HIGH PROBABILITY AND RISK-AVERSE GUARANTEES FOR STOCHASTIC SADDLE POINT PROBLEMS

Yassine Laguel
Department of Management Science and Information Systems
Rutgers Business School, Rutgers University
Piscataway, NJ, USA.
yassine.laguel@rutgers.edu

Necdet Serhat Aybat
Department of Industrial and Manufacturing Engineering
Pennsylvania State University
University Park, PA, USA.
nsa10@psu.edu

Mert Gürbüzbalaban
Department of Management Science and Information Systems
Rutgers Business School, Rutgers University
Piscataway, NJ, USA.
mert.gurbuzbalaban@rutgers.edu

March 05, 2023

ABSTRACT

We consider strongly-convex-strongly-concave (SCSC) saddle point (SP) problems which frequently arise in many applications from distributionally robust learning to game theory and fairness in machine learning. We focus on the recently developed stochastic accelerated primal-dual algorithm (SAPD), which admits optimal complexity in several settings as an accelerated algorithm. We provide high probability guarantees for convergence to a neighborhood of the saddle point that reflects accelerated convergence behavior. We also provide an analytical formula for the limit covariance matrix of the iterates for SCSC quadratic problems under Gaussian perturbations. This allows us to develop lower bounds for quadratic problems that show that our analysis is tight. We also provide a risk-averse convergence analysis characterizing the "Conditional Value at Risk" and the "Entropic Value at Risk" of the distance to the saddle point, highlighting the trade-offs between the bias and the risk associated to an approximate solution.

1 Introduction

We consider strongly convex/strongly concave (SCSC) saddle point problems of the form:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} \mathcal{L}(x,y) \triangleq f(x)+\Phi(x,y)-g(y), \tag{1.1}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are finite-dimensional Euclidean spaces, $f:\mathcal{X}\to\mathbb{R}$ and $g:\mathcal{Y}\times\mathbb{R}$ are closed, strongly convex functions, and $\Phi:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ is a smooth convex-concave function – see Assumption 1 for details.

SCSC problems can arise in many applications and contexts. In unconstrained and constrained optimization problems, saddle-point formulations arise naturally when the problems are formulated based on the Lagrangian duality. Furthermore, the SP formulation in (1.1) encompasses many key problems such as *robust optimization* [3] – here $\mathcal{Y}$ represents the uncertainty set from which nature (adversary) picks an uncertain model parameter y, and the objective is to choose

$x \in \mathcal{X}$ that minimizes the worst-case cost $\max_{y\in\mathcal{Y}} \mathcal{L}(x, y)$, i.e., a two-player zero-sum game. Other applications where SCSC problems arise include but are not limited to *supervised learning* with non-separable regularizers (where $\Phi(x, y)$ may not be bilinear) [25], *unsupervised learning* [25] and various *image processing* problems, e.g., denoising, [6].

In this work, we are interested in stochastic SCSC problems where the partial gradients $\nabla_x \Phi(x, y)$ and $\nabla_y \Phi(x, y)$ are not deterministically available, but instead we postulate access to their stochastic estimate. Such a setting arises frequently in large-scale optimization and machine learning applications where the gradients are estimated from either streaming data or from random samples of data (see e.g. [40, 15, 5]). In this work, we focus on stochastic first-order (FO) methods that relies on stochastic estimates of the gradient information which have been the leading computational approach for computing low-to-medium-accuracy solutions because of their cheap iterations and mild dependence on the problem dimension and data size.

Relevant work. Stochastic algorithms generate a sequence of primal and dual iterate pairs $z_k = (x_k, y_k) \in \mathcal{X} \times \mathcal{Y} \triangleq \mathcal{Z}$ starting from an initial point $(x_0, y_0) \in \mathbf{dom}\, f \times \mathbf{dom}\, g \triangleq Z$. Two popular metrics to access the quality of a random solution $(\hat{x}, \hat{y})$ returned by a stochastic optimization algorithm are the *expected gap* and the *expected distance squared* defined as

$$\mathcal{G}(\hat{x}, \hat{y}) \triangleq \mathbb{E}[\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \{\mathcal{L}(\hat{x}, y) - \mathcal{L}(x, \bar{y})\}], \quad \mathcal{D}(\hat{x}, \hat{y}) \triangleq \mathbb{E}[\|\hat{x} - x_*\|^2 + \|\hat{y} - y_*\|^2], \tag{1.2}$$

respectively where (x_*, y_*) is the saddle point which is unique due to the strong convexity of f and g. The iteration complexity in these two metrics depend naturally on the block Lipschitz constants L_{xx}, L_{xy} and L_{yy}, i.e. Lipschitz constants of $\nabla_x \Phi(\cdot, y)$, $\nabla_y \Phi(x, \cdot)$ and $\nabla_y \Phi(\cdot, y)$ as well as on the strong convexity constants μ_x and μ_y of the functions f and g. In particular, [11] shows that a multi-stage variant of Stochastic Gradient Descent Ascent (SGDA) algorithm achieves the guarantee $\mathcal{D}(x_k, y_k) \leq \epsilon$ after $k = \mathcal{O}(\kappa^2 \ln(1/\epsilon) + \frac{\delta^2}{\mu\epsilon})$ iterations where $\delta^2 = \max(\delta_x^2, \delta_y^2)$, δ_x^2 and δ_y^2 are bounds on the variance of the stochastic gradients with respect to x and y respectively while $\mu := \min(\mu_x, \mu_y)$ are $L := \max(L_{xx}, L_{xy}, L_{yy})$ are the worst-case strong convexity and Lipschitz constants and $\kappa = L/\mu$ is defined as the *condition number*. SGDA consists of Jacobi-style updates in the sense that stochastic gradient descent and ascent steps are taken simultaneously. In [36], it is shown that for deterministic SCSC problems, if gradient descent ascent (GDA) is modified with Gauss-Seidel-style updates where the dual variable is updated after the primal variable in an alternating fashion, than an accelerated convergence rate (where iteration complexity scales with κ instead of κ^2) can be obtained. However, as discussed in [36], this comes with the price that Gauss-Seidel style updates greatly complicate the analysis because every iteration of an alternating algorithm is a composition of two half updates. Simultaneous Jacobi-style updates are easier to analyze in general and can be viewed as solving a structured a variational inequality problem where many existing techniques directly apply [14, 7].

There are also other algorithms for stochastic SCSC problems. The authors in [11] show that the *Stochastic Optimistic Gradient Descent Ascent* (OGDA) algorithm achieves an iteration complexity of $\mathcal{O}(\kappa \ln(1/\epsilon) + \frac{\delta^2}{\mu\epsilon})$ in expected distance squared. In [38], a stochastic accelerated primal-dual (SAPD) algorithm which consists of Gauss-Seidel type updates is proposed. SAPD achieves $\mathcal{O}\Big(\Big(\frac{L_{xx}}{\mu_x} + \frac{L_{yx}}{\sqrt{\mu_x\mu_y}} + \frac{L_{yy}}{\mu_y} + \Big(\frac{\delta_x^2}{\mu_x} + \frac{\delta_y^2}{\mu_y}\Big)\frac{1}{\epsilon}\Big)\log(\frac{1}{\varepsilon})\Big)$, in a weighted expected distance squared metric. This complexity is optimal for bilinear problems. To our knowledge, SAPD is also the fastest single-loop algorithm for solving stochastic smooth SCSC problems that are non-bilinear.

Despite these results that provide guarantees in expectation based on the metrics (1.2), high probability bounds for the iterates of saddle-point problems are relatively much less studied. In particular results provided in the metrics (1.2) do not allow us to control tail events, i.e. the expected gap and distance can be smaller than a given target threshold ε, but the iterates can in principle still be arbitrarily far away from the saddle point with a non-zero probability. In this context, high probability guarantees are key in the sense that they allow us to control tail probabilities and quantify how many iterations are needed for the iterates to be in a neighborhood of the saddle point with a given probability level $p \in (0, 1)$.

While high-probability guarantees are available in the optimization setting for stochastic gradient descent-like methods [18, 28, 9], they are more limited in the SP setting. Among existing results, in [35], it is shown that the expected gap $\mathcal{G}(x_k, y_k) \leq \varepsilon$ with probability at least $p \in (0, 1)$ after $\mathcal{O}\left(\frac{1}{\varepsilon}\log(\frac{1}{1-p}) + \frac{\delta^2}{\mu\varepsilon}\log(\frac{1}{1-p})\right)$ iterations for possibly non-smooth, SCSC problems. In [34], high probability bounds are given for online algorithms applied to a stochastic saddle point problem where the objective is time-varying and is revealed in a sequential manner and and the data distribution over which stochastic gradients are estimated depends on the decision variables. However, these high-probability guarantees are obtained for non-accelerated algorithms; therefore the high probability bounds do not enjoy accelerated decay of the dependence to initialization with a linear rate that scales linearly with the condition number.

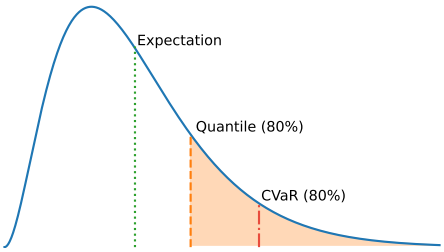

Figure 1: Representation of expectation, VaR (quantiles) and CVaR of a gamma distribution with shape parameters $k = 3$ and scale parameter $\theta = 5$.

Contributions. In this paper, we study the speed-accuracy trade-off of stochastic primal-dual methods from a risk-averse perspective. Our main focus is the SAPD method, for which we provide high-probability bounds where the dependence to the initialization decays in an accelerated manner with an accelerated linear rate. We also provide tight upper-bounds with respect to several risk measures illustrated in Figure 1, including the Value at Risk (corresponding to a high probability bound), the Conditional Value at Risk (CVaR), and the Entropic Value at Risk (EVaR). Our convergence analysis also captures the performance of the standard SGDA which is studied apart in our Appendix.

While our analysis builds upon concentration results developed in [18], and already utilized in a variety of works on stochastic first order methods [8, 39, 22], we address two technical challenges specific to SAPD: (i) the *ergodic* nature of the analysis based on which SAPD was previously shown to converge requires to adapt accordingly the recursive concentration inequality [18] and (ii) the Gauss-Seidel iteration based on which SAPD was developped significantly complicates the analysis, as one may observe in comparison to the the analysis of SGDA in the Appendix.

We complement our results with an in-depth analysis of the performance of SAPD on quadratic problems subject to Gaussian perturbations. This contribution, based on orthogonal arguments, is key to show the tightness of our general analysis.

Notations. Throughout this manuscript, $\mathcal{X} = \mathbb{R}^{\mathrm{n}}$ and $\mathcal{Y} = \mathbb{R}^{\mathrm{m}}$ denote finite dimensional vector spaces equipped with the Euclidean norm $\|u\| \triangleq \langle u, u\rangle^{\frac{1}{2}}$, and $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$. We adopted $\mathbb{Z}_{++}$ for counting numbers and $\mathbb{Z}_{+} = \mathbb{Z}_{++} \cup \{0\}$. For $A \in \mathbb{R}^{n\times n}$, $\|A\|_F = (\sum_{i,j=1}^{n} A_{i,j}^2)^{1/2}$ denotes the *Frobenius norm* of A and $\rho(A)$ denotes its *spectral radius*. We recall that for all A, $\rho(A) \leq \|A\|_F$. For any convex set $C \in \mathcal{X}$, $\mathcal{I}_C$ denotes the indicator function of C, i.e., $\mathcal{I}_C(x) = 0$ if $x \in C$, and equal to $+\infty$ otherwise. For a given proper, closed and convex function $\varphi\colon \mathcal{X} \to \mathbb{R}\cup\{+\infty\}$, $\mathrm{prox}_{\varphi}(\cdot)$ denotes the associated *proximal operator*: $x \mapsto \arg\min_{u\in\mathcal{X}} \varphi(u) + \frac{1}{2}\|u - x\|^2$. By strong convexity/strong concavity, the problem in (1.1) admits a unique saddle point [10], $\mathrm{z}^\star \triangleq (\mathrm{x}^\star, \mathrm{y}^\star)$ which satisfies:

$$\mathcal{L}(\mathrm{x}^\star, y) \leq \mathcal{L}(\mathrm{x}^\star, \mathrm{y}^\star) \leq \mathcal{L}(x, \mathrm{y}^\star) \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}. \tag{1.3}$$

We use the Landau notations o, O, and Θ to describe the asymptotic behavior of functions. For $u \in \mathbb{R}\cup\{\pm\infty\}$, a function $f(x) = o(g(x))$ in a neighborhood of u if $\frac{f(x)}{g(x)} \to 0$ as $x \to u$. $f(x) = \mathcal{O}(g(x))$ if there exist positive constants C such that $|f(x)| \leq C|g(x)|$ in some neighborhood of u. Finally $f(x) = \Theta(g(x))$, if $f(x) = \mathcal{O}(g(x))$ and $g(x) = \mathcal{O}(f(x))$. Given an m-dimensional vector $v = [v_1, v_2, \ldots, v_m]^T$, $\mathrm{Diag}(v)$ denotes the $m \times m$ matrix whose diagonal is the v vector.

2 Technical Background and Preliminaries

2.1 Stochastic Accelerated Primal-Dual (SAPD) Method

SAPD, displayed in Algorithm 1, is a stochastic accelerated primal-dual method developed in [38] which uses stochastic estimates $\tilde{\nabla}_x \Phi$ and $\tilde{\nabla}_y \Phi$ of the partial gradients $\nabla_x \Phi$ and $\nabla_y \Phi$. SAPD extends the accelerated primal dual method (APD) proposed in [16] to the stochastic setting. Given primal and dual stepsizes τ and σ and (number of iterations) horizon N, SAPD uses proximal-gradient-type updates while applying momentum averaging to the partial gradients with respect to y. It achieves accelerated rates and an optimal bias-variance trade-off for the SCSC scenario when stochastic gradients admit bounded variance [38]. When the coupling function is bilinear, it reduces to the Chambolle-Pock algorithm given in [6] for the deterministic setting.

Algorithm 1 SAPD Algorithm

Require: Parameters τ, σ, θ. Starting point (x_0, y_0). Horizon n.

Initialize:

$x_{-1} \leftarrow x_0, \quad y_{-1} \leftarrow y_0, \quad \tilde{\mathrm{q}}_0 \leftarrow \mathbf{0}$

for $k \geq 0$ **do**

$\tilde{\mathrm{s}}_k \leftarrow \widetilde{\nabla}_{\mathrm{y}} \Phi(x_k, y_k, \omega_k^y) + \theta\, \tilde{\mathrm{q}}_k$ ▷ Momentum averaging

$y_{k+1} \leftarrow \mathrm{prox}_{\sigma g}(y_k + \sigma\, \tilde{\mathrm{s}}_k)$

$x_{k+1} \leftarrow \mathrm{prox}_{\tau f}(x_k - \tau\, \widetilde{\nabla}_{\mathrm{x}} \Phi(x_k, y_{k+1}, \omega_k^x))$

$\tilde{\mathrm{q}}_{k+1} \leftarrow \widetilde{\nabla}_{\mathrm{y}} \Phi(x_{k+1}, y_{k+1}, \omega_{k+1}^y) - \widetilde{\nabla}_{\mathrm{y}} \Phi(x_k, y_k, \omega_k^y)$

return (x_n, y_n)

For the convergence analysis of SAPD, we next introduce the following assumption which basically says that the coupling function Φ is smooth. Such an assumption is standard for the analysis of first-order methods, see e.g. [23, 14, 37].

Assumption 1. *Let $f\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ and $g\colon \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ be two functions with convexity moduli $\mu_x > 0$ and $\mu_y > 0$, respectively. The coupling function $\Phi\colon \mathbb{R}^{\mathrm{d}} \times \mathbb{R}^{\mathrm{d}} \to \mathbb{R}$ is continuously differentiable on an open set containing $\mathbf{dom}\, f \times \mathbf{dom}\, g$ such that:*

(i) For all $y \in \mathbf{dom}\, g$, $\Phi(\cdot, y)$ is convex on $\mathbf{dom}\, f$.

(ii) For all $x \in \mathbf{dom}\, f$, $\Phi(x, \cdot)$ is concave on $\mathbf{dom}\, g$.

(iii) There exists $\mathrm{L}_{\mathrm{xx}}, \mathrm{L}_{\mathrm{yy}} \geq 0$ and $\mathrm{L}_{\mathrm{xy}}, \mathrm{L}_{\mathrm{yx}} > 0$ satisfying for all $(x, y), (\bar{x}, \bar{y}) \in \mathbf{dom}\, f \times \mathbf{dom}\, g$:

$$\|\nabla_{\mathrm{x}} \Phi(x, y) - \nabla_{\mathrm{x}} \Phi(\bar{x}, \bar{y})\| \leq \mathrm{L}_{\mathrm{xx}} \|x - \bar{x}\| + \mathrm{L}_{\mathrm{xy}} \|y - \bar{y}\|,$$
$$\|\nabla_{\mathrm{y}} \Phi(x, y) - \nabla_{\mathrm{y}} \Phi(\bar{x}, \bar{y})\| \leq \mathrm{L}_{\mathrm{yx}} \|x - \bar{x}\| + \mathrm{L}_{\mathrm{yy}} \|y - \bar{y}\|.$$

Following the literature on stochastic saddle-point algorithms [24, 20, 7], we assume that only (noisy) stochastic estimates $\widetilde{\nabla}_{\mathrm{y}} \Phi(x_k, y_k, \omega_k^y), \widetilde{\nabla}_{\mathrm{x}} \Phi(x_k, y_{k+1}, \omega_k^x)$ of the partial gradients $\nabla_{\mathrm{y}} \Phi(x_k, y_k), \nabla_{\mathrm{x}} \Phi(x_k, y_{k+1})$ are available, where ω_k^x, ω_k^y are random variables that are being revealed sequentially as we discuss next. First, we introduce some notation: Let $(\omega_k^x)_{k \geq 0}$, $(\omega_k^y)_{k \geq 0}$ be two sequences of random variables revealed sequentially, in the order:

$$\omega_0^y \to \omega_0^x \to \omega_1^y \to \omega_1^x \to \omega_2^y \to \cdots,$$

and let $(\mathcal{F}_k^y)_{k \geq 0}$ and $(\mathcal{F}_k^x)_{k \geq 0}$ denote the associated filtration:

$$\mathcal{F}_0^y = \sigma(\omega_0^y), \quad \mathcal{F}_0^x = \sigma(\omega_0^y, \omega_0^x),$$
$$\mathcal{F}_k^y = \sigma(\mathcal{F}_{k-1}^x, \sigma(\omega_k^y)), \quad \mathcal{F}_k^x = \sigma(\mathcal{F}_k^y, \sigma(\omega_k^x)), \quad \forall\, k \geq 1.$$

For any $k \geq 0$, we introduce the following random variables to represent the gradient noise:

$$\Delta_k^y \triangleq \widetilde{\nabla}_{\mathrm{y}} \Phi(x_k, y_k, \omega_k^y) - \nabla_{\mathrm{y}} \Phi(x_k, y_k), \quad \Delta_k^x \triangleq \widetilde{\nabla}_{\mathrm{x}} \Phi(x_k, y_{k+1}, \omega_k^x) - \nabla_{\mathrm{x}} \Phi(x_k, y_{k+1}).$$

Often times, stochastic gradients are assumed to be unbiased with a bounded variance conditional on the history of the iterates. Such an assumption is standard in the study of stochastic optimization algorithms and stochastic approximation theory [17] and frequently arises in the context of stochastic gradient methods that estimate the gradients from randomly sampled subsets of data [5].

Assumption 2. *For any $k \geq 0$, the gradient noise satisfies*

$$\mathbb{E}\left[\Delta_k^y | \mathcal{F}_{k-1}^x\right] = 0, \quad \mathbb{E}\left[\Delta_k^x | \mathcal{F}_k^y\right] = 0.$$

Assumption 3. *For any $k \geq 0$, there exists scalars $\delta_x, \delta_y > 0$ such that*

$$\mathbb{E}\left[\|\Delta_k^y\|^2 | \mathcal{F}_{k-1}^x\right] \leq \delta_y^2, \quad \mathbb{E}\left[\|\Delta_k^x\|^2 | \mathcal{F}_k^y\right] \leq \delta_x^2.$$

Based on Assumptions 1, 2, 3; [38] established bounds for the expected squared distance to the saddle point $\mathbb{E}\left[\|z_k - z_*\|^2\right]$. Such upper bounds consists of a "*bias*" term (that decays exponentially fast characterizing how fast the initial conditions are forgotten) and a *variance* term that scales with the noise variance δ_x^2, δ_y^2. In this setting,

the rate of decay of the bias coincides with the "convergence rate" of the underlying SAPD algorithm without noise (i.e. when $\delta_x^2 = \delta_y^2 = 0$).

In this paper, our focus is to obtain high probability guarantees as well as bounds on the risk of the distance to the saddle point $\|z_k - z_*\|$. For quantifying risk, we will resort to ϕ-divergence-based risk measures borrowed from the risk measure theory [4], including CVaR EVaR and χ^2-divergence. We introduce the following "light-tail" assumption, which basically says that the gradient noise is norm-subGaussian. Random vectors with Norm-subGaussian distribution were introduced in [19], and encompass a large class of random vectors including subGaussian random vectors. In the rest of the paper, through all our results, we will assume that Assumption 4 holds in addition to Assumptions 1 and 2.

Definition 2.1. A random vector $X\colon \Omega \to \mathbb{R}^{\mathrm{d}}$ is *norm-subGaussian* with proxy σ, denoted by $X \in nSG(\sigma)$, if

$$\mathbb{P}\left[\|X - \mathbb{E}[X]\| \ \geq t\right] \leq 2e^{\frac{-t^2}{2\sigma^2}}, \quad \forall t \in \mathbb{R}\,.$$

Assumption 4. *For any $k \geq 0$ the random variables Δ_k^x and Δ_k^y are conditionally norm-subGaussians with respective proxy parameters $\delta_x, \delta_y > 0$. More precisely, for all $t \geq 0$, we almost surely have*

$$\mathbb{P}[\|\Delta_k^y\| \geq t | F_{k-1}^x] \leq 2e^{\frac{-t^2}{2\delta_y^2}}, \quad \mathbb{P}[\|\Delta_k^x\| \geq t | F_k^y] \leq 2e^{\frac{-t^2}{2\delta_x^2}}.$$

Such subGaussian noise assumptions are common in large-scale stochastic optimization [27, 12, 18]. In machine learning applications, where stochastic gradients are often estimated on sampled batches, noisy estimates typically behave as Gaussians for moderately high sample sizes, as a consequence of the central limit theorem [26]. Furthermore, they also can be volontarily imposed from perturbations set by practitionners for privacy reasons [21, 32] We also recall some basic properties of norm-subGaussian random vectors which will be used repeatedly in the proof of our results.

2.2 Elementary Properties of Norm-subGaussian Vectors

In this section, we recall elementary properties of norm-subGaussian vectors. Proofs, which follow from standard arguments that can be found in textbooks such as [8, 6], are given in Section 6 of the Appendix for the sake of completeness. First, note that given arbitrary $\alpha > 0$ and $X\colon \Omega \to \mathbb{R}^{\mathrm{d}}$ such that $X \in nSG(\sigma)$ for some $\sigma > 0$, we immediately have the following implication:

$$X \in nSG(\sigma) \implies \alpha X \in nSG(\alpha\sigma)\,. \tag{2.1}$$

For instance, $X : \Omega \to \mathbb{R}^d$ is norm-subGaussian when X is subGaussian, or it is bounded, i.e., $\exists B > 0$ such that $\|X\| \leq B$ with probability 1. As remarked in [19, Lemma 3], the squared norm of a norm-subGaussian vector is sub-Exponential.

Definition 2.2. A random variable $\mathcal{U}\colon \Omega \to \mathbb{R}$ is subExponential with proxy $K > 0$ if it satisfies

$$\mathbb{E}\left[e^{\lambda|\mathcal{U}|}\right] \leq e^{\lambda K}, \quad \forall \lambda \in [0, \frac{1}{K}].$$

In particular, if we take $U = \|X\|^2$, with $X \in nSG(\sigma)$, the following lemma shows that U is subExponential with proxy $K = 8\sigma^2$.

Lemma 2.1. *Let $X \in nSG(\sigma)$ be such that $\mathbb{E}[X] = 0$. Then, for any $\lambda \in \left[0, \frac{1}{4\sigma^2}\right]$,*

$$\mathbb{E}\left[e^{\lambda\|X\|^2}\right] \leq 2e^{4\lambda\sigma^2} - 1 \leq e^{8\lambda\sigma^2}. \tag{2.2}$$

Proof of Lemma 2.1. We follow standard arguments from [33]. First note that, for any $k \in \mathbb{Z}_{++}$, we have

$$\begin{aligned}
\mathbb{E}[\|X\|^k] &= \int_{t=0}^{+\infty} \mathbb{P}[\|X\|^k \geq t]\,\mathrm{d}\,t = \int_{t=0}^{+\infty} \mathbb{P}[\|X\| \geq t^{1/k}]\,\mathrm{d}\,t \\
&\leq 2\int_{t=0}^{\infty} e^{-t^{\frac{2}{k}}/(2\sigma^2)}\,\mathrm{d}\,t \\
&= k(2\sigma^2)^{\frac{k}{2}} \int_{u=0}^{\infty} e^{-u} u^{\frac{k}{2}-1}\,\mathrm{d}\,u = k(2\sigma^2)^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right).
\end{aligned}$$

Hence, noting that $\Gamma(k) = (k-1)!$, by the monotone convergence theorem,

$$\mathbb{E}[e^{\lambda\|X\|^2}] = 1 + \sum_{k=1}^{\infty} \frac{\lambda^k}{k!}\,\mathbb{E}[\|X\|^{2k}] \leq 1 + \sum_{k=1}^{\infty} \frac{\lambda^k}{k!}(2k)(2\sigma^2)^k\Gamma(k)$$
$$\leq 1 + 2\sum_{k=1}^{\infty}\left(2\lambda\sigma^2\right)^k = \frac{2}{1-2\lambda\sigma^2} - 1,$$

with the last equality being valid for any $\lambda \in [0, \frac{1}{2\sigma^2})$. Finally, we get $\mathbb{E}\left[e^{\lambda\|X\|^2}\right] \leq 2e^{4\lambda\sigma^2} - 1$ for any $\lambda \in [0, \frac{1}{4\sigma^2}]$ since for any $u \in [0, \frac{1}{2}]$, $\frac{1}{1-u} \leq e^{2u}$. □

Lemma 2.2. *Let $X \in nSG(\sigma)$ such that $\mathbb{E}[X] = 0$. Then, for any $u \in \mathbb{R}^{\mathrm{d}}$ and $\lambda \geq 0$, it holds that*

$$\mathbb{E}\left[e^{\lambda\langle u, X\rangle}\right] \leq e^{8\lambda^2\|u\|^2\sigma^2}.$$

Proof of Lemma (2.2). For $u = 0$, the inequality to prove is trivial. Assume $u \neq 0$. From Lemma 2.1 and Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[e^{\lambda^2\langle u, X\rangle^2}\right] \leq \mathbb{E}[e^{\lambda^2\|u\|^2\|X\|^2}] \leq e^{8\lambda^2\|u\|^2\sigma^2}, \tag{2.3}$$

for all $\lambda \in [0, \frac{1}{2\sqrt{2}\sigma\|u\|}]$. Thus, for any such λ, noticing that $e^t \leq t + e^{t^2}$ for $t \in \mathbb{R}$, we obtain

$$\mathbb{E}\left[e^{\lambda\langle u, X\rangle}\right] \leq \mathbb{E}\left[\lambda\langle u, X\rangle + e^{\lambda^2\langle u, X\rangle^2}\right] \leq e^{8\lambda^2\|u\|^2\sigma^2},$$

where the second inequality follows from (2.3) and the assumption that $\mathbb{E}[X] = 0$. Moreover, for $\lambda \geq \frac{1}{2\sqrt{2}\|u\|\sigma}$, we have by Cauchy Schwarz's inequality and Lemma 2.1 that

$$\mathbb{E}[e^{\lambda\langle u, X\rangle}] \leq \mathbb{E}\left[e^{\frac{8\lambda^2\sigma^2\|u\|^2}{2} + \frac{\|X\|^2}{16\sigma^2}}\right] \leq e^{\frac{1}{2}\left(1 + 8\lambda^2\sigma^2\|u\|^2\right)} \leq e^{8\lambda^2\|u\|^2\sigma^2},$$

where the last inequality is due to $e^{\frac{1+t}{2}} \leq e^t$ for $t \geq 1$. □

2.3 Robustness via risk measures

In this paper, we investigate the robustness of SAPD under different convergence metrics, borrowed from the theory of risk measures [30]. Our motivation is to provide convergence guarantees under various types of distributional perturbations on the stochastic estimates of the gradients $\tilde{\nabla}\Phi_x, \tilde{\nabla}\Phi_y$. As we focus on smooth and SCSC problems, we can rely on the squared distance of the iterates (x_n, y_n) to the solution $(\mathrm{x}^\star, \mathrm{y}^\star)$ to quantify sub-optimality. Precisely, sub-optimality will be measured in terms of the weighted squared distance to the solution, i.e.,

$$\mathcal{D}_n \triangleq \frac{1}{2\tau}\|x_n - \mathrm{x}^\star\|^2 + \frac{1}{2}\Big(\frac{1}{\sigma} - \alpha\Big)\|y_n - \mathrm{y}^\star\|^2, \tag{2.4}$$

for some $\alpha \in [0, \sigma^{-1})$.

The first risk measure of interest is the quantile function, defined for a random variable $\mathcal{U}\colon \Omega \to \mathbb{R}$ as

$$Q_p(\mathcal{U}) \triangleq \inf_{t\in\mathbb{R}} \mathbb{P}[\mathcal{U} \leq t] \geq p.$$

Quantile upperbounds correspond to high-probability results, and quantile bounds have been already fairly studied to assess the robustness of stochastic algorithms [12, 27, 18]. One key contribution of this paper is the derivation of an upperbound on the quantiles of the distance metric $\mathcal{D}_n$, defined in (2.4), exhibiting a tight bias-variance trade-off –see Section 3.3.

Furthermore, we investigate the robustness of SAPD with respect to three convex risk measures, based on φ-divergences [4]. Generally speaking, for a given proper convex function $\varphi : \mathbb{R}_+ \to \mathbb{R}$ satisfying $\varphi(1) = 0$ and $\lim_{t\to 0^+}\varphi(t) = \varphi(0)$, the associated φ-divergence, is defined as

$$D_\varphi(\mathbb{Q}||\mathbb{P}) \triangleq \int_\Omega \varphi\left(\frac{\mathrm{d}\,\mathbb{Q}}{\mathrm{d}\,\mathbb{P}}\right)\mathrm{d}\,\mathbb{P},$$

for any input probability measures $\mathbb{Q}, \mathbb{P}$ such that $\mathbb{Q} \ll \mathbb{P}$, i.e. $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$. Different choices of φ-divergence result in different risk measures [4, 31].

Risk measure	Formulation	Divergence
CVaR_p, $p \in [0,1)$	$\frac{1}{1-p}\int_{p'=p}^{1} Q_{p'}(\mathcal{U})\mathrm{d}p'$	$\varphi(t) = \mathcal{I}_{[0,\frac{1}{1-p}]}(t)$
EVaR_p, $p \in [0,1)$	$\inf_{\eta>0}\left\{\frac{-\log(1-p)}{\eta} + \frac{1}{\eta}\log(\mathbb{E}(e^{\eta\mathcal{U}}))\right\}$	$\varphi(t) = t\log t - t + 1$
$\mathcal{R}_{\chi^2,r}$, $r \geq 0$	$\inf_{\eta\geq 0}\left\{\sqrt{1+2r}\sqrt{\mathbb{E}(\mathcal{U}-\eta)_+^2} + \eta\right\}$	$\varphi(t) = \frac{1}{2}(t-1)^2$

Table 1: Three examples of φ-divergence based risk measures studied in this paper.

Definition 2.3. For any $r \geq 0$, we define the φ-divergence based risk measure at level r as follows:

$$\mathcal{R}_{\varphi,r}(\mathcal{U}) \triangleq \sup_{\substack{\mathbb{Q}\ll\mathbb{P} \\ D_\varphi(\mathbb{Q}||\mathbb{P})\leq r}} \mathbb{E}_{\mathbb{Q}}[\mathcal{U}] \tag{2.5}$$

where $\mathbb{P}$ denotes an arbitrary reference probability measure.

In this paper, we investigate the performances of SAPD under three φ-divergence based risk measures, summarized in Table 1. First, given $p \in [0,1)$, we consider the conditional value at risk (CVaR_p), defined as

$$\mathrm{CVaR}_p(\mathcal{U}) \triangleq \frac{1}{1-p}\int_{p'=p}^{1} Q_{p'}(\mathcal{U})\ \mathrm{d}p'. \tag{2.6}$$

The CVaR admits the variational representation (2.5) with $\varphi : t \mapsto \mathcal{I}_{[0,(1-p)^{-1}]}(t)$ for any $r > 0$.Define the indicator function. As an average of the higher quantiles of $\mathcal{U}$, $\mathrm{CVaR}_p(\mathcal{U})$ holds intuitively as a statistical summary of the tail of $\mathcal{U}$, beyond its p-quantile. While high-probability bounds do not take into account the *price of failure* tied to the event $\mathcal{U} \geq Q_p(\mathcal{U})$, the CVaR presents the advantage of integrating the whole tail of the distribution, which may makes it more suitable to assess the robustness of a given random variable.

The second convex risk measure we investigate is the Entropic Value at Risk [1], denoted EVaR, and is defined as

$$\mathrm{EVaR}_p(\mathcal{U}) \triangleq \inf_{\eta>0}\left\{-\frac{\log(1-p)}{\eta} + \frac{1}{\eta}\log(\mathbb{E}(e^{\eta\mathcal{U}}))\right\}.$$

The EVaR admits the variational representation (2.5) with $\varphi : t \mapsto t\log(t) - t + 1$ and the parameter r is set to $-\log(1-p)$ for given $p \in [0,1)$ –see e.g. [31]. EVaR exhibits a higher tail-sensitivity than CVaR, in the sense that $\mathrm{CVaR}_p(\mathcal{U}) \leq \mathrm{EVaR}_p(\mathcal{U})$, for all $p \in [0,1)$ whenever $\mathrm{EVaR}_p(\mathcal{U}) < \infty$.

Finally we will also derive results in terms of the χ^2-divergence based risk measure, defined as (2.5) with $\varphi : t \mapsto \frac{1}{2}(t-1)^2$.

3 Main Results

In this section, we derive risk-sensitive bounds for the convergence of the SAPD algorithm (see Algorithm 1). We first recall in Section (3.1) a general class of hyperparameters for which `SAPD` is known to converge at an accelerated convergence rate in expectation [38]. We present in Section 3.2 the main results of this paper, which consists of convergence analysis of `SAPD`in high-probability and with respect to the three convex risk measures presented in Table 1. We finally demonstrate in Section 3.3 some tight characteristics of our analysis.

3.1 A class of admissible parameters for `SAPD`

`SAPD` was shown to converge in expectation in [38], at a linear rate $\rho \in (0,1)$ provided the inequality

$$\begin{pmatrix} \frac{1}{\tau}+\mu_x-\frac{1}{\rho\tau} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sigma}+\mu_y-\frac{1}{\rho\sigma} & \left(\frac{\theta}{\rho}-1\right)L_{yx} & \left(\frac{\theta}{\rho}-1\right)L_{yy} & 0 \\ 0 & \left(\frac{\theta}{\rho}-1\right)L_{yx} & \frac{1}{\tau}-L_{xx} & 0 & -\frac{\theta}{\rho}L_{yx} \\ 0 & \left(\frac{\theta}{\rho}-1\right)L_{yy} & 0 & \frac{1}{\sigma}-\alpha & -\frac{\theta}{\rho}L_{yy} \\ 0 & 0 & -\frac{\theta}{\rho}L_{yx} & -\frac{\theta}{\rho}L_{yy} & \frac{\alpha}{\rho} \end{pmatrix} \succeq 0, \tag{3.1}$$

holds for some $\alpha \in [0, \sigma^{-1})$. An important class of solutions to the matrix inequality in (3.1) takes the following form:

$$\tau = \frac{1-\theta}{\theta\mu_x}, \quad \sigma = \frac{1-\theta}{\theta\mu_y}, \quad \theta \geq \bar{\theta}, \tag{3.2}$$

for some $\bar{\theta} \in (0,1)$ that can be explicitly given. It is shown in [6] that for a particular value of θ[1], this parametric choice of primal-dual step sizes τ and σ in the momentum parameter θ ensures acceleration of the algorithm (CP) proposed in [6] for the deterministic case when Φ is bilinear. Indeed, CP Algorithm can be obtained as a special case of SAPD for bilinear coupling functions Φ. In other words, (3.1) describes a general set of parameters for which SAPD will converge in expectation; however risk-sensitive guarantees, including in high-probability are not known. In the forthcoming results, we study SAPD for parameters satisfying (3.1), and obtain convergence rates in high probability, in CVaR, in EVaR, and in the χ^2-divergence-based risk measure, as defined in Table 1. The special parameterization (3.2) will be key to demonstrate the sharpness of our convergence analysis.

3.2 Risk-averse convergence analysis

The main result of this section establishes the convergence of SAPD in high probability, and its proof will be provided in Section 4.

Theorem 3.1. *Suppose $(x_n, y_n)_{n\geq 1}$ are generated by SAPD, initialized at an arbitrary tuple $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$. For all $n \in \mathbb{N}$, $p \in (0,1)$ and $\tau, \sigma > 0$, and $\theta \geq 0$ satisfying* (3.1) *for some $\rho \in (0,1)$ and $\alpha \in [0, \sigma^{-1})$, it holds that*

$$\mathbb{P}\left[\mathcal{D}_{n+1} + (1-\rho)\mathcal{D}_n \leq 2\left(\frac{1+\rho}{2}\right)^n \left(\frac{5}{4}\mathcal{E}_{\tau,\sigma} + \Xi^{(1)}_{\tau,\sigma,\theta}\bar{\delta}^2\right) + \frac{4\Xi^{(1)}_{\tau,\sigma,\theta}\bar{\delta}^2}{1-\rho}\left(1 + \Xi^{(2)}_{\tau,\sigma,\theta}\log\left(\frac{1}{1-p}\right)\right)\right] \geq p, \tag{3.3}$$

where $\mathcal{D}_n = \frac{1}{2\tau}\|x_n - \mathrm{x}^\star\|^2 + \frac{1-\alpha\sigma}{2\sigma}\|y_n - \mathrm{y}^\star\|^2$, $\mathcal{E}_{\tau,\sigma} \triangleq \frac{1}{2\tau}\|x_0 - \mathrm{x}^\star\|^2 + \frac{1}{2\sigma}\|y_0 - \mathrm{y}^\star\|^2$, and $\bar{\delta} \triangleq \max\{\delta_x, \delta_y\}$; furthermore, $\Xi^{(1)}_{\tau,\sigma,\theta}$ and $\Xi^{(2)}_{\tau,\sigma,\theta}$ are constants that do not depend on n and p, they only depend on the problem and algorithm parameters.

Remark 3.1. *It is possible to choose the SAPD parameters satisfying* (3.1) *so that $\rho = 1 - \frac{c}{\kappa}$ for some constant $c > 0$ [38]. Therefore, the decay rate of the bias term $\frac{1+\rho}{2} = 1 - \frac{c/2}{\kappa}$ in Theorem 3.1 demonstrates an accelerated behavior that scales with κ instead of κ^2 dependence of SGDA methods [11]. It is worth noting that for any given $\rho \in (0,1)$, to check if there exists SAPD parameters τ, σ, θ such that the bias component of $\mathbb{E}[\mathcal{D}_n^2]$ decreases to 0 linearly with a rate coefficient at most ρ, one needs to solve a 5-dimensional SDP, i.e., after fixing ρ, checking the feasibility of* (3.1) *reduces to an SDP problem, see [38] for details.*

Using Theorem 3.1 and building on the representation of the CVaR in terms of the quantiles, we can deduce a bound on $\mathrm{CVaR}_p(\mathcal{D}_n^{\frac{1}{2}})$ as shown in Theorem 3.2, where we also provide bounds on the entropic value at risk and on the χ^2-based risk measure, as defined in Table 1.

Theorem 3.2 (Bounds on Risk Measures)**.** *Under the premise of Theorem 3.1, the following bounds hold for all $n \in \mathbb{N}$ and $p \in (0,1)$:*

$$\mathrm{CVaR}_p\left(\mathcal{D}_{n+1}^{\frac{1}{2}}\right) \leq \left(\frac{1+\rho}{2}\right)^{n/2}\left((1+\rho)\mathcal{E}_{\tau,\sigma} + 2\Xi^{(1)}_{\tau,\sigma,\theta}\bar{\delta}^2\right)^{1/2} + \sqrt{\frac{4\Xi^{(1)}_{\tau,\sigma,\theta}\bar{\delta}^2}{1-\rho}\left(1 + \Xi^{(2)}_{\tau,\sigma,\theta}\left(1 + \log\left(\frac{1}{1-p}\right)\right)\right)}, \tag{3.4}$$

$$\mathrm{EVaR}_p(\mathcal{D}_{n+1}^{\frac{1}{2}}) \leq \left(\frac{1+\rho}{2}\right)^{\frac{n}{2}}\left((1+\rho)\mathcal{E}_{\tau,\sigma} + 2\Xi^{(1)}_{\tau,\sigma,\theta}\bar{\delta}^2\right)^{1/2} + \sqrt{\frac{4\Xi^{(1)}_{\tau,\sigma,\theta}\bar{\delta}^2}{1-\rho}}\left(1 + \sqrt{\Xi^{(2)}_{\tau,\sigma,\theta}}\left(\sqrt{\log\left(\frac{1}{1-p}\right)} + \sqrt{\pi}\right)\right), \tag{3.5}$$

where $\mathcal{D}_n$, $\Xi^{(1)}_{\tau,\sigma,\theta}$, $\Xi^{(2)}_{\tau,\sigma,\theta}$ and $\bar{\delta}$ are as defined in Theorem 3.1. Furthermore, for all $n \in \mathbb{N}$ and $r > 0$, the right-hand side of (3.5) *with $p = 1 - \frac{1}{1+r}$ is an upper bound on $\mathcal{R}_{\chi^2,r}(\mathcal{D}_{n+1}^{1/2})$.*

Proof. The CVaR bound in (3.4) directly follows from Corollary 1 applied to the process V_n introduced in (4.19), with the associated constants as defined in (4.23). Furthermore, the EVaR bound in (3.5) directly follows from Corollary 2 applied to the same $(V_n)_{n\geq 0}$. Finally, the bound on $\mathcal{R}_{\chi^2,r}(\mathcal{D}_n^{1/2})$ follows from (3). □

[1] see [6, 48].

3.3 Tightness analysis

We will discuss in this section that the constants given in Theorem 3.1 are tight in the sense that under the Chambolle-Pock parameterization given in (3.2), which corresponds to a particular solution of the matrix inequality in (3.1), the dependency of these constants to θ and p cannot be improved. To this end, we consider quadratic problems subject to additive isotropic Gaussian noise for which we can do exact computations, i.e., both $\{\Delta_k^x\}$ and $\{\Delta_k^y\}$ are iid Gaussian random vector sequences with isotropic covariances, and these sequences are independent from each other as well.

In Section (C.1) of the appendix, under the isotropic Gaussian noise assumption, we show that the distribution π_n of the iterates $z_n \triangleq (x_n, y_n)$ converges to a Gaussian distribution π_∞ with mean (x_*, y_*) and a covariance matrix $\Sigma^\star$ for which we provide a formula in (C.6).

For simplicity of the discussion, consider running SAPD using (CP) parameters given in (3.2) on the following one-dimensional instance of the general quadratic problem (C.1) studied in depth in Section C.

$$\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} \frac{1}{2}x^2 + xy + \frac{1}{2}y^2, \tag{3.6}$$

where we assume that $\theta \geq \bar{\theta}$ as in (3.2). We also set the gradient noise variance as $\delta_x^2 = \delta_y^2 = 1$. In this example, the unique saddle point is at $(x_*, y_*) = (0, 0)$. By our Corollary 4, this ensures that the distribution π_n of the generated sequence $z_n \triangleq (x_n, y_n)$ converges to a centered Gaussian π_∞ in distribution with the covariance matrix $\Sigma^\star$ given in (C.6). If we let z_∞ denote a random variable with the stationary distribution π_∞, in the next proposition we provide lower bounds on the quantiles of $\|z_\infty\|^2$ and compare them to the upper bounds we derived in Theorem 3.1.

Theorem 3.3. *Let $(z_n)_{n\geq 0}$ be the sequence initialized at an arbitrary tuple $z_0 = (x_0, y_0)$ generated by SAPD on Problem* (3.6) *under the parameterization* (3.2)*. Then, for any $p(0, 1)$, the p-quantile $Q_p(\|z_\infty\|^2)$ of the squared norm of the limit $z_\infty = \lim_{n\to\infty} z_n$ satisfieshere, since we talk about convergence in distribution, maybe we should use another notation as one may read it as a.s. convergence.*

$$\psi_1(p, \theta) \leq Q_p(\|z_\infty\|^2) \leq \psi_2(p, \theta),$$

where ψ_1 and ψ_2 satisfy $\psi_1(p, \theta) = (1 - \theta)\log(1/(1 - p))\Theta(1)$ and $\psi_2(p, \theta) = (1 - \theta)\mathcal{O}(1 + \log(1/(1 - p)))$, as $\theta \to 1$.

Proof. See Appendix, section C.3 □

4 Proof of Main Results

4.1 Concentration inequalities through recursive control

In this section, we provide general concentration inequalities that will be specialized later for the analysis of SAPD and SGDA . The following proposition is a variant of the recursive control inequality derived in [8, 18] to analyze the dynamical system corresponding to the stochastic gradient descent (SGD) method for minimizing smooth and strongly convex functions with bounded domains.

Proposition 4.1. *Let $(\mathcal{F}_n)_{n\geq 0}$ be a filtration on $(\Omega, \mathcal{F}, \mathbb{P})$. Let $(V_n)_{n\geq 0}$, $(T_n)_{n\geq 0}$, and $(R_n)_{n\geq 0}$, be three scalar stochastic processes adapted to $(\mathcal{F}_n)_{n\geq 0}$ with following properties: there exist $\sigma_R, \sigma_T > 0$ such that for all $n \geq 0$,*

- *V_n is non-negative;*
- *$\mathbb{E}\left[e^{\lambda T_{n+1}}|\mathcal{F}_n\right] \leq e^{\lambda^2\sigma_T^2 V_n}$ for all $\lambda > 0$, i.e., T_{n+1} conditionally on $\mathcal{F}_n$ is subGaussian;*
- *$\mathbb{E}\left[e^{\lambda R_{n+1}}|\mathcal{F}_n\right] \leq e^{\lambda\sigma_R^2}$ for all $\lambda \in [0, 1/\sigma_R^2]$, i.e., R_{n+1} conditionally on $\mathcal{F}_n$ is subExponential.*

If there exists $\rho \in (0, 1)$ such that

$$V_{n+1} - T_{n+1} - R_{n+1} \leq \rho\, V_n, \quad \forall n \geq 0, \tag{4.1}$$

then for all $\lambda \in \left(0, \min\{\frac{1}{2\sigma_R^2}, \frac{1-\rho}{4\sigma_T^2}\}\right)$, it holds that

$$\mathbb{E}\left[e^{\lambda V_{n+1}}\right] \leq e^{\lambda\sigma_R^2}\mathbb{E}\left[e^{\frac{\lambda(1+\rho)}{2}V_n}\right], \quad \forall\, n \geq 0.$$

Proof. The proof follows closely the arguments of [18]. For any $\lambda \geq 0$, (4.1) together with Cauchy-Schwarz inequality implies that

$$\mathbb{E}\left[e^{\lambda V_{n+1}}|\mathcal{F}_n\right] \leq e^{\lambda\rho V_n}\mathbb{E}\left[e^{\lambda(T_{n+1}+R_{n+1})}|\mathcal{F}_n\right] \leq e^{\lambda\rho V_n}\mathbb{E}\left[e^{2\lambda T_{n+1}}|\mathcal{F}_n\right]^{1/2}\mathbb{E}\left[e^{2\lambda R_{n+1}}|\mathcal{F}_n\right]^{1/2}.$$

Thus for $\lambda \in \left(0, \frac{1}{2\sigma_R^2}\right]$, we have

$$\mathbb{E}\left[e^{\lambda V_{n+1}}|\mathcal{F}_n\right] \leq e^{\lambda\sigma_R^2}e^{\lambda(\rho+2\lambda\sigma_T^2)V_n}.$$

Setting $0 \leq \lambda \leq \min\left\{\frac{1}{2\sigma_R^2}, \frac{1-\rho}{4\sigma_T^2}\right\}$ and taking the non-conditional expectation, we ensure that

$$\mathbb{E}\left[e^{\lambda V_{n+1}}\right] \leq e^{\lambda\sigma_R^2}\mathbb{E}\left[e^{\lambda\frac{1+\rho}{2}V_n}\right].$$

This completes the proof. □

Unrolling the above recursive property on the moment generating function of V provides us with high probability results on $(V_n)_{n\geq 0}$.

Proposition 4.2. *Let V_n, T_n, R_n be defined as in Proposition 4.1. Then, for all $n \geq 0$ and $\lambda \in \left[0, \min\left\{\frac{1-\rho}{4\sigma_T^2}, \frac{1}{2\sigma_R^2}\right\}\right]$, we have*

$$\mathbb{E}\left[e^{\lambda V_n}\right] \leq e^{\frac{2\lambda\sigma_R^2}{1-\rho}}\mathbb{E}\left[e^{\lambda\left(\frac{1+\rho}{2}\right)^n V_0}\right]. \tag{4.2}$$

Furthermore, if $V_0 = C_0$ is constant, then

$$\mathbb{P}\left[V_n \leq \left(\frac{1+\rho}{2}\right)^n C_0 + \frac{2\sigma_R^2}{1-\rho}\left(1 + \max\left\{1, 2\,\frac{\sigma_T^2}{\sigma_R^2}\right\}\log\left(\frac{1}{1-p}\right)\right)\right] \geq p. \tag{4.3}$$

Alternatively, if V_0 can be expressed as $V_0 = C_0 + \mathcal{U}$ where $C_0 \geq 0$ is constant and $\mathcal{U}$ satisfies

$$\mathbb{E}\left[e^{\lambda\mathcal{U}}\right] \leq e^{\alpha\lambda+\beta\lambda^2}, \quad \forall\lambda \in \left[0, \frac{1}{\bar{\alpha}}\right]$$

for some constants $\alpha, \bar{\alpha}, \beta > 0$, then, for any $p \in [0,1)$ and $\lambda \in [0, \gamma]$,

$$\mathbb{P}\left(V_n \leq \left(\frac{1+\rho}{2}\right)^n (C_0 + \alpha) + \left(\frac{1+\rho}{2}\right)^{2n}\lambda\beta + \frac{2\sigma_R^2}{1-\rho} + \frac{1}{\lambda}\log\left(\frac{1}{1-p}\right)\right) \geq p, \tag{4.4}$$

where $\gamma \triangleq \frac{1-\rho}{\max\{\bar{\alpha}, 2\sigma_R^2, 4\sigma_T^2\}}$.

Proof. Let us first prove by induction on n that for all $\lambda \in \left(0, \min\left\{\frac{1-\rho}{4\sigma_T^2}, \frac{1}{2\sigma_R^2}\right\}\right)$,

$$\mathbb{E}\left[e^{\lambda V_n}\right] \leq \mathbb{E}\left[e^{\lambda\left(\frac{1+\rho}{2}\right)^n V_0 + \lambda\sigma_R^2\sum_{k=0}^{n-1}\left(\frac{1+\rho}{2}\right)^k}\right]. \tag{4.5}$$

For $n = 0$, this property holds trivially with the convention $\Sigma_{k=0}^{n-1} = 0$ when $n = 0$. Assuming the inequality holds for some $n \geq 0$, next we show it also holds for $n+1$. According to Proposition 4.1, we have

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda V_{n+1}}\right] &\leq e^{\lambda\sigma_R^2}\mathbb{E}\left[e^{\lambda\frac{1+\rho}{2}V_n}\right] \\
&\leq e^{\lambda\sigma_R^2}\mathbb{E}\left[e^{\lambda\frac{1+\rho}{2}\left(\frac{1+\rho}{2}\right)^n V_0 + \lambda\frac{1+\rho}{2}\sigma_R^2\sum_{k=0}^{n-1}\left(\frac{1+\rho}{2}\right)^k}\right] \\
&= \mathbb{E}\left[e^{\lambda\left(\frac{1+\rho}{2}\right)^{n+1}V_0 + \lambda\sigma_R^2\sum_{k=0}^{n}\left(\frac{1+\rho}{2}\right)^k}\right],
\end{aligned}$$

where the second inequality follows from the induction hypothesis since

$$0 < \lambda(1+\rho)/2 \leq \lambda \leq \min\left\{\frac{1-\rho}{4\sigma_T^2}, \frac{1}{2\sigma_R^2}\right\},$$

and this completes the induction. Thus, (4.2) follows from using $\sum_{k=0}^{n-1}\left(\frac{1+\rho}{2}\right)^k \leq \frac{2}{1-\rho}$ within (4.5). The remaining statements follow from a Chernoff bound; indeed, if $V_0 = C_0$ is constant, we obtain

$$\mathbb{P}\left[V_n \geq \left(\frac{1+\rho}{2}\right)^n C_0 + \frac{2\sigma_R^2}{1-\rho} + t\right] \leq e^{-\lambda t}$$

for $\lambda = \frac{1-\rho}{2\max\{\sigma_R^2, 2\sigma_T^2\}}$ and $t = \frac{1}{\lambda}\log(\frac{1}{1-p})$, which implies the desired result.

Next, suppose $V_0 = C_0 + \mathcal{U}$ for some constant C_0 and $\mathcal{U}$ as in the hypothesis. First, observe that for $\lambda \in \left(0, \min\left\{\frac{1-\rho}{4\sigma_T^2}, \frac{1}{2\sigma_R^2}, \frac{1}{\bar{\alpha}}\right\}, \right)$, we have

$$\begin{aligned}\mathbb{E}\left[e^{\lambda V_n}\right] &\leq e^{\frac{2\lambda\sigma_R^2}{1-\rho}}\mathbb{E}\left[e^{\lambda\left(\frac{1+\rho}{2}\right)^n(C_0+\mathcal{U})}\right] \\ &\leq e^{\lambda\left(\left(\frac{1+\rho}{2}\right)^n(C_0+\alpha)+\frac{2\sigma_R^2}{1-\rho}\right)+\lambda^2\left(\frac{1+\rho}{2}\right)^{2n}\beta}.\end{aligned} \tag{4.6}$$

Thus, for all $t \geq 0$,

$$\mathbb{P}\left(V_n > \left(\frac{1+\rho}{2}\right)^n(C_0+\alpha) + \frac{2\sigma_R^2}{1-\rho} + t\right) \leq e^{\lambda^2\left(\frac{1+\rho}{2}\right)^{2n}\beta - \lambda t}.$$

Fixing an arbitrary non-negative λ such that $\lambda \leq \frac{1-\rho}{\max\{\bar{\alpha}, 2\sigma_R^2, 4\sigma_T^2\}}$, we have $\exp(\lambda^2\left(\frac{1+\rho}{2}\right)^{2n}\beta - \lambda t) = 1-p \iff t = \lambda(\frac{1+\rho}{2})^{2n}\beta + \frac{1}{\lambda}\log(1/(1-p))$, which proves (4.4). □

Thanks to the recursive control property 4.2, one can derive convergence rates for the CVaR and EVaR risk measures of the scalar process $(V_n)_{n\geq 0}$.

Corollary 1. *Let V_n, T_n, R_n, γ be defined as in Proposition 4.2. Then, for any $p \in [0,1)$ and $\lambda \in [0,\gamma]$,*

$$\mathrm{CVaR}_p(V_n^{\frac{1}{2}}) \leq \left(\frac{1+\rho}{2}\right)^{\frac{n}{2}}\sqrt{C_0+\alpha+\lambda\beta} + \sqrt{\frac{2}{1-\rho}\sigma_R^2 + \frac{1}{\lambda}\left(1+\log\left(\frac{1}{1-p}\right)\right)}. \tag{4.7}$$

Proof. Note that the first and second terms on the right-hand side of (4.4) can be bounded by $\left(\frac{1+\rho}{2}\right)^n(C_0+\alpha+\gamma\beta)$; hence, by integrating the resulting looser bound with respect to p, and using CVaR's integral formulation in (2.6), we obtain

$$\mathrm{CVaR}_p(V_n) \leq \left(\frac{1+\rho}{2}\right)^n(C_0+\alpha+\lambda\beta) + \frac{2\sigma_R^2}{1-\rho} + \frac{1}{\lambda}\left(1+\log\left(\frac{1}{1-p}\right)\right),$$

which directly implies (4.7), due to Lemma (5.2) and the sub-additivity of $\sqrt{\cdot}$. □

Corollary 2. *Let V_n, T_n, R_n, γ be defined as in Proposition 4.2. Then, for any $p \in [0,1)$, and $\lambda \in [0,\gamma]$,*

$$\mathrm{EVaR}_p\left(V_n^{\frac{1}{2}}\right) \leq \left(\frac{1+\rho}{2}\right)^{n/2}\sqrt{C_0+\alpha+\lambda\beta} + \sqrt{\frac{2}{1-\rho}}\sigma_R + \left(\sqrt{\frac{1}{\lambda}\log(\frac{1}{1-p})} + \frac{\sqrt{\pi}}{\sqrt{\lambda}}\right). \tag{4.8}$$

Proof. The bound in (4.4) of Proposition 4.2 ensures that for all $p \in [0,1)$ and $\lambda \in [0,\gamma]$, the p-th quantile of V_n satisfies

$$Q_p\left(V_n\right) \leq \left(\frac{1+\rho}{2}\right)^n(C_0+\alpha+\lambda\beta) + \frac{2\sigma_R^2}{1-\rho} + \frac{1}{\lambda}\log\left(\frac{1}{1-p}\right);$$

hence, non-negativity of V_n, Lemma 5.2 and sub-additivity of $t \mapsto \sqrt{t}$ together imply that

$$Q_p\left(V_n^{1/2}\right) \leq \left(\frac{1+\rho}{2}\right)^{n/2}\sqrt{C_0+\alpha+\lambda\beta} + \sqrt{\frac{2}{1-\rho}}\sigma_R + \frac{1}{\sqrt{\lambda}}\log\left(\frac{1}{1-p}\right)^{1/2}. \tag{4.9}$$

For $n \geq 0$, let

$$U_n \triangleq V_n^{1/2} - \left(\frac{1+\rho}{2}\right)^{n/2}\sqrt{C_0+\alpha+\lambda\beta} - \sqrt{\frac{2}{1-\rho}}\sigma_R,$$

and note that (4.9) implies

$$\mathbb{P}\left(U_n > t\right) \leqslant e^{-\lambda t^2} \quad \forall\, t \geq 0. \tag{4.10}$$

Therefore, following standard arguments from [33], we have for any $\eta > 0$ that

$$\begin{aligned}
\mathbb{E}(e^{\eta U_n}) &= \int_0^\infty \mathbb{P}\left[e^{\eta U_n} > t\right] dt = \int_{-\infty}^\infty \mathbb{P}\left[e^{\eta U_n} > e^u\right] e^u du \\
&= \int_{-\infty}^0 \mathbb{P}\left[e^{\eta U_n} > e^u\right] e^u du + \int_0^\infty \mathbb{P}\left[e^{\eta U_n} > e^u\right] e^u du \\
&\leq \int_{-\infty}^0 e^u du + \int_0^\infty e^{-\frac{\lambda}{\eta^2}u^2} e^u du = 1 + e^{\frac{\eta^2}{4\lambda}} \int_0^\infty e^{-\frac{\lambda}{\eta^2}(u-\frac{\eta^2}{2\lambda})^2} du \\
&= 1 + e^{\frac{\eta^2}{4\lambda}} \int_{-\frac{\eta^2}{2\lambda}}^\infty e^{-\frac{\lambda}{\eta^2}s^2} ds \\
&\leq 1 + \eta e^{\frac{\eta^2}{4\lambda}} \sqrt{\frac{\pi}{\lambda}} \leq \left(1 + \eta\sqrt{\frac{\pi}{\lambda}}\right) e^{\frac{\eta^2}{4\lambda}},
\end{aligned}$$

where we used (4.10). On the other hand,

$$\begin{aligned}
\mathrm{EVaR}_p\left[U_n\right] = \inf_{\eta>0}\left\{\frac{-\log(1-p)}{\eta} + \frac{1}{\eta}\log \mathbb{E}\left[e^{\eta U_n}\right]\right\} &\leq \inf_{\eta>0} \frac{-\log(1-p)}{\eta} + \frac{1}{\eta}\left(\frac{\eta^2}{4\lambda} + \eta\frac{\sqrt{\pi}}{\sqrt{\lambda}}\right) \\
&= \frac{\sqrt{\log(\frac{1}{1-p})}}{\sqrt{\lambda}} + \frac{\sqrt{\pi}}{\sqrt{\lambda}},
\end{aligned}$$

where we used $\log(1+x) \leq x$ for $x \geq 0$. Finally, by translation invariance of the EVaR, we obtain

$$\mathrm{EVaR}_p\left[V_n^{\frac{1}{2}}\right] \leq \left(\frac{1+\rho}{2}\right)^{n/2} \sqrt{C_0 + \alpha + \lambda\beta} + \sqrt{\frac{2}{1-\rho}}\sigma_R + \left(\sqrt{\frac{1}{\lambda}\log\left(\frac{1}{1-p}\right)} + \frac{\sqrt{\pi}}{\sqrt{\lambda}}\right).$$

□

We finish with a bound on the χ^2-based risk measure, as defined in Table 1.

Corollary 3. *Let V_n, T_n, R_n, γ be defined as in Proposition 4.2. Then, for any $r > 0$, and $\lambda \in [0, \gamma]$,*

$$\mathcal{R}_{\chi^2,r}\left(V_n^{\frac{1}{2}}\right) \leq \left(\frac{1+\rho}{2}\right)^{n/2} \sqrt{C_0 + \alpha + \lambda\beta} + \sqrt{\frac{2}{1-\rho}}\sigma_R + \left(\sqrt{\frac{1}{\lambda}\log\left(1+r\right)} + \frac{\sqrt{\pi}}{\sqrt{\lambda}}\right). \tag{4.11}$$

Proof. By[13, Theorem 5], for all $\mathbb{Q} \ll \mathbb{P}$, we have

$$D_{\varphi_{\mathrm{KL}}}(\mathbb{Q}, \mathbb{P}) \leq \log\left(1 + D_{\varphi_{\chi^2}}(\mathbb{Q}, \mathbb{P})\right),$$

where $\varphi_{\mathrm{KL}}(t) = t\log(t) - t + 1$. Therefore, for any integrable random variable $U : \Omega \to \mathbb{R}$, we get

$$\sup_{\mathbb{Q}: D_{\varphi_{\chi^2}}(\mathbb{Q} \,||\, \mathbb{P}) \leq r} \mathbb{E}_{\mathbb{Q}}[U] \leq \sup_{\mathbb{Q}: D_{\varphi_{\mathrm{KL}}}(\mathbb{Q},\mathbb{P}) \leq \log(1+r)} \mathbb{E}_{\mathbb{Q}}[U] = \mathrm{EVaR}_{1-1/(1+r)}(U),$$

whenever $\mathrm{EVaR}_{1-1/(1+r)}(U) < \infty$, where we used the EVaR representation given in Table 1. The statement follows directly from Corollary 2. □

In the next section, we design scalar processes which satisfy the above properties while dominating the error on `SGDA` and `SAPD` iterates respectively.

4.2 Proof of Theorem 3.1

The aim of this section is to prove Theorem 3.1. Here the application of the recursive control inequality from Section 4.1 is not straightforward: the Gauss-Seidel iteration,peculiar to `SAPD` iterates significantly complicates their measurability properties, as illustrated in Figure 2. We circumvent this issue by introducing a stochastic process that almost surely upperbounds the distance to the saddle point while exhibiting simpler measurability characteristics.

Our proof combines several ingredients. Namely, the almost sure upper-bound derived in Proposition 4.3, the recursive control inequality from Proposition 4.2, some elementary concentration results on norm sub-Gaussian vector that we

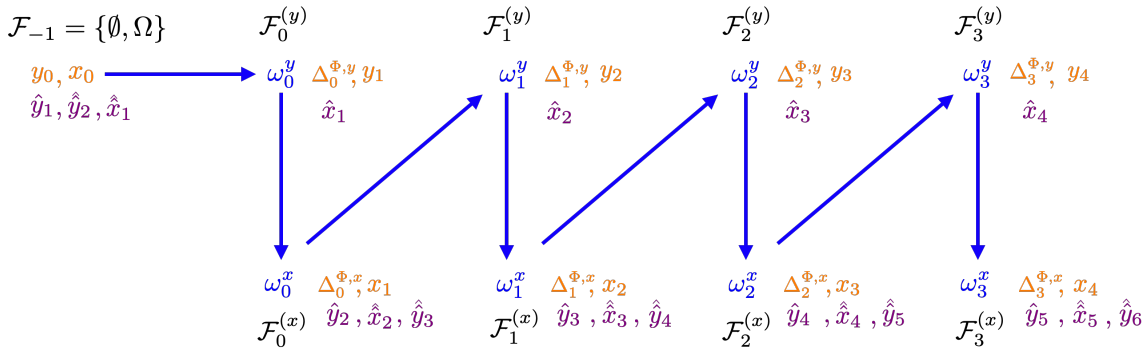

Figure 2: Measurability of sequences of interest in SAPD. Our analysis is made possible by the introduction of "noise-free" counterparts $\hat{\mathrm{x}}_k, \hat{\hat{\mathrm{x}}}_k, \hat{\mathrm{y}}_k, \hat{\hat{\mathrm{y}}}_k$ to the iterates x_k, y_k as defined in (4.13).

review in Section 2.2 and convex inequalities that we prove in Section 4.3.We should reorder things here – we should not refer to later parts of the paper in the proof.

First, observe that according to Proposition 4.3, we have

$$\begin{aligned}
\mathcal{E}_n &\leq \rho^{n-1}\mathcal{E}_{\tau,\sigma} + \sum_{k=0}^{n-1} \rho^{n-1-k} \left(\langle \Delta_k^x, \mathrm{x}^\star - x_{k+1} \rangle + \langle (1+\theta)\Delta_k^y - \theta \Delta_{k-1}^y, y_{k+1} - \mathrm{y}^\star \rangle \right) \\
&= \rho^{n-1}\mathcal{E}_{\tau,\sigma} + \sum_{k=0}^{n-1} \rho^{n-1-k} \left(\langle \Delta_k^x, \mathrm{x}^\star - \hat{\hat{\mathrm{x}}}_{k+1} \rangle + (1+\theta)\langle \Delta_k^y, \hat{\mathrm{y}}_{k+1} - \mathrm{y}^\star \rangle - \theta \langle \Delta_{k-1}^y, \hat{\hat{\mathrm{y}}}_{k+1} - \mathrm{y}^\star \rangle \right) \\
&\quad + \sum_{k=0}^{n-1} \rho^{n-1-k} \left(\langle \Delta_k^x, \hat{\hat{\mathrm{x}}}_{k+1} - x_{k+1} \rangle + (1+\theta)\langle \Delta_k^y, y_{k+1} - \hat{\mathrm{y}}_{k+1} \rangle - \theta \langle \Delta_{k-1}^y, y_{k+1} - \hat{\hat{\mathrm{y}}}_{k+1} \rangle \right)
\end{aligned} \tag{4.12}$$

with the pseudo-iterates $\hat{\mathrm{x}}_k, \hat{\mathrm{y}}_k, \hat{\hat{\mathrm{x}}}_k, \hat{\hat{\mathrm{y}}}_k$ are defined as follows:

$$\hat{\mathrm{x}}_0 \triangleq \hat{\hat{\mathrm{x}}}_0 \triangleq x_0, \qquad \hat{\mathrm{y}}_0 \triangleq \hat{\hat{\mathrm{y}}}_0 \triangleq y_0, \tag{4.13}$$

$$\hat{\mathrm{x}}_{k+1} \triangleq \mathrm{prox}_{\tau f}(x_k - \tau \nabla_{\mathrm{x}} \Phi(x_k, y_{k+1})), \qquad \hat{\mathrm{y}}_{k+1} \triangleq \mathrm{prox}_{\tau f}\Big(y_k + \sigma(1+\theta)\nabla_{\mathrm{y}} \Phi(x_k, y_k) - \sigma\theta \nabla_{\mathrm{y}} \Phi(x_{k-1}, y_{k-1})\Big), \tag{4.14}$$

$$\hat{\hat{\mathrm{x}}}_{k+1} \triangleq \mathrm{prox}_{\tau f}\left(x_k - \tau \nabla_{\mathrm{x}} \Phi(x_k, \hat{\mathrm{y}}_{k+1})\right), \qquad \hat{\hat{\mathrm{y}}}_{k+1} \triangleq \mathrm{prox}_{\tau g}\left(\hat{\mathrm{y}}_k + \sigma(1+\theta)\nabla_{\mathrm{y}} \Phi(\hat{\hat{\mathrm{x}}}_k, \hat{\mathrm{y}}_k) - \sigma\theta \nabla_{\mathrm{y}} \Phi(x_{k-1}, y_{k-1})\right). \tag{4.15}$$

(4.16)

These pseudo-iterates, which measurability properties are illustrated in Figure 2, will be key for the application of the high probability bound given in Proposition 4.2.

For $k \geq 0$, we also define

$$\begin{aligned}
P_k^{(1)} &\triangleq \langle \Delta_k^x, \mathrm{x}^\star - \hat{\hat{\mathrm{x}}}_{k+1} \rangle + (1+\theta)\langle \Delta_k^y, \hat{\mathrm{y}}_{k+1} - \mathrm{y}^\star \rangle, \\
P_k^{(2)} &\triangleq \frac{-\theta}{\rho} \langle \Delta_k^y, \hat{\hat{\mathrm{y}}}_{k+2} - \mathrm{y}^\star \rangle, \\
Q_k &\triangleq \langle \Delta_k^x, \hat{\hat{\mathrm{x}}}_{k+1} - x_{k+1} \rangle + (1+\theta)\langle \Delta_k^y, y_{k+1} - \hat{\mathrm{y}}_{k+1} \rangle - \theta \langle \Delta_{k-1}^y, y_{k+1} - \hat{\hat{\mathrm{y}}}_{k+1} \rangle.
\end{aligned}$$

Thus, rearranging the sums in (4.12) and using $\Delta_{-1}^y = \mathbf{0}$, we may write it equivalently as follows:

$$\mathcal{E}_n \leq \rho^{n-1}\mathcal{E}_{\tau,\sigma} + \sum_{k=0}^{n-1} \rho^{n-1-k} P_k^{(1)} + \sum_{k=0}^{n-2} \rho^{n-1-k} P_k^{(2)} + \sum_{k=0}^{n-1} \rho^{n-1-k} Q_k.$$

Now notice that for $n \geq 0$,

$$\begin{aligned}\mathcal{E}_{n+1} + (1-\rho)\mathcal{E}_n &\leq \rho^{n-1}\mathcal{E}_{\tau,\sigma} + \sum_{k=0}^{n-1}\rho^{n-1-k}P_k^{(1)} + \sum_{k=0}^{n-2}\rho^{n-1-k}P_k^{(2)} + \sum_{k=0}^{n-1}\rho^{n-1-k}Q_k \\ &\quad + P_n^{(1)} + \rho P_{n-1}^{(2)} + Q_n \\ &= \rho^{-1}\left(\rho^n\mathcal{E}_{\tau,\sigma} + \sum_{k=0}^{n}\rho^{n-k}P_k^{(1)} + \sum_{k=0}^{n}\rho^{n-k}P_k^{(2)} + \sum_{k=0}^{n-1}\rho^{n-k}Q_k\right) \\ &\quad + \left(1-\rho^{-1}\right)P_n^{(1)} - \rho^{-1}P_n^{(2)} + (\rho-1)\,P_{n-1}^{(2)} + Q_n.\end{aligned} \tag{4.17}$$

By Proposition 4.6, we obtain

$$\begin{aligned}\frac{\rho}{2}\left(\mathcal{E}_{n+1} + (1-\rho)\mathcal{E}_n\right) &\leq \rho^n\mathcal{E}_{\tau,\sigma} + \sum_{k=0}^{n}\rho^{n-k}P_k^{(1)} + \sum_{k=0}^{n}\rho^{n-k}P_k^{(2)} \\ &\quad + \sum_{k=0}^{n}\rho^{n-k}\mathcal{Q}\left(\|\Delta_k^x\|^2 + \|\Delta_k^y\|^2 + \rho\left(\|\Delta_{k-1}^x\|^2 + \|\Delta_{k-1}^y\|^2\right) + \rho^2\|\Delta_{k-2}^y\|^2\right),\end{aligned}$$

which implies

$$\frac{\rho}{2}\left(\mathcal{E}_{n+1} + (1-\rho)\mathcal{E}_n\right) \leq \rho^n\mathcal{E}_{\tau,\sigma} + \sum_{k=0}^{n}\rho^{n-k}P_k^{(1)} + \sum_{k=0}^{n}\rho^{n-k}P_k^{(2)} + 3\,\mathcal{Q}\sum_{k=0}^{n}\rho^{n-k}\left(\|\Delta_k^x\|^2 + \|\Delta_k^y\|^2\right), \tag{4.18}$$

which follows from $\Delta_{-1}^x \triangleq \mathbf{0}$ and $\Delta_{-1}^y = \Delta_{-2}^y \triangleq \mathbf{0}$. For $n \in \mathbb{N}$, we define V_n, T_{n+1} and R_{n+1} as follows:

$$\begin{aligned}V_n &\triangleq \rho^n\mathcal{E}_{\tau,\sigma} + \sum_{k=0}^{n}\rho^{n-k}P_k^{(1)} + \sum_{k=0}^{n}\rho^{n-k}P_k^{(2)} + 3\,\mathcal{Q}\sum_{k=0}^{n}\rho^{n-k}\left(\|\Delta_k^x\|^2 + \|\Delta_k^y\|^2\right), \\ T_{n+1} &\triangleq P_{n+1}^{(1)} + P_{n+1}^{(2)}, \\ R_{n+1} &\triangleq 3\,\mathcal{Q}\left(\|\Delta_{n+1}^x\|^2 + \|\Delta_{n+1}^y\|^2\right);\end{aligned} \tag{4.19}$$

therefore, (4.18) implies that

$$\frac{\rho}{2}\left(\mathcal{E}_{n+1} + (1-\rho)\mathcal{E}_n\right) \leq V_n, \quad \forall\, n \geq 0, \quad \text{a.s.} \tag{4.20}$$

Let us now show that V_n satisfies the assumptions of the recursive control inequality in (4.1). Indeed, for any $n \geq 0$,

$$V_{n+1} - V_n = (\rho-1)V_n + P_{n+1}^{(1)} + P_{n+1}^{(2)} + 3\,\mathcal{Q}\left(\|\Delta_{n+1}^x\|^2 + \|\Delta_{n+1}^y\|^2\right),$$

which is equivalent to $V_{n+1} \leq \rho V_n + T_{n+1} + R_{n+1}$. Let $(\mathcal{F}_n)_{n\geq -1}$ be the filtration defined as $\mathcal{F}_{-1} \triangleq \{\emptyset, \Omega\}$, and

$$\mathcal{F}_n = \sigma\left(\mathcal{F}_{n-1} \cup \sigma(\Delta_n^y) \cup \sigma(\Delta_n^x)\right), \quad \forall\, n \geq 0.$$

We first observe that for all $n \in \mathbb{N}$, V_n, T_n and R_n are $\mathcal{F}_n$-measurable; moreover, V_n is non-negative. Second, for any $n \geq 0$, we also note that since Δ_n^x and Δ_n^y are norm subGaussian conditioned on $\mathcal{F}_{n-1}$, we get for any $\lambda \geq 0$ that

$$\begin{aligned}\mathbb{E}\left[e^{\lambda T_{n+1}}|\mathcal{F}_n\right] &= \mathbb{E}\left[e^{\lambda\langle\Delta_{n+1}^x, \mathrm{x}^\star - \hat{\hat{\mathrm{x}}}_{n+2}\rangle + \lambda\left\langle\Delta_{n+1}^y,\, (1+\theta)\left(\hat{\mathrm{y}}_{n+2} - \mathrm{y}^\star\right) - \theta\rho^{-1}\left(\hat{\hat{\mathrm{y}}}_{n+3} - \mathrm{y}^\star\right)\right\rangle}\Big|\,\mathcal{F}_n\right] \\ &\leq \mathbb{E}\left[e^{2\lambda\langle\Delta_{n+1}^x, \mathrm{x}^\star - \hat{\hat{\mathrm{x}}}_{n+2}\rangle}|\mathcal{F}_n\right]^{\frac{1}{2}}\mathbb{E}\left[e^{2\lambda\langle\Delta_{n+1}^y, (1+\theta)\left(\hat{\mathrm{y}}_{n+2} - \mathrm{y}^\star\right) - \theta\rho^{-1}\left(\hat{\hat{\mathrm{y}}}_{n+3} - \mathrm{y}^\star\right)\rangle}|\mathcal{F}_n\right]^{\frac{1}{2}} \\ &\leq e^{16\lambda^2\left(\delta_x^2\,\|\hat{\hat{\mathrm{x}}}_{n+2} - \mathrm{x}^\star\|^2 + \delta_y^2\,\|(1+\theta)\left(\hat{\mathrm{y}}_{n+2} - \mathrm{y}^\star\right) - \theta\rho^{-1}\left(\hat{\hat{\mathrm{y}}}_{n+3} - \mathrm{y}^\star\right)\|^2\right)} \\ &\leq e^{16\lambda^2\left(\delta_x^2\,\|\hat{\hat{\mathrm{x}}}_{n+2} - \mathrm{x}^\star\|^2 + 2\delta_y^2\,(1+\theta)^2\|\hat{\mathrm{y}}_{n+2} - \mathrm{y}^\star\|^2 + \delta_y^2\frac{2\theta^2}{\rho^2}\|\hat{\hat{\mathrm{y}}}_{n+3} - \mathrm{y}^\star\|^2\right)},\end{aligned}$$

where in the first inequality we used Cauchy-Schwarz inequality, in the second inequality we used Lemma 2.1 together with $\hat{\hat{\mathrm{x}}}_{n+2}$, $\hat{\mathrm{y}}_{n+2}$, $\hat{\hat{\mathrm{y}}}_{n+3}$ all being $\mathcal{F}_n$-measurable. Hence, in view of Lemma 4.7, we have

$$\begin{aligned}\mathbb{E}\left[e^{\lambda T_{n+1}}|\mathcal{F}_n\right] &\leq e^{16\lambda^2\left(\|A_1\|^2\delta_x^2 + \left(\|A_2\|^2 + \|A_3\|^2\right)\delta_y^2\right)\left(\mathcal{E}_{n+1} + (1-\rho)\mathcal{E}_n\right)} \\ &\leq e^{\frac{32\lambda^2}{\rho}\left(\|A_1\|^2\delta_x^2 + \left(\|A_2\|^2 + \|A_3\|^2\right)\delta_y^2\right)V_n},\end{aligned} \tag{4.21}$$

where the second inequality follows from (4.20).

Third, for all $n \geq 0$ and $\lambda \in \left(0, \frac{1}{48\mathcal{Q}\max\{\delta_x^2, \delta_y^2\}}\right)$, we have in view of Lemma 2.1

$$\mathbb{E}\left[e^{\lambda R_n}\right] \leq \mathbb{E}\left[e^{6\lambda\mathcal{Q}\|\Delta_n^x\|^2}\right]^{\frac{1}{2}} \mathbb{E}\left[e^{6\lambda\mathcal{Q}\|\Delta_n^y\|^2}\right]^{\frac{1}{2}} \leq \exp\left(24\lambda\mathcal{Q}\ \left(\delta_x^2 + \delta_y^2\right)\right). \tag{4.22}$$

Finally, note that $\hat{\mathrm{y}}_1, \hat{\mathrm{x}}_1$ and $\hat{\mathrm{y}}_2$ are all deterministic quantities; hence, using $\mathbb{E}[|WXYZ|] \leq \mathbb{E}[W^4]^{1/4}\,\mathbb{E}[X^4]^{1/4}\,\mathbb{E}[Y^4]^{1/4}\,\mathbb{E}[Z^4]^{1/4}$ as a result of Hölder's inequality and invoking Lemma 2.2, we obtain for all $\lambda \in \left[0, \left(96\mathcal{Q}\max\{\delta_x^2, \delta_y^2\}\right)^{-1}\right]$ that

$$\begin{aligned}
&\mathbb{E}\left[e^{\lambda(V_0 - \mathcal{E}_{\tau,\sigma})}\right] = \mathbb{E}\left[e^{\lambda\left(P_0^{(1)} + P_0^{(2)} + 3\mathcal{Q}\left(\|\Delta_0^x\|^2 + \|\Delta_0^y\|^2\right)\right)}\right] \\
&= \mathbb{E}\left[e^{\lambda\langle\Delta_0^x, \mathrm{x}^\star - \hat{\mathrm{x}}_1\rangle + \lambda\langle\Delta_0^y, (1+\theta)(\hat{\mathrm{y}}_1 - \mathrm{y}^\star) - \frac{\theta}{\rho}(\hat{\mathrm{y}}_2 - \mathrm{y}^\star)\rangle + 3\lambda\mathcal{Q}\left(\|\Delta_0^x\|^2 + \|\Delta_0^y\|^2\right)}\right] \\
&\leq \mathbb{E}\left[e^{4\lambda\langle\Delta_0^x, \mathrm{x}^\star - \hat{\mathrm{x}}_1\rangle}\right]^{\frac{1}{4}} \mathbb{E}\left[e^{4\lambda\langle\Delta_0^y, (1+\theta)(\hat{\mathrm{y}}_1 - \mathrm{y}^\star) - \frac{\theta}{\rho}(\hat{\mathrm{y}}_2 - \mathrm{y}^\star)\rangle}\right]^{\frac{1}{4}} \mathbb{E}\left[e^{12\lambda\mathcal{Q}\|\Delta_0^x\|^2}\right]^{\frac{1}{4}} \mathbb{E}\left[e^{12\lambda\mathcal{Q}\|\Delta_0^y\|^2}\right]^{\frac{1}{4}} \\
&\leq e^{32\lambda^2\left(\|\hat{\mathrm{x}}_1 - \mathrm{x}^\star\|^2\delta_x^2 + 2(1+\theta)^2\|(\hat{\mathrm{y}}_1 - \mathrm{y}^\star)\|^2\delta_y^2 + \frac{2\theta^2}{\rho^2}\|\hat{\mathrm{y}}_2 - \mathrm{y}^\star\|^2\delta_y^2\right)} e^{24\lambda\mathcal{Q}\left(\delta_x^2 + \delta_y^2\right)} \\
&\leq e^{24\mathcal{Q}\left(\delta_x^2 + \delta_y^2\right)\lambda} \cdot e^{32(2-\rho)\left(\|A_1\|^2\delta_x^2 + \left(\|A_2\|^2 + \|A_3\|^2\right)\delta_y^2\right)\mathcal{E}_0\lambda^2},
\end{aligned}$$

where in the last inequality we used Lemma 4.7 and the relations $x_0 = x_{-1}$, and $y_0 = y_{-1}$. Hence, we can use Proposition4.2 with

$$\begin{aligned}
\sigma_T^2 &= \frac{32}{\rho}\left(\|A_1\|^2\delta_x^2 + \left(\|A_2\|^2 + \|A_3\|^2\right)\delta_y^2\right), \quad \sigma_R^2 = 24\mathcal{Q}(\delta_x^2 + \delta_y^2) \\
\alpha &= 24\mathcal{Q}\left(\delta_x^2 + \delta_y^2\right), \quad \bar{\alpha} = 96\mathcal{Q}\max\{\delta_x^2, \delta_y^2\}, \\
\beta &= 32\frac{(2-\rho)}{\rho}\Big(\|A_1\|^2\delta_x^2 + \left(\|A_2\|^2 + \|A_3\|^2\right)\delta_y^2\Big)\mathcal{E}_{\tau,\sigma},
\end{aligned} \tag{4.23}$$

where we used the fact that $\mathcal{E}_{\tau,\sigma} \geq \rho\mathcal{E}_0$ while setting the value for β. When we invoke Proposition4.2, we set $\lambda = \tilde{\gamma}$ within (4.4) for some particular $\tilde{\gamma} > 0$ such that $\tilde{\gamma} \leq \gamma$ as required by the proposition. Thus, for any $p \in (0, 1)$ and $n \geq 0$, the following inequality giving a bound on V_n,

$$\begin{aligned}
V_n \leq &\left(\frac{1+\rho}{2}\right)^n \left(\mathcal{E}_{\tau,\sigma} + 24\mathcal{Q}(\delta_x^2 + \delta_y^2)\right) \\
&+ \left(\frac{1+\rho}{2}\right)^{2n} \tilde{\gamma}\ \frac{32(2-\rho)}{\rho}\Big(\|A_1\|^2\delta_x^2 + \left(\|A_2\|^2 + \|A_3\|^2\right)\delta_y^2\Big)\mathcal{E}_{\tau,\sigma} \\
&+ \frac{1}{1-\rho}\left(48\mathcal{Q}(\delta_x^2 + \delta_y^2)\right) + \frac{1}{\tilde{\gamma}}\log\left(\frac{1}{1-p}\right)
\end{aligned}$$

holds with probability at least p, where

$$\tilde{\gamma} \triangleq \frac{1-\rho}{\max\{96\mathcal{Q}, 128\|A\|_F^2/\rho\}\bar{\delta}^2} \leq \gamma \triangleq \frac{1-\rho}{\max\{\bar{\alpha}, 2\sigma_R^2, 4\sigma_T^2\}}. \tag{4.24}$$

and $\bar{\delta} \triangleq \{\delta_x, \delta_y\}$. We can further simplify the above bound as follows:

$$\begin{aligned}
V_n \leq &\left(\frac{1+\rho}{2}\right)^n \left(\mathcal{E}_{\tau,\sigma} + 48\mathcal{Q}\bar{\delta}^2\right) + \left(\frac{1+\rho}{2}\right)^{2n} \frac{(1-\rho)(2-\rho)}{4}\mathcal{E}_{\tau,\sigma} \\
&+ \frac{1}{1-\rho}96\mathcal{Q}\bar{\delta}^2\left(1 + \max\left\{1,\ \frac{4}{3\rho}\frac{\|A\|_F^2}{\mathcal{Q}}\right\}\log\left(\frac{1}{1-p}\right)\right)
\end{aligned} \tag{4.25}$$

Finally, using the crude bound $\left(\frac{1+\rho}{2}\right)^n(1-\rho)(2-\rho)/4 \leq 1/4$ for $n \geq 1$, the desired result with $\Xi_{\tau,\sigma,\theta}^{(1)} = 48\mathcal{Q}$ and $\Xi_{\tau,\sigma,\theta}^{(2)} = \max\left\{1,\ \frac{4}{3\rho}\frac{\|A\|_F^2}{\mathcal{Q}}\right\}$ follows from (4.20).

4.3 Intermediate Results

Almost sure domination of SAPD iterates. Our analysis starts by leveraging an almost sure upper bound on the sequence of iterates' distances to the solution $(\mathrm{x}^\star, \mathrm{y}^\star)$. This bound, which appears in substance in [38], will serve as the foundation of our risk-averse analysis.

Proposition 4.3. *Let (x_n, y_n) be the sequence generated by SAPD, intialized at an arbitrary tuple $(x_{-1}, y_{-1}) = (x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$. Provided that there exists $\tau, \sigma > 0$, and $\theta \geq 0$ that satisfy (3.1) for some $\rho \in (0,1)$ and $\alpha \in [0, \sigma^{-1})$, we have*

$$\mathcal{E}_n \leq \rho^{n-1}\mathcal{E}_{\tau,\sigma} + \sum_{k=0}^{n-1} \rho^{n-1-k} \left(\langle \Delta_k^x, \mathrm{x}^\star - x_{k+1} \rangle + \langle (1+\theta)\Delta_k^y - \theta \Delta_{k-1}^y, y_{k+1} - \mathrm{y}^\star \rangle \right), \tag{4.26}$$

where $\mathcal{E}_n \triangleq \frac{1}{2\rho\tau}\|x_n - \mathrm{x}^\star\|^2 + \frac{1-\alpha\sigma}{2\rho\sigma}\|y_n - \mathrm{y}^\star\|^2$, and $\mathcal{E}_{\tau,\sigma} \triangleq \frac{1}{2\tau}\|x_0 - \mathrm{x}^\star\|^2 + \frac{1}{2\sigma}\|y_0 - \mathrm{y}^\star\|^2$.

Although this bound is already present in substance in [38], it does not appear explicitly. For completeness, we provide below a minimal proof based on various arguments developed in [38].

Proof. Letting $\bar{x}_n \triangleq K_n(\rho)^{-1} \sum_{k=0}^{n-1} \rho^{-k} x_{k+1}$, and $\bar{y}_n \triangleq K_n(\rho)^{-1} \sum_{k=0}^{n-1} \rho^{-k} y_{k+1}$, with $K_n(\rho) \triangleq \sum_{k=0}^{n-1} \rho^{-k} = \frac{1}{\rho^{n-1}} \cdot \frac{1-\rho^n}{1-\rho}$, by Jensen's inequality , we have for all $\rho \in (0,1]$,

$$K_n(\rho) \left(\mathcal{L}(\bar{x}_n, \mathrm{y}^\star) - \mathcal{L}(\mathrm{x}^\star, \bar{y}_n) \right) \leq \sum_{k=0}^{n-1} \rho^{-k} \left(\mathcal{L}(x_{k+1}, \mathrm{y}^\star) - \mathcal{L}(\mathrm{x}^\star, y_{k+1}) \right).$$

Hence, in view of Lemma 5.1,

$$\begin{aligned} &K_n(\rho) \left(\mathcal{L}(\bar{x}_n, \mathrm{y}^\star) - \mathcal{L}(\mathrm{x}^\star, \bar{y}_n) \right) \\ &\leq \sum_{k=0}^{n-1} \rho^{-k} \Big(- \langle q_{k+1}, y_{k+1} - \mathrm{y}^\star \rangle + \theta \langle q_k, y_k - \mathrm{y}^\star \rangle + \Lambda_k - \Sigma_{k+1} + \Gamma_{k+1} \\ &\quad + \langle \Delta_k^x, \mathrm{x}^\star - x_{k+1} \rangle + \langle (1+\theta)\Delta_k^y - \theta \Delta_{k-1}^y, y_{k+1} - \mathrm{y}^\star \rangle \Big), \end{aligned} \tag{4.27}$$

where $q_k \triangleq \nabla_{\mathrm{y}} \Phi(x_k, y_k) - \nabla_{\mathrm{y}} \Phi(x_{k-1}, y_{k-1})$. By Cauchy-Schwarz inequality, observe that

$$|\langle q_{k+1}, y_{k+1} - \mathrm{y}^\star \rangle| \leq S_{k+1} \triangleq L_{yx} \|x_{k+1} - x_k\| \|y_{k+1} - y\| + L_{yy} \|y_{k+1} - y_k\| \|y_{k+1} - \mathrm{y}^\star\|, \quad \forall k \geq 0.$$

Hence, using $q_0 = \mathbf{0}$ due to our initialization of $(x_{-1}, y_{-1}) = (x_0, y_0)$, we have

$$\begin{aligned} &\sum_{k=0}^{n-1} \rho^{-k} \left(- \langle q_{k+1}, y_{k+1} - \mathrm{y}^\star \rangle + \theta \langle q_k, y_k - \mathrm{y}^\star \rangle \right) = \sum_{k=0}^{n-2} \rho^{-k} \left(\frac{\theta}{\rho} - 1 \right) \langle q_{k+1}, y_{k+1} - \mathrm{y}^\star \rangle - \rho^{-n+1} \langle q_n, y_n - \mathrm{y}^\star \rangle \\ &\leq \sum_{k=0}^{n-2} \rho^{-k} \left| 1 - \frac{\theta}{\rho} \right| S_{k+1} + \rho^{-n+1} S_n \leq \sum_{k=0}^{n-1} \rho^{-k} \left| 1 - \frac{\theta}{\rho} \right| S_{k+1} + \rho^{-n+1} \frac{\theta}{\rho} S_n \end{aligned}$$

From (4.27), it follows that

$$K_n(\rho) \left(\mathcal{L}(\bar{x}_n, \mathrm{y}^\star) - \mathcal{L}(\mathrm{x}^\star, \bar{y}_n) \right) + \rho^{-n+1} \mathcal{E}_n \leq U_n + \sum_{k=0}^{n-1} \rho^{-k} \Big(\langle \Delta_k^x, \mathrm{x}^\star - x_{k+1} \rangle + \langle (1+\theta)\Delta_k^y - \theta \Delta_{k-1}^y, y_{k+1} - \mathrm{y}^\star \rangle \Big),$$

where

$$U_n \triangleq \sum_{k=0}^{n-1} \rho^{-k} \left(\Gamma_{k+1} + \Lambda_k - \Sigma_{k+1} + \left| 1 - \frac{\theta}{\rho} \right| S_{k+1} \right) - \rho^{-n+1} \left(-\mathcal{E}_n - \frac{\theta}{\rho} S_n \right).$$

Now, observe that for all $n \geq 1$,

$$\begin{aligned} U_n &= \frac{1}{2} \sum_{k=0}^{n-1} \rho^{-k} \left(\xi_k^\top A \xi_k - \xi_{k+1}^\top B \xi_{k+1} \right) - \rho^{-n+1} \left(-\mathcal{E}_n - \frac{\theta}{\rho} S_n \right) \\ &= \frac{1}{2} \xi_0^\top A \xi_0 - \frac{1}{2} \sum_{k=1}^{n-1} \rho^{-k+1} \left[\xi_k^\top \left(B - \frac{1}{\rho} A \right) \xi_k \right] - \rho^{-n+1} \left(\frac{1}{2} \xi_n^\top B \xi_n - \mathcal{E}_n - \frac{\theta}{\rho} S_n \right), \end{aligned}$$

where $A, B \in \mathbb{R}^{5\times 5}$ and $\xi_k \in \mathbb{R}^5$ are defined for $k \geq 0$ as

$$A \triangleq \begin{pmatrix} \frac{1}{\tau} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sigma} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta L_{yx} \\ 0 & 0 & 0 & 0 & \theta L_{yy} \\ 0 & 0 & \theta L_{yx} & \theta L_{yy} & -\alpha \end{pmatrix}, \quad \xi_k \triangleq \begin{pmatrix} \|x_k - \mathrm{x}^\star\| \\ \|y_k - \mathrm{y}^\star\| \\ \|x_k - x_{k-1}\| \\ \|y_k - y_{k-1}\| \\ \|y_{k+1} - y_k\| \end{pmatrix}$$

$$B \triangleq \begin{pmatrix} \frac{1}{\tau} + \mu_x & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sigma} + \mu_y & -\left|1 - \frac{\theta}{\rho}\right| L_{yx} & -\left|1 - \frac{\theta}{\rho}\right| L_{yy} & 0 \\ 0 & -\left|1 - \frac{\theta}{\rho}\right| L_{yx} & \frac{1}{\tau} - L_{xx} & 0 & 0 \\ 0 & -\left|1 - \frac{\theta}{\rho}\right| L_{yy} & 0 & \frac{1}{\sigma} - \alpha & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

By [38, Lemma 5], the matrix inequality condition (3.1) is equivalent to having $B - \rho^{-1}A \succeq 0$. In this case, we almost surely have

$$U_n \leq \frac{1}{2}\xi_0^\top A \xi_0 - \rho^{-n+1}\left(\frac{1}{2}\xi_n^\top B \xi_n - \mathcal{E}_n - \frac{\theta}{\rho}S_n\right). \tag{4.28}$$

Finally, denoting

$$G'' \triangleq \begin{pmatrix} \frac{1}{\sigma}\left(1 - \frac{1}{\rho}\right) + \mu_y + \frac{\alpha}{\rho} & \left(-\left|1 - \frac{\theta}{\rho}\right| - \frac{\theta}{\rho}\right) L_{yx} & \left(-\left|1 - \frac{\theta}{\rho}\right| - \frac{\theta}{\rho}\right) L_{yy} \\ \left(-\left|1 - \frac{\theta}{\rho}\right| - \frac{\theta}{\rho}\right) L_{yx} & \frac{1}{\tau} - L_{xx} & 0 \\ \left(-\left|1 - \frac{\theta}{\rho}\right| - \frac{\theta}{\rho}\right) L_{yy} & 0 & \frac{1}{\sigma} - \alpha \end{pmatrix},$$

we have $G'' \succeq 0$ in view of [38, Lemma 6]; thus,

$$\begin{aligned} &\frac{1}{2}\xi_n^\top B \xi_n - \frac{\theta}{\rho}S_n \\ =&\frac{1}{2\rho\tau}\|x_n - x\|^2 + \frac{1}{2}\left(\frac{1}{\rho\sigma} - \frac{\alpha}{\rho}\right)\|y_n - y\|^2 + \frac{1}{2}\xi_n^\top \begin{pmatrix} \frac{1}{\tau}\left(1 - \frac{1}{\rho}\right) + \mu_x & \mathbf{0}_{1\times 3} & 0 \\ \mathbf{0}_{3\times 1} & G'' & \mathbf{0}_{3\times 1} \\ 0 & \mathbf{0}_{1\times 3} & 0 \end{pmatrix} \xi_n \\ \geq&\frac{1}{2\rho\tau}\|x_n - x\|^2 + \frac{1}{2\rho\sigma}(1 - \alpha\sigma)\|y_n - y\|^2 = \mathcal{E}_n \end{aligned}$$

Therefore, using (4.28), we can conclude that $U_n \leq \frac{1}{2}\xi_0^\top A \xi_0 \leq \frac{1}{2\tau}\|x_0 - \mathrm{x}^\star\|^2 + \frac{1}{2\sigma}\|y_0 - \mathrm{y}^\star\|^2 = \mathcal{E}_{\tau,\sigma}$. Finally, by non-negativity of $\mathcal{L}(\bar{x}_n, \mathrm{y}^\star) - \mathcal{L}(\mathrm{x}^\star, \bar{y}_n)$, we obtain (4.26). □

Lemma 4.4. *For any $n \geq 0$, Check if it holds for $n = 0$.*

$$\|\hat{\mathrm{y}}_{n+1} - \mathrm{y}^\star\| \leq \|A_0\|(\mathcal{E}_n + (1-\rho)\mathcal{E}_{n-1})^{1/2} + \frac{1}{1 + \sigma\mu_y}\left((1 + \sigma(1+\theta)\,\mathrm{L_{yy}})\|y_n - \hat{\mathrm{y}}_n\| + \sigma(1+\theta)\,\mathrm{L_{yx}}\|x_n - \hat{\mathrm{x}}_n\|\right)$$

where A_0 as defined in Table 2.

Proof. In view of 4.30 and Lemma B.2, we have

$$\|\hat{\mathrm{y}}_{n+1} - \mathrm{y}^\star\| \leq \frac{1}{1 + \sigma\mu_y}\left\|\hat{\mathrm{y}}_n + \sigma(1+\theta)\,\nabla_{\mathrm{y}}\,\Phi(\hat{\mathrm{x}}_n, \hat{\mathrm{y}}_n) - \sigma\theta\,\nabla_{\mathrm{y}}\,\Phi(x_{n-1}, y_{n-1}) - \mathrm{y}^\star - \sigma\,\nabla_{\mathrm{y}}\,\Phi(\mathrm{x}^\star, \mathrm{y}^\star)\right\|$$

By the triangular inequality and smoothness assumptions on $\nabla_{\mathrm{y}}\,\Phi$, we deduce

$$\begin{aligned} \|\hat{\mathrm{y}}_{n+1} - \mathrm{y}^\star\| &\leq \frac{1}{1 + \sigma\mu_y}\left((1 + \sigma(1+\theta)\,\mathrm{L_{yy}})\|\hat{\mathrm{y}}_n - \mathrm{y}^\star\| + \sigma(1+\theta)\,\mathrm{L_{yx}}\|\hat{\mathrm{x}}_n - \mathrm{x}^\star\| + \sigma\theta\,\mathrm{L_{yx}}\|x_{n-1} - \mathrm{x}^\star\| + \sigma\theta\,\mathrm{L_{yy}}\|y_{n-1} - \mathrm{y}^\star\|\right) \\ &\leq \frac{1}{1 + \sigma\mu_y}\left((1 + \sigma(1+\theta)\,\mathrm{L_{yy}})\|y_n - \mathrm{y}^\star\| + \sigma(1+\theta)\,\mathrm{L_{yx}}\|x_n - \mathrm{x}^\star\| + \sigma\theta\,\mathrm{L_{yx}}\|x_{n-1} - \mathrm{x}^\star\| + \sigma\theta\,\mathrm{L_{yy}}\|y_{n-1} - \mathrm{y}^\star\|\right) \\ &\quad + \frac{1}{1 + \sigma\mu_y}\left((1 + \sigma(1+\theta)\,\mathrm{L_{yy}})\|\hat{\mathrm{y}}_n - y_n\| + \sigma(1+\theta)\,\mathrm{L_{yx}}\|\hat{\mathrm{x}}_n - x_n\|\right) \end{aligned}$$

The statement finally follows from Cauchy-Schwarz inequality. □

Lemma 4.5 (See Lemma 3 - [38]). *For any $n \in \mathbb{N}$, we have*

$$
\begin{aligned}
\|\hat{\mathrm{x}}_{n+1} - x_{n+1}\| &\leq \frac{\tau}{1+\tau\mu_x}\|\Delta_n^x\| \\
\|\hat{\mathrm{y}}_{n+1} - y_{n+1}\| &\leq \frac{\sigma}{1+\sigma\mu_y}\left((1+\theta)\|\Delta_n^y\| + \theta\|\Delta_{n-1}^y\|\right) \\
\|\hat{\hat{\mathrm{x}}}_{n+1} - x_{n+1}\| &\leq \frac{\tau}{1+\tau\mu_x}\left(\|\Delta_n^x\| + \mathrm{L_{xy}}\frac{\sigma}{1+\sigma\mu_y}\left((1+\theta)\|\Delta_n^y\| + \theta\|\Delta_{n-1}^y\|\right)\right) \\
\|\hat{\hat{\mathrm{y}}}_{n+1} - y_{n+1}\| &\leq \frac{\sigma}{1+\sigma\mu_y}\left(\frac{\tau(1+\theta)\,\mathrm{L_{yx}}}{1+\tau\mu_x}\|\Delta_{n-1}^x\|\right. \\
&\qquad + (1+\theta)\|\Delta_n^y\| \\
&\qquad + \left(\theta + (1+\theta)\left(\frac{1+\sigma(1+\theta)\,\mathrm{L_{yy}}}{1+\sigma\mu_y} + \frac{\tau\sigma(1+\theta)\,\mathrm{L_{yx}}\,\mathrm{L_{xy}}}{(1+\tau\mu_x)(1+\sigma\mu_y)}\right)\right)\|\Delta_{n-1}^y\| \\
&\qquad \left. + \theta\left(\frac{1+\sigma(1+\theta)\,\mathrm{L_{yy}}}{1+\sigma\mu_y} + \frac{\tau\sigma(1+\theta)\,\mathrm{L_{yx}}\,\mathrm{L_{xy}}}{(1+\tau\mu_x)(1+\sigma\mu_y)}\right)\|\Delta_{n-2}^y\|\right)
\end{aligned}
$$

The following proposition plays a key role for the introduction of the scalar process V_n (see Equation (4.19)) in the proof of Theorem 3.1.

Proposition 4.6. *For any $n \in \mathbb{N}$,is this bound defined for $n = 0$? Otherwise, we should say for $n \geq 1$.*

$$
\begin{aligned}
&(\rho-1)P_n^{(1)} - P_n^{(2)} + \rho\,(\rho-1)\,P_{n-1}^{(2)} + \rho Q_n \\
&\qquad \leq \frac{\rho}{2}(\mathcal{E}_{n+1} + (1-\rho)\mathcal{E}_n) + \mathcal{Q}\left(\|\Delta_n^x\|^2 + \|\Delta_n^y\|^2 + \rho\left(\|\Delta_{n-1}^x\|^2 + \|\Delta_{n-1}^y\|^2\right) + \rho^2\|\Delta_{n-2}^y\|^2\right)
\end{aligned}
$$

and for any $k \geq 0$,

$$
Q_k \leq \mathcal{Q}\left(\|\Delta_k^x\|^2 + \|\Delta_k^y\|^2 + \rho\left(\|\Delta_{k-1}^x\|^2 + \|\Delta_{k-1}^y\|^2\right) + \rho^2\|\Delta_{k-2}^y\|^2\right),
$$

where $\Delta_{-1}^x \triangleq \mathbf{0}$ and $\Delta_{-1}^y = \Delta_{-2}^y \triangleq \mathbf{0}$. and $\mathcal{Q}$ is given explicitly in Table 2.

Proof. By Young's inequality, we first note that that for any $\gamma_x, \gamma_y > 0$,

$$
\begin{aligned}
&(\rho-1)P_n^{(1)} + \rho\langle\Delta_n^x, \hat{\hat{\mathrm{x}}}_{n+1} - x_{n+1}\rangle + \rho(1+\theta)\,\langle\Delta_n^y, y_{n+1} - \hat{\mathrm{y}}_{n+1}\rangle \\
&\quad = (\rho-1)\,\left(\langle\Delta_n^x, \mathrm{x}^\star - \hat{\hat{\mathrm{x}}}_{n+1}\rangle + (1+\theta)\langle\Delta_n^y, \hat{\mathrm{y}}_{n+1} - \mathrm{y}^\star\rangle\right) \\
&\qquad\quad + \rho\langle\Delta_n^x, \hat{\hat{\mathrm{x}}}_{n+1} - x_{n+1}\rangle + \rho(1+\theta)\,\langle\Delta_n^y, y_{n+1} - \hat{\mathrm{y}}_{n+1}\rangle \\
&\quad = (\rho-1)\,\left(\langle\Delta_n^x, \mathrm{x}^\star - x_{n+1}\rangle + (1+\theta)\langle\Delta_n^y, y_{n+1} - \mathrm{y}^\star\rangle\right) \\
&\qquad\quad + \langle\Delta_n^x, \hat{\hat{\mathrm{x}}}_{n+1} - x_{n+1}\rangle + (1+\theta)\,\langle\Delta_n^y, y_{n+1} - \hat{\mathrm{y}}_{n+1}\rangle \\
&\quad \leq \frac{\gamma_x(1-\rho)}{2}\|\Delta_n^x\|^2 + \frac{(1-\rho)}{2\gamma_x}\|\mathrm{x}^\star - x_{n+1}\|^2 + \frac{(1-\rho)\gamma_y(1+\theta)}{2}\|\Delta_n^y\|^2 + \frac{(1-\rho)(1+\theta)}{2\gamma_y}\|y_{n+1} - \mathrm{y}^\star\|^2 \\
&\qquad\quad + \langle\Delta_n^x, \hat{\hat{\mathrm{x}}}_{n+1} - x_{n+1}\rangle + (1+\theta)\,\langle\Delta_n^y, y_{n+1} - \hat{\mathrm{y}}_{n+1}\rangle.
\end{aligned}
$$

Setting $\gamma_x \triangleq 8\tau(1-\rho)$ and $\gamma_y \triangleq 8\frac{\sigma(1-\rho)(1+\theta)}{1-\alpha\sigma}$, we ensure that

$$
\begin{aligned}
&(\rho-1)\,P_n^{(1)} + \rho\langle\Delta_n^x, \hat{\hat{\mathrm{x}}}_{n+1} - x_{n+1}\rangle + (1+\theta)\rho\,\langle\Delta_n^y, y_{n+1} - \hat{\mathrm{y}}_{n+1}\rangle \\
&\qquad \leq \frac{\rho}{8}\mathcal{E}_{n+1} + 4\tau(1-\rho)^2\|\Delta_n^x\|^2 + \frac{4\sigma(1+\theta)^2(1-\rho)^2}{(1-\alpha\sigma)}\|\Delta_n^y\|^2 \\
&\qquad\qquad + \langle\Delta_n^x, \hat{\hat{\mathrm{x}}}_{n+1} - x_{n+1}\rangle + (1+\theta)\,\langle\Delta_n^y, y_{n+1} - \hat{\mathrm{y}}_{n+1}\rangle.
\end{aligned}
$$

Similarly, we have

$$
\rho\,(\rho-1)\,P_{n-1}^{(2)} - \theta\rho\langle\Delta_{n-1}^y, y_{n+1} - \hat{\hat{\mathrm{y}}}_{n+1}\rangle \leq \rho\left((\rho-1)\frac{\theta}{\rho}\theta\langle\Delta_{n-2}^y, \mathrm{y}^\star - \hat{\hat{\mathrm{y}}}_n\rangle\right) - \theta\rho\langle\Delta_{n-1}^y, y_{n+1} - \hat{\hat{\mathrm{y}}}_{n+1}\rangle
$$

Hence,

$$\rho\,(\rho-1)\,P_{n-1}^{(2)}-\theta\rho\langle\Delta_{n-1}^{y},y_{n+1}-\hat{\hat{\mathrm{y}}}_{n+1}\rangle\leq\frac{\rho}{8}\mathcal{E}_{n+1}+\frac{4\sigma\theta^{2}(1-\rho)^{2}}{1-\alpha\sigma}\|\Delta_{n-1}^{y}\|^{2}-\theta\langle\Delta_{n-1}^{y},y_{n+1}-\hat{\hat{\mathrm{y}}}_{n+1}\rangle.$$

Finally, observe that for any $\gamma>0$,

$$\begin{aligned}-P_{n}^{(2)}&\leq\frac{\gamma}{2}\|\Delta_{n}^{y}\|^{2}+\frac{1}{2\gamma}\|\hat{\hat{\mathrm{y}}}_{n+2}-\mathrm{y}^{\star}\|^{2}\\&\leq\frac{\gamma}{2}\|\Delta_{n}^{y}\|^{2}+\frac{1}{2\gamma}\Bigg(3\|A_{0}\|^{2}(\mathcal{E}_{n+1}+(1-\rho)\mathcal{E}_{n})+3\Big(\frac{1+\sigma(1+\theta)\,\mathrm{L_{yy}}}{1+\sigma\mu_{y}}\Big)^{2}\|\hat{\mathrm{y}}_{n+1}-y_{n+1}\|^{2}\\&\qquad+3\Big(\frac{\sigma(1+\theta)\,\mathrm{L_{yx}}}{1+\sigma\mu_{y}}\Big)^{2}\|\hat{\hat{\mathrm{x}}}_{n+1}-x_{n+1}\|^{2}\Bigg),\end{aligned}$$

where the last inequality follows from Lemma 4.4 and Lemma B.1. Setting $\gamma\triangleq\frac{6\|A_{0}\|^{2}}{\rho}$ ensures that

$$\begin{aligned}-P_{n}^{(2)}\leq\frac{\rho}{4}(\mathcal{E}_{n+1}+(1-\rho)\mathcal{E}_{n})+\frac{3\|A_{0}\|^{2}}{\rho}\|\Delta_{n}^{y}\|^{2}+\frac{\rho}{4\|A_{0}\|^{2}}\left(\frac{1+\sigma(1+\theta)\,\mathrm{L_{yy}}}{1+\sigma\mu_{y}}\right)^{2}\|\hat{\mathrm{y}}_{n+1}-y_{n+1}\|^{2}\\+\frac{\rho}{4\|A_{0}\|^{2}}\left(\frac{\sigma(1+\theta)\,\mathrm{L_{yx}}}{1+\sigma\mu_{y}}\right)^{2}\|\hat{\hat{\mathrm{x}}}_{n+1}-x_{n+1}\|^{2}.\end{aligned}$$

Hence, combining all the bounds we obtained above, we getThis bound is correct; though, it uses a loose bound $\frac{\rho}{8}\mathcal{E}_{n+1}\leq\frac{\rho}{4}(\mathcal{E}_{n+1}+(1-\rho)\mathcal{E}_{n})$.

$$\begin{aligned}&(\rho-1)P_{n}^{(1)}-P_{n}^{(2)}+\rho\,(\rho-1)\,P_{n-1}^{(2)}+\rho Q_{n}\\&\quad\leq\frac{\rho}{2}(\mathcal{E}_{n+1}+(1-\rho)\mathcal{E}_{n})+\langle\Delta_{n}^{x},\hat{\hat{\mathrm{x}}}_{n+1}-x_{n+1}\rangle+(1+\theta)\,\langle\Delta_{n}^{y},y_{n+1}-\hat{\mathrm{y}}_{n+1}\rangle-\theta\langle\Delta_{n-1}^{y},y_{n+1}-\hat{\hat{\mathrm{y}}}_{n+1}\rangle\\&\qquad+4\tau(1-\rho)^{2}\|\Delta_{n}^{x}\|^{2}+\frac{4\sigma(1+\theta)^{2}(1-\rho)^{2}}{1-\alpha\sigma}\|\Delta_{n}^{y}\|^{2}+\frac{4\sigma\theta^{2}(1-\rho)^{2}}{1-\alpha\sigma}\|\Delta_{n-1}^{y}\|^{2}+\frac{3\|A_{0}\|^{2}}{\rho}\|\Delta_{n}^{y}\|^{2}\\&\qquad+\frac{\rho}{4\|A_{0}\|^{2}}\left(\frac{1+\sigma(1+\theta)\,\mathrm{L_{yy}}}{1+\sigma\mu_{y}}\right)^{2}\|\hat{\mathrm{y}}_{n+1}-y_{n+1}\|^{2}+\frac{\rho}{4\|A_{0}\|^{2}}\left(\frac{\sigma(1+\theta)\,\mathrm{L_{yx}}}{1+\sigma\mu_{y}}\right)^{2}\|\hat{\hat{\mathrm{x}}}_{n+1}-x_{n+1}\|^{2}.\end{aligned}$$

Let us now introduce $\zeta_{k}\triangleq\Big[\|\Delta_{k}^{x}\|,\ \rho^{1/2}\|\Delta_{k-1}^{x}\|,\ \|\Delta_{k}^{y}\|,\ \rho^{1/2}\|\Delta_{k-1}^{y}\|,\ \rho\|\Delta_{k-2}^{y}\|\Big]^{\top}\in\mathbb{R}^{5}$ for $k\geq0$; then, the following bounds follow from Lemma 4.5 and Cauchy-Schwartz inequality:

$$\begin{aligned}\Big(\langle\Delta_{n}^{x},\hat{\hat{\mathrm{x}}}_{n+1}-x_{n+1}\rangle+(1+\theta)\,\langle\Delta_{n}^{y},y_{n+1}-\hat{\mathrm{y}}_{n+1}\rangle-\theta\langle\Delta_{n-1}^{y},y_{n+1}-\hat{\hat{\mathrm{y}}}_{n+1}\rangle\Big)&\leq\zeta_{n}^{\top}B_{0}\zeta_{n}\\\frac{\rho}{4\|A_{0}\|^{2}}\left(\left(\frac{1+\sigma(1+\theta)\,\mathrm{L_{yy}}}{1+\sigma\mu_{y}}\right)^{2}\|\hat{\mathrm{y}}_{n+1}-y_{n+1}\|^{2}+\left(\frac{\sigma(1+\theta)\,\mathrm{L_{yx}}}{1+\sigma\mu_{y}}\right)^{2}\|\hat{\hat{\mathrm{x}}}_{n+1}-x_{n+1}\|^{2}\right)&\leq\zeta_{n}^{\top}B_{1}\zeta_{n}\\4\tau(1-\rho)^{2}\|\Delta_{n}^{x}\|^{2}+\frac{4\sigma(1+\theta)^{2}(1-\rho)^{2}}{1-\alpha\sigma}\|\Delta_{n}^{y}\|^{2}+\frac{4\sigma\theta^{2}(1-\rho)^{2}}{1-\alpha\sigma}\|\Delta_{n-1}^{y}\|^{2}+\frac{3\|A_{0}\|^{2}}{\rho}\|\Delta_{n}^{y}\|^{2}&\leq\zeta_{n}^{\top}B_{2}\zeta_{n},\end{aligned}$$

where $B_{0},B_{1},B_{2}\in\mathbb{R}^{5\times5}$ are defined in Table 2. Therefore,

$$\begin{aligned}(\rho-1)\,P_{n}^{(1)}-P_{n}^{(2)}+\rho\,(\rho-1)\,P_{n-1}^{(2)}+\rho Q_{n}&\leq\frac{\rho}{2}(\mathcal{E}_{n+1}+(1-\rho)\mathcal{E}_{n})+\zeta_{n}^{\top}(B_{0}+B_{1}+B_{2})\zeta_{n}\\&\leq\frac{\rho}{2}(\mathcal{E}_{n+1}+(1-\rho)\mathcal{E}_{n})+\|B_{0}+B_{1}+B_{2}\|_{F}\,\|\zeta_{n}\|^{2}.\end{aligned}$$

Notice finally that $\rho\leq1$ and B_{0}, B_{1} and B_{2} all have non-negative coefficients, so that for all $k\geq0$

$$Q_{k}\leq\zeta_{k}^{\top}B_{0}\zeta_{k}\leq\|B_{0}+B_{1}+B_{2}\|_{F}\,\|\zeta_{k}\|^{2}.$$

□

The following lemma will be used to verify the assumptions of Proposition 4.1 when analyzing SAPD.

Lemma 4.7. *For any $k \in \mathbb{N}$ and for $\rho \in (0,1)$, we have almost surely:*

$$\begin{pmatrix} \|\hat{\mathrm{x}}_{k+1} - \mathrm{x}^\star\| \\ \sqrt{2}(1+\theta)\|\hat{\mathrm{y}}_{k+1} - \mathrm{y}^\star\| \\ \sqrt{2}\frac{\theta}{\rho}\|\hat{\mathrm{y}}_{k+2} - \mathrm{y}^\star\| \end{pmatrix} \leq A \begin{pmatrix} \frac{1}{\sqrt{2\rho\tau}}\|x_k - \mathrm{x}^\star\| \\ \frac{\sqrt{1-\alpha\sigma}}{\sqrt{2\rho\sigma}}\|y_k - \mathrm{y}^\star\| \\ \frac{\sqrt{1-\rho}}{\sqrt{2\rho\tau}}\|x_{k-1} - \mathrm{x}^\star\| \\ \frac{\sqrt{(1-\rho)(1-\alpha\sigma)}}{\sqrt{2\rho\sigma}}\|y_{k-1} - \mathrm{y}^\star\| \end{pmatrix}, \tag{4.29}$$

where the inequality is taken element-wise and A is defined in Table 2.

Proof. Let $k \in \mathbb{N}$ be fixed. Fist note that since $(\mathrm{x}^\star, \mathrm{y}^\star)$ is solution of (1.1), $\mathrm{x}^\star$ and $\mathrm{y}^\star$ are respective fixed points of two deterministic proximal gradient steps:

$$\mathrm{x}^\star = \mathrm{prox}_{\tau f}\left(\mathrm{x}^\star - \tau\, \nabla_{\mathrm{x}}\, \Phi(\mathrm{x}^\star, \mathrm{y}^\star)\right), \quad \mathrm{y}^\star = \mathrm{prox}_{\sigma g}\left(\mathrm{y}^\star + \sigma\, \nabla_{\mathrm{y}}\, \Phi(\mathrm{x}^\star, \mathrm{y}^\star)\right). \tag{4.30}$$

Thus, using Lemma B.2 and Lemma B.1, we obtain

$$\begin{aligned}
\|\hat{\mathrm{y}}_{k+1} - \mathrm{y}^\star\| &\leq \frac{1}{1+\sigma\mu_y}\|y_k + \sigma(1+\theta)\,\nabla_{\mathrm{y}}\,\Phi(x_k, y_k) - \sigma\theta\,\nabla_{\mathrm{y}}\,\Phi(x_{k-1}, y_{k-1}) - \mathrm{y}^\star - \sigma\,\nabla_{\mathrm{y}}\,\Phi(\mathrm{x}^\star, \mathrm{y}^\star)\| \\
&\leq \frac{1}{1+\sigma\mu_y}\Big(\|y_k - \mathrm{y}^\star\| + \sigma(1+\theta)\|\nabla_{\mathrm{y}}\,\Phi(x_k, y_k) - \nabla_{\mathrm{y}}\,\Phi(\mathrm{x}^\star, \mathrm{y}^\star)\| + \theta\sigma\|\nabla_{\mathrm{y}}\,\Phi(x_{k-1}, y_{k-1}) - \nabla_{\mathrm{y}}\,\Phi(\mathrm{x}^\star, \mathrm{y}^\star)\|\Big) \\
&\leq \frac{1}{1+\sigma\mu_y}\Big(\sigma(1+\theta)\,\mathrm{L}_{\mathrm{yx}}\|x_k - \mathrm{x}^\star\| + (1+\sigma(1+\theta)\,\mathrm{L}_{\mathrm{yy}})\|y_k - \mathrm{y}^\star\| \\
&\qquad + \sigma\theta\,\mathrm{L}_{\mathrm{yx}}\|x_{k-1} - \mathrm{x}^\star\| + \sigma\theta\,\mathrm{L}_{\mathrm{yy}}\|y_{k-1} - \mathrm{y}^\star\|\Big),
\end{aligned}$$

where the third inequality follows from the smoothness assumptions on $\nabla_{\mathrm{x}}\,\Phi$ and $\nabla_{\mathrm{y}}\,\Phi$. Through analogous algebraic computations, we also obtain

$$\begin{aligned}
\|\hat{\mathrm{x}}_{k+1} - \mathrm{x}^\star\| &\leq \frac{1}{1+\tau\mu_x}\Big((1+\tau\,\mathrm{L}_{\mathrm{xx}})\|x_k - \mathrm{x}^\star\| + \tau\,\mathrm{L}_{\mathrm{xy}}\|\hat{\mathrm{y}}_{k+1} - \mathrm{y}^\star\|\Big) \\
&\leq \frac{1}{1+\tau\mu_x}\left(\left(1+\tau\,\mathrm{L}_{\mathrm{xx}} + \frac{\tau\sigma(1+\theta)\,\mathrm{L}_{\mathrm{yx}}\,\mathrm{L}_{\mathrm{xy}}}{1+\sigma\mu_y}\right)\|x_k - \mathrm{x}^\star\| + \frac{\tau\sigma\theta\,\mathrm{L}_{\mathrm{xy}}\,\mathrm{L}_{\mathrm{yx}}}{1+\sigma\mu_y}\|x_{k-1} - \mathrm{x}^\star\|\right. \\
&\qquad \left. + \frac{\tau\,\mathrm{L}_{\mathrm{xy}}(1+\sigma(1+\theta)\,\mathrm{L}_{\mathrm{yy}})}{1+\sigma\mu_y}\|y_k - \mathrm{y}^\star\| + \frac{\tau\sigma\theta\,\mathrm{L}_{\mathrm{xy}}\,\mathrm{L}_{\mathrm{yy}}}{1+\sigma\mu_y}\|y_{k-1} - \mathrm{y}^\star\|\right),
\end{aligned}$$

from which we deduce the following bound:

$$\begin{aligned}
\|\hat{\mathrm{y}}_{k+2} - \mathrm{y}^\star\| &\leq \frac{1}{1+\sigma\mu_y}\Big(\|\hat{\mathrm{y}}_{k+1} - \mathrm{y}^\star\| + \sigma(1+\theta)\,\mathrm{L}_{\mathrm{yx}}\|\hat{\mathrm{x}}_{k+1} - \mathrm{x}^\star\| + \sigma(1+\theta)\,\mathrm{L}_{\mathrm{yy}}\|\hat{\mathrm{y}}_{k+1} - \mathrm{y}^\star\| \\
&\qquad + \sigma\theta\,\mathrm{L}_{\mathrm{yx}}\|x_k - \mathrm{x}^\star\| + \sigma\theta\,\mathrm{L}_{\mathrm{yy}}\|y_k - \mathrm{y}^\star\|\Big) \\
&\leq \frac{1}{1+\sigma\mu_y}\left(\left((1+\sigma(1+\theta)\,\mathrm{L}_{\mathrm{yy}})\frac{\sigma(1+\theta)\,\mathrm{L}_{\mathrm{yx}}}{1+\sigma\mu_y} + \sigma(1+\theta)\,\mathrm{L}_{\mathrm{yx}}\,\frac{\left(1+\tau\,\mathrm{L}_{\mathrm{xx}} + \frac{\tau\sigma(1+\theta)\,\mathrm{L}_{\mathrm{yx}}\,\mathrm{L}_{\mathrm{xy}}}{1+\sigma\mu_y}\right)}{1+\tau\mu_x} + \sigma\theta\,\mathrm{L}_{\mathrm{yx}}\right)\|x_k - \mathrm{x}^\star\|\right. \\
&\qquad + \left(\frac{(1+\sigma(1+\theta)\,\mathrm{L}_{\mathrm{yy}})^2}{1+\sigma\mu_y} + \sigma(1+\theta)\,\mathrm{L}_{\mathrm{yx}}\,\frac{\tau\,\mathrm{L}_{\mathrm{xy}}(1+\sigma(1+\theta)\,\mathrm{L}_{\mathrm{yy}})}{(1+\sigma\mu_y)(1+\tau\mu_x)} + \sigma\theta\,\mathrm{L}_{\mathrm{yy}}\right)\|y_k - \mathrm{y}^\star\| \\
&\qquad + \left((1+\sigma(1+\theta)\,\mathrm{L}_{\mathrm{yy}})\frac{\theta\sigma\,\mathrm{L}_{\mathrm{yx}}}{(1+\sigma\mu_y)} + \sigma(1+\theta)\frac{\tau\sigma\theta\,\mathrm{L}_{\mathrm{xy}}\,\mathrm{L}_{\mathrm{yx}}{}^2}{(1+\sigma\mu_y)(1+\tau\mu_x)}\right)\|x_{k-1} - \mathrm{x}^\star\| \\
&\qquad \left. + \left(\frac{(1+\sigma(1+\theta)\,\mathrm{L}_{\mathrm{yy}})\sigma\theta\,\mathrm{L}_{\mathrm{yy}}}{1+\sigma\mu_y} + \sigma(1+\theta)\frac{\tau\sigma\theta\,\mathrm{L}_{\mathrm{xy}}\,\mathrm{L}_{\mathrm{yx}}\,\mathrm{L}_{\mathrm{yy}}}{(1+\sigma\mu_y)(1+\tau\mu_x)}\right)\|y_{k-1} - \mathrm{y}^\star\|\right)
\end{aligned}$$

□

5 Additional Lemmas

In relationship to high-probability convergence results. We first recall the following descent Lemma, used to derive the almost sure bound 4.3.

Lemma 5.1 (See [38]Lemma 1)**.** *The iterates (x_k, y_k) of SAPD satisfy for all $k \geq 0$*

$$\begin{aligned}\mathcal{L}\left(x_{k+1}, \mathrm{y}^{\star}\right) - \mathcal{L}\left(\mathrm{x}^{\star}, y_{k+1}\right) \leq & -\left\langle q_{k+1}, y_{k+1} - \mathrm{y}^{\star}\right\rangle + \theta\left\langle q_k, y_k - \mathrm{y}^{\star}\right\rangle + \Lambda_k(\mathrm{x}^{\star}, \mathrm{y}^{\star}) - \Sigma_{k+1}(\mathrm{x}^{\star}, \mathrm{y}^{\star}) + \Gamma_{k+1} \\ & + \left\langle \Delta_k^x, \mathrm{x}^{\star} - x_{k+1}\right\rangle + \left\langle (1+\theta)\Delta_k^y - \theta\Delta_{k-1}^y, y_{k+1} - \mathrm{y}^{\star}\right\rangle\end{aligned}$$

where

$$\begin{aligned}q_k &\triangleq \nabla_{\mathrm{y}}\Phi\left(x_k, y_k\right) - \nabla_{\mathrm{y}}\Phi\left(x_{k-1}, y_{k-1}\right) \\ \Lambda_k &\triangleq \frac{1}{2\tau}\left\|\mathrm{x}^{\star} - x_k\right\|^2 + \frac{1}{2\sigma}\left\|\mathrm{y}^{\star} - y_k\right\|^2 \\ \Sigma_{k+1} &\triangleq \left(\frac{1}{2\tau} + \frac{\mu_x}{2}\right)\left\|\mathrm{x}^{\star} - x_{k+1}\right\|^2 + \left(\frac{1}{2\sigma} + \frac{\mu_y}{2}\right)\left\|\mathrm{y}^{\star} - y_{k+1}\right\|^2 \\ \Gamma_{k+1} &\triangleq \left(\frac{L_{xx}}{2} - \frac{1}{2\tau}\right)\left\|x_{k+1} - x_k\right\|^2 - \frac{1}{2\sigma}\left\|y_{k+1} - y_k\right\|^2 + \theta L_{yx}\left\|x_k - x_{k-1}\right\|\left\|y_{k+1} - y_k\right\| \\ &\quad + \theta L_{yy}\left\|y_k - y_{k-1}\right\|\left\|y_{k+1} - y_k\right\|.\end{aligned}$$

In relationship to the CVaR upperbound.

Lemma 5.2. *For any non-negative random variable $U\colon \Omega \to \mathbb{R}_+$, we have for all $p \in [0,1)$:*

$$\begin{aligned}Q_p(U^2)^{\frac{1}{2}} &= Q_p(U) \\ \mathrm{CVaR}_p(U^2)^{\frac{1}{2}} &\geq \mathrm{CVaR}_p(U)\end{aligned}$$

Proof. We first show that $Q_p(X^2) = Q_p(X)^2$ for any $p \in (0,1)$. Indeed, for any $0 \leq t < Q_p(U)^2$, by non-negativity of U, we have:

$$\mathbb{P}[U^2 \leq t] = \mathbb{P}[U \leq \sqrt{t}] < p$$

by definition of $Q_p(U)$. This implies $t < Q_p(U^2)$; thus, $Q_p(U)^2 \leq Q_p(U^2)$. Conversely, we note that

$$p \leq \mathbb{P}[U \leq Q_p(U)] = \mathbb{P}[U^2 \leq Q_p(U)^2],$$

which implies $Q_p(U)^2 \geq Q_p(U^2)$; hence, $Q_p(X^2) = Q_p(X)^2$. Using this result, we get

$$\begin{aligned}\mathrm{CVaR}_p(U^2) &= \left(\frac{1}{1-p}\int_{p'=p}^{1} Q_{p'}(U^2)\mathrm{d}p'\right) \\ &= \mathbb{E}_{p' \sim \mathcal{U}[p,1]}[Q_{p'}(U)^2] \\ &\geq \mathbb{E}_{p' \sim \mathcal{U}[p,1]}[Q_{p'}(U)]^2 = \mathrm{CVaR}_p(U)^2,\end{aligned}$$

where $\mathcal{U}[p,1]$ denotes the uniform distribution on $[p,1]$, and the last inequality follows from the identity $\mathbb{E}[X^2] = \mathbb{E}[X]^2 + \mathbb{E}[(X - \mathbb{E}[X])^2]$. □

In relationship to the quadratic case analysis.

Lemma 5.3. *Let ν_1, ν_2, be the two complex conjugate eigenvalues of A^i, as specified in Lemma C.2. Then, the following relations are satisfied*

$$\begin{aligned}\nu_1\nu_2 &= \theta^2 - \theta(1-\theta)^2\kappa^2 \\ \nu_1 + \nu_2 &= 2\theta - (1-\theta)^2(1+\theta)\kappa^2 \\ \nu_1^2 + \nu_2^2 &= 2\theta^2 - 2\theta(1-\theta)^2(1+2\theta)\kappa^2 + (1-\theta)^4(1+\theta)^2\kappa^4 \\ v_1^3 + v_2^3 &= \left(2\theta - (1-\theta)^2(1+\theta)\kappa^2\right)\left(\theta^2 - \theta(1-\theta)^2(1+4\theta)\kappa^2 + (1-\theta)^4(1+\theta)^2\kappa^4\right) \\ v_1^4 + v_2^4 &= 2\theta^4 - (1-\theta)^2\kappa^2\theta^3(4+16\theta) + (1-\theta)^4\kappa^4\theta^2\left(6+24\theta+20\theta^2\right) \\ &\quad - (1-\theta)^6\kappa^6 4\theta\left(1+4\theta+5\theta^2+2\theta^3\right) + (1-\theta)^8\kappa^8(1+\theta)^4\end{aligned}$$

6 Omitted Proofs

We start with elementary proofs of the lemmas 2.1 and 2.2, following standard arguments that can be found in textbooks such as [33, 29].

A Symbols and constants used in the paper

The convergence analysis of SAPD relies on a series of convex inequalities that we wrote in matrix form for compactness. The matrix of coefficients introduced in Proposition 4.6, Lemma4.7 and Lemma 4.4 are made explicit in Table 2.

$$A_0 \triangleq \frac{1}{1+\sigma\mu_y}\begin{bmatrix} \sqrt{2\rho\tau}\sigma(1+\theta)\,\mathrm{L_{yx}} \\ \frac{\sqrt{2\rho\sigma}}{\sqrt{1-\alpha\sigma}}(1+\sigma(1+\theta)\,\mathrm{L_{yy}}) \\ \frac{\sqrt{2\rho\tau}}{\sqrt{1-\rho}}\cdot\sigma\theta\,\mathrm{L_{yx}} \\ \frac{\sqrt{2\rho\sigma}}{\sqrt{(1-\alpha\sigma)(1-\rho)}}\cdot\sigma\theta\,\mathrm{L_{yy}} \end{bmatrix}, \quad \hat{A}_1 \triangleq \begin{bmatrix} \left(1+\tau\,\mathrm{L_{xx}}+\frac{\tau\sigma(1+\theta)\,\mathrm{L_{yx}}\,\mathrm{L_{xy}}}{1+\sigma\mu_y}\right) \\ \frac{\tau\,\mathrm{L_{xy}}(1+\sigma(1+\theta)\,\mathrm{L_{yy}})}{1+\sigma\mu_y} \\ \frac{\sigma}{\sqrt{1-\rho}}\frac{\tau\theta\,\mathrm{L_{xy}}\,\mathrm{L_{yx}}}{(1+\sigma\mu_y)} \\ \frac{\sigma}{\sqrt{1-\rho}}\frac{\tau\theta\,\mathrm{L_{xy}}\,\mathrm{L_{yy}}}{(1+\sigma\mu_y)} \end{bmatrix}, \quad \hat{A}_2 \triangleq \begin{bmatrix} \sigma(1+\theta)\,\mathrm{L_{yx}} \\ 1+\sigma(1+\theta)\,\mathrm{L_{yy}} \\ \frac{\sigma}{\sqrt{1-\rho}}\theta\,\mathrm{L_{yx}} \\ \frac{\sigma}{\sqrt{1-\rho}}\theta\,\mathrm{L_{yy}} \end{bmatrix}$$

$$\hat{A}_3 \triangleq \begin{bmatrix} C_{\sigma,\theta}\sigma(1+\theta)(1+\tau\mu_x)\,\mathrm{L_{yx}} + \sigma(1+\theta)\,\mathrm{L_{yx}}\left((1+\tau\,\mathrm{L_{xx}})(1+\sigma\mu_y)+\tau\sigma(1+\theta)\,\mathrm{L_{yx}}\,\mathrm{L_{xy}}\right)+\sigma\theta\,\mathrm{L_{yx}}(1+\tau\mu_x)(1+\sigma\mu_y) \\ (1+\tau\mu_x)C_{\sigma,\theta}^2+\sigma(1+\theta)\,\mathrm{L_{yx}}\,\tau\,\mathrm{L_{xy}}\,C_{\sigma,\theta}+\sigma\theta\,\mathrm{L_{yy}}(1+\tau\mu_x)(1+\sigma\mu_y) \\ \frac{\sigma}{\sqrt{1-\rho}}\left(C_{\sigma,\theta}\theta\,\mathrm{L_{yx}}(1+\tau\mu_x)+(1+\theta)\tau\sigma\theta\,\mathrm{L_{xy}}\,\mathrm{L_{yx}}^2\right) \\ \frac{\sigma}{\sqrt{1-\rho}}\left(C_{\sigma,\theta}\theta\,\mathrm{L_{yy}}(1+\tau\mu_x)+(1+\theta)\tau\sigma\theta\,\mathrm{L_{xy}}\,\mathrm{L_{yx}}\,\mathrm{L_{yy}}\right) \end{bmatrix}.$$

$$A \triangleq \begin{bmatrix} A_1^\top \\ A_2^\top \\ A_3^\top \end{bmatrix} \triangleq \begin{bmatrix} \frac{1}{1+\tau\mu_x}\hat{A}_1^\top \\ \frac{\sqrt{2}(1+\theta)}{1+\sigma\mu_y}\hat{A}_2^\top \\ \frac{\sqrt{2}\theta\rho^{-1}}{(1+\sigma\mu_y)^2(1+\tau\mu_x)}\hat{A}_3^\top \end{bmatrix} \mathrm{Diag}\begin{bmatrix} \sqrt{2\rho\tau} \\ \sqrt{2\rho\sigma/(1-\alpha\sigma)} \\ \sqrt{2\rho\tau} \\ \sqrt{2\rho\sigma/(1-\alpha\sigma)} \end{bmatrix}.$$

$$B_0 \triangleq \begin{bmatrix} \frac{\tau}{1+\tau\mu_x} & 0 & \frac{\tau\sigma(1+\theta)\,\mathrm{L_{xy}}}{(1+\tau\mu_x)(1+\sigma\mu_y)} & \frac{\tau\sigma\theta\,\mathrm{L_{xy}}}{\sqrt{\rho}(1+\tau\mu_x)(1+\sigma\mu_y)} & 0 \\ 0 & 0 & 0 & \frac{\sigma\tau\theta(1+\theta)\,\mathrm{L_{yx}}}{\rho(1+\sigma\mu_y)(1+\tau\mu_x)} & 0 \\ 0 & 0 & \frac{\sigma(1+\theta)^2}{1+\sigma\mu_y} & \frac{2\theta\sigma(1+\theta)}{\sqrt{\rho}(1+\sigma\mu_y)} & 0 \\ 0 & 0 & 0 & \frac{\sigma\theta}{\rho(1+\sigma\mu_y)}\left(\theta+(1+\theta)S_{\tau,\sigma,\theta}\right) & \frac{\sigma\theta^2}{\rho^{3/2}(1+\sigma\mu_y)}S_{\tau,\sigma,\theta} \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$B_1 \triangleq \frac{\rho}{4\|A_0\|^2}\mathrm{Diag}\begin{bmatrix} \frac{3\tau^2}{(1+\tau\mu_x)^2}\left(\frac{\sigma(1+\theta)\,\mathrm{L_{yx}}}{1+\sigma\mu_y}\right)^2 \\ 0 \\ \frac{2\sigma^2(1+\theta)^2}{(1+\sigma\mu_y)^2}\left(\frac{1+\sigma(1+\theta)\,\mathrm{L_{yy}}}{1+\sigma\mu_y}\right)^2+\frac{3\,\tau^2\sigma^2(1+\theta)^2\,\mathrm{L_{xy}}^2}{(1+\tau\mu_x)^2(1+\sigma\mu_y)^2}\left(\frac{\sigma(1+\theta)\,\mathrm{L_{yx}}}{1+\sigma\mu_y}\right)^2 \\ \frac{2\sigma^2\theta^2}{\rho(1+\sigma\mu_y)^2}\left(\frac{1+\sigma(1+\theta)\,\mathrm{L_{yy}}}{1+\sigma\mu_y}\right)^2+\frac{3\,\tau^2\sigma^2\theta^2\,\mathrm{L_{xy}}^2}{\rho(1+\tau\mu_x)^2(1+\sigma\mu_y)^2}\left(\frac{\sigma(1+\theta)\,\mathrm{L_{yx}}}{1+\sigma\mu_y}\right)^2 \\ 0 \end{bmatrix}, \quad B_2 \triangleq \mathrm{Diag}\begin{bmatrix} 4\tau(1-\rho)^2 \\ 0 \\ \frac{4\sigma(1+\theta)^2(1-\rho)^2}{(1-\alpha\sigma)}+\frac{3\|A_0\|^2}{\rho} \\ \frac{4\sigma\theta^2(1-\rho)^2}{\rho(1-\alpha\sigma)} \\ 0 \end{bmatrix}$$

$$\mathcal{Q} \triangleq \|B_0+B_1+B_2\|_F, \qquad S_{\tau,\sigma,\theta} \triangleq \frac{1+\sigma(1+\theta)\,\mathrm{L_{yy}}}{1+\sigma\mu_y}+\frac{\tau\sigma(1+\theta)\,\mathrm{L_{yx}}\,\mathrm{L_{xy}}}{(1+\tau\mu_x)(1+\sigma\mu_y)}, \qquad C_{\sigma,\theta} \triangleq 1+\sigma(1+\theta)\,\mathrm{L_{yy}}$$

Table 2: Summary of the constants used throughout the analysis.

Furthermore, we provide the expressions of the same matrices under the (CP) parameterization 3.2 in Table 3

B Basic inequalities

Throughout this paper, we'll make extensive use of the following inequalities, without necessarily referring to them. We start with a simple convex inequality known as Cauchy-Schwartz's inequality when $M = 2$.

$$A_1 = \begin{bmatrix} \sqrt{1-\theta}\,\theta\sqrt{\frac{2}{\mu_x}}\left(1+\frac{(1-\theta)L_{xx}}{\theta\mu_x}+\frac{(1-\theta)^2(1+\theta)L_{xy}L_{yx}}{\theta\mu_x\mu_y}\right) \\ (1-\theta)^{3/2}\frac{2\theta}{\sqrt{\mu_y\left(1+2\frac{(1-\theta)}{\theta}\frac{L_{yy}}{\mu_y}\right)}}\frac{L_{xy}}{\mu_x}\left(1+\frac{(1-\theta)(1+\theta)}{\theta}\frac{L_{yy}}{\mu_x}\right) \\ (1-\theta)^2\theta\sqrt{\frac{2}{\mu_x}}\frac{L_{xy}L_{yx}}{\mu_x\mu_y} \\ (1-\theta)^2\frac{2\theta}{\sqrt{\mu_y\left(1+2\frac{(1-\theta)}{\theta^2}\frac{L_{yy}}{\mu_y}\right)}}\frac{L_{xy}L_{yy}}{\mu_x\mu_y} \end{bmatrix}, \quad A_2 = \begin{bmatrix} (1-\theta)^{3/2}\frac{2(1+\theta)^2L_{yx}}{\sqrt{\mu_x}\mu_y} \\ \sqrt{1-\theta}\frac{2\sqrt{2}\theta(1+\theta)}{\sqrt{\mu_y\left(1+2\frac{(1-\theta)}{\theta^{1/2}}\frac{L_{yy}}{\mu_y}\right)}}\left(1+\frac{(1-\theta)+(1+\theta)L_{yx}}{\theta\mu_y}\right) \\ (1-\theta)\frac{2\theta(1+\theta)L_{yx}}{\sqrt{\mu_x}\mu_y} \\ (1-\theta)\frac{2\sqrt{2}\theta(1+\theta)L_{yy}}{\mu_y\sqrt{\mu_y\left(1+2\frac{(1-\theta)}{\theta^{1/2}}\frac{L_{yy}}{\mu_y}\right)}} \end{bmatrix}$$

$$A_3 = \begin{bmatrix} (1-\theta)^{3/2}\frac{2\sqrt{2}\theta^2}{\sqrt{\mu_x}\mu_y}\left(\left(1+\frac{(1-\theta)(1+\theta)}{\theta\mu_y}L_{yy}\right)(1+\theta)\theta L_{yx}+(1+\theta)L_{yx}\left(\theta^2+\frac{(1-\theta)^2}{\theta^2\mu_x\mu_y}(1+\theta)L_{yx}L_{yy}\right)+\theta^3L_{yx}\right) \\ \sqrt{1-\theta}\frac{2\sqrt{2}\theta^3}{\sqrt{\mu_y\left(1+2\frac{(1-\theta)}{\theta^{1/2}}\frac{l_{yy}}{\mu_y}\right)}}\left[\theta\left(1+\frac{(1-\theta)(1+\theta)}{\theta\mu_y}L_{yy}\right)^2+\frac{(1-\theta)^2}{\theta^2}\left(1+\frac{(1-\theta)(1+\theta)}{\theta\mu_y}L_{yy}\right)\frac{L_{yx}L_{xy}}{\mu_x\mu_y}+(1-\theta)\theta^2\frac{L_{yy}}{\mu_y}\right] \\ (1-\theta)\frac{2\theta^2}{\sqrt{\mu_x}\mu_y}\left[\left(1+\frac{(1-\theta)(1+\theta)}{\theta\mu_y}L_{yy}\right)\theta^2L_{yx}+(1+\theta)\frac{(1-\theta)^2}{\theta}\frac{L_{xy}L_{yx}^2}{\mu_x\mu_y}\right] \\ (1-\theta)\frac{2\sqrt{2}\theta^2}{\mu_y\sqrt{\mu_y\left(1+2\frac{(1-\theta)}{\theta'^2}\frac{c_{yy}}{\mu_y}\right)}}\left[\left(1+\frac{(1-\theta)(1+\theta)}{\theta\mu_y}L_{yy}\right)\theta^2L_{yy}+(1+\theta)\frac{(1-\theta)^2}{\theta}\frac{L_{xy}L_{yx}L_{yy}}{\mu_x\mu_y}\right] \end{bmatrix}$$

$$B_0 \triangleq \begin{bmatrix} \frac{(1-\theta)}{\mu_x} & 0 & (1-\theta)^2\frac{(1+\theta)L_{xy}}{\mu_x\mu_y} & (1-\theta)^2\frac{\sqrt{\theta}L_{xy}}{\mu_x\mu_y} & 0 \\ 0 & 0 & 0 & (1-\theta)^2\frac{(1+\theta)L_{yx}}{\mu_x\mu_y} & 0 \\ 0 & 0 & 0 & (1-\theta)\frac{2\sqrt{\theta}(1+\theta)}{\mu_y} & 0 \\ 0 & 0 & 0 & \frac{(1-\theta)}{\mu_y}\left(\theta+(1+\theta)\tilde{S}_\theta\right) & (1-\theta)\sqrt{\theta}\tilde{S}_\theta \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$B_1 \triangleq \frac{\theta}{4\|A_0\|_F^2}\mathrm{Diag}\begin{bmatrix} \frac{3(1-\theta)^4(1+\theta)^2l_{yx}^2}{\mu_x^2\mu_y^2} \\ 0 \\ (1-\theta)^2 2\frac{(1+\theta)^2\theta^2}{\mu_y^2}\left(1+\left(1-\theta^2\right)\frac{L_{yy}}{\theta\mu_y}\right)^2+(1-\theta)^6\frac{3(1+\theta)^4L_{xy}^2L_{yx}^2}{\mu_x^2\mu_y^4} \\ \frac{2(1-\theta)^2\theta^3}{\mu_y^2}\cdot\left(1+\left(1-\theta^2\right)\frac{L_{yy}}{\theta\mu_y}\right)^2+(1-\theta)^6\frac{3\theta(1+\theta)^2L_{xy}^2L_{yx}^2}{\mu_x^2\mu_y^4} \\ 0 \end{bmatrix}, \quad B_2 \triangleq \mathrm{Diag}\begin{bmatrix} \frac{4(1-\theta)^3}{\theta\mu_x} \\ 0 \\ \frac{4(1+\theta)^2(1-\theta)^3}{\theta\mu_y\left(\frac{1}{2}+\sqrt{\theta}\,\mathrm{L}_{\mathrm{yy}}\,\sigma\right)}+\frac{3}{\theta}\|A_0\|^2 \\ \frac{4\sqrt{\theta}(1-\theta)^2}{\mathrm{L}_{\mathrm{yy}}} \\ 0 \end{bmatrix}$$

$$\tilde{S}_\theta \triangleq \theta+(1-\theta)(1+\theta)\frac{L_{yy}}{\mu_y}+(1-\theta)^2(1+\theta)\frac{L_{yx}L_{xy}}{\mu_x\mu_y}$$

Table 3: Simplifications of the matrices A_is and B_is from Table 2 under the (CP) parameterization 3.2.

Lemma B.1. *For any sequence $x_1, \ldots, x_M \in \mathbb{R}^{\mathrm{d}}$, $M \geq 1$,*

$$\|x_1+\cdots+x_M\|^2 \leq M\sum_{i=1}^{M}\|x_i\|^2. \tag{B.1}$$

We do not provide a proof of the following statement which may be seen as a direct extension of e.g. [2, Theorem 6.42].We have this result in SAPD paper, we can directly cite it.

Lemma B.2. *For any μ-strongly convex function $\varphi : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}\cup\{+\infty\}$, for any $x, y \in \mathbb{R}^{\mathrm{d}}$,*

$$\|prox_\varphi(x) - prox_\varphi(y)\| \leq \frac{1}{1+\mu}\|x-y\|. \tag{B.2}$$

C Analytical solution for quadratics

In this section, we study the behaviour of SAPD on quadratic problems subjects to Gaussian isotropic noise. That is, given a symmetric matrix $K \in \mathbb{R}^{\mathrm{d}}$, and two regularization parameters μ_x, μ_y, we aim at solving

$$\min_{x\in\mathbb{R}^{\mathrm{d}}}\max_{y\in\mathbb{R}^{\mathrm{d}}}\frac{\mu_x}{2}\|x\|^2+\langle Kx, y\rangle-\frac{\mu_y}{2}\|y\|^2, \tag{C.1}$$

while having access only to noisy estimates $\widetilde{\nabla}_{\mathrm{x}}\,\Phi, \widetilde{\nabla}_{\mathrm{y}}\,\Phi$ of the partial gradients of $\Phi : x, y \mapsto \langle Kx, y\rangle$. Precisely, we assume

$$\widetilde{\nabla}_{\mathrm{x}}\,\Phi(x,y) = K^\top y+\omega_k^x, \quad \widetilde{\nabla}_{\mathrm{y}}\,\Phi(x,y) = Kx+\omega_k^y$$

where the $(\omega_k^x)_{k\geq 0}$ and $(\omega_k^y)_{k\geq 0}$ denote independent and identically distributed centered Gaussian vectors satisfying

$$\mathbb{E}\left[\omega_k^x \omega_k^{x\top}\right] = \frac{\delta^2}{d} I, \quad \mathbb{E}\left[\omega_k^y \omega_k^{y\top}\right] = \frac{\delta^2}{d} I.$$

This problem was first studied in [38] where it was shown that the sequence of iterates $(x_k, y_k)_{k\geq 0}$ generated by SAPD converges in distribution to an centered gaussian which covariance matrix Σ^∞ is solution of a Lyapunov equation parameterizeds by τ, σ, θ, and K. Precisely, denoting $z_k = [x_{k-1}, y_k]^\top$ and $\omega_k = \left[\omega_{k-1}^x \omega_{k-1}^y; \omega_k^y\right]^\top$, the authors observe that $(z_k)_{k\geq 0}$ satisfies the recurrence relation $z_{k+1} = Az_k + B\omega_k$ where A and B are defined as

$$A \triangleq \begin{bmatrix} \frac{1}{1+\tau\mu_x} I_d & \frac{-\tau}{(1+\tau\mu_x)} K^\top \\ \frac{1}{1+\sigma\mu_y}\left(\frac{\sigma(1+\theta)}{1+\tau\mu_x} - \sigma\theta\right) K & \frac{1}{1+\sigma\mu_y}\left(I_d - \frac{\tau\sigma(1+\theta)}{1+\tau\mu_x} KK^\top\right) \end{bmatrix}, \quad B \triangleq \begin{bmatrix} \frac{-\tau}{1+\tau\mu_x} I_d & 0_d & 0_d \\ \frac{-\tau\sigma(1+\theta)}{(1+\tau\mu_x)(1+\sigma\mu_y)} K & \frac{-\sigma\theta}{1+\sigma\mu_y} I_d & \frac{\sigma(1+\theta)}{1+\sigma\mu_y} I_d \end{bmatrix}$$

As a result, the covariance matrix Σ_k of z_k satisfies for all k,

$$\Sigma_{k+1} = A\Sigma_k A^\top + R. \tag{C.2}$$

where $R = \frac{\delta^2}{d} BB^\top + A\mathbb{E}\left[z_k \omega_k^\top\right] B^\top + B\mathbb{E}\left[\omega_k z_k^\top\right] A^\top$. Using the independence assumptions on the ω_k^x's and ω_k^y's, elementary derivations lead to expressing R as

$$R = \frac{\delta^2}{d}\begin{bmatrix} \frac{\tau^2}{(1+\tau\mu_x)^2} & \left(\frac{\tau^2\sigma(1+\theta)}{(1+\tau\mu_x)^2(1+\sigma\mu_y)} + \frac{\tau\sigma^2\theta(1+\theta)}{(1+\sigma\mu_y)^2(1+\tau\mu_x)}\right) K \\ \left(\frac{\tau^2\sigma(1+\theta)}{(1+\tau\mu_x)^2(1+\sigma\mu_y)} + \frac{\tau\sigma^2\theta(1+\theta)}{(1+\sigma\mu_y)^2(1+\tau\mu_x)}\right) K & \frac{\sigma^2(1+\theta)^2}{(1+\sigma\mu_y)^2}\left(\frac{\tau^2}{(1+\tau\mu_x)^2} + \frac{2\tau\sigma\theta}{(1+\tau\mu_x)(1+\sigma\mu_y)}\right) KK + \frac{\sigma^2}{(1+\sigma\mu_y)^2}\left(1 + \frac{2\theta(1+\theta)\sigma\mu_y}{1+\sigma\mu_y}\right) \end{bmatrix}.$$

Provided that the spectral radius $\rho(A)$ of A is less than 1, the sequence $(\Sigma_k)_{k\geq 0}$ converges to a matrix Σ^∞ satisfying

$$\Sigma^\infty = A\,\Sigma^\infty\, A^\top + R \tag{C.3}$$

Leveraging the spectral theorem, it is shown in [38] that an orthogonal change of basis enables to reduce the $2d \times 2d$ Lyapunov equation to d systems of the form

$$\Sigma^{\infty,\lambda_i} = A^{\lambda_i}\Sigma^{\infty,\lambda_i} {A^{\lambda_i}}^\top + R^{\lambda_i} \tag{C.4}$$

with

$$A^{\lambda_i} \triangleq \begin{bmatrix} \frac{1}{1+\tau\mu_x} I_d & \frac{-\tau}{(1+\tau\mu_x)}\lambda_i \\ \frac{1}{1+\sigma\mu_y}\left(\frac{\sigma(1+\theta)}{1+\tau\mu_x} - \sigma\theta\right)\lambda_i & \frac{1}{1+\sigma\mu_y}\left(I_d - \frac{\tau\sigma(1+\theta)}{1+\tau\mu_x}\lambda_i^2\right) \end{bmatrix}$$

$$R^{\lambda_i} \triangleq \begin{bmatrix} \frac{\tau^2}{(1+\tau\mu_x)^2} & \left(\frac{\tau^2\sigma(1+\theta)}{(1+\tau\mu_x)^2(1+\sigma\mu_y)} + \frac{\tau\sigma^2\theta(1+\theta)}{(1+\sigma\mu_y)^2(1+\tau\mu_x)}\right)\lambda_i \\ \left(\frac{\tau^2\sigma(1+\theta)}{(1+\tau\mu_x)^2(1+\sigma\mu_y)} + \frac{\tau\sigma^2\theta(1+\theta)}{(1+\sigma\mu_y)^2(1+\tau\mu_x)}\right)\lambda_i & \frac{\sigma^2(1+\theta)^2}{(1+\sigma\mu_y)^2}\left[\frac{\tau^2}{(1+\tau\mu_x)^2} + \frac{2\tau\sigma\theta}{(1+\tau\mu_x)(1+\sigma\mu_y)}\right]\lambda_i^2 + \frac{\sigma^2}{(1+\sigma\mu_y)^2}\left(1 + \frac{2\theta(1+\theta)\sigma\mu_y}{1+\sigma\mu_y}\right) \end{bmatrix}$$

and λ_i denotes one of n real eigenvalues of A. From that point onwards, [38] proceed to a numerical solving of this Lyapunov equation for a randomly generated matrix K. In contrast, analytically solving (C.4) is a challenging problem that standard symbolic computation tools were not in position to address: we obtain several pages long answers for each coefficient of the equilibrium matrix. In the next section we provide answers along this direction under the Chambolle-Pock parameterization:

$$\tau = \frac{1-\theta}{\theta\mu_x}, \quad \sigma = \frac{1-\theta}{\theta\mu_y}. \tag{C.5}$$

A notable outcome of this development will be the sharpness of our analysis when we will consider the larger class of strongly-convex/strongly-concave problems subject to sub-gaussian perturbation (see Section 3).

C.1 Analytical hand-solving

In this section, we solve analytically (C.4) under the parameterization (C.5). Throughout, we let $\kappa \triangleq \frac{\lambda}{\sqrt{\mu_x\mu_y}}$. κ refers intuitively to a *local* condition number and will play a key role in the expression of the solution $\Sigma^{\infty,\lambda}$. The main result of this section holds as follows

Proposition C.1. *Under the Chambolle-Pock parameterization, the solutions $\Sigma^{\infty,\lambda_i}$ of* (C.4) *satisfy*

$$\Sigma^{\infty,0} = \frac{\delta^2}{d}\frac{(1-\theta)}{\mu_x^2\mu_y^2(1+\theta)}\begin{bmatrix} \mu_y^2 & \left(1-\theta^2\right)\left(\theta\mu_x + \mu_y\right)\lambda \\ \left(1-\theta^2\right)\left(\theta\mu_x + \mu_y\right)\lambda & (1-\theta)^2(1+\theta)^2\left(1 + 2\theta\frac{\mu_x}{\mu_y}\right)\lambda^2 + \mu_x^2\left(1 + 2\left(1-\theta^2\right)\theta\right) \end{bmatrix}$$

if $0 \in Sp(A)$*, and for any non null* $\lambda \in Sp(A)$*,*

$$\Sigma^{\infty,\lambda} = \frac{\delta^2(1-\theta)}{d\lambda_i^2 P_c(\theta,\kappa)} \begin{bmatrix} \frac{\lambda_i^2}{\mu_x^2}\left(P_{1,1}^{(1)}(\theta,\kappa) + \frac{\lambda_i^2}{\mu_y^2}P_{1,1}^{(2)}(\theta,\kappa)\right) & \frac{\lambda}{\mu_x}\left(P_{1,2}^{(1)}(\theta,\kappa) + \frac{\lambda_i^2}{\mu_y^2}P_{1,2}^{(2)}(\theta,\kappa)\right) \\ \frac{\lambda}{\mu_x}\left(P_{1,2}^{(1)}(\theta,\kappa) + \frac{\lambda_i^2}{\mu_y^2}P_{1,2}^{(2)}(\theta,\kappa)\right) & P_{2,2}^{(1)}(\theta,\kappa) + \frac{\lambda_i^2}{\mu_y^2}P_{2,2}^{(2)}(\theta,\kappa) \end{bmatrix}$$

where the $P_{q,\ell}^{(k)}$ *and* P_c *are polynomials in* θ, κ*, defined in Equations* (C.11) *to* (C.14)*.*

As a direct consequence, on can deduce the actual the covariance matrix of the limiting distribution $\lim_{N\infty}(x_N, y_N)$ (as opposed to $\lim_{n\to\infty} z_n = (x_{n-1}, y_n)$),

Corollary 4. *The covariance matrix of the limiting Gaussian vector* $\lim_N(x_N, y_N)$ *satisfies*

$$\Sigma^\star = \begin{bmatrix} \frac{1}{(1+\tau\mu_x)^2}\left(\Sigma_{11}^\infty + \tau^2 K\Sigma_{22}^\infty K^\top - 2\tau K\Sigma_{12}^\infty\right) & \frac{1}{1+\tau\mu_x}\left(\Sigma_{12}^\infty - \tau K\Sigma_{22}^\infty\right) \\ \frac{1}{1+\tau\mu_x}\left(\Sigma_{12}^\infty - \tau K\Sigma_{22}^\infty\right) & \Sigma_{22}^\infty \end{bmatrix} \tag{C.6}$$

Proof. Immediate, given the relation $x_n = \frac{1}{1+\tau\mu_x}(x_{n-1} - \tau K y_n)$. □

Proof of Proposition C.1. We first note that under the parameterization (C.5), the matrices A^λ and R^λ simplify to

$$A^\lambda = \begin{bmatrix} \theta & -(1-\theta)\frac{\lambda}{\mu_x} \\ (1-\theta)\theta^2\frac{\lambda}{\mu_y} & \theta - (1-\theta)^2(1+\theta)\kappa^2 \end{bmatrix}$$

$$R^\lambda = \frac{\delta^2}{d}\frac{(1-\theta)^2}{\mu_x^2\mu_y^2} \begin{bmatrix} \mu_y^2 & \left(1-\theta^2\right)\left(\theta\mu_x + \mu_y\right)\lambda \\ \left(1-\theta^2\right)\left(\theta\mu_x + \mu_y\right)\lambda & (1-\theta)^2(1+\theta)^2\left(1 + 2\theta\frac{\mu_x}{\mu_y}\right)\lambda^2 + \mu_x^2\left(1 + 2\left(1-\theta^2\right)\theta\right) \end{bmatrix}$$

If $\lambda = 0$, then $A^\lambda = \text{Diag}(\theta, \theta)$. Hence, using the relation $\text{Vec}(ABC) = (C^\top \otimes A)\,\text{Vec}(B)$, we have

$$\begin{aligned} \Sigma^{\infty,\lambda} = A^\lambda\Sigma^{\infty,\lambda}A^\lambda + R &\Leftrightarrow \text{Vec}\left(\Sigma^{\infty,\lambda}\right) = \left(A^\lambda \otimes A^\lambda\right)\text{Vec}\left(\Sigma^{\infty,\lambda}\right) + \text{Vec}(R^\lambda) \\ &\Leftrightarrow \text{Vec}\left(\Sigma^{\infty,\lambda}\right) = \left(I - A^\lambda \otimes A^\lambda\right)^{-1}\text{Vec}\left(R^\lambda\right). \end{aligned}$$

Noting that $\left(I - A^\lambda \otimes A^\lambda\right)^{-1} = \text{Diag}(\frac{1}{1-\theta^2}, \frac{1}{1-\theta^2}, \frac{1}{1-\theta^2}, \frac{1}{1-\theta^2})$, we obtain $\Sigma^{\infty,0} = \frac{1}{1-\theta^2}R^0$.

From now on, let us fix a specific $\lambda \in \{\lambda_1, \ldots, \lambda_n\}$ that we assume non null. We first start with a further reduction of the system (C.4) by diagonalizing A^λ.

Lemma C.2. *For any* $\theta > \sqrt{1 - \frac{2}{\kappa^2}\left(\sqrt{1+\kappa^2} - 1\right)}$*, the matrix* A^λ *introduced in* (C.4) *admits the diagonalization* $A^\lambda = VJV^{-1}$ *where*

$$J = \begin{bmatrix} \nu_1 & 0 \\ 0 & \nu_2 \end{bmatrix}, \quad V \triangleq \begin{bmatrix} -A_{1,2}^\lambda & -A_{1,2}^\lambda \\ \theta - \nu_1 & \theta - \nu_2 \end{bmatrix},$$

with

- $\nu_1 \triangleq \frac{\left(2\theta - (1-\theta)^2(1+\theta)\kappa^2\right) + i\sqrt{|\Delta|}}{2}$
- $\nu_2 \triangleq \frac{\left(2\theta - (1-\theta)^2(1+\theta)\kappa^2\right) - i\sqrt{|\Delta|}}{2}$
- $\Delta = (1-\theta)^4(1+\theta)^2\kappa^4 - 4\theta^2(1-\theta)^2\kappa^2$.

Proof. Noting that $\text{Tr}(A^\lambda) = 2\theta - (1-\theta)^2(1+\theta)\kappa^2$ and $\text{Det}(A^\lambda) = \theta^2 - (1-\theta)^2\theta\kappa^2$, the characteristic polynomial of A^λ has for discriminant

$$\Delta = \text{Tr}(A^\lambda) - 4\text{Det}(A^\lambda) = \left((1-\theta)^4(1+\theta)^2\kappa^4 - 4\theta^2(1-\theta)^2\kappa^2\right)$$

Note also that

$$\Delta < 0 \iff (1-\theta^2)^2 \leq \frac{4\theta^2}{\kappa^2} \iff \theta \geq \sqrt{1 - \frac{2}{\kappa^2}\left(\sqrt{1+\kappa^2} - 1\right)},$$

and in such case, A^λ admits the two complex conjugate values ν_1, ν_2 specified above. Furthermore, observe that for $\nu \in \mathbb{C}$, $x, y \in \mathbb{C}$,

$$\begin{aligned}(A^\lambda - vI)\begin{bmatrix} x \\ y \end{bmatrix} = 0 \Leftrightarrow \begin{cases} (\theta - v)\, x + A^\lambda_{1,2} y = 0 \\ a_{21} x + (a_{22} - v)\, y = 0 \end{cases} \\ \Leftrightarrow y = \frac{-(\theta - v)}{A^\lambda_{1,2}} x \\ \Leftrightarrow \quad (x, y) \in \mathrm{Span}\begin{pmatrix} 1 \\ -\frac{(\theta - v)}{A^\lambda_{1,2}} \end{pmatrix} = \mathrm{Span}\begin{pmatrix} -A^\lambda_{1,2} \\ \theta - v \end{pmatrix}\end{aligned}$$

where the second line follows from assuming $\theta < 1$ and $\lambda \neq 0$. This justifies setting V as in Lemma C.2. □

Lemma C.3. *The equilibrium covariance matrix $\Sigma^{\infty,\lambda}$ satisfies*

$$\mathit{Vec}\left(\widetilde{\Sigma}^{\infty,\lambda}\right) = \left(I_{d^2} - J \otimes J\right)^{-1} \mathit{Vec}(\widetilde{R}^\lambda) \tag{C.7}$$

where

$$\begin{aligned}\widetilde{\Sigma}^{\infty,\lambda} &\triangleq V^{-1} \Sigma^{\infty,\lambda} (V^{-1})^\top \\ \widetilde{R}^\lambda &\triangleq V^{-1} R^\lambda \left(V^{-1}\right)^\top\end{aligned}$$

and J, V denotes the diagonalization of A, as expressed in Lemma (C.2)*.*

Proof. Given Lemma C.2, (C.4) writes

$$\Sigma^{\infty,\lambda} = V J V^{-1} \Sigma^{\infty,\lambda} (V^{-1})^\top J^\top V^\top + R^\lambda,$$

which amounts to

$$V^{-1} \Sigma^{\infty,\lambda} (V^{-1})^\top = J V^{-1} \Sigma^{\infty,\lambda} (V^{-1})^\top J^\top + V^{-1} R (V^{-1})^\top,$$

i.e.

$$\widetilde{\Sigma}^{\infty,\lambda} = J \widetilde{\Sigma}^{\infty,\lambda} J^\top + \widetilde{R}^\lambda$$

Finally, noting the relation $\mathrm{Vec}(ABC) = (C^\top \otimes A)\mathrm{Vec}(B)$, it remains to show that $I - J \otimes J$ is invertible. Observing that

$$J \otimes J = \begin{pmatrix} \nu_1^2 & 0 & 0 & 0 \\ 0 & \nu_1 \nu_2 & 0 & 0 \\ 0 & 0 & \nu_1 \nu_2 & 0 \\ 0 & 0 & 0 & \nu_2^2 \end{pmatrix},$$

it suffices to show that $\nu_1 \nu_2 \neq 1$. Now, in view of Lemma 5.3, we have for any $\theta \geq 0$ such that $\theta \neq 1$

$$\begin{aligned}\nu_1 \nu_2 - 1 = 0 &\iff -(1 - \theta)\left(1 + \theta + (1 - \theta)\theta\kappa^2\right) = 0 \\ &\iff 1 + (1 - \kappa^2)\theta - \theta^2 \kappa^2 = 0 \\ &\iff \theta = \frac{\kappa^2 - 1 \pm \sqrt{(\kappa^2 - 1)^2 + 4\kappa^2}}{2\kappa^2} \\ &\iff \theta \in \{-1/\kappa^2, 1\}\end{aligned}$$

which completes the proof. □

Let us now compute each of the quantities involved in (C.7).

Computation of $\widetilde{R}^\lambda$. Using the Cramer rule, first observe that V^{-1} satisfies

$$V^{-1} = \frac{1}{A^\lambda_{1,2}(\nu_2 - \nu_1)} \begin{bmatrix} (\theta - \nu_2) & +A^\lambda_{1,2} \\ -(\theta - \nu_1) & -A^\lambda_{1,2} \end{bmatrix}$$

from which we deduce

$$\widetilde{R}^\lambda = V^{-1} R^\lambda V^{-1\top} = \frac{\delta^2 (1 - \theta)^2}{d (A^\lambda_{1,2})^2 (\nu_1 - \nu_2)^2 \mu_x^2 \mu_y^2} \begin{bmatrix} Q_{1,1} & Q_{1,2} \\ Q_{1,2} & Q_{2,2} \end{bmatrix}$$

where

$$
\begin{aligned}
Q_{1,1} \triangleq & \left(\theta-\nu_2\right)^2 \mu_y^2 \\
& +\left(A_{1,2}^\lambda\right)^2\left((1-\theta)^2(1+\theta)^2\left(1+2\theta\frac{\mu_x}{\mu_y}\right)\lambda^2+\mu_x^2\left(1+2\left(1-\theta^2\right)\theta\right)\right) \\
& +2A_{1,2}^\lambda\left(\theta-\nu_2\right)\left(1-\theta^2\right)\left(\theta\mu_x+y_y\right)\lambda \\
Q_{2,2} \triangleq & \left(\theta-\nu_1\right)^2 \mu_y^2 \\
& +\left(A_{1,2}^\lambda\right)^2\left((1-\theta)^2(1+\theta)^2\left(1+2\theta\frac{\mu_x}{\mu_y}\right)\lambda^2+\mu_x^2\left(1+2\left(1-\theta^2\right)\theta\right)\right) \\
& +2A_{1,2}^\lambda\left(\theta-\nu_1\right)\left(1-\theta^2\right)\left(\theta\mu_x+y_y\right)\lambda \\
Q_{1,2} \triangleq & -\left(\theta-\nu_2\right)\left(\theta-\nu_1\right)\mu_y^2 \\
& -\left(A_{1,2}^\lambda\right)^2\left((1-\theta)^2(1+\theta)^2\left(1+2\theta\frac{\mu_x}{\mu_y}\right)\lambda^2+{\mu_x}^2\left(1+2\left(1-\theta^2\right)\theta\right)\right) \\
& +\left(1-\theta^2\right)\left(\theta\mu_x+\mu_y\right)\lambda A_{1,2}^\lambda\left(\nu_1+\nu_2-2\theta\right)
\end{aligned}
\tag{C.8}
$$

Computation of $\widetilde{\Sigma}^{\infty,\lambda}$ and partial expression of $\Sigma^{\infty,\lambda}$. Noting that

$$
\left(I-J\otimes J\right)^{-1}=\begin{bmatrix} \frac{1}{1-\nu_1^2} & 0 & 0 & 0 \\ 0 & \frac{1}{1-\nu_1\nu_2} & 0 & 0 \\ 0 & 0 & \frac{1}{1-\nu_1\nu_2} & 0 \\ 0 & 0 & 0 & \frac{1}{1-\nu_2^2} \end{bmatrix}
$$

we deduce

$$
\widetilde{\Sigma}^{\infty,\lambda}=\frac{\delta^2(1-\theta)^2}{d(A_{1,2}^\lambda)^2\left(\nu_1-\nu_2\right)^2\mu_x^2\mu_y^2}\begin{bmatrix} \frac{1}{1-\nu_1^2}\,Q_{1,1} & \frac{1}{1-\nu_1\nu_2}\,Q_{1,2} \\ \frac{1}{1-\nu_1\nu_2}\,Q_{1,2} & \frac{1}{1-\nu_2^2}\,Q_{2,2} \end{bmatrix}
$$

Hence, we have

$$
\Sigma^{\infty,\lambda}=V\widetilde{\Sigma}^{\infty,\lambda}V^\top=\frac{\delta^2(1-\theta)^2}{d(A_{1,2}^\lambda)^2\left(\nu_1-\nu_2\right)^2\mu_x^2\mu_y^2}\begin{bmatrix} S_{1,1} & S_{1,2} \\ S_{1,2} & S_{2,2} \end{bmatrix}
$$

with

- $S_{1,1} \triangleq (A_{1,2}^\lambda)^2\left(\widetilde{\Sigma}_{1,1}^{\infty,i}+\widetilde{\Sigma}_{2,2}^{\infty,i}+2\widetilde{\Sigma}_{1,2}^{\infty,i}\right)$
- $S_{1,2}=-A_{1,2}^\lambda\left[\left(\theta-\nu_1\right)\widetilde{\Sigma}_{1,1}^{\infty,i}+\left(\theta-\nu_2\right)\widetilde{\Sigma}_{2,2}^{\infty,i}+\left(2\theta-\left(\nu_1+\nu_2\right)\right)\widetilde{\Sigma}_{1,2}^{\infty,i}\right]$
- $S_{2,2}=\left(\theta-\nu_1\right)^2\widetilde{\Sigma}_{1,1}^{\infty,i}+\left(\theta-\nu_2\right)^2\widetilde{\Sigma}_{1,2}^{\infty,i}+2\left(\theta-\nu_1\right)\left(\theta-\nu_2\right)\widetilde{\Sigma}_{1,2}^{\infty,i}$

Simplification of $S_{1,1}$. Given the expression of $\widetilde{\Sigma}^{\infty,i}$, we have

$$
\begin{aligned}
S_{1,1}=&\frac{(1-\theta)^2\lambda^2}{\mu_x^2}\frac{1}{\left(1-\nu_1^2\right)\left(1-\nu_2^2\right)\left(1-\nu_1\nu_2\right)}\times \\
&\left(\begin{array}{l}
\left(1-\nu_2^2\right)\left(1-\nu_1\nu_2\right)\left(\begin{array}{l}\left(\theta-\nu_2\right)^2\mu_y^2 \\ +\left(A_{1,2}^\lambda\right)^2\left(\left(1-\theta\right)^2(1+\theta)^2\left(1+2\theta\frac{\mu_x}{\mu_y}\right)\lambda^2+\mu_x^2\left(1+2\left(1-\theta^2\right)\theta\right)\right) \\ +2A_{1,2}^\lambda\left(\theta-\nu_2\right)\left(1-\theta^2\right)\left(\theta\mu_x+\mu_y\right)\lambda\end{array}\right) \\
+2\left(1-\nu_1^2\right)\left(1-\nu_2^2\right)\left(\begin{array}{l}-\left(\theta-\nu_2\right)\left(\theta-\nu_1\right)\mu_y^2 \\ -\left(A_{1,2}^\lambda\right)^2\left(\left(1-\theta\right)^2(1+\theta)^2\left(1+2\theta\frac{\mu_x}{\mu_y}\right)\lambda^2+\mu_x^2\left(1+2\left(1-\theta^2\right)\theta\right)\right) \\ +\left(1-\theta^2\right)\left(\theta\mu_x+\mu_y\right)\lambda A_{1,2}^\lambda\left(\nu_1+\nu_2-2\theta\right)\end{array}\right) \\
+\left(1-\nu_1\nu_2\right)\left(1-\nu_1^2\right)\left(\begin{array}{l}\left(\theta-\nu_1\right)^2\mu_y^2 \\ +\left(A_{1,2}^\lambda\right)^2\left(\left(1-\theta\right)^2(1+\theta)^2\left(1+2\theta\frac{\mu_x}{\mu_y}\right)\lambda^2+\mu_x^2\left(1+2\left(1-\theta^2\right)\theta\right)\right) \\ +2A_{1,2}^\lambda\left(\theta-\nu_1\right)\left(1-\theta^2\right)\left(\theta\mu_x+\mu_y\right)\lambda\end{array}\right)
\end{array}\right)
\end{aligned}
$$

First note that

$$\begin{aligned}
&\mu_y^2\left((1-\nu_1\nu_2)\left((\theta-\nu_2)^2\left(1-\nu_2^2\right)+(\theta-\nu_1)^2\left(1-\nu_1^2\right)\right)-2\left(1-\nu_1^2\right)\left(1-\nu_2^2\right)(\theta-\nu_1)(\theta-\nu_2)\right)\\
&=\mu_y^2\begin{pmatrix}\left(\nu_1^2+\nu_2^2\right)\left(1+\theta^2\right)-\left(\nu_1^4+\nu_2^4\right)\\ +\nu_1\nu_2\begin{pmatrix}\left(\nu_1^2+\nu_2^2\right)\left(1+\theta^2\right)+\left(\nu_1^4+\nu_2^4\right)-2\theta\left(\nu_1^3+\nu_2^3\right)\\ -2(1+\theta^2)-2\nu_1\nu_2\theta^2+2\theta\nu_1\nu_2\left(\nu_1+\nu_2\right)-2\left(\nu_1\nu_2\right)^2\end{pmatrix}\end{pmatrix}\\
&=\mu_y^2\begin{pmatrix}-4(1-\theta)^2\theta^2\left(1-\theta^2\right)^2\kappa^2\\ +(1-\theta)^4\left(1+2\theta-\theta^2-16\theta^3-17\theta^4+14\theta^5+9\theta^6\right)\kappa^4\\ +(1-\theta)^6\theta\left(3+14\theta+16\theta^2-12\theta^3-23\theta^4-6\theta^5\right)\kappa^6\\ +(1-\theta)^8(1+\theta)^2\left(-1-2\theta+2\theta^2+8\theta^3+\theta^4\right)\kappa^8\\ -(1-\theta)^{10}\theta(1+\theta)^4\kappa^{10}\end{pmatrix}
\end{aligned}$$

where the last line can be deduced from Lemma 5.3. Second, we observe that

$$\begin{aligned}
&\left(1-\nu_2^2\right)(1-\nu_1\nu_2)-2\left(1-\nu_1^2\right)\left(1-\nu_2^2\right)+(1-\nu_1\nu_2)\left(1-\nu_1^2\right)\\
&\qquad=\begin{pmatrix}(1-\theta)^2\kappa^2\left(2\theta-2\theta(1+2\theta)-2\theta^3(1+2\theta)+2\theta^3\right)\\ +(1-\theta)^4\kappa^4\left(\left(1+\theta^2\right)(1+\theta)^2+2\theta^2(1+2\theta)-2\theta^2\right)\\ -(1-\theta)^6\kappa^6\theta(1+\theta)^2\end{pmatrix}
\end{aligned}$$

and

$$\begin{aligned}
&(A_{1,2}^{\lambda})^2\left((1-\theta)^2(1+\theta)^2\left(1+2\theta\frac{\mu_x}{\mu_y}\right)\lambda^2+\mu_x^2\left(1+2\left(1-\theta^2\right)\theta\right)\right)\\
&\qquad=\mu_y^2\left((1-\theta)^4\kappa^4(1+\theta)^2+\left(2\theta(1-\theta)^4\kappa^2(1+\theta)^2+(1-\theta)^2(1+2(1-\theta^2)\theta)\right)\frac{\lambda^2}{\mu_y^2}\right)
\end{aligned}\tag{C.9}$$

and finally, noting that

$$2A_{12}\left(1-\theta^2\right)(\theta\mu_x+\mu_y)\lambda=-\mu_y^2\left(2(1-\theta)^2(1+\theta)\theta\frac{\lambda^2}{\mu_y^2}+(1-\theta)^2(1+\theta)2\kappa^2\right)\tag{C.10}$$

and using again Lemma 5.3, we obtain

$$\begin{aligned}
&\begin{pmatrix}\left(1-\nu_2^2\right)(1-\nu_1\nu_2) & 2A_{1,2}^{\lambda}(\theta-\nu_2)\left(1-\theta^2\right)(\theta\mu_2+\mu_y)\lambda\\ +2\left(1-\nu_1^2\right)\left(1-\nu_2^2\right)\left(1-\theta^2\right)(\theta\mu_2+y_y)\lambda & A_{1,2}^{\lambda}\left(\nu_1+\nu_2-2\theta\right)\\ +(1-\nu_1\nu_2)\left(1-\nu_1^2\right) & 2A_{1,2}^{\lambda}(\theta-\nu_1)\left(1-\theta^2\right)(\theta\mu_x+y_y)\lambda\end{pmatrix}\\
&=2A_{12}\left(1-\theta^2\right)(\theta\mu_x+\mu_y)\lambda\begin{pmatrix}-2\nu_1\nu_2\theta+\nu_1\nu_2\theta\left(\nu_1^2+\nu_2^2\right)-\nu_1\nu_2\left(\nu_1^3+\nu_2^2\right)\\ +\nu_1^2\nu_2^2\left(\nu_1+\nu_2\right)+\left(\nu_1^2+\nu_2^2\right)\theta-2\nu_1^2\nu_2^2\theta\end{pmatrix}\\
&=\mu_y^2\left(2(1-\theta)^2(1+\theta)\theta\frac{\lambda^2}{\mu_y^2}+(1-\theta)^2(1+\theta)2\kappa^2\right)\begin{pmatrix}4(1-\theta)^2\theta^3(\theta^2-1)\kappa^2\\ +(1-\theta)^4(\theta+2\theta^2-10\theta^4-5*\theta^5)\kappa^4\\ +(1-\theta)^6\theta^2*(2+9\theta+8\theta^2+\theta^3)\kappa^6\\ -(1-\theta)^8\theta*(1+\theta)^3\kappa^8\end{pmatrix}
\end{aligned}$$

Hence, grouping together the terms which have a λ^2/μ_y^2 factor and those which do not, we obtain

$$\begin{aligned}
S_{1,1}=&\frac{(1-\theta)^6\lambda^2\mu_y^2}{\mu_x^2\left(1-\nu_1^2\right)\left(1-\nu_2^2\right)(1-\nu_1\nu_2)}\times\\
&\left(\underbrace{\begin{pmatrix}-4\kappa^2\theta^2(1+\theta)^2\\ +\kappa^4\begin{pmatrix}1+2\theta-\theta^2-8\theta^3\\ -9\theta^4+6\theta^5+\theta^6\end{pmatrix}\\ +(1-\theta)^2\kappa^6\left(\theta+4\theta^2+4\theta^3-\theta^5\right)\end{pmatrix}}_{P_{1,1}^{(1)}(\theta,\kappa)\triangleq}+\frac{\lambda^2}{\mu_y^2}\underbrace{\begin{pmatrix}-4\kappa^2\theta^2\left(1+2\theta-\theta^2-2\theta^3+2\theta^4\right)\\ +(1-\theta)^2\kappa^4\begin{pmatrix}1+4\theta+4\theta^2-6\theta^3\\ -11\theta^4+2\theta^5+2\theta^6\end{pmatrix}\\ +(1-\theta)^4\kappa^6\theta(1+\theta)^2(1+2\theta)\end{pmatrix}}_{P_{1,1}^{(2)}(\theta,\kappa)\triangleq}\right)
\end{aligned}\tag{C.11}$$

Simplification of $S_{2,2}$. Similarly, (C.8), also gives

$$S_{2,2} = \frac{1}{(1-\nu_1^2)(1-\nu_2^2)(1-\nu_1\nu_2)} \times \left(\begin{array}{c} \mu_y^2 \begin{pmatrix} (\theta-\nu_1)^2(1-\nu_2^2)(1-\nu_1\nu_2)(\theta-\nu_2)^2 \\ (\theta-\nu_2)^2(1-\nu_1\nu_2)(1-\nu_1^2)(\theta-\nu_1)^2 \\ -2(\theta-\nu_1)(\theta-\nu_2)(1-\nu_1^2)(1-\nu_2^2)(\theta-\nu_2)(\theta-\nu_1) \end{pmatrix} \\ +(A_{1,2}^{\lambda})^2 \left((1-\theta)^2(1+\theta)^2\left(1+2\theta\frac{\mu_x}{\mu_y}\right)\lambda^2 + \mu_x^2\left(1+2\left(1-\theta^2\right)\theta\right)\right) \begin{pmatrix} (\theta-\nu_1)^2(1-\nu_2^2)(1-\nu_1\nu_2) \\ +(\theta-\nu_2)^2(1-\nu_1\nu_2)(1-\nu_1^2) \\ -2(\theta-\nu_1)(\theta-\nu_2)(1-\nu_1^2)(1-\nu_2^2) \end{pmatrix} \\ +A_{1,2}^{\lambda}(1-\theta^2)(\theta\mu_x+y_y)\lambda \begin{pmatrix} 2(\theta-\nu_2) & (\theta-\nu_1)^2(1-\nu_2^2)(1-\nu_1\nu_2) \\ +2(\theta-\nu_1)(\theta-\nu_2)^2(1-\nu_1\nu_2)(1-\nu_1^2) \\ +(\nu_1+\nu_2-2\theta)2(\theta-\nu_1)(\theta-\nu_2)(1-\nu_1^2)(1-\nu_2^2) \end{pmatrix} \end{array} \right)$$

which, combined with (C.9) and (C.10), gives

$$S_{2,2} = \frac{\mu_y^2}{(1-\nu_1^2)(1-\nu_2^2)(1-\nu_1\nu_2)} \times \left(\begin{array}{c} \left(\begin{array}{c} (\theta-\nu_1)^2(\theta-\nu_2)^2\left((1-\nu_1\nu_2)\left(2-\left(\nu_1^2+\nu_2^2\right)\right)-2\left(1-\nu_1^2\right)\left(1-\nu_2^2\right)\right) \end{array} \right) \\ + \begin{pmatrix} (1-\theta)^4\kappa^4(1-\theta)^2 \\ +\frac{\lambda^2}{\mu_y^2}\left(2\theta(1-\theta)^4\kappa^2(1+\theta)^2+(1-\theta)^2(1+2(1-\theta)\theta)\right) \end{pmatrix} \begin{pmatrix} (\theta-\nu_1)^2(1-\nu_2^2)(1-\nu_1\nu_2) \\ +(\theta-\nu_2)^2(1-\nu_1\nu_2)(1-\nu_1^2) \\ -2(\theta-\nu_1)(\theta-\nu_2)(1-\nu_1^2)(1-\nu_2^2) \end{pmatrix} \\ +\left((1-\theta)^2(1+\theta)\theta\frac{\lambda^2}{\mu_y^2}+(1-\theta)^2(1+\theta)\kappa^2\right) \begin{pmatrix} 2(\theta-\nu_2) & (\theta-\nu_1)^2(1-\nu_2^2)(1-\nu_1\nu_2) \\ +2(\theta-\nu_1)(\theta-\nu_2)^2(1-\nu_1\nu_2)(1-\nu_1^2) \\ +(\nu_1+\nu_2-2\theta)2(\theta-\nu_1)(\theta-\nu_2)(1-\nu_1^2)(1-\nu_2^2) \end{pmatrix} \end{array} \right).$$

Finally, using Lemma (5.3), and grouping again terms with a λ_i^2/μ_y^2 factor, we obtain

$$S_{2,2} = \frac{\mu_y^2(1-\theta)^6}{(1-\nu_1^2)(1-\nu_2^2)(1-\nu_1\nu_2)} \times \left(\begin{array}{c} \underbrace{\begin{pmatrix} -4\kappa^6\theta^2\left(1+2\theta-\theta^2-2\theta^3+2\theta^4\right) \\ +(1-\theta)^2\kappa^8\begin{pmatrix} 1+4\theta+4\theta^2-6\theta^3 \\ -11\theta^4+2\theta^5+2\theta^6 \end{pmatrix} \\ +(1-\theta)^4\kappa^{10}\theta(1+\theta)^2(1+2\theta) \end{pmatrix}}_{P_{2,2}^{(1)}(\theta,\kappa)\triangleq} + \frac{\lambda^2}{\mu^2} \underbrace{\begin{pmatrix} \kappa^2 4\theta^2(1+\theta)^2\left(-1-2\theta+2\theta^3\right) \\ +\kappa^4\begin{pmatrix} 1+4\theta+3\theta^2-20\theta^3 \\ -45\theta^4-2\theta^5+53\theta^6 \\ +20\theta^7-20\theta^8-2\theta^9 \end{pmatrix} \\ +(1-\theta)^2\kappa^6\theta\begin{pmatrix} 3+14\theta+20\theta^2 \\ -8\theta^3-47\theta^4 \\ -30\theta^5+4\theta^6+4\theta^7 \end{pmatrix} \\ +(1-\theta)^4\kappa^8 2\theta^2(1+\theta)^3(1+2\theta) \end{pmatrix}}_{P_{2,2}^{(2)}(\theta,\kappa)\triangleq} \end{array} \right) \tag{C.12}$$

Simplification of $S_{1,2}$. Going through the exact same steps leads to

$$S_{1,2} = \frac{\mu_y^2}{\mu_x}\frac{(1-\theta)^6\lambda}{(1-\nu_1^2)(1-\nu_2^2)(1-\nu_1\nu_2)} \left(\begin{array}{c} \underbrace{\begin{pmatrix} -4\kappa^4\theta^2\left(1+2\theta-\theta^3\right) \\ +(1-\theta)\kappa^6\begin{pmatrix} 1+3\theta+\theta^2-8\theta^3 \\ -11\theta^4+\theta^5+\theta^6 \end{pmatrix} \\ +(1-\theta)^3\kappa^8\theta(1+\theta)^2(1+2\theta) \end{pmatrix}}_{P_{1,2}^{(1)}(\theta,\kappa)\triangleq} + \frac{\lambda^2}{\mu^2} \underbrace{\begin{pmatrix} \kappa^2 4\theta^4\left(1+2\theta-\theta^3\right) \\ -(1-\theta)\kappa^4\theta^2\begin{pmatrix} 5+15\theta+5\theta^2-20\theta^3 \\ -11\theta^4+9\theta^5+\theta^6 \end{pmatrix} \\ +(1-\theta)^3\kappa^6\begin{pmatrix} 1+5\theta+8\theta^2-3\theta^3 \\ -21\theta^4-14\theta^5 \\ +2\theta^6+2\theta^7 \end{pmatrix} \\ +(1-\theta)^5\kappa^8\theta(1+\theta)^3(1+2\theta) \end{pmatrix}}_{P_{1,2}^{(2)}(\theta,\kappa)\triangleq} \end{array} \right) \tag{C.13}$$

Simplification of the common factor $\frac{\delta^2(1-\theta)^2}{d(A^\lambda_{1,2})^2(\nu_1-\nu_2)^2\mu_x^2\mu_y^2}\frac{1}{\left(1-\nu_1^2\right)\left(1-\nu_2^2\right)(1-\nu_1\nu_2)}$.

$$\begin{aligned}
&\frac{\delta^2(1-\theta)^2\mu_y^2}{d(A^\lambda_{1,2})^2\left(\nu_1-\nu_2\right)^2\mu_x^2\mu_y^2}\frac{1}{\left(1-\nu_1^2\right)\left(1-\nu_2^2\right)\left(1-\nu_1\nu_2\right)} = \frac{\delta^2}{d\lambda_i^2}\frac{1}{\left(\nu_1-\nu_2\right)^2\left(1-\nu_1^2\right)\left(1-\nu_2^2\right)\left(1-\nu_1\nu_2\right)} \\
&= \frac{\delta^2}{d\lambda_i^2(1-\theta)^5}\underbrace{\frac{1}{\left(\begin{array}{l}-4\kappa^2\theta^2\left(1+\theta\right)^3 \\ +\kappa^4\left(1+3\theta+\theta^2-17\theta^3-33\theta^4-3\theta^5+15\theta^6+\theta^7\right) \\ +(1-\theta)\kappa^6\theta\left(3+14\theta+13\theta^2-24\theta^3-35\theta^4+10\theta^5+3\theta^6\right) \\ +(1-\theta)^3\kappa^8\left(-1-4\theta-2\theta^2+14\theta^3+21\theta^4+2\theta^5-2\theta^6\right) \\ -(1-\theta)^5\kappa^{10}\theta(1+\theta)^2(1+2\theta)\end{array}\right)}}_{P_c(\theta,\kappa)\triangleq}
\end{aligned} \tag{C.14}$$

Conclusion. In view of the previous simplifications, the limit $\Sigma^{\infty,\lambda}$ satisfies

$$\Sigma^{\infty,\lambda} = \frac{\delta^2(1-\theta)}{d\lambda_i^2 P_c(\theta,\kappa)}\begin{pmatrix} \frac{\lambda_i^2}{\mu_x^2}\left(P^{(1)}_{1,1}(\theta,\kappa)+\frac{\lambda_i^2}{\mu_y^2}P^{(2)}_{1,1}(\theta,\kappa)\right) & \frac{\lambda}{\mu_x}\left(P^{(1)}_{1,2}(\theta,\kappa)+\frac{\lambda_i^2}{\mu_y^2}P^{(2)}_{1,2}(\theta,\kappa)\right) \\ \frac{\lambda}{\mu_x}\left(P^{(1)}_{1,2}(\theta,\kappa)+\frac{\lambda_i^2}{\mu_y^2}P^{(2)}_{1,2}(\theta,\kappa)\right) & P^{(1)}_{2,2}(\theta,\kappa)+\frac{\lambda_i^2}{\mu_y^2}P^{(2)}_{2,2}(\theta,\kappa)\end{pmatrix}$$

where the $P^{(k)}_{q,\ell}$ and P_c are polynomials in θ,κ, defined in Equations (C.11) to (C.14).

C.2 Illustrations

In this section, we illustrate and comment the properties of the equilibrium covariance derived in Section C.1.

Noise accumulates at a linear rate. First note that, though asymptotic, the equilibrium covariance matrix is reached at at linear rate given by the square of the spectral radius of A.

Lemma C.4. *For any* $\theta > \sqrt{1-\frac{2}{\kappa^2}\left(\sqrt{1+\kappa^2}-1\right)}$

$$\|\Sigma_n - \Sigma^\infty\| = \mathcal{O}\left(\rho(A)^{2n}\right) \tag{C.15}$$

Proof. Let V, J denote the Jordan decomposition of A (note that here, V is orthogonal). For $n\in\mathbb{N}$, let $\widetilde{\Sigma}_n := V^{-1}\Sigma_n\left(V^{-1}\right)^\top$, $\widetilde{\Sigma}^\infty := V^{-1}\Sigma^\infty\left(V^{-1}\right)^\top$, and $\widetilde{R}^i = V^{-1}R^\lambda\left(V^{-1}\right)^\top$. In view of the recursion (C.2), we have

$$\widetilde{\Sigma}_{n+1} = J\widetilde{\Sigma}_n J + \widetilde{R},$$

and vectorizing again this recursion lead to

$$\mathrm{Vec}(\widetilde{\Sigma}_{n+1}) = (J\otimes J)\,\mathrm{Vec}(\widetilde{\Sigma}_n) + \widetilde{R},$$

i.e.

$$\widetilde{\Sigma}_n = (J\otimes J)^{n-1}\,\widetilde{\Sigma}_1 + \sum_{k=1}^{n-1}(J\otimes J)^{k-1}\,\mathrm{Vec}(\widetilde{R})$$

Hence, noting that $\widetilde{\Sigma}^\infty = \sum_{k=0}^\infty (J\otimes J)^k\,\mathrm{Vec}(\widetilde{R})$ we obtain

$$\begin{aligned}
\|\Sigma_n - \Sigma^\infty\| = \|\widetilde{\Sigma_n} - \widetilde{\Sigma^\infty}\| &= \|\mathrm{Vec}(\widetilde{\Sigma_n}) - \mathrm{Vec}(\widetilde{\Sigma^\infty})\| \\
&= \|(J\otimes J)^{n-1}\,\widetilde{\Sigma}_1 + \sum_{k=n-1}^{\infty}(J\otimes J)^{k-1}\,\mathrm{Vec}(\widetilde{R})\| \\
&\le \rho(J\otimes J)^{n-1}\,\|\widetilde{\Sigma}_1 + \widetilde{\Sigma}^\infty\|
\end{aligned}$$

and the result directly follows from observing that $\rho\,(J\otimes J) = \rho(A)^2$. □

The convergence of covariances matrices can be observed on the toy instances displayed in Figure 3. We ran `SAPD` on three $1D$ problems which constants are given as

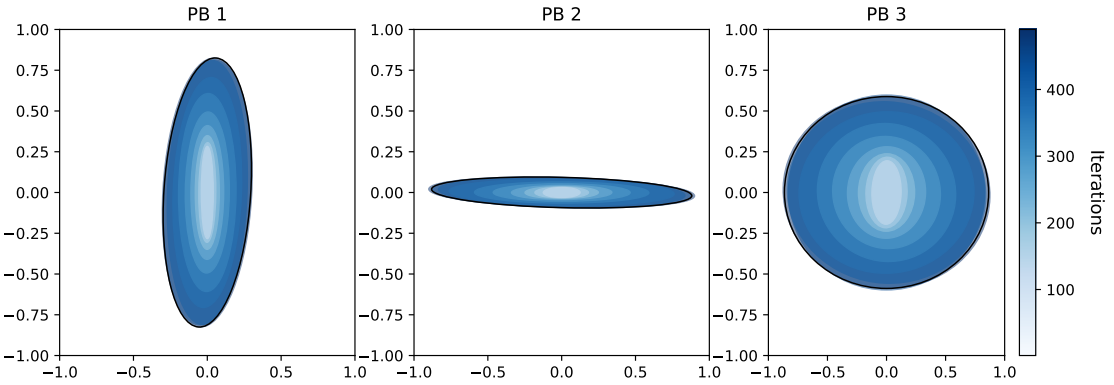

Figure 3: Noise accumulation over SAPD iterates: covariances matrices convergens to an equilibrium covariance as derived in Section C.1

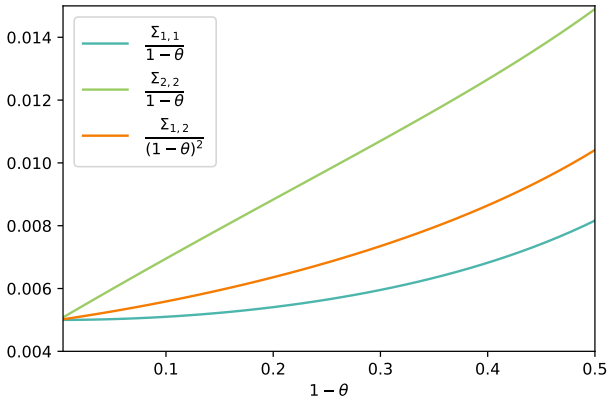

Figure 4: Scaling of the equilibrium covariance matrix under the Chambolle-Pock parameterization (C.5).

1. Pb 1 : $\lambda_1 = 1, \mu_x = 4.4, \mu_y = 1.5$, $\delta = 35$
2. Pb 2 : $\lambda_1 = 1, \mu_x = 2, \mu_y = 20$, $\delta = 50$.
3. Pb 3 : $\lambda_1 = 10e^{-3}, \mu_x = 0.205, \mu_y = 0.307$, $\delta = 5$.

SAPD was run 2000 times for 500 iterations under the Chambolle-Pock parameterization (C.5), with θ fixed to .99. We estimate the empirical covariance matrix on along iterations evenly distributed on a logscale, from 0 to 500. The theoretical covariance matrix derived in the previous section is represented in red on each plot. Figure 3 confirms the linear convergence of the matrices to the equilibrium matrix $\Sigma^\star$. Subsequently, we observe on these three examples how noise accumulates along iterations, producing covariance matrices that increase with respect to Lowner order. Note that this is a well-known property of Lyapunov recursions of the form C.3, as per [citation needed].

Equilibrium covariance scales with stepsizes. In figure 4, we display the dependency of the coefficients of the equilibrium covariance matrix with respect to the momentum θ under the Chambolle-Pock parameterization (C.5). We observe that both diagonal coefficients of the covariance matrix scale with $1-\theta$ for θ close to 1, while the non-diagonal coefficient scales with $(1-\theta)^2$.

C.3 Proof of Theorem 3.3

We start with proving the lower bound, and then we will proceed to the upper bound.

Lower bound. In view of Proposition C.1, z_∞ follows a centered Gaussian distribution with covariance matrix $\Sigma^\star$ as defined in (C.6). For $K = \mu_x = \mu_y = 1$ and $\theta \geq \bar{\theta}$, first observe that $\Sigma^{\infty,1}$ simplifies to

$$\Sigma^{\infty,1} = \frac{(1-\theta)}{P_c(\theta)} \begin{bmatrix} \begin{pmatrix} 2+6\theta-10\theta^2-42\theta^3-6\theta^4 \\ +42\theta^5-22\theta^6-6\theta^7+4\theta^8 \end{pmatrix} & (1-\theta)\begin{pmatrix} 2+8\theta-6\theta^2-56\theta^3-38\theta^4 \\ +60\theta^5+30\theta^6-32\theta^7-4\theta^8+4\theta^9 \end{pmatrix} \\ (1-\theta)\begin{pmatrix} 2+8\theta-6\theta^2-56\theta^3-38\theta^4 \\ +60\theta^5+30\theta^6-32\theta^7-4\theta^8+4\theta^9 \end{pmatrix} & \begin{pmatrix} 2+10\theta+2\theta^2-62\theta^3-98\theta^4+14\theta^5 \\ +118\theta^6+46\theta^7-64\theta^8-8\theta^9+8\theta^{10} \end{pmatrix} \end{bmatrix},$$

with $P_c(\theta) = 4\theta + 16\theta^2 - 16\theta^3 - 112\theta^4 - 40\theta^5 + 112\theta^6 - 16\theta^7 - 16\theta^8 + 4\theta^9$. We thus have

$$\begin{aligned}\Sigma^\star &= \begin{bmatrix} \theta^2\Sigma_{11}^{\infty,1} + (1-\theta)^2\Sigma_{22}^{\infty,1} - 2(1-\theta)\theta\Sigma_{12}^{\infty,1} & \theta\Sigma_{12}^{\infty,1} - (1-\theta)\Sigma_{22}^{\infty,1} \\ \theta\Sigma_{12}^{\infty,1} - (1-\theta)\Sigma_{22}^{\infty,1} & \Sigma_{22}^{\infty,1} \end{bmatrix} \\ &= \frac{(1-\theta)}{P_c(\theta)} \begin{bmatrix} \begin{pmatrix} 2+4\theta-18\theta^2-30\theta^3+54\theta^4+38\theta^5-94\theta^6 \\ -6\theta^7-28\theta^8+70\theta^9-16\theta^{10}-12\theta^{11}+4\theta^{12} \end{pmatrix} & \begin{pmatrix} -2-6\theta+14\theta^2+50\theta^3-14\theta^4-94\theta^5-6\theta^6 \\ +42\theta^7+48\theta^8-28\theta^9-8\theta^{10}+4\theta^{11} \end{pmatrix} \\ \begin{pmatrix} -2-6\theta+14\theta^2+50\theta^3-14\theta^4-94\theta^5-6\theta^6 \\ +42\theta^7+48\theta^8-28\theta^9-8\theta^10+4\theta^{11} \end{pmatrix} & \begin{pmatrix} 2+10\theta+2\theta^2-62\theta^3-98\theta^4+14\theta^5 \\ +118\theta^6+46\theta^7-64\theta^8-8\theta^9+8\theta^{10} \end{pmatrix} \end{bmatrix}.\end{aligned}$$

Since the eigenvalues of a 2×2 symmetric matrix $\begin{bmatrix} u & w \\ w & v \end{bmatrix}$ can be written in closed form $\frac{1}{2}\left(u+v \pm \sqrt{(u-v)^2+4w^2}\right)$, the smallest eigenvalue λ_1 of $\Sigma^\star$ satisfies

$$\lambda_1 = \frac{1}{2}\left(\Sigma_{11}^\star + \Sigma_{22}^\star - \sqrt{\left(\Sigma_{11}^\star - \Sigma_{12}^\star\right)^2 + 4{\Sigma_{12}^\star}^2}\right). \tag{C.16}$$

Furthermore, by sub-additivity of $\sqrt{\cdot}$, we have

$$u + v - \sqrt{(u-v)^2+4w^2} \geq (u+v) - |u-v| - 2|w| = 2(\min\{u,v\} - |w|);$$

therefore, we get

$$\lambda_1 \geq \min\{\Sigma_{11}^\star, \Sigma_{22}^\star\} - |\Sigma_{12}^\star|. \tag{C.17}$$

Let $\mathcal{U} \triangleq (\Sigma^\star)^{-1/2} z_\infty$, then $\mathcal{U}$ follows a multi-variate standard normal distribution in 2 dimension; hence, $\|\mathcal{U}\|^2$ follows χ^2 with 2-degrees of freedom, and we have $\lambda_1\|\mathcal{U}\|^2 \leq \|z_\infty\|^2$, i.e., $\|z_\infty\|^2$ can be lower bounded in a.s. sense by a Gamma random variable with shape-scale parameters $(1, 2\lambda_1)$. Thus, for any $t \geq 0$,

$$\mathbb{P}\left[\|z_\infty\|^2 \geq t\right] \geq \mathbb{P}\left[\lambda_1\|\mathcal{U}\|^2 \geq t\right] = \Gamma\left(1, \frac{t}{2\lambda_1}\right) = e^{\frac{-t}{2\lambda_1}},$$

where $\Gamma : a, x \mapsto \int_x^\infty s^{a-1}e^{-s}\,\mathrm{d}\,s$ denotes the upper incomplete Gamma function. Thus,

$$Q_p(\|z_\infty\|^2) \geq 2\lambda_1 \log\left(\frac{1}{1-p}\right).$$

Now, we provide an upper bound on $|\Sigma_{12}^\star|$ to be able to give a further lower bound on λ_1 based on (C.17). First, observe that $\Sigma_{12}^\star = \frac{1-\theta}{P_c(\theta)} H(\theta)$ for

$$\begin{aligned} H(\theta) &\triangleq -2 - 6\theta + 14\theta^2 + 50\theta^3 - 14\theta^4 - 94\theta^5 - 6\theta^6 + 42\theta^7 + 48\theta^8 - 28\theta^9 - 8\theta^{10} + 4\theta^{11} \\ &= 2(1-\theta)^2(\theta^2 - 2\theta - 1)(\theta^2 + 2\theta - 1)(-1 - 5\theta - 8\theta^2 - 4\theta^3 + 2\theta^5). \end{aligned}$$

Note that $(\theta^2 - 2\theta - 1)(\theta^2 + 2\theta - 1) = (1-\theta^2)^2 - 4\theta^2 \in [-4, 1]$ for $\theta \in [0,1]$; furthermore, $h(\theta) \triangleq (-1 - 5\theta - 8\theta^2 - 4\theta^3 + 2\theta^5)$ satisfies $h(0) < 0$ and $h'(\theta) = -5 - 16\theta - 12\theta^2 + 10\theta^4 < 0$ for $\theta \in [0,1]$, which implies that $|h(\theta)| \leq |h(1)| = 16$. Hence,

$$|\Sigma_{12}^\star| \leq 128\,\frac{(1-\theta)^3}{P_c(\theta)},$$

and using (C.16) and (C.17), we can conclude that

$$\lambda_1 \geq (1-\theta)\min\left(\frac{D_1(\theta) - 128(1-\theta)^2}{P_c(\theta)}, \frac{D_2(\theta) - 128(1-\theta)^2}{P_c(\theta)}\right),$$

with $D_1(\theta) \triangleq \frac{P_c(\theta)}{1-\theta}\Sigma_{11}^\star$ and $D_2(\theta) \triangleq \frac{P_c(\theta)}{1-\theta}\Sigma_{22}^\star$. Hence

Algorithm 2 SGDA Algorithm

Require: Stepsize α. Starting point (x_0, y_0). Horizon N

1: **for** $k \geq 0$ **do**

2: $\quad y_{k+1} \leftarrow y_k + \alpha \widetilde{\nabla}_{\mathrm{y}} \Phi(x_k, y_k, \omega_k^y)$

3: $\quad x_{k+1} \leftarrow x_k - \alpha \widetilde{\nabla}_{\mathrm{x}} \Phi(x_k, y_k, \omega_k^x)$

return (x_N, y_N)

$$Q_p(\|z_\infty\|^2) \geq 2(1-\theta)\min\left(\tfrac{D_1(\theta)-128(1-\theta)^2}{P_c(\theta)}, \tfrac{D_2(\theta)-128(1-\theta)^2}{P_c(\theta)}\right)\log\left(\tfrac{1}{1-p}\right) \triangleq \psi_1(p,\theta),$$

Noting that $D_1(\theta), D_2(\theta) \to -32$ as $\theta \to 1$, we conclude that for any fixed $p \in (0,1)$, $\psi_1(p,\theta) = \Theta(1-\theta)$.

Upper bound. The CP parametrization corresponds to choosing $\alpha = \frac{1}{2\sigma} - \sqrt{\theta}L_{yy}$ in the matrix inequality [38, Cor. 1]. Under this parameterization, the metric $\mathcal{E}_n = \mathcal{D}_n/\rho$ simplifies to

$$\mathcal{E}_n = \frac{\theta}{1-\theta}\left(\frac{1}{\mu_x}x_n^2 + \frac{1}{\mu_y}y^2\right).$$

By From (4.25), the p-quantile of $\|z_\infty\|^2$ satisfies

$$Q_p(\|z_\infty\|^2) \leq \underbrace{(1-\theta)\max(\mu_x,\mu_y)\ \left(\frac{1}{1-\theta}96\mathcal{Q}\bar{\delta}^2\left(1+\max\left\{1, \frac{4}{3\theta}\frac{\|A\|_F^2}{\mathcal{Q}}\right\}\log\left(\frac{1}{1-p}\right)\right)\right)}_{\psi_2(p,\theta)\,\triangleq},$$

for any $p \in (0,1)$. It suffices therefore to show that $\mathcal{Q} = \Theta(1-\theta)$, and $\|A\|_F^2 = \Theta(1-\theta)$, as $\theta \to 1$. Given Table, we readily observe that $\|A_0\|^2 = \Theta(1-\theta)$ and $\|B_0\|, \|B_1\|, \|B_2\|$ are $\Theta(1-\theta)$ as $\theta \to 1$. This implies that $\mathcal{Q} = \Theta(1-\theta)$ as $\theta \to 1$.

Similarly, for $\|A\|_F^2$, it suffices to show that $\|A_1\|^2, \|A_2\|^2, \|A_3\|^2 = \Theta(1-\theta)$, as $\theta \to 1$. Following the same line of arguments, one may observe that $\|A_1\|^2 = \Theta(32\mu_x^{-1}(1-\theta))$, $\|A_2\|^2 = \Theta(512\mu_y^{-1}(1-\theta))$, and $\|A_3\|^2 = \Theta(128\mu_y^{-1}(1-\theta))$, as $\theta \to 1$. This ensures overall that $\|A\|^2 = \Theta(1-\theta)$ as $\theta \to 1$.

D Preliminary results for SGDA

In this section, we derive non-asymptotic convergence rate for the Proximal Stochastic Gradient Descent Ascent Algorithm (SGDA) holding with high probability. Our analysis relies on a recent concentration inequality derived in [18] Mention how we customize it in this setting, and challenges etc. otherwise refereees may think our results is obvious. For the main result of this section, we consider the more general problem

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\Phi(x,y). \tag{D.1}$$

and we replace the structured Assumption 1 by the following standard setting.

Assumption 1'. $\Phi\colon \mathbb{R}^{\mathrm{d}} \times \mathbb{R}^{\mathrm{d}} \to \mathbb{R}$ *is a continuously differentiable function such that:*

(i) For all $y \in \mathcal{Y}$, $\Phi(\cdot, y)$ is μ-strongly convex and L-smooth.

(ii) For all $x \in \mathcal{X}$, $\Phi(x, \cdot)$ is μ-strongly concave and L-smooth.

For any $k \in \mathbb{N}$, we introduce $z_k \triangleq (x_k, y_k)^\top$ and denote $\mathrm{z}^\star \triangleq (\mathrm{x}^\star, \mathrm{y}^\star)$ the unique saddle point of the SP problem in (D.1).

Theorem D.1. *Let $(x_k, y_k)_{k\in\mathbb{Z}_+}$ be the sequence generated by SGDA, initialized at an arbitrary couple $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$, and for a stepsize $\alpha \leq \frac{\mu}{4L^2}$. Then, for all $k \in \mathbb{Z}_+$ and $\delta \in (0,1)$, with probability greater than p:*

$$\mathcal{E}_n \leq (1-\frac{\alpha\mu}{2})^n\ \mathcal{E}_0 + \frac{32\alpha\delta^2}{\mu}\left(1+10\log\left(\frac{4}{1-p}\right)\right), \tag{D.2}$$

where $\mathcal{E}_n \triangleq \|x_n - \mathrm{x}^\star\|^2 + \|y_n - \mathrm{y}^\star\|^2$.

The proof of this theorem is deferred to the section D of the appendix.

D.1 Convergence analysis of SGDA

We analyze in this section the convergence of SAPD in high probability. We first recall that SGDA is known to converge in expectation as per the following result, proved in [11, Theorem III.2.].

Proposition D.2. *Let $(z_n)_{n\geq 0}$ be the sequence generated by SAPD with a stepsize $\alpha \in (0, \frac{\mu}{4L^2}]$. We have for any $n \in \mathbb{N}$*

$$\mathbb{E}[\|z_n - \mathrm{z}^\star\|^2] \leq (1-\alpha\mu)^n \|z_0 - \mathrm{z}^\star\|^2 + \frac{2\alpha}{\mu}\mathcal{V}$$

where $\mathcal{V}$ denotes a uniform upperbound on $(\mathbb{V}ar(\Delta_i^\Phi))_{i\geq 0}$.

We introduce the concatenated gradient operator $A : z \mapsto (\nabla_{\mathrm{x}} \Phi, \nabla_{\mathrm{y}} \Phi)^\top$. We recall that, in view of Assumption 1' A satisfies the coercive and the cocoercive properties

$$\langle A(u) - A(v), u - v\rangle \geq \mu\|u - v\|^2, \quad \forall u, v \in \mathcal{X} \times \mathcal{Y} \tag{D.3}$$

$$\langle A(u) - A(v), u - v\rangle \geq \frac{\mu}{4L^2}\|A(u) - A(v)\|^2, \quad \forall u, v \in \mathcal{X} \times \mathcal{Y} \tag{D.4}$$

We first prove a standard result on the monotonic decrease of the distance to the solution, displayed by the (non-stochastic) gradient descent ascent.

Lemma D.3. *If $\alpha \leq \frac{\mu}{4L^2}$, we have for all $n \in \mathbb{N}$,*

$$\|z_i - \mathrm{z}^\star - \alpha A(z_i)\|^2 \leq (1-\alpha\mu)\|z_i - \mathrm{z}^\star\|^2. \tag{D.5}$$

Proof. Noting that $A(\mathrm{z}^\star) = 0$, we have, by Eq. (D.3) and Eq. (D.4), for any $i \in \mathbb{N}$

$$\begin{aligned}
\|z_i - \mathrm{z}^\star - \alpha A(z_i)\|^2 &= \|z_i - \mathrm{z}^\star\|^2 - 2\alpha\langle z_i - \mathrm{z}^\star, A(z_i)\rangle + \alpha^2\|A(z_i)\|^2 \\
&\leq \|z_i - \mathrm{z}^\star\|^2 - \alpha\mu\|z_i - \mathrm{z}^\star\|^2 - \frac{\alpha\mu}{4L^2}\|A(z_i)\|^2 + \alpha^2\|A(z_i)\|^2 \\
&\leq (1-\alpha\mu)\|z_i - \mathrm{z}^\star\|^2 + \alpha\left(\alpha - \frac{\mu}{4L^2}\right)\|A(z_i)\|^2 \\
&\leq (1-\alpha\mu)\|z_i - \mathrm{z}^\star\|^2
\end{aligned}$$

with the last inequality following from the stepsize condition. □

In what follows, we let $\Delta_i^\Phi \triangleq (\Delta_i^x, \Delta_i^y)$ be the oracle noise at step $i \in \mathbb{N}$. We note that, by Assumption 4, Δ_i^Φ is δ-subGaussian for some $\delta > 0$. We prove now Theorem D.1.

Proof of TheoremD.1. We wish to apply the recursive control property 4.2. We first observe, in view of Lemma D.3 that, for all $n \geq 0$

$$\begin{aligned}
\|z_{n+1} - \alpha A(z_{n+1}) - \mathrm{z}^\star\|^2 &\leq (1-\alpha\mu)\|z_{n+1} - \mathrm{z}^\star\|^2 = (1-\alpha\mu)\|z_n - \alpha A(z_n) - \mathrm{z}^\star - \alpha\Delta_n^\Phi\|^2 \\
&\leq (1-\alpha\mu)V_n + D_n + R_n.
\end{aligned}$$

where $V_n \triangleq \|z_n - \alpha A(z_n) - \mathrm{z}^\star\|^2$, $D_n \triangleq -2\alpha(1-\alpha\mu)\langle z_n - \alpha A(z_n) - \mathrm{z}^\star, \Delta_n^\Phi\rangle$, and $R_n \triangleq (1-\alpha\mu)\alpha^2\|\Delta_n^\Phi\|^2$. Let us now check that V_n, D_n, and R_n satisfy the requirements of Proposition 4.2. Let $(\mathcal{F}_n)_{n\geq -1}$ be the filtration generated by $(\Delta_n^\Phi)_{n\geq 0}$, with $\mathcal{F}_{-1} \triangleq \{\emptyset, \Omega\}$. For any $n \in \mathbb{N}$, V_n is non-negative and $\mathcal{F}_n$-measurable. For any $\lambda > 0$, we have

$$\begin{aligned}
\mathbb{E}[\exp(\lambda D_n)|\mathcal{F}_n] &\leq \mathbb{E}(\exp(32\lambda^2\alpha^2\delta^2(1-\alpha\mu)^2 V_n)) \quad \text{(by Lemma 2.2)} \\
\text{and } \mathbb{E}[\exp(\lambda R_n)|\mathcal{F}_n] &\leq \mathbb{E}(\exp(8\lambda\alpha^2(1-\alpha\mu)\delta^2)) \quad \text{(by Lemma 2.1).}
\end{aligned}$$

Hence, by Proposition 4.2, for any $n \in \mathbb{N}$, any $p \in (0,1)$, the estimate

$$\|z_n - \alpha A(z_n) - \mathrm{z}^\star\|^2 \leq \left(1 - \frac{\alpha\mu}{2}\right)^n \|z_0 - \alpha A(z_0) - \mathrm{z}^\star\|^2 + \frac{32\alpha(1-\alpha\mu)\delta^2}{\mu}\left(1 + (1 + 8(1-\alpha\mu))\left(1 + \log\left(\frac{2}{1-p}\right)\right)\right)$$

is satisfied with probability at least greater than $(1+p)/2$. Notice finally that $\|z_{n+1} - \mathrm{z}^\star\|^2 \leq 2\|z_n - \alpha A(z_n) - \mathrm{z}^\star\|^2 + 2\alpha^2\|\Delta_n^\Phi\|^2 \leq 2\|z_n - \alpha A(z_n) - \mathrm{z}^\star\|^2 + 4\alpha^2\delta^2\log\left(\frac{4}{1-p}\right)$. We conclude with a union bound, observing that

$1 - \alpha\mu \leq 1$: for any $n \in \mathbb{N}$, the estimate

$$\begin{aligned}\|z_{n+1} - \mathrm{z}^\star\|^2 &\leq 2\left(1 - \frac{\alpha\mu}{2}\right)^n \|z_0 - \mathrm{z}^\star\|^2 + \frac{32\alpha(1-\alpha\mu)\delta^2}{\mu}\left(1 + (1 + 8(1-\alpha\mu))\left(1 + \log\left(\frac{2}{1-p}\right)\right)\right) \\ &\quad + 4\alpha^2\delta^2 \log\left(\frac{4}{1-p}\right) \\ &\leq \left(1 - \frac{\alpha\mu}{2}\right)^n \; 2\|z_0 - \mathrm{z}^\star\|^2 + \frac{32\alpha\delta^2}{\mu}\left(1 + 10\log\left(\frac{4}{1-p}\right)\right)\end{aligned}$$

is satisfied with probability greater than p. □

References

[1] Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155:1105–1123, 2012.

[2] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[3] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

[4] Aharon Ben-Tal and Marc Teboulle. An old-new concept of convex risk measures: the optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.

[5] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

[6] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40:120–145, 2011.

[7] Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. Accelerated schemes for a class of variational inequalities. *Mathematical Programming*, 165:113–149, 2017.

[8] Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. Stochastic optimization under time drift: iterate averaging, step-decay schedules, and high probability guarantees. *Advances in Neural Information Processing Systems*, 34:11859–11869, 2021.

[9] Damek Davis and Dmitriy Drusvyatskiy. High probability guarantees for stochastic convex optimization. In *Conference on Learning Theory*, pages 1411–1427. PMLR, 2020.

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

[11] Alireza Fallah, Asuman Ozdaglar, and Sarath Pattathil. An optimal multistage stochastic gradient method for minimax problems. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 3573–3579. IEEE, 2020.

[12] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[13] Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[14] 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.

[15] Mert Gürbüzbalaban, Andrzej Ruszczyński, and Landi Zhu. A stochastic subgradient method for distributionally robust non-convex and non-smooth learning. *Journal of Optimization Theory and Applications*, 194(3):1014–1041, 2022.

[16] Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31(2):1299–1329, 2021.

[17] J Harold, G Kushner, and George Yin. Stochastic approximation and recursive algorithm and applications. *Application of Mathematics*, 35, 1997.

[18] Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613. PMLR, 2019.

[19] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

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

[21] Daniel Levy, Ziteng Sun, Kareem Amin, Satyen Kale, Alex Kulesza, Mehryar Mohri, and Ananda Theertha Suresh. Learning with user-level privacy. *Advances in Neural Information Processing Systems*, 34:12466–12479, 2021.

[22] Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive sgd with momentum. *arXiv preprint arXiv:2007.14294*, 2020.

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

[24] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[25] Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, pages 1416–1424, 2016.

[26] Abhishek Panigrahi, Raghav Somani, Navin Goyal, and Praneeth Netrapalli. Non-gaussianity of stochastic gradient noise. *arXiv preprint arXiv:1910.09626*, 2019.

[27] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

[28] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1571–1578, 2012.

[29] Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 813(814):46, 2015.

[30] Alexander Schied*. Risk measures and robust optimization problems. *Stochastic Models*, 22(4):753–831, 2006.

[31] Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

[32] Prateek Varshney, Abhradeep Thakurta, and Prateek Jain. (nearly) optimal private linear regression for sub-gaussian data via adaptive clipping. In *Conference on Learning Theory*, pages 1126–1166. PMLR, 2022.

[33] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[34] Killian Wood and Emiliano Dall'Anese. Online Saddle Point Tracking with Decision-Dependent Data. *arXiv e-prints*, page arXiv:2212.02693, December 2022.

[35] Yan Yan, Yi Xu, Qihang Lin, Wei Liu, and Tianbao Yang. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5789–5800. Curran Associates, Inc., 2020.

[36] Guodong Zhang, Yuanhao Wang, Laurent Lessard, and Roger B Grosse. Near-optimal local convergence of alternating gradient descent-ascent for minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 7659–7679. PMLR, 2022.

[37] Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, 194(1-2):901–935, jun 2021.

[38] Xuan Zhang, Necdet Serhat Aybat, and Mert Gürbüzbalaban. Robust accelerated primal-dual methods for computing saddle points, 2023. Available at https://arxiv.org/pdf/2111.12743.

[39] Dongruo Zhou, Jinghui Chen, Yuan Cao, Yiqi Tang, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

[40] Landi Zhu, Mert Gürbüzbalaban, and Andrzej Ruszczyński. Distributionally robust learning with weakly convex losses: Convergence rates and finite-sample guarantees. *arXiv preprint arXiv:2301.06619*, 2023.