A. Constrained Optimization and ASPG

Consider the constrained optimization problem $\min_{x\in\mathcal{C}} f(x)$, where $\mathcal{C} \subset \mathbb{R}^d$ is a compact set with a finite diameter $\mathcal{D}_\mathcal{C} := \sup_{x,y\in\mathcal{C}} \|x-y\|_2$ and $G_M := \max_{x\in\mathcal{C}} \|\nabla f(x)\|$. The accelerated stochastic projected gradient method (ASPG) consists of the iterations

$$\tilde{x}_{k+1} = \mathcal{P}_\mathcal{C}\left(\tilde{y}_k - \alpha(\nabla f(\tilde{y}_k) + \varepsilon_{k+1})\right), \tag{30}$$

$$\tilde{y}_k = (1+\beta)\tilde{x}_k - \beta\tilde{x}_{k-1}, \tag{31}$$

where ε_k is the random gradient error satisfying Assumption 2, $\alpha, \beta > 0$ are the stepsize and momentum parameter and $\mathcal{P}_\mathcal{C}(x)$ denotes the projection of a point x to the compact set $\mathcal{C}$. For constrained problems, algorithms based on projection steps that restricts the iterates to the constraint set are more natural compared to the standard AG algorithm primarily designed for the unconstrained optimization (Bubeck, 2014). Accelerated projected gradient methods can also be viewed as a special case of the accelerated proximal gradient methods as the proximal operator reduces to a projection in a special case (see e.g. Parikh et al. (2014)).

We will show in Proposition 28 that the metric d_ψ implies the standard p-Wasserstein metric in the sense that for any two probability measures μ_1, μ_2 on the product space $\mathcal{C}^2 := \mathcal{C}\times\mathcal{C}$,

$$\mathcal{W}_p(\mu_1,\mu_2) \le 2^{1/p}\mathcal{D}_{\mathcal{C}^2}\|\mu_1-\mu_2\|_{TV}^{1/p} \le \mathcal{D}_{\mathcal{C}^2} d_\psi^{1/p}(\mu_1,\mu_2),$$

where $\mathcal{D}_{\mathcal{C}^2} = \sqrt{2}D_C$ is the diameter of $\mathcal{C}^2$.

Under Assumption 2, $\tilde{\xi}_k = (\tilde{x}_k^T, \tilde{x}_{k-1}^T)^T$ forms a time-homogeneous Markov chain and we assume $\tilde{\xi}_0 \in \mathcal{C}^2$. In addition to Assumption 2, we also assume that the random gradient error ε_k admits a continuous density so that conditional on $\tilde{\xi}_k = (\tilde{x}_k^T, \tilde{x}_{k-1}^T)^T$, $\tilde{x}_{k+1}$ also admits a continuous density, i.e.

$$\mathbb{P}(\tilde{x}_{k+1} \in d\tilde{x} | \tilde{\xi}_k = \tilde{\xi}) = \tilde{p}(\tilde{\xi},\tilde{x})d\tilde{x},$$

where $\tilde{p}(\tilde{\xi},\tilde{x}) > 0$ is continuous in both $\tilde{\xi}$ and $\tilde{x}$.

For the ASPG method with any given α, β so that $\rho_{\alpha,\beta}$, $P_{\alpha,\beta}$ satisfy the LMI inequality (6), the next result gives a bound of k-th iterate to stationary distribution in the weighted total variation distance and standard p-Wasserstein distance, and also a bound on the expected suboptimality $\mathbb{E}[f(\tilde{x}_k)] - f(\tilde{x}_*)$ after k iterations.

Theorem 16. *Given any $\eta \in (0,1)$ and $R > 0$ so that*

$$\inf_{\tilde{x}\in\mathcal{C}:\tilde{\xi}\in\mathcal{C}^2, V_{P_{\alpha,\beta}}(\tilde{\xi})\le R} \frac{\tilde{p}(\tilde{\xi},\tilde{x})}{\tilde{p}(\tilde{\xi}_*,\tilde{x})} \ge \eta.$$

Consider the Markov chain generated by the iterates $\tilde{\xi}_k^T = (\tilde{x}_k^T, \tilde{x}_{k-1}^T)$ of the ASPG algorithm. Then the distribution $\tilde{\nu}_{k,\alpha,\beta}$ of $\tilde{\xi}_k$ converges linearly to a unique invariant distribution $\tilde{\pi}_{\alpha,\beta}$ satisfying

$$\mathcal{W}_p(\tilde{\nu}_{k,\alpha,\beta},\tilde{\pi}_{\alpha,\beta}) \le \mathcal{D}_{\mathcal{C}^2} d_{\tilde{\psi}}^{1/p}(\tilde{\nu}_{k,\alpha,\beta},\tilde{\pi}_{\alpha,\beta}) \le (1-\tilde{\eta})^k \mathcal{D}_{\mathcal{C}^2} d_{\tilde{\psi}}^{1/p}(\tilde{\nu}_{0,\alpha,\beta},\tilde{\pi}_{\alpha,\beta}), \tag{32}$$

where $\mathcal{W}_p$ is the standard p-Wasserstein metric ($p \ge 1$) and

$$\mathbb{E}[f(\tilde{x}_k)] - f(\tilde{x}_*) \le V_{P_{\alpha,\beta}}(\tilde{\xi}_0)\rho_{\alpha,\beta}^k + \frac{\tilde{K}_{\alpha,\beta}}{1-\rho_{\alpha,\beta}}, \tag{33}$$

where $\tilde{K}_{\alpha,\beta} := \alpha\sigma\left((\alpha\sigma + 2\mathcal{D}_\mathcal{C})\|P_{\alpha,\beta}\| + G_M + \frac{\alpha\sigma L}{2}\right)$, $\tilde{\eta} := \min\left\{\frac{\eta}{2}, \left(\frac{1}{2} - \frac{\rho_{\alpha,\beta}}{2} - \frac{\tilde{K}_{\alpha,\beta}}{R}\right)\frac{R\eta}{4\tilde{K}_{\alpha,\beta}+R\eta}\right\}$ and $\tilde{\psi} := \frac{\eta}{2\tilde{K}_{\alpha,\beta}}$.

We can see from (33) that the expected value of the objective with respect to the k-th iterate is close to the true minimum of the objective if k is large, and the stepsize α or the variance of the noise σ^2 is small. By choosing $(\alpha,\beta) = (\alpha_{AG}, \beta_{AG})$, we obtain the optimal convergence in the next theorem.

Proposition 17. *Given $(\alpha, \beta) = (\alpha_{AG}, \beta_{AG})$. Define R as in Theorem 16 with $\eta = 1/\kappa^{1/2}$. Also assume that the noise has small variance, i.e.*

$$\sigma^2 < \frac{1}{4a_1^2}\left(-b_1 + \sqrt{b_1^2 + (a_1 R/\sqrt{\kappa})}\right)^2,$$

where $a_1 := \frac{1}{L^2}\left(\frac{\mu}{2}((1-\sqrt{\kappa})^2+\kappa) + \frac{L}{2}\right)$ and $b_1 := \frac{1}{L}\left(\mathcal{D}_{\mathcal{C}}\mu((1-\sqrt{\kappa})^2+\kappa) + G_M\right)$. Then, we have

$$\mathcal{W}_p(\tilde{\nu}_{k,\alpha,\beta}, \tilde{\pi}_{\alpha,\beta}) \le \mathcal{D}_{\mathcal{C}^2} d_{\tilde{\psi}}^{1/p}(\tilde{\nu}_{k,\alpha,\beta}, \tilde{\pi}_{\alpha,\beta}) \le \left(1 - \frac{1}{8\sqrt{\kappa}}\right)^k \mathcal{D}_{\mathcal{C}^2} d_{\tilde{\psi}}^{1/p}(\tilde{\nu}_{0,\alpha,\beta}, \tilde{\pi}_{\alpha,\beta}), \tag{34}$$

where $\mathcal{W}_p$ is the standard p-Wasserstein metric ($p \ge 1$) and

$$\mathbb{E}[f(\tilde{x}_k)] - f(\tilde{x}_*) \le V_{P_{AG}}(\tilde{\xi}_0)\left(1 - \frac{1}{\sqrt{\kappa}}\right)^k + \sqrt{\kappa}\tilde{K}, \tag{35}$$

where $\tilde{K} := \frac{2\sigma\mathcal{D}_{\mathcal{C}}L+\sigma^2}{2L^2}\mu((1-\sqrt{\kappa})^2+\kappa) + \frac{\sigma G_M}{L} + \frac{\sigma^2}{2L}$ and $\tilde{\psi} := \frac{1}{2\sqrt{\kappa}\tilde{K}}$.

B. Weakly Convex Constrained Optimization

In this section, we extend the constrained optimization for the accelerated stochastic projected gradient method (ASPG) from the strongly convex objectives studied in Section A to the (weakly) convex objectives.

Consider the constrained optimization problem $\min_{x\in\mathcal{C}} f(x)$ for $f \in \mathcal{S}_{0,L}$ on the convex compact domain $\mathcal{C} \subseteq \mathbb{R}^d$ with diameter $\mathcal{D}_{\mathcal{C}}$. Consider the following (regularized) function

$$f_\varepsilon(x) = f(x) + \frac{\varepsilon}{2\mathcal{D}_{\mathcal{C}}^2}\|x\|^2,$$

which is strongly convex with parameter $\mu_\varepsilon = \varepsilon/\mathcal{D}_{\mathcal{C}}^2$ and smooth with parameter $L_\varepsilon = L + \varepsilon/\mathcal{D}_{\mathcal{C}}^2$, i.e. $f_\varepsilon \in \mathcal{S}_{\mu_\varepsilon, L_\varepsilon}$ with a condition number $\kappa_\varepsilon := L_\varepsilon/\mu_\varepsilon = 1 + L\mathcal{D}_{\mathcal{C}}^2/\varepsilon$. Let $\tilde{x}_k^\varepsilon$ denote iterates of ASPG defined by f_ε (i.e $f = f_\varepsilon(x)$) in (30) and (31)) with optimal value $\tilde{x}_*^\varepsilon$ and define $\tilde{x}_*$ to be one of the minimizers of $f(x)$ (the optimizer may not be unique). By applying Proposition 17, we can control the expected suboptimality after k iterations as follows:

$$\mathbb{E}[f_\varepsilon(\tilde{x}_k^\varepsilon)] - f_\varepsilon(\tilde{x}_*^\varepsilon) \le V_{P_{AG}^\varepsilon}(\tilde{\xi}_0)\left(1 - \frac{1}{\sqrt{\kappa_\varepsilon}}\right)^k + \sqrt{\kappa_\varepsilon}\tilde{K}_\varepsilon,$$

where

$$\tilde{K}_\varepsilon := \frac{2\sigma\mathcal{D}_{\mathcal{C}}L_\varepsilon + \sigma^2}{2L_\varepsilon^2}\mu_\varepsilon((1-\sqrt{\kappa_\varepsilon})^2 + \kappa_\varepsilon) + \frac{\sigma G_M^\varepsilon}{L_\varepsilon} + \frac{\sigma^2}{2L_\varepsilon}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}[f(\tilde{x}_k^\varepsilon)] - f(\tilde{x}_*) &= \mathbb{E}\left[f_\varepsilon(\tilde{x}_k^\varepsilon)\right] - f_\varepsilon(\tilde{x}_*) + \frac{\varepsilon}{2\mathcal{D}_{\mathcal{C}}^2}\left(\|\tilde{x}_*\|^2 - \mathbb{E}[\|\tilde{x}_k^\varepsilon\|^2]\right) \\
&\le \mathbb{E}\left[f_\varepsilon(\tilde{x}_k^\varepsilon)\right] - f_\varepsilon(\tilde{x}_*^\varepsilon) + \frac{\varepsilon}{2\mathcal{D}_{\mathcal{C}}^2}\left(\|\tilde{x}_*\|^2 - \mathbb{E}[\|\tilde{x}_k^\varepsilon\|^2]\right) \\
&\le V_{P_{AG}^\varepsilon}(\tilde{\xi}_0)\left(1 - \frac{1}{\sqrt{\kappa_\varepsilon}}\right)^k + \sqrt{\kappa_\varepsilon}\tilde{K}_\varepsilon + \frac{\varepsilon}{2},
\end{aligned}$$

where we used the fact that $\tilde{x}_k^\varepsilon, \tilde{x}_* \in \mathcal{C}$. Therefore, if the noise level σ is small enough such that $\sqrt{\kappa_\varepsilon}\tilde{K}_\varepsilon \le \frac{\varepsilon}{2}$ and if

$$k \ge \frac{|\log(\varepsilon) - \log(V_{P_{AG}^\varepsilon}(\tilde{\xi}_0))|}{|\log(1 - \frac{1}{\sqrt{\kappa_\varepsilon}})|} = O\left(\frac{1}{\sqrt{\varepsilon}}\log\left(\frac{1}{\varepsilon}\right)\right),$$

we obtain

$$\mathbb{E}[f(\tilde{x}_k^\varepsilon)] - f(\tilde{x}_*) \le 2\varepsilon. \tag{36}$$

This shows that if the noise is small is enough, it suffices to have

$$O\left(\frac{1}{\sqrt{\varepsilon}}\log\left(\frac{1}{\varepsilon}\right)\right)$$

many iterations to sample an ε-optimal point in expectation.

C. Proofs of Results in Section 3

In this section, we prove the results for Section 3, in which the objective is quadratic: $f(x) = \frac{1}{2}x^TQx + a^Tx + b$ and $f \in \mathcal{S}_{\mu,L}$, which satisfies the inequalities:

$$f(x) - f(y) \geq \nabla f(y)^T(x-y) + \frac{\mu}{2}\|x-y\|^2,$$
$$f(y) - f(x) \geq \nabla f(y)^T(y-x) - \frac{L}{2}\|x-y\|^2,$$

(see e.g. Nesterov (2004)).

C.1. Proofs of Results in Section 3.1

Before we proceed to the proofs of the results in Section 3.1, we first show that the matrix $S_{\alpha,\beta}$ defined in (17) is positive definite so that the weighted 2-Wasserstein metric $\mathcal{W}_{2,S_{\alpha,\beta}}$ given in (1) is well-defined.

Lemma 18. *The matrix $S_{\alpha,\beta} \in \mathbb{R}^{2d\times 2d}$ defined by (17) is positive definite if $\tilde{P}_{\alpha,\beta}(2,2) \neq 0$.*

Proof. For brevity of the notation, we will not explicitly write the dependency of the matrices to α, β and set $P = P_{\alpha,\beta}$ and $\tilde{P} = P_{\alpha,\beta}$ in our discussion. It is known that if $A \in R^{n\times n}$ is a symmetric matrix with eigenvalues $\{\lambda_i\}_{i=1}^m$ and eigenvectors $\{a_i\}_{i=1}^n$, and $B \in \mathbb{R}^{d\times d}$ is a symmetric matrix with eigenvalues $\{\mu_j\}_{j=1}^d$ and eigenvectors $\{b_j\}_{j=1}^n$, the eigenvalues of the Kronecker product $A \otimes B$ are exactly $\lambda_i\mu_j$ with corresponding eigenvectors $a_i \otimes b_j$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, d$. Since $P = \tilde{P} \otimes I_d$ and $\tilde{P}$ is positive-semi definite by assumption, this implies that P is positive semi-definite and in case P has a zero eigenvalue, any eigenvector z of P (corresponding to a zero eigenvalue of P) can be written as

$$z = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \otimes s = \begin{pmatrix} c_1 s \\ c_2 s \end{pmatrix} \in \mathbb{R}^{2d},$$

for some $s \in \mathbb{R}^d$, $s \neq 0$ where $c = [c_1\ c_2]^T$ is an eigenvector of $\tilde{P}$ corresponding to a zero eigenvalue. The symmetric matrix

$$S := P + \hat{Q}, \qquad \text{where} \quad \hat{Q} := \begin{pmatrix} \frac{1}{2}Q & 0_d \\ 0_d & 0_d \end{pmatrix}, \tag{37}$$

is the sum of two positive semi-definite matrices, therefore it is positive semi-definite by the eigenvalue interlacing property of the sum of symmetric matrices (see e.g. Golub & Van Loan (1996)). Thus, it suffices to show that S is non-singular, i.e. it does not have a zero eigenvalue. If $\tilde{P}$ is of full rank, then such a vector z cannot exist and P cannot have a zero eigenvalue. Therefore, P is positive definite and hence S is positive definite which completes the proof.

The remaining case is when $\tilde{P}$ is of rank one ($\tilde{P} = 0$ is excluded as $\tilde{P}_{22} \neq 0$) in which case we can write $\tilde{P} = uu^T$ for some $u = \begin{pmatrix} u_1 & u_2 \end{pmatrix}^T \in \mathbb{R}^{2d}$ and $u_2 \neq 0$. We will prove the claim by contradiction. Assume that there exists a non-zero $v \in \mathbb{R}^{2d}$ such that $Sv = 0$. Then,

$$0 = v^TSv = v^TPv + v^T\hat{Q}v.$$

Since both of the matrices P and $\hat{Q}$ are positive semi-definite, this is true if and only if $v^TPv = 0$ and $v^T\hat{Q}v = 0$. Since $v^T\hat{Q}v = 0$ and Q is positive definite, from the structure of $\hat{Q}$, it follows that the first d entries of v has to be zero, i.e. $v = [0 \quad v_2^T]^T$ for some $v_2 \in \mathbb{R}^d$.

It is easy to see that the eigenvalues of the two by two symmetric rank-one matrix $\tilde{P} = uu^T$ are $\lambda_1 = \|u\|^2 > 0$ and $\lambda_2 = 0$ with corresponding eigenvectors $\begin{pmatrix} u_1 & u_2 \end{pmatrix}^T$ and $\begin{pmatrix} u_2 & -u_1 \end{pmatrix}^T$ respectively. Since v is an eigenvector of P corresponding

to an eigenvalue zero (i.e. $Pv = 0$), then using (C.1) we can write

$$v = \begin{pmatrix} u_2 \\ -u_1 \end{pmatrix} \otimes s = \begin{pmatrix} u_2 s \\ -u_1 s \end{pmatrix} \in \mathbb{R}^{2d},$$

for some $s \in \mathbb{R}^d$, $s \neq 0$. Since $v = [0 \quad v_2^T]^T$ for some $v_2 \in \mathbb{R}^d$, this implies $u_2 = 0$ as $s \neq 0$. This is a contradiction. □

Next, before we proceed to the proofs of the results in Section 3.1, let us first recall that throughout Section 3, the noise ε_k are assumed to be i.i.d. Let us define the coupling

$$x_{k+1}^{(j)} = y_k^{(j)} - \alpha \left[\nabla f \left(y_k^{(j)} \right) + \varepsilon_{k+1} \right], \tag{38}$$

$$y_k^{(j)} = (1+\beta) x_k^{(j)} - \beta x_{k-1}^{(j)}, \tag{39}$$

with $j = 1, 2$. Then, we have

$$\xi_{k+1} = A\xi_k + Bw_k,$$

where $A = \tilde{A} \otimes I_d$, $B = \tilde{B} \otimes I_d$, for

$$\tilde{A} = \begin{pmatrix} 1+\beta & -\beta \\ 1 & 0 \end{pmatrix}, \qquad \tilde{B} = \begin{pmatrix} -\alpha \\ 0 \end{pmatrix},$$

and

$$\xi_k = \left(\left(x_k^{(1)} - x_k^{(2)} \right)^T, \left(x_{k-1}^{(1)} - x_{k-1}^{(2)} \right)^T \right)^T, \tag{40}$$

$$w_k = \nabla f \left((1+\beta) x_k^{(1)} - \beta x_{k-1}^{(1)} \right) - \nabla f \left((1+\beta) x_k^{(2)} - \beta x_{k-1}^{(2)} \right). \tag{41}$$

Let us define:

$$\tilde{X} = \rho \tilde{X}_1 + (1-\rho) \tilde{X}_2, \tag{42}$$

where

$$\tilde{X}_1 = \frac{1}{2} \begin{pmatrix} \beta^2\mu & -\beta^2\mu & -\beta \\ -\beta^2\mu & \beta^2\mu & \beta \\ -\beta & \beta & \alpha(2-L\alpha) \end{pmatrix}, \tag{43}$$

and

$$\tilde{X}_2 = \frac{1}{2} \begin{pmatrix} (1+\beta)^2\mu & -\beta(1+\beta)\mu & -(1+\beta) \\ -\beta(1+\beta)\mu & \beta^2\mu & \beta \\ -(1+\beta) & \beta & \alpha(2-L\alpha) \end{pmatrix}, \tag{44}$$

and $X = \tilde{X} \otimes I_d$, $X_1 = \tilde{X}_1 \otimes I_d$, $X_2 = \tilde{X}_2 \otimes I_d$.

Before we proceed, let us recall the following lemma from Hu & Lessard (2017).

Lemma 19 (Theorem 2 Hu & Lessard (2017))**.** *Let X be a symmetric matrix with $X \in \mathbb{R}^{(n_\varepsilon + n_w) \times (n_\varepsilon + n_w)}$. If there exists a matrix $P \in \mathbb{R}^{n_\varepsilon \times n_\varepsilon}$ with $P \geq 0$ so that*

$$\begin{pmatrix} A^T P A - \rho P & A^T P B \\ B^T P A & B^T P B \end{pmatrix} - X \preceq 0,$$

then, we have

$$V(\xi_{k+1}) - \rho V(\xi_k) \leq S(\xi_k, w_k),$$

where $V(\xi) := \xi^T P \xi$, and

$$S(\xi, w) := \begin{pmatrix} \xi \\ w \end{pmatrix}^T X \begin{pmatrix} \xi \\ w \end{pmatrix},$$

and

$$\xi_{k+1} = A\xi_k + Bw_k.$$

The proof of Theorem 4 relies on the following lemma.

Lemma 20. *Assume the coupling:*

$$x_{k+1}^{(j)} = y_k^{(j)} - \alpha\left[\nabla f\left(y_k^{(j)}\right) + \varepsilon_{k+1}\right], \tag{45}$$

$$y_k^{(j)} = (1+\beta)x_k^{(j)} - \beta x_{k-1}^{(j)}, \tag{46}$$

with $j = 1, 2$. Assume that f is quadratic and $f(x) = \frac{1}{2}x^TQx + a^Tx + b$, where Q is positive definite.

Let $\rho = \rho_{\alpha,\beta} \in (0,1)$ that can depend on α and β so that there exists some $P = P_{\alpha,\beta}$ symmetric and positive semi-definite that can depend on α and β such that

$$\begin{pmatrix} A^TPA - \rho P & A^TPB \\ B^TPA & B^TPB \end{pmatrix} - X \preceq 0, \tag{47}$$

where $X := \tilde{X} \otimes I_d$, where $\tilde{X}$ is defined in (42). Then, we have

$$\mathbb{E}\left[\begin{pmatrix} x_{k+1}^{(1)} - x_{k+1}^{(2)} \\ x_k^{(1)} - x_k^{(2)} \end{pmatrix}^T P_{\alpha,\beta} \begin{pmatrix} x_{k+1}^{(1)} - x_{k+1}^{(2)} \\ x_k^{(1)} - x_k^{(2)} \end{pmatrix} + \frac{1}{2}\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right)^T Q\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right)\right]$$
$$\leq \rho_{\alpha,\beta}\left(\mathbb{E}\left[\begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix}^T P_{\alpha,\beta} \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix} + \frac{1}{2}\left(x_k^{(1)} - x_k^{(2)}\right)^T Q\left(x_k^{(1)} - x_k^{(2)}\right)\right]\right).$$

Proof of Lemma 20. First of all, since f is L-smooth and μ-strongly convex, we have for every $x, y \in \mathbb{R}^d$:

$$f(x) - f(y) \geq \nabla f(y)^T(x-y) + \frac{\mu}{2}\|x-y\|^2, \tag{48}$$

$$f(y) - f(x) \geq \nabla f(y)^T(y-x) - \frac{L}{2}\|y-x\|^2. \tag{49}$$

Note that since f is L-smooth, we also have for every $x, y \in \mathbb{R}^d$:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|.$$

Let us first consider the simpler case $f(x) = \frac{1}{2}x^TQx$. Since f is quadratic, ∇f is linear. Applying (48) and the linearity of ∇f, we get

$$f\left(x_k^{(1)} - x_k^{(2)}\right) - f\left(y_k^{(1)} - y_k^{(2)}\right) \geq \left(\nabla f\left(y_k^{(1)}\right) - \nabla f\left(y_k^{(2)}\right)\right)^T\left(x_k^{(1)} - x_k^{(2)} - \left(y_k^{(1)} - y_k^{(2)}\right)\right)$$
$$+ \frac{\mu}{2}\left\|x_k^{(1)} - x_k^{(2)} - \left(y_k^{(1)} - y_k^{(2)}\right)\right\|^2.$$

Applying (49) and the linearity of ∇f, we get

$$f\left(y_k^{(1)} - y_k^{(2)}\right) - f\left(y_k^{(1)} - y_k^{(2)} - \alpha\nabla f\left(y_k^{(1)} - y_k^{(2)}\right)\right) \geq \frac{\alpha}{2}(2 - L\alpha)\left\|\nabla f\left(y_k^{(1)}\right) - \nabla f\left(y_k^{(2)}\right)\right\|^2.$$

Using the identity:

$$x_{k+1}^{(1)} - x_{k+1}^{(2)} = y_k^{(1)} - y_k^{(2)} - \alpha\nabla f\left(y_k^{(1)} - y_k^{(2)}\right),$$

we get

$$f\left(y_k^{(1)} - y_k^{(2)}\right) - f\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right) \geq \frac{\alpha}{2}(2 - L\alpha)\left\|\nabla f\left(y_k^{(1)}\right) - \nabla f\left(y_k^{(2)}\right)\right\|^2.$$

Hence, we get

$$f\left(x_k^{(1)} - x_k^{(2)}\right) - f\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right)$$
$$\geq \left(\nabla f\left(y_k^{(1)}\right) - \nabla f\left(y_k^{(2)}\right)\right)^T\left(x_k^{(1)} - x_k^{(2)} - \left(y_k^{(1)} - y_k^{(2)}\right)\right)$$
$$+ \frac{\mu}{2}\left\|x_k^{(1)} - x_k^{(2)} - \left(y_k^{(1)} - y_k^{(2)}\right)\right\|^2 + \frac{\alpha}{2}(2 - L\alpha)\left\|\nabla f\left(y_k^{(1)}\right) - \nabla f\left(y_k^{(2)}\right)\right\|^2.$$

By the definition of $\tilde{X}_1$ from (43), with $X_1 = \tilde{X}_1 \otimes I_d$, we get

$$\begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \\ \nabla f(y_k^{(1)}) - \nabla f(y_k^{(2)}) \end{pmatrix}^T X_1 \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \\ \nabla f(y_k^{(1)}) - \nabla f(y_k^{(2)}) \end{pmatrix} \leq f\left(x_k^{(1)} - x_k^{(2)}\right) - f\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right).$$

Similarly, by applying (48) with $(x, y) \mapsto (0, y_k^{(1)} - y_k^{(2)})$, by the definition of $\tilde{X}_2$ from (44), with $X_2 = \tilde{X}_2 \otimes I_d$, we get

$$\begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \\ \nabla f(y_k^{(1)}) - \nabla f(y_k^{(2)}) \end{pmatrix}^T X_2 \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \\ \nabla f(y_k^{(1)}) - \nabla f(y_k^{(2)}) \end{pmatrix} \leq f(0) - f\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right).$$

By using $\tilde{X} = \rho\tilde{X}_1 + (1-\rho)\tilde{X}_2$ and $X = \tilde{X} \otimes I_d$, we get

$$\begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \\ \nabla f(y_k^{(1)}) - \nabla f(y_k^{(2)}) \end{pmatrix}^T X \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \\ \nabla f(y_k^{(1)}) - \nabla f(y_k^{(2)}) \end{pmatrix} \leq -\left(f\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right) - f(0)\right) + \rho\left(f\left(x_k^{(1)} - x_k^{(2)}\right) - f(0)\right).$$

By Lemma 19 and the definition of $\rho_{\alpha,\beta}$, $P_{\alpha,\beta}$ the inequality (47) holds. Thus

$$\begin{pmatrix} x_{k+1}^{(1)} - x_{k+1}^{(2)} \\ x_k^{(1)} - x_k^{(2)} \end{pmatrix}^T P_{\alpha,\beta} \begin{pmatrix} x_{k+1}^{(1)} - x_{k+1}^{(2)} \\ x_k^{(1)} - x_k^{(2)} \end{pmatrix} + f\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right) - f(0)$$
$$\leq \rho_{\alpha,\beta}\left(\begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix}^T P_{\alpha,\beta} \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix} + f\left(x_k^{(1)} - x_k^{(2)}\right) - f(0)\right).$$

Since f is quadratic, and we assumed that $f(x) = \frac{1}{2}x^TQx$, where Q is positive definite, we get

$$\begin{pmatrix} x_{k+1}^{(1)} - x_{k+1}^{(2)} \\ x_k^{(1)} - x_k^{(2)} \end{pmatrix}^T P_{\alpha,\beta} \begin{pmatrix} x_{k+1}^{(1)} - x_{k+1}^{(2)} \\ x_k^{(1)} - x_k^{(2)} \end{pmatrix} + \frac{1}{2}\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right)^T Q\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right)$$
$$\leq \rho_{\alpha,\beta}\left(\begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix}^T P_{\alpha,\beta} \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix} + \frac{1}{2}\left(x_k^{(1)} - x_k^{(2)}\right)^T Q\left(x_k^{(1)} - x_k^{(2)}\right)\right).$$

Previously, we assumed $f(x) = \frac{1}{2}x^TQx$, so that $\nabla f(x-y) = \nabla f(x) - \nabla f(y)$. In general, the quadratic function takes the form

$$f(x) = \frac{1}{2}x^TQx + a^Tx + b.$$

In this case,

$$\nabla f(x-y) - (\nabla f(x) - \nabla f(y)) = a^T(x-y).$$

By the definition of $\tilde{X}_1$ from (43), with $X_1 = \tilde{X}_1 \otimes I_d$, we get

$$\begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \\ \nabla f(y_k^{(1)}) - \nabla f(y_k^{(2)}) \end{pmatrix}^T X_1 \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \\ \nabla f(y_k^{(1)}) - \nabla f(y_k^{(2)}) \end{pmatrix}$$
$$\leq f\left(x_k^{(1)} - x_k^{(2)}\right) - f\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right)$$
$$+ \left(\nabla f\left(y_k^{(1)} - y_k^{(2)}\right) - \nabla f\left(y_k^{(1)}\right) + \nabla f\left(y_k^{(2)}\right)\right)^T \left(x_{k+1}^{(1)} - x_{k+1}^{(2)} - \left(x_k^{(1)} - x_k^{(2)}\right)\right).$$
$$= f\left(x_k^{(1)} - x_k^{(2)}\right) - f\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right) + a^T\left(x_{k+1}^{(1)} - x_{k+1}^{(2)} - \left(x_k^{(1)} - x_k^{(2)}\right)\right).$$

By the definition of $\tilde{X}_2$ from (44), with $X_2 = \tilde{X}_2 \otimes I_d$, we get

$$\begin{aligned}
&\begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \\ \nabla f\left(y_k^{(1)}\right) - \nabla f\left(y_k^{(2)}\right) \end{pmatrix}^T X_2 \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \\ \nabla f\left(y_k^{(1)}\right) - \nabla f\left(y_k^{(2)}\right) \end{pmatrix} \\
&\leq f(0) - f\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right) + \left(\nabla f\left(y_k^{(1)} - y_k^{(2)}\right) - \nabla f\left(y_k^{(1)}\right) + \nabla f\left(y_k^{(2)}\right)\right)^T \left(x_{k+1}^{(1)} - x_{k+2}^{(2)}\right) \\
&= f(0) - f\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right) + a^T \left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right).
\end{aligned}$$

Using $\tilde{X} = \rho\tilde{X}_1 + (1-\rho)\tilde{X}_2$ and $X = \tilde{X} \otimes I_d$, we get

$$\begin{aligned}
&\begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \\ \nabla f(y_k^{(1)}) - \nabla f(y_k^{(2)}) \end{pmatrix}^T X \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \\ \nabla f(y_k^{(1)}) - \nabla f(y_k^{(2)}) \end{pmatrix} \\
&\leq -\left(f\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right) - f(0)\right) + \rho\left(f\left(x_k^{(1)} - x_k^{(2)}\right) - f(0)\right) \\
&\qquad + a^T\left(x_{k+1}^{(1)} - x_{k+1}^{(2)} - \rho\left(x_k^{(1)} - x_k^{(2)}\right)\right) \\
&= -\frac{1}{2}\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right)^T Q \left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right) + \rho\frac{1}{2}\left(x_k^{(1)} - x_k^{(2)}\right) Q \left(x_k^{(1)} - x_k^{(2)}\right).
\end{aligned}$$

Hence, by Lemma 19 and the definition of $\rho_{\alpha,\beta}$, $P_{\alpha,\beta}$ so that (47) holds, we get the same result as before:

$$\begin{aligned}
&\begin{pmatrix} x_{k+1}^{(1)} - x_{k+1}^{(2)} \\ x_k^{(1)} - x_k^{(2)} \end{pmatrix}^T P_{\alpha,\beta} \begin{pmatrix} x_{k+1}^{(1)} - x_{k+1}^{(2)} \\ x_k^{(1)} - x_k^{(2)} \end{pmatrix} + \frac{1}{2}\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right)^T Q \left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right) \\
&\leq \rho_{\alpha,\beta}\left(\begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix}^T P_{\alpha,\beta} \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix} + \frac{1}{2}\left(x_k^{(1)} - x_k^{(2)}\right)^T Q \left(x_k^{(1)} - x_k^{(2)}\right)\right).
\end{aligned}$$

□

By taking $\alpha = \alpha_{AG}$, $\beta = \beta_{AG}$, $\rho = \rho_{AG}$ and P_{AG} in definition (7), we recall the following result from Hu & Lessard (2017).

Lemma 21 (Hu & Lessard (2017))**.** *, With the choice*

$$\alpha = \alpha_{AG} = \frac{1}{L}, \qquad \beta = \beta_{AG} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}, \qquad \rho = \rho_{AG} = 1 - \frac{1}{\sqrt{\kappa}},$$

where $\kappa = L/\mu$ is the condition number, there exists a matrix $\tilde{P}_{AG} \in \mathbb{R}^{2\times 2}$ with $\tilde{P}_{AG} \geq 0$, where

$$\tilde{P}_{AG} := \tilde{u}\tilde{u}^T, \quad \tilde{u} = \begin{pmatrix} \sqrt{\frac{L}{2}} & \sqrt{\frac{\mu}{2}} - \sqrt{\frac{L}{2}} \end{pmatrix}^T,$$

such that $P_{AG} = \tilde{P}_{AG} \otimes I_d$ and

$$\begin{pmatrix} A^T P_{AG} A - \rho P_{AG} & A^T P_{AG} B \\ B^T P_{AG} A & B^T P_{AG} B \end{pmatrix} - X \preceq 0,$$

where $X := \tilde{X} \otimes I_d$, where $\tilde{X}$ is defined in (42).

We immediately obtain the following result.

Lemma 22. *Assume the coupling (45)-(46). Assume that f is quadratic and $f(x) = \frac{1}{2}x^TQx + a^Tx + b$, where Q is positive definite. Then, we have*

$$\mathbb{E}\left[\begin{pmatrix} x_{k+1}^{(1)} - x_{k+1}^{(2)} \\ x_k^{(1)} - x_k^{(2)} \end{pmatrix}^T P_{AG} \begin{pmatrix} x_{k+1}^{(1)} - x_{k+1}^{(2)} \\ x_k^{(1)} - x_k^{(2)} \end{pmatrix} + \frac{1}{2}\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right)^T Q \left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right)\right]$$
$$\leq \rho_{AG}\left(\mathbb{E}\left[\begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix}^T P_{AG} \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix} + \frac{1}{2}\left(x_k^{(1)} - x_k^{(2)}\right)^T Q \left(x_k^{(1)} - x_k^{(2)}\right)\right]\right),$$

where P is defined in (7).

Now, we are ready to state the proof of Theorem 4.

Proof of Theorem 4. Recall the iterates $\xi_k = (x_k^T, x_{k-1}^T)^T$, the Markov kernel $\mathcal{P}_{\alpha,\beta}$ and the definition of the weighted 2-Wasserstein distance (1) with the weighted norm (16)-(17) and $P = P_{\alpha,\beta}$. Then showing Theorem 4 is equivalent to show

$$\mathcal{W}^2_{2,S_{\alpha,\beta}}(R^k_{\alpha,\beta}((x_0, x_{-1}), \cdot), \pi_{\alpha,\beta}) \tag{50}$$
$$\leq \rho^k_{\alpha,\beta} \int_{\mathbb{R}^d \times \mathbb{R}^d} \left[\begin{pmatrix} x_0 - \hat{x}_0 \\ x_{-1} - \hat{x}_{-1} \end{pmatrix}^T P_{\alpha,\beta} \begin{pmatrix} x_0 - \hat{x}_0 \\ x_{-1} - \hat{x}_{-1} \end{pmatrix} + \frac{1}{2}(x_0 - \hat{x}_0)^T Q (x_0 - \hat{x}_0)\right] d\pi_{\alpha,\beta}(\hat{x}_0, \hat{x}_{-1}).$$

Let $(((x_k^{(i)})^T, (x_{k-1}^{(i)})^T)^T)_{k=0}^\infty$, $i = 1, 2$ be a coupling of $((x_k^T, x_{k-1}^T)^T)_{k=0}^\infty$ defined as before. We have shown before that for every k,

$$\begin{pmatrix} x_{k+1}^{(1)} - x_{k+1}^{(2)} \\ x_k^{(1)} - x_k^{(2)} \end{pmatrix}^T P_{\alpha,\beta} \begin{pmatrix} x_{k+1}^{(1)} - x_{k+1}^{(2)} \\ x_k^{(1)} - x_k^{(2)} \end{pmatrix} + \frac{1}{2}\left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right)^T Q \left(x_{k+1}^{(1)} - x_{k+1}^{(2)}\right)$$
$$\leq \rho_{\alpha,\beta}\left[\begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix}^T P_{\alpha,\beta} \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix} + \frac{1}{2}\left(x_k^{(1)} - x_k^{(2)}\right)^T Q \left(x_k^{(1)} - x_k^{(2)}\right)\right].$$

Using induction on k, we get

$$\begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix}^T P_{\alpha,\beta} \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix} + \frac{1}{2}\left(x_k^{(1)} - x_k^{(2)}\right)^T Q \left(x_k^{(1)} - x_k^{(2)}\right)$$
$$\leq \rho^k_{\alpha,\beta}\left[\begin{pmatrix} x_0^{(1)} - x_0^{(2)} \\ x_{-1}^{(1)} - x_{-1}^{(2)} \end{pmatrix}^T P_{\alpha,\beta} \begin{pmatrix} x_0^{(1)} - x_0^{(2)} \\ x_{-1}^{(1)} - x_{-1}^{(2)} \end{pmatrix} + \frac{1}{2}\left(x_0^{(1)} - x_0^{(2)}\right)^T Q \left(x_0^{(1)} - x_0^{(2)}\right)\right].$$

By taking expectation and since $\frac{1}{2}x^TQx \geq 0$ for any x, we get

$$\mathbb{E}\left[\begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix}^T P_{\alpha,\beta} \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix}\right]$$
$$\leq \rho^k_{\alpha,\beta}\mathbb{E}\left[\begin{pmatrix} x_0^{(1)} - x_0^{(2)} \\ x_{-1}^{(1)} - x_{-1}^{(2)} \end{pmatrix}^T P_{\alpha,\beta} \begin{pmatrix} x_0^{(1)} - x_0^{(2)} \\ x_{-1}^{(1)} - x_{-1}^{(2)} \end{pmatrix} + \frac{1}{2}\left(x_0^{(1)} - x_0^{(2)}\right)^T Q \left(x_0^{(1)} - x_0^{(2)}\right)\right].$$

Let $\lambda_1, \lambda_2 \in \mathcal{P}_{2,S_{\alpha,\beta}}(\mathbb{R}^{2d})$. There exist a couple of random vectors $(x_0^{(1)}, x_{-1}^{(1)})$, and $(x_0^{(2)}, x_{-1}^{(2)})$, independent of $(\varepsilon_k)_{k=0}^\infty$ such that

$$\mathcal{W}^2_{2,S_{\alpha,\beta}}(\lambda_1, \lambda_2) = \mathbb{E}\left[\begin{pmatrix} x_0^{(1)} - x_0^{(2)} \\ x_{-1}^{(1)} - x_{-1}^{(2)} \end{pmatrix}^T P_{\alpha,\beta} \begin{pmatrix} x_0^{(1)} - x_0^{(2)} \\ x_{-1}^{(1)} - x_{-1}^{(2)} \end{pmatrix} + \frac{1}{2}\left(x_0^{(1)} - x_0^{(2)}\right)^T Q \left(x_0^{(1)} - x_0^{(2)}\right)\right].$$

Then, we get

$$\mathcal{W}^2_{2,S_{\alpha,\beta}}\left(\mathcal{P}^k_{\alpha,\beta}\lambda_1, \mathcal{P}^k_{\alpha,\beta}\lambda_2\right) \leq \rho^k_{\alpha,\beta} I^2(\lambda_1,\lambda_2),$$

where

$$I^2(\lambda_1,\lambda_2) = \mathbb{E}_{(x_0^{(j)},x_{-1}^{(j)})\sim\lambda_j, j=1,2}\left[\begin{pmatrix} x_0^{(1)}-x_0^{(2)} \\ x_{-1}^{(1)}-x_{-1}^{(2)}\end{pmatrix}^T P_{\alpha,\beta}\begin{pmatrix} x_0^{(1)}-x_0^{(2)} \\ x_{-1}^{(1)}-x_{-1}^{(2)}\end{pmatrix} + \frac{1}{2}\left(x_0^{(1)}-x_0^{(2)}\right)^T Q\left(x_0^{(1)}-x_0^{(2)}\right)\right].$$

Therefore,

$$\sum_{k=1}^{\infty}\mathcal{W}^2_{2,S_{\alpha,\beta}}\left(\mathcal{P}^k_{\alpha,\beta}\lambda_1, \mathcal{P}^k_{\alpha,\beta}\lambda_2\right) < \infty.$$

By taking $\lambda_2 = \mathcal{P}_{\alpha,\beta}\lambda_1$, we get

$$\sum_{k=1}^{\infty}\mathcal{W}^2_{2,S_{\alpha,\beta}}\left(\mathcal{P}^k_{\alpha,\beta}\lambda_1, \mathcal{P}^{k+1}_{\alpha,\beta}\lambda_1\right) < \infty.$$

Hence $\mathcal{P}^k_{\alpha,\beta}\lambda_1$ is a Cauchy sequence and converges to a limit $\pi^{\lambda_1}_{\alpha,\beta}$:

$$\lim_{k\to\infty}\mathcal{W}_{2,S_{\alpha,\beta}}\left(\mathcal{P}^k_{\alpha,\beta}\lambda_1, \pi^{\lambda_1}_{\alpha,\beta}\right) = 0.$$

Next, let us show that $\pi^{\lambda_1}_{\alpha,\beta}$ does not depend on λ_1. Assume that there exists $\pi^{\lambda_2}_{\alpha,\beta}$ so that $\lim_{k\to\infty}\mathcal{W}_{2,S_{\alpha,\beta}}(\mathcal{P}^k_{\alpha,\beta}\lambda_2, \pi^{\lambda_2}_{\alpha,\beta}) = 0$. Since $\mathcal{W}_{2,S_{\alpha,\beta}}$ is a metric, by the triangle inequality,

$$\mathcal{W}_{2,S_{\alpha,\beta}}\left(\pi^{\lambda_1}_{\alpha,\beta}, \pi^{\lambda_2}_{\alpha,\beta}\right) \leq \mathcal{W}_{2,S_{\alpha,\beta}}\left(\pi^{\lambda_1}_{\alpha,\beta}, \mathcal{P}^k_{\alpha,\beta}\lambda_1\right) + \mathcal{W}_{2,S_{\alpha,\beta}}\left(\mathcal{P}^k_{\alpha,\beta}\lambda_1, \mathcal{P}^k_{\alpha,\beta}\lambda_2\right) + \mathcal{W}_{2,S_{\alpha,\beta}}\left(\pi^{\lambda_2}_{\alpha,\beta}, \mathcal{P}^k_{\alpha,\beta}\lambda_2\right),$$

which goes to zero as $k\to\infty$. Hence, $\pi^{\lambda_1}_{\alpha,\beta} = \pi^{\lambda_2}_{\alpha,\beta}$. The limit is therefore the same for any initial distributions and we can denote it by $\pi_{\alpha,\beta}$. Indeed,

$$\mathcal{W}_{2,S_{\alpha,\beta}}\left(\mathcal{P}_{\alpha,\beta}\pi_{\alpha,\beta}, \pi_{\alpha,\beta}\right) \leq \mathcal{W}_{2,S_{\alpha,\beta}}\left(\mathcal{P}_{\alpha,\beta}\pi_{\alpha,\beta}, \mathcal{P}^k_{\alpha,\beta}\pi_{\alpha,\beta}\right) + \mathcal{W}_{2,S_{\alpha,\beta}}\left(\mathcal{P}^k_{\alpha,\beta}\pi_{\alpha,\beta}, \pi_{\alpha,\beta}\right),$$

which goes to zero as $k\to\infty$. Hence $\mathcal{P}_{\alpha,\beta}\pi_{\alpha,\beta} = \pi_{\alpha,\beta}$ gives the invariant distribution. We can also show similarly as before that it is unique. □

Remark 23. *If $\alpha \in (0, 1/L]$ and $\beta = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$, then we can take the matrix $P_{\alpha,\beta}$ appearing in Theorem 4 according to the P_α matrix defined in Aybat et al. (2019, Theorem 2.3) to obtain $\rho(\alpha,\beta) = 1-\sqrt{\alpha\mu}$. For $\alpha = \frac{\log^2(k)}{\mu k^2}$, then this leads to $\mathcal{W}_{2,S_{\alpha,\beta}}\left(\nu_{k,\alpha,\beta}, \pi_{\alpha,\beta}\right) \leq \frac{1}{k}\mathcal{W}_{2,S_{\alpha,\beta}}(\nu_{0,\alpha,\beta}, \pi_{\alpha,\beta})$ and it can be shown with an analysis similar to that of Aybat et al. (2019) that the second moment of $\pi_{\alpha,\beta}$ is also $O(1/k)$; ignoring some logarithmic factors in k. Therefore, our results do not violate (and are in agreement with) the $\Omega(1/k)$ lower bounds studied in Chatterjee et al. (2016); Raginsky & Rakhlin (2011); Agarwal et al. (2009) for strongly convex stochastic optimization.*

Proof of Theorem 5. First let us recall the AG method:

$$\begin{aligned} x_{k+1} &= y_k - \alpha[\nabla f(y_k)], \\ y_k &= (1+\beta)x_k - \beta x_{k-1}, \end{aligned}$$

where $\alpha > 0$ is the step size and β is the momentum parameter. In the case when f is quadratic and $f(x) = \frac{1}{2}x^TQx + a^Tx + b$, we can compute that

$$\begin{aligned} x_{k+1} &= y_k - \alpha[Qy_k + a], \\ y_k &= (1+\beta)x_k - \beta x_{k-1}, \end{aligned}$$

and with the optimizer x_* we get

$$\begin{aligned} x_{k+1} - x_* &= y_k - x_* - \alpha[Q(y_k - x_*)], \\ y_k - y_* &= (1+\beta)(x_k - x_*) - \beta(x_{k-1} - x_*), \end{aligned}$$

which implies that

$$\begin{pmatrix} x_{k+1}-x_* \\ x_k - x_* \end{pmatrix} = \begin{pmatrix} (1+\beta)(I_d-\alpha Q) & -\beta(I_d-\alpha Q) \\ I_d & 0_d \end{pmatrix} \begin{pmatrix} x_k - x_* \\ x_{k-1}-x_* \end{pmatrix},$$

which yields that

$$\begin{pmatrix} x_k - x_* \\ x_{k-1}-x_* \end{pmatrix} = \begin{pmatrix} (1+\beta)(I_d-\alpha Q) & -\beta(I_d-\alpha Q) \\ I_d & 0_d \end{pmatrix}^k \begin{pmatrix} x_0 - x_* \\ x_{-1}-x_* \end{pmatrix},$$

and we aim to provide an upper bound to the 2-norm of the matrix, that is:

$$\left\| \begin{pmatrix} (1+\beta)(I_d-\alpha Q) & -\beta(I_d-\alpha Q) \\ I_d & 0_d \end{pmatrix}^k \right\|.$$

Let us assume that Q has the decomposition

$$Q = VDV^T,$$

where D is diagonal consisting of eigenvalues λ_i, $1 \leq i \leq d$ in increasing order:

$$\mu = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d = L,$$

then we have

$$I_d - \alpha Q = V\tilde{D}V^T,$$

where $\tilde{D} = I_d - \alpha D$ is diagonal matrix with entries

$$1 - \alpha\lambda_i, \qquad 1 \leq i \leq d.$$

Therefore, the matrix

$$\begin{pmatrix} (1+\beta)(I_d-\alpha Q) & -\beta(I_d-\alpha Q) \\ I_d & 0_d \end{pmatrix}$$

has the same eigenvalues as the matrix

$$\begin{pmatrix} (1+\beta)(I_d-\alpha D) & -\beta(I_d-\alpha D) \\ I_d & 0_d \end{pmatrix},$$

which has the same eigenvalues as the matrix:

$$\begin{pmatrix} T_1 & \cdots & 0 & 0 \\ 0 & T_2 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & T_d \end{pmatrix},$$

where

$$T_i = \begin{pmatrix} (1+\beta)(1-\alpha\lambda_i) & -\beta(1-\alpha\lambda_i) \\ 1 & 0 \end{pmatrix}, \qquad 1 \leq i \leq d,$$

are 2×2 matrices with eigenvalues:

$$\mu_{i,\pm} = \frac{(1+\beta)(1-\alpha\lambda_i) \pm \sqrt{(1+\beta)^2(1-\alpha\lambda_i)^2 - 4\beta(1-\alpha\lambda_i)}}{2},$$

where $1 \leq i \leq d$, and therefore

$$\left\| \begin{pmatrix} (1+\beta)(I_d-\alpha Q) & -\beta(I_d-\alpha Q) \\ I_d & 0_d \end{pmatrix}^k \right\| \leq \max_{1\leq i\leq d} \left\|T_i^k\right\|. \tag{51}$$

Next, we upper bound $\|T_i^k\|$. We recall the choice:

$$\alpha = \frac{4}{3L+\mu}, \qquad \beta = \frac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}+2}, \qquad \rho = 1 - \frac{2}{\sqrt{3\kappa+1}}. \tag{52}$$

We can compute that

$$\Delta_i := (1+\beta)^2(1-\alpha\lambda_i)^2 - 4\beta(1-\alpha\lambda_i) = 16\frac{(1-\alpha\lambda_i)}{(\sqrt{3\kappa+1}+2)^2}\left(1-\frac{\lambda_i}{\mu}\right). \tag{53}$$

Therefore $\Delta_i = 0$ if and only if $\lambda_i = \mu$ or $\lambda_i = \frac{3L+\mu}{4}$, and moreover $\Delta_i < 0$ for $\mu < \lambda_i < \frac{3L+\mu}{4}$ and $\Delta_i > 0$ for $\lambda_i > \frac{3L+\mu}{4}$.

(1) Consider the case $\mu < \lambda_i < \frac{3L+\mu}{4}$. Then $\Delta_i < 0$. It is known that the k-th power of a 2×2 matrix A with distinct eigenvalues $\mu_\pm$ is given by

$$A^k = \frac{\mu_+^k}{\mu_+ - \mu_-}(A - \mu_- I) + \frac{\mu_-^k}{\mu_- - \mu_+}(A - \mu_+ I),$$

where I is the 2×2 identity matrix (Williams, 1992). In our context, $A = T_i$ and $\mu_\pm = \mu_{i,\pm}$, we get

$$T_i^k = \frac{\mu_{i,+}^k}{\mu_{i,+} - \mu_{i,-}}(T_i - \mu_{i,-} I) + \frac{\mu_{i,-}^k}{\mu_{i,-} - \mu_{i,+}}(T_i - \mu_{i,+} I). \tag{54}$$

We can compute that

$$|\mu_{i,+}| = |\mu_{i,-}| = (\beta(1-\alpha\lambda_i))^{1/2} = \left(\frac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}+2}\frac{3L+\mu-4\lambda_i}{3L+\mu}\right)^{1/2} \le \left(\frac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}+2}\frac{3\kappa-3}{3\kappa+1}\right)^{1/2}, \tag{55}$$

and notice that

$$3\kappa - 3 = \left(\sqrt{3\kappa+1}+2\right)\left(\sqrt{3\kappa+1}-2\right), \tag{56}$$

and thus we get

$$|\mu_{i,+}| = |\mu_{i,-}| \le \left(\frac{(\sqrt{3\kappa+1}-2)^2}{3\kappa+1}\right)^{1/2} = 1 - \frac{2}{\sqrt{3\kappa+1}} = \rho. \tag{57}$$

Moreover,

$$\frac{1}{|\mu_{i,+} - \mu_{i,-}|} = \frac{1}{\sqrt{|\Delta_i|}} \le \frac{\sqrt{3\kappa+1}+2}{4} \max_{i:\mu<\lambda_i<\frac{3L+\mu}{4}} \frac{\sqrt{\mu}}{\sqrt{(\lambda_i-\mu)(1-\frac{4\lambda_i}{3L+\mu})}}. \tag{58}$$

Furthermore,

$$T_i - \mu_{i,-} I = \begin{pmatrix} \mu_{i,+} & -\beta(1-\alpha\lambda_i) \\ 1 & -\mu_{i,-} \end{pmatrix} = \begin{pmatrix} \mu_{i,+} \\ 1 \end{pmatrix} \begin{pmatrix} 1 & -\mu_{i,-} \end{pmatrix},$$

and

$$T_i - \mu_{i,+} I = \begin{pmatrix} \mu_{i,-} & -\beta(1-\alpha\lambda_i) \\ 1 & -\mu_{i,+} \end{pmatrix} = \begin{pmatrix} \mu_{i,-} \\ 1 \end{pmatrix} \begin{pmatrix} 1 & -\mu_{i,+} \end{pmatrix}.$$

Therefore,

$$\|T_i - \mu_{i,-} I\| \le \left\| \begin{pmatrix} \mu_{i,+} \\ 1 \end{pmatrix} \right\| \left\| \begin{pmatrix} 1 & -\mu_{i,-} \end{pmatrix} \right\| = \rho^2 + 1, \tag{59}$$

and

$$\|T_i - \mu_{i,+} I\| \le \left\| \begin{pmatrix} \mu_{i,-} \\ 1 \end{pmatrix} \right\| \left\| \begin{pmatrix} 1 & -\mu_{i,+} \end{pmatrix} \right\| = \rho^2 + 1. \tag{60}$$

Hence, it follows from (54), (57), (58), (59) and (60) that

$$\left\|T_i^k\right\| \le \frac{\sqrt{3\kappa+1}+2}{2} \max_{i:\mu<\lambda_i<\frac{3L+\mu}{4}} \frac{\sqrt{\mu}}{\sqrt{(\lambda_i-\mu)(1-\frac{4\lambda_i}{3L+\mu})}} \rho^k(\rho^2+1).$$

(2) Consider the case $\frac{3L+\mu}{4} < \lambda_i < L$. Then, $\Delta_i > 0$. As before, we have

$$T_i^k = \frac{\mu_{i,+}^k}{\mu_{i,+} - \mu_{i,-}}(T_i - \mu_{i,-}I) + \frac{\mu_{i,-}^k}{\mu_{i,-} - \mu_{i,+}}(T_i - \mu_{i,+}I). \tag{61}$$

We can compute that

$$\begin{aligned} |\mu_{i,+}| \leq |\mu_{i,-}| &= \frac{1}{2}(1+\beta)(\alpha\lambda_i - 1) + \frac{1}{2}\sqrt{\Delta_i} \\ &\leq \frac{1}{2}(1+\beta)(\alpha L - 1) + \frac{1}{2}\sqrt{16\frac{(\alpha L - 1)}{(\sqrt{3\kappa+1}+2)^2}\frac{L-\mu}{\mu}} \\ &= \frac{\sqrt{3\kappa+1}}{\sqrt{3\kappa+1}+2}\frac{\kappa-1}{3\kappa+1} + \frac{1}{2}\sqrt{16\frac{\kappa-1}{(\sqrt{3\kappa+1}+2)^2}\frac{\kappa-1}{3\kappa+1}} = 1 - \frac{2}{\sqrt{3\kappa+1}} = \rho. \end{aligned} \tag{62}$$

Moreover,

$$\frac{1}{|\mu_{i,+} - \mu_{i,-}|} = \frac{1}{\sqrt{\Delta_i}} \leq \frac{\sqrt{3\kappa+1}+2}{4} \max_{i:\frac{3L+\mu}{4}<\lambda_i<L} \frac{\sqrt{\mu}}{\sqrt{(\lambda_i - \mu)(\frac{4\lambda_i}{3L+\mu} - 1)}}. \tag{63}$$

Furthermore,

$$T_i - \mu_{i,-}I = \begin{pmatrix} \mu_{i,+} & -\beta(1-\alpha\lambda_i) \\ 1 & -\mu_{i,-} \end{pmatrix} = \begin{pmatrix} \mu_{i,+} \\ 1 \end{pmatrix} \begin{pmatrix} 1 & -\mu_{i,-} \end{pmatrix},$$

and

$$T_i - \mu_{i,+}I = \begin{pmatrix} \mu_{i,-} & -\beta(1-\alpha\lambda_i) \\ 1 & -\mu_{i,+} \end{pmatrix} = \begin{pmatrix} \mu_{i,-} \\ 1 \end{pmatrix} \begin{pmatrix} 1 & -\mu_{i,+} \end{pmatrix}.$$

Therefore,

$$\|T_i - \mu_{i,-}I\| \leq \left\| \begin{pmatrix} \mu_{i,+} \\ 1 \end{pmatrix} \right\| \left\| \begin{pmatrix} 1 & -\mu_{i,-} \end{pmatrix} \right\| \leq \rho^2 + 1, \tag{64}$$

and

$$\|T_i - \mu_{i,+}I\| \leq \left\| \begin{pmatrix} \mu_{i,-} \\ 1 \end{pmatrix} \right\| \left\| \begin{pmatrix} 1 & -\mu_{i,+} \end{pmatrix} \right\| \leq \rho^2 + 1. \tag{65}$$

Hence, it follows from (61), (62), (63), (64) and (65) that

$$\left\|T_i^k\right\| \leq \frac{\sqrt{3\kappa+1}+2}{2} \max_{i:\frac{3L+\mu}{4}<\lambda_i<L} \frac{\sqrt{\mu}}{\sqrt{(\lambda_i - \mu)(\frac{4\lambda_i}{3L+\mu} - 1)}} \rho^k(\rho^2+1).$$

(3) Consider the case $\lambda_i = \mu$. Then $\Delta_i = 0$. It is known that the k-th power of a 2×2 matrix A with two equal eigenvalues $\mu_+ = \mu_- = \mu$ is given by

$$A^k = \mu^{k-1}(kA - (k-1)\mu I),$$

where I is the 2×2 identity matrix (Williams, 1992). In our context, $A = T_i$ and

$$\mu = \mu_\pm = \mu_{i,\pm} = \frac{1}{2}(1+\beta)(1-\alpha\lambda_i) = 1 - \frac{2}{\sqrt{3\kappa+1}} = \rho. \tag{66}$$

Therefore, with $\lambda_i = \mu$, we have

$$\begin{aligned} T_i^k &= \rho^k(kT_i - (k-1)\rho I) \\ &= \rho^k \begin{pmatrix} k(1+\beta)(1-\alpha\lambda_i) - (k-1)\rho & -k\beta(1-\alpha\lambda_i) \\ k & -(k-1)\rho \end{pmatrix} \\ &= \begin{pmatrix} (k+1)\rho & -k\rho^2 \\ k & -(k-1)\rho \end{pmatrix}, \end{aligned}$$

and therefore

$$\|T_i^k\| \le \sqrt{\mathrm{Tr}\left(T_i^k (T_i^k)^T\right)} \tag{67}$$
$$= \rho^k \left((k+1)^2\rho^2 + (k-1)^2\rho^2 + k^2\rho^4 + k^2\right)^{1/2} \tag{68}$$
$$= \rho^k \sqrt{k^2(\rho^2+1)^2 + 2\rho^2}. \tag{69}$$

Furthermore, we see that the sequence T_i^k/k converges to a non-zero matrix. Therefore, $\|T_i^k\| \ge ck$ for some constant c for every k. This means that the linear dependency to k of our upper bound in (69) is tight. This behavior is expected due to the fact that T_i^k has double roots.

(4) Consider the case $\lambda_i = \frac{3L+\mu}{4}$. Then $\Delta_i = 0$. We can compute that

$$\mu_{i,\pm} = \frac{1}{2}(1+\beta)(1-\alpha\lambda_i) = 1 - \frac{2}{\sqrt{3\kappa+1}} = 0. \tag{70}$$

In this case, $T_i = 0$.

Finally, combining the three cases (1) $\mu < \lambda_i < \frac{3L+\mu}{4}$; (2) $\lambda_i > \frac{3L+\mu}{4}$; (3) $\lambda_i = \mu$; (4) $\lambda_i = \frac{3L+\mu}{4}$, and recall (51), we get

$$\begin{aligned}
&\left\| \begin{pmatrix} (1+\beta)(I_d - \alpha Q) & -\beta(I_d - \alpha Q) \\ I_d & 0_d \end{pmatrix}^k \right\| \\
&\le \max_{1\le i\le d} \left\|T_i^k\right\| \\
&\le \rho^k \max\left\{ \frac{\sqrt{3\kappa+1}+2}{2}(\rho^2+1), \max_{i:\mu<\lambda_i\neq\frac{3L+\mu}{4}} \frac{\sqrt{\mu}}{\sqrt{(\lambda_i-\mu)|1-\frac{4\lambda_i}{3L+\mu}|}}, \sqrt{k^2(\rho^2+1)^2+2\rho^2} \right\}.
\end{aligned}$$

The proof is complete. □

Proof of Theorem 7. First let us recall the ASG method:

$$\begin{aligned}
x_{k+1} &= y_k - \alpha[\nabla f(y_k) + \varepsilon_{k+1}], \\
y_k &= (1+\beta)x_k - \beta x_{k-1},
\end{aligned}$$

where $\alpha > 0$ is the step size and β is the momentum parameter. In the case when f is quadratic and $f(x) = \frac{1}{2}x^TQx + a^Tx + b$, we can compute that

$$\begin{aligned}
x_{k+1} &= y_k - \alpha[Qy_k + a + \varepsilon_{k+1}], \\
y_k &= (1+\beta)x_k - \beta x_{k-1},
\end{aligned}$$

so that with two couplings $x_k^{(1)}, x_k^{(2)}$:

$$\begin{aligned}
x_{k+1}^{(j)} &= y_k^{(j)} - \alpha\left[Qy_k^{(j)} + a + \varepsilon_{k+1}\right], \\
y_k^{(j)} &= (1+\beta)x_k^{(j)} - \beta x_{k-1}^{(j)},
\end{aligned}$$

with $j = 1, 2$, we get

$$\begin{aligned}
x_{k+1}^{(1)} - x_{k+1}^{(2)} &= y_k^{(1)} - y_k^{(2)} - \alpha Q\left(y_k^{(1)} - y_k^{(2)}\right), \\
y_k^{(1)} - y_k^{(2)} &= (1+\beta)(x_k^{(1)} - x_k^{(2)}) - \beta(x_{k-1}^{(1)} - x_{k-1}^{(2)}),
\end{aligned}$$

which implies that

$$\begin{pmatrix} x_{k+1}^{(1)} - x_{k+1}^{(2)} \\ x_k^{(1)} - x_k^{(2)} \end{pmatrix} = \begin{pmatrix} (1+\beta)(I_d - \alpha Q) & -\beta(I_d - \alpha Q) \\ I_d & 0_d \end{pmatrix} \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix},$$

which yields that

$$\left\| \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix} \right\| \leq \left\| \begin{pmatrix} (1+\beta)(I_d - \alpha Q) & -\beta(I_d - \alpha Q) \\ I_d & 0_d \end{pmatrix}^k \right\| \left\| \begin{pmatrix} x_0^{(1)} - x_0^{(2)} \\ x_{-1}^{(1)} - x_{-1}^{(2)} \end{pmatrix} \right\|.$$

Following from the proof of Theorem 4, we can show by constructing a Cauchy sequence that there exists a unique stationary distribution $\pi_{\alpha,\beta}$. Finally, we assume that $(x_0^{(1)}, x_{-1}^{(1)})$ starts from the given (x_0, x_{-1}) distributed as $\nu_{0,\alpha,\beta}$ and $(x_0^{(2)}, x_{-1}^{(2)})$ starts from the stationary distribution $\pi_{\alpha,\beta}$ so that their L_p distance is exactly the $\mathcal{W}_p$ distance. Then we get

$$\mathcal{W}_p^p\left(\nu_{k,\alpha,\beta}, \pi_{\alpha,\beta}\right) \leq \mathbb{E}\left\| \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix} \right\|^p \leq (C_k^*)^p (\rho_{AG}^*)^{pk} \mathcal{W}_p^p\left(\nu_{0,\alpha,\beta}, \pi_{\alpha,\beta}\right),$$

and the proof is complete by taking the power $1/p$ in the above equation. □

Before we state the proof of Theorem 8, let us spell out X and $V_{AG}^*(\xi_0)$ in the statement of Theorem 8 explicitly here. We will show that Theorem 8 holds with $V_{AG}^*(\xi_0)$ given by

$$V_{AG}^*(\xi_0) := \mathbb{E}\left[\left\|(\xi_0 - \xi_*)(\xi_0 - \xi_*)^T\right\|\right] + \frac{(\alpha_{AG}^*)^2 \|\Sigma\|}{1 - (\rho_{AG}^*)^2},$$

where $\Sigma := \mathbb{E}[\varepsilon_k \varepsilon_k^T]$ and $X_{AG}^* = \mathbb{E}[(\xi_\infty - \xi_*)(\xi_\infty - \xi_*)^T]$ satisfies the discrete Lyapunov equation:

$$X_{AG}^* = A_Q^* X_{AG}^* (A_Q^*)^T + \begin{pmatrix} (\alpha_{AG}^*)^2 \Sigma & 0_d \\ 0_d & 0_d \end{pmatrix},$$

and

$$A_Q^* := \begin{pmatrix} (1+\beta_{AG}^*)(I_d - \alpha_{AG}^* Q) & -\beta_{AG}^*(I_d - \alpha_{AG}^* Q) \\ I_d & 0_d \end{pmatrix}.$$

In the special case $\Sigma = c^2 I_d$ for some constant $c \geq 0$, it follows from Aybat et al. (2018) that

$$\mathrm{Tr}(X_{AG}^*) = c^2 \sum_{i=1}^{d} \frac{\alpha_{AG}^*}{\lambda_i(1 - \beta_{AG}^*(1 - \alpha_{AG}^* \lambda_i))}, \tag{71}$$

where $\{\lambda_i\}_{i=1}^d$ are the eigenvalues of Q.

Now, we are ready to prove Theorem 8.

Proof of Theorem 8. For the ASG method,

$$x_{k+1} = (1+\beta)x_k - \beta x_{k-1} - \alpha(\nabla f((1+\beta)x_k - \beta x_{k-1}) + \varepsilon_{k+1}),$$

where we consider the quadratic objective $f(x) = \frac{1}{2}x^T Q x + a^T x + b$ so that

$$x_{k+1} = (1+\beta)x_k - \beta x_{k-1} - \alpha(Q((1+\beta)x_k - \beta x_{k-1}) + a + \varepsilon_{k+1}),$$

and the minimizer x_* satisfies:

$$x_* = (1+\beta)x_* - \beta x_* - \alpha(Q((1+\beta)x_* - \beta x_*) + a),$$

so that

$$x_{k+1} - x_* = (1+\beta)(x_k - x_*) - \beta(x_{k-1} - x_*) - \alpha(Q((1+\beta)(x_k - x_*) - \beta(x_{k-1} - x_*)) + \varepsilon_{k+1}),$$

and

$$\begin{pmatrix} x_k - x_* \\ x_{k-1} - x_* \end{pmatrix} = \begin{pmatrix} (1+\beta)(I_d - \alpha Q) & -\beta(I_d - \alpha Q) \\ I_d & 0_d \end{pmatrix} \begin{pmatrix} x_{k-1} - x_* \\ x_{k-2} - x_* \end{pmatrix} + \begin{pmatrix} -\alpha\varepsilon_k \\ 0_d \end{pmatrix},$$

and with $\Sigma := \mathbb{E}[\varepsilon_k \varepsilon_k^T]$, we get

$$\mathbb{E}\left[(\xi_k - \xi_*)(\xi_k - \xi_*)^T\right] = A_Q^* \mathbb{E}\left[(\xi_{k-1} - x_*)(\xi_{k-1} - x_*)^T\right](A_Q^*)^T + \begin{pmatrix} \alpha^2\Sigma & 0_d \\ 0_d & 0_d \end{pmatrix}, \tag{72}$$

where

$$A_Q^* = \begin{pmatrix} (1+\beta)(I_d - \alpha Q) & -\beta(I_d - \alpha Q) \\ I_d & 0_d \end{pmatrix}.$$

Therefore,

$$X = \mathbb{E}\left[(\xi_\infty - \xi_*)(\xi_\infty - \xi_*)^T\right]$$

satisfies the discrete Lyapunov equation:

$$X = A_Q^* X (A_Q^*)^T + \begin{pmatrix} \alpha^2\Sigma & 0_d \\ 0_d & 0_d \end{pmatrix}.$$

Next by iterating equation (72) over k, we immediately obtain

$$\mathbb{E}\left[(\xi_k - \xi_*)(\xi_k - \xi_*)^T\right] = \left(A_Q^*\right)^k \mathbb{E}\left[(\xi_0 - \xi_*)(\xi_0 - \xi_*)^T\right]\left((A_Q^*)^T\right)^k + \sum_{j=0}^{k-1} \left(A_Q^*\right)^j \begin{pmatrix} \alpha^2\Sigma & 0_d \\ 0_d & 0_d \end{pmatrix} \left((A_Q^*)^T\right)^j,$$

so that

$$\begin{aligned}\mathbb{E}\left[(\xi_k - \xi_*)(\xi_k - \xi_*)^T\right] = \mathbb{E}\left[(\xi_\infty - \xi_*)(\xi_\infty - \xi_*)^T\right] + \left(A_Q^*\right)^k \mathbb{E}\left[(\xi_0 - \xi_*)(\xi_0 - \xi_*)^T\right]\left((A_Q^*)^T\right)^k \\ - \sum_{j=k}^{\infty} \left(A_Q^*\right)^j \begin{pmatrix} \alpha^2\Sigma & 0_d \\ 0_d & 0_d \end{pmatrix} \left((A_Q^*)^T\right)^j,\end{aligned}$$

which implies that

$$\begin{aligned}\mathrm{Tr}\left(\mathbb{E}\left[(\xi_k - \xi_*)(\xi_k - \xi_*)^T\right]\right) &= \mathrm{Tr}\left(\mathbb{E}\left[(\xi_\infty - \xi_*)(\xi_\infty - \xi_*)^T\right]\right) + \left(A_Q^*\right)^k \mathbb{E}\left[(\xi_0 - \xi_*)(\xi_0 - \xi_*)^T\right]\left((A_Q^*)^T\right)^k \\ &\quad - \sum_{j=k}^{\infty} \left(A_Q^*\right)^j \begin{pmatrix} \alpha^2\Sigma & 0_d \\ 0_d & 0_d \end{pmatrix} \left((A_Q^*)^T\right)^j \\ &\le \mathrm{Tr}(X) + \left\|(A_Q^*)^k\right\|^2 \mathbb{E}\left[\left\|(\xi_0 - \xi_*)(\xi_0 - \xi_*)^T\right\|\right] + \sum_{j=k}^{\infty} \left\|(A_Q^*)^j\right\|^2 \alpha^2 \|\Sigma\| \\ &\le \mathrm{Tr}(X) + (C_k^*)^2 (\rho_{AG}^*)^{2k} \mathbb{E}\left[\left\|(\xi_0 - \xi_*)(\xi_0 - \xi_*)^T\right\|\right] + \alpha^2 \|\Sigma\| (C_k^*)^2 \frac{(\rho_{AG}^*)^{2k}}{1 - (\rho_{AG}^*)^2},\end{aligned}$$

where we used the estimate $\|(A_Q^*)^k\| \le C_k^* (\rho_{AG}^*)^k$ from the proof of Theorem 5.

Finally, since ∇f is L-Lipschtiz,

$$\mathbb{E}[f(x_k)] - f(x_*) \le \frac{L}{2}\mathbb{E}\|x_k - x_*\|^2 \le \frac{L}{2}\mathbb{E}\|\xi_k - \xi_*\|^2 = \frac{L}{2}\mathrm{Tr}\left(\mathbb{E}\left[(\xi_k - \xi_*)(\xi_k - \xi_*)^T\right]\right).$$

The proof of (23) is complete. □

Remark 24. *Note that our results in p-Wasserstein distances would hold if there exists some $p \ge 1$ so that p-th moment of the noise is finite. For instance, the $p < 2$ case can arise in applications where the noise has heavy tail (see e.g. (Simsekli et al., 2019)).*

C.2. Proofs of Results in Section 3.2

Proof of Theorem 9. First let us recall the HB method:

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}),$$

where $\alpha > 0$ is the step size and β is the momentum parameter. In the case when f is quadratic and $f(x) = \frac{1}{2}x^T Q x + a^T x + b$, we can compute that

$$x_{k+1} = x_k - \alpha(Qx_k + a) + \beta(x_k - x_{k-1}),$$

and the minimizer x_* satisfies

$$x_* = x_* - \alpha(Qx_* + a) + \beta(x_* - x_*),$$

which implies that

$$\begin{pmatrix} x_{k+1} - x_* \\ x_k - x_* \end{pmatrix} = \begin{pmatrix} (1+\beta)I_d - \alpha Q & -\beta I_d \\ I_d & 0_d \end{pmatrix} \begin{pmatrix} x_k - x_* \\ x_{k-1} - x_* \end{pmatrix},$$

which yields that

$$\begin{pmatrix} x_k - x_* \\ x_{k-1} - x_* \end{pmatrix} = \begin{pmatrix} (1+\beta)I_d - \alpha Q & -\beta I_d \\ I_d & 0_d \end{pmatrix}^k \begin{pmatrix} x_0 - x_* \\ x_{-1} - x_* \end{pmatrix},$$

and we aim to provide an upper bound to the 2-norm of the matrix, that is:

$$\left\| \begin{pmatrix} (1+\beta)I_d - \alpha Q & -\beta I_d \\ I_d & 0_d \end{pmatrix}^k \right\|.$$

Let us assume that Q has the decomposition

$$Q = VDV^T,$$

where D is diagonal consisting of eigenvalues λ_i, $1 \le i \le d$ in increasing order:

$$\mu = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_d = L,$$

then we have

$$(1+\beta)I_d - \alpha Q = V\tilde{D}V^T,$$

where $\tilde{D} = (1+\beta)I_d - \alpha D$ is diagonal matrix with entries

$$1 + \beta - \alpha\lambda_i, \qquad 1 \le i \le d.$$

Therefore, the matrix

$$\begin{pmatrix} (1+\beta)I_d - \alpha Q & -\beta I_d \\ I_d & 0_d \end{pmatrix}$$

has the same eigenvalues as the matrix

$$\begin{pmatrix} (1+\beta)I_d - \alpha D & -\beta I_d \\ I_d & 0_d \end{pmatrix},$$

which has the same eigenvalues as the matrix:

$$\begin{pmatrix} T_1 & \cdots & 0 & 0 \\ 0 & T_2 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & T_d \end{pmatrix},$$

where

$$T_i = \begin{pmatrix} 1 + \beta - \alpha\lambda_i & -\beta \\ 1 & 0 \end{pmatrix}, \qquad 1 \le i \le d,$$

are 2×2 matrices with eigenvalues:

$$\mu_{i,\pm} = \frac{1+\beta-\alpha\lambda_i \pm \sqrt{(1+\beta-\alpha\lambda_i)^2 - 4\beta}}{2},$$

where $1 \leq i \leq d$, and therefore

$$\left\| \begin{pmatrix} (1+\beta)I_d - \alpha Q & -\beta I_d \\ I_d & 0_d \end{pmatrix}^k \right\| \leq \max_{1\leq i\leq d} \left\| T_i^k \right\|. \tag{73}$$

Next, we upper bound $\|T_i^k\|$. We consider three cases (1) $\mu < \lambda_i < L$; (2) $\lambda_i = \mu$; (3) $\lambda_i = L$.

(1) Consider the case $\mu < \lambda_i < L$. With the choice of α and β in (12), we can compute that for those $\mu < \lambda_i < L$, we have

$$1+\beta-\alpha\lambda_i < 1+\beta-\alpha\mu = 2\sqrt{\beta},$$

and

$$1+\beta-\alpha\lambda_i > 1+\beta-\alpha L = -2\sqrt{\beta},$$

and thus the eigenvalues are complex and

$$\mu_{i,\pm} = \frac{1+\beta-\alpha\lambda_i \pm \mathbf{i}\sqrt{4\beta-(1+\beta-\alpha\lambda_i)^2}}{2},$$

where $1 \leq i \leq d$. It is known that the k-th power of a 2×2 matrix A with distinct eigenvalues $\mu_\pm$ is given by

$$A^k = \frac{\mu_+^k}{\mu_+ - \mu_-}(A - \mu_- I) + \frac{\mu_-^k}{\mu_- - \mu_+}(A - \mu_+ I),$$

where I is the 2×2 identity matrix (Williams, 1992). In our context, $A = T_i$ and $\mu_\pm = \mu_{i,\pm}$, we get

$$T_i^k = \frac{\mu_{i,+}^k}{\mu_{i,+} - \mu_{i,-}}(T_i - \mu_{i,-} I) + \frac{\mu_{i,-}^k}{\mu_{i,-} - \mu_{i,+}}(T_i - \mu_{i,+} I). \tag{74}$$

We can compute that

$$|\mu_{i,+}| = |\mu_{i,-}| = \left(\frac{1}{4}\left[(1+\beta-\alpha\lambda_i)^2 + (4\beta-(1+\beta-\alpha\lambda_i)^2)\right]\right)^{1/2} = \sqrt{\beta}, \tag{75}$$

and

$$\begin{aligned} \frac{1}{|\mu_{i,+} - \mu_{i,-}|} &= \frac{1}{\sqrt{4\beta-(1+\beta-\alpha\lambda_i)^2}} \\ &= \frac{1}{\sqrt{(2\sqrt{\beta}-1-\beta+\alpha\lambda_i)(2\sqrt{\beta}+1+\beta-\alpha\lambda_i)}} \\ &= \frac{1}{\sqrt{(-(\sqrt{\beta}-1)^2+\alpha\lambda_i)((\sqrt{\beta}+1)^2-\alpha\lambda_i)}} \\ &= \frac{(\sqrt{\mu}+\sqrt{L})^2}{4\sqrt{(\lambda_i-\mu)(L-\lambda_i)}}. \end{aligned} \tag{76}$$

Moreover,

$$T_i - \mu_{i,-} I = \begin{pmatrix} \mu_{i,+} & -\beta \\ 1 & -\mu_{i,-} \end{pmatrix} = \begin{pmatrix} \mu_{i,+} \\ 1 \end{pmatrix} \begin{pmatrix} 1 & -\mu_{i,-} \end{pmatrix},$$

and

$$T_i - \mu_{i,+} I = \begin{pmatrix} \mu_{i,-} & -\beta \\ 1 & -\mu_{i,+} \end{pmatrix} = \begin{pmatrix} \mu_{i,-} \\ 1 \end{pmatrix} \begin{pmatrix} 1 & -\mu_{i,+} \end{pmatrix}.$$

Therefore,

$$\|T_i - \mu_{i,-} I\| \leq \left\| \begin{pmatrix} \mu_{i,+} \\ 1 \end{pmatrix} \right\| \left\| \begin{pmatrix} 1 & -\mu_{i,-} \end{pmatrix} \right\| = \beta + 1, \tag{77}$$

and

$$\|T_i - \mu_{i,+} I\| \leq \left\| \begin{pmatrix} \mu_{i,-} \\ 1 \end{pmatrix} \right\| \left\| \begin{pmatrix} 1 & -\mu_{i,+} \end{pmatrix} \right\| = \beta + 1. \tag{78}$$

Hence, it follows from (74), (75), (76), (77) and (78) that

$$\left\|T_i^k\right\| \leq (\sqrt{\beta})^k \frac{(\beta+1)(\sqrt{\mu}+\sqrt{L})^2}{4\sqrt{(\lambda_i-\mu)(L-\lambda_i)}} = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^k \frac{\mu+L}{2\sqrt{(\lambda_i-\mu)(L-\lambda_i)}}.$$

(2) Consider the case $\lambda_i = \mu$. With the choice of α and β in (12), we can compute that for those $\lambda_i = \mu$, we have

$$(1+\beta-\alpha\lambda_i)^2 = (1+\beta-\alpha\mu)^2 = 4\beta,$$

so we have double eigenvalues and indeed $1+\beta-\alpha\lambda_i = 2\sqrt{\beta}$, and

$$T_i = \begin{pmatrix} 2\sqrt{\beta} & -\beta \\ 1 & 0 \end{pmatrix}, \qquad 1 \leq i \leq d,$$

and by a direct computation (e.g. induction on k), we get:

$$T_i^k = (\sqrt{\beta})^k \begin{pmatrix} (k+1) & -k\beta^{1/2} \\ k\beta^{-1/2} & -(k-1) \end{pmatrix}, \quad 1 \leq i \leq d.$$

Thus,

$$\left\|T_i^k\right\| \leq \sqrt{\mathrm{Tr}\left(T_i^k (T_i^k)^T\right)} \tag{79}$$

$$= (\sqrt{\beta})^k \sqrt{2k^2 + 2 + k^2(\beta + \beta^{-1})} \tag{80}$$

$$= \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^k \sqrt{4k^2 \left(\frac{L+\mu}{L-\mu}\right)^2 + 2}. \tag{81}$$

Finally, we note that the matrix $T_i^k/(\sqrt{\beta}^k k)$ as k goes to infinity converges to the 2×2 matrix

$$M_{2,2}(\beta) := \begin{pmatrix} 1 & -\beta^{1/2} \\ \beta^{-1/2} & -1 \end{pmatrix}, \quad \|M_{2,2}(\beta)\| > 0.$$

Therefore, the linear dependency of our bound in (81) with respect to k is tight. This behavior is expected due to the fact that T_i^k has double roots.

(3) Consider the case $\lambda_i = L$. With the choice of α and β in (12), we can compute that for those $\lambda_i = L$, we have

$$(1+\beta-\alpha\lambda_i)^2 = (1+\beta-\alpha L)^2 = 4\beta,$$

so we have double eigenvalues and indeed $1+\beta-\alpha\lambda_i = -2\sqrt{\beta}$, and

$$T_i = \begin{pmatrix} -2\sqrt{\beta} & -\beta \\ 1 & 0 \end{pmatrix}, \qquad 1 \leq i \leq d,$$

and by a direct computation (e.g. induction on k), we get:

$$T_i^k = (\sqrt{\beta})^k \begin{pmatrix} (k+1) & k\beta^{1/2} \\ -k\beta^{-1/2} & -(k-1) \end{pmatrix}, \quad 1 \leq i \leq d.$$

Thus,

$$\begin{aligned}\left\|T_i^k\right\| &\leq \sqrt{\mathrm{Tr}\left(T_i^k(T_i^k)^T\right)}\\ &= (\sqrt{\beta})^k\sqrt{2k^2+2+k^2(\beta+\beta^{-1})}\\ &= \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^k\sqrt{4k^2\left(\frac{L+\mu}{L-\mu}\right)^2+2}.\end{aligned}$$

Finally, combining the three cases (1) $\mu<\lambda_i<L$; (2) $\lambda_i=\mu$; (3) $\lambda_i=L$, we get

$$\max_{1\leq i\leq d}\left\|T_i^k\right\| \leq \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^k \max\left\{\max_{i:\mu<\lambda_i<L}\frac{\mu+L}{2\sqrt{(\lambda_i-\mu)(L-\lambda_i)}}, \sqrt{4k^2\left(\frac{L+\mu}{L-\mu}\right)^2+2}\right\}. \tag{82}$$

Then it follows from (73) that

$$\begin{aligned}&\left\|\begin{pmatrix}(1+\beta)I_d-\alpha Q & -\beta I_d\\ I_d & 0_d\end{pmatrix}^k\right\| \\ &\leq \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^k \max\left\{\max_{i:\mu<\lambda_i<L}\frac{\mu+L}{2\sqrt{(\lambda_i-\mu)(L-\lambda_i)}}, \sqrt{4k^2\left(\frac{L+\mu}{L-\mu}\right)^2+2}\right\}.\end{aligned} \tag{83}$$

Recall that

$$\begin{pmatrix}x_k-x_*\\ x_{k-1}-x_*\end{pmatrix} = \begin{pmatrix}(1+\beta)I_d-\alpha Q & -\beta I_d\\ I_d & 0_d\end{pmatrix}^k\begin{pmatrix}x_0-x_*\\ x_{-1}-x_*\end{pmatrix},$$

and the proof is complete by applying (83). □

Before we state the proof of Theorem 11, let us state the following result, which is built on Theorem 9.

Lemma 25. *Let us consider two couplings $(x_k^{(1)})_{k\geq 0}$ and $(x_k^{(2)})_{k\geq 0}$ with the common noise $(\varepsilon_{k+1})_{k\geq 0}$ that starts from $x_0^{(1)}$ and $x_0^{(2)}$:*

$$x_{k+1}^{(1)} = x_k^{(1)} - \alpha\nabla f(x_k^{(1)}) + \beta(x_k^{(1)}-x_{k-1}^{(1)}) + \varepsilon_{k+1}, \tag{84}$$

$$x_{k+1}^{(2)} = x_k^{(2)} - \alpha\nabla f(x_k^{(2)}) + \beta(x_k^{(2)}-x_{k-1}^{(2)}) + \varepsilon_{k+1}, \tag{85}$$

where f is quadratic and $f(x)=\frac{1}{2}x^TQx+a^Tx+b$. Then, we have

$$\left\|\begin{pmatrix}x_{k+1}^{(1)}-x_{k+1}^{(2)}\\ x_k^{(1)}-x_k^{(2)}\end{pmatrix}\right\| \leq C_k\rho_{HB}^k\left\|\begin{pmatrix}x_1^{(1)}-x_1^{(2)}\\ x_0^{(1)}-x_0^{(2)}\end{pmatrix}\right\|,$$

where ρ_{HB} and C_k are defined by (13) and (25) respectively.

Proof of Lemma 25. We can compute that

$$\begin{pmatrix}x_{k+1}^{(1)}-x_{k+1}^{(2)}\\ x_k^{(1)}-x_k^{(2)}\end{pmatrix} = \begin{pmatrix}(1+\beta)I_d-\alpha Q & -\beta I_d\\ I_d & 0_d\end{pmatrix}^k\begin{pmatrix}x_1^{(1)}-x_1^{(2)}\\ x_0^{(1)}-x_0^{(2)}\end{pmatrix}.$$

It follows from the estimate (83) in the proof of Theorem 9 and the definitions of ρ_{HB} and C_k in (13) and (25) that we have

$$\left\|\begin{pmatrix}(1+\beta)I_d-\alpha Q & -\beta I_d\\ I_d & 0_d\end{pmatrix}^k\right\| \leq C_k\rho_{HB}^k.$$

The proof is complete. □

Proof of Theorem 11. We recall from Lemma 25 that for any coupling $x^{(1)}$ and $x^{(2)}$

$$\left\| \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix} \right\| \leq C_k \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^k \left\| \begin{pmatrix} x_0^{(1)} - x_0^{(2)} \\ x_{-1}^{(1)} - x_{-1}^{(2)} \end{pmatrix} \right\|.$$

Following from the proof of Theorem 4, we can show by constructing a Cauchy sequence that there exists a unique stationary distribution $\pi_{\alpha,\beta}$. Finally, we assume that $(x_0^{(1)}, x_{-1}^{(1)})$ starts from the given (x_0, x_{-1}) distributed as $\nu_{0,\alpha,\beta}$ and $(x_0^{(2)}, x_{-1}^{(2)})$ starts from the stationary distribution $\pi_{\alpha,\beta}$ so that their L_p distance is exactly the $\mathcal{W}_p$ distance. Then we get

$$\begin{aligned} \mathcal{W}_p^p \left(\nu_{k,\alpha,\beta}, \pi_{\alpha,\beta} \right) &\leq \mathbb{E} \left\| \begin{pmatrix} x_k^{(1)} - x_k^{(2)} \\ x_{k-1}^{(1)} - x_{k-1}^{(2)} \end{pmatrix} \right\|^p \\ &\leq C_k^p \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^{pk} \mathcal{W}_p^p \left(\nu_{0,\alpha,\beta}, \pi_{\alpha,\beta} \right), \end{aligned}$$

and the proof is complete by taking the power $1/p$ in the above equation. □

Before we state the proof of Theorem 12, let us spell out X and $V_{HB}(\xi_0)$ in the statement of Theorem 12 explicitly here. We will show that Theorem 12 holds with $V_{HB}(\xi_0)$ given by

$$V_{HB}(\xi_0) := \mathbb{E}\left[\|(\xi_0 - \xi_*)(\xi_0 - \xi_*)^T\| \right] + \frac{\alpha_{HB}^2 \|\Sigma\|}{1 - \rho_{HB}^2},$$

where $\Sigma := \mathbb{E}[\varepsilon_k \varepsilon_k^T]$ and $X_{HB} = \mathbb{E}[(\xi_\infty - \xi_*)(\xi_\infty - \xi_*)^T]$ satisfies the discrete Lyapunov equation:

$$X_{HB} = A_Q X_{HB} A_Q^T + \begin{pmatrix} \alpha_{HB}^2 \Sigma & 0_d \\ 0_d & 0_d \end{pmatrix}.$$

and

$$A_Q := \begin{pmatrix} (1 + \beta_{HB}) I_d - \alpha_{HB} Q & -\beta_{HB} I_d \\ I_d & 0_d \end{pmatrix}.$$

In the special case $\Sigma = c^2 I_d$ for some constant $c \geq 0$, we obtain

$$\mathrm{Tr}(X_{HB}) = c^2 \sum_{i=1}^{d} \frac{2\alpha_{HB}(1 + \beta_{HB})}{(1 - \beta_{HB})\lambda_i (2 + 2\beta_{HB} - \alpha_{HB}\lambda_i)}, \tag{86}$$

where $\{\lambda_i\}_{i=1}^d$ are the eigenvalues of Q.

Now, we are ready to prove Theorem 12.

Proof of Theorem 12. For the stochastic heavy ball method

$$x_{k+1} = x_k - \alpha(\nabla f(x_k) + \varepsilon_{k+1}) + \beta(x_k - x_{k-1}),$$

where we consider the quadratic objective $f(x) = \frac{1}{2} x^T Q x + a^T x + b$ so that

$$x_{k+1} = x_k - \alpha(Q x_k + a + \varepsilon_{k+1}) + \beta(x_k - x_{k-1}),$$

and the minimizer x_* satisfies:

$$x_* = x_* - \alpha(Q x_* + a) + \beta(x_* - x_*),$$

so that

$$(x_{k+1} - x_*) = (x_k - x_*) - \alpha(Q(x_k - x_*) + \varepsilon_{k+1}) + \beta((x_k - x_*) - (x_{k-1} - x_*)),$$

and

$$\begin{pmatrix} x_k - x_* \\ x_{k-1} - x_* \end{pmatrix} = \begin{pmatrix} (1 + \beta) I_d - \alpha Q & -\beta I_d \\ I_d & 0_d \end{pmatrix} \begin{pmatrix} x_{k-1} - x_* \\ x_{k-2} - x_* \end{pmatrix} + \begin{pmatrix} -\alpha \varepsilon_k \\ 0_d \end{pmatrix},$$

and with $\Sigma := \mathbb{E}[\varepsilon_k \varepsilon_k^T]$, we get

$$\mathbb{E}\left[(\xi_k - \xi_*)(\xi_k - \xi_*)^T\right] = A_Q \mathbb{E}\left[(\xi_{k-1} - x_*)(\xi_{k-1} - x_*)^T\right] A_Q^T + \begin{pmatrix} \alpha^2\Sigma & 0_d \\ 0_d & 0_d \end{pmatrix}, \tag{87}$$

where

$$A_Q = \begin{pmatrix} (1+\beta)I_d - \alpha Q & -\beta I_d \\ I_d & 0_d \end{pmatrix}.$$

Therefore,

$$X = \mathbb{E}\left[(\xi_\infty - \xi_*)(\xi_\infty - \xi_*)^T\right]$$

satisfies the discrete Lyapunov equation:

$$X = A_Q X A_Q^T + \begin{pmatrix} \alpha^2\Sigma & 0_d \\ 0_d & 0_d \end{pmatrix}.$$

Next by iterating equation (87) over k, we immediately obtain

$$\mathbb{E}\left[(\xi_k - \xi_*)(\xi_k - \xi_*)^T\right] = (A_Q)^k \mathbb{E}\left[(\xi_0 - \xi_*)(\xi_0 - \xi_*)^T\right] \left(A_Q^T\right)^k + \sum_{j=0}^{k-1} (A_Q)^j \begin{pmatrix} \alpha^2\Sigma & 0_d \\ 0_d & 0_d \end{pmatrix} \left(A_Q^T\right)^j,$$

so that

$$\begin{aligned}
&\mathbb{E}\left[(\xi_k - \xi_*)(\xi_k - \xi_*)^T\right] \\
&= \mathbb{E}\left[(\xi_\infty - \xi_*)(\xi_\infty - \xi_*)^T\right] + (A_Q)^k \mathbb{E}\left[(\xi_0 - \xi_*)(\xi_0 - \xi_*)^T\right] \left(A_Q^T\right)^k - \sum_{j=k}^{\infty} (A_Q)^j \begin{pmatrix} \alpha^2\Sigma & 0_d \\ 0_d & 0_d \end{pmatrix} \left(A_Q^T\right)^j,
\end{aligned}$$

which implies that

$$\begin{aligned}
&\mathrm{Tr}\left(\mathbb{E}\left[(\xi_k - \xi_*)(\xi_k - \xi_*)^T\right]\right) \\
&= \mathrm{Tr}\left(\mathbb{E}\left[(\xi_\infty - \xi_*)(\xi_\infty - \xi_*)^T\right]\right) + (A_Q)^k \mathbb{E}\left[(\xi_0 - \xi_*)(\xi_0 - \xi_*)^T\right] \left(A_Q^T\right)^k - \sum_{j=k}^{\infty} (A_Q)^j \begin{pmatrix} \alpha^2\Sigma & 0_d \\ 0_d & 0_d \end{pmatrix} \left(A_Q^T\right)^j \\
&\le \mathrm{Tr}(X) + \left\|A_Q^k\right\|^2 \mathbb{E}\left[\|(\xi_0 - \xi_*)(\xi_0 - \xi_*)^T\|\right] + \sum_{j=k}^{\infty} \left\|A_Q^j\right\|^2 \alpha^2 \|\Sigma\| \\
&\le \mathrm{Tr}(X) + C_k^2 \rho_{HB}^{2k} \mathbb{E}\left[\|(\xi_0 - \xi_*)(\xi_0 - \xi_*)^T\|\right] + \alpha^2 \|\Sigma\| C_k^2 \frac{\rho_{HB}^{2k}}{1 - \rho_{HB}^2},
\end{aligned}$$

where we used the estimate $\|A_Q^k\| \le C_k \rho_{HB}^k$ from the proof of Theorem 9.

Finally, since ∇f is L-Lipschtiz,

$$\mathbb{E}[f(x_k)] - f(x_*) \le \frac{L}{2}\mathbb{E}\|x_k - x_*\|^2 \le \frac{L}{2}\mathbb{E}\|\xi_k - \xi_*\|^2 = \frac{L}{2}\mathrm{Tr}\left(\mathbb{E}\left[(\xi_k - \xi_*)(\xi_k - \xi_*)^T\right]\right).$$

The proof of (27) is complete. To show (86), we can adapt the proof technique of Aybat et al. (2018, Proposition 3.2) for gradient descent to HB. Without loss of generality, due to the scaling of the Lyapunov equation, we can assume $c = 1$. Consider the eigenvalue decomposition $A_Q = V \Lambda V^T$ where Q is orthogonal and Λ is diagonal with $\Lambda(i,i) = \lambda_i$. We can write

$$A_Q = \bar{V} A_\Lambda \bar{V}^T,$$

where

$$\bar{V} = \begin{pmatrix} V & 0_d \\ 0_d & V \end{pmatrix}, \quad A_\Lambda = \begin{pmatrix} (1+\beta)I_d - \alpha\Lambda & -\beta I_d \\ I_d & 0_d \end{pmatrix}.$$

Futhermore, following Recht (2012), let $P \in \mathbb{R}^{2d\times 2d}$ be the permutation matrix with entries

$$P(i,j) = \begin{cases} 1 & \text{if } i \text{ is odd}, j = i, \\ 1 & \text{if } i \text{ is even}, j = 2d + i, \\ 0 & \text{otherwise.} \end{cases}$$

Then, we have

$$A_M := PA_\Lambda P^T = \begin{pmatrix} M_1 & 0_d & \dots & 0_d \\ 0_d & M_2 & \dots & 0_d \\ \vdots & \vdots & \ddots & \vdots \\ 0_d & 0_d & \dots & M_d \end{pmatrix} \quad \text{where} \quad M_i = \begin{pmatrix} (1+\beta) - \alpha\lambda_i & -\beta \\ 1 & 0 \end{pmatrix} \in \mathbb{R}^{2\times 2}.$$

If we define $Y := UXU^{-1}$ for the orthogonal matrix $U = P\bar{V}^T$, it solves

$$A_M Y A_M^T - Y + S = 0, \quad S := P \begin{pmatrix} \alpha^2 I_d & 0_d \\ 0_d & 0_d \end{pmatrix} P^T,$$

where the latter matrix S is a $2d \times 2d$ diagonal matrix with entries $S(i,i) = \alpha^2$ if i is odd, and zero if i is even. Due to the special structure of S and A_M, the solution Y has the structure

$$Y = \begin{pmatrix} Y_1 & 0_d & \dots & 0_d \\ 0_d & Y_2 & \dots & 0_d \\ \vdots & \vdots & \ddots & \vdots \\ 0_d & 0_d & \dots & Y_d \end{pmatrix},$$

where Y_i solves the 2×2 Lyapunov equation

$$M_i Y_i M_i^T - Y_i + \begin{pmatrix} \alpha^2 & 0 \\ 0 & 0 \end{pmatrix} = 0.$$

If we write

$$Y_i = \begin{pmatrix} x_i & y_i \\ y_i & w_i \end{pmatrix}$$

with scalars x_i, y_i and w_i, this equation is equivalent to the linear system

$$\begin{pmatrix} a^2 - 1 & 2ab & b^2 \\ a & b-1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} x_i \\ y_i \\ w_i \end{pmatrix} = \begin{pmatrix} -\alpha^2 \\ 0 \\ 0 \end{pmatrix},$$

with

$$a = 1 + \beta - \alpha\lambda_i, \quad b = -\beta.$$

After a simple computation, we obtain

$$x_i = w_i = \frac{\alpha^2(b-1)}{(b+1)(a-b+1)(a+b-1)} = \frac{\alpha(1+\beta)}{(1-\beta)\lambda_i(2+2\beta-\alpha\lambda_i)}.$$

Therefore we obtain

$$\mathrm{Tr}(X) = \mathrm{Tr}(Y) = \sum_{i=1}^{d} \mathrm{Tr}(Y_i) = 2\sum_{i=1}^{d} x_i = \sum_{i=1}^{d} \frac{2\alpha(1+\beta)}{(1-\beta)\lambda_i(2+2\beta-\alpha\lambda_i)},$$

which completes the proof.

□

D. Proofs of Results in Section 4

Before we proceed to prove the main results in Section 4, let us first show that the weighted total variation distance d_ψ upper bounds the standard 1-Wasserstein distance.

Proposition 26. *Assume $\tilde{P}(2,2) \neq 0$. Then,*

$$\mathcal{W}_1(\mu_1, \mu_2) \leq c_0^{-1} d_\psi(\mu_1, \mu_2),$$

where $\mathcal{W}_1$ is the standard 1-Wasserstein distance and

$$c_0 := \min\{\hat{c}_0 \psi, 1\}, \tag{88}$$

where $\hat{c}_0$ is the smallest positive eigenvalue of

$$\tilde{P} \otimes I_d + \begin{pmatrix} \frac{\mu}{2} I_d & 0_d \\ 0_d & 0_d \end{pmatrix}.$$

Proof. By applying the Kantorovich-Rubinstein duality for the Wasserstein metric (see e.g. Villani (2009)), we get

$$\begin{aligned}
\mathcal{W}_1(\mu_1, \mu_2) &= \sup_{\phi \in L^1(d\mu_1)} \left\{ \int_{\mathbb{R}^{2d}} \phi(\xi)(\mu_1 - \mu_2)(d\xi) : \phi \text{ is 1-Lipschitz} \right\} \\
&= \sup_{\phi \in L^1(d\mu_1)} \left\{ \int_{\mathbb{R}^{2d}} (\phi(\xi) - \phi(\xi_*))(\mu_1 - \mu_2)(d\xi) : \phi \text{ is 1-Lipschitz} \right\} \\
&\leq \int_{\mathbb{R}^{2d}} \|\xi - \xi_*\| |\mu_1 - \mu_2|(d\xi) \\
&\leq c_0^{-1} \int_{\mathbb{R}^{2d}} (1 + \psi V_P(\xi)) |\mu_1 - \mu_2|(d\xi) = c_0^{-1} d_\psi(\mu_1, \mu_2),
\end{aligned}$$

where we used $1 + \psi V_P(\xi) \geq c_0 \|\xi - \xi_*\|$ from Lemma 27. □

Lemma 27. *Assume $\tilde{P}(2,2) \neq 0$. Then,*

$$1 + \psi V_P(\xi) \geq c_0 \|\xi - \xi_*\|,$$

for any $\xi \in \mathbb{R}^{2d}$, where $c_0 = \min\{\hat{c}_0 \psi, 1\}$, where $\hat{c}_0$ is the smallest positive eigenvalue of

$$\tilde{P} \otimes I_d + \begin{pmatrix} \frac{\mu}{2} I_d & 0_d \\ 0_d & 0_d \end{pmatrix}.$$

Proof. Let $\xi^T = (x^T, y^T)$. If $\|\xi - \xi_*\| \leq 1$, then $c_0 = 1$ works. Otherwise,

$$\begin{aligned}
V_P(\xi) &= f(x) - f(x_*) + (\xi - \xi_*)^T P(\xi - \xi_*) \\
&\geq (\xi - \xi_*)^T P(\xi - \xi_*) + \frac{\mu}{2} \|x - x_*\|^2 \\
&= (\xi - \xi_*)^T \tilde{P} \otimes I_d (\xi - \xi_*) + (\xi - \xi_*)^T \begin{pmatrix} \frac{\mu}{2} I_d & 0_d \\ 0_d & 0_d \end{pmatrix} (\xi - \xi_*).
\end{aligned}$$

The proof is complete. □

For constrained optimization on a compact set $\mathcal{C}$, we have the following result.

Proposition 28. *For any μ_1, μ_2 on the product space $\mathcal{C}^2 := \mathcal{C} \times \mathcal{C}$,*

$$\mathcal{W}_p(\mu_1, \mu_2) \leq 2^{1/p} \mathcal{D}_{\mathcal{C}^2} \|\mu_1 - \mu_2\|_{TV}^{1/p} \leq \mathcal{D}_{\mathcal{C}^2} d_\psi^{1/p}(\mu_1, \mu_2),$$

where $\mathcal{D}_{\mathcal{C}^2}$ is the diameter of $\mathcal{C}^2$.

Proof. The second inequality in Proposition 28 follows from $d_\psi(\mu_1,\mu_2) \geq 2\|\mu_1-\mu_2\|_{TV}$. So it suffices to prove the first inequality. We can compute that

$$\begin{aligned}
\mathcal{W}_p^p(\mu_1,\mu_2) &= \inf_{X_1\sim\mu_1, X_2\sim\mu_2} \mathbb{E}\left[\|X_1-X_2\|^p\right] \\
&\leq \mathcal{D}_{\mathcal{C}^2}^{p-1} \inf_{X_1\sim\mu_1, X_2\sim\mu_2} \mathbb{E}\left[\|X_1-X_2\|\right] \\
&= \mathcal{D}_{\mathcal{C}^2}^{p-1} \mathcal{W}_1(\mu_1,\mu_2) \\
&= \mathcal{D}_{\mathcal{C}^2}^{p-1} \sup_{\phi\in L^1(d\mu_1)} \left\{ \int_{\mathbb{R}^{2d}} (\phi(\xi)-\phi(\xi_*))(\mu_1-\mu_2)(d\xi) : \phi \text{ is 1-Lipschitz} \right\} \\
&\leq \mathcal{D}_{\mathcal{C}^2}^{p-1} \int_{\mathbb{R}^{2d}} \|\xi-\xi_*\| |\mu_1-\mu_2|(d\xi) \leq 2\mathcal{D}_{\mathcal{C}^2}^{p} \|\mu_1-\mu_2\|_{TV}.
\end{aligned}$$

□

D.1. Proofs of Results in Section 4.1

Throughout Section 4, the noise ε_k are assumed to satisfy Assumption 2. Our proof of Theorem 13 relies on the geometric ergodicity and convergence theory of Markov chains. Geometric ergodicity and convergence of Markov chains has been well studied in the literature. Harris' ergodic theorem of Markov chains essentially states that a Markov chain is ergodic if it admits a small set that is visited infinitely often (Harris, 1956). Such a result often relies on finding an appropriate Lyapunov function (Meyn & Tweedie, 1993). The transition probabilities converge exponentially fast towards the unique invariant measure, and the prefactor is controlled by the Lyapunov function (Meyn & Tweedie, 1993). Computable bounds for geometric convergence rates of Markov chains has been obtained in e.g. Meyn & Tweedie (1994); Hairer & Mattingly (2011). In the following, we state the results from Hairer & Mattingly (2011). Before we proceed, let us introduce some definitions and notations.

Let $\mathbb{X}$ be a measurable space and $\mathcal{P}(x,\cdot)$ be a Markov transition kernel on $\mathbb{X}$. For any measurable function $\varphi : \mathbb{X} \to [0,+\infty]$, we define:

$$(\mathcal{P}\varphi)(x) = \int_{\mathbb{X}} \varphi(y)\mathcal{P}(x,dy).$$

Assumption 29 (Drift Condition)**.** *There exists a function $V : \mathbb{X} \to [0,\infty)$ and some constants $K \geq 0$ and $\gamma \in (0,1)$ so that*

$$(\mathcal{P}V)(x) \leq \gamma V(x) + K,$$

for all $x \in \mathbb{X}$.

Assumption 30 (Minorization Condition)**.** *There exists some constant $\eta \in (0,1)$ and a probability measure ν so that*

$$\inf_{x\in\mathbb{X}:V(x)\leq R} \mathcal{P}(x,\cdot) \geq \eta\nu(\cdot),$$

for some $R > 2K/(1-\gamma)$.

Let us recall the definition of the weighted total variation distance:

$$d_\psi(\mu_1,\mu_2) = \int_{\mathbb{X}} (1+\psi V(x))|\mu_1-\mu_2|(dx).$$

It is noted in Hairer & Mattingly (2011) that d_ψ has the following alternative expression. Define the weighted supremum norm for any $\psi > 0$:

$$\|\varphi\|_\psi := \sup_{x\in\mathbb{X}} \frac{|\varphi(x)|}{1+\psi V(x)},$$

and its associated dual metric d_ψ on probability measures:

$$d_\psi(\mu_1,\mu_2) = \sup_{\varphi:\|\varphi\|_\psi\leq 1} \int_{\mathbb{X}} \varphi(x)(\mu_1-\mu_2)(dx).$$

It is also noted in Hairer & Mattingly (2011) that d_ψ can also be expressed as:

$$d_\psi(\mu_1,\mu_2)=\sup_{\varphi:|||\varphi|||_\psi\le 1}\int_{\mathbb{X}}\varphi(x)(\mu_1-\mu_2)(dx),$$

where

$$|||\varphi|||_\psi:=\sup_{x\neq y}\frac{|\varphi(x)-\varphi(y)|}{2+\psi V(x)+\psi V(y)}.$$

Lemma 31 (Theorem 1.3. Hairer & Mattingly (2011))**.** *If the drift condition (Assumption 29) and minorization condition (Assumption 30) hold, then there exists $\bar\eta\in(0,1)$ and $\psi>0$ so that*

$$d_\psi(\mathcal{P}\mu_1,\mathcal{P}\mu_2)\le\bar\eta d_\psi(\mu_1,\mu_2)$$

for any probability measures μ_1,μ_2 on $\mathbb{X}$. In particular, for any $\eta_0\in(0,\eta)$ and $\gamma_0\in(\gamma+2K/R,1)$ one can choose $\psi=\eta_0/K$ and $\bar\eta=(1-(\eta-\eta_0))\vee(2+R\psi\gamma_0)/(2+R\psi)$.

Lemma 32 (Theorem 1.2. Hairer & Mattingly (2011))**.** *If the drift condition (Assumption 29) and minorization condition (Assumption 30) hold, then $\mathcal{P}$ admits a unique invariant measure μ_*, i.e. $\mathcal{P}\mu_*=\mu_*$.*

The drift condition has indeed been obtained in Aybat et al. (2018). The AG method follows the dynamics

$$\xi_{k+1}=A\xi_k+B(\nabla f(y_k)+\varepsilon_{k+1}),\tag{89}$$
$$y_k=C\xi_k,\tag{90}$$

where

$$A:=\begin{pmatrix}(1+\beta)I_d & -\beta I_d\\ I_d & 0_d\end{pmatrix},\quad B:=\begin{pmatrix}-\alpha I_d\\ 0_d\end{pmatrix},\quad C:=\begin{pmatrix}(1+\beta)I_d & -\beta I_d\end{pmatrix}.$$

Define $\tilde y_k:=y_k-x_*$ and $\tilde\xi_k:=\xi_k-\xi_*$, where $\xi_*=A\xi_*$ and $x_*=C\xi_*$. Let us recall the Lyapunov function from (5)

$$V_P(\xi_k)=(\xi_k-\xi_*)^TP(\xi_k-\xi_*)+f(x_k)-f_*,$$

where $\xi_*=(x_*,x_*)$.

Next, let us prove that the drift condition holds. The proof is mainly built on Corollary 4.2. and Lemma 4.5. in Aybat et al. (2018).

Lemma 33.

$$(\mathcal{P}_{\alpha,\beta}V_{P_{\alpha,\beta}})(\xi)\le\gamma_{\alpha,\beta}V_{P_{\alpha,\beta}}(\xi)+K_{\alpha,\beta},$$

where

$$\gamma_{\alpha,\beta}:=\rho_{\alpha,\beta},\qquad K_{\alpha,\beta}:=\left(\frac{L}{2}+\tilde P_{\alpha,\beta}(1,1)\right)\alpha^2\sigma^2.$$

Proof. By Corollary 4.2. and its proof in Aybat et al. (2018) (In Aybat et al. (2018), the noise are assumed to be independent. But a closer look at the proof of Corollary 4.2. reveals that our Assumption 2 suffices), we have

$$\begin{aligned}&\mathbb{E}[V(\xi_{k+1})]-\rho\mathbb{E}[V(\xi_k)]\\&=\mathbb{E}\left[\begin{pmatrix}\tilde\xi_k\\ \nabla f(y_k)\end{pmatrix}^T\begin{pmatrix}A^TPA-\rho P & A^TPB\\ B^TPA & B^TPB\end{pmatrix}\begin{pmatrix}\tilde\xi_k\\ \nabla f(y_k)\end{pmatrix}\right]+\mathbb{E}\left[\varepsilon_{k+1}^TB^TPB\varepsilon_{k+1}\right],\end{aligned}\tag{91}$$

where

$$V(\xi):=(\xi-\xi_*)^TP(\xi-\xi_*).$$

A closer look at the proof of Corollary 4.2. in Aybat et al. (2018) reveals that the following equality also holds:

$$\begin{aligned}&\mathbb{E}[V(\xi_{k+1})|\xi_k]-\rho V(\xi_k)\\&=\begin{pmatrix}\tilde\xi_k\\ \nabla f(y_k)\end{pmatrix}^T\begin{pmatrix}A^TPA-\rho P & A^TPB\\ B^TPA & B^TPB\end{pmatrix}\begin{pmatrix}\tilde\xi_k\\ \nabla f(y_k)\end{pmatrix}+\mathbb{E}\left[\varepsilon_{k+1}^TB^TPB\varepsilon_{k+1}\right].\end{aligned}\tag{92}$$

When $f \in \mathcal{S}_{\mu,L}$ is strongly convex, Lemma 4.5. in Aybat et al. (2018) states that for any $\rho \in (0,1)$,

$$\begin{aligned} &\begin{pmatrix} \tilde{\xi}_k \\ \nabla f(y_k) \end{pmatrix}^T X \begin{pmatrix} \tilde{\xi}_k \\ \nabla f(y_k) \end{pmatrix} \\ &\leq \rho(f(x_k) - f_*) - (f(x_{k+1}) - f_*) + \frac{L\alpha^2}{2}\|\varepsilon_{k+1}\|^2 - \alpha(1 - L\alpha)\nabla f(y_k)^T \varepsilon_{k+1}, \end{aligned} \tag{93}$$

where $X := \rho X_1 + (1-\rho)X_2$, where

$$X_1 := \frac{1}{2}\begin{pmatrix} \beta^2\mu I_d & -\beta^2\mu I_d & -\beta I_d \\ -\beta^2\mu I_d & \beta^2\mu I_d & \beta I_d \\ -\beta I_d & \beta I_d & \alpha(2 - L\alpha) I_d \end{pmatrix}, \tag{94}$$

$$X_2 := \frac{1}{2}\begin{pmatrix} (1+\beta)^2\mu I_d & -\beta(1+\beta)\mu I_d & -(1+\beta) I_d \\ -\beta(1+\beta)\mu I_d & \beta^2\mu I_d & \beta I_d \\ -(1+\beta) I_d & \beta I_d & \alpha(2 - L\alpha) I_d \end{pmatrix}. \tag{95}$$

Taking expectation w.r.t. the noise ε_{k+1} only in (93), we get

$$\begin{pmatrix} \tilde{\xi}_k \\ \nabla f(y_k) \end{pmatrix}^T X \begin{pmatrix} \tilde{\xi}_k \\ \nabla f(y_k) \end{pmatrix} \leq \rho(f(x_k) - f_*) - (f(x_{k+1}) - f_*) + \frac{L\alpha^2}{2}\sigma^2. \tag{96}$$

With the definition of $\rho_{\alpha,\beta}$, $P_{\alpha,\beta}$ by Lemma 21, we get

$$\begin{pmatrix} A^T P_{\alpha,\beta} A - \rho_{\alpha,\beta} P_{\alpha,\beta} & A^T P B \\ B^T P_{\alpha,\beta} A & B^T P_{\alpha,\beta} B \end{pmatrix} - X \preceq 0. \tag{97}$$

Then, combining (92) and (96), applying (97) and the definition of $V_{P_{\alpha,\beta}}$, we get

$$\begin{aligned} \mathbb{E}[V_{P_{\alpha,\beta}}(\xi_{k+1})|\xi_k] &\leq \rho_{\alpha,\beta} V_{P_{\alpha,\beta}}(\xi_k) + \mathbb{E}\left[\varepsilon_{k+1}^T B^T P_{\alpha,\beta} B \varepsilon_{k+1}\right] + \frac{L\alpha^2}{2}\sigma^2 \\ &= \rho_{\alpha,\beta} V_{P_{\alpha,\beta}}(\xi_k) + \mathbb{E}\left[\varepsilon_{k+1}^T \alpha^2 \tilde{P}_{\alpha,\beta}(1,1) I_d \varepsilon_{k+1}\right] + \frac{L\alpha^2}{2}\sigma^2 \\ &\leq \rho_{\alpha,\beta} V_{P_{\alpha,\beta}}(\xi_k) + \alpha^2 \tilde{P}_{\alpha,\beta}(1,1)\sigma^2 + \frac{L\alpha^2}{2}\sigma^2 \end{aligned}$$

It follows that

$$(\mathcal{P}_{\alpha,\beta} V_{P_{\alpha,\beta}})(\xi) \leq \rho_{\alpha,\beta} V_{P_{\alpha,\beta}}(\xi) + \left(\frac{L}{2} + \tilde{P}_{\alpha,\beta}(1,1)\right)\alpha^2\sigma^2.$$

□

In the special case $(\alpha, \beta) = (\alpha_{AG}, \beta_{AG})$, we obtain the following result.

Lemma 34. *Given* $(\alpha, \beta) = (\alpha_{AG}, \beta_{AG})$.

$$(\mathcal{P}_{\alpha,\beta} V_{P_{AG}})(\xi) \leq \gamma V_{P_{AG}}(\xi) + K,$$

where

$$\gamma := \rho_{AG}, \qquad K := \frac{\sigma^2}{L},$$

where $\rho_{AG} = 1 - 1/\sqrt{\kappa}$.

Proof. By letting $(\alpha, \beta) = (\alpha_{AG}, \beta_{AG})$ in Lemma 33, we get

$$(\mathcal{P}_{\alpha,\beta} V_{P_{AG}})(\xi) \leq \gamma V_{P_{AG}}(\xi) + K,$$

where

$$\gamma = \rho_{AG}, \qquad K = \left(\frac{L}{2} + \tilde{P}_{AG}(1,1)\right)\alpha_{AG}^2\sigma^2,$$

where $\rho_{AG} = 1 - 1/\sqrt{\kappa}$ and $\tilde{P}_{AG}(1,1)$ is the $(1,1)$-entry of $\tilde{P}_{AG}$. Notice that

$$\tilde{P}_{AG} = \begin{pmatrix} \sqrt{\frac{L}{2}} \\ \sqrt{\frac{\mu}{2}} - \sqrt{\frac{L}{2}} \end{pmatrix} \begin{pmatrix} \sqrt{\frac{L}{2}} & \sqrt{\frac{\mu}{2}} - \sqrt{\frac{L}{2}} \end{pmatrix},$$

and hence

$$P_{AG} = \tilde{P}_{AG} \otimes I_d = \begin{pmatrix} \frac{L}{2} I_d & \left(\frac{\sqrt{L\mu}}{2} - \frac{L}{2}\right) I_d \\ \left(\frac{\sqrt{\mu L}}{2} - \frac{L}{2}\right) I_d & \frac{(\sqrt{\mu}-\sqrt{L})^2}{2} I_d \end{pmatrix},$$

which implies that $\tilde{P}_{AG}(1,1) = \frac{L}{2}$. □

Next, let us verify the minorization condition. Assume that the noise admits a continuous probability density function, then the Markov transition kernel $\mathcal{P}_{\alpha,\beta}$ also admits a continuous probability density function for x_{k+1} conditional on x_k and x_{k-1}, which we denote by $p(\xi, x)$, that is, $\mathbb{P}(x_{k+1} \in dx | (x_k^T, x_{k-1}^T) = \xi^T) = p(\xi, x)dx$. Also note that when we transit from $(x_k^T, x_{k-1}^T)^T$ to (x_{k+1}, x_k), the value of x_k follows a Dirac delta distribution. We aim to show that for any Borel measurable sets A, B

$$\inf_{(x_k, x_{k-1}) \in \mathbb{R}^{2d} : V_P((x_k, x_{k-1})) \le R} \mathcal{P}((x_k, x_{k-1}), (x_{k+1}, x_k) \in A \times B) \ge \eta \nu_2(A \times B),$$

for some probability measure ν_2. Let us define:

$$B_R := \left\{ x \in \mathbb{R}^d : \exists\, y \in \mathbb{R}^d, V_P(x, y) \le R \right\}.$$

We define ν_2 such that $\nu_2(A \times B) = 0$ for any B that does not contain B_R, and $\nu_2(A \times B) = \nu_1(A)$ for some probability measure ν_1 and for any B that contains B_R.Then, it suffices for us to show that

$$\inf_{\xi \in \mathbb{R}^{2d}, V_P(\xi) \le R} p(\xi, x) \ge \eta \nu(x),$$

where $\nu(x)$ is the probability density function for some probability measure $\nu_1(\cdot)$.

Lemma 35. *For any $\eta \in (0,1)$, there exists some $R > 0$ such that*

$$\inf_{\xi \in \mathbb{R}^{2d}, V_P(\xi) \le R} p(\xi, x) \ge \eta \nu(x).$$

Proof. Let us take:

$$\nu(x) = p(\xi_*, x) \cdot \frac{1_{\|x - x_*\| \le M}}{\int_{\|x - x_*\| \le M} p(\xi_*, x) dx},$$

where $M > 0$ is sufficiently large so that the denominator in the above equation is positive.When $\|x - x_*\| > M$, $\inf_{\xi \in \mathbb{R}^{2d}, V_P(\xi) \le R} p(\xi, x) \ge 0$ automatically holds. Thus, we only need to focus on $\|x - x_*\| \le M$.

Note that for sufficiently large M, $\int_{\|x - x_*\| \le M} p(\xi_*, x) dx$ can get arbitrarily close to 1. Fix M, by the continuity of $p(\xi, x)$ in both ξ and x, we can find $\eta' \in (0,1)$ such that uniformly in $\|x - x_*\| \le M$,

$$\inf_{\xi \in \mathbb{R}^{2d}, V_P(\xi) \le R} p(\xi, x) \ge \eta' p(\xi_*, x) = \eta \nu(x),$$

where we can take

$$\eta := \eta' \int_{\|x - x_*\| \le M} p(\xi_*, x) dx,$$

which can be arbitrarily close to 1 if we take $R > 0$ to be sufficiently small. In particular, if we fix $\eta \in (0,1)$, then we can take $M > 0$ such that

$$\int_{\|x-x_*\| \le M} p(\xi_*, x)dx \ge \sqrt{\eta},$$

and similarly with fixed η and M, we take $R > 0$ such that uniformly in $\|x - x_*\| \le M$,

$$\inf_{\xi \in \mathbb{R}^{2d}, V_P(\xi) \le R} p(\xi, x) \ge \sqrt{\eta} p(\xi_*, x).$$

□

Finally, we are ready to state the proof of Theorem 13 and Proposition 14.

Proof of Theorem 13. According to the proof of Lemma 35, for any fixed $\eta > 0$, we can define:

$$M \ge \inf \left\{ m > 0 : \int_{\|x-x_*\| \le m} p(\xi_*, x)dx = \sqrt{\eta} \right\},$$

and

$$R \le \sup \left\{ r > 0 : \inf_{\xi \in \mathbb{R}^{2d}, V_{P_{\alpha,\beta}}(\xi) \le R} p(\xi, x) \ge \sqrt{\eta} p(\xi_*, x) \text{ for every } \|x - x_*\| \le M \right\}.$$

Then, we have

$$\inf_{\xi \in \mathbb{R}^{2d}, V_{P_{\alpha,\beta}}(\xi) \le R} p(\xi, x) \ge \eta \nu(x).$$

Let us recall that

$$(\mathcal{P}_{\alpha,\beta} V_{P_{\alpha,\beta}})(\xi) \le \gamma_{\alpha,\beta} V_{P_{\alpha,\beta}}(\xi) + K_{\alpha,\beta}.$$

By Lemma 31 and Lemma 32,

$$d_\psi(\nu_{k,\alpha,\beta}, \pi_{\alpha,\beta}) \le \bar{\eta}^k d_\psi(\nu_{0,\alpha,\beta}, \pi_{\alpha,\beta})$$

where $\bar{\eta} = (1 - (\eta - \eta_0)) \vee (2 + R\psi\gamma_0)/(2 + R\psi)$ and $\psi = \eta_0 / K_{\alpha,\beta}$, where $\eta_0 \in (0, \eta)$ and $\gamma_0 \in (\gamma_{\alpha,\beta} + 2K_{\alpha,\beta}/R, 1)$. In particular, we can choose

$$\eta_0 = \frac{\eta}{2}, \qquad \gamma_0 = \frac{1}{2}\gamma_{\alpha,\beta} + \frac{1}{2} + \frac{K_{\alpha,\beta}}{R}.$$

Therefore,

$$\bar{\eta} = \max \left\{ 1 - \frac{\eta}{2}, 1 - \left(\frac{1}{2} - \frac{1}{2}\gamma_{\alpha,\beta} - \frac{K_{\alpha,\beta}}{R} \right) \frac{R\psi}{2 + R\psi} \right\},$$

where $\psi := \frac{\eta}{2K_{\alpha,\beta}}$ so that

$$\bar{\eta} = \max \left\{ 1 - \frac{\eta}{2}, 1 - \left(\frac{1}{2} - \frac{1}{2}\gamma_{\alpha,\beta} - \frac{K_{\alpha,\beta}}{R} \right) \frac{R\eta}{4K_{\alpha,\beta} + R\eta} \right\}.$$

The proof is complete. □

Proof of Proposition 14. Let us recall that $\gamma = \rho = 1 - \frac{1}{\sqrt{\kappa}}$ and $K = \frac{\sigma^2}{L}$. Recall that γ_0 satisfies $\gamma_0 \in (\gamma + 2K/R, 1)$ and let us assume that K is sufficiently small so that $K \le \frac{R}{4\sqrt{\kappa}}$, then we can take

$$\gamma_0 = 1 - \frac{1}{4\sqrt{\kappa}}.$$

We also recall that $\psi = \eta_0 / K$ and

$$\bar{\eta} = \max \left\{ 1 - \eta + \eta_0, \frac{2 + R\psi\gamma_0}{2 + R\psi} \right\} = \max \left\{ 1 - \eta + \eta_0, \frac{K + R\eta_0\gamma_0}{K + R\eta_0} \right\}.$$

We have discussed before that we can take η to be arbitrarily close to 1 by taking M sufficiently large, and for fixed M take R sufficiently small. Let us take

$$\eta = 1 - \rho = \frac{1}{\sqrt{\kappa}}, \qquad \eta_0 = \frac{1}{2}\eta = \frac{1}{2\sqrt{\kappa}},$$

and then

$$1 - \eta + \eta_0 = 1 - \frac{1}{2\sqrt{\kappa}}.$$

If we take $K < R\eta_0 = \frac{R}{2\sqrt{\kappa}}$, then

$$\frac{K + R\eta_0\gamma_0}{K + R\eta_0} \leq 1 - \frac{1}{8\sqrt{\kappa}}.$$

Hence, we can take $K \leq \frac{R}{4\sqrt{\kappa}}$, that is,

$$\sigma^2 \leq \frac{RL}{4\sqrt{\kappa}},$$

so that

$$\bar{\eta} \leq 1 - \frac{1}{8\sqrt{\kappa}}.$$

Finally, we want to take $R > 0$ and $M > 0$ such that

$$\inf_{\xi \in \mathbb{R}^{2d}, V_{P_{AG}}(\xi) \leq R} p(\xi, x) \geq \eta\nu(x) = \frac{\nu(x)}{\sqrt{\kappa}}$$

holds for the choice of

$$\nu(x) = p(\xi_*, x) \cdot \frac{1_{\|x - x_*\| \leq M}}{\int_{\|x - x_*\| \leq M} p(\xi_*, x) dx}.$$

It is easy to see that we can take M so that

$$\int_{\|x - x_*\| \leq M} p(\xi_*, x) dx \geq \frac{1}{\kappa^{1/4}},$$

and take R such that for any $\|x - x_*\| \leq M$,

$$\inf_{\xi \in \mathbb{R}^{2d}, V_{P_{AG}}(\xi) \leq R} p(\xi, x) \geq \frac{1}{\kappa^{1/4}} p(\xi_*, x).$$

Hence, by applying Lemma 31, we conclude that for any two probability measures μ_1, μ_2 on $\mathbb{R}^{2d}$:

$$d_\psi(\mathcal{P}^k_{\alpha,\beta}\mu_1, \mathcal{P}^k_{\alpha,\beta}\mu_2) \leq \left(1 - \frac{1}{8\sqrt{\kappa}}\right)^k d_\psi(\mu_1, \mu_2).$$

Recall that $\nu_{k,\alpha,\beta}$ denotes the law of the iterates ξ_k. By Lemma 32, the Markov chain ξ_k admits a unique invariant distribution $\pi_{\alpha,\beta}$. By letting $\mu_1 = \nu_{0,\alpha,\beta}$ and $\mu_2 = \pi_{\alpha,\beta}$, we conclude that

$$d_\psi(\nu_{k,\alpha,\beta}, \pi_{\alpha,\beta}) \leq \left(1 - \frac{1}{8\sqrt{\kappa}}\right)^k d_\psi(\nu_{0,\alpha,\beta}, \pi_{\alpha,\beta}),$$

where

$$\psi = \frac{\eta_0}{K} = \frac{1}{2\sqrt{\kappa}K} = \frac{L}{2\sqrt{\kappa}\sigma^2}.$$

Finally, let us prove (29). Given $(\alpha, \beta) = (\alpha_{AG}, \beta_{AG})$, we have $\rho_{\alpha,\beta} = 1 - \frac{1}{\sqrt{\kappa}}$, $\alpha = \frac{1}{L}$. It follows from Lemma 34 and its proof that

$$\mathbb{E}[V_{P_{AG}}(\xi_{k+1})] \leq \rho_{AG}\mathbb{E}[V_{P_{AG}}(\xi_k)] + \frac{1}{L}\sqrt{\kappa}\sigma^2.$$

By induction on k, we can show that for every k,

$$\mathbb{E}[V_{P_{AG}}(\xi_{k+1})] \leq V_{P_{AG}}(\xi_0)\rho_{AG}^{k+1} + \frac{1}{L}\sqrt{\kappa}\sigma^2.$$

By the definition of V_P, it follows that

$$\mathbb{E}[f(x_{k+1})] - f(x_*) \leq V_{P_{AG}}(\xi_0)\rho_{AG}^{k+1} + \frac{1}{L}\sqrt{\kappa}\sigma^2 = V_{P_{AG}}(\xi_0)\rho_{AG}^{k+1} + \frac{1}{L}\sqrt{\kappa}\sigma^2.$$

Thus, we get

$$\mathbb{E}[f(x_k)] - f(x_*) \leq V_{P_{AG}}(\xi_0)\left(1 - \frac{1}{\sqrt{\kappa}}\right)^k + \frac{1}{L}\sqrt{\kappa}\sigma^2.$$

The proof is complete. □

Remark 36. *In Proposition 14, the amount of noise that can be tolerated is limited. Nevertheless, in applications where the gradient is estimated from noisy measurements, such results would be applicable if the noise level is mild (Birand et al., 2013).*

Proof of Corollary 15. If the noise ε_k are i.i.d. Gaussian $\mathcal{N}(0, \Sigma)$, then conditional on $x_k = x_{k-1} = x_*$ in the AG method, with stepsize $\alpha = 1/L$, x_{k+1} is distributed as $\mathcal{N}(x_*, L^{-2}\Sigma)$ with $\Sigma \preceq L^2 I_d$. Therefore, for $\gamma > 0$ sufficiently small,

$$\mathbb{E}\left[e^{\gamma\|x_{k+1}-x_*\|^2}\middle| x_k = x_{k-1} = x_*\right] = \frac{1}{\sqrt{\det\left(I_d - 2\gamma L^{-2}\Sigma\right)}}.$$

By Chebychev's inequality, letting $\gamma = 1/2$, for any $m \geq 0$, we get

$$\mathbb{P}\left(\|x_{k+1} - x_*\| \geq m | x_k = x_{k-1} = x_*\right) \leq \frac{e^{-\frac{1}{2}m^2}}{\sqrt{\det(I_d - L^{-2}\Sigma)}}.$$

Hence, we can take

$$M = \left(-2\log\left(\left(1 - \frac{1}{\kappa^{1/4}}\right)\sqrt{\det(I_d - L^{-2}\Sigma)}\right)\right)^{1/2}.$$

Conditional on $(x_k^T, x_{k-1}^T)^T = \xi = (\xi_{(1)}^T, \xi_{(2)}^T)^T$, where $V_P(\xi) \leq r$ for some $r > 0$, then, x_{k+1} is Gaussian distributed:

$$x_{k+1}|(x_k, x_{k-1}) = (\xi_{(1)}, \xi_{(2)}) \sim \mathcal{N}\left(\mu_\xi, L^{-2}\Sigma\right),$$

where

$$\mu_\xi = \frac{2\sqrt{\kappa}}{\sqrt{\kappa}+1}\xi_{(1)} - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\xi_{(2)} - L^{-1}\nabla f\left(\frac{2\sqrt{\kappa}}{\sqrt{\kappa}+1}\xi_{(1)} - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\xi_{(2)}\right). \tag{98}$$

Thus, uniformly in $\|x - x_*\| \leq M$,

$$\frac{p(\xi, x)}{p(\xi_*, x)} = e^{-\frac{1}{2}(x-\mu_\xi)^T L^2\Sigma^{-1}(x-\mu_\xi) + \frac{1}{2}(x-x_*)^T L^2\Sigma^{-1}(x-x_*)}.$$

Note that $V_{P_{AG}}(\xi) \leq r$ implies that

$$\begin{pmatrix} \xi_{(1)} - x_* \\ \xi_{(2)} - x_* \end{pmatrix}^T P_{AG} \begin{pmatrix} \xi_{(1)} - x_* \\ \xi_{(2)} - x_* \end{pmatrix} \leq r.$$

By the definition of P_{AG}, we get

$$\begin{pmatrix} \xi_{(1)} - x_* \\ \xi_{(2)} - x_* \end{pmatrix}^T \begin{pmatrix} \sqrt{\frac{L}{2}} I_d \\ \left(\sqrt{\frac{\mu}{2}} - \sqrt{\frac{L}{2}}\right) I_d \end{pmatrix} \begin{pmatrix} \sqrt{\frac{L}{2}} I_d \\ \left(\sqrt{\frac{\mu}{2}} - \sqrt{\frac{L}{2}}\right) I_d \end{pmatrix}^T \begin{pmatrix} \xi_{(1)} - x_* \\ \xi_{(2)} - x_* \end{pmatrix} \leq r,$$

so that

$$\frac{L}{2}\|\xi_{(1)} - x_*\|^2 + \frac{(\sqrt{\mu} - \sqrt{L})^2}{2}\|\xi_{(2)} - x_*\|^2 \le r,$$

which implies that

$$\|\xi_{(1)} - x_*\| \le \frac{\sqrt{2r}}{\sqrt{L}}, \qquad \|\xi_{(2)} - x_*\| \le \frac{\sqrt{2r}}{\sqrt{L} - \sqrt{\mu}}.$$

Moreover,

$$\begin{aligned}
\mu_\xi - x_* &= \frac{2\sqrt{\kappa}}{\sqrt{\kappa}+1}\xi_{(1)} - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\xi_{(2)} - L^{-1}\nabla f\left(\frac{2\sqrt{\kappa}}{\sqrt{\kappa}+1}\xi_{(1)} - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\xi_{(2)}\right) \\
&\quad - \left(\frac{2\sqrt{\kappa}}{\sqrt{\kappa}+1}x_* - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}x_* - L^{-1}\nabla f\left(\frac{2\sqrt{\kappa}}{\sqrt{\kappa}+1}x_* - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}x_*\right)\right) \\
&= \frac{2\sqrt{\kappa}}{\sqrt{\kappa}+1}(\xi_{(1)} - x_*) - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}(\xi_{(2)} - x_*) \\
&\quad - L^{-1}\left(\nabla f\left(\frac{2\sqrt{\kappa}}{\sqrt{\kappa}+1}\xi_{(1)} - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\xi_{(2)}\right) - \nabla f\left(\frac{2\sqrt{\kappa}}{\sqrt{\kappa}+1}x_* - \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}x_*\right)\right).
\end{aligned}$$

Since ∇f is L-Lipschitz,

$$\begin{aligned}
\|\mu_\xi - x_*\| &\le (1 + L^{-1}L)\frac{2\sqrt{\kappa}}{\sqrt{\kappa}+1}\|\xi_{(1)} - x_*\| + (1 + L^{-1}L)\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\|\xi_{(2)} - x_*\| \\
&\le 2\frac{2\sqrt{\kappa}}{\sqrt{\kappa}+1}\frac{\sqrt{2r}}{\sqrt{L}} + 2\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\frac{\sqrt{2r}}{\sqrt{L}-\sqrt{\mu}} \\
&\le 2\frac{2\sqrt{\kappa}}{\sqrt{\kappa}+1}\frac{\sqrt{2r}}{\sqrt{L}-\sqrt{\mu}} + 2\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\frac{\sqrt{2r}}{\sqrt{L}-\sqrt{\mu}} \\
&= 2\frac{3\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\frac{\sqrt{2r}}{\sqrt{L}-\sqrt{\mu}}.
\end{aligned} \tag{99}$$

Thus, uniformly in $\|x - x_*\| \le M$,

$$\begin{aligned}
\frac{p(\xi, x)}{p(\xi_*, x)} &= \exp\left\{-\frac{1}{2}(x-\mu_\xi)^T L^2 \Sigma^{-1}(x-\mu_\xi) + \frac{1}{2}(x-x_*)^T L^2\Sigma^{-1}(x-x_*)\right\} \\
&\ge \exp\left\{-\frac{1}{2}\|\mu_\xi - x_*\|L^2\|\Sigma^{-1}\|(\|x-\mu_\xi\| + \|x-x_*\|)\right\} \\
&\ge \exp\left\{-\frac{1}{2}\|\mu_\xi - x_*\|L^2\|\Sigma^{-1}\|(\|\mu_\xi - x_*\| + 2\|x-x_*\|)\right\} \\
&\ge \exp\left\{-\frac{1}{2}L^2\|\Sigma^{-1}\|(\|\mu_\xi - x_*\|^2 + 2M\|\mu_\xi - x_*\|)\right\} \ge \frac{1}{\kappa^{1/4}},
\end{aligned}$$

if we have

$$\|\mu_\xi - x_*\| \le -M + \sqrt{M^2 + \frac{\log(\kappa)}{2L^2\|\Sigma^{-1}\|}}. \tag{100}$$

Combining (99) and (100), we can take

$$\begin{aligned}
R &= \frac{1}{8}\left(-M + \sqrt{M^2 + \frac{\log(\kappa)}{2L^2\|\Sigma^{-1}\|}}\right)^2 \frac{(\sqrt{\kappa}+1)^2(\sqrt{L}-\sqrt{\mu})^2}{(3\sqrt{\kappa}-1)^3} \\
&= \left(-M + \sqrt{M^2 + \frac{\log(L/\mu)}{2L^2\|\Sigma^{-1}\|}}\right)^2 \frac{(L-\mu)^2}{8(3\sqrt{L}-\sqrt{\mu})^3}.
\end{aligned}$$

For the remaining of the proof, without loss of generality assume that $\mu = \Theta(1)$ and $L = \Theta(\kappa)$.[2] It is straightforward to see from the Taylor expansion of M that $M = O(\kappa^{-1/8})$ and

$$\begin{aligned} R &= \frac{\left(\frac{\log(L/\mu)}{2L^2\|\Sigma^{-1}\|}\right)^2}{\left(M + \sqrt{M^2 + \frac{\log(L/\mu)}{2L^2\|\Sigma^{-1}\|}}\right)^2} \frac{(L-\mu)^2}{8(3\sqrt{L} - \sqrt{\mu})^3} \\ &= O\left(\frac{1}{M^2}\left(\frac{\log(L/\mu)}{2L^2\|\Sigma^{-1}\|}\right)^2 \frac{(L-\mu)^2}{8(3\sqrt{L} - \sqrt{\mu})^3}\right) \\ &= O\left(\kappa^{-13/4}\log^2(\kappa)\right). \end{aligned}$$

□

D.2. Proofs of Results in Section A

Consider the constrained optimization problem

$$\min_{x\in\mathcal{C}} f(x),$$

where $\mathcal{C} \subset \mathbb{R}^d$ is compact. The projected AG method consists of the iterations

$$\tilde{x}_{k+1} = \mathcal{P}_{\mathcal{C}}\left(\tilde{y}_k - \alpha(\nabla f(\tilde{y}_k) + \varepsilon_{k+1})\right), \tag{101}$$

$$\tilde{y}_k = (1+\beta)\tilde{x}_k - \beta\tilde{x}_{k-1}, \tag{102}$$

where ε_k is the random gradient error satisfying Assumption 2, $\alpha, \beta > 0$ are the stepsize and momentum parameter and the projection onto the convex compact set C with diameter $\mathcal{D}_{\mathcal{C}}$ can be written as

$$\mathcal{P}_{\mathcal{C}}(x) := \arg\min_{y\in\mathbb{R}^d}\left(\frac{1}{2\alpha}\|x-y\|^2 + h(y)\right)$$

where the function $h : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is the indicator function, defined to be zero if $y \in \mathcal{C}$ and infinity otherwise. Let us recall that we assumed that the random gradient error ε_k admits a continuous density so that conditional on $\tilde{\xi}_k = (\tilde{x}_k^T, \tilde{x}_{k-1}^T)^T$, $\tilde{x}_{k+1}$ also admits a continuous density, i.e.

$$\mathbb{P}(\tilde{x}_{k+1} \in d\tilde{x} | \tilde{\xi}_k = \tilde{\xi}) = \tilde{p}(\tilde{\xi}, \tilde{x})d\tilde{x},$$

where $\tilde{p}(\tilde{\xi}, \tilde{x}) > 0$ is continuous in both $\tilde{\xi}$ and $\tilde{x}$.

For the function $f(x)$, the gradient mapping $g : \mathbb{R}^d \to \mathbb{R}$ which replaces the gradient for constrained optimization problems is defined as

$$g(y) = \frac{1}{\alpha}\left(y - \mathcal{P}_{\mathcal{C}}(y - \alpha\nabla f(y))\right), \quad \alpha > 0.$$

Due to the noise in the gradients, we also define the perturbed gradient mapping, $g_\varepsilon(y) : \mathbb{R}^d \to \mathbb{R}$ as

$$g_\varepsilon(y) = \frac{1}{\alpha}\left(y - \mathcal{P}_{\mathcal{C}}\big(y - \alpha(\nabla f(y) + \varepsilon)\big)\right), \quad \alpha > 0, \quad \varepsilon \in \mathbb{R}^d.$$

Due to the non-expansiveness property of the projection operator, we have (see e.g. Combettes & Wajs (2005, Lemma 2.4))

$$\Delta_\varepsilon(y) := g_\varepsilon(y) - g(y), \quad \|\Delta_\varepsilon(y)\|^2 \le \|\varepsilon\|^2, \quad \text{for every } y \in \mathbb{R}^d. \tag{103}$$

Following a similar approach to Hu & Lessard (2017); Fazlyab et al. (2017), we reformulate the projected AG iterations as a linear dynamical system as

$$\begin{aligned} \tilde{x}_{k+1} &= (1+\beta)\tilde{x}_k - \beta\tilde{x}_{k-1} - \alpha g_{\epsilon_{k+1}}(\tilde{y}_k), \\ \tilde{y}_k &= (1+\beta)\tilde{x}_k - \beta\tilde{x}_{k-1}, \end{aligned}$$

[2] Given two scalar-valued functions f and g, we say $f = \Theta(g)$, if the ratio $f(x)/g(x)$ lies in an interval $[c_1, c_2]$ for every x and some$c_1, c_2 > 0$.

which is equivalent to

$$\tilde{\xi}_{k+1} = A\tilde{\xi}_k + B\tilde{u}_k, \tag{104}$$
$$\tilde{y}_k = C\xi_k, \quad \tilde{x}_k = E\tilde{\xi}_k, \tag{105}$$
$$\tilde{u}_k = g(\tilde{y}_k) + \Delta_{\varepsilon_{k+1}}(\tilde{y}_k), \tag{106}$$

with $\tilde{\xi}_k = [\tilde{x}_k^T \ \tilde{x}_{k-1}^T]^T$, and

$$A = \begin{pmatrix} (1+\beta)I_d & -\beta I_d \\ I_d & 0_d \end{pmatrix}, \quad B = \begin{pmatrix} -\alpha I_d \\ 0_d \end{pmatrix}, \quad C = \begin{pmatrix} (1+\beta)I_d & -\beta I_d \end{pmatrix}, \quad E = \begin{pmatrix} I_d & 0_d \end{pmatrix}. \tag{107}$$

We see that $\tilde{\xi}_k$ forms a time-homogeneous Markov chain. To this chain, we can associate a Markov kernel $\tilde{\mathcal{P}}_{\alpha,\beta}$, following a similar approach to the Markov kernel $\mathcal{P}_{\alpha,\beta}$ we defined for AG. We have the following result.

Lemma 37.

$$(\tilde{\mathcal{P}}_{\alpha,\beta} V_{P_{\alpha,\beta}})(\tilde{\xi}) \leq \rho_{\alpha,\beta} V_{P_{\alpha,\beta}}(\tilde{\xi}) + \tilde{K}_{\alpha,\beta},$$

where

$$\tilde{K}_{\alpha,\beta} := \alpha\sigma(2\mathcal{D}_{\mathcal{C}}\|P_{\alpha,\beta}\| + G_M) + \alpha^2\sigma^2\left(\|P_{\alpha,\beta}\| + \frac{L}{2}\right),$$

if there exists a matrix $P_{\alpha,\beta} \in \mathbb{R}^{2d\times 2d}$ such that

$$-\rho_{\alpha,\beta}X_1 - (1-\rho_{\alpha,\beta})X_2 + X_3 \preceq 0, \tag{108}$$

where

$$X_1 = \frac{1}{2}\begin{pmatrix} \beta^2\mu I_d & -\beta^2\mu I_d & -\beta I_d \\ -\beta^2\mu I_d & \beta^2\mu I_d & \beta I_d \\ -\beta I_d & \beta I_d & \alpha(2-L\alpha)I_d \end{pmatrix}, \ X_2 = \frac{1}{2}\begin{pmatrix} (1+\beta)^2\mu I_d & -\beta(1+\beta)\mu I_d & -(1+\beta)I_d \\ -\beta(1+\beta)\mu I_d & \beta^2\mu I_d & \beta I_d \\ -(1+\beta)I_d & \beta I_d & \alpha(2-L\alpha)I_d \end{pmatrix},$$

and

$$X_3 = \begin{pmatrix} A^T P_{\alpha,\beta}A - \tilde{\rho}_{\alpha,\beta}P_{\alpha,\beta} & A^T P_{\alpha,\beta}B \\ B^T P_{\alpha,\beta}A & B^T P_{\alpha,\beta}B \end{pmatrix},$$

where $G_M := \max_{x\in\mathcal{C}} \|\nabla f(x)\|$.

In particular, with $\rho = 1 - \frac{1}{\sqrt{\kappa}}$, $\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$, $\alpha = \frac{1}{L}$ where $\kappa = \frac{L}{\mu}$. Then (108) holds with the matrix

$$P = \frac{\mu}{2}\begin{pmatrix} (1-\sqrt{\kappa})I_d & \sqrt{\kappa}I_d \end{pmatrix}^T \begin{pmatrix} (1-\sqrt{\kappa})I_d & \sqrt{\kappa}I_d \end{pmatrix}.$$

Proof. We follow the proof technique of Fazlyab et al. (2017) for deterministic proximal AG which is based on Nesterov (2004, Lemma 2.4) and adapt this proof technique to accelerated stochastic projected gradient. Defining the error at step k

$$\tilde{e}_k := [(\tilde{\xi}_k - \tilde{\xi}_*)^T \ (g(\tilde{y}_k) - g(\tilde{y}_*))^T]^T,$$

where $\tilde{\xi}_* := [x_*^T \ x_*^T]^T$ and $g(\tilde{y}_*) = 0$ due to the first order optimality conditions where $\tilde{y}_* := \tilde{x}_*$ is the unique minimum of f over C. Let $\mathcal{F}_k$ be the natural filtration for the iterations of the algorithm until and including step k so that x_k, y_k and $\tilde{e}_k$

are $\mathcal{F}_k$-measurable. Similar to the analysis of AG, we estimate

$$\begin{aligned}
\mathbb{E}\left[f\left(\tilde{x}_{k+1}\right)-f\left(\tilde{x}_k\right)\middle|\mathcal{F}_k\right] &= \mathbb{E}\left[f\left(\tilde{y}_k-\alpha g_{\varepsilon_{k+1}}\left(\tilde{y}_k\right)\right)-f\left(\tilde{x}_k\right)\middle|\mathcal{F}_k\right] && (109)\\
&= \mathbb{E}\left[f\left(\tilde{y}_k-\alpha g\left(\tilde{y}_k\right)-\alpha\Delta_{\varepsilon_{k+1}}\left(\tilde{y}_k\right)\right)-f\left(\tilde{x}_k\right)\middle|\mathcal{F}_k\right] && (110)\\
&\leq \mathbb{E}\Big[f\left(\tilde{y}_k-\alpha g\left(\tilde{y}_k\right)\right)+\nabla f\left(\tilde{y}_k-\alpha g\left(\tilde{y}_k\right)\right)^T\alpha\Delta_{\varepsilon_{k+1}}\left(\tilde{y}_k\right) && (111)\\
&\qquad +\frac{\alpha^2 L}{2}\|\Delta_{\varepsilon_{k+1}}(\tilde{y}_k)\|^2-f\left(\tilde{x}_k\right)\Big|\mathcal{F}_k\Big] && (112)\\
&\leq f\left(\tilde{y}_k-\alpha g\left(\tilde{y}_k\right)\right)-f\left(\tilde{x}_k\right)+\mathbb{E}\left[\alpha G_M\|\Delta_{\varepsilon_{k+1}}(\tilde{y}_k)\|+\frac{\alpha^2 L}{2}\|\varepsilon_{k+1}\|^2\middle|\mathcal{F}_k\right] && (113)\\
&\leq f\left(\tilde{y}_k-\alpha g\left(\tilde{y}_k\right)\right)-f\left(\tilde{x}_k\right)+\alpha G_M\sigma+\frac{\alpha^2 L}{2}\sigma^2, && (114)
\end{aligned}$$

where in the first inequality we used the fact that the gradient of f is L-smooth which implies that

$$f(y)-f(z)\leq\nabla f(z)^T(y-z)+\frac{L}{2}\|y-z\|^2,\quad \text{for every } y,z\in\mathbb{R}^d$$

(see e.g. (Bubeck, 2014)) and second inequality follows from Jensen's inequality.Finally, the last step is a consequence of (103) and Assumption 2 on the noise. It follows from a similar computation that

$$\mathbb{E}\left[f(\tilde{x}_{k+1})-f(\tilde{x}_*)\middle|\mathcal{F}_k\right]\leq f\big(\tilde{y}_k-\alpha g(\tilde{y}_k)\big)-f(\tilde{x}_*)+\alpha G_M\sigma+\frac{\alpha^2 L}{2}\sigma^2. \tag{115}$$

We note that the matrices X_1 and X_2 can be written as

$$X_1=\frac{-1}{2}\begin{pmatrix}-\mu(C-E)^T(C-E) & (C-E)^T\\ C-E & (L\alpha^2-2\alpha)I_d\end{pmatrix},\quad X_2 = \frac{-1}{2}\begin{pmatrix}-\mu C^TC & C^T\\ C & (L\alpha^2-2\alpha)I_d\end{pmatrix}, \tag{116}$$

where A, B, C, E are defined by (107). Using Fazlyab et al. (2017, eqn. (36)–(37)) and Lemma 38, we have

$$\begin{aligned}
f\big(\tilde{y}_k-\alpha g(\tilde{y}_k)\big)-f(\tilde{x}_k)&\leq-\tilde{e}_k^TX_1\tilde{e}_k, && (117)\\
f\big(\tilde{y}_k-\alpha g(\tilde{y}_k)\big)-f(\tilde{x}_*)&\leq-\tilde{e}_k^TX_2\tilde{e}_k. && (118)
\end{aligned}$$

Plugging these into (114) and (115), we obtain

$$\begin{aligned}
\mathbb{E}\left[f(\tilde{x}_{k+1})-f(\tilde{x}_k)\middle|\mathcal{F}_k\right] &\leq -\tilde{e}_k^TX_1\tilde{e}_k+\alpha G_M\sigma+\frac{\alpha^2 L}{2}\sigma^2, && (119)\\
\mathbb{E}\left[f(\tilde{x}_{k+1})-f(\tilde{x}_*)\middle|\mathcal{F}_k\right] &\leq -\tilde{e}_k^TX_2\tilde{e}_k+\alpha G_M\sigma+\frac{\sigma^2 L}{2}\sigma^2. && (120)
\end{aligned}$$

It also follows from (104)– (106) and the facts that $A\tilde{\xi}_*=\tilde{\xi}_*$ and $B\tilde{u}_*=0$ that

$$\tilde{\xi}_{k+1}-\tilde{\xi}_*=A\left(\tilde{\xi}_k-\tilde{\xi}_*\right)+B\left(\tilde{u}_k-\tilde{u}_*\right)+B\Delta_{\varepsilon_{k+1}}(\tilde{y}_k)=\zeta_k+B\Delta_{\varepsilon_{k+1}}(\tilde{y}_k), \tag{121}$$

where

$$\zeta_k:=A\left(\tilde{\xi}_k-\tilde{\xi}_*\right)+B\left(\tilde{u}_k-\tilde{u}_*\right).$$

For any symmetric positive semi-definite matrix $P_{\alpha,\beta}\in\mathbb{R}^{2d\times 2d}$, we define the quadratic function

$$Q_{P_{\alpha,\beta}}(\tilde{\xi})=\tilde{\xi}^TP_{\alpha,\beta}\tilde{\xi}.$$

We can estimate that

$$\begin{aligned}
&\mathbb{E}\left[Q_{P_{\alpha,\beta}}\left(\tilde{\xi}_{k+1}\right)|\mathcal{F}_k\right] \\
&= \mathbb{E}\left[\left(\tilde{\xi}_{k+1}-\tilde{\xi}_*\right)^T P_{\alpha,\beta}\left(\tilde{\xi}_{k+1}-\tilde{\xi}_*\right)|\mathcal{F}_k\right] \\
&= \zeta_k^T P_{\alpha,\beta}\zeta_k^T + \mathbb{E}\left[2(\tilde{\xi}_{k+1}-\tilde{\xi}_*)^T P_{\alpha,\beta}B\Delta_{\varepsilon_{k+1}}(\tilde{y}_k) + B^T\Delta_{\varepsilon_{k+1}}(\tilde{y}_k)^T P_{\alpha,\beta}B\Delta_{\varepsilon_{k+1}}(\tilde{y}_k)\big|\mathcal{F}_k\right] \\
&\leq \tilde{e}_k^T\begin{pmatrix}A^TP_{\alpha,\beta}A & A^TP_{\alpha,\beta}B\\ B^TP_{\alpha,\beta}A & B^TP_{\alpha,\beta}B\end{pmatrix}\tilde{e}_k + \mathbb{E}\Big[2\alpha\mathcal{D}_{\mathcal{C}}\cdot\|P_{\alpha,\beta}\|\cdot\|\varepsilon_{k+1}\| + \alpha^2\|P_{\alpha,\beta}\|\cdot\|\varepsilon_{k+1}\|^2|\mathcal{F}_k\Big] \\
&= \tilde{e}_k^T\begin{pmatrix}A^TP_{\alpha,\beta}A & A^TP_{\alpha,\beta}B\\ B^TP_{\alpha,\beta}A & B^TP_{\alpha,\beta}B\end{pmatrix}\tilde{e}_k + 2\mathcal{D}_{\mathcal{C}}\alpha\sigma\|P_{\alpha,\beta}\| + \alpha^2\sigma^2\|P_{\alpha,\beta}\|.
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[Q_{P_{\alpha,\beta}}\left(\tilde{\xi}_{k+1}\right) - Q_{P_{\alpha,\beta}}\left(\tilde{\xi}_k\right)\Big|\mathcal{F}_k\right] = \tilde{e}_k^T X_3\tilde{e}_k + 2\mathcal{D}_{\mathcal{C}}\alpha\sigma\|P_{\alpha,\beta}\| + \alpha^2\sigma^2\|P_{\alpha,\beta}\|. \tag{122}$$

Considering the Lyapunov function $V_{P_{\alpha,\beta}}(\tilde{\xi}_k) = f(\tilde{x}_k) - f(\tilde{x}_*) + \tilde{\xi}_k^T P_{\alpha,\beta}\tilde{\xi}_k$, we have

$$\begin{aligned}
V_{P_{\alpha,\beta}}\left(\tilde{\xi}_{k+1}\right) - \tilde{\rho}_{\alpha,\beta}V_{P_{\alpha,\beta}}\left(\tilde{\xi}_k\right) &= \tilde{\rho}_{\alpha,\beta}\left(f\left(\tilde{\xi}_{k+1}\right) - f\left(\tilde{\xi}_*\right)\right) + (1-\tilde{\rho}_{\alpha,\beta})\left(f\left(\tilde{\xi}_{k+1}\right) - f\left(\tilde{\xi}_*\right)\right) && (123)\\
&\quad + Q_{P_{\alpha,\beta}}\left(\tilde{\xi}_{k+1}-\tilde{\xi}_*\right) - Q_{P_{\alpha,\beta}}\left(\tilde{\xi}_k-\tilde{\xi}_*\right). && (124)\\
& && (125)
\end{aligned}$$

Taking conditional expectations and inserting (119)–(120),

$$\begin{aligned}
\mathbb{E}\left[V_{P_{\alpha,\beta}}\left(\tilde{\xi}_{k+1}\right)\Big|\mathcal{F}_k\right] &\leq \tilde{\rho}_{\alpha,\beta}V_{P_{\alpha,\beta}}\left(\tilde{\xi}_k\right) + \tilde{e}_k^T\Big(-\tilde{\rho}_{\alpha,\beta}X_1 - (1-\tilde{\rho}_{\alpha,\beta})X_2 + X_3\Big)\tilde{e}_k && (126)\\
&\quad + 2\mathcal{D}_{\mathcal{C}}\alpha\sigma\|P_{\alpha,\beta}\| + \alpha^2\sigma^2\left(\|P_{\alpha,\beta}\| + \frac{L}{2}\right) && (127)\\
&\leq \tilde{\rho}_{\alpha,\beta}V_{P_{\alpha,\beta}}\left(\tilde{\xi}_k\right) + \alpha\sigma(2\mathcal{D}_{\mathcal{C}}\|P_{\alpha,\beta}\| + G_M) + \alpha^2\sigma^2\left(\|P_{\alpha,\beta}\| + \frac{L}{2}\right), && (128)
\end{aligned}$$

which completes the proof. □

Lemma 38 (Fazlyab et al. 2017)**.** *Using the notations as in the proof of Lemma 37, we have the following two inequalities:*

$$f\big(\tilde{y}_k - \alpha g(\tilde{y}_k)\big) - f(\tilde{x}_k) \leq -\tilde{e}_k^T X_1\tilde{e}_k, \tag{129}$$
$$f\big(\tilde{y}_k - \alpha g(\tilde{y}_k)\big) - f(\tilde{x}_*) \leq -\tilde{e}_k^T X_2\tilde{e}_k. \tag{130}$$

Proof. Recall that f satisfies following inequalities,

$$f(z) - f(y) \leq \nabla f(y)^T(z-y) + \frac{L}{2}\|y-z\|^2, \tag{131}$$
$$f(y) - f(x) \leq \nabla f(y)^T(y-x) - \frac{\mu}{2}\|y-x\|^2. \tag{132}$$

Choosing $z = \tilde{y}_k - \alpha g(\tilde{y}_k),\ y = \tilde{y}_k$ and $x = \tilde{x}_k$ yields,

$$f(y_k - \alpha g(y_k)) - f(x_k) \leq \nabla f(y_k)^T\big(y_k - x_k - \alpha g(y_k)\big) + \frac{L}{2}\|\alpha g(y_k)\|^2 - \frac{\mu}{2}\|y_k - x_k\|^2. \tag{133}$$

Additionally let $\partial h(x) := \{v \in \mathbb{R}^d : h(x) - h(y) \leq v^T(x-y)\forall y \in \mathbb{R}^d\}$ then by optimality condition, $0 \in \partial(\mathcal{P}_C(w)) - \frac{1}{\alpha}(\mathcal{P}_C(w) - w)$ (e.g. (Beck, 2017) theorem 6.39). In particular there exists a $T_h(w) \in \partial h(x)$ such that $g(w) = \nabla f(w) + T_h(w)$. Choose $w = y_k$ and note that $y_k = (1+\beta)x_k - \beta x_{k-1}$ and C is a convex set thus $y_k \in C$. So

if $T_h(y_k) \in \partial h(y_k)$ then either $0 \leq T_h(y_k)^T(y_k - x)$ or $-\infty \leq T_h(y_k)^T(y_k - x)$ therefore $0 \leq T_h(y_k)^T(y_k - x)$ implying that $\nabla f(y)^T(y - z) \leq g(y)^T(y - x)$ for all $x \in \mathbb{R}^d$. Combining this result with (133) we obtain,

$$\begin{aligned} f(y_k - \alpha g(y_k)) - f(x_k) &\leq \nabla f(y_k)^T\big(y_k - x_k - \alpha g(y_k)\big) + \frac{L}{2}\alpha^2\|g(y_k)\|^2 - \frac{\mu}{2}\beta^2\|x_k - x_{k-1}\|^2 \\ f(y_k - \alpha g(y_k)) - f(x_k) &\leq \beta g(y_k)^T(x_k - x_{k-1}) + \left(\frac{L}{2}\alpha^2 - \alpha\right)\|g(y_k)\|^2 \\ &\quad -\frac{\mu}{2}\beta^2\left(\|x_k - x_*\|^2 - 2(x_k - x_*)^T(x_{k-1} - x_*) + \|x_{k-1} - x_*\|^2\right). \end{aligned}$$

This proves (117). Finally, (118) can also be obtained if we take $x = x_*$ and follow similar steps. □

Lemma 39. *Given $\alpha = \frac{1}{L}$, $\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$, where $\kappa = L/\mu$, we have*

$$(\tilde{\mathcal{P}}_{\alpha,\beta}V_{P_{\alpha,\beta}})(\tilde{\xi}) \leq \tilde{\gamma}V_{P_{\alpha,\beta}}(\tilde{\xi}) + \tilde{K},$$

where

$$\tilde{\gamma} := 1 - \frac{1}{\sqrt{\kappa}}, \qquad \tilde{K} := \frac{\sigma}{L}\left(\mathcal{D}_{\mathcal{C}}\mu((1-\sqrt{\kappa})^2 + \kappa) + G_M\right) + \frac{\sigma^2}{L^2}\left(\frac{\mu}{2}((1-\sqrt{\kappa})^2 + \kappa) + \frac{L}{2}\right).$$

Proof. Note that

$$(\tilde{\mathcal{P}}_{\alpha,\beta}V_{P_{\alpha,\beta}})(\tilde{\xi}) \leq \tilde{\rho}_{\alpha,\beta}V_{P_{\alpha,\beta}}(\tilde{\xi}) + \tilde{K}_{\alpha,\beta},$$

where

$$\tilde{K}_{\alpha,\beta} := \alpha\sigma(2\mathcal{D}_{\mathcal{C}}\|P_{\alpha,\beta}\| + G_M) + \alpha^2\sigma^2\left(\|P_{\alpha,\beta}\| + \frac{L}{2}\right),$$

and with $\alpha = \frac{1}{L}$, $\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$, we have

$$P_{\alpha,\beta} = \frac{\mu}{2}\begin{pmatrix}(1-\sqrt{\kappa})I_d & \sqrt{\kappa}I_d\end{pmatrix}^T\begin{pmatrix}(1-\sqrt{\kappa})I_d & \sqrt{\kappa}I_d\end{pmatrix},$$

so that

$$\|P_{\alpha,\beta}\| \leq \frac{\mu}{2}\left\|\begin{pmatrix}(1-\sqrt{\kappa})I_d & \sqrt{\kappa}I_d\end{pmatrix}^T\right\| \cdot \left\|\begin{pmatrix}(1-\sqrt{\kappa})I_d & \sqrt{\kappa}I_d\end{pmatrix}\right\| = \frac{\mu}{2}((1-\sqrt{\kappa})^2 + \kappa).$$

Hence,

$$\tilde{K}_{\alpha,\beta} \leq \frac{\sigma}{L}\left(\mathcal{D}_{\mathcal{C}}\mu((1-\sqrt{\kappa})^2 + \kappa) + G_M\right) + \frac{\sigma^2}{L^2}\left(\frac{\mu}{2}((1-\sqrt{\kappa})^2 + \kappa) + \frac{L}{2}\right).$$

□

Proof of Theorem 16. The proof is similar to the proof of Theorem 13 and the proof of (29). We obtain

$$\mathbb{E}[f(\tilde{x}_k)] - f(\tilde{x}_*) \leq V_{P_{\alpha,\beta}}(\tilde{\xi}_0)\tilde{\gamma}_{\alpha,\beta}^k + \frac{\tilde{K}_{\alpha,\beta}}{1 - \tilde{\gamma}_{\alpha,\beta}}.$$

The conclusion then follows from the defintiion of $\tilde{\gamma}_{\alpha,\beta}$ and $\tilde{K}_{\alpha,\beta}$. □

Proof of Proposition 17. The proof is similar as the proof of Proposition 14. We can take $\tilde{K} \leq \frac{R}{4\sqrt{\kappa}}$, that is,

$$\frac{\sigma}{L}\left(\mathcal{D}_{\mathcal{C}}\mu((1-\sqrt{\kappa})^2 + \kappa) + G_M\right) + \frac{\sigma^2}{L^2}\left(\frac{\mu}{2}((1-\sqrt{\kappa})^2 + \kappa) + \frac{L}{2}\right) \leq \frac{R}{4\sqrt{\kappa}},$$

which implies

$$\sigma \leq \frac{-b_1}{2a_1} + \frac{1}{2a_1}\sqrt{b_1^2 + a_1\frac{R}{\sqrt{\kappa}}},$$

where

$$a_1 = \frac{1}{L^2}\left(\frac{\mu}{2}((1-\sqrt{\kappa})^2+\kappa)+\frac{L}{2}\right), \qquad b_1 = \frac{1}{L}\left(\mathcal{D}_{\mathcal{C}}\mu((1-\sqrt{\kappa})^2+\kappa)+G_M\right).$$

As in the proof of Proposition 14, we can take

$$\tilde{\psi} = \frac{1}{2\sqrt{\kappa}\tilde{K}}.$$

Finally, the proof of (35) is similar as the proof of (33). We obtain

$$\mathbb{E}[f(\tilde{x}_k)] - f(\tilde{x}_*) \leq V_{P_{AG}}(\tilde{\xi}_0)\tilde{\gamma}^k + \frac{\tilde{K}}{1-\tilde{\gamma}}.$$

The conclusion then follows from the definition of $\tilde{K}$ and $\tilde{\gamma}$. □

E. Numerical Illustrations

In this section, we illustrate some of our theoretical results over some simple functions with numerical experiments. On the left panel of Figure 1, we compare ASG for the quadratic objective $f(x) = x^2/2$ in dimension one with additive i.i.d. Gaussian noise on the gradients for different noise levels $\sigma \in \{0.01, 0.1, 1, 2\}$. The plots show performance with respect to expected suboptimality using 10^4 sample paths. As expected, the performance deteriorates when σ increases. The fact that the performance stabilizes after a certain number of iterations supports the claim that a stationary distribution exists, a claim that was proved in Theorem 4. In the middle panel, we repeat the experiment in dimension $d = 10$ over the quadratic objective $f(x) = \frac{1}{2}x^T Qx$, where Q is a diagonal matrix with diagonal entries $Q_{ii} = 1/i$. We observe similar patterns.

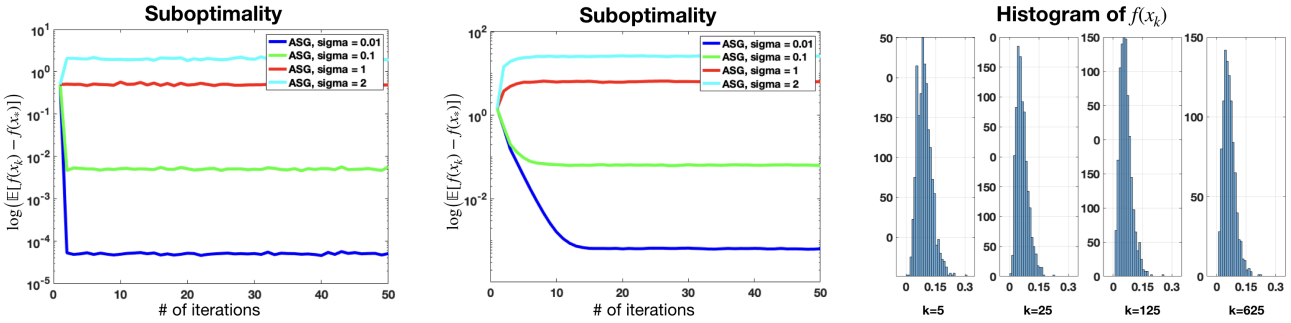

Figure 1. Performance comparison of ASG for different noise levels σ on quadratic functions. *Left panel*: $f(x) = \frac{1}{2}x^2$ in dimension one. *Middle panel*: $f(x) = \frac{1}{2}x^T Qx$ in dimension $d = 10$. *Right panel*: Histogram of $f(x_k)$ for different values of k where $f(x) = \frac{1}{2}x^T Qx$ in dimension $d = 10$.

Finally, on the right panel of Figure 1, we estimate the distribution of $f(x_k)$ for $k \in \{5, 25, 125, 625\}$. For this purpose, we plot the histograms of $f(x_k)$ over 10^4 sample paths for every fixed k. We observe that the histograms for $k = 125$ and 625 are similar, illustrating the fact that ASG admits a stationary distribution.