Supplementary Material

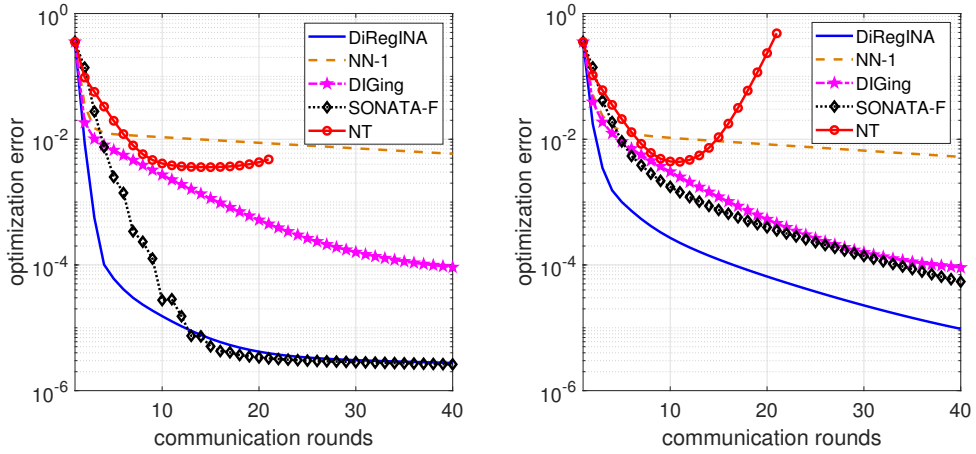
This supplementary material is organized as follows. Sec. A provides additional numerical experiments, complementing those in Sec. 5 of the main paper. In Sec. C, we establish asymptotic convergence of DiRegINA and prove some intermediate results that are instrumental for our rate analysis. Sec. D-G are devoted to prove Sec. 4 of the paper, namely: Theorem 7 is proved in Sec. D; Theorem 9 and Corollary 11 are proved in Sec. E; and finally, Theorem 12 is proved in Sec. F.

Furthermore, there are some convergence results stated in Table 1 that could not be stated in the paper because of space limit; they are reported here in the following sections: i) the case of quadratic functions f_i in the setting of Theorem 9 is stated in Theorem 18 in Sec. E.4 while the case of quadratic f_i's in the setting of Theorem 12 is stated in Theorem 19, Sec. G.

A. Additional Numerical Experiments

Convex (non-strongly convex) objective

We consider a (non-strongly) convex instance of the regression problem. Specifically, we have: $f_i(x) = (1/2n)\,\|A_i x - b_i\|^2$ and $\mathcal{K} = \mathbb{R}^d$, where A_i and b_i are determined by the scaled LIBSVM dataset `space-ga` ($N = 3107$, $d = 6$, and $\beta = 0.6353$). The network is simulated as the Erdős-Rényi network model, with $m = 30$ and two connectivity values, $\rho = 0.3843$ and $\rho = 0.8032$. We compared DiRegINA with the algorithms described in Sec. 4, namely: NN-1, NT, DIGing and SONATA-F. Note that NN-1 and NT are not guaranteed to converge when applied to convex (non-strongly convex) functions. The tuning of the algorithm is the same as the one described in Sec. 5.1. In Fig. 4, we plot the optimization error versus the communication rounds achieved by the aforementioned algorithms in the two network settings, $\rho = 0.3843$ and $\rho = 0.8032$. As already observed for the other simulated problems (cf. Sec. 5.1), SONATA-F shows similar performance of DiRegINA when running on well-connected networks while its performance deteriorates in poorly connected network. NT seems to be non-convergent while NN1 and DIGing converge, yet slow, to acceptable accuracy.



(a) (b)

Figure 4. Distributed ridge regression on `space-ga` dataset and Erdős-Rényi graph with (a) $\rho = 0.3843$ (b) $\rho = 0.8032$.

$O(1/\sqrt{mn})$-regularized logistic regression

We train logistic regression models, regularized by an additive ℓ_2-norm (with coefficient $\lambda > 0$). The problem is an instance of (P), with each $f_i(x) = -(1/n)\sum_{j=1}^{n}[\xi_i^{(j)} \ln(z_i^{(j)}) + (1 - \xi_i^{(j)}) \ln(1 - z_i^{(j)})] + (\lambda/2)||x||^2$ and $\mathcal{K} = \mathbb{R}^d$, where $z_i^{(j)} \triangleq 1/(1 + e^{-\langle a_i^{(j)}, x\rangle})$ and binary class labels $\xi_i^{(j)} \in \{0, 1\}$ and vectors $a_i^{(j)}$, $i = 1, \dots m$ and $j = 1, \dots, n$ are determined by the data set. We considered the LIBSVM `a4a` ($N = 4,781$, $d = 123$) and we set $\lambda = 1/\sqrt{mn}$. The Network is simulated according to the Erdős-Rényi model with $m = 30$ and connectivity $\rho = 0.3372$ and $\rho = 0.7387$.

We compare DiRegINA , NN-1, DIGing, SONATA-F and NT, all initialized from the same random point. The free parameters of the algorithms are tuned manually; the best practical performance are observed with the following tuning: DiRegINA is tuned as described in Sec. 5.2, i.e., $\tau = 1$, $M = 1e-3$, and $K = 1$; NN-1, $\alpha = 1e-3$ and $\epsilon = 1$; DIGing, stepsize equal to 1; SONATA-F, $\tau = 0.1$; NT, $\epsilon = 0.2$ and $\alpha = 0.05$.

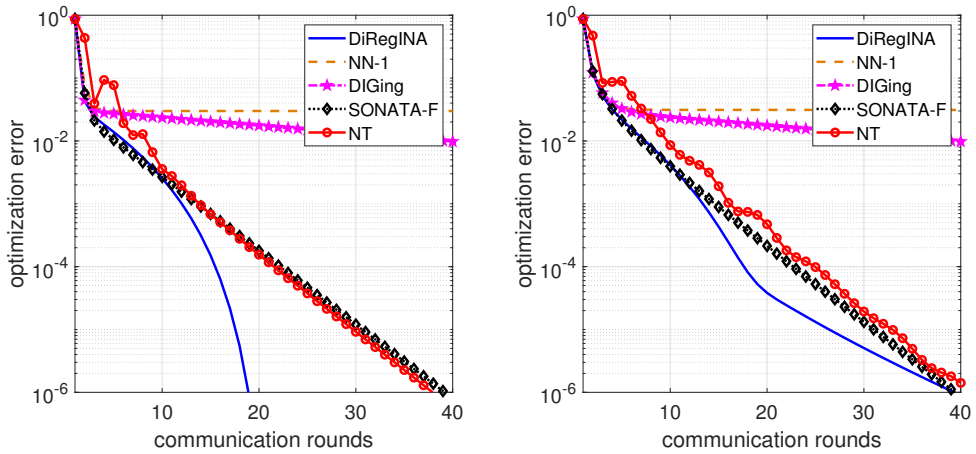


Figure 5. Distributed logistic regression on `a4a` dataset and Erdős-Rényi graph with (a) $\rho = 0.3372$ (b) $\rho = 0.7387$.

In Fig. 4, we plot the optimization error versus the communication rounds achieved by the aforementioned algorithms in two network settings corresponding to $\rho = 0.3372$ and $\rho = 0.7387$. In both settings (panels (a)-(b)), NN-1 and DIGing still exhibits slow convergence, with a slight advantage of DIGing over NN-1. DiRegINA , NT and SONATA-F, perform similarly, with DiRegINA showing some improvements when the network is better connected [panel (a)].

B. Notations and Preliminary Results

We begin introducing some notation which will be used in all the proofs, along with some preliminary results.

Define

$$\delta_i^\nu \triangleq s_i^\nu - \nabla F(x_i^\nu) \quad \text{and} \quad B_i^\nu \triangleq \nabla^2 f_i(x_i^\nu) - \nabla^2 F(x_i^\nu), \tag{17}$$

The local surrogate function $\widetilde{F}_i(y; x_i^\nu)$ in (7a) can be rewritten as

$$\widetilde{F}_i(y; x_i^\nu) \triangleq F(x_i^\nu) + \langle \nabla F(x_i^\nu) + \delta_i^\nu, y - x_i^\nu \rangle + \frac{1}{2} \left\langle \left[\nabla^2 F(x_i^\nu) + B_i^\nu + \tau_i I\right](y - x_i^\nu), y - x_i^\nu \right\rangle + \frac{M_i}{6} \|y - x_i^\nu\|^3 . \tag{18}$$

Let us recall the following basic result, which is a consequence of Assumption 3.

Lemma 1 (Nesterov (2018, Lemma 1.2.4)). *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a twice-differentiable function satisfying Assumption 3. Then, for all $x, y \in \mathbb{R}^d$,*

$$\left| F(y) - F(x) - \langle \nabla F(x), y - x \rangle - \frac{1}{2} \left\langle \nabla^2 F(x)(y - x), y - x \right\rangle \right| \leq \frac{L}{6} \|y - x\|^3 . \tag{19}$$

$$\left\| \nabla F(y) - \nabla F(x) - \nabla^2 F(x)(y - x) \right\| \leq \frac{L}{2} \|y - x\|^2 . \tag{20}$$

Setting $x = x_i^\nu$ in (19) implies

$$F(x_i^\nu) + \langle \nabla F(x_i^\nu), y - x_i^\nu \rangle + \frac{1}{2} \left\langle \nabla^2 F(x_i^\nu)(y - x_i^\nu), y - x_i^\nu \right\rangle \leq F(y) + \frac{L}{6} \|y - x_i^\nu\|^3 , \quad \forall y \in \mathbb{R}^d,$$

which, together with (18), gives the following upper bound for the surrogate function $\widetilde{F}_i$ defined in (18):

$$\widetilde{F}_i(y; x_i^\nu) \leq F(y) + \frac{1}{2} \|y - x_i^\nu\|^2_{(\beta + \tau_i)I} + \frac{M_i + L}{6} \|y - x_i^\nu\|^3 + \langle \delta_i^\nu, y - x_i^\nu \rangle , \quad \forall y \in \mathbb{R}^d, \tag{21}$$

where for a positive semidefinite matrix A, $\|x\|_A^2 \triangleq \langle Ax, x\rangle$. We also denote

$$\Delta x_i^\nu \triangleq x_i^{\nu+} - x_i^\nu, \quad \delta^\nu \triangleq (\delta_i^\nu)_{i=1}^m, \quad J \triangleq 11^\top/m, \tag{22}$$

where we remind that $x_i^{\nu+}$ is obtained by the minimization of the local surrogate function $\widetilde{F}_i(y; x_i^\nu)$. The rest of the symbols and notations are as defined in the main manuscript.

C. Asymptotic convergence of DiRegINA

In this section we prove the following theorem stating asymptotic convergence of DiRegINA .

Theorem 13. *Let Assumptions 1 and 3-5 hold, $M_i \geq L$ and $\tau_i = 2\beta$ for all $i = 1, \ldots, m$. If a reference matrix $\overline{W}$ satisfying Assumption 6 is used in steps (7b)-(7c), with $\rho \triangleq \lambda_{\max}(\overline{W} - J) < 1$ and $K = \widetilde{\mathcal{O}}(1/\sqrt{1-\rho})$ (explicit condition is provided in eq. (41)), then $p^\nu \to 0$ and $||x_i^\nu - x_j^\nu|| \to 0$, as $\nu \to \infty$ for all $i, j = 1, \ldots, m$.*

We prove the theorem in three main steps:

Step 1 (Sec. C.1): Deriving optimization bounds on the per-iteration decrease of p^ν;

Step 2 (Sec. C.2): Bounding the gradient tracking error δ^ν, which in turn affects the per-iteration decrease of p^ν;

Step 3 (Sec. C.3): Constructing a proper Lyapunov function based on the error terms in the previous two steps, whose dynamics imply asymptotic convergence of DiRegINA .

To simplify the derivations, we study the case of strongly convex or nonstronlgy convex F together, by setting $\mu = 0$ in the latter case.

C.1. Optimization error bounds

In this subsection we establish an upper bound for $p^{\nu+1} - p^\nu$ [cf. (32)]. We begin with two technical intermediate results–Lemma 2 and Lemma 3.

Lemma 2. *Under Assumption 1, there holds*

$$\widetilde{F}_i(x_i^{\nu+}; x_i^\nu) \leq \widetilde{F}_i(x_i^\nu; x_i^\nu) - \frac{M_i}{3}\|\Delta x_i^\nu\|^3 - \frac{\mu_i + \tau_i}{2}\|\Delta x_i^\nu\|^2. \tag{23}$$

Proof. By the optimality of $x_i^{\nu+}$ in (18), we infer

$$\left\langle s_i^\nu + \left[\nabla^2 f_i(x_i^\nu) + \tau_i I\right]\Delta x_i^\nu, \Delta x_i^\nu \right\rangle \leq -\frac{M_i}{2}\|\Delta x_i^\nu\|^3. \tag{24}$$

Since $\widetilde{F}_i(x_i^\nu; x_i^\nu) = F(x_i^\nu)$, we have

$$\begin{aligned}
&\widetilde{F}_i(x_i^{\nu+}; x_i^\nu) - \widetilde{F}_i(x_i^\nu; x_i^\nu) \\
&\overset{(18)}{=} \left\langle s_i^\nu, x_i^{\nu+} - x_i^\nu \right\rangle + \frac{1}{2}\left\langle \left[\nabla^2 f_i(x_i^\nu) + \tau_i I\right]\Delta x_i^\nu, \Delta x_i^\nu \right\rangle + \frac{M_i}{6}\left\|x_i^{\nu+} - x_i^\nu\right\|^3 \\
&\overset{(24)}{\leq} -\frac{1}{2}\left\langle \left[\nabla^2 f_i(x_i^\nu) + \tau_i I\right]\Delta x_i^\nu, \Delta x_i^\nu \right\rangle - \frac{M_i}{3}\|\Delta x_i^\nu\|^3 \\
&\leq -\frac{M_i}{3}\left\|x_i^{\nu+} - x_i^\nu\right\|^3 - \frac{\mu_i + \tau_i}{2}\left\|x_i^{\nu+} - x_i^\nu\right\|^2.
\end{aligned}$$

□

Lemma 3. *Let Assumptions 1 and 3-4 hold. Then, any arbitrary $\epsilon > 0$, we have*

$$F(x_i^{\nu+}) - \widetilde{F}_i(x_i^{\nu+}; x_i^\nu) \leq -\frac{M_i - L}{6}||\Delta x_i^\nu||^3 - \frac{\tau_i - \beta - \epsilon}{2}||\Delta x_i^\nu||^2 + \frac{1}{2\epsilon}\|\delta_i^\nu\|^2. \tag{25}$$

Proof. Taylor's theorem applied to functions $\widetilde{F}_i(\cdot; x_i^\nu)$ and $F(\cdot)$ around x_i^ν yields

$$F(x_i^{\nu+}) = F(x_i^\nu) + \langle \nabla F(x_i^\nu), \Delta x_i^\nu \rangle + \Delta x_i^{\nu\top} H_i^\nu \Delta x_i^\nu, \tag{26a}$$

$$\widetilde{F}_i(x_i^{\nu+}; x_i^\nu) = \widetilde{F}_i(x_i^\nu; x_i^\nu) + \left\langle \nabla \widetilde{F}_i(x_i^\nu; x_i^\nu), \Delta x_i^\nu \right\rangle + \Delta x_i^{\nu\top} \widetilde{H}_i^\nu \Delta x_i^\nu, \tag{26b}$$

where

$$H_i^\nu = \int_0^1 (1-\theta)\nabla^2 F(\theta x_i^{\nu+} + (1-\theta)x_i^\nu) d\theta,$$
$$\widetilde{H}_i^\nu = \int_0^1 (1-\theta)\nabla^2 \widetilde{F}_i(\theta x_i^{\nu+} + (1-\theta)x_i^\nu; x_i^\nu) d\theta.$$

Since $\widetilde{F}_i(x_i^\nu; x_i^\nu) = F(x_i^\nu)$ and $\nabla \widetilde{F}_i(x_i^\nu; x_i^\nu) = \nabla F(x_i^\nu) + \delta_i^\nu$, subtracting (26a)-(26b) gives

$$F(x_i^{\nu+}) - \widetilde{F}_i(x_i^{\nu+}; x_i^\nu) = \left\langle \left(H_i^\nu - \widetilde{H}_i^\nu\right)\Delta x_i^\nu, \Delta x_i^\nu \right\rangle - \langle \delta_i^\nu, \Delta x_i^\nu \rangle . \tag{27}$$

Now let us simplify (27). Note that the hessian of $\widetilde{F}_i(\cdot; x_i^\nu)$ is

$$\nabla^2 \widetilde{F}_i(x_i; x_i^\nu) = \nabla^2 F(x_i^\nu) + B_i^\nu + \tau_i I + M_i G(x_i; x_i^\nu), \tag{28}$$

where

$$G(x_i; x_i^\nu) \triangleq \frac{1}{2}\left(\|x_i - x_i^\nu\| I + \frac{(x_i - x_i^\nu)(x_i - x_i^\nu)^\top}{\|x_i - x_i^\nu\|} \right).$$

Hence,

$$\begin{aligned}
&H_i^\nu - \widetilde{H}_i^\nu \\
&= \int_0^1 (1-\theta)\nabla^2 F\left(\theta x_i^{\nu+} + (1-\theta)x_i^\nu\right) d\theta - \int_0^1 (1-\theta)\nabla^2 \widetilde{F}_i\left(\theta x_i^{\nu+} + (1-\theta)x_i^\nu; x_i^\nu\right) d\theta \\
&\overset{(28)}{=} \int_0^1 (1-\theta)\nabla^2 F\left(\theta x_i^{\nu+} + (1-\theta)x_i^\nu\right) d\theta - \int_0^1 (1-\theta)\left[\nabla^2 F(x_i^\nu) + B_i^\nu\right] d\theta - \int_0^1 (1-\theta)\tau_i I d\theta \\
&\quad - \int_0^1 (1-\theta) M_i \theta G(x_i^{\nu+}; x_i^\nu) d\theta \\
&= \int_0^1 (1-\theta)\left(\nabla^2 F\left(\theta x_i^{\nu+} + (1-\theta)x_i^\nu\right) - \nabla^2 F(x_i^\nu)\right) d\theta \\
&\quad - \int_0^1 (1-\theta) B_i^\nu d\theta - \int_0^1 (1-\theta)\tau_i I d\theta - \int_0^1 (1-\theta) M_i \theta G(x_i^{\nu+}; x_i^\nu) d\theta \\
&\overset{(a)}{\preceq} \int_0^1 (1-\theta) L\theta ||x_i^{\nu+} - x_i^\nu|| I d\theta \\
&\quad - \int_0^1 (1-\theta) B_i^\nu d\theta - \int_0^1 (1-\theta)\tau_i I d\theta - \int_0^1 (1-\theta) M_i \theta G(x_i^{\nu+}; x_i^\nu) d\theta \\
&= -\frac{M_i}{6} G(x_i^{\nu+}; x_i^\nu) + \frac{L}{6}||x_i^{\nu+} - x_i^\nu|| I - \frac{\tau_i}{2} I - \frac{B_i^\nu}{2}
\end{aligned} \tag{29}$$

where (a) holds since $\nabla^2 F$ is L-Lipschitz continuous. Combining (27) and (29), we conclude

$$\begin{aligned}
F(x_i^{\nu+}) - \widetilde{F}_i(x_i^{\nu+}; x_i^\nu) &\le -\frac{M_i - L}{6}||\Delta x_i^\nu||^3 - \frac{\tau_i}{2}||\Delta x_i^\nu||^2 - \frac{1}{2}\langle B_i^\nu \Delta x_i^\nu, \Delta x_i^\nu \rangle - \langle \delta_i^\nu, \Delta x_i^\nu \rangle \\
&\le -\frac{M_i - L}{6}||\Delta x_i^\nu||^3 - \frac{\tau_i - \beta - \epsilon}{2}||\Delta x_i^\nu||^2 + \frac{1}{2\epsilon}\|\delta_i^\nu\|^2,
\end{aligned}$$

for arbitrary $\epsilon > 0$, where the last inequality is due to the Cauchy-Schwarz inequality and $|\langle B_i^\nu \Delta x_i^\nu, \Delta x_i^\nu \rangle| \le \beta ||\Delta x_i^\nu||^2$, which is a consequence of (17) and Assumption 4. □

We are now in a position to prove the main result of this subsection.

Combining (23) in Lemma 3 with (25) in Lemma 2, and using $\widetilde{F}_i(x_i^\nu; x_i^\nu) = F(x_i^\nu)$, yields

$$F(x_i^{\nu+}) - F(x_i^\nu) \leq -\left(\frac{M_i}{2} - \frac{L}{6}\right)||\Delta x_i^\nu||^3 - \left(\frac{\mu_i}{2} + \tau_i - \frac{\beta+\epsilon}{2}\right)||\Delta x_i^\nu||^2 + \frac{1}{2\epsilon}\left\|\delta_i^\nu\right\|^2.$$

Since under either Assumption 1 or Assumption 2 combined with Assumption 4 it holds that $\mu_i \geq \max\{0, \mu - \beta\}$, we obtain

$$F(x_i^{\nu+}) - F(x_i^\nu) \leq -\left(\frac{M_i}{2} - \frac{L}{6}\right)||\Delta x_i^\nu||^3 - \left(\frac{\max(0,\mu-\beta)}{2} + \tau_i - \frac{\beta+\epsilon}{2}\right)||\Delta x_i^\nu||^2 + \frac{1}{2\epsilon}\left\|\delta_i^\nu\right\|^2. \tag{30}$$

Denoting $p^{\nu+} \triangleq (1/m)\sum_{i=1}^m \left\{F(x_i^{\nu+}) - F(\widehat{x})\right\}$, we derive a simple relation with $p^{\nu+1}$:

$$\begin{aligned} p^{\nu+1} + F(\widehat{x}) &= \frac{1}{m}\sum_{i=1}^m F\Big(x_i^{\nu+1}\Big) \overset{(7b)}{=} \frac{1}{m}\sum_{i=1}^m F\Big(\sum_{j=1}^m (W_K)_{i,j}\, x_j^{\nu+}\Big) \\ &\overset{(a)}{\leq} \frac{1}{m}\sum_{i,j=1}^m (W_K)_{i,j} F\Big(x_j^{\nu+}\Big) \overset{(b)}{=} \frac{1}{m}\sum_{j=1}^m F\Big(x_j^{\nu+}\Big) = p^{\nu+} + F(\widehat{x}), \end{aligned} \tag{31}$$

where (a) is due to convexity of F (cf. Assumptions 1 and 2) and $\sum_{j=1}^m (W_K)_{ij} = 1$ (cf. Assumption 6); and in (b) we used $\sum_{i=1}^m (W_K)_{ij} = 1$ (cf. Assumption 6). Summing (30) over i while setting $\epsilon = \beta$, $\tau_i = 2\beta$ and $M_i \geq L/3$ (recall that it is assumed $M_i \geq L$), gives the desired per-iteration decrease of p^ν when $\|\delta^\nu\|$ is sufficiently small:

$$p^{\nu+1} - p^\nu \overset{(31)}{\leq} p^{\nu+} - p^\nu \leq -\frac{\max(\mu,\beta)}{2}\cdot\frac{1}{m}\sum_{i=1}^m \|\Delta x_i^\nu\|^2 + \frac{1}{2m\beta}\|\delta^\nu\|^2. \tag{32}$$

C.2. Network error bounds

The goal of this subsection is to prove an upper bound for $\|\delta^\nu\|$ in terms of the number of communication steps K, implying that this error can be made sufficiently small by choosing sufficiently large K. For notation simplicity and without loss of generality, we assume $d = 1$; the case $d > 1$ follows trivially.

Recall that $x^\nu \triangleq (x_i^\nu)_{i=1}^m$, $s^\nu \triangleq (s_i^\nu)_{i=1}^m$, $J \triangleq (1/m)1_m 1_m^\top$, and

$$x_\perp^\nu \triangleq (I-J)x^\nu = x^\nu - 1_m\frac{1_m^\top x^\nu}{m}, \quad s_\perp^\nu \triangleq (I-J)s^\nu = s^\nu - 1_m\frac{1_m^\top s^\nu}{m}, \quad \Delta x^\nu \triangleq (\Delta x_i^\nu)_{i=1}^m.$$

Note that the vectors $x_\perp^\nu$ and $s_\perp^\nu$ are the consensus and gradient-tracking errors; when $\|x_\perp^\nu\| = \|s_\perp^\nu\| = 0$, we have $x_i^\nu = x_j^\nu$ and $s_i^\nu = s_j^\nu$ for all $i, j = 1, \ldots, m$. The following holds for $x_\perp^\nu$ and $s_\perp^\nu$.

Lemma 4 (Proposition 3.5 in Sun et al. (2019))**.** *Under Assumptions 1 and 5-6, for all $\nu \geq 0$,*

$$\|x_\perp^{\nu+1}\| \leq \rho_K\|x_\perp^\nu\| + \rho_K\|\Delta x^\nu\|, \tag{33a}$$

$$\|s_\perp^{\nu+1}\| \leq \rho_K\|s_\perp^\nu\| + 2Q_{\max}\rho_K\|x_\perp^\nu\| + Q_{\max}\rho_K\|\Delta x^\nu\|, \tag{33b}$$

where $\rho_K = \lambda_{\max}(W_K - J) < 1$. Note that in case of K-rounds of communications using a reference matrix $\overline{W}$ with $\rho \triangleq \lambda_{\max}(\overline{W} - J) < 1$, we have $\rho_K = \rho^K$; if Chebyshev acceleration is employed, we have $\rho_K = \left(1 - \sqrt{1-\rho}\right)^K$.

Now let us bound δ_i^ν defined in (17). Note that by column-stochasticity of W_K and initialization rule $s_i^0 = \nabla f_i(x_i^0)$, it can be trivially concluded from (7c) that

$$1_m^\top s^\nu = \sum_{j=1}^m \nabla f_j(x_j^\nu).$$

Hence,

$$\begin{aligned}\|\delta_i^\nu\|^2 =&\left\|s_i^\nu - \frac{1_m^\top s^\nu}{m} + \frac{1}{m}\sum_{j=1}^m \nabla f_j(x_j^\nu) - \nabla F(x_i^\nu)\right\|^2\\ \overset{(a)}{\leq}& 2\left\|s_i^\nu - \frac{1_m^\top s^\nu}{m}\right\|^2 + \frac{2Q_{\max}^2}{m}\left(\sum_{j=1}^m \left\|x_i^\nu \pm \frac{1_m^\top x^\nu}{m} - x_j^\nu\right\|^2\right)\\ \leq& 2\left\|s_i^\nu - \frac{1_m^\top s^\nu}{m}\right\|^2 + \frac{4Q_{\max}^2}{m}\left(\|x_\perp^\nu\|^2 + m\left\|x_i^\nu - \frac{1_m^\top x^\nu}{m}\right\|^2\right),\end{aligned} \tag{34}$$

where (a) is due to $Q_{\max}$-Lipschitz continuity of ∇f_i. Summing (34) over i and taking the square root, gives

$$\|\delta^\nu\| \leq \tilde{\delta}^\nu \triangleq \sqrt{2}\left(\|s_\perp^\nu\| + 2Q_{\max}\|x_\perp^\nu\|\right). \tag{35}$$

It remains to bound $\tilde{\delta}^\nu$ defined above:

$$\begin{aligned}\tilde{\delta}^{\nu+1} = \sqrt{2}\left(\|s_\perp^{\nu+1}\| + 2Q_{\max}\|x_\perp^{\nu+1}\|\right) \overset{(a)}{\leq}& \rho_K\sqrt{2}\left(\|s_\perp^\nu\| + 4Q_{\max}\|x_\perp^\nu\|\right) + 3\sqrt{2}Q_{\max}\rho_K\|\Delta x^\nu\|\\ \leq& 2\rho_K\tilde{\delta}^\nu + 3\sqrt{2}Q_{\max}\rho_K\|\Delta x^\nu\|,\end{aligned}$$

where in (a) we used Lemma 4 [cf. (33a)-(33b)]. Consequently,

$$(\tilde{\delta}^{\nu+1})^2 \leq 8\rho_K^2(\tilde{\delta}^\nu)^2 + 36Q_{\max}^2\rho_K^2\|\Delta x^\nu\|^2. \tag{36}$$

Since ρ_K decreases as K increases, the latter inequality provides a leverage to make $\tilde{\delta}^{\nu+1}$ sufficiently small by choosing K sufficiently large.

C.3. Asymptotic convergence

We combine the results of the previous two subsections to finally prove Theorem 13. Combining (32) and (35), we obtain

$$p^{\nu+1} \leq\ p^\nu - \frac{\max(\beta,\mu)}{2m}\|\Delta x^\nu\|^2 + \frac{1}{2m\beta}(\tilde{\delta}^\nu)^2. \tag{37}$$

Next, we combine (36) with (37) multiplied by some weight $w > 0$ to obtain

$$wp^{\nu+1} + (\tilde{\delta}^{\nu+1})^2 \leq wp^\nu + \left(8\rho_K^2 + \frac{w}{2m\beta}\right)(\tilde{\delta}^\nu)^2 - w\left(\frac{\max(\beta,\mu)}{2m} - \frac{36}{w}Q_{\max}^2\rho_K^2\right)||\Delta x^\nu||^2. \tag{38}$$

Let $w = c_w\beta$, for some $0 < c_w \leq 1$. Then, if

$$8\rho_K^2 + \frac{w}{2m\beta} \leq c_w, \quad \frac{\max(\beta,\mu)}{4m} \geq \frac{36}{w}Q_{\max}^2\rho_K^2, \tag{39}$$

(38) becomes

$$wp^{\nu+1} + (\tilde{\delta}^{\nu+1})^2 \leq wp^\nu + c_w(\tilde{\delta}^\nu)^2 - \frac{w\max(\beta,\mu)}{4m}||\Delta x^\nu||^2. \tag{40}$$

Note that by Lemma 4, condition (39) holds if

$$K \geq \frac{1}{\sqrt{1-\rho}}\log\left(\max\left\{\frac{2\sqrt{2}}{\sqrt{c_w(1-\frac{1}{2m})}}, \frac{12\sqrt{m}Q_{\max}}{\sqrt{c_w\beta\max(\beta,\mu)}}\right\}\right). \tag{41}$$

Denoting

$$\xi^\nu \triangleq wp^\nu + (\tilde{\delta}^\nu)^2, \tag{42}$$

let us show that $\xi^\nu \to 0$ as $\nu \to \infty$, which implies that the optimization error p^ν and network error $\tilde{\delta}^\nu$ asymptotically vanish. Since $\xi^\nu \geq 0$, inequality (40) implies $\sum_{\nu=0}^{\infty} ||\Delta x^\nu||^2 < \infty$. Thus, $||\Delta x^\nu|| \to 0$; and $||\Delta x^\nu|| \leq D_1$, for some $D_1 > 0$ and all $\nu \geq 0$. Further, $\{\xi^\nu\}_\nu$ is non-increasing and $||\xi^\nu|| \leq D_2$ for some $D_2 > 0$ and all $\nu \geq 0$. Thus, $p^\nu \leq D_2/w$, which together with Assumption 1(iv) and Assumption 2, also implies $||x_i^\nu|| \leq D_3$ for some D_3, all i and $\nu \geq 0$. Using $||\Delta x^\nu|| \to 0$ and (36), if $8\rho_K^2 < 1$ (which holds under (41)), we obtain that $\tilde{\delta}^\nu \to 0$. Finally, it remains to show that $p^\nu \to 0$. Using optimality condition of $x_i^{\nu+}$ defined in (7a), we get

$$\left\langle \nabla F(x_i^\nu) + \delta_i^\nu + \left[\nabla^2 F(x_i^\nu) + B_i^\nu + \tau_i I\right] \Delta x_i^\nu + \frac{M_i}{2}||\Delta x_i^\nu||\Delta x_i^\nu, \widehat{x} - x_i^{\nu+} \right\rangle \geq 0.$$

Rearranging terms gives

$$\begin{aligned} &\left\langle \nabla F(x_i^\nu) + \nabla^2 F(x_i^\nu)\Delta x_i^\nu, \widehat{x} - x_i^{\nu+} \right\rangle \\ \geq &\left\langle \frac{M_i}{2}||\Delta x_i^\nu||\Delta x_i^\nu, x_i^{\nu+} - \widehat{x} \right\rangle + \left\langle \delta_i^\nu, x_i^{\nu+} - \widehat{x} \right\rangle + \left\langle \tilde{B}_i^\nu \Delta x_i^\nu, x_i^{\nu+} - \widehat{x} \right\rangle, \end{aligned} \tag{43}$$

where $\tilde{B}_i^\nu \triangleq B_i^\nu + \tau_i I$. By convexity of F, we can write

$$\begin{aligned} 0 \geq &F(\widehat{x}) - F(x_i^{\nu+}) \\ \geq &\left\langle \nabla F(x_i^{\nu+}), \widehat{x} - x_i^{\nu+} \right\rangle \\ = &\left\langle \nabla F(x_i^{\nu+}) - \nabla F(x_i^\nu) - \nabla^2 F(x_i^\nu)\Delta x_i^\nu, \widehat{x} - x_i^{\nu+} \right\rangle + \left\langle \nabla F(x_i^\nu) + \nabla^2 F(x_i^\nu)\Delta x_i^\nu, \widehat{x} - x_i^{\nu+} \right\rangle \\ \overset{(43)}{\geq} &\left\langle \nabla F(x_i^{\nu+}) - \nabla F(x_i^\nu) - \nabla^2 F(x_i^\nu)\Delta x_i^\nu, \widehat{x} - x_i^{\nu+} \right\rangle + \left\langle \frac{M_i}{2}||\Delta x_i^\nu||\Delta x_i^\nu, x_i^{\nu+} - \widehat{x} \right\rangle \\ &+ \left\langle \delta_i^\nu, x_i^{\nu+} - \widehat{x} \right\rangle + \left\langle \tilde{B}_i^\nu \Delta x_i^\nu, x_i^{\nu+} - \widehat{x} \right\rangle. \end{aligned} \tag{44}$$

Using Lipschitz continuity of ∇F, $||\Delta x_i^\nu|| \to 0$ and $\tilde{\delta}^\nu \to 0$ (hence $||\delta_i^\nu|| \to 0$), we conclude that the RHS of (44) asymptotically vanishes, for all $i = 1, \ldots, m$. Hence, $F(x_i^{\nu+}) - F(\widehat{x}) \to 0$, for all $i = 1, \ldots, m$. Using (31), we finally obtain $p^\nu \to 0$.

Finally, by (35) and $\tilde{\delta}^\nu \to 0$, we obtain $\|s_\perp^\nu\| \to 0$ and $\|x_\perp^\nu\| \to 0$, implying $||x_i^\nu - x_j^\nu|| \to 0$, for all $i, j = 1, \ldots, m$ as $\nu \to \infty$. This concludes the proof of Theorem 13.

Remark 14. *Note that* (36) *implies*

$$(\tilde{\delta}^\nu)^2 \leq \rho_K^2 \bar{D}_\delta, \quad \bar{D}_\delta \triangleq 8D_2 + 36Q_{\max}^2 D_1^2, \quad \forall \nu \geq 0, \tag{45}$$

since $(\tilde{\delta}^\nu)^2 \leq \xi^\nu \leq D_2$ *and* $||\Delta x^\nu|| \leq D_1$, *for all* $\nu \geq 0$.

D. Proof of Theorem 7

We first prove a detailed "region-based" complexity of DiRegINA (cf. Theorem 15, Subsec. D.1) for the prevalent scenario $0 < \beta \leq 1$ [recall that typically $\beta = \mathcal{O}(1/\sqrt{n})$]. For the sake of completeness, the case $\beta \geq 1$ is studied in Theorem 16 (cf. Subsec. D.2). Building on Theorems 15-16, we can finally prove the main result, Theorem 7 (cf. Subsec. D.3).

D.1. Complexity Analysis when $0 < \beta \leq 1$

Theorem 15 ($0 < \beta \leq 1$ and $L > 0$)**.** *Let Assumptions 1 and 3 -5 hold along with* $0 < \beta \leq 1$. *Let* $M_i = L > 0$, $\tau_i = 2\beta$, *and recall the definition of* $D > 0$ *implying* $||x_i^0 - \widehat{x}|| \leq D$, *for all* $i = 1, \ldots, m$. *W.l.o.g. assume* $D \geq 2/L$. *Pick an accuracy* $\varepsilon > 0$. *If a reference matrix* $\overline{W}$ *satisfying Assumption 6 is used in steps* (7b)-(7c), *with* $\rho \triangleq \lambda_{\max}(\overline{W} - J) < 1$ *and* $K = \widetilde{\mathcal{O}}(\log(1/\varepsilon)/\sqrt{1-\rho})$ *(the explicit expression of* K *can be found in* (63)), *then the sequence* $\{p^\nu\}$ *generated by DiRegINA satisfies the following:*

(a) if $p^\nu \geq 2LD^3$,

$$p^{\nu+1} \leq \frac{5}{6}\, p^\nu,$$

(b) if $\beta^2 \cdot (2LD^3) \leq p^\nu \leq 2LD^3$,

$$p^\nu \leq \frac{244 \cdot LD^3}{\nu^2},$$

(c) if $\varepsilon \leq p^\nu \leq \beta^2 \cdot (2LD^3)$,

$$p^\nu \leq 24^2 \cdot (LD^3)^2 \cdot \frac{\beta^2}{\varepsilon} \cdot \frac{1}{\nu^2}.$$

Proof. Recalling Lemma 3 from the proof of Theorem 13, we can write

$$F(x_i^{\nu+}) \leq \widetilde{F}_i(x_i^{\nu+}; x_i^\nu) + \frac{1}{2\epsilon} \left\|\delta_i^\nu\right\|^2, \tag{46}$$

for arbitrary $\epsilon > 0$, $M_i \geq L$, and $\tau_i \geq \beta + \epsilon$. In addition, by the upperbound approximation of $\widetilde{F}_i(\cdot; x_i^\nu)$ in (21), there holds

$$\widetilde{F}_i(y; x_i^\nu) \leq F(y) + \frac{1}{2} \left\|y - x_i^\nu\right\|^2_{(\beta+\tau_i+\epsilon)I} + \frac{M_i + L}{6} \left\|y - x_i^\nu\right\|^3 + \frac{1}{2\epsilon} \left\|\delta_i^\nu\right\|^2, \quad \forall y \in \mathcal{K}. \tag{47}$$

Let $\alpha_0 \in (0, 1]$. Set $\epsilon = \beta$ and $\tau_i = 2\beta$. By (46)-(47) and $x_i^{\nu+}$ being the minimizer of $\widetilde{F}(\cdot; x_i^\nu)$ [see (7a)], we obtain

$$\begin{aligned}
&F(x_i^{\nu+}) - F(\widehat{x}) \\
\leq &\min_{y \in \mathcal{K}} \left\{ F(y) - F(\widehat{x}) + 2\beta \left\|y - x_i^\nu\right\|^2 + \frac{M_i + L}{6} \left\|y - x_i^\nu\right\|^3 + \frac{1}{\beta} \left\|\delta_i^\nu\right\|^2 \right\} \\
\leq &\min_{\alpha \in [0, \alpha_0]} \Big\{ F(y) - F(\widehat{x}) + 2\beta \left\|y - x_i^\nu\right\|^2 + \frac{M_i + L}{6} \left\|y - x_i^\nu\right\|^3 + \frac{1}{\beta} \left\|\delta_i^\nu\right\|^2 \\
&\qquad\qquad : y = \alpha\widehat{x} + (1 - \alpha) x_i^\nu \Big\} \\
\leq &\min_{\alpha \in [0, \alpha_0]} \Big\{ (1 - \alpha) \left(F(x_i^\nu) - F(\widehat{x})\right) \\
&\qquad + 2\beta\alpha^2 \left\|\widehat{x} - x_i^\nu\right\|^2 + \frac{M_i + L}{6} \alpha^3 \left\|\widehat{x} - x_i^\nu\right\|^3 + \frac{1}{\beta} \left\|\delta_i^\nu\right\|^2 \Big\},
\end{aligned} \tag{48}$$

where the last inequality holds by the convexity of F. Note that, by definition, $||x_i^0 - \widehat{x}|| \leq D$, for all $i = 1, \ldots, m$. Assuming $||x_i^\nu - \widehat{x}|| \leq D$, for all $i = 1, \ldots, m$, we prove descent at iteration $\nu + 1$, i.e. $p^{\nu+1} < p^\nu$, unless $p^\nu = 0$. Note that by Assumption 1(iv), if $\{p^\nu\}_\nu$ is non-increasing, then $||x_i^\nu - \widehat{x}|| \leq D$ for all $\nu \geq 0$ and $i = 1, \ldots, m$. Now set $M_i = L$ in (48) and compute the mean over $i = 1, \ldots, m$, which yields

$$p^{\nu+1} \overset{(31)}{\leq} p^{\nu+} \leq \min_{\alpha \in [0, \alpha_0]} \left\{ (1 - \alpha) p^\nu + 2\beta\alpha^2 D^2 + \frac{LD^3}{3} \alpha^3 + \frac{1}{m\beta} \left\|\delta^\nu\right\|^2 \right\}. \tag{49}$$

Denote

$$C_1 \triangleq \frac{LD^3}{3}. \tag{50}$$

Since $D \geq \frac{2}{L}$, it holds $2\beta D^2 \leq 3\beta C_1$. Then, setting $\alpha_0 = \min\{1, p^\nu/(6\beta C_1)\}$ in (49) yields

$$\begin{aligned}
p^{\nu+1} &\leq \min_{\alpha \in [0, \min\{1, \frac{p^\nu}{6\beta C_1}\}]} \left\{ (1 - \alpha) p^\nu + 3\beta C_1 \alpha^2 + C_1 \alpha^3 + \frac{1}{m\beta} \left\|\delta^\nu\right\|^2 \right\} \\
&\leq \min_{\alpha \in [0, \min\{1, \frac{p^\nu}{6\beta C_1}\}]} \left\{ (1 - \alpha/2) p^\nu + C_1 \alpha^3 + \frac{1}{m\beta} \left\|\delta^\nu\right\|^2 \right\}.
\end{aligned} \tag{51}$$

Let us assess (51) over the following "regions". Denoting by α^* the minimizer of the optimization problem at the RHS of (51), we have the following:

(a) If $p^\nu \geq 6C_1$, then $\alpha^* = 1$ and

$$p^{\nu+1} \leq \frac{1}{2} p^\nu + C_1 + \frac{1}{m\beta} \left\|\delta^\nu\right\|^2 \leq \left(\frac{1}{2} + \frac{1}{6}\right) p^\nu + \frac{1}{m\beta} \left\|\delta^\nu\right\|^2, \tag{52}$$

and under the condition

$$\frac{1}{m\beta}\left\|\delta^{\nu}\right\|^{2} \leq \frac{1}{6}p^{\nu} \impliedby \frac{1}{m\beta}\left\|\delta^{\nu}\right\|^{2} \leq C_1, \tag{53}$$

(52) yields

$$p^{\nu+1} \leq \frac{5}{6}\,p^{\nu}.$$

Note that, by (45) and Lemma 4, condition (53) holds if

$$K \geq \frac{1}{\sqrt{1-\rho}} \cdot \frac{1}{2}\log\left(\frac{\bar{D}_{\delta}}{m\beta C_1}\right). \tag{54}$$

(b) If $6\beta^2 C_1 \leq p^{\nu} \leq 6C_1$, then $\alpha^{*} = \sqrt{\frac{p^{\nu}}{6C_1}}$ and

$$p^{\nu+1} \leq p^{\nu} - \frac{(p^{\nu})^{3/2}}{3\sqrt{6C_1}} + \frac{1}{m\beta}\left\|\delta^{\nu}\right\|^{2}, \tag{55}$$

and if (similar to derivation of (54))

$$K \geq \frac{1}{\sqrt{1-\rho}} \cdot \frac{1}{2}\log\left(\frac{\bar{D}_{\delta}}{m\beta^4 C_1}\right) \implies \frac{1}{m\beta}\left\|\delta^{\nu}\right\|^{2} \leq \beta^3 C_1 \implies \frac{1}{m\beta}\left\|\delta^{\nu}\right\|^{2} \leq \frac{(p^{\nu})^{3/2}}{6\sqrt{6C_1}}, \tag{56}$$

(55) implies

$$p^{\nu+1} \leq p^{\nu} - \frac{(p^{\nu})^{3/2}}{6\sqrt{6C_1}}. \tag{57}$$

Finally, since p^{ν} is non-increasing,

$$\begin{aligned}\frac{1}{\sqrt{p^{\nu+1}}} - \frac{1}{\sqrt{p^{\nu}}} =& \frac{p^{\nu} - p^{\nu+1}}{\left(\sqrt{p^{\nu}} + \sqrt{p^{\nu+1}}\right)\sqrt{p^{\nu}p^{\nu+1}}} \overset{(57)}{\geq} \frac{\frac{1}{6\sqrt{6C_1}}(p^{\nu})^{3/2}}{\left(\sqrt{p^{\nu}} + \sqrt{p^{\nu+1}}\right)\sqrt{p^{\nu}p^{\nu+1}}} \\ \geq& c_0 \triangleq \frac{1}{12}\sqrt{\frac{1}{6C_1}},\end{aligned}$$

and consequently,

$$p^{\nu} \leq \frac{1}{c_0^2\left(\nu + \frac{1}{c_0\sqrt{p^0}}\right)^2} \leq \frac{1}{c_0^2\nu^2}.$$

(c) If $\varepsilon \leq p^{\nu} \leq 6\beta^2 C_1$, then $\alpha^{*} = \frac{p^{\nu}}{6\beta C_1}$ and

$$p^{\nu+1} \leq p^{\nu} - \frac{(p^{\nu})^2}{18\beta C_1} + \frac{1}{m\beta}\left\|\delta^{\nu}\right\|^{2}, \tag{58}$$

and if (similar to derivation of (54))

$$K \geq \frac{1}{\sqrt{1-\rho}} \cdot \frac{1}{2}\log\left(\frac{36C_1\bar{D}_{\delta}}{m\varepsilon^2}\right) \implies \frac{1}{m\beta}\left\|\delta^{\nu}\right\|^{2} \leq \frac{\varepsilon^2}{36\beta C_1} \implies \frac{1}{m\beta}\left\|\delta^{\nu}\right\|^{2} \leq \frac{(p^{\nu})^2}{36\beta C_1}, \tag{59}$$

we deduce from (58)

$$p^{\nu+1} \leq p^{\nu} - \frac{(p^{\nu})^2}{36\beta C_1}. \tag{60}$$

Since p^{ν} is non-increasing,

$$\begin{aligned}\frac{1}{\sqrt{p^{\nu+1}}} - \frac{1}{\sqrt{p^{\nu}}} =& \frac{p^{\nu} - p^{\nu+1}}{\left(\sqrt{p^{\nu}} + \sqrt{p^{\nu+1}}\right)\sqrt{p^{\nu}p^{\nu+1}}} \overset{(60)}{\geq} \frac{\frac{1}{36\beta C_1}(p^{\nu})^2}{\left(\sqrt{p^{\nu}} + \sqrt{p^{\nu+1}}\right)\sqrt{p^{\nu}p^{\nu+1}}} \\ \geq& \tilde{c}_0 \triangleq \frac{\sqrt{\varepsilon}}{72\beta C_1},\end{aligned} \tag{61}$$

and consequently,

$$p^\nu \leq \frac{1}{\tilde{c}_0^2\left(\nu + \frac{1}{\tilde{c}_0\sqrt{p^0}}\right)^2} \leq \frac{1}{\tilde{c}_0^2\nu^2} = 72^2 \cdot C_1^2 \cdot \frac{\beta^2}{\varepsilon} \cdot \frac{1}{\nu^2}. \tag{62}$$

Finally, combining all the conditions (41), (54),(56), and (59), the requirement on K reads

$$K \geq \frac{1}{\sqrt{1-\rho}} \cdot \frac{1}{2} \log\left(\max\left\{\frac{16}{c_w}, \frac{12^2 m Q_{\max}^2}{c_w \beta \max(\beta,\mu)}, \frac{\bar{D}_\delta}{\min\left\{m\beta C_1, m\beta^4 C_1, \frac{m}{36C_1}\varepsilon^2\right\}}\right\}\right), \tag{63}$$

where $\bar{D}_\delta$ and C_1 are defined in (45) and (50), respectively. □

D.2. Complexity Analysis when $\beta \geq 1$

Theorem 16 ($\beta \geq 1$ and $L > 0$)**.** *Let Assumptions 1 and 3-5 hold and $\beta \geq 1$. Let $M_i = L > 0$, $\tau_i = 2\beta$, and recall the definition of $D > 0$ implying $\max_{i\in[m]} ||x_i^0 - \widehat{x}|| \leq D$. W.l.o.g. assume $D \geq 2/L$. Pick an arbitrary $\varepsilon > 0$. If a reference matrix $\overline{W}$ satisfying Assumption 6 is used in steps (7b)-(7c), with $\rho \triangleq \lambda_{\max}(\overline{W} - J) < 1$ and $K = \widetilde{\mathcal{O}}(\log(1/\varepsilon)/\sqrt{1-\rho})$ (the explicit expression is given in (63)), then the sequence $\{p^\nu\}$ generated by DiRegINA satisfies the following:*

(a) if $p^\nu \geq \beta \cdot (2LD^3)$,

$$p^{\nu+1} \leq \frac{5}{6}\, p^\nu,$$

(b) if $\varepsilon \leq p^\nu \leq \beta \cdot (2LD^3)$,

$$p^\nu \leq 24^2 \cdot (LD^3)^2 \cdot \frac{\beta^2}{\varepsilon} \cdot \frac{1}{\nu^2}.$$

Proof. Excluding β, the parameter setting is identical to Theorem 15. Recall (51), i.e.,

$$p^{\nu+1} \leq \min_{\alpha\in[0,\min\{1,\frac{p^\nu}{6\beta C_1}\}]} \left\{(1-\alpha/2)p^\nu + C_1\alpha^3 + \frac{1}{m\beta}\left\|\delta^\nu\right\|^2\right\}, \tag{64}$$

where C_1 is defined in (50). Denoting by α^* the minimizer of the optimization problem at the RHS of (51), we have:

(a) If $p^\nu \geq 6\beta C_1$, then $\alpha^* = 1$ and under (63), (64) yields

$$p^{\nu+1} \leq \frac{4 + 1/\beta}{6} p^\nu \leq \frac{5}{6}\, p^\nu.$$

(b) If $\varepsilon \leq p^\nu \leq 6\beta C_1$, then $\alpha^* = \frac{p^\nu}{6\beta C_1}$. Under (63), (64) yields

$$p^{\nu+1} \leq p^\nu - \frac{(p^\nu)^2}{36\beta C_1},$$

and following similar steps as in derivation of (62), we obtain

$$p^\nu \leq \frac{1}{\tilde{c}_0^2\nu^2} = 72^2 \cdot C_1^2 \cdot \frac{\beta^2}{\varepsilon} \cdot \frac{1}{\nu^2}.$$

□

D.3. Proof of main theorem

We proceed to prove Theorem 7. Given an accuracy $0 < \varepsilon \ll 1$, when $0 < \beta \leq 1$, Theorem 15 gives the following expression of rate: to achieve $p^\nu \leq \varepsilon$, DiRegINA requires

$$O\left(\log\left(\frac{1}{6C_1}\right) + \sqrt{\frac{LD^3}{\varepsilon}} + \frac{\beta\left(LD^3\right)}{\varepsilon}\right) = \tilde{O}\left(\sqrt{\frac{LD^3}{\varepsilon}} + \frac{\beta\left(LD^3\right)}{\varepsilon}\right), \tag{65}$$

iterations, while if $\beta \geq 1$, by Theorem 16, DiRegINA requires

$$O\left(\log\left(\frac{1}{2\beta LD^3}\right) + \frac{\beta\left(LD^3\right)}{\varepsilon}\right) = \tilde{O}\left(\frac{\beta\left(LD^3\right)}{\varepsilon}\right)$$

iterations. Therefore, (65) is a valid rate complexity expression (in terms of iterations) in both discussed cases (i.e. $0 < \beta \leq 1$ and $\beta \geq 1$). Now, recall that every iteration requires K rounds of communications, with K satisfying (41) and (63); hence $K = \tilde{O}\left(1/\sqrt{1-\rho}\cdot\log(1/\varepsilon)\right) = \tilde{O}\left(1/\sqrt{1-\rho}\cdot\varepsilon^{-\alpha/2}\right)$, for any arbitrary small $\alpha > 0$. Therefore the final communication complexity reads

$$\tilde{O}\left(\frac{1}{\sqrt{1-\rho}}\cdot\left\{\sqrt{\frac{LD^3}{\varepsilon^{1+\alpha}}} + \frac{\beta\left(LD^3\right)}{\varepsilon^{1+\frac{\alpha}{2}}}\right\}\right).$$

E. Proof of Theorem 9 and Corollary 11

We begin introducing some intermediate technical results, instrumental to proving the main theorems, namely: i) Lemmata 6-5 in Sec. E.1; and ii) a detailed "region-based" complexity of DiRegINA as in in Theorem 17 (cf. Sec. E.2). We prove Theorem 9 and the improved rates in case of quadratic functions in Sec. E.3 and Sec. E.4, respectively. Finally, Corollary 11 is proved in Sec. E.5.

E.1. Preliminary results

We establish necessary connections between the optimization error p^ν, the network error $\|\delta^\nu\|$ and $||\Delta x^\nu||$ in Lemmata 5-6:

Lemma 5. *Let Assumptions 2-4 hold, $\tau_i = 2\beta$, and $M_i \geq L/3$. Then*

$$\frac{1}{m}\sum_{i=1}^m \|\Delta x_i^\nu\|^2 \leq \frac{8}{\mu}\, p^\nu + \frac{2}{m\beta\mu}\,\|\delta^\nu\|^2, \tag{66}$$

where p^ν is defined in (9).

Proof. By μ-strongly convexity of F and optimality of $\widehat{x}$,

$$\begin{aligned} F(x_i^{\nu+}) - F(\widehat{x}) \geq& \frac{\mu}{2}\left\|x_i^{\nu+} - \widehat{x}\right\|^2 \geq \frac{\mu}{4}\left\|x_i^{\nu+} - x_i^\nu\right\|^2 - \frac{\mu}{2}\left\|x_i^\nu - \widehat{x}\right\|^2 \\ \geq& \frac{\mu}{4}\left\|x_i^{\nu+} - x_i^\nu\right\|^2 - \left(F(x_i^\nu) - F(\widehat{x})\right). \end{aligned}$$

Averaging the above inequalities over $i = 1, \ldots, m$, yields

$$\frac{1}{m}\sum_{i=1}^m \|\Delta x_i^\nu\|^2 \leq \frac{4}{\mu}\left(p^{\nu+} + p^\nu\right),$$

where $p^{\nu+} = (1/m)\sum_{i=1}^m \left\{F(x_i^{\nu+}) - F(\widehat{x})\right\}$. Using (32) proves (66). □

Lemma 6. *Let Assumptions 2-4 hold and set $\tau_i = 2\beta$. Define*

$$\omega_0 \triangleq \frac{12\beta}{\sqrt{L^2 + 4M_{\max}^2}}, \quad M_{\max} \triangleq \max_{i\in[m]} M_i.$$

Then

$$\frac{1}{m}\sum_{i=1}^{m}\left\{F(x_i^{\nu+})-F(\widehat{x})\right\}\leq\varphi\left(\{x_i^{\nu+}\}_i,\{x_i^{\nu}\}_i\right)+\frac{8}{m\mu}\left\|\delta^{\nu}\right\|^2, \tag{67}$$

where

$$\varphi\left(\{x_i^{\nu+}\}_i,\{x_i^{\nu}\}_i\right)=\begin{cases}\frac{L^2+4M_{\max}^2}{m\mu}\left(\sum_{i=1}^m\left\|x_i^{\nu+}-x_i^{\nu}\right\|^2\right)^2, & \text{if}\quad C\text{:}\ \sqrt{\sum_{i=1}^m\left\|x_i^{\nu+}-x_i^{\nu}\right\|^2}\geq\omega_0;\\ \frac{144\beta^2}{m\mu}\sum_{i=1}^m\left\|x_i^{\nu+}-x_i^{\nu}\right\|^2, & \text{if}\quad \overline{C}\text{:}\ \sqrt{\sum_{i=1}^m\left\|x_i^{\nu+}-x_i^{\nu}\right\|^2}<\omega_0.\end{cases}$$

Proof. Recall (43), a consequence of optimality of $x_i^{\nu+}$ (defined in (7a)), reads

$$\begin{aligned}&\left\langle\nabla F(x_i^{\nu})+\nabla^2F(x_i^{\nu})\Delta x_i^{\nu},\widehat{x}-x_i^{\nu+}\right\rangle\\ \geq&\left\langle\frac{M_i}{2}||\Delta x_i^{\nu}||\Delta x_i^{\nu},x_i^{\nu+}-\widehat{x}\right\rangle+\left\langle\delta_i^{\nu},x_i^{\nu+}-\widehat{x}\right\rangle+\left\langle\tilde{B}_i^{\nu}\Delta x_i^{\nu},x_i^{\nu+}-\widehat{x}\right\rangle,\end{aligned} \tag{68}$$

where $\tilde{B}_i^{\nu}=B_i^{\nu}+\tau_iI$ and recall $\Delta x_i^{\nu}=x_i^{\nu+}-x_i^{\nu}$ [cf. (22)]. By μ-strongly convexity of F,

$$\begin{aligned}&F(\widehat{x})-F(x_i^{\nu+})\\ \geq&\left\langle\nabla F(x_i^{\nu+}),\widehat{x}-x_i^{\nu+}\right\rangle+\frac{\mu}{2}\left\|x_i^{\nu+}-\widehat{x}\right\|^2\\ =&\left\langle\nabla F(x_i^{\nu+})-\nabla F(x_i^{\nu})-\nabla^2F(x_i^{\nu})\Delta x_i^{\nu},\widehat{x}-x_i^{\nu+}\right\rangle+\frac{\mu}{2}\left\|x_i^{\nu+}-\widehat{x}\right\|^2\\ &+\left\langle\nabla F(x_i^{\nu})+\nabla^2F(x_i^{\nu})\Delta x_i^{\nu},\widehat{x}-x_i^{\nu+}\right\rangle\\ \overset{(68)}{\geq}&\left\langle\nabla F(x_i^{\nu+})-\nabla F(x_i^{\nu})-\nabla^2F(x_i^{\nu})\Delta x_i^{\nu},\widehat{x}-x_i^{\nu+}\right\rangle+\frac{\mu}{2}\left\|x_i^{\nu+}-\widehat{x}\right\|^2\\ &+\left\langle\frac{M_i}{2}||\Delta x_i^{\nu}||\Delta x_i^{\nu},x_i^{\nu+}-\widehat{x}\right\rangle+\left\langle\delta_i^{\nu},x_i^{\nu+}-\widehat{x}\right\rangle+\left\langle\tilde{B}_i^{\nu}\Delta x_i^{\nu},x_i^{\nu+}-\widehat{x}\right\rangle\\ \geq&-\frac{1}{2\mu}\left\|\nabla F(x_i^{\nu+})-\nabla F(x_i^{\nu})-\nabla^2F(x_i^{\nu})\Delta x_i^{\nu}\right\|^2\\ &+\left\langle\frac{M_i}{2}||\Delta x_i^{\nu}||\Delta x_i^{\nu},x_i^{\nu+}-\widehat{x}\right\rangle+\left\langle\delta_i^{\nu},x_i^{\nu+}-\widehat{x}\right\rangle+\left\langle\tilde{B}_i^{\nu}\Delta x_i^{\nu},x_i^{\nu+}-\widehat{x}\right\rangle,\end{aligned} \tag{69}$$

and by applying Lemma 1 (cf. inequality (20)) to the first term on the RHS of (69) along with Cauchy-schwarz inequality, yield

$$\begin{aligned}&F(\widehat{x})-F(x_i^{\nu+})\\ \geq&-\left(\frac{L^2}{8\mu}+\frac{M_i}{4\epsilon_0}\right)\left\|\Delta x_i^{\nu}\right\|^4-\frac{M_i\epsilon_0}{4}\left\|x_i^{\nu+}-\widehat{x}\right\|^2-\frac{1}{2\epsilon_1}\left\|\delta_i^{\nu}\right\|^2-\frac{\epsilon_1}{2}\left\|x_i^{\nu+}-\widehat{x}\right\|^2+\left\langle\tilde{B}_i^{\nu}\Delta x_i^{\nu},x_i^{\nu+}-\widehat{x}\right\rangle\\ \overset{(a)}{\geq}&-\left(\frac{L^2}{8\mu}+\frac{M_i}{4\epsilon_0}\right)\left\|\Delta x_i^{\nu}\right\|^4-\left(\frac{M_i\epsilon_0}{2\mu}+\frac{\epsilon_1}{\mu}\right)\left(F(x_i^{\nu+})-F(\widehat{x})\right)-\frac{1}{2\epsilon_1}\left\|\delta_i^{\nu}\right\|^2+\left\langle\tilde{B}_i^{\nu}\Delta x_i^{\nu},x_i^{\nu+}-\widehat{x}\right\rangle,\end{aligned} \tag{70}$$

for arbitrary $\epsilon_0,\epsilon_1>0$, where (a) is due to the μ-strongly convexity of F and optimality of $\widehat{x}$. By Assumption 4 and some algebraic manipulations, the last term on the RHS of (70) is lower-bounded as

$$\begin{aligned}\left\langle\Delta x_i^{\nu},x_i^{\nu+}-\widehat{x}\right\rangle_{\tilde{B}_i^{\nu}}\geq&-\frac{\beta+\tau_i}{2\epsilon_2}\left\|\Delta x_i^{\nu}\right\|^2-\frac{\epsilon_2(\beta+\tau_i)}{2}\left\|x_i^{\nu+}-\widehat{x}\right\|^2\\ \overset{(a)}{\geq}&-\frac{\beta+\tau_i}{2\epsilon_2}\left\|\Delta x_i^{\nu}\right\|^2-\frac{\epsilon_2(\beta+\tau_i)}{\mu}\left(F(x_i^{\nu+})-F(\widehat{x})\right),\end{aligned} \tag{71}$$

with arbitrary $\epsilon_2>0$, where (a) follows from the μ-strong convexity of F and optimality of $\widehat{x}$. Set

$$\epsilon_0=\frac{\mu}{2M_{\max}},\quad \epsilon_1=\frac{\mu}{4},\quad \epsilon_2=\frac{\mu}{4(\beta+\tau_{\max})},$$

where $\tau_{\max} \triangleq \max_{i\in[m]}\ \tau_i$; then combining (70)-(71) and averaging over $i = 1, \dots, m$, lead to

$$\frac{1}{m}\sum_{i=1}^{m}\left(F(x_i^{\nu+}) - F(\widehat{x})\right) \leq \quad \frac{L^2 + 4M_{\max}^2}{2m\mu}\sum_{i=1}^{m}\|\Delta x_i^\nu\|^4 + \frac{8\left(\beta+\tau_{\max}\right)^2}{m\mu}\sum_{i=1}^{m}\|\Delta x_i^\nu\|^2 + \frac{8}{m\mu}\|\delta^\nu\|^2. \tag{72}$$

The bound (67) is a direct consequence of (72), with $\tau_i = 2\beta$, for all $i = 1, \dots, m$. □

E.2. Preliminary complexity results

Theorem 17. *Let Assumptions 2-5 hold. Let also $M_i \geq L$ and $\tau_i = 2\beta$, for all $i = 1, \dots, m$, and denote*

$$C_2 \triangleq \xi \cdot \frac{(M_{\max}+L)\sqrt{2m}}{3\mu^{3/2}}, \quad M_{\max} \triangleq \max_{i\in[m]} M_i,$$

for some arbitrary $\xi \geq 1$. If a reference matrix $\overline{W}$ satisfying Assumption 6 is used in steps (7b)-(7c), with $\rho \triangleq \lambda_{\max}(\overline{W} - J) < 1$ and $K = \widetilde{\mathcal{O}}(1/\sqrt{1-\rho})$ (the explicit expression of K is given in (97)), then the sequence $\{p^\nu\}$ generated by DiRegINA satisfies the following:

(a) If

$$p^\nu \geq \underline{p}_1 \triangleq \frac{\mu^3}{2m(M_{\max}+L)^2\xi^2}\left(1 + \frac{4\beta}{\mu}\right)^4,$$

then

$$(p^\nu)^{1/4} \leq (p^0)^{1/4} - \frac{\nu}{12\sqrt{3C_2}}.$$

(b) Assume [exclusively in this case (b)] $\beta \leq \mu$ and denote

$$\tilde{p}^\nu \triangleq p^\nu/c^2, \quad c \triangleq \frac{\mu\sqrt{\mu}}{8\sqrt{m(L^2+4M_{\max}^2)}}, \quad \underline{p}_2 \triangleq \frac{2\cdot 12^4}{L^2+4M_{\max}^2}\cdot\frac{\beta^2\mu}{m}.$$

If $p^\nu \geq \underline{p}_2$ and $p^{\nu-1} \leq c^2$, then $\tilde{p}^\nu \leq (\tilde{p}^{\nu-1})^2$.

(c) If

$$p^\nu < \underline{p}_3 \triangleq \frac{9}{L^2+4M_{\max}^2}\cdot\frac{\beta^2\mu}{m}, \tag{73}$$

then $\{p^\nu\}$ converges Q-linearly to zero with rate

$$\left(1 + \frac{\max(\beta,\mu)}{4mb_2}\right)^{-1} = \left(1 + \frac{1}{576}\cdot\frac{\mu\max(\beta,\mu)}{\beta^2}\right)^{-1}. \tag{74}$$

Proof. We organize the proof into three parts, **(a)-(c)**, in accordance with the three cases in the statement of the theorem.

(a) Recall Lemma 3 from the proof of Theorem 13:

$$F(x_i^{\nu+}) \leq \widetilde{F}_i(x_i^{\nu+}; x_i^\nu) + \frac{1}{2\epsilon}\|\delta_i^\nu\|^2, \tag{75}$$

for arbitrary $\epsilon > 0$, where $M_i \geq L$ and $\tau_i \geq \beta + \epsilon$. In addition, by the upperbound approximation of $\widetilde{F}_i(\cdot; x_i^\nu)$ in (21), there holds

$$\widetilde{F}_i(y; x_i^\nu) \leq F(y) + \frac{1}{2}\|y - x_i^\nu\|^2_{(\beta+\tau_i+\epsilon)I} + \frac{M_i+L}{6}\|y - x_i^\nu\|^3 + \frac{1}{2\epsilon}\|\delta_i^\nu\|^2, \quad \forall y \in \mathcal{K}. \tag{76}$$

Set $\tau_i = 2\beta$ and $\epsilon = \beta$, then by (75)-(76) and $x_i^{\nu+}$ being the minimizer of $\widetilde{F}(\cdot; x_i^\nu)$,

$$\begin{aligned}
&F(x_i^{\nu+}) - F(\widehat{x}) \\
\leq &\min_{y\in\mathcal{K}} \left\{ F(y) - F(\widehat{x}) + 2\beta \left\|y - x_i^\nu\right\|^2 + \frac{M_i + L}{6} \left\|y - x_i^\nu\right\|^3 + \frac{1}{\beta} \left\|\delta_i^\nu\right\|^2 \right\} \\
\leq &\min_{\alpha\in[0,\alpha_0]} \left\{ F(y) - F(\widehat{x}) + 2\beta \left\|y - x_i^\nu\right\|^2 + \frac{M_i + L}{6} \left\|y - x_i^\nu\right\|^3 + \frac{1}{\beta} \left\|\delta_i^\nu\right\|^2 : y = \alpha\widehat{x} + (1-\alpha)x_i^\nu \right\} \\
\overset{(a)}{\leq} &\min_{\alpha\in[0,\alpha_0]} \Big\{ (1-\alpha)\left(F(x_i^\nu) - F(\widehat{x})\right) - \frac{\alpha(1-\alpha)\mu}{2} \left\|x_i^\nu - \widehat{x}\right\|^2 \\
&\qquad + 2\beta\alpha^2 \left\|\widehat{x} - x_i^\nu\right\|^2 + \frac{M_i + L}{6}\alpha^3 \left\|\widehat{x} - x_i^\nu\right\|^3 + \frac{1}{\beta} \left\|\delta_i^\nu\right\|^2 \Big\},
\end{aligned} \tag{77}$$

where (a) is due to the μ-strong convexity of F. If $\alpha_0 = 1/(1 + 4\beta/\mu)$, (77) implies

$$F(x_i^{\nu+}) - F(\widehat{x}) \leq \min_{\alpha\in[0,\alpha_0]} \left\{ (1-\alpha)\left(F(x_i^\nu) - F(\widehat{x})\right) + \frac{M_i + L}{6}\alpha^3 \left\|\widehat{x} - x_i^\nu\right\|^3 + \frac{1}{\beta} \left\|\delta_i^\nu\right\|^2 \right\},$$

where by the μ-strongly convexity of F and optimality of $\widehat{x}$, we also deduce

$$\begin{aligned}
&F(x_i^{\nu+}) - F(\widehat{x}) \\
\leq &\min_{\alpha\in[0,\alpha_0]} \left\{ (1-\alpha)\left(F(x_i^\nu) - F(\widehat{x})\right) + \frac{M_i + L}{6}\alpha^3 \left(\frac{2}{\mu}\left(F(x_i^\nu) - F(\widehat{x})\right) \right)^{3/2} + \frac{1}{\beta} \left\|\delta_i^\nu\right\|^2 \right\}.
\end{aligned} \tag{78}$$

Averaging (78) over $i = 1, 2, \ldots, m$ while using (31), yields

$$p^{\nu+1} \leq \min_{\alpha\in[0,\alpha_0]} \left\{ (1-\alpha)p^\nu + C_2\alpha^3 \left(p^\nu\right)^{3/2} + \frac{1}{m\beta} \left\|\delta^\nu\right\|^2 \right\}, \quad C_2 \triangleq \xi \cdot \frac{(M_{\max} + L)\sqrt{2m}}{3\mu^{3/2}}, \tag{79}$$

where $M_{\max} = \max_{i\in[m]} M_i$ and $\xi \geq 1$ is arbitrary.

Denote by α^* the minimizer of the RHS of (79); then if $p^\nu \geq \underline{p}_1 \triangleq 1/(9C_2^2\alpha_0^4)$, we have $\alpha^* = 1/\sqrt{3C_2\sqrt{p^\nu}}$, and

$$p^{\nu+1} \leq p^\nu - \frac{2(p^\nu)^{3/4}}{3\sqrt{3C_2}} + \frac{1}{m\beta} \left\|\delta^\nu\right\|^2. \tag{80}$$

If

$$\frac{1}{m\beta} \left\|\delta^\nu\right\|^2 \leq \frac{1}{3\sqrt{3C_2}} (\underline{p}_1)^{3/4} \implies \frac{1}{m\beta} \left\|\delta^\nu\right\|^2 \leq \frac{1}{3\sqrt{3C_2}} (p^\nu)^{3/4}, \tag{81}$$

(80) yields

$$p^{\nu+1} \leq p^\nu - \tilde{c}\left(p^\nu\right)^{3/4}, \quad \forall \nu \geq 0, \quad \tilde{c} \triangleq \frac{1}{3\sqrt{3C_2}}. \tag{82}$$

Note that, by (45) and Lemma 4, condition (81) holds if

$$K \geq \frac{1}{\sqrt{1-\rho}} \cdot \frac{1}{2} \log\left(\frac{3\bar{D}_\delta\sqrt{3C_2}}{m\beta\underline{p}_1^{3/4}} \right). \tag{83}$$

We now prove by induction that (82) implies

$$(p^\nu)^{1/4} \leq l_\nu \triangleq (p^0)^{1/4} - \frac{\tilde{c}}{4}\nu, \quad \forall \nu \geq 0. \tag{84}$$

Clearly, (84) holds for $\nu = 0$. Since the RHS of (82) is increasing (as a function of p^ν) when $p^\nu \geq (3\tilde{c}/4)^4 = 1/(9 \cdot 2^8 C_2^2)$ (which holds since $p^\nu \geq \underline{p}_1$), then $p^\nu \leq l_\nu^4$ implies

$$p^{\nu+1} \leq l_\nu^4 - \tilde{c}l_\nu^3,$$

which also implies $p^{\nu+1} \leq l_{\nu+1}^4$, as by definition of l^ν in (84),

$$l_\nu^4 - l_{\nu+1}^4 = (l_\nu - l_{\nu+1})\,(l_\nu + l_{\nu+1})\left(l_\nu^2 + l_{\nu+1}^2\right) = \frac{\tilde{c}}{4}\,(l_\nu + l_{\nu+1})\left(l_\nu^2 + l_{\nu+1}^2\right) \leq \tilde{c}\, l_\nu^3.$$

(b) Recall (40) (from the proof of Theorem 13), which under Assumptions 2-6 and condition (41), reads

$$wp^{\nu+1} + (\tilde{\delta}^{\nu+1})^2 \leq wp^\nu + c_w(\tilde{\delta}^\nu)^2 - \frac{w\mu}{4m}||\Delta x^\nu||^2. \tag{85}$$

Recall also Lemma 6 when condition C is satisfied, which together with (31), implies

$$p^{\nu+1} \leq b_1 \left(\sum_{i=1}^m \left\|x_i^{\nu+} - x_i^\nu\right\|^2\right)^2 + \frac{8}{m\mu}\left\|\delta^\nu\right\|^2, \quad b_1 \triangleq \frac{L^2 + 4M_{\max}^2}{m\mu}. \tag{86}$$

Note that $p^{\nu+1} \geq \underline{p}_2$ implies that condition C in Lemma 6 holds, as proved next by contradiction. Suppose $p^{\nu+1} \geq \underline{p}_2$ but $||\Delta x^\nu|| < \omega_0$. Then Lemma 6 yields

$$\underline{p}_2 \leq p^{\nu+1} \overset{(31)}{\leq} p^{\nu+} < \frac{144\beta^2}{m\mu}\cdot\omega_0^2 + \frac{8}{m\mu}\left\|\delta^\nu\right\|^2 \overset{(a)}{\leq} \frac{2\cdot 12^4}{L^2 + 4M_{\max}^2}\cdot\frac{\beta^4}{m\mu},$$

implying $\beta > \mu$, which is in contradiction with the assumption; note that (a) holds under (similar to derivation of (83))

$$K \geq \frac{1}{\sqrt{1-\rho}}\cdot\frac{1}{2}\log\left(\frac{\bar{D}_\delta}{18\beta^2\omega_0^2}\right) \implies \frac{8}{m\mu}\left\|\delta^\nu\right\|^2 \leq \frac{144\beta^2\omega_0^2}{m\mu}. \tag{87}$$

Now since $x \mapsto x^h$ is subadditive for $0 \leq h \leq 1$, i.e. $(a+b)^h \leq a^h + b^h$ for any $a, b \geq 0$, (86) together with (35) imply

$$-\sum_{i=1}^m \left\|\Delta x_i^\nu\right\|^2 \leq -\,b_1^{-\frac{1}{2}}\left(p^{\nu+1}\right)^{\frac{1}{2}} + \sqrt{\frac{8}{m\mu b_1}}\,\tilde{\delta}^\nu. \tag{88}$$

Combining (85) with (88) yields

$$wp^{\nu+1} + (\tilde{\delta}^{\nu+1})^2 \leq wp^\nu + c_w(\tilde{\delta}^\nu)^2 - \frac{w\mu}{4m\sqrt{b_1}}\sqrt{p^{\nu+1}} + \frac{w\mu}{4m}\sqrt{\frac{8}{m\mu b_1}}\tilde{\delta}^\nu,$$

and since $\tilde{\delta}^\nu \leq \sqrt{\xi^\nu} \leq \sqrt{D_2}, \forall \nu \geq 0$ (see the discussion in Subsec. C.3, proof of Theorem 13), we get

$$wp^{\nu+1} + (\tilde{\delta}^{\nu+1})^2 \leq wp^\nu - \frac{w\mu}{4m\sqrt{b_1}}\sqrt{p^{\nu+1}} + C_3\tilde{\delta}^\nu, \quad C_3 \triangleq \left(c_w\sqrt{D_2} + \frac{c_w\beta\mu}{4m}\sqrt{\frac{8}{m\mu b_1}}\right). \tag{89}$$

Since $p^{\nu+1} \geq \underline{p}_2$, under (similar to derivation of (83))

$$K \geq \frac{1}{\sqrt{1-\rho}}\cdot\frac{1}{2}\log\left(\frac{64\bar{D}_\delta m^2 b_1 C_3^2}{c_w^2\beta^2\mu^2\underline{p}_2}\right) \implies C_3\tilde{\delta}^\nu \leq \frac{w\mu\sqrt{\underline{p}_2}}{8m\sqrt{b_1}}, \tag{90}$$

(89) yields

$$p^{\nu+1} + c\sqrt{p^{\nu+1}} \leq p^\nu, \quad c \triangleq \frac{\mu}{8m\sqrt{b_1}}.$$

Denote by $\tilde{p}^\nu \triangleq p^\nu/c^2$, then we get $\tilde{p}^{\nu+1} + \sqrt{\tilde{p}^{\nu+1}} \leq \tilde{p}^\nu$ which implies quadratic convergence when $p^{\nu+1} \geq \underline{p}_2$ and $\tilde{p}^\nu \leq 1 \equiv p^\nu \leq c^2$.

(c) Again recall (40):

$$wp^{\nu+1} + (\tilde{\delta}^{\nu+1})^2 \leq wp^\nu + c_w(\tilde{\delta}^\nu)^2 - \frac{w\max(\beta,\mu)}{4m}||\Delta x^\nu||^2. \tag{91}$$

Invoking Lemma 6 under condition $\bar{\mathsf{C}}$ and $\tau_i = 2\beta$, along with (31) and (35), we have

$$p^{\nu+1} \leq b_2 \sum_{i=1}^{m} \left\| x_i^{\nu+} - x_i^{\nu} \right\|^2 + \frac{8}{m\mu} (\tilde{\delta}^{\nu})^2, \quad b_2 \triangleq \frac{144\beta^2}{m\mu}. \tag{92}$$

Combining (91) and (92) yields

$$w \left(1 + \frac{\max(\beta,\mu)}{4mb_2}\right) p^{\nu+1} + (\tilde{\delta}^{\nu+1})^2 \leq w p^{\nu} + \left(c_w + \frac{2w\max(\beta,\mu)}{m^2\mu b_2}\right) (\tilde{\delta}^{\nu})^2, \tag{93}$$

where by choosing c_w to satisfy

$$\left(c_w + \frac{2w\max(\beta,\mu)}{m^2\mu b_2}\right) \leq \left(1 + \frac{\max(\beta,\mu)}{4mb_2}\right)^{-1} \overset{(a)}{\equiv} c_w \leq \left(1 + \frac{2\beta\max(\beta,\mu)}{m^2\mu b_2}\right)^{-1} \left(1 + \frac{\max(\beta,\mu)}{4mb_2}\right)^{-1}, \tag{94}$$

[where (a) is due to $w = c_w\beta$ defined in Sec. C.3], (93) becomes

$$w \left(1 + \frac{\max(\beta,\mu)}{4mb_2}\right) p^{\nu+1} + (\tilde{\delta}^{\nu+1})^2 \leq w p^{\nu} + \left(1 + \frac{\max(\beta,\mu)}{4mb_2}\right)^{-1} (\tilde{\delta}^{\nu})^2,$$

implying linear convergence of $\{\xi^{\nu}\}_{\nu}$ where

$$\zeta^{\nu} \triangleq w \left(1 + \frac{\max(\beta,\mu)}{4mb_2}\right) p^{\nu} + (\tilde{\delta}^{\nu})^2,$$

and decay rate

$$\left(1 + \frac{\max(\beta,\mu)}{4mb_2}\right)^{-1} = \left(1 + \frac{1}{576} \cdot \frac{\mu\max(\beta,\mu)}{\beta^2}\right)^{-1}. \tag{95}$$

Therefore, $\{p^{\nu}\}_{\nu}$ converges Q-linearly with rate (95).

Now let us derive (73) that defines this region. The goal is to identify the region where $\bar{\mathsf{C}}$ (cf. Lemma 6) holds. Under the condition (similar to derivation of (83))

$$K \geq \frac{1}{\sqrt{1-\rho}} \cdot \frac{1}{2} \log\left(\frac{4\bar{D}_{\delta}}{\beta\mu\omega_0^2}\right) \implies \frac{2(\tilde{\delta}^{\nu})^2}{\beta\mu} \leq \frac{\omega_o^2}{2}, \tag{96}$$

and Lemma 5, there holds

$$\frac{1}{m} \sum_{i=1}^{m} \|\Delta x_i^{\nu}\|^2 \leq \frac{8}{\mu} p^{\nu} + \frac{\omega_0^2}{2m},$$

which implies that $\bar{\mathsf{C}}$ is necessarily satisfied when

$$p^{\nu} < \frac{\omega_0^2\mu}{16m} = \frac{9}{L^2 + 4M_{\max}^2} \cdot \frac{\beta^2\mu}{m}.$$

Finally, unifying the conditions on K derived in (41). (83), (87), (90), (96), K must satisfy

$$K \geq \frac{1}{\sqrt{1-\rho}} \cdot \frac{1}{2} \log\left(\bar{D}_{\delta} \cdot \max\left\{\frac{16}{\bar{D}_{\delta}c_w}, \frac{12^2 m Q_{\max}^2}{\bar{D}_{\delta}c_w\beta\max(\beta,\mu)}, \frac{3\sqrt{3C_2}}{m\beta\underline{p}_1^{3/4}}, \frac{1}{18\beta^2\omega_0^2}, \frac{64m^2b_1C_3^2}{c_w^2\beta^2\mu^2\underline{p}_2}, \frac{4}{\beta\mu\omega_0^2}\right\}\right), \tag{97}$$

where recall that $c_w > 0$ must satisfy (94). □

E.3. Proof of Theorem 9

Let $M_i = L$ for all $i = 1, \dots, m$, and set the free parameter $\xi \geq 1$ (defined in Theorem 17) to $\xi = 100\sqrt{5}$, and define the regions of convergence,

$$\begin{aligned} &\text{(R0)}: \quad \Omega_0 \leq p^\nu, \\ &\text{(R1)}: \quad \Omega_1 \leq p^\nu < \Omega_0, \\ &\text{(R2)}: \quad \max(\varepsilon, \Omega_2) \leq p^\nu < \Omega_1, \\ &\text{(R3)}: \quad \varepsilon \leq p^\nu < \max(\varepsilon, \Omega_2), \end{aligned}$$

where

$$\Omega_0 = 244 \cdot D^2 \mu, \quad \Omega_1 = c^2/2 = \frac{1}{640 L^2} \cdot \frac{\mu^3}{m}, \quad \Omega_2 = \underline{p}_2 = \frac{2 \cdot 12^4}{5L^2} \cdot \frac{\beta^2 \mu}{m},$$

and c and $\underline{p}_2$ are defined in Theorem 17.

Using Theorem 15, region (R0) takes at most $\sqrt{\frac{LD}{\mu}}$ iterations. Now using Theorem 17, region (R1) lasts at most ν_1 iterations satisfying

$$(\Omega_1)^{1/4} \geq (\Omega_0)^{1/4} - \frac{\nu_1}{12\sqrt{3C_2}} \quad \Longleftarrow \quad \nu_1 \geq 480\sqrt{3\sqrt{5}} \cdot m^{1/4} \cdot \sqrt{\frac{LD}{\mu}}.$$

Let us conservatively consider scenarios $\Omega_1 \geq \varepsilon \geq \Omega_2$ and $\varepsilon < \Omega_2$, then the region of quadratic convergence (R2) lasts for at most

$$2\log\left(2\log\left(\min\left\{\frac{c^2}{\Omega_2}, \frac{c^2}{\varepsilon}\right\}\right)\right) \leq 2\log\left[2\log\left[\min\left\{\frac{1}{128 \cdot 12^4} \cdot \frac{\mu^2}{\beta^2}, \frac{\mu^3}{320 m L^2} \cdot \frac{1}{\varepsilon}\right\}\right]\right]: \quad c^2 \geq \Omega_2, \varepsilon \leq c^2,$$

iterations. Note that conditions $p^\nu \geq \underline{p}_2$ and $p^\nu < \underline{p}_3$ in Theorem 17 are sufficient conditions identifying the region of quadratic and linear rate (or more specifically $\mathtt{C}$ and $\overline{\mathtt{C}}$ in Lemma 6); note that $\underline{p}_2$ and $\underline{p}_3$ are identical up to multiplying constants. Hence, to obtain a valid complexity of overall performance, we pessimistically associate the region of linear rate (R3) with $\varepsilon < p^\nu \leq \max(\varepsilon, \Omega_2)$ rather than $\varepsilon < p^\nu \leq \max(\varepsilon, \underline{p}_3)$; therefore, this region at most lasts for $O(\beta/\mu \cdot \log(\max(\varepsilon, \Omega_2)/\varepsilon))$ iterations. Thus, since the number of communications per iteration is $\tilde{O}\left(1/\sqrt{1-\rho}\right)$ [cf. (41), (63), (97) and note that $\varepsilon = \Omega_0$ in (63)], the overall complexity reads

$$\tilde{O}\left(\frac{1}{\sqrt{1-\rho}}\left\{\sqrt{\frac{LD}{\mu}}\left(1 + m^{1/4}\right) + \log\left[\log\left[\frac{\mu^2}{\beta^2} \cdot \min\left\{1, \frac{\beta^2 \mu}{mL^2} \cdot \frac{1}{\varepsilon}\right\}\right]\right] + \frac{\beta}{\mu}\log\left[\max\left(1, \frac{\beta^2 \mu}{mL^2} \cdot \frac{1}{\varepsilon}\right)\right]\right\}\right)$$

communications.

E.4. The case of quadratic f_i in Theorem 9

Here we refine the proof of Theorem 9 to enhance the rate when $L = 0$:

Theorem 18. *Let Assumptions 2-5 hold with $L = 0$ and $\beta < \mu$. Denote by D_p an upperbound of p^0, i.e. $p^0 \leq D_p$ for all $\nu \geq 0$. Also choose $M_i = \Theta(\mu^{3/2}/\sqrt{mD_p})$ sufficiently small (explicit condition is provided in (98)) and $\tau_i = 2\beta$ for all $i = 1, \dots, m$. If a reference matrix $\overline{W}$ satisfying Assumption 6 is used in steps (7b)-(7c), with $\rho \triangleq \lambda_{\max}(\overline{W} - J) < 1$ and $K = \tilde{O}\left(1/\sqrt{1-\rho}\right)$ (explicit condition is provided in (97)), then for any given $\varepsilon > 0$, DiRegINA returns a solution with $p^\nu \leq \varepsilon$ after total*

$$\tilde{O}\left(\frac{1}{\sqrt{1-\rho}} \cdot \left\{\log\log\left(\frac{D_p}{\varepsilon}\right) + \frac{\beta}{\mu}\log\left(\frac{D_p \beta^2}{\mu^2 \varepsilon}\right)\right\}\right)$$

communications. Note that when $\beta = O(1/\sqrt{n})$, $\epsilon = \Omega(V_N)$ and $n \geq m$, the above communication complexity reduces to

$$\tilde{O}\left(\frac{1}{\sqrt{1-\rho}} \cdot \left\{\log\log\left(\frac{D_p}{V_N}\right)\right\}\right).$$

Proof. Let us specialize the results established in Theorem 17 (in particular case (b)-(c)). Note that, since $L = 0$, we can impose $p^0 \leq c^2/2$ by a proper choice of M_i, allowing DiRegINA to circumvent the first region (associated with case (a) in Theorem 17) and start off in the quadratic rate region. Hence we only need to derive a sufficient condition for $p^0 \leq c^2/2$. Let us first consider case (b): if $M_i = \Theta(\mu^{3/2}/\sqrt{mD_p}), \forall i$, sufficiently small,

$$M_i \leq \frac{\mu^{3/2}}{16\sqrt{2mD_p}},\ \forall i \implies p^0 \leq \frac{\mu^3}{512mM_{\max}^2} \implies p^0 \leq c^2/2, \tag{98}$$

where $M_{\max} \triangleq \max_{i\in[m]} M_i$. Let us also evaluate the precision achieved in case (b), i.e. $\underline{p}_2$: denote by $\underline{C}_M$ such that $M_i \geq \underline{C}_M \mu^{3/2}/\sqrt{mD_p}, \forall i$, then

$$\underline{p}_2 \triangleq \frac{12^4}{2M_{\max}^2} \cdot \frac{\beta^2\mu}{m} \leq \frac{12^4}{2\underline{C}_M^2} \cdot \frac{\beta^2 D_p}{\mu^2}.$$

Therefore the number of iterations to reach $\varepsilon = \Omega(\underline{p}_2)$ is $O(\log\log(c^2/\underline{p}_2)) = \log\log(D_p/\varepsilon)$, and since $K = \tilde{O}\left(1/\sqrt{1-\rho}\right)$, the total number of communication will be $\tilde{O}\left(1/\sqrt{1-\rho}\cdot\log\log(D_p/\varepsilon)\right)$.

Now let us derive the complexity when $\varepsilon = O(\underline{p}_2)$ (i.e. case (c) in Theorem 17). Setting $L = 0$ and following similar arguments, for arbitrary precision $\varepsilon > 0$, we obtain a communication complexity $\tilde{O}\left(1/\sqrt{1-\rho}\cdot\left\{\log\log(D_p/\varepsilon) + \beta/\mu\log(\beta^2 D_p/(\mu^2\varepsilon))\right\}\right)$. □

E.5. Proof of Corollary 11

Let us customize the rate established in Theorem 17 (in particular case (b)-(c)). We derive a sufficient condition for $p^0 \leq c^2/2$ which guarantees that the initial point is in the region of quadratic convergence. Using initialization policy (8), there holds $p^0 \leq C_\Delta/n$ for some $C_\Delta > 0$. Hence, under

$$n \geq \frac{640C_\Delta L^2}{\mu^3}\cdot m \implies p^0 \leq \frac{\mu^3}{640mL^2} \implies p^0 \leq c^2/2,$$

DiRegINA converges quadratically to the precision

$$\underline{p}_2 \triangleq \frac{2\cdot 12^4}{5L^2}\cdot\frac{\beta^2\mu}{m}.$$

By $\beta = O(1/\sqrt{n})$, $\underline{p}_2 = O(V_N)$. Hence, since $K = \tilde{O}\left(1/\sqrt{1-\rho}\right)$, the total number of communication will be $\tilde{O}\left(1/\sqrt{1-\rho}\cdot\log\log(\mu^3/(mL^2V_N))\right)$.

F. Proof of Theorem 12

Let $M_i = L$ for all $i = 1,\ldots,m$, and set the free parameter $\xi = 50\beta/(3\mu)$ (defined in Theorem 17) and define the regions of convergence,

$$\begin{aligned}
(\overline{\text{R0}}):&\quad \overline{\Omega}_0 \leq p^\nu,\\
(\overline{\text{R1}}):&\quad \overline{\Omega}_1 \leq p^\nu < \overline{\Omega}_0,\\
(\overline{\text{R2}}):&\quad \varepsilon \leq p^\nu < \overline{\Omega}_1,
\end{aligned}$$

where

$$\overline{\Omega}_0 = 244\cdot D^2\mu, \quad \overline{\Omega}_1 = \frac{0.9}{L^2}\cdot\frac{\beta^2\mu}{m}.$$

Using Theorem 15, region $(\overline{\text{R0}})$ takes at most $\sqrt{\frac{LD}{\mu}}$ iteration; note that $\mu = \Omega(\beta^2)$ by assumption $n \geq m$, thus $\overline{\Omega}_0 = \Omega(\beta^2\cdot 2LD^3)$. Now using Theorem 17, region $(\overline{\text{R1}})$ lasts at most ν_1 iteration satisfying

$$\left(\overline{\Omega}_1\right)^{1/4} \geq (\overline{\Omega}_0)^{1/4} - \frac{\nu_1}{12\sqrt{3C_2}} \iff \nu_1 \geq 240\sqrt{2}\cdot\frac{\sqrt{\beta LD\sqrt{m}}}{\mu}.$$

Finally, by case (c) in Theorem 17, region ($\overline{\text{R2}}$) lasts for $O(\beta/\mu \cdot \log(\overline{\Omega}_1/\varepsilon))$. Thus, since communication cost per iteration is $\tilde{O}\left(1/\sqrt{1-\rho}\right)$ [cf. (41), (97)], the overall complexity is

$$\tilde{O}\left(\frac{1}{\sqrt{1-\rho}}\left\{\sqrt{\frac{LD}{\mu}}\left(1+m^{1/4}\cdot\sqrt{\frac{\beta}{\mu}}\right)+\frac{\beta}{\mu}\log\left(\frac{\beta^2\mu}{mL^2}\cdot\frac{1}{\varepsilon}\right)\right\}\right).$$

G. The case of quadratic f_i in Theorem 12

Theorem 19. *Instate the setting of Theorem 12 where $L=0$. Then, the total number of communications for DiRegINA to make $p^\nu \le \varepsilon$ reads*

$$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}}\cdot\frac{\beta}{\mu}\log\left(\frac{1}{\varepsilon}\right)\right).$$

When $\beta = O(1/\sqrt{n})$, $\epsilon = \Omega(V_N)$ and $n \ge m$, the above communication complexity reduces to

$$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{1-\rho}}\cdot m^{1/2}\cdot\log\left(\frac{1}{V_N}\right)\right).$$

Proof. We customize case **(c)** in Theorem 17, when $L=0$. Note that $\bar{\mathbb{C}}$ in Lemma 6 holds for all $\nu \ge 0$ and condition (96) is no longer required. Therefore, the algorithm converges linearly with rate (74) and returns a solution within ε precision within $O\left(\beta/\mu \cdot \log(1/\varepsilon)\right)$ iterations and since $K = \tilde{O}\left(1/\sqrt{1-\rho}\right)$ [cf. (41)] , the total number of required communications is $\tilde{O}\left(1/\sqrt{1-\rho}\cdot\beta/\mu\cdot\log(1/\varepsilon)\right)$. □