A Eigen-analysis of G

In this section, we give a thorough analysis of the spectral properties of the matrix

$$G = \begin{bmatrix} \rho I & -\beta^{1/2}\widehat{A}^T \\ \beta^{1/2}\widehat{A} & \beta\widehat{C} \end{bmatrix}, \tag{20}$$

which is critical in analyzing the convergence of the PDBG, SAGA and SVRG algorithms for policy evaluation. Here $\beta = \sigma_w/\sigma_\theta$ is the ratio between the dual and primal step sizes in these algorithms. For convenience, we use the following notation:

$$L \triangleq \lambda_{\max}(\widehat{A}^T\widehat{C}^{-1}\widehat{A}),$$
$$\mu \triangleq \lambda_{\min}(\widehat{A}^T\widehat{C}^{-1}\widehat{A}).$$

Under Assumption 1, they are well defined and we have $L \geq \mu > 0$.

A.1 Diagonalizability of G

First, we examine the condition of β that ensures the diagonalizability of the matrix G. We cite the following result from (Shen et al., 2008).

Lemma 1. *Consider the matrix $\mathcal{A}$ defined as*

$$\mathcal{A} = \begin{bmatrix} A & -B^\top \\ B & C \end{bmatrix}, \tag{21}$$

where $A \succeq 0$, $C \succ 0$, and B is full rank. Let $\tau = \lambda_{\min}(C)$, $\delta = \lambda_{\max}(A)$ and $\sigma = \lambda_{\max}(B^\top C^{-1}B)$. If $\tau > \delta + 2\sqrt{\tau\sigma}$ holds, then $\mathcal{A}$ is diagonalizable with all its eigenvalues real and positive.

Applying this lemma to the matrix G in (20), we have

$$\tau = \lambda_{\min}(\beta\widehat{C}) = \beta\lambda_{\min}(\widehat{C}),$$
$$\delta = \lambda_{\max}(\rho I) = \rho,$$
$$\sigma = \lambda_{\max}\big(\beta^{1/2}\widehat{A}^\top(\beta\widehat{C})^{-1}\beta^{1/2}\widehat{A}\big) = \lambda_{\max}(\widehat{A}^\top\widehat{C}^{-1}\widehat{A}).$$

The condition $\tau > \delta + 2\sqrt{\tau\sigma}$ translates into

$$\beta\lambda_{\min}(\widehat{C}) > \rho + 2\sqrt{\beta\lambda_{\min}(\widehat{C})\lambda_{\max}(\widehat{A}^\top\widehat{C}^{-1}\widehat{A})},$$

which can be solved as

$$\sqrt{\beta} > \frac{\sqrt{\lambda_{\max}(\widehat{A}^\top\widehat{C}^{-1}\widehat{A})} + \sqrt{\rho + \lambda_{\max}(\widehat{A}^\top\widehat{C}^{-1}\widehat{A})}}{\sqrt{\lambda_{\min}(\widehat{C})}}.$$

In the rest of our discussion, we choose β to be

$$\beta = \frac{8\Big(\rho + \lambda_{\max}\big(\widehat{A}^\top\widehat{C}^{-1}\widehat{A}\big)\Big)}{\lambda_{\min}(\widehat{C})} = \frac{8(\rho + L)}{\lambda_{\min}(\widehat{C})}, \tag{22}$$

which satisfies the inequality above.

A.2 Analysis of eigenvectors

If the matrix G is diagonalizable, then it can be written as

$$G = Q\Lambda Q^{-1},$$

where Λ is a diagonal matrix whose diagonal entries are the eigenvalues of G, and Q consists of it eigenvectors (each with unit norm) as columns. Our goal here is to bound $\kappa(Q)$, the condition number of the matrix Q. Our analysis is inspired by Liesen & Parlett (2008). The core is the following fundamental result from linear algebra.

Theorem 4 (Theorem 5.1.1 of Gohberg et al. (2006))**.** *Suppose G is diagonalizable. If H is a symmetric positive definite matrix and HG is symmetric, then there exist a complete set of eigenvectors of G, such that they are orthonormal with respect to the inner product induced by H:*

$$Q^\top HQ = I. \tag{23}$$

If H satisfies the conditions in Theorem 4, then we have $H = Q^{-\top}Q^{-1}$, which implies $\kappa(H) = \kappa^2(Q)$. Therefore, in order to bound $\kappa(Q)$, we only need to find such an H and analyze its conditioning. To this end, we consider the matrix of the following form:

$$H = \begin{bmatrix} (\delta-\rho)I & \sqrt{\beta}\widehat{A}^\top \\ \sqrt{\beta}\widehat{A} & \beta\widehat{C} - \delta I \end{bmatrix}. \tag{24}$$

It is straightforward to check that HG is a symmetric matrix. The following lemma states the conditions for H being positive definite.

Lemma 2. *If $\delta - \rho > 0$ and $\beta\widehat{C} - \delta I - \frac{\beta}{\delta-\rho}\widehat{A}\widehat{A}^\top \succ 0$, then H is positive definite.*

Proof. The matrix H in (24) admits the following Schur decomposition:

$$H = \begin{bmatrix} I & 0 \\ \frac{\sqrt{\beta}}{\delta-\rho}\widehat{A} & I \end{bmatrix} \begin{bmatrix} (\delta-\rho)I & \\ & S \end{bmatrix} \begin{bmatrix} I & \frac{\sqrt{\beta}}{\delta-\rho}\widehat{A}^\top \\ 0 & I \end{bmatrix},$$

where $S = \beta\widehat{C} - \delta I - \frac{\beta}{\delta-\rho}\widehat{A}\widehat{A}^\top$. Thus H is congruence to the block diagonal matrix in the middle, which is positive definite under the specified conditions. Therefore, the matrix H is positive definite under the same conditions. □

In addition to the choice of β in (22), we choose δ to be

$$\delta = 4(\rho + L). \tag{25}$$

It is not hard to verify that this choice ensures $\delta - \rho > 0$ and $\beta\widehat{C} - \delta I - \frac{\beta}{\delta-\rho}\widehat{A}\widehat{A}^\top \succ 0$ so that H is positive definite. We now derive an upper bound on the condition number of H. Let λ be an eigenvalue of H and $[x^T y^T]^T$ be its associated eigenvector, where $\|x\|^2 + \|y\|^2 > 0$. Then it holds that

$$(\delta-\rho)x + \sqrt{\beta}\widehat{A}^T y = \lambda x, \tag{26}$$

$$\sqrt{\beta}\widehat{A}x + (\beta\widehat{C} - \delta I)y = \lambda y. \tag{27}$$

From (26), we have

$$x = \frac{\sqrt{\beta}}{\lambda - \delta + \rho}\widehat{A}^T y. \tag{28}$$

Note that $\lambda - \delta + \rho \neq 0$ because if $\lambda - \delta + \rho = 0$ we have $\widehat{A}^T y = 0$ so that $y = 0$ since $\widehat{A}$ is full rank. With $y = 0$ in (27), we will have $\widehat{A}x = 0$ so that $x = 0$, which contradicts the assumption that $\|x\|^2 + \|y\|^2 > 0$.

Substituting (28) into (27) and multiplying both sides with y^T, we obtain the following equation after some algebra

$$\lambda^2 - p\lambda + q = 0, \tag{29}$$

where

$$p \triangleq \delta - \rho + \frac{y^T(\beta\widehat{C} - \delta I)y}{\|y\|^2},$$
$$q \triangleq (\delta - \rho)\frac{y^T(\beta\widehat{C} - \delta I)y}{\|y\|^2} - \beta\frac{y^T\widehat{A}\widehat{A}^T y}{\|y\|^2}.$$

We can verify that both p and q are positive with our choice of δ and β. The roots of the quadratic equation in (29) are given by

$$\lambda = \frac{p \pm \sqrt{p^2 - 4q}}{2}. \tag{30}$$

Therefore, we can upper bound the largest eigenvalue as

$$\begin{aligned}
\lambda_{\max}(H) &\leq \frac{p + \sqrt{p^2 - 4q}}{2} \\
&\leq p = \delta - \rho - \delta + \beta\frac{y^T\widehat{C}y}{\|y\|^2} \\
&\leq -\rho + \beta\lambda_{\max}(\widehat{C}) \\
&= -\rho + \frac{8(\rho + L)}{\lambda_{\min}(\widehat{C})}\lambda_{\max}(\widehat{C}) \\
&\leq 8(\rho + L)\kappa(\widehat{C}).
\end{aligned} \tag{31}$$

Likewise, we can lower bound the smallest eigenvalue:

$$\begin{aligned}
\lambda_{\min}(H) &\geq \frac{p - \sqrt{p^2 - 4q}}{2} \geq \frac{p - p + 2q/p}{2} = \frac{q}{p} \\
&= \frac{\beta\left((\delta - \rho)\frac{y^T\widehat{C}y}{\|y\|^2} - \frac{y^T\widehat{A}\widehat{A}^T y}{\|y\|^2}\right) - \delta(\delta - \rho)}{-\rho + \beta\frac{y^T\widehat{C}y}{\|y\|^2}} \\
&\overset{(a)}{\geq} \frac{\beta\left((\delta - \rho)\frac{y^T\widehat{C}y}{\|y\|^2} - \frac{y^T\widehat{A}\widehat{A}^T y}{\|y\|^2}\right) - \delta(\delta - \rho)}{\beta\frac{y^T\widehat{C}y}{\|y\|^2}} \\
&= \delta - \rho - \frac{y^T\widehat{A}\widehat{A}^T y}{y^T\widehat{C}y} - \frac{\delta(\delta - \rho)}{\beta} \cdot \frac{1}{\frac{y^T\widehat{C}y}{\|y\|^2}} \\
&\overset{(b)}{\geq} \delta - \rho - L - \frac{\delta(\delta - \rho)}{\beta\lambda_{\min}(\widehat{C})} \\
&\overset{(c)}{=} (\rho + L)\left(3 - \frac{3\rho + 4L}{2(\rho + L)}\right) \\
&\geq \rho + L,
\end{aligned} \tag{32}$$

where step (a) uses the fact that both the numerator and denominator are positive, step (b) uses the fact

$$L \triangleq \lambda_{\max}\left(\widehat{A}^T\widehat{C}^{-1}\widehat{A}\right) \geq \frac{y^T\widehat{A}\widehat{A}^T y}{y^T\widehat{C}y},$$

and step (c) substitutes the expressions of δ and β. Therefore, we can upper bound the condition number of H, and thus that of Q, as follows:

$$\kappa^2(Q) = \kappa(H) \leq \frac{8(\rho + L)\kappa(\widehat{C})}{\rho + L} = 8\kappa(\widehat{C}). \tag{33}$$

A.3 Analysis of eigenvalues

Suppose λ is an eigenvalue of G and let $\left(\xi^\top, \eta^\top\right)^\top$ be its corresponding eigenvector. By definition, we have

$$G\begin{bmatrix}\xi \\ \eta\end{bmatrix} = \lambda\begin{bmatrix}\xi \\ \eta\end{bmatrix},$$

which is equivalent to the following two equations:

$$\rho\xi - \sqrt{\beta}\widehat{A}^\top\eta = \lambda\xi,$$
$$\sqrt{\beta}\widehat{A}\xi + \beta\widehat{C}\eta = \lambda\eta.$$

Solve ξ in the first equation in terms of η, then plug into the second equation, we obtain:

$$\lambda^2\eta - \lambda(\rho\eta + \beta\widehat{C}\eta) + \beta(\widehat{A}\widehat{A}^\top\eta + \rho\widehat{C}\eta) = 0.$$

Now left multiply $\eta^\top$, then divide by the $\|\eta\|_2^2$, we have:

$$\lambda^2 - p\lambda + q = 0.$$

where p and q are defined as

$$p \triangleq \rho + \beta\frac{\eta^\top\widehat{C}\eta}{\|\eta\|^2},$$
$$q \triangleq \beta\left(\frac{\eta^T\widehat{A}\widehat{A}^\top\eta}{\|\eta\|^2} + \rho\frac{\eta^T\widehat{C}\eta}{\|\eta\|^2}\right). \tag{34}$$

Therefore the eigenvalues of G satisfy:

$$\lambda = \frac{p \pm \sqrt{p^2 - 4q}}{2}. \tag{35}$$

Recall that our choice of β ensures that G is diagonalizable and has positive real eigenvalues. Indeed, we can verify that the diagonalization condition guarantees $p^2 \geq 4q$

so that all eigenvalues are real and positive. Now we can obtain upper and lower bounds based on (35). For upper bound, notice that

$$\begin{aligned}\lambda_{\max}(G) &\leq p \leq \rho + \beta\lambda_{\max}(\widehat{C})\\ &= \rho + \frac{8(\rho+L)}{\lambda_{\min}(\widehat{C}}\lambda_{\max}(\widehat{C})\\ &= \rho + 8(\rho+L)\kappa(\widehat{C})\\ &\leq 9\kappa(\widehat{C})\big(\rho+L\big)\\ &= 9\kappa(\widehat{C})\lambda_{\max}\big(\rho I + \widehat{A}^T\widehat{C}^{-1}\widehat{A}\big). \qquad (36)\end{aligned}$$

For lower bound, notice that

$$\begin{aligned}\lambda_{\min}(G) &\geq \frac{p-\sqrt{p^2-4q}}{2} \geq \frac{p-p+2q/p}{2} = q/p\\ &= \frac{\beta\left(\frac{\eta^T\widehat{A}\widehat{A}^T\eta}{\eta^T\widehat{C}\eta}+\rho\right)}{\rho\frac{\|\eta\|^2}{\eta^T\widehat{C}\eta}+\beta}\\ &\overset{(a)}{\geq} \frac{\beta(\rho+\mu)}{\rho/\lambda_{\min}(\widehat{C})+\beta} = \frac{\beta\lambda_{\min}(\widehat{C})(\rho+\mu)}{\rho+\beta\lambda_{\min}(\widehat{C})}\\ &\overset{(b)}{=} \frac{8(\rho+L)(\rho+\mu)}{\rho+8(\rho+L)}\\ &\geq \frac{8}{9}(\rho+\mu)\\ &= \frac{8}{9}(\rho+\lambda_{\min}(\widehat{A}^T\widehat{C}^{-1}\widehat{A}))\\ &= \frac{8}{9}\lambda_{\min}(\rho I+\widehat{A}^T\widehat{C}^{-1}\widehat{A}), \qquad (37)\end{aligned}$$

where the second inequality is by the concavity property of the square root function, step (a) used the fact

$$\mu \triangleq \lambda_{\min}\left(\widehat{A}^T\widehat{C}^{-1}\widehat{A}\right) \leq \frac{y^T\widehat{A}\widehat{A}^Ty}{y^T\widehat{C}y},$$

and step (b) substitutes the expressions of β.

Since G is not a normal matrix, we cannot use their eigenvalue bounds to bound its condition number $\kappa(G)$.

B Linear convergence of PDBG

Recall the saddle-point problem we need to solve:

$$\min_{\theta}\max_{w}\ \mathcal{L}(\theta,w),$$

where the Lagrangian is defined as

$$\mathcal{L}(\theta,w) = \frac{\rho}{2}\|\theta\|^2 - w^\top\widehat{A}\theta - \frac{1}{2}w^\top\widehat{C}w + \widehat{b}^\top w. \qquad (38)$$

Our assumption is that $\widehat{C}$ is positive definite and $\widehat{A}$ has full rank. The optimal solution can be expressed as

$$\theta_\star = \left(\widehat{A}^\top\widehat{C}^{-1}\widehat{A}+\rho I\right)^{-1}\widehat{A}^\top\widehat{C}^{-1}\widehat{b},$$

$$w_\star = \widehat{C}^{-1}\left(\widehat{b}-\widehat{A}^\top\theta_\star\right).$$

The gradients of the Lagrangian with respect to θ and w, respectively, are

$$\begin{aligned}\nabla_\theta\mathcal{L}(\theta,w) &= \rho\theta - \widehat{A}^\top w\\ \nabla_w\mathcal{L}(\theta,w) &= -\widehat{A}\theta - \widehat{C}w + \widehat{b}.\end{aligned}$$

The first-order optimality condition is obtained by setting them to zero, which is satisfied by $(\theta_\star, w_\star)$:

$$\begin{bmatrix}\rho I & -\widehat{A}^\top\\ \widehat{A} & \widehat{C}\end{bmatrix}\begin{bmatrix}\theta_\star\\ w_\star\end{bmatrix} = \begin{bmatrix}0\\ \widehat{b}\end{bmatrix}. \qquad (39)$$

The PDBG method in Algorithm 1 takes the following iteration:

$$\begin{bmatrix}\theta_{m+1}\\ w_{m+1}\end{bmatrix} = \begin{bmatrix}\theta_m\\ w_m\end{bmatrix} - \begin{bmatrix}\sigma_\theta & 0\\ 0 & \sigma_w\end{bmatrix}B(\theta_m,w_m),$$

where

$$B(\theta,w) = \begin{bmatrix}\nabla_\theta L(\theta,w)\\ -\nabla_w L(\theta,w)\end{bmatrix} = \begin{bmatrix}\rho I & -\widehat{A}^\top\\ \widehat{A} & \widehat{C}\end{bmatrix}\begin{bmatrix}\theta\\ w\end{bmatrix} - \begin{bmatrix}0\\ \widehat{b}\end{bmatrix}.$$

Letting $\beta = \sigma_w/\sigma_\theta$, we have

$$\begin{bmatrix}\theta_{m+1}\\ w_{m+1}\end{bmatrix} = \begin{bmatrix}\theta_m\\ w_m\end{bmatrix} - \sigma_\theta\left(\begin{bmatrix}\rho I & -\widehat{A}^\top\\ \beta\widehat{A} & \beta\widehat{C}\end{bmatrix}\begin{bmatrix}\theta_m\\ w_m\end{bmatrix} - \begin{bmatrix}0\\ \beta\widehat{b}\end{bmatrix}\right).$$

Subtracting both sides of the above recursion by $(\theta_\star, w_\star)$ and using (39), we obtain

$$\begin{bmatrix}\theta_{m+1}-\theta_\star\\ w_{m+1}-w_\star\end{bmatrix} = \begin{bmatrix}\theta_m-\theta_\star\\ w_m-w_\star\end{bmatrix} - \sigma_\theta\begin{bmatrix}\rho I & -\widehat{A}^T\\ \beta\widehat{A} & \beta\widehat{C}\end{bmatrix}\begin{bmatrix}\theta_m-\theta_\star\\ w_m-w_\star\end{bmatrix}.$$

We analyze the convergence of the algorithms by examining the differences between the current parameters to the optimal solution. More specifically, we define a scaled residue vector

$$\Delta_m \triangleq \begin{bmatrix}\theta_m-\theta_\star\\ \frac{1}{\sqrt{\beta}}(w_m-w_\star)\end{bmatrix}, \qquad (40)$$

which obeys the following iteration:

$$\Delta_{m+1} = (I-\sigma_\theta G)\,\Delta_m, \qquad (41)$$

where G is exactly the matrix defined in (20). As analyzed in Section A.1, if we choose β sufficiently large, such as in (22), then G is diagonalizable with all its eigenvalues real and positive. In this case, we let Q be the matrix of eigenvectors in the eigenvalue decomposition $G = Q\Lambda Q^{-1}$, and use the potential function

$$P_m \triangleq \left\|Q^{-1}\Delta_m\right\|_2^2$$

in our convergence analysis. We can bound the usual Euclidean distance by P_m as

$$\|\theta_m - \theta_\star\|^2 + \|w_m - w_\star\|^2 \leq (1+\beta)\sigma_{\max}^2(Q)P_m.$$

If we have linear convergence in P_m, then the extra factor $(1+\beta)\sigma_{\max}^2(Q)$ will appear inside a logarithmic term.

Remark: This potential function has an intrinsic geometric interpretation. We can view column vectors of Q^{-1} a basis for the vector space, which is *not* orthogonal. Our goal is to show that in this coordinate system, the distance to optimal solution shrinks at every iteration.

We proceed to bound the growth of P_m:

$$\begin{aligned}
P_{m+1} &= \left\|Q^{-1}\Delta_{m+1}\right\|_2^2 \\
&= \left\|Q^{-1}\left(I - \sigma_\theta G\right)\Delta_m\right\|_2^2 \\
&= \left\|Q^{-1}\left(QQ^{-1} - \sigma_\theta Q\Lambda Q^{-1}\right)\Delta_m\right\|_2^2 \\
&= \left\|\left(I - \sigma_\theta \Lambda\right)Q^{-1}\Delta_m\right\|_2^2 \\
&\leq \left\|I - \sigma_\theta \Lambda\right\|_2^2 \left\|Q^{-1}\Delta_m\right\|_2^2 \\
&= \left\|I - \sigma_\theta \Lambda\right\|_2^2 P_m
\end{aligned} \tag{42}$$

The inequality above uses sub-multiplicity of spectral norm. We choose σ_θ to be

$$\sigma_\theta = \frac{1}{\lambda_{\max}(\Lambda)} = \frac{1}{\lambda_{\max}(G)}, \tag{43}$$

Since all eigenvalues of G are real and positive, we have

$$\begin{aligned}
\|I - \sigma_\theta \Lambda\|^2 &= \left(1 - \frac{\lambda_{\min}(G)}{\lambda_{\max}(G)}\right)^2 \\
&\leq \left(1 - \frac{8}{81}\cdot\frac{1}{\kappa(\widehat{C})\kappa(\rho I + \widehat{A}^T\widehat{C}^{-1}\widehat{A})}\right)^2,
\end{aligned}$$

where we used the bounds on the eigenvalues $\lambda_{\max}(G)$ and $\lambda_{\min}(G)$ in (36) and (37) respectively. Therefore, we can achieve an ϵ-close solution with

$$m = O\left(\kappa(\widehat{C})\kappa(\rho I + \widehat{A}^T\widehat{C}^{-1}\widehat{A})\log\left(\frac{P_0}{\epsilon}\right)\right)$$

iterations of the PDBG algorithm.

In order to minimize $\|I - \sigma_\theta\Lambda\|$, we can choose

$$\sigma_\theta = \frac{2}{\lambda_{\max}(G) + \lambda_{\min}(G)},$$

which results in $\|I - \sigma_\theta\Lambda\| = 1 - 2/(1+\kappa(\Lambda))$ instead of $1 - 1/\kappa(\Lambda)$. The resulting complexity stays the same order.

The step sizes stated in Theorem 1 is obtained by replacing $\lambda_{\max}$ in (43) with its upper bound in (36) and setting σ_w through the ratio $\beta = \sigma_w/\sigma_\theta$ as in (22).

C Analysis of SVRG

Here we establish the linear convergence of the SVRG algorithm for policy evaluation described in Algorithm 2.

Recall the finite sum structure in $\widehat{A}$, $\widehat{b}$ and $\widehat{C}$:

$$\widehat{A} = \frac{1}{n}\sum_{t=1}^{n} A_t, \quad \widehat{b} = \frac{1}{n}\sum_{t=1}^{n} b_t, \quad \widehat{C} = \frac{1}{n}\sum_{t=1}^{n} C_t.$$

This structure carries over to the Lagrangian $\mathcal{L}(\theta, w)$ as well as the gradient operator $B(\theta, w)$, so we have

$$B(\theta, w) = \frac{1}{n}\sum_{t=1}^{n} B_t(\theta, w),$$

where

$$B_t(\theta, w) = \begin{bmatrix} \rho I & -A_t^\top \\ A_t & C_t \end{bmatrix}\begin{bmatrix} \theta \\ w \end{bmatrix} - \begin{bmatrix} 0 \\ b_t \end{bmatrix}. \tag{44}$$

Algorithm 2 has both an outer loop and an inner loop. We use the index m for the outer iteration and j for the inner iteration. Fixing the outer loop index m, we look at the inner loop of Algorithm 2. Similar to full gradient method, we first simplify the dynamics of SVRG.

$$\begin{aligned}
\begin{bmatrix} \theta_{m,j+1} \\ w_{m,j+1} \end{bmatrix} &= \begin{bmatrix} \theta_{m,j} \\ w_{m,j} \end{bmatrix} - \begin{bmatrix} \sigma_\theta & \\ & \sigma_w \end{bmatrix} \times \Big(B(\theta_{m-1}, w_{m-1}) \\
&\qquad + B_{t_j}(\theta_{m,j}, w_{m,j}) - B_t(\theta_{m-1}, w_{m-1}) \Big) \\
&= \begin{bmatrix} \theta_{m,j} \\ w_{m,j} \end{bmatrix} - \begin{bmatrix} \sigma_\theta & \\ & \sigma_w \end{bmatrix} \\
&\quad \times \left(\begin{bmatrix} \rho I & -\widehat{A}^\top \\ \widehat{A} & \widehat{C} \end{bmatrix}\begin{bmatrix} \theta_{m-1} \\ w_{m-1} \end{bmatrix} - \begin{bmatrix} 0 \\ \widehat{b} \end{bmatrix} \right. \\
&\qquad + \begin{bmatrix} \rho I & -A_t^\top \\ A_t & C_t \end{bmatrix}\begin{bmatrix} \theta_{m,j} \\ w_{m,j} \end{bmatrix} - \begin{bmatrix} 0 \\ b_t \end{bmatrix} \\
&\qquad \left. - \begin{bmatrix} \rho I & -A_t^\top \\ A_t & C_t \end{bmatrix}\begin{bmatrix} \theta_{m-1} \\ w_{m-1} \end{bmatrix} + \begin{bmatrix} 0 \\ b_t \end{bmatrix} \right).
\end{aligned}$$

Subtracting $(\theta_\star, w_\star)$ from both sides and using the optimality condition (39), we have

$$\begin{aligned}
\begin{bmatrix} \theta_{m,j+1} - \theta_\star \\ w_{m,j+1} - w_\star \end{bmatrix} &= \begin{bmatrix} \theta_{m,j} - \theta_\star \\ w_{m,j} - w_\star \end{bmatrix} - \begin{bmatrix} \sigma_\theta & \\ & \sigma_w \end{bmatrix} \\
&\quad \times \left(\begin{bmatrix} \rho I & -\widehat{A}^\top \\ \widehat{A} & \widehat{C} \end{bmatrix}\begin{bmatrix} \theta_{m-1} - \theta_\star \\ w_{m-1} - w_\star \end{bmatrix} \right. \\
&\qquad + \begin{bmatrix} \rho I & -A_t^\top \\ A_t & C_t \end{bmatrix}\begin{bmatrix} \theta_{m,j} - \theta_\star \\ w_{m,j} - w_\star \end{bmatrix} \\
&\qquad \left. - \begin{bmatrix} \rho I & -A_t^\top \\ A_t & C_t \end{bmatrix}\begin{bmatrix} \theta_{m-1} - \theta_\star \\ w_{m-1} - w_\star \end{bmatrix} \right).
\end{aligned}$$

Multiplying both sides of the above recursion by $\mathrm{diag}(I, 1/\sqrt{\beta}I)$, and using a residue vector $\Delta_{m,j}$ defined similarly as in (40), we obtain

$$\begin{aligned}\Delta_{m,j+1} &= \Delta_{m,j} - \sigma_\theta(G\Delta_{m-1} + G_{t_j}\Delta_{m,j} - G_{t_j}\Delta_{m-1}) \\ &= (I - \sigma_\theta G)\,\Delta_{m,j} \\ &\quad + \sigma_\theta \left(G - G_{t_j}\right)(\Delta_{m,j} - \Delta_{m-1}), \end{aligned} \tag{45}$$

where G_{t_j} is defined in (18).

For SVRG, we use the following potential functions to facilitate our analysis:

$$P_m \triangleq \mathbb{E}\left[\left\|Q^{-1}\Delta_m\right\|^2\right], \tag{46}$$

$$P_{m,j} \triangleq \mathbb{E}\left[\left\|Q^{-1}\Delta_{m,j}\right\|^2\right]. \tag{47}$$

Unlike the analysis for the batch gradient methods, the non-orthogonality of the eigenvectors will lead to additional dependency of the iteration complexity on the condition number of Q, for which we give a bound in (33).

Multiplying both sides of Eqn. (45) by Q^{-1}, taking squared 2-norm and taking expectation, we obtain

$$\begin{aligned}P_{m,j+1} &= \mathbb{E}\Big[\big\|Q^{-1}\big[\,(I - \sigma_\theta G)\,\Delta_{m,j} \\ &\qquad + \sigma_\theta\left(G - G_{t_j}\right)(\Delta_{m,j} - \Delta_{m-1})\,\big]\big\|^2\Big] \\ &\overset{(a)}{=} \mathbb{E}\Big[\,\big\|(I - \sigma_\theta\Lambda)\,Q^{-1}\Delta_{m,j}\big\|^2\,\Big] \\ &\quad + \sigma_\theta^2\,\mathbb{E}\Big[\,\big\|Q^{-1}\left(G - G_{t_j}\right)(\Delta_{m,j} - \Delta_{m-1})\big\|^2\,\Big] \\ &\overset{(b)}{\le} \|I - \sigma_\theta\Lambda\|^2\,\mathbb{E}\Big[\big\|Q^{-1}\Delta_{m,j}\big\|^2\Big] \\ &\quad + \sigma_\theta^2\,\mathbb{E}\Big[\,\big\|Q^{-1}G_{t_j}(\Delta_{m,j} - \Delta_{m-1})\big\|^2\,\Big] \\ &\overset{(c)}{=} \|I - \sigma_\theta\Lambda\|^2\,P_{m,j} \\ &\quad + \sigma_\theta^2\,\mathbb{E}\Big[\,\big\|Q^{-1}G_{t_j}(\Delta_{m,j} - \Delta_{m-1})\big\|^2\,\Big]. \end{aligned} \tag{48}$$

where step (a) used the facts that G_{t_j} is independent of $\Delta_{m,j}$ and Δ_{m-1} and $\mathbb{E}[G_{t_j}] = G$ so the cross terms are zero, step (b) used again the same independence and that the variance of a random variable is less than its second moment, and step (c) used the definition of $P_{m,j}$ in (47). To bound the last term in the above inequality, we use the simple notation $\delta = \Delta_{m,j} - \Delta_{m-1}$ and have

$$\begin{aligned}\big\|Q^{-1}G_{t_j}\delta\big\|^2 &= \delta^T G_{t_j}^T Q^{-T}Q^{-1}G_{t_j}\delta \\ &\le \lambda_{\max}(Q^{-T}Q^{-1})\delta^T G_{t_j}^T G_{t_j}\delta. \end{aligned}$$

Therefore, we can bound the expectation as

$$\begin{aligned}&\mathbb{E}\big[\big\|Q^{-1}G_{t_j}\delta\big\|^2\big] \\ \le&\lambda_{\max}(Q^{-T}Q^{-1})\mathbb{E}\big[\delta^T G_{t_j}^T G_{t_j}\delta\big] \\ =&\lambda_{\max}(Q^{-T}Q^{-1})\mathbb{E}\big[\delta^T \mathbb{E}[G_{t_j}^T G_{t_j}]\delta\big] \\ \le&\lambda_{\max}(Q^{-T}Q^{-1})L_G^2\mathbb{E}\big[\delta^T\delta\big] \\ =&\lambda_{\max}(Q^{-T}Q^{-1})L_G^2\mathbb{E}\big[\delta^T Q^{-T}Q^T Q Q^{-1}\delta\big] \\ =&\lambda_{\max}(Q^{-T}Q^{-1})\lambda_{\max}(Q^TQ)L_G^2\mathbb{E}\big[\delta^T Q^{-T}Q^{-1}\delta\big] \\ \le&\kappa(Q)^2L_G^2\mathbb{E}\big[\|Q^{-1}\delta\|^2\big], \end{aligned} \tag{49}$$

where in the second inequality we used the definition of L_G^2 in (18), i.e., $L_G^2 = \|\mathbb{E}[G_{t_j}^T G_{t_j}]\|$. In addition, we have

$$\begin{aligned}\mathbb{E}\big[\|Q^{-1}\delta\|^2\big] =&\mathbb{E}\big[\big\|Q^{-1}(\Delta_{m,j} - \Delta_{m-1})\big\|^2\big] \\ \le&2\,\mathbb{E}\big[\big\|Q^{-1}\Delta_{m,j}\big\|^2\big] + 2\,\mathbb{E}\big[\big\|Q^{-1}\Delta_{m-1}\big\|^2\big] \\ =&\,2P_{m,j} + 2P_{m-1}. \end{aligned}$$

Then it follows from (48) that

$$\begin{aligned}P_{m,j+1} \le&\|I - \sigma_\theta\Lambda\|^2 P_{m,j} \\ &+ 2\sigma_\theta^2\kappa^2(Q)L_G^2(P_{m,j} + P_{m-1}). \end{aligned}$$

Next, let $\lambda_{\max}$ and $\lambda_{\min}$ denote the largest and smallest diagonal elements of Λ (eigenvalues of G), respectively. Then we have

$$\begin{aligned}\|I - \sigma_\theta\Lambda\|^2 &= \max\big\{(1 - \sigma_\theta\lambda_{\min})^2,\ (1 - \sigma_\theta\lambda_{\min})^2\big\} \\ &\le 1 - 2\sigma_\theta\lambda_{\min} + \sigma_\theta^2\lambda_{\max}^2 \\ &\le 1 - 2\sigma_\theta\lambda_{\min} + \sigma_\theta^2\kappa^2(Q)L_G^2, \end{aligned}$$

where the last inequality uses the relation

$$\lambda_{\max}^2 \le \|G\|^2 = \|\mathbb{E}G_t\|^2 \le \|\mathbb{E}G_t^T G_t\| = L_G^2 \le \kappa^2(Q)L_G^2.$$

It follows that

$$\begin{aligned}P_{m,j+1} &\le \big(1 - 2\sigma_\theta\lambda_{\min} + \sigma_\theta^2\kappa^2(Q)\,L_G^2\big)P_{m,j} \\ &\quad + 2\sigma_\theta^2\,\kappa^2(Q)\,L_G^2(P_{m,j} + P_{m-1}) \\ &= \big[1 - 2\sigma_\theta\lambda_{\min} + 3\sigma_\theta^2\kappa^2(Q)\,L_G^2\big]\,P_{m,j} \\ &\quad + 2\sigma_\theta^2\,\kappa^2(Q)\,L_G^2 P_{m-1} \end{aligned}$$

If we choose σ_θ to satisfy

$$0 < \sigma_\theta \le \frac{\lambda_{\min}}{3\kappa^2(Q)\,L_G^2}, \tag{50}$$

then $3\sigma_\theta^2\kappa^2(Q)\,L_G^2 < \sigma_\theta\lambda_{\min}$, which implies

$$P_{m,j+1} \le (1 - \sigma_\theta\lambda_{\min})\,P_{m,j} + 2\sigma_\theta^2\,\kappa^2(Q)\,L_G^2 P_{m-1}.$$

Iterating the above inequality over $j = 1, \cdots, N-1$ and using $P_{m,0} = P_{m-1}$ and $P_{m,N} = P_m$, we obtain

$$\begin{aligned}P_m &= P_{m,N} \\ &\le \left[\big(1 - \sigma_\theta\lambda_{\min}\big)^N + 2\sigma_\theta^2\kappa^2(Q)L_G^2\sum_{j=0}^{N-1}\big(1 - \sigma_\theta\lambda_{\min}\big)^j\right]P_{m-1} \end{aligned}$$

$$= \left[\left(1-\sigma_\theta\lambda_{\min}\right)^N + 2\sigma_\theta^2\kappa^2(Q)L_G^2 \frac{1-(1-\sigma_\theta\lambda_{\min})^N}{1-(1-\sigma_\theta\lambda_{\min})}\right] P_{m-1}$$
$$\leq \left[\left(1-\sigma_\theta\lambda_{\min}\right)^N + \frac{2\sigma_\theta^2\kappa^2(Q)\,L_G^2}{\sigma_\theta\lambda_{\min}}\right] P_{m-1}$$
$$= \left[\left(1-\sigma_\theta\lambda_{\min}\right)^N + \frac{2\sigma_\theta\kappa^2(Q)\,L_G^2}{\lambda_{\min}}\right] P_{m-1}. \qquad (51)$$

We can choose

$$\sigma_\theta = \frac{\lambda_{\min}}{5\kappa^2(Q)L_G^2}, \quad N = \frac{1}{\sigma_\theta\lambda_{\min}} = \frac{5\kappa^2(Q)L_G^2}{\lambda_{\min}^2}, \quad (52)$$

which satisfies the condition in (50) and results in

$$P_m \leq (e^{-1} + 2/5)P_{m-1} \leq (4/5)P_{m-1}.$$

There are many other similar choices, for example,

$$\sigma_\theta = \frac{\lambda_{\min}}{3\kappa^2(Q)L_G^2}, \quad N = \frac{3}{\sigma_\theta\lambda_{\min}} = \frac{9\kappa^2(Q)L_G^2}{\lambda_{\min}^2},$$

which results in

$$P_m \leq (e^{-3} + 2/3)P_{m-1} \leq (3/4)P_{m-1}.$$

These results imply that the number of outer iterations needed to have $\mathbb{E}[P_m] \leq \epsilon$ is $\log(P_0/\epsilon)$. For each outer iteration, the SVRG algorithm need $O(nd)$ operations to compute the full gradient operator $B(\theta, w)$, and then $N = O(\kappa^2(Q)L_G^2/\lambda_{\min}^2)$ inner iterations with each costing $O(d)$ operations. Therefore the overall computational cost is

$$O\left(\left(n + \frac{\kappa^2\,(Q)\,L_G^2}{\lambda_{\min}^2}\right) d\ \log\left(\frac{P_0}{\epsilon}\right)\right).$$

Substituting (33) and (37) in the above bound, we get the overall cost estimate

$$O\left(\left(n + \frac{\kappa(\widehat{C})L_G^2}{\lambda_{\min}^2(\rho I + \widehat{A}^T\widehat{C}^{-1}\widehat{A})}\right) d\ \log\left(\frac{P_0}{\epsilon}\right)\right).$$

Finally, substituting the bounds in (33) and (37) into (52), we obtain the σ_θ and N stated in Theorem 2:

$$\sigma_\theta = \frac{\lambda_{\min}(\rho I + \widehat{A}^T\widehat{C}^{-1}\widehat{A})}{48\kappa(\widehat{C})L_G^2},$$
$$N = \frac{51\kappa^2(\widehat{C})L_G^2}{\lambda_{\min}^2(\rho I + \widehat{A}^T\widehat{C}^{-1}\widehat{A})},$$

which achieves the same complexity.

D Analysis of SAGA

SAGA in Algorithm 3 maintains a table of previously computed gradients. Notation wise, we use ϕ_t^m to denote that at m-th iteration, g_t is computed using $\theta_{\phi_t^m}$ and $w_{\phi_t^m}$. With this definition, ϕ_t^m has the following dynamics:

$$\phi_t^{m+1} = \begin{cases} \phi_t^m & \text{if } t_m \neq t, \\ m & \text{if } t_m = t. \end{cases} \qquad (53)$$

We can write the m-th iteration's full gradient as

$$B = \frac{1}{n}\sum_{t=1}^n B_t\left(\theta_{\phi_t^m}, w_{\phi_t^m}\right).$$

For convergence analysis, we define the following quantity:

$$\Delta_{\phi_t^m} \triangleq \begin{bmatrix} \theta_{\phi_t^m} - \theta_\star \\ \frac{1}{\sqrt{\beta}}(w_{\phi_t^m} - w_\star) \end{bmatrix}. \qquad (54)$$

Similar to (53), it satisfies the following iterative relation:

$$\Delta_{\phi_t^{m+1}} = \begin{cases} \Delta_{\phi_t^m} & \text{if} \quad t_m \neq t, \\ \Delta_m & \text{if} \quad t_m = t. \end{cases}$$

With these notations, we can express the vectors used in SAGA as

$$B_m = \frac{1}{n}\sum_{t=1}^n \begin{bmatrix} \rho I & -A_t^T \\ A_t & C_t \end{bmatrix} \begin{bmatrix} \theta_{\phi_t^m} \\ w_{\phi_t^m} \end{bmatrix} - \frac{1}{n}\sum_{t=1}^n \begin{bmatrix} 0 \\ b_t \end{bmatrix},$$
$$h_{t_m} = \begin{bmatrix} \rho I & -A_{t_m}^T \\ A_{t_m} & C_{t_m} \end{bmatrix} \begin{bmatrix} \theta_m \\ w_m \end{bmatrix} - \begin{bmatrix} 0 \\ b_{t_m} \end{bmatrix},$$
$$g_{t_m} = \begin{bmatrix} \rho I & -A_{t_m}^T \\ A_{t_m} & C_{t_m} \end{bmatrix} \begin{bmatrix} \theta_{\phi_t^m} \\ w_{\phi_t^m} \end{bmatrix} - \begin{bmatrix} 0 \\ b_{t_m} \end{bmatrix}.$$

The dynamics of SAGA can be written as

$$\begin{bmatrix} \theta_{m+1} \\ w_{m+1} \end{bmatrix} = \begin{bmatrix} \theta_m \\ w_m \end{bmatrix} - \begin{bmatrix} \sigma_\theta & \\ & \sigma_w \end{bmatrix} (B_m + h_{t_m} - g_{t_m})$$
$$= \begin{bmatrix} \theta_m \\ w_m \end{bmatrix} - \begin{bmatrix} \sigma_\theta & \\ & \sigma_w \end{bmatrix} \left\{ \frac{1}{n}\sum_{t=1}^n \begin{bmatrix} \rho I & -A_t^T \\ A_t & C_t \end{bmatrix} \begin{bmatrix} \theta_{\phi_t^m} \\ w_{\phi_t^m} \end{bmatrix} + \frac{1}{n}\sum_{t=1}^n \begin{bmatrix} 0 \\ b_t \end{bmatrix} + \begin{bmatrix} \rho I & -A_{t_m}^T \\ A_{t_m} & C_{t_m} \end{bmatrix} \begin{bmatrix} \theta_m \\ w_m \end{bmatrix} - \begin{bmatrix} \rho I & -A_{t_m}^T \\ A_{t_m} & C_{t_m} \end{bmatrix} \begin{bmatrix} \theta_{\phi_{t_m}^m} \\ w_{\phi_{t_m}^m} \end{bmatrix} \right\}$$

Subtracting $(\theta_\star, w_\star)$ from both sides, and using the optimality condition in (39), we obtain

$$\begin{bmatrix} \theta_{m+1} - \theta_\star \\ w_{m+1} - w_\star \end{bmatrix} = \begin{bmatrix} \theta_m - \theta_\star \\ w_m - w_\star \end{bmatrix} - \begin{bmatrix} \sigma_\theta & \\ & \sigma_w \end{bmatrix} \left\{ \frac{1}{n}\sum_{t=1}^n \begin{bmatrix} \rho I & -A_t^T \\ A_t & C_t \end{bmatrix} \begin{bmatrix} \theta_{\phi_t^m} - \theta_\star \\ w_{\phi_t^m} - w_\star \end{bmatrix} + \begin{bmatrix} \rho I & -A_{t_m}^T \\ A_{t_m} & C_{t_m} \end{bmatrix} \begin{bmatrix} \theta_m - \theta_\star \\ w_m - w_\star \end{bmatrix} \right.$$

$$- \begin{bmatrix} \rho I & -A_{t_m}^T \\ A_{t_m} & C_{t_m} \end{bmatrix} \begin{bmatrix} \theta_{\phi_{t_m}^m} - \theta_\star \\ w_{\phi_{t_m}^m} - w_\star \end{bmatrix} \Bigg\}.$$

Multiplying both sides by $\mathrm{diag}(I, 1/\sqrt{\beta} I)$, we get

$$\Delta_{m+1} = \Delta_m - \left(\frac{\sigma_\theta}{n} \sum_{t=1}^{n} G_t \Delta_{\phi_t^m} \right) - \sigma_\theta G_{t_m} \left(\Delta_m - \Delta_{\phi_{t_m}^m} \right). \tag{55}$$

where G_{t_m} is defined in (18).

For SAGA, we use the following two potential functions:

$$P_m = \mathbb{E} \left\| Q^{-1} \Delta_m \right\|_2^2,$$
$$Q_m = \mathbb{E}\Big[\frac{1}{n} \sum_{t=1}^{n} \left\| Q^{-1} G_t \Delta_{\phi_t^m} \right\|_2^2 \Big] = \mathbb{E}\Big[\Big\| Q^{-1} G_{t_m} \Delta_{\phi_{t_m}^m} \Big\|_2^2 \Big].$$

The last equality holds because we use uniform sampling. We first look at how P_m evolves. To simplify notation, let

$$v_m = \left(\frac{\sigma_\theta}{n} \sum_{t=1}^{n} G_t \Delta_{\phi_t^m} \right) + \sigma_\theta G_{t_m} \left(\Delta_m - \Delta_{\phi_{t_m}^m} \right),$$

so that (55) becomes $\Delta_{m+1} = \Delta_m - v_m$. We have

$$\begin{aligned} P_{m+1} &= \mathbb{E}\left[\left\| Q^{-1} \Delta_{m+1} \right\|_2^2 \right] \\ &= \mathbb{E}\Big[\left\| Q^{-1} (\Delta_m - v_m) \right\|^2 \Big] \\ &= \mathbb{E}\Big[\left\| Q^{-1} \Delta_m \right\|_2^2 - 2 \Delta_m^\top Q^{-\top} Q^{-1} v_m + \left\| Q^{-1} v_m \right\|_2^2 \Big] \\ &= P_m - \mathbb{E}\big[2 \Delta_m^\top Q^{-\top} Q^{-1} v_m \big] + \mathbb{E}\Big[\left\| Q^{-1} v_m \right\|_2^2 \Big]. \end{aligned}$$

Since Δ_m is independent of t_m, we have

$$\mathbb{E}\Big[2 \Delta_m^\top Q^{-\top} Q^{-1} v_m \Big] = \mathbb{E}\Big[2 \Delta_m^\top Q^{-\top} Q^{-1} \mathbb{E}_{t_m}[v_m] \Big],$$

where the inner expectation is with respect to t_m conditioned on all previous random variables. Notice that

$$\mathbb{E}_{t_m} \left[G_{t_m} \Delta_{\phi_{t_m}^m} \right] = \frac{1}{n} \sum_{t=1}^{n} G_t \Delta_{\phi_t^m},$$

which implies $\mathbb{E}_{t_m}[v_m] = \sigma_\theta \mathbb{E}_{t_m}[G_{t_m}] \Delta_m = \sigma_\theta G \Delta_m$. Therefore, we have

$$\begin{aligned} P_{m+1} &= P_m - \mathbb{E}\Big[2 \sigma_\theta \Delta_m^T Q^{-T} Q^{-1} G \Delta_m \Big] + \mathbb{E}\Big[\left\| Q^{-1} v_m \right\|_2^2 \Big] \\ &= P_m - \mathbb{E} 2 \sigma_\theta \Big[\Delta_m^T Q^{-T} \Lambda Q^{-1} \Delta_m \Big] + \mathbb{E}\Big[\left\| Q^{-1} v_m \right\|_2^2 \Big] \\ &\le P_m - 2 \sigma_\theta \lambda_{\min} \mathbb{E}\Big[\left\| Q^{-1} \Delta_m \right\|^2 \Big] + \mathbb{E}\Big[\left\| Q^{-1} v_m \right\|_2^2 \Big] \\ &= (1 - 2 \sigma_\theta \lambda_{\min}) P_m + \mathbb{E}\left[\left\| Q^{-1} v_m \right\|_2^2 \right], \end{aligned} \tag{56}$$

where the inequality used $\lambda_{\min} \triangleq \lambda_{\min}(\Lambda) = \lambda_{\min}(G) > 0$, which is true under our choice of $\beta = \sigma_w / \sigma_\theta$ in Section A.1. Next, we bound the last term of Eqn. (56):

$$\begin{aligned} &\mathbb{E}\Big[\left\| Q^{-1} v_m \right\|_2^2 \Big] \\ &= \mathbb{E}\Bigg[\Big\| Q^{-1} \Big(\frac{\sigma_\theta}{n} \sum_{t=1}^{n} G_t \Delta_{\phi_t^m} + \sigma_\theta G_{t_m} \Big(\Delta_m - \Delta_{\phi_{t_m}^m} \Big) \Big) \Big\|^2 \Bigg] \\ &\le 2 \sigma_\theta^2 \mathbb{E}\left[\left\| Q^{-1} G_{t_m} \Delta_m \right\|_2^2 \right] \\ &\quad + 2 \sigma_\theta^2 \mathbb{E}\Bigg[\Big\| Q^{-1} \Big(\frac{1}{n} \sum_{t=1}^{n} G_t \Delta_{\phi_t^m} - G_{t_m} \Delta_{\phi_{t_m}^m} \Big) \Big\|^2 \Bigg] \\ &\le 2 \sigma_\theta^2 \mathbb{E}\Big[\left\| Q^{-1} G_{t_m} \Delta_m \right\|_2^2 \Big] + 2 \sigma_\theta^2 \mathbb{E}\Big[\| Q^{-1} G_{t_m} \Delta_{\phi_{t_m}^m} \|^2 \Big] \\ &= 2 \sigma_\theta^2 \mathbb{E}\Big[\left\| Q^{-1} G_{t_m} \Delta_m \right\|_2^2 \Big] + 2 \sigma_\theta^2 Q_m, \end{aligned}$$

where the first inequality uses $\|a + b\|_2^2 \le 2 \|a\|_2^2 + 2 \|b\|_2^2$, and the second inequality holds because for any random variable ξ, $\mathbb{E} \|\xi - \mathbb{E}[\xi]\|_2^2 = \mathbb{E} \|\xi\|_2^2 - \|\mathbb{E}\xi\|_2^2 \le \mathbb{E} \|\xi\|_2^2$. Using similar arguments as in (49), we have

$$\mathbb{E}\Big[\left\| Q^{-1} G_{t_m} \Delta_m \right\|_2^2 \Big] \le \kappa^2(Q) L_G^2 P_m, \tag{57}$$

Therefore, we have

$$P_{m+1} \le \left(1 - 2 \sigma_\theta \lambda_{\min} + 2 \sigma_\theta^2 \kappa^2(Q) L_G^2 \right) P_m + 2 \sigma_\theta^2 Q_m. \tag{58}$$

The inequality (58) shows that the dynamics of P_m depends on both P_m itself and Q_m. So we need to find another iterative relation for P_m and Q_m. To this end, we have

$$\begin{aligned} Q_{m+1} &= \mathbb{E}\left[\frac{1}{n} \sum_{t=1}^{n} \left\| Q^{-1} G_t \Delta_{\phi_t^{m+1}} \right\|_2^2 \right] \\ &= \mathbb{E}\Big[\frac{1}{n} \| Q^{-1} G_{t_m} \Delta_{\phi_{t_m}^{m+1}} \|^2 + \frac{1}{n} \sum_{t \ne t_m} \| Q^{-1} G_t \Delta_{\phi_t^{m+1}} \|^2 \Big] \\ &\overset{(a)}{=} \mathbb{E}\Big[\frac{1}{n} \| Q^{-1} G_{t_m} \Delta_m \|^2 + \frac{1}{n} \sum_{t \ne t_m} \| Q^{-1} G_t \Delta_{\phi_t^m} \|^2 \Big] \\ &= \mathbb{E}\Big[\frac{1}{n} \| Q^{-1} G_{t_m} \Delta_m \|^2 - \frac{1}{n} \| Q^{-1} G_{t_m} \Delta_{\phi_{t_m}^m} \|^2 + \frac{1}{n} \sum_{t=1}^{n} \| Q^{-1} G_t \Delta_{\phi_t^m} \|^2 \Big] \\ &= \frac{1}{n} \mathbb{E}[\| Q^{-1} G_{t_m} \Delta_m \|^2] - \frac{1}{n} \mathbb{E}[\| Q^{-1} G_{t_m} \Delta_{\phi_{t_m}^m} \|^2] \\ &\quad + \mathbb{E}\Big[\frac{1}{n} \sum_{t=1}^{n} \| Q^{-1} G_t \Delta_{\phi_t^m} \|^2 \Big] \end{aligned}$$

$$= \frac{1}{n}\mathbb{E}[\|Q^{-1}G_{t_m}\Delta_m\|^2] - \frac{1}{n}\mathbb{E}[\|Q^{-1}G_{t_m}\Delta_{\phi_{t_m}^m}\|^2] + \mathbb{E}\left[\|Q^{-1}G_{t_m}\Delta_{\phi_{t_m}^m}\|^2\right]$$
$$= \frac{1}{n}\mathbb{E}[\|Q^{-1}G_{t_m}\Delta_m\|^2] + \frac{n-1}{n}Q_m$$
$$\overset{(b)}{\leq} \frac{\kappa^2(Q)L_G^2}{n}P_m + \frac{n-1}{n}Q_m. \tag{59}$$

where step (a) uses (53) and step (b) uses (57).

To facilitate our convergence analysis on P_m, we construct a new Lyapunov function which is a linear combination of Eqn. (58) and Eqn. (59). Specifically, consider

$$T_m = P_m + \frac{n\sigma_\theta\lambda_{\min}(1-\sigma_\theta\lambda_{\min})}{\kappa^2(Q)L_G^2}Q_m.$$

Now consider the dynamics of T_m. We have

$$\begin{aligned}T_{m+1} &= P_{m+1} + \frac{n\sigma_\theta\lambda_{\min}(1-\sigma_\theta\lambda_{\min})}{\kappa^2(Q)L_G^2}Q_{m+1}\\
&\leq \left(1-2\sigma_\theta\lambda_{\min}+2\sigma_\theta^2\kappa^2(Q)L_G^2\right)P_m + 2\sigma_\theta^2 Q_m\\
&\quad + \frac{n\sigma_\theta\lambda_{\min}(1-\sigma_\theta\lambda_{\min})}{\kappa^2(Q)L_G^2}\left(\frac{\kappa^2(Q)L_G^2}{n}P_m + \frac{n-1}{n}Q_m\right)\\
&= \left(1-\sigma_\theta\lambda_{\min}+2\sigma_\theta^2\kappa^2(Q)L_G^2-\sigma_\theta^2\lambda_{\min}^2\right)P_m\\
&\quad + \frac{2\sigma_\theta^2\kappa^2(Q)L_G^2+(n-1)\sigma_\theta\lambda_{\min}(1-\sigma_\theta\lambda_{\min})}{\kappa^2(Q)L_G^2}Q_m.\end{aligned}$$

Let's define

$$\rho = \sigma_\theta\lambda_{\min} - 2\sigma_\theta^2\kappa^2(Q)L_G^2.$$

The coefficient for P_m in the previous inequality can be upper bounded by $1-\rho$ because $1-\rho-\sigma_\theta^2\lambda_{\min}^2 \leq 1-\rho$. Then we have

$$\begin{aligned}&T_{m+1}\\
&\leq (1-\rho)P_m + \frac{2\sigma_\theta^2\kappa^2(Q)L_G^2+(n-1)\sigma_\theta\lambda_{\min}(1-\sigma_\theta\lambda_{\min})}{\kappa^2(Q)L_G^2}Q_m\\
&= (1-\rho)\left(P_m + \frac{n\sigma_\theta\lambda_{\min}(1-\sigma_\theta\lambda_{\min})}{\kappa^2(Q)L_G^2}Q_m\right)\\
&\quad + \sigma_\theta\frac{2\sigma_\theta\kappa^2(Q)L_G^2+(n\rho-1)\lambda_{\min}(1-\sigma_\theta\lambda_{\min})}{\kappa^2(Q)L_G^2}Q_m\\
&= (1-\rho)T_m\\
&\quad + \sigma_\theta\frac{2\sigma_\theta\kappa^2(Q)L_G^2+(n\rho-1)\lambda_{\min}(1-\sigma_\theta\lambda_{\min})}{\kappa^2(Q)L_G^2}Q_m.\end{aligned} \tag{60}$$

Next we show that with the step size

$$\sigma_\theta = \frac{\lambda_{\min}}{3(\kappa^2(Q)L_G^2+n\lambda_{\min}^2)} \tag{61}$$

(or smaller), the second term on the right-hand side of (60) is non-positive. To see this, we first notice that with this choice of σ_θ, we have

$$\frac{\lambda_{\min}^2}{9(\kappa^2(Q)L_G^2+n\lambda_{\min}^2)} \leq \rho \leq \frac{\lambda_{\min}^2}{3(\kappa^2(Q)L_G^2+n\lambda_{\min}^2)},$$

which implies

$$n\rho - 1 \leq \frac{n\lambda_{\min}^2}{3(\kappa^2(Q)L_G^2+n\lambda_{\min}^2)} - 1 \leq \frac{1}{3} - 1 = -\frac{2}{3}.$$

Then, it holds that

$$\begin{aligned}&2\sigma_\theta\kappa^2(Q)L_G^2 + (n\rho-1)\lambda_{\min}(1-\sigma_\theta\lambda_{\min})\\
\leq& 2\sigma_\theta\kappa^2(Q)L_G^2 - \frac{2}{3}\lambda_{\min}(1-\sigma_\theta\lambda_{\min})\\
=& -\frac{(6n-2)\lambda_{\min}^3}{9(\kappa^2(Q)L_G^2+n\lambda_{\min}^2)} < 0.\end{aligned}$$

Therefore (60) implies

$$T_{m+1} \leq (1-\rho)T_m.$$

Notice that $P_m \leq T_m$ and $Q_0 = P_0$. Therefore we have $T_0 \leq 2P_0$ and

$$P_m \leq 2(1-\rho)^m P_0.$$

Using (61), we have

$$\rho = \sigma_\theta\lambda_{\min}(G) - 2\sigma_\theta^2\kappa^2(Q)L_G^2 \geq \frac{\lambda_{\min}^2}{9\left(\kappa^2(Q)L_G^2+n\lambda_{\min}^2\right)}.$$

To achieve $P_m \leq \epsilon$, we need at most

$$m = O\left(\left(n + \frac{\kappa^2(Q)L_G^2}{\lambda_{\min}^2}\right)\log\left(\frac{P_0}{\epsilon}\right)\right)$$

iterations. Substituting (37) and (33) in the above bound, we get the desired iteration complexity

$$O\left(\left(n + \frac{\kappa(\widehat{C})L_G^2}{\lambda_{\min}^2(\rho I + \widehat{A}^T\widehat{C}^{-1}\widehat{A})}\right)\log\left(\frac{P_0}{\epsilon}\right)\right).$$

Finally, using the bounds in (33) and (37), we can replace the step size in (61) by

$$\sigma_\theta = \frac{\mu_\rho}{3\left(8\kappa^2(\widehat{C})L_G^2+n\mu_\rho^2\right)},$$

where $\mu_\rho = \lambda_{\min}^2(\rho I + \widehat{A}^T\widehat{C}^{-1}\widehat{A})$ as defined in (14).