Journal of Machine Learning Research 25 (2024) 1-56 Submitted 7/23; Revised 2/24; Published 5/24

High Probability and Risk-Averse Guarantees for a Stochastic Accelerated Primal-Dual Method

Yassine Laguel YASSINE.LAGUEL@UNIV-COTEDAZUR.FR
Laboratoire Jean Alexandre Dieudonné
Université Côte d'Azur
Nice, France.

Necdet Serhat Aybat NSA10@PSU.EDU
Department of Industrial and Manufacturing Engineering
Pennsylvania State University
University Park, PA, USA

Mert Gürbüzbalaban MGURBUZBALABAN@BUSINESS.RUTGERS.EDU
Department of Management Science and Information Systems
Rutgers Business School, Rutgers University.
Piscataway, NJ, USA.

Editor: Martin Jaggi

Abstract

We consider stochastic strongly-convex-strongly-concave (SCSC) saddle point (SP) problems which frequently arise in applications ranging 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 limiting covariance matrix of the iterates for a class of stochastic SCSC quadratic problems where the gradient noise is additive and Gaussian. This allows us to develop lower bounds for this class of quadratic problems which show that our analysis is tight in terms of the high probability bound dependence on the problem parameters. We also provide a risk-averse convergence analysis characterizing the "Conditional Value at Risk", the "Entropic Value at Risk", and the χ^2-divergence of the distance to the saddle point for the iterate sequence, highlighting the trade-offs between the bias and the risk associated with an approximate solution obtained by terminating the algorithm at any iteration.

Keywords: stochastic min-max optimization, high-probability guarantees, risk measures, accelerated primal-dual methods

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, $f : \mathcal{X} \to \mathbb{R}\cup\{+\infty\}$ and $g : \mathcal{Y}\to\mathbb{R}\cup\{+\infty\}$ are closed convex functions, and $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a smooth convex-concave function such that $\mathcal{L}(x, y)$ is strongly convex in x and strongly concave in y, i.e., SCSC – see Assumption 1 for details. Throughout the paper, we assume that f and g are strongly convex; indeed, whenever $\mathcal{L}$ is SCSC, this assumption holds without loss of generality as strong convexity/concavity can be transferred from Φ to f and g by adding and subtracting simple quadratics (see also Remark 1 for details).

Such saddle point (SP) problems arise in many applications and different contexts. In constrained optimization problems, saddle-point formulations arise naturally when the problems are reformulated as a minimax problem based on the Lagrangian duality. Furthermore, the SP formulation in (1.1) encompasses many key problems such as *robust optimization* (Ben-Tal et al., 2009) – here g is selected to be the indicator function of an 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 involving SCSC problems include but are not limited to *supervised learning* with non-separable regularizers (where $\Phi(x, y)$ may not be bilinear) (Palaniappan and Bach, 2016), *fairness* in machine learning (Liu et al., 2022), *unsupervised learning* (Palaniappan and Bach, 2016) and various *image processing* problems, e.g., denoising, (Chambolle and Pock, 2011, 2016).

In this work, we are interested in the setting where the partial gradients $\nabla_x\Phi$ and $\nabla_y\Phi$ are not deterministically available; but, instead we postulate that their stochastic estimates $\tilde{\nabla}_x\Phi$ and $\tilde{\nabla}_y\Phi$ are accessible. 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 (Zhu et al., 2023; Gürbüzbalaban et al., 2022; Bottou et al., 2018). First-order (FO) methods that rely on stochastic estimates of the gradient information have been the leading computational approach for computing low-to-medium-accuracy solutions for these problems because of their cheap iterations and mild dependence on the problem dimension and data size. In this paper, our focus is on the first-order primal-dual algorithms that rely on stochastic gradient estimates for solving (1.1).

Existing relevant work. Stochastic primal-dual algorithms for solving SP problems generate a sequence of primal and dual iterate pairs $z_n = (x_n, y_n) \in \mathcal{X} \times \mathcal{Y} \triangleq \mathcal{Z}$ starting from an initial point $(x_0, y_0) \in \mathbf{dom}\, f \times \mathbf{dom}\, g \triangleq Z$. Two popular metrics to assess the quality of a random solution $(\hat{x}, \hat{y})$ returned by a stochastic algorithm are the *expected gap* and the *expected squared distance* defined as

$$\mathcal{G}(\hat{x}, \hat{y}) \triangleq \mathbb{E}\big[\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \big\{\mathcal{L}(\hat{x}, y) - \mathcal{L}(x, \hat{y})\big\}\big], \qquad \mathcal{D}(\hat{x}, \hat{y}) \triangleq \mathbb{E}[\|\hat{x} - \mathrm{x}^\star\|^2 + \|\hat{y} - \mathrm{y}^\star\|^2], \quad (1.2)$$

respectively, where $(\mathrm{x}^\star, \mathrm{y}^\star)$ denotes the *unique* saddle point of (1.1), due to the strong convexity of f and g. The iteration complexity of FO-methods in these two metrics depend naturally on the block Lipschitz constants L_{xx}, L_{xy}, L_{yy} and L_{yx}, i.e., Lipschitz constants of $\nabla_x\Phi(\cdot, y)$, $\nabla_x\Phi(x, \cdot)$, $\nabla_y\Phi(x, \cdot)$ 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 –see Assumption 1 for precise definition of these constants. In particular, Fallah et al. (2020) show that a multi-stage variant of Stochastic Gradient Descent Ascent (SGDA) algorithm generates (x_ϵ, y_ϵ) such that $\mathcal{D}(x_\epsilon, y_\epsilon) \leqslant \epsilon$ within $\mathcal{O}(\kappa^2 \ln(1/\epsilon) + \frac{\delta^2}{\mu^2}\frac{1}{\epsilon})$ gradient oracle calls, where $\delta^2 = \max\{\delta_x^2, \delta_y^2\}$, while δ_x^2 and δ_y^2 are bounds on the variance of the stochastic gradients $\tilde{\nabla}_x\Phi$ and $\tilde{\nabla}_y\Phi$, respectively; $\mu \triangleq \min\{\mu_x, \mu_y\}$ and

$L \triangleq \max\{L_{xx}, L_{xy}, L_{yx}, L_{yy}\}$ are the worst-case strong convexity and Lipschitz constants, and $\kappa \triangleq L/\mu$ is defined as the *condition number*. SGDA analyzed in (Fallah et al., 2020) employs Jacobi-type updates, i.e., stochastic gradient descent and ascent steps are taken simultaneously. Jacobi-type updates are easier to analyze than Gauss-Seidel updates in general, and can be viewed as solving a structured variational inequality (VI) problem, for which there are many existing techniques that directly apply, e.g., (Gidel et al., 2018; Chen et al., 2017). For deterministic SCSC problems, Zhang et al. (2022) consider gradient descent ascent (GDA) with Gauss-Seidel-type updates i.e., the primal and dual variables are updated in an alternating fashion using the most recent information obtained from the previous update step. Their results show that an accelerated asymptotic convergence rate can be obtained for the Gauss-Seidel variant of GDA, i.e., iteration complexity scales linearly with κ instead of κ^2. However, as discussed in (Zhang et al., 2022), this comes at a price: Gauss-Seidel style updates greatly complicate the analysis because every iteration of the algorithm turns out to be a composition of two half updates. Furthermore, extending the acceleration result to non-asymptotic rates requires using a momentum term in either the primal or the dual updates, and this further complicates the convergence analysis. Fallah et al. (2020) also considered using momentum terms both in the primal and dual updates, and show that the multi-stage *Stochastic Optimistic Gradient Descent Ascent* (OGDA) algorithm using Jacobi-type updates achieves an iteration complexity of $\mathcal{O}(\kappa \ln(1/\epsilon) + \frac{\delta^2}{\mu^2\epsilon})$ to guarantee an ϵ-solution in terms of expected squared distance. There are also several other algorithms that can achieve the accelerated rate, i.e., $\log(1/\epsilon)$ has the coefficient κ instead of κ^2 - see, e.g. (Beznosikov et al., 2022). We call this term that depends on the condition number as *initialization bias* since it captures how fast the error due to initial conditions decay and reflects the behavior of the algorithm in the noiseless setting. Among the algorithms that achieve an accelerated rate, the most closely related work to ours is (Zhang et al., 2024) which develops a stochastic accelerated primal-dual (SAPD) algorithm with Gauss-Seidel type updates. SAPD using a momentum acceleration only in the dual variable can generate (x_ϵ, y_ϵ) such that $\mathbb{E}[\mu_x\|x_\epsilon - \mathrm{x}^\star\|^2 + \mu_y\|y_\epsilon - \mathrm{y}^\star\|^2] \leqslant \epsilon$ within $\mathcal{O}\Big(\Big(\frac{L_{xx}}{\mu_x} + \frac{L_{yx}}{\sqrt{\mu_x\mu_y}} + \frac{L_{yy}}{\mu_y} + \Big(\frac{\delta_x^2}{\mu_x} + \frac{\delta_y^2}{\mu_y}\Big)\frac{1}{\epsilon}\Big)\log(\frac{1}{\varepsilon})\Big)$ iterations; this result implies $\tilde{\mathcal{O}}(\kappa \ln(\varepsilon^{-1}) + \mu^{-2}\delta^2\varepsilon^{-1})$ bound in terms of the expected squared distance – this complexity is optimal for bilinear problems. To our knowledge, SAPD is also the fastest single-loop algorithm for solving stochastic SCSC problems in the form of (1.1) that are non-bilinear; furthermore, using acceleration only in one update, as opposed to in both variables (Fallah et al., 2020), leads to smaller variance accumulation (see Zhang et al. (2024) for more details).

While the aforementioned results provide performance guarantees in expectation based on the metrics defined in (1.2) and their variants, unfortunately having guarantees in these metrics do not allow us to control tail events, i.e., the expected gap and distance can be smaller than a given target threshold $\varepsilon > 0$; but, the iterates can 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 $p \in (0, 1)$. This is illustrated Figure 1 (**Left**) where we run the SAPD algorithm for two different values of the momentum parameter $\theta = 0.95$ and $\theta = 0.99$ for a toy problem $\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} \frac{1}{2}x^2 + xy + \frac{1}{2}y^2$ with strong convexity parameters $\mu_x = \mu_y = 1$

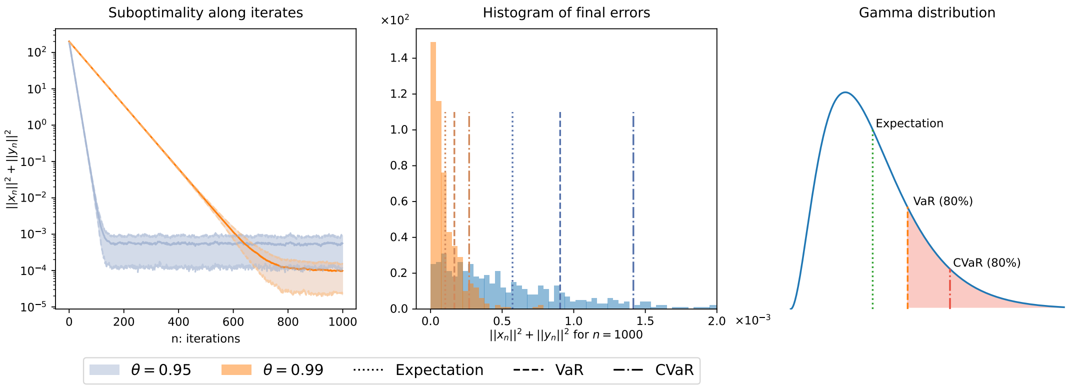

Figure 1: **(Left)** Convergence of SAPD on the saddle point problem $\min_{x \in \mathbb{R}} \max_{y \in \mathbb{R}} x^2/2 + xy + y^2/2$, initialized at $x_0 = y_0 = 10$ with momentum parameters $\theta = 0.95$ and $\theta = 0.99$. **(Middle)** Histogram of the distribution of the SAPD iterates (x_n, y_n) after $n = 1000$ iterations for 500 runs, with corresponding momentum parameters $\theta = 0.95$ and $\theta = 0.99$. **(Right)** Illustration of the expectation $\mathbb{E}(X)$, p-th quantile $(\mathrm{VaR}_p(X) = Q_p(X))$ and $\mathrm{CVaR}_p(X)$ for $p = 80\%$, where X is a gamma-distributed random variable with shape parameter 3 and scale parameter 5.

admitting a saddle point at $\mathrm{x}^\star = \mathrm{y}^\star = 0$. SAPD is initialized at $(x_0, y_0) = (10, 10)$, primal and dual stepsizes are chosen according to the Chambolle-Pock parametrization as suggested in (Zhang et al., 2024). In this specific example, for simplicity, the stochastic gradients $\tilde{\nabla}_x \Phi$ and $\tilde{\nabla}_y \Phi$ are set to $\nabla_x \Phi$ and $\nabla_y \Phi$ perturbed with additive i.i.d. Gaussian $\mathcal{N}(0, 0.1)$ noise, and we assume that $\tilde{\nabla}_x \Phi$ and $\tilde{\nabla}_y \Phi$ are independent from each other and also from the past history of the algorithm. For each parameter choice, we call SAPD for 500 runs, and for each run we compute $n = 1000$ iterations –see Figure 1 (**Left**), and plot the distribution of the squared distance of (x_n, y_n) to the unique saddle point $(\mathrm{x}^\star, \mathrm{y}^\star)$, i.e., $E_n \triangleq \|x_n - \mathrm{x}^\star\|^2 + \|y_n - \mathrm{y}^\star\|^2$, in Figure 1 (**Middle**).

As we can see in the middle plot, the random error E_n can take significantly larger values than its expectation $\mathcal{D}(x_n, y_n) = \mathbb{E}[E_n]$. This motivates estimating the p-th quantile of the error E_n, which is also called the *value at risk* at level p, traditionally abbreviated as $Q_p(E_n)$ in the financial literature. While quantiles represent a worst-case error E_n associated with a probability p, they do not capture the behavior if that worst-case threshold is ever crossed. Conditional value at risk (CVaR) at level p, on the other hand, is an alternative risk measure that can be used characterizing the expected error if that worst-case threshold is ever crossed. CVaR is in fact a coherent risk measure with some desirable properties (Rockafellar and Royset, 2013). In Fig. 1(**Middle**), we report the average, the quantile, and the CVaR at the safety threshold $p = 80\%$ for $n = 1000$. We can see that quantile and the conditional value at risk capture the tail behavior better than the expectation. Similar behavior can be seen on other distributions, e.g., in Figure 1(**Right**), we illustrate the expectation, the p-th quantile, and the p-conditional value at risk of a gamma-distributed random variable with shape parameter 3, and scale parameter 5 corresponding to $p = \%80$. Since CVaR at level $p \in (0, 1)$ estimates the average of the tails after the p-th quantile, it is useful for capturing the average risk associated to tail events beyond the p-th quantile. In addition to CVaR, there are also other coherent risk measures such as entropic value at risk (EVaR) and χ^2-divergence which have been of interest in the study of stochastic optimization algorithms as they can provide risk-averse guarantees capturing the worst-case tail behavior and deviations from the mean performance (Can and Gürbüzbalaban, 2022).

While high-probability guarantees (quantile guarantees) and risk guarantees in terms of risk measures such as CVaR and EVaR are available in the optimization setting for the iterates of stochastic gradient descent-like methods (Harvey et al., 2019; Rakhlin et al., 2012; Davis and Drusvyatskiy, 2020; Can and Gürbüzbalaban, 2022), results in a similar nature are considerably more limited in the SP setting. Among existing results, Juditsky et al. (2011) obtained high-probability guarantees for the stochastic mirror-prox algorithm for solving stochastic VIs with Lipschitz and monotone operators. This algorithm can be used to solve smooth stochastic convex/concave SP problems, corresponding to the case $f = g = 0$ with $\Phi(x, y)$ being smooth, convex in x and concave in y, and implies that with probability $p \in (0, 1)$, after n iterations, the gap metric for the VI will be bounded by $\mathcal{O}(\frac{1}{\sqrt{n}} + \log(\frac{1}{1-p})\frac{1}{n})$ assuming that the domain is bounded and the stochastic gradient noise is light-tailed with a sub-Gaussian distribution. In (Gorbunov et al., 2022), it is shown that the same high-probability results as in (Juditsky et al., 2011) can be attained by clipping the gradients properly without resorting to the sub-Gaussian and bounded domain assumptions. It should be emphasized that (Gorbunov et al., 2022) is designed for solving stochastic variational inequalities (stochastic VIs), uses the Lipschitz constant of the gradient operator in determining step sizes, and does not exploit the block Lipschitz structure of the minmax problem; furthermore, it also does not handle closed convex functions f and g. In another line of work (Yan et al., 2020), it is shown for the SGDA algorithm that the expected gap guarantee $\mathcal{G}(x_n, y_n) \leqslant \varepsilon$ can be achieved with probability at least $p \in (0, 1)$ after $n = \mathcal{O}\left(\frac{1}{\varepsilon}\log(\frac{1}{1-p}) + \frac{\delta^2}{\mu\varepsilon}\log(\frac{1}{1-p})\right)$ oracle calls for possibly non-smooth SCSC problems. Moreover, in (Wood and Dall'Anese, 2022), high probability bounds are given for online algorithms applied to stochastic saddle point problems where the objective is time-varying and revealed in a sequential manner –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 with Jacobi-style updates; therefore, the high probability bounds do not exhibit accelerated decay of the initialization bias, and scale as κ^2, i.e., quadratically with the condition number κ, instead of a linear scaling. To our knowledge, high-probability bounds for algorithms with Gauss-Seidel style updates are not available in the literature on SP problems even if they do not incorporate momentum, see the survey by Beznosikov et al. (2022). Similarly, we are not aware of any risk guarantees (in terms of CVaR and EVaR of the performance metric over iterations) for any primal-dual algorithm for solving stochastic SP problems.

Contributions. In this paper, we present a risk-averse analysis of the `SAPD` method (Zhang et al., 2024) to solve saddle point problems of the form (1.1). A key novelty of our work lies in providing the first analysis of an accelerated algorithm for SCSC problems with high probability guarantees, where our bounds reflect the accelerated decay of the initialization bias scaling linearly with the condition number κ. Indeed, our high-probability bounds provided in Section 3 imply that given target accuracy $\varepsilon > 0$, `SAPD`, with a proper choice of parameters that we state explicitly, can generate a solution (x_n, y_n) satisfying $\mu_x\|x_n - \mathrm{x}^\star\|^2 + \mu_y\|y_n - \mathrm{y}^\star\|^2 \leqslant \varepsilon$

with probability $p \in (0,1)$ after n iterations for $n \in \mathbb{Z}_+$ satisfying

$$n = \mathcal{O}\left(\left[\frac{\mathrm{L}_{\mathrm{xx}}}{\mu_x} + \frac{\mathrm{L}_{\mathrm{yx}}}{\sqrt{\mu_x \mu_y}} + \frac{\mathrm{L}_{\mathrm{yy}}}{\mu_y} + \left(1 + \frac{\mathrm{L}_{\mathrm{xy}}}{\mathrm{L}_{\mathrm{yx}}} + \frac{\mathrm{L}_{\mathrm{xy}}^2}{\mathrm{L}_{\mathrm{yx}}^2}\right) \max\left(\frac{\delta_x^2}{\mu_x}, \frac{\delta_y^2}{\mu_y}\right) \frac{1 + \log\left(\frac{1}{1-p}\right)}{\varepsilon}\right] \cdot \log\left(\frac{(1 + \frac{\mathrm{L}_{\mathrm{xy}}^2}{\mathrm{L}_{\mathrm{yx}}^2})\mathcal{S}_0}{\varepsilon}\right)\right), \tag{1.3}$$

where $\mathcal{S}_0 \triangleq \mu_x \|x_0 - \mathrm{x}^\star\|^2 + \mu_y \|y_0 - \mathrm{y}^\star\|^2$. When the partial gradients $\nabla_x \Phi$ and $\nabla_y \Phi$ are continuously differentiable, which is the case for bilinear problems and for many SCSC problems arising in practice (Zhang et al., 2024; Palaniappan and Bach, 2016; Chambolle and Pock, 2011, 2016), we can take $\mathrm{L}_{\mathrm{xy}} = \mathrm{L}_{\mathrm{yx}}$ (as discussed in Remark 13) and the complexity in (1.3) simplifies to

$$n = \mathcal{O}\left(\left[\tfrac{\mathrm{L}_{\mathrm{xx}}}{\mu_x} + \tfrac{\mathrm{L}_{\mathrm{yx}}}{\sqrt{\mu_x \mu_y}} + \tfrac{\mathrm{L}_{\mathrm{yy}}}{\mu_y} + \max\left(\tfrac{\delta_x^2}{\mu_x}, \tfrac{\delta_y^2}{\mu_y}\right) \tfrac{1 + \log\left(\frac{1}{1-p}\right)}{\varepsilon}\right] \log\left(\tfrac{1}{\varepsilon}\right)\right),$$

hiding constants depending on the initialization. Simplifying the terms further, we can conclude that $n = \mathcal{O}\left(\kappa \log(\frac{1}{\varepsilon}) + (1 + \log(\frac{1}{1-p}))\frac{\delta^2 \log(1/\varepsilon)}{\mu\varepsilon}\right)$ iterations are sufficient, where $\delta^2 = \max(\delta_x^2, \delta_y^2)$. To achieve this, under a light-tail (subGaussian) assumption on the norm of the gradient noise, we develop concentration inequalities tailored to the specific *Gauss-Seidel* structure of `SAPD`. The Gauss-Seidel type updates and the use of a momentum term within `SAPD` complicate the analysis significantly as the evolution of the iterates and the performance metric over the iterations need to be studied with respect to a non-standard filtration for having the right measurability properties (as discussed in Section 5.2 in detail). Deriving these results requires construction of a new Lyapunov function, V_n, with some favorable contraction properties. To our knowledge, SP problems, these are the first high-probability guarantees for an algorithm with Gauss-Seidel updates, and first high-probability guarantees involving acceleration in the sense that the iteration complexity's dependence on the initialization bias scaling linearly with the condition number κ. Indeed, one of our main contributions is to show that this desirable κ-dependence can be preserved when extending the guarantees in expectation to high probability bounds for the SAPD algorithm.

Table 1 provides a summary of our results, providing a comparison with the existing relevant work. To the best of our knowledge, our work is the only one that can provide high-probability bounds in the presence of non-smooth regularizers, i.e., when there are closed convex functions f and g. In addition, existing high-probability guarantees are obtained only for Jacobi-style algorithms, our results are the first high-probability results for a primal-dual algorithm with Gauss-Seidel-type updates in the minimax setting –although Gauss-Seidel (GS) type updates are harder to analyze, adopting GS updates often leads to faster and theoretically better methods than those using Jacobi-type updates Zhang et al. (2022). Finally, our analysis exploits the structure information of the minimax problems better than the analyses provided for stochastic VI methods; indeed, our high-probability results (in Corollary 12) are the only ones that capture the precise effect of the block Lipschitz/strong convexity constants $(\mathrm{L}_{\mathrm{xx}}, \mathrm{L}_{\mathrm{xy}}, \mathrm{L}_{\mathrm{yx}}, \mathrm{L}_{\mathrm{yy}}, \mu_x, \mu_y)$ on the iteration complexity. The previous high-probability bounds for min-max problems in the literature were given in terms of worst-case constants $L = \max(\mathrm{L}_{\mathrm{xx}}, \mathrm{L}_{\mathrm{xy}}, \mathrm{L}_{\mathrm{yx}}, \mathrm{L}_{\mathrm{yy}})$ and $\mu = \min(\mu_x, \mu_y)$ which result in a significantly loose analysis as the block Lipschitz constants $\mathrm{L}_{\mathrm{xx}}, \mathrm{L}_{\mathrm{xy}}, \mathrm{L}_{\mathrm{yx}}, \mathrm{L}_{\mathrm{yy}}$ and strong convexity constants μ_x, μ_y are generally not of the same order. Basically, our primal and dual stepsizes, τ and σ, as well as the momentum parameter, θ, can adapt to the block Lipschitz

Algorithm	**Complexity — Exploits Structure**	**f, g**	**Acceleration**	**High Prob.**	**Metric**
Hsieh et al. (2019)	$\mathcal{O}\left(\frac{1}{\varepsilon}+\frac{\delta^2}{\mu}\varepsilon\right)$ — ✗	✗	✗	✗	$\mathcal{D}$
Fallah et al. (2020)	$\mathcal{O}\left(\kappa\log\left(\frac{1}{\varepsilon}\right)+\frac{\delta^2}{\mu}\frac{\log(1/\varepsilon)}{\varepsilon}\right)$ — ✗	✗	✓	✗	$\mathcal{D}$
Zhang et al. (2024)	$\mathcal{O}\left(\kappa\log\left(\frac{1}{\varepsilon}\right)+\frac{\delta^2}{\mu}\frac{\log(1/\varepsilon)}{\varepsilon}\right)$ — ✓	✓	✓	✗	$\mathcal{D}$
Yan et al. (2020)[†]	$\mathcal{O}\left(\frac{\delta^2}{\mu\varepsilon}\log(1/(1-p))\right)$ — ✗	✗	✗	✓	$\mathcal{G}$
Gorbunov et al. (2022)[‡]	$\mathcal{O}\left(\max\left\{\kappa,\frac{\delta^2}{\mu\varepsilon}\right\}\log\left(\frac{1}{\varepsilon}\right)\ln\left(\frac{\kappa}{1-p}\right)\right)$ — ✗	✗	✓	✓	$\mathcal{D}$
Our Paper (Cor. 12)	$\mathcal{O}\left(\kappa\log\left(\frac{1}{\varepsilon}\right)+\left(1+\log\left(\frac{1}{1-p}\right)\right)\frac{\delta^2\log(1/\varepsilon)}{\mu\varepsilon}\right)$ — ✓	✓	✓	✓	$\mathcal{D}$

Table 1: Summary of relevant work for SCSC problems. The second column reports the complexity (number of iterations needed) to achieve ε-accuracy and whether the results exploit the block primal-dual structure (by specifying the dependence to constants $\mu_x, \mu_y, \mathrm{L_{xx}}, \mathrm{L_{xy}}, \mathrm{L_{yx}}, \mathrm{L_{yy}}$ explicitly); if the block structure is not exploited, that means the results are given in terms of the worst-case constants $\mu = \min(\mu_x, \mu_y)$ and $L = \max(\mathrm{L_{xx}}, \mathrm{L_{xy}}, \mathrm{L_{yx}}, \mathrm{L_{yy}})$ instead. The third column reports whether possibly non-smooth, closed convex f and g are supported in the analysis. The fourth column is about acceleration, whether the bias term proportional to $\log(1/\varepsilon)$ has linear dependence to κ. The fifth column indicates whether the results provide high-probability bounds. The sixth column indicates the metric used to quantify convergence, as defined in (1.2). **Table notes:** [†] Yan et al. (2020) requires two nested loops. [‡] Gorbunov et al. (2022) employ gradient clipping techniques.

structure; therefore, SAPD can leverage the structure specific to the minimax problems that VI problems do not possess. Furthermore, our analysis technique based on the concentration inequality derived in Proposition 20 can be of independent interest: in principle, it can be used to analyze various different primal-dual methods for solving minimax problems, including SAPD and GS-style stochastic gradient descent ascent[1] (SGDA) methods.

We also provide finite-time risk guarantees, in terms of the CVaR, EVaR and χ^2-divergence of the distance to the saddle point. In addition, we provide an in-depth analysis of the behavior of SAPD on a class of quadratic problems subject to i.i.d. isotropic Gaussian noise where we can characterize the behavior of the distribution of the iterates explicitly. In particular, we derive an analytical formula for the asymptotic covariance matrix of SAPD's iterates, which demonstrates the tightness of our high probability bounds with respect to several parameter choices in SAPD. To our knowledge, these are the first risk-averse guarantees that quantify the risk associated with an *approximate* solution generated by a primal-dual algorithm for SP problems.

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 positive integers and $\mathbb{Z}_{+} = \mathbb{Z}_{++} \cup \{0\}$. For $A, B \in \mathbb{R}^{n\times n}$, we denote $\mathrm{Vec}(A) \in \mathbb{R}^{n^2}$ the vector composed of the vertical concatenation of the columns of A, and $A \otimes B$ the Kronecker product of A and B. We let $\|A\|$ denote the spectral norm of A and let $\rho(A)$ denote the *spectral radius* of A, i.e., the largest modulus of the eigenvalues of A. For a finite sequence of reals $x_1, \ldots, x_n$ (resp. matrices $X_1, \ldots, X_n$), we denote $\mathrm{Diag}(x_1 \ldots, x_n)$ (resp. $\mathrm{Diag}(X_1 \ldots, X_n)$) the associated (block) diagonal matrix. If A is diagonalizable, $\mathrm{Sp}(A)$ denotes the set of the eigenvalues of A. For any convex set C, $\mathcal{I}_C$ denotes the indicator

[1] SGDA can be seen as a special case of SAPD with the momentum parameter $\theta = 0$.

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 $\varphi\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$, $\operatorname{prox}_\varphi(\cdot)$ denotes the associated *proximal operator*: $x \mapsto \arg\min_{u \in \mathcal{X}} \varphi(u) + \frac{1}{2}\|u - x\|^2$. We use the Landau notation to describe the asymptotic behavior of functions. That is, 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 $x \to u$, whereas $f(x) = \mathcal{O}(g(x))$ if there exist a positive constant C such that $|f(x)| \leqslant C|g(x)|$ in some neighborhood of u. Similarly, we say $f(x) = \Theta(g(x))$, if $f(x) = \mathcal{O}(g(x))$ and $g(x) = \mathcal{O}(f(x))$. Given random vectors $U_n : \Omega \to \mathbb{R}^d$ for $n \geqslant 0$, we let $U_n \xrightarrow{\mathcal{D}} U$ if U_n converges in distribution to another random vector $U : \Omega \to \mathbb{R}^d$. In Appendix A, we provide a table summarizing the key notation used within the paper together with references to where they are introduced in the text.

2. Preliminaries and Background

2.1 Stochastic Accelerated Primal-Dual (SAPD) Method

SAPD, displayed in Algorithm 1, is a stochastic accelerated primal-dual method developed in (Zhang et al., 2024) 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 (Hamedani and Aybat, 2021) to the stochastic setting, which itself is an extension of the Chambolle-Pock (CP) method (Chambolle and Pock, 2011, 2016) developed for bilinear couplings Φ. Given primal and dual stepsizes τ and σ and a number of iterations n, SAPD applies momentum averaging to the partial gradients with respect to the dual variable, and updates the primal and the dual variables in an alternating fashion computing proximal-gradient steps. While the gradients of Φ are stochastic and inexact, it is assumed that proximal steps with respect to non-smooth terms f and g are computed in an exact fashion.[2]

Algorithm 1 SAPD Algorithm

```
Require: Parameters $\tau, \sigma, \theta$. 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 $n = 0, \dots, N-1$ do
    $\tilde{\mathrm{s}}_n \leftarrow \tilde{\nabla}_{\mathrm{y}}\, \Phi(x_n, y_n, \omega_n^y) + \theta\, \tilde{\mathrm{q}}_n$        ▷ Momentum averaging
    $y_{n+1} \leftarrow \operatorname{prox}_{\sigma g}(y_n + \sigma\, \tilde{\mathrm{s}}_n)$
    $x_{n+1} \leftarrow \operatorname{prox}_{\tau f}(x_n - \tau\, \tilde{\nabla}_{\mathrm{x}}\, \Phi(x_n, y_{n+1}, \omega_n^x))$
    $\tilde{\mathrm{q}}_{n+1} \leftarrow \tilde{\nabla}_{\mathrm{y}}\, \Phi(x_{n+1}, y_{n+1}, \omega_{n+1}^y) - \tilde{\nabla}_{\mathrm{y}}\, \Phi(x_n, y_n, \omega_n^y)$
return $z_N = (x_N, y_N)$
```

The high-probability convergence guarantees of SAPD derived in this paper rely on several standard assumptions on f, g, Φ, and on the noisy estimates $\tilde{\nabla}_x \Phi$ and $\tilde{\nabla}_y \Phi$ of the partial gradients of Φ. The first assumption on the smoothness properties of the coupling

[2] In many cases, these proximal steps are easy to compute exactly or up to high accuracy; for instance when f and g are taken as indicator functions of some convex sets, or when f and g are ℓ_1-regularizers; see also Parikh et al. (2014) for other examples where the proximal operators $\operatorname{prox}_{\sigma g}(\cdot)$ and $\operatorname{prox}_{\tau f}(\cdot)$ admit an explicit formula.

function Φ is standard for first-order methods (see e.g.. (Mokhtari et al., 2020; Gidel et al., 2018; Zhang et al., 2021)).

Assumption 1 *$f\colon \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ and $g\colon \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ are strongly convex with convexity modulii $\mu_x, \mu_y > 0$, respectively; and $\Phi\colon\ \mathbb{R}^{\mathrm{d}} \times \mathbb{R}^{\mathrm{d}} \to \mathbb{R}$ is continuously differentiable on an open set containing $\mathbf{dom}\, f \times \mathbf{dom}\, g$ such that*

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

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

(iii) there exist $\mathrm{L_{xx}}, \mathrm{L_{yy}} \geqslant 0$ and $\mathrm{L_{xy}}, \mathrm{L_{yx}} > 0$ such that

$$\begin{aligned}\| \nabla_{\mathrm{x}} \Phi(x,y) - \nabla_{\mathrm{x}} \Phi(\bar{x},\bar{y})\| &\leqslant \mathrm{L_{xx}}\, \|x-\bar{x}\| + \mathrm{L_{xy}}\, \|y-\bar{y}\|,\\ \| \nabla_{\mathrm{y}} \Phi(x,y) - \nabla_{\mathrm{y}} \Phi(\bar{x},\bar{y})\| &\leqslant \mathrm{L_{yx}}\, \|x-\bar{x}\| + \mathrm{L_{yy}}\, \|y-\bar{y}\|,\end{aligned}$$

for all $(x,y), (\bar{x},\bar{y}) \in \mathbf{dom}\, f \times \mathbf{dom}\, g$.

By strong convexity/strong concavity of $\mathcal{L}$ from Assumption 1, the problem in (1.1) admits a unique saddle point $\mathrm{z}^\star \triangleq (\mathrm{x}^\star, \mathrm{y}^\star)$ which satisfies

$$\mathcal{L}(\mathrm{x}^\star, y) \leqslant \mathcal{L}(\mathrm{x}^\star, \mathrm{y}^\star) \leqslant \mathcal{L}(x, \mathrm{y}^\star), \quad \forall (x,y) \in \mathcal{X} \times \mathcal{Y}. \tag{2.1}$$

Remark 1 *In case Φ is μ_x-strongly convex with respect to x and μ_y-strongly concave with respect to y, one can redefine the problem so that Assumption 1 holds. Indeed, consider $\tilde{\Phi}(x,y) \triangleq \Phi(x,y) - \frac{\mu_x}{2}\|x\|_2 + \frac{\mu_y}{2}\|y\|_2$, $\tilde{f}(x) \triangleq f(x) + \frac{\mu_x}{2}\|x\|_2^2$ and $\tilde{g} \triangleq g(y) + \frac{\mu_y}{2}\|y\|_2^2$ and reformulate the problem in (1.1) as $\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \tilde{f}(x) + \tilde{\Phi}(x,y) - \tilde{g}(y)$. Note that $\tilde{\Phi}, \tilde{f}$ and $\tilde{g}$ satisfy Assumption 1, i.e., $\tilde{\Phi}(x,y)$ is convex/concave, $\tilde{f}$ is μ_x-strongly convex and $\tilde{g}$ is μ_y-strong convex, and these two formulations are equivalent. Furthermore, the evaluation of $\mathrm{prox}_{\tilde{f}}$ for $\tilde{f} = f + \frac{\mu}{2}\|\cdot\|_2^2$ is computationally similar to evaluating prox_f, see (Zhang et al., 2024, Remark 1) for further details.*

Following the literature on stochastic saddle-point algorithms (Nemirovski et al., 2009; Juditsky et al., 2011; Chen et al., 2017), we assume that only (noisy) stochastic estimates $\tilde{\nabla}_{\mathrm{y}} \Phi(x_n, y_n, \omega_n^y), \tilde{\nabla}_{\mathrm{x}} \Phi(x_n, y_{n+1}, \omega_n^x)$ of the partial gradients $\nabla_{\mathrm{y}} \Phi(x_n, y_n), \nabla_{\mathrm{x}} \Phi(x_n, y_{n+1})$ are available, where ω_n^x, ω_n^y are random variables that are being revealed sequentially. Specifically, we let $(\omega_n^x)_{n\geqslant 0}$, $(\omega_n^y)_{n\geqslant 0}$ be two sequences of random variables revealed in the following order in time which is the natural order for the SAPD updates:

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

and we let $(\mathcal{F}_n^y)_{n\geqslant 0}$ and $(\mathcal{F}_n^x)_{n\geqslant 0}$ denote the associated filtrations, i.e., $\mathcal{F}_0^y = \sigma(\omega_0^y)$ is the sigma algebra generated by the random variable ω_0^y, $\mathcal{F}_0^x = \sigma(\omega_0^y, \omega_0^x)$ and

$$\mathcal{F}_n^y = \sigma(\mathcal{F}_{n-1}^x, \sigma(\omega_n^y)), \quad \mathcal{F}_n^x = \sigma(\mathcal{F}_n^y, \sigma(\omega_n^x)), \quad \forall\ k \geqslant 1.$$

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

$$\Delta_n^y \triangleq \tilde{\nabla}_{\mathrm{y}} \Phi(x_n, y_n, \omega_n^y) - \nabla_{\mathrm{y}} \Phi(x_n, y_n), \quad \Delta_n^x \triangleq \tilde{\nabla}_{\mathrm{x}} \Phi(x_n, y_{n+1}, \omega_n^x) - \nabla_{\mathrm{x}} \Phi(x_n, y_{n+1}).$$

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 (Harold et al., 1997) and frequently arises in the context of stochastic gradient methods that estimate the gradients from randomly sampled subsets of data (Bottou et al., 2018).

Assumption 2 *There exists scalars $\nu_x, \nu_y > 0$ such that for all $n \geqslant 0$,*

$$\mathbb{E}\left[\Delta_n^y \middle| \mathcal{F}_{n-1}^x\right] = 0, \quad \mathbb{E}\left[\Delta_n^x \middle| \mathcal{F}_n^y\right] = 0, \quad \mathbb{E}\left[\|\Delta_n^y\|^2 \middle| \mathcal{F}_{n-1}^x\right] \leqslant {\nu_y}^2, \quad \mathbb{E}\left[\|\Delta_n^x\|^2 \middle| \mathcal{F}_n^y\right] \leqslant {\nu_x}^2.$$

Under Assumptions 1 and 2, given a set of parameters (τ, σ, θ), `SAPD` iterates (x_n, y_n) were shown to converge to a neighborhood of the solution linearly in expectation where the size of the neighborhood gets smaller when the gradient noise levels $\nu_x, \nu_y \geqslant 0$ gets smaller (Zhang et al., 2024); in particular, in the absence of noise (when $\nu_x = \nu_y = 0$), the iterates (x_n, y_n) converges to $(\mathrm{x}^\star, \mathrm{y}^\star)$ at a linear rate $\rho \in (0,1)$ provided that there exists some $\alpha \in [0, \sigma^{-1})$ for which the following inequality holds:

$$\begin{pmatrix} \frac{1}{\tau} + \mu_x - \frac{1}{\rho\tau} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sigma} + \mu_y - \frac{1}{\rho\sigma} & \left(\frac{\theta}{\rho} - 1\right) L_{yx} & \left(\frac{\theta}{\rho} - 1\right) L_{yy} & 0 \\ 0 & \left(\frac{\theta}{\rho} - 1\right) L_{yx} & \frac{1}{\tau} - L_{xx} & 0 & -\frac{\theta}{\rho} L_{yx} \\ 0 & \left(\frac{\theta}{\rho} - 1\right) L_{yy} & 0 & \frac{1}{\sigma} - \alpha & -\frac{\theta}{\rho} L_{yy} \\ 0 & 0 & -\frac{\theta}{\rho} L_{yx} & -\frac{\theta}{\rho} L_{yy} & \frac{\alpha}{\rho} \end{pmatrix} \succeq 0. \tag{2.2}$$

An important class of solutions to the matrix inequality in (2.2) takes the following form: given an arbitrary $c \in (0, 1]$, choose

$$\tau = \frac{1-\theta}{\theta\mu_x}, \quad \sigma = \frac{1-\theta}{\theta\mu_y}, \quad \theta \geqslant \bar{\theta}_c, \tag{2.3}$$

for some $\bar{\theta}_c \in (0,1)$ explicitly given in (Zhang et al., 2024, Corollary 1) – (τ, σ, θ) satisfying (2.3) solves (2.2) with $\rho = \theta$ and $\alpha = \frac{c}{\sigma} - \sqrt{\theta} L_{yy}$ with $c \in (0, 1]$. SAPD generalizes the primal-dual algorithm CP proposed in (Chambolle and Pock, 2011) – CP algorithm can solve SP problems with a *bilinear* coupling function Φ when a deterministic first-order oracle for Φ exists; indeed, for bilinear coupling functions with deterministic first-order oracles, SAPD reduces to the CP algorithm. It is shown in (Chambolle and Pock, 2016) that for a particular value[3] of θ, the choice of primal and stepsizes (τ, σ) according to (2.3) achieves the accelerated rate. For SAPD, Zhang et al. (2024) study the squared distance of iterates to the saddle point in expectation and extends the same acceleration result to the case when Φ is not bilinear and when one has only access to a stochastic first-order oracle rather than a deterministic one.

As we focus on 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 a weighted squared distance to the solution, i.e.,

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

[3] see (Chambolle and Pock, 2016, Eq.(48)).

for some $\alpha \in [0, \sigma^{-1})$. This weighted metric turns out to be more convenient for the convergence analysis of SAPD, but it is clearly equivalent to the unweighted squared distance $E_n = \|x_n - \mathrm{x}^\star\|^2 + \|y_n - \mathrm{y}^\star\|^2$ up to a constant that depends on the choice of (τ, σ, α). For the sake of completeness of the paper, we first recall the convergence of `SAPD` in expected weighted squared distance, established in (Zhang et al., 2024).

Theorem 2 ((Zhang et al., 2024), Theorem 1) *Suppose Assumptions 1 and 2 hold and let $z_n = (x_n, y_n)_{n \geqslant 1}$ be the iterates generated by* `SAPD`*, starting from an arbitrary tuple $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$. For all $n \in \mathbb{N}$, $\tau, \sigma > 0$, and $\theta \geqslant 0$ satisfying* (2.2) *for some $\rho \in (0,1)$ and $\alpha \in [0, \sigma^{-1})$, it holds that*

$$\mathbb{E}\left[\mathcal{W}_n\right] \leqslant \rho^n \, \mathcal{W}_{\tau,\sigma} + \frac{\rho}{1-\rho}\left(\frac{\tau}{1+\tau\mu_x}\Xi^x_{\tau,\sigma,\theta}\nu_x^2 + \frac{\sigma}{1+\sigma\mu_y}\Xi^y_{\tau,\sigma,\theta}\nu_y^2\right), \tag{2.5}$$

where $\mathcal{W}_{\tau,\sigma} \triangleq \frac{1}{2\tau}\|x_0 - \mathrm{x}^\star\|^2 + \frac{1}{2\sigma}\|y_n - \mathrm{y}^\star\|^2$ denotes the initial bias, and $\Xi^x_{\tau,\sigma,\theta} \triangleq 1 + \frac{\sigma\theta(1+\theta)L_{yx}}{2(1+\sigma\mu_y)}$ $\Xi^y_{\tau,\sigma,\theta} \triangleq \frac{\tau\theta(1+\theta)L_{yx}}{2(1+\tau\mu_x)} + \left(1 + 2\theta + \frac{\theta+\sigma\theta(1+\theta)L_{yy}}{1+\sigma\mu_y} + \frac{\tau\sigma\theta(1+\theta)L_{yx}L_{xy}}{(1+\tau\mu_x)(1+\sigma\mu_y)}\right)(1 + 2\theta)$ are noise related constants that depend on the problem and algorithm parameters.

As stated above, the convergence of SAPD in expected squared distance presents the classical bias-variance trade-off, which can be controlled through adjusting the `SAPD` parameter choice. The bias term $\rho^n \mathcal{W}_{\tau,\sigma}$, captures the rate at which the error due to initialization (bias) decays, ignoring the noise. It is shown in (Zhang et al., 2024) that for certain choice of parameters, convergence of initialization bias to 0 occurs at an accelerated rate $\rho = 1 - \Theta(\frac{1}{\kappa})$ instead of the non-accelerated rate $\rho = 1 - \Theta(\frac{1}{\kappa^2})$ of methods such as (Jacobi-style) SGDA. The variance term constitutes the (remaining part) second term at the right-hand side of (2.5) and is due to noise accumulation that scales with the stepsize and the noise variance. For a particular choice of SAPD parameters, it is shown that in expectation SAPD exhibits the optimal complexity of $\tilde{\mathcal{O}}(1/\varepsilon)$ up to a logarithmic factor, and achieves an accelerated decay rate for the bias term; however, in a number of risk-sensitive situations, convergence in expectation can prove to be insufficient. In this paper, we further investigate the properties of `SAPD` for several measures of risks, that we detail in Section 2.3.

2.2 Assumptions on the gradient noise

Although according to Theorem 2, (2.2) describes a general set of parameters for which `SAPD` will admit guarantees in terms of the *expected* weighted squared distance to the solution, risk-sensitive guarantees for SAPD, including high-probability bounds are not known. In the forthcoming sections, we study `SAPD` for parameters satisfying (2.2), and we obtain convergence guarantees in high probability, in CVaR, in EVaR, and also in the χ^2-divergence-based risk measures, which are properly defined in Section 2.3. In other words, our focus here is to obtain high probability guarantees as well as bounds on the risk associated with $E_n^{1/2} = \|z_n - \mathrm{z}^\star\|$. To this end, we will make a "light-tail" assumption on the magnitude of the gradient noise, adopting a subGaussian structure. Before giving our assumption on the gradient noise precisely, we start with introducing the family of norm-subGaussian random variables, and recall their basic properties.

Definition 3 *A random vector $X\colon \Omega \to \mathbb{R}^{\mathrm{d}}$ is norm-subGaussian with proxy $\delta > 0$, denoted by $X \in nSG\left(\delta\right)$, if we have $\mathbb{P}\left[\|X - \mathbb{E}[X]\| \; \geqslant t\right] \leqslant 2e^{\frac{-t^2}{2\delta^2}}, \quad \forall t \in \mathbb{R}$.*

Random vectors with norm-subGaussian distribution were introduced in (Jin et al., 2019), and encompass a large class of random vectors including subGaussian random vectors. First, note that given an arbitrary $\alpha > 0$ and a random variable $X\colon \Omega \to \mathbb{R}^{\mathrm{d}}$ such that $X \in nSG\left(\delta\right)$ for some $\delta > 0$, we immediately have the following implication:

$$X \in nSG\left(\delta\right) \implies \alpha X \in nSG\left(\alpha\delta\right). \tag{2.6}$$

Moreover, $X : \Omega \to \mathbb{R}^d$ is norm-subGaussian when X is subGaussian, or it is bounded, i.e., $\exists B > 0$ such that $\|X\| \leqslant B$ with probability 1. As discussed in (Jin et al., 2019, Lemma 3), the squared norm of a norm-subGaussian vector admits a sub-Exponential distribution, which is defined next.

Definition 4 *A random variable $\mathcal{U}\colon \Omega \to \mathbb{R}$ is subExponential with proxy $K > 0$ if it satisfies*

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

In particular, if we take $U = \|X\|^2$ for $X \in nSG\left(\delta\right)$ with $\mathbb{E}[X] = 0$, then the following result shows that U is subExponential with proxy $K = 8\delta^2$.

Lemma 5 *Let $X \in nSG\left(\delta\right)$ be such that $\mathbb{E}[X] = 0$. Then, for any $\lambda \in \left[0, \frac{1}{4\delta^2}\right]$,*

$$\mathbb{E}\left[e^{\lambda\|X\|^2}\right] \leqslant 2e^{4\lambda\delta^2} - 1 \leqslant e^{8\lambda\delta^2}. \tag{2.7}$$

Lemma 6 *Let $X \in nSG\left(\delta\right)$ such that $\mathbb{E}[X] = 0$. Then, for any $u \in \mathbb{R}^{\mathrm{d}}$ and $\lambda \geqslant 0$, it holds that $\mathbb{E}\left[e^{\lambda\langle u, X\rangle}\right] \leqslant e^{8\lambda^2\|u\|^2\delta^2}$.*

For completeness, the proofs of these two elementary results are provided in Section B.1 of the appendix. Next, we will introduce an assumption which says that gradient noise terms Δ_n^x and Δ_n^y are light-tailed admitting a norm-subGaussian structure when conditioned on the natural filtration of the past iterates.

Assumption 3 *For any $n \geqslant 0$ the random vectors Δ_n^x and Δ_n^y are conditionally unbiased and norm-subGaussian with respective proxy parameters $\delta_x, \delta_y > 0$. More precisely, for all $t \geqslant 0$, we have $\mathbb{E}\left[\Delta_n^y \middle| \mathcal{F}_{n-1}^x\right] = 0, \mathbb{E}\left[\Delta_n^x \middle| \mathcal{F}_n^y\right] = 0$, and*

$$\mathbb{P}[\|\Delta_n^y\| \geqslant t|\mathcal{F}_{n-1}^x] \leqslant 2e^{\frac{-t^2}{2\delta_y^2}}, \quad \mathbb{P}[\|\Delta_n^x\| \geqslant t|\mathcal{F}_n^y] \leqslant 2e^{\frac{-t^2}{2\delta_x^2}}.$$

We note that such subGaussian noise assumptions are common in the study of stochastic optimization algorithms (Rakhlin et al., 2012; Ghadimi and Lan, 2012; Harvey et al., 2019). In machine learning applications, where stochastic gradients are often estimated on sampled batches, noisy estimates typically behave Gaussian for moderately high sample sizes, as a consequence of the central limit theorem (Panigrahi et al., 2019). Furthermore, there are applications in data privacy where i.i.d. subGaussian noise is added to the gradients for privacy reasons (Levy et al., 2021; Varshney et al., 2022). In such settings, we expect Assumption 3 to hold naturally. In the rest of the paper, together with Assumption 1, we will assume that Assumption 3 holds in lieu of Assumption 2.

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

Table 2: Three examples of φ-divergence based risk measures studied in this paper.

2.3 VaR, CVaR, EVaR and χ^2-divergences

For any given $n \geqslant 0$, to quantify the risk associated with $\|z_n - \mathrm{z}^\star\|$, i.e., the distance to the unique saddle point, we will resort to ϕ-divergence-based risk measures borrowed from the risk measure theory (Ben-Tal and Teboulle, 2007), including CVaR, EVaR and χ^2-divergence. The first risk measure of interest is the quantile function, also known as value at risk, defined for any random variable $\mathcal{U}\colon \Omega \to \mathbb{R}$ as follows:

$$Q_p(\mathcal{U}) \triangleq \inf\{t \in \mathbb{R} : \mathbb{P}[\mathcal{U} \leqslant t] \geqslant p\}. \tag{2.8}$$

Quantile upper bounds correspond to high-probability results, which have been already fairly studied to assess the robustness of stochastic algorithms (Ghadimi and Lan, 2012; Rakhlin et al., 2012; Harvey et al., 2019). One key contribution of this paper is the derivation of an upper bound on the quantiles of the weighted distance metric $\mathcal{W}_n$, defined in (2.4), such that this upper bound exhibits a tight bias-variance trade-off –see Section 4.2.

Furthermore, we investigate the robustness of `SAPD` with respect to three convex risk measures based on φ-divergences (Ben-Tal and Teboulle, 2007). 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 as discussed next.

Definition 7 *For any $r \geqslant 0$, the φ-divergence based risk measure at level r is defined as*

$$\mathcal{R}_{\varphi,r}(\mathcal{U}) \triangleq \sup_{\mathbb{Q}\ll\mathbb{P},\ D_\varphi(\mathbb{Q}\|\mathbb{P})\leqslant r} \mathbb{E}_{\mathbb{Q}}\left[\mathcal{U}\right], \tag{2.9}$$

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

We refer the reader to (Ben-Tal and Teboulle, 2007; Shapiro, 2017) for more on φ-divergence based risk measures. In this paper, we investigate the performances of SAPD under three φ-divergence based risk measures, summarized in Table 2.

First, given $p \in [0,1)$, we consider the conditional value at risk at level p, i.e., CVaR_p, defined as

$$\mathrm{CVaR}_p(\mathcal{U}) \triangleq \frac{1}{1-p}\int_{p'=p}^{1} Q_{p'}(\mathcal{U})\,\mathrm{d}p'. \tag{2.10}$$

The CVaR measure admits the variational representation (2.9) with $\varphi : t \mapsto \mathcal{I}_{[0,(1-p)^{-1}]}(t)$ for any $r > 0$. 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 $\mathcal{U} \geqslant Q_p(\mathcal{U})$, the CVaR presents the advantage of averaging the whole tail of the distribution; therefore, it can quantify the risk associated with tail events in a robust fashion.

The second convex risk measure we investigate is the Entropic Value at Risk (Ahmadi-Javid, 2012), denoted by EVaR, and is defined as $\mathrm{EVaR}_p(\mathcal{U}) \triangleq \inf_{\eta>0}\left\{-\eta^{-1}\log(1-p) + \eta^{-1}\log(\mathbb{E}(\exp(\eta\mathcal{U})))\right\}$. The EVaR admits the variational representation (2.9) 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. (Shapiro, 2017). EVaR exhibits a higher tail-sensitivity than CVaR, in the sense that $\mathrm{CVaR}_p(\mathcal{U}) \leqslant \mathrm{EVaR}_p(\mathcal{U})$ 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 (2.9) with $\varphi : t \mapsto \frac{1}{2}(t-1)^2$.

3. Main Results

In this section, we present the main results of this paper, which consist of convergence analyses of `SAPD` in high-probability and provide guarantees in terms of the three convex risk measures presented in Table 2. Later in Section 4, we derive analytical expressions related to convergence behavior of `SAPD` applied on quadratic SP problems, and in Section 4.2 we discuss some tight characteristics of our main results provided in this section. Finally, in Section 5, we provide the proofs of our main result stated in Theorem 8.

Theorem 8 *Suppose Assumption 1 and Assumption 3 hold. Given $\tau, \sigma > 0$, and $\theta \geqslant 0$ satisfying* (2.2) *for some $\rho \in (0,1)$ and $\alpha \in [0, \sigma^{-1})$, let $(x_n, y_n)_{n\geqslant 1}$ denote the corresponding `SAPD` iterates, initialized at an arbitrary tuple $(x_0, y_0) \in \mathbf{dom}\, f \times \mathbf{dom}\, g$. For all $n \in \mathbb{N}$, $p \in [0,1)$, it holds that*

$$\mathbb{P}\Big[\mathcal{W}_{n+1} + \mathcal{W}_n \leqslant q_{p,n+1}\Big] \geqslant p, \quad \textit{where} \tag{3.1}$$

$$q_{p,n+1} \triangleq \left(\frac{1+\rho}{2}\right)^n \left(\mathcal{C}_{\tau,\sigma,\theta}\, \mathcal{W}_{\tau,\sigma} + \Xi^{(1)}_{\tau,\sigma,\theta}\right) + \Xi^{(2)}_{\tau,\sigma,\theta} + \Xi^{(3)}_{\tau,\sigma,\theta} \log\left(\frac{1}{1-p}\right), \tag{3.2}$$

where $\mathcal{W}_n = \frac{1}{2\tau}\|x_n - \mathrm{x}^\star\|^2 + \frac{1-\alpha\sigma}{2\sigma}\|y_n - \mathrm{y}^\star\|^2$, $\mathcal{W}_{\tau,\sigma} \triangleq \frac{1}{2\tau}\|x_0 - \mathrm{x}^\star\|^2 + \frac{1}{2\sigma}\|y_0 - \mathrm{y}^\star\|^2$, $\mathcal{C}_{\tau,\sigma,\theta}$ *and* $\Xi^{(i)}_{\tau,\sigma,\theta} \triangleq \Xi^{(i,x)}_{\tau,\sigma,\theta}\delta_x^2 + \Xi^{(i,y)}_{\tau,\sigma,\theta}\delta_y^2$ *for $i = 1,2,3$ depend only on the algorithm parameters (τ, σ, θ) and the problem parameters $(\mu_x, \mu_y, \mathrm{L_{xx}}, \mathrm{L_{yy}}, \mathrm{L_{xy}}, \mathrm{L_{yx}})$. Furthermore, all these constants can be made explicit*[4] *and in particular, under the CP parameterization in* (2.3)*, they satisfy $\mathcal{C}_{\tau,\sigma,\theta} = \Theta(1)$, $\Xi^{(i,x)}_{\tau,\sigma,\theta} = \Theta(1)$, and $\Xi^{(i,y)}_{\tau,\sigma,\theta} = \Theta(1)$ for all $i = 1,2,3$ as $\theta \to 1$, which implies that*

$$\limsup_{n\to\infty} Q_p(\|z_n - \mathrm{z}^\star\|^2) = \mathcal{O}\Big((1-\theta)\delta^2\Big(1 + \log\Big(\frac{1}{1-p}\Big)\Big)\Big).$$

Proof The proof is provided in Section 5.2.1. ■

Remark 9 *Under the premise of Theorem 8,* (3.1) *implies that for all $p \in [0,1)$ and $n \geqslant 0$,*

$$Q_p\left(\mathcal{W}_{n+1}^{1/2}\right) \leqslant Q_p\left((\mathcal{W}_{n+1} + \mathcal{W}_n)^{1/2}\right) \leqslant q_{p,n+1}^{1/2}. \tag{3.3}$$

[4] These constants are explicitly given within the proof, provided in Section 5.

Remark 10 *For any given $\rho \in (0,1)$, to check if there exists* `SAPD` *parameters τ, σ, θ such that the bias component of $\mathbb{E}[\mathcal{W}_{n+1} + \mathcal{W}_n]$ decreases to 0 linearly with a rate coefficient bounded above by $\rho \in (0,1)$, one needs to solve a 5-dimensional SDP, i.e., after fixing ρ, checking the feasibility of (2.2) reduces to an SDP problem, see (Zhang et al., 2024) for details.* Below in Corollary 12, we provide a particular solution to (2.2), in the form of (2.3), for which the choice of ρ leads to an accelerated behavior with a complexity of $\mathcal{O}(\kappa \log(\varepsilon^{-1}) + \mu^{-1}\delta^2(1 + \log((1-p)^{-1}))\varepsilon^{-1}\log(\varepsilon^{-1}))$ where $\delta^2 = \max(\delta_x^2, \delta_y^2)$. Thus, the bias term in Theorem 8 decays at an *accelerated* rate, which differs from the standard decay of non-accelerated Jacobi-style algorithms where the initialization (bias) error scales proportionally to κ^2 (Fallah et al., 2020).

Remark 11 *Our bound for the p-th quantile, $q_{p,n+1}$, is tight, i.e., under the parameter choice in (2.3), the dependency of $q_{p,n+1}$ to θ and p cannot be improved when n is large enough. See Theorem 19 for further details.*

Next, we provide the oracle complexity of `SAPD` in high probability, which can be derived as a corollary to our main Theorem 8.

Corollary 12 *For $p \in [0,1)$ and $\varepsilon > 0$, set $\tau, \sigma, \theta, \rho$ as*

$$\tau = \frac{1-\theta}{\theta\mu_x}, \quad \sigma = \frac{1-\theta}{\theta\mu_y}, \quad \rho = \theta = \max\left(1/2, \bar{\theta}_1, \bar{\theta}_2, \bar{\bar{\theta}}_x, \bar{\bar{\theta}}_y\right), \quad \textit{where} \tag{3.4}$$

$$\bar{\theta}_1 \triangleq 1 - \frac{\beta\,(\mathrm{L_{xx}} + \mu_x)\,\mu_y}{4\,\mathrm{L_{yx}}^2}\left(\sqrt{1 + \frac{8\mu_x\,\mathrm{L_{yx}}^2}{\beta\mu_y\,(\mathrm{L_{xx}} + \mu_x)^2}} - 1\right), \quad \bar{\theta}_2 \triangleq \begin{cases} 1 - \frac{(1-\beta)^2}{32}\frac{\mu_y^2}{\mathrm{L_{yy}}^2}\left(\sqrt{1 + \frac{64\,\mathrm{L_{yy}}^2}{(1-\beta)^2\mu_y^2}} - 1\right) & \textit{if } \mathrm{L_{yy}} > 0 \\ 0 & \textit{if } \mathrm{L_{yy}} = 0 \end{cases},$$

$$\begin{aligned} \bar{\bar{\theta}}_x &\triangleq 1 - \frac{\mu_x\varepsilon}{\left(c_x^{(1)} + c_x^{(2)}\frac{\mathrm{L_{xy}}}{\mathrm{L_{yx}}} + c_x^{(3)}\frac{\mathrm{L_{xy}}^2}{\mathrm{L_{yx}}^2}\right)\delta_x^2\left(1 + \log(1/(1-p))\right)}, \\ \bar{\bar{\theta}}_y &\triangleq 1 - \frac{\mu_y\varepsilon}{\left(c_y^{(1)} + c_y^{(2)}\frac{\mathrm{L_{xy}}}{\mathrm{L_{yx}}} + c_y^{(3)}\frac{\mathrm{L_{xy}}^2}{\mathrm{L_{yx}}^2}\right)\delta_y^2\left(1 + \log(1/(1-p))\right)}, \end{aligned}$$

with $\beta = \min\{1/2, \mu_x/\mu_y, \mu_y/\mu_x\}$, and for universal positive constants $c_x^{(i)}, c_y^{(i)}$ for $i = 1,2,3$ that are large enough[5]*. Then, (2.2) is satisfied for $\alpha = \frac{1}{2\sigma} - \sqrt{\theta}L_{yy}$ and* `SAPD` *guarantees that $\mathbb{P}\left[\mu_x\|x_n - \mathrm{x}^\star\|^2 + \mu_y\|y_n - \mathrm{y}^\star\|^2 \leqslant \varepsilon\right] \geqslant p$ for n satisfying (1.3).*

Proof The result follows directly from Theorem 8 after plugging in our choice of parameters based on tedious but straightforward computations. For the sake of completeness, we provide the details of these computations in Appendix G of the online-only supplementary material - see the extended version of the paper Laguel et al. (2023). ■

Remark 13 *By Assumption 1, the partial gradients $\nabla_x\Phi$ and $\nabla_y\Phi$ are Lipschitz continuous; therefore, they are almost everywhere differentiable by Rademacher's Theorem. If we assume slightly more, i.e., if $\nabla_x\Phi$ and $\nabla_y\Phi$ are continuously differentiable, then the partial derivatives commute and we have $\nabla_y\nabla_x\Phi(x,y) = \nabla_x\nabla_y\Phi(x,y)$, as a consequence of Schwarz's Theorem. In this case, we can take $\mathrm{L_{xy}} = \mathrm{L_{yx}}$ in Assumption 1 and in Corollary 12.*

[5] For simplicity of the presentation, we do not provide these universal constants explicitly here; that said, the constants can be made explicit in a straightforward manner following the step-by-step computations in our proof provided in the extended version of the paper Laguel et al. (2023).

Remark 14 *Our (τ, σ, θ) choice in Corollary 12 depend on the convexity moduli $\mu_x, \mu_y > 0$ and Lipschitz constants $L_{xx}, L_{xy}, L_{yx}, L_{yy}$, the noise level δ_x^2 and δ_y^2 and the target probability $p \in (0,1)$. Our θ choice here, which depends on the given probability $p \in (0,1)$, is different from the existing literature on* `SAPD` *given in (Zhang et al., 2024, Theorem 2) for analysis of* `SAPD` *iterate sequence in expectation. Since strong convexity and Lipschitz constants corresponding to the stochastic minimax problems of interest may not be available, designing adaptive (stepsize) methods with high-probability guarantees and that are oblivious to these problem parameters is an interesting open research problem; but, this is beyond the scope of this work. That being said, for some problems these constants may be known or can be estimated. For example, strong convexity constants are known for ℓ_2-regularized problems of the form: $\min_x \max_y \Phi(x,y) + \frac{\lambda_x}{2}\|x\|^2 - \frac{\lambda_y}{2}\|y\|^2$ where Φ is convex in x and concave in y, and $\lambda_x, \lambda_y > 0$ are regularization parameters; in this case we can simply take $\mu_x = \lambda_x$ and $\mu_y = \lambda_y$, this setting may arise for instance in distributionally robust learning Zhang et al. (2024). If the strong convexity constants are not known, we could also rely on estimation techniques such as Barré et al. (2020); Malitsky and Mishchenko (2019). When Φ is twice continuously differentiable, Lipschitz constants can also be estimated by approximately maximizing the norm of $\nabla^2\Phi$; but, this can be computationally expensive in high dimensions. Alternatively, for some structured minimax problems, Lipschitz constants can be estimated explicitly, e.g., distributionally robust logistic regression Zhang et al. (2024), and distributionally robust linear regression, where the dual domain is the probability simplex, in which case the norm of $\nabla^2\Phi$ can be characterized explicitly and one can obtain precise estimates for the Lipschitz constants.*

Using Theorem 8 and building on the representation of the CVaR in terms of the quantiles, we can deduce a bound on $\mathrm{CVaR}_p(\mathcal{W}_n^{\frac{1}{2}})$ as shown in Theorem 15, where we also provide bounds on the entropic value at risk and on the χ^2-based risk measure, as defined in Table 2.

Theorem 15 (Bounds on Risk Measures) *Under the premise of Theorem 8,*

$$\mathrm{CVaR}_p\left(\mathcal{W}_{n+1}^{\frac{1}{2}}\right) \leqslant \sqrt{\left(\frac{1+\rho}{2}\right)^{n/2}\left(\mathcal{C}_{\tau,\sigma,\theta}\mathcal{W}_{\tau,\sigma} + \Xi^{(1)}_{\tau,\sigma,\theta}\right) + \Xi^{(2)}_{\tau,\sigma,\theta}} + \sqrt{\Xi^{(3)}_{\tau,\sigma,\theta}\Big(1 + \log\Big(\frac{1}{1-p}\Big)\Big)}, \tag{3.5}$$

$$\mathrm{EVaR}_p(\mathcal{W}_{n+1}^{\frac{1}{2}}) \leqslant \sqrt{\left(\frac{1+\rho}{2}\right)^{n/2}\left(\mathcal{C}_{\tau,\sigma,\theta}\mathcal{W}_{\tau,\sigma} + \Xi^{(1)}_{\tau,\sigma,\theta}\right) + \Xi^{(2)}_{\tau,\sigma,\theta}} + \sqrt{\Xi^{(3)}_{\tau,\sigma,\theta}}\left(\log\Big(\frac{1}{1-p}\Big)^{1/2} + \sqrt{\pi}\right), \tag{3.6}$$

hold for all $n \in \mathbb{N}$ and $p \in [0,1)$, where $\mathcal{W}_n$, $\Xi^{(1)}_{\tau,\sigma,\theta}$, $\Xi^{(2)}_{\tau,\sigma,\theta}$ and $\bar{\delta}$ are as defined in Theorem 8. Furthermore, for all $n \in \mathbb{N}$ and $r > 0$, the right-hand side of (3.6) with $p = 1 - \frac{1}{1+r}$ is an upper bound on $\mathcal{R}_{\chi^2,r}(\mathcal{W}_{n+1}^{1/2})$.

Proof The proof is provided in Section 5.2.2. ■

In the next section, we discuss a family of quadratic SP problems for which we can compute the asymptotic covariance matrix of the iterates in expectation explicitly, assuming additive i.i.d. Gaussian noise on the partial gradients. This will allow us to gain insights about the effect of parameter choices and argue about the tightness of our analysis.

Remark 16 *There are notable differences among the risk measures considered in this work. First, their domain of definition differs significantly. Indeed, the quantile function $Q_p(\cdot)$*

is defined on any random variable X satisfying $\mathbb{P}(|X| < \infty) = 1$, while CVaR *and* EVaR *require $|X|$ to be integrable, and $|X|$ to be sub-exponential, respectively. Second, while high probability bounds do not account for the price of failure associated with the tail event when we are above the p-th quantile of the error $\mathcal{W}_n$, convex risk measures such as CVaR integrate these high-values, and are as such more sensitive to the worst-case scenarios; this can also be observed from our example given in Fig. 1. While CVaR, EVaR and χ^2-divergences belong to the general class of ϕ-divergence-based risk measures, each of these risk measures differ in terms of how they consider or average the tail events (see Section 2.3). Note that these risk measures are all coherent (possessing favorable properties as a risk measure such as monotonicity, sub-additivity, homogeneity, and translational invariance), and are of interest in the study and design of stochastic optimization algorithms (Can and Gürbüzbalaban, 2022; Ahmadi-Javid, 2012; Chouzenoux et al., 2019). In our context, obtaining risk-averse guarantees in these risk measures are relevant because they allow us to obtain finer characterizations of the tail events associated to the optimization error $\mathcal{W}_n$, quantifying deviations from the average performance (expected error $\mathbb{E}[\mathcal{W}_n]$) in different ways.*

Our results in Theorem 8 and Corollary 15 prove that under the norm-subGaussian assumption, the upper bounds for these risk measures, i.e., $Q_p(\cdot)$, $\text{CVaR}_p(\cdot)$ and $\text{EVaR}_p(\cdot)$, behave similarly in terms of their dependence on the problem parameters. This is due to the tight control we manage to get over the moment generating function corresponding to the error $\mathcal{W}_n$ at stage n in our analysis given in Section 5.2. However, we suspect that beyond the subGaussian regime, in the heavy-tail scenario, the error bounds for these risk measures may exhibit different decay properties. Indeed, existence of a finite EVaR is more restrictive than CVaR being finite as it requires for the moment generating function of the underlying random variable to be well defined, whereas CVaR requires less restrictive conditions on its moments (see also Remark 16).

4. Analytical solution for quadratics

In this section, we study the behaviour of `SAPD` on quadratic problems subject to additive isotropic Gaussian noise. More specifically, we consider

$$\min_{x\in\mathbb{R}^{\mathrm{d}}}\max_{y\in\mathbb{R}^{\mathrm{d}}} \frac{\mu_x}{2}\|x\|^2 + \langle Kx, y\rangle - \frac{\mu_y}{2}\|y\|^2, \tag{4.1}$$

where $K \in \mathbb{R}^{d\times d}$ is a symmetric matrix, and $\mu_x, \mu_y > 0$ are two regularization parameters. The unique saddle point of (4.1) is the origin $\mathrm{z}^\star = (\mathrm{x}^\star, \mathrm{y}^\star) = (\mathbf{0}, \mathbf{0})$. At each iteration $n \geqslant 0$, suppose we have access to noisy estimates $\tilde{\nabla}_{\mathrm{y}}\,\Phi(x_n, y_n) = Kx_n + \omega_n^y$ and $\tilde{\nabla}_{\mathrm{x}}\,\Phi(x_n, y_{n+1}) = K^\top y_{n+1} + \omega_n^x$ of the partial gradients of $\Phi : (x, y) \mapsto \langle Kx, y\rangle$, where the $(\omega_n^x)_{n\geqslant 0}$ and $(\omega_n^y)_{n\geqslant 0}$ denote i.i.d. centered Gaussian vectors satisfying $\mathbb{E}[\omega_n^x {\omega_n^x}^\top] = \mathbb{E}[\omega_n^y {\omega_n^y}^\top] = d^{-1}\delta^2 I$ for some $\delta \geqslant 0$. In this special case, we have the gradient noise vectors $\Delta_n^x = \omega_n^x$, $\Delta_n^y = \omega_n^y$. Our main motivation to study this toy problem is to gain some insights into the sample paths `SAPD` generates, and use these insights while studying the tightness of high-probability bounds provided in Section 3.

This problem was first studied in (Zhang et al., 2024) where it was shown that, under certain conditions on τ, σ, θ, the sequence of iterates $(\tilde{z}_n)_{n\geqslant 0}$ generated by `SAPD`, where

$\tilde{z}_n = (x_{n-1}, y_n)$, converges in distribution to a zero-mean multi-variate Gaussian random vector whose covariance matrix $\widetilde{\Sigma}^{\infty}$ satisfies a certain Lyapunov equation of dimension $2d \times 2d$. The authors of (Zhang et al., 2024) manage to split this equation into d many 2×2 Lyapunov equations:

$$\widetilde{\Sigma}^{\infty,\lambda} = A^{\lambda}\ \widetilde{\Sigma}^{\infty,\lambda}\ (A^{\lambda})^{\top} + R^{\lambda} \in \mathbb{R}^{2\times 2}, \quad \forall \lambda \in \mathrm{Sp}(K), \tag{4.2}$$

i.e., for each eigenvalue λ of K, there is a 2×2 Lyapunov equation to be solved, where $A^{\lambda}, R^{\lambda} \in \mathbb{R}^{2\times 2}$ depend only on τ, σ, θ, δ^2 and $\lambda \in \mathrm{Sp}(K)$ –for completeness, we provide these steps in detail in Section D.1 of the appendix.

Given an arbitrary symmetric matrix K, in (Zhang et al., 2024), the small-dimensional Lyapunov equation in (4.2) is solved numerically. On the other hand, to establish the tightness of our high-probability bounds for the class of SCSC problems with a non-bilinear Φ subject to noisy gradients with subGaussian tails (see Section 3), we need to analytically solve (4.2). However, analytically solving (4.2) for general parameters satisfying the matrix inequality (2.2) is a challenging problem that standard symbolic computation tools were not in a position to properly address. That said, as we shall discuss next, we can provide analytical solutions for (4.2) under the Chambolle-Pock (CP) parameterization in (2.3), where primal and dual stepsizes are parameterized in θ, i.e., the momentum parameter. We recall that in this case, the momentum parameter value coincides with the convergence rate bound for `SAPD`, i.e., $\theta = \rho$. We should note that CP parameterization represents a rich enough class of admissible SAPD parameters in the sense that under this parameterization SAPD can achieve accelerated bias decay in the expected squared distance metric (Zhang et al., 2024) and the accelerated high probability results we derived in Corollary 12.

4.1 Covariance matrix of the iterates under the CP parameterization

The main result of this section is an explicit analytical formula for the asymptotic covariance matrix Σ^{∞} of `SAPD`'s iterates (x_n, y_n) as $n \to \infty$. Our proof yields closed-form solutions under the Chambolle-Pock parameterization (2.3), useful for understanding the effect of parameters on the solution. Due to lengthy calculations involved, the proof is provided in Section D.2 of the Appendix. Our proof technique is based on identifying the conditions on the parameters so that the Lyapunov matrix in (4.2) admits a unique solution and we solve it as a function of λ by diagonalizing A^{λ} given in (4.2) with a proper change of basis.

Theorem 17 *Let $\kappa_{max} \triangleq \rho(K)/\sqrt{\mu_x \mu_y}$. For any given $\theta > (\sqrt{1+\kappa_{max}^2} - 1)/\kappa_{max}$ fixed, set $\tau = (1-\theta)/(\theta\mu_x)$ and $\sigma = (1-\theta)/(\theta\mu_y)$. Suppose gradient noise sequences are i.i.d. centered Gaussian satisfying $\mathbb{E}[\omega_n^x {\omega_n^x}^{\top}] = \mathbb{E}[\omega_n^y {\omega_n^y}^{\top}] = \frac{\delta^2}{d} I_d$. Then, the iterates $z_n = (x_n, y_n)$ generated by `SAPD` applied to the SCSC problem in (4.1) with the given parameters $\tau, \sigma > 0$ and $\theta \in (0,1)$ converge in distribution to a centered Gaussian distribution with covariance matrix Σ^{∞} satisfying $\Sigma^{\infty} = V^{\top}\Sigma^{\infty,\Lambda}V$ where V is orthogonal, and $\Sigma^{\infty,\Lambda}$ is block diagonal with d blocs $\Sigma^{\infty,\lambda_i} \in \mathbb{R}^{2\times 2}$ for $i = 1, \ldots, d$, where $(\lambda_i)_{1\leqslant i\leqslant d}$ denote the eigenvalues of K. Specifically, for each $\lambda \in Sp(K)$, block $\Sigma^{\infty,\lambda}$ has the following form: For $\lambda = 0$,*

$$\Sigma^{\infty,0} = \frac{\delta^2}{d}\frac{(1-\theta)}{\mu_x^2\mu_y^2(1+\theta)}\begin{bmatrix} \theta^2\mu_y^2 & 0 \\ 0 & \mu_x^2\left(1 + 2\left(1-\theta^2\right)\theta\right) \end{bmatrix}, \tag{4.3}$$

otherwise, for $\lambda \neq 0$,

$$\Sigma^{\infty,\lambda} = \frac{\delta^2}{d}\frac{1-\theta}{P_c(\theta,\kappa)}\begin{bmatrix} \frac{1}{\mu_x^2}\left(P_{1,1}^{(\infty,1)}(\theta,\kappa) + \frac{\lambda^2}{\mu_y^2}P_{1,1}^{(\infty,2)}(\theta,\kappa)\right) & \frac{1}{\lambda\mu_x}\left(P_{1,2}^{(\infty,1)}(\theta,\kappa) + \frac{\lambda^2}{\mu_y^2}P_{1,2}^{(\infty,2)}(\theta,\kappa)\right) \\ \frac{1}{\lambda\mu_x}\left(P_{1,2}^{(\infty,1)}(\theta,\kappa) + \frac{\lambda^2}{\mu_y^2}P_{1,2}^{(\infty,2)}(\theta,\kappa)\right) & \frac{1}{\lambda^2}\left(P_{2,2}^{(\infty,1)}(\theta,\kappa) + \frac{\lambda^2}{\mu_y^2}P_{2,2}^{(\infty,2)}(\theta,\kappa)\right) \end{bmatrix} \tag{4.4}$$

where $P_{1,1}^{(\infty,k)}$, $P_{1,2}^{(\infty,k)}$ and $P_{2,2}^{(\infty,k)}$ for $k = 1, 2$, and P_c are polynomials of (θ, κ) that can be made explicit and are provided in Table 7 of Appendix E. Moreover, for any $\lambda \in Sp(K)$, all elements of the matrix $\Sigma^{\infty,\lambda} \in \mathbb{R}^{4\times4}$ scale with $(1-\theta)$ as $\theta \to 1$.

Proof The proof is given in Section D.2. ■

According to Theorem 17, the matrix $\Sigma^{\infty,\lambda}$ has the property that it scales with $(1-\theta)$ as $\theta \to 1$; we leverage this fact to establish the tightness of our analysis in the next section (Section 4.2).

In Figure 2, we illustrate Theorem 17 on a simple quadratic problem where primal and dual iterates are scalar, i.e., $d = 1$ and $K = c$ is a scalar. In the three panels of Figure 2, we consider three problems P1, P2, P3 from left to right where the problem constants, $(c, \mu_x, \mu_y, \delta)$, are chosen as P1: $(1, 4.4, 1.5, 35)$ - P2: $(1, 2, 20, 50)$ - P3: $(10^{-3}, 0.205, 0.307, 5)$. SAPD was run 2000 times for 500 iterations using CP parameterization in (2.3) with $\theta = 0.99$. For each problem, we estimate the empirical covariance matrix Σ_n for $n \in \{2^k : \ k = 0, \ldots, \log_2(500)\}$. The level set $\{z : \ z^\top\Sigma^\infty z = 1\}$ for the theoretical covariance matrix Σ^∞ derived in Theorem 17 is represented by a brown edged ellipse on each plot. Figure 2 suggests the linear convergence of the matrices to the equilibrium matrix Σ^∞. Subsequently, we observe on these three examples how noise accumulates along iterations, producing covariance matrices that are non-decreasing in the sense of the Loewner ordering. This monotonicity behavior is intuitively expected as the noise accumulates over the iterations, but can also be proven using the fact that covariance matrix Σ_n of $z_n = [x_n^\top, y_n^\top]^\top$ follows a Lyapunov recursion (Laub et al., 1990; Hassibi et al., 1999). We elaborate further on this property by showing below that convergence of Σ_n to Σ^∞ happens at a linear rate characterized by the spectral radius of a particular matrix related to the SAPD iterations. The proof builds on the spectral characterizations of the covariance matrix Σ^∞ obtained in the proof of Theorem 17.

Corollary 18 *In the premise of Theorem 17, for any $\theta \in \left(\kappa_{max}^{-1}(\sqrt{1+\kappa_{max}^2} - 1), 1\right)$, the sequence of covariance matrices $\Sigma_n \triangleq \mathbb{E}[z_n z_n^\top]$ satisfies*

$$\|\Sigma_n - \Sigma^\infty\| = \mathcal{O}\left(\rho(A)^{2n}\right), \tag{4.5}$$

where $z_n \triangleq (x_n^\top, y_n^\top)^\top$, $A \triangleq \begin{bmatrix} \frac{1}{1+\tau\mu_x}I_d & \frac{-\tau}{(1+\tau\mu_x)}K \\ \frac{1}{1+\sigma\mu_y}\left(\frac{\sigma(1+\theta)}{1+\tau\mu_x} - \sigma\theta\right)K & \frac{1}{1+\sigma\mu_y}\left(I_d - \frac{\tau\sigma(1+\theta)}{1+\tau\mu_x}K^2\right) \end{bmatrix}$ with $\rho(A) < 1$.

Proof The proof is provided in Appendix D.3. ■

4.2 Tightness analysis

We discuss in this section that the constants given in Theorem 8 are tight in the sense that under the CP parameterization given in (2.3), which corresponds to a particular solution

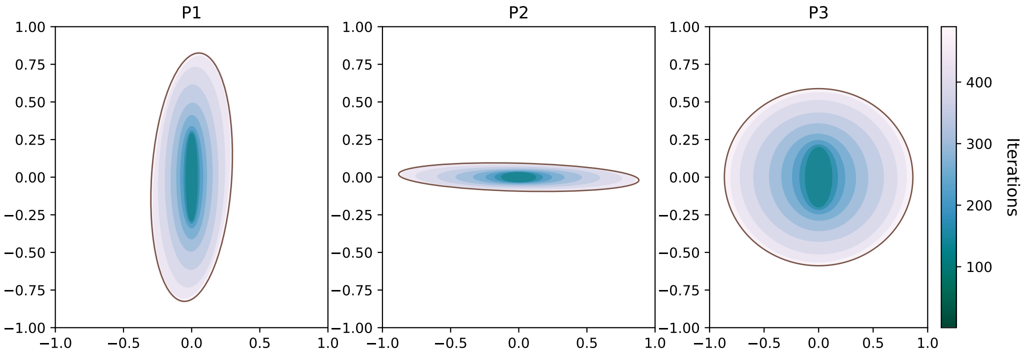

Figure 2: Noise accumulation over SAPD iterates: $\{\Sigma_n\}$ converges to an equilibrium covariance Σ^∞ due to the convergence in distribution of $z_n \xrightarrow{\mathcal{D}} z^\infty$ derived in Thereom 17. By Corollary 18, convergence occurs at a linear rate which is plotted for the level sets $\{z \in \mathbb{R}^2 : z^\top \Sigma_n z = 1\}$ for $d = 1$. The constants $(c, \mu_x, \mu_y, \delta)$ for each problem from left to right: P1: $(1, 4.4, 1.5, 35)$ - P2: $(1, 2, 20, 50)$ - P3: $(10^{-3}, 0.205, 0.307, 5)$, where $K = c \in \mathbb{R}$.

of the matrix inequality in (2.2), the dependency of these constants to θ and p cannot be improved when the number of iterations n is sufficiently large. To this end, we consider quadratic problems subject to additive isotropic Gaussian noise for which we can do exact computations, i.e., both $\{\Delta_n^x\}$ and $\{\Delta_n^y\}$ are i.i.d zero-mean Gaussian random vector sequences with isotropic covariances, and these sequences are independent from each other as well.

In Section 4.1, 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 $\mathrm{z}^\star = (\mathrm{x}^\star, \mathrm{y}^\star)$ and a covariance matrix Σ^∞ for which we provide a formula in (4.4). If we let z_∞ denote a random variable with the stationary distribution π_∞, Theorem 8 implies for any $p \in [0, 1)$ that

$$Q_p(\|z_\infty - \mathrm{z}^\star\|^2) = \limsup_{n\to\infty} Q_p(\|z_n - \mathrm{z}^\star\|^2) = \mathcal{O}\Big((1-\theta)\delta^2\Big(1 + \log\Big(\frac{1}{1-p}\Big)\Big)\Big), \tag{4.6}$$

as $\theta \to 1$. This upper bound (grows) scales linearly with respect to $1-\theta$ and $\log(\frac{1}{1-p})$, and a natural question is whether this scaling can be improved. In the next proposition we provide lower bounds on the quantiles of $\|z_\infty\|^2$ that also grows linearly with respect to $1-\theta$ and $\log(\frac{1}{1-p})$, matching the upper bound in (4.6). Therefore, we conclude that our analysis is tight in the sense that we cannot expect to improve our bound in (4.6) in terms of its dependency to p and $1-\theta$.

Theorem 19 *Let $(z_n)_{n\geqslant 0}$ be the sequence initialized at an arbitrary tuple $z_0 = (x_0, y_0)$ generated by* **SAPD** *under the parameterization* (2.3) *on the quadratic problem* (4.1) *where* $\mathrm{z}^\star = (0,0)$. *Then, the sequence $(z_n)_{n\geqslant 0}$ converges in distribution to a Gaussian vector z_∞. Furthermore, for any $p \in (0,1)$, p-th quantile $Q_p(\|z_\infty - \mathrm{z}^\star\|^2)$ admits the bound*

$$\psi_1(p,\theta) \leqslant Q_p(\|z_\infty - \mathrm{z}^\star\|^2) \leqslant \psi_2(p,\theta),$$

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

Proof The proof is provided in Section D.4 of the appendix. ■

5. Proof of Main Results

5.1 Concentration inequalities through recursive control

This section presents general concentration inequalities that will be specialized later for the analysis of `SAPD`. The first result is a recursive concentration inequality extending the result provided in (Cutler et al., 2021, Proposition 6.7), which is used in the analysis of the stochastic gradient descent (SGD) method for minimization of a smooth strongly convex function in (Cutler et al., 2021). Our variant of this inequality enables us to analyze saddle point problems with acceleration, providing new insights on their robustness properties.

Proposition 20 *Let* $(\mathcal{F}_n)_{n\geqslant 0}$ *be a filtration on* $(\Omega,\mathcal{F},\mathbb{P})$. *Let* $(V_n)_{n\geqslant 0}$, $(T_n)_{n\geqslant 0}$, *and* $(R_n)_{n\geqslant 0}$, *be three scalar stochastic processes adapted to* $(\mathcal{F}_n)_{n\geqslant 0}$ *with following properties: there exist* $\sigma_R, \sigma_T > 0$ *such that for all* $n \geqslant 0$,

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

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

$$V_{n+1} - T_{n+1} - R_{n+1} \leqslant \rho\, V_n, \quad \forall n \geqslant 0, \tag{5.1}$$

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* $\mathbb{E}\left[e^{\lambda V_{n+1}}\right] \leqslant e^{\lambda\sigma_R^2}\mathbb{E}\left[e^{\frac{\lambda(1+\rho)}{2}V_n}\right]$, *for* $n \geqslant 0$.

Proof Our proof follows closely the arguments of (Cutler et al., 2021, Proposition 6.7). The main difference is in the term T_{n+1} which takes the specific form $T_{n+1} = G_{n+1}\sqrt{V_n}$ in (Cutler et al., 2021), where G_{n+1} conditioned on $\mathcal{F}_n$ is assumed to be subGaussian. For any $\lambda \geqslant 0$, (5.1) together with Cauchy-Schwarz inequality implies that

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

Thus for $\lambda \in \left(0, \frac{1}{2\sigma_R^2}\right]$, we have $\mathbb{E}\left[e^{\lambda V_{n+1}} | \mathcal{F}_n\right] \leqslant e^{\lambda\sigma_R^2}e^{\lambda(\rho+2\lambda\sigma_T^2)V_n}$. Setting $0 \leqslant \lambda \leqslant \min\left\{\frac{1}{2\sigma_R^2}, \frac{1-\rho}{4\sigma_T^2}\right\}$ and taking the non-conditional expectation, we ensure that $\mathbb{E}\left[e^{\lambda V_{n+1}}\right] \leqslant e^{\lambda\sigma_R^2}\mathbb{E}\left[e^{\lambda\frac{1+\rho}{2}V_n}\right]$. This completes the proof. ■

Unrolling the above recursive property on the moment generating function of V_n provides us with high probability results on $(V_n)_{n\geqslant 0}$, given in the next result.

Proposition 21 *Let V_n, T_n, R_n be defined as in Proposition 20. Then, for all $n \geqslant 0$ and $\lambda \in \left[0, \min\left\{\frac{1-\rho}{4\sigma_T^2}, \frac{1}{2\sigma_R^2}\right\}\right]$,*

$$\mathbb{E}\left[e^{\lambda V_n}\right] \leqslant e^{\frac{2\lambda\sigma_R^2}{1-\rho}} \mathbb{E}\left[e^{\lambda\left(\frac{1+\rho}{2}\right)^n V_0}\right]. \tag{5.2}$$

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

$$\mathbb{P}\left[V_n \leqslant \left(\frac{1+\rho}{2}\right)^n C_0 + \frac{2\sigma_R^2}{1-\rho}\left(1 + \max\left\{1, 2\,\frac{\sigma_T^2}{\sigma_R^2}\right\}\log\left(\frac{1}{1-p}\right)\right)\right] \geqslant p. \tag{5.3}$$

Alternatively, if V_0 can be expressed as $V_0 = C_0 + \mathcal{U}$ such that $C_0 \geqslant 0$ is constant and $\mathcal{U}$ satisfies $\mathbb{E}\left[e^{\lambda\mathcal{U}}\right] \leqslant e^{\alpha\lambda + \beta\lambda^2}$, for all $\lambda \in \left[0, \frac{1}{\bar{\alpha}}\right]$, for some constants $\alpha, \bar{\alpha} > 0$ and $\beta \geqslant 0$, then for any $p \in [0, 1)$ and $\lambda \in [0, \gamma]$ where $\gamma \triangleq \frac{1-\rho}{\max\{\bar{\alpha}, 2\sigma_R^2, 4\sigma_T^2\}}$, we have

$$\mathbb{P}\left(V_n \leqslant \left(\frac{1+\rho}{2}\right)^n (C_0 + \alpha) + \left(\frac{1+\rho}{2}\right)^{2n} \lambda\beta + \frac{2\sigma_R^2}{1-\rho} + \frac{1}{\lambda}\log\left(\frac{1}{1-p}\right)\right) \geqslant p. \tag{5.4}$$

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

$$\mathbb{E}\left[e^{\lambda V_n}\right] \leqslant \mathbb{E}\left[e^{\lambda\left(\frac{1+\rho}{2}\right)^n V_0 + \lambda\sigma_R^2 \sum_{k=0}^{n-1}\left(\frac{1+\rho}{2}\right)^k}\right]. \tag{5.5}$$

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 $n \geqslant 0$, next we show it also holds for $n + 1$. According to Proposition 20,

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

where the second inequality follows from the induction hypothesis since

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

and this completes the induction. Thus, (5.2) follows from using $\sum_{k=0}^{n-1}\left(\frac{1+\rho}{2}\right)^k \leqslant \frac{2}{1-\rho}$ within (5.5). The remaining statements follow from a Chernoff bound; indeed, if $V_0 = C_0$ is constant, we obtain

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

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 $\lambda \in \left(0, \min\left\{\frac{1-\rho}{4\sigma_T^2}, \frac{1}{2\sigma_R^2}, \frac{1}{\bar{\alpha}}\right\}\right)$,

$$\mathbb{E}\left[e^{\lambda V_n}\right] \leqslant e^{\frac{2\lambda\sigma_R^2}{1-\rho}}\mathbb{E}\left[e^{\lambda\left(\frac{1+\rho}{2}\right)^n (C_0 + \mathcal{U})}\right] \leqslant e^{\lambda\left(\left(\frac{1+\rho}{2}\right)^n (C_0+\alpha) + \frac{2\sigma_R^2}{1-\rho}\right) + \lambda^2\left(\frac{1+\rho}{2}\right)^{2n}\beta}. \tag{5.6}$$

Thus, for all $t \geqslant 0$,

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

Fixing an arbitrary non-negative λ such that $\lambda \leqslant \frac{1-\rho}{\max\{\bar{\alpha}, 2\sigma_R^2, 4\sigma_T^2\}}$, 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 (5.4). ■

Based on Proposition 21, as a corollary, one can derive convergence rates for the CVaR and EVaR risk measures of the scalar process $(V_n)_{n\geqslant 0}$.

Corollary 22 *Let V_n, T_n, R_n, γ be defined as in Proposition 21, with V_0 of the form $V_0 = C_0 + \mathcal{U}$. Then, for any $p \in [0,1)$ and $\lambda \in [0,\gamma]$,*

$$\mathrm{CVaR}_p(V_n^{\frac{1}{2}}) \leqslant \left(\frac{1+\rho}{2}\right)^{\frac{n}{2}} \sqrt{C_0+\alpha+\lambda\beta} + \sqrt{\frac{2}{1-\rho}\sigma_R^2 + \frac{1}{\lambda}\left(1+\log\left(\frac{1}{1-p}\right)\right)}. \tag{5.7}$$

Proof Note that the first and second terms on the right-hand side of (5.4) satisfy

$$\left(\frac{1+\rho}{2}\right)^n (C_0+\alpha) + \left(\frac{1+\rho}{2}\right)^{2n}\lambda\beta \leqslant \left(\frac{1+\rho}{2}\right)^n (C_0+\alpha+\lambda\beta).$$

Hence, by integrating the resulting looser bound with respect to p, and using CVaR's integral formulation in (2.10), we obtain

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

which directly implies (5.7), due to Lemma (28) and the sub-additivity of $t \mapsto \sqrt{t}$. ■

Corollary 23 *Let V_n, T_n, R_n, γ be defined as in Proposition 21. Then, for any $p \in [0,1)$, and $\lambda \in [0,\gamma]$,*

$$\mathrm{EVaR}_p\left(V_n^{\frac{1}{2}}\right) \leqslant \left(\frac{1+\rho}{2}\right)^{n/2} \sqrt{C_0+\alpha+\lambda\beta} + \sqrt{\frac{2}{1-\rho}}\sigma_R + \left(\sqrt{\frac{1}{\lambda}}\log(\frac{1}{1-p}) + \frac{\sqrt{\pi}}{\sqrt{\lambda}}\right). \tag{5.8}$$

Proof The bound in (5.4) of Proposition 21 ensures that for all $p \in [0,1)$ and $\lambda \in [0,\gamma]$, the p-th quantile of V_n satisfies

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

hence, non-negativity of V_n, Lemma 28 and sub-additivity of $t \mapsto \sqrt{t}$ together imply that

$$Q_p\left(V_n^{1/2}\right) \leqslant \left(\frac{1+\rho}{2}\right)^{n/2} \sqrt{C_0+\alpha+\lambda\beta} + \sqrt{\frac{2}{1-\rho}}\sigma_R + \frac{1}{\sqrt{\lambda}}\log\left(\frac{1}{1-p}\right)^{1/2}. \tag{5.9}$$

For $n \geqslant 0$, 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 (5.9) implies

$$\mathbb{P}\left(U_n > t\right) \leqslant e^{-\lambda t^2} \quad \forall\ t \geqslant 0. \tag{5.10}$$

Therefore, following standard arguments from (Vershynin, 2018), we have for any $\eta > 0$ that

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

where we used (5.10). On the other hand,

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

where we used $\log(1+x) \leqslant x$ for $x \geqslant 0$. Finally, by translation invariance of the EVaR, we obtain $\mathrm{EVaR}_p\left[V_n^{\frac{1}{2}}\right] \leqslant \left(\frac{1+\rho}{2}\right)^{n/2} \sqrt{C_0 + \alpha + \lambda\beta} + \sqrt{\frac{2}{1-\rho}}\sigma_R + \left(\sqrt{\frac{1}{\lambda}\log\left(\frac{1}{1-p}\right)} + \frac{\sqrt{\pi}}{\sqrt{\lambda}}\right)$. ■

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

Corollary 24 *Let V_n, T_n, R_n, γ be defined as in Proposition 21. Then, for any $r > 0$, and $\lambda \in [0, \gamma]$,*

$$\mathcal{R}_{\chi^2, r}\left(V_n^{\frac{1}{2}}\right) \leqslant \left(\frac{1+\rho}{2}\right)^{n/2} \sqrt{C_0 + \alpha + \lambda\beta} + \sqrt{\frac{2}{1-\rho}}\sigma_R + \left(\sqrt{\frac{1}{\lambda}\log\left(1+r\right)} + \frac{\sqrt{\pi}}{\sqrt{\lambda}}\right). \tag{5.11}$$

Proof By(Gibbs and Su, 2002, Theorem 5), for all $\mathbb{Q} \ll \mathbb{P}$, we have $D_{\varphi_{\mathrm{KL}}}(\mathbb{Q}\|\,\mathbb{P}) \leqslant \log\Big(1 + D_{\varphi_{\chi^2}}(\mathbb{Q}\|\,\mathbb{P})\Big)$, where $\varphi_{\mathrm{KL}}(t) = t\log(t) - t + 1$. Therefore, for any integrable random variable $U : \Omega \to \mathbb{R}$,

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

whenever $\mathrm{EVaR}_{1-1/(1+r)}(U) < \infty$, where we used the EVaR representation given in Table 2. The statement follows directly from Corollary 23. ■

In the next section, we design scalar processes V_n, T_n, R_n which satisfy the above assumptions while dominating the error $\mathcal{W}_n$ on `SAPD` iterates, so that Proposition 20, Corollaries 23 and 24 will allow us to prove our main results, stated in Theorem 8 and Theorem 15.

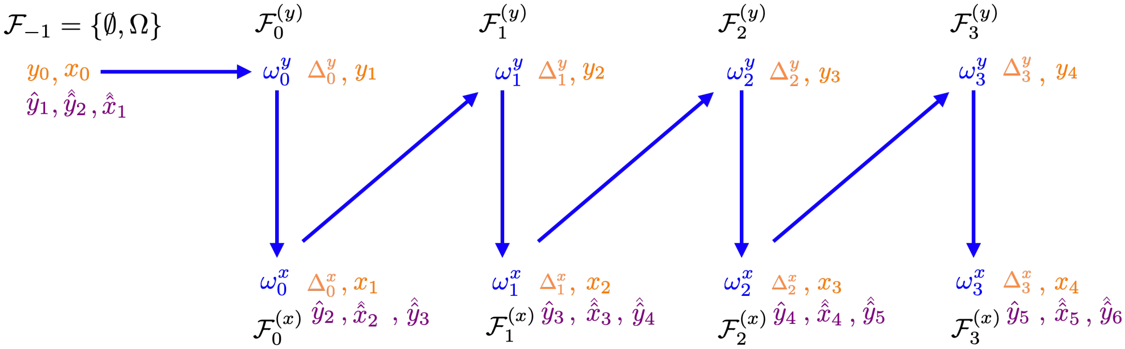
$\mathcal{F}_{-1} = \{\emptyset, \Omega\}$
y_0, x_0
$\hat{y}_1, \hat{\hat{y}}_2, \hat{\hat{x}}_1$
$\mathcal{F}_0^{(y)}$
ω_0^y Δ_0^y, y_1
ω_0^x Δ_0^x, x_1
$\mathcal{F}_0^{(x)}$ $\hat{y}_2, \hat{\hat{x}}_2, \hat{\hat{y}}_3$
$\mathcal{F}_1^{(y)}$
ω_1^y Δ_1^y, y_2
ω_1^x Δ_1^x, x_2
$\mathcal{F}_1^{(x)}$ $\hat{y}_3, \hat{\hat{x}}_3, \hat{\hat{y}}_4$
$\mathcal{F}_2^{(y)}$
ω_2^y Δ_2^y, y_3
ω_2^x Δ_2^x, x_3
$\mathcal{F}_2^{(x)}$ $\hat{y}_4, \hat{\hat{x}}_4, \hat{\hat{y}}_5$
$\mathcal{F}_3^{(y)}$
ω_3^y Δ_3^y, y_4
ω_3^x Δ_3^x, x_4
$\mathcal{F}_3^{(x)}$ $\hat{y}_5, \hat{\hat{x}}_5, \hat{\hat{y}}_6$

Figure 3: Measurability of `SAPD` sequences. Our analysis is made possible by the introduction of predictable counterparts $\hat{\hat{\mathrm{x}}}_n, \hat{\mathrm{y}}_n, \hat{\hat{\mathrm{y}}}_n$ to the iterates x_n, y_n as defined in (5.12).

5.2 Proofs of Theorem 8 and Theorem 15

For proving the main results of this paper, namely Theorem 8 and Theorem 15, the application of the recursive control inequality from Section 5.1 is not straightforward. In particular, Gauss-Seidel type updates within `SAPD` significantly complicate the measurability properties of `SAPD` iterate sequence, as illustrated in Figure 3: the iterates x_n and y_n are measurable with respect to different filtrations $\mathcal{F}_{n-1}^{(x)}$ and $\mathcal{F}_{n-1}^{(y)}$. We circumvent this issue by introducing a stochastic process $\{V_n\}_{n\geqslant 0}$ that almost surely upper bounds the distance to the saddle point while exhibiting simpler measurability characteristics as discussed next. We note that even though algorithms with Gauss-Seidel type updates, such as `SAPD`, are significantly more complicated to analyze than their Jacobi counterparts, such an analysis is rewarding in the sense that algorithms with Gauss-Seidel type updates can often be faster than those using Jacobi type updates, see (Zhang et al., 2024, 2022). Indeed, our analysis for `SAPD` allowed us to obtain high-probability bounds that demonstrate an accelerated behavior for a stochastic primal-dual algorithm for SP problems.

5.2.1 Proof of Theorem 8

Our proof combines several ingredients. Let $(\tau, \sigma, \theta, \rho, \alpha)$ be a solution to the matrix inequality in (2.2). Recall the weighted distance square metric $\mathcal{W}_n{=}\frac{1}{2\tau}\|x_n - \mathrm{x}^\star\| + \frac{1-\alpha\sigma}{2\sigma}\|y_n - \mathrm{y}^\star\|^2$ we introduced in (2.4). In the proof, we use a scaled version $\mathcal{E}_n \triangleq \mathcal{W}_n/\rho$ for $n \geqslant 0$ that simplifies the analysis. We first introduce the following auxiliary iterates $\hat{\mathrm{y}}_n, \hat{\hat{\mathrm{x}}}_n, \hat{\hat{\mathrm{y}}}_n$ which can be interpreted as the "noise-free counterparts" to the actual iterates x_n, y_n in the sense they represent roughly how the algorithm would behave if the gradients were deterministic in lieu of being stochastic at step n:

$$\hat{\hat{\mathrm{x}}}_0 \triangleq x_0, \qquad \hat{\hat{\mathrm{x}}}_{n+1} \triangleq \mathrm{prox}_{\tau f}\left(x_n - \tau\, \nabla_{\mathrm{x}} \Phi(x_n, \hat{\mathrm{y}}_{n+1})\right), \tag{5.12}$$

$$\hat{\mathrm{y}}_0 \triangleq \hat{\hat{\mathrm{y}}}_0 \triangleq y_0, \quad \hat{\mathrm{y}}_{n+1} \triangleq \mathrm{prox}_{\sigma g}\Big(y_n + \sigma(1+\theta)\,\nabla_{\mathrm{y}} \Phi(x_n, y_n) - \sigma\theta\,\nabla_{\mathrm{y}} \Phi(x_{n-1}, y_{n-1})\Big), \tag{5.13}$$

$$\hat{\hat{\mathrm{y}}}_{n+1} \triangleq \mathrm{prox}_{\sigma g}\Big(\hat{\mathrm{y}}_n + \sigma(1+\theta)\,\nabla_{\mathrm{y}} \Phi(\hat{\hat{\mathrm{x}}}_n, \hat{\mathrm{y}}_n) - \sigma\theta\,\nabla_{\mathrm{y}} \Phi(x_{n-1}, y_{n-1})\Big). \tag{5.14}$$

where we recall that $x_{-1} = x_0$ and $y_{-1} = y_0$ (see Algorithm 1). These auxiliary iterates were first introduced in Zhang et al. (2024) to derive convergence rates for `SAPD` in terms of expected weighted distance square, to establish Theorem 2. Here, we further leverage their measurability properties, as illustrated in Figure 3, in order to apply Proposition 21 and to obtain high-probability results for SAPD. Our proof is based on establishing an almost sure upper bound of the quantity $\mathcal{E}_{n+1} + \mathcal{E}_n$ by a scalar process V_n, and then showing that our choice of V_n satisfies the assumptions of Proposition 20. This will then directly yield the desired high-probability estimates for SAPD. We start with a proposition that provides an almost sure bound to the scaled squared distance metric $\mathcal{E}_n$. Although this bound is already present in substance in (Zhang et al., 2024), it does not appear explicitly. For completeness, in Appendix C.1, we provide its proof based on various arguments developed in (Zhang et al., 2024).

Proposition 25 *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 $\theta \geqslant 0$ that satisfy (2.2) for some $\rho \in (0,1)$ and $\alpha \in [0, \sigma^{-1})$, the following almost sure bound on $\mathcal{E}_n$,*

$$\mathcal{E}_n \leqslant \rho^{n-1}\mathcal{W}_{\tau,\sigma} + \sum_{k=0}^{n-1} \rho^{n-1-k}\Big(\langle \Delta_k^x, \mathrm{x}^\star - x_{k+1}\rangle + \langle (1+\theta)\Delta_k^y - \theta\Delta_{k-1}^y, y_{k+1} - \mathrm{y}^\star\rangle\Big), \quad (5.15)$$

holds for all $n \geqslant 1$, where $\mathcal{E}_n = \frac{1}{2\rho\tau}\|x_n - \mathrm{x}^\star\|^2 + \frac{1-\alpha\sigma}{2\rho\sigma}\|y_n - \mathrm{y}^\star\|^2$, and $\mathcal{W}_{\tau,\sigma} = \frac{1}{2\tau}\|x_0 - \mathrm{x}^\star\|^2 + \frac{1}{2\sigma}\|y_0 - \mathrm{y}^\star\|^2$.

Proof The proof is provided in Appendix C.1. ■

Now, equipped with Proposition 25, we can write

$$\begin{aligned}
\mathcal{E}_n &\leqslant \rho^{n-1}\mathcal{W}_{\tau,\sigma} + \sum_{k=0}^{n-1} \rho^{n-1-k}\Big(\langle \Delta_k^x, \mathrm{x}^\star - x_{k+1}\rangle + \langle (1+\theta)\Delta_k^y - \theta\Delta_{k-1}^y, y_{k+1} - \mathrm{y}^\star\rangle\Big) \\
&= \rho^{n-1}\mathcal{W}_{\tau,\sigma} + \sum_{k=0}^{n-1} \rho^{n-1-k}\Big(\langle \Delta_k^x, \mathrm{x}^\star - \hat{\hat{\mathrm{x}}}_{k+1}\rangle + (1+\theta)\langle \Delta_k^y, \hat{\mathrm{y}}_{k+1} - \mathrm{y}^\star\rangle - \theta\langle \Delta_{k-1}^y, \hat{\hat{\mathrm{y}}}_{k+1} - \mathrm{y}^\star\rangle\Big) \\
&\quad + \sum_{k=0}^{n-1} \rho^{n-1-k}\Big(\langle \Delta_k^x, \hat{\hat{\mathrm{x}}}_{k+1} - x_{k+1}\rangle + (1+\theta)\langle \Delta_k^y, y_{k+1} - \hat{\mathrm{y}}_{k+1}\rangle - \theta\langle \Delta_{k-1}^y, y_{k+1} - \hat{\hat{\mathrm{y}}}_{k+1}\rangle\Big).
\end{aligned} \quad (5.16)$$

For $k \geqslant 0$, introducing the scalar quantities

$$P_k^{(1)} \triangleq \langle \Delta_k^x, \mathrm{x}^\star - \hat{\hat{\mathrm{x}}}_{k+1}\rangle + (1+\theta)\langle \Delta_k^y, \hat{\mathrm{y}}_{k+1} - \mathrm{y}^\star\rangle, \qquad P_k^{(2)} \triangleq \frac{-\theta}{\rho}\langle \Delta_k^y, \hat{\hat{\mathrm{y}}}_{k+2} - \mathrm{y}^\star\rangle, \quad (5.17a)$$

$$Q_k \triangleq \langle \Delta_k^x, \hat{\hat{\mathrm{x}}}_{k+1} - x_{k+1}\rangle + (1+\theta)\langle \Delta_k^y, y_{k+1} - \hat{\mathrm{y}}_{k+1}\rangle - \theta\langle \Delta_{k-1}^y, y_{k+1} - \hat{\hat{\mathrm{y}}}_{k+1}\rangle, \quad (5.17b)$$

rearranging the sums in (5.16) and using $\Delta_{-1}^y = \mathbf{0}$, we may write (5.16) equivalently as follows:

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

Now notice that for $n \geqslant 0$,

$$\begin{aligned}\mathcal{E}_{n+1} + \mathcal{E}_n &\leqslant \left(1 + \frac{1}{\rho}\right)\left(\rho^n \mathcal{W}_{\tau,\sigma} + \sum_{k=0}^{n-1} \rho^{n-k} P_k^{(1)} + \sum_{k=0}^{n-2} \rho^{n-k} P_k^{(2)} + \sum_{k=0}^{n-1} \rho^{n-k} Q_k\right) + P_n^{(1)} + \rho P_{n-1}^{(2)} + Q_n \\ &= \left(1 + \frac{1}{\rho}\right)\left(\rho^n \mathcal{W}_{\tau,\sigma} + \sum_{k=0}^{n} \rho^{n-k} P_k^{(1)} + \sum_{k=0}^{n} \rho^{n-k} P_k^{(2)}\right) \\ &\qquad + \left(1 + \frac{1}{\rho}\right)\left(\frac{-1}{1+\rho} P_n^{(1)} - P_n^{(2)} - \frac{\rho}{1+\rho} P_{n-1}^{(2)} - \frac{1}{1+\rho} Q_n + \sum_{k=0}^{n} \rho^{n-k} Q_k\right). \end{aligned} \tag{5.18}$$

We next present a lemma which bounds the terms on the right-hand side of the above equality.

Lemma 26 *Let $P_n^{(1)}, P_n^{(2)}$ and Q_n be defined as in (5.17). Then, for any $n \geqslant 0$,*

$$\begin{aligned}&\frac{-1}{1+\rho} P_n^{(1)} - P_n^{(2)} - \frac{\rho}{1+\rho} P_{n-1}^{(2)} - \frac{1}{1+\rho} Q_n + \sum_{k=0}^{n} \rho^{n-k} Q_k \\ &\qquad\qquad \leqslant \frac{\rho}{2(1+\rho)}(\mathcal{E}_{n+1} + \mathcal{E}_n) + \sum_{k=0}^{n} \rho^{n-k}\left(\mathcal{Q}_x \|\Delta_k^x\|^2 + \mathcal{Q}_y \|\Delta_k^y\|^2\right),\end{aligned}$$

for some positive constants $\mathcal{Q}_x$ and $\mathcal{Q}_y$, which depend only on the algorithm and problem parameters, and are provided explicitly in Table 4 of Appendix E.

Proof The proof is provided in Appendix C.2.1. ■

Applying Lemma 26 to the inequality (5.18), we obtain

$$\frac{\rho}{2(1+\rho)}(\mathcal{E}_{n+1} + \mathcal{E}_n) \leqslant \rho^n \mathcal{W}_{\tau,\sigma} + \sum_{k=0}^{n} \rho^{n-k} P_k^{(1)} + \sum_{k=0}^{n} \rho^{n-k} P_k^{(2)} + \sum_{k=0}^{n} \rho^{n-k}\Big(\mathcal{Q}_x \|\Delta_k^x\|^2 + \mathcal{Q}_y \|\Delta_k^y\|^2\Big). \tag{5.19}$$

For $n \in \mathbb{N}$, we define V_n, T_{n+1} and R_{n+1} as follows:

$$\begin{aligned} V_n &\triangleq \rho^n \mathcal{W}_{\tau,\sigma} + \sum_{k=0}^{n} \rho^{n-k}\Big(P_k^{(1)} + P_k^{(2)}\Big) + \sum_{k=0}^{n} \rho^{n-k}\Big(\mathcal{Q}_x \|\Delta_k^x\|^2 + \mathcal{Q}_y \|\Delta_k^y\|^2\Big), \\ T_{n+1} &\triangleq P_{n+1}^{(1)} + P_{n+1}^{(2)}, \qquad R_{n+1} \triangleq \mathcal{Q}_x \|\Delta_{n+1}^x\|^2 + \mathcal{Q}_y \|\Delta_{n+1}^y\|^2; \end{aligned} \tag{5.20}$$

therefore, (5.19) implies that

$$\frac{\rho}{2(1+\rho)}(\mathcal{E}_{n+1} + \mathcal{E}_n) \leqslant V_n, \quad \forall\ n \geqslant 0, \quad \text{a.s.} \tag{5.21}$$

Next, we argue that V_n satisfies the assumptions of the recursive control inequality in (20). To achieve this goal, we will use the following lemma.

Lemma 27 *For any $n \in \mathbb{N}$ and $\rho \in (0,1)$, the following inequalities,*

$$\begin{aligned}8\| \hat{\hat{\mathrm{x}}}_{n+1} - \mathrm{x}^\star \|^2 &\leqslant \|A_1\|^2 \, \frac{\rho}{2(1+\rho)} \left(\mathcal{E}_n + \mathcal{E}_{n-1}\right), \\ 16(1+\theta)^2 \| \hat{\mathrm{y}}_{n+1} - \mathrm{y}^\star \|^2 &\leqslant \|A_2\|^2 \, \frac{\rho}{2(1+\rho)} \left(\mathcal{E}_n + \mathcal{E}_{n-1}\right), \\ 16\frac{\theta^2}{\rho^2} \| \hat{\hat{\mathrm{y}}}_{n+2} - \mathrm{y}^\star \|^2 &\leqslant \|A_3\|^2 \, \frac{\rho}{2(1+\rho)} \left(\mathcal{E}_n + \mathcal{E}_{n-1}\right),\end{aligned} \tag{5.22}$$

hold almost surely with the convention that $\mathcal{E}_{-1} \triangleq \mathcal{E}_0$, for some vectors $A_1, A_2, A_3 \in \mathbb{R}^4$ which are explicitly provided in Table 5 of Appendix E.

Proof The proof is provided in Appendix C.2.2. ■

Let us now show that V_n satisfies the assumptions of the recursive control inequality in (20). Indeed, for any $n \geqslant 0$, $V_{n+1} - V_n = (\rho - 1)V_n + P^{(1)}_{n+1} + P^{(2)}_{n+1} + \mathcal{Q}_x \|\Delta^x_{n+1}\|^2 + \mathcal{Q}_y \|\Delta^y_{n+1}\|^2$, which is equivalent to $V_{n+1}{=}\rho V_n{+}T_{n+1} + R_{n+1}$. Let $(\mathcal{F}_n)_{n \geqslant -1}$ be the filtration defined as $\mathcal{F}_{-1} \triangleq \{\varnothing, \Omega\}$, and $\mathcal{F}_n \triangleq \mathcal{F}^x_n = \sigma\left(\mathcal{F}_{n-1} \cup \sigma(\Delta^y_n) \cup \sigma(\Delta^x_n)\right)$, for all $n \geqslant 0$.

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 due to (5.21). Second, for any $n \geqslant 0$, since Δ^x_n and Δ^y_n are norm-subGaussian conditioned respectively on $\mathcal{F}^y_n$ and $\mathcal{F}^x_{n-1}$, for any $\lambda \geqslant 0$, we get that

$$\begin{aligned}\mathbb{E}\left[e^{\lambda T_{n+1}} \middle| \mathcal{F}_n\right] &= \mathbb{E}\left[e^{\lambda \left\langle \Delta^y_{n+1},\ (1+\theta)\left(\hat{\mathrm{y}}_{n+2} - \mathrm{y}^\star\right) - \theta\rho^{-1}(\hat{\hat{\mathrm{y}}}_{n+3} - \mathrm{y}^\star) \right\rangle} \mathbb{E}\left[e^{\lambda\langle \Delta^x_{n+1}, \mathrm{x}^\star - \hat{\hat{\mathrm{x}}}_{n+2}\rangle} \middle| \mathcal{F}^y_{n+1}\right] \middle| \mathcal{F}_n\right] \\ &\leqslant e^{8\lambda^2\left(\| \hat{\hat{\mathrm{x}}}_{n+2} - \mathrm{x}^\star \|^2 \delta_x^2 + \|(1+\theta)\left(\hat{\mathrm{y}}_{n+2} - \mathrm{y}^\star\right) - \theta\rho^{-1}\left(\hat{\hat{\mathrm{y}}}_{n+3} - \mathrm{y}^\star\right)\|^2 \delta_y^2\right)} \\ &\leqslant e^{8\lambda^2\left(\| \hat{\hat{\mathrm{x}}}_{n+2} - \mathrm{x}^\star \|^2 \delta_x^2 + 2(1+\theta)^2 \| \hat{\mathrm{y}}_{n+2} - \mathrm{y}^\star \|^2 + 2\theta^2\rho^{-2} \| \hat{\hat{\mathrm{y}}}_{n+3} - \mathrm{y}^\star \|^2 \delta_y^2\right)},\end{aligned}$$

where we used Lemma 6 and the inequality $(a+b)^2 \leqslant 2a^2 + 2b^2$ for scalars a, b in the last step, noting that $\hat{\hat{\mathrm{x}}}_{n+2}$, $\hat{\mathrm{y}}_{n+2}$, $\hat{\hat{\mathrm{y}}}_{n+3}$ are all $\mathcal{F}_n$-measurable. Hence, in view of Lemma 27 provided above and the bound in (5.21), we have

$$\mathbb{E}\left[e^{\lambda T_{n+1}} \middle| \mathcal{F}_n\right] \leqslant e^{\lambda^2\left(\|A_1\|^2\delta_x^2 + \left(\|A_2\|^2 + \|A_3\|^2\right)\delta_y^2\right) \frac{\rho}{2(1+\rho)} \left(\mathcal{E}_{n+1} + \mathcal{E}_n\right)} \leqslant e^{\lambda^2 \left(\|A_1\|^2\delta_x^2 + \left(\|A_2\|^2 + \|A_3\|^2\right)\delta_y^2 \right) V_n}, \tag{5.23}$$

where we used (5.21) to obtain the second inequality. Third, for all $n \geqslant 0$ and $\lambda \in \left(0, \frac{1}{4\max\{\mathcal{Q}_x\delta_x^2, \mathcal{Q}_y\delta_y^2\}}\right)$, we have in view of Lemma 5

$$\mathbb{E}\left[e^{\lambda R_{n+1}} \middle| \mathcal{F}_n\right] = \mathbb{E}\left[e^{\lambda \mathcal{Q}_y \|\Delta^y_{n+1}\|^2} \mathbb{E}\left[e^{\lambda \mathcal{Q}_x \|\Delta^x_{n+1}\|^2} \middle| \mathcal{F}^y_{n+1}\right] \middle| \mathcal{F}_n\right] \leqslant \exp\left(8\lambda\left(\mathcal{Q}_x\delta_x^2 + \mathcal{Q}_y\delta_y^2\right)\right). \tag{5.24}$$

Finally, we next argue that V_0 can be expressed as $V_0 = \mathcal{W}_{\tau,\sigma} + \mathcal{U}$ for some $\mathcal{U}$ satisfying $\mathbb{E}\left[e^{\lambda\mathcal{U}}\right] \leqslant e^{\alpha\lambda + \beta\lambda^2}$, $\quad \forall \lambda \in \left[0, \frac{1}{\bar{\alpha}}\right]$, for some constants $\alpha, \bar{\alpha} > 0$ and $\beta \geqslant 0$. First, note that $\hat{\mathrm{y}}_1, \hat{\hat{\mathrm{x}}}_1$ and $\hat{\hat{\mathrm{y}}}_2$ are all deterministic quantities as they depend only on the initialization; hence, using the inequality $u^\top v \leqslant \frac{\gamma}{2(1-\rho)}\|u\|^2 + \frac{(1-\rho)}{2\gamma}\|v\|^2$ for any $\gamma > 0$, we observe that for all $\lambda \in \left[0, \left(4\max\{\left(\frac{1-\rho}{2\mu_x} + \mathcal{Q}_x\right)\delta_x^2, \left(\frac{(1-\rho)}{\mu_y} + \mathcal{Q}_y\right)\delta_y^2\}\right)^{-1}\right]$, we have

$$\begin{aligned}
&\mathbb{E}\left[e^{\lambda(V_0-\mathcal{W}_{\tau,\sigma})}\right] = \mathbb{E}\left[e^{\lambda\left(P_0^{(1)}+P_0^{(2)}+\mathcal{Q}_x\|\Delta_0^x\|^2+\mathcal{Q}_y\|\Delta_0^y\|^2\right)}\right]\\
&= \mathbb{E}\left[e^{\lambda\langle\Delta_0^x,\mathrm{x}^\star-\hat{\hat{\mathrm{x}}}_1\rangle+\lambda\langle\Delta_0^y,(1+\theta)(\hat{\mathrm{y}}_1-\mathrm{y}^\star)-\frac{\theta}{\rho}(\hat{\hat{\mathrm{y}}}_2-\mathrm{y}^\star)\rangle+\lambda\mathcal{Q}_x\|\Delta_0^x\|^2+\lambda\mathcal{Q}_y\|\Delta_0^y\|^2}\right]\\
&\leqslant \mathbb{E}\Bigg[e^{\frac{\lambda}{2(1-\rho)}\left(\mu_x\|\hat{\hat{\mathrm{x}}}_1-\mathrm{x}^\star\|^2+(1+\theta)^2\mu_y\|\hat{\mathrm{y}}_1-\mathrm{y}^\star\|^2+\theta^2\rho^{-2}\mu_y\|\hat{\hat{\mathrm{y}}}_2-\mathrm{y}^\star\|^2\right)}\\
&\qquad e^{\frac{\lambda(1-\rho)}{2}\left(\frac{1}{\mu_x}\|\Delta_0^x\|^2+\frac{1}{\mu_y}\|\Delta_0^y\|^2+\frac{1}{\mu_y}\|\Delta_0^y\|^2\right)+\lambda\mathcal{Q}_x\|\Delta_0^x\|^2+\lambda\mathcal{Q}_y\|\Delta_0^y\|^2}\Bigg]\\
&= e^{\frac{\lambda}{2(1-\rho)}\left(\mu_x\|\hat{\hat{\mathrm{x}}}_1-\mathrm{x}^\star\|^2+(1+\theta)^2\mu_y\|\hat{\mathrm{y}}_1-\mathrm{y}^\star\|^2+\theta^2\rho^{-2}\mu_y\|\hat{\hat{\mathrm{y}}}_2-\mathrm{y}^\star\|^2\right)}\,\mathbb{E}\left[e^{\lambda\left(\left(\frac{1-\rho}{2\mu_x}+\mathcal{Q}_x\right)\|\Delta_0^x\|^2+\left(\frac{1-\rho}{\mu_y}+\mathcal{Q}_y\right)\|\Delta_0^y\|^2\right)}\right].
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{E}\left[e^{\lambda(V_0-\mathcal{W}_{\tau,\sigma})}\right]\\
&\leqslant e^{\frac{\lambda}{2(1-\rho)}\left(\mu_x\|\hat{\hat{\mathrm{x}}}_1-\mathrm{x}^\star\|^2+(1+\theta)^2\mu_y\|\hat{\mathrm{y}}_1-\mathrm{y}^\star\|^2+\theta^2\rho^{-2}\mu_y\|\hat{\hat{\mathrm{y}}}_2-\mathrm{y}^\star\|^2\right)}e^{8\lambda\left(\left(\frac{1-\rho}{2\mu_x}+\mathcal{Q}_x\right)\delta_x^2+\left(\frac{1-\rho}{\mu_y}+\mathcal{Q}_y\right)\delta_y^2\right)}\\
&\leqslant e^{\frac{\lambda}{2(1-\rho)}\left(\frac{\mu_x}{8}\|A_1\|^2+\frac{\mu_y}{16}(\|A_2\|^2+\|A_3\|^2)\right)\frac{\rho}{2(1+\rho)}\frac{2}{\rho}\mathcal{W}_{\tau,\sigma}}e^{8\lambda\left(\left(\frac{1-\rho}{2\mu_x}+\mathcal{Q}_x\right)\delta_x^2+\left(\frac{1-\rho}{\mu_y}+\mathcal{Q}_y\right)\delta_y^2\right)}\\
&\leqslant e^{\frac{\lambda}{2(1-\rho)}\left(\frac{\mu_x}{8}\|A_1\|^2+\frac{\mu_y}{16}(\|A_2\|^2+\|A_3\|^2)\right)\mathcal{W}_{\tau,\sigma}}e^{8\lambda\left(\left(\frac{1-\rho}{2\mu_x}+\mathcal{Q}_x\right)\delta_x^2+\left(\frac{1-\rho}{\mu_y}+\mathcal{Q}_y\right)\delta_y^2\right)},
\end{aligned}$$

where the first inequality follows from Lemma 5, in the second inequality we used Lemma 27 given above, and the relation $\mathcal{E}_0 + \mathcal{E}_{-1} = 2\mathcal{E}_0{=}2\mathcal{W}_0/\rho \leqslant 2\mathcal{W}_{\tau,\sigma}/\rho$, which follows from $1-\alpha\sigma < 1$ and the relations $x_0 = x_{-1}$, $y_0 = y_{-1}$. Hence, we can apply Proposition 21 to the V_n, R_n, T_n sequence defined in (5.20) with the following choice of parameter values,

$$\begin{aligned}
&C_0 = \mathcal{W}_{\tau,\sigma}, \quad \mathcal{U} = P_0^{(1)} + P_0^{(2)} + \mathcal{Q}_x\|\Delta_0^x\|^2 + \mathcal{Q}_y\|\Delta_0^y\|^2,\\
&\sigma_T^2 = \|A_1\|^2\delta_x^2 + \left(\|A_2\|^2 + \|A_3\|^2\right)\delta_y^2, \quad \sigma_R^2 = 8\,(\mathcal{Q}_x\delta_x^2 + \mathcal{Q}_y\delta_y^2),\\
&\alpha = \frac{1}{16(1-\rho)}\left(\mu_x\|A_1\|^2 + \frac{\mu_y}{2}(\|A_2\|^2 + \|A_3\|^2)\right)\mathcal{W}_{\tau,\sigma}\\
&\qquad\qquad + \left(4\frac{(1-\rho)}{\mu_x} + 8\mathcal{Q}_x\right)\delta_x^2 + \left(8\frac{(1-\rho)}{\mu_y} + 8\mathcal{Q}_y\right)\delta_y^2,\\
&\bar{\alpha} = 4\max\left(\left(\frac{1-\rho}{2\mu_x} + \mathcal{Q}_x\right)\delta_x^2, \left(\frac{(1-\rho)}{\mu_y} + \mathcal{Q}_y\right)\delta_y^2\right), \quad \beta = 0,
\end{aligned} \tag{5.25}$$

where $\mathcal{W}_{\tau,\sigma}$ is defined in the statement of Theorem 8. When we invoke Proposition 21, we set $\lambda = \tilde{\gamma}$ within (5.4) for some particular $\tilde{\gamma} > 0$ such that $\tilde{\gamma} \leqslant \gamma$ as required by the proposition. Thus, for any $p \in (0,1)$ and $n \geqslant 0$, the following inequality

$$\begin{aligned}
V_n \leqslant \left(\frac{1+\rho}{2}\right)^n\Bigg[&\left(1 + \frac{1}{16(1-\rho)}\left(\mu_x\|A_1\|^2 + \frac{\mu_y}{2}(\|A_2\|^2 + \|A_3\|^2)\right)\right)\mathcal{W}_{\tau,\sigma}\\
&+ \left(\frac{4(1-\rho)}{\mu_x} + 8\mathcal{Q}_x\right)\delta_x^2 + \left(8\frac{(1-\rho)}{\mu_y}{+}8\mathcal{Q}_y\right)\delta_y^2\Bigg] + \frac{16}{1-\rho}\left(\mathcal{Q}_x\delta_x^2 + \mathcal{Q}_y\delta_y^2\right) + \frac{1}{\tilde{\gamma}}\log\left(\frac{1}{1-p}\right),
\end{aligned}$$

holds with probability at least p, with the choice of

$$\tilde{\gamma} \triangleq \frac{1-\rho}{\gamma_x \delta_x^2 + \gamma_y \delta_x^2} \leqslant \gamma \triangleq \frac{1-\rho}{\max\{\bar{\alpha}, 2\sigma_R^2, 4\sigma_T^2\}}, \tag{5.26}$$

where

$$\gamma_x = 2\frac{(1-\rho)}{\mu_x} + 16\mathcal{Q}_x + 4\|A_1\|^2, \quad \gamma_y = 4\frac{(1-\rho)}{\mu_y} + 16\mathcal{Q}_y + 4(\|A_2\|^2 + \|A_3\|^2). \tag{5.27}$$

In view of (5.21), and noting that $\mathcal{W}_n = \rho\mathcal{E}_n$, we obtain $\mathcal{W}_{n+1} + \mathcal{W}_n \leqslant 2(1+\rho)V_n \leqslant 4V_n$. Therefore,

$$\begin{gathered}\mathcal{C}_{\tau,\sigma,\theta} = \left(4 + \frac{1}{4(1-\rho)}\left(\mu_x\|A_1\|^2 + \frac{\mu_y}{2}(\|A_2\|^2 + \|A_3\|^2)\right)\right), \\ \Xi^{(x,1)}_{\tau,\sigma,\theta} = 16\frac{(1-\rho)}{\mu_x} + 32\mathcal{Q}_x, \qquad \Xi^{(y,1)}_{\tau,\sigma,\theta} = 32\frac{(1-\rho)}{\mu_y} + 32\mathcal{Q}_y, \\ \Xi^{(x,2)}_{\tau,\sigma,\theta} = \frac{64\mathcal{Q}_x}{(1-\rho)}, \quad \Xi^{(x,3)}_{\tau,\sigma,\theta} = \frac{4\gamma_x}{1-\rho}, \qquad \Xi^{(y,2)}_{\tau,\sigma,\theta} = \frac{64\mathcal{Q}_y}{(1-\rho)}, \quad \Xi^{(y,3)}_{\tau,\sigma,\theta} = \frac{4\gamma_y}{1-\rho},\end{gathered} \tag{5.28}$$

completes the proof of (3.1). The remaining items to prove regarding the asymptotic properties of $\Xi^{(1)}_{\tau,\sigma,\theta}$ and $\Xi^{(2)}_{\tau,\sigma,\theta}$ as $\theta \to 1$ follows from straightforward but tedious computations; for completeness, we provide the details in the online-only supplementary material (see Lemma 36 in Appendix G.1 of Laguel et al. (2023)).

5.2.2 Proof of Theorem 15

We can deduce Theorem 15 from the above analysis. Indeed, the CVaR bound in (3.5) directly follows from Corollary 22 applied to the process V_n introduced in (5.20), with the associated constants defined in (5.25). Furthermore, the EVaR bound in (3.6) follows from Corollary 23 applied to the same $(V_n)_{n\geqslant 0}$. Finally, the bound on $\mathcal{R}_{\chi^2,r}(\mathcal{W}_n^{1/2})$ follows from Corollary 24.

6. Numerical Results

In this section, we illustrate the robustness properties of SAPD when solving bilinear games and distributionally robust learning problems involving both synthetic and real data. First we consider the regularized bilinear game presented in (4.1),

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^d} \frac{\mu_x}{2}\|x\|^2 + x^\top K y - \frac{\mu_y}{2}\|y\|^2, \quad \text{for } K \triangleq 10\tilde{K}/\|\tilde{K}\|, \quad \tilde{K} \triangleq (M + M^\top)/2,$$

where $M = (M_{i,j})$ is a 30×30 matrix with entries sampled from i.i.d standard normal variables. We set the regularization variables as $\mu_x = \mu_y = 1$. We explore two values of the momentum parameter θ as $\bar{\theta}$ and $1 - (1-\bar{\theta})^2$, with $\bar{\theta} \triangleq (1+\kappa_{\max}^2)^{1/2} - 1$ computed based on the threshold value from Theorem 17. We then determine the stepsizes τ, σ according to the CP parameterization (2.3) where $\rho = \theta$. Finally, SAPD is initialized at a random tuple $(x_0, y_0) = 50(\tilde{x}_0, \tilde{y}_0)$, where $\tilde{x}_0, \tilde{y}_0 \in \mathbb{R}^{30}$ have entries sampled from i.i.d. standard normal distributions. In Figure 4, we report the histogram of the distance squared

$E_n = \|x_n - \mathrm{x}^\star\|^2 + \|y_n - \mathrm{y}^\star\|^2$ to the saddle point $\mathrm{z}^\star = \mathbf{0}$ after $n = 2000$ (top, middle panel) and $n = 5000$ iterations (top, right panel) based on 500 sample paths and for both choice of (momentum) parameter values. The expected distance $\mathbb{E}[E_n]$ over iterations is also reported on the top, left panel along with the error bars around it. The continuous vertical line in the convergence plots represents the sample average (estimating the expectation $\mathbb{E}[E_n]$), while the dashed vertical line represents the p-quantile of E_n for $p = 0.90$, i.e., the 90^{th} percentile of the error E_n. We observe that the performance is sensitive to the choice of parameters and there are bias/risk trade-offs in the choice of parameters; indeed, when the number of steps is smaller (for $n = 2000$), the noise accumulation is not dominant and a smaller rate parameter $\rho = \theta$ allows faster decay of the initialization bias, resulting in better guarantees for the value at risk with $p = 0.90$ or equivalently for the 90-th quantile. On the other hand when the number of steps is larger (for $n = 5000$), there is more risk associated to accumulation of noise and a larger choice of $\rho = \theta$ close to 1 is preferable, as this results in smaller primal and dual stepsizes which allows to control the tail risk at the expense of a slower decay of the initialization bias.

Next, we aim to solve the following distributionally robust logistic regression problem introduced in (Zhang et al., 2024): $\min_{x\in\mathbb{R}^\mathrm{d}}\max_{y\in\mathcal{P}_r} \frac{\mu_x}{2}\|x\|^2 + \sum_{i=1}^m y_i\phi_i(x) - \frac{\mu_y}{2}\|y\|^2$, where $\phi_i(x) \triangleq \log(1+\exp(-b_i a_i^\top x))$, and $\mathcal{P}_r \triangleq \{y \in \mathbb{R}_m^+ : \mathbf{1}^\top y = 1, \|y - \mathbf{1}/m\|^2 \leqslant \frac{r}{m^2}\}$, with $r = 2\sqrt{m}$. We consider two datasets from the UCI Repository[6], `DryBean`, and `Arcene`, and follow the preprocessing protocol outlined in (Zhang et al., 2024). For each dataset, we run `SAPD` with two values θ_1, θ_2 that are greater than the threshold value $\bar{\theta}$ given in (Zhang et al., 2024, Corollary 1). `SAPD` is initialized for both datasets at $x_0 = [2, \dots, 2]$ and $y_0 = \mathbf{1}/m$. In the middle and bottom panels of Figure 4, we display the average of the error E_n over the course of the iterations as well as the error histogram for `SAPD` over 500 runs as we did in the previous experiment.

Our numerical findings are similar to the bilinear case, i.e., to obtain the best risk guarantees, one needs to choose the algorithm parameters in a careful fashion –which is inline with our theoretical results, where obtaining the accelerated iteration complexity in Corollary 12 requires choosing the parameters in an optimized fashion over the class of admissible CP parameters. However, our parameter choice in Corollary 12 optimizes the complexity bounds in the worst-case, i.e., these bounds apply to any SCSC problem; moreover, these worst-case bounds involve some universal $\mathcal{O}(1)$ constants that are not fully optimized in our complexity results. In practice choosing θ specific to the problem at hand is beneficial. Indeed, a practical alternative to our θ choice in Corollary 12 would be to implement a grid search on θ and to set τ and σ according to the Chambolle-Pock parameterization in (2.3) for each θ choice in the grid.

7. Conclusion

We consider a first-order primal-dual method that relies on stochastic estimates of the gradients for solving SCSC saddle point problems. We focused on the stochastic accelerated primal dual (`SAPD`) method Zhang et al. (2024). We obtained high-probability bounds for the iterates to lie in a given neighborhood of the saddle point that reflects accelerated behavior. For a class of quadratic SCSC problems subject to i.i.d. isotropic Gaussian noise

[6] `https://archive.ics.uci.edu/ml/index.php`

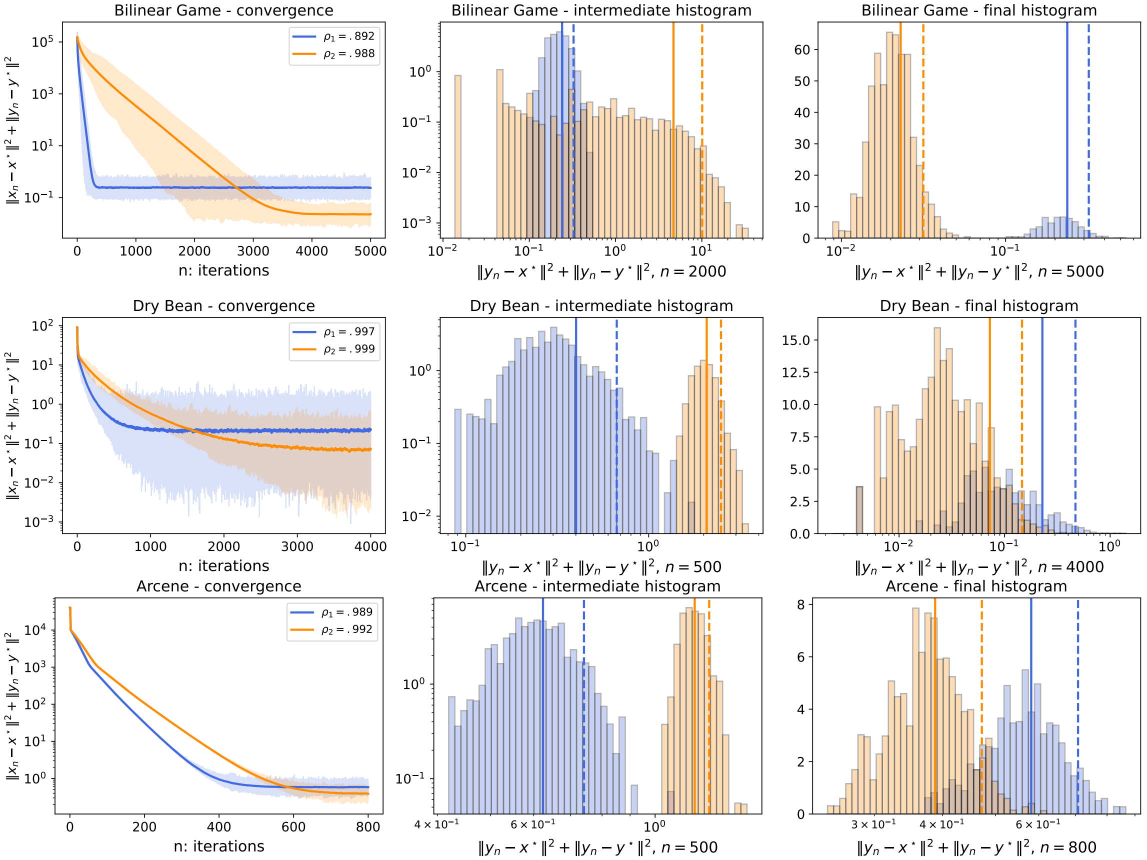

Figure 4: The figure shows the convergence behavior and distribution of performance scores for SAPD across three datasets. The left column displays the expected distance squared E_n of SAPD iterates to the solution over iterations, while the middle and right columns show histograms of E_n at fixed iterations. The continuous line in the convergence plots represents the average score $\mathcal{E}(E_n)$, while the dashed line represents the 90^{th} percentile. The datasets include a synthetically generated bilinear game, and Dry Bean and Arcene from the UCI repository.

and under a particular parameterization of the SAPD parameters, we were able to compute the distribution of the SAPD iterates exactly in closed form. We used this result to show that our high-probability bound is tight in terms of its dependency to target probability p, primal and dual stepsizes and the momentum parameter θ. We also provide a risk-averse convergence analysis characterizing the "Conditional Value at Risk", χ^2-divergence and the "Entropic Value at Risk" of the distance to the saddle point, highlighting the trade-offs between the bias and the risk associated with an approximate solution. In a follow-up to this work, Laguel et al. (2024) demonstrates that the concentration inequality-based techniques developed here can be applied to achieve high-probability guarantees for stochastic non-convex minimax problems. This highlights the potential of our techniques beyond convex saddle-point problems.

For light-tailed gradient noise, under the norm-subGaussian assumption, our results show that all the risk measures we considered behave similarly in the sense that they admit similar iteration complexity bounds. However, when the gradient noise has heavier tails beyond the subGaussian regime, we suspect that these measures can exhibit significantly different behaviors, and we leave investigating this as future work. In the future, we also plan to consider the extension of our results to the "online" setting, where the coupling function can be time-varying instead of being fixed.

Acknowledgements

Yassine Laguel and Mert Gürbüzbalaban acknowledge support from the grants Office of Naval Research N00014-21-1-2244 and N00014-24-1-2628, National Science Foundation (NSF) CCF-1814888, NSF DMS1723085, NSF DMS-2053485. Necdet Serhat Aybat's work was supported in part by the grant Office of Naval Research Award N00014-21-1-2271 and N00014-24-1-2666.

References

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

Mathieu Barré, Adrien Taylor, and Alexandre d'Aspremont. Complexity guarantees for polyak steps with momentum. In *Conference on Learning Theory*, pages 452–478. PMLR, 2020.

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.

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

Aleksandr Beznosikov, Boris Polyak, Eduard Gorbunov, Dmitry Kovalev, and Alexander Gasnikov. Smooth monotone stochastic variational inequalities and saddle point problems–survey. *arXiv preprint arXiv:2208.13592*, 2022.

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

Bugra Can and Mert Gürbüzbalaban. Entropic risk-averse generalized momentum methods. *arXiv preprint arXiv:2204.11292*, 2022.

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.

Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1):253–287, 2016.

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

Emilie Chouzenoux, Henri Gérard, and Jean-Christophe Pesquet. General risk measures for robust machine learning. *Foundations of Data Science*, 1(3):249–269, 2019.

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.

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

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.

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.

Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

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.

Eduard Gorbunov, Marina Danilova, David Dobre, Pavel Dvurechensky, Alexander Gasnikov, and Gauthier Gidel. Clipped stochastic methods for variational inequalities with heavy-tailed noise. *arXiv preprint arXiv:2206.01095*, 2022.

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.

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.

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

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.

Babak Hassibi, Ali H Sayed, and Thomas Kailath. *Indefinite-Quadratic estimation and control: a unified approach to H 2 and H∞ theories*. SIAM, 1999.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Tadeusz Inglot. Inequalities for quantiles of the chi-square distribution. *Probability and Mathematical Statistics*, 30(2):339–351, 2010.

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.

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

Yassine Laguel, Necdet Serhat Aybat, and Mert Gürbüzbalaban. High probability and risk-averse guarantees for a stochastic accelerated primal-dual method. *arXiv preprint arXiv:2304.00444*, 2023.

Yassine Laguel, Yasa Syed, Necdet Aybat, and Mert Gurbuzbalaban. High-probability complexity bounds for stochastic non-convex minimax optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems, To Appear*, 2024. URL `https://openreview.net/forum?id=XMQTNzlgTJ`.

P Gahinet A Laub, Ch Kenney, and G Hewer. Sensitivity of the stable discrete-time Lyapunov equation. *IEEE Trans. Automat. Control*, 35:1209–1217, 1990.

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.

Chengchang Liu, Shuxian Bi, Luo Luo, and John CS Lui. Partial-Quasi-Newton methods: Efficient algorithms for minimax optimization problems with unbalanced dimensionality. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1031–1041, 2022.

Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. *arXiv preprint arXiv:1910.09529*, 2019.

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.

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.

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

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

Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014.

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.

R Tyrrell Rockafellar and Johannes O Royset. Superquantiles and their applications to risk, random variables, and regression. In *Theory Driven by Influential Applications*, pages 151–167. Informs, 2013.

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

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.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Killian Wood and Emiliano Dall'Anese. Online Saddle Point Tracking with Decision-Dependent Data. *arXiv e-prints*, art. arXiv:2212.02693, December 2022. doi: 10.48550/arXiv.2212.02693.

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. URL `https://proceedings.neurips.cc/paper/2020/file/3f8b2a81da929223ae025fcec26dde0d-Paper.pdf`.

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.

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. doi: 10.1007/s10107-021-01660-z.

Xuan Zhang, Necdet Serhat Aybat, and Mert Gürbüzbalaban. Robust accelerated primal-dual methods for computing saddle points. *SIAM Journal on Optimization*, 34(1):1097–1130, 2024.

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.

Appendix A. Index of Notations

Category	**Notation**	**Reference**
Problem	$\mathcal{L}$, Φ, f, g, $\mathcal{X}$, $\mathcal{Y}$	Eqn. (1.1)
	μ_x, μ_y, $L_{xx}, L_{xy}, L_{yx}, L_{yy}$	Assumption 1
	$\tilde{\nabla}_x \Phi$, $\tilde{\nabla}_y \Phi$	Above Assumption 2
	δ_x, δ_y	Assumption 3
Algorithm	$x_n, y_n, z_n, \tau, \sigma, \theta$	Algorithm 1
	ρ, α, c	Eqn. (2.2)
Evaluation Metrics	$\mathcal{G}$, $\mathcal{D}$	Eqn. (1.2)
	$\mathcal{W}_{\tau,\sigma}$, $\mathcal{W}_n, \mathcal{E}_n$	Theorem 8 - Above Eqn. (5.12)
	Q_p, CVaR, EVaR, $\mathcal{R}_{\chi^2,r}$	Eqn. (2.8) - Table 2
Auxiliary iterates	Δ_n^x, Δ_n^y	Assumption 2
	$V_n, T_n, R_n, \sigma_T, \sigma_R$	Prop. 20 - Eqn. (5.20), (5.25)
	$C_0, \mathcal{U}, \alpha, \bar{\alpha}, \beta, \bar{\gamma}$	Prop. 21 - Eqn. (5.25), (5.26)
	$P_k^{(i)}, Q_k$	Eqn. (5.17)
Convergence rate-related	$(\Xi^{(i)}_{\tau,\sigma,\theta})_{1 \leqslant i \leqslant 3}$	Theorem 8, Eqn. (5.28)
	$\mathcal{Q}_x$, $\mathcal{Q}_y$	Lemma 26, Table 4
	A_1, A_2, A_3	Lemma 27, Table 5
Quadratic case	z_∞	Above Eqn. (4.6)
	$K, d, \omega_n^{x,y}, \delta$	Eqn. 4.1 and below
	$\tilde{\Sigma}^{\infty,\lambda}$	Eqn. (4.2)
	$\Sigma^\infty, \Sigma^{\infty,\Lambda}, V, \Sigma^{\infty,\lambda}$	Theorem 17
	$P_{(1,1)}$, $P_{(1,2)}$, $P_{(2,2)}, P_c$	Theorem 17, Table 7

Table 3: Key notations and references to where they are defined in the text.

Appendix B. Elementary proofs for subGaussians and convex risk measures

We provide in this section proofs of elementary properties of subGaussian vectors and convex risk measures.

B.1 Elementary Properties of Norm-subGaussian Vectors

In this section, we provide elementary proof of Lemma 5 and Lemma 6. The proofs follow from standard arguments that can be found in textbooks such as [8, 6].

B.1.1 Proof of Lemma 5

We follow standard arguments from (Vershynin, 2018). First note that, for any $k > 0$, we have

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

where Γ denotes the gamma function. Hence, noting that $\Gamma(k) = (k-1)!$, by the monotone convergence theorem,

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

the last equality being valid for any $\lambda \in [0, \frac{1}{2\sigma^2})$. Since for any $u \in [0, \frac{1}{2}]$, $\frac{1}{1-u} \leqslant e^{2u}$, we obtain $\mathbb{E}\left[e^{\lambda\|X\|^2}\right] \leqslant 2e^{4\lambda\sigma^2} - 1$ for any $\lambda \in [0, \frac{1}{4\sigma^2}]$. Finally, last inequality follows from $2e^{2u} - 1 \leqslant e^{4u}$, where we chose $u = 2\lambda\sigma^2$. □

B.1.2 Proof of Lemma 6

For $u = 0$, the inequality to prove is trivial. Assume $u \neq 0$. From Lemma 5 and Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[e^{\lambda^2\langle u,X\rangle^2}\right] \leqslant \mathbb{E}[e^{\lambda^2\|u\|^2\|X\|^2}] \leqslant e^{8\lambda^2\|u\|^2\sigma^2}, \tag{B.1}$$

for all $\lambda \in [0, \frac{1}{2\sqrt{2}\sigma\|u\|}]$. Thus, for any such λ, noticing that $e^t \leqslant t + e^{t^2}$ for $t \in \mathbb{R}$, we obtain $\mathbb{E}\left[e^{\lambda\langle u,X\rangle}\right] \leqslant \mathbb{E}\left[\lambda\langle u, X\rangle + e^{\lambda^2\langle u,X\rangle^2}\right] \leqslant e^{8\lambda^2\|u\|^2\sigma^2}$, where the second inequality follows from (B.1) and the assumption that $\mathbb{E}[X] = 0$. Moreover, for $\lambda \geqslant \frac{1}{2\sqrt{2}\|u\|\sigma}$, we have by Cauchy Schwarz's inequality and Lemma 5 that $\mathbb{E}[e^{\lambda\langle u,X\rangle}] \leqslant \mathbb{E}\left[e^{\frac{8\lambda^2\sigma^2\|u\|^2}{2}+\frac{\|X\|^2}{16\sigma^2}}\right] \leqslant e^{\frac{1}{2}\left(1+8\lambda^2\sigma^2\|u\|^2\right)} \leqslant e^{8\lambda^2\|u\|^2\sigma^2}$, where the last inequality is due to $e^{\frac{1+t}{2}} \leqslant e^t$ for $t \geqslant 1$. □

B.2 Elementary properties of Convex Risk Measures

The following lemma is used in the derivation of CVaR and EVaR bounds.

Lemma 28 *For any non-negative random variable $U\colon \Omega \to \mathbb{R}_+$, we have for all $p \in [0, 1)$:*

$$Q_p(U^2)^{\frac{1}{2}} = Q_p(U), \qquad \mathrm{CVaR}_p(U^2)^{\frac{1}{2}} \geqslant \mathrm{CVaR}_p(U).$$

Proof We first show that $Q_p(X^2) = Q_p(X)^2$ for any $p \in (0, 1)$. Indeed, for any $0 \leqslant t < Q_p(U)^2$, we have $\mathbb{P}[U^2 \leqslant t] = \mathbb{P}[U \leqslant \sqrt{t}] < p$ which follows from non-negativity of U and definition of $Q_p(U)$. This implies $t < Q_p(U^2)$; thus, $Q_p(U)^2 \leqslant Q_p(U^2)$. Conversely, we note that $p \leqslant \mathbb{P}[U \leqslant Q_p(U)] = \mathbb{P}[U^2 \leqslant Q_p(U)^2]$, which implies $Q_p(U)^2 \geqslant Q_p(U^2)$; hence, $Q_p(X^2) = Q_p(X)^2$. Using this result,

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

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

Appendix C. Intermediate results and proofs for the non-quadratic case

To start with, for the sake of completeness, we cite two results from (Zhang et al., 2024). The first lemma is used to derive the almost sure bound result of Proposition 25, which is provided below in Appendix C.1, while the second lemma is used for deriving the convex inequalities provided in Appendix C.2.

Lemma 29 (See (Zhang et al., 2024, Lemma 1)) *The iterates (x_n, y_n) of SAPD satisfy*

$$\begin{aligned}\mathcal{L}(x_{n+1}, \mathrm{y}^\star) - \mathcal{L}(\mathrm{x}^\star, y_{n+1}) \leqslant &- \langle q_{n+1}, y_{n+1} - \mathrm{y}^\star\rangle + \theta\langle q_n, y_n - \mathrm{y}^\star\rangle + \Lambda_n - \Sigma_{n+1} + \Gamma_{n+1} \\ &+ \langle \Delta_n^x, \mathrm{x}^\star - x_{n+1}\rangle + \langle (1+\theta)\Delta_n^y - \theta\Delta_{n-1}^y, y_{n+1} - \mathrm{y}^\star\rangle,\end{aligned}$$

for all $n \geqslant 0$, where

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

Lemma 30 (See (Zhang et al., 2024, Lemma 3)) *Let $(x_n, y_n)_{n\geqslant 0}$ denote the SAPD iterate sequence. Then, the following inequalities hold for all $n \in \mathbb{N}$,*

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

C.1 Proof of Proposition 25 (Almost sure domination of SAPD iterates)

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 $K_n(\rho) \triangleq \sum_{k=0}^{n-1}\rho^{-k} = \frac{1}{\rho^{n-1}} \times \frac{1-\rho^n}{1-\rho}$, by Jensen's inequality , we have for all $\rho \in (0,1]$,

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

Hence, in view of Lemma 29,

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

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

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

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

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

From (C.1), it follows that

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

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 $n \geqslant 1$,

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

where $A, B \in \mathbb{R}^{5\times5}$ and $\xi_k \in \mathbb{R}^5$ are defined for $k \geqslant 0$ as

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

and $\xi_k \triangleq \left(\ \|x_k - \mathrm{x}^\star\|, \ \ \|y_k - \mathrm{y}^\star\|. \ \ \|x_k - x_{k-1}\|, \ \ \|y_k - y_{k-1}\|, \ \ \|y_{k+1} - y_k\| \ \right)^\top \in \mathbb{R}^5$. By (Zhang et al., 2024, Lemma 5), the matrix inequality condition (2.2) is equivalent to having $B - \rho^{-1}A \succeq 0$. In this case, we almost surely have

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

Finally, denoting

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

we have $G'' \succeq 0$ in view of (Zhang et al., 2024, Lemma 6); thus,

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

Therefore, using (C.2), we can conclude that $U_n \leqslant \frac{1}{2}\xi_0^\top A\xi_0 \leqslant \frac{1}{2\tau}\|x_0 - \mathrm{x}^\star\|^2 + \frac{1}{2\sigma}\|y_0 - \mathrm{y}^\star\|^2 = \mathcal{W}_{\tau,\sigma}$. Finally, by non-negativity of $\mathcal{L}(\bar{x}_n, \mathrm{y}^\star) - \mathcal{L}(\mathrm{x}^\star, \bar{y}_n)$, we obtain (5.15). □

C.2 Convex inequalities

C.2.1 Proof of Lemma 26

We first start with a technical result we will use in the proof of Lemma 26.

Lemma 31 *For any $n \geqslant 1$,*

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

where $\hat{\hat{\mathrm{y}}}_{n+1}, \hat{\mathrm{y}}_n, \hat{\hat{\mathrm{x}}}_n$ are defined in (5.12)*, and $A_0 \in \mathbb{R}^4$ is defined as*

$$A_0 \triangleq \frac{1}{1+\sigma\mu_y}\begin{bmatrix} \sqrt{2\rho\tau}\sigma(1+\theta)\,\mathrm{L}_{\mathrm{yx}} \\ \frac{\sqrt{2\rho\sigma}}{\sqrt{1-\alpha\sigma}}(1+\sigma(1+\theta)\,\mathrm{L}_{\mathrm{yy}}) \\ \sqrt{2\rho\tau}\cdot\sigma\theta\,\mathrm{L}_{\mathrm{yx}} \\ \frac{\sqrt{2\rho\sigma}}{\sqrt{(1-\alpha\sigma)}}\cdot\sigma\theta\,\mathrm{L}_{\mathrm{yy}} \end{bmatrix} \in \mathbb{R}^4.$$

Proof Since $(\mathrm{x}^\star, \mathrm{y}^\star)$ is a solution of (1.1), $\mathrm{x}^\star$ and $\mathrm{y}^\star$ are fixed points of two deterministic proximal gradient maps, i.e.,

$$\mathrm{x}^\star = \mathrm{prox}_{\tau f}\left(\mathrm{x}^\star - \tau\nabla_{\mathrm{x}}\Phi(\mathrm{x}^\star, \mathrm{y}^\star)\right), \quad \mathrm{y}^\star = \mathrm{prox}_{\sigma g}\left(\mathrm{y}^\star + \sigma\nabla_{\mathrm{y}}\Phi(\mathrm{x}^\star, \mathrm{y}^\star)\right). \tag{C.3}$$

Thus, by the contraction properties of the prox for strongly convex functions, and convexity of the squared norm, we have

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

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

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

The statement finally follows from Cauchy-Schwarz inequality. ■

Now we are ready to prove Lemma 26. By Young's inequality, for any $\gamma_x, \gamma_y > 0$,

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

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

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

where $\mathcal{E}_n = \mathcal{W}_n/\rho$ and $\mathcal{W}_n$ is defined in (2.4). Moreover, we also have $\frac{-\rho}{1+\rho} P_{n-1}^{(2)} + \frac{\theta}{1+\rho} \langle \Delta_{n-1}^y, y_{n+1} - \hat{\hat{\mathrm{y}}}_{n+1} \rangle = \frac{\theta}{1+\rho} \langle \Delta_{n-1}^y, y_{n+1} - \mathrm{y}^\star \rangle \leqslant \frac{\theta}{1+\rho} \Big(\frac{\gamma_y'}{2} \|\Delta_{n-1}^y\|^2 + \frac{1}{2\gamma_y'} \| \mathrm{y}^\star - y_{n+1} \|^2 \Big)$ for any $\gamma_y' > 0$. Hence, setting $\gamma_y' = \frac{8\theta\sigma}{1-\alpha\sigma}$ leads to

$$\frac{-\rho}{1+\rho} P_{n-1}^{(2)} + \frac{\theta}{1+\rho} \langle \Delta_{n-1}^y, y_{n+1} - \hat{\hat{\mathrm{y}}}_{n+1} \rangle \leqslant \frac{\rho}{8(1+\rho)} \mathcal{E}_{n+1} + \frac{4\sigma\theta^2}{(1+\rho)(1-\alpha\sigma)} \|\Delta_{n-1}^y\|^2. \tag{C.5}$$

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

$$\begin{aligned}
-P_n^{(2)} &\leqslant \frac{\theta\gamma}{2\rho} \|\Delta_n^y\|^2 + \frac{\theta}{2\gamma\rho} \| \hat{\hat{\mathrm{y}}}_{n+2} - \mathrm{y}^\star \|^2 \\
&\leqslant \frac{\theta\gamma}{2\rho} \|\Delta_n^y\|^2 + \frac{\theta}{2\gamma\rho} \Bigg(3\|A_0\|^2 (\mathcal{E}_{n+1} + \mathcal{E}_n) + 3\Big(\frac{1+\sigma(1+\theta)\,\mathrm{L}_{\mathrm{yy}}}{1+\sigma\mu_y} \Big)^2 \| \hat{\mathrm{y}}_{n+1} - y_{n+1} \|^2 \\
&\qquad\qquad\qquad\qquad\qquad\qquad + 3\Big(\frac{\sigma(1+\theta)\,\mathrm{L}_{\mathrm{yx}}}{1+\sigma\mu_y} \Big)^2 \| \hat{\hat{\mathrm{x}}}_{n+1} - x_{n+1} \|^2 \Bigg),
\end{aligned}$$

where the last inequality follows from Lemma 31 and the simple inequality $(a+b+c)^2 \leqslant 3a^2+3b^2+3c^2$ for any $a,b,c \in \mathbb{R}$. Setting $\gamma \triangleq \frac{6\theta\|A_0\|^2(1+\rho)}{\rho^2}$ ensures that

$$
\begin{aligned}
-P_n^{(2)} \leqslant \frac{\rho}{4(1+\rho)}(\mathcal{E}_{n+1}+\mathcal{E}_n) &+ \frac{3\theta^2\|A_0\|^2(1+\rho)}{\rho^3}\|\Delta_n^y\|^2 \\
&+ \frac{\rho}{4\|A_0\|^2(1+\rho)}\left(\frac{1+\sigma(1+\theta)\,\mathrm{L_{yy}}}{1+\sigma\mu_y}\right)^2\|\,\hat{\mathrm{y}}_{n+1}-y_{n+1}\|^2 \\
&+ \frac{\rho}{4\|A_0\|^2(1+\rho)}\left(\frac{\sigma(1+\theta)\,\mathrm{L_{yx}}}{1+\sigma\mu_y}\right)^2\|\,\hat{\hat{\mathrm{x}}}_{n+1}-x_{n+1}\|^2.
\end{aligned} \tag{C.6}
$$

Hence, using the trivial upper bound $\mathcal{E}_{n+1} \leqslant (\mathcal{E}_{n+1}+\mathcal{E}_n)$ and combining the bounds eqs. (C.4) to (C.6) we obtained above, we get

$$
\begin{aligned}
&\frac{-1}{1+\rho}P_n^{(1)} - P_n^{(2)} - \frac{\rho}{1+\rho}P_{n-1}^{(2)} - \frac{1}{1+\rho}Q_n \\
&\quad\leqslant \frac{\rho}{2(1+\rho)}(\mathcal{E}_{n+1}+\mathcal{E}_n) \\
&\qquad+ \frac{4\tau}{1+\rho}\|\Delta_n^x\|^2 + \frac{4\sigma(1+\theta)^2}{(1+\rho)(1-\alpha\sigma)}\|\Delta_n^y\|^2 + \frac{4\sigma\theta^2}{(1+\rho)(1-\alpha\sigma)}\|\Delta_{n-1}^y\|^2 + \frac{3\theta^2(1+\rho)\|A_0\|^2}{\rho^3}\|\Delta_n^y\|^2 \\
&\qquad+ \frac{\rho}{4\|A_0\|^2(1+\rho)}\left(\frac{1+\sigma(1+\theta)\,\mathrm{L_{yy}}}{1+\sigma\mu_y}\right)^2\|\,\hat{\mathrm{y}}_{n+1}-y_{n+1}\|^2 \\
&\qquad+ \frac{\rho}{4\|A_0\|^2(1+\rho)}\left(\frac{\sigma(1+\theta)\,\mathrm{L_{yx}}}{1+\sigma\mu_y}\right)^2\|\,\hat{\hat{\mathrm{x}}}_{n+1}-x_{n+1}\|^2.
\end{aligned}
$$

Let us now introduce $\zeta_n \triangleq \left[\|\Delta_n^x\|,\ \|\Delta_n^y\|,\ \rho^{1/2}\|\Delta_{n-1}^y\|\right]^\top \in \mathbb{R}^3$ for $n \geqslant 0$; then, by similar computations the following bounds follow from Lemma 30:

$$
\begin{aligned}
\|\,\hat{\mathrm{y}}_{n+1}-y_{n+1}\|^2 &\leqslant \zeta_n^\top \operatorname{Diag}\left[0, \frac{2\sigma^2(1+\theta)^2}{(1+\sigma\mu_y)^2}, \frac{2\sigma^2\theta^2\rho^{-1}}{(1+\sigma\mu_y)^2}\right]\zeta_n \\
\|\,\hat{\hat{\mathrm{x}}}_{n+1}-x_{n+1}\|^2 &\leqslant \zeta_n^\top \operatorname{Diag}\left[\frac{3\tau^2}{(1+\tau\mu_x)^2}, \frac{3\tau^2\sigma^2(1+\theta)^2\,\mathrm{L_{xy}}^2}{(1+\tau\mu_x)^2(1+\sigma\mu_y)^2}, \frac{3\tau^2\sigma^2\theta^2\rho^{-1}\,\mathrm{L_{xy}}^2}{(1+\tau\mu_x)^2(1+\sigma\mu_y)^2}\right]\zeta_n,
\end{aligned}
$$

and we deduce that

$$
\frac{-1}{1+\rho}P_n^{(1)} - P_n^{(2)} - \frac{\rho}{1+\rho}P_{n-1}^{(2)} - \frac{1}{1+\rho}Q_n \leqslant \frac{\rho}{2(1+\rho)}(\mathcal{E}_{n+1}+\mathcal{E}_n) + \zeta_n^\top \operatorname{Diag}\left[B^x, B^y, B_{-1}^y\right]\zeta_n,
$$

where B^x, B^y, B_{-1}^y are constants specified in Table 4 of Appendix E.

We now treat the sum $\sum_{k=0}^{n} \rho^{n-k} Q_k$. Observe first that for all $n \in \mathbb{N}$, Lemma 30,

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

which, after organizing the terms and using $ab \leqslant a^2/2 + b^2/2$ for any scalars a, b, becomes

$$\begin{aligned}
Q_n &\leqslant \frac{\tau}{1+\tau\mu_x}\|\Delta_n^x\|^2 + \frac{\sigma(1+\theta)^2}{1+\sigma\mu_y}\|\Delta_n^y\|^2 \qquad . \\
&\quad + \frac{\sigma\theta}{1+\sigma\mu_{\mathbf{y}}} \left(\theta + (1+\theta) \left(\frac{1+\sigma(1+\theta)\mathrm{L_{yy}}}{1+\sigma\mu_y} + \frac{\tau\sigma(1+\theta)\mathrm{L_{yx}}\mathrm{L_{xy}}}{(1+\tau\mu_x)\,(1+\sigma\mu_y)} \right) \right) \|\Delta_{n-1}^y\|^2 \\
&\quad + \frac{\sigma\theta}{2(1+\sigma\mu_y)} \quad \frac{\tau(1+\theta)\,\mathrm{L_{yx}}}{1+\tau\mu x}\|\Delta_{n-1}^y\|^2 + \frac{\sigma\theta}{2(1+\sigma\mu_y)} \frac{\tau(1+\theta)\,\mathrm{L_{yx}}}{1+\tau\mu_x}\|\Delta_{n-1}^x\|^2 \\
&\quad + \frac{\sigma\theta}{1+\sigma\mu y}(1+\theta)\|\Delta_n^y\|^2 + \frac{\sigma\theta}{1+\sigma\mu_y}(1+\theta)\|\Delta_{n-1}^y\|^2 \\
&\quad + \frac{\sigma\theta}{1+\sigma\mu_y}\frac{\theta}{2} \left(\frac{1+\sigma(1+\theta)\mathrm{L_{yy}}}{1+\sigma\mu_y} + \frac{\tau\sigma(1+\theta)\mathrm{L_{yx}}\mathrm{L_{xy}}}{(1+\tau\mu_x)\,(1+\sigma\mu_y)} \right) \|\Delta_{n-1}^y\|^2 \\
&\qquad + \frac{\sigma\theta}{1+\sigma\mu_y}\frac{\theta}{2} \left(\frac{1+\sigma(1+\theta)\mathrm{L_{yy}}}{1+\sigma\mu_y} + \frac{\tau\sigma(1+\theta)\mathrm{L_{yx}}\mathrm{L_{xy}}}{(1+\tau\mu_x)\,(1+\sigma\mu_y)} \right) \|\Delta_{n-2}^y\|^2 \\
&\quad + \frac{\tau\sigma(1+\theta)\,\mathrm{L_{xy}}}{2\,(1+\tau\mu_x)\,(1+\tau\mu_y)}\|\Delta_n^x\|^2 + \frac{\tau\sigma(1+\theta)\,\mathrm{L_{xy}}}{2\,(1+\tau\mu_x)\,(1+\tau\mu_y)}\|\Delta_n^y\|^2 \\
&\quad + \frac{\tau\sigma\theta\,\mathrm{L_{xy}}}{2\,(1+\tau\mu_x)\,(1+\tau\mu_y)}\|\Delta_n^x\|^2 + \frac{\tau\sigma\theta\,\mathrm{L_{xy}}}{2\,(1+\tau\mu_x)\,(1+\tau\mu_y)}\|\Delta_{n-1}^y\|^2
\end{aligned}$$

Thus, we obtain $Q_n \leqslant C^x \; \|\Delta_n^x\|^2 \; + \; C_{-1}^x \; \rho\|\Delta_{n-1}^x\|^2 \; + \; C^y \; \|\Delta_n^y\|^2 \; + \; C_{-1}^y \; \rho\|\Delta_{n-1}^y\|^2 \; + \; C_{-2}^y \; \rho^2\|\Delta_{n-2}^y\|^2$ for some constants $C^x, C_{-1}^x, C^y, C_{-1}^y, C_{-2}^y$ (that are explicitly given in Table 4 of Appendix E). Hence, setting $\Delta_{-1}^x = \Delta_{-1}^y = \Delta_{-2}^x = 0$, we obtain,

$$\begin{aligned}
\sum_{k=0}^{n} \rho^{n-k} Q_k - \frac{1}{1+\rho} P_n^{(1)} - P_n^{(2)} - \frac{\rho}{1+\rho} P_{n-1}^{(2)} - \frac{1}{1+\rho} Q_n& \\
\leqslant \frac{\rho}{2(1+\rho)} (\mathcal{E}_{n+1} + \mathcal{E}_n) + \sum_{k=0}^{n} \rho^{n-k} &(C^x\|\Delta_k^x\|^2 + C_{-1}^x \rho \|\Delta_{k-1}^x\|^2 \\
&+ C^y\|\Delta_k^y\|^2 + C_{-1}^y \rho\|\Delta_{k-1}^y\|^2 + C_{-2}^y \rho^2 \|\Delta_{k-2}^y\|^2) \\
&+ B^x\|\Delta_n^x\|^2 + B^y\|\Delta_n^y\|^2 + B_{-1}^y \rho \|\Delta_{n-1}^y\|^2;
\end{aligned}$$

therefore, rearranging the terms together we get

$$\begin{aligned}
&\sum_{k=0}^{n} \rho^{n-k} Q_k - \frac{1}{1+\rho} P_n^{(1)} - P_n^{(2)} - \frac{\rho}{1+\rho} P_{n-1}^{(2)} - \frac{1}{1+\rho} Q_n \\
&\qquad \leqslant \frac{\rho}{2(1+\rho)} (\mathcal{E}_{n+1} + \mathcal{E}_n) \\
&\qquad\quad + C^x \sum_{k=0}^{n} \rho^{n-k} \|\Delta_k^x\|^2 + C_{-1}^x \sum_{k=0}^{n-1} \rho^{n-k} \|\Delta_k^x\|^2 + C^y \sum_{k=0}^{n} \rho^{n-k} \|\Delta_k^y\|^2 \\
&\qquad\quad + C_{-1}^y \sum_{k=0}^{n-1} \rho^{n-k} \|\Delta_k^y\|^2 + C_{-2}^y \sum_{k=0}^{n-2} \rho^{n-k} \|\Delta_k^y\|^2 + B^x \|\Delta_n^x\|^2 + B^y \|\Delta_n^y\|^2 + B_{-1}^y \rho \|\Delta_{n-1}^y\|^2 \\
&\qquad \leqslant \frac{\rho}{2(1+\rho)} (\mathcal{E}_{n+1} + \mathcal{E}_n) + \mathcal{Q}_x \sum_{k=0}^{n} \rho^{n-k} \|\Delta_k^x\|^2 + \mathcal{Q}_y \sum_{k=0}^{n} \rho^{n-k} \|\Delta_k^y\|^2,
\end{aligned}$$

where $\mathcal{Q}_x \triangleq B^x + C^x + C_{-1}^x$ and $\mathcal{Q}_y \triangleq B^y + B_{-1}^y + C^y + C_{-1}^y + C_{-2}^y$. This completes the proof. □

C.2.2 Proof of Lemma 27

Let $n \in \mathbb{N}$ be fixed. In view of (C.3), we have

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

where the third inequality follows from the smoothness assumptions on $\nabla_{\mathrm{x}} \Phi$ and $\nabla_{\mathrm{y}} \Phi$, and for the $n = 0$ case, we have $x_{-1} = x_0$ and $y_{-1} = y_0$. Using similar arguments, we also obtain

$$\begin{aligned}
\|\hat{\hat{\mathrm{x}}}_{n+1} - \mathrm{x}^\star\| &\leqslant \frac{1}{1+\tau\mu_x} \Big((1 + \tau \mathrm{L}_{\mathrm{xx}}) \|x_n - \mathrm{x}^\star\| + \tau \mathrm{L}_{\mathrm{xy}} \|\hat{\mathrm{y}}_{n+1} - \mathrm{y}^\star\| \Big) \\
&\leqslant \frac{1}{1+\tau\mu_x} \Bigg(\left(1 + \tau \mathrm{L}_{\mathrm{xx}} + \frac{\tau\sigma(1+\theta) \mathrm{L}_{\mathrm{yx}} \mathrm{L}_{\mathrm{xy}}}{1+\sigma\mu_y} \right) \|x_n - \mathrm{x}^\star\| + \frac{\tau\sigma\theta \mathrm{L}_{\mathrm{xy}} \mathrm{L}_{\mathrm{yx}}}{1+\sigma\mu_y} \|x_{n-1} - \mathrm{x}^\star\| \\
&\qquad\qquad + \frac{\tau \mathrm{L}_{\mathrm{xy}} (1 + \sigma(1+\theta) \mathrm{L}_{\mathrm{yy}})}{1+\sigma\mu_y} \|y_n - \mathrm{y}^\star\| + \frac{\tau\sigma\theta \mathrm{L}_{\mathrm{xy}} \mathrm{L}_{\mathrm{yy}}}{1+\sigma\mu_y} \|y_{n-1} - \mathrm{y}^\star\| \Bigg),
\end{aligned}$$

from which we deduce the following bound:

$$
\begin{aligned}
\|\hat{\mathrm{y}}_{n+2}-\mathrm{y}^\star\| \leqslant{}& \frac{1}{1+\sigma\mu_y}\Big(\|\hat{\mathrm{y}}_{n+1}-\mathrm{y}^\star\| + \sigma(1+\theta)\,\mathrm{L_{yx}}\,\|\hat{\mathrm{x}}_{n+1}-\mathrm{x}^\star\| + \sigma(1+\theta)\,\mathrm{L_{yy}}\,\|\hat{\mathrm{y}}_{n+1}-\mathrm{y}^\star\| \\
&\qquad\qquad + \sigma\theta\,\mathrm{L_{yx}}\,\|x_n-\mathrm{x}^\star\| + \sigma\theta\,\mathrm{L_{yy}}\,\|y_n-\mathrm{y}^\star\|\Big) \\
\leqslant{}& \frac{1}{1+\sigma\mu_y}\Bigg(\Bigg((1+\sigma(1+\theta)\,\mathrm{L_{yy}})\frac{\sigma(1+\theta)\,\mathrm{L_{yx}}}{1+\sigma\mu_y} \\
&\qquad + \sigma(1+\theta)\,\mathrm{L_{yx}}\,\frac{\Big(1+\tau\,\mathrm{L_{xx}}+\frac{\tau\sigma(1+\theta)\,\mathrm{L_{yx}}\,\mathrm{L_{xy}}}{1+\sigma\mu_y}\Big)}{1+\tau\mu_x} + \sigma\theta\,\mathrm{L_{yx}}\Bigg)\|x_n-\mathrm{x}^\star\| \\
&\qquad + \left(\frac{(1+\sigma(1+\theta)\,\mathrm{L_{yy}})^2}{1+\sigma\mu_y} + \sigma(1+\theta)\,\mathrm{L_{yx}}\,\frac{\tau\,\mathrm{L_{xy}}(1+\sigma(1+\theta)\,\mathrm{L_{yy}})}{(1+\sigma\mu_y)(1+\tau\mu_x)} + \sigma\theta\,\mathrm{L_{yy}}\right)\|y_n-\mathrm{y}^\star\| \\
&\qquad + \left((1+\sigma(1+\theta)\,\mathrm{L_{yy}})\frac{\theta\sigma\,\mathrm{L_{yx}}}{(1+\sigma\mu_y)} + \sigma(1+\theta)\frac{\tau\sigma\theta\,\mathrm{L_{xy}}\,\mathrm{L_{yx}}^2}{(1+\sigma\mu_y)(1+\tau\mu_x)}\right)\|x_{n-1}-\mathrm{x}^\star\| \\
&\qquad + \left(\frac{(1+\sigma(1+\theta)\,\mathrm{L_{yy}})\sigma\theta\,\mathrm{L_{yy}}}{1+\sigma\mu_y} + \sigma(1+\theta)\frac{\tau\sigma\theta\,\mathrm{L_{xy}}\,\mathrm{L_{yx}}\,\mathrm{L_{yy}}}{(1+\sigma\mu_y)(1+\tau\mu_x)}\right)\|y_{n-1}-\mathrm{y}^\star\|\Bigg).
\end{aligned}
$$

Combining the above bounds with Cauchy-Schwarz inequality implies (5.22) and we conclude.

□

Appendix D. Details and proofs for the quadratic setting

D.1 Properties of SAPD on the quadratic SP problem given in (4.1)

In this section, we briefly recall the discussion in (Zhang et al., 2024) regarding the convergence behaviour of SAPD on the SP problem in (4.1). Precisely, denoting $\tilde{z}_n = [x_{n-1}, y_n]^\top$ and $\omega_n = \left[\omega^x_{n-1}\omega^y_{n-1};\omega^y_n\right]^\top$, the authors observe that $(\tilde{z}_n)_{n\geqslant 0}$ satisfies the recurrence relation $\tilde{z}_{n+1} = A\tilde{z}_n + B\omega_n$ where A and B are defined as

$$
A=\begin{bmatrix} \frac{1}{1+\tau\mu_x}I_d & \frac{-\tau}{(1+\tau\mu_x)}K \\ \frac{1}{1+\sigma\mu_y}\left(\frac{\sigma(1+\theta)}{1+\tau\mu_x}-\sigma\theta\right)K & \frac{1}{1+\sigma\mu_y}\left(I_d - \frac{\tau\sigma(1+\theta)}{1+\tau\mu_x}K^2\right) \end{bmatrix},\quad B=\begin{bmatrix} \frac{-\tau}{1+\tau\mu_x}I_d & 0_d & 0_d \\ \frac{-\tau\sigma(1+\theta)}{(1+\tau\mu_x)(1+\sigma\mu_y)}K & \frac{-\sigma\theta}{1+\sigma\mu_y}I_d & \frac{\sigma(1+\theta)}{1+\sigma\mu_y}I_d \end{bmatrix}. \tag{D.1}
$$

As a result, the covariance matrix $\tilde{\Sigma}_n$ of $\tilde{z}_n$ satisfies for all $n \geqslant 0$,

$$
\tilde{\Sigma}_{n+1} = A\tilde{\Sigma}_n A^\top + R, \tag{D.2}
$$

where $R = \frac{\delta^2}{d}BB^\top + A\mathbb{E}\left[\tilde{z}_n\omega_n^\top\right]B^\top + B\mathbb{E}\left[\omega_n\tilde{z}_n^\top\right]A^\top$. Using the independence assumptions on the ω^x_n's and ω^y_n's, elementary derivations lead to expressing R as

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

Provided that the spectral radius $\rho(A)$ of A is less than 1, the sequence $(\tilde{\Sigma}_n)_{n\geqslant 0}$ converges to a matrix $\tilde{\Sigma}^\infty$ satisfying

$$
\tilde{\Sigma}^\infty = A\tilde{\Sigma}^\infty A^\top + R. \tag{D.3}
$$

Leveraging the spectral theorem, it is shown in (Zhang et al., 2024) that an orthogonal change of basis enables to reduce the $2d \times 2d$ Lyapunov equation to d systems of the following form for each $\lambda \in \mathrm{Sp}(K)$:

$$\tilde{\Sigma}^{\infty,\lambda} = A^{\lambda}\tilde{\Sigma}^{\infty,\lambda}{A^{\lambda}}^{\top} + R^{\lambda}, \tag{D.4}$$

such that A^{λ} and R^{λ} are 2×2 matrices defined for each $\lambda \in \mathrm{Sp}(K)$ as

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

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

and A is similar to the matrix $\mathrm{Diag}(A^{\lambda_1}, \cdots, A^{\lambda_d})$. Therefore, we have $\rho(A) = \max_{i=1,2,\ldots,d}\rho(A^{\lambda_i})$.

D.2 Proof of Theorem 17

In this section, we solve the Lyapunov equations (D.4) analytically under the parameterization in (2.3). Throughout, given $\lambda \in \mathrm{Sp}(K)$, we introduce the quantity $\kappa_\lambda \triangleq \frac{\lambda}{\sqrt{\mu_x\mu_y}}$ which is closely related to the condition number $\kappa = \max\{L_{xx}, L_{yx}, L_{yy}\}/\min\{\mu_y, \mu_x\}$. Indeed, for $\mu_x = \mu_y$, we have $\kappa = \max\{|\kappa_\lambda| : \lambda \in \mathrm{Sp}(K)\}$. For each $\lambda \in \mathrm{Sp}(K)$, let $\tilde{\Sigma}^{\infty,\lambda}$ be a solution to (D.4), i.e., $\tilde{\Sigma}^{\infty,\lambda}$ solves the following 2×2 Lyapunov equation:

$$\tilde{\Sigma}^{\infty,\lambda} = A^{\lambda}\tilde{\Sigma}^{\infty,\lambda}{A^{\lambda}}^{\top} + R^{\lambda}. \tag{D.5}$$

Furthermore, such a solution is unique if $\rho(A^{\lambda}) < 1$ (Laub et al., 1990; Hassibi et al., 1999). The following result provides an explicit formula to $\tilde{\Sigma}^{\infty,\lambda}$ whose proof is deferred to Section D.5.

Proposition 32 *Under the Chambolle-Pock parameterization in* (2.3), *for* $\lambda \neq 0$ *and* $\theta \in \left(\frac{1}{\kappa_\lambda}\left(\sqrt{1 + (\kappa_\lambda)^2} - 1\right), 1\right)$, *we have* $\rho(A^{\lambda}) < 1$, *and the unique solution* $\Sigma^{\infty,\lambda}$ *to the equation in* (D.5) *is given by*

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

where P_c *and* $P_{q,\ell}^{(k)}$ *for* $q = 1, 2$ *and* $\ell = 1, 2$ *are polynomials in* θ *and* κ_λ, *defined in the top part of Table 7. Otherwise, for* $\lambda = 0$ *and* $\theta \in [0, 1)$, *we also have* $\rho(A^{\lambda}) < 1$ *and the unique solution* $\tilde{\Sigma}^{\infty,\lambda}$ *of* (D.5) *is given by*

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

Theorem 17 will then follow directly from Proposition 32.

Proof [Proof of Theorem 17] Proposition 32 characterizes the asymptotic covariance matrix of (x_{n-1}, y_n) in the limit as $n \to \infty$, which we will use to deduce the covariance matrix Σ^{∞} of (x_n, y_n) in the limit as $n \to \infty$. First, recall from (Zhang et al., 2024) that the orthogonal matrix leading to the reduced Lyapunov (D.4) is given by $Z = PV$ where P is the permutation matrix associated to the permutation $\mathcal{P}$ of $\{1, \dots, 2d\}$ defined as $\mathcal{P}(qd + r) = q + 2r - 1 \bmod [d]$, for all $q \in \{0, 1\}, r \in \{1, \dots, d\}$, and $V \triangleq \mathrm{Diag}(U, U) \in \mathbb{R}^{2d \times 2d}$ where U describes an orthogonal basis for K with $K = U \, \mathrm{Diag}(\lambda_1, \dots, \lambda_d) U^{\top}$. Now, since $x_n = \frac{1}{1+\tau\mu_x}(x_{n-1} - \tau K y_n)$, we have $[x_n^{\top}, y_n^{\top}]^{\top} = T[x_{n-1}^{\top}, y_n^{\top}]^{\top}$ where

$$T \triangleq \begin{bmatrix} \frac{1}{1+\tau\mu_x} I & \frac{-\tau}{1+\tau\mu_x} K \\ 0 & I \end{bmatrix}.$$

Thus, $\Sigma^{\infty} = T \Sigma^{\infty} T^{\top}$, and noting that T admits the block diagonalization $T = Z T^{(\Lambda)} Z^{\top}$ where $T^{(\Lambda)} = \mathrm{Diag}(T^{(\lambda_1)}, \dots, T^{(\lambda_d)})$ and

$$T^{(\lambda_i)} = \begin{bmatrix} \frac{1}{1+\tau\mu_x} & \frac{-\tau\lambda_i}{1+\tau\mu_x} \\ 0 & 1 \end{bmatrix} \quad \forall i \in \{1, \dots, d\},$$

we obtain $\Sigma^{\infty} = Z T^{(\Lambda)} \tilde{\Sigma}^{\infty,\Lambda} (T^{(\Lambda)})^{\top} Z^{\top}$, where $\tilde{\Sigma}^{\infty,\Lambda} = \mathrm{Diag}(\tilde{\Sigma}^{\infty,\lambda_1}, \dots, \tilde{\Sigma}^{\infty,\lambda_d})$. Finally, we observe that $T^{(\Lambda)} \tilde{\Sigma}^{\infty,\Lambda} (T^{(\Lambda)})^{\top} = \mathrm{Diag}(\Sigma^{\infty,,\lambda_1}, \dots, \Sigma^{\infty,,\lambda_d})$ where

$$\Sigma^{\infty,\lambda_i} \triangleq \begin{bmatrix} \theta^2 \tilde{\Sigma}_{11}^{\infty,\lambda_i} - 2\theta(1-\theta)\frac{\lambda_i}{\mu_x}\tilde{\Sigma}_{12}^{\infty,\lambda_i} + \frac{(1-\theta)^2\lambda_i^2}{\mu_x^2}\tilde{\Sigma}_{22}^{\infty,\lambda_i} & \theta\tilde{\Sigma}_{12}^{\infty,\lambda_i} - \theta(1-\theta)\frac{\lambda_i}{\mu_x}\tilde{\Sigma}_{22}^{\infty,\lambda_i} \\ \theta\tilde{\Sigma}_{12}^{\infty,\lambda_i} - \theta(1-\theta)\frac{\lambda_i}{\mu_x}\tilde{\Sigma}_{22}^{\infty,\lambda_i} & \tilde{\Sigma}_{22}^{\infty,\lambda_i} \end{bmatrix}.$$

Plugging $\lambda = \lambda_i$ into the expression of $\tilde{\Sigma}^{\infty,\lambda}$ computed in Proposition 32, we obtain $\Sigma^{\infty,0} = \frac{\delta^2}{d}\frac{(1-\theta)}{\mu_x^2\mu_y^2(1+\theta)} \mathrm{Diag}\left[\theta^2\mu_y^2,\ \mu_x^2\left(1 + 2\left(1-\theta^2\right)\theta\right)\right]$, if $\lambda_i = 0$; otherwise,

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

where the polynomials $P_{i,j}^{(\infty,k)}$ and P_c are given explicitly in the bottom part of Table 7 of Appendix E. From the closed-form expressions of these polynomials, the fact that the elements of the matrix $\Sigma^{\infty,\lambda}$ scale with $(1-\theta)$ as $\theta \to 1$ can be checked in a straightforward manner. ■

D.3 Proof of Corollary 18

Let V, J denote the Jordan decomposition of A. For $n \in \mathbb{N}$, let $\check{\Sigma}_n \triangleq V^{-1}\Sigma_n\left(V^{-1}\right)^{\top}$, $\check{\Sigma}^{\infty} \triangleq V^{-1}\Sigma^{\infty}\left(V^{-1}\right)^{\top}$, and $\check{R}^i = V^{-1}R\left(V^{-1}\right)^{\top}$. In view of the recursion (D.2), we have $\check{\Sigma}_{n+1} = J\check{\Sigma}_n J + \check{R}$, and vectorizing again this recursion lead to $\mathrm{Vec}(\check{\Sigma}_{n+1}) = (J \otimes J)\,\mathrm{Vec}(\check{\Sigma}_n) + \check{R}$, i.e., $\check{\Sigma}_n = (J \otimes J)^{n-1}\,\check{\Sigma}_1 + \sum_{k=1}^{n-1}(J \otimes J)^{k-1}\,\mathrm{Vec}(\check{R})$. Hence, noting that $\check{\Sigma}^{\infty} = \sum_{k=0}^{\infty}(J \otimes J)^{k}\,\mathrm{Vec}(\check{R})$ we obtain

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

and the claimed convergence rate follows from observing that $\rho\left(J \otimes J\right) = \rho(A)^2$. Note that here $\rho(A) < 1$ because by Proposition 32 we have $\rho(A^{\lambda_i}) < 1$ for every i and $\rho(A) = \max_i \rho(A^{\lambda_i})$.

D.4 Proof of Theorem 19

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

D.4.1 Lower bound

In view of Theorem 17, z_∞ follows a centered Gaussian distribution with covariance matrix Σ^∞ as defined in (4.4). Hence, let $X \sim \mathcal{N}(0, I_d)$ be such that $\|z_\infty\|^2 = X^\top \Sigma^\infty X$. We almost surely have $\|z_\infty\|^2 \geqslant \|X\|^2 \ \min \mathrm{Sp}(\Sigma^\infty) \triangleq \psi_1(p,\theta)$. By (Inglot, 2010), we have $Q_p(\|X\|^2) \geqslant 2d + 2\log\left(1/(1-p)\right)) - 5/2$, where we used $\mathrm{z}^\star = (0,0)$. Thus, it suffices to show that $\min \mathrm{Sp}(\Sigma^\infty) = \Theta(1-\theta)$ as $\theta \to 1$. Given the bloc decomposition of Σ^∞ in (4.4), we have

$$\min \mathrm{Sp}(\Sigma^\infty) = \min_{i \in \{1,\dots,d\}} \mathrm{Sp}(\Sigma^{\infty,\lambda_i}) = \min_{i \in \{1,\dots,d\}} \frac{1}{2}\left(\Sigma_{11}^{\infty,,\lambda_i} + \Sigma_{22}^{\infty,,\lambda_i} - \sqrt{\left(\Sigma_{11}^{\infty,,\lambda_i} - \Sigma_{22}^{\infty,,\lambda_i}\right)^2 + 4{\Sigma_{12}^{\infty}}^2}\right).$$

We will now show that for all $\lambda \in \mathrm{Sp}(K)$, $\Sigma_{11}^{\infty,,\lambda} + \Sigma_{22}^{\infty,,\lambda} - \sqrt{\left(\Sigma_{11}^{\infty,,\lambda} - \Sigma_{22}^{\infty,,\lambda}\right)^2 + 4{\Sigma_{12}^{\infty}}^2} = \Theta(1-\theta)$, as $\theta \to 1$. If $0 \in \mathrm{Sp}(\Sigma^\infty)$, given (4.3), we have

$$\begin{aligned}
&\Sigma_{1,1}^{\infty,0} + \Sigma_{2,2}^{\infty,0} - \sqrt{\left(\Sigma_{1,1}^{\infty,0} - \Sigma_{2,2}^{\infty,0}\right)^2 + 4(\Sigma_{1,2}^{\infty,0})^2} \\
&\qquad = \frac{\delta^2}{d}\frac{(1-\theta)}{\mu_x^2\mu_y^2(1+\theta)}\left(\theta^2\mu_y^2 + \mu_x^2\left(1 + 2\left(1-\theta^2\right)\theta\right) - \mid \theta_{\mu_y^2}^2 - \mu_x^2\left(1 + 2\left(1-\theta^2\right)\theta\right) \mid\right) \\
&\qquad = \frac{\delta^2}{d}\frac{2\theta(1-\theta)}{\mu_x^2\mu_y^2(1+\theta)}\min\left(\mu_y^2, \mu_x^2\left(1 + 2\left(1-\theta^2\right)\right)\right) = \Theta(1-\theta),
\end{aligned}$$

where second equality follows from having $a + b - |a-b| = 2\min(a,b)$. If $\lambda \neq 0$ is in $\mathrm{Sp}(K)$, in view of Table 7 of Appendix E, we have as $\theta \to 1$,

$$\begin{aligned}
&P_{1,1}^{\infty,1}(\theta,\kappa) = -16\kappa^2 + o(1-\theta), && P_{1,2}^{\infty,1}(\theta,\kappa) = -8\kappa^4 + o(1-\theta), && P_{2,2}^{\infty,1}(\theta,\kappa) = -8\kappa^6 + o(1-\theta),\\
&P_{1,1}^{\infty,2}(\theta,\kappa) = -8\kappa^2 + o(1-\theta), && P_{1,2}^{\infty,2}(\theta,\kappa) = 8\kappa^2 + o(1-\theta), && P_{2,2}^{\infty,2}(\theta,\kappa) = -16\kappa^2 + o(1-\theta),
\end{aligned}$$

and $P_c(\theta,\kappa) = -32, \kappa^2 + o(1-\theta)$, so that

$$\begin{aligned}
\Sigma_{1,1}^{\infty,\lambda} &= \tfrac{(1-\theta)\delta^2}{d(-32\kappa^2)\lambda^2}(-8\kappa^2)\left(\tfrac{\lambda^2}{\mu_x^2} + \tfrac{\lambda^4}{\mu_x^2\mu_y^2}\right) + o(1-\theta), \quad & \Sigma_{1,2}^{\infty,\lambda} &= \tfrac{(1-\theta)\delta^2}{d(-32\kappa^2)\lambda^2}(8\kappa^2)\left(\tfrac{\lambda^3}{\mu_y^2\mu_x} - \tfrac{\lambda^3}{\mu_x^2\mu_y}\right) + o(1-\theta),\\
& & \Sigma_{2,2}^{\infty,\lambda} &= \tfrac{(1-\theta)\delta^2}{d(-32\kappa^2)\lambda^2}(-8\kappa^2)\left(\tfrac{\lambda^2}{\mu_y^2} + \tfrac{\lambda^4}{\mu_x^2\mu_y^2}\right) + o(1-\theta).
\end{aligned}$$

Hence, we deduce that

$$\begin{aligned}
&\Sigma_{1,1}^{\infty,\lambda} + \Sigma_{2,2}^{\infty,\lambda} - \sqrt{\left(\Sigma_{1,1}^{\infty,\lambda} - \Sigma_{2,2}^{\infty,\lambda}\right)^2 + 4(\Sigma_{1,2}^{\infty,\lambda})^2} \\
&\qquad = \frac{(1-\theta)\delta^2}{2d\mu_x^2\mu_y^2}\left(\lambda^2 + \mu_x^2 + \mu_y^2 - \sqrt{(\mu_x^2-\mu_y^2)^2 + \lambda^2(\mu_x-\mu_y)^2}\right) + o(1-\theta),
\end{aligned}$$

and it suffices to show that $\lambda^2 + \mu_x^2 + \mu_y^2 - \sqrt{(\mu_x^2 - \mu_y^2)^2 + \lambda^2(\mu_x - \mu_y)^2} > 0$. Now given the identity $a - b = \frac{a^2 - b^2}{a+b}$ for $a + b \neq 0$, we have

$$\lambda^2 + \mu_x^2 + \mu_y^2 - \sqrt{(\mu_x^2 - \mu_y^2)^2 + \lambda^2(\mu_x - \mu_y)^2} = \frac{\lambda^2(\mu_x - \mu_y)^2 + \lambda^4}{\lambda^2 + \mu_x^2 + \mu_y^2 + \sqrt{(\mu_x^2 - \mu_y^2)^2 + \lambda^2(\mu_x - \mu_y)^2}} > 0,$$

which completes the proof.

D.4.2 Upper bound

The CP parametrization corresponds to choosing $\alpha = \frac{1}{2\sigma} - \sqrt{\theta} L_{yy}$ in the matrix inequality (Zhang et al., 2024, Cor. 1). Under this parameterization, since $1 - \alpha\sigma \geqslant 1/2$, we have $\mathcal{W}_n \geqslant \frac{\theta}{4(1-\theta)} \left(\mu_x \|x_n - \mathrm{x}^\star\|^2 + \mu_y \|y_n - \mathrm{y}^\star\|^2\right)$ and we have $\mathrm{z}^\star = (\mathrm{x}^\star, \mathrm{y}^\star) = (0, 0)$. Since $\{z_n\}_{n \geqslant 0}$ converges in distribution to z_∞, (3.1) implies that the p-quantile of $\|z_\infty\|^2$ satisfies $Q_p(\|z_\infty\|^2) \leqslant \psi_2(p, \theta)$ for any $p \in (0, 1)$, where $\psi_2(p, \theta) \triangleq \frac{4(1-\theta)}{\theta \min\{\mu_x, \mu_y\}} \left(\Xi^{(1)}_{\tau,\sigma,\theta} + \Xi^{(2)}_{\tau,\sigma,\theta} \log\left(\frac{1}{1-p}\right)\right)$. Thus, the asymptotic property of our upper bound follows from Lemma **??**.

D.5 Proof of Proposition 32

We first note that under the parameterization (2.3), the matrices A^λ and R^λ simplify to

$$A^\lambda = \begin{bmatrix} \theta & -(1-\theta)\frac{\lambda}{\mu_x} \\ (1-\theta)\theta^2 \frac{\lambda}{\mu_y} & \theta - (1-\theta)^2(1+\theta)\kappa^2 \end{bmatrix},$$
$$R^\lambda = \frac{\delta^2}{d} \frac{(1-\theta)^2}{\mu_x^2 \mu_y^2} \begin{bmatrix} \mu_y^2 & \left(1-\theta^2\right)\left(\theta\mu_x + \mu_y\right)\lambda \\ \left(1-\theta^2\right)\left(\theta\mu_x + \mu_y\right)\lambda & (1-\theta)^2(1+\theta)^2\left(1 + 2\theta\frac{\mu_x}{\mu_y}\right)\lambda^2 + \mu_x^2\left(1 + 2\left(1-\theta^2\right)\theta\right) \end{bmatrix}. \tag{D.6}$$

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

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

Noting that $\left(I - A^\lambda \otimes A^\lambda\right)^{-1} = \mathrm{Diag}(\frac{1}{1-\theta^2}, \frac{1}{1-\theta^2}, \frac{1}{1-\theta^2}, \frac{1}{1-\theta^2})$, we obtain $\Sigma^{\infty,0} = \frac{1}{1-\theta^2} R^0$ for any $\theta \in [0, 1)$. It remains to consider the case when $\lambda \neq 0$. We first provide an eigenvalue decomposition to the matrix A^λ.

Lemma 33 *For any $\theta \in \left(\left(\sqrt{1+\kappa_\lambda^2} - 1\right)/|\kappa_\lambda|, 1\right)$ and $\lambda \neq 0$, the matrix A^λ introduced in* (D.4) *admits the diagonalization $A^\lambda = V^\lambda J^\lambda (V^\lambda)^{-1}$ where*

$$J^\lambda = \begin{bmatrix} \nu_{1,\lambda} & 0 \\ 0 & \nu_{2,\lambda} \end{bmatrix}, \quad V^\lambda \triangleq \begin{bmatrix} -A_{1,2} & -A_{1,2} \\ \theta - \nu_{1,\lambda} & \theta - \nu_{2,\lambda} \end{bmatrix}, \tag{D.7}$$

with complex eigenvalues $\nu_{1,\lambda} \triangleq \frac{\left(2\theta - (1-\theta)^2(1+\theta)\kappa_\lambda^2\right) + i\sqrt{|\Delta|}}{2}$, $\nu_{2,\lambda} \triangleq \frac{\left(2\theta - (1-\theta)^2(1+\theta)\kappa_\lambda^2\right) - i\sqrt{|\,\Delta_\lambda\,|}}{2}$, and $\Delta_\lambda = (1-\theta)^4(1+\theta)^2\kappa_\lambda^4 - 4\theta^2(1-\theta)^2\kappa_\lambda^2$. Moreover, in this case, $\Delta_\lambda < 0$ and $\rho(A^\lambda) < 1$.

Proof Noting that $\mathrm{Tr}(A^\lambda) = 2\theta - (1-\theta)^2(1+\theta)\kappa_\lambda^2$ and $\mathrm{Det}(A^\lambda) = \theta^2 - (1-\theta)^2\theta\kappa_\lambda^2$, the characteristic polynomial of A^λ has for discriminant $\Delta_\lambda = \mathrm{Tr}(A^\lambda) - 4\mathrm{Det}(A^\lambda) = (1-\theta)^4(1+\theta)^2\kappa_\lambda^4 - 4\theta^2(1-\theta)^2\kappa_\lambda^2$. Note also that $\kappa_\lambda \neq 0$ since $\lambda \neq 0$ by assumption, and

$$\Delta_\lambda < 0 \iff (1-\theta^2)^2 \leqslant \frac{4\theta^2}{\kappa_\lambda^2} \iff (1-\theta^2) \leqslant \frac{2\theta}{|\kappa_\lambda|} \iff \theta \geqslant \frac{1}{|\kappa_\lambda|}(\sqrt{1+\kappa_\lambda^2} - 1),$$

and in such case, it is straightforward to check that A^λ admits the two complex conjugate values $\nu_{1,\lambda}, \nu_{2,\lambda}$. Furthermore, observe that $A_{12}^{(\lambda} \neq 0$ as $\lambda \neq 0$ and $\theta < 1$ and for $\nu \in \mathbb{C}$, $x, y \in \mathbb{C}$,

$$(A^\lambda - vI)\begin{bmatrix} x \\ y \end{bmatrix} = 0 \Leftrightarrow \begin{cases} (\theta - v)\, x + A^\lambda_{1,2} y = 0 \\ A^\lambda_{2,1} x + \left(A^\lambda_{2,2} - v\right) y = 0 \end{cases} \Leftrightarrow y = \frac{-\,(\theta - v)}{A^\lambda_{1,2}} x$$

$$\Leftrightarrow \quad (x, y) \in \operatorname{Span}\begin{pmatrix} 1 \\ -\frac{(\theta - v)}{A^\lambda_{1,2}} \end{pmatrix} = \operatorname{Span}\begin{pmatrix} -A_{1,2} \\ \theta - v \end{pmatrix}.$$

Therefore, the columns of the V^λ matrix are in fact eigenvectors corresponding to the complex conjugate eigenvalues $\nu_{1,\lambda}$ and $\nu_{2,\lambda}$, and we conclude that the eigenvalue decomposition $A^\lambda = V^\lambda J^\lambda (V^\lambda)^{-1}$ holds. Finally, $\rho(A^\lambda)^2 = |\nu_{1,\lambda}|^2 = \operatorname{Det}(A^\lambda) = \theta^2 - \theta(1-\theta)^2\kappa_\lambda^2$ so that we have $\rho(A^\lambda)^2 - 1 = \operatorname{Det}(A^\lambda) - 1 = -(1-\theta)\left(1 + \theta\left(1 + \kappa_\lambda^2\right) - \theta^2\kappa_\lambda^2\right)$, and $\rho(A^\lambda)^2 = 1$ if and only if $\theta \in \left\{1, \frac{1}{2} + \frac{1}{2\kappa_\lambda^2} \pm \frac{1}{2\kappa_\lambda^2}\sqrt{(1+\kappa_\lambda^2)^2 + 4\kappa_\lambda^2}\right\}$. Observing that $\sqrt{(1+\kappa_\lambda^2)^2 + 4\kappa_\lambda^2} \geqslant 1 + \kappa_\lambda^2$, we deduce that $\frac{1}{2} + \frac{1}{2\kappa_\lambda^2} + \frac{1}{2\kappa_\lambda^2}\sqrt{(1+\kappa_\lambda^2)^2 + 4\kappa_\lambda^2} > 1$ and $\frac{1}{2} + \frac{1}{2\kappa_\lambda^2} - \frac{1}{2\kappa_\lambda^2}\sqrt{(1+\kappa_\lambda^2)^2 + 4\kappa_\lambda^2} < 0$. Hence, we conclude $\rho(A^\lambda) < 1$ for any $\theta \in \left(\frac{1}{\kappa_\lambda}(\sqrt{1+\kappa_\lambda^2} - 1), 1\right)$. ■

In the following lemma, we also provide basic identities satisfied by the eigenvalues $\nu_{1,\lambda}$ and $\nu_{2,\lambda}$ which will be key for the exact computation of $\tilde{\Sigma}^{\infty,\lambda}$. The proof of this lemma is omitted as it follows from straightforward calculations.

Lemma 34 *Let $\nu_{1,\lambda}, \nu_{2,\lambda}$, be the two complex conjugate eigenvalues of A^λ, as specified in Lemma 33. Then,*

$$\begin{aligned}
\nu_{1,\lambda}\nu_{2,\lambda} &= \theta^2 - \theta(1-\theta)^2\kappa_\lambda^2, \\
\nu_{1,\lambda} + \nu_{2,\lambda} &= 2\theta - (1-\theta)^2(1+\theta)\kappa_\lambda^2, \\
\nu_{1,\lambda}^2 + \nu_{2,\lambda}^2 &= 2\theta^2 - 2\theta(1-\theta)^2(1+2\theta)\kappa_\lambda^2 + (1-\theta)^4(1+\theta)^2\kappa_\lambda^4, \\
\nu_{1,\lambda}^3 + \nu_{2,\lambda}^3 &= \left(2\theta - (1-\theta)^2(1+\theta)\kappa_\lambda^2\right)\left(\theta^2 - \theta(1-\theta)^2(1+4\theta)\kappa_\lambda^2 + (1-\theta)^4(1+\theta)^2\kappa_\lambda^4\right), \\
\nu_{1,\lambda}^4 + \nu_{2,\lambda}^4 &= 2\theta^4 - (1-\theta)^2\kappa_\lambda^2\theta^3(4+16\theta) + (1-\theta)^4\kappa_\lambda^4\theta^2\left(6 + 24\theta + 20\theta^2\right) \\
&\quad - (1-\theta)^6\kappa_\lambda^6 4\theta\left(1 + 4\theta + 5\theta^2 + 2\theta^3\right) + (1-\theta)^8\kappa_\lambda^8(1+\theta)^4,
\end{aligned}$$

where $\kappa_\lambda = \frac{\lambda}{\sqrt{\mu_x\mu_y}}$.

The following lemma says that the solution $\tilde{\Sigma}^{\infty,\lambda}$ of (D.5) can be computed by solving 4-dimensional linear equations.

Lemma 35 *For any $\lambda \neq 0$ and $\theta \in \left(\frac{1}{|\kappa_\lambda|}(\sqrt{1+\kappa_\lambda^2} - 1), 1\right)$, the solution $\tilde{\Sigma}^{\infty,\lambda}$ of* (D.5) *satisfies*

$$Vec\left(\check{\Sigma}^{\infty,\lambda}\right) = \left(I_4 - J^\lambda \otimes J^\lambda\right)^{-1} Vec(\check{R}^\lambda), \tag{D.8}$$

where $\check{\Sigma}^{\infty,\lambda} \triangleq (V^\lambda)^{-1}\tilde{\Sigma}^{\infty,\lambda}\left((V^\lambda)^{-1}\right)^\top$, $\check{R}^\lambda \triangleq (V^\lambda)^{-1}R^\lambda\left((V^\lambda)^{-1}\right)^\top$ and J^λ, V^λ are the matrices arising in the eigenvalue decomposition of A^λ, as in Lemma 33.

Proof Given Lemma 33, the equation (D.5) can be rewritten as $\tilde{\Sigma}^{\infty,\lambda} = V^\lambda J^\lambda (V^\lambda)^{-1}\tilde{\Sigma}^{\infty,\lambda}((V^\lambda)^{-1})^\top (J^\lambda)^\top (V^\lambda)^\top + R^\lambda$, which amounts to

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

or equivalently $\check{\Sigma}^{\infty,\lambda} = J^\lambda \check{\Sigma}^{\infty,\lambda}(J^\lambda)^\top + \check{R}$. Taking $\mathrm{Vec}(\cdot)$ of both sides, and noting the relation $\mathrm{Vec}(ABC) = (C^\top \otimes A)\mathrm{Vec}(B)$, it suffices to show that the 4×4 matrix $I_4 - J^\lambda \otimes J^\lambda$ is invertible. Observing that $\mathrm{Diag}\left[\nu_{1,\lambda}^2, \nu_{1,\lambda}\nu_{2,\lambda}, \nu_{1,\lambda}\nu_{2,\lambda}, \nu_{2,\lambda}^2\right]$, it is sufficient to show that $\nu_{1,\lambda}\nu_{2,\lambda} \neq 1$ for $\theta \in (\frac{1}{|\kappa_\lambda|}(\sqrt{1+\kappa_\lambda^2} - 1), 1)$, and this directly follows from the proof of Lemma 33. ■

Equipped with the representation (D.8), we complete the proof of Proposition 32 in three steps: (I) explicit computation of $\check{R}^\lambda$, (II) explicit computation of $\check{\Sigma}^{\infty,\lambda}$ from $\check{R}^\lambda$ based on (D.8), (III) explicit computation of $\tilde{\Sigma}^{\infty,\lambda}$ from $\check{\Sigma}^{\infty,\lambda}$ based on the relationship given in Lemma 35.

(I) **Computation of $\check{R}$.** Using the Cramer rule, first observe that V^{-1} satisfies

$$(V^\lambda)^{-1} = \frac{1}{A_{1,2}\left(\nu_{2,\lambda} - \nu_{1,\lambda}\right)} \begin{bmatrix} (\theta - \nu_{2,\lambda}) & +A_{1,2} \\ -(\theta - \nu_{1,\lambda}) & -A_{1,2} \end{bmatrix},$$

from which we deduce

$$\check{R}^{(\lambda)} = (V^\lambda)^{-1} R^\lambda ((V^\lambda)^{-1})^\top = \frac{\delta^2(1-\theta)^2}{d(A_{1,2}^\lambda)^2\left(\nu_{1,\lambda} - \nu_{2,\lambda}\right)^2 \mu_x^2\mu_y^2} \begin{bmatrix} Q_{1,1}^\lambda & Q_{1,2}^\lambda \\ Q_{1,2}^\lambda & Q_{2,2}^\lambda \end{bmatrix},$$

where

$$\begin{aligned} Q_{1,1}^\lambda \triangleq\ & \left(\theta - \nu_{2,\lambda}\right)^2 \mu_y^2 \\ & + (A_{1,2}^\lambda)^2 \left((1-\theta)^2(1+\theta)^2\left(1 + 2\theta\frac{\mu_x}{\mu_y}\right)\lambda^2 + \mu_x^2\left(1 + 2\left(1-\theta^2\right)\theta\right)\right) \\ & + 2A_{1,2}^\lambda\left(\theta - \nu_{2,\lambda}\right)\left(1-\theta^2\right)\left(\theta\mu_x + y_y\right)\lambda, \\ Q_{2,2}^\lambda \triangleq\ & \left(\theta - \nu_{1,\lambda}\right)^2 \mu_y^2 \\ & + (A_{1,2}^\lambda)^2 \left((1-\theta)^2(1+\theta)^2\left(1 + 2\theta\frac{\mu_x}{\mu_y}\right)\lambda^2 + \mu_x^2\left(1 + 2\left(1-\theta^2\right)\theta\right)\right) \\ & + 2A_{1,2}^\lambda\left(\theta - \nu_{1,\lambda}\right)\left(1-\theta^2\right)\left(\theta\mu_x + y_y\right)\lambda, \\ Q_{1,2}^\lambda \triangleq\ & -\left(\theta - \nu_{2,\lambda}\right)\left(\theta - \nu_{1,\lambda}\right)\mu_y^2 \\ & - (A_{1,2}^\lambda)^2 \left((1-\theta)^2(1+\theta)^2\left(1 + 2\theta\frac{\mu_x}{\mu_y}\right)\lambda^2 + {\mu_x}^2\left(1 + 2\left(1-\theta^2\right)\theta\right)\right) \\ & + \left(1-\theta^2\right)\left(\theta\mu_x + \mu_y\right)\lambda A_{1,2}^\lambda\left(\nu_{1,\lambda} + \nu_{2,\lambda} - 2\theta\right). \end{aligned} \tag{D.9}$$

(II) **Computation of $\check{\Sigma}^{\infty,\lambda}$.** From the definition of J^λ given in (D.7), observing that $\left(I_4 - J^\lambda \otimes J^\lambda\right)^{-1} = \mathrm{Diag}\left[\frac{1}{1-\nu_{1,\lambda}^2}, \frac{1}{1-\nu_{1,\lambda}\nu_{2,\lambda}}, \frac{1}{1-\nu_{1,\lambda}\nu_{2,\lambda}}, \frac{1}{1-\nu_{2,\lambda}^2}\right]$, we deduce from Lemma 35 that

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

(III) **Computation of $\tilde{\Sigma}^{\infty,\lambda}$.** From Lemma 35, we also have

$$\tilde{\Sigma}^{\infty,\lambda} = V^\lambda \check{\Sigma}^{\infty,\lambda}(V^\lambda)^\top = \frac{\delta^2(1-\theta)^2}{d(A_{1,2}^\lambda)^2\left(\nu_{1,\lambda} - \nu_2\right)^2 \mu_x^2\mu_y^2} \begin{bmatrix} S_{1,1}^\lambda & S_{1,2}^\lambda \\ S_{1,2}^\lambda & S_{2,2}^\lambda \end{bmatrix}, \tag{D.10}$$

with

$$
\begin{aligned}
S_{1,1}^{\lambda} &\triangleq (A_{1,2}^{\lambda})^2 \left(\frac{Q_{1,1}^{\lambda}}{1-\nu_{1,\lambda}^2} + \frac{Q_{2,2}^{\lambda}}{1-\nu_{2,\lambda}^2} + 2\frac{Q_{1,2}^{\lambda}}{1-\nu_{1,\lambda}\nu_2} \right), \\
S_{1,2}^{\lambda} &\triangleq -A_{1,2}^{\lambda} \left[(\theta-\nu_{1,\lambda})\frac{Q_{1,1}^{\lambda}}{1-\nu_{1,\lambda}^2} + (\theta-\nu_2)\frac{Q_{2,2}^{\lambda}}{1-\nu_{2,\lambda}^2} + (2\theta-(\nu_{1,\lambda}+\nu_2))\frac{Q_{1,2}^{\lambda}}{1-\nu_{1,\lambda}\nu_2} \right], \qquad \text{(D.11)} \\
S_{2,2}^{\lambda} &\triangleq (\theta-\nu_{1,\lambda})^2 \frac{Q_{1,1}^{\lambda}}{1-\nu_{1,\lambda}^2} + (\theta-\nu_2)^2\frac{Q_{2,2}^{\lambda}}{1-\nu_{2,\lambda}^2} + 2(\theta-\nu_{1,\lambda})(\theta-\nu_2)\frac{Q_{1,2}^{\lambda}}{1-\nu_{1,\lambda}\nu_2}.
\end{aligned}
$$

While (D.10) provides a formula for $\Sigma^{\lambda,\infty}$, the dependence of this formula to κ_λ and θ are not very clear. We provide in Section F of the extended version of this paper Laguel et al. (2023) a simplification of the terms in (D.10) in terms of their dependence to κ_λ and θ. As a consequence, in view of (D.10), we may write

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

where the $\tilde{P}_{q,\ell}^{(k)}$ and P_c are polynomials in θ, κ_λ, defined in the top part of Table 7. This proves Proposition 32.

Appendix E. 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. All the constants arising in these inequalities (including those mentioned in the statement of Lemma 26 and Lemma27) are made explicit as follows in Table 4 and Table 5. For convenience of the reader, in Section H of the online supplementary material Laguel et al. (2023), we also provide the expressions of these constants under the CP parameterization (2.3) which is a particular class of parameters where our complexity results can be achieved. We finally detail in Table 7 the polynomials involved in the entries of the covariance matrix $\Sigma^{\infty,\lambda}$ given in Theorem 17.

$$B^x = \tfrac{4\tau}{1+\rho} + \tfrac{\sigma^2(1+\theta)^2\,\mathrm{L_{yx}}^2}{(1+\sigma\mu_y)^2}\tfrac{\rho}{4\|A_0\|^2(1+\rho)}\tfrac{3\tau^2}{(1+\tau\mu_x)^2}, \quad C^x = \tfrac{\tau}{1+\tau\mu_x} + \tfrac{\tau\sigma(1+2\theta)\,\mathrm{L_{xy}}}{2(1+\tau\mu_x)(1+\sigma\mu_y)}, \quad C^x_{-1} = \tfrac{\tau\sigma\theta}{2\rho(1+\sigma\mu_y)}\tfrac{(1+\theta)\,\mathrm{L_{yx}}}{1+\tau\mu_x},$$

$$\begin{aligned} B^y = \tfrac{4(1+\theta)^2\sigma}{(1+\rho)(1-\alpha\sigma)} + \tfrac{3\|A_0\|^2(1+\rho)\theta^2}{\rho^3} &+ \frac{\rho}{4\left\|A_0\right\|^2(1+\rho)}\frac{(1+\sigma(1+\theta)\,\mathrm{L_{yy}})^2}{(1+\sigma\mu_y)^2}\frac{2\sigma^2(1+\theta)^2}{\left(1+\sigma\mu_y\right)^2} \\ &+ \frac{\sigma^2(1+\theta)^2\,\mathrm{L_{yx}}^2}{\left(1+\sigma\mu_y\right)^2}\frac{\rho}{4\left\|A_0\right\|^2(1+\rho)}\frac{3\tau^2\sigma^2(1+\theta)^2\,\mathrm{L_{xy}}^2}{\left(1+\tau\mu_x\right)^2\left(1+\sigma\mu_y\right)^2} \end{aligned}$$

$$B^y_{-1} = \tfrac{4\sigma\theta^2}{\rho(1+\rho)(1-\alpha\sigma)} + \tfrac{\rho}{4\|A_0\|^2(1+\rho)}\tfrac{(1+\sigma(1+\theta)\,\mathrm{L_{yy}})^2}{(1+\sigma\mu_y)^2}\tfrac{2\sigma^2\theta^2\rho^{-1}}{(1+\sigma\mu_y)^2} + \tfrac{\sigma^2(1+\theta)^2\,\mathrm{L_{yx}}^2}{(1+\sigma\mu_y)^2}\tfrac{\rho}{4\|A_0\|^2(1+\rho)}\tfrac{3\tau^2\sigma^2\theta^2\rho^{-1}\,\mathrm{L_{xy}}^2}{(1+\tau\mu_x)^2(1+\sigma\mu_y)^2},$$

$$C^y = \tfrac{\sigma(1+2\theta)(1+\theta)}{1+\sigma\mu_y} + \tfrac{\tau\sigma(1+\theta)\,\mathrm{L_{xy}}}{2(1+\tau\mu_x)(1+\sigma\mu_y)}, \quad C^y_{-2} = \tfrac{\sigma\theta^2}{2\rho^2(1+\sigma\mu_y)}\left(\tfrac{1+\sigma(1+\theta)\mathrm{L_{yy}}}{1+\sigma\mu_y} + \tfrac{\tau\sigma(1+\theta)\mathrm{L_{yx}}\mathrm{L_{xy}}}{(1+\tau\mu_x)(1+\sigma\mu_y)}\right),$$

$$C^y_{-1} = \tfrac{\sigma\theta}{\rho(1+\sigma\mu_y)}\Big(1 + 2\theta + \tfrac{\tau}{2(1+\tau\mu_x)}\left((1+\theta)\,\mathrm{L_{yx}} + \mathrm{L_{xy}}\right) + (1+\tfrac{3\theta}{2})\Big(\tfrac{1+\sigma(1+\theta)\mathrm{L_{yy}}}{1+\sigma\mu_y} + \tfrac{\tau\sigma(1+\theta)\mathrm{L_{yx}}\mathrm{L_{xy}}}{(1+\tau\mu_x)(1+\sigma\mu_y)}\Big)\Big),$$

$$\mathcal{Q}_x \triangleq B^x + C^x + C^x_{-1}, \quad \mathcal{Q}_y = B^y + B^y_{-1} + C^y + C^y_{-1} + C^y_{-2}.$$

Table 4: Summary of the constants $B^x, C^x, C^x_{-1}, \mathcal{Q}^x, B^y, B^y_{-1}, C^y, C^y_{-1}, C^y_{-2}$, and $\mathcal{Q}^y$ used throughout the analysis.

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

$$\hat{A}_3 \triangleq \begin{bmatrix} \begin{bmatrix} C_{\sigma,\theta}\sigma(1+\theta)(1+\tau\mu_x)\,\mathrm{L_{yx}} + \sigma(1+\theta)\,\mathrm{L_{yx}}\Big((1+\tau\,\mathrm{L_{xx}})(1+\sigma\mu_y) + \tau\sigma(1+\theta)\,\mathrm{L_{yx}}\,\mathrm{L_{xy}}\Big) \\ + \sigma\theta\,\mathrm{L_{yx}}(1+\tau\mu_x)(1+\sigma\mu_y) \end{bmatrix} \\ (1+\tau\mu_x)C^2_{\sigma,\theta} + \sigma(1+\theta)\,\mathrm{L_{yx}}\,\tau\,\mathrm{L_{xy}}\,C_{\sigma,\theta} + \sigma\theta\,\mathrm{L_{yy}}(1+\tau\mu_x)(1+\sigma\mu_y) \\ \sigma\Big(C_{\sigma,\theta}\theta\,\mathrm{L_{yx}}(1+\tau\mu_x) + (1+\theta)\tau\sigma\theta\,\mathrm{L_{xy}}\,\mathrm{L_{yx}}^2\Big) \\ \sigma\Big(C_{\sigma,\theta}\theta\,\mathrm{L_{yy}}(1+\tau\mu_x) + (1+\theta)\tau\sigma\theta\,\mathrm{L_{xy}}\,\mathrm{L_{yx}}\,\mathrm{L_{yy}}\Big) \end{bmatrix},$$

$$A \triangleq \begin{bmatrix} A_1^\top \\ A_2^\top \\ A_3^\top \end{bmatrix} \triangleq \tfrac{\sqrt{1+\rho}}{\sqrt{\rho}} \begin{bmatrix} \frac{4}{1+\tau\mu_x}\hat{A}_1^\top \\ \frac{4\sqrt{2}(1+\theta)}{1+\sigma\mu_y}\hat{A}_2^\top \\ \frac{4\sqrt{2}\theta\rho^{-1}}{(1+\sigma\mu_y)^2(1+\tau\mu_x)}\hat{A}_3^\top \end{bmatrix} \mathrm{Diag}\begin{bmatrix} \frac{\sqrt{2\rho\tau}}{\sqrt{2\rho\sigma/(1-\alpha\sigma)}} \\ \frac{\sqrt{2\rho\tau}}{\sqrt{2\rho\sigma/(1-\alpha\sigma)}} \end{bmatrix}, \quad C_{\sigma,\theta} \triangleq 1 + \sigma(1+\theta)\,\mathrm{L_{yy}}.$$

Table 5: Summary of the constants A_1, A_2, A_3 used throughout the analysis.

$\begin{aligned} \tilde{P}^{(1)}_{1,1}(\theta,\kappa) = & -4\kappa_\lambda^2\theta^2\left(1+\theta\right)^2 \\ & +\kappa_\lambda^4\begin{pmatrix} 1+2\theta-\theta^2-8\theta^3 \\ -9\theta^4+6\theta^5+\theta^6 \end{pmatrix} \\ & +(1-\theta)^2\kappa_\lambda^6\left(\theta+4\theta^2+4\theta^3-\theta^5\right) \end{aligned}$,	$\begin{aligned} \tilde{P}^{(2)}_{1,1}(\theta,\kappa) = & -4\kappa_\lambda^2\theta^2\left(1+2\theta-\theta^2-2\theta^3+2\theta^4\right) \\ & +(1-\theta)^2\kappa_\lambda^4\begin{pmatrix} 1+4\theta+4\theta^2-6\theta^3 \\ -11\theta^4+2\theta^5+2\theta^6 \end{pmatrix} \\ & +(1-\theta)^4\kappa_\lambda^6\theta(1+\theta)^2(1+2\theta) \end{aligned}$,
$\begin{aligned} \tilde{P}^{(1)}_{1,2}(\theta,\kappa) = & -4\kappa_\lambda^4\theta^2\left(1+2\theta-\theta^3\right) \\ & +(1-\theta)\kappa_\lambda^6\begin{pmatrix} 1+3\theta+\theta^2-8\theta^3 \\ -11\theta^4+\theta^5+\theta^6 \end{pmatrix} \\ & +(1-\theta)^3\kappa_\lambda^8\theta(1+\theta)^2(1+2\theta) \end{aligned}$,	$\begin{aligned} \tilde{P}^{(2)}_{1,2}(\theta,\kappa) = & \kappa_\lambda^2 4\theta^4\left(1+2\theta-\theta^3\right) \\ & -(1-\theta)\kappa_\lambda^4\theta^2\begin{pmatrix} 5+15\theta+5\theta^2-20\theta^3 \\ -11\theta^4+9\theta^5+\theta^6 \end{pmatrix} \\ & +(1-\theta)^3\kappa_\lambda^6\begin{pmatrix} 1+5\theta+8\theta^2-3\theta^3 \\ -21\theta^4-14\theta^5 \\ +2\theta^6+2\theta^7 \end{pmatrix} \\ & +(1-\theta)^5\kappa_\lambda^8\theta(1+\theta)^3(1+2\theta) \end{aligned}$,
$\begin{aligned} \tilde{P}^{(1)}_{2,2}(\theta,\kappa) = & -4\kappa_\lambda^6\theta^2\left(1+2\theta-\theta^2-2\theta^3+2\theta^4\right) \\ & +(1-\theta)^2\kappa_\lambda^8\begin{pmatrix} 1+4\theta+4\theta^2-6\theta^3 \\ -11\theta^4+2\theta^5+2\theta^6 \end{pmatrix} \\ & +(1-\theta)^4\kappa_\lambda^{10}\theta(1+\theta)^2(1+2\theta) \end{aligned}$,	$\begin{aligned} \tilde{P}^{(2)}_{2,2}(\theta,\kappa) = & \kappa_\lambda^2 4\theta^2\left(1+\theta\right)^2\left(-1-2\theta+2\theta^3\right) \\ & +\kappa_\lambda^4\begin{pmatrix} 1+4\theta+3\theta^2-20\theta^3 \\ -45\theta^4-2\theta^5+53\theta^6 \\ +20\theta^7-20\theta^8-2\theta^9 \end{pmatrix} \\ & +(1-\theta)^2\kappa_\lambda^6\theta\begin{pmatrix} 3+14\theta+20\theta^2 \\ -8\theta^3-47\theta^4 \\ -30\theta^5+4\theta^6+4\theta^7 \end{pmatrix} \\ & +(1-\theta)^4\kappa_\lambda^8 2\theta^2(1+\theta)^3(1+2\theta) \end{aligned}$,

Table 6: Polynomials involved in the description of the equilibrium covariance matrix $\tilde{\Sigma}^\infty$ of $\lim_{n\to\infty}[x_{n-1}, y_n]$.

$$
\begin{aligned}
P_{1,1}^{(\infty,1)}(\theta,\kappa) = {} & -4\kappa^2\theta^4(1+\theta)^2 \\
& + \kappa^4 \begin{pmatrix} \theta^2 + 10\theta^3 + 7\theta^4 - 24\theta^5 \\ -17\theta^6 + 14\theta^7 + \theta^8 \end{pmatrix} \\
& - \kappa^6(1-\theta)^2\theta \begin{pmatrix} 2 + 10\theta + 9\theta^2 - 24\theta^3 \\ -34\theta^4 + 10\theta^5 + 3\theta^6 \end{pmatrix} \\
& + \kappa^8(1-\theta)^4 \begin{pmatrix} 1 + 4\theta + 2\theta^2 - 14\theta^3 - 21\theta^4 \\ -2\theta^5 + 2\theta^6 \end{pmatrix} \\
& + \kappa^{10}(1-\theta)^6\theta(1+\theta)^2(1+2\theta)
\end{aligned}
\quad , \quad
\begin{aligned}
P_{1,1}^{(\infty,2)}(\theta,\kappa) = {} & \kappa^2 \begin{pmatrix} -4\theta^2 - 8\theta^3 + 4\theta^4 \\ +8\theta^5 - 8\theta^6 \end{pmatrix} \\
& + \kappa^4(1-\theta)^2 \begin{pmatrix} 1 + 4\theta + 4\theta^2 - 6\theta^3 \\ -11\theta^4 + 2\theta^5 + 2\theta^6 \end{pmatrix} \\
& + \kappa^6(1-\theta)^4\theta(1+\theta)^2(1+2\theta)
\end{aligned}
\quad ,
$$

$$
\begin{aligned}
P_{1,2}^{(\infty,1)}(\theta,\kappa) = {} & +4\kappa^4\theta^3(-1-2\theta+\theta^3) \\
& + \kappa^6(1-\theta)\theta \begin{pmatrix} 1 + 3\theta + 5\theta^2 \\ -15\theta^4 - 7\theta^5 + 9\theta^6 \end{pmatrix} \\
& - \kappa^8(1-\theta)^3\theta \begin{pmatrix} 1 + 3\theta - 11\theta^3 \\ -13\theta^4 + 2\theta^5 + 2\theta^6 \end{pmatrix} \\
& - \kappa^{10}(1-\theta)^5\theta^2(1+\theta)^2(1+2\theta)
\end{aligned}
\quad , \quad
\begin{aligned}
P_{1,2}^{(\infty,2)}(\theta,\kappa) = {} & \kappa^2(4\theta^3 + 12\theta^4 + 8\theta^5 - 12\theta^6 - 16\theta^7 + 4\theta^8 + 8\theta^9), \\
& + \kappa^4(1-\theta)\theta \begin{pmatrix} -1 - 4\theta - 8\theta^2 + 5\theta^3 \\ +40\theta^4 + 22\theta^5 - 42\theta^6 \\ -29\theta^7 + 19\theta^8 + 2\theta^9 \end{pmatrix} \\
& + \kappa^6(1-\theta)^3 \begin{pmatrix} \theta + 2\theta^2 - 6\theta^3 - 23\theta^4 \\ -13\theta^5 + 33\theta^6 \\ +32\theta^7 - 2\theta^8 - 4\theta^9 \end{pmatrix} \\
& + \kappa^8(1-2\theta)(1-\theta)^5\theta^2(1+\theta)^3(1+2\theta)
\end{aligned}
$$

$$
\begin{aligned}
P_{2,2}^{\infty,1}(\theta,\kappa) = {} & -4\kappa^6\theta^2\left(1 + 2\theta - \theta^2 - 2\theta^3 + 2\theta^4\right) \\
& + (1-\theta)^2\kappa^8 \begin{pmatrix} 1 + 4\theta + 4\theta^2 - 6\theta^3 \\ -11\theta^4 + 2\theta^5 + 2\theta^6 \end{pmatrix} \\
& + (1-\theta)^4\kappa^{10}\theta(1+\theta)^2(1+2\theta)
\end{aligned}
\quad , \quad
\begin{aligned}
P_{2,2}^{\infty,2}(\theta,\kappa) = {} & \kappa^2 4\theta^2\left(1+\theta\right)^2\left(-1 - 2\theta + 2\theta^3\right) \\
& + \kappa^4 \begin{pmatrix} 1 + 4\theta + 3\theta^2 - 20\theta^3 - 45\theta^4 - 2\theta^5 \\ +53\theta^6 + 20\theta^7 - 20\theta^8 - 2\theta^9 \end{pmatrix} \\
& + (1-\theta)^2\kappa^6\theta \begin{pmatrix} 3 + 14\theta + 20\theta^2 - 8\theta^3 \\ -47\theta^4 - 30\theta^5 + 4\theta^6 + 4\theta^7 \end{pmatrix} \\
& + (1-\theta)^4\kappa^8 2\theta^2(1+\theta)^3(1+2\theta)
\end{aligned}
\quad ,
$$

$$
\begin{aligned}
P_c(\theta,\kappa) = {} & -4\kappa^2\theta^2\left(1+\theta\right)^3 + \kappa^4\left(1 + 3\theta + \theta^2 - 17\theta^3 - 33\theta^4 - 3\theta^5 + 15\theta^6 + \theta^7\right). \\
& + (1-\theta)\kappa^6\theta\left(3 + 14\theta + 13\theta^2 - 24\theta^3 - 35\theta^4 + 10\theta^5 + 3\theta^6\right) \\
& + (1-\theta)^3\kappa^8\left(-1 - 4\theta - 2\theta^2 + 14\theta^3 + 21\theta^4 + 2\theta^5 - 2\theta^6\right) \\
& - (1-\theta)^5\kappa^{10}\theta(1+\theta)^2(1+2\theta)
\end{aligned}
$$

Table 7: Polynomials involved in the description of the equilibrium covariance matrices Σ^{∞} of $\lim_{n\to\infty}[x_n, y_n]$ (bottom).