On the Convergence of Stochastic Extragradient for Bilinear Games using Restarted Iteration Averaging

Chris Junchi Li[◇,⋆] **Yaodong Yu**[◇,⋆] **Nicolas Loizou**[◁] **Gauthier Gidel**[†,‡]
Yi Ma[◇] **Nicolas Le Roux**[†,‡,§,□] **Michael I. Jordan**[◇]

University of California, Berkeley[◇] Johns Hopkins University[◁] Mila[†]
Université de Montréal[‡] McGill University[§] Microsoft Research[□]

Abstract

We study the stochastic bilinear minimax optimization problem, presenting an analysis of the same-sample Stochastic ExtraGradient (SEG) method with constant step size, and presenting variations of the method that yield favorable convergence. In sharp contrasts with the basic SEG method whose last iterate only contracts to a fixed neighborhood of the Nash equilibrium, SEG augmented with iteration averaging provably converges to the Nash equilibrium under the same standard settings, and such a rate is further improved by incorporating a scheduled restarting procedure. In the interpolation setting where noise vanishes at the Nash equilibrium, we achieve an optimal convergence rate up to tight constants. We present numerical experiments that validate our theoretical findings and demonstrate the effectiveness of the SEG method when equipped with iteration averaging and restarting.

1 INTRODUCTION

The *minimax optimization* framework provides solution concepts useful in game theory [Morgenstern and Von Neumann, 1944], statistics [Bach, 2019] and online learning [Blackwell, 1956, Cesa-Bianchi and Lugosi, 2006]. It has recently been prominent in the deep learning community due to its application to generative modeling [Goodfellow et al., 2014, Arjovsky et al., 2017] and robust prediction [Madry et al., 2018, Zhang et al., 2019a]. There remains, however, a gap between minimax characterizations of solutions and algorithmic frameworks that provably converge to such solutions in practice.

In standard single-objective machine learning applications, the traditional algorithmic realization of optimization frameworks is stochastic gradient descent (SGD, or one of its variants), where the full gradient is formulated as an expectation over the data-generating mechanism. In general minimax optimization problems, however, naive use of SGD leads to pathological behavior due to the presence of rotational dynamics [Goodfellow, 2016, Balduzzi et al., 2018].

One way to overcome these rotations is to use gradient-based methods specifically designed for the minimax setting (or more generally for the multi-player game setting). A key example of such a method is the celebrated *extragradient method*. Originally introduced by [Korpelevich, 1976], it addresses general minimax optimization problems and yields optimal convergence guarantees in the batch setting [Azizian et al., 2020b]. In the stochastic setting, however, it has only been analyzed in special cases, such as the constrained case [Juditsky et al., 2011], the bounded-noise case [Hsieh et al., 2020], and the interpolatory case [Vaswani et al., 2019b]. In the current paper, we study the general stochastic bilinear minimax optimization problem, also known as the bilinear saddle-point problem,

$$\min_{\mathbf{x}} \max_{\mathbf{y}} \ \mathbf{x}^\top \mathbb{E}_\xi[\mathbf{B}_\xi]\mathbf{y} + \mathbf{x}^\top \mathbb{E}_\xi[\mathbf{g}_\xi^{\mathbf{x}}] + \mathbb{E}_\xi[(\mathbf{g}_\xi^{\mathbf{y}})^\top]\mathbf{y}\,, \quad (1)$$

where the index ξ denotes the randomness associated with stochastic sampling. Following standard practice we assume that the expected coupling matrix $\mathbf{B} = \mathbb{E}[\mathbf{B}_\xi]$ is nonsingular, and that the intercept vectors $\mathbf{g}_\xi^{\mathbf{x}}$ and $\mathbf{g}_\xi^{\mathbf{y}}$ have zero mean: $\mathbb{E}[\mathbf{g}_\xi^{\mathbf{x}}] = \mathbf{0}_n$ and $\mathbb{E}[\mathbf{g}_\xi^{\mathbf{y}}] = \mathbf{0}_m$. Thus the Nash equilibrium point is $[\mathbf{x}^*; \mathbf{y}^*] = [\mathbf{0}_n; \mathbf{0}_m]$. Such assumptions are standard in the literature on bilinear optimization [see, e.g., Vaswani et al., 2019b,

[⋆] Equal contribution. Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

Mishchenko et al., 2020].[1]

In this work, we present theoretical results in the general setting of bilinear minimax games for a version of the Stochastic ExtraGradient (SEG) method that incorporates iteration averaging and scheduled restarting. The introduction of stochasticity in the matrix $\mathbf{B}_\xi$ together with an unbounded domain presents technical challenges that have been a major stumbling block in earlier work [cf. Dieuleveut et al., 2016]. Here we show how to surmount these challenges. Formally, we introduce the following SEG method composed of an extrapolation step (half-iterates) and an update step:

$$\begin{aligned}
\mathbf{x}_{t-1/2} &= \mathbf{x}_{t-1} - \eta_t \left[\mathbf{B}_{\xi,t}\mathbf{y}_{t-1} + \mathbf{g}^{\mathbf{x}}_{\xi,t}\right], \\
\mathbf{y}_{t-1/2} &= \mathbf{y}_{t-1} + \eta_t \left[\mathbf{B}^\top_{\xi,t}\mathbf{x}_{t-1} + \mathbf{g}^{\mathbf{y}}_{\xi,t}\right], \\
\mathbf{x}_{t} &= \mathbf{x}_{t-1} - \eta_t \left[\mathbf{B}_{\xi,t}\mathbf{y}_{t-1/2} + \mathbf{g}^{\mathbf{x}}_{\xi,t}\right], \\
\mathbf{y}_{t} &= \mathbf{y}_{t-1} + \eta_t \left[\mathbf{B}^\top_{\xi,t}\mathbf{x}_{t-1/2} + \mathbf{g}^{\mathbf{y}}_{\xi,t}\right].
\end{aligned} \tag{2}$$

Here and throughout we adopt a ***same-sample-and-step-size*** notation in which the extrapolation and extragradient steps share the same stochastic sample [Gidel et al., 2019, Mishchenko et al., 2020] and step size η_t; i.e., the updates in Eq. (2) use the same samples of $\mathbf{B}_\xi$, $\mathbf{g}^{\mathbf{x}}_\xi$ and $\mathbf{g}^{\mathbf{y}}_\xi$. Note that there exist counterexamples [see, e.g., Chavdarova et al., 2019, Theorem 1] where the SEG iteration [Juditsky et al., 2011] persistently diverges when using independent samples. The same-sample stochastic extra gradient (SEG) method aims to address this issue [Gidel et al., 2019, Mishchenko et al., 2020]. In practice, for the bilinear game problems we consider in this paper as well as other application problems, including generative adversarial networks and adversarial training, it is easy to perform the same-sample SEG updates: in most machine learning applications one can re-use a sample without significant extra cost.

Main contributions. We provide an in-depth study of SEG on bilinear games and we show that, unlike in the minimization-only setting, in the minimax optimization setting the last-iterate SEG algorithm with the same sample and step sizes *cannot* converge in general even when the step sizes are diminishing to zero [Theorems A.1 and A.2]. This motivates our study of averaging and restarting in order to obtain meaningful convergence rates:

(i) We prove that in the bilinear game setting, under mild assumptions, iteration averaging allows SEG to converge at the rate of $1/\sqrt{K}$ [Theorem 3.1], K being the number of samples the algorithm has processed. This rate is statistically optimal up to a constant multiplier. Additionally, we can further boost the convergence rate when we combine iteration averaging with scheduled restarting [Theorem 3.2] when the lower bound of the smallest eigenvalue in the coupling matrix is known to the system. In this case, exponential forgetting of the initialization and an optimal statistical rate are achieved.

(ii) In the special case of the interpolation setting, we are able to show that SEG with iteration averaging and scheduled restarting achieves an accelerated rate of convergence, faster than (last-iterate) SEG [Theorem 3.3], reducing the dependence of the rate on the condition number to a dependence on its square root. We achieve state-of-the-art rates comparable to the full batch optimal rate [Azizian et al., 2020b], with access only to a stochastic estimate of the gradient, improving upon Vaswani et al. [2019b].

(iii) We provide the first convergence result on SEG with unbounded noise. The only existing result of which we are aware of for the unbounded noise setting is the work of Vaswani et al. [2019b] in the interpolation setting. Our theoretical results are further validated by experiments on synthetic data.

1.1 Related Work

Bilinear minimax optimization. The study of the bilinear example as a tool to understand minimax optimization originated with Daskalakis et al. [2018], who studied an optimistic gradient descent-ascent (OGDA) algorithm to solve that minimax problem. They were able to prove sublinear convergence for this method. Later, Mokhtari et al. [2020] proposed to analyze OGDA and the related ExtraGradient (EG) method as perturbations of the Proximal Point (PP) method. They were able to prove a linear convergence rate for both EG and OGDA with an iteration complexity of $O(\kappa \log(1/\epsilon))$, where $\kappa \equiv \lambda_{\max}(\mathbf{B}^\top\mathbf{B})/\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)$ is the condition number of problem Eq. (1). Highly related to the current work is that of Gidel et al. [2019], who studied the bilinear case and proved an $O(\kappa \log(1/\epsilon))$ iteration complexity for EG with a better constant than Mokhtari et al. [2020]. Wei et al. [2021] studied Optimistic Multiplicative Weights Update (OMWU) for solving constrained bilinear games and established the linear last-iterate convergence.

Regarding optimal methods, a combination of Ibrahim et al. [2020] and Zhang et al. [2019b] established a gen-

[1] In the case of a square, nonsingular coupling matrix $\mathbf{B}$ this assumption is feasible without loss of generality, while in the rectangular matrix case we simply restrict ourselves to the nonsingular $\min(n, m)$-dimensional subspace induced by singular value decomposition. The nonzero component of the intercept vectors $[\mathbf{g}^{\mathbf{x}}_\xi; \mathbf{g}^{\mathbf{y}}_\xi]$ projected onto such a subspace is not taken into account in the SEG dynamics.

eral lower bound, which specializes to a lower bound of $\Omega(\sqrt{\kappa}\log(1/\epsilon))$ for the case of the bilinear minimax game setting. Azizian et al. [2020b] proved linear convergence results for a series of algorithms that achieve this lower bound and also provided an alternative proof for this lower bound by using spectral arguments. However, Azizian et al. [2020b] did not provide accelerated rates for OGDA and provided an accelerated rate for EG with momentum but with an unknown constant. In this work, we completely close that gap by providing accelerated convergence rates for (stochastic) EG with relatively tight constants. In another work, Azizian et al. [2020a] proved a full-regime result for EG without momentum where they show that the $O(\kappa\log(1/\epsilon))$ iteration complexity for EG is optimal among the methods using a fixed number of composed gradient evaluations and only the last iterate (excluding momentum and restarting). A similar iteration complexity, (with an unknown constant) can be derived from the seminal work by Tseng [1995] on EG.

Stochastic bilinear minimax and variational inequalities. The standard assumptions made in the literature on stochastic variational inequalities [Nemirovski et al., 2009, Juditsky et al., 2011] is that the set of parameters and the variance of the stochastic estimate of the vector field are bounded. These two assumptions do not hold in the stochastic bilinear case, because it is unconstrained and the noise increases with the norm of the parameters. Recently, Hsieh et al. [2020] provided results on stochastic EG with different step sizes, without the bounded domain assumption but still requiring the bounded noise assumption. Iusem et al. [2017] and Bot et al. [2019] studied the independent-sample, minibatch setting where the summation of inverse batchsize converges. Mishchenko et al. [2020] discussed how using the same mini-batch for the two gradients in stochastic EG gives stronger guarantees. Using a Hamiltonian viewpoint, Loizou et al. [2020] provided the first set of global non-asymptotic last-iterate convergence guarantees for a stochastic game over a non-compact domain, in the absence of strong monotonicity assumptions. In particular, their stochastic Hamiltonian gradient methods come with last-iterate convergence guarantees in the finite-sum stochastic bilinear game as well. In our work, we provide an accelerated convergence rate for EG in the bilinear setting with unbounded domain and unbounded noise.

Restarting and acceleration. Restarting has long been introduced as an effective approach to accelerate first-order methods in the optimization literature [O'Donoghue and Candes, 2015, Roulet and d'Aspremont, 2020, Renegar and Grimmer, 2021]. In particular, O'Donoghue and Candes [2015] proposed an adaptive restarting technique that significantly improves the convergence rate of Nesterov's accelerated gradient descent method. Roulet and d'Aspremont [2020] developed optimal restarting methods for solving convex optimization problems that satisfy the sharpness assumption. Renegar and Grimmer [2021] considered a more general set of problems than Roulet and d'Aspremont [2020] and presented a simple and near-optimal restarting scheme. Our variant restarting achieves acceleration via a fundamentally different idea that is inspired by modern variance-reduction ideas.

Averaging in convex-concave games. Golowich et al. [2020] studied the effect of averaging for EG in the smooth convex-concave setting. They showed that the last iterate converges at a rate of $O(1/\sqrt{K})$ in terms of the square root of the Hamiltonian (and also the duality gap), while it is known that iteration averaging enjoys an $O(1/K)$ rate [Nemirovski, 2004]. A tight lower bound was also proved to justify an assertion of optimality in the last-iterate setting. Such a result provides a convincing argument in favor of restarting the algorithm from an average of the iterates. This is a theme that we pursue in the current paper.

Stability of limit points in minimax games. GDA dynamics often encounter limit cycles or non-Nash stable limiting points [Daskalakis and Panageas, 2018, Adolphs et al., 2019, Berard et al., 2020, Mazumdar et al., 2019]. To mitigate this, Adolphs et al. [2019] and Mazumdar et al. [2019] proposed to exploit the curvature associated with the stable limit points that are not Nash equilibria. While appealing theoretically, such methods generally involve costly inversion of Jacobian matrices at each step.

Over-parameterized models and interpolation. Recently it was shown that popular stochastic gradient methods, like SGD and its momentum variants, converge considerably faster when the underlying model is sufficiently over-parameterized as to interpolate the data [Gower et al., 2019, 2021, Loizou and Richtárik, 2020, Vaswani et al., 2019a, Loizou et al., 2021b, Sebbouh et al., 2020]. In the minimax optimization setting, an analysis that also covers the interpolation regime is rare. To the best of our knowledge the only paper that provides convergence guarantees for SEG in this setting is Vaswani et al. [2019b], where SEG with line search is proposed and analyzed. In our work we provide convergence guarantees in the interpolation regime as corollaries of our main theorems but with a tight $1/e$-prefactor in the linear convergence.

Organization. The remainder of this paper is organized as follows. §2 details the basic setup and assumptions for our main results. §3 presents our convergence results for SEG with averaging and restarting.

§4 provides experiments that validate our theory. §5 concludes this paper with future directions. All technical analyses along with auxiliary results are relegated to later sections in the supplementary materials.

Notation. Throughout this paper we use the following notation. For two real symmetric matrices, $\mathbf{B}_1, \mathbf{B}_2$, we denote $\mathbf{B}_1 \preceq \mathbf{B}_2$ when $\mathbf{v}^\top \mathbf{B}_1 \mathbf{v} \le \mathbf{v}^\top \mathbf{B}_2 \mathbf{v}$ holds for all vectors $\mathbf{v}$. Let $\lambda_{\max}(\mathbf{B})$ (resp. $\lambda_{\min}(\mathbf{B})$) be the largest (resp. smallest) eigenvalue of a generic (real symmetric) matrix $\mathbf{B}$. Let $\|\mathbf{B}\|_{op}$ denotes the operator norm of $\mathbf{B}$. Let $\mathcal{F}_t$ be the filtration generated by the stochastic samples, $\mathbf{B}_{\xi,s}, \mathbf{g}_{\xi,s}$, $s = 1, \ldots, t$, in the bilinear game. Let $\max(a, b)$ or $a \vee b$ denote the maximum value of $a, b \in \mathbb{R}$, and let $\min(a; b)$ or $a \wedge b$ denote the minimum. For two real sequences, (a_n) and (b_n), we write $a_n = O(b_n)$ to mean that $|a_n| \le C b_n$ for a positive, numerical constant C, for all $n \ge 1$, and let $a_n = \tilde{O}(b_n)$ mean that $|a_n| \le C b_n$ where C hides a logarithmic factor in relevant parameters. We also denote $\widehat{\mathbf{M}}_\xi \equiv \mathbf{B}_\xi^\top \mathbf{B}_\xi$ and $\mathbf{M}_\xi \equiv \mathbf{B}_\xi \mathbf{B}_\xi^\top$ for brevity, each being positive semi-definite for each realization of ξ. Finally, let $[n] = \{1, \ldots, n\}$ for n being a natural number.

2 SETUP FOR MAIN RESULTS

In this section, we introduce the basic setup and assumptions needed for our statement of the convergence of the stochastic extragradient (SEG) algorithm. We first make the following assumptions on $\mathbf{B}_\xi$. Let us recall that $\widehat{\mathbf{M}} \equiv \mathbb{E}_\xi \widehat{\mathbf{M}}_\xi \equiv \mathbb{E}_\xi[\mathbf{B}_\xi^\top \mathbf{B}_\xi]$ and $\mathbf{M} \equiv \mathbb{E}_\xi \mathbf{M}_\xi \equiv \mathbb{E}_\xi[\mathbf{B}_\xi \mathbf{B}_\xi^\top]$.

Assumption 2.1 (Assumption on $\mathbf{B}_\xi$) *Denote $\mathbf{B} = \mathbb{E}_\xi[\mathbf{B}_\xi]$ for $\mathbf{B} \in \mathbb{R}^{n \times m}$ and impose the following regularity conditions: $\lambda_{\max}(\mathbf{B}^\top \mathbf{B}) > 0$ and $\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}}) > 0$. We assume that there exist $\sigma_\mathbf{B}, \sigma_{\mathbf{B},2} \in [0, \infty)$ such that*

$$\begin{aligned} \|\mathbb{E}_\xi[(\mathbf{B}_\xi - \mathbf{B})^\top (\mathbf{B}_\xi - \mathbf{B})]\|_{op} &\le \sigma_\mathbf{B}^2, \\ \|\mathbb{E}_\xi\left[(\mathbf{B}_\xi - \mathbf{B})(\mathbf{B}_\xi - \mathbf{B})^\top\right]\|_{op} &\le \sigma_\mathbf{B}^2, \end{aligned} \tag{3}$$

and

$$\begin{aligned} \|\mathbb{E}_\xi[\mathbf{B}_\xi^\top \mathbf{B}_\xi - \widehat{\mathbf{M}}]^2\|_{op} &\le \sigma_{\mathbf{B},2}^2, \\ \|\mathbb{E}_\xi[\mathbf{B}_\xi \mathbf{B}_\xi^\top - \mathbf{M}]^2\|_{op} &\le \sigma_{\mathbf{B},2}^2. \end{aligned} \tag{4}$$

The assumption of $n \ge m$ (i.e. $\mathbf{B}$ is tall) is without loss of generality; we can convert the SEG iterates with a wide coupling matrix to that of its transpose. Note also $\sigma_\mathbf{B} = 0$ corresponds to the nonrandom $\mathbf{B}_\xi = \mathbf{B}$ case. The stochasticity introduced in $\mathbf{B}_\xi$ allows us to conclude the first convergence result under the unbounded noise condition.[2] Next we impose an assumption on the intercept vector $\mathbf{g}_\xi$.

Assumption 2.2 (Assumption on $\mathbf{g}_\xi$) *There exists a $\sigma_\mathbf{g} \in [0, \infty)$ such that*

$$\mathbb{E}_\xi\left[\|\mathbf{g}_\xi^\mathbf{x}\|^2 + \|\mathbf{g}_\xi^\mathbf{y}\|^2\right] \le \sigma_\mathbf{g}^2 < \infty.$$

Furthermore, we let $\mathbb{E}_\xi[\mathbf{g}_\xi^\mathbf{x}] = \mathbf{0}_n$, $\mathbb{E}_\xi[\mathbf{g}_\xi^\mathbf{y}] = \mathbf{0}_m$ and assume independence between the stochastic matrix $\mathbf{B}_\xi$ and the vector $[\mathbf{g}_\xi^\mathbf{x}; \mathbf{g}_\xi^\mathbf{y}]$.

We remark that the independence assumption in Assumption 2.2 significantly simplifies our analysis.[3] In particular, it ensures $\mathbb{E}[\mathbf{B}_\xi \mathbf{g}_\xi^\mathbf{y}] = \mathbf{0}_n$ and $\mathbb{E}[\mathbf{B}_\xi^\top \mathbf{g}_\xi^\mathbf{x}] = \mathbf{0}_m$, so the Nash equilibrium is the equilibrium point that the last-iterate SEG oscillates around. The independence structure of $\mathbf{B}_\xi$ and $[\mathbf{g}_\xi^\mathbf{x}; \mathbf{g}_\xi^\mathbf{y}]$ in Assumption 2.2 is crucial for our analysis, which is satisfied in certain statistical models. Specially, when one of the $\mathbf{B}_\xi$ and $[\mathbf{g}_\xi^\mathbf{x}; \mathbf{g}_\xi^\mathbf{y}]$ is nonrandom this is always satisfied. Our analysis can be further generalized to more relaxed assumptions on zero correlation between $[\mathbf{g}_\xi^\mathbf{x}; \mathbf{g}_\xi^\mathbf{y}]$ and the first three moments of $\mathbf{B}_\xi$, with a second-moment condition similar to $\mathbb{E}_\xi[\|\mathbf{B}_\xi \mathbf{g}_\xi^\mathbf{y}\|^2 + \|\mathbf{B}_\xi^\top \mathbf{g}_\xi^\mathbf{x}\|^2] \le C(\lambda_{\max}(\mathbf{M}) \vee \lambda_{\max}(\widehat{\mathbf{M}}))\sigma_\mathbf{g}^2$. We defer the full development of this extension to future work. With Assumptions 2.1 and 2.2 at hand, we are ready to state our main results on the convergence of SEG variants.

3 SEG WITH AVERAGING AND RESTARTING

Recall that in contrast to SGD theory in convex optimization, the last iterate of SEG does *not* converge to an arbitrarily small neighborhood of the Nash equilibrium even for the case of a converging step size [Hsieh et al., 2020]. We accordingly turn to an analysis of the *averaged iterate* of $\mathbf{x}_t$ and $\mathbf{y}_t$, $t = 0, 1, \ldots, K$, denoted as

$$\bar{\mathbf{x}}_K \equiv \frac{1}{K+1}\sum_{t=0}^{K} \mathbf{x}_t, \quad \bar{\mathbf{y}}_K \equiv \frac{1}{K+1}\sum_{t=0}^{K} \mathbf{y}_t. \tag{5}$$

For simplicity we focus on the case in which $\mathbf{B}_\xi, \mathbf{B}$ are square matrices. Let us define $\eta_\mathbf{M}$ as follows, which is the maximal step size that the SEG algorithm analysis takes:

$$\eta_\mathbf{M} \equiv \frac{1}{\sqrt{\rho_1 \vee \rho_2}}, \tag{6}$$

[2] As a comparison, Hsieh et al. [2020] only provides a proof for the bounded noise case.

[3] In practice, such independence can be *approximately* achieved via the following decoupling argument: we formulate the random Jacobian-vector product and the random intercept using two independent random samples, separately. Note an approximate knowledge of the Nash equilibrium is required in this decoupling argument.

where $\rho_1 = \lambda_{\max}\big(\mathbf{M}^{-1/2}[\mathbb{E}_\xi \mathbf{M}_\xi^2]\mathbf{M}^{-1/2}\big)$ and $\rho_2 = \lambda_{\max}\big(\widehat{\mathbf{M}}^{-1/2}[\mathbb{E}_\xi \widehat{\mathbf{M}}_\xi^2]\widehat{\mathbf{M}}^{-1/2}\big)$. We introduce the following variants:

$$\begin{aligned}\hat{\eta}_{\mathbf{M}}(\alpha) &\equiv \frac{\eta_{\mathbf{M}}}{\sqrt{2}} \wedge \frac{\alpha\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}{2\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}},\\ \bar{\eta}_{\mathbf{M}}(\alpha) &\equiv \eta_{\mathbf{M}} \wedge \frac{\alpha\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}{2\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}},\end{aligned} \tag{7}$$

which reduce to $1/\sqrt{2\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}$ and $1/\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}$ when $\mathbf{B}_\xi$ is nonrandom. We state our first main result on SEG with iteration averaging, Theorem 3.1, whose proof is provided in §D.1:

Theorem 3.1 (SEG Averaged Iterate) *Let Assumptions 2.1 and 2.2 hold with $n = m$. Prescribing an $\alpha \in (0, 1)$, when the step size η is chosen as $\hat{\eta}_{\mathbf{M}}(\alpha)$ as defined in Eq. (7), we have for all $K \geq 1$ the following convergence bound for the averaged iterate:*

$$\begin{aligned}&\mathbb{E}\big[\|\bar{\mathbf{x}}_K\|^2 + \|\bar{\mathbf{y}}_K\|^2\big]\\ &\leq \tau_1 \cdot \frac{\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2}{(K+1)^2} + \tau_2 \cdot \frac{\sigma_{\mathbf{g}}^2}{K+1},\end{aligned} \tag{8}$$

where τ_1, τ_2 depending on $\sigma_{\mathbf{B}}, \sigma_{\mathbf{B},2}$ are defined as

$$\begin{aligned}\tau_1 &= \frac{16 + 8\kappa_\zeta}{(1-\alpha)\hat{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)},\\ \tau_2 &= \frac{18 + 12\kappa_\zeta}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)},\end{aligned}$$

and $\kappa_\zeta \equiv \frac{\sigma_{\mathbf{B}}^2 + \hat{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2}{\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})}$ denotes the effective noise condition number of problem Eq. (1).

Measured by the Euclidean metric, Theorem 3.1 indicates an $O(1/\sqrt{K})$ leading-order convergence rate for the averaged iterate of SEG in the general stochastic setting, which is known to be statistically optimal up to a constant multiplier. We provide detailed comparisons with previous related work in §B. Nevertheless, the iteration slowly forgets initial conditions at a polynomial rate, and this result can be improved if we utilize a restarting scheme and take advantage of the knowledge of the smallest eigenvalue of $\mathbf{B}\mathbf{B}^\top$. Indeed, in the following result, we boost the convergence rate shown in Eq. (8), when the smallest eigenvalue $\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)$ is available to the system, via a novel restarting procedure at specific times. The rationale behind this analysis is akin to that used in boosting sublinear convergence in convex optimization to linear convergence when the designer has (an estimate of) the strong convexity parameter.

We now develop this argument in detail. We continue to assume the case of square matrices $\mathbf{B}_\xi, \mathbf{B}$. In Algorithm 1 we run SEG with averaging and restart the iteration at chosen timestamps, $\{\mathcal{T}_i\}_{i\in[\mathsf{Epoch}-1]} \subseteq [K]$, initializing at the averaged iterate of the previous epoch. The principle behind our choice of parameters in this algorithm is that we trigger the restarting when the expected squared Euclidean metric $\mathbb{E}\left[\|\mathbf{x}_K\|^2 + \|\mathbf{y}_K\|^2\right]$ decreases by a factor of $1/e^2$, and we halt the restarting procedure once the last iterate reaches stationarity in squared Euclidean metric in the sense of Theorem A.1:[4]

$$\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2 \approx \frac{3\sigma_{\mathbf{g}}^2}{\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})}.$$

Algorithm 1 Iteration Averaged SEG with Scheduled Restarting

```
Require: Initialization $\mathbf{x}_0$, step sizes $\eta_t$, total number of iterates $K$, restarting timestamps $\{\mathcal{T}_i\}_{i\in[\mathsf{Epoch}-1]} \subseteq [K]$ with the total number of epoches $\mathsf{Epoch} \geq 1$, index $s \leftarrow 0$
for $t = 1, 2, \ldots, K$ do
    $s \leftarrow s + 1$
    Update $\mathbf{x}_t$, $\mathbf{y}_t$ via Eq. (2)
    Update $\hat{\mathbf{x}}_t$, $\hat{\mathbf{y}}_t$ via
        $\hat{\mathbf{x}}_t \leftarrow \frac{s-1}{s}\hat{\mathbf{x}}_{t-1} + \frac{1}{s}\mathbf{x}_t, \quad \hat{\mathbf{y}}_t \leftarrow \frac{s-1}{s}\hat{\mathbf{y}}_{t-1} + \frac{1}{s}\mathbf{y}_t$
    if $t \in \{\mathcal{T}_i\}_{i\in[\mathsf{Epoch}-1]}$ then
        Overload $\mathbf{x}_t \leftarrow \hat{\mathbf{x}}_t$, $\mathbf{y}_t \leftarrow \hat{\mathbf{y}}_t$, and set $s \leftarrow 0$
        //restarting procedure is triggered
    end if
end for
Output: $\hat{\mathbf{x}}_K, \hat{\mathbf{y}}_K$
```

Given these choices, summarized in Algorithm 1, we obtain the following theorem:

Theorem 3.2 (SEG with Averaging/Restarting) *Let Assumptions 2.1 and 2.2 hold with $n = m$. For any prescribed $\alpha \in (0, 1)$, choose the step size $\hat{\eta}_{\mathbf{M}}(\alpha)$ as in Eq. (7) and assume a proper restarting schedule. For all $K \geq K_{\text{complexity}} + 1$ we have the following convergence bound for the output $\hat{\mathbf{x}}_K, \hat{\mathbf{y}}_K$ of Algorithm 1:*

$$\mathbb{E}\left[\|\hat{\mathbf{x}}_K\|^2 + \|\hat{\mathbf{y}}_K\|^2\right] \leq C_1 \cdot \frac{\sigma_{\mathbf{g}}^2}{K - K_{\text{complexity}} + 1}, \tag{9}$$

where

$$C_1 \equiv \frac{18}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \left[1 + \underbrace{\frac{O(\sigma_{\mathbf{B}}^2 + \hat{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2)}{\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})}}_{\textit{higher-order term } O(\kappa_\zeta)}\right],$$

[4] The choice of the discount factor $1/e^2$ is to be consistent with our optimal choice in the interpolation setting, where in the $\sigma_{\mathbf{B}} = 0$ case the total complexity is minimized to $e\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})/\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}$.

where $K_{\text{complexity}}$ is the fixed burn-in complexity *defined as*

$$\frac{\textit{logarithmic factor}}{\frac{1}{e}\sqrt{(1-\alpha)\bar{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} - C_2}, \tag{10}$$

with C_2 being

$$O\Big(\bar{\eta}_{\mathbf{M}}(\alpha)^{3/2}(\lambda_{\min}(\mathbf{B}\mathbf{B}^\top))^{1/4}\sqrt{\sigma_{\mathbf{B}}^2 + \bar{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2}\Big).$$

The proof of Theorem 3.2 is provided in §D.2. Here we not only achieve the optimal $O(1/\sqrt{K})$ convergence rate for the averaged iterate, but the proper restarting schedule allows us to achieve a convergence rate bound for iteration-averaged SEG in Eq. (9) that forgets the initialization at an exponential rate instead of the polynomial rate that is obtained without restarting [cf. Theorem 3.2].

Finally, we consider the interpolation setting, where the noise vanishes at the Nash equilibrium. That is, $\mathbf{g}_\xi^{\mathbf{x}} = \mathbf{0}_n$ and $\mathbf{g}_\xi^{\mathbf{y}} = \mathbf{0}_m$; i.e. $\sigma_{\mathbf{g}} = 0$ in Assumption 2.2. In that setting, we prove that SEG with iteration averaging achieves an accelerated linear convergence rate. Set the (constant) interval length of restarting timestamps $K_{\text{thres}}(\alpha)$ as

$$\frac{2}{\frac{1}{e}\sqrt{(1-\alpha)\bar{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} - C_3}, \tag{11}$$

with C_3 being

$$O\left(\bar{\eta}_{\mathbf{M}}(\alpha)^{3/2}(\lambda_{\min}(\mathbf{B}\mathbf{B}^\top))^{1/4}\sqrt{\sigma_{\mathbf{B}}^2 + \bar{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2}\right).$$

We present an analysis of this algorithm in the following theorem, which can be seen as a corollary of Theorem 3.2 but benefits from a refined analysis where tight constant prefactor sits in each term of the bound:

Theorem 3.3 (Interpolation Setting) *Let Assumptions 2.1 and 2.2 hold with $n = m$ and $\sigma_{\mathbf{g}} = 0$. For any prescribed $\alpha \in (0,1)$ choosing the step size $\eta = \bar{\eta}_{\mathbf{M}}(\alpha)$ as in Eq. (7) and the restarting timestamps $\mathcal{T}_i = i \cdot K_{\text{thres}}(\alpha)$ where $K_{\text{thres}}(\alpha)$ was defined as in Eq. (11), we conclude for all $K \geq 1$ that is divisible by $K_{\text{thres}}(\alpha)$ the following convergence rate for the output $\hat{\mathbf{x}}_K, \hat{\mathbf{y}}_K$ of Algorithm 1:*

$$\begin{aligned}&\mathbb{E}\left[\|\hat{\mathbf{x}}_K\|^2 + \|\hat{\mathbf{y}}_K\|^2\right] \\ &\leq e^{-\frac{K}{e}\sqrt{(1-\alpha)\bar{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}+C_4}\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right],\end{aligned} \tag{12}$$

with C_4 being

$$O\left(K\bar{\eta}_{\mathbf{M}}(\alpha)^{3/2}(\lambda_{\min}(\mathbf{B}\mathbf{B}^\top))^{1/4}\sqrt{\sigma_{\mathbf{B}}^2 + \bar{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2}\right).$$

The proof of Theorem 3.3 is provided in §D.3. The idea behind Theorem 3.3 is, in plain words, to trigger restarting whenever the last-iterate SEG has travelled through a full cycle, giving insights on the design of $K_{\text{thres}}(\alpha)$ in the restarting mechanism. Compared with Eq. (13) in Theorem A.1 with $\sigma_{\mathbf{g}}$ equal to zero, the contraction rate (in terms of its exponent) to the Nash equilibrium $-\frac{\eta_{\mathbf{M}}^2}{4} \cdot \left(\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})\right)$ improves to $-\frac{1}{e}\sqrt{(1-\alpha)\bar{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}$ plus higher-order moment terms involving $\mathbf{B}_\xi$. It is worth mentioning that Algorithm 1 achieves this accelerated convergence rate in Eq. (12) via simple restarting and does *not* require an explicit Polyak- or Nesterov-type momentum update rule [Nesterov, 2018]. In the case of nonrandom $\mathbf{B}_\xi$, this rate matches the lower bound [Ibrahim et al., 2020, Zhang et al., 2019b],[5] and the only algorithm that achieves this optimal rate to our best knowledge is Azizian et al. [2020b] without an explicit $1/e$-prefactor on the right hand of Eq. (12).

We end this section with some remarks. For the results in this section, we can forgo fully optimizing the prefactor over α and simply set a step size η as in Eq. (7). Both the analyses of Theorems 3.1 and 3.2 adopt a step size of $\eta_{\mathbf{M}}/\sqrt{2}$, capped by some α-dependent threshold, due to the fact that our analysis relies heavily on the last-iterate convergence to stationarity. In the meantime, Theorem 3.3 does not rely on such an argument and accommodates the larger (thresholded) $\eta_{\mathbf{M}}$ as the step size. Lastly, we emphasize that the knowledge of $\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)$ is required for the algorithm to achieve the accelerated rate. Considerations regarding such knowledge are related to the topic of adaptivity of stochastic gradient algorithms [see, e.g., Lei and Jordan, 2020].

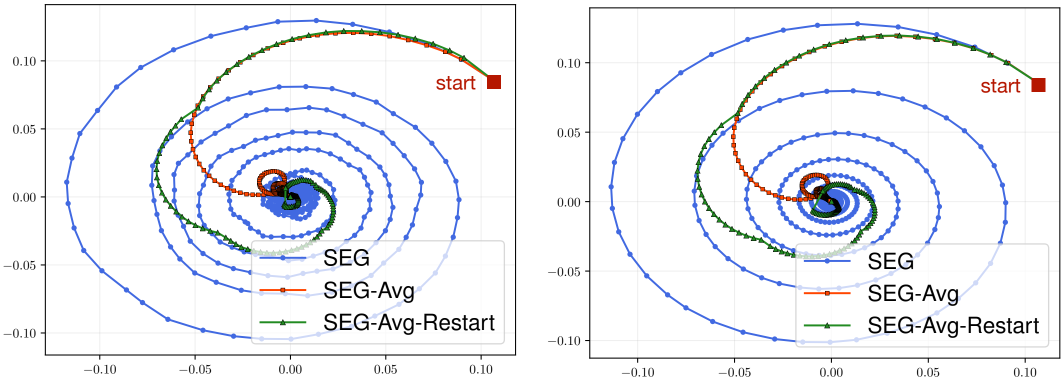
4 EXPERIMENTS

In this section, we present the results of numerical experiments on stochastic bilinear minimax optimization problems, including both the general setting and the interpolation setting (i.e., zero noise at the Nash equilibrium). The objective function we study remains the same as Eq. (1), repeated here for convenience:

$$\min_{\mathbf{x}}\max_{\mathbf{y}} \ \mathbf{x}^\top\mathbb{E}_\xi[\mathbf{B}_\xi]\mathbf{y} + \mathbf{x}^\top\mathbb{E}_\xi[\mathbf{g}_\xi^{\mathbf{x}}] + \mathbb{E}_\xi[(\mathbf{g}_\xi^{\mathbf{y}})^\top]\mathbf{y}. \tag{1}$$

Here we assume $\mathbf{B}_\xi$ is a square matrix of dimension $d \times d$ where $d = m = n$. To generate $\mathbf{B}_\xi$ for each ξ, where ξ corresponds to one iteration in our experiments, we first generate a random vector $\mathbf{u} \in \mathbb{R}^d$, where each

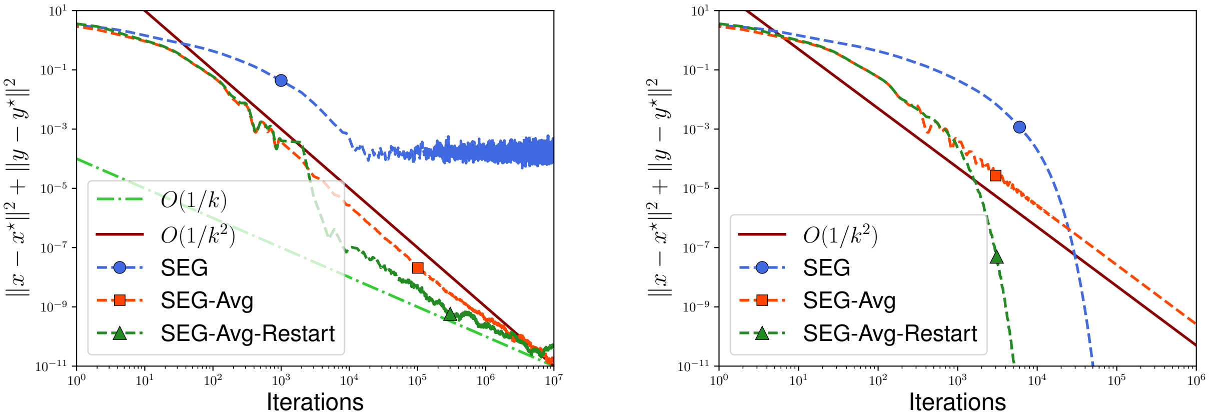
[5] Ibrahim et al. [2020] paper provides the stated lower bound $\sqrt{\kappa}\log(1/\epsilon)$. Although the argument in Zhang et al. [2019b] does not achieve this bound directly (since they did not consider the bilinear-coupling case), modifying their arguments easily extends it to the same lower bound in the bilinear-coupling case. Theorem 3.3 matches this lower bound in the nonrandom case.



(a) General setting.

(b) Interpolation setting.

Figure 1: Illustration (in two dimensions) of the stochastic extragradient (SEG) algorithm, stochastic extragradient with iteration averaging (SEG-Avg), and stochastic extragradient with restarted iteration averaging (SEG-Avg-Restart) on the stochastic minimax optimization problem defined in Eq. (1). Here the Nash equilibrium is $[\mathbf{x}^*; \mathbf{y}^*] = [\mathbf{0}_n; \mathbf{0}_m]$. **(a)** General setting. **(b)** Interpolation setting, where noise vanishes at the Nash equilibrium.



(a) General setting.

(b) Interpolation setting.

Figure 2: Comparing SEG, SEG-Avg, and SEG-Avg-Restart on a stochastic bilinear optimization problem. The horizontal axis represents the iteration number, and vertical axis represents the square ℓ_2-distance to the Nash equilibrium. **(a)** General setting ($d = 100, \text{std}_\mathbf{B} = 0.1, \text{std}_\mathbf{g} = 0.01$). **(b)** Interpolation setting ($d = 100, \text{std}_\mathbf{B} = 0.1, \text{std}_\mathbf{g} = 0.0$).

element of the vector $\mathbf{u}$ is sampled from a uniform distribution, $\mathbf{u}_j \sim \text{Unif}\,[1, d+1]$, for $j \in [d]$. Then we define $\mathbf{B} = \mathsf{Diag}(\mathbf{u})$ and generate $\mathbf{B}_\xi \in \mathbb{R}^{d\times d}$ as follows:

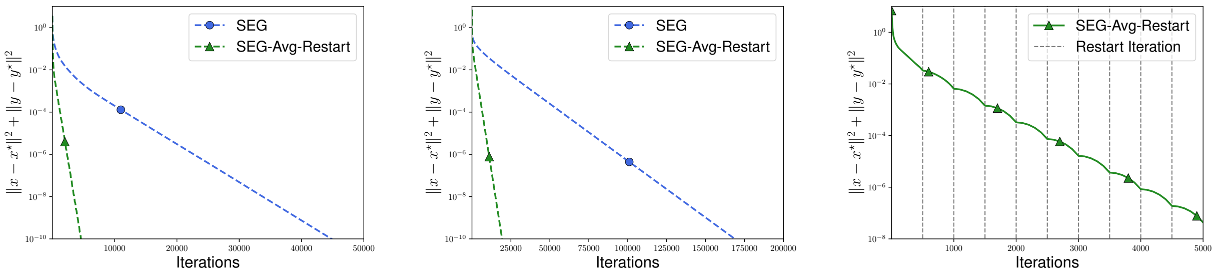
$$\mathbf{B}_\xi = \mathbf{B} + \mathbf{E}_\xi, \quad \text{and } [\mathbf{E}_\xi]_{ij} \sim \mathcal{N}(0, \text{std}_\mathbf{B}^2),$$

where $\mathbb{E}_\xi[\mathbf{B}_\xi] = \mathbf{B}$, and $\mathbf{B}$ is a fixed matrix for all $\mathbf{B}_\xi$. We generate the noise vectors $\mathbf{g}_\xi^\mathbf{x} \sim \mathcal{N}(\mathbf{g}^\mathbf{x}, \text{std}_\mathbf{g}^2\,\mathbf{I}_{d\times d})$ and $\mathbf{g}_\xi^\mathbf{y} \sim \mathcal{N}(\mathbf{g}^\mathbf{y}, \text{std}_\mathbf{g}^2\,\mathbf{I}_{d\times d})$, where we generate the means as follows: $\mathbf{g}^\mathbf{x}$, $\mathbf{g}^\mathbf{y} \sim \mathcal{N}(0, 0.1\cdot\mathbf{I}_{d\times d})$ (note that $\mathbf{g}^\mathbf{x}$, $\mathbf{g}^\mathbf{y}$ are fixed for all $\mathbf{g}_\xi^\mathbf{x}$, $\mathbf{g}_\xi^\mathbf{y}$). More specifically, for each iteration, we randomly generate $\{\mathbf{B}_\xi, \mathbf{g}_\xi^\mathbf{x}, \mathbf{g}_\xi^\mathbf{y}\}$ according to the above procedure. When $\text{std}_\mathbf{B} = \text{std}_\mathbf{g} = 0$, the objective in Eq. (1) equals $\mathbf{x}^\top\mathbf{B}\mathbf{y} + \mathbf{x}^\top\mathbf{g}^\mathbf{x} + (\mathbf{g}^\mathbf{y})^\top\mathbf{y}$, where the Nash equilibrium is $\mathbf{x}^\star = -(\mathbf{B}^\top)^{-1}\mathbf{g}^\mathbf{y}$ and $\mathbf{y}^\star = -\mathbf{B}^{-1}\mathbf{g}^\mathbf{x}$.

We study three algorithms in this section: Stochastic ExtraGradient (**SEG**), Stochastic ExtraGradient with iteration averaging (**SEG-Avg**), and Stochastic ExtraGradient with Restarted iteration averaging (**SEG-Avg-Restart**).[6]

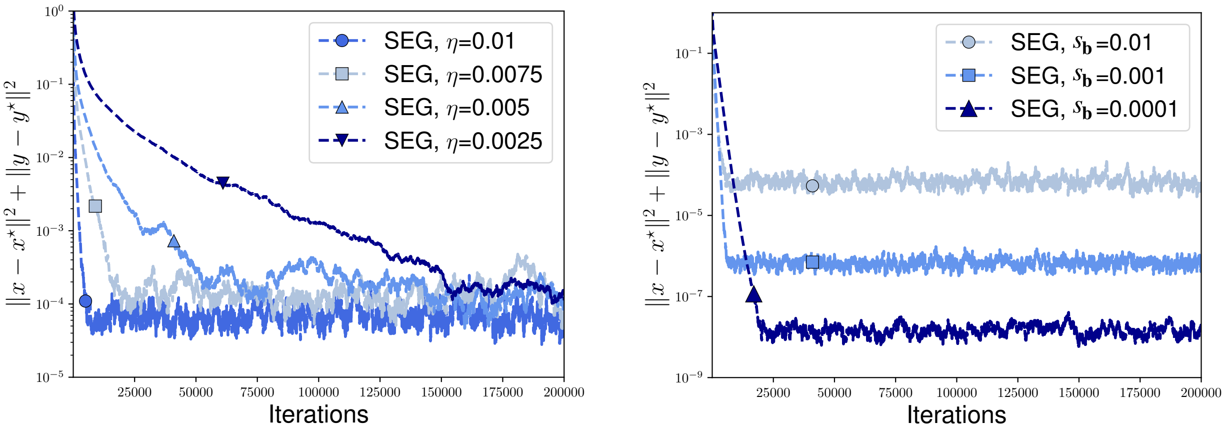
General setting ($\sigma_\mathbf{g} > 0$). We first set $\text{std}_\mathbf{g} = 0.01$ and $\text{std}_\mathbf{B} = 0.1$. The results comparing the three algorithms are shown in Figure 2(a). We find that SEG can only converge to a neighborhood of the Nash equilibrium, whereas SEG-Avg and SEG-Avg-Restart can converge to the equilibrium. From Figure 2(a), we also observe that the convergence rate of SEG-Avg is

[6] Straightforward calculation gives $\sigma_\mathbf{B} = \text{std}_\mathbf{B}\sqrt{d}$ and $\sigma_\mathbf{g} = \text{std}_\mathbf{g}\sqrt{2d}$ in our example, as in Assumptions 2.1, 2.2.



(a) $d = 100$ (b) $d = 200$ (c) SEG-Avg-Restart ($d = 200$)

Figure 3: Comparing SEG and SEG-Avg-Restart on a stochastic bilinear optimization problem in the interpolation setting. The horizontal axis represents the iteration number, and the vertical axis represents the squared ℓ_2-distance to the Nash equilibrium. (**a**) Comparison on dimension $d = 100$ ($\text{std}_{\mathbf{B}} = 0.1, \text{std}_{\mathbf{g}} = 0.0$). (**b**) Comparison on dimension $d = 200$ ($\text{std}_{\mathbf{B}} = 0.1, \text{std}_{\mathbf{g}} = 0.0$). (**c**). Zoomed-in visualization of SEG-Avg-Restart on dimension $d = 200$ ($\text{std}_{\mathbf{B}} = 0.1, \text{std}_{\mathbf{g}} = 0.0$).



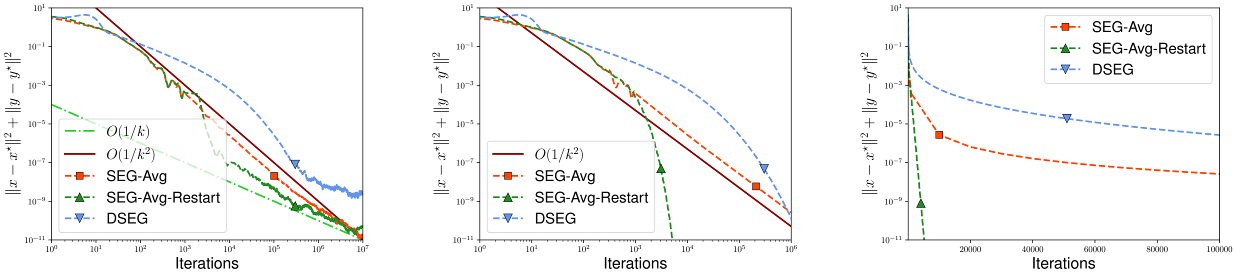
(a) Different step size η. (b) Different noise $\text{std}_{\mathbf{g}}$.

Figure 4: Comparison of SEG (without averaging) with different step sizes η and noise magnitudes $\text{std}_{\mathbf{g}}$ on a stochastic bilinear optimization problem in the general setting. The horizontal axis represents the iteration number, and the vertical axis represents the squared ℓ_2-distance to the Nash equilibrium. (**a**) Comparison with respect to varying step size $\eta \in \{0.01, 0.0075, 0.005, 0.0025\}$ ($\text{std}_{\mathbf{B}} = 0.1, \text{std}_{\mathbf{g}} = 0.01$). (**b**) Comparison with respect to varying noise $\text{std}_{\mathbf{g}} \in \{0.01, 0.001, 0.0001\}$ with step size $\eta = 0.01$ ($\text{std}_{\mathbf{B}} = 0.1$).

$O(1/K^2)$ at the beginning, and then the convergence rate of SEG-Avg changes to $O(1/K)$. Similar to the interpolation setting, SEG-Avg-Restart converges faster than both SEG and SEG-Avg. We also study the effect of the step size η and the noise parameter $\text{std}_{\mathbf{g}}$ for SEG. As shown in Figure 4(a), we observe that SEG cannot converge to a smaller neighborhood of the Nash equilibrium with smaller step size η, which aligns well with our theoretical results. We summarize the varying noise experimental results in Figure 4(b), where we observe that SEG converges to a smaller neighborhood of the Nash equilibrium when we decrease the noise parameter $\text{std}_{\mathbf{g}}$.

Comparisons with DSEG. As shown in Figures 5(a), 5(b), and 5(c), we provide experimental results on comparing SEG-Avg, SEG-Avg-Restart with the *Double Stepsize Extragradient* (DSEG) method, proposed in Hsieh et al. [2020], which allows the step sizes of the extrapolation step and gradient step admitting different scales. We follow the optimized hyperparameter setup described in Hsieh et al. [2020] and select the step size constants to achieve faster convergence. From Figure 5(a), for the general setting, we find that the convergence rate of DSEG is $O(1/K)$ and both SEG-Avg and SEG-Avg-Restart converge faster than DSEG. For the interpolation setting in Figures 5(b) and 5(c), we observe that the convergence rate of DSEG is significantly slower than SEG-Avg-Restart.

Interpolation setting ($\sigma_{\mathbf{g}} = 0$). We first set the noise parameter $\text{std}_{\mathbf{g}} = 0$, and set $\text{std}_{\mathbf{B}} = 0.1$. The performance of SEG, SEG-Avg, and SEG-Avg-Restart is summarized in Figure 2(b), where we set the dimen-



(a) General setting.

(b) Interpolation setting.

(c) Interpolation setting.

Figure 5: Comparing SEG-Avg, SEG-Avg-Restart, and DSEG methods [Hsieh et al., 2020] on the stochastic bilinear optimization problem. The horizontal axis represents the iteration number, and the vertical axis represents the square ℓ_2-distance to the Nash equilibrium. (**a**) General setting ($d = 100, \mathrm{std}_{\mathbf{B}} = 0.1, \mathrm{std}_{\mathbf{g}} = 0.01$). (**b**) Interpolation setting ($d = 100, \mathrm{std}_{\mathbf{B}} = 0.1, \mathrm{std}_{\mathbf{g}} = 0.0$). (**c**). Interpolation setting ($d = 100, \mathrm{std}_{\mathbf{B}} = 0.1, \mathrm{std}_{\mathbf{g}} = 0.0$) under the semi-log scale in the vertical.

sion $d = 100$. We observe that the convergence rate of SEG-Avg is $O(1/K^2)$, which aligns with our theoretical analysis. Meanwhile, we find that SEG-Avg-Restart converges faster than SEG under this interpolation setting. As shown in Figures 3(a) and 3(b), we compare the convergence rate of SEG and SEG-Avg-Restart on a semi-log plot, since both algorithms converge exponentially to the Nash equilibrium in the interpolation setting. We observe that SEG-Avg-Restart converges faster than SEG (for both $d = 100$ and $d = 200$) as suggested by our theory. We also present a zoomed-in plot of SEG-Avg-Restart in Figure 3(c).

5 CONCLUSIONS

We have presented an analysis of the classical Stochastic ExtraGradient (SEG) method for stochastic bilinear minimax optimization. Despite that the last iterate only contracts to a fixed neighborhood of the Nash equilibrium and the diameter of the neighborhood is independent of the step size, we show that SEG accompanied by iteration averaging converges to Nash equilibria at a sublinear rate. Moreover, the forgetting of the initialization is optimal when we use a scheduled restarting procedure in both the general and interpolation settings. Numerical experiments further validate this use of iteration averaging and restarting in the SEG setting.

Further directions for research include justification of the optimality of our convergence result, improvement of the convergence of SEG for nonlinear convex-concave optimization problems with relaxed assumptions, and connection to the Hamiltonian viewpoint for bilinear minimax optimization.

Acknowledgements

We would like to acknowledge support from the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764. Gauthier Gidel and Nicolas Le Roux are supported by Canada CIFAR AI Chairs. Part of this work was done while Nicolas Loizou was a postdoctoral research fellow at Mila, Université de Montréal, supported by the IVADO Postdoctoral Funding Program. Yi Ma acknowledges support from ONR grant N00014-20-1-2002 and the joint Simons Foundation-NSF DMS grant number 2031899.

References

Leonard Adolphs, Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Local saddle point optimization: A curvature exploitation approach. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 486–495. PMLR, 2019.

Ahmet Alacaoglu and Yura Malitsky. Stochastic variance reduction for variational inequality methods. *arXiv preprint arXiv:2102.08352*, 2021.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *International Conference on Artificial Intelligence and Statistics*, pages 2863–2873. PMLR, 2020a.

Waïss Azizian, Damien Scieur, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. Accelerating smooth games by manipulating spectral shapes.

In *International Conference on Artificial Intelligence and Statistics*, pages 1705–1715. PMLR, 2020b.

Francis Bach. The “η-trick” or the effectiveness of reweighted least-squares, 2019.

David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. In *International Conference on Machine Learning*, pages 354–363. PMLR, 2018.

Hugo Berard, Gauthier Gidel, Amjad Almahairi, Pascal Vincent, and Simon Lacoste-Julien. A closer look at the optimization landscapes of generative adversarial networks. In *International Conference on Learning Representations*, 2020.

David Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8, 1956.

Radu Ioan Bot, Panayotis Mertikopoulos, Mathias Staudigl, and Phan Tu Vuong. Forward-backward-forward methods with variance reduction for stochastic variational inequalities. *arXiv preprint arXiv:1902.03355*, 2019.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. In *Advances in Neural Information Processing Systems*, volume 32, pages 393–403, 2019.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, volume 31, pages 9236–9246, 2018.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *International Conference on Learning Representations*, 2018.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *arXiv preprint arXiv:1602.05419*, 2016.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2019.

Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *Conference on Learning Theory*, pages 1758–1784. PMLR, 2020.

Ian Goodfellow. NIPS2016 Tutorial: Generative Adversarial Networks. *arXiv preprint arXiv:1701.00160*, 2016.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Advances in Neural Information Processing Systems*, 2014.

Eduard Gorbunov, Hugo Berard, Gauthier Gidel, and Nicolas Loizou. Stochastic extragradient: General analysis and improved rates. *arXiv preprint arXiv:2111.08611*, 2021a.

Eduard Gorbunov, Nicolas Loizou, and Gauthier Gidel. Extragradient method: $o(1/k)$ last-iterate convergence for monotone variational inequalities and connections with cocoercivity. *arXiv preprint arXiv:2110.04261*, 2021b.

Robert Gower, Othmane Sebbouh, and Nicolas Loizou. SGD for structured nonconvex functions: Learning rates, minibatching and interpolation. In *International Conference on Artificial Intelligence and Statistics*, pages 1315–1323. PMLR, 2021.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR, 2019.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In *Advances in Neural Information Processing Systems*, volume 33, pages 16223–16234, 2020.

Adam Ibrahim, Waıss Azizian, Gauthier Gidel, and Ioannis Mitliagkas. Linear lower bounds and conditioning of differentiable games. In *International Conference on Machine Learning*, pages 4583–4593. PMLR, 2020.

Alfredo N Iusem, Alejandro Jofré, Roberto Imbuzeiro Oliveira, and Philip Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017.

Anatoli Juditsky and Arkadii S Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

G.M. Korpelevich. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 12:747–756, 1976.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Lihua Lei and Michael I Jordan. On the adaptivity of stochastic gradient-based optimization. *SIAM Journal on Optimization*, 30(2):1473–1500, 2020.

Nicolas Loizou and Peter Richtárik. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, 77(3): 653–710, 2020.

Nicolas Loizou, Hugo Berard, Alexia Jolicoeur-Martineau, Pascal Vincent, Simon Lacoste-Julien, and Ioannis Mitliagkas. Stochastic Hamiltonian gradient methods for smooth games. In *International Conference on Machine Learning*, pages 6370–6381. PMLR, 2020.

Nicolas Loizou, Hugo Berard, Gauthier Gidel, Ioannis Mitliagkas, and Simon Lacoste-Julien. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *Advances in Neural Information Processing Systems*, 34, 2021a.

Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic Polyak stepsize for SGD: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1306–1314. PMLR, 2021b.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Eric V Mazumdar, Michael I Jordan, and S Shankar Sastry. On finding local Nash equilibria (and only local Nash equilibria) in zero-sum games. *arXiv preprint arXiv:1901.00838*, 2019.

Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pages 4573–4582. PMLR, 2020.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020.

Oskar Morgenstern and John Von Neumann. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

Arkadi Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19:1574–1609, 2009.

Yurii Nesterov. *Lectures on Convex Optimization*, volume 137. Springer, 2018.

Brendan O'Donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732, 2015.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

James Renegar and Benjamin Grimmer. A simple nearly optimal restart scheme for speeding up first-order methods. *Foundations of Computational Mathematics*, pages 1–46, 2021.

Vincent Roulet and Alexandre d'Aspremont. Sharpness, restart, and acceleration. *SIAM Journal on Optimization*, 30(1):262–289, 2020.

Mark Schmidt. Convergence rate of stochastic gradient with constant step size. *Technical Report, University of British Columbia*, 2014.

Othmane Sebbouh, Robert M Gower, and Aaron Defazio. On the convergence of the stochastic heavy ball method. *arXiv preprint arXiv:2006.07867*, 2020.

Lloyd N Trefethen and David Bau III. *Numerical Linear Algebra*. SIAM, 1997.

Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2): 237–252, 1995.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR, 2019a.

Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. In *Advances in Neural Information Processing Systems*, volume 32, pages 3732–3745, 2019b.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *International Conference on Learning Representations*, 2021.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019a.

Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the saddle point problems. *arXiv preprint arXiv:1912.07481*, 2019b.

Supplementary Material: On the Convergence of Stochastic Extragradient for Bilinear Games using Restarted Iteration Averaging

A ISSUES WITH LAST-ITERATE CONVERGENCE

In this section, we revisit the last-iterate convergence of SEG under our setting. In contrast with minimization problems where stochastic gradient methods with a constant step size converge to a neighborhood of the optimum whose size depends on the step size [Schmidt, 2014], solving the stochastic bilinear minimax optimization problem with Stochastic ExtraGradient (SEG) method under standard settings leads to a last iterate contracting to a fixed neighborhood of the Nash equilibrium whose diameter is independent of the step size. Hence, a classical diminishing step size strategy is not sufficient.

We recall the following notations. Let ξ be an abstract random variable that is equi-distributed as ξ_t. The expectations, also positive semi-definite, are denoted by $\widehat{\mathbf{M}} = \mathbb{E}_\xi \widehat{\mathbf{M}}_\xi \equiv \mathbb{E}_\xi[\mathbf{B}_\xi^\top \mathbf{B}_\xi]$ and $\mathbf{M} = \mathbb{E}_\xi \mathbf{M}_\xi \equiv \mathbb{E}_\xi[\mathbf{B}_\xi \mathbf{B}_\xi^\top]$. It is easy to verify that both matrices are symmetric and positive semi-definite. Recall that $\eta_{\mathbf{M}}$ is the *maximal step size* the SEG algorithm analysis takes, defined earlier as

$$\eta_{\mathbf{M}} \equiv \frac{1}{\sqrt{\lambda_{\max}\left(\mathbf{M}^{-1/2}[\mathbb{E}_\xi \mathbf{M}_\xi^2]\mathbf{M}^{-1/2}\right) \vee \lambda_{\max}\left(\widehat{\mathbf{M}}^{-1/2}[\mathbb{E}_\xi \widehat{\mathbf{M}}_\xi^2]\widehat{\mathbf{M}}^{-1/2}\right)}}. \tag{6}$$

When $\mathbf{B}_\xi$ is nonrandom the value $\eta_{\mathbf{M}}$ simply reduces to $1/\sqrt{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}$. It is worth highlighting that the spectral knowledge of matrices involving moments of $\mathbf{B}_\xi$ that we assume is mild, as analogous spectral information has been traditionally assumed in the online stochastic optimization literature.

We remind the readers of the following result on last-iterate SEG (extension of Hsieh et al. [2020]):

Theorem A.1 (SEG Last Iterate) *Under proper assumptions [e.g., Assumptions 2.1 and 2.2 in §2], if η is chosen as $\eta_{\mathbf{M}}/\sqrt{2}$ where $\eta_{\mathbf{M}}$ is defined as in Eq. (6), we have the following upper bound for the last iterate, $(\mathbf{x}_K, \mathbf{y}_K)$ generated by the algorithm in Eq. (2), for all $K \ge 1$:*

$$\mathbb{E}\left[\|\mathbf{x}_K\|^2 + \|\mathbf{y}_K\|^2\right] \le e^{-\frac{\eta_{\mathbf{M}}^2}{4}\cdot(\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}}))\cdot K}\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right] + \frac{3\sigma_{\mathbf{g}}^2}{\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})}. \tag{13}$$

With our chosen step size, as $K \to \infty$ the expected squared Euclidean norm converges linearly in Eq. (13), i.e.,

$$\limsup_{K\to\infty} \mathbb{E}\left[\|\mathbf{x}_K\|^2 + \|\mathbf{y}_K\|^2\right] \le \frac{3\sigma_{\mathbf{g}}^2}{\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})},$$

which is, in the $\sigma_{\mathbf{g}} > 0$ case, bounded away from zero.[7] A version of Theorem A.1 was provided by Hsieh et al. [2020], where a two-timescale method was proposed to remedy this lack of convergence to zero, with a large step size update of gradient step followed by a smaller step size update of the extragradient step. In this case the asymptotic neighborhood size is proportional to the square root of their ratio. However, Hsieh et al. [2020] only provide a proof under an assumption of bounded noise. In the interpolation case where $\sigma_{\mathbf{g}} = 0$, Vaswani et al. [2019b] showed a weaker version of Theorem A.1 that incorporates an exact line-search step. To the best of our knowledge, the statement of Theorem A.1 is the first to identify the maximal step size $\eta_{\mathbf{M}}$ that can be taken by the SEG method in Eq. (6). For completeness, we provide the proof of our version of Theorem A.1 in §C.1.

[7] In contrast to $\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)$ being zero when $\mathbf{B} \in \mathbb{R}^{n\times m}$ with $n > m$, the $\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})$ can be positive when $\mathbf{B}_\xi$ is random in general. A standard instance will be $\mathbf{B}_\xi$ being $n \times m$ a Gaussian random matrix consisting of independent standard normals.

$\sigma_{\mathbf{B}} > 0$ and $\sigma_{\mathbf{g}} = 0$	Convergence Rate
Juditsky et al. [2011]	$O\left(\max\left\{\frac{R^2}{K^2}, \frac{1}{K}\right\}\right)^{\dagger}$
This work	$O\left(\frac{\|\mathbf{x}_0\|^2+\|\mathbf{y}_0\|^2}{K^2}\right)$
$\sigma_{\mathbf{B}} = 0$ and $\sigma_{\mathbf{g}} > 0$	Convergence Rate
Hsieh et al. [2020]	$O\left(\frac{\sigma_{\mathbf{g}}^2}{K}\right) + o\left(\frac{1}{K}\right)$
This work	$O\left(\frac{\sigma_{\mathbf{g}}^2}{K}\right) + O\left(\frac{\|\mathbf{x}_0\|^2+\|\mathbf{y}_0\|^2}{K^2}\right)$

Table 1: Comparing convergence rates with Juditsky and Nemirovski [2008] and Hsieh et al. [2020]. $^{\dagger}R$ is the squared domain radius.

In addition, we introduce the following negative result that establishes a lower bound that accommodates a broader range of step sizes. This result shows that the upper bound on SEG convergence rate in Theorem A.1 is not improvable even for the case of diminishing step sizes, which limits the applicability of the last-iterate output of the SEG algorithm in general [Hsieh et al., 2020].

Theorem A.2 (Lower bound for SEG, extension of Hsieh et al. [2020]) *Under the assumptions of Theorem A.1, there exist $n, m \geq 1$, a distribution $\mathbb{P}$ supported on $\mathbb{R}^{n\times m} \times \mathbb{R}^n \times \mathbb{R}^m$ for $\{(\mathbf{B}_\xi, \mathbf{g}_\xi^{\mathbf{x}}, \mathbf{g}_\xi^{\mathbf{y}})\}$, and an initialization $(\mathbf{x}_0, \mathbf{y}_0)$ satisfying $\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2 \geq C_1\sigma_{\mathbf{g}}^2$ such that, for any sequence of step sizes $\eta_t \in [0, \eta_{\mathbf{M}}]$, the last-iterate SEG $(\mathbf{x}_K, \mathbf{y}_K)$ generated by Eq. (2) satisfies $\mathbb{E}\left[\|\mathbf{x}_K\|^2 + \|\mathbf{y}_K\|^2\right] \geq C_2\sigma_{\mathbf{g}}^2$ for any $K \geq 1$, where C_1, C_2 are positive, numerical constants.*

In this work, we remedy the lack of convergence that this results indicate via a convergence analysis of the *averaged* iterates. We show in the main text of this paper that SEG with properly scheduled restarting and iteration averaging achieves a statistically optimal rate of convergence, as well as an exponentially mixing (forgetting) of the initialization; see §3.

B COMPARISON OF THEOREM 3.1 WITH EXISTING WORK

In this section, we compare our results with existing work. We first provide a few remarks regarding the convergence rate in Theorem 3.1:

(i) In the general stochastic setting ($\sigma_{\mathbf{g}} > 0$), the step size of our algorithm is not sensitive to the number of iteration (K), i.e., simply picking the constant step size would guarantee the sharp convergence of (same-sample) SEG to the optimal solution, which benefits from the intrinsic linearity of our problem. In comparison, the algorithms in [Juditsky et al., 2011, Mishchenko et al., 2020] rigidly select the step size $\eta = O(1/\sqrt{K})$. Meanwhile, our algorithm does not require the projection step compared with Juditsky et al. [2011], Mishchenko et al. [2020].

(ii) Our analysis of Theorem 3.1 indicates that the "forgetting rate" of the dependency on initialization $\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2$ can be improved to $O(1/K^2)$, achieving an optimal overall rate that is faster than existing work. Mathematically we concluded (8) which is recapped here:

$$\mathbb{E}\left[\|\overline{\mathbf{x}}_K\|^2 + \|\overline{\mathbf{y}}_K\|^2\right] \leq \frac{16 + 8\kappa_\zeta}{(1-\alpha)\hat{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2}{(K+1)^2} + \frac{18 + 12\kappa_\zeta}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{\sigma_{\mathbf{g}}^2}{K+1}, \tag{8}$$

where we recall that $\kappa_\zeta \equiv \frac{\sigma_{\mathbf{B}}^2 + \hat{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2}{\lambda_{\min}(\mathbf{M})\wedge\lambda_{\min}(\widehat{\mathbf{M}})}$ denotes the effective noise condition number of problem Eq. (1). Nevertheless in our upcoming restarting analysis, we present an alternative convergence rate bound for the averaged iterate as follows: for arbitrary $\gamma \in (0, \infty)$

$$\begin{aligned}\mathbb{E}\left[\|\overline{\mathbf{x}}_K\|^2 + \|\overline{\mathbf{y}}_K\|^2\right] &\leq \frac{8(1+\gamma)}{(1-\alpha)\hat{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2}{(K+1)^2} \\ &\frac{2\left(1+\frac{1}{\gamma}\right)(\sigma_{\mathbf{B}}^2 + \hat{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2)\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2 + \frac{3\sigma_{\mathbf{g}}^2}{\lambda_{\min}(\mathbf{M})\wedge\lambda_{\min}(\widehat{\mathbf{M}})}\right] + 9(1+\gamma)\sigma_{\mathbf{g}}^2}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{1}{K+1},\end{aligned} \tag{14}$$

which is slightly better (in the case of $\gamma = 1$) for our restarting analysis. See the discussion paragraph on pp. 24 for more on this.

Comparison with Hsieh et al. [2020] Hsieh et al. [2020] considered the independent-sample double-stepsize SEG, and our work focuses on the same-sample extra-gradient methods. The convergence rate of DSEG Hsieh et al. [2020] in the general stochastic bilinear minimax optimization problem ($\sigma_{\mathbf{B}} = \sigma_{\mathbf{B},2} = 0$ and $\sigma_{\mathbf{g}} > 0$) is

$$\mathbb{E}\left[\|\mathbf{x}_K\|^2 + \|\mathbf{y}_K\|^2\right] \le \frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}^2(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{\sigma_{\mathbf{g}}^2}{K} + o\left(\frac{1}{K}\right).$$

In contrast, our convergence rate is, in a coarse manner, (by setting $\sigma_{\mathbf{B}} = \sigma_{\mathbf{B},2} = 0$ in Eq. (33))

$$\mathbb{E}\left[\|\overline{\mathbf{x}}_K\|^2 + \|\overline{\mathbf{y}}_K\|^2\right] \lesssim \frac{1}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{\sigma_{\mathbf{g}}^2}{K} + \frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2}{K^2}.$$

We observe in the above two displays that our rate is sharper than the rate of Hsieh et al. [2020] in terms of the coefficient of $\sigma_{\mathbf{g}}^2/K$, which is the dominant term in both bounds. In particular, when the step size is chosen properly our convergence rate bound is sharper in the interpolation setting where $\sigma_{\mathbf{B}} > 0$ and $\sigma_{\mathbf{g}} = 0$.

Comparison with Juditsky et al. [2011] We first provide the connection between restricted gap and distance to the Nash equilibrium. Suppose we consider the bounded domain setting for the bilinear minimax optimization problem where $Z = \{\|\mathbf{x}\| \le R, \|\mathbf{y}\| \le R\}$ and R is the domain radius, and the variational inequality with monotone mapping $F(\mathbf{z}) = \begin{bmatrix} \mathbf{B}\mathbf{y} \\ -\mathbf{B}^\top\mathbf{x} \end{bmatrix}$ where $\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$. Then the restricted gap (i.e., merit function) can be expressed as

$$\mathrm{Err}_{\mathrm{vi}}(\mathbf{z}_K) = \max_{\mathbf{z}\in Z} \langle F(\mathbf{z}), \mathbf{z}_K - \mathbf{z}\rangle = \max_{\|\mathbf{y}\|\le R} \mathbf{x}_K^\top \mathbf{B}\mathbf{y} - \min_{\|\mathbf{x}\|\le R} \mathbf{y}_K^\top \mathbf{B}^\top \mathbf{x} = R\left(\|\mathbf{B}^\top\mathbf{x}_K\| + \|\mathbf{B}\mathbf{y}_K\|\right).$$

Therefore, the restricted gap can be lower bounded as

$$\mathrm{Err}_{\mathrm{vi}}(\mathbf{z}_K) \ge R\sqrt{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}\left(\|\mathbf{x}_K\| + \|\mathbf{y}_K\|\right).$$

With this relation at hand, the convergence rate in Juditsky et al. [2011] when calibrated to the interpolation setting ($\sigma_{\mathbf{B}} > 0$ and $\sigma_{\mathbf{g}} = 0$) is

$$\mathbb{E}\left[\|\overline{\mathbf{x}}_K\| + \|\overline{\mathbf{y}}_K\|\right]^2 \lesssim \frac{1}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{\sigma_{\mathbf{B}}^2}{K} + \frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{R^2}{K^2}.$$

In comparison with Juditsky et al. [2011] our convergence rate in Eq. (8) spells

$$\mathbb{E}\left[\|\overline{\mathbf{x}}_K\|^2 + \|\overline{\mathbf{y}}_K\|^2\right] \lesssim \left(\frac{1}{\hat\eta_{\mathbf{M}}(\alpha)^2} + \frac{\sigma_{\mathbf{B}}^2 + \hat\eta_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2}{\hat\eta_{\mathbf{M}}(\alpha)^2(\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}}))}\right) \cdot \frac{1}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2}{K^2},$$

where our convergence rate is significantly better in terms of the $\sigma_{\mathbf{B}}$-dependency.[8]

Other related work on stochastic min-max problems Alacaoglu and Malitsky [2021] proposed stochastic variance reduced algorithms for solving variational inequalities with the finite-sum structure. For more recent results on stochastic iterative methods for solving min-max problems we refer the interested reader to Loizou et al. [2021a], Gorbunov et al. [2021a,b] and the references therein.

C TECHNICAL ANALYSIS OF LAST-ITERATE SEG

In this section we present the technical details of our theoretical results in §A, focusing on the last-iterate Theorem A.1.[9] We first introduce a lemma without proof, which is a standard result in linear algebra [Trefethen and Bau III, 1997, Lecture 5] stating the relations between spectrum of relevant matrices:

[8] In the above five displays, $a_n \lesssim b_n$ denotes $a_n = O(b_n)$ for the two positive sequences.

[9] The proof of Theorem A.2 can be found in Hsieh et al. [2020] and hence we omit it in this work.

Lemma C.1 (Spectral properties) *For our coupling matrix $\mathbf{B} \in \mathbb{R}^{n\times m}$ with $n \ge m$ (tall matrix), $\mathbf{B}^\top\mathbf{B}$ and $\mathbf{B}\mathbf{B}^\top$ share the same spectrum or eigenvalues except zeroes:*

$$\sigma(\mathbf{B}\mathbf{B}^\top) = \sigma(\mathbf{B}^\top\mathbf{B}) \cup (\underbrace{0,\dots,0}_{n-m}).$$

Furthermore, both $\sigma(\mathbf{B}^\top\mathbf{B})$ and $\sigma(\mathbf{B}\mathbf{B}^\top)$ are subsets of the nonnegative reals, so we always have

$$\lambda_{\max}(\mathbf{B}\mathbf{B}^\top) = \lambda_{\max}(\mathbf{B}^\top\mathbf{B}),$$

and

$$\lambda_{\min}\left(\mathbf{B}^\top\mathbf{B}\right) \ge \lambda_{\min}\left(\mathbf{B}\mathbf{B}^\top\right). \tag{15}$$

In special, when $n = m$ we have $\lambda_{\min}(\mathbf{B}^\top\mathbf{B}) = \lambda_{\min}(\mathbf{B}\mathbf{B}^\top)$. The $\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)$ might be different from $\lambda_{\min}(\mathbf{B}^\top\mathbf{B})$ when $\mathbf{B}$ is nonsquare, in which case $\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)$ simply reduces to 0 whenever $n > m$.

Next, we introduce the *contraction parameter* that plays a key role in our analysis.

$$\lambda^*(\eta) \equiv \lambda_{\min}\left(\mathbf{M} - \eta^2[\mathbb{E}_\xi \mathbf{M}_\xi^2]\right) \wedge \lambda_{\min}\left(\widehat{\mathbf{M}} - \eta^2[\mathbb{E}_\xi \widehat{\mathbf{M}}_\xi^2]\right). \tag{16}$$

Note that $\lambda^*(\eta)$ is *not* necessarily nonnegative for positive ηs. We have the following lemma establishing various inequalities regarding $\lambda^*(\eta)$ and $\eta_{\mathbf{M}}$ as in Eq. (6):

Lemma C.2 *Under Assumption 2.1 we have*

(1) For all $\eta > 0$, it holds that

$$\eta^2\lambda^*(\eta) \le 1/4. \tag{17}$$

(2) For all $\eta \in (0, \eta_{\mathbf{M}}]$ where $\eta_{\mathbf{M}}$ is defined as in Eq. (6), it holds that

$$\lambda^*(\eta) \ge \left(1 - \frac{\eta^2}{\eta_{\mathbf{M}}^2}\right)\left(\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})\right) \ge 0. \tag{18}$$

(3) $\eta_{\mathbf{M}}$ defined as in Eq. (6) satisfies, for any $\eta > 0$ such that $\lambda^*(\eta) \ge 0$,

$$0 < \eta_{\mathbf{M}} \le \frac{1}{\sqrt{\lambda_{\max}(\mathbf{M}) \vee \lambda_{\max}(\widehat{\mathbf{M}})}} \le \frac{1}{\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}. \tag{19}$$

When $\mathbf{B}_\xi = \mathbf{B}$ a.s., both equalities hold in the above Eq. (19).

The proof of Lemma C.2 is detailed in §C.3.1.

C.1 Analysis of Theorem A.1

Theorem A.1, Full Version *Let Assumptions 2.1 and 2.2 hold. For any positive η we have for all $K \ge 1$*

$$\begin{aligned} &\mathbb{E}\left[\|\mathbf{x}_K\|^2 + \|\mathbf{y}_K\|^2\right] \\ &\le \left(1 - \eta^2\lambda^*(\eta)\right)^K \left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right] + \eta^2 \mathcal{Q}_K(\eta)\left[1 + \eta^2\left(\lambda_{\max}(\mathbf{M}) \vee \lambda_{\max}(\widehat{\mathbf{M}})\right)\right]\sigma_{\mathbf{g}}^2, \end{aligned} \tag{20}$$

where we denote

$$\mathcal{Q}_K(\eta) \equiv \sum_{t=1}^{K}\left(1 - \eta^2\lambda^*(\eta)\right)^{t-1} \quad \textit{which is upper bounded by } K \wedge \frac{1}{\eta^2\lambda^*(\eta)}, \tag{21}$$

and $\lambda^*(\eta)$ was earlier defined as in Eq. (16). For all $\eta \in (0, \eta_{\mathbf{M}}]$ where $\eta_{\mathbf{M}}$ is defined as in Eq. (6), we have for all $K \geq 1$

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{x}_K\|^2 + \|\mathbf{y}_K\|^2\right] \\
&\leq \left(1 - \eta^2\left(1 - \frac{\eta^2}{\eta_{\mathbf{M}}^2}\right)\left(\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})\right)\right)^K \left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right] \\
&\qquad + \eta^2 \mathcal{Q}_K(\eta)\left[1 + \eta^2\left(\lambda_{\max}(\mathbf{M}) \vee \lambda_{\max}(\widehat{\mathbf{M}})\right)\right]\sigma_{\mathbf{g}}^2 \\
&\leq \exp\left(-\frac{\eta_{\mathbf{M}}^2}{4}\left(\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})\right) \cdot K\right)\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right] + \frac{3\sigma_{\mathbf{g}}^2}{\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})} \\
&\qquad \textit{(when } \eta = \eta_{\mathbf{M}}/\sqrt{2}\textit{)}.
\end{aligned} \tag{22}$$

Analogous to our remarks immediately following the statement of Theorem A.1 in §A, for a given range of step size such that $\lambda^*(\eta)$ is positive, $\mathcal{Q}_K(\eta) \to 1/(\eta^2\lambda^*(\eta))$ as $K \to \infty$ the squared Euclidean norm approaches

$$\mathbb{E}\left[\|\mathbf{x}_K\|^2 + \|\mathbf{y}_K\|^2\right] \to \frac{1}{\lambda^*(\eta)}\left[1 + \eta^2\left(\lambda_{\max}(\mathbf{M}) \vee \lambda_{\max}(\widehat{\mathbf{M}})\right)\right]\sigma_{\mathbf{g}}^2,$$

which is bounded below, due to Eq. (17), by $\frac{\sigma_{\mathbf{g}}^2}{\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})}$ due to $\lambda^*(\eta) \leq \lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})$ and hence bounded away from 0. Optimizing the choice of η achieves, as observed in Eq. (13), a limiting upper bound that triples the above display so the bandwidth of the limiting points is rather narrow (within a triple bandwidth).

We now turn to prove Theorem A.1.

Proof.[Proof of Theorem A.1] We denote for short $\mathbf{M}_\xi \equiv \mathbf{B}_\xi\mathbf{B}_\xi^\top$ and $\widehat{\mathbf{M}}_\xi \equiv \mathbf{B}_\xi^\top\mathbf{B}_\xi$, and $\mathbf{x} \equiv \mathbf{x}_t$, $\mathbf{y} \equiv \mathbf{y}_t$, $\mathbf{x}^- \equiv \mathbf{x}_{t-1}$, $\mathbf{y}^- \equiv \mathbf{y}_{t-1}$, $\mathbf{B}_\xi \equiv \mathbf{B}_{\xi,t}$, $\mathbf{g}_\xi \equiv \mathbf{g}_{\xi,t}$, as well as the conditional expectation $\mathbb{E}_\xi[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. Recall Eq. (2) combined gives the SEG update rules is in total

$$\begin{aligned}
\mathbf{x} &= \mathbf{x}^- - \eta^2\mathbf{B}_\xi\mathbf{B}_\xi^\top\mathbf{x}^- - \eta\left[\mathbf{B}_\xi\mathbf{y}^- + \mathbf{g}_\xi^{\mathbf{x}}\right] - \eta^2\mathbf{B}_\xi\mathbf{g}_\xi^{\mathbf{y}} \\
\mathbf{y} &= \mathbf{y}^- - \eta^2\mathbf{B}_\xi^\top\mathbf{B}_\xi\mathbf{y}^- + \eta\left[\mathbf{B}_\xi^\top\mathbf{x}^- + \mathbf{g}_\xi^{\mathbf{y}}\right] - \eta^2\mathbf{B}_\xi^\top\mathbf{g}_\xi^{\mathbf{x}}.
\end{aligned} \tag{23}$$

By analyzing equation Eq. (23) we derive

$$\begin{aligned}
\mathbb{E}_\xi\left[\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2\right] &= \mathbb{E}_\xi\left\|\left(\mathbf{I} - \eta^2\mathbf{B}_\xi\mathbf{B}_\xi^\top\right)\mathbf{x}^- - \eta\mathbf{B}_\xi\mathbf{y}^- - \eta\mathbf{g}_\xi^{\mathbf{x}} - \eta^2\mathbf{B}_\xi\mathbf{g}_\xi^{\mathbf{y}}\right\|^2 \\
&\quad + \mathbb{E}_\xi\left\|\left(\mathbf{I} - \eta^2\mathbf{B}_\xi^\top\mathbf{B}_\xi\right)\mathbf{y}^- + \eta\mathbf{B}_\xi^\top\mathbf{x}^- + \eta\mathbf{g}_\xi^{\mathbf{y}} - \eta^2\mathbf{B}_\xi^\top\mathbf{g}_\xi^{\mathbf{x}}\right\|^2 \\
&= \mathbb{E}_\xi\left\|\left(\mathbf{I} - \eta^2\mathbf{B}_\xi\mathbf{B}_\xi^\top\right)\mathbf{x}^- - \eta\mathbf{B}_\xi\mathbf{y}^-\right\|^2 + \mathbb{E}_\xi\left\|-\eta\mathbf{g}_\xi^{\mathbf{x}} - \eta^2\mathbf{B}_\xi\mathbf{g}_\xi^{\mathbf{y}}\right\|^2 \\
&\quad \underbrace{+2\mathbb{E}_\xi\left\langle\left(\mathbf{I} - \eta^2\mathbf{B}_\xi\mathbf{B}_\xi^\top\right)\mathbf{x}^- - \eta\mathbf{B}_\xi\mathbf{y}^-, -\eta\mathbf{g}_\xi^{\mathbf{x}} - \eta^2\mathbf{B}_\xi\mathbf{g}_\xi^{\mathbf{y}}\right\rangle}_{\text{cross term}} \\
&\quad + \mathbb{E}_\xi\left\|\left(\mathbf{I} - \eta^2\mathbf{B}_\xi^\top\mathbf{B}_\xi\right)\mathbf{y}^- + \eta\mathbf{B}_\xi^\top\mathbf{x}^-\right\|^2 + \mathbb{E}_\xi\left\|\eta\mathbf{g}_\xi^{\mathbf{y}} - \eta^2\mathbf{B}_\xi^\top\mathbf{g}_\xi^{\mathbf{x}}\right\|^2 \\
&\quad \underbrace{+2\mathbb{E}_\xi\left\langle\left(\mathbf{I} - \eta^2\mathbf{B}_\xi^\top\mathbf{B}_\xi\right)\mathbf{y}^- + \eta\mathbf{B}_\xi^\top\mathbf{x}^-, \eta\mathbf{g}_\xi^{\mathbf{y}} - \eta^2\mathbf{B}_\xi^\top\mathbf{g}_\xi^{\mathbf{x}}\right\rangle}_{\text{cross term}},
\end{aligned}$$

where by independence we have the cross terms being

$$\begin{aligned}
2\mathbb{E}_\xi\left\langle\left(\mathbf{I} - \eta^2\mathbf{B}_\xi\mathbf{B}_\xi^\top\right)\mathbf{x}^- - \eta\mathbf{B}_\xi\mathbf{y}^-, -\eta\mathbf{g}_\xi^{\mathbf{x}} - \eta^2\mathbf{B}_\xi\mathbf{g}_\xi^{\mathbf{y}}\right\rangle &= 0 \\
2\mathbb{E}_\xi\left\langle\left(\mathbf{I} - \eta^2\mathbf{B}_\xi^\top\mathbf{B}_\xi\right)\mathbf{y}^- + \eta\mathbf{B}_\xi^\top\mathbf{x}^-, \eta\mathbf{g}_\xi^{\mathbf{y}} - \eta^2\mathbf{B}_\xi^\top\mathbf{g}_\xi^{\mathbf{x}}\right\rangle &= 0.
\end{aligned}$$

Therefore

$$
\begin{aligned}
&\mathbb{E}_\xi\left[\|\mathbf{x}\|^2+\|\mathbf{y}\|^2\right]\\
&=\mathbb{E}_\xi\left\|\left(\mathbf{I}-\eta^2\mathbf{B}_\xi\mathbf{B}_\xi^\top\right)\mathbf{x}^- -\eta\mathbf{B}_\xi\mathbf{y}^-\right\|^2+\mathbb{E}_\xi\left\|-\eta\mathbf{g}_\xi^{\mathbf{x}}-\eta^2\mathbf{B}_\xi\mathbf{g}_\xi^{\mathbf{y}}\right\|^2\\
&\quad+\mathbb{E}_\xi\left\|\left(\mathbf{I}-\eta^2\mathbf{B}_\xi^\top\mathbf{B}_\xi\right)\mathbf{y}^- +\eta\mathbf{B}_\xi^\top\mathbf{x}^-\right\|^2+\mathbb{E}_\xi\left\|\eta\mathbf{g}_\xi^{\mathbf{y}}-\eta^2\mathbf{B}_\xi^\top\mathbf{g}_\xi^{\mathbf{x}}\right\|^2\\
&=\mathbb{E}_\xi\left\|\left(\mathbf{I}-\eta^2\mathbf{B}_\xi\mathbf{B}_\xi^\top\right)\mathbf{x}^-\right\|^2+\mathbb{E}_\xi\left\|-\eta\mathbf{B}_\xi\mathbf{y}^-\right\|^2+\mathbb{E}_\xi\left\|\left(\mathbf{I}-\eta^2\mathbf{B}_\xi^\top\mathbf{B}_\xi\right)\mathbf{y}^-\right\|^2+\mathbb{E}_\xi\left\|\eta\mathbf{B}_\xi^\top\mathbf{x}^-\right\|^2\\
&\quad\underbrace{+2\mathbb{E}_\xi\left\langle\left(\mathbf{I}-\eta^2\mathbf{B}_\xi\mathbf{B}_\xi^\top\right)\mathbf{x}^-,-\eta\mathbf{B}_\xi\mathbf{y}^-\right\rangle+2\mathbb{E}_\xi\left\langle\left(\mathbf{I}-\eta^2\mathbf{B}_\xi^\top\mathbf{B}_\xi\right)\mathbf{y}^-,\eta\mathbf{B}_\xi^\top\mathbf{x}^-\right\rangle}_{\text{cross term}}\\
&\quad+\mathbb{E}_\xi\left\|-\eta\mathbf{g}_\xi^{\mathbf{x}}\right\|^2+\mathbb{E}_\xi\left\|-\eta^2\mathbf{B}_\xi\mathbf{g}_\xi^{\mathbf{y}}\right\|^2+\mathbb{E}_\xi\left\|\eta\mathbf{g}_\xi^{\mathbf{y}}\right\|^2+\mathbb{E}_\xi\left\|-\eta^2\mathbf{B}_\xi^\top\mathbf{g}_\xi^{\mathbf{x}}\right\|^2\\
&\quad\underbrace{+2\mathbb{E}_\xi\left\langle-\eta\mathbf{g}_\xi^{\mathbf{x}},-\eta^2\mathbf{B}_\xi\mathbf{g}_\xi^{\mathbf{y}}\right\rangle+2\mathbb{E}_\xi\left\langle\eta\mathbf{g}_\xi^{\mathbf{y}},-\eta^2\mathbf{B}_\xi^\top\mathbf{g}_\xi^{\mathbf{x}}\right\rangle}_{\text{cross term}},
\end{aligned}
$$

where it is again easy to verify the cross terms are zero due to the identities

$$
\mathbb{E}_\xi\left\langle\left(\mathbf{I}-\eta^2\mathbf{B}_\xi\mathbf{B}_\xi^\top\right)\mathbf{x}^-,-\eta\mathbf{B}_\xi\mathbf{y}^-\right\rangle+\mathbb{E}_\xi\left\langle\left(\mathbf{I}-\eta^2\mathbf{B}_\xi^\top\mathbf{B}_\xi\right)\mathbf{y}^-,\eta\mathbf{B}_\xi^\top\mathbf{x}^-\right\rangle=0,
$$

$$
\mathbb{E}_\xi\left\langle-\eta\mathbf{g}_\xi^{\mathbf{x}},-\eta^2\mathbf{B}_\xi\mathbf{g}_\xi^{\mathbf{y}}\right\rangle+\mathbb{E}_\xi\left\langle\eta\mathbf{g}_\xi^{\mathbf{y}},-\eta^2\mathbf{B}_\xi^\top\mathbf{g}_\xi^{\mathbf{x}}\right\rangle=0.
$$

Finally

$$
\begin{aligned}
&\mathbb{E}_\xi\left[\|\mathbf{x}\|^2+\|\mathbf{y}\|^2\right]\\
&=\mathbb{E}_\xi\left\|\left(\mathbf{I}-\eta^2\mathbf{B}_\xi\mathbf{B}_\xi^\top\right)\mathbf{x}^-\right\|^2+\mathbb{E}_\xi\left\|-\eta\mathbf{B}_\xi\mathbf{y}^-\right\|^2+\mathbb{E}_\xi\left\|\left(\mathbf{I}-\eta^2\mathbf{B}_\xi^\top\mathbf{B}_\xi\right)\mathbf{y}^-\right\|^2+\mathbb{E}_\xi\left\|\eta\mathbf{B}_\xi^\top\mathbf{x}^-\right\|^2\\
&\quad+\mathbb{E}_\xi\left\|-\eta\mathbf{g}_\xi^{\mathbf{x}}\right\|^2+\mathbb{E}_\xi\left\|-\eta^2\mathbf{B}_\xi\mathbf{g}_\xi^{\mathbf{y}}\right\|^2+\mathbb{E}_\xi\left\|\eta\mathbf{g}_\xi^{\mathbf{y}}\right\|^2+\mathbb{E}_\xi\left\|-\eta^2\mathbf{B}_\xi^\top\mathbf{g}_\xi^{\mathbf{x}}\right\|^2\\
&=\left(\mathbf{x}^-\right)^\top\mathbb{E}\left(\mathbf{I}-\eta^2\mathbf{B}_\xi\mathbf{B}_\xi^\top+\left(\eta^2\mathbf{B}_\xi\mathbf{B}_\xi^\top\right)^2\right)\mathbf{x}^-+\left(\mathbf{y}^-\right)^\top\mathbb{E}\left(\mathbf{I}-\eta^2\mathbf{B}_\xi^\top\mathbf{B}_\xi+\left(\eta^2\mathbf{B}_\xi^\top\mathbf{B}_\xi\right)^2\right)\mathbf{y}^-\\
&\quad+\eta^2\mathbb{E}_\xi\left[\left(\mathbf{g}_\xi^{\mathbf{x}}\right)^\top\left(\mathbf{I}+\eta^2\mathbf{B}_\xi\mathbf{B}_\xi^\top\right)\mathbf{g}_\xi^{\mathbf{x}}\right]+\eta^2\mathbb{E}_\xi\left[\left(\mathbf{g}_\xi^{\mathbf{y}}\right)^\top\left(\mathbf{I}+\eta^2\mathbf{B}_\xi^\top\mathbf{B}_\xi\right)\mathbf{g}_\xi^{\mathbf{y}}\right],
\end{aligned}
\tag{24}
$$

and in the last equality we use the *independence* assumption of $\mathbf{B}_\xi$ and $[\mathbf{g}_\xi^{\mathbf{x}};\mathbf{g}_\xi^{\mathbf{y}}]$ as in Assumption 2.2, so we have since $\mathbb{E}_\xi[\mathbf{B}_\xi^\top\mathbf{B}_\xi]\preceq\lambda_{\max}(\widehat{\mathbf{M}})\mathbf{I}_m$ and $\mathbb{E}_\xi[\mathbf{B}_\xi\mathbf{B}_\xi^\top]\preceq\lambda_{\max}(\mathbf{M})\mathbf{I}_n$ that

$$
\mathbb{E}_\xi\left[\|\mathbf{B}_\xi\mathbf{g}_\xi^{\mathbf{y}}\|^2\mid\mathbf{g}_\xi\right]=\mathbf{g}_\xi^{\mathbf{y}}\mathbb{E}_\xi[\mathbf{B}_\xi^\top\mathbf{B}_\xi](\mathbf{g}_\xi^{\mathbf{y}})^\top\le\mathbf{g}_\xi^{\mathbf{y}}\left[\lambda_{\max}(\widehat{\mathbf{M}})\mathbf{I}_m\right](\mathbf{g}_\xi^{\mathbf{y}})^\top=\lambda_{\max}(\widehat{\mathbf{M}})\|\mathbf{g}_\xi^{\mathbf{y}}\|^2,
$$

and analogously

$$
\mathbb{E}_\xi\left[\|\mathbf{B}_\xi^\top\mathbf{g}_\xi^{\mathbf{x}}\|^2\mid\mathbf{g}_\xi\right]\le\lambda_{\max}(\mathbf{M})\|\mathbf{g}_\xi^{\mathbf{x}}\|^2,
$$

so summing up the above two and taking expectation gives, due to Assumption 2.2,

$$
\begin{aligned}
\mathbb{E}_\xi\left[\|\mathbf{B}_\xi\mathbf{g}_\xi^{\mathbf{y}}\|^2\right]+\mathbb{E}_\xi\left[\|\mathbf{B}_\xi^\top\mathbf{g}_\xi^{\mathbf{x}}\|^2\right]&\le\lambda_{\max}(\widehat{\mathbf{M}})\mathbb{E}_\xi\|\mathbf{g}_\xi^{\mathbf{y}}\|^2+\lambda_{\max}(\mathbf{M})\mathbb{E}_\xi\|\mathbf{g}_\xi^{\mathbf{x}}\|^2\\
&\le\left(\lambda_{\max}(\widehat{\mathbf{M}})\vee\lambda_{\max}(\mathbf{M})\right)\mathbb{E}\left[\|\mathbf{g}_\xi^{\mathbf{y}}\|^2+\|\mathbf{g}_\xi^{\mathbf{x}}\|^2\right]=\left(\lambda_{\max}(\widehat{\mathbf{M}})\vee\lambda_{\max}(\mathbf{M})\right)\sigma_{\mathbf{g}}^2.
\end{aligned}
$$

Therefore Eq. (24) gives, for any positive η, that

$$
\begin{aligned}
\mathbb{E}_\xi\left[\|\mathbf{x}\|^2+\|\mathbf{y}\|^2\right]&=\left(\mathbf{x}^-\right)^\top\left(\mathbf{I}-\eta^2\left(\mathbf{M}-\eta^2[\mathbb{E}_\xi\mathbf{M}_\xi^2]\right)\right)\mathbf{x}^-+\left(\mathbf{y}^-\right)^\top\left(\mathbf{I}-\eta^2\left(\widehat{\mathbf{M}}-\eta^2[\mathbb{E}_\xi\widehat{\mathbf{M}}_\xi^2]\right)\right)\mathbf{y}^-\\
&\quad+\eta^2\left[1+\eta^2\left(\lambda_{\max}(\mathbf{M})\vee\lambda_{\max}(\widehat{\mathbf{M}})\right)\right]\sigma_{\mathbf{g}}^2\\
&\le\left(1-\eta^2\lambda^*(\eta)\right)\left(\|\mathbf{x}^-\|^2+\|\mathbf{y}^-\|^2\right)+\eta^2\left[1+\eta^2\left(\lambda_{\max}(\mathbf{M})\vee\lambda_{\max}(\widehat{\mathbf{M}})\right)\right]\sigma_{\mathbf{g}}^2,
\end{aligned}
$$

where $\lambda^*(\eta)$ was earlier defined in Eq. (16). Recursively applying this allows us to conclude

$$\mathbb{E}\left[\|\mathbf{x}_K\|^2 + \|\mathbf{y}_K\|^2\right] \le \left(1 - \eta^2\lambda^*(\eta)\right)^K \left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right] + \underbrace{\left[\sum_{t=1}^{K} \left(1 - \eta^2\lambda^*(\eta)\right)^{t-1} \eta^2\right]}_{= \eta^2 \mathcal{Q}_K(\eta) \text{ due to Eq. (21)}} \left[1 + \eta^2\left(\lambda_{\max}(\mathbf{M}) \vee \lambda_{\max}(\widehat{\mathbf{M}})\right)\right] \sigma_{\mathbf{g}}^2,$$

and hence concludes Eq. (20). The rest of the proof, under the condition $\eta \in (0, \eta_{\mathbf{M}}]$, follows from $\lambda^*(\eta)$'s definition in Eq. (16) along with Lemma C.2. □

C.2 Proof of Theorem A.2

Theorem A.2 is, in fact, a variant of Proposition 1 of Hsieh et al. [2020] under our assumptions. Hence the proof therein applies, and we provide the statement in our work mainly for the sake of completeness.

C.3 Auxiliary Proofs

C.3.1 Proof of Lemma C.2

Proof.[Proof of Lemma C.2]

(1) Since $\mathbb{E}_\xi \mathbf{M}_\xi^2 - \mathbf{M}^2 = \mathbb{E}_\xi\left(\mathbf{M}_\xi - \mathbf{M}\right)^2 \succeq \mathbf{0}$ a simple discriminant argument of a quadratic $1 - t + t^2$, which is greater than or equal to $3/4$ for all reals t, concludes

$$\mathbf{I} - \eta^2\left(\mathbf{M} - \eta^2[\mathbb{E}_\xi \mathbf{M}_\xi^2]\right) \succeq \mathbf{I} - \eta^2\left(\mathbf{M} - \eta^2\mathbf{M}^2\right) \succeq \frac{3}{4}\mathbf{I},$$

and

$$\mathbf{I} - \eta^2\left(\widehat{\mathbf{M}} - \eta^2[\mathbb{E}_\xi \widehat{\mathbf{M}}_\xi^2]\right) \succeq \mathbf{I} - \eta^2\left(\widehat{\mathbf{M}} - \eta^2\widehat{\mathbf{M}}^2\right) \succeq \frac{3}{4}\mathbf{I},$$

and hence

$$1 - \eta^2\lambda^*(\eta) = \lambda_{\max}\left(\mathbf{I} - \eta^2\left(\mathbf{M} - \eta^2[\mathbb{E}_\xi \mathbf{M}_\xi^2]\right)\right) \vee \lambda_{\max}\left(\mathbf{I} - \eta^2\left(\widehat{\mathbf{M}} - \eta^2[\mathbb{E}_\xi \widehat{\mathbf{M}}_\xi^2]\right)\right) \ge \frac{3}{4},$$

proving Eq. (17).

(2) The definition of $\eta_{\mathbf{M}}$ as in Eq. (6) gives for all $\eta \in (0, \eta_{\mathbf{M}}]$

$$\eta^2\mathbf{M}^{-1/2}[\mathbb{E}_\xi \mathbf{M}_\xi^2]\mathbf{M}^{-1/2} \preceq \mathbf{I} \qquad \text{and} \qquad \eta^2\widehat{\mathbf{M}}^{-1/2}[\mathbb{E}_\xi \widehat{\mathbf{M}}_\xi^2]\widehat{\mathbf{M}}^{-1/2} \preceq \mathbf{I},$$

and we have

$$\mathbb{E}_\xi \mathbf{M}_\xi^2 \preceq \frac{1}{\eta_{\mathbf{M}}^2}\mathbf{M} \qquad \text{and} \qquad \mathbb{E}_\xi \widehat{\mathbf{M}}_\xi^2 \preceq \frac{1}{\eta_{\mathbf{M}}^2}\widehat{\mathbf{M}}$$

hold, which concludes when η satisfies $\eta \in (0, \eta_{\mathbf{M}}]$ both

$$\mathbf{M} - \eta^2[\mathbb{E}_\xi \mathbf{M}_\xi^2] \succeq \left(1 - \frac{\eta^2}{\eta_{\mathbf{M}}^2}\right)\mathbf{M} \qquad \text{and} \qquad \widehat{\mathbf{M}} - \eta^2[\mathbb{E}_\xi \widehat{\mathbf{M}}_\xi^2] \succeq \left(1 - \frac{\eta^2}{\eta_{\mathbf{M}}^2}\right)\widehat{\mathbf{M}},$$

and hence via Eq. (16)

$$\lambda^*(\eta) = \lambda_{\min}\left(\mathbf{M} - \eta^2[\mathbb{E}_\xi \mathbf{M}_\xi^2]\right) \wedge \lambda_{\min}\left(\widehat{\mathbf{M}} - \eta^2[\mathbb{E}_\xi \widehat{\mathbf{M}}_\xi^2]\right) \ge \left(1 - \frac{\eta^2}{\eta_{\mathbf{M}}^2}\right)\left(\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})\right) \ge 0$$

holds for all $\eta \in (0, \eta_{\mathbf{M}}]$, which proves Eq. (18).

(3) Note $\mathbf{M} = \mathbb{E}_\xi \mathbf{M}_\xi$ and $\widehat{\mathbf{M}} = \mathbb{E}_\xi \widehat{\mathbf{M}}_\xi$, and $\eta_{\mathbf{M}} > 0$ is due to the finiteness of $\lambda_{\max}\left(\mathbf{M}^{-1/2}[\mathbb{E}_\xi \mathbf{M}_\xi^2]\mathbf{M}^{-1/2}\right)$ under Assumption 2.1. For the second inequality note by the part (1) of the proof

$$\mathbf{M}^{-1/2}[\mathbb{E}_\xi \mathbf{M}_\xi^2]\mathbf{M}^{-1/2} \succeq \mathbf{M}^{-1/2}\mathbf{M}^2\mathbf{M}^{-1/2} = \mathbf{M},$$

and

$$\widehat{\mathbf{M}}^{-1/2}[\mathbb{E}_\xi \widehat{\mathbf{M}}_\xi^2]\widehat{\mathbf{M}}^{-1/2} \succeq \widehat{\mathbf{M}}^{-1/2}\widehat{\mathbf{M}}^2\widehat{\mathbf{M}}^{-1/2} = \widehat{\mathbf{M}},$$

hold due to Eq. (4), and it is straightforward to check that all equalities hold in the $\mathbf{B}_\xi = \mathbf{B}$ a.s. case, proving the second inequality of Eq. (19). For the third inequality, we have

$$\mathbf{M} - \mathbf{B}\mathbf{B}^\top = \mathbb{E}\left[(\mathbf{B}_\xi - \mathbf{B})(\mathbf{B}_\xi - \mathbf{B})^\top\right] \succeq 0,$$

and

$$\widehat{\mathbf{M}} - \mathbf{B}^\top\mathbf{B} = \mathbb{E}\left[(\mathbf{B}_\xi - \mathbf{B})^\top(\mathbf{B}_\xi - \mathbf{B})\right] \succeq 0,$$

with equality holds when $\mathbf{B}_\xi = \mathbf{B}$ a.s. Hence

$$\lambda_{\max}(\mathbf{M}) \geq \lambda_{\max}(\mathbf{B}\mathbf{B}^\top) \qquad \text{and} \qquad \lambda_{\max}(\widehat{\mathbf{M}}) \geq \lambda_{\max}(\mathbf{B}^\top\mathbf{B}).$$

Note $\lambda_{\max}(\mathbf{B}\mathbf{B}^\top) = \lambda_{\max}(\mathbf{B}^\top\mathbf{B})$ as indicated by Lemma C.1 gives the third inequality and the whole lemma.

□

D TECHNICAL ANALYSIS IN §3

In this section, we collects the technical analyses and proofs of our main theoretical results. The study of SEG in general stochastic setting §3 for the averaged-iterate Theorem 3.1 and restarted-averaged-iterate Theorem 3.2. When narrowing down to the interpolation setting in §3, we state Theorem 3.3. For each of the theorems we first detail their full versions and accompany them with proofs, separately.

D.1 Analysis of Theorem 3.1

Theorem 3.1, Full Version *Let Assumptions 2.1 and 2.2 hold and we assume that $\lambda^*(\eta) > 0$. Under the condition on step size $\eta \in (0, \eta_{\mathbf{M}}]$ where $\eta_{\mathbf{M}}$ was earlier defined as in Eq. (6), we have for all $K \geq 0$ the following convergence rate holds for the averaged iterate $\overline{\mathbf{x}}_K, \overline{\mathbf{y}}_K$ defined in Theorem 3.1:*

$$\begin{aligned}
&\left(\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\left(1 + \eta^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\right) - 2\eta\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}\right)\mathbb{E}\left[\|\overline{\mathbf{x}}_K\|^2 + \|\overline{\mathbf{y}}_K\|^2\right] \\
&\leq \mathbb{E}\left[\|\mathbf{B}\overline{\mathbf{y}}_K + \eta\mathbf{M}\overline{\mathbf{x}}_K\|^2 + \left\|\mathbf{B}^\top\overline{\mathbf{x}}_K - \eta\widehat{\mathbf{M}}\overline{\mathbf{y}}_K\right\|^2\right] \\
&\leq \left(\frac{8(1+\gamma)}{\eta^2(K+1)^2} + \frac{2\left(1+\frac{1}{\gamma}\right)(\sigma_{\mathbf{B}}^2 + \eta^2\sigma_{\mathbf{B},2}^2)}{K+1}\right)\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right] \\
&\quad + \frac{6(1+\gamma) + 2\left(1+\frac{1}{\gamma}\right)(\sigma_{\mathbf{B}}^2 + \eta^2\sigma_{\mathbf{B},2}^2)\lambda^*(\eta)^{-1}}{K+1}\left[1 + \eta^2\left(\lambda_{\max}(\mathbf{M}) \vee \lambda_{\max}(\widehat{\mathbf{M}})\right)\right]\sigma_{\mathbf{g}}^2,
\end{aligned} \tag{25}$$

where $\gamma \in (0,\infty)$ is arbitrary. In addition when $\mathbf{B}_\xi, \mathbf{B}$ are square matrices, we have

$$\mathbb{E}\left[\|\overline{\mathbf{x}}_K\|^2 + \|\overline{\mathbf{y}}_K\|^2\right] \leq \widehat{\mathcal{P}}_{K+1}(\eta) \cdot \frac{\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2}{(K+1)^2}. \tag{26}$$

In above the prefactor is defined as[10]

$$\begin{aligned}
&\widehat{\mathcal{P}}_{K+1}(\eta) \\
&\equiv \begin{cases} +\infty & \textit{if } \lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\left(1 + \eta^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\right) \leq 2\eta\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})} \\ \dfrac{8(1+\gamma) + \left(2\left(1+\frac{1}{\gamma}\right)\eta^2(\sigma_{\mathbf{B}}^2+\eta^2\sigma_{\mathbf{B},2}^2) + \frac{6(1+\gamma)+2\left(1+\frac{1}{\gamma}\right)(\sigma_{\mathbf{B}}^2+\eta^2\sigma_{\mathbf{B},2}^2)\lambda^*(\eta)^{-1}}{\|\mathbf{x}_0\|^2+\|\mathbf{y}_0\|^2}\cdot\eta^2\left[1+\eta^2\left(\lambda_{\max}(\mathbf{M})\vee\lambda_{\max}(\widehat{\mathbf{M}})\right)\right]\sigma_{\mathbf{g}}^2\right)\cdot(K+1)}{\eta^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)(1+\eta^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)) - 2\eta^3\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}} & \textit{otherwise} \end{cases},
\end{aligned}$$

[10] Here we interpret $0 \cdot (+\infty)$ as $+\infty$ whenever it appears.

and by setting $\eta = \hat{\eta}_{\mathbf{M}}(\alpha)$ defined earlier as in Eq. (7), we have

$$\widehat{\mathcal{P}}_{K+1}(\hat{\eta}_{\mathbf{M}}(\alpha)) \le \frac{8(1+\gamma)}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{1}{\hat{\eta}_{\mathbf{M}}(\alpha)^2} + \frac{2\left(1+\frac{1}{\gamma}\right)(\sigma_{\mathbf{B}}^2 + \hat{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2)\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2 + \frac{3\sigma_{\mathbf{g}}^2}{\lambda_{\min}(\mathbf{M})\wedge\lambda_{\min}(\widehat{\mathbf{M}})}\right] + 9(1+\gamma)\sigma_{\mathbf{g}}^2}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{K+1}{\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2}, \tag{27}$$

which recovers Eq. (14).

Proof.[Proof of Theorem 3.1] First, as long as $\eta^2\lambda^*(\eta) \le 1/4$ the Eq. (20) is further bounded as

$$\begin{aligned} &\mathbb{E}\left[\|\mathbf{x}_K\|^2 + \|\mathbf{y}_K\|^2\right] \\ &\le \left(1-\eta^2\lambda^*(\eta)\right)^K\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right] + \eta^2\mathcal{Q}_K(\eta)\left[1 + \eta^2\left(\lambda_{\max}(\mathbf{M}) \vee \lambda_{\max}(\widehat{\mathbf{M}})\right)\right]\sigma_{\mathbf{g}}^2. \end{aligned} \tag{28}$$

Depending on the behavior of $\mathcal{Q}_K(\eta)$, the expected squared Euclidean norm admits two different upper bounds: (i) when $\lambda^*(\eta)$ is bounded away from 0, uniform bound holds with its limit being bounded by a quantity that is inverse proportional to $\lambda^*(\eta)$; (ii) when $\lambda^*(\eta)$ approaches zero, the quantity eventually grows linearly at a rate that does not depend on $\lambda^*(\eta)$. In our analysis we will assume that $\lambda^*(\eta)$ is bounded away from zero while applying two different bounds interchangeably.

Returning to the SEG update Eq. (23) which we repeat as

$$\begin{aligned} \mathbf{x}_t &= \mathbf{x}_{t-1} - \eta^2\mathbf{B}_{\xi,t}\mathbf{B}_{\xi,t}^\top\mathbf{x}_{t-1} - \eta\left[\mathbf{B}_{\xi,t}\mathbf{y}_{t-1} + \mathbf{g}_{\xi,t}^{\mathbf{x}}\right] - \eta^2\mathbf{B}_{\xi,t}\mathbf{g}_{\xi,t}^{\mathbf{y}} \\ \mathbf{y}_t &= \mathbf{y}_{t-1} - \eta^2\mathbf{B}_{\xi,t}^\top\mathbf{B}_{\xi,t}\mathbf{y}_{t-1} + \eta\left[\mathbf{B}_{\xi,t}^\top\mathbf{x}_{t-1} + \mathbf{g}_{\xi,t}^{\mathbf{y}}\right] - \eta^2\mathbf{B}_{\xi,t}^\top\mathbf{g}_{\xi,t}^{\mathbf{x}}. \end{aligned} \tag{23}$$

Setting $\eta = \hat{\eta}_{\mathbf{M}}(\alpha)$ as in Eq. (7) and telescoping both sides of the update rule Eq. (23) for $t = 1, \ldots, K$ gives

$$\begin{aligned} \mathbf{x}_K - \mathbf{x}_0 &= -\eta^2\sum_{t=1}^K \mathbf{B}_{\xi,t}\mathbf{B}_{\xi,t}^\top\mathbf{x}_{t-1} - \eta\sum_{t=1}^K\left[\mathbf{B}_{\xi,t}\mathbf{y}_{t-1} + \mathbf{g}_{\xi,t}^{\mathbf{x}}\right] - \eta^2\sum_{t=1}^K\mathbf{B}_{\xi,t}\mathbf{g}_{\xi,t}^{\mathbf{y}} \\ \mathbf{y}_K - \mathbf{y}_0 &= -\eta^2\sum_{t=1}^K \mathbf{B}_{\xi,t}^\top\mathbf{B}_{\xi,t}\mathbf{y}_{t-1} + \eta\sum_{t=1}^K\left[\mathbf{B}_{\xi,t}^\top\mathbf{x}_{t-1} + \mathbf{g}_{\xi,t}^{\mathbf{y}}\right] - \eta^2\sum_{t=1}^K\mathbf{B}_{\xi,t}^\top\mathbf{g}_{\xi,t}^{\mathbf{x}}. \end{aligned}$$

Manipulating gives

$$\begin{aligned} \frac{1}{K}\sum_{t=1}^K\mathbf{B}_{\xi,t}\mathbf{y}_{t-1} + \frac{\eta}{K}\sum_{t=1}^K\mathbf{B}_{\xi,t}\mathbf{B}_{\xi,t}^\top\mathbf{x}_{t-1} &= \frac{\mathbf{x}_K - \mathbf{x}_0}{-\eta K} - \frac{1}{K}\sum_{t=1}^K\mathbf{g}_{\xi,t}^{\mathbf{x}} - \frac{\eta}{K}\sum_{t=1}^K\mathbf{B}_{\xi,t}\mathbf{g}_{\xi,t}^{\mathbf{y}}, \\ \frac{1}{K}\sum_{t=1}^K\mathbf{B}_{\xi,t}^\top\mathbf{x}_{t-1} - \frac{\eta}{K}\sum_{t=1}^K\mathbf{B}_{\xi,t}^\top\mathbf{B}_{\xi,t}\mathbf{y}_{t-1} &= \frac{\mathbf{y}_K - \mathbf{y}_0}{\eta K} - \frac{1}{K}\sum_{t=1}^K\mathbf{g}_{\xi,t}^{\mathbf{y}} + \frac{\eta}{K}\sum_{t=1}^K\mathbf{B}_{\xi,t}^\top\mathbf{g}_{\xi,t}^{\mathbf{x}}. \end{aligned} \tag{29}$$

Now we try to bound the norm of the left hands in the above two displays. Young's inequality gives that for fixed $\gamma > 0$, $\|a+b\|^2 \le (1+\gamma)\|a\|^2 + (1+\frac{1}{\gamma})\|b\|^2$ so $\|a\|^2 \ge \frac{1}{1+\gamma}\|a+b\|^2 - \frac{1}{\gamma}\|b\|^2$ holds for two vectors a, b of same dimensions,

$$\begin{aligned} &\mathbb{E}\left\|\frac{1}{K}\sum_{t=1}^K\mathbf{B}_{\xi,t}\mathbf{y}_{t-1} + \frac{\eta}{K}\sum_{t=1}^K\mathbf{B}_{\xi,t}\mathbf{B}_{\xi,t}^\top\mathbf{x}_{t-1}\right\|^2 \\ &= \mathbb{E}\left\|\mathbf{B}\overline{\mathbf{y}}_{K-1} + \eta\mathbf{M}\overline{\mathbf{x}}_{K-1} + \frac{1}{K}\sum_{t=1}^K(\mathbf{B}_{\xi,t} - \mathbf{B})\mathbf{y}_{t-1} + \frac{\eta}{K}\sum_{t=1}^K\left(\mathbf{B}_{\xi,t}\mathbf{B}_{\xi,t}^\top - \mathbf{M}\right)\mathbf{x}_{t-1}\right\|^2 \\ &\ge \frac{1}{1+\gamma}\mathbb{E}\left\|\mathbf{B}\overline{\mathbf{y}}_{K-1} + \eta\mathbf{M}\overline{\mathbf{x}}_{K-1}\right\|^2 - \frac{1}{\gamma}\mathbb{E}\left\|\frac{1}{K}\sum_{t=1}^K(\mathbf{B}_{\xi,t} - \mathbf{B})\mathbf{y}_{t-1} + \frac{\eta}{K}\sum_{t=1}^K\left(\mathbf{B}_{\xi,t}\mathbf{B}_{\xi,t}^\top - \mathbf{M}\right)\mathbf{x}_{t-1}\right\|^2. \end{aligned} \tag{30}$$

Analogously,

$$\begin{aligned}
&\mathbb{E}\left\|\frac{1}{K}\sum_{t=1}^{K}\mathbf{B}_{\xi,t}^{\top}\mathbf{x}_{t-1}-\frac{\eta}{K}\sum_{t=1}^{K}\mathbf{B}_{\xi,t}^{\top}\mathbf{B}_{\xi,t}\mathbf{y}_{t-1}\right\|^2\\
&=\mathbb{E}\left\|\mathbf{B}^{\top}\overline{\mathbf{x}}_{K-1}-\eta\widehat{\mathbf{M}}\overline{\mathbf{y}}_{K-1}+\frac{1}{K}\sum_{t=1}^{K}(\mathbf{B}_{\xi,t}-\mathbf{B})^{\top}\mathbf{x}_{t-1}-\frac{\eta}{K}\sum_{t=1}^{K}\left(\mathbf{B}_{\xi,t}^{\top}\mathbf{B}_{\xi,t}-\widehat{\mathbf{M}}\right)\mathbf{y}_{t-1}\right\|^2 \qquad (31)\\
&\geq\frac{1}{1+\gamma}\mathbb{E}\left\|\mathbf{B}^{\top}\overline{\mathbf{x}}_{K-1}-\eta\widehat{\mathbf{M}}\overline{\mathbf{y}}_{K-1}\right\|^2-\frac{1}{\gamma}\mathbb{E}\left\|\frac{1}{K}\sum_{t=1}^{K}(\mathbf{B}_{\xi,t}-\mathbf{B})^{\top}\mathbf{x}_{t-1}-\frac{\eta}{K}\sum_{t=1}^{K}\left(\mathbf{B}_{\xi,t}^{\top}\mathbf{B}_{\xi,t}-\widehat{\mathbf{M}}\right)\mathbf{y}_{t-1}\right\|^2.
\end{aligned}$$

Combining the above two displays Eq. (30), Eq. (31) with Eq. (29) we have

$$\begin{aligned}
&\frac{1}{1+\gamma}\mathbb{E}\left[\left\|\mathbf{B}\overline{\mathbf{y}}_{K-1}+\eta\mathbf{M}\overline{\mathbf{x}}_{K-1}\right\|^2+\left\|\mathbf{B}^{\top}\overline{\mathbf{x}}_{K-1}-\eta\widehat{\mathbf{M}}\overline{\mathbf{y}}_{K-1}\right\|^2\right]\\
&\quad-\frac{1}{\gamma}\mathbb{E}\left\|\frac{1}{K}\sum_{t=1}^{K}(\mathbf{B}_{\xi,t}-\mathbf{B})\mathbf{y}_{t-1}+\frac{\eta}{K}\sum_{t=1}^{K}\left(\mathbf{B}_{\xi,t}\mathbf{B}_{\xi,t}^{\top}-\mathbf{M}\right)\mathbf{x}_{t-1}\right\|^2\\
&\quad-\frac{1}{\gamma}\mathbb{E}\left\|\frac{1}{K}\sum_{t=1}^{K}(\mathbf{B}_{\xi,t}-\mathbf{B})^{\top}\mathbf{x}_{t-1}-\frac{\eta}{K}\sum_{t=1}^{K}\left(\mathbf{B}_{\xi,t}^{\top}\mathbf{B}_{\xi,t}-\widehat{\mathbf{M}}\right)\mathbf{y}_{t-1}\right\|^2\\
&\leq\mathbb{E}\left\|\frac{\mathbf{x}_K-\mathbf{x}_0}{-\eta K}-\frac{1}{K}\sum_{t=1}^{K}\mathbf{g}_{\xi,t}^{\mathbf{x}}-\frac{\eta}{K}\sum_{t=1}^{K}\mathbf{B}_{\xi,t}\mathbf{g}_{\xi,t}^{\mathbf{y}}\right\|^2+\mathbb{E}\left\|\frac{\mathbf{y}_K-\mathbf{y}_0}{\eta K}-\frac{1}{K}\sum_{t=1}^{K}\mathbf{g}_{\xi,t}^{\mathbf{y}}+\frac{\eta}{K}\sum_{t=1}^{K}\mathbf{B}_{\xi,t}^{\top}\mathbf{g}_{\xi,t}^{\mathbf{x}}\right\|^2 \qquad (32)\\
&\leq(1+\beta)\mathbb{E}\left\|\frac{\mathbf{x}_K-\mathbf{x}_0}{-\eta K}\right\|^2+(1+\beta)\mathbb{E}\left\|\frac{\mathbf{y}_K-\mathbf{y}_0}{\eta K}\right\|^2\\
&\quad+\left(1+\frac{1}{\beta}\right)\mathbb{E}\left\|-\frac{1}{K}\sum_{t=1}^{K}\mathbf{g}_{\xi,t}^{\mathbf{x}}-\frac{\eta}{K}\sum_{t=1}^{K}\mathbf{B}_{\xi,t}\mathbf{g}_{\xi,t}^{\mathbf{y}}\right\|^2+\left(1+\frac{1}{\beta}\right)\mathbb{E}\left\|-\frac{1}{K}\sum_{t=1}^{K}\mathbf{g}_{\xi,t}^{\mathbf{y}}+\frac{\eta}{K}\sum_{t=1}^{K}\mathbf{B}_{\xi,t}^{\top}\mathbf{g}_{\xi,t}^{\mathbf{x}}\right\|^2,
\end{aligned}$$

where the last inequality is an application of Young's that involves an arbitrary fixed number $\beta\in(0,\infty)$. The rest of this proof follows in three steps:

(i) As a first step, we have from Eq. (28) along with Lemma C.2

$$\begin{aligned}
&\mathbb{E}\left\|\frac{\mathbf{x}_K-\mathbf{x}_0}{-\eta K}\right\|^2+\mathbb{E}\left\|\frac{\mathbf{y}_K-\mathbf{y}_0}{\eta K}\right\|^2\\
&\leq\frac{2}{\eta^2K^2}\left(\|\mathbf{x}_K\|^2+\|\mathbf{x}_0\|^2\right)+\frac{2}{\eta^2K^2}\left(\|\mathbf{y}_K\|^2+\|\mathbf{y}_0\|^2\right)\\
&\leq\frac{2}{\eta^2K^2}\left(\left(1-\eta^2\lambda^*(\eta)\right)^K\left[\|\mathbf{x}_0\|^2+\|\mathbf{y}_0\|^2\right]+\eta^2\mathcal{Q}_K(\eta)\left[1+\eta^2\left(\lambda_{\max}(\mathbf{M})\vee\lambda_{\max}(\widehat{\mathbf{M}})\right)\right]\sigma_{\mathbf{g}}^2\right)\\
&\quad+\frac{2}{\eta^2K^2}\left[\|\mathbf{x}_0\|^2+\|\mathbf{y}_0\|^2\right]\\
&\leq\frac{2}{\eta^2K^2}\left[2\|\mathbf{x}_0\|^2+2\|\mathbf{y}_0\|^2+\eta^2\mathcal{Q}_K(\eta)\left[1+\eta^2\left(\lambda_{\max}(\mathbf{M})\vee\lambda_{\max}(\widehat{\mathbf{M}})\right)\right]\sigma_{\mathbf{g}}^2\right]\\
&\leq\frac{4}{\eta^2K^2}[\|\mathbf{x}_0\|^2+\|\mathbf{y}_0\|^2]+\frac{2\left[1+\eta^2\left(\lambda_{\max}(\mathbf{M})\vee\lambda_{\max}(\widehat{\mathbf{M}})\right)\right]}{K}\sigma_{\mathbf{g}}^2.
\end{aligned}$$

(ii) A second step is, due to Assumption 2.2, standard L^2 martingale analysis gives

$$\begin{aligned}
&\mathbb{E}\left\|-\frac{1}{K}\sum_{t=1}^{K}\mathbf{g}^{\mathbf{x}}_{\xi,t}-\frac{\eta}{K}\sum_{t=1}^{K}\mathbf{B}_{\xi,t}\mathbf{g}^{\mathbf{y}}_{\xi,t}\right\|^2+\mathbb{E}\left\|-\frac{1}{K}\sum_{t=1}^{K}\mathbf{g}^{\mathbf{y}}_{\xi,t}+\frac{\eta}{K}\sum_{t=1}^{K}\mathbf{B}^\top_{\xi,t}\mathbf{g}^{\mathbf{x}}_{\xi,t}\right\|^2\\
&=\frac{1}{K^2}\sum_{t=1}^{K}\mathbb{E}\left\|-\mathbf{g}^{\mathbf{x}}_{\xi,t}-\eta\mathbf{B}_{\xi,t}\mathbf{g}^{\mathbf{y}}_{\xi,t}\right\|^2+\frac{1}{K^2}\sum_{t=1}^{K}\mathbb{E}\left\|-\mathbf{g}^{\mathbf{y}}_{\xi,t}+\eta\mathbf{B}^\top_{\xi,t}\mathbf{g}^{\mathbf{x}}_{\xi,t}\right\|^2\\
&=\frac{1}{K}\mathbb{E}_\xi\left\|-\mathbf{g}^{\mathbf{x}}_{\xi}-\eta\mathbf{B}_{\xi}\mathbf{g}^{\mathbf{y}}_{\xi}\right\|^2+\frac{1}{K}\mathbb{E}_\xi\left\|-\mathbf{g}^{\mathbf{y}}_{\xi}+\eta\mathbf{B}^\top_{\xi}\mathbf{g}^{\mathbf{x}}_{\xi}\right\|^2\\
&=\frac{1}{K}\mathbb{E}_\xi\|\mathbf{g}^{\mathbf{x}}_{\xi}\|^2+\frac{\eta^2}{K}\mathbb{E}_\xi\|\mathbf{B}_\xi\mathbf{g}^{\mathbf{y}}_{\xi}\|^2+\frac{1}{K}\mathbb{E}_\xi\|\mathbf{g}^{\mathbf{y}}_{\xi}\|^2+\frac{\eta^2}{K}\mathbb{E}_\xi\|\mathbf{B}^\top_\xi\mathbf{g}^{\mathbf{x}}_{\xi}\|^2\\
&\quad\underbrace{+\frac{2\eta}{K}\mathbb{E}_\xi\langle\mathbf{g}^{\mathbf{x}}_{\xi},\mathbf{B}_\xi\mathbf{g}^{\mathbf{y}}_{\xi}\rangle-\frac{2\eta}{K}\mathbb{E}_\xi\langle\mathbf{g}^{\mathbf{y}}_{\xi},\mathbf{B}^\top_\xi\mathbf{g}^{\mathbf{x}}_{\xi}\rangle}_{\text{cross term}=0}\\
&\le\frac{1}{K}\mathbb{E}_\xi\left[\|\mathbf{g}^{\mathbf{x}}_{\xi}\|^2+\|\mathbf{g}^{\mathbf{y}}_{\xi}\|^2\right]+\frac{\eta^2}{K}\mathbb{E}_\xi\left[\|\mathbf{B}_\xi\mathbf{g}^{\mathbf{y}}_{\xi}\|^2+\|\mathbf{B}^\top_\xi\mathbf{g}^{\mathbf{x}}_{\xi}\|^2\right]\\
&\le\frac{1+\eta^2\left(\lambda_{\max}(\mathbf{M})\vee\lambda_{\max}(\widehat{\mathbf{M}})\right)}{K}\sigma^2_{\mathbf{g}},
\end{aligned}$$

where a similar analysis as in the proof of Theorem A.1 was adopted.

(iii) A third step is that, due to $\lambda^*(\eta)\ge 0$ of Lemma C.2,

$$\begin{aligned}
&\mathbb{E}\left\|\frac{1}{K}\sum_{t=1}^{K}(\mathbf{B}_{\xi,t}-\mathbf{B})\mathbf{y}_{t-1}+\frac{\eta}{K}\sum_{t=1}^{K}\left(\mathbf{B}_{\xi,t}\mathbf{B}^\top_{\xi,t}-\mathbf{M}\right)\mathbf{x}_{t-1}\right\|^2\\
&\quad+\mathbb{E}\left\|\frac{1}{K}\sum_{t=1}^{K}(\mathbf{B}_{\xi,t}-\mathbf{B})^\top\mathbf{x}_{t-1}-\frac{\eta}{K}\sum_{t=1}^{K}\left(\mathbf{B}^\top_{\xi,t}\mathbf{B}_{\xi,t}-\widehat{\mathbf{M}}\right)\mathbf{y}_{t-1}\right\|^2\\
&\le\frac{2}{K^2}\sum_{t=1}^{K}\mathbb{E}\left\|(\mathbf{B}_{\xi,t}-\mathbf{B})\mathbf{y}_{t-1}\right\|^2+\frac{2\eta^2}{K^2}\sum_{t=1}^{K}\mathbb{E}\left\|\left(\mathbf{B}_{\xi,t}\mathbf{B}^\top_{\xi,t}-\mathbf{M}\right)\mathbf{x}_{t-1}\right\|^2\\
&\quad+\frac{2}{K^2}\sum_{t=1}^{K}\mathbb{E}\left\|(\mathbf{B}_{\xi,t}-\mathbf{B})^\top\mathbf{x}_{t-1}\right\|^2+\frac{2\eta^2}{K^2}\sum_{t=1}^{K}\mathbb{E}\left\|\left(\mathbf{B}^\top_{\xi,t}\mathbf{B}_{\xi,t}-\widehat{\mathbf{M}}\right)\mathbf{y}_{t-1}\right\|^2\\
&\le\frac{2(\sigma^2_{\mathbf{B}}+\eta^2\sigma^2_{\mathbf{B},2})}{K^2}\sum_{t=1}^{K}\mathbb{E}\left[\|\mathbf{x}_{t-1}\|^2+\|\mathbf{y}_{t-1}\|^2\right]\\
&\le\frac{2(\sigma^2_{\mathbf{B}}+\eta^2\sigma^2_{\mathbf{B},2})}{K^2}\sum_{t=1}^{K}\bigg(\left(1-\eta^2\lambda^*(\eta)\right)^{t-1}\left[\|\mathbf{x}_0\|^2+\|\mathbf{y}_0\|^2\right]\\
&\qquad\qquad+\eta^2\mathcal{Q}_{t-1}(\eta)\left[1+\eta^2\left(\lambda_{\max}(\mathbf{M})\vee\lambda_{\max}(\widehat{\mathbf{M}})\right)\right]\sigma^2_{\mathbf{g}}\bigg)\\
&\le\frac{2(\sigma^2_{\mathbf{B}}+\eta^2\sigma^2_{\mathbf{B},2})}{K}\left(\|\mathbf{x}_0\|^2+\|\mathbf{y}_0\|^2+\left[1+\eta^2\left(\lambda_{\max}(\mathbf{M})\vee\lambda_{\max}(\widehat{\mathbf{M}})\right)\right]\lambda^*(\eta)^{-1}\sigma^2_{\mathbf{g}}\right),
\end{aligned}$$

where since $\eta^2\lambda^*(\eta)\in[0,1/4]$ we applied the result of Eq. (28).

Putting the above pieces together, along with Eq. (32), yields for any $K \geq 0$ we have[11]

$$\begin{aligned}
&\frac{1}{1+\gamma}\mathbb{E}\left[\left\|\mathbf{B}\overline{\mathbf{y}}_{K-1}+\eta\mathbf{M}\overline{\mathbf{x}}_{K-1}\right\|^2+\left\|\mathbf{B}^\top\overline{\mathbf{x}}_{K-1}-\eta\widehat{\mathbf{M}}\overline{\mathbf{y}}_{K-1}\right\|^2\right]\\
&\le 2\mathbb{E}\left\|\frac{\mathbf{x}_K-\mathbf{x}_0}{-\eta K}\right\|^2+2\mathbb{E}\left\|\frac{\mathbf{y}_K-\mathbf{y}_0}{\eta K}\right\|^2\\
&\quad+2\mathbb{E}\left\|-\frac{1}{K}\sum_{t=1}^{K}\mathbf{g}^{\mathbf{x}}_{\xi,t}-\frac{\eta}{K}\sum_{t=1}^{K}\mathbf{B}_{\xi,t}\mathbf{g}^{\mathbf{y}}_{\xi,t}\right\|^2+2\mathbb{E}\left\|-\frac{1}{K}\sum_{t=1}^{K}\mathbf{g}^{\mathbf{y}}_{\xi,t}+\frac{\eta}{K}\sum_{t=1}^{K}\mathbf{B}^\top_{\xi,t}\mathbf{g}^{\mathbf{x}}_{\xi,t}\right\|^2\\
&\quad+\frac{1}{\gamma}\mathbb{E}\left\|\frac{1}{K}\sum_{t=1}^{K}(\mathbf{B}_{\xi,t}-\mathbf{B})\mathbf{y}_{t-1}+\frac{\eta}{K}\sum_{t=1}^{K}\left(\mathbf{B}_{\xi,t}\mathbf{B}^\top_{\xi,t}-\mathbf{M}\right)\mathbf{x}_{t-1}\right\|^2\\
&\quad+\frac{1}{\gamma}\mathbb{E}\left\|\frac{1}{K}\sum_{t=1}^{K}(\mathbf{B}_{\xi,t}-\mathbf{B})^\top\mathbf{x}_{t-1}-\frac{\eta}{K}\sum_{t=1}^{K}\left(\mathbf{B}^\top_{\xi,t}\mathbf{B}_{\xi,t}-\widehat{\mathbf{M}}\right)\mathbf{y}_{t-1}\right\|^2,
\end{aligned}$$

which is further bounded by

$$\begin{aligned}
&\frac{8}{\eta^2K^2}[\|\mathbf{x}_0\|^2+\|\mathbf{y}_0\|^2]+\frac{4\left[1+\eta^2\left(\lambda_{\max}(\mathbf{M})\vee\lambda_{\max}(\widehat{\mathbf{M}})\right)\right]}{K}\sigma^2_{\mathbf{g}}\\
&+\frac{2\left[1+\eta^2\left(\lambda_{\max}(\mathbf{M})\vee\lambda_{\max}(\widehat{\mathbf{M}})\right)\right]}{K}\sigma^2_{\mathbf{g}}\\
&+\frac{1}{\gamma}\cdot\frac{2(\sigma^2_{\mathbf{B}}+\eta^2\sigma^2_{\mathbf{B},2})}{K}\left([\|\mathbf{x}_0\|^2+\|\mathbf{y}_0\|^2]+\left[1+\eta^2\left(\lambda_{\max}(\mathbf{M})\vee\lambda_{\max}(\widehat{\mathbf{M}})\right)\right]\lambda^*(\eta)^{-1}\sigma^2_{\mathbf{g}}\right)\\
\le&\left(\frac{8}{\eta^2K^2}+\frac{2(\sigma^2_{\mathbf{B}}+\eta^2\sigma^2_{\mathbf{B},2})}{\gamma K}\right)[\|\mathbf{x}_0\|^2+\|\mathbf{y}_0\|^2]\\
&+\frac{6+\frac{2}{\gamma}(\sigma^2_{\mathbf{B}}+\eta^2\sigma^2_{\mathbf{B},2})\lambda^*(\eta)^{-1}}{K}\left[1+\eta^2\left(\lambda_{\max}(\mathbf{M})\vee\lambda_{\max}(\widehat{\mathbf{M}})\right)\right]\sigma^2_{\mathbf{g}},
\end{aligned}$$

and by rearranging the terms in the last display along with Eq. (43) we have in finale (and shifting the time index forward by one)

$$\begin{aligned}
&\left(\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\left(1+\eta^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\right)-2\eta\sigma^2_{\mathbf{B}}\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}\right)\mathbb{E}\left[\|\overline{\mathbf{x}}_K\|^2+\|\overline{\mathbf{y}}_K\|^2\right]\\
&\le\mathbb{E}\left[\|\mathbf{B}\overline{\mathbf{y}}_K+\eta\mathbf{M}\overline{\mathbf{x}}_K\|^2+\left\|\mathbf{B}^\top\overline{\mathbf{x}}_K-\eta\widehat{\mathbf{M}}\overline{\mathbf{y}}_K\right\|^2\right]\\
&\le\left(\frac{8(1+\gamma)}{\eta^2(K+1)^2}+\frac{2\left(1+\frac{1}{\gamma}\right)(\sigma^2_{\mathbf{B}}+\eta^2\sigma^2_{\mathbf{B},2})}{K+1}\right)[\|\mathbf{x}_0\|^2+\|\mathbf{y}_0\|^2]\\
&\quad+\frac{6(1+\gamma)+2\left(1+\frac{1}{\gamma}\right)(\sigma^2_{\mathbf{B}}+\eta^2\sigma^2_{\mathbf{B},2})\lambda^*(\eta)^{-1}}{K+1}\left[1+\eta^2\left(\lambda_{\max}(\mathbf{M})\vee\lambda_{\max}(\widehat{\mathbf{M}})\right)\right]\sigma^2_{\mathbf{g}},
\end{aligned}$$

where we used the iterated laws of expectation at multiple occasions, as well as the property of L^2 martingale differences as well as the definitions Eq. (3) and Eq. (4) in Assumption 2.1. This concludes Eq. (25). The rest of the proof sits upon the application of Lemma C.2, esp. Eq. (18) and the fact that $1+\hat{\eta}_{\mathbf{M}}(\alpha)^2\left(\lambda_{\max}(\mathbf{M})\vee\lambda_{\max}(\widehat{\mathbf{M}})\right)\le\frac{3}{2}$, concluding the whole proof of Theorem 3.1.

□

Discussion We remark that the magnitude of $\mathcal{Q}_{K+1}(\eta)$ can be either $O(1)$ or $O(K)$, depending on whether $\lambda^*(\eta)$ is bounded away from zero or sufficiently close to zero. When applying iteration average, one needs to

[11] For simplicity we optimize the numerical constants on γ and take $\beta = 1$.

maximize the step size to achieve a sharp bound in which case it is sufficient to replace $\mathcal{Q}_{K+1}(\eta)$ in the bound by $K+1$ instead of $\frac{1}{\eta^2\lambda^*(\eta)}$. In special, the dependency on $\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right]$ can be improved to $O(\frac{1}{(K+1)^2})$ if we adopt the $\frac{1}{\eta^2\lambda^*(\eta)}$ bound for $\mathcal{Q}_{K+1}(\eta)$, achieving

$$\begin{aligned}
&(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\left(1+\eta^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\right)\mathbb{E}\left[\|\bar{\mathbf{x}}_K\|^2 + \|\bar{\mathbf{y}}_K\|^2\right] \\
&\le \frac{16 + 4(\sigma_{\mathbf{B}}^2 + \eta^2\sigma_{\mathbf{B},2}^2)\lambda^*(\eta)^{-1}}{\eta^2(K+1)^2}\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right] \\
&\quad + \frac{12 + 4(\sigma_{\mathbf{B}}^2 + \eta^2\sigma_{\mathbf{B},2}^2)\lambda^*(\eta)^{-1}}{K+1}\left[1+\eta^2\left(\lambda_{\max}(\mathbf{M}) \vee \lambda_{\max}(\widehat{\mathbf{M}})\right)\right]\sigma_{\mathbf{g}}^2,
\end{aligned} \tag{33}$$

which recovers (8). In the upcoming technical analysis for restarting, we do not, however, utilize this upper bound.

D.2 Analysis of Theorem 3.2

Theorem 3.2, Full Version *Under Assumptions 2.1 and 2.2 and assume that $\mathbf{B}_\xi, \mathbf{B}$ are square matrices, we apply restarting at* $\mathsf{epoch} = 1, 2, \dots, \mathsf{Epoch}$ *after $K \ge K_{\mathsf{epoch}}$ steps where*

$$K_{\mathsf{epoch}} = \left\lceil \frac{q_2 + \sqrt{q_2^2 + 4q_1q_3}}{2q_3} \right\rceil - 1, \tag{34}$$

with

$$\begin{aligned}
q_1 &\equiv \frac{16}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{e^{2-2\mathsf{epoch}}[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2]}{\hat{\eta}_{\mathbf{M}}(\alpha)^2} \\
q_2 &\equiv \frac{4(\sigma_{\mathbf{B}}^2 + \hat{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2)\left[e^{2-2\mathsf{epoch}}[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2] + \frac{3\sigma_{\mathbf{g}}^2}{\lambda_{\min}(\mathbf{M})\wedge\lambda_{\min}(\widehat{\mathbf{M}})}\right] + 18\sigma_{\mathbf{g}}^2}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \\
q_3 &\equiv \frac{\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2}{e^{2\mathsf{epoch}}},
\end{aligned} \tag{35}$$

where we denote

$$\mathsf{Epoch} = \left\lceil \frac{1}{2}\log\left(\frac{\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})}{3\sigma_{\mathbf{g}}^2}[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2]\right)\right\rceil.$$

Then for $\widehat{\mathsf{Epoch}} = 1, \dots, \mathsf{Epoch}$ *where* $K = \sum_{\mathsf{epoch}=1}^{\widehat{\mathsf{Epoch}}} K_{\mathsf{epoch}}$ *the iteration has the* expected squared Euclidean metric *that is discounted by a factor of* $1/e^{2\widehat{\mathsf{Epoch}}}$*:*

$$\mathbb{E}\left[\|\hat{\mathbf{x}}_K\|^2 + \|\hat{\mathbf{y}}_K\|^2\right] \le \frac{1}{e^{2\widehat{\mathsf{Epoch}}}}\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right],$$

and for $K = \sum_{\mathsf{epoch}=1}^{\mathsf{Epoch}} K_{\mathsf{epoch}} + \hat{K}$, $\hat{K} = 0, 1, \dots$ *where SEG with aforementioned restarting and (tail-) iteration averaging achieves*

$$\begin{aligned}
\mathbb{E}\left[\|\bar{\mathbf{x}}_K\|^2 + \|\bar{\mathbf{y}}_K\|^2\right] &\le \frac{16}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{\frac{3\sigma_{\mathbf{g}}^2}{\lambda_{\min}(\mathbf{M})\wedge\lambda_{\min}(\widehat{\mathbf{M}})}}{\hat{\eta}_{\mathbf{M}}(\alpha)^2(\hat{K}+1)^2} \\
&\quad + \frac{4(\sigma_{\mathbf{B}}^2 + \hat{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2) \cdot \frac{6\sigma_{\mathbf{g}}^2}{\lambda_{\min}(\mathbf{M})\wedge\lambda_{\min}(\widehat{\mathbf{M}})} + 18\sigma_{\mathbf{g}}^2}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{1}{\hat{K}+1} \\
&= \left[1 + \frac{\frac{8}{3(\hat{K}+1)} + O(\hat{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B}}^2 + \hat{\eta}_{\mathbf{M}}(\alpha)^4\sigma_{\mathbf{B},2}^2)}{\hat{\eta}_{\mathbf{M}}(\alpha)^2\left(\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})\right)}\right] \cdot \frac{18\sigma_{\mathbf{g}}^2}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{1}{\hat{K}+1},
\end{aligned} \tag{36}$$

which recovers Eq. (9).

Proof.[Proof of Theorem 3.2] Without loss of generality we consider the first epoch initialized at $\mathbf{x}_0, \mathbf{y}_0$. Recall from Eq. (8) we have

$$\begin{aligned}\mathbb{E}\left[\|\bar{\mathbf{x}}_K\|^2 + \|\bar{\mathbf{y}}_K\|^2\right] &\le \frac{16}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2}{\hat{\eta}_{\mathbf{M}}(\alpha)^2 (K+1)^2} \\ &+ \frac{4(\sigma_{\mathbf{B}}^2 + \hat{\eta}_{\mathbf{M}}(\alpha)^2 \sigma_{\mathbf{B},2}^2)\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2 + \frac{3\sigma_{\mathbf{g}}^2}{\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})}\right] + 18\sigma_{\mathbf{g}}^2}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{1}{K+1},\end{aligned} \tag{8}$$

so after K steps the iteration has a *squared Euclidean metric* that is discounted by a factor of $1/e^2$, in the sense that the right hand of the above display is $\le \frac{1}{e^2}\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right]$. This reduces to finding the solutions to the quadratic inequality

$$\frac{q_1}{(K+1)^2} + \frac{q_2}{K+1} - q_3 \le 0, \tag{37}$$

where q_1, q_2, q_3 were earlier defined as in Eq. (35) in the $\mathsf{epoch} = 1$ case, repeated as

$$\begin{aligned} q_1 &\equiv \frac{16}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2}{\hat{\eta}_{\mathbf{M}}(\alpha)^2} \\ q_2 &\equiv \frac{4(\sigma_{\mathbf{B}}^2 + \hat{\eta}_{\mathbf{M}}(\alpha)^2 \sigma_{\mathbf{B},2}^2)\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2 + \frac{3\sigma_{\mathbf{g}}^2}{\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})}\right] + 18\sigma_{\mathbf{g}}^2}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \\ q_3 &\equiv \frac{\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2}{e^2}. \end{aligned}$$

The root formula gives (and omitting the infeasible solutions)

$$\frac{1}{K+1} \le \frac{-q_2 + \sqrt{q_2^2 + 4q_1q_3}}{2q_1} \quad \text{or equivalently} \quad K \ge \left\lceil \frac{q_2 + \sqrt{q_2^2 + 4q_1q_3}}{2q_3} \right\rceil - 1.$$

This gives the epoch number Eq. (34).

To get a sensible bound on time complexity, instead of solving the quadratic formula Eq. (37) we instead consider to upper bound of K_{epoch} and its summation $\sum_{\mathsf{epoch}=1}^{\mathsf{Epoch}} K_{\mathsf{epoch}}$. Recall first we set $\eta = \hat{\eta}_{\mathbf{M}}(\alpha)$ defined earlier as in Eq. (7). From time to time we omit the ceilings for simplicity (which does not affect the magnitude as the terms grow large). Since $\mathbf{B}_\xi$ is a square matrix, we set $\eta = \eta_{\mathbf{M}}$ as in Eq. (7) again and apply the following restarting schedule: run SEG at $\mathsf{epoch} = 1, 2, \ldots$ for an iteration number of K_{epoch} defined as in Eq. (34), one can upper bound the iterate number at $\mathsf{epoch} = 1, 2, \ldots$ a maximal over two terms where

$$\begin{aligned} q_1 &\equiv \frac{16}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \cdot \frac{e^{2-2\mathsf{epoch}}[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2]}{\hat{\eta}_{\mathbf{M}}(\alpha)^2} \\ q_2 &\equiv \frac{4(\sigma_{\mathbf{B}}^2 + \hat{\eta}_{\mathbf{M}}(\alpha)^2 \sigma_{\mathbf{B},2}^2)\left[e^{2-2\mathsf{epoch}}[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2] + \frac{3\sigma_{\mathbf{g}}^2}{\lambda_{\min}(\mathbf{M}) \wedge \lambda_{\min}(\widehat{\mathbf{M}})}\right] + 18\sigma_{\mathbf{g}}^2}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} \\ q_3 &\equiv \frac{\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2}{e^{2\mathsf{epoch}}}, \end{aligned} \tag{35}$$

which allows the following bound[12]

$$
\begin{aligned}
K_{\text{complexity}} &\le \sum_{\mathsf{epoch}=1}^{\mathsf{Epoch}} \left[\frac{2q_2}{q_3} + \sqrt{\frac{2q_1}{q_3}}\right] \\
&\le \sum_{\mathsf{epoch}=1}^{\mathsf{Epoch}} \left[\frac{4e^2(\sigma_{\mathbf{B}}^2 + \hat\eta_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2)}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} + \frac{18\sigma_{\mathbf{g}}^2 + 4(\sigma_{\mathbf{B}}^2 + \hat\eta_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2)\frac{3\sigma_{\mathbf{g}}^2}{\lambda_{\min}(\mathbf{M})\wedge\lambda_{\min}(\widehat{\mathbf{M}})}}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2]} \cdot e^{2\mathsf{epoch}}\right] \\
&\quad + \sum_{\mathsf{epoch}=1}^{\mathsf{Epoch}} \sqrt{\frac{16e^2}{(1-\alpha)\hat\eta_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}},
\end{aligned}
$$

which is further bounded by

$$
\begin{aligned}
K_{\text{complexity}} &\le \left(\sqrt{\frac{16e^2}{(1-\alpha)\hat\eta_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}} + \frac{4e^2(\sigma_{\mathbf{B}}^2 + \hat\eta_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2)}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}\right) \cdot \mathsf{Epoch} \\
&\quad + \frac{18\sigma_{\mathbf{g}}^2 + 4(\sigma_{\mathbf{B}}^2 + \hat\eta_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2)\frac{3\sigma_{\mathbf{g}}^2}{\lambda_{\min}(\mathbf{M})\wedge\lambda_{\min}(\widehat{\mathbf{M}})}}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2]} \cdot \sum_{\mathsf{epoch}=1}^{\mathsf{Epoch}} e^{2\mathsf{epoch}} \\
&= \left(\sqrt{\frac{16e^2}{(1-\alpha)\hat\eta_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}} + \frac{4e^2(\sigma_{\mathbf{B}}^2 + \hat\eta_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2)}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}\right) \cdot \left\lceil \frac{1}{2}\mathsf{LOG} \right\rceil \\
&\quad + \frac{6(\lambda_{\min}(\mathbf{M})\wedge\lambda_{\min}(\widehat{\mathbf{M}})) + 4(\sigma_{\mathbf{B}}^2 + \hat\eta_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2)}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2]} \cdot \frac{\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2}{1-e^{-2}},
\end{aligned}
$$

where

$$
\mathsf{LOG} \equiv \log\left(\frac{\lambda_{\min}(\mathbf{M})\wedge\lambda_{\min}(\widehat{\mathbf{M}})}{3\sigma_{\mathbf{g}}^2}[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2]\right),
$$

and we applied the fact

$$
\sum_{\mathsf{epoch}=1}^{\mathsf{Epoch}} e^{2\mathsf{epoch}} = \frac{e^{2\mathsf{Epoch}-2} - 1}{1-e^{-2}} \le \frac{\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2}{1-e^{-2}} \cdot \frac{\lambda_{\min}(\mathbf{M})\wedge\lambda_{\min}(\widehat{\mathbf{M}})}{3\sigma_{\mathbf{g}}^2},
$$

hence concluding Eq. (10) and the theorem. □

D.3 Analysis of Theorem 3.3

Theorem 3.3, Full Version *Let Assumptions 2.1 and 2.2 hold with $\sigma_{\mathbf{g}} = 0$. When $\mathbf{B}_\xi, \mathbf{B}$ are square matrices, for any prescribed $\alpha \in (0,1)$ choosing the step size $\eta = \bar\eta_{\mathbf{M}}(\alpha)$ defined as in Eq. (7), for an iteration number of $K \ge K_{\text{thres}}(\alpha)$ defined as in Eq. (11). Then we have for all $K \ge 1$ that is divisible by $K_{\text{thres}}(\alpha)$ the following convergence rate for $\hat{\mathbf{x}}_K, \hat{\mathbf{y}}_K$ (outputs of Algorithm 1) holds*

$$
\mathbb{E}\left[\|\hat{\mathbf{x}}_K\|^2 + \|\hat{\mathbf{y}}_K\|^2\right] \le \exp\left(-\frac{2}{K_{\text{thres}}(\alpha)} \cdot K\right)\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right]. \tag{38}
$$

In above we adopt the fine-grained iteration number per epoch

$$
K_{\text{thres}}(\alpha) \equiv \left(\frac{4}{\sqrt{2\bar\eta_{\mathbf{M}}(\alpha)^2\left(\sigma_{\mathbf{B}}^2 + \bar\eta_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2\right) + 8\mathsf{\mathit{RATE}}} - \sqrt{2\bar\eta_{\mathbf{M}}(\alpha)^2\left(\sigma_{\mathbf{B}}^2 + \bar\eta_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2\right)}}\right)^2, \tag{11'}
$$

[12] Note it is easy to verify $\left\lceil \frac{q_2 + \sqrt{q_2^2 + 4q_1q_3}}{2q_3} \right\rceil - 1 \le \max\left(\frac{2q_2}{q_3}, \sqrt{\frac{2q_1}{q_3}}\right) \le \frac{2q_2}{q_3} + \sqrt{\frac{2q_1}{q_3}}$ by considering the two cases of $\frac{2q_2}{q_3} \le \sqrt{\frac{2q_1}{q_3}}$ and $\frac{2q_2}{q_3} \ge \sqrt{\frac{2q_1}{q_3}}$, separately.

where

$$\mathit{RATE} \equiv \sqrt{\frac{(1-\alpha)\bar{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}{e^2}}. \tag{39}$$

In the regime of $\sigma_{\mathbf{B}}, \sigma_{\mathbf{B},2} \to 0^+$ we do asymptotic expansion and get[13]

$$\begin{aligned}\frac{2}{K_{\text{thres}}} &= \frac{1}{8}\left(\sqrt{2\bar{\eta}_{\mathbf{M}}(\alpha)^2\left(\sigma_{\mathbf{B}}^2+\bar{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2\right)+8\mathsf{RATE}} - \sqrt{2\bar{\eta}_{\mathbf{M}}(\alpha)^2\left(\sigma_{\mathbf{B}}^2+\bar{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2\right)}\right)^2 \\ &= \frac{1}{e}\sqrt{(1-\alpha)\bar{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)} - O\left(\sqrt[4]{\bar{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}\cdot\sqrt{\bar{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B}}^2+\bar{\eta}_{\mathbf{M}}(\alpha)^4\sigma_{\mathbf{B},2}^2}\right),\end{aligned}$$

which along with Eq. (38) gives Eq. (12) of Theorem 3.3.

Now for the interpolation setting, in the case of square matrices $\mathbf{B}_\xi, \mathbf{B}$, we turn to consider the convergence rate of SEG with iteration averaging, stating the following lemma:

Lemma D.1 *Let Assumptions 2.1 and 2.2 hold with $\sigma_{\mathbf{g}} = 0$. Under the condition on step size $\eta \in (0, \eta_{\mathbf{M}}]$ where $\eta_{\mathbf{M}}$ was earlier defined as in Eq. (6), we conclude for all $K \geq 0$ the following convergence rate for $\overline{\mathbf{x}}_K, \overline{\mathbf{y}}_K$ holds*

$$\begin{aligned}&\left(\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\left(1+\eta^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\right) - 2\eta\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}\right)\mathbb{E}\left[\left\|\overline{\mathbf{x}}_K\right\|^2+\left\|\overline{\mathbf{y}}_K\right\|^2\right] \\ &\leq \mathbb{E}\left[\left\|\mathbf{B}\overline{\mathbf{y}}_K+\eta\mathbf{M}\overline{\mathbf{x}}_K\right\|^2+\left\|\mathbf{B}^\top\overline{\mathbf{x}}_K-\eta\widehat{\mathbf{M}}\overline{\mathbf{y}}_K\right\|^2\right] \\ &\leq \left(\frac{2}{\eta}+\sqrt{2(\sigma_{\mathbf{B}}^2+\eta^2\sigma_{\mathbf{B},2}^2)\mathcal{Q}_{K+1}(\eta)}\right)^2\frac{\|\mathbf{x}_0\|^2+\|\mathbf{y}_0\|^2}{(K+1)^2}.\end{aligned} \tag{40}$$

In addition when $\mathbf{B}_\xi, \mathbf{B}$ are square matrices, we have

$$\mathbb{E}\left[\|\overline{\mathbf{x}}_K\|^2+\|\overline{\mathbf{y}}_K\|^2\right] \leq \mathcal{P}_{K+1}(\eta)\cdot\frac{\|\mathbf{x}_0\|^2+\|\mathbf{y}_0\|^2}{(K+1)^2}. \tag{41}$$

In above the prefactor is defined as[14]

$$\mathcal{P}_{K+1}(\eta) \equiv \begin{cases} +\infty & \textit{if } \lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\left(1+\eta^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\right) \leq 2\eta\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})} \\ \dfrac{\left(2+\sqrt{2\eta^2\left(\sigma_{\mathbf{B}}^2+\eta^2\sigma_{\mathbf{B},2}^2\right)\mathcal{Q}_{K+1}(\eta)}\right)^2}{\eta^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)(1+\eta^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top))-2\eta^3\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}} & \textit{otherwise}\end{cases},$$

where $\mathcal{Q}_K(\eta)$ was earlier defined as in Eq. (21), and by setting η as

$$\bar{\eta}_{\mathbf{M}}(\alpha) \equiv \eta_{\mathbf{M}} \wedge \frac{\alpha\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}{2\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}, \tag{7}$$

we have

$$\mathcal{P}_{K+1}(\bar{\eta}_{\mathbf{M}}(\alpha)) \leq \frac{2}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}\left(\sqrt{\frac{2}{\bar{\eta}_{\mathbf{M}}(\alpha)^2}}+\sqrt{\left(\sigma_{\mathbf{B}}^2+\bar{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2\right)(K+1)}\right)^2. \tag{42}$$

Lemma D.1 can be seen as a fine-grained version of Theorem 3.1, and its proof is provided in §D.4.1. To understand it consider Theorem D.1 in the case where $\mathbf{B}_\xi$ is nonstochastic so $\sigma_{\mathbf{B}} = \sigma_{\mathbf{B},2} = 0$, taking η as the maximal $\eta_{\mathbf{M}} = \frac{1}{\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}$ then Eq. (42) achieves the optimal prefactor which is bounded by the quadruple condition number of $\mathbf{B}^\top\mathbf{B}$. In the general case where $\mathbf{B}_\xi$ is stochastic, the convergence rate upper bound Eq. (41) has the nonrandom component as $O(1/K^2)$ as well as the random component of $O(1/K)$.

To prepare the proof we first introduce the following "metric conversion" lemma that translates bounds between two metrics:

[13] We used the Taylor's asymptotic expansion $(\sqrt{x+a}-\sqrt{x})^2 = a(\sqrt{1+x/a}-\sqrt{x/a})^2 = a - O(\sqrt{ax})$ as $x \to 0^+$ for fixed positive a.

[14] Here we interpret $0 \cdot (+\infty)$ as $+\infty$ whenever it occurs.

Lemma D.2 *We have for any* $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$ *that*

$$\begin{aligned}
&\|\mathbf{B}\mathbf{y} + \eta\mathbf{M}\mathbf{x}\|^2 + \left\|\mathbf{B}^\top\mathbf{x} - \eta\widehat{\mathbf{M}}\mathbf{y}\right\|^2 = \left\|\begin{bmatrix}\mathbf{B}^\top & -\eta\widehat{\mathbf{M}} \\ \eta\mathbf{M} & \mathbf{B}\end{bmatrix}\begin{bmatrix}\mathbf{x} \\ \mathbf{y}\end{bmatrix}\right\|^2 \\
\geq &\left(\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\left(1 + \eta^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\right) - 2\eta\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}\right)\left[\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2\right].
\end{aligned} \tag{43}$$

Lemma D.2 establishes a lower bound on a modified version of the Hamiltonian metric by (a constant multiple of) the squared Euclidean norm metric. The proof of the above inequality is due to an estimation of the spectral lower bound of a matrix. To take a first glance note in the nonrandom case $\sigma_{\mathbf{B}} = 0$, $\widehat{\mathbf{M}} = \mathbf{B}^\top\mathbf{B}$ and $\mathbf{M} = \mathbf{B}\mathbf{B}^\top$ and we conclude Eq. (43) in the form

$$\begin{aligned}
&\|\mathbf{B}\mathbf{y} + \eta\mathbf{M}\mathbf{x}\|^2 + \left\|\mathbf{B}^\top\mathbf{x} - \eta\widehat{\mathbf{M}}\mathbf{y}\right\|^2 = \left\|\mathbf{B}\mathbf{y} + \eta\mathbf{B}\mathbf{B}^\top\mathbf{x}\right\|^2 + \left\|\mathbf{B}^\top\mathbf{x} - \eta\mathbf{B}^\top\mathbf{B}\mathbf{y}\right\|^2 \\
= &\,\mathbf{x}^\top\left(\mathbf{B}\mathbf{B}^\top + \eta^2(\mathbf{B}\mathbf{B}^\top)^2\right)\mathbf{x} + \mathbf{y}^\top\left(\mathbf{B}^\top\mathbf{B} + \eta^2(\mathbf{B}^\top\mathbf{B})^2\right)\mathbf{y} \\
\geq &\,\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\left(1 + \eta^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\right)\left[\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2\right].
\end{aligned}$$

We prove the lemma for the general stochastic $\mathbf{B}_\xi$ case; details are deferred to §D.4.2.

We are ready for the proof of Theorem 3.3.

Proof.[Proof of Theorem 3.3] From Lemma D.1 we have

$$\mathbb{E}\left[\|\overline{\mathbf{x}}_K\|^2 + \|\overline{\mathbf{y}}_K\|^2\right] \leq \frac{2}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}\left(\sqrt{\frac{2}{\overline{\eta}_{\mathbf{M}}(\alpha)^2(K+1)^2}} + \sqrt{\frac{\sigma_{\mathbf{B}}^2 + \overline{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2}{K+1}}\right)^2\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right],$$

so after K steps the metric $\mathbb{E}\left[\|\overline{\mathbf{x}}_K\|^2 + \|\overline{\mathbf{y}}_K\|^2\right] \leq \frac{1}{e^2}\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right]$, i.e. we only need

$$\sqrt{\frac{2}{(1-\alpha)\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}}\left(\sqrt{\frac{2}{\overline{\eta}_{\mathbf{M}}(\alpha)^2(K+1)^2}} + \sqrt{\frac{\sigma_{\mathbf{B}}^2 + \overline{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2}{K+1}}\right) \leq \frac{1}{e},$$

i.e.

$$\frac{2}{K+1} + \sqrt{2\overline{\eta}_{\mathbf{M}}(\alpha)^2\left(\sigma_{\mathbf{B}}^2 + \overline{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2\right)} \cdot \frac{1}{\sqrt{K+1}} \leq \sqrt{\frac{(1-\alpha)\overline{\eta}_{\mathbf{M}}(\alpha)^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}{e^2}} \equiv \mathsf{RATE},$$

solving the above inequality and ignoring the infeasible solutions gives for any prescribed $\alpha \in (0,1)$

$$K_{\text{thres}}(\alpha) + 1 = \left(\frac{\sqrt{2\overline{\eta}_{\mathbf{M}}(\alpha)^2\left(\sigma_{\mathbf{B}}^2 + \overline{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2\right)} + \sqrt{2\overline{\eta}_{\mathbf{M}}(\alpha)^2\left(\sigma_{\mathbf{B}}^2 + \overline{\eta}_{\mathbf{M}}(\alpha)^2\sigma_{\mathbf{B},2}^2\right) + 8\mathsf{RATE}}}{2\mathsf{RATE}}\right)^2,$$

which reduces to (11') after rationalizing the numerator. □

D.4 Auxiliary Proofs

D.4.1 Proof of Lemma D.1

For the proof of Lemma D.1, our analysis lends help of Young's inequality via coefficients $1+\gamma$, $1+\frac{1}{\gamma}$ with optimized coefficient $\gamma \in (0,\infty)$.

Proof.[Proof of Lemma D.1] Turning to Eq. (8), setting η as in Eq. (7) and telescoping both sides of the update rule Eq. (23) with $\mathbf{g}_{\xi,t}^{\mathbf{x}} = 0$ and $\mathbf{g}_{\xi,t}^{\mathbf{y}} = 0$ for $t = 1, \ldots, K$ gives

$$\begin{aligned}
-\eta^2\sum_{t=1}^{K}\mathbf{B}_{\xi,t}\mathbf{B}_{\xi,t}^\top\mathbf{x}_{t-1} - \eta\sum_{t=1}^{K}\mathbf{B}_{\xi,t}\mathbf{y}_{t-1} &= \mathbf{x}_K - \mathbf{x}_0 \\
-\eta^2\sum_{t=1}^{K}\mathbf{B}_{\xi,t}^\top\mathbf{B}_{\xi,t}\mathbf{y}_{t-1} + \eta\sum_{t=1}^{K}\mathbf{B}_{\xi,t}^\top\mathbf{x}_{t-1} &= \mathbf{y}_K - \mathbf{y}_0.
\end{aligned}$$

Manipulating gives

$$\begin{aligned}
&\mathbf{B}\overline{\mathbf{y}}_{K-1} + \eta\mathbf{M}\overline{\mathbf{x}}_{K-1} + \frac{1}{K}\sum_{t=1}^{K}(\mathbf{B}_{\xi,t} - \mathbf{B})\mathbf{y}_{t-1} + \frac{\eta}{K}\sum_{t=1}^{K}\left(\mathbf{B}_{\xi,t}\mathbf{B}_{\xi,t}^\top - \mathbf{M}\right)\mathbf{x}_{t-1} \\
&= \frac{1}{K}\sum_{t=1}^{K}\mathbf{B}_{\xi,t}\mathbf{y}_{t-1} + \frac{\eta}{K}\sum_{t=1}^{K}\mathbf{B}_{\xi,t}\mathbf{B}_{\xi,t}^\top\mathbf{x}_{t-1} = \frac{\mathbf{x}_K - \mathbf{x}_0}{-\eta K},
\end{aligned}$$

and

$$\begin{aligned}
&\mathbf{B}^\top\overline{\mathbf{x}}_{K-1} - \eta\widehat{\mathbf{M}}\overline{\mathbf{y}}_{K-1} + \frac{1}{K}\sum_{t=1}^{K}(\mathbf{B}_{\xi,t} - \mathbf{B})^\top\mathbf{x}_{t-1} - \frac{\eta}{K}\sum_{t=1}^{K}\left(\mathbf{B}_{\xi,t}^\top\mathbf{B}_{\xi,t} - \widehat{\mathbf{M}}\right)\mathbf{y}_{t-1} \\
&= \frac{1}{K}\sum_{t=1}^{K}\mathbf{B}_{\xi,t}^\top\mathbf{x}_{t-1} - \frac{\eta}{K}\sum_{t=1}^{K}\mathbf{B}_{\xi,t}^\top\mathbf{B}_{\xi,t}\mathbf{y}_{t-1} = \frac{\mathbf{y}_K - \mathbf{y}_0}{\eta K}.
\end{aligned}$$

Now we try to bound the sum of squared norms of the first part (i.e. first two terms) on the left hands in the above two displays: applying Young's inequality gives for any fixed $\gamma \in (0, \infty)$

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{B}\overline{\mathbf{y}}_{K-1} + \eta\mathbf{M}\overline{\mathbf{x}}_{K-1}\right\|^2 + \left\|\mathbf{B}^\top\overline{\mathbf{x}}_{K-1} - \eta\widehat{\mathbf{M}}\overline{\mathbf{y}}_{K-1}\right\|^2\right] \\
&\le (1+\gamma)\mathbb{E}\left\|\frac{\mathbf{x}_K - \mathbf{x}_0}{-\eta K}\right\|^2 + (1+\gamma)\mathbb{E}\left\|\frac{\mathbf{y}_K - \mathbf{y}_0}{\eta K}\right\|^2 \\
&\quad + \left(1 + \frac{1}{\gamma}\right)\mathbb{E}\left\|\frac{1}{K}\sum_{t=1}^{K}(\mathbf{B}_{\xi,t} - \mathbf{B})\mathbf{y}_{t-1} + \frac{\eta}{K}\sum_{t=1}^{K}\left(\mathbf{B}_{\xi,t}\mathbf{B}_{\xi,t}^\top - \mathbf{M}\right)\mathbf{x}_{t-1}\right\|^2 \\
&\quad + \left(1 + \frac{1}{\gamma}\right)\mathbb{E}\left\|\frac{1}{K}\sum_{t=1}^{K}(\mathbf{B}_{\xi,t} - \mathbf{B})^\top\mathbf{x}_{t-1} - \frac{\eta}{K}\sum_{t=1}^{K}\left(\mathbf{B}_{\xi,t}^\top\mathbf{B}_{\xi,t} - \widehat{\mathbf{M}}\right)\mathbf{y}_{t-1}\right\|^2 \\
&\le \frac{4(1+\gamma)}{\eta^2K^2}\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right] \\
&\quad + \frac{2\left(1+\frac{1}{\gamma}\right)}{K^2}\sum_{t=1}^{K}\left[\mathbb{E}\left\|(\mathbf{B}_{\xi,t} - \mathbf{B})\mathbf{y}_{t-1}\right\|^2 + \mathbb{E}\left\|(\mathbf{B}_{\xi,t} - \mathbf{B})^\top\mathbf{x}_{t-1}\right\|^2\right] \\
&\quad + \frac{2\left(1+\frac{1}{\gamma}\right)\eta^2}{K^2}\sum_{t=1}^{K}\left[\mathbb{E}\left\|\left(\mathbf{B}_{\xi,t}\mathbf{B}_{\xi,t}^\top - \mathbf{M}\right)\mathbf{x}_{t-1}\right\|^2 + \mathbb{E}\left\|\left(\mathbf{B}_{\xi,t}^\top\mathbf{B}_{\xi,t} - \widehat{\mathbf{M}}\right)\mathbf{y}_{t-1}\right\|^2\right],
\end{aligned}$$

where for each $t = 1, \ldots, K$ we have, by applying Eq. (3) and Eq. (4) in Assumption 2.1 on operator norms, that

$$\begin{aligned}
&\mathbb{E}_\xi\left\|(\mathbf{B}_{\xi,t} - \mathbf{B})\mathbf{y}_{t-1}\right\|^2 + \mathbb{E}_\xi\left\|(\mathbf{B}_{\xi,t} - \mathbf{B})^\top\mathbf{x}_{t-1}\right\|^2 \\
&= (\mathbf{y}_{t-1})^\top\mathbb{E}_\xi\left[(\mathbf{B}_{\xi,t} - \mathbf{B})^\top(\mathbf{B}_{\xi,t} - \mathbf{B})\right]\mathbf{y}_{t-1} + (\mathbf{x}_{t-1})^\top\mathbb{E}_\xi\left[(\mathbf{B}_{\xi,t} - \mathbf{B})(\mathbf{B}_{\xi,t} - \mathbf{B})^\top\right]\mathbf{x}_{t-1} \\
&\le \left\|\mathbb{E}_\xi\left[(\mathbf{B}_{\xi,t} - \mathbf{B})^\top(\mathbf{B}_{\xi,t} - \mathbf{B})\right]\right\|_{op}\|\mathbf{y}_{t-1}\|^2 + \left\|\mathbb{E}_\xi\left[(\mathbf{B}_{\xi,t} - \mathbf{B})(\mathbf{B}_{\xi,t} - \mathbf{B})^\top\right]\right\|_{op}\|\mathbf{x}_{t-1}\|^2 \\
&\le \sigma_{\mathbf{B}}^2\left[\|\mathbf{x}_{t-1}\|^2 + \|\mathbf{y}_{t-1}\|^2\right],
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}_\xi\left\|\left(\mathbf{B}_{\xi,t}\mathbf{B}_{\xi,t}^\top - \mathbf{M}\right)\mathbf{x}_{t-1}\right\|^2 + \mathbb{E}_\xi\left\|\left(\mathbf{B}_{\xi,t}^\top\mathbf{B}_{\xi,t} - \widehat{\mathbf{M}}\right)\mathbf{y}_{t-1}\right\|^2 \\
&= (\mathbf{x}_{t-1})^\top\mathbb{E}_\xi\left(\mathbf{B}_{\xi,t}\mathbf{B}_{\xi,t}^\top - \mathbf{M}\right)^2\mathbf{x}_{t-1} + (\mathbf{y}_{t-1})^\top\mathbb{E}_\xi\left(\mathbf{B}_{\xi,t}^\top\mathbf{B}_{\xi,t} - \widehat{\mathbf{M}}\right)^2\mathbf{y}_{t-1} \\
&\le \left\|\mathbb{E}_\xi\left(\mathbf{B}_{\xi,t}^\top\mathbf{B}_{\xi,t} - \widehat{\mathbf{M}}\right)^2\right\|_{op}\|\mathbf{y}_{t-1}\|^2 + \left\|\mathbb{E}_\xi\left(\mathbf{B}_{\xi,t}\mathbf{B}_{\xi,t}^\top - \mathbf{M}\right)^2\right\|_{op}\|\mathbf{x}_{t-1}\|^2 \\
&\le \sigma_{\mathbf{B},2}^2\left[\|\mathbf{x}_{t-1}\|^2 + \|\mathbf{y}_{t-1}\|^2\right].
\end{aligned}$$

Taking expectation once again gives, by applying Eq. (13) that for $t = 1, \ldots, K$

$$\begin{aligned}
&\mathbb{E}\left\|(\mathbf{B}_{\xi,t} - \mathbf{B})\mathbf{y}_{t-1}\right\|^2 + \mathbb{E}\left\|(\mathbf{B}_{\xi,t} - \mathbf{B})^\top \mathbf{x}_{t-1}\right\|^2 \\
&\le \sigma_{\mathbf{B}}^2 \mathbb{E}\left[\|\mathbf{x}_{t-1}\|^2 + \|\mathbf{y}_{t-1}\|^2\right] \le \sigma_{\mathbf{B}}^2 \left(1 - \eta^2 \lambda^*(\eta)\right)^{t-1} \left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right],
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left\|\left(\mathbf{B}_{\xi,t}\mathbf{B}_{\xi,t}^\top - \mathbf{M}\right)\mathbf{x}_{t-1}\right\|^2 + \mathbb{E}\left\|\left(\mathbf{B}_{\xi,t}^\top \mathbf{B}_{\xi,t} - \widehat{\mathbf{M}}\right)\mathbf{y}_{t-1}\right\|^2 \\
&\le \sigma_{\mathbf{B},2}^2 \mathbb{E}\left[\|\mathbf{x}_{t-1}\|^2 + \|\mathbf{y}_{t-1}\|^2\right] \le \sigma_{\mathbf{B},2}^2 \left(1 - \eta^2 \lambda^*(\eta)\right)^{t-1} \left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right],
\end{aligned}$$

so denoting $\mathcal{Q}_K(\eta) \equiv \sum_{t=1}^K \left(1 - \eta^2\lambda^*(\eta)\right)^{t-1}$ as in Eq. (21) concludes

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{B}\overline{\mathbf{y}}_{K-1} + \eta\mathbf{M}\overline{\mathbf{x}}_{K-1}\right\|^2 + \left\|\mathbf{B}^\top\overline{\mathbf{x}}_{K-1} - \eta\widehat{\mathbf{M}}\overline{\mathbf{y}}_{K-1}\right\|^2\right] \\
&\le \frac{4(1+\gamma)}{\eta^2 K^2}\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right] + \frac{2\left(1+\frac{1}{\gamma}\right)}{K^2}\sum_{t=1}^K \left[\mathbb{E}\left\|(\mathbf{B}_{\xi,t} - \mathbf{B})\mathbf{y}_{t-1}\right\|^2 + \mathbb{E}\left\|(\mathbf{B}_{\xi,t} - \mathbf{B})^\top\mathbf{x}_{t-1}\right\|^2\right] \\
&\quad + \frac{2\left(1+\frac{1}{\gamma}\right)\eta^2}{K^2}\sum_{t=1}^K \left[\mathbb{E}\left\|\left(\mathbf{B}_{\xi,t}\mathbf{B}_{\xi,t}^\top - \mathbf{M}\right)\mathbf{x}_{t-1}\right\|^2 + \mathbb{E}\left\|\left(\mathbf{B}_{\xi,t}^\top\mathbf{B}_{\xi,t} - \widehat{\mathbf{M}}\right)\mathbf{y}_{t-1}\right\|^2\right] \\
&\le \frac{4(1+\gamma)}{\eta^2 K^2}\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right] + \frac{2\left(1+\frac{1}{\gamma}\right)}{K^2}\sum_{t=1}^K \sigma_{\mathbf{B}}^2\left(1 - \eta^2\lambda^*(\eta)\right)^{t-1}\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right] \\
&\quad + \frac{2\left(1+\frac{1}{\gamma}\right)\eta^2}{K^2}\sum_{t=1}^K \sigma_{\mathbf{B},2}^2\left(1 - \eta^2\lambda^*(\eta)\right)^{t-1}\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right] \\
&\le \frac{4(1+\gamma)}{\eta^2 K^2}\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right] + \frac{2\left(1+\frac{1}{\gamma}\right)}{K^2}\left(\sigma_{\mathbf{B}}^2 + \eta^2\sigma_{\mathbf{B},2}^2\right)\mathcal{Q}_K(\eta)\left[\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2\right].
\end{aligned}$$

In above we used the iterated laws of expectation as well as the property of L^2 martingale at multiple occasions. Therefore

$$\begin{aligned}
&\left(\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\left(1 + \eta^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\right) - 2\eta\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}\right)\mathbb{E}\left[\left\|\overline{\mathbf{x}}_{K-1}\right\|^2 + \left\|\overline{\mathbf{y}}_{K-1}\right\|^2\right] \\
&\le \mathbb{E}\left[\left\|\mathbf{B}\overline{\mathbf{y}}_{K-1} + \eta\mathbf{M}\overline{\mathbf{x}}_{K-1}\right\|^2 + \left\|\mathbf{B}^\top\overline{\mathbf{x}}_{K-1} - \eta\widehat{\mathbf{M}}\overline{\mathbf{y}}_{K-1}\right\|^2\right] \\
&\le \inf_{\gamma\in(0,\infty)}\left(\frac{4(1+\gamma)}{\eta^2} + 2\left(1 + \frac{1}{\gamma}\right)\left(\sigma_{\mathbf{B}}^2 + \eta^2\sigma_{\mathbf{B},2}^2\right)\mathcal{Q}_K(\eta)\right)\frac{\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2}{K^2}.
\end{aligned} \tag{44}$$

Note by optimizing over $\gamma \in (0, \infty)$ in above the identity is

$$\inf_{\gamma\in(0,\infty)}\left(\frac{4\gamma}{\eta^2} + \frac{2}{\gamma}\left(\sigma_{\mathbf{B}}^2 + \eta^2\sigma_{\mathbf{B},2}^2\right)\mathcal{Q}_K(\eta)\right) = 2\sqrt{\frac{8}{\eta^2}\left(\sigma_{\mathbf{B}}^2 + \eta^2\sigma_{\mathbf{B},2}^2\right)\mathcal{Q}_K(\eta)},$$

so the prefactor on the right hand of Eq. (44) reduces to

$$\frac{4}{\eta^2} + 2\left(\sigma_{\mathbf{B}}^2 + \eta^2\sigma_{\mathbf{B},2}^2\right)\mathcal{Q}_K(\eta) + 2\sqrt{\frac{8}{\eta^2}\left(\sigma_{\mathbf{B}}^2 + \eta^2\sigma_{\mathbf{B},2}^2\right)\mathcal{Q}_K(\eta)} = \left(\frac{2}{\eta} + \sqrt{2\left(\sigma_{\mathbf{B}}^2 + \eta^2\sigma_{\mathbf{B},2}^2\right)\mathcal{Q}_K(\eta)}\right)^2,$$

concluding Eq. (40) by replacing K by $K+1$.

Now to finish the proof, by setting η as $\bar{\eta}_{\mathbf{M}}(\alpha)$ defined as in Eq. (7), we have a tight upper bound of the prefactor as the step size η over interval $(0, \bar{\eta}_{\mathbf{M}}(\alpha)]$ for a prescribed $\alpha \in (0,1)$, as

$$\begin{aligned}
\mathcal{P}_{K+1}(\eta) &= \frac{\left(2 + \sqrt{2\eta^2\left(\sigma_{\mathbf{B}}^2 + \eta^2\sigma_{\mathbf{B},2}^2\right)\mathcal{Q}_{K+1}(\eta)}\right)^2}{\eta^2\lambda_{\min}(\mathbf{BB}^\top)\left(1 + \eta^2\lambda_{\min}(\mathbf{BB}^\top)\right) - 2\eta^3\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}} \\
&\le \frac{\left(2 + \sqrt{2\eta^2\left(\sigma_{\mathbf{B}}^2 + \eta^2\sigma_{\mathbf{B},2}^2\right)(K+1)}\right)^2}{(1-\alpha)\eta^2\lambda_{\min}(\mathbf{BB}^\top)\left(1 + \eta^2\lambda_{\min}(\mathbf{BB}^\top)\right)} \\
&\le \frac{2}{(1-\alpha)\lambda_{\min}(\mathbf{BB}^\top)}\left(\sqrt{\frac{2}{\eta^2}} + \sqrt{\left(\sigma_{\mathbf{B}}^2 + \eta^2\sigma_{\mathbf{B},2}^2\right)(K+1)}\right)^2,
\end{aligned}$$

which is just

$$\mathcal{P}_{K+1}(\eta) \le \frac{2}{(1-\alpha)\lambda_{\min}(\mathbf{BB}^\top)}\left(\underbrace{\frac{2}{\eta^2} + \frac{2\sqrt{2(K+1)}}{\eta}\cdot\sqrt{\sigma_{\mathbf{B}}^2 + \eta^2\sigma_{\mathbf{B},2}^2}}_{\text{linearization}} + \underbrace{(K+1)\cdot\left(\sigma_{\mathbf{B}}^2 + \eta^2\sigma_{\mathbf{B},2}^2\right)}_{\text{higher-order term}}\right). \tag{42}$$

It is straightforward to verify that for a prescribed $\alpha \in (0,1)$, $\bar{\eta}_{\mathbf{M}}(\alpha)$ simply minimizes the upper bound in the last line of Eq. (42) when dropping the higher-order term in $\sqrt{\sigma_{\mathbf{B}}^2 + \eta^2\sigma_{\mathbf{B},2}^2}$, since such a *linearized prefactor* is nonincreasing over $\eta \in (0, \bar{\eta}_{\mathbf{M}}(\alpha)]$. This completes the proof of Eq. (40) and the full version of Lemma D.1.[15] □

[15] Note one can further optimize the above prefactor over $\alpha \in (0,1)$ (so that the convergence rate upper bound is minimized), but finding an interpretable closed-form solution can be unrealistic. An initial attempt on this thread is to optimize a surrogate function $\frac{\eta_{\mathbf{M}}^2}{(1-\alpha)\bar{\eta}_{\mathbf{M}}(\alpha)^2}$, $\alpha \in (0,1)$, which is upper-bounded by

$$\frac{1}{1-\alpha} + \frac{1}{\alpha^2(1-\alpha)}\left(\frac{2\eta_{\mathbf{M}}\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}{\lambda_{\min}(\mathbf{BB}^\top)}\right)^2 \equiv \frac{1}{1-\alpha} + \frac{\mathcal{A}}{\alpha^2(1-\alpha)},$$

In the regime of $\mathcal{A} \to 0^+$, its closed-form solution is available but hard to interpret, and standard asymptotic analysis indicates that $\alpha \sim \mathcal{A}^{1/3} = \left(\frac{2\eta_{\mathbf{M}}\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}{\lambda_{\min}(\mathbf{BB}^\top)}\right)^{2/3}$ minimizes the above display.

D.4.2 Proof of Lemma D.2

Proof.[Proof of Lemma D.2] The left hand of Eq. (43) reads

$$\begin{aligned}
&\|\mathbf{B}\mathbf{y}+\eta\mathbf{M}\mathbf{x}\|^2+\left\|\mathbf{B}^\top\mathbf{x}-\eta\widehat{\mathbf{M}}\mathbf{y}\right\|^2=\left\|\begin{bmatrix}\mathbf{B}^\top & -\eta\widehat{\mathbf{M}}\\ \eta\mathbf{M} & \mathbf{B}\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\right\|^2\\
&=\begin{bmatrix}\mathbf{x}^\top & \mathbf{y}^\top\end{bmatrix}\begin{bmatrix}\mathbf{B} & \eta\mathbf{M}\\ -\eta\widehat{\mathbf{M}} & \mathbf{B}^\top\end{bmatrix}\begin{bmatrix}\mathbf{B}^\top & -\eta\widehat{\mathbf{M}}\\ \eta\mathbf{M} & \mathbf{B}\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\\
&=\begin{bmatrix}\mathbf{x}^\top & \mathbf{y}^\top\end{bmatrix}\begin{bmatrix}\mathbf{B}\mathbf{B}^\top+\eta^2\mathbf{M}^2 & -\eta\mathbf{B}\widehat{\mathbf{M}}+\eta\mathbf{M}^\top\mathbf{B}\\ -\eta\widehat{\mathbf{M}}\mathbf{B}^\top+\eta\mathbf{B}^\top\mathbf{M} & \mathbf{B}^\top\mathbf{B}+\eta^2\widehat{\mathbf{M}}^2\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\\
&=\begin{bmatrix}\mathbf{x}^\top & \mathbf{y}^\top\end{bmatrix}\begin{bmatrix}\mathbf{B}\mathbf{B}^\top+\eta^2(\mathbf{B}\mathbf{B}^\top)^2 & 0\\ 0 & \mathbf{B}^\top\mathbf{B}+\eta^2(\mathbf{B}^\top\mathbf{B})^2\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\\
&\quad+\begin{bmatrix}\mathbf{x}^\top & \mathbf{y}^\top\end{bmatrix}\begin{bmatrix}\eta^2\mathbf{M}^2-\eta^2(\mathbf{B}\mathbf{B}^\top)^2 & 0\\ 0 & \eta^2\widehat{\mathbf{M}}^2-\eta^2(\mathbf{B}^\top\mathbf{B})^2\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\\
&\quad+\begin{bmatrix}\mathbf{x}^\top & \mathbf{y}^\top\end{bmatrix}\begin{bmatrix}0 & -\eta\mathbf{B}\widehat{\mathbf{M}}+\eta\mathbf{M}^\top\mathbf{B}\\ -\eta\widehat{\mathbf{M}}\mathbf{B}^\top+\eta\mathbf{B}^\top\mathbf{M} & 0\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\\
&=\mathbf{x}^\top\left[\mathbf{B}\mathbf{B}^\top+\eta^2(\mathbf{B}\mathbf{B}^\top)^2\right]\mathbf{x}+\mathbf{y}^\top\left[\mathbf{B}^\top\mathbf{B}+\eta^2(\mathbf{B}^\top\mathbf{B})^2\right]\mathbf{y}\\
&\quad+\mathbf{x}^\top\left[\eta^2\mathbf{M}^2-\eta^2(\mathbf{B}\mathbf{B}^\top)^2\right]\mathbf{x}+\mathbf{y}^\top\left[\eta^2\widehat{\mathbf{M}}^2-\eta^2(\mathbf{B}^\top\mathbf{B})^2\right]\mathbf{y}\\
&\quad+2\mathbf{y}^\top\left(-\eta\widehat{\mathbf{M}}\mathbf{B}^\top+\eta\mathbf{B}^\top\mathbf{M}\right)\mathbf{x}\\
&\equiv \mathrm{I}+\mathrm{II}+\mathrm{III},
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{I}&=\mathbf{x}^\top\left[\mathbf{B}\mathbf{B}^\top+\eta^2(\mathbf{B}\mathbf{B}^\top)^2\right]\mathbf{x}+\mathbf{y}^\top\left[\mathbf{B}^\top\mathbf{B}+\eta^2(\mathbf{B}^\top\mathbf{B})^2\right]\mathbf{y}\\
&=\left\|\mathbf{B}\mathbf{y}+\eta\mathbf{B}\mathbf{B}^\top\mathbf{x}\right\|^2+\left\|\mathbf{B}^\top\mathbf{x}-\eta\mathbf{B}^\top\mathbf{B}\mathbf{y}\right\|^2.
\end{aligned}$$

In addition, we have from Assumption 2.1 that $\mathbf{M}-\mathbf{B}\mathbf{B}^\top=\mathbb{E}\left[\mathbf{B}_\xi\mathbf{B}_\xi^\top\right]-\mathbf{B}\mathbf{B}^\top=\mathbb{E}\left[(\mathbf{B}_\xi-\mathbf{B})(\mathbf{B}_\xi-\mathbf{B})^\top\right]\succeq\mathbf{0}$ and analogously $\widehat{\mathbf{M}}-\mathbf{B}^\top\mathbf{B}=\mathbb{E}\left[\mathbf{B}_\xi^\top\mathbf{B}_\xi\right]-\mathbf{B}^\top\mathbf{B}=\mathbb{E}\left[(\mathbf{B}_\xi-\mathbf{B})^\top(\mathbf{B}_\xi-\mathbf{B})\right]\succeq 0$ that almost surely

$$\mathrm{II}\geq\mathbf{x}^\top\left[\eta^2\mathbf{M}^2-\eta^2(\mathbf{B}\mathbf{B}^\top)^2\right]\mathbf{x}+\mathbf{y}^\top\left[\eta^2\widehat{\mathbf{M}}^2-\eta^2(\mathbf{B}^\top\mathbf{B})^2\right]\mathbf{y}\geq 0.$$

The third term

$$\mathrm{III}=2\mathbf{y}^\top\left(-\eta\widehat{\mathbf{M}}\mathbf{B}^\top+\eta\mathbf{B}^\top\mathbf{M}\right)\mathbf{x}$$

satisfies

$$\begin{aligned}
|\mathrm{III}|&=\left|2\mathbf{y}^\top\left(-\eta\widehat{\mathbf{M}}\mathbf{B}^\top+\eta\mathbf{B}^\top\mathbf{M}\right)\mathbf{x}\right|\\
&=2\left|\mathbf{y}^\top(-\eta\widehat{\mathbf{M}}\mathbf{B}^\top+\eta\mathbf{B}^\top\mathbf{B}\mathbf{B}^\top+\eta\mathbf{B}^\top\mathbf{M}-\eta\mathbf{B}^\top\mathbf{B}\mathbf{B}^\top)\mathbf{x}\right|\\
&\leq 2\eta\left|-\mathbf{y}^\top\left(\widehat{\mathbf{M}}-\mathbf{B}^\top\mathbf{B}\right)\mathbf{B}^\top\mathbf{x}\right|+2\eta\left|\mathbf{y}^\top\mathbf{B}^\top\left(\mathbf{M}-\mathbf{B}\mathbf{B}^\top\right)\mathbf{x}\right|\\
&\leq 2\eta\left\|\left(\widehat{\mathbf{M}}-\mathbf{B}^\top\mathbf{B}\right)\mathbf{y}\right\|\left\|\mathbf{B}^\top\mathbf{x}\right\|+2\eta\left\|\left(\mathbf{M}-\mathbf{B}\mathbf{B}^\top\right)\mathbf{x}\right\|\left\|\mathbf{B}\mathbf{y}\right\|\\
&\leq 2\eta\sigma_{\mathbf{B}}^2\left\|\mathbf{y}\right\|\left\|\mathbf{B}^\top\mathbf{x}\right\|+2\eta\sigma_{\mathbf{B}}^2\left\|\mathbf{x}\right\|\left\|\mathbf{B}\mathbf{y}\right\|\\
&\leq 2\eta\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}\left[\|\mathbf{x}\|^2+\|\mathbf{y}\|^2\right],
\end{aligned}$$

where we have from Eq. (3) and Eq. (4) of Assumption 2.1 that for all $\mathbf{x}\in\mathbb{R}^n,\mathbf{y}\in\mathbb{R}^m$ that $\left\|(\widehat{\mathbf{M}}-\mathbf{B}^\top\mathbf{B})\mathbf{y}\right\|\leq\sigma_{\mathbf{B}}^2\left\|\mathbf{y}\right\|$ and $\left\|(\mathbf{M}-\mathbf{B}\mathbf{B}^\top)\mathbf{x}\right\|\leq\sigma_{\mathbf{B}}^2\left\|\mathbf{x}\right\|$.

One final piece is that for any given $\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m$ and $\eta > 0$

$$\begin{aligned} &\left\|\mathbf{B}\mathbf{y} + \eta\mathbf{B}\mathbf{B}^\top\mathbf{x}\right\|^2 + \left\|\mathbf{B}^\top\mathbf{x} - \eta\mathbf{B}^\top\mathbf{B}\mathbf{y}\right\|^2 \\ &= \mathbf{x}^\top\left[\mathbf{B}\mathbf{B}^\top + \eta^2(\mathbf{B}\mathbf{B}^\top)^2\right]\mathbf{x} + \mathbf{y}^\top\left[\mathbf{B}^\top\mathbf{B} + \eta^2(\mathbf{B}^\top\mathbf{B})^2\right]\mathbf{y} \\ &\geq \left(1 + \eta^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\right)\left[\left\|\mathbf{B}^\top\mathbf{x}\right\|^2 + \|\mathbf{B}\mathbf{y}\|^2\right] \geq \lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\left(1 + \eta^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\right)\left[\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2\right]. \end{aligned} \tag{45}$$

Now to conclude Eq. (43), we apply Eq. (45), and the above analysis (taking the expectation) gives the following result

$$\begin{aligned} \|\mathbf{B}\mathbf{y} + \eta\mathbf{M}\mathbf{x}\|^2 + \left\|\mathbf{B}^\top\mathbf{x} - \eta\widehat{\mathbf{M}}\mathbf{y}\right\|^2 &= \mathrm{I} + \mathrm{II} + \mathrm{III} \geq \mathrm{I} - |\mathrm{III}| \\ = \left\|\mathbf{B}\mathbf{y} + \eta\mathbf{B}\mathbf{B}^\top\mathbf{x}\right\|^2 + \left\|\mathbf{B}^\top\mathbf{x} - \eta\mathbf{B}^\top\mathbf{B}\mathbf{y}\right\|^2 &- 2\eta\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}\left[\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2\right], \end{aligned}$$

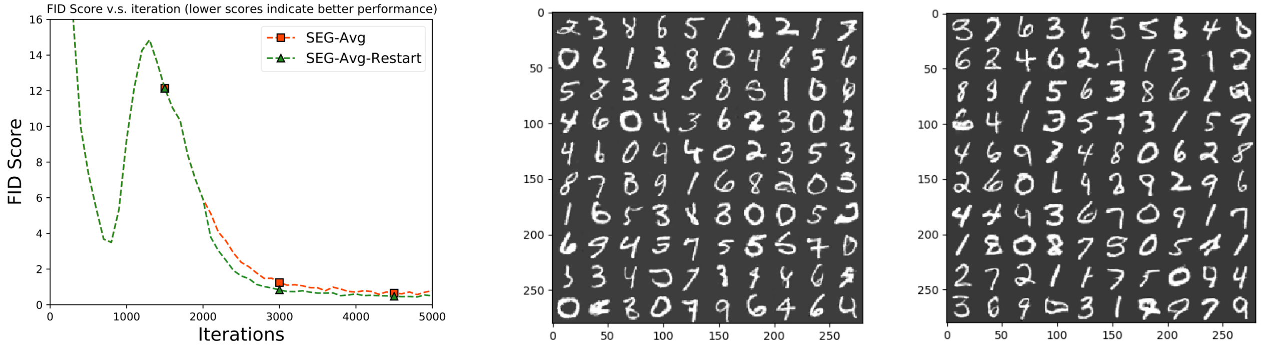
which is no less than

$$\left(\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\left(1 + \eta^2\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\right) - 2\eta\sigma_{\mathbf{B}}^2\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}\right)\left[\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2\right],$$

again due to Eq. (45). □

E ADDITIONAL EXPERIMENTAL RESULTS

Experiments on GANs. We conduct GANs experiments on MNIST [LeCun, 1998] understand more the empirical performance of our restarted iteration-averaged SEG in non-convex non-concave minimax optimization problems. Specifically, we adopt the DCGAN network architecture proposed in Radford et al. [2015] and use the loss proposed in Goodfellow et al. [2014]. We adopt ExtraAdam [Gidel et al., 2019] as the optimizer and denote the ExtraAdam with averaging by SEG-Avg. Meanwhile, we apply restarting at iteration 2000 (halfway the total run) for SEG-Avg and denote the restarting method by SEG-Avg-Restart. We apply Fréchet Inception distance (FID) [Heusel et al., 2017] score for measuring GAN performances, and compare the performance of SEG-Avg and SEG-Avg-Restart in Figure 6. We observe that proper restarting schedule improves the model performance w.r.t. FID score (lower scores indicate better performance).



(a) FID comparison.

(b) SEG-Avg.

(c) SEG-Avg-Restart.

Figure 6: GAN experimental results. (**a**). Comparing GAN performance (measure by FID score) of SEG-Avg and SEG-Avg-Restart. (**b**). Images generated by SEG-Avg. (**c**). Images generated by SEG-Avg-Restart.