\star)\right] \leq \mathbb{E}[T_\ell] - \mathbb{E}[T_{\ell+1}].$$

Summing over $\ell = 1, \ldots, Q$ and telescoping the right hand of the sum we arrive at

$$\begin{aligned} QP\gamma^2\mathbb{E}\Big[\Phi\big(\tfrac{1}{QP}\textstyle\sum_{\ell=1}^{Q}\sum_{j=1}^{P} x_{\ell,j}\big) - \Phi(x^\star)\Big] &\le QP\gamma^2\mathbb{E}\Big[\tfrac{1}{Q}\textstyle\sum_{\ell=1}^{Q}\Phi\left(\tfrac{1}{P}\sum_{j=1}^{P} x_{\ell,j}\right) - \Phi(x^\star)\Big] \\ &\le \mathbb{E}[T_1] - \mathbb{E}[T_{Q+1}], \end{aligned}$$

where the first inequality follows from Jensen's inequality and convexity of Φ. Dividing both sides by $kP\gamma^2$ gives the required error bound. The convergence of $\varepsilon_k^{\mathrm{SVRG}}$ is a straightforward consequence of the convergence of $x_{l,p}$.

Now we prove the second claim of the theorem. Taking expectation of both sides of (2.3) in Lemma 2.2 and summing from $p = 1, \cdots, P$ yields

$$\begin{aligned} &(1-\gamma\alpha_R)\mathbb{E}\big[\|x_{\ell,P} - x^\star\|^2\big] + 2\gamma\mathbb{E}\big[\Phi(x_{\ell,P}) - \Phi(x^\star)\big] \\ &\le -(\alpha_F + \alpha_R)\textstyle\sum_{p=1}^{P}\mathbb{E}\big[\|x_{\ell,p} - x^\star\|^2\big] - \big(2\gamma - 8\gamma^2 L\big)\sum_{p=1}^{P-1}\mathbb{E}\big[\Phi(x_{\ell,p}) - \Phi(x^\star)\big] \\ &\quad + (1-\gamma\alpha_F)\mathbb{E}\big[\|x_{\ell,0} - x^\star\|^2\big] + 8\gamma^2 L\mathbb{E}\big[\Phi(x_{\ell,0}) - \Phi(x^\star)\big] + 8\gamma^2 LP\mathbb{E}\big[\Phi(\tilde{x}_\ell) - \Phi(x^\star)\big]. \end{aligned}$$

Since $\gamma L < \frac{1}{4(P+1)} < \frac{1}{4}$, we have $2\gamma - 8\gamma^2 L > 0$, and we have from the above

$$\begin{aligned} &(1-\gamma\alpha_R)\mathbb{E}\big[\|x_{\ell,P} - x^\star\|^2\big] + 2\gamma\mathbb{E}\big[\Phi(x_{\ell,P}) - \Phi(x^\star)\big] \\ &\le (1-\gamma\alpha_F)\mathbb{E}\big[\|x_{\ell,0} - x^\star\|^2\big] + 8\gamma^2 L(P+1)\mathbb{E}\big[\Phi(\tilde{x}_\ell) - \Phi(x^\star)\big]. \end{aligned}$$

Define

$$T_\ell \overset{\text{def}}{=} (1-\gamma\alpha_R)\mathbb{E}\big[\|x_{\ell,P} - x^\star\|^2\big] + 2\gamma\mathbb{E}\big[\Phi(x_{\ell,P}) - \Phi(x^\star)\big],$$

then there holds

$$\mathbb{E}[T_\ell] \le \max\left\{\frac{1-\gamma\alpha_F}{1+\gamma\alpha_R}, 4L\gamma(P+1)\right\}\mathbb{E}[T_{\ell-1}]$$

which implies the desired result. □

3. Finite manifold identification of SAGA/Prox-SVRG

3.1. Proofs for Theorem 3.2

Proof of Theorem 3.2. First of all, the definition of proximity operator (2) and the update of x_{k+1} (4) entail that

$$\frac{x_k - x_{k+1}}{\gamma_k} - \nabla F(x_k) - \varepsilon_k \in \partial R(x_{k+1}), \tag{3.1}$$

from which we get

$$\begin{aligned} \operatorname{dist}\big(-\nabla F(x^\star), \partial R(x_{k+1})\big) &\le \|\tfrac{1}{\gamma_k}\big(x_k - x_{k+1}\big) - \nabla F(x_k) - \varepsilon_k + \nabla F(x^\star)\| \\ &\le \tfrac{1}{\gamma_k}\|x_k - x_{k+1}\| + \|\nabla F(x_k) - \nabla F(x^\star)\| + \|\varepsilon_k\| \\ &\le \tfrac{1}{\underline{\gamma}}\|x_{k+1} - x_k\| + L_F\|x_k - x^\star\| + \|\varepsilon_k\|, \end{aligned}$$

where lower boundedness of γ_k and the L_F-Lipschitz continuity of ∇F (see assumption (**A.2**)) is applied to get the last inequality. We have:

- The almost sure convergence of $\{x_k\}_{k\in\mathbb{N}}$ (condition (**B.3**)) ensures that $L_F\|x_k - x^\star\|$ converges to 0 almost surely. Owing to assumption (**A.1**), R is sub-differentially continuous at all the points of its domain, typically at $x^\star$ for $-\nabla F(x^\star)$, hence we have $R(x_k) \to R(x^\star)$ almost surely;
- Combine the almost sure convergence of $\{x_k\}_{k\in\mathbb{N}}$ and (**B.1**) the bounded from below property of $\{\gamma_k\}_{k\in\mathbb{N}}$, we have that $\frac{1}{\underline{\gamma}}\|x_{k+1} - x_k\|$ converges to 0 almost surely.
- Condition (**B.2**) asserts that $\|\varepsilon_k\| \to 0$ almost surely.

Altogether, we have that

$$\operatorname{dist}\big(-\nabla F(x^\star), \partial R(x_{k+1})\big) \to 0 \ \text{ almost surely.}$$

To this point, all the conditions of Theorem 5.3 of (Hare & Lewis, 2004) are fulfilled almost surely on function $\langle \nabla F(x^\star), \cdot\rangle + R$, hence the identification result follows. □

3.2. When non-degeneracy condition fails

In Theorem 3.2, besides the partial smoothness assumption of R, the non-degeneracy condition (ND) is crucial to the identification of the sequence $\{x_k\}_{k\in\mathbb{N}}$. Owing to the result of (Lewis & Zhang, 2013; Hare & Lewis, 2004; 2007), it is a necessary condition for identification of the manifold $\mathcal{M}_{x^\star}$, and moreover ensures that the manifold $\mathcal{M}_{x^\star}$ is minimal and unique.

Recently, efforts are made to relax the non-degeneracy condition. In (Fadili et al., 2017), under a so-called "mirror stratification condition", the authors manage to relax the non-degeneracy condition, however at the price that the manifold to be identified is no longer unique. More precisely, there will be another manifold $\overline{\mathcal{M}}_{x^\star}$, which includes $\mathcal{M}_{x^\star}$ and is determined by how (ND) is violated. The sequence $\{x_k\}_{k\in\mathbb{N}}$ will identify a manifold $\widetilde{\mathcal{M}}_{x^\star}$ such that

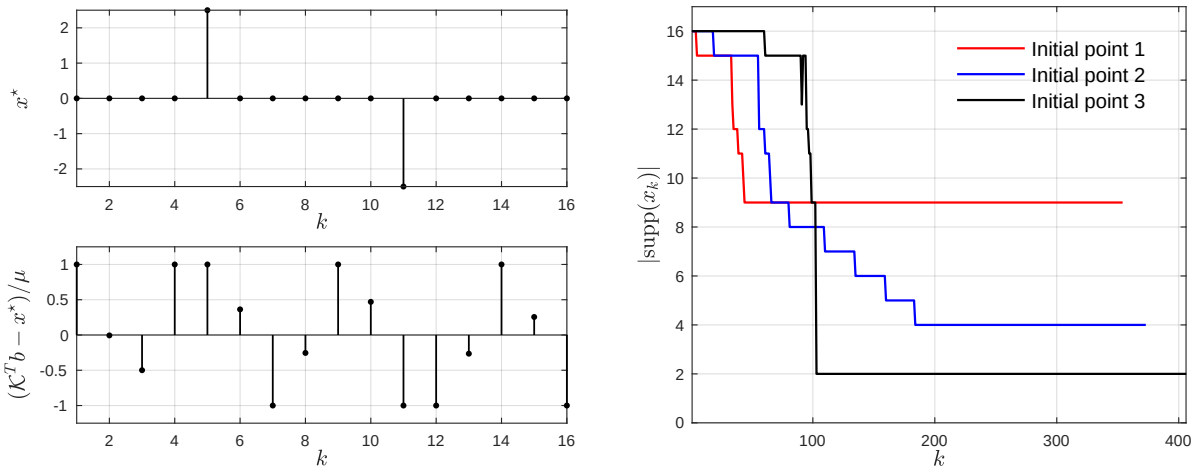
$$\mathcal{M}_{x^\star} \subseteq \widetilde{\mathcal{M}}_{x^\star} \subseteq \overline{\mathcal{M}}_{x^\star}.$$

Furthermore, the identification of $\{x_k\}_{k\in\mathbb{N}}$ could be unstable, that is $\{x_k\}_{k\in\mathbb{N}}$ may identify several different manifolds which are between $\mathcal{M}_{x^\star}$ and $\overline{\mathcal{M}}_{x^\star}$.

A degenerate LASSO problem We present a simple example of LASSO problem to demonstrate the unstable identification behaviour of $\{x_k\}_{k\in\mathbb{N}}$ when the non-degeneracy conditions fails. Consider the problem

$$\min_{x\in\mathbb{R}^n} \mu\|x\|_1 + \frac{1}{2}\|\mathcal{K}x - b\|^2, \tag{3.2}$$

where $\mu > 0$ is the penalty parameter, $\mathcal{K} \in \mathbb{R}^{n\times n}$ is a unitary matrix, and $b \in \mathbb{R}^n$ is a vector.



(a) $x^\star$ and $(\mathcal{K}^T b - x^\star)/\mu$ (b) $|\mathrm{supp}(x_k)|$ under different starting point

Figure 3.1: Identification properties of deterministic Forward–Backward splitting method when the non-degeneracy condition (ND) fails.

Since $\mathcal{K}$ is a unitary matrix, the solution of (3.2) is unique and can be given explicitly, which is

$$x^\star = \mathrm{sign}(\mathcal{K}^T b) \odot \max\left\{|\mathcal{K}^T b| - \mu, 0\right\}, \tag{3.3}$$

and $\odot$ denotes point-wise product. Moreover, we have the gradient at $x^\star$

$$-\nabla\left(\frac{1}{2}\|\mathcal{K}x^\star - b\|^2\right) = -\mathcal{K}^T(\mathcal{K}x^\star - b) = \mathcal{K}^T b - x^\star.$$

In the experiments, we set $\mu = 0.5$ and $n = 16$, and moreover the vector b is designed such that the non-degeneracy condition (ND) is violated. The two vectors $x^\star$ and $(\mathcal{K}^T b - x^\star)/\mu$ are shown in Figure 3.1(a), and it can be observed that

$x^\star$ has only *two* non-zero elements, while $(\mathcal{K}^T b - x^\star)/\mu$ has *nine* saturated elements (the saturation means that the absolute value of corresponding element is equal to μ).

Though the solution $x^\star$ can be provided in closed form (3.3), we choose to solve (3.2) with deterministic Forward–Backward splitting with fixed step-size $\gamma = 0.1$, which is the following iteration

$$x_{k+1} = \operatorname{sign}(w_k) \odot \max\left\{|w_k| - \gamma\mu, 0\right\} \quad \text{where} \quad w_k = (1-\gamma)x_k - \mathcal{K}^T b. \tag{3.4}$$

Three different initial points for (3.4) are considered. For each starting point, the size of support of the sequence $\{x_k\}_{k\in\mathbb{N}}$, *i.e.* $\{|\operatorname{supp}(x_k)|\}_{k\in\mathbb{N}}$, is plotted in Figure 3.1(b). For all three cases, the iterations are ran until machine accuracy is reached. We obtain the following observations from the comparisons:

- "Initial point 1" and "Initial point 2" are unable to identify the support of the solution $x^\star$;
- "Initial point 1" identifies the largest manifold, *i.e.* $\overline{\mathcal{M}}_{x^\star}$. For "Initial point 2", the identification is not stable in the early iterations (*e.g.* $k \leq 190$) compared to the other cases, and eventually (*e.g.* $k \geq 190$) stabilises onto a manifold $\widetilde{\mathcal{M}}_{x^\star}$ with $\mathcal{M}_{x^\star} \subset \widetilde{\mathcal{M}}_{x^\star} \subset \overline{\mathcal{M}}_{x^\star}$;
- "Initial point 3" manages to identify the smallest manifold, *i.e.* $\mathcal{M}_{x^\star}$.

We can conclude that the starting point is very crucial when the non-degeneracy condition (ND) fails.

4. Local linear convergence of SAGA/Prox-SVRG

4.1. An overdetermined LASSO problem

Below we present a example of overdetermined LASSO problem, to show that the practical performance of SAGA/Prox-SVRG could be much worse than ρ_{FBS}.

Consider again the LASSO problem,

$$\min_{x\in\mathbb{R}^n} \mu\|x\|_1 + \frac{1}{m}\sum_{i=1}^m \frac{1}{2}\|\mathcal{K}_i x - b_i\|^2,$$

where now $\mathcal{K} \in \mathbb{R}^{m\times n}$ is a random Gaussian matrix with zero means and $b \in \mathbb{R}^m$. Moreover, we choose $m = 256, n = 32$, that is much more measurements than the size of the vector.

For the test example, we have $L = 0.2239$ and the local quadratic grow parameter $\alpha = 0.0032$. The parameter choices of SAGA and Prox-SVRG with "Option II" are:

$$\text{SAGA} : \gamma = \frac{1}{3L}; \qquad \text{Prox-SVRG} : \gamma = \frac{1}{10L}, \ \ P = \frac{100L}{\alpha}.$$

We have $P \approx 27m$ which is quite large. As discussion in the original work (Xiao & Zhang, 2014), with the above parameters choices, $\rho_{\text{SVRG}} \approx \frac{5}{6}$.

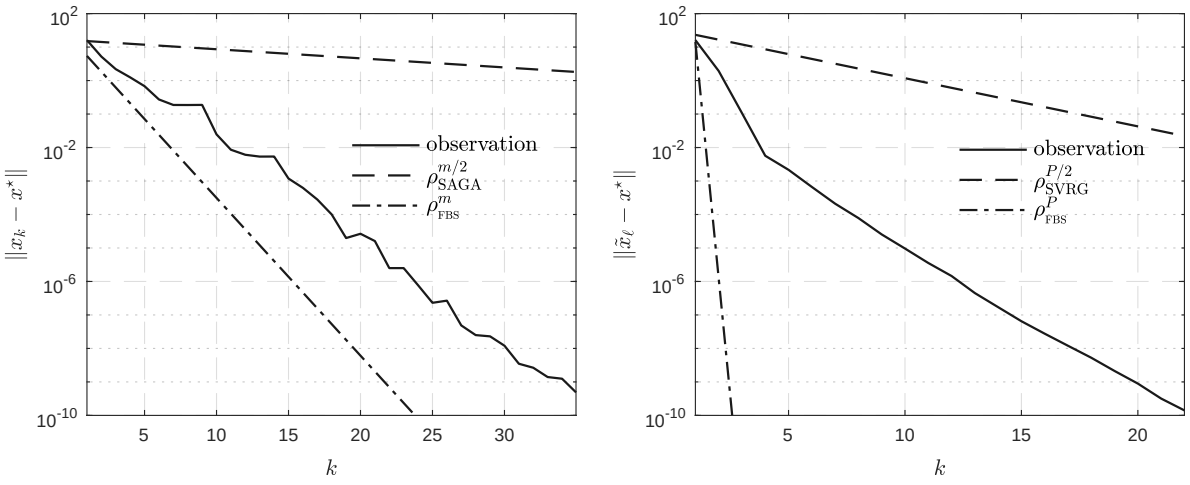
The outcomes of the numerical experiments are shown in Figure 4.1, where the observation of $\{\|x_k - x^\star\|\}_{k\in\mathbb{N}}$ is provided for SAGA and $\{\|\tilde{x}_\ell - x^\star\|\}_{k\in\mathbb{N}}$ for Prox-SVRG. The *solid* lines stand for practical observations of the methods, the *dashed* lines are the theoretical estimation from Proposition 4.3 and 4.4, the *dot-dashed* lines are the estimation from ρ_{FBS}. All the lines are sub-sampled, one out of every m points for SAGA and P points for Prox-SVRG. Note also that the observation is not in norm square.

For this example, both the convergence speeds of SAGA and Prox-SVRG are slower than ρ_{FBS}. Empirically, the reason for SAGA could be that the ratio of m/n is much larger than 1, while for Prox-SVRG the reason is that P/m is too large.

4.2. Large-scale datasets

Now we consider binary classification problem with two large-scale datasets obtained from LIBSVM[1], the size of the datasets can be found in the table below.

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/



(a) SAGA, $\|x_k - x^\star\|$ (b) Prox-SVRG "Option II", $\|\tilde{x}_\ell - x^\star\|$

Figure 4.1: Convergence rate of SAGA and Prox-SVRG when solving an overdetermined LASSO problem. (a) convergence behaviour of $\|x_k - x^\star\|$ of SAGA; (b) convergence behaviour of $\|\tilde{x}_\ell - x^\star\|$ of Prox-SVRG. The *solid* lines stands for practical observations of the methods, the *dashed* lines are the theoretical estimation from Proposition 4.3 and 4.4, the *dot-dashed* lines are the estimation from ρ_{FBS}. All the lines are sub-sampled, one out of every m points for SAGA and P points for Prox-SVRG.

Table 1: Considered datasets, number of samples m, and size of each sample n

Name	mushrooms	rcv1.binary
m	8124	20,242
n	112	47,236
μ	5×10^{-3}	10^{-2}

Let $(h_i, l_i), i = 1, ..., m$ be sample and label, the LASSO problem (3.2) can be then formulated as

$$\min_{x \in \mathbb{R}^n} \mu \|x\|_1 + \frac{1}{2m} \sum_{i=1}^m (h_i^T x - l_i)^2, \tag{4.1}$$

where $\mu > 0$ is again a trade-off parameter whose value can be found in the last row of Table 1.

The outcomes of the numerical experiments are illustrated in Figure 4.2. The step-size of both algorithms for this experiment are chosen as the same, which is $\gamma = \frac{1}{3L}$. It can be observed that the performance of both algorithms (*i.e.* $\Phi(x_k) - \Phi(x^\star)$) in terms of number of paths are quite close (see Figure 4.2 (b) and (d)), while Prox-SVRG shows slightly better identification than SAGA, for instance in Figure 4.2 (a) the small jump at $k/m = 35$ of the black line for SAGA.

5. Local acceleration of SAGA/Prox-SVRG

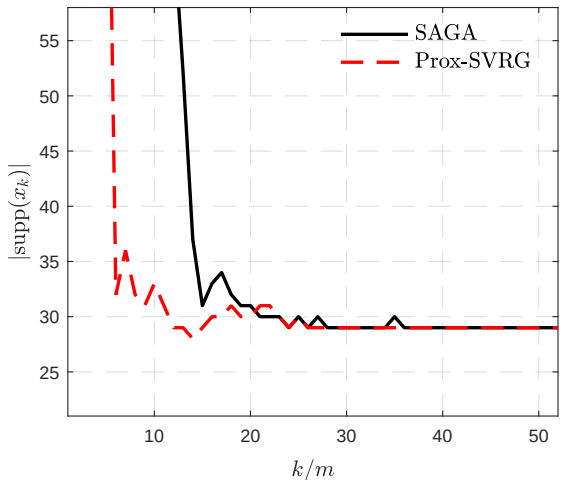
In this section, we provide details on how to implement the local acceleration technique based on the local Lipschitz constants. The result focuses on the functions whose partly smooth manifold $\mathcal{M}_{x^\star}$ is an affine subspace, such functions include $\ell_1, \ell_{1,2}, \ell_\infty$-norms and total variation; see Table 1.

Recall the original optimisation problem, which reads

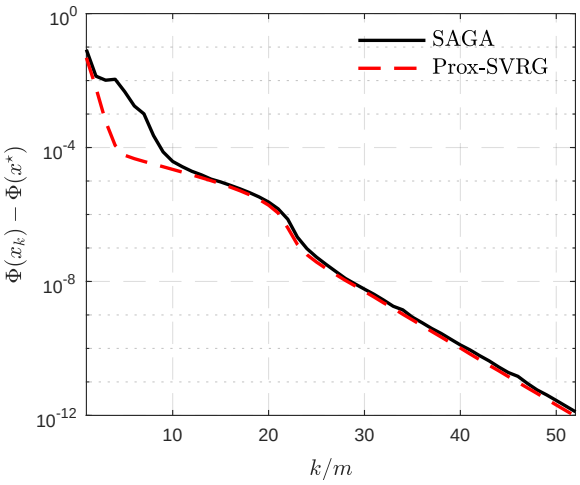
$$\min_{x \in \mathbb{R}^n} \Phi(x) \stackrel{\text{def}}{=} R(x) + F(x). \tag{5.1}$$

Let $\mathcal{T}_{x^\star} \subseteq \mathbb{R}^n$ be an affine subspace, and $\mathcal{M}_{x^\star} = \mathcal{T}_{x^\star}$. Once the manifold $\mathcal{M}_{x^\star}$ of an $x^\star \in \operatorname{argmin}(\Phi)$ is identified, then the problem locally becomes

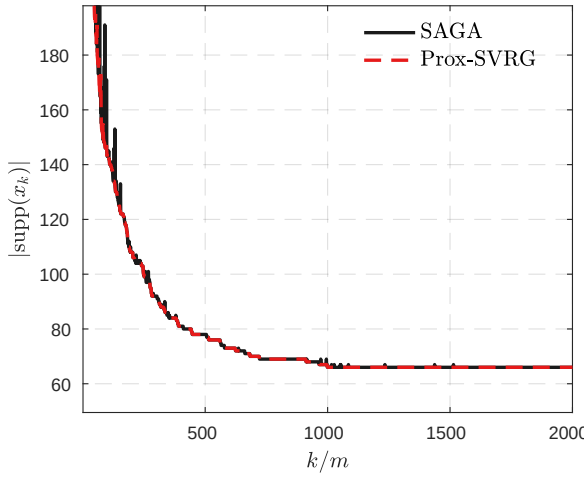
$$\min_{x \in \mathcal{T}_{x^\star}} \Phi_{\mathcal{T}_{x^\star}}(x) \stackrel{\text{def}}{=} R(x) + F(\mathrm{P}_{\mathcal{T}_{x^\star}}(x)). \tag{5.2}$$



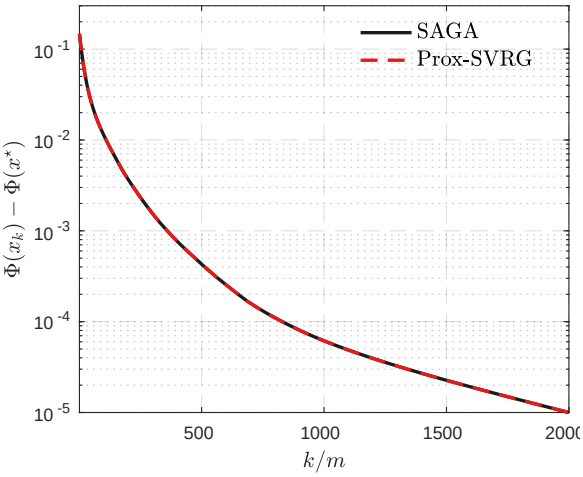
(a) mushrooms: manifold identification



(b) mushrooms: convergence rate



(c) rcv1: manifold identification



(d) rcv1: convergence rate

Figure 4.2: Experiments on large-scale datasets.

This means that locally, it is a manifold constrained optimisation problem, and the properties of the functions, typically F, could become much better.

Denote $\mathrm{P}_{\mathcal{T}_{x^\star}}$ the projection operator on to $\mathcal{T}_{x^\star}$. Clearly, given any $x \in \mathcal{T}_{x^\star}$ the gradient of $F(\mathrm{P}_{\mathcal{T}_{x^\star}}(x))$ reads

$$\nabla F(\mathrm{P}_{\mathcal{T}_{x^\star}}(x)) = \mathrm{P}_{\mathcal{T}_{x^\star}} \circ \nabla F \circ \mathrm{P}_{\mathcal{T}_{x^\star}}(x),$$

indicating that we only need to consider the Lipschitz constant of $\mathrm{P}_{\mathcal{T}_{x^\star}} \circ \nabla F \circ \mathrm{P}_{\mathcal{T}_{x^\star}}$. Since the projection operator $\mathrm{P}_{\mathcal{T}_{x^\star}}$ is non-expansive, we have for any $x, y \in \mathcal{T}_{x^\star}$

$$\begin{aligned}\|\mathrm{P}_{\mathcal{T}_{x^\star}} \circ \nabla F \circ \mathrm{P}_{\mathcal{T}_{x^\star}}(x) - \mathrm{P}_{\mathcal{T}_{x^\star}} \circ \nabla F \circ \mathrm{P}_{\mathcal{T}_{x^\star}}(y)\| &\leq \|\nabla F \circ \mathrm{P}_{\mathcal{T}_{x^\star}}(x) - \nabla F \circ \mathrm{P}_{\mathcal{T}_{x^\star}}(y)\| \\ &\leq L\|\mathrm{P}_{\mathcal{T}_{x^\star}}(x) - \mathrm{P}_{\mathcal{T}_{x^\star}}(y)\| \\ &\leq L\|x - y\|.\end{aligned}$$

Similar result can be derived for each function $f_i, i = 1, ..., m$.

In the following, we use $R = \|x\|_1$ and $F = \frac{1}{2}\|Ax - y\|^2$, where $A \in \mathbb{R}^{m \times n}$ is a linear operator and $y \in \mathbb{R}^m$ is label vector, to demonstrate the computation. Since ℓ_1-norm is polyhedral, we have that $\mathcal{T}_{x^\star}$ is a subspace (Table 1),

$$\mathcal{T}_{x^\star} = \left\{z \in \mathbb{R}^n : \mathcal{I}_z \subseteq \mathcal{I}_{x^\star}\right\}, \quad \mathcal{I}_{x^\star} = \left\{i : x_i^\star \neq 0\right\}.$$

The projection operator $\mathrm{P}_{\mathcal{T}_{x^\star}}$ then reads

$$\mathrm{P}_{\mathcal{T}_{x^\star}} = \mathtt{diag}\big(\mathtt{sign}(x^\star)\big),$$

which is a diagonal matrix with $i^{\text{th}}, i \in \mathcal{I}_{x^\star}$ diagonal element being 1 and 0 otherwise. Here, `diag` and `sign` are built-in functions of MATLAB. Clearly, $\mathrm{P}_{\mathcal{T}_{x^\star}}$ is a column selection. Denote $A_{\mathcal{T}_{x^\star}} = A \circ \mathrm{P}_{\mathcal{T}_{x^\star}}$, then the local Lipschitz constant of ∇F is

$$L_{F,\mathcal{T}_{x^\star}} = \|A_{\mathcal{T}_{x^\star}}\|.$$

While the global Lipschitz constant is $L_F = \|A\|$. Then for each function $f_i, i = 1, ..., m$, denote $A_{\mathcal{T}_{x^\star},i}$ the i^{th} row of $A_{\mathcal{T}_{x^\star}}$, then

$$L_{\mathcal{T}_{x^\star},i} = \|A_{\mathcal{T}_{x^\star},i}\|.$$

Finally, we have $L_{\mathcal{T}_{x^\star}} = \max_{i=1,\cdots,m} L_{\mathcal{T}_{x^\star},i}$.

References

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pp. 1646–1654, 2014.

Fadili, J., Malick, J., and Peyré, G. Sensitivity analysis for mirror-stratifiable convex functions. *arXiv preprint arXiv:1707.03194*, 2017.

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

Hare, W. L. and Lewis, A. S. Identifying active manifolds. *Algorithmic Operations Research*, 2(2):75–82, 2007.

Lewis, A. S. and Zhang, S. Partial smoothness, tilt stability, and generalized Hessians. *SIAM Journal on Optimization*, 23 (1):74–94, 2013.

Neveu, J. *Discrete-parameter martingales*, volume 10. Elsevier, 1975.

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