SUPPLEMENTARY DOCUMENT

Stochastic Gauss-Newton Algorithms for Nonconvex Compositional Optimization

A. The Proof of Technical Results in Section 2: Mathematical Tools

This section provides the full proof of technical results in Section 2. Let us first recall the bound (10). The proof of this bound can be found, e.g., in Nesterov (2007). However, for completeness, we prove it here.

The proof of (10). Since F' is L_F-Lipschitz continuous with a Lipschitz constant L_F, we have $\|F(y) - F(x) - F'(x)(y - x)\| \leq \frac{L_F}{2}\|y - x\|^2$ for any $x, y \in \mathbb{R}^p$. On the other hand, since ϕ is M_ϕ-Lipschitz continuous, we have $\phi(u) \leq \phi(v) + M_\phi\|u - v\|$ for any $u, v \in \mathbb{R}^q$. Hence, we have

$$\begin{array}{lcl} \phi(F(y)) & \leq & \phi(F(x) + F'(x)(y-x)) + M_\phi\|F(y) - F(x) - F'(x)(y-x)\| \\ & \leq & \phi(F(x) + F'(x)(y-x)) + \frac{M_\phi L_F}{2}\|y - x\|^2, \end{array}$$

which proves (10). □

A.1. The Proof of Lemma 2.1: Approximate Optimality Condition

Lemma. 2.1. *Suppose that Assumption 1.1 holds. Let $\widetilde{T}_M(x)$ be computed by* (11) *and $\widetilde{G}_M(x)$ be defined by* (13)*. Then, $\mathcal{E}(\widetilde{T}_M(x), y)$ of* (1) *or* (2) *defined by* (7) *with $y \in \partial\phi(F(\widetilde{T}_M(x)))$ is bounded by*

$$\begin{array}{lcl} \mathcal{E}(\widetilde{T}_M(x), y) & := & \mathrm{dist}\left(0, -F(\widetilde{T}_M(x)) + \partial\phi^*(y)\right) + \|F'(\widetilde{T}_M(x))^\top y\| \\ & \leq & \left(1 + \frac{M_\phi L_F}{M}\right)\|\widetilde{G}_M(x)\| + \frac{(1+L_F)}{2M^2}\|\widetilde{G}_M(x)\|^2 + \|\widetilde{F}(x) - F(x)\| + \frac{1}{2}\|\widetilde{J}(x) - F'(x)\|^2. \end{array} \tag{14}$$

Proof. First, the optimality condition of (11) becomes

$$0 \in \widetilde{J}(x)^\top\partial\phi(\widetilde{F}(x) + \widetilde{J}(x)(\widetilde{T}_M(x) - x)) + M(\widetilde{T}_M(x) - x). \tag{37}$$

We can rewrite this optimality condition as

$$r_F(x) = F'(\widetilde{T}_M(x))^\top y \qquad \text{and} \qquad r_D(x) \in -F(\widetilde{T}_M(x)) + \partial\phi^*(y),$$

where

$$\left\{\begin{array}{ll} r_F(x) & := M(x - \widetilde{T}_M(x)) + (F'(\widetilde{T}_M(x)) - \widetilde{J}(x))^\top y, \\ r_D(x) & := \widetilde{F}(x) + \widetilde{J}(x)(\widetilde{T}_M(x) - x) - F(\widetilde{T}_M(x)). \end{array}\right.$$

Next, since $y \in \partial\phi(\widetilde{F}(x) + \widetilde{J}(x)(\widetilde{T}_M(x) - x))$ and ϕ is M_ϕ-Lipschitz continuous, we can bound y as $\|y\| \leq M_\phi$. Now, we need to bound r_F as follows:

$$\begin{array}{lcl} \|r_F(x)\| & = & \|M(x - \widetilde{T}_M(x)) + (F'(\widetilde{T}_M(x)) - \widetilde{J}(x))^\top y\| \\ & = & \|M(x - \widetilde{T}_M(x)) + (F'(\widetilde{T}_M(x)) - F'(x))^\top y + (F'(x) - \widetilde{J}(x))^\top y\| \\ & \leq & M\|x - \widetilde{T}_M(x)\| + \|F'(\widetilde{T}_M(x)) - F'(x)\|_F\|y\| + \|F'(x) - \widetilde{J}(x)\|_F\|y\| \\ & \leq & \|\widetilde{G}_M(x)\| + M_\phi\|F'(\widetilde{T}_M(x)) - F'(x)\|_F + M_\phi\|F'(x) - \widetilde{J}(x)\| \\ & \leq & \left(1 + \frac{M_\phi L_F}{M}\right)\|\widetilde{G}_M(x)\| + M_\phi\|F'(x) - \widetilde{J}(x)\|. \end{array}$$

Similarly, we can also bound r_D as

$$\begin{array}{lcl} \|r_D(x)\| & = & \|\widetilde{F}(x) + \widetilde{J}(x)(\widetilde{T}_M(x) - x) - F(\widetilde{T}_M(x))\| \\ & = & \|\widetilde{F}(x) - F(x) + F(x) + F'(x)(\widetilde{T}_M(x) - x) - F(\widetilde{T}_M(x)) + [\widetilde{J}(x) - F'(x)](\widetilde{T}_M(x) - x)\| \\ & \leq & \|\widetilde{F}(x) - F(x)\| + \|F(x) + F'(x)(\widetilde{T}_M(x) - x) - F(\widetilde{T}_M(x))\| + \|[\widetilde{J}(x) - F'(x)](\widetilde{T}_M(x) - x)\| \\ & \leq & \|\widetilde{F}(x) - F(x)\| + \frac{L_F}{2}\|\widetilde{T}_M(x) - x\|^2 + \frac{1}{2}\|F'(x) - \widetilde{J}(x)\|^2 + \frac{1}{2}\|\widetilde{T}_M(x) - x\|^2 \\ & = & \|\widetilde{F}(x) - F(x)\| + \frac{1}{2}\|F'(x) - \widetilde{J}(x)\|^2 + \frac{(1+L_F)}{2M^2}\|\widetilde{G}_M(x)\|^2. \end{array}$$

Combining these bounds, we can show that

$$\begin{array}{lcl} \mathcal{E}(\widetilde{T}_M(x), y) & := & \|F'(\widetilde{T}_M(x))^\top y\| + \mathrm{dist}\left(0, -F(\widetilde{T}_M(x)) + \partial\phi^*(y)\right) \\ & \leq & \left(1 + \frac{M_\phi L_F}{M}\right)\|\widetilde{G}_M(x)\| + \frac{(1+L_F)}{2M^2}\|\widetilde{G}_M(x)\|^2 + \|\widetilde{F}(x) - F(x)\| + \frac{1}{2}\|F'(x) - \widetilde{J}(x)\|^2, \end{array}$$

which is exactly (14). □

B. The Proof of Technical Results in Section 3: Convergence of Inexact GN Framework

This appendix provides the full proof of technical results in Section 3 on convergence of the inexact Gauss-Newton framework, Algorithm 1.

B.1. The Proof of Lemma 3.1: Descent Property

Lemma. 3.1. *Let Assumption 1.1 hold, $\widetilde{T}_M(x)$ be computed by (11), and $\widetilde{G}_M(x) := M(x - \widetilde{T}_M(x))$ be the prox-gradient mapping of F. Then, for any $z \in \mathbb{R}^p$, we have*

$$\phi(\widetilde{F}(x) + \widetilde{J}(x)(\widetilde{T}_M(x) - x)) \leq \phi(\widetilde{F}(x) + \widetilde{J}(x)(z - x)) - \langle \widetilde{G}_M(x), z - x\rangle - \frac{1}{M}\|\widetilde{G}_M(x)\|^2. \tag{38}$$

For any $\beta_d > 0$, we also have

$$\begin{array}{ll} \phi(F(\widetilde{T}_M(x))) & \leq \phi(F(x)) + 2L_\phi\|F(x) - \widetilde{F}(x)\| + M_\phi\|F'(x) - \widetilde{J}(x)\|\|x - \widetilde{T}_M(x)\| - \frac{(2M - M_\phi L_F)}{2}\|\widetilde{T}_M(x) - x\|^2 \\ & \leq \phi(F(x)) + 2L_\phi\|F(x) - \widetilde{F}(x)\| + \frac{M_\phi}{2\beta_d}\|F'(x) - \widetilde{J}(x)\|_F^2 - \frac{(2M - M_\phi L_F - \beta_d L_\phi)}{2M^2}\|\widetilde{G}_M(x)\|^2. \end{array} \tag{15}$$

Proof. The optimality condition (37) can be written as

$$\widetilde{J}(x)^\top y = M(x - \widetilde{T}_M(x)) \quad \text{and} \quad y \in \partial\phi(\widetilde{F}(x) + \widetilde{J}(x)(\widetilde{T}_M(x) - x)).$$

By convexity of ϕ, using the above relations, we have

$$\begin{array}{ll} \phi(\widetilde{F}(x) + \widetilde{J}(x)(z - x)) & \geq \phi(\widetilde{F}(x) + \widetilde{J}(x)(\widetilde{T}_M(x) - x)) + \langle y, \widetilde{F}(x) + \widetilde{J}(x)(z - x) - (\widetilde{F}(x) + \widetilde{J}(x)(\widetilde{T}_M(x) - x))\rangle \\ & \geq \phi(\widetilde{F}(x) + \widetilde{J}(x)(\widetilde{T}_M(x) - x)) + \langle \widetilde{J}(x)^\top y, z - \widetilde{T}_M(x)\rangle \\ & = \phi(\widetilde{F}(x) + \widetilde{J}(x)(\widetilde{T}_M(x) - x)) + M\langle z - \widetilde{T}_M(x), x - \widetilde{T}_M(x)\rangle \\ & = \phi(\widetilde{F}(x) + \widetilde{J}(x)(\widetilde{T}_M(x) - x)) + M\langle x - \widetilde{T}_M(x), z - x\rangle + M\|x - \widetilde{T}_M(x)\|^2 \\ & = \phi(\widetilde{F}(x) + \widetilde{J}(x)(\widetilde{T}_M(x) - x)) + \langle \widetilde{G}_M(x), z - x\rangle + \frac{1}{M}\|\widetilde{G}_M(x)\|^2, \end{array}$$

which implies (38).

Now, combining (10) and (38), we can show that

$$\begin{aligned}
\phi(F(\widetilde{T}_M(x))) &\overset{(10)}{\leq} \phi(F(x) + F'(x)(\widetilde{T}_M(x) - x)) + \tfrac{M_\phi L_F}{2}\|\widetilde{T}_M(x) - x\|^2 \\
&\leq \phi(\widetilde{F}(x) + \widetilde{J}(x)(\widetilde{T}_M(x) - x)) + \tfrac{M_\phi L_F}{2}\|\widetilde{T}_M(x) - x\|^2 \\
&\quad + |\phi(F(x) + F'(x)(\widetilde{T}_M(x) - x)) - \phi(\widetilde{F}(x) + \widetilde{J}(x)(\widetilde{T}_M(x) - x))| \\
&\overset{(38)}{\leq} \phi(\widetilde{F}(x) + \widetilde{J}(x)(z - x)) - M\langle x - \widetilde{T}_M(x), z - x\rangle - \tfrac{(2M - M_\phi L_F)}{2}\|\widetilde{T}_M(x) - x\|^2 \\
&\quad + M_\phi\|F(x) - \widetilde{F}(x) + [F'(x) - \widetilde{J}(x)](\widetilde{T}_M(x) - x)\| \\
&\leq \phi(F(x)) - \tfrac{(2M - M_\phi L_F)}{2}\|\widetilde{T}_M(x) - x\|^2 + M_\phi\|F(x) - \widetilde{F}(x)\| \\
&\quad + M_\phi\|F(x) - \widetilde{F}(x) - \widetilde{J}(x)(z - x)\| - M\langle x - \widetilde{T}_M(x), z - x\rangle \\
&\quad + M_\phi\|(F'(x) - \widetilde{J}(x))(\widetilde{T}_M(x) - x)\|.
\end{aligned}$$

Substituting $z = x$ into this estimate, we obtain

$$\begin{aligned}
\phi(F(\widetilde{T}_M(x))) \leq{}& \phi(F(x)) - \tfrac{(2M - M_\phi L_F)}{2}\|\widetilde{T}_M(x) - x\|^2 + 2M_\phi\|F(x) - \widetilde{F}(x)\| \\
&+ M_\phi\|(F'(x) - \widetilde{J}(x))(\widetilde{T}_M(x) - x)\|.
\end{aligned} \tag{39}$$

Using the Cauchy-Schwarz inequality, we have

$$\|(F'(x) - \widetilde{J}(x))(\widetilde{T}_M(x) - x)\| \leq \|F'(x) - \widetilde{J}(x)\|\|\widetilde{T}_M(x) - x\|.$$

Next, applying Young's inequality to the right hand side of this inequality, for any $\beta_d > 0$, we obtain

$$\|(F'(x) - \widetilde{J}(x))(\widetilde{T}_M(x) - x)\| \leq \|F'(x) - \widetilde{J}(x)\|_F\|\widetilde{T}_M(x) - x\| \leq \frac{1}{2\beta_d}\|F'(x) - \widetilde{J}(x)\|^2 + \frac{\beta_d}{2}\|\widetilde{T}_M(x) - x\|^2. \tag{40}$$

Finally, plugging (40) into (39), we have

$$\begin{aligned}
\phi(F(\widetilde{T}_M(x))) &\leq \phi(F(x)) - \tfrac{(2M - M_\phi L_F)}{2}\|\widetilde{T}_M(x) - x\|^2 + 2L_\phi\|F(x) - \widetilde{F}(x)\| + M_\phi\|F'(x) - \widetilde{J}(x)\|\|\widetilde{T}_M(x) - x\| \\
&\leq \phi(F(x)) - \tfrac{(2M - M_\phi L_F - \beta_d L_\phi)}{2}\|\widetilde{T}_M(x) - x\|^2 + 2L_\phi\|F(x) - \widetilde{F}(x)\| + \tfrac{M_\phi}{2\beta_d}\|F'(x) - \widetilde{J}(x)\|^2,
\end{aligned}$$

for any $\beta_d > 0$, which exactly implies (15). □

B.2. The Proof of Theorem 3.1: Convergence Rate of Algorithm 1

Theorem. 3.1. *Assume that Assumptions 1.1 and 1.2 are satisfied. Let $\{x_t\}$ be generated by Algorithm 1 to solve either (1) or (2). Then, the following statements hold:*

(a) *If (16) holds for some $\varepsilon \geq 0$, then*

$$\min_{0 \leq t \leq T}\|\widetilde{G}_M(x_t)\|^2 \leq \frac{1}{(T+1)}\sum_{t=0}^{T}\|\widetilde{G}_M(x_t)\|^2 \leq \frac{2M^2\left[\Psi(x_0) - \Psi^\star\right]}{C_g(T+1)} + \frac{\varepsilon^2}{2}, \tag{19}$$

where $C_g := 2M - M_\phi(L_F + \beta_d)$ for $M > \frac{1}{2}M_\phi(L_F + \beta_d)$.

(b) *If (17) and (18) hold for given $C_a > 0$, then*

$$\min_{0 \leq t \leq T}\|\widetilde{G}_M(x_t)\|^2 \leq \frac{1}{(T+1)}\sum_{t=0}^{T}\|\widetilde{G}_M(x_t)\|^2 \leq \frac{2M^2\left[\Psi(x_0) - \Psi^\star\right]}{C_a(T+1)} + \frac{\varepsilon^2}{2}. \tag{20}$$

Consequently, with $\varepsilon > 0$, the total number of iterations T to achieve $\frac{1}{(T+1)}\sum_{t=0}^{T}\|\widetilde{G}_M(x_t)\|^2 \leq \varepsilon^2$ is at most

$$T := \left\lfloor \frac{4M^2\left[\Psi(x_0) - \Psi^\star\right]}{D\varepsilon^2} \right\rfloor = \mathcal{O}\left(\frac{\left[\Psi(x_0) - \Psi^\star\right]}{\varepsilon^2}\right),$$

where $D := C_g$ for the case (a) *and $D := C_a$ for the case* (b).

Proof. Using the second inequality of (15) with $x := x_t$ and $T_M(x) = x_{t+1}$, we have

$$\phi(F(x_{t+1})) \le \phi(F(x_t)) - \frac{(2M - M_\phi(L_F + \beta_d))}{2}\|x_{t+1} - x_t\|^2 + 2M_\phi\|F(x_t) - \widetilde{F}_t\| + \frac{M_\phi\|F'(x_t) - \widetilde{J}_t\|^2}{2\beta_d}. \quad (41)$$

(a) If (16) holds for some $\varepsilon \ge 0$, then using (16) into (41), we have

$$\phi(F(x_{t+1})) \le \phi(F(x_t)) - \frac{C_g}{2}\|x_{t+1} - x_t\|^2 + 2M_\phi \cdot \frac{C_g\varepsilon^2}{16M_\phi M^2} + \frac{M_\phi}{2\beta_d} \cdot \frac{\beta_d C_g \varepsilon^2}{4M_\phi M^2},$$

where $C_g := 2M - M_\phi(L_F + \beta_d) > 0$. Since $\Psi(x) = \phi(F(x))$, the last estimate leads to

$$\Psi(x_{t+1}) \le \Psi(x_t) - \frac{C_g}{2}\|x_{t+1} - x_t\|^2 + \frac{C_g\varepsilon^2}{4M^2}.$$

By induction, $\widetilde{G}_M(x_t) := M(x_t - \widetilde{T}_M(x_t))$, and $\Psi(x_{T+1}) \ge \Psi^\star$, we can show that

$$\frac{1}{M^2(T+1)}\sum_{t=0}^{T}\|\widetilde{G}_M(x_t)\|^2 = \frac{1}{T+1}\sum_{t=0}^{T}\|x_{t+1} - x_t\|^2 \le \frac{2\left[\Psi(x_0) - \Psi^\star\right]}{C_g(T+1)} + \frac{\varepsilon^2}{2M^2}, \quad (42)$$

which leads to (19).

(b) If (17) and (18) are used, then from (41) and (18), we have

$$\phi(F(x_{t+1})) \le \phi(F(x_t)) - \frac{C_1}{2}\|x_{t+1} - x_t\|^2 + \frac{C_2}{2}\|x_t - x_{t-1}\|^2, \quad \forall t \ge 1.$$

where $C_1 := 2M - M_\phi L_F - \beta_d M_\phi$ and $C_2 := 2M_\phi\sqrt{C_f} + \frac{M_\phi C_d}{2\beta_d}$. For $t = 0$, it follows from (41) and (17) that

$$\phi(F(x_1)) \le \phi(F(x_0)) - \frac{C_1}{2}\|x_1 - x_0\|^2 + \frac{(C_1 - C_2)\varepsilon^2}{4M^2}.$$

Now, note that $\Psi(x) = \phi(F(x))$, the last two estimates respectively become

$$\Psi(x_{t+1}) \le \Psi(x_t) - \frac{C_1}{2}\|x_{t+1} - x_t\|^2 + \frac{C_2}{2}\|x_t - x_{t-1}\|^2, \quad \forall t \ge 1,$$

and for $t = 0$, it holds that

$$\Psi(x_1) \le \Psi(x_0) - \frac{C_1}{2}\|x_1 - x_0\|^2 + \frac{(C_1 - C_2)\varepsilon^2}{4M^2}.$$

By induction and $\Psi^\star \le \Psi(x_{T+1})$, this estimate leads to

$$\begin{aligned}\Psi^\star \le \Psi(x_{T+1}) \quad &\le \Psi(x_0) - \tfrac{(C_1 - C_2)}{2}\textstyle\sum_{t=0}^{T}\|x_{t+1} - x_t\|^2 + \tfrac{(C_1 - C_2)\varepsilon^2}{4M^2} \\ &- \tfrac{C_2}{2}\|x_{T+1} - x_T\|^2.\end{aligned}$$

Since $C_1 > C_2$, if we define $C_a := C_1 - C_2 > 0$, then the last inequality implies

$$\frac{1}{M^2(T+1)}\sum_{t=0}^{T}\|\widetilde{G}_M(x_t)\|^2 = \frac{1}{(T+1)}\sum_{t=0}^{T}\|x_{t+1} - x_t\|^2 \le \frac{2\left[\Psi(x_0) - \Psi^\star\right]}{C_a(T+1)} + \frac{\varepsilon^2}{4M^2},$$

which leads to (20). The last statement of this theorem is a direct consequence of either (19) or (20), and we omit the detailed derivation here. □

C. High Probability Inequalities and Variance Bounds

Since our methods are stochastic, we recall some mathematical tools from high probability and concentration theory, as well as variance bounds that will be used for our analysis. First, we need the following lemmas to estimate sample complexity of our algorithms.

Lemma C.1 (Matrix Bernstein inequality (Tropp, 2012)(Theorem 1.6)). *Let $X_1, X_2, \cdots, X_n$ be independent random matrices in $\mathbb{R}^{p_1 \times p_2}$. Assume that $\mathbb{E}\left[X_i\right] = 0$ and $\|X_i\| \leq R$ a.s. for $i = 1, \cdots, n$ and given $R > 0$, where $\|\cdot\|$ is the spectral norm. Define $\sigma_X^2 := \max\left\{\left\|\sum_{i=1}^n \mathbb{E}\left[X_i X_i^\top\right]\right\|, \left\|\sum_{i=1}^n \mathbb{E}\left[X_i^\top X_i\right]\right\|\right\}$. Then, for any $\epsilon > 0$, we have*

$$\mathbf{Prob}\left(\Big\|\sum_{i=1}^n X_i\Big\| \geq \epsilon\right) \leq (p_1 + p_2)\exp\left(-\frac{3\epsilon^2}{6\sigma_X^2 + 2R\epsilon}\right).$$

As a consequence, if $\sigma_X^2 \leq \bar{\sigma}_X^2$ for a given $\bar{\sigma}_X^2 > 0$, then

$$\mathbf{Prob}\left(\Big\|\sum_{i=1}^n X_i\Big\| \leq \epsilon\right) \geq 1 - (p_1 + p_2)\exp\left(-\frac{3\epsilon^2}{6\bar{\sigma}_X^2 + 2R\epsilon}\right).$$

Lemma C.2 (Lohr (2009)). *Let $\widetilde{F}(x_t)$ and $\widetilde{J}(x_t)$ be the mini-batch stochastic estimators of $F(x_t)$ and $F'(x_t)$ defined by (21), respectively, and $\mathcal{F}_t := \sigma(x_0, x_1, \cdots, x_{t-1})$ be the σ-field generated by $\{x_0, x_1, \cdots, x_{t-1}\}$. Then, these are unbiased estimators, i.e., $\mathbb{E}\left[\widetilde{F}(x_t) \mid \mathcal{F}_t\right] = F(x_t)$ and $\mathbb{E}\left[\widetilde{J}(x_t) \mid \mathcal{F}_t\right] = F'(x_t)$. Moreover, under Assumption 1.2, we have*

$$\mathbb{E}\left[\|\widetilde{F}(x_t) - F(x_t)\|^2 \mid \mathcal{F}_t\right] \leq \frac{\sigma_F^2}{b_t} \quad \text{and} \quad \mathbb{E}\left[\|\widetilde{J}(x_t) - F'(x_t)\|^2 \mid \mathcal{F}_t\right] \leq \frac{\sigma_D^2}{\hat{b}_t}. \tag{43}$$

Lemma C.3 (Nguyen et al. (2017); Pham et al. (2020)). *Let $\widetilde{F}_t$ and $\widetilde{J}_t$ be the mini-batch SARAH estimators of $F(x_t)$ and $F'(x_t)$, respectively defined by (27), and $\mathcal{F}_t := \sigma(x_0, x_1, \cdots, x_{t-1})$ be the σ-field generated by $\{x_0, x_1, \cdots, x_{t-1}\}$. Then, we have the following estimate*

$$\mathbb{E}\left[\|\widetilde{F}_t - F(x_t)\|^2 \mid \mathcal{F}_t\right] = \|\widetilde{F}_{t-1} - F(x_{t-1})\|^2 + \rho_t \mathbb{E}_\xi\left[\|\mathbf{F}(x_t, \xi) - \mathbf{F}(x_{t-1}, \xi)\|^2\right] - \rho_t\|F(x_t) - F(x_{t-1})\|^2, \tag{44}$$

where $\rho_t := \frac{n - b_t}{(n-1)b_t}$ if $F(x) := \frac{1}{n}\sum_{i=1}^n F_i(x)$, and $\rho_t := \frac{1}{b_t}$, otherwise, i.e., $F(x) = \mathbb{E}_\xi\left[\mathbf{F}(x, \xi)\right]$.

Similarly, we also have

$$\mathbb{E}\left[\|\widetilde{J}_t - F'(x_t)\|^2 \mid \mathcal{F}_t\right] = \|\widetilde{J}_{t-1} - F'(x_{t-1})\|^2 + \hat{\rho}_t \mathbb{E}_\xi\left[\|\mathbf{F}'(x_t, \xi) - \mathbf{F}'(x_{t-1}, \xi)\|^2\right] - \hat{\rho}_t\|F'(x_t) - F'(x_{t-1})\|^2, \tag{45}$$

where $\hat{\rho}_t := \frac{n - \hat{b}_t}{(n-1)\hat{b}_t}$ if $F(x) := \frac{1}{n}\sum_{i=1}^n F_i(x)$, and $\hat{\rho}_t := \frac{1}{\hat{b}_t}$, otherwise, i.e., $F(x) = \mathbb{E}_\xi\left[\mathbf{F}(x, \xi)\right]$.

D. The Proof of Technical Results in Section 4

This appendix provides the full proof of technical results in Section 4 on our stochastic Gauss-Newton methods.

D.1. The Proof of Theorem 4.1: Convergence of The Stochastic Gauss-Newton Method for Solving (1)

Theorem. 4.1. *Suppose that Assumptions 1.1 and 1.2 hold for (1). Let $\widetilde{F}_t$ and $\widetilde{J}_t$ defined by (21) be mini-batch stochastic estimators of $F(x_t)$ and $F'(x_t)$, respectively. Let $\{x_t\}$ be generated by Algorithm 1 (called **SGN**) to solve (1). For a given tolerance $\varepsilon > 0$, assume that b_t and $\hat{b}_t$ in (21) are chosen as*

$$\begin{cases} b_t & := \left\lfloor \dfrac{256 M_\phi^2 M^4 \sigma_F^2}{C_g^2 \varepsilon^4} \right\rfloor & = \mathcal{O}\left(\dfrac{\sigma_F^2}{\varepsilon^4}\right), \\ \hat{b}_t & := \left\lfloor \dfrac{2 M_\phi M^2 \sigma_D^2}{\beta_d C_g \varepsilon^2} \right\rfloor & = \mathcal{O}\left(\dfrac{\sigma_D^2}{\varepsilon^2}\right). \end{cases} \tag{22}$$

Furthermore, let $\widehat{x}_T$ be chosen uniformly at random in $\{x_t\}_{t=0}^T$ as the output of Algorithm 1 after T iterations. Then

$$\mathbb{E}\left[\|\widetilde{G}_M(\widehat{x}_T)\|^2\right] = \frac{1}{(T+1)}\sum_{t=0}^{T}\mathbb{E}\left[\|\widetilde{G}_M(x_t)\|^2\right] \le \frac{2M^2\left[\Psi(x_0)-\Psi^\star\right]}{C_g(T+1)} + \frac{\varepsilon^2}{2}, \tag{23}$$

where $C_g := 2M - M_\phi(L_F+\beta_d)$ with $M > \frac{1}{2}M_\phi(L_F+\beta_d)$. Moreover, the total number $\mathcal{T}_f$ of function evaluations $\mathbf{F}(x_t,\xi)$ and the total number $\mathcal{T}_d$ of Jacobian evaluations $\mathbf{F}'(x_t,\zeta)$ to achieve $\mathbb{E}\left[\|\widetilde{G}_M(\widehat{x}_T)\|^2\right] \le \varepsilon^2$ do not exceed

$$\begin{cases} \mathcal{T}_f & := & \left\lfloor \dfrac{1024M^6M_\phi^2\sigma_F^2\left[\Psi(x_0)-\Psi^\star\right]}{C_g^3\varepsilon^6} \right\rfloor & = & \mathcal{O}\left(\dfrac{\sigma_F^2}{\varepsilon^6}\right), \\ \mathcal{T}_d & := & \left\lfloor \dfrac{8M^4M_\phi\sigma_D^2\left[\Psi(x_0)-\Psi^\star\right]}{\beta_d C_g^2\varepsilon^4} \right\rfloor & = & \mathcal{O}\left(\dfrac{\sigma_D^2}{\varepsilon^4}\right). \end{cases} \tag{24}$$

Proof. Let $\mathcal{F}_t := \sigma(x_0, x_1, \cdots, x_{t-1})$ be the σ-field generated by $\{x_0, x_1, \cdots, x_{t-1}\}$. By repeating a similar proof as of (19), but taking the full expectation overall the randomness with $\mathbb{E}[\cdot] = \mathbb{E}[\mathbb{E}[\cdot] \mid \mathcal{F}_{t+1}]$, we have

$$\frac{1}{(T+1)}\sum_{t=0}^{T}\mathbb{E}\left[\|\widetilde{G}_M(x_t)\|^2\right] \le \frac{2M^2\left[\Psi(x_0)-\Psi^\star\right]}{C_g(T+1)} + \frac{\varepsilon^2}{2}, \tag{46}$$

where $C_g := 2M - M_\phi(L_F+\beta_d)$ with $M > \frac{1}{2}M_\phi(L_F+\beta_d)$. Moreover, by the choice of $\widehat{x}_T$, we have $\mathbb{E}\left[\|\widetilde{G}_M(\widehat{x}_T)\|^2\right] = \dfrac{1}{(T+1)}\displaystyle\sum_{t=0}^{T}\mathbb{E}\left[\|\widetilde{G}_M(x_t)\|^2\right]$. Combining this relation and (46), we proves (23).

Next, by Lemma C.2, to guarantee the condition (16) in expectation, i.e.:

$$\begin{cases} \mathbb{E}\left[\|\widetilde{F}(x_t) - F(x_t)\|^2 \mid \mathcal{F}_t\right] & \le & \dfrac{C_g^2\varepsilon^4}{256M_\phi^2M^4}, \\ \mathbb{E}\left[\|\widetilde{J}(x_t) - F'(x_t)\|^2 \mid \mathcal{F}_t\right] & \le & \dfrac{\beta_d C_g\varepsilon^2}{2M^2M_\phi}, \end{cases}$$

we have to choose $\frac{\sigma_F^2}{b_t} \le \frac{C_g^2\varepsilon^4}{256M_\phi^2M^4}$ and $\frac{\sigma_D^2}{\hat{b}_t} \le \frac{\beta_d C_g\varepsilon^2}{2M_\phi M^2}$, which respectively lead to

$$b_t \ge \frac{256M_\phi^2M^4\sigma_F^2}{C_g^2\varepsilon^4} \quad \text{and} \quad \hat{b}_t \ge \frac{2M_\phi M^2\sigma_D^2}{\beta_d C_g\varepsilon^2}.$$

By rounding to the nearest integer, we obtain (22). Using (19), we can see that since $\mathbb{E}\left[\|\widetilde{G}_M(\hat{x}_T)\|^2\right] = \frac{1}{(T+1)}\sum_{t=0}^{T}\mathbb{E}\left[\|\widetilde{G}_M(x_t)\|^2\right]$, to guarantee $\mathbb{E}\left[\|\widetilde{G}_M(\hat{x}_T)\|^2\right] \le \varepsilon^2$, we impose $\frac{2M^2[\Psi(x_0)-\Psi^\star]}{C_g(T+1)} \le \frac{\varepsilon^2}{2}$, which leads to $T := \left\lfloor \frac{4M^2[\Psi(x_0)-\Psi^\star]}{C_g\varepsilon^2} \right\rfloor$. Hence, the total number $\mathcal{T}_f$ of stochastic function evaluations $\mathbf{F}(x_t,\xi)$ can be bounded by

$$\mathcal{T}_f := Tb_t = \left\lfloor \frac{1024M^6M_\phi^2\sigma_F^2\left[\Psi(x_0)-\Psi^\star\right]}{C_g^3\varepsilon^6} \right\rfloor = \mathcal{O}\left(\frac{\sigma_F^2}{\varepsilon^6}\right).$$

Similarly, the total number $\mathcal{T}_d$ of stochastic Jacobian evaluations $\mathbf{F}'(x_t,\zeta)$ can be bounded by

$$\mathcal{T}_d := T\hat{b}_t = \left\lfloor \frac{8M^4M_\phi\sigma_D^2\left[\Psi(x_0)-\Psi^\star\right]}{\beta_d C_g^2\varepsilon^4} \right\rfloor = \mathcal{O}\left(\frac{\sigma_D^2}{\varepsilon^4}\right).$$

These two last estimates prove (24). □

D.2. The Proof of Theorem 4.2: Convergence of The Stochastic Gauss-Newton Method for Solving (2)

Theorem. 4.2. *Suppose that Assumptions 1.1 and 1.2 hold for (2). Let $\widetilde{F}_t$ and $\widetilde{J}_t$ defined by (21) be mini-batch stochastic estimators to approximate $F(x_t)$ and $F'(x_t)$, respectively. Let $\{x_t\}$ be generated by Algorithm 1 for solving (2). Assume that b_t and $\hat{b}_t$ in (21) are chosen such that $b_t := \min\left\{n, \bar{b}_t\right\}$ and $\hat{b}_t := \min\left\{n, \hat{\bar{b}}_t\right\}$ for $t \geq 0$, where*

$$\begin{cases} \bar{b}_0 & := & \left\lfloor \dfrac{32M_{\phi}M^2\sigma_F\left[48\sigma_F M_{\phi}M^2 + C_a\varepsilon^2\right]}{3C_a^2\varepsilon^4} \cdot \log\left(\dfrac{p+1}{\delta}\right)\right\rfloor, \\ \hat{\bar{b}}_0 & := & \left\lfloor \dfrac{4M\sqrt{2M_{\phi}}\sigma_D\left(3M\sqrt{2M_{\phi}}\sigma_D + \sqrt{\beta_d C_a}\varepsilon\right)}{\beta_d C_a\varepsilon^2} \cdot \log\left(\dfrac{p+q}{\delta}\right)\right\rfloor, \\ \bar{b}_t & := & \left\lfloor \dfrac{\left(6\sigma_F^2 + 2\sigma_F\sqrt{C_f}\|x_t - x_{t-1}\|^2\right)}{3C_f^2\|x_t - x_{t-1}\|^4} \cdot \log\left(\dfrac{p+1}{\delta}\right)\right\rfloor \quad (t \geq 1), \\ \hat{\bar{b}}_t & := & \left\lfloor \dfrac{\left(6\sigma_D^2 + 2\sigma_D\sqrt{C_d}\|x_t - x_{t-1}\|\right)}{3C_d\|x_t - x_{t-1}\|^2} \cdot \log\left(\dfrac{p+q}{\delta}\right)\right\rfloor \quad (t \geq 1), \end{cases} \tag{25}$$

*for $\delta \in (0, 1)$, and C_f and C_d given in **Condition 2**, where $\varepsilon > 0$ is a given tolerance.*
Then, we have the following conclusions:

- *With probability at least $1 - \delta$, the bound (20) in Theorem 3.1 still holds.*
- *Moreover, the total number $\mathcal{T}_f$ of stochastic function evaluations $\mathbf{F}(x_t, \xi)$ and the total number $\mathcal{T}_d$ of stochastic Jacobian evaluations $\mathbf{F}'(x_t, \zeta)$ to guarantee $\frac{1}{(T+1)}\sum_{t=0}^{T}\|\widetilde{G}_M(x_t)\|^2 \leq \varepsilon^2$ do not exceed*

$$\begin{cases} \mathcal{T}_f & := & \mathcal{O}\left(\dfrac{\sigma_F^2\left[\Psi(x_0) - \Psi^{\star}\right]}{\varepsilon^6} \cdot \log\left(\dfrac{p+1}{\delta}\right)\right), \\ \mathcal{T}_d & := & \mathcal{O}\left(\dfrac{\sigma_D^2\left[\Psi(x_0) - \Psi^{\star}\right]}{\varepsilon^4} \cdot \log\left(\dfrac{p+q}{\delta}\right)\right). \end{cases} \tag{26}$$

Proof. We first use Lemma C.1 to estimate the total number of samples for $F(x_t)$ and $F'(x_t)$. Let $\mathcal{F}_t := \sigma(x_0, x_1, \cdots, x_{t-1})$ be the σ-field generated by $\{x_0, x_1, \cdots, x_{t-1}\}$. We define $X_i := F_i(x_t) - F(x_t) \in \mathbb{R}^p$ for $i \in \mathcal{B}_t$. Conditioned on $\mathcal{F}_t$, due to the choice of $\mathcal{B}_t$, $\{X_i\}_{i\in\mathcal{B}_t}$ are independent vector-valued random variables and $\mathbb{E}\left[X_i\right] = 0$. Moreover, by Assumption 1.2, we have $\|F_i(x) - F(x)\| \leq \sigma_F$ for all $i \in [n]$. This implies that $\|X_i\| \leq \sigma_F$ a.s. and $\mathbb{E}\left[\|X_i\|^2\right] \leq \sigma_F^2$. Hence, the conditions of Lemma C.1 hold. In addition, we have

$$\sigma_X^2 := \max\left\{\Big\|\sum_{i\in\mathcal{B}_t}\mathbb{E}\left[X_iX_i^{\top}\right]\Big\|, \Big\|\sum_{i\in\mathcal{B}_t}\mathbb{E}\left[X_i^{\top}X_i\right]\Big\|\right\} \leq \sum_{i\in\mathcal{B}_t}\mathbb{E}\left[\|X_i\|^2\right] \leq b_t\sigma_F^2.$$

Since $\widetilde{F}_t := \frac{1}{b_t}\sum_{i\in\mathcal{B}_t}F_i(x_t)$, by Lemma C.1, we have

$$\begin{array}{lcl} \mathbf{Prob}\left(\|\widetilde{F}_t - F(x_t)\| \leq \epsilon\right) & = & \mathbf{Prob}\left(\left\|\sum_{i\in\mathcal{B}_t}X_i\right\| \leq b_t\epsilon\right) \\ & \geq & 1 - (p+1)\exp\left(-\frac{3b_t^2\epsilon^2}{6b_t\sigma_F^2 + 2\sigma_F b_t\epsilon}\right) \\ & = & 1 - (p+1)\exp\left(-\frac{3b_t\epsilon^2}{6\sigma_F^2 + 2\sigma_F\epsilon}\right). \end{array}$$

Let us choose $\delta \in (0, 1]$ such that $\delta \geq (p+1)\exp\left(-\frac{3b_t\epsilon^2}{6\sigma_F^2 + 2\sigma_F\epsilon}\right)$ and $\delta \leq 1$, then $\mathbf{Prob}\left(\|\widetilde{F}_t - F(x_t)\| \leq \epsilon\right) \geq 1 - \delta$. Hence, we have $b_t \geq \left(\frac{6\sigma_F^2 + 2\sigma_F\epsilon}{3\epsilon^2}\right) \cdot \log\left(\frac{p+1}{\delta}\right)$.

To guarantee the first condition of (17), we choose $\epsilon := \frac{C_a\varepsilon^2}{16M_{\phi}M^2}$. Then, the condition on b_0 leads to $b_0 \geq \frac{32M_{\phi}M^2\sigma_F\left(48\sigma_F M_{\phi}M^2 + C_a\varepsilon^2\right)}{3C_g^2\varepsilon^4} \cdot \log\left(\frac{p+1}{\delta}\right)$. To guarantee the first condition of (18), we choose $\epsilon := \sqrt{C_f}\left\|x_t - x_{t-1}\right\|^2$.

Then, the condition on b_t leads to $b_t \geq \frac{(6\sigma_F^2+2\sigma_F\sqrt{C_f}\|x_t-x_{t-1}\|^2}{3C^2\|x_t-x_{t-1}\|^4} \cdot \log\left(\frac{p+1}{\delta}\right)$. Rounding both b_0 and b_t, we obtain

$$\begin{cases} \bar{b}_0 & := & \left\lfloor \frac{32M_\phi M^2\sigma_F\left[48\sigma_F M_\phi M^2+C_a\varepsilon^2\right]}{3C_a^2\varepsilon^4} \cdot \log\left(\frac{p+1}{\delta}\right) \right\rfloor & = & \mathcal{O}\left(\frac{\sigma_F^2}{\varepsilon^4} \cdot \log\left(\frac{p}{\delta}\right)\right), \\ \bar{b}_t & := & \left\lfloor \frac{(6\sigma_F^2+2\sigma_F\sqrt{C_f}\|x_t-x_{t-1}\|^2}{3C^2\|x_t-x_{t-1}\|^4} \cdot \log\left(\frac{p+1}{\delta}\right) \right\rfloor & = & \mathcal{O}\left(\frac{\sigma_F^2}{\|x_t-x_{t-1}\|^4} \cdot \log\left(\frac{p}{\delta}\right)\right), \quad \forall t \geq 1. \end{cases}$$

Since $b_t \leq n$ for all $t \geq 0$, we have $b_t := \min\left\{n, \bar{b}_t\right\}$ for $t \geq 0$, which proves the first part of (25).

Next, we estimate a sample size for $\widetilde{J}_t$. Let us define $Y_i := F_i'(x_t) - F'(x_t)$. Then, similar to the above proof of X_i for F, we have $\widetilde{J}_t - F'(x_t) = \frac{1}{\hat{b}_t}\sum_{i\in\hat{\mathcal{B}}_t}(F_i'(x_t) - F'(x_t)) = \frac{1}{\hat{b}_t}\sum_{i\in\hat{\mathcal{B}}_t} Y_i$. Under Assumption 1.2, the sequence $\{Y_i\}$ satisfies all conditions of Lemma C.1. Hence, we obtain

$$\mathbf{Prob}\left(\|\widetilde{J}_t - F'(x_t)\| \leq \epsilon\right) \geq 1 - (p+q)\exp\left(\frac{-3\hat{b}_t\epsilon^2}{6\sigma_D^2 + 2\sigma_D\epsilon}\right).$$

Hence, we can choose $\hat{b}_t \geq \left[\frac{6\sigma_D^2+2\sigma_D\epsilon}{3\epsilon^2}\right] \cdot \log\left(\frac{p+q}{\delta}\right)$. From the second condition of (17), if we choose $\epsilon := \frac{\sqrt{\beta_d C_a}\varepsilon}{M\sqrt{2M_\phi}}$, then we have $\hat{b}_0 \geq \frac{4M\sqrt{2M_\phi}\sigma_D\left(3M\sqrt{2M_\phi}\sigma_D+\sqrt{\beta_d C_a}\varepsilon\right)}{\beta_d C_a\varepsilon^2} \cdot \log\left(\frac{p+q}{\delta}\right)$. From the second condition of (18), if we choose $\epsilon := \sqrt{C_d}\|x_t - x_{t-1}\|$, then we have $\hat{b}_t \geq \frac{\left(6\sigma_D^2+2\sigma_D\sqrt{C_d}\|x_t-x_{t-1}\|\right)}{3C_d\|x_t-x_{t-1}\|^2} \cdot \log\left(\frac{p+q}{\delta}\right)$. Rounding $\hat{b}_t$, we obtain

$$\begin{array}{lclcl} \hat{\bar{b}}_0 & := & \left\lfloor \frac{4M\sqrt{2M_\phi}\sigma_D\left(3M\sqrt{2M_\phi}\sigma_D+\sqrt{\beta_d C_a}\varepsilon\right)}{\beta_d C_a\varepsilon^2} \cdot \log\left(\frac{p+q}{\delta}\right) \right\rfloor & = & \mathcal{O}\left(\frac{\sigma_D^2}{\varepsilon^2} \cdot \log\left(\frac{p+q}{\delta}\right)\right), \\ \hat{\bar{b}}_t & := & \left\lfloor \frac{\left(6\sigma_D^2+2\sigma_D\sqrt{C_d}\|x_t-x_{t-1}\|\right)}{3C_d\|x_t-x_{t-1}\|^2} \cdot \log\left(\frac{p+q}{\delta}\right) \right\rfloor & = & \mathcal{O}\left(\frac{\sigma_D^2}{\|x_t-x_{t-1}\|^2} \cdot \log\left(\frac{p+q}{\delta}\right)\right), \quad t \geq 1. \end{array}$$

Since $\hat{b}_t \leq n$ for all $t \geq 0$, combining these conditions, we obtain $\hat{b}_t := \min\{n, \hat{\bar{b}}_t\}$ for $t \geq 0$, which proves the second part of (25).

For $t \geq 1$, we have $\|\widetilde{G}_M(x_{t-1})\| = M\|x_t - x_{t-1}\| > \varepsilon$. Otherwise, the algorithm has been terminated. Therefore, we can even bound b_t and $\hat{b}_t$ as

$$b_t \leq \frac{2M^2\sigma_F(3M^2\sigma_F + \sqrt{C_f}\varepsilon^2)}{3C^2\varepsilon^4} \cdot \log\left(\frac{p+1}{\delta}\right) \quad \text{and} \quad \hat{b}_t \leq \frac{M\left(6M\sigma_D^2 + 2\sigma_D\sqrt{C_d}\varepsilon\right)}{3C_d\varepsilon^2} \cdot \log\left(\frac{p+q}{\delta}\right).$$

From (20), to guarantee $\frac{1}{(T+1)}\sum_{t=0}^{T}\|\widetilde{G}_M(x_t)\|^2 \leq \varepsilon^2$, we impose $\frac{2M^2[\Psi(x_0)-\Psi^\star]}{C_a(T+1)} \leq \frac{\varepsilon^2}{2}$, which leads to $T := \left\lfloor \frac{4M^2[\Psi(x_0)-\Psi^\star]}{C_a\varepsilon^2} \right\rfloor$. Hence, the total number $\mathcal{T}_f$ of stochastic function evaluations $\mathbf{F}(\cdot, \xi)$ can be bounded by

$$\begin{array}{lcl} \mathcal{T}_f & := & b_0 + (T-1)b_t \\ & \leq & \left[\frac{32M_\phi M^2\sigma_F\left(48\sigma_F M_\phi M^2+C_a\varepsilon^2\right)}{3C_a^2\varepsilon^4} + \frac{8M^4\sigma_F(3M^2\sigma_F+\sqrt{C_f}\varepsilon^2)[\Psi(x_0)-\Psi^\star]}{3C^2C_a\varepsilon^6}\right] \cdot \log\left(\frac{p+1}{\delta}\right) \\ & = & \mathcal{O}\left(\frac{\sigma_F^2[\Psi(x_0)-\Psi^\star]}{\varepsilon^6} \cdot \log\left(\frac{p+1}{\delta}\right)\right). \end{array}$$

Similarly, the total number $\mathcal{T}_d$ of stochastic Jacobian evaluations $\mathbf{F}'(\cdot, \zeta)$ can be bounded by

$$\begin{array}{lcl} \mathcal{T}_d & := & \hat{b}_0 + (T-1)\hat{b}_t \\ & \leq & \left[\frac{4M\sqrt{2M_\phi}\sigma_D\left(3M\sqrt{2M_\phi}\sigma_D+\sqrt{\beta_d C_a}\varepsilon\right)}{\beta_d C_a\varepsilon^2} + \frac{4M^3[\Psi(x_0)-\Psi^\star]\left(6M\sigma_D^2+2\sigma_D\sqrt{C_d}\varepsilon\right)}{3C_d C_a\varepsilon^4}\right] \cdot \log\left(\frac{p+q}{\delta}\right) \\ & = & \mathcal{O}\left(\frac{\sigma_D^2[\Psi(x_0)-\Psi^\star]}{\varepsilon^4} \cdot \log\left(\frac{p+q}{\delta}\right)\right). \end{array}$$

Taking the upper bounds, these two last estimates prove (26). □

D.3. The Proof of Theorem 4.3: Convergence and Complexity Analysis of Algorithm 2 for (1)

Theorem. 4.3. *Suppose that Assumptions 1.1 and 1.2, and 4.1 are satisfied for* (1)*. Let* $\{x_t^{(s)}\}_{t=0\to m}^{s=1\to S}$ *be generated by Algorithm 2 to solve* (1)*. Let* θ_F *and* m *be chosen by* (28)*, and the mini-batches* b_s*,* $\hat{b}_s$*,* $b_t^{(s)}$*, and* $\hat{b}_t^{(s)}$ *be set as in* (29)*. Assume that the output* $\widehat{x}_T$ *of Algorithm 2 is chosen uniformly at random in* $\{x_t^{(s)}\}_{t=0\to m}^{s=1\to S}$*. Then:*

(a) *For a given tolerance* $\varepsilon > 0$*, the following bound holds*

$$\mathbb{E}\left[\|\widetilde{G}_M(\widehat{x}_T)\|^2\right] = \frac{1}{S(m+1)}\sum_{s=1}^{S}\sum_{t=0}^{m}\mathbb{E}\left[\|\widetilde{G}_M(x_t)\|^2\right] \le \varepsilon^2. \tag{30}$$

(b) *The total number of iterations* T *to obtain* $\mathbb{E}\left[\|\widetilde{G}_M(\widehat{x}_T)\|^2\right] \le \varepsilon^2$ *is at most*

$$T := S(m+1) = \left\lfloor \frac{8M^2\left[\Psi(\widetilde{x}^0) - \Psi^{\star}\right]}{\theta_F\varepsilon^2} \right\rfloor = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right).$$

Moreover, the total numbers $\mathcal{T}_f$ *and* $\mathcal{T}_d$ *of stochastic function evaluations* $\mathbf{F}(x_t, \xi)$ *and stochastic Jacobian evaluations* $\mathbf{F}'(x_t, \zeta)$*, respectively do not exceed:*

$$\left\{\begin{array}{lcl} \mathcal{T}_f & := & \mathcal{O}\left(\dfrac{M_{\phi}^2\sigma_F^2}{\theta_F^2\varepsilon^4} + \dfrac{M^4M_{\phi}^2\left[\Psi(\widetilde{x}^0) - \Psi^{\star}\right]}{\theta_F^2\varepsilon^5}\right), \\ \mathcal{T}_d & := & \mathcal{O}\left(\dfrac{M_{\phi}\sigma_D^2}{\theta_F\varepsilon^2} + \dfrac{M^2M_{\phi}\left[\Psi(\widetilde{x}^0) - \Psi^{\star}\right]}{\theta_F\varepsilon^3}\right). \end{array}\right. \tag{31}$$

Proof. We first analyze the inner loop. Using (15) with $x := x_t^{(s)}$ and $T_M(x) = x_{t+1}^{(s)}$, and then taking the expectation conditioned on $\mathcal{F}_{t+1}^{(s)} := \sigma(x_0^{(s)}, x_1^{(s)}, \cdots, x_t^{(s)})$, we have

$$\begin{array}{lcl} \mathbb{E}\left[\phi(F(x_{t+1}^{(s)})) \mid \mathcal{F}_{t+1}^{(s)}\right] & \le & \phi(F(x_t^{(s)})) - \frac{(2M - M_{\phi}(L_F+\beta_d))}{2}\mathbb{E}\left[\|x_{t+1}^{(s)} - x_t^{(s)}\|^2 \mid \mathcal{F}_{t+1}^{(s)}\right] \\ & & + \frac{L_{\phi}}{\xi_t^s}\mathbb{E}\left[\|F(x_t^{(s)}) - \widetilde{F}(x_t^{(s)})\|^2 \mid \mathcal{F}_{t+1}^{(s)}\right] \\ & & + \frac{M_{\phi}}{2\beta_d}\mathbb{E}\left[\|F'(x_t^{(s)}) - \widetilde{J}(x_t^{(s)})\|^2 \mid \mathcal{F}_{t+1}^{(s)}\right] + M_{\phi}\xi_t^s, \end{array}$$

for any $\xi_t^s > 0$, where we use $2ab \le a^2 + b^2$ and the Jensen inequality $\left(\mathbb{E}\left[\|F(x_t^{(s)}) - \widetilde{F}(x_t^{(s)})\| \mid \mathcal{F}_{t+1}^{(s)}\right]\right)^2 \le \mathbb{E}\left[\|F(x_t^{(s)}) - \widetilde{F}(x_t^{(s)})\|^2 \mid \mathcal{F}_{t+1}^{(s)}\right]$ in the second line. Taking the full expectation both sides of the last inequality, and noting that $\Psi(x) = \phi(F(x))$, we obtain

$$\begin{array}{lcl} \mathbb{E}\left[\Psi(x_{t+1}^{(s)})\right] & \le & \mathbb{E}\left[\Psi(x_t^{(s)})\right] - \frac{C_g}{2}\mathbb{E}\left[\|x_{t+1}^{(s)} - x_t^{(s)}\|^2\right] + \frac{L_{\phi}}{\xi_t^s}\mathbb{E}\left[\|F(x_t^{(s)}) - \widetilde{F}(x_t^{(s)})\|^2\right] + M_{\phi}\xi_t^s \\ & & + \frac{M_{\phi}}{2\beta_d}\mathbb{E}\left[\|F'(x_t^{(s)}) - \widetilde{J}(x_t^{(s)})\|^2\right], \end{array} \tag{47}$$

where $C_g := 2M - M_{\phi}(L_F + \beta_d) > 0$, and $\beta_d > 0$ and $\xi_t^s > 0$ are given.

Next, from Lemma C.3, using the Lipschitz continuity of F' in Assumption 1.2, we have

$$\mathbb{E}\left[\|\widetilde{J}_t^{(s)} - F'(x_t^{(s)})\|^2\right] \le \mathbb{E}\left[\|\widetilde{J}_{t-1}^{(s)} - F'(x_{t-1}^{(s)})\|^2\right] + \frac{L_F^2}{\hat{b}_t^{(s)}}\mathbb{E}\left[\|x_t^{(s)} - x_{t-1}^{(s)}\|^2\right]. \tag{48}$$

Similarly, using Lemma C.3, we also have

$$\mathbb{E}\left[\|\widetilde{F}_t^{(s)} - F(x_t^{(s)})\|^2 \mid \mathcal{F}_{t+1}^{(s)}\right] \le \|\widetilde{F}_{t-1}^{(s)} - F(x_{t-1}^{(s)})\|^2 + \frac{1}{b_t}\mathbb{E}_{\xi}\left[\|\mathbf{F}(x_t, \xi) - \mathbf{F}(x_{t-1}, \xi)\|^2\right].$$

Taking the full expectation both sides of this inequality, and using Assumption 4.1, we obtain

$$\mathbb{E}\left[\|\widetilde{F}_t^{(s)} - F(x_t^{(s)})\|^2\right] \leq \mathbb{E}\left[\|\widetilde{F}_{t-1}^{(s)} - F(x_{t-1}^{(s)})\|^2\right] + \frac{M_F^2}{b_t^{(s)}}\mathbb{E}\left[\|x_t^{(s)} - x_{t-1}^{(s)}\|^2\right]. \tag{49}$$

Let us define a Lyapunov function as

$$\mathcal{L}(x_t^{(s)}) := \mathbb{E}\left[\Psi(x_t^{(s)})\right] + \frac{a_t^s}{2}\mathbb{E}\left[\|\widetilde{F}_t^{(s)} - F(x_t^{(s)}\|^2\right] + \frac{c_t^s}{2}\mathbb{E}\left[\|\widetilde{J}_t^{(s)} - F'(x_t^{(s)}\|^2\right], \tag{50}$$

for some $a_t^s > 0$ and $c_t^s > 0$.

Combining (47), (48), and (49), and then using the definition of $\mathcal{L}$ in (50), we have

$$\begin{aligned}
\mathcal{L}(x_{t+1}^{(s)}) &= \mathbb{E}\left[\Psi(x_{t+1}^{(s)})\right] + \tfrac{a_{t+1}^s}{2}\mathbb{E}\left[\|\widetilde{F}_{t+1}^{(s)} - F(x_{t+1}^{(s)}\|^2\right] + \tfrac{c_{t+1}^s}{2}\mathbb{E}\left[\|\widetilde{J}_{t+1}^{(s)} - F'(x_{t+1}^{(s)}\|^2\right] \\
&\leq \mathbb{E}\left[\Psi(x_t^{(s)})\right] - \left[\tfrac{C_g}{2} - \tfrac{M_F^2 a_{t+1}^s}{2b_{t+1}^{(s)}} - \tfrac{L_F^2 c_{t+1}^s}{2\hat{b}_{t+1}^{(s)}}\right]\mathbb{E}\left[\|x_{t+1}^{(s)} - x_t^{(s)}\|^2\right] + M_\phi \xi_t^s \\
&\quad + \left(\tfrac{a_{t+1}^s}{2} + \tfrac{L_\phi}{\xi_t^s}\right)\mathbb{E}\left[\|F(x_t^{(s)}) - \widetilde{F}(x_t^{(s)})\|^2\right] + \left(\tfrac{c_{t+1}^s}{2} + \tfrac{M_\phi}{2\beta_d}\right)\mathbb{E}\left[\|F'(x_t^{(s)}) - \widetilde{J}(x_t^{(s)})\|^2\right].
\end{aligned} \tag{51}$$

If we assume that

$$a_t^s \geq a_{t+1}^s + \frac{M_\phi}{\xi_t^s} \quad \text{and} \quad c_t^s \geq c_{t+1}^s + \frac{M_\phi}{\beta_d}, \tag{52}$$

then, from (51), we have

$$\mathcal{L}(x_{t+1}^{(s)}) \leq \mathcal{L}(x_t^{(s)}) - \frac{\rho_{t+1}^s}{2}\mathbb{E}\left[\|x_{t+1}^{(s)} - x_t^{(s)}\|^2\right] + M_\phi \xi_t^s, \tag{53}$$

where $\rho_{t+1}^s := C_g - \frac{M_F^2 a_{t+1}^s}{b_{t+1}^{(s)}} - \frac{L_F^2 c_{t+1}^s}{\hat{b}_{t+1}^{(s)}}$.

Let us first fix $\xi_t^s := \xi > 0$. Next, we choose $a_t^s := (m+1-t)\frac{M_\phi}{\xi}$ and $c_t^s := (m+1-t)\frac{M_\phi}{\beta_d}$. Clearly, $a_{m+1}^s = c_{m+1}^s = 0$ and they both satisfy the condition (52). Then, we choose $b_t^{(s)} := \frac{1}{\gamma_1}a_t^s = (m+1-t)\frac{M_\phi}{\gamma_1\xi}$ and $\hat{b}_t^{(s)} = \frac{1}{\gamma_2}c_t^s = \frac{M_\phi}{\beta_d\gamma_2}(m+1-t)$ for some $\gamma_1 > 0$ and $\gamma_2 > 0$. In this case, we have $\rho_t^s = C_g - M_F^2\gamma_1 - L_F^2\gamma_2 \equiv \theta_F > 0$ due to (28) by appropriately choosing γ_1 and γ_2. Consequently, (53) reduces to

$$\mathcal{L}(x_{t+1}^{(s)}) \leq \mathcal{L}(x_t^{(s)}) - \frac{\theta_F}{2}\mathbb{E}\left[\|x_{t+1}^{(s)} - x_t^{(s)}\|^2\right] + M_\phi \xi.$$

Summing up this inequality from $t = 0$ to $t = m$, we obtain

$$\frac{\theta_F}{2}\sum_{t=0}^{m}\mathbb{E}\left[\|x_{t+1}^{(s)} - x_t^{(s)}\|^2\right] \leq \mathcal{L}(x_0^{(s)}) - \mathcal{L}(x_{m+1}^{(s)}) + (m+1)M_\phi \xi.$$

Using the fact that $\widetilde{x}^{s-1} = x_0^{(s)}$ and $\widetilde{x}^s = x_{m+1}^{(s)}$, we have

$$\frac{\theta_F}{2}\sum_{t=0}^{m}\mathbb{E}\left[\|x_{t+1}^{(s)} - x_t^{(s)}\|^2\right] \leq \mathcal{L}(\widetilde{x}^{s-1}) - \mathcal{L}(\widetilde{x}^s) + (m+1)M_\phi \xi.$$

Summing up this inequality from $s = 1$ to S and multiplying the result by $\frac{2}{\theta_F S(m+1)}$, we obtain

$$\frac{1}{S(m+1)}\sum_{s=1}^{S}\sum_{t=0}^{m}\mathbb{E}\left[\|x_{t+1}^{(s)} - x_t^{(s)}\|^2\right] \leq \frac{2\left[\mathcal{L}(\widetilde{x}^0) - \mathcal{L}(\widetilde{x}^S)\right]}{\theta_F S(m+1)} + \frac{2M_\phi \xi}{\theta_F}. \tag{54}$$

Since $\mathcal{L}(\widetilde{x}^0) = \Psi(\widetilde{x}^0) + \frac{(m+1)M_\phi}{2\xi}\mathbb{E}\left[\|\widetilde{F}_0 - F(\widetilde{x}^0)\|^2\right] + \frac{(m+1)M_\phi}{2\beta_d}\mathbb{E}\left[\|\widetilde{J}_0 - F'(\widetilde{x}^0)\|^2\right]$ and $\mathcal{L}(\widetilde{x}^S) = \mathbb{E}\left[\Psi(\widetilde{x}^S\right] \geq \Phi^\star$, we obtain from (54) that

$$\begin{aligned}
\frac{1}{S(m+1)}\sum_{s=1}^{S}\sum_{t=0}^{m}\mathbb{E}\left[\|x_{t+1}^{(s)} - x_t^{(s)}\|^2\right] &\leq \frac{2\left[\Psi(\widetilde{x}^0) - \Psi^\star\right]}{\theta_F S(m+1)} + \frac{M_\phi}{\xi\theta_F S}\mathbb{E}\left[\|\widetilde{F}_0 - F(\widetilde{x}^0)\|^2\right] \\
&\quad + \frac{M_\phi}{\theta_F\beta_d S}\mathbb{E}\left[\|\widetilde{J}_0 - F'(\widetilde{x}^0)\|^2\right] + \frac{2M_\phi \xi}{\theta_F}.
\end{aligned} \tag{55}$$

Note that $\mathbb{E}\left[\|\widetilde{F}_0 - F(\widetilde{x}^0)\|^2\right] \le \frac{\sigma_F^2}{b}$ and $\mathbb{E}\left[\|\widetilde{J}_0 - F'(\widetilde{x}^0)\|^2\right] \le \frac{\sigma_D^2}{\hat{b}}$ due to the choice of $b_s = b > 0$ and $\hat{b}_s = \hat{b} > 0$ at Step 3 of Algorithm 2. Hence, we can further bound (55) as

$$\frac{1}{S(m+1)}\sum_{s=1}^{S}\sum_{t=0}^{m}\mathbb{E}\left[\|x_{t+1}^{(s)} - x_t^{(s)}\|^2\right] \le \frac{2\big[\Psi(\widetilde{x}^0) - \Psi^{\star}\big]}{\theta_F S(m+1)} + \frac{M_\phi\sigma_F^2}{\xi\theta_F Sb} + \frac{M_\phi\sigma_D^2}{\theta_F\beta_d S\hat{b}} + \frac{2M_\phi\xi}{\theta_F}.$$

Since $\|\widetilde{G}_M(x_t^{(s)})\| = M\|x_{t+1}^{(s)} - x_t^{(s)}\|$, to guarantee $\frac{1}{S(m+1)}\sum_{s=1}^{S}\sum_{t=0}^{m}\mathbb{E}\left[\|\widetilde{G}_M(x_t^{(s)})\|^2\right] \le \varepsilon^2$ for a given tolerance $\varepsilon > 0$, we need to set

$$\frac{2\big[\Psi(\widetilde{x}^0) - \Psi^{\star}\big]}{\theta_F S(m+1)} + \frac{M_\phi\sigma_F^2}{\xi\theta_F Sb} + \frac{M_\phi\sigma_D^2}{\theta_F\beta_d S\hat{b}} + \frac{2M_\phi\xi}{\theta_F} = \frac{\varepsilon^2}{M^2}.$$

Let us break this condition into

$$\frac{2\big[\Psi(\widetilde{x}^0) - \Psi^{\star}\big]}{\theta_F S(m+1)} = \frac{\varepsilon^2}{4M^2} \quad \text{and} \quad \frac{M_\phi\sigma_F^2}{\xi\theta_F Sb} = \frac{M_\phi\sigma_D^2}{\theta_F\beta_d S\hat{b}} = \frac{2M_\phi\xi}{\theta_F} = \frac{\varepsilon^2}{4M^2}.$$

Hence, we can choose $\xi := \frac{\theta_F\varepsilon^2}{8M^2M_\phi}$, $\hat{b} := \frac{4M_\phi\sigma_D^2}{\theta_F\beta_d M^2 S\varepsilon^2}$, $b := \frac{2M_\phi^2\sigma_F^2}{\theta_F^2 M^2 S\varepsilon^4}$, and $S(m+1) = \frac{8M^2\left[\Psi(\widetilde{x}^0) - \Psi^{\star}\right]}{\theta_F\varepsilon^2}$.

Now, let us choose $m + 1 := \frac{\hat{C}}{\varepsilon}$ for some constant $\hat{C} > 0$. Then, we can estimate the total number $\mathcal{T}_f$ of stochastic function evaluations $\mathbf{F}(x_t^{(s)}, \xi)$ as follows:

$$\begin{array}{lcl}
\mathcal{T}_f & := & \sum_{s=1}^{S} b_s + \sum_{s=1}^{S}\sum_{t=0}^{m} b_t^{(s)} = Sb + \frac{M_\phi}{\gamma_1\xi}\sum_{s=1}^{S}\sum_{t=0}^{m}(m+1-t) \\
& = & \frac{2M_\phi^2\sigma_F^2}{\theta_F^2 M^2\varepsilon^4} + \frac{8M^2M_\phi^2}{\gamma_1\theta_F\varepsilon^2}\cdot\frac{S(m+1)(m+2)}{2} \\
& = & \frac{2M_\phi^2\sigma_F^2}{\theta_F^2 M^2\varepsilon^4} + \frac{8M^2M_\phi^2}{\gamma_1\theta_F\varepsilon^2}\cdot\frac{8M^2\left[\Psi(\widetilde{x}^0) - \Psi^{\star}\right]}{\theta_F\varepsilon^2}\cdot\frac{\hat{C}+\varepsilon}{2\varepsilon} \\
& = & \mathcal{O}\left(\frac{M_\phi^2\sigma_F^2}{\theta_F^2\varepsilon^4} + \frac{M^4M_\phi^2\left[\Psi(\widetilde{x}^0) - \Psi^{\star}\right]}{\theta_F^2\varepsilon^5}\right).
\end{array}$$

Similarly, the total number $\mathcal{T}_d$ of stochastic Jacobian evaluations $\mathbf{F}'(x_t^{(s)}, \zeta)$ can be bounded as

$$\begin{array}{lcl}
\mathcal{T}_d & := & \sum_{s=1}^{S}\hat{b}_s + \sum_{s=1}^{S}\sum_{t=0}^{m}\hat{b}_t^{(s)} = S\hat{b} + \frac{M_\phi S}{\beta_d\gamma_2}\sum_{t=0}^{m}(m+1-t) \\
& \le & \frac{4M_\phi\sigma_D^2}{\theta_F\beta_d M^2\varepsilon^2} + \frac{8M^2M_\phi\left[\Psi(\widetilde{x}^0) - \Psi^{\star}\right]}{\beta_d\gamma_2\theta_F\varepsilon^2}\cdot\frac{\hat{C}+\varepsilon}{2\varepsilon} \\
& = & \mathcal{O}\left(\frac{M_\phi\sigma_D^2}{\theta_F\varepsilon^2} + \frac{M^2M_\phi\left[\Psi(\widetilde{x}^0) - \Psi^{\star}\right]}{\theta_F\varepsilon^3}\right).
\end{array}$$

Hence, taking the upper bounds, we have proven (31). □

E. Solution Routines for Computing Gauss-Newton Search Directions

One main step of SGN methods is to compute the Gauss-Newton direction by solving the subproblem (11). This subproblem is also called a **prox-linear operator**, which can be rewritten as

$$\min_{d\in\mathbb{R}^p}\left\{\phi(\widetilde{F}_t + \widetilde{J}_t d) + \hat{g}(d) + \tfrac{M}{2}\|d\|_2^2\right\}, \tag{56}$$

where $\widetilde{F}_t \approx F(x_t)$, $\widetilde{J}_t \approx F'(x_t)$, $d := x - x_t$, ϕ is convex, $\hat{g}(d) := g(x_t + d)$, and $M > 0$ is given. This is a basic convex problem, and we can apply different methods to solve it. Here, we describe two methods for solving (56).

E.1. Accelerated Dual Proximal-Gradient Method

For accelerated dual proximal-gradient method, we consider the case $\hat{g}(d) = 0$ for simplicity. Using Fenchel's conjugate of ϕ, we can write $\phi(\widetilde{F}_t + \widetilde{J}_t d) = \max\left\{\langle \widetilde{F}_t + \widetilde{J}_t d, u\rangle - \phi^*(u)\right\}$. Assume that strong duality holds for (56), then using this expression, we can write it as

$$\min_d \max_u \left\{\langle \widetilde{F}_t + \widetilde{J}_t d, u\rangle - \phi^*(u) + \frac{M}{2}\|d\|_2^2\right\} \quad \Leftrightarrow \quad \max_u \left\{\min_d \left\{\langle \widetilde{F}_t + \widetilde{J}_t d, u\rangle + \frac{M}{2}\|d\|_2^2\right\} - \phi^*(u)\right\}.$$

Solving the inner problem $\min_d \left\{\langle \widetilde{F}_t + \widetilde{J}_t d, u\rangle + \frac{M}{2}\|d\|_2^2\right\}$, we obtain $d^*(u) := -\frac{1}{M}\widetilde{J}^\top u$. Substituting it into the objective, we eventually obtain the dual problem as follows:

$$\min_u \left\{\frac{1}{2M}\|\widetilde{J}_t^\top u\|_2^2 - \langle \widetilde{F}_t, u\rangle + \phi^*(u)\right\}. \tag{57}$$

We can solve this problem by an accelerated proximal-gradient method (Beck & Teboulle, 2009; Nesterov, 2004), which is described as follows.

Algorithm 3 (Accelerated Dual Proximal-Gradient **(ADPG)**)

```
Initialization: Choose $u_0 \in \mathbb{R}^m$. Set $\tau_0 := 1$ and $\hat{u}_0 := u_0$. Evaluate $L := \frac{1}{M}\|\widetilde{J}_t^\top \widetilde{J}_t\|$.
For $k := 0, \cdots, k_{\max}$ do
    $u_{k+1} := \mathrm{prox}_{(1/L)\phi^*}\left(\hat{u}_k - \frac{1}{L}(\frac{1}{M}\widetilde{J}_t\widetilde{J}_t^\top \hat{u}_k - \widetilde{F}_t)\right)$.
    $\tau_{k+1} := \frac{1+\sqrt{1+4\tau_k^2}}{2}$.
    $\hat{u}_{k+1} := u_{k+1} + \left(\frac{\tau_k - 1}{\tau_{k+1}}\right)(u_{k+1} - u_k)$.
End For
Output: Reconstruct $d^* := -\frac{1}{M}\widetilde{J}_t^\top u_k$ as an approximate solution of (56).
```

Note that in Algorithm 3, we use the proximal operator $\mathrm{prox}_{\lambda\phi^*}$ of ϕ^*. However, by Moreau's identity, $\mathrm{prox}_{\lambda\phi^*}(v) + \lambda\mathrm{prox}_{\phi/\lambda}(v/\lambda) = v$, we can again use the proximal operator $\mathrm{prox}_{\phi/\lambda}$ of ϕ.

E.2. Primal-Dual First-Order Methods

We can apply any primal-dual algorithm from the literature (Bauschke & Combettes, 2017; Chambolle & Pock, 2011; Esser, 2010; Goldstein et al., 2013; Tran-Dinh et al., 2018; Tran-Dinh, 2019) to solve (56). Here, we describe the well-known Chambolle-Pock's primal-dual method (Chambolle & Pock, 2011) to solve (56).

Let us define $\hat{\phi}(z) := \phi(z + F_k)$ and $\hat{\psi}(d) := \hat{g}(d) + \frac{M}{2}\|d\|^2$. Since (56) is strongly convex with the strong convexity parameter $\mu_{\hat{\psi}} := M$, we can apply the strongly convex primal-dual variant as follows.

Choose $\sigma_0 > 0$ and $\tau_0 > 0$ such that $\tau_0\sigma_0 \le \frac{1}{\|\widetilde{J}_t^\top \widetilde{J}_t\|}$. For example, we can choose $\sigma_0 = \tau_0 = \frac{1}{\|\widetilde{J}_t\|}$, or we choose $\sigma_0 > 0$ first, and choose $\tau_0 := \frac{1}{\sigma_0\|\widetilde{J}_t^\top \widetilde{J}_t\|}$. Choose $d_0 \in \mathbb{R}^p$ and $u_0 \in \mathbb{R}^m$ and set $\bar{d}_0 := d_0$. Then, at each iteration $k \ge 0$, we update

$$\left\{\begin{array}{lcl} u_{k+1} & := & \mathrm{prox}_{\sigma_k\hat{\phi}^*}\left(u_k + \sigma_k\widetilde{J}_t\bar{d}_k\right), \\ d_{k+1} & := & \mathrm{prox}_{\tau_k\hat{\psi}}\left(d_k - \tau_k\widetilde{J}_t^\top u_{k+1}\right), \\ \theta_k & := & 1/\sqrt{1 + 2M\tau_k}, \\ \tau_{k+1} & := & \theta_k\tau_k, \\ \sigma_{k+1} & := & \sigma_k/\theta_k, \\ \bar{d}_{k+1} & := & d_{k+1} + \theta_k(d_{k+1} - d_k). \end{array}\right. \tag{58}$$

Alternatively to the Accelerated Dual Proximal-Gradient and the primal-dual methods, we can also apply the alternating direction method of multipliers (ADMM) to solve (56). However, this method requires to solve a linear system, that may not scale well when the dimension p is large.

F. Details of The Experiments in Section 5

In this supplementary document, we provide the details of our experiments in Section 5, including modeling, data generating routines, and experiment configurations. We also provide more experiments for both examples. All algorithms are implemented in Python 3.6 running on a Macbook Pro with 2.3 GHz Quad-Core, 8 GB RAM and on a Linux-based computing node, called Longleaf, where each node has 24 physical cores, 2.50 GHz processors, and 256 GB RAM.

F.1. Stochastic Nonlinear Equations

Our goal is to solve the following nonlinear equation in expectation as described in Subsection 5.1:

$$F(x) = 0, \ \text{where} \ F(x) := \mathbb{E}_{\xi}\left[\mathbf{F}(x,\xi)\right]. \tag{59}$$

Here, $\mathbf{F}$ is a stochastic vector function from $\mathbb{R}^p \times \Omega \to \mathbb{R}^q$. As discussed in the main text, (59) covers the first-order optimality condition $\mathbb{E}_{\xi}\left[\nabla_x \mathbf{G}(x,\xi)\right] = 0$ of a stochastic optimization problem $\min_x \mathbb{E}_{\xi}\left[\mathbf{G}(x,\xi)\right]$ as a special case. More generally, it also covers the KKT condition of a stochastic optimization problem with equality constraints. However, these problems may not have stationary point, which leads to an inconsistency of (59). As a remedy, we can instead consider

$$\min_x \left\{\Psi(x) := \left\|\mathbb{E}_{\xi}\left[\mathbf{F}(x,\xi)\right]\right\|\right\}, \tag{60}$$

for a given norm $\|\cdot\|$ (e.g., ℓ_1-norm or ℓ_2-norm). Problem (59) also covers the expectation formulation of stochastic nonlinear equations such as stochastic ODEs or PDEs.

In our experiment from Subsection 5.1, we only consider one instance of (60) by choosing $q = 4$ and $\mathbf{F}_j$ $(j = 1, \cdots, q)$ as

$$\begin{cases} \mathbf{F}_1(x,\xi_i) & := & (1 - \tanh(y_i(a_i^{\top}x + b_i)), \\ \mathbf{F}_2(x,\xi_i) & := & \left(1 - \frac{1}{1+\exp(-y_i(a_i^{\top}x+b_i))}\right)^2, \\ \mathbf{F}_3(x,\xi_i) & := & \log(1+\exp(-y_i(a_i^{\top}x+b_i))) - \log(1+\exp(-y_i(a_i^{\top}x+b_i)-1)), \\ \mathbf{F}_4(x,\xi_i) & := & \log(1+(y_i(a_i^{\top}x+b_i)-1)^2), \end{cases} \tag{61}$$

where a_i is the i-row of an input matrix $A \in \mathbb{R}^{n\times p}$, $y \in \{-1,1\}^n$ is a vector of labels, $b \in \mathbb{R}^n$ is a bias vector in binary classification, and $\xi_i := (a_i, b_i, y_i)$. Note that the binary classification problem with nonconvex loss has been widely studied in the literature, including Zhao et al. (2010), where one aims at solving:

$$\min_{x\in\mathbb{R}^p}\left\{H(x) := \frac{1}{n}\sum_{i=1}^{n}\ell(y_i(a_i^T x + b_i))\right\}, \tag{62}$$

for a given loss function ℓ. If ℓ is nonnegative, then instead of solving (62), we can solve $\min_x |H(x)|$. If we have q different losses ℓ_j for $j = 1, \cdots, q$ and we want to solve q problems of the form (62) for different losses simultaneously, then we can formulate such a problem into (60) to have $\min_x \|\mathbf{H}(x)\|$, where $\mathbf{H}(x) := (H_1(x), H_2(x), \cdots, H_q(x))^{\top}$. Since we use different losses, under the formulation (60), we can view it as a binary classification task with an averaging loss.

Table 1. Hyper-parameter configurations for the two algorithms on all datasets when using the $\|\cdot\|_2$ loss.

Algorithm	w8a			ijcnn1			covtype			url_combined		
	$\hat{b}_t$	b_t	Inner Iterations	$\hat{b}_t$	b_t	Inner Iterations	$\hat{b}_t$	b_t	Inner Iterations	$\hat{b}_t$	b_t	Inner Iterations
SGN	256	512		512	1,024		1,024	4,096		20,000	50,000	
SGN2	64	128	2,000	128	256	1,000	256	512	2000	5,000	10,000	5,000
	a9a			rcv1_train.binary			real-sim			skin_nonskin		
	$\hat{b}_t$	b_t	Inner Iterations	$\hat{b}_t$	b_t	Inner Iterations	$\hat{b}_t$	b_t	Inner Iterations	$\hat{b}_t$	b_t	Inner Iterations
SGN	512	1,024		512	1,024		1,024	4,096		512	1024	
SGN2	64	128	2000	128	256	1,000	256	512	2,000	128	256	5,000

Datasets. We test three algorithms: GN, SGN, and SGN2 on four real datasets: `w8a` $(\boldsymbol{n = 49,749; p = 300})$, `ijcnn1` $(\boldsymbol{n = 91,701; p = 22})$, `covtype` $(\boldsymbol{n = 581,012; p = 54})$, and `url_combined` $(\boldsymbol{n = 2,396,130; p = 3,231,961})$ from LIBSVM.

Table 2. Hyper-parameter configurations for the four algorithms on 4 datasets when using the Huber loss.

Algorithm	w8a			ijcnn1			covtype			url_combined		
	$\hat{b}_t$	b_t	Inner Iterations	$\hat{b}_t$	b_t	Inner Iterations	$\hat{b}_t$	b_t	Inner Iterations	$\hat{b}_t$	b_t	Inner Iterations
SGN	256	512		512	1,024		512	1,024		20,000	50,000	
SCGD	256	512		512	1,024		512	1,024		20,000	50,000	
SGN2	64	128	5,000	128	256	2,000	128	256	5,000	5,000	10,000	5,000
N-SPIDER	64	128	5,000	128	256	2,000	128	256	5,000	5,000	10,000	5,000
	a9a			rcv1_train.binary			real-sim			news20.binary		
	$\hat{b}_t$	b_t	Inner Iterations	$\hat{b}_t$	b_t	Inner Iterations	$\hat{b}_t$	b_t	Inner Iterations	$\hat{b}_t$	b_t	Inner Iterations
SGN	128	256		128	512		256	512		128	512	
SCGD	1,024	2,048		128	512		256	512		128	512	
SGN2	64	128	2,000	64	128	5,000	64	128	5,000	64	128	5,000
N-SPIDER	64	128	2,000	64	128	5,000	64	128	5,000	64	128	5,000

Parameter configuration. We can easily check that F defined by (61) satisfies Assumption 1.1 and Assumption 2. However, we do not accurately estimate the Lipschitz constant of F' since it depends on the dataset. We were instead experimenting with different choices of the parameter M and ρ, and eventually fix $\rho := 1$ and $M := 1$ for our tests. We also choose the mini-batch sizes for both $\widetilde{F}$ and $\widetilde{J}$ in SGN and SGN2 by sweeping over the set of $\{64, 128, 256, 512, 1024, 2048, 4096, 8192\}$ to estimate the best ones. Table 1 presents the chosen parameters for the instance when $\phi = \|\cdot\|_2$.

In the case of smooth ϕ, i.e., using Huber loss, we add two competitors: N-SPIDER (Yang et al., 2019, Algorithm 3) and SCGD Wang et al. (2017a, Algorithm 1). The learning rates of N-SPIDER and SCGD are tuned from a set of different values: $\{0.01, 0.05, 0.1, 0.5, 1, 2\}$. Eventually we obtain $\eta := 1.0$ and set $\varepsilon := 10^{-1}$ for N-SPIDER, see (Yang et al., 2019, Algorithm 3). For SCGD, we use $\beta_k := 1$ and $\alpha_k := 1$, see Wang et al. (2017a, Algorithm 1). The mini-batch sizes of these algorithm are chosen using similar search as in the previous case. Table 2 reveals the parameter configuration of the algorithms when using the Huber loss.

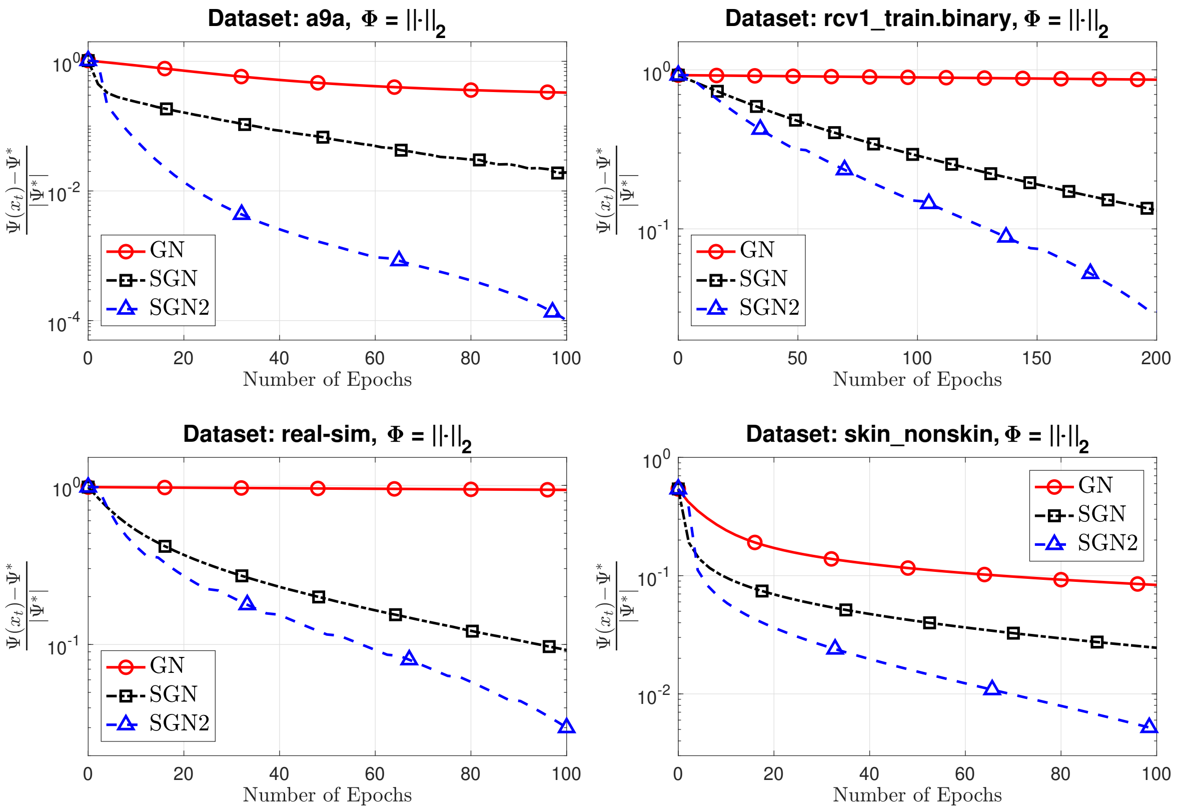

Figure 6. The performance of three algorithms on additional real datasets when $\phi(\cdot) = \|\cdot\|_2$.

Additional Experiments. When $\phi(\cdot) = \|\cdot\|_2$, we also run these algorithms on other classification datasets from LIBSVM: a9a ($\boldsymbol{n = 32,561; p = 123}$), rcv1_train.binary ($\boldsymbol{n = 20,242; p = 47,236}$), real-sim ($\boldsymbol{n = 72,309; p = 20,958}$), and skin_nonskin ($\boldsymbol{n = 245,057; p = 3}$). We set $M := 1$ and $\rho := 1$ for three datasets. Other parameters are obtained via grid search and the results are shown in Table 1. The performance of three algorithms on these datasets are presented in Figure 6.

SGN2 appears to be the best among the 3 algorithms while SGN is much better than the baseline GN. SGN appears to have advantage in the early stage but SGN2 makes better progress later on.

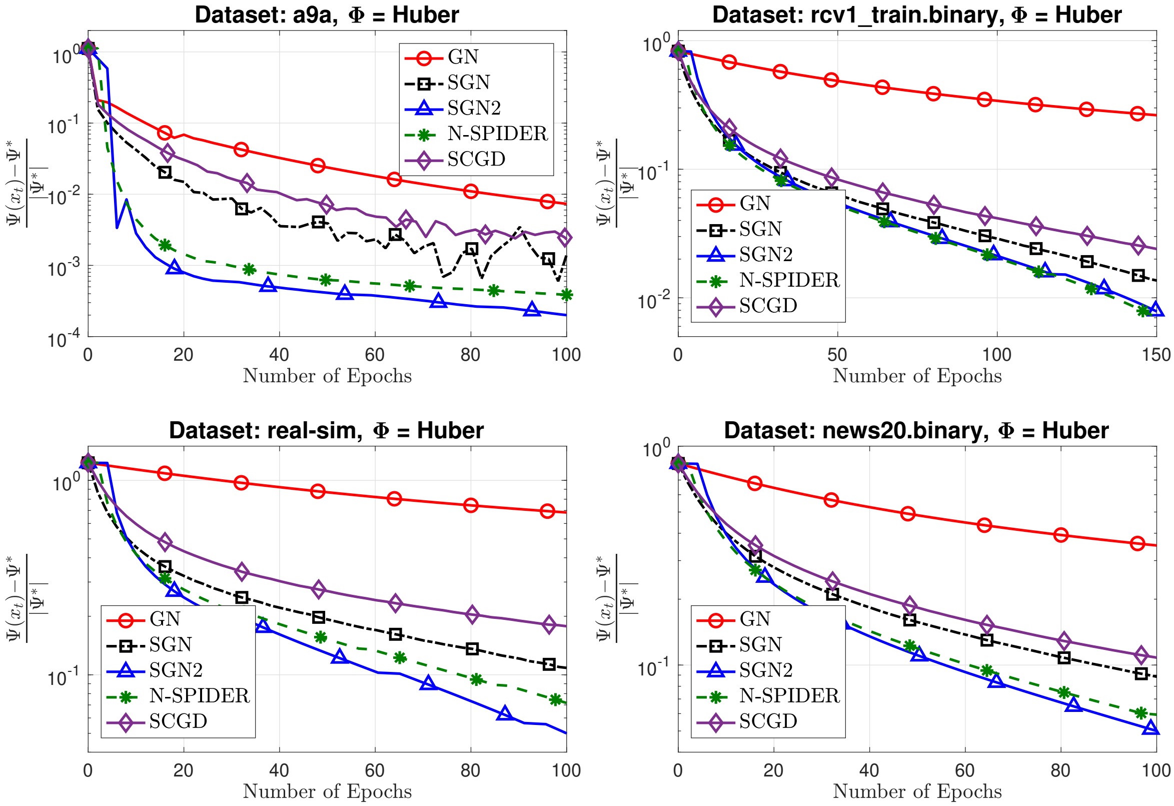

Figure 7. The performance of three algorithms on additional real datasets when using Huber loss.

In addition, we also run 5 algorithms on these datasets in the smooth case when using the Huber loss. We still tune the parameters for these algorithms and obtain the learning rate of 1.0 for both N-SPIDER and SCGD. We again use $\varepsilon = 10^{-1}$ for N-SPIDER. More details about other parameters selection are presented in Table 2 and the performance of these algorithms are shown in Figure 7.

From Figure 7, SGN2 performs better than other algorithms in most cases while N-SPIDER is better than SGN and somewhat comparable with SGN2 in the rcv1_train.binary and news20.binary datasets. SGN and SCGD appear to have similar behavior, but SGN is slightly better than SCGD in these datasets.

F.2. Optimization Involving Expectation Constraints

We consider an optimization problem involving expectation constraints as described in (34). As mentioned, this problem has various applications in different fields, including optimization with conditional value at risk (CVaR) constraints and metric learning, see, e.g., Lan & Zhou (2016) for detailed discussion.

Instead of solving the constrained setting (34), we consider its exact penalty formulation (35):

$$\min_{x\in\mathbb{R}^p}\left\{\Psi(x) := g(x) + \phi(\mathbb{E}_{\xi}\left[\mathbf{F}(x,\xi)\right])\right\}, \tag{35}$$

where $\phi(u) := \rho\sum_{i=1}^{q}[u_i]_+$ with $[u]_+ := \max\{0, u\}$ is a penalty function, and $\rho > 0$ is a given penalty parameter. It is well-known that under mild conditions and ρ sufficiently large (e.g., $\rho > \|y^{\star}\|^{*}$, the dual norm of the optimal Lagrange multiplier $y^{\star}$), if $x^{\star}$ is a stationary point of (35) and it is feasible to (34), then it is also a stationary point of (34).

As a concrete instance of (34), we solve the following asset allocation problem studied in Rockafellar & Uryasev (2000); Lan & Zhou (2016):

$$\begin{cases} \min\limits_{z\in\mathbb{R}^p, \tau\in[\underline{\tau},\bar{\tau}]} & -c^{\top}z \\ \text{s.t} & \tau + \frac{1}{\beta n}\sum_{i=1}^{n}[-\xi_i^{\top}z - \tau]_{+} \leq 0, \\ & z \in \Delta_p := \left\{\hat{z}\in\mathbb{R}^p_{+} \mid \sum_{i=1}^{p}\hat{z}_i = 1\right\}. \end{cases} \tag{63}$$

Here, Δ_p denotes the standard simplex in $\mathbb{R}^p$, and $[\underline{\tau}, \bar{\tau}]$ is a given range of τ. The exact penalty formulation of (63) is given by (36):

$$\min_{z\in\Delta^p, \tau\in[\underline{\tau},\bar{\tau}]} \left\{ -c^{\top}z + \phi\left(\tau + \frac{1}{\beta n}\sum_{i=1}^{n}[-\xi_i^{\top}z - \tau]_{+}\right)\right\}, \tag{36}$$

where $\phi(u) := \rho[u]_{+}$ with given $\rho > 0$. However, since $[-\xi_i^{\top}z - \tau]_{+}$ is nonsmooth, we smooth it by $\sqrt{(\xi_i^{\top}z + \tau)^2 + \gamma^2} - \gamma - \xi_i^{\top}z - \tau$ for sufficiently small value of $\gamma > 0$. Hence, (36) can be approximated by

$$\min_{z\in\Delta_p, \tau\in[\underline{\tau},\bar{\tau}]} \left\{ -c^{\top}z + \phi\left(\tau + \frac{1}{\beta n}\sum_{i=1}^{n}\left[\sqrt{(\xi_i^{\top}z + \tau)^2 + \gamma^2} - \gamma - \xi_i^{\top}z - \tau\right]\right)\right\}. \tag{64}$$

If we introduce $x := (z, \tau)$, $\mathbf{F}(x, \xi) := \tau + \frac{1}{2\beta}\left(\sqrt{(\xi_i^{\top}z + \tau)^2 + \gamma^2} - \gamma - \xi_i^{\top}z - \tau\right)$ for $i = 1, \cdots, n$, and $g(x) = -c^{\top}z + \delta_{\Delta_p\times[\underline{\tau},\bar{\tau}]}(x)$, where $\delta_{\mathcal{X}}$ is the indicator of $\mathcal{X}$, then we can reformulate (64) into (3). It is obvious to check that $\mathbf{F}(\cdot, \xi)$ is Lipschitz continuous with $M_i := 1 + \frac{\|\xi_i\|+1}{\beta\gamma}$ and its gradient $\mathbf{F}'(\cdot, \zeta)$ is also Lipschitz continuous with $L_i := \frac{\|\xi_i\|^2}{2\beta\gamma}$. Hence, Assumptions 1.1 and 4.1 hold.

Datasets. We consider both synthetic and US stock datasets. For the synthetic datasets, we follow the procedures from Lan et al. (2012) to generate the data with $n = 10^5$ and $p \in \{300, 500, 700\}$. We obtain real datasets of US stock prices for 889, 865, and 500 types of stocks as described, e.g., Sun & Tran-Dinh (2019). Then, we apply a bootstrap strategy to resample in order to obtain three corresponding new datasets of sizes $n = 10^5$.

Table 3. Hyper-parameter configuration of the two algorithms on 6 datasets in the asset allocation example.

Algorithm	Synthetic: p = 300			Synthetic: p = 500			Synthetic: p = 700		
	$\hat{b}_t$	b_t	Inner Iterations	$\hat{b}_t$	b_t	Inner Iterations	$\hat{b}_t$	b_t	Inner Iterations
SGN	1,024	2,048		1,024	2,048		1,024	2,048	
SGN2	128	256	5,000	128	256	2,000	256	512	2,000
Algorithm	US Stock 1: p = 889			US Stock 1: p = 865			US Stock 1: p = 500		
	$\hat{b}_t$	b_t	Inner Iterations	$\hat{b}_t$	b_t	Inner Iterations	$\hat{b}_t$	b_t	Inner Iterations
SGN	512	1,024		512	1,024		512	1,024	
SGN2	128	256	5,000	128	256	5,000	128	256	5,000

Parameter selection. We fix the smoothness parameter $\gamma := 10^{-3}$ and choose the range $[\underline{\tau}, \bar{\tau}]$ to be $[0, 1]$. The parameter $\beta := 0.1$ as discussed in Lan & Zhou (2016). Note that we do not use the theoretical values for M as in our theory since that value is obtained in the worst-case. We were instead experimenting different values for the penalty parameter ρ and M, and eventually get $\rho := 5$ and $M := 5$ as default values for this example.

Experiment setup. We implement our algorithms: SGN and SGN2, and also a baseline variant, the deterministic GN scheme (i.e., we exactly evaluate F and its Jacobian using the full batches) as in the first example. Similar to the first example, we sweep over the same set of possible mini-batch sizes, and the chosen parameters are reported in Table 3.

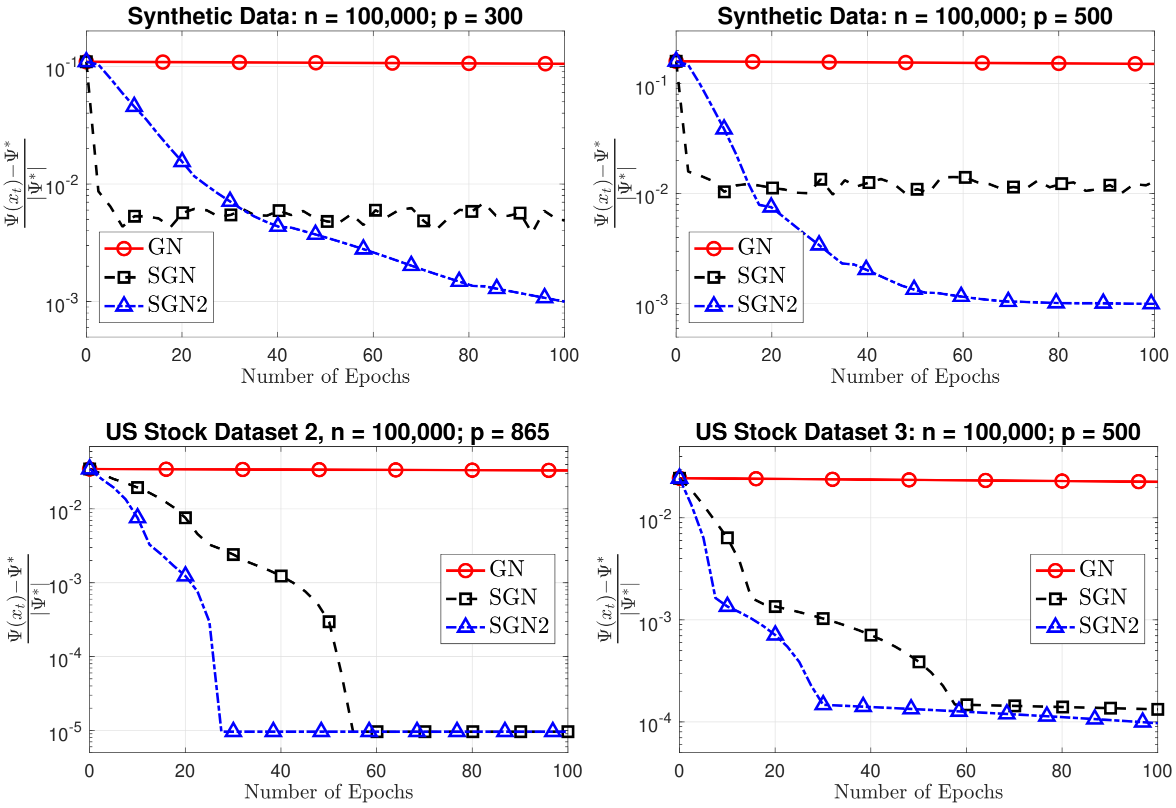

Figure 8. The performance of the three algorithms on two synthetic and two real datasets.

Additional experiments. We run three algorithms: GN, SGN, and SGN2 with 3 synthetic datasets, where the first one was reported in Figure 2 of the main text. We also use two other US Stock datasets and the performance of three algorithms on these synthetic and real datasets are revealed in Figure 8.

Clearly, SGN2 is the best, while SGN still outperforms GN in these two datasets. We believe that this experiment confirms our theoretical results presented in the main text.