STOCHASTIC EXTRAGRADIENT METHODS BEYOND LIPSCHITZ CONTINUITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Stochastic variational inequality problems have received significant attention due to their ability to model large-scale machine learning tasks. Typically, the operators involved are assumed to be globally Lipschitz, with only a handful of works extending beyond the standard setting to the so-called (L_0, L_1)-Lipschitz setting. In this work, we propose a novel stochastic past extragradient algorithm for variational inequality problems under a broad regularity assumption that is even less restrictive than (L_0, L_1)-Lipschitz continuity and provide convergence guarantees under bounded variance of the stochastic oracle. More precisely, we show a.s. convergence with suitable stepsize schedules and obtain convergence rates in expectation and in probability for fixed horizons. Finally, in numerical simulations we compare our proposed algorithm against standard methods.

1 INTRODUCTION

Variational inequalities (VI) and root-finding problems (Facchinei & Pang, 2003) provide a common language for modeling a broad range of learning problems including constrained minimization, multiplayer games and minimax optimization. Typical applications include adversarial network training, distributionally robust optimization and optimal transport. In their simplest (unconstrained) form, VI problems require finding the zeroes of an operator F, i.e., finding $x^\star$ such that $F(x^\star) = 0$.

Probably the most well-known forward methods for tackling such problems are the extragradient (EG) (Korpelevich, 1977) and the past extragradient (PEG) method (Popov, 1980), which require additional gradient (forward) steps per iteration, with EG also requiring an extra computation of F. Traditional assumptions for this type of problems are the monotonicity and Lipschitz continuity of F, which allow for a clean analysis but can be restrictive in modern learning problems. For that reason, a large part of the recent literature has been devoted to lifting these strict assumptions.

A prominent direction for relaxing monotonicity is based on *Minty-type* conditions. Instead of requiring the pairwise inequality defining monotonicity, the Minty variational inequality (MVI) only assumes the existence of a solution $x^\star$ such that

$$\langle F(x), x - x^\star \rangle \geq 0 \qquad \text{for all } x.$$

More recently, Diakonikolas et al. (2021) introduced the *weak Minty* condition, which allows a controlled violation of this inequality proportional to $\|F(x)\|^2$, and showed that a suitably relaxed extragradient method, EG+, still converges under this substantially weaker assumption. This line of work was further developed by Pethick et al. (2023b), who enlarged the admissible range of non-monotonicity and extended the theory to constrained and regularized problems, as well as by Pethick et al. (2023a) in the stochastic setting. More recent results have pushed the weak-Minty framework further through multi-step exploration (Fan et al., 2024), refined fixed-point analyses (Alacaoglu et al., 2024), and schemes interpolating between extragradient and proximal-point methods (Pethick et al., 2025a).

A complementary line of work has sought to relax the global Lipschitz continuity assumption. Motivated in part by the observation that the local smoothness of modern machine learning objectives can vary substantially over the optimization trajectory, (L_0, L_1)-smoothness allows the local Lipschitz constant to grow with the gradient norm (Zhang et al., 2019; 2020). An analogous notion for general operators is the α-symmetric (L_0, L_1)-Lipschitz condition, under which the variation of F

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

is controlled by a quantity that depends on the magnitude of the operator itself. Vankov et al. (2024) recently extended this framework to variational inequalities and established convergence guarantees for adaptive versions of projection, extragradient, and optimistic methods under generalized smoothness. Most recently, Choudhury & Loizou (2025) provided a refined analysis of extragradient for (L_0, L_1)-Lipschitz root-finding problems, obtaining guarantees for monotone and strongly monotone operators as well as local convergence under weak Minty structure. Nevertheless, much less attention has been devoted to the stochastic setting, where to the best of our knowledge only the work of Vankov et al. (2025a) studies (L_0, L_1)-Lipschitz type problems. Their analysis however relies on a p-quasi-sharpness condition, which imposes a global error bound and a uniform alignment of the operator with the solution set. In particular, this condition is sufficiently strong to yield convergence guarantees even for suitably clipped projection/gradient descent-ascent (GDA) schemes.

In this paper we aim at addressing this gap by proposing and analyzing nonlinearly preconditioned stochastic past extragradient methods under a general notion of continuity that is in fact even less restrictive than (L_0, L_1)-Lipschitz continuity. More precisely, our contributions are the following:

- We propose a stochastic past extragradient method for solving VI problems involving operators that are not Lipschitz continuous. The class of operators we consider, described in Definition 3.1, is larger than that of the recently introduced class of (L_0, L_1)-Lipschitz continuous operators (Vankov et al., 2025a; Choudhury & Loizou, 2025). Assuming bounded variance of the stochastic oracle, we show almost sure convergence to solutions of the VI in Theorem 4.8 and obtain convergence rates in expectation for a natural preconditioned residual in Corollary 4.6. Moreover, we get convergence guarantees for the standard residual in probability in Theorems 4.9 and 4.10, complementing our in expectation results.
- We present various examples of operators satisfying our regularity assumptions, describing both theoretical constructions and potential applications. Furthermore, we compare our proposed Algorithm 1 against standard solvers such as the two-stepsize versions of the stochastic extragradient and stochastic optimistic gradient methods in various numerical experiments ranging from simple quartic games to neural optimal transport and GAN training.

1.1 Notation

Let $\mathbb{N} = \{0, 1, 2, \dots\}$ denote the naturals including 0. We denote by $\langle \cdot, \cdot \rangle$ the standard Euclidean inner product on $\mathbb{R}^n$ and by $\| \cdot \|$ the standard Euclidean norm on $\mathbb{R}^n$ as well as the spectral norm for matrices. For a closed, convex set $C \subseteq \mathbb{R}^n$ and $\bar{x} \in C$ we denote its normal cone $N_C(\bar{x}) = \{y \in \mathbb{R}^n \mid \langle y, x - \bar{x} \rangle \leq 0 : \forall x \in C\}$. The squared distance to a set E will be denoted as $\operatorname{dist}^2(x, E) = \inf_{\bar{x} \in E} \|x - \bar{x}\|^2$. For $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ a set-valued mapping, we define its graph as $\operatorname{gph} T := \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n \mid y \in T(x)\}$ and its inverse as $T^{-1}(y) := \{x \in \mathbb{R}^n \mid (x, y) \in \operatorname{gph} T\}$.

2 Related work

Generalized smoothness. Our work is closely related to works on generalized smoothness conditions in optimization. Starting from Zhang et al. (2019) that introduced the notion of (L_0, L_1)-smoothness to better describe functions where the Hessian norm grows with the gradient norm, there has been a substantial body of work devoted to this topic. In Wang et al. (2023); Faw et al. (2023); Koloskova et al. (2023); Pethick et al. (2025b); Riabinin et al. (2025), the authors focused on analyzing the convergence properties of existing algorithms under this new condition, while Gorbunov et al. (2025); Vankov et al. (2025b) suggest new algorithms tailored to this smoothness condition. Our proposed method Algorithm 1 utilizes independent sampling for the update direction and the normalization, which has been used in the context of generalized smoothness in Gaash et al. (2025) and for variational inequality problems in Vankov et al. (2025a).

Nonlinear (dual space) preconditioning. The generalized Lipschitz continuity inequality we utilize in Definition 3.1 is influenced by the literature on dual space or nonlinear preconditioning for gradient descent and the corresponding anisotropic descent inequality. In Maddison et al. (2021) dual space preconditioning for convex optimization was introduced based on a notion of generalized cocoercivity. The method was extended to the nonconvex additive nonsmooth setting in Laude & Patrinos (2025) and was analyzed under the notion of anisotropic smoothness which was itself first

introduced in Laude et al. (2023). The connection of this new descent condition with (L_0, L_1)-smoothness that we build upon in this work was first established in Oikonomidis et al. (2025a). Oikonomidis et al. (2025b) analyzed the method with Polyak momentum and under simple noise assumptions, while Bodard & Patrinos (2025) studied the saddle point escape properties of nonlinearly preconditioned gradient methods and provided various new examples of (L_0, L_1)-smooth functions. A nonlinearly preconditioned (anisotropic) proximal point algorithm for finding zeroes of monotone operators was extensively studied in Laude & Patrinos (2023).

Stochastic methods for VI problems. In Juditsky et al. (2011) the Stochastic Mirror-Prox was analyzed and finite-time guarantees for monotone variational inequalities using stochastic prediction–correction steps were obtained. Subsequently, stochastic extragradient type methods have been studied under increasingly general conditions, including pseudomonotonicity, unbounded feasible sets, and nonuniform stochastic noise (Iusem et al., 2017). In Hsieh et al. (2020) the convergence of a two-timescale version of EG was established and in Bohm et al. (2022) the convergence of the stochastic forward-backward-forward method for convex-concave minimax problems was studied. Moreover, an extension of the stochastic EG method without large batch sizes to the weak Minty setting was proposed in Pethick et al. (2023a), while Gorbunov et al. (2022a; 2023) established the convergence of clipped versions of EG under heavy-tailed noise. A major line of work, that is not directly related to our paper, focuses on variance reduction techniques for finite-sum VIs (Palaniappan & Bach, 2016; Chavdarova et al., 2019; Alacaoglu & Malitsky, 2022; Emmanouilidis et al., 2024). A significant part of the related literature originates from learning in games with a particular emphasis on training GANs (Daskalakis et al., 2017; Mertikopoulos et al., 2018; Gidel et al., 2018) and reinforcement learning (Daskalakis et al., 2020; Kalogiannis & Farina, 2025).

3 Problem Statement and Algorithm

3.1 Problem Statement

We study stochastic methods for tackling VI problems of the form: find $x^\star \in C$ such that

$$\langle F(x^\star), x - x^\star \rangle \geq 0 \qquad \forall x \in C, \tag{VIP}$$

where $C \subseteq \mathbb{R}^n$ is a closed, convex set. We assume that such an $x^\star$ always exists and denote by S the set of all such points. Recall that (VIP) is also equivalent to finding the zeroes of the operator $F + N_C$, so that $S = (F + N_C)^{-1}(0)$.

Throughout the text, the jointly measurable map $G : \mathbb{R}^n \times \Xi \to \mathbb{R}^n$, satisfying $\mathbb{E}[\|G(x, \xi)\|] < +\infty$, $\forall x \in \mathbb{R}^n$, denotes a stochastic oracle for the population operator $F(x) = \mathbb{E}[G(x, \xi)]$. Regarding the regularity of F we do not follow the standard definition of (L_0, L_1)-Lipschitz continuity of Vankov et al. (2024, Assumption 2.1) but rather an extension of the preconditioned Lipschitz continuity inequality of Oikonomidis et al. (2025b, Assumption 2.5).

Definition 3.1. We say that F is $\mathcal{K}$-Lipschitz continuous with $\mathcal{K}(y) = \frac{y}{L_0 + L_1\|y\|}, L_0 > 0, L_1 \geq 0$ if the following inequality holds for all $x, \bar{x} \in \mathbb{R}^n$:

$$\|\mathcal{K}(F(x)) - \mathcal{K}(F(\bar{x}))\| \leq \|x - \bar{x}\|. \tag{1}$$

We denote $T(x) \coloneqq \mathcal{K}(F(x))$ and note that (1) is equivalent to the 1-Lipschitz continuity of T. This condition lies at the center of our analysis, since the Lipschitz continuity of T, together with the fact that T and F share zeroes, allows us to design algorithms acting directly on T. Nevertheless, in the stochastic setting we consider, this is not trivial: applying the preconditioner $\mathcal{K}(G(x, \xi))$ leads to a biased estimate of $T(x)$, i.e., $\mathbb{E}[\mathcal{K}(G(x, \xi))] \neq T(x)$. We explore this issue in more detail in the following subsection.

At the same time, Definition 3.1 serves another purpose: it defines an operator class that is larger than the class of (L_0, L_1)-Lipschitz operators. In the case where $F \in \mathcal{C}^1(\mathbb{R}^n)$, this can be seen by the differential characterizations. Indeed, for (L_0, L_1)-Lipschitz continuity, the necessary and sufficient condition is that

$$\|\nabla F(x)\| \leq L_0 + L_1\|F(x)\| \qquad \forall x \in \mathbb{R}^n, \tag{2}$$

while for Definition 3.1, $\|\nabla T(x)\| \leq 1$, i.e.,

$$\frac{1}{L_0 + L_1\|F(x)\|} \left\| \left(I - \frac{L_1}{L_0 + L_1\|F(x)\|} \frac{F(x)F(x)^\top}{\|F(x)\|} \right) \nabla F(x) \right\| \leq 1, \tag{3}$$

for $x \in \mathbb{R}^n \setminus F^{-1}(0)$, while by continuity $\nabla T(x) = \nabla F(x)/L_0$ for $x \in F^{-1}(0)$. It is straightforward to check that (2) implies (3). The reverse implication does not hold, however, with a standard example being from Bodard & Patrinos (2025): considering $F(x) = 2x\exp(x^2)$, there cannot be any L_0, L_1 such that (2) holds globally. On the other hand, for $L_0 = 2, L_1 = 1$, $\|\nabla T(x)\| \leq 1$ and thus Definition 3.1 holds. This relation between the two classes holds without requiring $F \in \mathcal{C}^1(\mathbb{R}^n)$, as is detailed in Appendix A. Examples of preconditioned Lipschitz operators

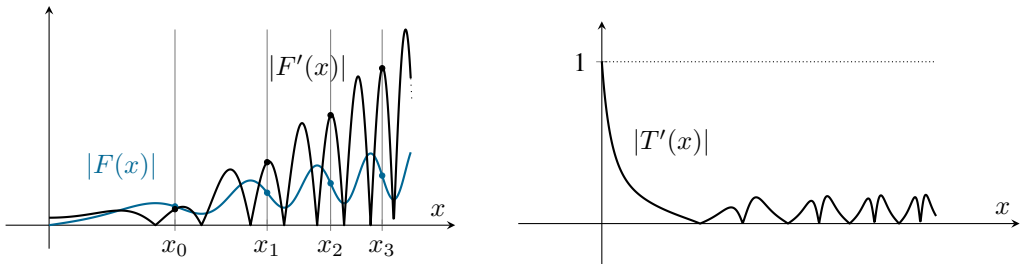

Figure 1: Comparison between preconditioned Lipschitz continuity and (L_0, L_1)-Lipschitz continuity for Example 3.2. *(Left)* (L_0, L_1)-Lipschitz continuity fails although $\|F(x)\|$ grows at most linearly. *(Right)* The gradient norm of the preconditioned operator T remains bounded by 1.

can be found in the related literature, with Maddison et al. (2021); Vankov et al. (2025b); Gorbunov et al. (2025); Oikonomidis et al. (2025a); Bodard & Patrinos (2025) covering the functional case $F = \nabla f$ for $f : \mathbb{R}^n \to \mathbb{R}$ and Vankov et al. (2024); Choudhury & Loizou (2025) covering the root-finding case. We provide some more interesting examples and applications in what follows, with more detailed derivations and further examples in Appendix A. We start with an example that satisfies Definition 3.1 but is not (L_0, L_1)-Lipschitz although it has at most linear growth, in contrast to $2x\exp(x^2)$ that we considered before.

Example 3.2 (Failure of (L_0, L_1)-Lipschitz with at most linear growth)**.** Let F be the gradient of the function $f(x) = \frac{x^2}{2} + \frac{1-\cos(x^2/2)}{2}$, i.e.,

$$F(x) = x\big(1 + \tfrac{1}{2}\sin(x^2/2)\big).$$

F satisfies Definition 3.1 with constants $L_0 = 1$ and $L_1 = 2$. Indeed, note that $|F(x)| \geq \frac{1}{2}|x|$, and since $F'(x) = 1 + \frac{1}{2}\sin(x^2/2) + \frac{x^2}{2}\cos(x^2/2)$, using $|\sin x| \leq |x|$, we have $|F'(x)| \leq 1 + \frac{3}{4}x^2 \leq 1 + x^2$. Therefore, $|T'(x)| = \frac{|F'(x)|}{(1+2|F(x)|)^2} \leq \frac{1+x^2}{(1+|x|)^2} \leq 1$. On the other hand, F is not (L_0, L_1)-Lipschitz for any finite combination of L_0 and L_1. This can be seen by taking a sequence $(x_\nu)_{\nu\in\mathbb{N}}$ of points $x_\nu = \sqrt{2(2\nu+1)\pi}$. Then, $\sin(x_\nu^2/2) = 0$ and $\cos(x_\nu^2/2) = -1$, and thus, $F(x_\nu) = x_\nu$ and $|F'(x_\nu)| = \frac{x_\nu^2}{2} - 1$. Therefore, $\frac{|F'(x_\nu)|}{|F(x_\nu)|} \to \infty$ and (2) cannot hold for finite $L_0 > 0$ and $L_1 \geq 0$. A visualization of this discrepancy can be found in Figure 1.

Another interesting example is a saddle operator where the coupling between the two variables is also nonlinear.

Example 3.3 (Quartic game with nonlinear cross-player interaction)**.** Let $x \in \mathbb{R}^n, y \in \mathbb{R}^m$ and $B \in \mathbb{R}^{n\times m}$. Consider the saddle function $L(x, y) = \frac{1}{4}\|x\|^4 - \frac{1}{4}\|y\|^4 + \frac{\lambda}{2}\|x\|^2\|y\|^2 + \langle x, By\rangle$ with $\lambda \in \mathbb{R}$. In this setting, the two players interact not only via the linear coupling but also via $\frac{\lambda}{2}\|x\|^2\|y\|^2$, leading to higher-order growth in the shared cost, rather than only in the player-specific terms. The corresponding saddle operator,

$$F(x, y) = \begin{pmatrix} (\|x\|^2 + \lambda\|y\|^2)x + By \\ (\|y\|^2 - \lambda\|x\|^2)y - B^\top x \end{pmatrix}$$

satisfies Definition 3.1 with $L_0 = \|B\| + 3 + 2|\lambda|$ and $L_1 = 6 + 4|\lambda|$.

As is now standard in the literature, our analysis assumes the Minty variational inequality, an extension of monotonicity requiring that the involved inequality only holds between general points and the solution set, rather than all pairs of points.

Algorithm 1 Stochastic Past Extra Preconditioned Gradient (SPEPG)

Input: Stochastic oracle G.
Input: Initial point $x^0 \in C$; $L_0 > 0$, $L_1 \geq 0$.
Input: Deterministic $0 < \alpha_k \leq 1, \gamma, \omega_k > 0$, with $\alpha_0 = 1$.

$m^{-1} \leftarrow 0$, $z^0 \leftarrow x^0$
Draw ξ^0, ζ^0 i.i.d. and compute $\widehat{T}(z^0; \xi^0, \zeta^0) = \frac{G(z^0, \xi^0)}{L_0 + L_1 \|G(z^0, \zeta^0)\|}$.
for $k = 0, 1, \ldots,$ **do**
  $m^k \leftarrow (1 - \alpha_k) m^{k-1} + \alpha_k \widehat{T}(z^k; \xi^k, \zeta^k)$
  $z^{k+1} \leftarrow P_C(x^k - \gamma m^k)$
  Draw ξ^{k+1}, ζ^{k+1} i.i.d. and compute $\widehat{T}(z^{k+1}; \xi^{k+1}, \zeta^{k+1}) = \frac{G(z^{k+1}, \xi^{k+1})}{L_0 + L_1 \|G(z^{k+1}, \zeta^{k+1})\|}$
  $x^{k+1} \leftarrow P_C(x^k - \omega_k \widehat{T}(z^{k+1}; \xi^{k+1}, \zeta^{k+1}))$

Definition 3.4. We say that F satisfies the Minty variational inequality if for all $x^\star \in S$, the following inequality holds:

$$\langle y, x - x^\star \rangle \geq 0 \qquad \forall (x, y) \in \operatorname{gph}(F + N_C). \tag{4}$$

Specializing to the case where $C = \mathbb{R}^n$, it is clear that if F satisfies Definition 3.4, then so does T. This observation motivates the design of a stochastic EG method by applying standard techniques to T instead of F. In general, this estimator is biased, that is, $\mathbb{E}[\mathcal{K}(G(x, \xi))] \neq T(x)$, so directly applying an analysis that requires an unbiased oracle for T does not guarantee convergence to a solution of the original problem. Our algorithm design therefore utilizes a more careful approach, which we detail in the next section.

Remark 3.5. In the related literature, the Minty variational inequality typically requires only one $x^\star \in S$ in (4) and not the whole set S as in Definition 3.4. Most of our analysis goes through with this more relaxed assumption, except for the a.s. full sequence convergence result in Theorem 4.8. We thus choose to keep this more restrictive assumption to make the presentation clear and concise and avoid different conditions between convergence theorems.

3.2 Algorithm

Our proposed extension of the past extragradient method is presented in Algorithm 1. The intuition behind our construction is relatively simple: assuming for simplicity that $C = \mathbb{R}^n$ and F is Lipschitz and monotone, the two-stepsize past extragradient (PEG) iteration takes the form:

$$z^{k+1} = x^k - \gamma F(z^k)$$
$$x^{k+1} = x^k - \omega F(z^{k+1}).$$

Since in our case, T is Lipschitz and satisfies the Minty VI, the next step is to replace F with T in the recursion displayed above. However, as discussed in the previous section, since in our stochastic setting we only have access to unbiased stochastic oracles G of F and not T, directly utilizing $\mathcal{K} \circ G$ as the oracle would not lead to convergence. Instead, we construct an estimator of $T(z^k)$ using two independently drawn samples ξ, ζ: given any $z \in \mathbb{R}^n$, we define

$$\widehat{T}(z; \xi, \zeta) \coloneqq \frac{G(z, \xi)}{L_0 + L_1 \|G(z, \zeta)\|}. \tag{5}$$

Despite the separate samples, in general $\mathbb{E}[\widehat{T}(z; \xi, \zeta)] \neq T(z)$. Nevertheless, due to the independence of the samples, it is easy to show that $\mathbb{E}[\widehat{T}(z; \xi, \zeta)] = \mathbb{E}\left[\frac{L_0 + L_1 \|F(z)\|}{L_0 + L_1 \|G(z, \zeta)\|}\right] T(z)$. Denoting $c(z) \coloneqq \mathbb{E}\left[\frac{L_0 + L_1 \|F(z)\|}{L_0 + L_1 \|G(z, \zeta)\|}\right]$, we have that $c(z) > 0$ which implies that in expectation the direction of the estimator is correct. In fact, under bounded variance (Assumption 4.1), we have uniform lower and upper bounds for c as we state formally in the following section. Finally, we use Polyak momentum in the update of z^{k+1} instead of using $\widehat{T}(z^k; \xi^k, \zeta^k)$ directly in order to make the most of its variance reduction properties.

Remark 3.6. Algorithm 1 can be extended to a standard extragradient type method where in the update of z^{k+1} we would use two more samples to compute $\widehat{T}(x^k; \xi_x^k, \xi_x^k)$. Nevertheless, since

this approach adds significant computational cost to the method, we decided not to include it in the current paper. In the unconstrained case, and with the same stepsize $\omega = \gamma$, it is well known that PEG and the optimistic gradient methods have equivalent forms (Gorbunov et al., 2022b, Equation (OG)). Since we are interested in the general VI problem and we utilize the two-sequence form of PEG in our analysis, we choose the name stochastic past extragradient for our proposed method.

4 Convergence analysis

In this section we present our main theoretical results regarding the convergence properties of our proposed scheme. The proofs can be found in Appendix B. We next formulate the main assumptions. We begin with the bounded variance assumption for the stochastic oracle, a standard assumption in the related literature, as well as the aforementioned regularity conditions on the operator F.

Assumption 4.1. For every $x \in \mathbb{R}^n$, we have $F(x) = \mathbb{E}[G(x, \xi)]$ and

$$\mathbb{E}[\|G(x, \xi) - F(x)\|^2] \leq \sigma^2.$$

Assumption 4.2. F is continuous and satisfies Definition 3.1 and Definition 3.4.

Under the bounded variance assumption we can formalize our results on the estimation of the preconditioned operator:

Lemma 4.3. *Let ξ, ζ be i.i.d. random variables, and suppose that Assumption 4.1 holds. Define*

$$c(x) := \mathbb{E}\left[\frac{L_0 + L_1\|F(x)\|}{L_0 + L_1\|G(x, \zeta)\|}\right], \quad x \in \mathbb{R}^n$$

and $\delta := \frac{L_1\sigma}{L_0}$. Then, for all $x \in \mathbb{R}^n$,

$$\mathbb{E}[\widehat{T}(x; \xi, \zeta)] = c(x)T(x), \qquad \frac{1}{1+\delta} \leq c(x) \leq 1 + \delta, \tag{6}$$

$$\mathbb{E}\left[\|\widehat{T}(x; \xi, \zeta) - c(x)T(x)\|^2\right] \leq \frac{2\sigma^2}{L_0^2}, \qquad \mathbb{E}\left[\|\widehat{T}(x; \xi, \zeta)\|^2\right] \leq c(x)^2\|T(x)\|^2 + \frac{2\sigma^2}{L_0^2}. \tag{7}$$

Finally, we assume the following step-size conditions, which follow directly from our analysis.

Assumption 4.4. The deterministic parameters satisfy

$$0 < \gamma \leq \frac{1}{4(1+\delta)}, \quad 0 < \theta \leq \frac{\gamma}{3(1+\delta)^2}, \quad \omega_k = \theta\alpha_{k+1}, \quad 0 < \alpha_{k+1} \leq 1, \quad k \in \mathbb{N}$$

where $\delta = \frac{L_1\sigma}{L_0}$ with σ as in Assumption 4.1 and L_0, L_1 as in Assumption 4.2.

Under these assumptions, we obtain the following main result.

Theorem 4.5. *Let Assumptions 4.1, 4.2 and 4.4 hold true. Let, moreover, $(z^k)_{k\in\mathbb{N}}$ be generated by Algorithm 1. Then, for every $K \geq 1$ and every $x^\star \in S$,*

$$\frac{1}{S_K}\sum_{k=0}^{K-1} \alpha_{k+1}\mathbb{E}[\mathrm{dist}^2(0, (T + N_C)(z^{k+1}))] \leq \frac{18}{7}\frac{(1+\delta)^2}{\theta\gamma S_K}\overline{V}_0(x^\star) + \frac{159}{14}(1+\delta)^4\frac{2\sigma^2}{L_0^2}\frac{Q_K}{S_K},$$

where $\overline{V}_0(x^\star)$ is some constant that depends on $x^\star$ and $S_K := \sum_{k=0}^{K-1}\alpha_{k+1}$ and $Q_K := \sum_{k=0}^{K-1}\alpha_{k+1}^2$.

Corollary 4.6. *Let Assumptions 4.1, 4.2 and 4.4 hold true. Let, moreover, $(z^k)_{k=1}^K$ be the sequence of iterates generated by running Algorithm 1 for $K \geq 1$ iterations with $\alpha_{k+1} = K^{-1/2}$ for all $k \in \{0, 1, \ldots, K-1\}$. Then,*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\mathrm{dist}^2(0, (T + N_C)(z^{k+1}))] = O(K^{-1/2}).$$

Remark 4.7. Some remarks are in order regarding Corollary 4.6. To begin with, our analysis leads to obtaining a rate on the mixed residual involving T instead of the usual F. This is a natural outcome since our regularity condition is the Lipschitzness of T which we combine with the fact that it is Minty. Moreover, $\operatorname{dist}^2(0,(T+N_C)(z))=0$ for some $z\in\mathbb{R}^n$ implies $-T(z)\in N_C(z)$ and thus that $-F(z)\in N_C(z)$ due to the cone properties of $N_C(z)$. It seems that the current analysis does not transfer these in expectation rates to F, unless some strict assumption such as boundedness of $\|F\|$ over the iterates is imposed. Nevertheless, we complement our results by showing a.s. full sequence convergence under standard momentum schedules and obtaining convergence rates in probability for the standard metrics.

Theorem 4.8. *Let Assumptions 4.1, 4.2 and 4.4 hold true. Let, moreover, $(x^k)_{k\in\mathbb{N}}$ be the sequence of iterates generated by Algorithm 1 with*

$$\sum_{k=0}^{\infty}\alpha_{k+1}=+\infty,\qquad \sum_{k=0}^{\infty}\alpha_{k+1}^2<+\infty.$$

Then, the sequence $(x^k)_{k\in\mathbb{N}}$ converges almost surely to some possibly random $x^\star\in S$.

Theorem 4.9. *Let Assumptions 4.1, 4.2 and 4.4 hold true. Let, moreover, $(z^k)_{k=1}^K$ be the sequence of iterates generated by running Algorithm 1 for $K\geq 1$ iterations with $\alpha_{k+1}=K^{-1/2}$ for all $k\in\{0,1,\dots,K-1\}$. Let J be uniformly distributed on $\{0,\dots,K-1\}$ and independent of the iterates. Then, for every $\eta\in(0,1)$, there exists a constant $C_\eta>0$, independent of K, such that, for every $K\geq 1$,*

$$\mathbb{P}\left\{\operatorname{dist}^2\left(0,(F+N_C)(z^{J+1})\right)>C_\eta K^{-1/2}\right\}\leq\eta.$$

In the unconstrained case (i.e., $C=\mathbb{R}^n$), we can strengthen the conclusions of Theorem 4.9 and additionally obtain an $O(K^{-1/2})$ bound on the failure probability given a tolerance $\varepsilon>0$.

Theorem 4.10. *Let the hypotheses of Theorem 4.9 hold true. If, additionally, $C=\mathbb{R}^n$, then, there exists a constant C_T, such that for all $\varepsilon>0$,*

$$\mathbb{P}\{\|F(z^{J+1})\|>\varepsilon\}\leq\frac{C_T(L_0+L_1\varepsilon)^2}{\sqrt{K}\varepsilon^2}.$$

To the best of our knowledge, our results are the first under (L_0,L_1)-Lipschitz continuity, and in fact under the less restrictive condition Definition 3.1, that merely require the Minty variational inequality Definition 3.4 to hold. The work Vankov et al. (2025a) that also tackles stochastic VI problems under generalized continuity assumptions requires a strong growth assumption in the form of Vankov et al. (2025a, Assumption 2.4).

5 Numerical Experiments

In this section we compare our proposed Algorithm 1 with standard algorithms like stochastic extra-gradient (SEG) and stochastic optimistic gradient (SOG). Moreover, for SEG and SOG we use two timescales in the stepsizes to allow for more flexibility. Further experimental results including GAN training on MNIST and neural optimal transport are presented in Appendix C.

5.1 Quartic game with nonlinear interaction

We first consider the setting of Example 3.3. We define $F(x,y)=\frac{1}{N}\sum_{i=1}^N F_i(x,y)$, $x,y\in\mathbb{R}^d$ with

$$F_i(x,y)=\begin{pmatrix}(\|x\|^2+\lambda\|y\|^2)x+B_iy\\(\|y\|^2-\lambda\|x\|^2)y-B_i^\top x\end{pmatrix},$$

where each $B_i=B+\sigma UD_iV^\top$ with $D_i=\operatorname{diag}(\varepsilon_{i,1},\dots,\varepsilon_{i,d})$ and $\varepsilon_{i,j}\in\{-1,1\}$ such that $\frac{1}{N}\sum_{i=1}^N B_i=B$ and $\sigma=0.5$. B is generated by taking singular values $\sigma_i\in[1/\kappa_B,1]$ and setting $B=U\operatorname{diag}(\sigma_i)_{i=1}^dV^\top$. We set $d=30$, $\lambda=L_1=1$ and tune the step-sizes γ,ω using a coarse logarithmic grid followed by a refinement over a linear grid around the selected parameters. More information on the tuning procedure is detailed in Appendix C.2.

We ran ten instances of the experiment. The results can be found in Figure 2 for $\kappa_B \in \{10, 100\}$. In both cases, we observe a similar trade-off as L_0 increases: For smaller values of L_0, the method progresses more slowly, but tends to achieve higher accuracy. As L_0 increases, the effect of the normalization diminishes so the methods behave more similarly to the baselines without preconditioning.

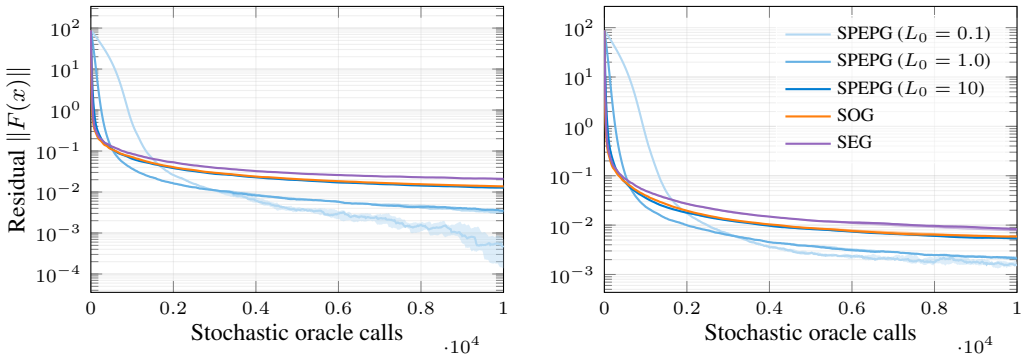

Figure 2: Convergence plots for the example in Section 5.1. The lines and shaded areas represent the median and interquartile range, respectively. (*left*) $\kappa(B) = 10$; (*right*) $\kappa(B) = 100$.

5.2 Cubic Game

We consider the two-player game with saddle point operator

$$F(x, y) = \begin{pmatrix} \sqrt{\langle x, Ax \rangle}\, Ax + By \\ \sqrt{\langle y, Cy \rangle}\, Cy - B^\top x \end{pmatrix},$$

where $A, C \in \mathbb{R}^{d \times d}$ are positive definite matrices, as described in Choudhury & Loizou (2025), and $B \in \mathbb{R}^{d \times d}$ is constructed similarly to the previous experiment. We consider two constructions for the stochastic oracle. Details are provided in Appendix C.2.

Finite sum case. Similarly to the example in Section 5.1, let $F_i(x, y) = \begin{pmatrix} \alpha_i \sqrt{x^\top Ax}\, Ax + B_i y \\ \delta_i \sqrt{y^\top Cy}\, Cy - B_i^\top x \end{pmatrix}$, $i = 1, \ldots, N$, where α_i, δ_i and B_i are constructed randomly to ensure $F(x, y) = \frac{1}{N} \sum_{i=1}^N F_i(x, y)$.

Student-t noise. To investigate the effect of heavy-tailed noise, we keep the nonlinear terms deterministic and perturb only the coupling matrix B for simplicity. At each oracle evaluation, we sample a matrix H with independent standard Student-t entries, $H_{jk} \sim t_\nu$, and replace B by $\widehat{B} = B + \frac{\sigma_B}{\sqrt{d}} H$, to obtain the oracle.

The results are presented in Figure 3. As in the previous experiment, we observe a similar trade-off as L_0 is varied. In particular, it is notable that for smaller values of L_0, our method is able to achieve far greater accuracy than the baselines, even in the presence of heavy-tailed noise.

5.3 Synthetic WGAN

We evaluate the solvers on a two-dimensional Wasserstein generative adversarial network with gradient penalty (WGAN-GP) (Gulrajani et al., 2017). The target distribution is a two-dimensional anisotropic Gaussian, $\mathcal{N}([2, -1], \Sigma_\star)$ where $\Sigma_\star$ has eigenvalues $2, 0.2$ and is rotated by 45 degrees. The latent variable is two-dimensional and sampled according to $z \sim \mathcal{N}(0, 1.5^2 I_2)$. The generator G_ϕ and critic D_ψ are chosen as two-hidden-layer MLPs with width 8 and Leaky-ReLU activations, and the critic initialization scale is chosen in $s_D \in \{1, 4\}$, i.e. the weights are sampled as $W_{ij} \sim \mathcal{N}(0, s_D^2/d_{\text{in}})$ with d_{in} the input dimension of each linear layer. The form of the optimization problem we are solving is the following:

$$\min_\phi \max_\psi \; L(\phi, \psi),$$

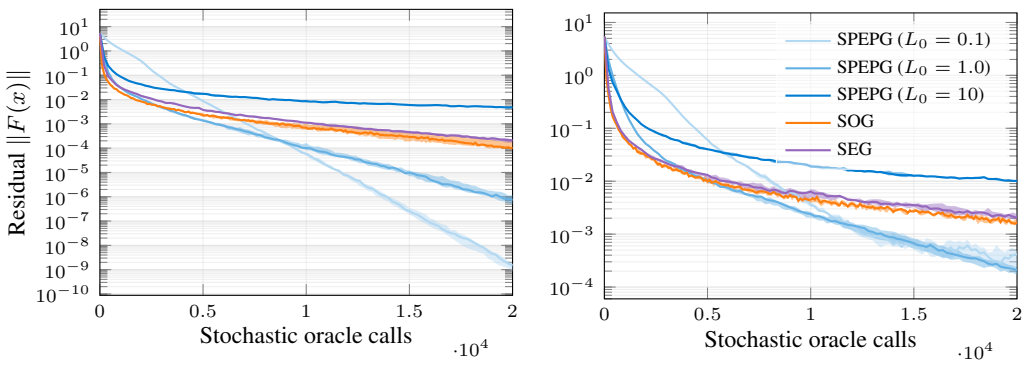

Figure 3: Convergence plots for the example in Section 5.2. The lines and shaded areas represent the median and interquartile range, respectively. (*left*) Finite-sum case; (*right*) Student-t noise.

Table 1: Results on the 2D Gaussian WGAN-GP problem over 10 runs. Lower is better. It can be seen that our proposed method is more stable across runs and more robust towards initialization.

s_D	Method	Last-5 SWD ↓	Mean error ↓	Covariance error ↓
1	SEG	0.3437 ± 0.2719	0.3662	0.6591
	SOG	0.3035 ± 0.1714	0.2772	0.9186
	SPEPG	$\mathbf{0.1673 \pm 0.0599}$	**0.1461**	**0.2865**
4	SEG	0.8033 ± 0.1738	0.9025	4.2742
	SOG	0.7350 ± 0.4496	1.1089	2.5993
	SPEPG	$\mathbf{0.2676 \pm 0.1481}$	**0.2354**	**0.4026**

where

$$L(\phi, \psi) = \mathbb{E}_{x \sim p_{\text{data}}}\left[D_\psi(x)\right] - \mathbb{E}_{z \sim p_z}\left[D_\psi(G_\phi(z))\right] - \lambda_{\text{gp}} \mathbb{E}_{x,z,\varepsilon}\left[\left(\|\nabla_{\hat{x}} D_\psi(\hat{x})\|_2 - 1\right)^2\right] - \frac{\lambda_D}{2}\|\psi\|_2^2 + \frac{\lambda_G}{2}\|\phi\|_2^2.$$

The regularization parameters are set to $\lambda_{\text{gp}} = 5$, $\lambda_D = 10^{-4}$ and $\lambda_G = 10^{-6}$. Moreover, $\hat{x} = \varepsilon x + (1 - \varepsilon)\,\text{sg}(G_\phi(z)), \varepsilon \sim \mathcal{U}[0, 1]$ where the symbol sg denotes stop-gradient, so the gradient-penalty term contributes only to the critic update and not to the generator update.

We set the budget at 8000 batch operator evaluations with batch size 64. The results of this experiment can be found in Table 1. We can see that our proposed method outperforms the rest of the methods in the SWD metric, especially in the case $s_D = 4$ where the difference is more pronounced, while it also achieves lower mean and covariance errors. Both SEG and SOG seem to have high variance between runs, while SPEPG is more stable and more robust to the initialization. The tuning procedure can be found in Appendix C.2 and the chosen hyperparameters in Table 4.

6 Conclusion

In this paper we introduced a stochastic independently normalized past extragradient algorithm for problems beyond Lipschitz continuity, assuming merely the Minty variational inequality. We obtained convergence guarantees in expectation and in probability and we showed that the proposed method converges a.s., under bounded variance assumptions for the stochastic oracle.

There are several open directions for future work. To begin with, analyzing the convergence of the method under more extreme noise assumptions such as heavy-tailed noise is an important research direction that is tightly connected to the normalization-like preconditioning the proposed algorithm performs. Moreover, extending the approach to include coordinate-wise preconditioning for stochastic VI problems is an important research avenue that is still largely underexplored.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs (for example in obtaining better constants), creating or editing software code, implementing methods. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims and generating synthetic data sets, supporting qualitative and thematic data analysis, interpreting results, designing or providing feedback on research methodology or experiments. Assisting with translation, cleaning and reformatting datasets are not applicable to this work. Additionally, we used generative AI tools for suggesting experimental parameters, editing the research paper to improve readability and checking mathematical proofs. We have reviewed all AI-assisted work. We have performed a thorough manual literature survey to ensure proper attribution. We have manually verified all AI-generated output used in this work. None of the contents of the manuscript was generated by AI. Software code was independently reproduced by two authors to ensure consistency of the results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

In the appendix, we provide complete proofs for all the theoretical results in the text, as well as detailed descriptions of the experiments (including selected hyperparameters). A software implementation for all the experiments is included in the supplementary material.

REFERENCES

Ahmet Alacaoglu and Yura Malitsky. Stochastic variance reduction for variational inequality methods. In *Conference on Learning Theory*, pp. 778–816, 2022.

Ahmet Alacaoglu, Donghwan Kim, and Stephen Wright. Revisiting inexact fixed-point iterations for min-max problems: Stochasticity and structured nonconvexity. In *International Conference on Machine Learning*, pp. 840–878, 2024.

Güzin Bayraksan and David K Love. Data-driven stochastic programming using phi-divergences. In *The operations research revolution*, pp. 1–19. Informs, 2015.

Dimitri P Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129(2):163–195, 2011.

Alexander Bodard and Panagiotis Patrinos. Escaping saddle points without Lipschitz smoothness: the power of nonlinear preconditioning. In *Advances in Neural Information Processing Systems*, pp. 124173–124208, 2025.

Axel Bohm, Michael Sedlmayer, Erno Robert Csetnek, and Radu Ioan Bot. Two steps at a time—taking GAN training in stride with Tseng's method. *SIAM Journal on Mathematics of Data Science*, 4(2):750–771, 2022.

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*, 2019.

Sayantan Choudhury and Nicolas Loizou. Extragradient method for (L_0, L_1)-Lipschitz root-finding problems. In *Advances in Neural Information Processing Systems*, pp. 70738–70781, 2025.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 5527–5540, 2020.

Li Deng. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Jelena Diakonikolas, Constantinos Daskalakis, and Michael I Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2746–2754, 2021.

Konstantinos Emmanouilidis, René Vidal, and Nicolas Loizou. Stochastic extragradient with random reshuffling: Improved convergence for variational inequalities. In *International Conference on Artificial Intelligence and Statistics*, pp. 3682–3690, 2024.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

Yifeng Fan, Yongqiang Li, and Bo Chen. Weaker MVI condition: Extragradient methods with multi-step exploration. In *International Conference on Learning Representations*, pp. 9353–9378, 2024.

Matthew Faw, Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Beyond uniform smoothness: A stopped analysis of adaptive SGD. In *Conference on Learning Theory*, pp. 89–160, 2023.

Ofir Gaash, Kfir Y. Levy, and Yair Carmon. Convergence of clipped SGD on convex (L_0, L_1)-smooth functions. In *Advances in Neural Information Processing Systems*, pp. 120410–120442, 2025.

Wilfrid Gangbo and Robert J. McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Eduard Gorbunov, Marina Danilova, David Dobre, Pavel Dvurechenskii, Alexander Gasnikov, and Gauthier Gidel. Clipped stochastic methods for variational inequalities with heavy-tailed noise. In *Advances in Neural Information Processing Systems*, pp. 31319–31332, 2022a.

Eduard Gorbunov, Adrien Taylor, and Gauthier Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. In *Advances in Neural Information Processing Systems*, pp. 21858–21870, 2022b.

Eduard Gorbunov, Abdurakhmon Sadiev, Marina Danilova, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability convergence for composite and distributed stochastic minimization and variational inequalities with heavy-tailed noise. *arXiv preprint arXiv:2310.01860*, 2023.

Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex (L_0, L_1)-smooth optimization: Clipping, acceleration, and adaptivity. In *International Conference on Learning Representations*, pp. 97261–97305, 2025.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, 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*, pp. 16223–16234, 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, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with Stochastic Mirror-Prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

Fivos Kalogiannis and Gabriele Farina. Policy gradient methods converge globally in imperfect-information extensive-form games. In *Advances in Neural Information Processing Systems*, pp. 29162–29224, 2025.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pp. 17343–17363, 2023.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. *arXiv preprint arXiv:2201.12220*, 2022.

Galina M Korpelevich. Extragradient method for finding saddle points and other problems. *Matekon*, 13(4):35–49, 1977.

Emanuel Laude and Panagiotis Patrinos. Anisotropic proximal point algorithm. *arXiv preprint arXiv:2312.09834*, 2023.

Emanuel Laude and Panagiotis Patrinos. Anisotropic proximal gradient. *Mathematical Programming*, 214:801–845, 2025.

Emanuel Laude, Andreas Themelis, and Panagiotis Patrinos. Dualities for non-Euclidean smoothness and strong convexity under the light of generalized conjugacy. *SIAM Journal on Optimization*, 33(4):2721–2749, 2023.

Chris J. Maddison, Daniel Paulin, Yee Whye Teh, and Arnaud Doucet. Dual space preconditioning for gradient descent. *SIAM Journal on Optimization*, 31(1):991–1016, 2021.

Yura Malitsky. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 184(1):383–410, November 2020.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv preprint arXiv:1807.02629*, 2018.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

Jiawang Nie and Xindong Tang. Nash equilibrium problems of polynomials. *Mathematics of Operations Research*, 49(2):1065–1090, 2024.

Konstantinos Oikonomidis, Jan Quan, Emanuel Laude, and Panagiotis Patrinos. Nonlinearly preconditioned gradient methods under generalized smoothness. In *International Conference on Machine Learning*, pp. 47132–47154, 2025a.

Konstantinos Oikonomidis, Jan Quan, and Panagiotis Patrinos. Nonlinearly preconditioned gradient methods: Momentum and stochastic analysis. In *Advances in Neural Information Processing Systems*, pp. 38957–38988, 2025b.

Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, 2016.

Thomas Pethick, Olivier Fercoq, Puya Latafat, Panagiotis Patrinos, and Volkan Cevher. Solving stochastic weak Minty variational inequalities without increasing batch size. *arXiv preprint arXiv:2302.09029*, 2023a.

Thomas Pethick, Puya Latafat, Panagiotis Patrinos, Olivier Fercoq, and Volkan Cevher. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2302.09831*, 2023b.

Thomas Pethick, Ioannis Mavrothalassitis, and Volkan Cevher. Efficient interpolation between extragradient and proximal methods for weak MVIs. In *International Conference on Learning Representations*, pp. 50896–50923, 2025a.

Thomas Pethick, Wanyun Xie, Mete Erdogan, Kimon Antonakopoulos, Antonio Silveti-Falls, and Volkan Cevher. Generalized gradient norm clipping & non-Euclidean (L_0, L_1)-smoothness. In *Advances in Neural Information Processing Systems*, pp. 21170–21208, 2025b.

Leonid Denisovich Popov. A modification of the Arrow–Hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making Muon & Scion great again! (Bridging theory and practice of LMO-based optimizers for LLMs). *arXiv preprint arXiv:2505.13416*, 2025.

Daniil Vankov, Angelia Nedich, and Lalitha Sankar. Generalized smooth variational inequalities: Methods with adaptive stepsizes. In *International Conference on Machine Learning*, pp. 49137–49170, 2024.

Daniil Vankov, Angelia Nedich, and Lalitha Sankar. Generalized smooth stochastic variational inequalities: Almost sure convergence and convergence rates. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856.

Daniil Vankov, Anton Rodomanov, Angelia Nedich, Lalitha Sankar, and Sebastian Stich. Optimizing (L_0, L_1)-smooth functions by gradient methods. In *International Conference on Learning Representations*, pp. 15953–15979, 2025b.

Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of Adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *Conference on Learning Theory*, pp. 161–190, 2023.

David Williams. *Probability with martingales*. Cambridge University Press, 1991.

Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. In *Advances in Neural Information Processing Systems*, pp. 15511–15521, 2020.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

A PRECONDITIONED LIPSCHITZ CONTINUITY

A.1 PRECONDITIONED LIPSCHITZ CONTINUITY AND (L_0, L_1)-LIPSCHITZ CONTINUITY

The regularity condition we work with is related to (L_0, L_1)-Lipschitz continuity but is in fact a less restrictive condition. In this section we explore the differences between the two notions and provide examples of operators that satisfy Definition 3.1. To begin with, we state the definition of (L_0, L_1)-Lipschitz continuity as adapted from Vankov et al. (2024, Assumption 2.1) (with $\alpha = 1$), see also Choudhury & Loizou (2025, Assumption 1.1). We focus on the case $L_0 > 0$.

Definition A.1. We say that a continuous $F : \mathbb{R}^n \to \mathbb{R}^n$ is (L_0, L_1)-Lipschitz if for some $L_0 > 0, L_1 \geq 0$, the following inequality holds:

$$\|F(x) - F(\bar{x})\| \leq \Big(L_0 + L_1 \max_{\theta\in[0,1]} \|F(\theta x + (1-\theta)\bar{x})\|\Big)\|x - \bar{x}\| \qquad \forall x, \bar{x} \in \mathbb{R}^n. \tag{8}$$

On the other hand, the preconditioned Lipschitz condition we consider in Definition 3.1 imposes 1-Lipschitz continuity of the operator $T(x) = \frac{F(x)}{L_0 + L_1\|F(x)\|}$, that is,

$$\left\| \frac{F(x)}{L_0 + L_1\|F(x)\|} - \frac{F(\bar{x})}{L_0 + L_1\|F(\bar{x})\|} \right\| \leq \|x - \bar{x}\| \qquad \forall x, \bar{x} \in \mathbb{R}^n.$$

We next show that every (L_0, L_1)-Lipschitz operator is also preconditioned Lipschitz by standard analytical arguments. To this end, we first show a preliminary result that will be used in the proof.

Lemma A.2. *Let $y, \bar{y} \in \mathbb{R}^n$, $L_0 > 0$, $L_1 \geq 0$. Then,*

$$\left\| \frac{y}{L_0 + L_1\|y\|} - \frac{\bar{y}}{L_0 + L_1\|\bar{y}\|} \right\| \leq \frac{\|y - \bar{y}\|}{\sqrt{(L_0 + L_1\|y\|)(L_0 + L_1\|\bar{y}\|)}}.$$

Proof. Denote $Y = L_0 + L_1\|y\|$ and $\bar{Y} = L_0 + L_1\|\bar{y}\|$. It suffices to show that $\|\bar{Y}y - Y\bar{y}\|^2 \leq Y\bar{Y}\|y - \bar{y}\|^2$. Indeed, we have

$$\begin{aligned} Y\bar{Y}\|y - \bar{y}\|^2 - \|\bar{Y}y - Y\bar{y}\|^2 &= (Y - \bar{Y})(\bar{Y}\|y\|^2 - Y\|\bar{y}\|^2) \\ &= L_1(\|y\| - \|\bar{y}\|)^2(L_0(\|y\| + \|\bar{y}\|) + L_1\|y\|\|\bar{y}\|) \geq 0, \end{aligned}$$

hence $\|\bar{Y}y - Y\bar{y}\| \leq \sqrt{Y\bar{Y}}\|y - \bar{y}\|$, which yields the desired result after dividing by $Y\bar{Y} \neq 0$. □

We now move on to proving the claim.

Proposition A.3. *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ satisfy Definition A.1. Then F satisfies Definition 3.1.*

Proof. Let $x, \bar{x} \in \mathbb{R}^n$ and denote $\gamma(t) = (1-t)x + t\bar{x}$ for $t \in [0, 1]$. Since F is continuous, $L_0 + L_1\|F(\gamma(t))\|$ is uniformly continuous on $[0, 1]$. Now let $\varepsilon > 0$ be arbitrary, and choose N sufficiently large, so that on each interval $[i/N, (i+1)/N]$ for $i = 0, \ldots, N-1$, we have

$$\begin{aligned} \max_{t\in[i/N,(i+1)/N]} (L_0 + L_1\|F(\gamma(t))\|) &\leq \min_{t\in[i/N,(i+1)/N]} (L_0 + L_1\|F(\gamma(t))\|) + \varepsilon L_0 \\ &\leq (1+\varepsilon) \min_{t\in[i/N,(i+1)/N]} (L_0 + L_1\|F(\gamma(t))\|). \end{aligned}$$

Let $x_i = \gamma(i/N)$ denote the endpoints of the intervals. Then, by Lemma A.2, we have

$$\begin{aligned} \|T(x_{i+1}) - T(x_i)\| &\leq \frac{\|F(x_{i+1}) - F(x_i)\|}{\sqrt{(L_0 + L_1\|F(x_{i+1})\|)(L_0 + L_1\|F(x_i)\|)}} \\ &\leq \frac{\max_{t\in[i/N,(i+1)/N]}(L_0 + L_1\|F(\gamma(t))\|)}{\sqrt{(L_0 + L_1\|F(x_{i+1})\|)(L_0 + L_1\|F(x_i)\|)}}\|x_{i+1} - x_i\| \\ &\leq \frac{\max_{t\in[i/N,(i+1)/N]}(L_0 + L_1\|F(\gamma(t))\|)}{\min_{t\in[i/N,(i+1)/N]}(L_0 + L_1\|F(\gamma(t))\|)}\|x_{i+1} - x_i\| \leq \frac{1+\varepsilon}{N}\|x - \bar{x}\| \end{aligned}$$

where we used Definition A.1 in the second inequality.

Summing over $i = 0, 1, \ldots, N-1$ and using the triangle inequality yields

$$\|T(x) - T(\bar{x})\| \leq \sum_{i=0}^{N-1} \|T(x_{i+1}) - T(x_i)\| \leq (1+\varepsilon)\|x - \bar{x}\|.$$

As $\varepsilon > 0$ is arbitrary, the desired result follows. □

Having established Proposition A.3 the claim about the class of preconditioned Lipschitz operators being larger follows from Bodard & Patrinos (2025, Theorem A.3), see also Examples 3.2 and A.10.

A.2 Examples

In this section we present examples of operators that satisfy Definition 3.1 but need not be globally Lipschitz continuous. A list of examples can also be found in Choudhury & Loizou (2025, Appendix C). We start with a minimax problem where the interaction between the two players is higher-order nonlinear.

Example A.4. Let $\lambda \in \mathbb{R}, x \in \mathbb{R}^n, y \in \mathbb{R}^m$ and $B \in \mathbb{R}^{n\times m}$. Consider $\mathcal{L}(x,y) = \frac{1}{4}\|x\|^4 - \frac{1}{4}\|y\|^4 + \frac{\lambda}{2}\|x\|^2\|y\|^2 + \langle x, By\rangle$ which leads to the saddle operator:

$$F(x,y) = \begin{pmatrix} (\|x\|^2 + \lambda\|y\|^2)x + By \\ (\|y\|^2 - \lambda\|x\|^2)y - B^\top x \end{pmatrix}.$$

F satisfies Definition 3.1 with constants $L_0 = \|B\| + 3 + 2|\lambda|, L_1 = 6 + 4|\lambda|$.

Proof. Consider $(x,y) \neq (0,0)$. To begin with we lower bound $\|F(x,y)\|$. From Cauchy–Schwarz,

$$\|F(x,y)\| \geq \langle F(x,y), (x,y)\rangle / \|(x,y)\| = (\|x\|^4 + \|y\|^4)/\|(x,y)\| \geq (\|x\|^2 + \|y\|^2)^{3/2}/2$$

where we used (22) with $a = \|x\|^2, b = \|y\|^2$ and $e = 1$ in the last inequality. The inequality above holds also at the origin by inspection. At the same time, for the Jacobian of F we have:

$$\nabla F(x,y) = \begin{pmatrix} (\|x\|^2 + \lambda\|y\|^2)I + 2xx^\top & 2\lambda xy^\top + B \\ -2\lambda yx^\top - B^\top & (\|y\|^2 - \lambda\|x\|^2)I + 2yy^\top \end{pmatrix}.$$

We may decompose this matrix as

$$\nabla F(x,y) = \begin{pmatrix} 0 & B \\ -B^\top & 0 \end{pmatrix} + \begin{pmatrix} \|x\|^2 I + 2xx^\top & 0 \\ 0 & \|y\|^2 I + 2yy^\top \end{pmatrix} + \begin{pmatrix} \lambda\|y\|^2 I & 2\lambda xy^\top \\ -2\lambda yx^\top & -\lambda\|x\|^2 I \end{pmatrix}$$

where the spectral norms of the components can be shown to be bounded by $\|B\|$, $3(\|x\|^2 + \|y\|^2)$ and $2|\lambda|(\|x\|^2 + \|y\|^2)$ respectively. Therefore,

$$\|\nabla F(x,y)\| \leq \|B\| + (3 + 2|\lambda|)(\|x\|^2 + \|y\|^2) \leq (\|B\| + 3 + 2|\lambda|) + (6 + 4|\lambda|)\|F(x,y)\|$$

where we also used the standard bound $t^2 \leq 1 + t^3$ for $t = (\|x\|^2 + \|y\|^2)^{1/2} \geq 0$. The result follows from (2) and Proposition A.3. □

The next example goes beyond polynomial growth of the operator, involving the hyperbolic sine.

Example A.5. Consider

$$F(x) = B^\top \sinh(Bx - b),$$

where $B \in \mathbb{R}^{m\times n}$ has full row rank, $b \in \mathbb{R}^m, x \in \mathbb{R}^n$ and sinh is considered componentwise. This operator is the gradient of $f(x) = \sum_{i=1}^m \cosh((Bx-b)_i)$, so F is monotone as the gradient of a convex function. Then F satisfies Definition 3.1 with constants $L_0 = \|B\|^2$ and $L_1 = \frac{\|B\|^2}{\sigma_{\min}(B)}$, where $\sigma_{\min}(B) \coloneqq \min_{u\in\mathbb{R}^m, \|u\|=1} \|B^\top u\| > 0$.

Proof. We have $\nabla F(x) = B^\top \operatorname{diag}(\cosh(Bx - b))B$. Using submultiplicativity of the spectral norm, we get

$$\|\nabla F(x)\| \leq \|B\|^2 \|\operatorname{diag}(\cosh(Bx-b))\| \leq \|B\|^2(1 + \|\sinh(Bx-b)\|)$$

where we also used $\cosh(t) = \sqrt{1+\sinh^2(t)} \leq 1+|\sinh(t)|$. By definition of $\sigma_{\min}(B)$, $\|F(x)\| \geq \sigma_{\min}(B)\|\sinh(Bx-b)\|$ and therefore

$$\|\nabla F(x)\| \leq \|B\|^2(1+\|\sinh(Bx-b)\|) \leq \|B\|^2 + \frac{\|B\|^2}{\sigma_{\min}(B)}\|F(x)\|,$$

leading to the claimed result in light of (2) and Proposition A.3. □

Another setting where these problems arise naturally is in distributionally robust optimization (DRO). In particular, for a DRO problem with a χ^2-ambiguity set (Namkoong & Duchi, 2016; Bayraksan & Love, 2015), replacing the ambiguity set constraint with a soft penalty leads to a problem of the following form.

Example A.6 (Higher-order distributionally robust optimization)**.** Consider the following distributionally robust optimization problem

$$\min_{x\in\mathbb{R}^n}\max_{y\in\Delta^m} \mathcal{L}(x,y) = \sum_{i=1}^m y_i \frac{(\langle a_i, x\rangle - b_i)^4}{4} - \frac{\mu}{2}\|y\|^2,$$

where $b \in \mathbb{R}^m$, $a_i^\top$ are the rows of the data matrix $A \in \mathbb{R}^{m\times n}$, Δ^m is the probability simplex and $\mu > 0$ is a regularization parameter. The corresponding saddle operator is

$$F(x,y) = \begin{pmatrix} \sum_{i=1}^m y_i(\langle a_i, x\rangle - b_i)^3 a_i \\ -l(x) + \mu y \end{pmatrix},$$

where $l : \mathbb{R}^n \to \mathbb{R}^m$ is given by $l(x) = (\frac{(\langle a_i,x\rangle - b_i)^4}{4})_{i=1}^m$. Then, F satisfies Definition 3.1 on $\mathbb{R}^n \times \Delta^m$ with $L_0 = \mu + 6(\max_i \|a_i\|)^2(2+\mu^{3/4}) + 4^{3/4}(\sum_{i=1}^m \|a_i\|^8)^{1/8}(1+\mu^{3/4})$ and $L_1 = 6(\max_i \|a_i\|)^2 + 4^{3/4}(\sum_{i=1}^m \|a_i\|^8)^{1/8}$.

Proof. Denote $r(x) \coloneqq (\langle a_i, x\rangle - b_i)_{i=1}^m$. Then,

$$\nabla F(x,y) = \begin{pmatrix} H(x,y) & B(x) \\ -B(x)^\top & \mu I_m \end{pmatrix}$$

with $H(x,y) = 3\sum_{i=1}^m y_i r_i(x)^2 a_i a_i^\top$ and $B(x) = [r_1(x)^3 a_1 \dots r_m(x)^3 a_m]$. We now proceed by bounding the operator norm of each submatrix in $\nabla F(x,y)$. Since $y \in \Delta^m$, it follows that $\|H(x,y)\| \leq 3\big(\max_i \|a_i\|\big)^2 \|r(x)\|_\infty^2$. Moreover, for $v \in \mathbb{R}^n$ with $\|v\| = 1$, we have

$$\|B(x)^\top v\|^2 = \sum_{i=1}^m |r_i(x)|^6 |\langle a_i, v\rangle|^2 \leq \left(\sum_{i=1}^m |r_i(x)|^8\right)^{3/4} \left(\sum_{i=1}^m |\langle a_i, v\rangle|^8\right)^{1/4}$$

where we used Hölder's inequality. Note that $\|l(x)\|^2 = \frac{1}{16}\sum_{i=1}^m |r_i(x)|^8$, hence taking square roots and then the supremum over $\|v\| = 1$, we obtain

$$\|B(x)\| \leq 4^{3/4}\left(\sum_{i=1}^m \|a_i\|^8\right)^{1/8} \|l(x)\|^{3/4}.$$

From the triangle inequality on the decomposition of ∇F we have

$$\|\nabla F(x,y)\| \leq \max\{\|H(x,y)\|, \mu\} + \|B(x)\|.$$

We now aim to associate the obtained bounds with $\|F(x,y)\|$. Note that $l(x) = \mu y - F_y(x,y)$ and thus $\|l(x)\| \leq \mu + \|F(x,y)\|$ since $\|y\| \leq 1$ due to the simplex constraint. Moreover, $|r_i(x)|^4 \leq 4(\mu + \|F(x,y)\|)$ for all $i \in \{1,\dots,m\}$ by the same reasoning, leading to $\|r(x)\|_\infty^2 \leq 2\sqrt{\|F(x,y)\| + \mu}$. We are thus led to

$$\|\nabla F(x,y)\| \leq \mu + 6\big(\max_i \|a_i\|\big)^2\sqrt{\|F(x,y)\| + \mu} + 4^{3/4}\left(\sum_{i=1}^m \|a_i\|^8\right)^{1/8} (\|F(x,y)\| + \mu)^{3/4}.$$

Using the standard bounds $t^{1/2} \le 1 + t^{3/4}$ and $(c+d)^{3/4} \le c^{3/4} + d^{3/4}$ for nonnegative c, d we get

$$\|\nabla F(x,y)\| \le \mu + 6\big(\max_i \|a_i\|\big)^2(1+\mu^{3/4}) + 4^{3/4}\left(\sum_{i=1}^m \|a_i\|^8\right)^{1/8}\mu^{3/4} + \left[6(\max_i \|a_i\|)^2 + 4^{3/4}\left(\sum_{i=1}^m \|a_i\|^8\right)^{1/8}\right]\|F(x,y)\|^{3/4}.$$

The claimed result now follows from $t^{3/4} \le 1 + t$. □

In what follows, we focus mostly on multiplayer games. We denote by p the number of players, where each player i controls $x_i \in \mathbb{R}^{n_i}$, and concatenate the actions in the variable $x = (x_1, x_2, \dots, x_p) \in \mathbb{R}^n$ with $n = \sum_{i=1}^p n_i$. Each player has cost $f_i(x_i, x_{-i})$ where x_{-i} encodes the actions of the rest of the players. The game mapping is given by

$$F(x) = \begin{pmatrix} \nabla_{x_1} f_1(x_1, x_{-1}) \\ \nabla_{x_2} f_2(x_2, x_{-2}) \\ \vdots \\ \nabla_{x_p} f_p(x_p, x_{-p}) \end{pmatrix}$$

and in the unconstrained setting under additional convexity assumptions, the first-order equilibrium conditions are $F(x^\star) = 0$.

Example A.7 (Three-player polynomial game)**.** Consider Nie & Tang (2024, Example 5.8) where a polynomial game with three players is described and assume that all $x_i \in \mathbb{R}^d$ such that $n = 3d$ for simplicity. Each player has cost

$$f_i(x) = \sum_{j=1}^d x_{i,j}^4 + \frac{1}{d^2} \sum_{0 \le m \le j \le k \le d} x_{i,m} x_{i,j} (x_{i,k} + x_{i+1,m} + x_{i+2,j}),$$

where $x_{i,0} = 1$ and we interpret player indices modulo three. Note that for $1 \le l \le d$, $F_{i,l}(x) = 4x_{i,l}^3 + p_{i,l}(x)$ where $p_{i,l}$ is a polynomial of degree at most 2. In this case, for any given $L_1 > 0$, there exists a finite $L_0 > 0$ such that F satisfies Definition 3.1.

Proof. We use the proof technique of Bodard & Patrinos (2025) and aim to bound the Jacobian by polynomials of degree at most two and show that $\|F(x)\|$ has cubic growth up to lower-order terms. Let H_3 be the leading term in F, i.e. $H_3(x) = 4x^{[3]}$, where $x^{[3]}$ denotes componentwise cubing. We have

$$\|H_3(x)\| = 4\|x\|_6^3 \ge \frac{4}{n}\|x\|^3.$$

Moreover, $\nabla H_3(x) = 12\,\mathrm{diag}(x_j^2)_{j=1}^n$ so that $\|\nabla H_3(x)\| \le 12\|x\|^2$. Regarding the remainder $P_2(x) \coloneqq F(x) - H_3(x)$, we note that it is a polynomial of degree at most 2 and thus $\|P_2(x)\| \le C_0 + C_1\|x\| + C_2\|x\|^2$ and $\|\nabla P_2(x)\| \le C_1 + C_2\|x\|$ for some $C_0, C_1, C_2 \ge 0$. Therefore, $\|\nabla F(x)\| \le 12\|x\|^2 + C_2\|x\| + C_1$ by the triangle inequality. At the same time, the reverse triangle inequality ensures that $\|F(x)\| \ge \frac{4}{n}\|x\|^3 - C_2\|x\|^2 - C_1\|x\| - C_0$. For fixed $L_1 > 0$, we are led to

$$\|\nabla F(x)\| - L_1\|F(x)\| \le -\frac{4L_1}{n}\|x\|^3 + (12 + L_1C_2)\|x\|^2 + (C_2 + L_1C_1)\|x\| + C_1 + L_1C_0.$$

For $\|x\| \to \infty$, the r.h.s. tends to $-\infty$ and thus admits a finite supremum. Choosing L_0 larger than this supremum concludes the proof. □

Example A.8 (A restriction of the Nash–Cournot game)**.** Consider the Nash–Cournot oligopolistic equilibrium model as described in Malitsky (2020). We explicitly restrict the nonnegative orthant constraints to the form $C = \prod_{i=1}^p [l_i, +\infty)$ with $l_i > 0$ so that the operator F is well-defined. We consider the equilibrium operator $F(x) = (F_i(x))_{i=1}^p$ where each coordinate is given by

$$F_i(x) = f_i'(x_i) - \pi\left(\sum_{j=1}^p x_j\right) - x_i\pi'\left(\sum_{j=1}^p x_j\right), \qquad i = 1, \dots, p.$$

The inverse demand function π and each firm's cost function f_i are given by

$$\pi(q) = a^\rho q^{-\rho} \quad \text{and} \quad f_i(x_i) = c_i x_i + \frac{d_i}{\nu_i + 1} x_i^{\nu_i + 1},$$

where $a, \nu_i, d_i, \rho > 0$ and $c_i \geq 0$. In this setting, F satisfies Definition 3.1 on C, with constants $L_0 = \rho a^\rho (\sum_{i=1}^p l_i)^{-\rho-1}[p + 1 + (\rho+1)\sqrt{p}] + a^\rho (\max_i \nu_i/l_i)(\sum_{i=1}^p l_i)^{-\rho}$ and $L_1 = \max_i \nu_i/l_i$.

Proof. We work in the constraint set C where all the functions involved are differentiable. We proceed in a similar manner to the previous proofs by bounding the higher-order terms in the Jacobian in terms of $\|F(x)\|$. To begin with, denoting $q = \sum_{j=1}^p x_j$, we have $F_i(x) = c_i + d_i x_i^{\nu_i} - a^\rho q^{-\rho} + \rho a^\rho x_i q^{-\rho-1}$. Therefore, we can write $F = c + D + H$ where $D_i(x) = d_i x_i^{\nu_i}$ and $H_i(x) = -a^\rho q^{-\rho}(1 - \rho x_i/q)$.

Since $x_i \geq l_i$, we have

$$H_i(x) \geq -a^\rho \left(\sum_{j=1}^p l_j \right)^{-\rho}, \qquad D_i(x) \leq |F_i(x)| + a^\rho \left(\sum_{j=1}^p l_j \right)^{-\rho} \leq \|F(x)\| + a^\rho \left(\sum_{j=1}^p l_j \right)^{-\rho},$$

where we also used $c_i \geq 0$. It follows that

$$\|\nabla D(x)\| = \max_i \frac{\nu_i}{x_i} D_i(x) \leq \left(\max_i \frac{\nu_i}{l_i} \right) \left(\|F(x)\| + a^\rho \left(\sum_{j=1}^p l_j \right)^{-\rho} \right).$$

Moreover,

$$\|\nabla H(x)\| \leq \rho a^\rho q^{-\rho-1} \left[\|I + \mathbf{1}\mathbf{1}^\top\| + (\rho+1)\frac{\|x\|}{q}\|\mathbf{1}\| \right] \leq \rho a^\rho \left(\sum_{j=1}^p l_j \right)^{-\rho-1} [p+1+(\rho+1)\sqrt{p}],$$

where $\mathbf{1}$ is the vector of all ones and I denotes the identity matrix. Here, we used that $\|I + \mathbf{1}\mathbf{1}^\top\| = p + 1$, $\|x\| \leq q$, $\|\mathbf{1}\| = \sqrt{p}$ and $q \geq \sum_{j=1}^p l_j$. Therefore, $\|\nabla F(x)\| \leq \|\nabla D(x)\| + \|\nabla H(x)\| \leq L_0 + L_1 \|F(x)\|$ with the stated L_0, L_1 yielding the claimed result. □

Finally, we move on to provide another example that satisfies Definition 3.1 but fails to be (L_0, L_1)-Lipschitz. To this end, we first prove the following helper result.

Lemma A.9. *Let $J \in \mathbb{R}^{n\times n}$ satisfy $J^\top J = I$, and let $g : [0,\infty) \to [0,\infty)$ satisfy $g(0) = 0$ and $g \in \mathcal{C}^1((0,\infty))$. Define $F : \mathbb{R}^n \to \mathbb{R}^n$ by $F(x) = g(\|x\|)\frac{Jx}{\|x\|}$ for $x \neq 0$ and 0 otherwise. Let $L_0 > 0$, $L_1 \geq 0$, and define $h(r) = \frac{g(r)}{L_0 + L_1 g(r)}$ for $r > 0$. If $|h'(r)| \leq 1$ and $\frac{|h(r)|}{r} \leq 1$ for all $r > 0$, then F satisfies Definition 3.1 with constants L_0, L_1.*

Proof. For $x \neq 0$, let $r = \|x\|$. Since $J^\top J = I$, $\|F(x)\| = g(r)$, and hence

$$T(x) = h(r)\frac{Jx}{r}, \qquad \nabla T(x) = J\left[h'(r)\frac{xx^\top}{r^2} + \frac{h(r)}{r}\left(I - \frac{xx^\top}{r^2} \right) \right].$$

The two matrices above are orthogonal projections and $J^\top J = I$, so we have that $\|\nabla T(x)\| \leq \max\left\{ |h'(r)|, \frac{|h(r)|}{r} \right\} \leq 1$. Moreover, $\|T(x)\| = |h(\|x\|)| \leq \|x\|$, so T is continuous at the origin. It follows that T is 1-Lipschitz continuous, from which the desired result follows. □

Example A.10. Let $J \in \mathbb{R}^{n\times n}$ satisfy $J^\top J = I$, and define the super-exponential operator $F : \mathbb{R}^n \to \mathbb{R}^n$ by

$$F(x) = \frac{\exp(\|x\|^2) - 1}{\|x\|} Jx, \qquad x \neq 0$$

and $F(0) = 0$. Then F satisfies Definition 3.1 with constants $L_0 = L_1 = 1$. Moreover, F is not (L_0, L_1)-Lipschitz for any $L_0 > 0$ and $L_1 \geq 0$.

Proof. Let $g(r) = e^{r^2} - 1$. For $L_0 = L_1 = 1$, $h(r) = \frac{g(r)}{1+g(r)} = 1 - e^{-r^2}$, and, using $1 - e^{-r^2} \leq \min\{r^2, 1\}$, we have $\frac{h(r)}{r} \leq \min\{r, 1/r\} \leq 1$ and $|h'(r)| = 2re^{-r^2} \leq \sqrt{\frac{2}{e}} < 1$. Hence, the first claim follows from Lemma A.9.

For the second claim, suppose that F is (L_0, L_1)-Lipschitz for some $L_0 > 0$ and $L_1 \geq 0$. Fix a unit vector u and apply Definition A.1 to $x = (r + t)u$ and $\bar{x} = ru$. Since g is increasing, $\frac{g(r+t)-g(r)}{t} \leq L_0 + L_1 g(r + t)$. Letting $t \downarrow 0$ and using $g(r) = e^{r^2} - 1$ gives

$$2re^{r^2} \leq L_0 + L_1(e^{r^2} - 1) \implies 2r \leq L_0 e^{-r^2} + L_1(1 - e^{-r^2}) \leq L_0 + L_1$$

which is impossible as $r \to \infty$. Hence F is not (L_0, L_1)-Lipschitz for any $L_0 > 0$ and $L_1 \geq 0$. □

B Analysis of stochastic past extra preconditioned gradient

Throughout this appendix, we analyze the past-extragradient method described in Algorithm 1 under Assumptions 4.1 and 4.2 and the parameter choices in Assumption 4.4. The main difficulty is that the conditional mean of the oracle is a positive, iteration-dependent rescaling of T, cf. Lemma 4.3, and the momentum estimator averages these oracle evaluations at past points. To address this challenge, we introduce an expected mean tracker and bound the momentum tracking error by an averaged noise term and a history-dependent term in Proposition B.5. Propositions B.3 and B.4 establish a one-step descent estimate for the distance to the solution, involving this momentum tracking error. Propositions B.6 to B.8 then control the noise term and the history-dependent term, leading to the total Lyapunov recursion in Proposition B.9. Using the resulting Lyapunov recursion, we then move on to prove the results stated in the main text.

B.1 Setup

For convenience, we repeat the algorithm here and establish some basic facts:

$$m^k = (1 - \alpha_k)m^{k-1} + \alpha_k \widehat{T}(z^k; \xi^k, \zeta^k) \tag{9}$$

$$z^{k+1} = P_C(x^k - \gamma m^k) \tag{10}$$

$$x^{k+1} = P_C(x^k - \underbrace{\theta\alpha_{k+1}}_{=:\omega_k} \widehat{T}(z^{k+1}; \xi^{k+1}, \zeta^{k+1})) \tag{11}$$

with $x^0 \in C$ deterministic, $z^0 = x^0$, $m^{-1} = 0$, $\alpha_0 = 1$. We define the filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$ corresponding to our sampling scheme:

$$\mathcal{F}_{-1} \coloneqq \sigma(\emptyset), \qquad \mathcal{F}_k = \mathcal{F}_{k-1} \vee \sigma(\xi^k, \zeta^k), \qquad k \in \mathbb{N}.$$

It is clear that x^k, m^k, z^{k+1} are $\mathcal{F}_k$-measurable. For ease of notation, we also define the error between the estimator $\widehat{T}$ and its conditional mean at z^k, see Lemma 4.3

$$\varepsilon^k \coloneqq \widehat{T}(z^k; \xi^k, \zeta^k) - c(z^k)T(z^k). \tag{12}$$

Note that since z^{k+1} is $\mathcal{F}_k$-measurable and ξ^{k+1}, ζ^{k+1} are independent of $\mathcal{F}_k$, by standard conditioning under independence (Williams, 1991, §9.10), Lemma 4.3 gives

$$\mathbb{E}[\varepsilon^k \mid \mathcal{F}_{k-1}] = 0, \qquad \mathbb{E}[\|\varepsilon^k\|^2 \mid \mathcal{F}_{k-1}] \leq \frac{2\sigma^2}{L_0^2}, \qquad k \in \mathbb{N}. \tag{13}$$

In the predictor step, we use Polyak momentum to reduce variance and as such we will have to bound the difference between $T(x^k)$ and its momentum approximation m^k. Since the conditional mean of $\widehat{T}(z^k; \xi^k, \zeta^k)$ is not $T(z^k)$ but rather $c(z^k)T(z^k)$ in light of Lemma 4.3, we build the expected mean tracker $\lambda_k T(x^k)$ where

$$\lambda_0 \coloneqq c(x^0) = c(z^0), \qquad \lambda_{k+1} \coloneqq (1 - \alpha_{k+1})\lambda_k + \alpha_{k+1}c(z^{k+1}), \qquad k \in \mathbb{N} \tag{14}$$

and define the momentum tracking error

$$e^k \coloneqq m^k - \lambda_k T(x^k). \tag{15}$$

For our analysis we will also require a scaled version of the subgradient that was chosen when computing z^{k+1} in (10):

$$n^{-1} := 0, \qquad n^k := \frac{x^k - \gamma m^k - z^{k+1}}{\gamma \lambda_k} \in N_C(z^{k+1}), \qquad k \in \mathbb{N}. \tag{16}$$

The scaling is chosen to ensure cancellation in the proof of Proposition B.4. Moreover, proceeding in the standard way of stochastic (past) extragradient methods leads to very large constants and suboptimal bounds for the stepsizes. For that reason we try to utilize (13) as much as possible in order to cancel out cross-products, which leads to the following quantity showing up:

$$r^k := x^k - z^{k+1} - \omega_k c(z^{k+1})(T(z^{k+1}) + n^k). \tag{17}$$

Before providing the full proof, we record some trivial inequalities due to (14), (6) and the parameter choices in Assumption 4.4.

Lemma B.1. *In the setting of Theorem 4.5, we have for every $k \geq 0$ that λ_k is $\mathcal{F}_{k-1}$-measurable. Moreover, almost surely,*

$$\frac{1}{1+\delta} \leq c(z^{k+1}) \leq 1 + \delta, \qquad \frac{1}{1+\delta} \leq \lambda_k \leq 1 + \delta, \tag{18}$$

$$\gamma \leq \gamma(1+\delta) \leq \frac{1}{4}, \qquad \theta(1+\delta) \leq \theta(1+\delta)^2 \leq \frac{\gamma}{3} \leq \frac{1}{12} \tag{19}$$

$$(1-\alpha_k)\left(1 + \frac{3\alpha_k}{4} + \frac{\alpha_k^2}{2}\right) \leq 1 - \frac{\alpha_k}{4} - \frac{\alpha_k^2}{4}, \qquad (1-\alpha_k)\left(1 + \frac{\alpha_k}{3}\right) \leq 1. \tag{20}$$

B.2 Useful results

We will make use of several standard results which we state here for convenience. For $a, b, d \in \mathbb{R}^n$ and $e > 0$, we have

$$2\langle a, b\rangle \leq e\|a\|^2 + \frac{1}{e}\|b\|^2, \tag{21}$$

$$\|a+b\|^2 \leq (1+e)\|a\|^2 + (1+\tfrac{1}{e})\|b\|^2, \tag{22}$$

$$2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2, \tag{23}$$

$$-\|a\|^2 - 2\langle d, a+b\rangle = -\|a+d\|^2 - \|b\|^2 + \|b-d\|^2. \tag{24}$$

Let $C \subseteq \mathbb{R}^n$ be a closed, convex, nonempty set and $y \in \mathbb{R}^n$. Then, for $P_C(y) := \arg\min_{x \in C} \|x - y\|^2$ the (necessarily unique) projection of y onto C, the following standard inequality holds:

$$\|P_C(y) - x\|^2 \leq \|y - x\|^2 - \|y - P_C(y)\|^2 \qquad \forall x \in C. \tag{25}$$

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For any two events $A, B \in \mathcal{F}$,

$$\mathbb{P}(A) \leq \mathbb{P}(B^c) + \mathbb{P}(A \cap B). \tag{26}$$

B.3 Oracle bounds

We start by bounding the statistical error induced by the stochastic estimator $\widehat{T}$.

B.4 Proof of Lemma 4.3

Proof. By definition of $\widehat{T}$ in (5) and the hypotheses:

$$\mathbb{E}[\widehat{T}(x;\xi,\zeta)] = F(x)\mathbb{E}\left[\frac{1}{L_0 + L_1\|G(x,\zeta)\|}\right] = T(x)\mathbb{E}\left[\frac{L_0 + L_1\|F(x)\|}{L_0 + L_1\|G(x,\zeta)\|}\right] = c(x)T(x).$$

To recover the bounds on $c(x)$ in (6), note that the triangle inequality ensures $\|F(x)\| \leq \|G(x,\zeta)\| + \|F(x) - G(x,\zeta)\|$. Therefore,

$$c(x) = \mathbb{E}\left[\frac{L_0 + L_1\|F(x)\|}{L_0 + L_1\|G(x,\zeta)\|}\right] \leq \mathbb{E}\left[\frac{L_0 + L_1\|G(x,\zeta)\| + L_1\|F(x) - G(x,\zeta)\|}{L_0 + L_1\|G(x,\zeta)\|}\right]$$
$$\leq 1 + \frac{L_1}{L_0}\mathbb{E}[\|F(x) - G(x,\zeta)\|] \leq 1 + \frac{L_1}{L_0}\sqrt{\mathbb{E}[\|F(x) - G(x,\zeta)\|^2]} \leq 1 + \delta,$$

where we used $L_0 \leq L_0 + L_1\|G(x,\zeta)\|$ in the second inequality, Jensen's inequality in the third inequality and Assumption 4.1 in the last inequality. Regarding the lower bound on $c(x)$, it follows from the triangle inequality that $L_0 + L_1\|G(x,\zeta)\| \leq L_0 + L_1\|F(x)\| + L_1\|G(x,\zeta) - F(x)\|$, which leads to

$$\mathbb{E}[L_0 + L_1\|G(x,\zeta)\|] \leq L_0 + L_1\|F(x)\| + L_1\sigma$$

by a similar argument. Applying Jensen's inequality for the convex function $t \mapsto 1/t$ on $(0, +\infty)$ then yields

$$c(x) \geq \frac{L_0 + L_1\|F(x)\|}{\mathbb{E}[L_0 + L_1\|G(x,\zeta)\|]} \geq \frac{1}{1+\delta},$$

concluding the proof for the bounds in (6).

We now move on to proving the bounded variance of $\widehat{T}$ in (7). By the least-squares property of the mean and (6), we have

$$\mathbb{E}[\|\widehat{T}(x;\xi,\zeta) - c(x)T(x)\|^2] \leq \mathbb{E}[\|\widehat{T}(x;\xi,\zeta) - T(x)\|^2],$$

so it suffices to upper bound the right-hand side. Adding and subtracting $\frac{F(x)}{L_0 + L_1\|G(x,\zeta)\|}$ yields

$$\widehat{T}(x;\xi,\zeta) - T(x) = \frac{G(x,\xi) - F(x)}{L_0 + L_1\|G(x,\zeta)\|} + \left(\frac{F(x)}{L_0 + L_1\|G(x,\zeta)\|} - \frac{F(x)}{L_0 + L_1\|F(x)\|}\right).$$

It follows that

$$\mathbb{E}[\|\widehat{T}(x;\xi,\zeta) - T(x)\|^2] = \mathbb{E}\left[\left\|\frac{G(x,\xi) - F(x)}{L_0 + L_1\|G(x,\zeta)\|}\right\|^2\right]$$
$$+ \mathbb{E}\left[\left\|\frac{F(x)}{L_0 + L_1\|G(x,\zeta)\|} - \frac{F(x)}{L_0 + L_1\|F(x)\|}\right\|^2\right]$$

where the cross-term vanishes by conditioning first on $\sigma(\zeta)$, and using Assumption 4.1 and the independence hypotheses. The first term on the right-hand side is upper bounded by $\frac{1}{L_0^2}\mathbb{E}[\|G(x,\xi) - F(x)\|^2] \leq \frac{\sigma^2}{L_0^2}$ by Assumption 4.1. For the second term, we calculate

$$\left\|\frac{F(x)}{L_0 + L_1\|G(x,\zeta)\|} - \frac{F(x)}{L_0 + L_1\|F(x)\|}\right\| = \frac{L_1\|F(x)\|\,|\|F(x)\| - \|G(x,\zeta)\||}{(L_0 + L_1\|G(x,\zeta)\|)(L_0 + L_1\|F(x)\|)}.$$

By the reverse triangle inequality, we have $|\|F(x)\| - \|G(x,\zeta)\|| \leq \|F(x) - G(x,\zeta)\|$. Further, since $\frac{L_1\|F(x)\|}{L_0 + L_1\|F(x)\|} \leq 1$ and $L_0 \leq L_0 + L_1\|G(x,\zeta)\|$, it follows that

$$\left\|\frac{F(x)}{L_0 + L_1\|G(x,\zeta)\|} - \frac{F(x)}{L_0 + L_1\|F(x)\|}\right\|^2 \leq \frac{1}{L_0^2}\|F(x) - G(x,\zeta)\|^2.$$

Taking expectation, and using Assumption 4.1 yields

$$\mathbb{E}\left[\left\|\frac{F(x)}{L_0 + L_1\|G(x,\zeta)\|} - \frac{F(x)}{L_0 + L_1\|F(x)\|}\right\|^2\right] \leq \frac{1}{L_0^2}\mathbb{E}[\|F(x) - G(x,\zeta)\|^2] \leq \frac{\sigma^2}{L_0^2}.$$

We conclude that

$$\mathbb{E}[\|\widehat{T}(x;\xi,\zeta) - c(x)T(x)\|^2] \leq \mathbb{E}[\|\widehat{T}(x;\xi,\zeta) - T(x)\|^2] \leq \frac{2\sigma^2}{L_0^2},$$

showing the first part of (7). Finally, write $\widehat{T}(x;\xi,\zeta) = (\widehat{T}(x;\xi,\zeta) - c(x)T(x)) + c(x)T(x)$, so that

$$\mathbb{E}[\|\widehat{T}(x;\xi,\zeta)\|^2] = \mathbb{E}[\|\widehat{T}(x;\xi,\zeta) - c(x)T(x)\|^2] + c(x)^2\|T(x)\|^2$$

where the cross-term vanishes due to (6). The second part of (7) now follows from the first part of (7). □

As a consequence of Lemma 4.3, we next establish integrability of several quantities of interest, ensuring the existence of all expectations in the subsequent analysis.

Lemma B.2. *Let $x^\star \in S$ be fixed. Under the standing assumptions and independent sampling scheme, for every $k \geq 0$,*

$$\mathbb{E}\left[\|x^k - x^\star\|^2 + \|z^k - x^\star\|^2 + \|m^k\|^2 + \|\widehat{T}(z^k;\xi^k,\zeta^k)\|^2 + \|n^k\|^2\right] < \infty.$$

Proof. For any random vector X with $\mathbb{E}[\|X - x^\star\|^2] < \infty$ and i.i.d. samples ξ, ζ independent of X, Lemma 4.3 and the 1-Lipschitz continuity of T give

$$\begin{aligned}\mathbb{E}[\|\widehat{T}(X;\xi,\zeta)\|^2] &\leq (1+\delta)^2\mathbb{E}[\|T(X)\|^2] + \frac{2\sigma^2}{L_0^2} \\ &\leq 2(1+\delta)^2\left(\mathbb{E}[\|X - x^\star\|^2] + \|T(x^\star)\|^2\right) + \frac{2\sigma^2}{L_0^2} < \infty.\end{aligned}$$

From this, starting from deterministic $x^0 = z^0$ and $m^{-1} = 0$, the momentum recursion (9), and nonexpansiveness of P_C in (10)–(11) establish finite second moments of the iterates, momentum vectors, and oracle estimates by induction. □

By Lemma B.2, all quantities appearing inside expectations in the analysis below are integrable. Indeed, Lipschitz continuity of T and the bounds (18) imply that ε^k, e^k, and r^k, defined in (12), (15), and (17), respectively, have finite second moments, as do $T(z^{k+1}) + n^k$ and the finite weighted average $\bar{\varepsilon}^k$ defined in Proposition B.5. Therefore, also $h_{k,i}$, H_k, and $V_k(x^\star)$, defined in (29) and (31), have finite expectations. Moreover, all inner products appearing inside expectations are integrable by Cauchy-Schwarz.

B.5 Lyapunov recursion

We start by establishing a one-step descent estimate for the distance to a solution.

Proposition B.3. *Let $x^\star \in S$. In the setting of Theorem 4.5, for $k \geq 0$, we have*

$$\mathbb{E}[\|x^{k+1} - x^\star\|^2 \mid \mathcal{F}_k] \leq \|x^k - x^\star\|^2 - \|x^k - z^{k+1}\|^2 + \|r^k\|^2 + \omega_k^2\frac{2\sigma^2}{L_0^2}.$$

Proof. From the projection inequality (25) with $x = x^\star, y = x^k - \omega_k\widehat{T}(z^{k+1};\xi^{k+1},\zeta^{k+1})$ and $P_C(y) \overset{(11)}{=} x^{k+1}$ along with the Pythagorean theorem, we have:

$$\begin{aligned}\|x^{k+1} - x^\star\|^2 &\leq \|x^k - x^\star\|^2 - 2\omega_k\langle\widehat{T}(z^{k+1};\xi^{k+1},\zeta^{k+1}), x^{k+1} - x^\star\rangle - \|x^{k+1} - x^k\|^2 \\ &\overset{(12)}{=} \|x^k - x^\star\|^2 - \|x^{k+1} - x^k\|^2 \underbrace{-2\omega_k\langle\varepsilon^{k+1} + c(z^{k+1})T(z^{k+1}), x^{k+1} - x^\star\rangle}_{A_k}.\end{aligned}$$

For the inner product term A_k, adding and subtracting z^{k+1}, we get:

$$A_k \leq -2\omega_k\langle\varepsilon^{k+1} + c(z^{k+1})T(z^{k+1}), x^{k+1} - z^{k+1}\rangle - 2\omega_k\langle\varepsilon^{k+1}, z^{k+1} - x^\star\rangle,$$

where we have also used the nonnegativity of $c(z^{k+1})\langle T(z^{k+1}), z^{k+1} - x^\star\rangle$. Indeed, $c(z^{k+1})$ is trivially nonnegative. Further, Assumption 4.2 ensures that Definition 3.4 holds. Taking the zero normal in Definition 3.4 gives $\langle F(z^{k+1}), z^{k+1} - x^\star\rangle \geq 0$, hence $\langle T(z^{k+1}), z^{k+1} - x^\star\rangle \geq 0$ due to $\frac{1}{L_0 + L_1\|F(z^{k+1})\|}$ being positive.

Further, by adding and subtracting n^k as defined in (16), we obtain

$$\begin{aligned}-2\omega_k c(z^{k+1})\langle T(z^{k+1}), x^{k+1} - z^{k+1}\rangle = 2\omega_k c(z^{k+1})[&-\langle T(z^{k+1}) + n^k, x^{k+1} - z^{k+1}\rangle \\ &+ \langle n^k, x^{k+1} - z^{k+1}\rangle].\end{aligned}$$

Since $n^k \in N_C(z^{k+1})$ by (16) and $x^{k+1} \in C$, it follows from the definition of the normal cone that the second inner product term on the r.h.s. of the equality above is nonpositive. Therefore,

$$\|x^{k+1} - x^\star\|^2 \leq \|x^k - x^\star\|^2 - \|x^{k+1} - x^k\|^2 - 2\omega_k\langle \varepsilon^{k+1}, z^{k+1} - x^\star\rangle \\ \underbrace{-2\omega_k\langle c(z^{k+1})(T(z^{k+1}) + n^k) + \varepsilon^{k+1}, x^{k+1} - z^{k+1}\rangle}_{B_k}.$$

Now applying (24) with $a = x^{k+1} - x^k$, $b = x^k - z^{k+1}$ and $d = \omega_k(c(z^{k+1})(T(z^{k+1}) + n^k) + \varepsilon^{k+1})$, we have:

$$-\|x^{k+1} - x^k\|^2 + B_k \overset{(24),(17)}{=} -\|x^{k+1} - x^k + \omega_k(c(z^{k+1})(T(z^{k+1}) + n^k) + \varepsilon^{k+1})\|^2 \\ - \|x^k - z^{k+1}\|^2 + \|r^k - \omega_k\varepsilon^{k+1}\|^2 \\ \leq -\|x^k - z^{k+1}\|^2 + \|r^k - \omega_k\varepsilon^{k+1}\|^2.$$

Putting everything together, and using the Pythagorean theorem once more, we get:

$$\|x^{k+1} - x^\star\|^2 \leq \|x^k - x^\star\|^2 - \|x^k - z^{k+1}\|^2 + \|r^k\|^2 - 2\omega_k\langle r^k, \varepsilon^{k+1}\rangle \\ + \omega_k^2\|\varepsilon^{k+1}\|^2 - 2\omega_k\langle \varepsilon^{k+1}, z^{k+1} - x^\star\rangle.$$

Now take conditional expectation with respect to $\mathcal{F}_k$ and note that since r^k and $z^{k+1} - x^\star$ are $\mathcal{F}_k$-measurable,

$$\mathbb{E}[\langle \varepsilon^{k+1}, z^{k+1} - x^\star\rangle \mid \mathcal{F}_k] = \mathbb{E}[\langle r^k, \varepsilon^{k+1}\rangle \mid \mathcal{F}_k] \overset{(13)}{=} 0.$$

Furthermore, in light of (13), $\mathbb{E}[\|\varepsilon^{k+1}\|^2 \mid \mathcal{F}_k] \leq \frac{2\sigma^2}{L_0^2}$. Combining these facts leads to the claimed result. □

Proposition B.4. *In the setting of Theorem 4.5, for $k \geq 0$, we have*

$$\|r^k\|^2 \leq \|x^k - z^{k+1}\|^2 - \frac{11}{16}\frac{\omega_k c(z^{k+1})}{\gamma\lambda_k}\|x^k - z^{k+1}\|^2 \\ - \frac{2}{3}\omega_k c(z^{k+1})\gamma\lambda_k\|T(z^{k+1}) + n^k\|^2 + \frac{5}{4}\omega_k\gamma\frac{c(z^{k+1})}{\lambda_k}\|e^k\|^2.$$

Proof. By the Pythagorean theorem, we have

$$\|r^k\|^2 \overset{(17)}{=} \|x^k - z^{k+1}\|^2 - 2\omega_k c(z^{k+1})\langle x^k - z^{k+1}, T(z^{k+1}) + n^k\rangle + \omega_k^2 c(z^{k+1})^2\|T(z^{k+1}) + n^k\|^2.$$

We now proceed to bound $\langle x^k - z^{k+1}, T(z^{k+1}) + n^k\rangle$. By (23) with $a = x^k - z^{k+1}$ and $b = \gamma\lambda_k(T(z^{k+1}) + n^k)$, we have

$$2\gamma\lambda_k\langle x^k - z^{k+1}, T(z^{k+1}) + n^k\rangle = \|x^k - z^{k+1}\|^2 + \gamma^2\lambda_k^2\|T(z^{k+1}) + n^k\|^2 \\ - \|x^k - z^{k+1} - \gamma\lambda_k(T(z^{k+1}) + n^k)\|^2.$$

Further, the definitions of n^k and e^k (16) and (15) yield $x^k - z^{k+1} - \gamma\lambda_k(T(z^{k+1}) + n^k) = \gamma e^k - \gamma\lambda_k(T(z^{k+1}) - T(x^k))$. Hence, by Young's inequality (22) with weight $\gamma\lambda_k$, and the Lipschitz continuity of T, we obtain

$$\|x^k - z^{k+1} - \gamma\lambda_k(T(z^{k+1}) + n^k)\|^2 = \|\gamma e^k - \gamma\lambda_k(T(z^{k+1}) - T(x^k))\|^2 \\ \overset{(22),(1)}{\leq} \gamma^2(1 + \gamma\lambda_k)\|e^k\|^2 + \gamma\lambda_k(1 + \gamma\lambda_k)\|x^k - z^{k+1}\|^2.$$

It follows that

$$-2\gamma\lambda_k\langle x^k - z^{k+1}, T(z^{k+1}) + n^k\rangle \leq -(1 - \gamma\lambda_k - \gamma^2\lambda_k^2)\|x^k - z^{k+1}\|^2 \\ - \gamma^2\lambda_k^2\|T(z^{k+1}) + n^k\|^2 + \gamma^2(1 + \gamma\lambda_k)\|e^k\|^2.$$

Multiplying by $\frac{\omega_k c(z^{k+1})}{\gamma\lambda_k} > 0$ and combining this with our preceding decomposition of $\|r^k\|^2$ leads to

$$\|r^k\|^2 \le \|x^k - z^{k+1}\|^2 - \frac{\omega_k c(z^{k+1})}{\gamma\lambda_k}(1 - \gamma\lambda_k - \gamma^2\lambda_k^2)\|x^k - z^{k+1}\|^2 \\ - \omega_k c(z^{k+1})(\gamma\lambda_k - \omega_k c(z^{k+1}))\|T(z^{k+1}) + n^k\|^2 + \omega_k\gamma\frac{c(z^{k+1})}{\lambda_k}(1 + \gamma\lambda_k)\|e^k\|^2.$$

Moreover, (18), (19), $\omega_k = \theta\alpha_{k+1}$, and $0 < \alpha_{k+1} \le 1$ ensure that $\gamma\lambda_k \le \gamma(1+\delta) \le \frac{1}{4}$ and $\frac{\omega_k c(z^{k+1})}{\gamma\lambda_k} \le \frac{\alpha_{k+1}}{\gamma}\theta(1+\delta)^2 \le \frac{\alpha_{k+1}}{3} \le \frac{1}{3}$. Therefore,

$$1 - \gamma\lambda_k - \gamma^2\lambda_k^2 \ge 1 - \frac{1}{4} - \frac{1}{16} = \frac{11}{16}, \quad \gamma\lambda_k - \omega_k c(z^{k+1}) = \gamma\lambda_k\left(1 - \frac{\omega_k c(z^{k+1})}{\gamma\lambda_k}\right) \ge \frac{2}{3}\gamma\lambda_k.$$

Inserting these bounds in our preceding bound on $\|r^k\|^2$, and using $1 + \gamma\lambda_k \le \frac{5}{4}$, yields the desired result. □

In the r.h.s. of Proposition B.4 we have two controllable terms and the one involving $\|e^k\|^2$. Using a recursion directly on this quantity does not seem to lead to convergence and we thus proceed to obtain upper bounds in Proposition B.5.

First, we unroll the momentum recursion. Define the standard momentum weights

$$w_{0,0} := 1, \qquad w_{k,0} := \prod_{l=1}^{k}(1 - \alpha_l), \qquad w_{k,i} := \alpha_i \prod_{l=i+1}^{k}(1 - \alpha_l), \quad 1 \le i \le k, \tag{27}$$

where we adopt the usual convention that the empty product equals one. Due to the convex combination, $\sum_{i=0}^{k} w_{k,i} = 1$ for each $k \in \mathbb{N}$. It follows from (9) that

$$\begin{aligned} m^0 &= \widehat{T}(z^0;\xi^0,\zeta^0) = w_{0,0}\widehat{T}(z^0;\xi^0,\zeta^0) \\ m^1 &= (1-\alpha_1)m^0 + \alpha_1\widehat{T}(z^1;\xi^1,\zeta^1) = w_{1,0}\widehat{T}(z^0;\xi^0,\zeta^0) + w_{1,1}\widehat{T}(z^1;\xi^1,\zeta^1) \\ &\vdots \\ m^k &= (1-\alpha_k)m^{k-1} + \alpha_k\widehat{T}(z^k;\xi^k,\zeta^k) = \sum_{i=0}^{k} w_{k,i}\widehat{T}(z^i;\xi^i,\zeta^i). \end{aligned}$$

Similarly, we have $\lambda_k = \sum_{i=0}^{k} w_{k,i}c(z^i)$ by (14) due to satisfying the same recursion. We are now ready to bound the momentum tracking error by an averaged noise term and a history-dependent term.

Proposition B.5. *Define $\bar{\varepsilon}^k := \sum_{i=0}^{k} w_{k,i}\varepsilon^i$. In the setting of Theorem 4.5, for $k \ge 0$, we have*

$$\|e^k\|^2 \le 2\|\bar{\varepsilon}^k\|^2 + 2\lambda_k\sum_{i=0}^{k} w_{k,i}c(z^i)\|x^k - z^i\|^2. \tag{28}$$

Proof. By our preceding recursions for m^k and λ_k, and (15) and (12), we obtain

$$e^k = \sum_{i=0}^{k} w_{k,i}(\varepsilon^i + c(z^i)T(z^i)) - \sum_{i=0}^{k} w_{k,i}c(z^i)T(x^k) = \bar{\varepsilon}^k + \sum_{i=0}^{k} w_{k,i}c(z^i)(T(z^i) - T(x^k)).$$

It follows that

$$\|e^k\|^2 \overset{(22)}{\le} 2\|\bar{\varepsilon}^k\|^2 + 2\lambda_k\sum_{i=0}^{k} w_{k,i}c(z^i)\|T(z^i) - T(x^k)\|^2 \overset{(1)}{\le} 2\|\bar{\varepsilon}^k\|^2 + 2\lambda_k\sum_{i=0}^{k} w_{k,i}c(z^i)\|z^i - x^k\|^2,$$

where we also used Jensen's inequality in the first inequality in light of $\sum_{i=0}^{k} w_{k,i}c(z^i)/\lambda_k = 1$. □

Proposition B.6. *In the setting of Theorem 4.5, for $k \geq 0$, we have*

$$\mathbb{E}[\|\bar{\varepsilon}^{k+1}\|^2 \mid \mathcal{F}_k] \leq (1-\alpha_{k+1})\,\|\bar{\varepsilon}^k\|^2 + \alpha_{k+1}^2 \frac{2\sigma^2}{L_0^2}.$$

Proof. Note that $\bar{\varepsilon}^{k+1} = (1-\alpha_{k+1})\bar{\varepsilon}^k + \alpha_{k+1}\varepsilon^{k+1}$ by the definition of the momentum weights $w_{k,i}$. Therefore, by the Pythagorean theorem

$$\|\bar{\varepsilon}^{k+1}\|^2 = (1-\alpha_{k+1})^2\|\bar{\varepsilon}^k\|^2 + 2\alpha_{k+1}(1-\alpha_{k+1})\langle\bar{\varepsilon}^k, \varepsilon^{k+1}\rangle + \alpha_{k+1}^2\|\varepsilon^{k+1}\|^2.$$

Taking conditional expectation with respect to $\mathcal{F}_k$, noting that $\bar{\varepsilon}^k$ is $\mathcal{F}_k$-measurable and using (13) and $(1-\alpha_{k+1})^2 \leq 1-\alpha_{k+1}$ due to $\alpha_{k+1} \leq 1$ leads to the claimed result. □

In Proposition B.6 we obtained a recursive bound for $\bar{\varepsilon}^k$. We now move on to bounding $\sum_{i=0}^k w_{k,i}c(z^i)(T(z^i) - T(x^k))$. Due to the Lipschitz continuity of T, the important part is $\|x^k - z^i\|^2$. We thus define our partial Lyapunov function in the following way:

$$H_k \coloneqq \sum_{i=0}^k w_{k,i}c(z^i)h_{k,i}, \quad \text{with} \quad h_{k,i} \coloneqq \|x^k - z^i\|^2 + 4\theta^2(1+\delta)^2\|T(z^i) + n^{i-1}\|^2. \tag{29}$$

Proposition B.7. *In the setting of Theorem 4.5, we have the following inequality for $k \geq 0$ and $0 \leq i \leq k$:*

$$(1-\alpha_{k+1})\mathbb{E}[h_{k+1,i} \mid \mathcal{F}_k] \leq \left(1 - \tfrac{\alpha_{k+1}(1+\alpha_{k+1})}{4}\right)h_{k,i} + \omega_k c(z^{k+1})\|x^k - z^{k+1}\|^2 + (1-\alpha_{k+1})\omega_k^2\frac{2\sigma^2}{L_0^2}.$$

Proof. We begin by bounding $\|x^{k+1} - z^i\|^2$ since the other term in (29) does not depend on k. Note that $x^{k+1} \overset{(11)}{=} P_C(x^k - \omega_k\widehat{T}(z^{k+1}, \xi^{k+1}, \zeta^{k+1}))$ and since $n^{i-1} \in N_C(z^i)$, we have $z^i = P_C(z^i + \omega_k c(z^{k+1})n^{i-1})$. Therefore, using the nonexpansiveness of the projection,

$$\begin{aligned}
\|x^{k+1} - z^i\|^2 &\leq \|x^k - z^i - \omega_k\widehat{T}(z^{k+1}, \xi^{k+1}, \zeta^{k+1}) - \omega_k c(z^{k+1})n^{i-1}\|^2 \\
&\overset{(12)}{=} \|x^k - z^i - \omega_k c(z^{k+1})(T(z^{k+1}) + n^{i-1}) - \omega_k\varepsilon^{k+1}\|^2 \\
&= \|\hat{r}^{k,i}\|^2 + \omega_k^2\|\varepsilon^{k+1}\|^2 - 2\langle\hat{r}^{k,i}, \omega_k\varepsilon^{k+1}\rangle
\end{aligned}$$

where we defined $\hat{r}^{k,i} = x^k - z^i - \omega_k c(z^{k+1})(T(z^{k+1}) + n^{i-1})$ for ease of notation. Taking conditional expectation with respect to $\mathcal{F}_k$ and using (13) we are led to the following bound:

$$\mathbb{E}[\|x^{k+1} - z^i\|^2 \mid \mathcal{F}_k] \leq \|\hat{r}^{k,i}\|^2 + \omega_k^2\frac{2\sigma^2}{L_0^2}. \tag{30}$$

Moreover, it follows from the Pythagorean theorem that

$$\begin{aligned}
\|\hat{r}^{k,i}\|^2 = \|x^k - z^i\|^2 &+ \omega_k^2 c(z^{k+1})^2\|T(z^{k+1}) + n^{i-1}\|^2 \\
&- 2\omega_k c(z^{k+1})\langle x^k - z^i, T(z^{k+1}) - T(z^i)\rangle - 2\omega_k c(z^{k+1})\langle x^k - z^i, T(z^i) + n^{i-1}\rangle,
\end{aligned}$$

where we have also added and subtracted $T(z^i)$ in the inner product term. We now proceed by bounding all the terms in the r.h.s. of the equality above. First,

$$\begin{aligned}
\omega_k^2 c(z^{k+1})^2\|T(z^{k+1}) + n^{i-1}\|^2 &\overset{(22)}{\leq} 2\omega_k^2 c(z^{k+1})^2[\|T(z^{k+1}) - T(z^i)\|^2 + \|T(z^i) + n^{i-1}\|^2] \\
&\overset{(22),(1),(6)}{\leq} 4\omega_k^2 c(z^{k+1})^2[\|z^{k+1} - x^k\|^2 + \|x^k - z^i\|^2] \\
&\quad + 2\omega_k^2(1+\delta)^2\|T(z^i) + n^{i-1}\|^2.
\end{aligned}$$

Second,

$$\begin{aligned}
-2\omega_k c(z^{k+1})\langle x^k - z^i, T(z^{k+1}) - T(z^i)\rangle &\leq 2\omega_k c(z^{k+1})[\|x^k - z^i\|\|x^k - z^{k+1}\| + \|x^k - z^i\|^2] \\
&\overset{(21)}{\leq} \omega_k c(z^{k+1})\|x^k - z^{k+1}\|^2 + 3\omega_k c(z^{k+1})\|x^k - z^i\|^2,
\end{aligned}$$

where the first inequality follows by Cauchy–Schwarz, (1) and the triangle inequality. Third, we use $\omega_k = \theta\alpha_{k+1}$, (21) with $e = \frac{\alpha_{k+1}}{2}$ and (18) to obtain

$$-2\omega_k c(z^{k+1})\langle x^k - z^i, T(z^i) + n^{i-1}\rangle \leq \frac{\alpha_{k+1}}{2}\|x^k - z^i\|^2 + 2\alpha_{k+1}\theta^2(1+\delta)^2\|T(z^i) + n^{i-1}\|^2.$$

Putting everything together and using $\omega_k = \theta\alpha_{k+1}$ and $\omega_k c(z^{k+1}) \leq \alpha_{k+1}\theta(1+\delta) \leq \frac{\alpha_{k+1}}{12}$, see (19), we obtain

$$\mathbb{E}[h_{k+1,i} \mid \mathcal{F}_k] \leq \left(1 + \frac{3\alpha_{k+1}}{4} + \frac{\alpha_{k+1}^2}{36}\right)\|x^k - z^i\|^2 + \omega_k c(z^{k+1})\left(1 + \frac{\alpha_{k+1}}{3}\right)\|x^k - z^{k+1}\|^2 + \left(1 + \frac{\alpha_{k+1}}{2} + \frac{\alpha_{k+1}^2}{2}\right)4\theta^2(1+\delta)^2\|T(z^i) + n^{i-1}\|^2 + \omega_k^2\frac{2\sigma^2}{L_0^2}.$$

We note that the coefficients of the first and third terms can both be bounded by $1 + \frac{3\alpha_{k+1}}{4} + \frac{\alpha_{k+1}^2}{2}$, hence

$$\mathbb{E}[h_{k+1,i} \mid \mathcal{F}_k] \leq \left(1 + \frac{3\alpha_{k+1}}{4} + \frac{\alpha_{k+1}^2}{2}\right)h_{k,i} + \omega_k c(z^{k+1})\left(1 + \frac{\alpha_{k+1}}{3}\right)\|x^k - z^{k+1}\|^2 + \omega_k^2\frac{2\sigma^2}{L_0^2}.$$

Using (20) yields the claimed result. □

Proposition B.8. *In the setting of Theorem 4.5, we have the following inequality for $k \geq 0$:*

$$\mathbb{E}[H_{k+1} \mid \mathcal{F}_k] \leq \left(1 - \tfrac{\alpha_{k+1}(1+\alpha_{k+1})}{4}\right)H_k + \omega_k c(z^{k+1})\lambda_k\|x^k - z^{k+1}\|^2 + \alpha_{k+1}c(z^{k+1})\|r^k\|^2 + 4\omega_k\theta c(z^{k+1})(1+\delta)^2\|T(z^{k+1}) + n^k\|^2 + \omega_k^2\lambda_{k+1}\frac{2\sigma^2}{L_0^2}.$$

Proof. Based on the definition of the momentum weights (27), we have for $0 \leq i \leq k$ that $w_{k+1,i} = (1-\alpha_{k+1})w_{k,i}$ and $w_{k+1,k+1} = \alpha_{k+1}$, hence

$$\mathbb{E}[H_{k+1} \mid \mathcal{F}_k] = \sum_{i=0}^{k} c(z^i)w_{k,i}(1-\alpha_{k+1})\mathbb{E}[h_{k+1,i} \mid \mathcal{F}_k] + \alpha_{k+1}c(z^{k+1})\mathbb{E}[h_{k+1,k+1} \mid \mathcal{F}_k].$$

Applying Proposition B.7, and using (29) and $\lambda_k = \sum_{i=0}^{k} w_{k,i}c(z^i)$, we obtain

$$\mathbb{E}[H_{k+1} \mid \mathcal{F}_k] \leq \left(1 - \tfrac{\alpha_{k+1}(1+\alpha_{k+1})}{4}\right)H_k + \omega_k c(z^{k+1})\lambda_k\|x^k - z^{k+1}\|^2 + (1-\alpha_{k+1})\omega_k^2\lambda_k\frac{2\sigma^2}{L_0^2} + \alpha_{k+1}c(z^{k+1})\mathbb{E}[h_{k+1,k+1} \mid \mathcal{F}_k].$$

Regarding the term $h_{k+1,k+1} = \|x^{k+1} - z^{k+1}\|^2 + 4\theta^2(1+\delta)^2\|T(z^{k+1}) + n^k\|^2$, we argue as in the derivation of (30), now with $i = k+1$, to obtain

$$\mathbb{E}[\|x^{k+1} - z^{k+1}\|^2 \mid \mathcal{F}_k] \leq \|x^k - z^{k+1} - \omega_k c(z^{k+1})(T(z^{k+1}) + n^k)\|^2 + \omega_k^2\frac{2\sigma^2}{L_0^2} \overset{(17)}{=} \|r^k\|^2 + \omega_k^2\frac{2\sigma^2}{L_0^2}.$$

It follows that

$$\alpha_{k+1}c(z^{k+1})\mathbb{E}[h_{k+1,k+1} \mid \mathcal{F}_k] \leq \alpha_{k+1}c(z^{k+1})\|r^k\|^2 + \alpha_{k+1}c(z^{k+1})\omega_k^2\frac{2\sigma^2}{L_0^2} + 4\omega_k\theta c(z^{k+1})(1+\delta)^2\|T(z^{k+1}) + n^k\|^2.$$

Putting everything together and using the definition of λ_{k+1} leads to the claimed result. □

Having sufficient bounds for all the terms showing up in our analysis, we can now define our Lyapunov function:

$$V_k(x^\star) := \|x^k - x^\star\|^2 + 4\theta\gamma(1+\delta)^2\|\bar{\varepsilon}^k\|^2 + 10\theta\gamma(1+\delta)H_k, \tag{31}$$

for some $x^\star \in S$.

Proposition B.9. *In the setting of Theorem 4.5, we have the following inequality for $k \geq 0$ and for any $x^\star \in S$:*

$$\mathbb{E}[V_{k+1}(x^\star) \mid \mathcal{F}_k] \leq V_k(x^\star) - \frac{\theta\alpha_{k+1}}{96\gamma(1+\delta)^2}\|x^k - z^{k+1}\|^2 - \frac{7}{18}\frac{\theta\gamma\alpha_{k+1}}{(1+\delta)^2}\|T(z^{k+1}) + n^k\|^2 + \frac{53}{12}\theta\gamma\alpha_{k+1}^2(1+\delta)^2\frac{2\sigma^2}{L_0^2}.$$

Proof. By the definition of V_k in (31), we have

$$\mathbb{E}[V_{k+1}(x^\star) \mid \mathcal{F}_k] = \mathbb{E}[\|x^{k+1} - x^\star\|^2 + 4\theta\gamma(1+\delta)^2\|\bar{\varepsilon}^{k+1}\|^2 + 10\theta\gamma(1+\delta)H_{k+1} \mid \mathcal{F}_k].$$

Applying linearity of expectation and Propositions B.3, B.6 and B.8 yields

$$\mathbb{E}[V_{k+1}(x^\star) \mid \mathcal{F}_k] \leq V_k(x^\star) - \|x^k - z^{k+1}\|^2 + C_1\|r^k\|^2 + C_2\|x^k - z^{k+1}\|^2 + C_3\|T(z^{k+1}) + n^k\|^2 - C_4\|\bar{\varepsilon}^k\|^2 - C_5H_k + C_6\frac{2\sigma^2}{L_0^2}$$

where

$$C_1 = 1 + 10\theta\gamma(1+\delta)\alpha_{k+1}c(z^{k+1}), \quad C_2 = 10\theta\gamma(1+\delta)\omega_k c(z^{k+1})\lambda_k,$$
$$C_3 = 40\theta^2\gamma(1+\delta)^3\omega_k c(z^{k+1}), \quad C_4 = 4\theta\gamma(1+\delta)^2\alpha_{k+1},$$
$$C_5 = \frac{5}{2}\theta\gamma(1+\delta)\alpha_{k+1}(1+\alpha_{k+1}), \quad C_6 = \omega_k^2 + 4\theta\gamma(1+\delta)^2\alpha_{k+1}^2 + 10\theta\gamma(1+\delta)\omega_k^2\lambda_{k+1}.$$

By our bound on $\|r^k\|^2$ from Proposition B.4 and collecting like terms, we obtain

$$\mathbb{E}[V_{k+1}(x^\star) \mid \mathcal{F}_k] \leq V_k(x^\star) + \left(C_1 - 1 + C_2 - \frac{11}{16}C_1\frac{\omega_k c(z^{k+1})}{\gamma\lambda_k}\right)\|x^k - z^{k+1}\|^2 + \left(C_3 - \frac{2}{3}C_1\omega_k c(z^{k+1})\gamma\lambda_k\right)\|T(z^{k+1}) + n^k\|^2 + \frac{5}{4}C_1\omega_k\gamma\frac{c(z^{k+1})}{\lambda_k}\|e^k\|^2 - C_4\|\bar{\varepsilon}^k\|^2 - C_5H_k + C_6\frac{2\sigma^2}{L_0^2}.$$

We first show that the terms corresponding to $\|e^k\|^2$, $\|\bar{\varepsilon}^k\|^2$ and H_k together become nonpositive, so they may be discarded. Indeed, (28) and (29) ensure that $\|e^k\|^2 \leq 2\|\bar{\varepsilon}^k\|^2 + 2\lambda_k H_k$. Moreover, $10\theta\gamma(1+\delta)c(z^{k+1}) \leq 10\gamma\theta(1+\delta)^2 \leq \frac{10}{12}\gamma \leq \frac{5}{24} \leq \frac{1}{4}$ by (18) and (19), hence $C_1 \leq 1 + \frac{\alpha_{k+1}}{4}$ and

$$\frac{5}{4}C_1\omega_k\gamma\frac{c(z^{k+1})}{\lambda_k}\|e^k\|^2 \leq \frac{5}{2}\left(1 + \frac{\alpha_{k+1}}{4}\right)\omega_k\gamma((1+\delta)^2\|\bar{\varepsilon}^k\|^2 + (1+\delta)H_k)$$

where we also used (18).

Therefore, utilizing the definitions of C_4 and C_5, and $\omega_k = \theta\alpha_{k+1}$, the group consisting of $\|e^k\|^2$, $\|\bar{\varepsilon}^k\|^2$ and H_k can be bounded by

$$\omega_k\gamma(1+\delta)^2\left(\frac{5}{2}\left(1 + \frac{\alpha_{k+1}}{4}\right) - 4\right)\|\bar{\varepsilon}^k\|^2 + \frac{5}{2}\omega_k\gamma(1+\delta)\left(1 + \frac{\alpha_{k+1}}{4} - (1+\alpha_{k+1})\right)H_k$$

where the terms inside the parentheses are nonpositive due to $0 < \alpha_{k+1} \leq 1$. It follows that we may discard the terms corresponding to $\|e^k\|^2$, $\|\bar{\varepsilon}^k\|^2$ and H_k.

For the coefficient corresponding to $\|x^k - z^{k+1}\|^2$, we may bound it as

$$\left(C_1 - 1 + C_2 - \frac{11}{16}C_1\frac{\omega_k c(z^{k+1})}{\gamma\lambda_k}\right) \leq -\frac{\omega_k c(z^{k+1})}{\gamma\lambda_k}\left(\frac{11}{16} - 10\gamma^2(1+\delta)\lambda_k(1+\theta\lambda_k)\right)$$

by utilizing $C_1 \geq 1$ and $C_1 - 1 + C_2 = 10\gamma(1+\delta)\omega_k c(z^{k+1})(1+\theta\lambda_k)$. Similarly, we can bound the coefficient corresponding to $\|T(z^{k+1}) + n^k\|^2$ via

$$\left(C_3 - \frac{2}{3}C_1\omega_k c(z^{k+1})\gamma\lambda_k\right) \leq -\omega_k c(z^{k+1})\gamma\lambda_k\left(\frac{2}{3} - \frac{40\theta^2(1+\delta)^3}{\lambda_k}\right),$$

In light of (18) and (19), we have that $10\gamma^2(1+\delta)\lambda_k(1+\theta\lambda_k) \leq 10(\gamma(1+\delta))^2(1+\theta(1+\delta)) \leq 10 \cdot \frac{1}{16} \cdot \frac{13}{12} = \frac{65}{96}$ and $\frac{40\theta^2(1+\delta)^3}{\lambda_k} \leq 40(\theta(1+\delta)^2)^2 \leq 40 \cdot \frac{1}{144} = \frac{5}{18}$. Since $\frac{11}{16} - \frac{65}{96} = \frac{1}{96}$ and $\frac{2}{3} - \frac{5}{18} = \frac{7}{18}$, and using (18), the desired coefficients corresponding to $\|x^k - z^{k+1}\|^2$ and $\|T(z^{k+1}) + n^k\|^2$ follow.

Lastly, we can write $C_6 = \omega_k^2(1 + 10\theta\gamma(1+\delta)\lambda_{k+1}) + 4\theta\gamma(1+\delta)^2\alpha_{k+1}^2$. By (18) and (19), we have $1 + 10\theta\gamma(1+\delta)\lambda_{k+1} \leq 1 + 10\gamma\theta(1+\delta)^2 \leq 1 + \frac{10}{3}\gamma^2 \leq 1 + \frac{10}{3} \cdot \frac{1}{16} = \frac{29}{24} \leq \frac{5}{4}$. Similarly, $\omega_k^2 = \omega_k\theta\alpha_{k+1} \leq \frac{1}{3}\omega_k\gamma\alpha_{k+1}(1+\delta)^2$ and $4\theta\gamma(1+\delta)^2\alpha_{k+1}^2 = 4\omega_k\gamma\alpha_{k+1}(1+\delta)^2$. Therefore, $C_6 \leq \left(\frac{5}{12} + 4\right)\omega_k\gamma\alpha_{k+1}(1+\delta)^2 = \frac{53}{12}\omega_k\gamma\alpha_{k+1}(1+\delta)^2$, finishing the proof. □

We now proceed to prove the results stated in the main text.

B.6 Proof of Theorem 4.5

Proof. We start from Proposition B.9, take total expectation, use the tower property and sum from $k = 0$ to $k = K - 1$ to obtain

$$\mathbb{E}[V_K(x^\star)] \leq \mathbb{E}[V_0(x^\star)] - \frac{\theta}{96\gamma(1+\delta)^2}\sum_{k=0}^{K-1}\alpha_{k+1}\mathbb{E}[\|x^k - z^{k+1}\|^2] - \frac{7\theta\gamma}{18(1+\delta)^2}\sum_{k=0}^{K-1}\alpha_{k+1}\mathbb{E}[\|T(z^{k+1}) + n^k\|^2] + \frac{53}{12}\theta\gamma(1+\delta)^2\frac{2\sigma^2}{L_0^2}\sum_{k=0}^{K-1}\alpha_{k+1}^2.$$

Since $V_K(x^\star) \geq 0$, and using $S_K = \sum_{k=0}^{K-1}\alpha_{k+1}$ and $Q_K = \sum_{k=0}^{K-1}\alpha_{k+1}^2$, it follows that

$$\frac{1}{S_K}\sum_{k=0}^{K-1}\alpha_{k+1}\mathbb{E}[\|T(z^{k+1}) + n^k\|^2] \leq \frac{18}{7}\frac{(1+\delta)^2}{\theta\gamma S_K}\mathbb{E}[V_0(x^\star)] + \frac{159}{14}(1+\delta)^4\frac{2\sigma^2}{L_0^2}\frac{Q_K}{S_K} \tag{32}$$

where we also dropped the nonpositive term corresponding to $\mathbb{E}[\|x^k - z^{k+1}\|^2]$. By (31), $\bar{\varepsilon}^0 = \varepsilon^0$, (13), (29), $z^0 = x^0$, (16) and (18), the desired result follows with

$$\overline{V}_0(x^\star) = \|x^0 - x^\star\|^2 + 4\theta\gamma(1+\delta)^2\tfrac{2\sigma^2}{L_0^2} + 40\theta^3\gamma(1+\delta)^4\|T(x^0)\|^2. \tag{33}$$

□

B.7 Proof of Corollary 4.6

Proof. For $\alpha_{k+1} = K^{-1/2}$, we have that $S_K = \sum_{k=0}^{K-1}\alpha_{k+1} = \sqrt{K}$ and $Q_K = \sum_{k=0}^{K-1}\alpha_{k+1}^2 = 1$. Thus, by (32), together with the bound on $\mathbb{E}[V_0(x^\star)]$ in (33),

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|T(z^{k+1}) + n^k\|^2] \leq \frac{18}{7}\frac{(1+\delta)^2}{\theta\gamma}\frac{1}{\sqrt{K}}\left(\overline{V}_0(x^\star) + \frac{53}{12}(1+\delta)^2\theta\gamma\frac{2\sigma^2}{L_0^2}\right). \tag{34}$$

Since $n^k \in N_C(z^{k+1})$ by (16), the desired claim is proven due to $\operatorname{dist}^2(0, (T + N_C)(z^{k+1})) \leq \|T(z^{k+1}) + n^k\|^2$. □

B.8 Proof of Theorem 4.8

Before moving to the convergence analysis we will require the supermartingale convergence theorem as adapted from Bertsekas (2011, Proposition 2).

Proposition B.10. *Let $(Y_k)_{k\in\mathbb{N}}$, $(Z_k)_{k\in\mathbb{N}}$ and $(W_k)_{k\in\mathbb{N}}$ be sequences of random variables and let $\mathcal{P}_k$ be a filtration. Suppose that:*

1. *The random variables are nonnegative and are $\mathcal{P}_k$-measurable.*
2. *For each $k \geq 0$ we have*

$$\mathbb{E}[Y_{k+1} \mid \mathcal{P}_k] \leq Y_k - Z_k + W_k.$$

3. *Almost surely,* $\sum_{k=0}^{\infty} W_k < +\infty$.

Then, we have $\sum_{k=0}^{\infty} Z_k < +\infty$ *and the sequence* Y_k *converges to a nonnegative random variable* Y *almost surely.*

We now move on to the proof of Theorem 4.8.

Proof. Fix $x^\star \in S$. We aim at using the standard almost supermartingale theorem to obtain almost sure convergence of $\|x^k - x^\star\|^2$. To this end, we first show that $\bar{\varepsilon}^k$ and H_k vanish almost surely, and then show convergence of $\|x^k - x^\star\|^2$.

From Proposition B.6 and $\sum_{k=0}^{\infty} \alpha_{k+1}^2 < \infty$, we may apply Proposition B.10 with $Y_k = \|\bar{\varepsilon}^k\|^2$, $Z_k = \alpha_{k+1}\|\bar{\varepsilon}^k\|^2$ and $W_k = \alpha_{k+1}^2 \frac{2\sigma^2}{L_0^2}$ to obtain $\sum_{k=0}^{\infty} \alpha_{k+1}\|\bar{\varepsilon}^k\|^2 < +\infty$ almost surely. Furthermore, $\|\bar{\varepsilon}^k\|^2$ converges to some finite value almost surely. Since $\sum_{k=0}^{\infty} \alpha_{k+1} = +\infty$, $\|\bar{\varepsilon}^k\|^2 \to 0$ almost surely.

Next, from Proposition B.9 and Proposition B.10 we get $(V_k(x^\star))_{k\in\mathbb{N}}$ converges to some finite limit a.s., and

$$\sum_{k=0}^{\infty} \alpha_{k+1}(\|T(z^{k+1}) + n^k\|^2 + \|x^k - z^{k+1}\|^2) < +\infty, \qquad \text{a.s.} \tag{35}$$

Since $V_k(x^\star) \geq \|x^k - x^\star\|^2$ by (31), and $V_k(x^\star)$ converges to some finite limit a.s., it follows that $(x^k)_{k\in\mathbb{N}}$ is bounded almost surely. For H_k, starting from Proposition B.8 we need to show that the positive error terms in the recursion are summable almost surely in order to apply Proposition B.10. We have

$$\|r^k\|^2 \overset{(17),(22),(18)}{\leq} 2\|x^k - z^{k+1}\|^2 + 2\omega_k^2(1+\delta)^2\|T(z^{k+1}) + n^k\|^2.$$

Therefore, Proposition B.8 becomes

$$\begin{aligned}\mathbb{E}[H_{k+1} \mid \mathcal{F}_k] \leq \left(1 - \tfrac{\alpha_{k+1}(1+\alpha_{k+1})}{4}\right) H_k + (\theta(1+\delta)^2 + 2(1+\delta))\alpha_{k+1}\|x^k - z^{k+1}\|^2 \\ + 6\theta^2(1+\delta)^3\alpha_{k+1}\|T(z^{k+1}) + n^k\|^2 + \theta^2(1+\delta)\frac{2\sigma^2}{L_0^2}\alpha_{k+1}^2\end{aligned}$$

where we used (18), $\alpha_{k+1} \leq 1$ and $\omega_k = \theta\alpha_{k+1}$. Then, define $Y_k = H_k$, $Z_k = \frac{\alpha_{k+1}(1+\alpha_{k+1})}{4} H_k$ and

$$W_k = \alpha_{k+1}(C_1\|x^k - z^{k+1}\|^2 + C_2\|T(z^{k+1}) + n^k\|^2) + \theta^2(1+\delta)\frac{2\sigma^2}{L_0^2}\alpha_{k+1}^2,$$

for some positive deterministic constants C_1, C_2. Note that $\sum_{k=0}^{\infty} W_k < +\infty$ almost surely in light of (35) and $\sum_{k=0}^{\infty} \alpha_{k+1}^2 < \infty$, hence we may invoke Proposition B.10 to obtain that $(H_k)_{k\in\mathbb{N}}$ converges to a finite limit a.s. and $\sum_{k=0}^{\infty} \alpha_{k+1} H_k < +\infty$ a.s., hence $H_k \to 0$ a.s. Therefore, by (31) and the fact that $V_k(x^\star)$ converges to some finite limit a.s., $\|x^k - x^\star\|^2$ converges to some finite limit almost surely.

From (35) and $\sum_{k=0}^{\infty} \alpha_{k+1} = +\infty$, it follows that

$$\liminf_{k\to\infty} (\|T(z^{k+1}) + n^k\|^2 + \|x^k - z^{k+1}\|^2) = 0 \qquad \text{a.s.}$$

Hence, almost surely, there exists a subsequence such that $\|T(z^{k_j+1}) + n^{k_j}\| \to 0$ and $\|x^{k_j} - z^{k_j+1}\| \to 0$. Since $(x^k)_{k\in\mathbb{N}}$ is bounded a.s., passing to a subsubsequence, if needed, we have $x^{k_j} \to \tilde{x}$. Since $z^{k_j+1} \to \tilde{x}$ and T is continuous, we have $T(z^{k_j+1}) \to T(\tilde{x})$ and $n^{k_j} \to -T(\tilde{x}) \in N_C(\tilde{x})$ a.s. by closedness of the graph of N_C. Therefore, $\tilde{x} \in S$ due to $N_C(\tilde{x})$ being a cone. The rest of the proof now follows by the standard arguments found in Bertsekas (2011, Proposition 9). □

B.9 Convergence in probability

Throughout this section, we adopt the hypotheses of Corollary 4.6, and fix a deterministic $x^\star \in S$. Our main goal is to prove Theorem 4.9. The central challenge is that the established bounds on

the residual with respect to the preconditioned operator T do not straightforwardly carry over to the original operator F, as they are related by the random factor $L_0 + L_1\|F(z^{J+1})\|$, which need not be uniformly bounded. To address this difficulty, we establish in Lemma B.12 a second-moment bound uniform in K. This ensures that the randomly selected iterate lies in a sufficiently large ball with high probability. On this ball, continuity of F bounds the scaling factor. Combining the residual bound inside the ball with the small probability of lying outside it yields the desired result.

In the unconstrained case, matters simplify significantly as the residuals with respect to T and F are related by a strictly increasing map. The proof for Theorem 4.10 therefore follows straightforwardly from the bounds on T.

Lemma B.11. *We have the following inequality for $K \geq 1$:*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|x^k - z^{k+1}\|^2] \leq \frac{96\gamma(1+\delta)^2}{\theta\sqrt{K}}\left(\overline{V}_0(x^\star) + \frac{53}{12}\theta\gamma(1+\delta)^2\frac{2\sigma^2}{L_0^2}\right)$$

Proof. The proof follows analogously to the proof of (34), where in the proof of Theorem 4.5 we keep the term corresponding to $\mathbb{E}[\|x^k - z^{k+1}\|^2]$ and drop the nonpositive term corresponding to $\mathbb{E}[\|T(z^{k+1}) + n^k\|^2]$ instead. □

Lemma B.12. *Let random variable J be uniformly distributed on $\{0, \dots, K-1\}$ and independent from $\mathcal{F}_K$. Then,*

$$\mathbb{E}\left[\|T(z^{J+1}) + n^J\|^2\right] \leq \frac{18(1+\delta)^2}{7\theta\gamma\sqrt{K}}\left(\overline{V}_0(x^\star) + \frac{53}{12}(1+\delta)^2\theta\gamma\frac{2\sigma^2}{L_0^2}\right) \tag{36}$$

and, for any solution $x^\star \in S$,

$$\mathbb{E}[\|z^{J+1} - x^\star\|^2] \leq \left(2 + \frac{192\gamma(1+\delta)^2}{\theta\sqrt{K}}\right)\left(\overline{V}_0(x^\star) + \frac{53}{12}\theta\gamma(1+\delta)^2\frac{2\sigma^2}{L_0^2}\right). \tag{37}$$

Proof. Since J is uniform on $\{0, \dots, K-1\}$ and independent of $\mathcal{F}_K$, and $\|T(z^{k+1}) + n^k\|^2$ is $\mathcal{F}_K$-measurable for all $0 \leq k \leq K-1$,

$$\mathbb{E}[\|T(z^{J+1}) + n^J\|^2 \mid \mathcal{F}_K] = \frac{1}{K}\sum_{k=0}^{K-1}\|T(z^{k+1}) + n^k\|^2.$$

Thus, applying the tower property,

$$\mathbb{E}[\|T(z^{J+1}) + n^J\|^2] = \mathbb{E}[\mathbb{E}[\|T(z^{J+1}) + n^J\|^2 \mid \mathcal{F}_K]] = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|T(z^{k+1}) + n^k\|^2].$$

Invoking (34) immediately establishes (36). By the same argument, Lemma B.11 implies

$$\mathbb{E}[\|x^J - z^{J+1}\|^2] \leq \frac{96\gamma(1+\delta)^2}{\theta\sqrt{K}}\left(\overline{V}_0(x^\star) + \frac{53}{12}\theta\gamma(1+\delta)^2\frac{2\sigma^2}{L_0^2}\right). \tag{38}$$

To obtain (37), first take the total expectation in Proposition B.9 and drop the nonpositive terms to obtain $\mathbb{E}[V_{k+1}(x^\star)] \leq \mathbb{E}[V_k(x^\star)] + \frac{53}{12}\theta\gamma K^{-1}(1+\delta)^2\frac{2\sigma^2}{L_0^2}$ for $0 \leq k \leq K-1$. Hence, by induction, $\mathbb{E}[V_k(x^\star)] \leq \mathbb{E}[V_0(x^\star)] + \frac{53}{12}\frac{k\theta\gamma K^{-1}(1+\delta)^2 2\sigma^2}{L_0^2}$ for $0 \leq k \leq K$. Since $kK^{-1} \leq 1$, we have

$$\mathbb{E}[\|x^k - x^\star\|^2] \leq \mathbb{E}[V_k(x^\star)] \leq \overline{V}_0(x^\star) + \frac{53}{12}\frac{\theta\gamma(1+\delta)^2 2\sigma^2}{L_0^2}, \qquad 0 \leq k \leq K. \tag{39}$$

To convert this bound to (37), add and subtract x^J and use Young's inequality (22),

$$\|z^{J+1} - x^\star\|^2 \leq 2\|x^J - x^\star\|^2 + 2\|z^{J+1} - x^J\|^2.$$

Taking expectations and combining (39) with (38) yields the claimed result. □

B.9.1 Proof of Theorem 4.9

Proof. Starting from Lemma B.12 and applying Markov's inequality, we obtain for any $\varepsilon > 0$,

$$\mathbb{P}\{\|T(z^{J+1}) + n^J\| > \varepsilon\} \leq \frac{C_T}{\sqrt{K}\varepsilon^2}, \tag{40}$$

with $C_T \coloneqq \frac{18(1+\delta)^2}{7\theta\gamma}\Big(\overline{V}_0(x^\star) + \frac{53}{12}\frac{\theta\gamma(1+\delta)^2 2\sigma^2}{L_0^2}\Big)$. Our goal is to convert this to a bound on the residual with respect to F. For any $z \in \mathbb{R}^n$, $F(z) = (L_0 + L_1\|F(z)\|)T(z)$, and moreover, if $n \in N_C(z)$, then also $(L_0+L_1\|F(z)\|)n \in N_C(z)$, since $N_C(z)$ is a cone and $L_0+L_1\|F(z)\| > 0$. Let $\tilde{n}^J = (L_0 + L_1\|F(z^{J+1})\|)n^J$, then

$$\|F(z^{J+1}) + \tilde{n}^J\| = (L_0 + L_1\|F(z^{J+1})\|)\|T(z^{J+1}) + n^J\|. \tag{41}$$

Note that since the factor $L_0 + L_1\|F(z^{J+1})\|$ is a random variable, which need not be uniformly bounded, (36) does not suffice to bound $\|F(z^{J+1}) + \tilde{n}^J\|$.

To address this, let $R > 0$ and define the event $B_R = \{\|z^{J+1} - x^\star\| \leq R\}$. Moreover, define

$$f_R \coloneqq \sup_{z \in C}\big\{L_0 + L_1\|F(z)\| \mid \|z - x^\star\| \leq R\big\}.$$

Since C is closed, the set $\{z \in C : \|z - x^\star\| \leq R\}$ is compact, so by continuity of F, $0 < f_R < \infty$. Now, decomposing according to B_R,

$$\begin{aligned}\mathbb{P}\{\|F(z^{J+1}) + \tilde{n}^J\| > \varepsilon\} &\overset{(26)}{\leq} \mathbb{P}(B_R^c) + \mathbb{P}(\{\|F(z^{J+1}) + \tilde{n}^J\| > \varepsilon\} \cap B_R) \\ &\overset{(41)}{=} \mathbb{P}(B_R^c) + \mathbb{P}(\{(L_0 + L_1\|F(z^{J+1})\|)\|T(z^{J+1}) + n^J\| > \varepsilon\} \cap B_R).\end{aligned}$$

The first term is bounded by applying Markov's inequality to (37), and using the fact that $K \geq 1$,

$$\mathbb{P}(B_R^c) \leq \frac{C_z}{R^2}, \quad \text{with } C_z = \Big(2 + \frac{192\gamma(1+\delta)^2}{\theta}\Big)\Big(\overline{V}_0(x^\star) + \frac{53}{12}\frac{\theta\gamma(1+\delta)^2 2\sigma^2}{L_0^2}\Big). \tag{42}$$

For the second term, note that by definition, $B_R \subseteq \big\{L_0 + L_1\|F(z^{J+1})\| \leq f_R\big\}$. Therefore,

$$\begin{aligned}\mathbb{P}\Big(\big\{(L_0 + L_1\|F(z^{J+1})\|)\|T(z^{J+1}) + n^J\| > \varepsilon\big\} \cap B_R\Big) &\leq \mathbb{P}\{f_R\|T(z^{J+1}) + n^J\| > \varepsilon\} \\ &= \mathbb{P}\big\{\|T(z^{J+1}) + n^J\| > \frac{\varepsilon}{f_R}\big\} \\ &\overset{(40)}{\leq} \frac{C_T f_R^2}{\sqrt{K}\varepsilon^2}.\end{aligned} \tag{43}$$

Plugging in the bounds (42) and (43), we obtain

$$\mathbb{P}\left\{\|F(z^{J+1}) + \tilde{n}^J\| > \varepsilon\right\} \leq \frac{C_z}{R^2} + \frac{C_T f_R^2}{\sqrt{K}\varepsilon^2}. \tag{44}$$

Finally, fix $\eta \in (0,1)$. Choose $R_\eta > 0$ such that $\frac{C_z}{R_\eta^2} \leq \frac{\eta}{2}$, and choose $C_\eta > 0$ satisfying $C_\eta \geq \frac{2}{\eta}C_T f_{R_\eta}^2$. These choices are independent of K. Since $\tilde{n}^J \in N_C(z^{J+1})$, we have

$$\operatorname{dist}\big(0, (F + N_C)(z^{J+1})\big) \leq \|F(z^{J+1}) + \tilde{n}^J\|.$$

Applying (44) with $\varepsilon = \sqrt{C_\eta}K^{-1/4}$ therefore yields

$$\mathbb{P}\left\{\operatorname{dist}^2\big(0, (F + N_C)(z^{J+1})\big) > C_\eta K^{-1/2}\right\} \leq \frac{C_z}{R_\eta^2} + \frac{C_T f_{R_\eta}^2}{C_\eta} \leq \eta.$$

□

B.9.2 Proof of Theorem 4.10

Proof. Since $C = \mathbb{R}^n$, $N_C \equiv \{0\}$, $n^J = 0$ and applying Markov's inequality to Lemma B.12 yields, for any $\varepsilon > 0$,

$$\mathbb{P}\{\|T(z^{J+1})\| > \varepsilon\} \leq \frac{C_T}{\sqrt{K}\varepsilon^2}, \tag{45}$$

Since $r \mapsto \frac{r}{L_0 + L_1 r}$ is strictly increasing on $[0, \infty)$, it holds for all $z \in \mathbb{R}^n$, $\varepsilon > 0$ that

$$\|F(z)\| > \varepsilon \iff \|T(z)\| = \frac{\|F(z)\|}{L_0 + L_1\|F(z)\|} > \frac{\varepsilon}{L_0 + L_1\varepsilon}.$$

Thus, the claim follows directly from (45). □

C Experimental details and further results

C.1 Updates of compared methods

SEG has the following standard update: sample ξ^k and compute $G(x^k, \xi^k)$. Then update the predictor step:

$$z^k = x^k - \gamma G(x^k, \xi^k). \tag{PS}$$

Sample ζ^k and compute $G(z^k, \zeta^k)$. Then update the corrector step:

$$x^{k+1} = x^k - \omega G(z^k, \zeta^k). \tag{CS}$$

For SOG the update is simpler, since we only use one sample per iterate. Sample ξ^k and compute $G(x^k, \xi^k)$. Then, perform the update:

$$x^{k+1} = x^k - \omega G(x^k, \xi^k) - \gamma(G(x^k, \xi^k) - G(x^{k-1}, \xi^{k-1})) \tag{46}$$

C.2 Detailed setup and tuning procedures

Tuning procedure for the experiments in Sections 5.1 and 5.2 For the proposed method, we consider $L_0 \in \{0.1, 1, 10\}$, and fix $L_1 = 1$. $\alpha_k = 0.1$. We tune the step sizes with a two-stage grid search over 100 parameter configurations, each on three seeds. Half of the runs are divided over a logarithmic grid over $[10^{-4}, 1]^2$. The remaining half is used for refinement pass over a linear grid within the intervals defined by the coarse winner's neighboring grid points. Each candidate is evaluated using three sampling seeds and a maximum of 20000 oracle calls per run, keeping the problem instance and initial point fixed. The same seeds are reused across candidates. We score each candidate by averaging the last 100 recorded residuals within each run and then averaging across the three runs.

Table 2: Selected step sizes (γ, ω) for the experiments of Section 5.1.

κ_B	SEG (γ, ω)	SOG (γ, ω)	SPEPG (γ, ω)		
			$L_0 = 0.1$	$L_0 = 1$	$L_0 = 10$
10	(0.0464, 0.0464)	(0.0442, 0.0243)	(0.215, 0.0100)	(1, 0.0464)	(1, 0.364)
100	(0.0464, 0.0464)	(0.0442, 0.0243)	(0.364, 0.00953)	(1, 0.0442)	(1, 0.364)

Table 3: Selected step sizes (γ, ω) for the experiments of Section 5.2.

Oracle	SEG (γ, ω)	SOG (γ, ω)	SPEPG (γ, ω)		
			$L_0 = 0.1$	$L_0 = 1$	$L_0 = 10$
Finite sum	(0.523, 0.0785)	(0.364, 0.0785)	(0.215, 0.0100)	(1, 0.113)	(1, 0.523)
Student-t noise	(0.523, 0.0442)	(0.364, 0.0317)	(0.205, 0.00953)	(1, 0.0464)	(1, 0.215)

Experimental details on the setup in Section 5.2 The problems are generated as $A = U_A D_A U_A^\top, C = U_C D_C U_C^\top$, where U_A, U_C are orthogonal matrices obtained from QR decompositions. D_A, D_C are diagonal matrices with strictly positive entries, geometrically spaced between

1 and $1/\kappa_A$, or $1/\kappa_C$. The code ensures that the condition-number parameters are at least 1. B is generated similarly to the example in Section 5.1. We set $\kappa(A) = \kappa(B) = \kappa(C) = 10$.

In the finite-sum case, the stochastic oracle is generated as

$$F_i(x, y) = \begin{pmatrix} \alpha_i \sqrt{x^\top Ax}\, Ax + B_i y \\ \delta_i \sqrt{y^\top Cy}\, Cy - B_i^\top x \end{pmatrix}, \qquad i = 1, \dots, N,$$

where $\alpha_i = 1 + \sigma_\alpha \xi_i$, $\delta_i = 1 + \sigma_\alpha \eta_i$, $B_i = B + \sigma_B G_i$. For $i = 1, \dots, N/2$, we independently draw $\xi_i, \eta_i \sim \mathcal{U}[-1, 1]$ and matrices G_i with independent entries $(G_i)_{jk} \sim \mathcal{N}(0, 1/d)$. The remaining components are constructed to be their negatives: $\xi_{i+N/2} = -\xi_i$, $\eta_{i+N/2} = -\eta_i$, $G_{i+N/2} = -G_i$. We then apply a common random permutation to the indices. Each oracle evaluation independently samples an index i uniformly from $\{1, \dots, N\}$ and returns $F_i(x, y)$. In our experiments, we set $N = 300$, $\sigma_\alpha = 0.3$, and $\sigma_B = 0.5$.

For the Student-t noise, we set $\nu = 3$, $d = 30$, and $\sigma_B = 0.5$. Since $\nu > 1$, the perturbations have zero mean and the oracle remains unbiased. Note that in both noise models, every oracle realization vanishes at the solution $(x, y) = (0, 0)$. For both experiments, we set $d = 30$.

Tuning procedure for the WGAN GP experiment. For our proposed method we choose $L_0 = 0.1$ and $L_1 = 1$, $\alpha = 0.1$, $\omega = \theta\alpha$. We tune the methods using as metric the mean of the SWD over the last five checkpoints taken at $90\%, 92.5\%, 95\%, 97.5\%$, and 100% of the training process. The initial tuning phase consists of 100 runs (10×10 for γ and ω) over 3 seeds with a logarithmic grid and it is followed by a refinement round with a linear grid for 100 more runs. The initial grids for SEG and SOG are $\gamma \in [10^{-7}, 10^{-1}]$ and $\omega \in [10^{-6}, 10^{-1}]$. For SPEPG, $\gamma \in [10^{-3}, 5]$ and $\theta \in [5 \times 10^{-4}, 5]$.

Table 4: Selected hyperparameters for the WGAN GP experiments.

s_D	SEG (γ, ω)	SOG (γ, ω)	SPEPG (γ, θ)
1	(1×10^{-3}, 7.7×10^{-3})	(4.7×10^{-3}, 7.8×10^{-4})	(2.17, 3.96×10^{-1})
4	(2.1×10^{-4}, 1.6×10^{-4})	(1.2×10^{-5}, 1.27×10^{-3})	(7.53×10^{-1}, 7.54×10^{-1})

C.3 Neural optimal transport

We consider the problem of estimating the optimal transport (OT) map $T_\star$ between two probability measures P and Q with transport cost $c(x, y) = h(x - y)$ for a strictly convex $h \in \mathcal{C}^1(\mathbb{R}^n)$, i.e. the setting of Gangbo & McCann (1996). In this case, the map has the form $T_\star(x) = x - \nabla h^*(\nabla g(x))$ where h^* is the convex conjugate of h and g is a c-concave (Gangbo & McCann, 1996, Definition 1.1) Kantorovich potential. We choose specifically $h(x) = \sum_{i=1}^{16} \cosh(x_i)$ to go beyond the standard quadratic cost and make the problem potentially harder, and the c-concave potential $g(x) = \sum_{i=1}^{16} \left(a_i \log\cosh((Rx)_i) - \frac{\lambda}{4}(Rx)_i^4\right)$ where R is a fixed orthogonal matrix, $a_i = 1 - \mu_i$, $\mu_i = \varepsilon(\mu_{\max}/\varepsilon)^{(i-1)/15}$, $\varepsilon = 0.05$ and $\mu_{\max} = 0.8$.

We generate the problem by setting $P = \mathcal{N}(0, I_{16})$, computing the exact form of the transport map and getting the target measure via the pushforward, $Q = T_{\star\#}P$. We consider the saddle formulation of the OT problem

$$\inf_T \sup_f \mathbb{E}_P[c(X, T(X)) - f(T(X))] + \mathbb{E}_Q[f(Y)],$$

which is obtained by using the definition of c-conjugates. We parametrize the transport map by T_ϕ and the potential by f_ψ with neural networks, similarly to Korotin et al. (2022) and solve the associated minimax problem. We set $T_\phi(x) = x + r_\phi(x)$ where r_ϕ is a fully connected network with two $\tanh$ hidden layers: $16 \to 64 \to 64 \to 16$ while for f the layers are of the form: $16 \to 64 \to 64 \to 1$. We estimate the transport map and use as a metric the MSE $\frac{1}{N}\sum_{i=1}^{N} \|T_\phi(x^i) - T_\star(x^i)\|^2$, $x_i \sim P$.

We set minibatch size of 256. For $\lambda = 0$ we give a budget of 10000 minibatch oracle calls while for $\lambda = 0.1$, 20000. Note that all the methods see an equal amount of data, but SEG and SPEPG perform half the iterations of SOG. We run the experiment ten times and present the results in Table 6. It can

be seen that in terms of MSE SPEPG performs worse on the easier task with $\lambda = 0$ but much better than the rest of the methods for $\lambda = 0.1$. The performance in the $\lambda = 0$ case seems to also depend on the specific instance in some cases performing better and in others much worse. SPEPG also has the best SWD from all the compared methods.

Tuning. For our proposed method we choose $L_0 = 0.1$ and $L_1 = 1$, $\alpha = 0.1$, $\omega = \theta\alpha$. We use three seeds for tuning the hyperparameters. We tune the methods using as metric the mean of the MSE over the last five checkpoints. The initial tuning consists of 100 runs (10×10 for γ and ω) over 3 seeds with a logarithmic grid and it is followed by a refinement round with a linear grid for 100 more runs. The chosen parameters are presented in Table 5.

Table 5: Selected hyperparameters for the neural OT experiments.

λ	SEG (γ, ω)	SOG (γ, ω)	SPEPG (γ, θ)
0	$(4.64\times10^{-4},\ 2.15\times10^{-1})$	$(3.64\times10^{-3},\ 1.52\times10^{-1})$	$(3.48,\ 9.57\times10^{-1})$
0.1	$(1.29\times10^{-2},\ 1.29\times10^{-2})$	$(4.26\times10^{-2},\ 2.35\times10^{-3})$	$(2.741,\ 7.21\times10^{-1})$

Table 6: Final held-out performance on the synthetic neural OT benchmark. "Last-5 MSE" denotes the mean map MSE over the final five evaluation checkpoints, while "Final MSE" denotes the map MSE at the end of training. Values are reported as mean ± standard deviation across runs. For $\lambda = 0.1$, **SEG was successful on** 5 **of the** 10 **runs**; its reported statistics are computed over those five successful runs.

λ	Method	Last-5 MSE ↓	Final MSE ↓	Final SWD ↓
0	SEG	0.04094 ± 0.00427	0.03614 ± 0.00412	0.02836 ± 0.00530
	SOG	$\mathbf{0.02605 \pm 0.00170}$	$\mathbf{0.02282 \pm 0.00193}$	0.02309 ± 0.00281
	SPEPG	0.06064 ± 0.03642	0.07151 ± 0.05210	$\mathbf{0.02033 \pm 0.00222}$
0.1	SEG	0.97276 ± 0.20715	0.85461 ± 0.13734	0.06242 ± 0.01867
	SOG	1.87128 ± 0.02249	1.86712 ± 0.02475	0.06527 ± 0.00305
	SPEPG	$\mathbf{0.62409 \pm 0.07215}$	$\mathbf{0.62707 \pm 0.08187}$	$\mathbf{0.03962 \pm 0.00247}$

C.4 GAN TRAINING ON MNIST

We compare the methods for training GANs (Goodfellow et al., 2020) to generate MNIST images (Deng, 2012). The MNIST training set is partitioned into 50000 training examples and 10000 validation examples using a fixed random split. We use a budget of 2048 minibatch oracle calls with batch size 528 and compare using Fréchet distance between Gaussian approximations to real and generated classifier features (mFID). For G_ψ the generator and D_ϕ the discriminator's logit, $l(t) \coloneqq \ln(1 + \exp(t))$, the two players minimize their losses:

$$L_G(\psi, \phi) = \mathbb{E}_z[l(-D_\phi(G_\psi(z)))],$$

$$L_D(\psi, \phi) = \mathbb{E}_u[l(-D_\phi(u))] + \mathbb{E}_z[l(D_\phi(G_\psi(z)))] + \tfrac{\lambda_{\text{out}}}{2}(\mathbb{E}_u[D_\phi(u)^2] + \mathbb{E}_z[D_\phi(G_\psi(z))^2]) + \tfrac{\lambda_{\text{wd}}}{2}\|\phi\|^2,$$

where we set $\lambda_{\text{out}} = 10^{-3}$ and $\lambda_{\text{wd}} = 10^{-6}$. The operator is simply defined as

$$F(\psi, \phi) = \begin{pmatrix} \nabla_\psi L_G(\psi, \phi) \\ \nabla_\phi L_D(\psi, \phi) \end{pmatrix}.$$

The generator and discriminator architectures can be found in the supplementary material. The results are presented in Table 8 and a visualization of the outputs can be found in Figure 4. It can be seen that our proposed method achieves the lowest mFID.

Tuning. For our proposed method we choose $L_0 = 0.1$ and $L_1 = 1$, $\alpha = 0.1$. We use 3 seeds for tuning the hyperparameters and choose the best setting by averaging the mFID across the three

seeds and five late training checkpoints. For this experiment due to the larger scale we followed a different tuning procedure: for all the methods we used a common range for $\gamma \in [10^{-4}, 0.3]$ and kept the ratio ω/γ in a range $[0.1, 4]$. In the first phase we use 32 configurations for each method obtained from a logarithmically spaced grid, which we then refine with 32 more configurations after we obtain the best parameters from the first run. The selected parameters can be found in Table 7.

Table 7: Selected hyperparameters for the GAN training on MNIST experiments.

SEG (γ, ω)	SOG (γ, ω)	SPEPG (γ, ω)
$(7.7\times10^{-3},\ 2.57\times10^{-2})$	$(8.1\times10^{-3},\ 2.75\times10^{-2})$	$(3.12\times10^{-2},\ 2.77\times10^{-1})$

Table 8: Final MNIST generation performance after 2048 stochastic-oracle calls. Results are reported as mean $\pm$ standard deviation across held-out seeds. Lower mFID is better, while higher digit-classifier score is better. **SEG failed numerically on 2 of the 5 seeds**; consequently, its reported mean and standard deviation are computed only over the 3 successful runs. SOG and SPEPG completed all 5 runs.

Method	mFID $\downarrow$	Digit-classifier score $\uparrow$
SEG	266.78 ± 10.39	5.793 ± 0.052
SOG	82.77 ± 11.05	6.871 ± 0.057
SPEPG	$\mathbf{48.12 \pm 4.47}$	$\mathbf{7.407 \pm 0.046}$



(a) SEG (b) SOG (c) SPEPG

Figure 4: Representative MNIST samples generated by SEG, SOG, and SPEPG after 2048 stochastic-oracle calls for one test seed. Each panel shows 64 samples from the final generator. All three methods produce recognizable digit-like images, although the qualitative sample quality differs across optimization methods. SEG exhibits more irregular and distorted digit shapes, while SOG and especially SPEPG produce cleaner and more consistently recognizable samples.