High Probability Convergence Guarantees of Stochastic Gradient Descent Ascent in Structured Nonconvex Min-Max Games

Junsoo Ha JUNSOO.HA.CONTACT@GMAIL.COM
Independent Researcher

Editors: Steve Hanneke and Tor Lattimore

Abstract

Nonconvex min-max optimization is a cornerstone of modern machine learning. However, its theoretical foundations remain largely limited to in-expectation convergence guarantees, which fail to capture the failure probability of individual training trajectories, particularly in the presence of heavy-tailed noise. In this work, we bridge this gap by establishing the first high-probability convergence guarantees of stochastic gradient descent-ascent (SGDA) in structured nonconvex games, specifically nonconvex-PL (NC-PL) and nonconvex-concave (NC-C) problems. We derive high-probability convergence rates of SGDA matching the best known in-expectation rates in the sub-gaussian noise regime. Then, we investigate the heavy-tailed noise regime and prove that SGDA cannot guarantee high-probability convergence in general. Finally, we analyze a gradient-clipped variant, $\text{SGDA}_{\text{Clip}}$, and show that it recovers high-probability convergence guarantees in both NC-PL and NC-C games. Our analysis is based on novel progress quantities that simultaneously bound stationarity and primal-dual martingale terms, which yield self-bounding concentration bounds.

Keywords: min-max games, stochastic gradient descent-ascent, high probability convergence

1. Introduction

Nonconvex min-max games have become a cornerstone of modern machine learning, underpinning the success of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), adversarial training (Madry et al., 2018), and multi-agent reinforcement learning (MARL) (Wu et al., 2025). A key workhorse algorithm for solving min-max games is stochastic gradient descent-ascent (SGDA) Lin et al. (2019), a generalization of stochastic gradient descent (SGD) for min-max problems. The prevailing theoretical landscape, however, is dominated by in-expectation guarantees (Gidel et al., 2019; Lin et al., 2019; Yang et al., 2022; Cho and Yun, 2023), except for a few works that focus on variational inequalities (Gorbunov et al., 2022b; Sadiev et al., 2023) or specialized algorithms under light-tailed noise distributions (Laguel et al., 2024). While in-expectation results provide coarse guarantees, they fail to capture the inherent volatility in individual training. Empirical observations also suggest neural networks often encounter heavy-tailed noise (Simsekli et al., 2019; Zhang et al., 2020), which can lead to substantial deviations from in-expectation behavior. The fact that in-expectation results offer no assurance against failures in each run is a critical deficit for large-scale training (Wu et al., 2025). To bridge this gap, one must move beyond in-expectation arguments and establish high probability guarantees.

Contributions. We establish the first high-probability convergence guarantees of SGDA in structured nonconvex min-max games, specifically nonconvex-PL (NC-PL) and nonconvex-concave (NC-C) problems. Our main contributions are two-fold:

- We prove high-probability convergence guarantees of SGDA in NC-PL and NC-C games under a subgaussian noise model, and match the best known in-expectation rates.
- We demonstrate that while SGDA can lose general high-probability guarantees under a heavy-tailed noise model, its gradient-clipped variant, $\text{SGDA}_{\text{Clip}}$, can recover high-probability guarantees even under heavy-tailed noise.

The key technical ingredient of our analysis is a two-player self-bounding concentration architecture for SGDA. Unlike SGD, the stochastic terms in SGDA couple primal and dual errors and are adapted to different stages of the half-step filtration. We therefore design progress quantities that simultaneously dominate stationarity, dual suboptimality, and the martingale variance proxies. Under heavy-tailed noise, clipping introduces additional bias terms whose bounds are valid only when both players remain in a low-signal regime; we close this circularity through a pathwise normalization/bootstrap argument. In NC-C games, the same idea must also be combined with a blockwise analysis of the dual maximizer and partition-level martingale control. We defer all proofs to Appendix B and Appendix C, and summarize our results in Table 1 and Table 2.

2. Related Work

Convex–concave games and monotone variational inequalities are now classic and well-studied (Nemirovski, 2004; Juditsky et al., 2011; Du and Hu, 2019; Gidel et al., 2019; Azizian et al., 2020b,a; Mokhtari et al., 2020; Golowich et al., 2020; Gorbunov et al., 2022c,d). The most well-studied algorithms for convex–concave games include extragradient (EG) (Korpelevich, 1976b) and optimistic gradient descent ascent (OGDA) (Popov, 1980; Daskalakis et al., 2018; Gidel et al., 2019; Mokhtari et al., 2020). Recent work established their last-iterate convergence rates (Golowich et al., 2020; Gorbunov et al., 2022c,d), and their in-expectation rates of stochastic counterparts (Gidel et al., 2019; Hsieh et al., 2019; Mishchenko et al., 2020; Gorbunov et al., 2022a). However, while EG and OGDA have shown strong theoretical guarantees, SGDA-type algorithms remain main workhorses in practice (Cheng et al., 2024a,b) due to their simplicity and efficiency, and have been shown to exhibit strong theoretical guarantees as well (Zhang et al., 2022; Beznosikov et al., 2023). For instance, Zhang et al. (2022) proved that SGDA can enjoy the same optimal local convergence rate as EG and OGDA in strongly-convex–strongly-concave games. Beznosikov et al. (2023) showed that SGDA can converge in monotone variational inequalities, and enjoys similar rates as EG and OGDA under quasi-strong monotonicity and star-cocoercivity. These results set baselines for convex regimes, but they do not address the nonconvexity studied in this paper.

Nonconvex games are computationally intractable in general (Papadimitriou, 1994; Daskalakis et al., 2006), and often exhibit cycles that induce non-convergence of iterative algorithms (Mertikopoulos et al., 2018). Such computational hardness motivated the study of *structured* nonconvex games, where one imposes computationally tractable structures over one of the players. The most widely studied structured nonconvex games are nonconvex–PL (NC-PL) games (Nouiehed et al., 2019; Yang et al., 2022; Cho and Yun, 2023; Huang et al., 2025) and nonconvex–concave (NC-C) games (Lin et al., 2019). NC-PL and NC-C games assume a Polyak–Łojasiewicz (PL) condition and concavity in the maximization variable, respectively, and both admit efficient algorithms with guarantees to approximate stationary points. For instance, Yang et al. (2020, 2022) established the first in-expectation convergence guarantees of SGDA in NC-PL games, and Lin et al. (2019) provided the first in-expectation convergence guarantees of SGDA in NC-C games. Subsequent

works improved upon these results using without-replacement sampling (Cho and Yun, 2023) and a momentum technique (Huang et al., 2025). Nevertheless, all these works focus on in-expectation convergence guarantees, and do not provide high-probability guarantees over individual runs.

High-probability guarantees are less studied in min-max optimization, especially in structured nonconvex min-max games. Laguel et al. (2024) is one of the few works that establish high-probability convergence guarantees in NC-PL games, but focuses on a smoothed variant of SGDA under subgaussian noise. Gorbunov et al. (2022b); Sadiev et al. (2023) study high-probability convergence guarantees of clipped variants of stochastic extragradient methods under both subgaussian and heavy-tailed noise models, but their results focus on variational inequalities. In this work, we fill this gap by establishing the first high-probability convergence guarantees of (clipped) SGDA in both NC-PL and NC-C games under subgaussian and heavy-tailed noise.

3. Preliminaries

3.1. Structured Nonconvex Games

Throughout this paper, we make the following standard smoothness assumption on the game f.

Assumption 3.1 (Smoothness) *The game $f : \mathbb{R}^{d_x} \times \mathcal{Y} \to \mathbb{R}$ is differentiable and ℓ-smooth as:*

$$\|\nabla_x f(x_1, y_1) - \nabla_x f(x_2, y_2)\| \le \ell(\|x_1 - x_2\| + \|y_1 - y_2\|), \quad \forall x_1, x_2 \in \mathbb{R}^{d_x}, \forall y_1, y_2 \in \mathcal{Y}$$
$$\|\nabla_y f(x_1, y_1) - \nabla_y f(x_2, y_2)\| \le \ell(\|x_1 - x_2\| + \|y_1 - y_2\|), \quad \forall x_1, x_2 \in \mathbb{R}^{d_x}, \forall y_1, y_2 \in \mathcal{Y}.$$

We refer to the setting where the function $y \mapsto -f(x, y)$ satisfies a PL inequality as *nonconvex–PL* (NC-PL) games, and define the condition number $\kappa \overset{\text{def}}{=} \ell/\mu$, where $\mu > 0$ is a PL parameter.

Assumption 3.2 (NC-PL) *Let $\mathcal{Y} = \mathbb{R}^{d_y}$ and $g(x) \overset{def}{=} \max_{y \in \mathcal{Y}} f(x, y) < \infty$. $\mathcal{Y}^\star(x) \overset{def}{=} \arg\max_{y \in \mathcal{Y}} f(x, y)$ is nonempty for every $x \in \mathbb{R}^{d_x}$. Moreover, there exists $\mu > 0$ such that, for each $x \in \mathbb{R}^{d_x}$,*

$$\frac{1}{2}\|\nabla_y f(x, y)\|^2 \;\ge\; \mu\big(g(x) - f(x, y)\big), \;\; \forall y \in \mathcal{Y}, \tag{1}$$

We refer to the setting where the function $y \mapsto f(x, y)$ is concave as *nonconvex–concave* (NC-C).

Assumption 3.3 (NC-C) *For each $x \in \mathbb{R}^{d_x}$, $f(x, \cdot)$ is concave on $\mathcal{Y}$, and there exists a maximizer $y^*(x) \in \arg\max_{y \in \mathcal{Y}} f(x, y)$ with $\nabla_y f(x, y^*(x)) = 0$. For each $y \in \mathcal{Y}$, the function $f(\cdot, y)$ is L-Lipschitz on $\mathbb{R}^{d_x}$. The domain $\mathcal{Y}$ is convex and closed with $D \overset{def}{=} \operatorname{diam}(\mathcal{Y}) < \infty$.*

3.2. Algorithms

SGDA We define two-time-scale stochastic alternating gradient descent-ascent (SGDA) as:

$$x_{t+1} = x_t - \eta_{x,t}\, G_x(x_t, y_t, \zeta_t^x), \quad y_{t+1} = \Pi_{\mathcal{Y}}\big(y_t + \eta_{y,t}\, G_y(x_{t+1}, y_t, \zeta_t^y)\big), \tag{2}$$

where $\Pi_{\mathcal{Y}}(\cdot)$ denotes the orthogonal projection from $\mathbb{R}^{d_y}$ to $\mathcal{Y}$. G_x, G_y are stochastic gradient oracles and ζ_t^x, ζ_t^y are random seeds. We focus on two-time-scale regime where y moves faster than x, *i.e.*, $0 < \eta_{x,t} \ll \eta_{y,t}$. We omit the random seeds ζ_t^x, ζ_t^y when they are clear from the context.

SGDA$_{\text{Clip}}$ We define a gradient-clipped variant of SGDA as:

$$x_{t+1} = x_t - \eta_{x,t}\ \mathrm{clip}_{\tau_x}\left(G_x(x_t, y_t, \zeta_t^x)\right), \quad y_{t+1} = \Pi_{\mathcal{Y}}\big(y_t + \eta_{y,t}\ \mathrm{clip}_{\tau_y}\left(G_y(x_{t+1}, y_t, \zeta_t^y)\right)\big). \quad (3)$$

where $\mathrm{clip}_\tau : \mathbb{R}^d \to \mathbb{R}^d$ denotes the standard norm clipping $\mathrm{clip}_\tau(v) := \begin{cases} v, & \|v\| \le \tau, \\ \tau\, v/\|v\|, & \|v\| > \tau. \end{cases}$

3.3. Noise Models

We define the stochastic gradient noise of each player at time $t \ge 0$ as:

$$\xi_t^x \overset{\text{def}}{=} G_x(x_t, y_t, \zeta_t^x) - \nabla_x f(x_t, y_t), \quad \xi_t^y \overset{\text{def}}{=} G_y(x_{t+1}, y_t, \zeta_t^y) - \nabla_y f(x_{t+1}, y_t), \quad (4)$$

and define the natural filtrations $(\mathcal{F}_t)_{t\ge0} \subseteq (\mathcal{F}_{t+1/2})_{t\ge0}$ as:

$$\mathcal{F}_t \overset{\text{def}}{=} \sigma\left(\{x_k, y_k \mid k = 0, \dots, t\}\right), \quad \mathcal{F}_{t+1/2} \overset{\text{def}}{=} \sigma\left(\{x_{t+1}\} \cup \mathcal{F}_t\right).$$

Throughout the paper, we make the following standard unbiasedness assumption.

Assumption 3.4 (Unbiased stochastic gradients) *The gradient noise $\xi_t^x \in \mathbb{R}^{d_x}, \xi_t^y \in \mathbb{R}^{d_y}$ satisfy:*

$$\mathbb{E}\left[\xi_t^x \mid \mathcal{F}_t\right] = 0, \quad \mathbb{E}\left[\xi_t^y \mid \mathcal{F}_{t+1/2}\right] = 0, \quad \forall t \ge 0$$

The following two are standard models in the literature: norm-subgaussian and heavy-tailed noise.

Assumption 3.5 (Norm-subgaussian noise) *There exists $\sigma > 0$ such that, for any $t \ge 0$, the stochastic gradient noise $\xi_t^x \in \mathbb{R}^{d_x}, \xi_t^y \in \mathbb{R}^{d_y}$ satisfy*

$$\mathbb{E}\left[\exp\left(\frac{\|\xi_t^x\|^2}{\sigma^2}\right) \middle| \mathcal{F}_t\right] \le e, \qquad \mathbb{E}\left[\exp\left(\frac{\|\xi_t^y\|^2}{\sigma^2}\right) \middle| \mathcal{F}_{t+1/2}\right] \le e.$$

Assumption 3.6 (Heavy-tailed noise) *For each $p \in (1, 2]$, there exists $\sigma > 0$ such that, for any $t \ge 0$, the gradient noise of each player $\xi_t^x \in \mathbb{R}^{d_x}, \xi_t^y \in \mathbb{R}^{d_y}$ satisfy:*

$$\mathbb{E}\left[\|\xi_t^x\|^p \mid \mathcal{F}_t\right] \le \sigma^p, \quad \mathbb{E}\left[\|\xi_t^y\|^p \mid \mathcal{F}_{t+1/2}\right] \le \sigma^p. \quad (5)$$

4. Nonconvex–PL Games

We follow Yang et al. (2022) and define a potential function P_t as:

$$P_t \overset{\text{def}}{=} a_t + \lambda_t b_t, \quad a_t \overset{\text{def}}{=} g(x_t) - g_*, \quad b_t \overset{\text{def}}{=} g(x_t) - f(x_t, y_t), \quad (6)$$

with a tunable $\lambda_t \ge 0$ and $g_* \overset{\text{def}}{=} \min_{x\in\mathbb{R}^{d_x}} g(x)$. Intuitively, a_t and b_t represent the primal and dual suboptimalities at time t, respectively. We fix $\lambda_t = \frac{1}{2}$ throughout our NC-PL analysis.

Table 1: Comparison of existing convergence guarantees in NC-PL games. Complexity is measured by the number of gradient evaluations to achieve an ε-stationary point, *i.e.*, $\min_{t\in[0,T)} \|\nabla g(x_t)\| \le \varepsilon$. Time horizon indicates whether the algorithm requires knowledge of a time horizon T in advance. H.P. stands for high-probability guarantees. Our results in Theorems 4.5, 4.6, 4.11 and 4.12 establish the first high-probability convergence guarantees of SGDA in NC-PL games under subgaussian and heavy-tailed noise models.

ALGORITHM	CITATION	NOISE MODEL	COMPLEXITY	TIME HORIZON	H.P.
SGDA	YANG ET AL. (2022)	BOUNDED σ^2	$\mathcal{O}(\kappa^4\ell\varepsilon^{-4})$	KNOWN	✗
SGDA-RR	CHO AND YUN (2023)	FINITE-SUM	$\mathcal{O}(\kappa^3\ell n^{0.5}\epsilon^{-3})$	KNOWN	✗
MSGDA	HUANG ET AL. (2025)	BOUNDED σ^2	$\tilde{\mathcal{O}}(\varepsilon^{-3})$[1]	KNOWN	✗
ADAMSGDA	HUANG ET AL. (2025)	BOUNDED σ^2	$\tilde{\mathcal{O}}(\varepsilon^{-3})$[1]	KNOWN	✗
SM-SGDA	YANG ET AL. (2022)	BOUNDED σ^2	$\mathcal{O}(\kappa^2\ell\varepsilon^{-4})$	KNOWN	✗
SM-SGDA	LAGUEL ET AL. (2024)	SUBGAUSSIAN	$\mathcal{O}(\kappa^2\ell\varepsilon^{-4})$	KNOWN	✓
SGDA	THEOREM 4.5	SUBGAUSSIAN	$\mathcal{O}(\kappa^4\ell\varepsilon^{-4})$	KNOWN	✓
SGDA	THEOREM 4.6	SUBGAUSSIAN	$\tilde{\mathcal{O}}(\kappa^4\ell\varepsilon^{-4})$	UNKNOWN	✓
$\text{SGDA}_{\text{CLIP}}$	THEOREM 4.11	HEAVY-TAILED	$\mathcal{O}\left(\kappa^{\frac{3p-2}{p-1}}\ell\varepsilon^{-\frac{3p-2}{p-1}}\right)$	KNOWN	✓
$\text{SGDA}_{\text{CLIP}}$	THEOREM 4.12	HEAVY-TAILED	$\tilde{\mathcal{O}}\left(\kappa^{\frac{3p-2}{p-1}}\ell\varepsilon^{-\frac{3p-2}{p-1}}\right)$	UNKNOWN	✓

4.1. Subgaussian NC-PL Games

Following (Nouiehed et al., 2019; Lin et al., 2019), we interpret SGDA as an inexact descent on the outer objective g, and obtain the following estimate of primal progress a_t.

Lemma 4.1 (Primal progress of SGDA in NC-PL) *Under Assumptions 3.1 and 3.2, if $\eta_{x,t} \le 1/(12\kappa\ell)$, SGDA iterates satisfy the following for all $t \ge 0$:*

$$a_{t+1} \le a_t - \frac{\eta_{x,t}}{2}\|\nabla g(x_t)\|^2 + 4\eta_{x,t}\kappa\ell b_t - \eta_{x,t}\langle \nabla g(x_t), \xi_t^x\rangle + 3\eta_{x,t}^2\kappa\ell\|\xi_t^x\|^2. \tag{7}$$

To control the potential P_t, we also need to estimate the dual progress b_t. We define the shorthands:

$$\delta_t \overset{\text{def}}{=} \nabla_x f(x_t, y_t) - \nabla g(x_t), \quad \phi_t(y) \overset{\text{def}}{=} g(x_{t+1}) - f(x_{t+1}, y), \quad b_{t+1/2} \overset{\text{def}}{=} \phi_t(y_t).$$

Namely, δ_t is the bias of $\nabla_x f(x_t, y_t)$ as an estimate of $\nabla g(x_t)$, and $b_{t+1/2}$ is the half-step dual gap.

Lemma 4.2 (Dual progress of SGDA in NC-PL) *Suppose Assumptions 3.1 and 3.2, and let $\alpha_t \overset{def}{=} 1 + 4\kappa\ell\eta_{x,t} + 12\kappa^2\ell^2\eta_{x,t}^2$. Then, SGDA satisfies the following $\forall t \ge 0$:*

$$b_{t+1} \le b_{t+1/2} - (\eta_{y,t} - \frac{\ell}{2}\eta_{y,t}^2)\|\nabla\phi_t(y_t)\|^2 + (\eta_{y,t} - \ell\eta_{y,t}^2)\langle\nabla\phi_t(y_t), \xi_t^y\rangle + \frac{\ell}{2}\eta_{y,t}^2\|\xi_t^y\|^2, \tag{8}$$

$$b_{t+1/2} \le \alpha_t b_t + (\frac{1}{4}\eta_{x,t} + 6\kappa\ell\eta_{x,t}^2)\|\nabla g(x_t)\|^2 + \eta_{x,t}\langle\delta_t, \xi_t^x\rangle + 6\kappa\ell\eta_{x,t}^2\|\xi_t^x\|^2. \tag{9}$$

1. Huang et al. (2025) state the $\tilde{\mathcal{O}}(\varepsilon^{-3})$ rate, but do not clearly expose the dependence on κ and ℓ.

A direct combination of Lemma 4.1 and Lemma 4.2 yields the following potential estimate.

Lemma 4.3 (Potential improvement of SGDA in NC-PL) *Suppose Assumptions 3.1 and 3.2, and let* $\gamma_{y,t} \overset{def}{=} \frac{1}{4}\left(\eta_{y,t} - \frac{\ell}{2}\eta_{y,t}^2\right)$. *If* $\eta_{y,t} \le \frac{1}{\ell}$, $\eta_{x,t} \le \frac{1}{64\kappa^2}\eta_{y,t}$, *then, SGDA satisfies the following* $\forall t \ge 0$*:*

$$\frac{1}{8}\eta_{x,t}\left\|\nabla g(x_t)\right\|^2 + \gamma_{y,t}\left\|\nabla \phi_t(y_t)\right\|^2 + 4\kappa\ell\eta_{x,t}b_t \le P_t - P_{t+1} \tag{10}$$

$$- \eta_{x,t}\left\langle \nabla g(x_t), \xi_t^x\right\rangle + \left(\frac{1}{2} - 2\mu\gamma_{y,t}\right)\eta_{x,t}\left\langle \delta_t, \xi_t^x\right\rangle + 6\kappa\ell\eta_{x,t}^2\left\|\xi_t^x\right\|^2 \tag{11}$$

$$+ \frac{1}{2}(\eta_{y,t} - \ell\eta_{y,t}^2)\left\langle \nabla\phi_t(y_t), \xi_t^y\right\rangle + \frac{\ell}{4}\eta_{y,t}^2\left\|\xi_t^y\right\|^2. \tag{12}$$

A naive approach to derive a high-probability bound from Lemma 4.3 would be directly invoking a standard concentration inequality, *e.g.*, Azuma–Hoeffding, to the martingale difference terms ξ_t^x, ξ_t^y. However, each term on the left-hand side of Lemma 4.3 is in fact an upper bound of variance proxies of ξ_t^x, ξ_t^y terms, hence prevents a direct application of standard concentration inequalities. To tackle this problem, we follow the recently adopted self-bounding concentration inequalities in optimization literature (Harvey et al., 2019; Nguyen et al., 2023; Liu et al., 2023). The following lemma identifies and exploits the self-bounding property of the martingale difference terms in Lemma 4.3.

Lemma 4.4 (Self-bounding subgaussian martingales) *Assume martingale differences* $\xi_t^x \in \mathbb{R}^{d_x}, \xi_t^y \in \mathbb{R}^{d_y}$ *satisfy Assumptions 3.4 and 3.5. Let* $\eta_{x,t}, \eta_{y,t} \ge 0$ *be deterministic. Let* $d_t^x, d_t^y \ge 0$ *be deterministic. Suppose we have nonnegative processes* $G_t, P_t \ge 0$ *and real-valued random variables* r_t *such that, for any* $t \ge 0$*, the following holds almost surely:*

$$G_t \le P_t - P_{t+1} + \eta_{x,t}\left\langle c_t^x, \xi_t^x\right\rangle + d_t^x\eta_{x,t}^2\left\|\xi_t^x\right\|^2 + \eta_{y,t}\left\langle c_t^y, \xi_t^y\right\rangle + d_t^y\eta_{y,t}^2\left\|\xi_t^y\right\|^2 + r_t. \tag{13}$$

Additionally, suppose that $c_t^x \in \mathbb{R}^{d_x}$ *is* $\mathcal{F}_t$*-measurable and* $c_t^y \in \mathbb{R}^{d_y}$ *is* $\mathcal{F}_{t+1/2}$*-measurable. No measurability condition is imposed on* r_t*; it is carried pathwise. Assume the self-bounding property*

$$\eta_{x,t}\left\|c_t^x\right\|^2 + \eta_{y,t}\left\|c_t^y\right\|^2 \le C_1 G_t, \quad \forall t \ge 0, \tag{14}$$

where $C_1 \ge 0$ *is some constant. Then, for any* $T \ge 1$ *and* $\delta \in (0,1)$*, with probability at least* $1-\delta$*:*

$$\sum_{t=0}^{T-1} G_t \le 2P_0 + 2\sum_{t=0}^{T-1} r_t + 16\sigma^2 C_1 \eta_{\max}\Gamma_\delta + 4\sigma^2 S_2(T)\Gamma_\delta \tag{15}$$

holds with $S_2(T) \overset{def}{=} \sum_{t=0}^{T-1}(d_t^x\eta_{x,t}^2 + d_t^y\eta_{y,t}^2)$, $\eta_{\max} \overset{def}{=} \max_{t\in[0,T)}\max\{\eta_{x,t}, \eta_{y,t}\}$, $\Gamma_\delta \overset{def}{=} \max\{1, \log(2/\delta)\}$.

The key point is that the left-hand side of Lemma 4.3 is engineered to be both algorithmic progress and a variance proxy upper bound of the stochastic terms ξ_t^x and ξ_t^y. In particular, it simultaneously controls the primal gradient $\nabla g(x_t)$, the dual gradient $\nabla\phi(y_t)$, and the gradient mismatch b_t. This two-player self-bounding closure is what allows high-probability control of SGDA through Lemma 4.4. Combining Lemma 4.3 and Lemma 4.4 with suitable step size schedules, we obtain our main high-probability convergence guarantees of SGDA in NC-PL games under subgaussian noise.

Theorem 4.5 (Convergence of SGDA in subgaussian NC-PL with fixed step sizes) *Suppose Assumptions 3.1, 3.2, 3.4 and 3.5 hold. Let $T \geq 1$, $\delta \in (0,1)$, and $\Gamma_\delta \overset{def}{=} \max\{1, \log(2/\delta)\}$. Then, if $\eta_{x,t} = \min\left\{\frac{1}{64\kappa^2\ell}, \sqrt{\frac{P_0}{A_{\kappa,\ell}\sigma^2 T\Gamma_\delta}}\right\}$, $\eta_{y,t} = 64\kappa^2\eta_x$, the following holds for SGDA with probability at least $1-\delta$:*

$$\min_{t\in[0,T)} \|\nabla g(x_t)\|^2 = \mathcal{O}\left(\frac{\kappa^2\ell P_0}{T} + \frac{\kappa^2\sigma\sqrt{\ell P_0\Gamma_\delta}}{\sqrt{T}} + \frac{\kappa^2\sigma^2\Gamma_\delta}{T}\right)$$

Theorem 4.6 (Convergence of SGDA in subgaussian NC-PL with decaying step sizes) *Suppose Assumptions 3.1, 3.2, 3.4 and 3.5. Let $T \geq 1$, $\delta \in (0,1)$, and $\Gamma_\delta \overset{def}{=} \max\{1, \log(2/\delta)\}$. Then, if $\eta_{x,t} = \min\left\{\frac{1}{64\kappa^2\ell}, \sqrt{\frac{P_0}{(6\kappa+1024\kappa^4)\ell\sigma^2\Gamma_\delta(t+1)}}\right\}$, $\eta_{y,t} = 64\kappa^2\eta_{x,t}$, the following holds for SGDA with probability at least $1-\delta$:*

$$\min_{t\in[0,T)} \|\nabla g(x_t)\|^2 = \mathcal{O}\left(\frac{\kappa^2\ell P_0}{T} + \frac{\kappa^2\sigma\sqrt{\ell P_0\Gamma_\delta}(1+\log T)}{\sqrt{T}} + \frac{\kappa^2\sigma^2\Gamma_\delta}{\sqrt{T}}\right)$$

4.2. Heavy-Tailed NC-PL Games

Recent observations suggest that modern training often encounters heavy-tailed noise (Zhang et al., 2020; Gorbunov et al., 2022b). Hence, a natural question would be: can we establish general high-probability guarantees even under heavy-tailed noise? The following result gives a negative answer.

Proposition 4.7 (Impossibility of a general high-probability convergence guarantee of SGDA) *Fix any $p \in (1,2]$. There exists a function $f : \mathbb{R}^{d_x} \times \mathcal{Y} \to \mathbb{R}$ satisfying Assumptions 3.1 and 3.2, with $f(x,\cdot)$ strongly concave, such that the following holds. For any deterministic step size schedules $\eta_{x,t}, \eta_{y,t} \geq 0$, any horizon $T \geq 2$ with at least one active x-update before time T,*

$$\exists\, t \in \{0,\dots,T-2\} \text{ such that } \eta_{x,t} > 0,$$

and any confidence level $\delta \in (0,1)$, there exists a stochastic gradient oracle satisfying Assumptions 3.4 and 3.6 with exponent p for which the corresponding SGDA iterates satisfy

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla g(x_t)\|^2 \geq \frac{1}{\delta T}\right) \geq \delta.$$

Consequently, any general high-probability guarantee of the form

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla g(x_t)\|^2 \leq \varepsilon\right) \geq 1-\delta$$

for this class of oracles must satisfy $T = \Omega\left(\frac{1}{\varepsilon\delta}\right)$.

However, Theorem 4.7 only applies to SGDA, and it has been shown that *gradient clipping* can efficiently counteract extreme fluctuations of heavy-tailed noise in convex optimization and variational inequalities (Gorbunov et al., 2022b; Sadiev et al., 2023; Nguyen et al., 2023). We show that the

same holds for NC-PL games, and $\text{SGDA}_{\text{Clip}}$ can be tuned to counteract such heavy-tailed noise. The key technical challenge lies in controlling the clipping biases. Define:

$$\begin{aligned}
\tilde{G}_t^x &\overset{\text{def}}{=} \text{clip}_{\tau_{x,t}}(G_x(x_t, y_t)), & \tilde{G}_t^y &\overset{\text{def}}{=} \text{clip}_{\tau_{y,t}}(G_y(x_{t+1}, y_t)),\\
\tilde{\xi}_t^x &\overset{\text{def}}{=} \tilde{G}_t^x - \mathbb{E}[\tilde{G}_t^x \mid \mathcal{F}_t], & \beta_t^x &\overset{\text{def}}{=} \mathbb{E}[\tilde{G}_t^x \mid \mathcal{F}_t] - \nabla_x f(x_t, y_t),\\
\tilde{\xi}_t^y &\overset{\text{def}}{=} \tilde{G}_t^y - \mathbb{E}[\tilde{G}_t^y \mid \mathcal{F}_{t+1/2}], & \beta_t^y &\overset{\text{def}}{=} \mathbb{E}[\tilde{G}_t^y \mid \mathcal{F}_{t+1/2}] - \nabla_y f(x_{t+1}, y_t).
\end{aligned}$$

Here $\tilde{G}_t^\bullet$ denotes the stochastic gradient clipped at threshold $\tau_{\bullet,t}$, $\tilde{\xi}_t^\bullet$ is its centered residual, and $\beta_t^\bullet$ is the corresponding clipping bias. The following bias decomposition lemma is now standard in the literature (Nguyen et al., 2023; Sadiev et al., 2023).

Lemma 4.8 (Clipping bias decomposition) *Suppose Assumptions 3.4 and 3.6 hold. Then, for each $t \geq 0$, the following holds almost surely:*

$$\left\|\tilde{\xi}_t^x\right\| \leq 2\tau_{x,t}, \quad \left\|\tilde{\xi}_t^y\right\| \leq 2\tau_{y,t}. \tag{16}$$

Additionally, if the following inequalities hold at time t:

$$\|\nabla_x f(x_t, y_t)\| \leq \frac{\tau_{x,t}}{2}, \quad \|\nabla_y f(x_{t+1}, y_t)\| \leq \frac{\tau_{y,t}}{2}, \tag{17}$$

then, the following inequalities hold, almost surely:

$$\begin{aligned}
&\|\beta_t^x\| \leq 4\sigma^p \tau_{x,t}^{1-p}, \;\; \mathbb{E}\left[\left\|\tilde{\xi}_t^x\right\|^2 \mid \mathcal{F}_t\right] \leq 16\sigma^p \tau_{x,t}^{2-p}, \\
&\|\beta_t^y\| \leq 4\sigma^p \tau_{y,t}^{1-p}, \;\; \mathbb{E}\left[\left\|\tilde{\xi}_t^y\right\|^2 \mid \mathcal{F}_{t+1/2}\right] \leq 16\sigma^p \tau_{y,t}^{2-p}.
\end{aligned} \tag{18}$$

Now we are ready to derive a potential lemma for $\text{SGDA}_{\text{Clip}}$.

Lemma 4.9 (Potential improvement of $\text{SGDA}_{\text{Clip}}$ in NC-PL) *Assume Assumptions 3.1, 3.2, 3.4 and 3.6 with $\lambda = 1/2$. Let $\gamma_{y,t} \overset{def}{=} \frac{1}{8}(\eta_{y,t} - \frac{\ell}{2}\eta_{y,t}^2)$. If the step sizes satisfy $\eta_{x,t} \leq \eta_{y,t}/(128\kappa^2)$, $\eta_{y,t} \leq 1/\ell$, and $\text{SGDA}_{\text{Clip}}$ satisfies (17) at a given $t \geq 0$, the following holds:*

$$\begin{aligned}
&\frac{1}{8}\eta_{x,t}\left\|\nabla g(x_t)\right\|^2 + \gamma_{y,t}\left\|\nabla \phi_t(y_t)\right\|^2 + 4\kappa\ell\eta_{x,t}b_t \leq P_t - P_{t+1} \\
&\quad - \eta_{x,t}\left\langle \nabla g(x_t), \tilde{\xi}_t^x \right\rangle + \left(\frac{1}{2} - 2\mu\gamma_{y,t}\right)\eta_{x,t}\left\langle \delta_t, \tilde{\xi}_t^x \right\rangle + 6\kappa\ell\eta_{x,t}^2\left\|\tilde{\xi}_t^x\right\|^2 \\
&\quad + \frac{1}{2}(\eta_{y,t} - \ell\eta_{y,t}^2)\left\langle \nabla \phi_t(y_t), \tilde{\xi}_t^y \right\rangle + \frac{\ell}{2}\eta_{y,t}^2\left\|\tilde{\xi}_t^y\right\|^2 + 6\eta_{x,t}\left\|\beta_t^x\right\|^2 + 2\eta_{y,t}\left\|\beta_t^y\right\|^2.
\end{aligned} \tag{19}$$

A key challenge for deriving a high-probability bound from Lemma 4.9 lies in controlling the bias terms β_t^x, β_t^y. Controlling such bias terms requires the low-signal conditions, *i.e.*, Equation (17) of Lemma 4.8, and therefore, we need to establish a uniform control over P_t. The following variant of Lemma 4.4 incorporates clipping biases, and yields a uniform control over the potential process P_t.

Lemma 4.10 (Self-bounding clipped martingales) *Suppose martingale differences $\tilde{\xi}_t^x \in \mathbb{R}^{d_x}, \tilde{\xi}_t^y \in \mathbb{R}^{d_y}$ satisfy Assumption 3.4, and admit deterministic bounds $v_t^x, v_t^y \geq 0$ and thresholds $\tau_{x,t}, \tau_{y,t} > 0$ such that, for all $t \geq 0$, the following holds almost surely:*

$$\left\|\tilde{\xi}_t^x\right\| \leq 2\tau_{x,t}, \quad \left\|\tilde{\xi}_t^y\right\| \leq 2\tau_{y,t}, \quad \mathbb{E}\left[\left\|\tilde{\xi}_t^x\right\|^2 \mid \mathcal{F}_t\right] \leq v_t^x, \quad \mathbb{E}\left[\left\|\tilde{\xi}_t^y\right\|^2 \mid \mathcal{F}_{t+1/2}\right] \leq v_t^y. \tag{20}$$

Let $\eta_{x,t}, \eta_{y,t} \geq 0$ be deterministic, and suppose we have nonnegative processes $G_t, P_t \geq 0$ and a predictable envelope $\bar{G}_t \in \mathcal{F}_t$ such that $G_t \leq \bar{G}_t$ almost surely. Assume that, for any $t \geq 0$, the following holds almost surely:

$$G_t \leq P_t - P_{t+1} + \eta_{x,t} \left\langle c_t^x, \tilde{\xi}_t^x \right\rangle + d_t^x \eta_{x,t}^2 \left\| \tilde{\xi}_t^x \right\|^2 + \eta_{y,t} \left\langle c_t^y, \tilde{\xi}_t^y \right\rangle + d_t^y \eta_{y,t}^2 \left\| \tilde{\xi}_t^y \right\|^2 + r_t, \tag{21}$$

where c_t^x is $\mathcal{F}_t$-measurable, c_t^y and r_t are $\mathcal{F}_{t+1/2}$-measurable, $d_t^x, d_t^y \geq 0$ are deterministic, $r_t \geq 0$, and the coefficients satisfy the self-bounding property

$$\eta_{x,t} \left\| c_t^x \right\|^2 + \eta_{y,t} \left\| c_t^y \right\|^2 \leq C_1 G_t, \quad \forall t \geq 0, \tag{22}$$

with a deterministic constant $C_1 \geq 0$. Define

$$B_t \overset{def}{=} \max\left\{1, \max_{i \in [0,t]} \max\{B_i^x, B_i^y, B_i^v\}\right\}, \quad z_t \overset{def}{=} \frac{1}{B_t},$$
$$B_i^x \overset{def}{=} 2\tau_{x,i}\sqrt{C_1 \eta_{x,i} \bar{G}_i} + 8 d_i^x \eta_{x,i}^2 \tau_{x,i}^2, \quad B_i^y \overset{def}{=} 2\tau_{y,i}\sqrt{C_1 \eta_{y,i} \bar{G}_i} + 8 d_i^y \eta_{y,i}^2 \tau_{y,i}^2,$$
$$B_i^v \overset{def}{=} 6C_1(\eta_{x,i} v_i^x + \eta_{y,i} v_i^y), \quad D_t \overset{def}{=} d_t^x \eta_{x,t}^2 v_t^x + d_t^y \eta_{y,t}^2 v_t^y.$$

Then, for any $\delta \in (0,1)$, with probability at least $1 - \delta$, the following holds for all $k \in \mathbb{N}$:

$$\sum_{t=0}^{k-1} z_t G_t + z_{k-1} P_k \leq 2 z_0 P_0 + 2 \sum_{t=0}^{k-1} (z_t r_t + 2 z_t D_t) + 2 \max\{\log(2/\delta), 1\}.$$

Now we are ready to present our main results for $\text{SGDA}_{\text{Clip}}$ in NC-PL games under heavy-tailed noise. Unlike the subgaussian case, the clipping bias estimates are conditional on both players remaining in a low-signal regime, *i.e.*, Equation (17). Thus the proof cannot simply apply a one-step bias bound and sum it: it must first establish a uniform pathwise control that keeps the primal and dual clipping conditions valid simultaneously. The normalization factor in Lemma 4.10 is designed precisely for this bootstrap. Lemma 4.9 and Lemma 4.10 yield the following results.

Theorem 4.11 (Convergence of $\text{SGDA}_{\text{Clip}}$ in heavy-tailed NC-PL with fixed step sizes) *Suppose Assumptions 3.1, 3.2, 3.4 and 3.6 hold with $p \in (1,2]$. Fix $T \geq 1$, $\delta \in (0,1)$. Let $\Gamma_\delta \overset{def}{=} \max\{\log(2/\delta), 1\}$, $A_{\kappa,\ell} \overset{def}{=} (6\kappa + 8192\kappa^4)\ell$, and $B_{\kappa,\ell} \overset{def}{=} 6 + 256\kappa^2$. Then, if we choose*

$$\eta_{x,t} = c_\eta \left(\frac{P_0 + \Gamma_\delta}{\sigma^2 B_{\kappa,\ell}^{(2-p)/p} A_{\kappa,\ell}^{2(p-1)/p} T} \right)^{\frac{p}{3p-2}}, \quad \eta_{y,t} = 128\kappa^2 \eta_{x,t}, \quad \tau_{x,t} = \tau_{y,t} = \left(\frac{\sigma^p B_{\kappa,\ell}}{2 A_{\kappa,\ell} \eta_x} \right)^{1/p},$$

there exists a universal constant $c_\eta > 0$ and a threshold $T_0 = T_0(p, \kappa, \ell, \sigma, P_0, \delta)$ such that, whenever $T \geq T_0$, the following holds for $\text{SGDA}_{\text{Clip}}$ with probability at least $1 - \delta$:

$$\min_{0 \leq t < T} \left\| \nabla g(x_t) \right\|^2 = \mathcal{O}\left(\sigma^{\frac{2p}{3p-2}} B_{\kappa,\ell}^{\frac{2-p}{3p-2}} A_{\kappa,\ell}^{\frac{2(p-1)}{3p-2}} \left(\frac{P_0 + \Gamma_\delta}{T} \right)^{\frac{2(p-1)}{3p-2}} \right).$$

Table 2: Comparison of existing convergence guarantees in NC-C games. Complexity is measured by the number of stochastic gradient evaluations to achieve an ε-stationary point, *i.e.*, $\min_{t\in[0,T)} \|\nabla\Phi(x_t)\| \le \varepsilon$. Time horizon indicates whether the algorithm requires knowledge of the time horizon T in advance. H.P. stands for high-probability guarantees. Our results in Theorems 5.6, 5.7, 5.9 and 5.10 establish the first high-probability convergence guarantees of SGDA in NC-C games under subgaussian and heavy-tailed noise models.

ALGORITHM	CITATION	NOISE MODEL	COMPLEXITY	TIME HORIZON	H.P.
SGDA	LIN ET AL. (2019)	BOUNDED σ^2	$\mathcal{O}(L^2\ell^3\varepsilon^{-8})$	KNOWN	✗
SGDA	THEOREM 5.6	SUBGAUSSIAN	$\mathcal{O}\left(L^2\ell^3\varepsilon^{-8}\right)$	KNOWN	✓
SGDA	THEOREM 5.7	SUBGAUSSIAN	$\tilde{\mathcal{O}}\left(L^2\ell^3\varepsilon^{-8}\right)$	UNKNOWN	✓
$\text{SGDA}_{\text{CLIP}}$	THEOREM 5.9	HEAVY-TAILED	$\mathcal{O}\left(L^2\ell^{\frac{2p-1}{p-1}}\varepsilon^{-\frac{2(3p-2)}{p-1}}\right)$	KNOWN	✓
$\text{SGDA}_{\text{CLIP}}$	THEOREM 5.10	HEAVY-TAILED	$\tilde{\mathcal{O}}\left(L^2\ell^{\frac{2p-1}{p-1}}\varepsilon^{-\frac{2(3p-2)}{p-1}}\right)$	UNKNOWN	✓

Theorem 4.12 (Convergence of $\text{SGDA}_{\text{Clip}}$ in heavy-tailed NC-PL with decaying step sizes) *Suppose Assumptions 3.1, 3.2, 3.4 and 3.6 hold with $p \in (1,2]$. Fix $\delta \in (0,1)$. Let $\Gamma_\delta \overset{def}{=} \max\{\log(2/\delta), 1\}$, $A_{\kappa,\ell} \overset{def}{=} (6\kappa + 8192\kappa^4)\ell$, and $B_{\kappa,\ell} \overset{def}{=} 6 + 256\kappa^2$. Then, there exist universal constants $c_\eta > 0$ and $s_0 = s_0(p,\kappa,\ell,\sigma,P_0,\delta)$ independent of T such that, if we choose*

$$\eta_{x,t} = c_\eta \left(\frac{P_0 + \Gamma_\delta}{\sigma^2 B_{\kappa,\ell}^{(2-p)/p} A_{\kappa,\ell}^{2(p-1)/p}(t+s_0)}\right)^{\frac{p}{3p-2}}, \quad \eta_{y,t} = 128\kappa^2\eta_{x,t}, \quad \tau_{x,t} = \tau_{y,t} = \left(\frac{\sigma^p B_{\kappa,\ell}}{2A_{\kappa,\ell}\eta_{x,t}}\right)^{1/p}.$$

then, the following holds for $\text{SGDA}_{\text{Clip}}$ with probability at least $1-\delta$:

$$\min_{0\le t<T} \|\nabla g(x_t)\|^2 = \mathcal{O}\left(\sigma^{\frac{2p}{3p-2}} B_{\kappa,\ell}^{\frac{2-p}{3p-2}} A_{\kappa,\ell}^{\frac{2(p-1)}{3p-2}} \left(\frac{P_0+\Gamma_\delta}{T}\right)^{\frac{2(p-1)}{3p-2}} (1+\log T)\right).$$

5. Nonconvex–Concave Games

5.1. Moreau Envelope and Stationarity

In NC-C games, the outer objective g can be *nondifferentiable* even when f is smooth (Lin et al., 2019). To define a principled notion of stationarity and a smooth surrogate objective, we follow Lin et al. (2019) and adopt the Moreau envelope, which can be viewed as a smoothing operator.

Definition 5.1 (Moreau envelope) *The λ-Moreau envelope of a proper lower semicontinuous function $f : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$, denoted by $f_\lambda(\cdot)$, and the proximal mapping $\text{prox}_{\lambda f}(\cdot)$ are defined as:*

$$f_\lambda(x) \overset{def}{=} \inf_{u\in\mathbb{R}^d}\left\{f(u) + \frac{1}{2\lambda}\|u-x\|^2\right\}, \quad \text{prox}_{\lambda f}(x) \overset{def}{=} \arg\min_{u\in\mathbb{R}^d}\left\{f(u) + \frac{1}{2\lambda}\|u-x\|^2\right\}.$$

Definition 5.2 (Weak convexity) *$f : \mathbb{R}^d \to \mathbb{R}$ is ρ-weakly convex if $f(\cdot) + \frac{\rho}{2}\|\cdot\|^2$ is convex.*

A standard result gives differentiability of the Moreau envelope of a weakly convex function.

Lemma 5.1 (Moreau envelope of a weakly convex function is smooth) *Let $f : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ be proper, closed, and ρ-weakly convex, and let $\lambda \in (0, 1/\rho)$. Then $\operatorname{prox}_{\lambda f}(x)$ is single-valued, f_λ is continuously differentiable, and $\nabla f_\lambda(x) = \lambda^{-1}(x - \operatorname{prox}_{\lambda f}(x))$. If $\lambda \le 1/(2\rho)$, f_λ is $1/\lambda$-smooth.*

Notably, Lin et al. (2019) show that the outer objective g is ℓ-weakly convex in NC-C games, and therefore, its Moreau envelope is differentiable, smooth, and hence, computationally tractable.

Lemma 5.2 (Moreau envelope of the outer objective, (Lin et al., 2019)) *Suppose Assumptions 3.1 and 3.3 hold, and let $\Phi(x) := g_{1/(2\ell)}(x)$. Then g is ℓ-weakly convex, Φ is continuously differentiable and 2ℓ-smooth, with $\nabla\Phi(x) = 2\ell(x - \operatorname{prox}_{g/(2\ell)}(x))$. In addition, we have $\inf_x \Phi(x) = \inf_x g(x)$.*

We follow Lin et al. (2019) and measure the stationarity of $\Phi(\cdot) \overset{\text{def}}{=} g_{1/(2\ell)}(\cdot)$, as $\|\nabla\Phi(x)\| \le \varepsilon$ implies a nearby point $\hat{x} \in \mathbb{R}^{d_x}$ with $\|\hat{x} - x\| \le \varepsilon/(2\ell)$ and $\min_{g' \in \partial g(\hat{x})} \|g'\| \le \varepsilon$ (Lin et al., 2019).

5.2. SGDA in Subgaussian NC-C Games

We use the same form of potential P_t as defined in Equation (6), but with $a_t := \Phi(x_t) - \Phi^\star$, $\Phi^\star := \inf_x \Phi(x)$, and $\lambda_t := 2\ell\eta_{x,t}$. The following lemma estimates the primal progress.

Lemma 5.3 (Primal Progress of SGDA in NC-C) *Under Assumptions 3.1 and 3.3, SGDA iterates satisfy the following for all $t \ge 0$:*

$$a_{t+1} \le a_t - \frac{\eta_{x,t}}{4}\|\nabla\Phi(x_t)\|^2 + 2\ell\eta_{x,t}b_t - \eta_{x,t}\langle\nabla\Phi(x_t), \xi_t^x\rangle + 2L^2\ell\eta_{x,t}^2 + 2\ell\eta_{x,t}^2\|\xi_t^x\|^2.$$

Now we derive dual gap estimate and potential improvement lemmas as in Section 4.1.

Lemma 5.4 (Block-wise dual gap of SGDA in NC-C) *Suppose Assumptions 3.1 and 3.3 hold. Let $B \in \mathbb{N}$ be a block size, and fix a block start index $s \in \mathbb{Z}_+$. Choose an $\mathcal{F}_s$-measurable maximizer $y_s^* \in \arg\max_{y\in\mathcal{Y}} f(x_s, y)$. If $\eta_{y,t} \le \frac{1}{2\ell}$, then, for every $t \in \{s, \dots, s+B-1\}$, the following holds for SGDA almost surely:*

$$b_{t+1} \le 2L\|x_{t+1} - x_s\| + \frac{\|y_t - y_s^*\|^2 - \|y_{t+1} - y_s^*\|^2}{2\eta_{y,t}} + \langle y_t - y_s^*, \xi_t^y\rangle + \eta_{y,t}\|\xi_t^y\|^2. \tag{23}$$

Lemma 5.5 (Potential improvement of SGDA in NC-C) *Suppose Assumptions 3.1 and 3.3, and let $B \in \mathbb{N}$ be a block size, and let $s \in \mathbb{Z}_+$ be a starting index of a block. For each block, denote a maximizer $y_s^* \in \arg\max_{y\in\mathcal{Y}} f(x_s, y)$. If $\eta_{y,t} \le \frac{1}{2\ell}$, then, for every $t \in \{s, \dots, s+B-1\}$, the following holds for SGDA almost surely:*

$$\begin{aligned}\frac{\eta_{x,t}}{4}\|\nabla\Phi(x_t)\|^2 &\le P_t - P_{t+1} + 2L^2\ell\eta_{x,t}^2 - \eta_{x,t}\langle\nabla\Phi(x_t), \xi_t^x\rangle + 2\ell\eta_{x,t}^2\|\xi_t^x\|^2 \\ &\quad + \lambda_{t+1}\Big(2L\|x_{t+1} - x_s\| + \frac{\|y_t - y_s^*\|^2 - \|y_{t+1} - y_s^*\|^2}{2\eta_{y,t}} + \langle y_t - y_s^*, \xi_t^y\rangle + \eta_{y,t}\|\xi_t^y\|^2\Big)\end{aligned} \tag{24}$$

Note that a similar block-wise analysis has been previously used in Lin et al. (2019). The motivation is that a maximizer of each x_t may be nonunique and discontinuous in NC-C games, and therefore, we need to fix a target maximizer y_s^* over a block of iterations to control the dual gap. For high-probability analysis, this blockwise device must be combined with concentration of the stochastic block drifts. Our proof therefore augments the potential with a blockwise dual-distance term and controls the resulting drifts through a partition-level martingale argument, rather than treating each block independently. A careful combination of Lemma 5.3 and Lemma 5.4 with martingale control over the blockwise drifts yields Theorems 5.6 and 5.7. We define a step-independent initial potential $\bar{P}_0 \overset{\text{def}}{=} \Phi(x_0) - \Phi^\star + \frac{1}{4}\big(g(x_0) - f(x_0, y_0)\big)$, which admits $P_0 \le \bar{P}_0$ for all the schedules below.

Theorem 5.6 (Convergence of SGDA in subgaussian NC-C games) *Suppose Assumptions 3.1 and 3.3 to 3.5. Let $\delta \in (0,1)$, $\Gamma_\delta \overset{def}{=} \max\{1, \log(2/\delta)\}$, $Q_\delta \overset{def}{=} L^2 + \sigma^2\Gamma_\delta$. Then, there exists a threshold $T_0 = T_0(\ell, L, D, \sigma, \bar{P}_0, \delta) > 0$ such that, for every $T \ge T_0$, SGDA with stepsizes $\eta_{x,t} = \frac{\bar{P}_0^{3/4}}{\left(\ell^3 D^2 \sigma^2 \Gamma_\delta Q_\delta\right)^{1/4} T^{3/4}}$, $\eta_{y,t} = \frac{\bar{P}_0^{1/4} D^{1/2} Q_\delta^{1/4}}{\ell^{1/4}\sigma^{3/2}\Gamma_\delta^{3/4} T^{1/4}}$ satisfies the following with probability at least $1-\delta$:*

$$\min_{0\le t<T} \|\nabla\Phi(x_t)\|^2 = \mathcal{O}\left(\left(\frac{\ell^3 D^2\sigma^2\Gamma_\delta Q_\delta \bar{P}_0}{T}\right)^{1/4} + \frac{\ell D\sigma\sqrt{\Gamma_\delta}}{\sqrt{T}} + \frac{\sigma^2\Gamma_\delta}{T}\right).$$

Theorem 5.7 (Convergence of SGDA in subgaussian NC-C games with decaying stepsizes) *Suppose Assumptions 3.1 and 3.3 to 3.5. Fix $\delta \in (0,1)$, and define $\Gamma_\delta \overset{def}{=} \max\{1, \log(c/\delta)\}$, $Q_\delta \overset{def}{=} L^2 + \sigma^2\Gamma_\delta$. Then, there exist universal constants $c > 0$ and $s_0 = s_0(\ell, L, D, \sigma, \bar{P}_0, \delta) > 0$ independent of T such that SGDA with step sizes $\eta_{x,t} = \frac{\bar{P}_0^{3/4}}{\left(\ell^3 D^2 \sigma^2 \Gamma_\delta Q_\delta\right)^{1/4} (t+s_0)^{3/4}}$, $\eta_{y,t} = \frac{\bar{P}_0^{1/4} D^{1/2} Q_\delta^{1/4}}{\ell^{1/4}\sigma^{3/2}\Gamma_\delta^{3/4}(t+s_0)^{1/4}}$ satisfies the following for every $T \ge \max\{2, s_0\}$ with probability at least $1-\delta$:*

$$\min_{0\le t<T} \|\nabla\Phi(x_t)\|^2 = \mathcal{O}\left((1+\log T)\left(\frac{\ell^3 D^2\sigma^2\Gamma_\delta Q_\delta \bar{P}_0}{T}\right)^{1/4}\right).$$

5.3. Heavy-Tailed NC-C Games

We follow a similar approach as in Section 4.2 and derive a variant of Lemma 5.5 for $\text{SGDA}_{\text{Clip}}$.

Lemma 5.8 (Potential improvement of $\text{SGDA}_{\text{Clip}}$ in NC-C) *Suppose Assumptions 3.1 and 3.3. Let $B \in \mathbb{N}$ be a block size, let s be the starting index of a block, and choose $y_s^\star \in \arg\max_{y\in\mathcal{Y}} f(x_s, y)$. If $\eta_{x,t} \le 1/(2\ell)$ and $\eta_{y,t} \le 1/(2\ell)$, then, for every $t \in \{s, \dots, s+B-1\}$, SGDA_{Clip} satisfies:*

$$\frac{\eta_{x,t}}{8}\|\nabla\Phi(x_t)\|^2 \le P_t - P_{t+1} + 2L^2\ell\eta_{x,t}^2 - \eta_{x,t}\left\langle\nabla\Phi(x_t), \tilde{\xi}_t^x\right\rangle + 4\ell\eta_{x,t}^2\left\|\tilde{\xi}_t^x\right\|^2 + 4\eta_{x,t}\|\beta_t^x\|^2$$
$$+ \lambda_{t+1}\left[2L\|x_{t+1} - x_s\| + \frac{\|y_t - y_s^\star\|^2 - \|y_{t+1} - y_s^\star\|^2}{2\eta_{y,t}} + \left\langle y_t - y_s^\star, \tilde{\xi}_t^y + \beta_t^y\right\rangle + 2\eta_{y,t}\left(\left\|\tilde{\xi}_t^y\right\|^2 + \|\beta_t^y\|^2\right)\right].$$

Lemma 5.8 and Lemma 4.10 produce our main results for heavy-tailed NC-C games.

Theorem 5.9 (Convergence of $\text{SGDA}_{\text{Clip}}$ in heavy-tailed NC-C games with fixed step sizes) *Suppose Assumptions 3.1, 3.3, 3.4 and 3.6 with $p \in (1,2]$. Let $\delta \in (0,1)$ and $\Gamma_\delta \overset{def}{=} \max\{1, \log(c/\delta)\}$*

for some constant $c > 0$. *Set* $\Delta_\delta \overset{def}{=} \bar{P}_0 + \Gamma_\delta$, $A_\delta \overset{def}{=} \ell^3 D^2 \sigma^p \Gamma_\delta (L^2 + \sigma^p \Gamma_\delta)$, *and* $\tau_0 \overset{def}{=} \left(\frac{\ell D \sigma^p}{(A_\delta \Delta_\delta)^{1/4}}\right)^{\frac{4}{3p-2}}$. *Then, there exist universal constants* $c, c_{x,p}, c_{y,p} > 0$, *and there exists a finite threshold* $T_0 = T_0(p, \ell, L, D, \sigma, \bar{P}_0, \delta) > 0$ *such that, for every* $T \geq T_0$, *SGDA*$_{Clip}$ *with*

$$\eta_{x,t} = c_{x,p} \frac{\Delta_\delta^{3/4}}{A_\delta^{1/4} \tau_0^{(2-p)/4}} T^{-\frac{2p-1}{3p-2}}, \quad \eta_{y,t} = c_{y,p} \frac{(A_\delta \Delta_\delta)^{1/4}}{\ell \sigma^p \Gamma_\delta \, \tau_0^{3(2-p)/4}} T^{-\frac{1}{3p-2}}, \quad \tau_{x,t} = \tau_{y,t} = \tau_0 T^{\frac{1}{3p-2}}$$

satisfies the following with probability at least $1 - \delta$:

$$\min_{0 \leq t < T} \|\nabla \Phi(x_t)\|^2 = \mathcal{O}\left(\left(\frac{\ell^3 D^2 \sigma^p \Gamma_\delta (L^2 + \sigma^p \Gamma_\delta)(\bar{P}_0 + \Gamma_\delta)(\ell D \sigma^p)^{\frac{2-p}{p-1}}}{T}\right)^{\frac{p-1}{3p-2}}\right).$$

Theorem 5.10 (Convergence of SGDA$_{\text{Clip}}$ in heavy-tailed NC-C games with decaying step sizes) *Suppose Assumptions 3.1, 3.3, 3.4 and 3.6 with* $p \in (1, 2]$. *Let* $\delta \in (0, 1)$ *and* $\Gamma_\delta := \max\{1, \log(c/\delta)\}$ *for some constant* $c > 0$. *Set* $\Delta_\delta \overset{def}{=} \bar{P}_0 + \Gamma_\delta$, $A_\delta \overset{def}{=} \ell^3 D^2 \sigma^p \Gamma_\delta (L^2 + \sigma^p \Gamma_\delta)$, *and* $\tau_0 \overset{def}{=} \left(\frac{\ell D \sigma^p}{(A_\delta \Delta_\delta)^{1/4}}\right)^{\frac{4}{3p-2}}$. *Then, there exist universal constants* $c, c_{x,p}, c_{y,p} > 0$, *and a finite shift* $s_0 = s_0(p, \ell, L, D, \sigma, \bar{P}_0, \delta) > 0$ *independent of* T, *such that, for every* $T \geq \max\{2, s_0\}$, *SGDA*$_{Clip}$ *with*

$$\eta_{x,t} = c_{x,p} \frac{\Delta_\delta^{3/4}}{A_\delta^{1/4} \tau_0^{(2-p)/4}} (t + s_0)^{-\frac{2p-1}{3p-2}}, \quad \eta_{y,t} = c_{y,p} \frac{(A_\delta \Delta_\delta)^{1/4}}{\ell \sigma^p \Gamma_\delta \, \tau_0^{3(2-p)/4}} (t + s_0)^{-\frac{1}{3p-2}}, \quad \tau_{x,t} = \tau_{y,t} = \tau_0 (t + s_0)^{\frac{1}{3p-2}}.$$

satisfies the following with probability at least $1 - \delta$:

$$\min_{0 \leq t < T} \|\nabla \Phi(x_t)\|^2 \leq C_p (1 + \log T) \left(\frac{\ell^3 D^2 \sigma^p \Gamma_\delta (L^2 + \sigma^p \Gamma_\delta)(\bar{P}_0 + \Gamma_\delta)(\ell D \sigma^p)^{\frac{2-p}{p-1}}}{T}\right)^{\frac{p-1}{3p-2}}.$$

6. Conclusion

In this paper, we established the first high-probability convergence guarantees of SGDA and SGDA$_{\text{Clip}}$ in NC-PL and NC-C games under subgaussian and heavy-tailed noise. Our unified analysis leverages a Lyapunov approach and self-bounding martingale techniques to handle coupled martingale differences that naturally arise in our analysis. We leave extending our analysis to extragradient-type algorithms (Korpelevich, 1976a; Popov, 1980) and to more general classes of games, such as min-max Markov games (Littman, 1994; Wei et al., 2021; Zeng et al., 2022), as future work.

References

Waiss Azizian, Ioannis Mitliagkas, S. Lacoste-Julien, and Gauthier Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *AISTATS*, 2020a.

Waiss Azizian, Damien Scieur, Ioannis Mitliagkas, S. Lacoste-Julien, and Gauthier Gidel. Accelerating smooth games by manipulating spectral shapes. In *AISTATS*, 2020b.

Aleksandr Beznosikov, Eduard Gorbunov, Hugo Berard, and Nicolas Loizou. Stochastic gradient descent-ascent: Unified theory and new efficient methods. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 172–235. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/beznosikov23a.html.

Pengyu Cheng, Tianhao Hu, Han Xu, Zhisong Zhang, Yong Dai, Lei Han, nan du, and Xiaolong Li. Self-playing adversarial language game enhances LLM reasoning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=oCGkSH7ys2.

Pengyu Cheng, Yifan Yang, Jian Li, Yong Dai, Tianhao Hu, Peixin Cao, Nan Du, and Xiaolong Li. Adversarial preference optimization: Enhancing your alignment via RM-LLM game. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3705–3716, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.221. URL https://aclanthology.org/2024.findings-acl.221/.

Hanseul Cho and Chulhee Yun. Sgda with shuffling: Faster convergence for nonconvex-pł minimax optimization. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6xXtM8bFFJ.

C. Daskalakis, P. Goldberg, and C. Papadimitriou. The Complexity of Computing a Nash Equilibrium. *Electron. Colloquium Comput. Complex.*, 2006.

C. Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with Optimism. In *ICLR*, 2018.

Simon S Du and Wei Hu. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. In *International Conference on Artificial Intelligence and Statistics*, pages 196–205. PMLR, 2019.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1laEnA5Ym.

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

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf.

Eduard Gorbunov, Hugo Berard, Gauthier Gidel, and Nicolas Loizou. Stochastic extragradient: General analysis and improved rates. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7865–7901. PMLR, 28–30 Mar 2022a. URL https://proceedings.mlr.press/v151/gorbunov22b.html.

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

Eduard Gorbunov, Nicolas Loizou, and Gauthier Gidel. Extragradient method: $o(1/k)$ last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In *International Conference on Artificial Intelligence and Statistics*, pages 366–402. PMLR, 2022c.

Eduard Gorbunov, Adrien Taylor, and Gauthier Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 21858–21870. Curran Associates, Inc., 2022d. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/893cd874ba98afa54ae9e385a24a83ac-Paper-Conference.pdf.

Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the 32nd International Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1579–1613. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/harvey19a.html.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *Advances in Neural Information Processing Systems*, volume 32, pages 6938–6948. Curran Associates, Inc., 2019.

Feihu Huang, Chunyu Xuan, Xinrui Wang, Siqi Zhang, and Songcan Chen. Enhanced adaptive gradient algorithms for nonconvex-PL minimax optimization. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pages 3439–3447. PMLR, 2025. URL https://proceedings.mlr.press/v258/huang25d.html.

Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 2011.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 1976a.

G. M. Korpelevich. The Extragradient Method for Finding Saddle Points and Other Problems. *Ekonomika i Matematicheskie Metody*, 12:747–756, 1976b.

Yassine Laguel, Yasa Syed, Necdet Serhat Aybat, and Mert Gürbüzbalaban. High-probability complexity bounds for stochastic non-convex minimax optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/fec946957ce1af51a61e8f2d851ac98f-Abstract-Conference.html.

Tianyi Lin, Chi Jin, and Michael I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, 2019. URL https://api.semanticscholar.org/CorpusID:173991107.

Michael L. Littman. Markov Games as a Framework for Multi-Agent Reinforcement Learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 157–163. Morgan Kaufmann, 1994. URL https://dl.acm.org/doi/10.5555/3091574.3091594.

Zijian Liu, Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy L. Nguyen. High probability convergence of stochastic gradient methods. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 21884–21914. PMLR, 2023. URL https://proceedings.mlr.press/v202/liu23aa.html.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

P. Mertikopoulos, C. Papadimitriou, and G. Piliouras. Cycles in Adversarial Regularized Learning. In *SODA*, 2018.

Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtarik, and Yura Malitsky. Revisiting stochastic extragradient. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4573–4582. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/mishchenko20a.html.

Aryan Mokhtari, A. Ozdaglar, and Sarath Pattathil. A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach. In *AISTATS*, 2020.

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

Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy L. Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In *Advances in Neural Information Processing Systems*, volume 36, pages 24191–24222, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4c454d34f3a4c8d6b4ca85a918e5d7ba-Abstract-Conference.html.

Maher Nouiehed, Maziar Sanjabi, Tianyi Lin, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, volume 32, pages 14934–14942, 2019.

C. Papadimitriou. On the complexity of the parity argument and other inefficient proofs of existence. *J. Comput. Syst. Sci.*, 48:498–532, 1994.

L. Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28:845–848, 1980.

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horvath, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtarik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *International Conference on Machine Learning*, 2023.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, 2019.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4259–4299. PMLR, 2021. URL https://proceedings.mlr.press/v134/wei21a.html.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=a3PmRgAB5T.

Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. In *Advances in Neural Information Processing Systems*, volume 33, pages 1153–1165, 2020.

Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 5485–5517. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/yang22b.html.

Sihan Zeng, Thinh T. Doan, and Justin Romberg. Regularized gradient descent ascent for two-player zero-sum markov games. In *Advances in Neural Information Processing Systems*, volume 35, pages 34546–34558, 2022. URL https://arxiv.org/abs/2205.13746.

Guodong Zhang, Yuanhao Wang, Laurent Lessard, and Roger B. Grosse. Near-optimal local convergence of alternating gradient descent-ascent for minimax optimization. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning*

Research, pages 7659–7679. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/zhang22e.html.

Jikai Zhang, Sai Praneeth Karimireddy, Veerapen Veerapen, and Virginia Smith. Why are adaptive methods good for attention models? In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15367–15378. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/b05b57f6add810d3b7490866d74c0053-Paper.pdf.

Appendix: Table of Contents

Appendix A. Preliminary Lemmas

A.1. Elementary Tools

Lemma A.1 *For every real $x < 3$,*

$$e^x \;\le\; 1 + x + \frac{x^2}{2\,(1 - x/3)}.$$

Proof Using the Taylor series,

$$e^x = 1 + x + \sum_{n=2}^{\infty} \frac{x^n}{n!}.$$

For $n \ge 2$, we have

$$n! = 2 \cdot 3 \cdot 4 \cdots n \;\ge\; 2 \cdot 3^{\,n-2},$$

hence

$$\frac{x^n}{n!} \le \frac{x^n}{2 \cdot 3^{\,n-2}} = \frac{x^2}{2} \left(\frac{x}{3}\right)^{n-2}.$$

Therefore, for $x < 3$,

$$\sum_{n=2}^{\infty} \frac{x^n}{n!} \le \frac{x^2}{2} \sum_{m=0}^{\infty} \left(\frac{x}{3}\right)^m = \frac{x^2}{2} \cdot \frac{1}{1 - x/3},$$

which proves the claim. ■

A.2. Properties of NC-PL Games

Lemma A.2 (The outer function is smooth in NC-PL games) *Under Assumptions 3.1 and 3.2, the function $g(\cdot) \overset{def}{=} \max_{y \in \mathcal{Y}} f(\cdot, y)$ is differentiable, and its gradient $\nabla g(\cdot)$ is $2\kappa\ell$-Lipschitz.*

Proof For each $x \in \mathbb{R}^{d_x}$, denote the (nonempty) maximizer set

$$\mathcal{Y}^\star(x) \;\overset{\text{def}}{=}\; \arg\max_{y \in \mathcal{Y}} f(x, y), \qquad g(x) = \max_{y \in \mathcal{Y}} f(x, y).$$

For NC-PL games, we work in the $\mathcal{Y} = \mathbb{R}^{d_y}$ regime, so the first-order condition $\nabla_y f(x, y^\star) = 0$ holds for any $y^\star \in \mathcal{Y}^\star(x)$.

Step 1: $\nabla_x f(x, \cdot)$ is single-valued on $\mathcal{Y}^\star(x)$. Fix x and let $y_1, y_2 \in \mathcal{Y}^\star(x)$. Define

$$d \;\overset{\text{def}}{=}\; \nabla_x f(x, y_1) - \nabla_x f(x, y_2).$$

We show $d = 0$. Suppose $d \neq 0$ and set $u = d/\|d\|$. For $\varepsilon > 0$ define $x_\varepsilon \overset{\text{def}}{=} x + \varepsilon u$. By ℓ-smoothness of $x \mapsto f(x, y_i)$ (Assumption 3.1),

$$f(x_\varepsilon, y_1) \ge f(x, y_1) + \varepsilon \langle \nabla_x f(x, y_1), u \rangle - \frac{\ell}{2}\varepsilon^2,$$

while

$$f(x_\varepsilon, y_2) \le f(x, y_2) + \varepsilon \langle \nabla_x f(x, y_2), u \rangle + \frac{\ell}{2}\varepsilon^2.$$

Since $f(x, y_1) = f(x, y_2) = g(x)$, subtracting the two bounds yields

$$f(x_\varepsilon, y_1) - f(x_\varepsilon, y_2) \;\geq\; \varepsilon\langle d, u\rangle - \ell\varepsilon^2 = \varepsilon\|d\| - \ell\varepsilon^2.$$

Choose $\varepsilon \in \big(0, \|d\|/(2\ell)\big)$. Then

$$f(x_\varepsilon, y_1) - f(x_\varepsilon, y_2) \;\geq\; \frac{\varepsilon}{2}\|d\|.$$

Hence, using $g(x_\varepsilon) \geq f(x_\varepsilon, y_1)$,

$$b(x_\varepsilon, y_2) \overset{\text{def}}{=} g(x_\varepsilon) - f(x_\varepsilon, y_2) \;\geq\; f(x_\varepsilon, y_1) - f(x_\varepsilon, y_2) \;\geq\; \frac{\varepsilon}{2}\|d\|.$$

On the other hand, since $y_2 \in \mathcal{Y}^\star(x)$ is a maximizer, $\nabla_y f(x, y_2) = 0$. By ℓ-Lipschitzness of $\nabla_y f(\cdot, y_2)$ (Assumption 3.1),

$$\|\nabla_y f(x_\varepsilon, y_2)\| = \|\nabla_y f(x_\varepsilon, y_2) - \nabla_y f(x, y_2)\| \leq \ell\|x_\varepsilon - x\| = \ell\varepsilon.$$

Applying the NC–PL inequality (Assumption 3.2) at (x_ε, y_2) gives

$$\frac{1}{2}\|\nabla_y f(x_\varepsilon, y_2)\|^2 \;\geq\; \mu\, b(x_\varepsilon, y_2) \;\geq\; \mu \cdot \frac{\varepsilon}{2}\|d\|.$$

Thus $\frac{1}{2}(\ell\varepsilon)^2 \geq \mu\frac{\varepsilon}{2}\|d\|$, i.e. $\ell^2\varepsilon \geq \mu\|d\|$. Letting $\varepsilon \downarrow 0$ forces $\|d\| = 0$, contradiction. Therefore $d = 0$, i.e.

$$\nabla_x f(x, y_1) = \nabla_x f(x, y_2) \qquad \forall\, y_1, y_2 \in \mathcal{Y}^\star(x).$$

Step 2: Differentiability of g and gradient formula. By Danskin's theorem, $\partial g(x) = \mathrm{conv}\{\nabla_x f(x, y) : y \in \mathcal{Y}^\star(x)\}$. Step 1 shows this set is a singleton, hence g is differentiable and

$$\nabla g(x) = \nabla_x f(x, y^\star) \qquad \forall\, y^\star \in \mathcal{Y}^\star(x).$$

Step 3: PL-to-distance bound in y. Fix x and define $\psi_x(y) \overset{\text{def}}{=} b(x, y) = g(x) - f(x, y) \geq 0$. Then ψ_x is ℓ-smooth in y (Assumption 3.1) and satisfies the PL inequality

$$\frac{1}{2}\|\nabla\psi_x(y)\|^2 = \frac{1}{2}\|\nabla_y f(x, y)\|^2 \geq \mu\,\psi_x(y) \qquad \forall y.$$

We claim that for every y,

$$\mathrm{dist}(y, \mathcal{Y}^\star(x))^2 \;\leq\; \frac{2}{\mu}\,\psi_x(y) = \frac{2}{\mu} b(x, y). \tag{25}$$

To prove (25), consider the gradient flow $\dot{y}(t) = -\nabla\psi_x(y(t))$ with $y(0) = y$. Since $\nabla\psi_x$ is Lipschitz, the flow exists and is unique. Let $\phi(t) \overset{\text{def}}{=} \psi_x(y(t))$. Then $\phi'(t) = \langle\nabla\psi_x(y(t)), \dot{y}(t)\rangle = -\|\nabla\psi_x(y(t))\|^2 \leq -2\mu\phi(t)$, so $\phi(t) \downarrow 0$. Moreover, the path length satisfies

$$\int_0^\infty \|\dot{y}(t)\|\,dt = \int_0^\infty \|\nabla\psi_x(y(t))\|\,dt = \int_0^\infty \frac{-\phi'(t)}{\|\nabla\psi_x(y(t))\|}\,dt \leq \int_0^\infty \frac{-\phi'(t)}{\sqrt{2\mu\phi(t)}}\,dt = \sqrt{\frac{2}{\mu}}\sqrt{\phi(0)}.$$

Hence $y(t)$ has finite length and converges to some $y_\infty \in \mathcal{Y}^\star(x)$, and $\mathrm{dist}(y, \mathcal{Y}^\star(x)) \leq \|y - y_\infty\| \leq \sqrt{2/\mu}\sqrt{\psi_x(y)}$, proving (25).

Step 4: Lipschitzness of ∇g. Fix x, x', pick $y^\star \in \mathcal{Y}^\star(x)$, and apply PL at $(x', y^\star)$. Since $\nabla_y f(x, y^\star) = 0$ and $\nabla_y f(\cdot, y^\star)$ is ℓ-Lipschitz (Assumption 3.1),

$$\|\nabla_y f(x', y^\star)\| = \|\nabla_y f(x', y^\star) - \nabla_y f(x, y^\star)\| \le \ell \|x' - x\|.$$

Thus, by Assumption 3.2,

$$b(x', y^\star) \le \frac{1}{2\mu} \|\nabla_y f(x', y^\star)\|^2 \le \frac{\ell^2}{2\mu} \|x' - x\|^2.$$

Using (25) at x' and $y^\star$, there exists $y^{\star\prime} \in \mathcal{Y}^\star(x')$ such that

$$\|y^\star - y^{\star\prime}\| \le \operatorname{dist}(y^\star, \mathcal{Y}^\star(x')) \le \frac{\ell}{\mu} \|x' - x\| = \kappa \|x' - x\|.$$

Finally, using $\nabla g(x) = \nabla_x f(x, y^\star)$ and $\nabla g(x') = \nabla_x f(x', y^{\star\prime})$, and ℓ-Lipschitzness of $\nabla_x f$ (Assumption 3.1),

$$\|\nabla g(x) - \nabla g(x')\| = \|\nabla_x f(x, y^\star) - \nabla_x f(x', y^{\star\prime})\| \le \ell \|x - x'\| + \ell \|y^\star - y^{\star\prime}\| \le (1 + \kappa) \ell \|x - x'\|.$$

Since $\mu \le \ell$ for an ℓ-smooth PL function, $\kappa = \ell/\mu \ge 1$, hence $1 + \kappa \le 2\kappa$. Therefore ∇g is $2\kappa\ell$-Lipschitz. ■

Lemma A.3 (The outer gradient can be approximated in NC-PL games) *Under Assumptions 3.1 and 3.2, the following holds for $g(x) \overset{def}{=} \max_{y \in \mathcal{Y}} f(x, y)$ and $b(x, y) \overset{def}{=} g(x) - f(x, y) \ge 0$:*

$$\|\nabla g(x) - \nabla_x f(x, y)\|^2 \le 2\kappa\ell \cdot b(x, y), \quad \forall (x, y) \in \mathbb{R}^{d_x} \times \mathcal{Y}. \tag{26}$$

Proof Fix (x, y) and let $\mathcal{Y}^\star(x) = \arg\max_{u \in \mathcal{Y}} f(x, u)$. Pick $y^\star \in \mathcal{Y}^\star(x)$ such that $\|y - y^\star\| = \operatorname{dist}(y, \mathcal{Y}^\star(x))$. By Lemma A.2, g is differentiable and $\nabla g(x) = \nabla_x f(x, y^\star)$. Hence,

$$\|\nabla g(x) - \nabla_x f(x, y)\| = \|\nabla_x f(x, y^\star) - \nabla_x f(x, y)\| \le \ell \|y^\star - y\|,$$

where we used ℓ-Lipschitzness of $\nabla_x f(x, \cdot)$ from Assumption 3.1. Squaring and using the PL-to-distance bound (25) established in the proof of Lemma A.2 yields

$$\|\nabla g(x) - \nabla_x f(x, y)\|^2 \le \ell^2 \operatorname{dist}(y, \mathcal{Y}^\star(x))^2 \le \ell^2 \cdot \frac{2}{\mu} b(x, y) = 2\frac{\ell^2}{\mu} b(x, y) = 2\kappa\ell \cdot b(x, y).$$

■

Lemma A.4 (The outer gradient upper bound in NC-PL games) *Under Assumptions 3.1 and 3.2, the following holds for $g(x) \overset{def}{=} \max_{y \in \mathcal{Y}} f(x, y)$ and $g_* \overset{def}{=} \min_{x \in \mathbb{R}^{d_x}} g(x)$:*

$$\|\nabla g(x)\|^2 \le 4\kappa\ell \left(g(x) - g_*\right), \quad \forall x \in \mathbb{R}^{d_x}. \tag{27}$$

Proof By Lemma A.2, g is L_g-smooth with $L_g = 2\kappa\ell$. Since g is bounded below by g_*, the standard smoothness descent inequality gives

$$g\left(x - \frac{1}{L_g}\nabla g(x)\right) \le g(x) - \frac{1}{2L_g}\|\nabla g(x)\|^2.$$

Because $g_* \le g(x - \nabla g(x)/L_g)$, we obtain

$$g_* \le g(x) - \frac{1}{2L_g}\|\nabla g(x)\|^2.$$

Therefore

$$\|\nabla g(x)\|^2 \le 2L_g(g(x) - g_*) = 4\kappa\ell(g(x) - g_*).$$

■

A.3. Properties of NC-C Games

The following result justifies Moreau-envelope stationarity as a valid convergence criterion in NC-C games, and a similar result was shown by Lin et al. (2019).

Lemma A.5 (Moreau envelope stationarity implies approximate stationarity) *Suppose Assumptions 3.1 and 3.3, and let* $g(x) \overset{def}{=} \max_{y\in\mathcal{Y}} f(x,y)$, *and let* $\Phi(x) \overset{def}{=} g_{1/(2\ell)}(x)$ *be the Moreau envelope of* g *with parameter* $1/(2\ell)$. *Then, if* $x \in \mathbb{R}^{d_x}$ *satisfies* $\|\nabla\Phi(x)\| \le \varepsilon$, *then there exists* $x' \in \mathbb{R}^{d_x}$ *such that* $\|x' - x\| \le \frac{\varepsilon}{2\ell}$ *with* $\min_{g'\in\partial g(x')} \|g'\| \le \varepsilon$.

Proof Let $x' = \mathrm{prox}_{g/(2\ell)}(x)$. Then, by the definition of the proximal operator, we have

$$x' = \arg\min_{u\in\mathbb{R}^{d_x}} \left\{ g(u) + \ell\,\|u - x\|^2 \right\}.$$

From Lemma 5.2, we have $\nabla\Phi(x) = 2\ell(x - x')$, which implies $\|x' - x\| = \frac{\|\nabla\Phi(x)\|}{2\ell} \le \frac{\varepsilon}{2\ell}$. Furthermore, the first-order optimality condition of the proximal operator gives

$$0 \in \partial g(x') + 2\ell(x' - x),$$

which implies there exists $g' \in \partial g(x')$ such that $\|g'\| = 2\ell\,\|x - x'\| = \|\nabla\Phi(x)\| \le \varepsilon$. ■

A.4. Martingale Tools

The following is a standard Markov-type maximal concentration that we will use in our analysis.

Lemma A.6 (Ville's inequality) *Let* $(M_n)_{n\ge0}$ *be a nonnegative supermartingale with* $\mathbb{E}[M_0] \le 1$. *Then for any* $a > 0$,

$$\mathbb{P}\left(\sup_{n\ge0} M_n \ge a\right) \le \frac{1}{a}.$$

We adopt the following lemma from Nguyen et al. (2023) to control the conditional moment generating function (MGF) of bounded mean-zero random variables.

Lemma A.7 (Conditional MGF bound for bounded mean-zero variables) *Let $\mathcal{G}$ be a σ-field and let X be a random variable such that $X \le 1$ almost surely and $\mathbb{E}[X \mid \mathcal{G}] = 0$. Then*

$$\mathbb{E}\big[e^X \mid \mathcal{G}\big] \le \exp\left(\frac{3}{4}\,\mathbb{E}[X^2 \mid \mathcal{G}]\right).$$

Proof Apply Lemma A.1 with $x = X$. Since $X \le 1$, we have $1 - X/3 \ge 2/3$, and thus

$$e^X \le 1 + X + \frac{X^2}{2(1 - X/3)} \le 1 + X + \frac{3}{4}X^2.$$

Taking conditional expectation and using $\mathbb{E}[X \mid \mathcal{G}] = 0$ gives

$$\mathbb{E}[e^X \mid \mathcal{G}] \le 1 + \frac{3}{4}\,\mathbb{E}[X^2 \mid \mathcal{G}] \le \exp\left(\frac{3}{4}\,\mathbb{E}[X^2 \mid \mathcal{G}]\right),$$

where we used $1 + u \le e^u$ for all real u. ■

Lemma A.8 (Norm-subgaussian projections) *Let $\mathcal{G}$ be a σ-field and let $X \in \mathbb{R}^d$ satisfy*

$$\mathbb{E}[X \mid \mathcal{G}] = 0, \qquad \mathbb{E}\left[\exp\left(\frac{\|X\|^2}{\sigma^2}\right)\middle|\mathcal{G}\right] \le e.$$

Then, for every $\mathcal{G}$-measurable vector $v \in \mathbb{R}^d$ and every $\lambda \in \mathbb{R}$,

$$\mathbb{E}[\exp(\lambda\langle v, X\rangle)|\mathcal{G}] \le \exp\big(4\lambda^2\sigma^2\|v\|^2\big).$$

Proof The claim is trivial when $v = 0$, so assume $v \ne 0$ and set $Y := \langle v, X\rangle$ and $s := \sigma\|v\|$. Then $Y^2/s^2 \le \|X\|^2/\sigma^2$, hence

$$\mathbb{E}\big[\exp(Y^2/s^2) \mid \mathcal{G}\big] \le e.$$

We use the elementary inequality $e^z \le z + e^{z^2}$ for all $z \in \mathbb{R}$. If $|\lambda|s \le 1$, conditional Jensen's inequality gives

$$\begin{aligned}\mathbb{E}\Big[e^{\lambda Y} \mid \mathcal{G}\Big] &\le \mathbb{E}\Big[\lambda Y + e^{\lambda^2 Y^2} \mid \mathcal{G}\Big] \\ &= \mathbb{E}\Big[e^{\lambda^2 Y^2} \mid \mathcal{G}\Big] \\ &\le \mathbb{E}\Big[e^{Y^2/s^2} \mid \mathcal{G}\Big]^{\lambda^2 s^2} \le e^{\lambda^2 s^2}.\end{aligned}$$

If $|\lambda|s > 1$, Young's inequality gives

$$\lambda Y \le |\lambda Y| \le \frac{Y^2}{s^2} + \frac{\lambda^2 s^2}{4}.$$

Therefore

$$\mathbb{E}\Big[e^{\lambda Y} \mid \mathcal{G}\Big] \le \exp\left(\frac{\lambda^2 s^2}{4}\right)\mathbb{E}\Big[e^{Y^2/s^2} \mid \mathcal{G}\Big] \le \exp\left(1 + \frac{\lambda^2 s^2}{4}\right) \le \exp\big(4\lambda^2 s^2\big),$$

where the last inequality uses $|\lambda|s > 1$. Combining the two cases proves the claim. ■

Lemma A.9 (Weighted norm-square concentration) *Let $(X_t)_{t=0}^{T-1}$ be adapted to a filtration $(\mathcal{G}_t)$ and satisfy*

$$\mathbb{E}\left[\exp\left(\frac{\|X_t\|^2}{\sigma^2}\right)\middle|\mathcal{G}_t\right] \le e.$$

Let $a_t \ge 0$ be deterministic weights and set $A := \sum_{t=0}^{T-1} a_t$. Then for a universal constant $C > 0$, with probability at least $1-\delta$,

$$\sum_{t=0}^{T-1} a_t\|X_t\|^2 \le C\sigma^2\Gamma_\delta A, \qquad \Gamma_\delta := \max\{1, \log(c/\delta)\}.$$

The same conclusion holds for predictable weights that are dominated by deterministic weights, after replacing A by the sum of the deterministic dominators.

Proof If $A = 0$, the claim is trivial. Let $a_{\max} := \max_{0\le t<T} a_t$ and $\theta := 1/(\sigma^2 a_{\max})$. Since $a_t/a_{\max} \in [0,1]$, Jensen's inequality gives

$$\mathbb{E}\big[\exp\big(\theta a_t\|X_t\|^2\big)\big|\mathcal{G}_t\big] = \mathbb{E}\left[\exp\left(\frac{a_t}{a_{\max}}\frac{\|X_t\|^2}{\sigma^2}\right)\middle|\mathcal{G}_t\right] \le e^{a_t/a_{\max}}.$$

Iterating over the filtration yields

$$\mathbb{E}\exp\left(\theta\sum_{t=0}^{T-1} a_t\|X_t\|^2\right) \le \exp(A/a_{\max}).$$

By Markov's inequality, with probability at least $1-\delta$,

$$\sum_{t=0}^{T-1} a_t\|X_t\|^2 \le \sigma^2\left(A + a_{\max}\log(1/\delta)\right) \le 2\sigma^2 A\max\{1, \log(1/\delta)\},$$

using $a_{\max} \le A$. Enlarging the universal constants gives the displayed form. The dominated predictable case follows by monotonicity from the deterministic upper bounds. ■

Lemma A.10 (Hilbert-valued block Freedman maximal inequality) *Let $I = [s, e)$ be a deterministic interval and let $(\zeta_i)_{i=s}^{e-1}$ be a Hilbert-valued martingale difference sequence adapted to $(\mathcal{F}_i)$. Suppose there are deterministic numbers $R_i \ge 0$ and $V_I \ge 0$ such that, almost surely,*

$$\|\zeta_i\| \le R_i, \qquad s \le i < e,$$

and

$$\sum_{i=s}^{e-1}\mathbb{E}[\|\zeta_i\|^2 \mid \mathcal{F}_i] \le V_I.$$

Let

$$R_I := \max_{s\le i<e} R_i, \qquad \Gamma_\delta := \max\{1, \log(c/\delta)\}.$$

Then, with probability at least $1-\delta$,

$$\max_{s\le t<e}\left\|\sum_{i=s}^{t}\zeta_i\right\| \le C\left(\sqrt{V_I\Gamma_\delta} + R_I\Gamma_\delta\right),$$

where $C, c > 0$ are universal constants. The same bound holds conditionally on $\mathcal{F}_s$.

Proof This is the standard Hilbert-valued Pinelis–Freedman maximal inequality applied to the stopped partial sums over the deterministic interval I. The bounded increment scale is R_I and the conditional quadratic variation is at most V_I. The conditional version follows by applying the same inequality after conditioning on $\mathcal{F}_s$ and enlarging the universal constants. ■

Lemma A.11 (Weighted block maximal inequality) *Let $0 = s_0 < s_1 < \cdots < s_M = T$ be a deterministic partition. For each block $I_m = [s_m, s_{m+1})$, let $(\zeta_i)_{i \in I_m}$ be Hilbert-valued martingale differences adapted to $(\mathcal{F}_i)$. Suppose there are deterministic numbers $R_i \geq 0$ and $V_m \geq 0$ such that, almost surely,*

$$\|\zeta_i\| \leq R_i, \qquad \sum_{i=s_m}^{s_{m+1}-1} \mathbb{E}[\|\zeta_i\|^2 \mid \mathcal{F}_i] \leq V_m.$$

Define

$$R_m := \max_{i \in I_m} R_i, \qquad K_m := \sqrt{V_m} + R_m.$$

Let $\alpha_m \geq 0$ be deterministic block weights. Then, with probability at least $1 - \delta$,

$$\sum_{m=0}^{M-1} \alpha_m \max_{s_m \leq t < s_{m+1}} \left\| \sum_{i=s_m}^{t} \zeta_i \right\| \leq C \left[\sum_{m=0}^{M-1} \alpha_m K_m + \Gamma_\delta \max_{0 \leq m < M} \alpha_m K_m \right],$$

where $C, c > 0$ are universal constants and $\Gamma_\delta := \max\{1, \log(c/\delta)\}$.

Proof Let

$$M_m := \max_{s_m \leq t < s_{m+1}} \left\| \sum_{i=s_m}^{t} \zeta_i \right\|.$$

By Lemma A.10, conditionally on $\mathcal{F}_{s_m}$, for all $u \geq 1$,

$$\Pr\left(M_m > C K_m u \mid \mathcal{F}_{s_m}\right) \leq e^{-u},$$

because $\sqrt{V_m u} + R_m u \leq (\sqrt{V_m} + R_m)u = K_m u$. Tail integration gives, after enlarging C,

$$\mathbb{E}\left[\exp\left(\frac{M_m}{C K_m} \right) \middle| \mathcal{F}_{s_m} \right] \leq 2$$

whenever $K_m > 0$. Equivalently, for universal constants $c_0, C_0 > 0$, whenever $\alpha_m K_m > 0$ and

$$0 \leq \theta \leq \frac{c_0}{\alpha_m K_m},$$

we have

$$\mathbb{E}[\exp(\theta \alpha_m M_m) | \mathcal{F}_{s_m}] \leq \exp(C_0 \theta \alpha_m K_m).$$

Blocks with $\alpha_m K_m = 0$ contribute zero and may be ignored. Set

$$K_\star := \max_{0 \leq m < M} \alpha_m K_m.$$

If $K_\star = 0$, the claim is trivial. For $0 \le \theta \le c_0/K_\star$, iterating the conditional moment-generating-function estimates over the deterministic blocks yields

$$\mathbb{E}\exp\left(\theta \sum_{m=0}^{M-1} \alpha_m M_m\right) \le \exp\left(C_0\theta \sum_{m=0}^{M-1} \alpha_m K_m\right).$$

Taking $\theta = c_0/K_\star$ and applying Markov's inequality yields, with probability at least $1-\delta$,

$$\sum_{m=0}^{M-1} \alpha_m M_m \le C \sum_{m=0}^{M-1} \alpha_m K_m + CK_\star \log(1/\delta).$$

Absorbing constants into Γ_δ proves the claim. ■

Appendix B. Proofs of Nonconvex–PL Games

B.1. SGDA in Subgaussian NC-PL Games

B.1.1. LEMMA 4.1

Proof Fix t and write $\eta_x = \eta_{x,t}$, $\xi^x = \xi_t^x$, and $\delta_t := \nabla_x f(x_t, y_t) - \nabla g(x_t)$. Then, we have:

$$x_{t+1} = x_t - \eta_x\big(\nabla_x f(x_t, y_t) + \xi^x\big) = x_t - \eta_x\big(\nabla g(x_t) + \delta_t + \xi^x\big).$$

By Lemma A.2, g is $2\kappa\ell$-smooth. Therefore, it follows:

$$\begin{aligned}
g(x_{t+1}) &\le g(x_t) + \langle \nabla g(x_t), x_{t+1} - x_t\rangle + \kappa\ell\|x_{t+1} - x_t\|^2 \\
&= g(x_t) - \eta_x\|\nabla g(x_t)\|^2 - \eta_x\langle \nabla g(x_t), \delta_t\rangle - \eta_x\langle \nabla g(x_t), \xi^x\rangle + \kappa\ell\eta_x^2\|\nabla g(x_t) + \delta_t + \xi^x\|^2 \\
&\le g(x_t) - \eta_x\|\nabla g(x_t)\|^2 + \frac{\eta_x}{4}\|\nabla g(x_t)\|^2 + \eta_x\|\delta_t\|^2 - \eta_x\langle \nabla g(x_t), \xi^x\rangle \\
&\quad + 3\kappa\ell\eta_x^2\Big(\|\nabla g(x_t)\|^2 + \|\delta_t\|^2 + \|\xi^x\|^2\Big) \\
&= g(x_t) - \eta_x\Big(\frac{3}{4} - 3\kappa\ell\eta_x\Big)\|\nabla g(x_t)\|^2 + \big(\eta_x + 3\kappa\ell\eta_x^2\big)\|\delta_t\|^2 - \eta_x\langle \nabla g(x_t), \xi^x\rangle + 3\kappa\ell\eta_x^2\|\xi^x\|^2 \\
&\le g(x_t) - \frac{\eta_x}{2}\|\nabla g(x_t)\|^2 + 2\eta_x\|\delta_t\|^2 - \eta_x\langle \nabla g(x_t), \xi^x\rangle + 3\kappa\ell\eta_x^2\|\xi^x\|^2.
\end{aligned}$$

The second inequality uses $-\langle u, v\rangle \le \frac{1}{4}\|u\|^2 + \|v\|^2$ and $\|u+v+w\|^2 \le 3(\|u\|^2 + \|v\|^2 + \|w\|^2)$. The last line uses $\eta_x \le 1/(12\kappa\ell)$.

Finally, Lemma A.3 gives $\|\delta_t\|^2 \le 2\kappa\ell b_t$. Substituting this and subtracting g_* from both sides yields:

$$a_{t+1} \le a_t - \frac{\eta_x}{2}\|\nabla g(x_t)\|^2 + 4\eta_x\kappa\ell b_t - \eta_x\langle \nabla g(x_t), \xi^x\rangle + 3\kappa\ell\eta_x^2\|\xi^x\|^2,$$

which is the desired claim. ■

B.1.2. LEMMA 4.2

Proof Fix t and write $\eta_x = \eta_{x,t}$ and $\eta_y = \eta_{y,t}$. Since

$$\phi_t(y) = g(x_{t+1}) - f(x_{t+1}, y), \qquad \nabla\phi_t(y_t) = -\nabla_y f(x_{t+1}, y_t).$$

Therefore, the y-update can be written as

$$y_{t+1} = y_t + \eta_y\big(\nabla_y f(x_{t+1}, y_t) + \xi_t^y\big) = y_t - \eta_y\nabla\phi_t(y_t) + \eta_y\xi_t^y.$$

By ℓ-smoothness of ϕ_t,

$$\begin{aligned} b_{t+1} &= \phi_t(y_{t+1}) \\ &\le \phi_t(y_t) + \langle\nabla\phi_t(y_t), -\eta_y\nabla\phi_t(y_t) + \eta_y\xi_t^y\rangle + \frac{\ell}{2}\eta_y^2\| - \nabla\phi_t(y_t) + \xi_t^y\|^2 \\ &= b_{t+1/2} - \Big(\eta_y - \frac{\ell}{2}\eta_y^2\Big)\|\nabla\phi_t(y_t)\|^2 + \big(\eta_y - \ell\eta_y^2\big)\langle\nabla\phi_t(y_t), \xi_t^y\rangle + \frac{\ell}{2}\eta_y^2\|\xi_t^y\|^2. \end{aligned}$$

This proves the first claim. For the half-step bound, define $b(x, y) := g(x) - f(x, y)$. By Lemma A.2 and Assumption 3.1, the function $x \mapsto b(x, y_t)$ is $4\kappa\ell$-smooth. Moreover, $\nabla_x b(x_t, y_t) = \nabla g(x_t) - \nabla_x f(x_t, y_t) = -\delta_t$ and $x_{t+1} - x_t = -\eta_x(\nabla g(x_t) + \delta_t + \xi_t^x)$. Therefore, it follows:

$$\begin{aligned} b_{t+1/2} = b(x_{t+1}, y_t) &\le b(x_t, y_t) + \langle\nabla_x b(x_t, y_t), x_{t+1} - x_t\rangle + 2\kappa\ell\|x_{t+1} - x_t\|^2 \\ &\le b_t + \eta_x\langle\delta_t, \nabla g(x_t) + \delta_t + \xi_t^x\rangle + 2\kappa\ell\eta_x^2\|\nabla g(x_t) + \delta_t + \xi_t^x\|^2 \\ &\le b_t + \Big(\frac{1}{4}\eta_x + 6\kappa\ell\eta_x^2\Big)\|\nabla g(x_t)\|^2 + \big(2\eta_x + 6\kappa\ell\eta_x^2\big)\|\delta_t\|^2 + \eta_x\langle\delta_t, \xi_t^x\rangle + 6\kappa\ell\eta_x^2\|\xi_t^x\|^2 \\ &\le \big(1 + 4\kappa\ell\eta_x + 12\kappa^2\ell^2\eta_x^2\big)b_t + \Big(\frac{1}{4}\eta_x + 6\kappa\ell\eta_x^2\Big)\|\nabla g(x_t)\|^2 + \eta_x\langle\delta_t, \xi_t^x\rangle + 6\kappa\ell\eta_x^2\|\xi_t^x\|^2. \end{aligned}$$

The second inequality uses $\langle u, v\rangle \le \frac{1}{4}\|v\|^2 + \|u\|^2$ and $\|u + v + w\|^2 \le 3(\|u\|^2 + \|v\|^2 + \|w\|^2)$. The last inequality uses Lemma A.3, namely $\|\delta_t\|^2 \le 2\kappa\ell b_t$. This is the half-step claim. ■

B.1.3. LEMMA 4.3

Proof Fix t and write $\eta_x = \eta_{x,t}$ and $\eta_y = \eta_{y,t}$. Let $\theta_t \overset{\text{def}}{=} \frac{1}{2} - 2\mu\gamma_{y,t} = \frac{1-\mu(\eta_y - \ell\eta_y^2/2)}{2}$. Decompose the potential at the half step and insert Lemmas 4.1 and 4.2 as:

$$\begin{aligned} P_t - P_{t+1} &= (a_t - a_{t+1}) + \frac{1}{2}(b_t - b_{t+1/2}) + \frac{1}{2}(b_{t+1/2} - b_{t+1}) \\ &\ge \frac{\eta_x}{2}\|\nabla g(x_t)\|^2 - 4\kappa\ell\eta_x b_t + \frac{1}{2}b_t - \frac{1}{2}b_{t+1/2} + \frac{1}{2}\Big(\eta_y - \frac{\ell}{2}\eta_y^2\Big)\|\nabla\phi_t(y_t)\|^2 \\ &\quad + \eta_x\langle\nabla g(x_t), \xi_t^x\rangle - 3\kappa\ell\eta_x^2\|\xi_t^x\|^2 - \frac{1}{2}(\eta_y - \ell\eta_y^2)\langle\nabla\phi_t(y_t), \xi_t^y\rangle - \frac{\ell}{4}\eta_y^2\|\xi_t^y\|^2. \end{aligned}$$

At fixed x_{t+1}, apply the NC-PL condition to $\phi_t(y) = g(x_{t+1}) - f(x_{t+1}, y)$:

$$\|\nabla\phi_t(y_t)\|^2 = \|\nabla_y f(x_{t+1}, y_t)\|^2 \ge 2\mu b_{t+1/2}.$$

Since $\gamma_{y,t} = \frac{1}{4}(\eta_y - \ell\eta_y^2/2)$, this converts the half-step term as

$$\frac{1}{2}b_t - \frac{1}{2}b_{t+1/2} + \frac{1}{2}\Big(\eta_y - \frac{\ell}{2}\eta_y^2\Big)\|\nabla\phi_t(y_t)\|^2 \geq \gamma_{y,t}\|\nabla\phi_t(y_t)\|^2 + \frac{1}{2}b_t - \theta_t b_{t+1/2}.$$

Moreover, $\eta_y \leq 1/\ell$ and $\mu \leq \ell$ imply $0 \leq \mu(\eta_y - \ell\eta_y^2/2) \leq 1$, hence $0 \leq \theta_t \leq 1/2$. Using the half-step estimate from Lemma 4.2 and $\theta_t \geq 0$,

$$\begin{aligned} P_t - P_{t+1} \geq \Big[\frac{\eta_x}{2} - \theta_t\Big(\frac{1}{4}\eta_x + 6\kappa\ell\eta_x^2\Big)\Big]\|\nabla g(x_t)\|^2 + \gamma_{y,t}\|\nabla\phi_t(y_t)\|^2 + \Big[\frac{1}{2} - \theta_t\alpha_t - 4\kappa\ell\eta_x\Big]b_t \\ + \eta_x\langle\nabla g(x_t), \xi_t^x\rangle - \theta_t\eta_x\langle\delta_t, \xi_t^x\rangle - (3 + 6\theta_t)\kappa\ell\eta_x^2\|\xi_t^x\|^2 - \frac{1}{2}(\eta_y - \ell\eta_y^2)\langle\nabla\phi_t(y_t), \xi_t^y\rangle - \frac{\ell}{4}\eta_y^2\|\xi_t^y\|^2. \end{aligned}$$

It remains to lower bound the two deterministic coefficients. For the gradient coefficient, $\theta_t \leq 1/2$ and $\eta_x \leq \eta_y/(64\kappa^2) \leq 1/(12\kappa\ell)$ give

$$\frac{\eta_x}{2} - \theta_t\Big(\frac{1}{4}\eta_x + 6\kappa\ell\eta_x^2\Big) \geq \frac{\eta_x}{2} - \frac{1}{2}\Big(\frac{1}{4}\eta_x + 6\kappa\ell\eta_x^2\Big) = \frac{3}{8}\eta_x - 3\kappa\ell\eta_x^2 \geq \frac{1}{8}\eta_x.$$

For the b_t coefficient, the step-size assumptions give

$$\mu\Big(\eta_y - \frac{\ell}{2}\eta_y^2\Big) = \frac{\ell\eta_y}{\kappa}\left(1 - \frac{\ell\eta_y}{2}\right) \geq \frac{\ell\eta_y}{2\kappa} \geq 32\kappa\ell\eta_x.$$

Using $\alpha_t = 1 + 4\kappa\ell\eta_x + 12\kappa^2\ell^2\eta_x^2$,

$$\begin{aligned} \frac{1}{2} - \theta_t\alpha_t - 4\kappa\ell\eta_x &= \frac{\mu(\eta_y - \ell\eta_y^2/2)\alpha_t - (\alpha_t - 1)}{2} - 4\kappa\ell\eta_x \\ &\geq \frac{\mu(\eta_y - \ell\eta_y^2/2)}{2} - 6\kappa\ell\eta_x - 6\kappa^2\ell^2\eta_x^2 \\ &\geq 16\kappa\ell\eta_x - 6\kappa\ell\eta_x - 6\kappa^2\ell^2\eta_x^2 \geq 4\kappa\ell\eta_x, \end{aligned}$$

where the last inequality uses $\kappa\ell\eta_x \leq 1$. Finally, $\theta_t \leq 1/2$ implies $(3 + 6\theta_t)\kappa\ell\eta_x^2 \leq 6\kappa\ell\eta_x^2$. Substituting these coefficient bounds into the lower bound on $P_t - P_{t+1}$ and moving the martingale and quadratic-noise terms to the right-hand side gives the claim. ■

B.1.4. LEMMA 4.4

Proof Let $S_T := \sum_{t=0}^{T-1} G_t$ and $R_T := \sum_{t=0}^{T-1} r_t$. The term R_T is carried pathwise throughout the proof; no conditional expectation or moment-generating-function argument is applied to r_t. Summing the assumed one-step inequality and using $P_T \geq 0$ gives

$$S_T \leq P_0 + M_T + Q_T + R_T,$$

where $M_T := \sum_{t=0}^{T-1} \eta_{x,t}\langle c_t^x, \xi_t^x\rangle + \sum_{t=0}^{T-1} \eta_{y,t}\langle c_t^y, \xi_t^y\rangle$ is the linear martingale term, and $Q_T := \sum_{t=0}^{T-1} d_t^x\eta_{x,t}^2\|\xi_t^x\|^2 + \sum_{t=0}^{T-1} d_t^y\eta_{y,t}^2\|\xi_t^y\|^2$ is the quadratic noise term. We first control M_T. By Theorem A.8 applied conditionally with Assumptions 3.4 and 3.5, for every $\lambda > 0$,

$$\mathbb{E}\left[\exp\left(\lambda M_T - 4\lambda^2\sigma^2\sum_{t=0}^{T-1}\big(\eta_{x,t}^2\|c_t^x\|^2 + \eta_{y,t}^2\|c_t^y\|^2\big)\right)\right] \leq 1.$$

Thus, by Markov's inequality, with probability at least $1-\delta/2$,

$$M_T \le 4\lambda\sigma^2 \sum_{t=0}^{T-1} \big(\eta_{x,t}^2\|c_t^x\|^2 + \eta_{y,t}^2\|c_t^y\|^2\big) + \frac{1}{\lambda}\log\frac{2}{\delta}.$$

By self-bounding,

$$\sum_{t=0}^{T-1} \big(\eta_{x,t}^2\|c_t^x\|^2 + \eta_{y,t}^2\|c_t^y\|^2\big) \le C_1\eta_{\max} S_T.$$

If $C_1\eta_{\max} = 0$, then $M_T = 0$ almost surely. Otherwise, choose $\lambda = 1/(8\sigma^2 C_1\eta_{\max})$. Then, with probability at least $1-\delta/2$,

$$M_T \le \frac{1}{2}S_T + 8\sigma^2 C_1\eta_{\max}\log\frac{2}{\delta}.$$

Next control Q_T. Write $a_t^x := d_t^x\eta_{x,t}^2$, $a_t^y := d_t^y\eta_{y,t}^2$, and $a_{\max} := \max_{0\le t<T}\max\{a_t^x, a_t^y\}$. If $a_{\max} = 0$, then $Q_T = 0$. Otherwise, let $\theta := 1/(\sigma^2 a_{\max})$. We use the squared-norm exponential moment in Assumption 3.5 directly. Since $a_t^x/a_{\max} \in [0,1]$, Jensen's inequality applied to the convex function $u \mapsto e^u$ and the weights $a_t^x/a_{\max}$ and $1 - a_t^x/a_{\max}$ gives

$$\mathbb{E}\big[\exp\big(\theta a_t^x\|\xi_t^x\|^2\big) \mid \mathcal{F}_t\big] = \mathbb{E}\left[\exp\left(\frac{a_t^x}{a_{\max}}\frac{\|\xi_t^x\|^2}{\sigma^2}\right) \mid \mathcal{F}_t\right] \le e^{a_t^x/a_{\max}}.$$

The same argument gives

$$\mathbb{E}\big[\exp\big(\theta a_t^y\|\xi_t^y\|^2\big) \mid \mathcal{F}_{t+1/2}\big] \le e^{a_t^y/a_{\max}}.$$

Iterating over the half-step filtration yields

$$\mathbb{E}\exp(\theta Q_T) \le \exp\left(\frac{S_2(T)}{a_{\max}}\right).$$

By Markov's inequality, with probability at least $1-\delta/2$,

$$Q_T \le \sigma^2 S_2(T) + \sigma^2 a_{\max}\log\frac{2}{\delta}.$$

Because $a_{\max} \le S_2(T)$ and $\Gamma_\delta \ge 1$,

$$Q_T \le 2\sigma^2 S_2(T)\Gamma_\delta.$$

On the intersection of the two events, which has probability at least $1-\delta$,

$$S_T \le P_0 + \frac{1}{2}S_T + 8\sigma^2 C_1\eta_{\max}\log\frac{2}{\delta} + 2\sigma^2 S_2(T)\Gamma_\delta + R_T.$$

Rearranging and using $\log(2/\delta) \le \Gamma_\delta$ gives

$$S_T \le 2P_0 + 2R_T + 16\sigma^2 C_1\eta_{\max}\Gamma_\delta + 4\sigma^2 S_2(T)\Gamma_\delta.$$

■

B.1.5. THEOREM B.1

Theorem B.1 (Convergence of SGDA in Subgaussian NC-PL games) *Suppose Assumptions 3.1, 3.2, 3.4 and 3.5 hold. Let $T \geq 1$ and $\delta \in (0,1)$. Suppose the deterministic step sizes satisfy*

$$\eta_{y,t} = 64\kappa^2 \eta_{x,t}, \qquad 0 \leq \eta_{x,t} \leq \frac{1}{64\kappa^2 \ell}, \qquad t = 0, \ldots, T-1.$$

Define

$$\Gamma_\delta := \max\left\{1, \log\frac{2}{\delta}\right\}, \qquad S_1(T) := \sum_{t=0}^{T-1} \eta_{x,t},$$

$$A_{\kappa,\ell} := (6\kappa + 1024\kappa^4)\ell, \qquad S_2(T) := A_{\kappa,\ell} \sum_{t=0}^{T-1} \eta_{x,t}^2,$$

and

$$\eta_{\max} := \max_{0 \leq t < T} \eta_{y,t} = 64\kappa^2 \max_{0 \leq t < T} \eta_{x,t}.$$

If $S_1(T) > 0$, then with probability at least $1-\delta$,

$$\min_{0 \leq t < T} \|\nabla g(x_t)\|^2 \leq \frac{16 P_0}{S_1(T)} + \frac{32\sigma^2\big(4C_1\eta_{\max} + S_2(T)\big)\Gamma_\delta}{S_1(T)},$$

where $C_1 := 73/8$.

Proof For each t, write $\phi_t(y) := g(x_{t+1}) - f(x_{t+1}, y)$, $\delta_t := \nabla_x f(x_t, y_t) - \nabla g(x_t)$, and $\gamma_{y,t} := \frac{1}{4}\eta_{y,t} - \frac{\ell}{8}\eta_{y,t}^2$. Define

$$G_t := \frac{1}{8}\eta_{x,t}\|\nabla g(x_t)\|^2 + \gamma_{y,t}\|\nabla \phi_t(y_t)\|^2 + 4\kappa\ell\eta_{x,t} b_t.$$

By Lemma 4.3, with $\theta_t := \frac{1}{2} - 2\mu\gamma_{y,t} = \{1 - \mu(\eta_{y,t} - \ell\eta_{y,t}^2/2)\}/2 \in [0, 1/2]$,

$$\begin{aligned} G_t \leq P_t - P_{t+1} - \eta_{x,t}\langle \nabla g(x_t), \xi_t^x\rangle + \theta_t \eta_{x,t}\langle \delta_t, \xi_t^x\rangle + 6\kappa\ell\eta_{x,t}^2\|\xi_t^x\|^2 \\ + \frac{1}{2}(\eta_{y,t} - \ell\eta_{y,t}^2)\langle \nabla\phi_t(y_t), \xi_t^y\rangle + \frac{\ell}{4}\eta_{y,t}^2\|\xi_t^y\|^2, \end{aligned}$$

This is in the form of Lemma 4.4 with $c_t^x := -\nabla g(x_t) + \theta_t\delta_t$, $d_t^x := 6\kappa\ell$, $c_t^y := \frac{1}{2}(1 - \ell\eta_{y,t})\nabla\phi_t(y_t)$, $d_t^y := \ell/4$, and $r_t = 0$, since $\eta_{y,t}\langle c_t^y, \xi_t^y\rangle = \frac{1}{2}(\eta_{y,t} - \ell\eta_{y,t}^2)\langle \nabla\phi_t(y_t), \xi_t^y\rangle$. We verify self-bounding. Since $0 \leq \theta_t \leq 1/2$, for any $\beta > 0$,

$$\begin{aligned} \|c_t^x\|^2 &\leq (1+\beta)\|\nabla g(x_t)\|^2 + \left(1 + \frac{1}{\beta}\right)\theta_t^2\|\delta_t\|^2 \\ &\leq (1+\beta)\|\nabla g(x_t)\|^2 + \left(1 + \frac{1}{\beta}\right)\frac{\kappa\ell}{2} b_t, \end{aligned}$$

where the last line uses Lemma A.3, $\|\delta_t\|^2 \leq 2\kappa\ell b_t$. Therefore

$$\eta_{x,t}\|c_t^x\|^2 \leq (1+\beta)\eta_{x,t}\|\nabla g(x_t)\|^2 + \left(1 + \frac{1}{\beta}\right)\frac{\kappa\ell}{2}\eta_{x,t} b_t.$$

Choosing $\beta = 1/64$ and comparing with the $\frac{1}{8}\eta_{x,t}\|\nabla g(x_t)\|^2$ and $4\kappa\ell\eta_{x,t}b_t$ terms inside G_t gives

$$\eta_{x,t}\|c_t^x\|^2 \le \frac{65}{8}G_t.$$

For the y-part, set $u_t := \ell\eta_{y,t} \in [0,1]$. Since $\frac{1}{4}(1-u_t)^2 \le \frac{1}{8}(2-u_t)$ on $[0,1]$,

$$\begin{aligned}\eta_{y,t}\|c_t^y\|^2 &= \frac{\eta_{y,t}}{4}(1-u_t)^2\|\nabla\phi_t(y_t)\|^2 \\ &\le \frac{\eta_{y,t}}{8}(2-u_t)\|\nabla\phi_t(y_t)\|^2 = \gamma_{y,t}\|\nabla\phi_t(y_t)\|^2 \le G_t.\end{aligned}$$

Combining gives

$$\eta_{x,t}\|c_t^x\|^2 + \eta_{y,t}\|c_t^y\|^2 \le \frac{73}{8}G_t.$$

Also,

$$d_t^x\eta_{x,t}^2 + d_t^y\eta_{y,t}^2 = 6\kappa\ell\eta_{x,t}^2 + \frac{\ell}{4}(64\kappa^2)^2\eta_{x,t}^2 = (6\kappa + 1024\kappa^4)\ell\,\eta_{x,t}^2.$$

Thus the quadratic normalization in Lemma 4.4 is $S_2(T) = A_{\kappa,\ell}\sum_{t=0}^{T-1}\eta_{x,t}^2$. Applying Lemma 4.4, with probability at least $1-\delta$,

$$\sum_{t=0}^{T-1} G_t \le 2P_0 + 16\sigma^2 C_1\eta_{\max}\Gamma_\delta + 4\sigma^2 S_2(T)\Gamma_\delta.$$

Since $G_t \ge \frac{1}{8}\eta_{x,t}\|\nabla g(x_t)\|^2$, we have

$$\sum_{t=0}^{T-1} G_t \ge \frac{1}{8}S_1(T)\min_{0\le t<T}\|\nabla g(x_t)\|^2.$$

Combining and multiplying by $8/S_1(T)$ gives

$$\min_{0\le t<T}\|\nabla g(x_t)\|^2 \le \frac{16P_0}{S_1(T)} + \frac{128\sigma^2 C_1\eta_{\max}\Gamma_\delta}{S_1(T)} + \frac{32\sigma^2 S_2(T)\Gamma_\delta}{S_1(T)}.$$

This is equivalent to the displayed bound. ■

B.1.6. THEOREM 4.5

Proof The chosen step sizes satisfy $\eta_x \le 1/(64\kappa^2\ell)$ and $\eta_y = 64\kappa^2\eta_x$, so Theorem B.1 applies. For fixed steps, $S_1(T) = T\eta_x$, $S_2(T) = A_{\kappa,\ell}T\eta_x^2$, and $\eta_{\max} = 64\kappa^2\eta_x$. Substituting these fixed-step quantities into Theorem B.1 gives, with probability at least $1-\delta$,

$$\min_{0\le t<T}\|\nabla g(x_t)\|^2 \le \frac{16P_0}{T\eta_x} + \frac{8192C_1\kappa^2\sigma^2\Gamma_\delta}{T} + 32A_{\kappa,\ell}\sigma^2\eta_x\Gamma_\delta.$$

The first and third terms are the only ones that depend on the balancing choice of η_x. Let $h := 1/(64\kappa^2\ell)$ and $a_T := \sqrt{P_0/(A_{\kappa,\ell}\sigma^2 T\Gamma_\delta)}$; since $\eta_x = \min\{h, a_T\}$, we have $1/\eta_x \le 1/h + 1/a_T$ and $\eta_x \le a_T$. Therefore

$$\begin{aligned}\frac{16P_0}{T\eta_x} + 32A_{\kappa,\ell}\sigma^2\eta_x\Gamma_\delta &\le 16\left(\frac{P_0}{Th} + \frac{P_0}{Ta_T}\right) + 32A_{\kappa,\ell}\sigma^2 a_T\Gamma_\delta \\ &= \frac{1024\kappa^2\ell P_0}{T} + \frac{48\sigma\sqrt{A_{\kappa,\ell}P_0\Gamma_\delta}}{\sqrt{T}}.\end{aligned}$$

The inequality estimates the initial-error term using the cap h and the variance-balancing scale a_T, while the last stochastic-accumulation term uses only $\eta_x \le a_T$. Substituting this estimate into the master bound gives

$$\begin{aligned}\min_{0\le t<T}\|\nabla g(x_t)\|^2 &\le \frac{1024\kappa^2\ell P_0}{T} + \frac{48\sigma\sqrt{A_{\kappa,\ell}P_0\Gamma_\delta}}{\sqrt{T}} + \frac{8192C_1\kappa^2\sigma^2\Gamma_\delta}{T} \\ &= \mathcal{O}\left(\frac{\kappa^2\ell P_0}{T} + \frac{\kappa^2\sigma\sqrt{\ell P_0\Gamma_\delta}}{\sqrt{T}} + \frac{\kappa^2\sigma^2\Gamma_\delta}{T}\right),\end{aligned}$$

because $C_1 = 73/8$ is numerical and, for $\kappa \ge 1$, $A_{\kappa,\ell} = (6\kappa + 1024\kappa^4)\ell \le 1030\kappa^4\ell$. ■

B.1.7. THEOREM 4.6

Proof Let $h := 1/(64\kappa^2\ell)$ and $a := \sqrt{P_0/(A_{\kappa,\ell}\sigma^2\Gamma_\delta)}$, so $\eta_{x,t} = \min\{h, a/\sqrt{t+1}\}$. The step-size assumptions of Theorem B.1 hold because $\eta_{x,t} \le h = 1/(64\kappa^2\ell)$ and $\eta_{y,t} = 64\kappa^2\eta_{x,t}$. Write $S_1(T) := \sum_{t=0}^{T-1}\eta_{x,t}$ and $Q_2(T) := \sum_{t=0}^{T-1}\eta_{x,t}^2$. Theorem B.1 gives, with probability at least $1-\delta$,

$$\min_{0\le t<T}\|\nabla g(x_t)\|^2 \le \frac{16P_0}{S_1(T)} + \frac{128\sigma^2C_1\eta_{\max}\Gamma_\delta}{S_1(T)} + \frac{32\sigma^2A_{\kappa,\ell}Q_2(T)\Gamma_\delta}{S_1(T)},$$

where $\eta_{\max} = 64\kappa^2\max_{0\le t<T}\eta_{x,t}$. First lower bound $S_1(T)$. Since $a/\sqrt{t+1} \ge a/\sqrt{T}$ for every $t < T$,

$$S_1(T) = \sum_{t=0}^{T-1}\min\left\{h, \frac{a}{\sqrt{t+1}}\right\} \ge T\min\left\{h, \frac{a}{\sqrt{T}}\right\} = \min\{hT, a\sqrt{T}\} \implies \frac{1}{S_1(T)} \le \frac{1}{hT} + \frac{1}{a\sqrt{T}}.$$

Consequently

$$\frac{P_0}{S_1(T)} \le \frac{P_0}{hT} + \frac{P_0}{a\sqrt{T}} = \frac{64\kappa^2\ell P_0}{T} + \frac{\sigma\sqrt{A_{\kappa,\ell}P_0\Gamma_\delta}}{\sqrt{T}}.$$

Next control the $\eta_{\max}$ term. Let $\eta_{\max}^x := \max_{0\le t<T}\eta_{x,t}$. If $h \le a/\sqrt{T}$, then all steps are capped, so $\eta_{\max}^x/S_1(T) = h/(hT) = 1/T \le 1/\sqrt{T}$. If $h > a/\sqrt{T}$, then $S_1(T) \ge a\sqrt{T}$ and $\eta_{\max}^x \le a$, so again $\eta_{\max}^x/S_1(T) \le 1/\sqrt{T}$. Therefore

$$\frac{\eta_{\max}}{S_1(T)} = 64\kappa^2\frac{\eta_{\max}^x}{S_1(T)} \le \frac{64\kappa^2}{\sqrt{T}}.$$

Thus $\sigma^2\Gamma_\delta\eta_{\max}/S_1(T) = O(\kappa^2\sigma^2\Gamma_\delta/\sqrt{T})$. Finally control the quadratic-noise term. If $h \le a/\sqrt{T}$, then all steps are capped and $Q_2(T)/S_1(T) = Th^2/(Th) = h \le a/\sqrt{T}$. If $h > a/\sqrt{T}$, then $S_1(T) \ge a\sqrt{T}$ and $Q_2(T) \le a^2\sum_{t=0}^{T-1}(t+1)^{-1} \le a^2(1+\log T)$. Therefore, in both cases,

$$\frac{Q_2(T)}{S_1(T)} \le \frac{a(1+\log T)}{\sqrt{T}}.$$

Hence

$$\frac{\sigma^2 A_{\kappa,\ell}\Gamma_\delta Q_2(T)}{S_1(T)} \le \frac{\sigma^2 A_{\kappa,\ell}\Gamma_\delta a(1+\log T)}{\sqrt{T}} = \frac{\sigma\sqrt{A_{\kappa,\ell}P_0\Gamma_\delta}(1+\log T)}{\sqrt{T}},$$

where the last equality substitutes $a = \sqrt{P_0/(A_{\kappa,\ell}\sigma^2\Gamma_\delta)}$. Combining the three bounds gives

$$\min_{0\le t<T}\|\nabla g(x_t)\|^2 = O\left(\frac{\kappa^2\ell P_0}{T} + \frac{\sigma\sqrt{A_{\kappa,\ell}P_0\Gamma_\delta}(1+\log T)}{\sqrt{T}} + \frac{\kappa^2\sigma^2\Gamma_\delta}{\sqrt{T}}\right).$$

Since $A_{\kappa,\ell} \le 1030\kappa^4\ell$ for $\kappa \ge 1$, we have $\sigma\sqrt{A_{\kappa,\ell}P_0\Gamma_\delta} = O(\kappa^2\sigma\sqrt{\ell P_0\Gamma_\delta})$. This proves the claimed rate. ■

B.2. SGDA in Heavy-Tailed NC-PL Games

B.2.1. PROPOSITION 4.7

Proof Fix any $p \in (1,2]$, deterministic step size schedules $\{\eta_{x,t}\}_{t\ge0}, \{\eta_{y,t}\}_{t\ge0}$, a horizon $T \ge 2$ with at least one active x-update before time T, and a confidence level $\delta \in (0,1)$. We construct a one-dimensional NC–PL game and an unbiased heavy-tailed oracle, which may depend on (T,δ), for which $\mathbb{P}\Big(T^{-1}\sum_{t=0}^{T-1}\|\nabla g(x_t)\|^2 \ge 1/(\delta T)\Big) \ge \delta$. This immediately implies $T = \Omega(1/(\varepsilon\delta))$ for any general high-probability guarantee over the oracle class satisfying Assumptions 3.4 and 3.6.

Step 1: An NC–PL game with strongly concave inner problem. Let $d_x = d_y = 1$, $\mathcal{Y} = \mathbb{R}$, and $f(x,y) := x^2/2 - y^2/2$. Then

$$g(x) = \max_{y\in\mathbb{R}} f(x,y) = \frac{1}{2}x^2, \quad g(x) - f(x,y) = \frac{1}{2}y^2,$$
$$\nabla g(x) = x, \quad \nabla_x f(x,y) = x, \quad \nabla_y f(x,y) = -y, \frac{1}{2}\|\nabla_y f(x,y)\|^2 = \frac{1}{2}y^2.$$

Thus Assumption 3.1 holds with $\ell = 1$, the inner problem is 1-strongly concave with unique maximizer $y^\star(x) = 0$, and Assumption 3.2 holds with $\mu = 1$.

Step 2: SGDA reduces to noisy SGD on a quadratic. Run SGDA from $(x_0,y_0) = (0,0)$, take the y-oracle to be exact, $G_y(x,y,\zeta_t^y) \equiv -y$, and take the x-oracle to be $G_x(x_t,y_t,\zeta_t^x) = x_t + \xi_t$, where $\{\xi_t\}$ is specified below. Then $\xi_t^y \equiv 0$, hence $y_t \equiv 0$ for all $t \ge 0$, and the x-recursion is

$$x_{t+1} = x_t - \eta_{x,t}(x_t + \xi_t) = (1-\eta_{x,t})x_t - \eta_{x,t}\xi_t.$$

Step 3: A heavy-tailed unbiased noise at *all* steps up to the horizon. Let $S := \{0 \le t \le T-2 : \eta_{x,t} > 0\}$ and $m := |S|$, so $m \ge 1$ by assumption. Set $q := 1-(1-\delta)^{1/m} \in (0,1)$; since $(1-\delta)^{1/m} \ge 1-\delta$, we have $q \le \delta$. For $t \in S$, define independent noises by

$$\xi_t := \begin{cases} \dfrac{1}{\eta_{x,t}\sqrt{\delta}}, & \text{with probability } \dfrac{q}{2}, \\ -\dfrac{1}{\eta_{x,t}\sqrt{\delta}}, & \text{with probability } \dfrac{q}{2}, \\ 0, & \text{with probability } 1-q, \end{cases}$$

and $\xi_t := 0$ almost surely for $t \notin S$. The distribution is symmetric, and for $t \in S$,

$$\mathbb{E}[\xi_t \mid \mathcal{F}_t] = 0, \quad \mathbb{E}[|\xi_t|^p \mid \mathcal{F}_t] = q\left(\frac{1}{\eta_{x,t}\sqrt{\delta}}\right)^p = \frac{q}{\eta_{x,t}^p \delta^{p/2}} \le \frac{\delta^{1-p/2}}{\eta_{x,t}^p} \le \frac{1}{\eta_{x,t}^p}.$$

The same bounds are trivial for $t \notin S$. Hence Assumptions 3.4 and 3.6 hold with, for example, $\sigma := \max_{t\in S} \eta_{x,t}^{-1}$. The oracle is nontrivial at every active x-oracle call that can affect the horizon, namely every $t \in S \subseteq \{0,\dots,T-2\}$.

Step 4: A rare outlier at an arbitrary step forces $\Omega(1/\delta)$ stationarity. Let $\tau := \min\{t \in S : \xi_t \ne 0\}$, with $\tau = +\infty$ if no such index exists. Independence and the choice of q give

$$\mathbb{P}(\tau < \infty) = 1-(1-q)^m = 1-(1-\delta) = \delta.$$

On $\{\tau < \infty\}$, all earlier noises that can affect the iterate are zero; since $x_0 = 0$, the recursion gives $x_t = 0$ for all $t \le \tau$ and hence

$$x_{\tau+1} = -\eta_{x,\tau}\xi_\tau, \quad |x_{\tau+1}| = \eta_{x,\tau}\cdot\frac{1}{\eta_{x,\tau}\sqrt{\delta}} = \frac{1}{\sqrt{\delta}}, \quad \|\nabla g(x_{\tau+1})\|^2 = x_{\tau+1}^2 = \frac{1}{\delta}.$$

Because $\tau \le T-2$, the index $\tau+1$ lies inside the averaging window, and therefore

$$\begin{aligned} &\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla g(x_t)\|^2 \ge \frac{1}{T}\|\nabla g(x_{\tau+1})\|^2 = \frac{1}{\delta T} \quad \text{on} \quad \{\tau < \infty\} \\ &\implies \mathbb{P}\left(\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla g(x_t)\|^2 \ge \frac{1}{\delta T}\right) \ge \mathbb{P}(\tau < \infty) = \delta. \end{aligned}$$

Thus any claim of the form $\mathbb{P}\Big(T^{-1}\sum_{t=0}^{T-1}\|\nabla g(x_t)\|^2 \le \varepsilon\Big) \ge 1-\delta$ must fail whenever $1/(\delta T) > \varepsilon$, because the constructed oracle gives the complementary event probability at least δ. Equivalently, any such guarantee requires $T \ge 1/(\varepsilon\delta)$, *i.e.*, $T = \Omega(1/(\varepsilon\delta))$. Finally, the constructed game satisfies that $f(x,\cdot)$ is 1-strongly concave in y, as claimed. ■

B.3. SGDA$_{\text{Clip}}$ in Heavy-Tailed NC-PL Games

B.3.1. LEMMA 4.8

Proof While this result is standard in the clipped-gradient literature (Nguyen et al., 2023; Sadiev et al., 2023), we include the proof below for completeness. We only prove the x-part, as the y-part is identical with $\tau_{y,t}$, and $g_y = \nabla_y f(x_{t+1}, y_t)$. Write $\tau \overset{\text{def}}{=} \tau_{x,t}$, $g \overset{\text{def}}{=} \nabla_x f(x_t, y_t)$, $\xi \overset{\text{def}}{=} G_t^x - g$, and $\widetilde{G} \overset{\text{def}}{=} \text{clip}_\tau(g + \xi)$, so that $\mathbb{E}[\xi \mid \mathcal{F}_t] = 0$, $\widetilde{\xi}_t^x = \widetilde{G} - \mathbb{E}[\widetilde{G} \mid \mathcal{F}_t]$, and $\beta_t^x = \mathbb{E}[\widetilde{G} \mid \mathcal{F}_t] - g$.

First, the norm bound follows directly from clipping and Jensen's inequality:

$$\|\widetilde{\xi}_t^x\| = \|\widetilde{G} - \mathbb{E}[\widetilde{G} \mid \mathcal{F}_t]\| \le \|\widetilde{G}\| + \|\mathbb{E}[\widetilde{G} \mid \mathcal{F}_t]\| \le \tau + \mathbb{E}[\|\widetilde{G}\| \mid \mathcal{F}_t] \le 2\tau.$$

Now assume the low-signal condition $\|g\| \le \tau/2$. Since clipping changes $g + \xi$ only on the event $\{\|g + \xi\| > \tau\}$, and since $\|g + \xi\| > \tau$ implies $\|\xi\| > \tau/2$, we have:

$$\begin{aligned}
\|\beta_t^x\| &= \|\mathbb{E}[\text{clip}_\tau(g + \xi) - (g + \xi) \mid \mathcal{F}_t]\| \\
&\le \mathbb{E}[\|\text{clip}_\tau(g + \xi) - (g + \xi)\| \mid \mathcal{F}_t] \\
&\le \mathbb{E}[\|g + \xi\| \mathbb{1}\{\|g + \xi\| > \tau\} \mid \mathcal{F}_t] \\
&\le 2\mathbb{E}[\|\xi\| \mathbb{1}\{\|\xi\| > \tau/2\} \mid \mathcal{F}_t] \\
&\le 2(\tau/2)^{1-p} \mathbb{E}[\|\xi\|^p \mid \mathcal{F}_t] \le 2^p \sigma^p \tau^{1-p} \le 4\sigma^p \tau^{1-p}.
\end{aligned}$$

The fourth line uses $\|g + \xi\| \le \|g\| + \|\xi\| \le 2\|\xi\|$ on $\{\|\xi\| > \tau/2\}$; the fifth uses $u\mathbb{1}\{u > a\} \le u^p a^{1-p}$ for $u \ge 0$, $a > 0$, and $p > 1$.

It remains to bound the conditional second moment. Since conditional expectation minimizes conditional mean-square deviation, we have:

$$\mathbb{E}[\|\widetilde{\xi}_t^x\|^2 \mid \mathcal{F}_t] = \mathbb{E}[\|\widetilde{G} - \mathbb{E}[\widetilde{G} \mid \mathcal{F}_t]\|^2 \mid \mathcal{F}_t] \le \mathbb{E}[\|\widetilde{G} - g\|^2 \mid \mathcal{F}_t].$$

Under $\|g\| \le \tau/2$, clipping gives $\|\widetilde{G} - g\| \le \|\widetilde{G}\| + \|g\| \le 3\tau/2$. Also, because $\text{clip}_\tau(\cdot)$ is the Euclidean projection onto the ball of radius τ and $\text{clip}_\tau(g) = g$, it is non-expansive and gives $\|\widetilde{G} - g\| \le \|\xi\|$. Therefore,

$$\mathbb{E}[\|\widetilde{G} - g\|^2 \mid \mathcal{F}_t] \le \left(\frac{3}{2}\tau\right)^{2-p} \mathbb{E}[\|\widetilde{G} - g\|^p \mid \mathcal{F}_t] \le \left(\frac{3}{2}\right)^{2-p} \tau^{2-p} \mathbb{E}[\|\xi\|^p \mid \mathcal{F}_t] \le 16\sigma^p \tau^{2-p}.$$

Combining the two yields the desired inequality $\mathbb{E}[\|\widetilde{\xi}_t^x\|^2 \mid \mathcal{F}_t] \le 16\sigma^p \tau^{2-p}$. ■

B.3.2. LEMMA 4.9

Proof Fix t and write $\eta_x = \eta_{x,t}$, $\eta_y = \eta_{y,t}$, and $\gamma_y = \gamma_{y,t}$. Abbreviate $\xi_t^x = \widetilde{\xi}_t^x$ and $\xi_t^y = \widetilde{\xi}_t^y$. By the definitions of the centered residuals and clipping biases,

$$\widetilde{G}_t^x = \nabla_x f(x_t, y_t) + \beta_t^x + \xi_t^x, \qquad \widetilde{G}_t^y = \nabla_y f(x_{t+1}, y_t) + \beta_t^y + \xi_t^y.$$

Also, for $\phi_t(y) := g(x_{t+1}) - f(x_{t+1}, y)$, we have $\nabla\phi_t(y_t) = -\nabla_y f(x_{t+1}, y_t)$ and hence $y_{t+1} = y_t - \eta_y(\nabla\phi_t(y_t) - \beta_t^y - \xi_t^y)$. The same smoothness calculations used in Lemmas 4.1 and 4.2, with

the bias terms handled by Young's inequality, give the three one-step estimates as:

$$\begin{aligned}
a_{t+1} &\le a_t - \frac{\eta_x}{2}\|\nabla g(x_t)\|^2 + 4\kappa\ell\eta_x b_t - \eta_x\langle \nabla g(x_t), \xi_t^x\rangle + 3\kappa\ell\eta_x^2\|\xi_t^x\|^2 + 4\eta_x\|\beta_t^x\|^2,\\
b_{t+1/2} &\le \alpha_t b_t + \left(\frac{1}{4}\eta_x + 6\kappa\ell\eta_x^2\right)\|\nabla g(x_t)\|^2 + \eta_x\langle \delta_t, \xi_t^x\rangle + 6\kappa\ell\eta_x^2\|\xi_t^x\|^2 + 4\eta_x\|\beta_t^x\|^2,\\
b_{t+1} &\le b_{t+1/2} - \left(\eta_y - \frac{\ell}{2}\eta_y^2\right)\|\nabla\phi_t(y_t)\|^2 + (\eta_y - \ell\eta_y^2)\langle\nabla\phi_t(y_t), \beta_t^y + \xi_t^y\rangle + \frac{\ell}{2}\eta_y^2\|\beta_t^y + \xi_t^y\|^2\\
&\le b_{t+1/2} - 4\gamma_y\|\nabla\phi_t(y_t)\|^2 + (\eta_y - \ell\eta_y^2)\langle\nabla\phi_t(y_t), \xi_t^y\rangle + \ell\eta_y^2\|\xi_t^y\|^2 + 4\eta_y\|\beta_t^y\|^2,
\end{aligned}$$

where $\alpha_t := 1 + 4\kappa\ell\eta_x + 12\kappa^2\ell^2\eta_x^2$. In the last line, we used $\eta_y \le 1/\ell$ to absorb $(\eta_y - \ell\eta_y^2)\langle\nabla\phi_t(y_t), \beta_t^y\rangle + \ell\eta_y^2\|\beta_t^y\|^2$ into $\eta_y\|\nabla\phi_t(y_t)\|^2/4 + 2\eta_y\|\beta_t^y\|^2$, and then used $\eta_y - \ell\eta_y^2/2 - \eta_y/4 \ge 4\gamma_y$.

Let $\theta_t := 1/2 - 2\mu\gamma_y$. Since $\eta_y \le 1/\ell$ and $\kappa = \ell/\mu \ge 1$, we have $0 \le \theta_t \le 1/2$. The NC-PL condition applied to ϕ_t gives $\|\nabla\phi_t(y_t)\|^2 \ge 2\mu b_{t+1/2}$. Combining this with the last display yields

$$\begin{aligned}
\frac{1}{2}(b_t - b_{t+1}) \ge \gamma_y\|\nabla\phi_t(y_t)\|^2 &+ \left(\frac{1}{2} - \theta_t\alpha_t\right) b_t - \theta_t\left(\frac{1}{4}\eta_x + 6\kappa\ell\eta_x^2\right)\|\nabla g(x_t)\|^2\\
&- \theta_t\eta_x\langle\delta_t, \xi_t^x\rangle - 6\theta_t\kappa\ell\eta_x^2\|\xi_t^x\|^2 - 4\theta_t\eta_x\|\beta_t^x\|^2\\
&- \frac{1}{2}(\eta_y - \ell\eta_y^2)\langle\nabla\phi_t(y_t), \xi_t^y\rangle - \frac{\ell}{2}\eta_y^2\|\xi_t^y\|^2 - 2\eta_y\|\beta_t^y\|^2.
\end{aligned}$$

Here the half-step estimate was substituted with a negative coefficient $-\theta_t$.

Now it remains to check the deterministic coefficients. The step-size assumptions imply:

$$\begin{aligned}
2\mu\gamma_y &= \frac{\mu}{4}\left(\eta_y - \frac{\ell}{2}\eta_y^2\right) = \frac{\ell\eta_y}{4\kappa}\left(1 - \frac{\ell\eta_y}{2}\right) \ge \frac{\ell\eta_y}{8\kappa} \ge 16\kappa\ell\eta_x,\\
\frac{1}{2} - \theta_t\alpha_t - 4\kappa\ell\eta_x &= 2\mu\gamma_y - \theta_t(\alpha_t - 1) - 4\kappa\ell\eta_x \ge 2\mu\gamma_y - 2\kappa\ell\eta_x - 6\kappa^2\ell^2\eta_x^2 - 4\kappa\ell\eta_x \ge 4\kappa\ell\eta_x,\\
\frac{\eta_x}{2} - \theta_t\left(\frac{1}{4}\eta_x + 6\kappa\ell\eta_x^2\right) &\ge \frac{3}{8}\eta_x - 3\kappa\ell\eta_x^2 \ge \frac{1}{8}\eta_x.
\end{aligned}$$

The second line uses $\theta_t \le 1/2$, $\alpha_t - 1 = 4\kappa\ell\eta_x + 12\kappa^2\ell^2\eta_x^2$, and $\kappa\ell\eta_x \le 1$; the last line uses $\eta_x \le 1/(12\kappa\ell)$, which follows from the step-size assumptions. Adding the primal estimate to the lower bound on $(b_t - b_{t+1})/2$, using the coefficient bounds, and moving the noise and bias terms to the right-hand side gives

$$\begin{aligned}
\frac{1}{8}\eta_x\|\nabla g(x_t)\|^2 + \gamma_y\|\nabla\phi_t(y_t)\|^2 + 4\kappa\ell\eta_x b_t &\le P_t - P_{t+1}\\
&- \eta_x\langle\nabla g(x_t), \xi_t^x\rangle + \left(\frac{1}{2} - 2\mu\gamma_y\right)\eta_x\langle\delta_t, \xi_t^x\rangle + 6\kappa\ell\eta_x^2\|\xi_t^x\|^2\\
&+ \frac{1}{2}(\eta_y - \ell\eta_y^2)\langle\nabla\phi_t(y_t), \xi_t^y\rangle + \frac{\ell}{2}\eta_y^2\|\xi_t^y\|^2 + 6\eta_x\|\beta_t^x\|^2 + 2\eta_y\|\beta_t^y\|^2.
\end{aligned}$$

■

B.3.3. LEMMA 4.10

Proof By construction, (B_t) is nondecreasing, so (z_t) is nonincreasing and $0 < z_t \le 1$. Since $\bar{G}_t \in \mathcal{F}_t$ and all other quantities entering B_t are deterministic or $\mathcal{F}_t$-measurable, z_t is predictable for the x-increment and $\mathcal{F}_{t+1/2}$-measurable for the y-increment. Moreover,

$$z_t B_t^x \le 1, \qquad z_t B_t^y \le 1, \qquad z_t B_t^v \le 1. \tag{28}$$

Fix $k \ge 1$. Multiplying the assumed one-step inequality by z_t, summing over $t < k$, using the monotonicity of (z_t), and separating the predictable quadratic drift from its centered fluctuation gives

$$\begin{aligned}
&\sum_{t=0}^{k-1} z_t G_t + z_{k-1} P_k \le z_0 P_0 + \sum_{t=0}^{k-1} z_t r_t + \sum_{t=0}^{k-1} z_t D_t + S_k, \qquad (29)\\
&S_k \overset{\text{def}}{=} \sum_{t=0}^{k-1} z_t \Big[\eta_{x,t} \langle c_t^x, \tilde{\xi}_t^x \rangle + d_t^x \eta_{x,t}^2 \big(\|\tilde{\xi}_t^x\|^2 - \mathbb{E}[\|\tilde{\xi}_t^x\|^2 \mid \mathcal{F}_t] \big) \\
&\qquad\qquad + \eta_{y,t} \langle c_t^y, \tilde{\xi}_t^y \rangle + d_t^y \eta_{y,t}^2 \big(\|\tilde{\xi}_t^y\|^2 - \mathbb{E}[\|\tilde{\xi}_t^y\|^2 \mid \mathcal{F}_{t+1/2}] \big) \Big].
\end{aligned}$$

The telescoping part used here is $\sum_{t<k} z_t (P_t - P_{t+1}) = z_0 P_0 - z_{k-1} P_k + \sum_{t=1}^{k-1} (z_t - z_{t-1}) P_t \le z_0 P_0 - z_{k-1} P_k$, and the drift separation uses $d_t^\bullet \ge 0$ and the conditional second-moment bounds.

Define the half-step filtration by $\mathcal{H}_{2t} := \mathcal{F}_t$ and $\mathcal{H}_{2t+1} := \mathcal{F}_{t+1/2}$, and let Δ_{2t} and Δ_{2t+1} be the x- and y-summands in S_k, respectively. Then $S_k = \sum_{s=0}^{2k-1} \Delta_s$ and $\mathbb{E}[\Delta_s \mid \mathcal{H}_s] = 0$.

We next verify the bounded-increment and variance conditions. It suffices to write the x-case; the y-case is identical. By Cauchy–Schwarz, clipping, and the self-bounding condition,

$$\begin{aligned}
&\eta_{x,t} |\langle c_t^x, \tilde{\xi}_t^x \rangle| \le 2\tau_{x,t} \eta_{x,t} \|c_t^x\| \le 2\tau_{x,t} \sqrt{C_1 \eta_{x,t} G_t} \le 2\tau_{x,t} \sqrt{C_1 \eta_{x,t} \bar{G}_t},\\
&d_t^x \eta_{x,t}^2 \left| \|\tilde{\xi}_t^x\|^2 - \mathbb{E}[\|\tilde{\xi}_t^x\|^2 \mid \mathcal{F}_t] \right| \le 8 d_t^x \eta_{x,t}^2 \tau_{x,t}^2.
\end{aligned}$$

Thus $|\Delta_{2t}| \le z_t B_t^x \le 1$, and similarly $|\Delta_{2t+1}| \le z_t B_t^y \le 1$.

For the conditional variance, use $(a+b)^2 \le 2a^2 + 2b^2$ directly on the two terms inside Δ_{2t}. The linear part is controlled by self-bounding and $z_t B_t^v \le 1$, while the centered quadratic part is controlled by clipping and $z_t B_t^x \le 1$:

$$\begin{aligned}
\mathbb{E}[\Delta_{2t}^2 \mid \mathcal{F}_t] &\le 2 z_t^2 \eta_{x,t}^2 \|c_t^x\|^2 v_t^x + 2 z_t^2 (d_t^x)^2 \eta_{x,t}^4 \mathbb{E}\left[\left(\|\tilde{\xi}_t^x\|^2 - \mathbb{E}[\|\tilde{\xi}_t^x\|^2 \mid \mathcal{F}_t] \right)^2 \middle| \mathcal{F}_t \right] \\
&\le 2 z_t G_t \left(z_t C_1 \eta_{x,t} v_t^x \right) + 2 z_t d_t^x \eta_{x,t}^2 v_t^x \left(4 z_t d_t^x \eta_{x,t}^2 \tau_{x,t}^2 \right) \\
&\le \frac{1}{3} z_t G_t + z_t d_t^x \eta_{x,t}^2 v_t^x, \qquad (30)\\
\mathbb{E}[\Delta_{2t+1}^2 \mid \mathcal{F}_{t+1/2}] &\le \frac{1}{3} z_t G_t + z_t d_t^y \eta_{y,t}^2 v_t^y. \qquad (31)
\end{aligned}$$

The second line uses $\mathbb{E}[(U - \mathbb{E}[U \mid \mathcal{F}_t])^2 \mid \mathcal{F}_t] \le \mathbb{E}[U^2 \mid \mathcal{F}_t] \le 4\tau_{x,t}^2 v_t^x$ for $U = \|\tilde{\xi}_t^x\|^2$; the last line uses $z_t C_1 \eta_{x,t} v_t^x \le 1/6$ from $z_t B_t^v \le 1$ and $4 z_t d_t^x \eta_{x,t}^2 \tau_{x,t}^2 \le 1/2$ from $z_t B_t^x \le 1$.

Since each Δ_s is a mean-zero martingale difference bounded above by 1, Lemma A.7 and Ville's inequality give, with probability at least $1-\delta$, simultaneously for all $n \geq 0$,

$$\sum_{s=0}^{n-1} \Delta_s \leq \frac{3}{4} \sum_{s=0}^{n-1} \mathbb{E}[\Delta_s^2 \mid \mathcal{H}_s] + \max\{\log(2/\delta), 1\}. \tag{32}$$

Indeed, the process $\exp\{\sum_{s<n} \Delta_s - (3/4)\sum_{s<n} \mathbb{E}[\Delta_s^2 \mid \mathcal{H}_s]\}$ is a nonnegative supermartingale, and $\exp(-\max\{\log(2/\delta), 1\}) \leq \delta$. Applying (32) with $n = 2k$ and using (30)–(31) gives, on the same event,

$$\begin{aligned} S_k &\leq \frac{3}{4} \sum_{t=0}^{k-1} \left(\mathbb{E}[\Delta_{2t}^2 \mid \mathcal{F}_t] + \mathbb{E}[\Delta_{2t+1}^2 \mid \mathcal{F}_{t+1/2}]\right) + \max\{\log(2/\delta), 1\} \\ &\leq \frac{1}{2} \sum_{t=0}^{k-1} z_t G_t + \sum_{t=0}^{k-1} z_t D_t + \max\{\log(2/\delta), 1\}, \end{aligned}$$

where $D_t = d_t^x \eta_{x,t}^2 v_t^x + d_t^y \eta_{y,t}^2 v_t^y$ and $D_t \geq 0$. Substituting this estimate for S_k into (29) yields:

$$\frac{1}{2} \sum_{t=0}^{k-1} z_t G_t + z_{k-1} P_k \leq z_0 P_0 + \sum_{t=0}^{k-1} z_t r_t + 2 \sum_{t=0}^{k-1} z_t D_t + \max\{\log(2/\delta), 1\}.$$

Multiplying by 2 and weakening the left side from $\sum_{t<k} z_t G_t + 2 z_{k-1} P_k$ to $\sum_{t<k} z_t G_t + z_{k-1} P_k$ proves the claimed bound. Since the Ville event holds for all n simultaneously, the conclusion holds for all $k \in \mathbb{N}$ simultaneously. ■

B.3.4. THEOREM B.2

We derive a master theorem for $\text{SGDA}_{\text{Clip}}$ in heavy-tailed NC-PL games, from which we deduce Theorem 4.11 and Theorem 4.12.

Theorem B.2 (Convergence of $\text{SGDA}_{\text{Clip}}$ in heavy-tailed NC-PL) *Suppose Assumptions 3.1, 3.2, 3.4 and 3.6 hold with exponent $p \in (1, 2]$. Assume $\eta_{y,t} = 128\kappa^2 \eta_{x,t}$ and $\eta_{x,t} \leq 1/(128\kappa^2 \ell)$ for all $t \geq 0$. Let*

$$\gamma_{y,t} \overset{def}{=} \frac{1}{8}\left(\eta_{y,t} - \frac{\ell}{2}\eta_{y,t}^2\right), \qquad \phi_t(y) \overset{def}{=} g(x_{t+1}) - f(x_{t+1}, y),$$

and define

$$\begin{aligned} G_t &\overset{def}{=} \frac{1}{8}\eta_{x,t} \left\|\nabla g(x_t)\right\|^2 + \gamma_{y,t} \left\|\nabla \phi_t(y_t)\right\|^2 + 4\kappa\ell\eta_{x,t} b_t, \\ \bar{G}_t &\overset{def}{=} \frac{1}{8}\eta_{x,t} \left\|\nabla g(x_t)\right\|^2 + 2\gamma_{y,t} \left\|\nabla_y f(x_t, y_t)\right\|^2 + 4\kappa\ell\eta_{x,t} b_t + 2\gamma_{y,t}\ell^2 \eta_{x,t}^2 \tau_{x,t}^2. \end{aligned}$$

Let

$$C_1 \overset{def}{=} \frac{81}{8}, \qquad d_t^x \overset{def}{=} 6\kappa\ell, \qquad d_t^y \overset{def}{=} \frac{\ell}{2}, \qquad v_t^x \overset{def}{=} 16\sigma^p \tau_{x,t}^{2-p}, \qquad v_t^y \overset{def}{=} 16\sigma^p \tau_{y,t}^{2-p}.$$

Define

$$B_t \overset{def}{=} \max\left\{1, \max_{i\in[0,t]} \max\{B_i^x, B_i^y, B_i^v\}\right\}, \qquad z_t \overset{def}{=} \frac{1}{B_t},$$
$$B_i^x \overset{def}{=} 2\tau_{x,i}\sqrt{C_1\eta_{x,i}\bar{G}_i} + 48\kappa\ell\eta_{x,i}^2\tau_{x,i}^2,$$
$$B_i^y \overset{def}{=} 2\tau_{y,i}\sqrt{C_1\eta_{y,i}\bar{G}_i} + 4\ell\eta_{y,i}^2\tau_{y,i}^2,$$
$$B_i^v \overset{def}{=} 96C_1\sigma^p(\eta_{x,i}\tau_{x,i}^{2-p} + \eta_{y,i}\tau_{y,i}^{2-p}),$$

$$S_{1,\tau}(k) \overset{def}{=} \sum_{t=0}^{k-1} z_t\left(6\eta_{x,t}\tau_{x,t}^{2-2p} + 2\eta_{y,t}\tau_{y,t}^{2-2p}\right), \quad S_{2,\tau}(k) \overset{def}{=} \sum_{t=0}^{k-1} z_t\left(6\kappa\ell\eta_{x,t}^2\tau_{x,t}^{2-p} + \frac{\ell}{2}\eta_{y,t}^2\tau_{y,t}^{2-p}\right).$$

Then, for each $\delta \in (0,1)$, there exists an event $\mathcal{E}_\delta$ with $\mathbb{P}(\mathcal{E}_\delta) \geq 1-\delta$ such that, on $\mathcal{E}_\delta$, the following implication holds for all integers $k \geq 1$ simultaneously:

$$\left[\|\nabla_x f(x_t,y_t)\| \leq \frac{\tau_{x,t}}{2}, \quad \|\nabla_y f(x_t,y_t)\| \leq \frac{\tau_{y,t}}{4}, \quad \ell\eta_{x,t}\tau_{x,t} \leq \frac{\tau_{y,t}}{4}, \quad \forall t \in \{0,1,\dots,k-1\}\right]$$
$$\Longrightarrow \left[z_{k-1}P_k + \sum_{t=0}^{k-1} z_t G_t \leq 2z_0P_0 + 32\sigma^{2p}S_{1,\tau}(k) + 64\sigma^p S_{2,\tau}(k) + 2\Gamma_\delta\right], \quad \Gamma_\delta \overset{def}{=} \max\{\log(2/\delta), 1\}$$

Proof We combine Lemmas 4.9 and 4.10.

Step 1: Predictable envelope and predictable stopping. Since $\nabla\phi_t(y_t) = -\nabla_y f(x_{t+1}, y_t)$ and the x-update is clipped,

$$\|x_{t+1} - x_t\| \leq \eta_{x,t}\tau_{x,t}.$$

By ℓ-smoothness,

$$\|\nabla_y f(x_{t+1},y_t) - \nabla_y f(x_t,y_t)\| \leq \ell\eta_{x,t}\tau_{x,t},$$

and hence

$$\|\nabla\phi_t(y_t)\|^2 = \|\nabla_y f(x_{t+1},y_t)\|^2 \leq 2\|\nabla_y f(x_t,y_t)\|^2 + 2\ell^2\eta_{x,t}^2\tau_{x,t}^2.$$

Thus $G_t \leq \bar{G}_t$ and $\bar{G}_t \in \mathcal{F}_t$. Define the predictable validity event

$$\mathcal{V}_t \overset{\text{def}}{=} \left\{\|\nabla_x f(x_t,y_t)\| \leq \frac{\tau_{x,t}}{2}, \quad \|\nabla_y f(x_t,y_t)\| \leq \frac{\tau_{y,t}}{4}, \quad \ell\eta_{x,t}\tau_{x,t} \leq \frac{\tau_{y,t}}{4}\right\},$$

and the stopping time $T_{\text{stop}} \overset{\text{def}}{=} \inf\{t \geq 0 : \mathcal{V}_t \text{ fails}\}$, with $T_{\text{stop}} = +\infty$ if the set is empty. Let $I_t \overset{\text{def}}{=} \mathbf{1}\{t < T_{\text{stop}}\}$. Then I_t is $\mathcal{F}_t$-measurable. On $\mathcal{V}_t$,

$$\|\nabla_y f(x_{t+1},y_t)\| \leq \|\nabla_y f(x_t,y_t)\| + \ell\|x_{t+1}-x_t\| \leq \frac{\tau_{y,t}}{4} + \ell\eta_{x,t}\tau_{x,t} \leq \frac{\tau_{y,t}}{2},$$

so Lemmas 4.8 and 4.9 apply whenever $I_t = 1$. Now define:

$$P_t' \overset{\text{def}}{=} P_{t\wedge T_{\text{stop}}}, \qquad G_t' \overset{\text{def}}{=} I_t G_t, \qquad \widetilde{\xi}_t^{x\prime} \overset{\text{def}}{=} I_t\widetilde{\xi}_t^x, \qquad \widetilde{\xi}_t^{y\prime} \overset{\text{def}}{=} I_t\widetilde{\xi}_t^y.$$

Since $I_t \in \mathcal{F}_t \subseteq \mathcal{F}_{t+1/2}$, indeed:

$$\mathbb{E}[\widetilde{\xi}_t^{x\prime} \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\widetilde{\xi}_t^{y\prime} \mid \mathcal{F}_{t+1/2}] = 0.$$

Step 2: Invoke Lemma 4.9. Whenever $I_t = 1$, Lemma 4.9 yields

$$G_t \le P_t - P_{t+1} + \eta_{x,t}\left\langle c_t^x, \widetilde{\xi}_t^x\right\rangle + d_t^x \eta_{x,t}^2 \|\widetilde{\xi}_t^x\|^2 + \eta_{y,t}\left\langle c_t^y, \widetilde{\xi}_t^y\right\rangle + d_t^y \eta_{y,t}^2 \|\widetilde{\xi}_t^y\|^2 + r_t, \tag{33}$$

where we set

$$c_t^x := -\nabla g(x_t) + \Big(\frac{1}{2} - 2\mu\gamma_{y,t}\Big)\delta_t, \qquad d_t^x := 6\kappa\ell,$$

$$c_t^y := \tfrac{1}{2}(1 - \ell\eta_{y,t})\,\nabla\phi_t(y_t), \qquad d_t^y := \frac{\ell}{2},$$

$$r_t := 6\eta_{x,t}\|\beta_t^x\|^2 + 2\eta_{y,t}\|\beta_t^y\|^2.$$

Set $c_t^{x\prime} \overset{\text{def}}{=} I_t c_t^x$, $c_t^{y\prime} \overset{\text{def}}{=} I_t c_t^y$, and $r_t' \overset{\text{def}}{=} I_t r_t$. If $I_t = 0$, then $G_t' = 0$, $P_t' = P_{t+1}'$, $\widetilde{\xi}_t^{x\prime} = \widetilde{\xi}_t^{y\prime} = 0$, and $r_t' = 0$, so (33) holds trivially with primes. Therefore, the stopped processes satisfy for all $t \ge 0$:

$$G_t' \le P_t' - P_{t+1}' + \eta_{x,t}\left\langle c_t^{x\prime}, \widetilde{\xi}_t^{x\prime}\right\rangle + d_t^x \eta_{x,t}^2 \|\widetilde{\xi}_t^{x\prime}\|^2 + \eta_{y,t}\left\langle c_t^{y\prime}, \widetilde{\xi}_t^{y\prime}\right\rangle + d_t^y \eta_{y,t}^2 \|\widetilde{\xi}_t^{y\prime}\|^2 + r_t'. \tag{34}$$

Step 3: Verify the clipped-noise bounds. Lemma 4.8 always gives $\|\widetilde{\xi}_t^x\| \le 2\tau_{x,t}$ and $\|\widetilde{\xi}_t^y\| \le 2\tau_{y,t}$. Hence, we have:

$$\|\widetilde{\xi}_t^{x\prime}\| \le 2\tau_{x,t}, \qquad \|\widetilde{\xi}_t^{y\prime}\| \le 2\tau_{y,t} \qquad \text{a.s. for all } t.$$

Moreover, if $I_t = 1$ then the predictable validity condition holds at t, so Lemma 4.8 implies

$$\mathbb{E}\big[\|\widetilde{\xi}_t^x\|^2 \mid \mathcal{F}_t\big] \le 16\sigma^p \tau_{x,t}^{2-p}, \qquad \mathbb{E}\big[\|\widetilde{\xi}_t^y\|^2 \mid \mathcal{F}_{t+\frac{1}{2}}\big] \le 16\sigma^p \tau_{y,t}^{2-p},$$

and therefore for all $t \ge 0$,

$$\mathbb{E}\big[\|\widetilde{\xi}_t^{x\prime}\|^2 \mid \mathcal{F}_t\big] \le v_t^x, \qquad \mathbb{E}\big[\|\widetilde{\xi}_t^{y\prime}\|^2 \mid \mathcal{F}_{t+\frac{1}{2}}\big] \le v_t^y,$$

with the deterministic choices $v_t^x := 16\sigma^p \tau_{x,t}^{2-p}$, $v_t^y := 16\sigma^p \tau_{y,t}^{2-p}$.

Step 4: Verify the self-bounding condition. We show that for all $t \ge 0$,

$$\eta_{x,t}\|c_t^{x\prime}\|^2 + \eta_{y,t}\|c_t^{y\prime}\|^2 \le C_1 G_t' \quad \text{with} \quad C_1 = \frac{81}{8}. \tag{35}$$

For $t \ge T_{\text{stop}}$ both sides are zero by definition, so it suffices to consider $t < T_{\text{stop}}$.

(x-part). Let $\theta_t \overset{\text{def}}{=} \frac{1}{2} - 2\mu\gamma_{y,t} \in [0, 1/2]$. For any $\beta > 0$, $\|u+v\|^2 \le (1+\beta)\|u\|^2 + (1+\beta^{-1})\|v\|^2$. With $u = -\nabla g(x_t)$ and $v = \theta_t\delta_t$,

$$\|c_t^x\|^2 \le (1+\beta)\|\nabla g(x_t)\|^2 + \Big(1+\beta^{-1}\Big)\frac{1}{4}\|\delta_t\|^2.$$

By Lemma A.3, $\|\delta_t\|^2 = \|\nabla_x f(x_t, y_t) - \nabla g(x_t)\|^2 \le 2\kappa\ell\, b_t$. Thus

$$\eta_{x,t}\|c_t^x\|^2 \le (1+\beta)\eta_{x,t}\|\nabla g(x_t)\|^2 + \Big(1+\beta^{-1}\Big)\frac{\kappa\ell}{2}\,\eta_{x,t} b_t.$$

Comparing with the definition of G_t, we can bound the RHS by $C_x G_t$ whenever

$$(1+\beta)\eta_{x,t}\|\nabla g(x_t)\|^2 \le C_x \cdot \frac{1}{8}\eta_{x,t}\|\nabla g(x_t)\|^2, \qquad \Big(1+\beta^{-1}\Big)\frac{\kappa\ell}{2}\eta_{x,t} b_t \le C_x \cdot 4\kappa\ell\,\eta_{x,t} b_t,$$

i.e. whenever $C_x \geq 8(1+\beta)$ and $C_x \geq \frac{1+\beta^{-1}}{8}$. Choosing $\beta = \frac{1}{64}$ gives $8(1+\beta) = \frac{65}{8}$ and $\frac{1+\beta^{-1}}{8} = \frac{65}{8}$, hence

$$\eta_{x,t}\|c_t^x\|^2 \leq \frac{65}{8}\, G_t. \tag{36}$$

(y-part). We have

$$\eta_{y,t}\|c_t^y\|^2 = \eta_{y,t} \cdot \frac{1}{4}(1-\ell\eta_{y,t})^2\, \|\nabla\phi_t(y_t)\|^2.$$

Let $u := \ell\eta_{y,t} \in [0,1]$ (since $\eta_{y,t} \leq \frac{1}{\ell}$). Then

$$\gamma_{y,t} = \frac{\eta_{y,t}}{16}(2-u) \quad \text{and} \quad \frac{1}{4}(1-u)^2 \leq \frac{2-u}{8} \qquad \forall u \in [0,1]$$

(the latter is equivalent to $u(2u-3) \leq 0$). Therefore

$$\eta_{y,t}\|c_t^y\|^2 \leq 2\gamma_{y,t}\|\nabla\phi_t(y_t)\|^2 \leq 2G_t. \tag{37}$$

Combining (36)–(37) yields $\eta_{x,t}\|c_t^x\|^2 + \eta_{y,t}\|c_t^y\|^2 \leq \frac{81}{8}G_t$, which gives (35).

Step 5: Apply Lemma 4.10 to the stopped process. We now invoke Lemma 4.10 with:

$$(\widetilde{\xi}_t^{x\prime}, \widetilde{\xi}_t^{y\prime}),\ (v_t^x, v_t^y),\ (\tau_{x,t}, \tau_{y,t}),\ (G_t', P_t'),\ (c_t^{x\prime}, c_t^{y\prime}),\ (d_t^x, d_t^y),\ r_t',$$

with the predictable envelope $\bar{G}_t$ and with the self-bounding constant $C_1 = \frac{81}{8}$ verified above. Lemma 4.10 yields an event $\mathcal{E}_\delta$ with $\Pr(\mathcal{E}_\delta) \geq 1-\delta$ such that on $\mathcal{E}_\delta$, for all integers $k \geq 1$ simultaneously,

$$\sum_{t=0}^{k-1} z_t G_t' + z_{k-1}P_k' \leq 2z_0P_0' + 2\sum_{t=0}^{k-1}\big(z_t r_t' + 2z_t D_t\big) + 2\Gamma_\delta, \tag{38}$$

where z_t is the theorem's weight sequence and $\Gamma_\delta = \max\{\log(2/\delta), 1\}$.

Step 6: Specialize to times $k \leq T_{\text{stop}}$ and bound r_t, D_t. Fix any integer $k \geq 1$ and assume the premise of the theorem for this k, i.e. the predictable validity event $\mathcal{V}_t$ holds for all $t \in \{0,\dots,k-1\}$. Equivalently, $T_{\text{stop}} \geq k$, and therefore, for all $t \in \{0,\dots,k-1\}$, we have:

$$G_t' = G_t, \quad P_k' = P_k, \quad P_0' = P_0, \quad r_t' = r_t.$$

Hence, on $\mathcal{E}_\delta$, inequality (38) becomes

$$\sum_{t=0}^{k-1} z_t G_t + z_{k-1}P_k \leq 2z_0P_0 + 2\sum_{t=0}^{k-1}\big(z_t r_t + 2z_t D_t\big) + 2\Gamma_\delta. \tag{39}$$

It remains to upper bound r_t and D_t via Lemma 4.8 (valid for $t < k$ under the premise). Indeed, for such t, Lemma 4.8 gives

$$\|\beta_t^x\| \leq 4\sigma^p\tau_{x,t}^{1-p}, \qquad \|\beta_t^y\| \leq 4\sigma^p\tau_{y,t}^{1-p}.$$

Therefore

$$r_t = 6\eta_{x,t}\|\beta_t^x\|^2 + 2\eta_{y,t}\|\beta_t^y\|^2 \leq 6\eta_{x,t}\cdot 16\sigma^{2p}\tau_{x,t}^{2-2p} + 2\eta_{y,t}\cdot 16\sigma^{2p}\tau_{y,t}^{2-2p},$$

so we have:

$$2r_t \leq 32\sigma^{2p}\Big(6\tau_{x,t}^{2-2p}\eta_{x,t} + 2\tau_{y,t}^{2-2p}\eta_{y,t}\Big). \tag{40}$$

Also by Lemma 4.8, we have:

$$v_t^x = 16\sigma^p\tau_{x,t}^{2-p}, \qquad v_t^y = 16\sigma^p\tau_{y,t}^{2-p},$$

and with $d_t^x = 6\kappa\ell$, $d_t^y = \ell/2$,

$$D_t = d_t^x\eta_{x,t}^2 v_t^x + d_t^y\eta_{y,t}^2 v_t^y = 6\kappa\ell\,\eta_{x,t}^2\cdot 16\sigma^p\tau_{x,t}^{2-p} + \frac{\ell}{2}\eta_{y,t}^2\cdot 16\sigma^p\tau_{y,t}^{2-p}$$

holds, and we have:

$$4D_t = 64\sigma^p\Big(6\kappa\ell\,\tau_{x,t}^{2-p}\eta_{x,t}^2 + \frac{\ell}{2}\tau_{y,t}^{2-p}\eta_{y,t}^2\Big). \tag{41}$$

Plugging (40)–(41) into (39) yields:

$$\begin{aligned} z_{k-1}P_k + \sum_{t=0}^{k-1} z_t G_t \leq 2z_0P_0 &+ 32\sigma^{2p}\sum_{t=0}^{k-1} z_t\Big(6\tau_{x,t}^{2-2p}\eta_{x,t} + 2\tau_{y,t}^{2-2p}\eta_{y,t}\Big) \\ &+64\sigma^p\sum_{t=0}^{k-1} z_t\Big(6\kappa\ell\,\tau_{x,t}^{2-p}\eta_{x,t}^2 + \frac{\ell}{2}\tau_{y,t}^{2-p}\eta_{y,t}^2\Big) + 2\Gamma_\delta. \end{aligned}$$

Recalling the definition of G_t and the theorem's notation

$$S_{1,\tau}(k) \coloneqq \sum_{t=0}^{k-1} z_t\Big(6\tau_{x,t}^{2-2p}\eta_{x,t} + 2\tau_{y,t}^{2-2p}\eta_{y,t}\Big), \qquad S_{2,\tau}(k) \coloneqq \sum_{t=0}^{k-1} z_t\Big(6\kappa\ell\,\tau_{x,t}^{2-p}\eta_{x,t}^2 + \frac{\ell}{2}\tau_{y,t}^{2-p}\eta_{y,t}^2\Big),$$

we obtain exactly the claimed inequality. Since $\Pr(\mathcal{E}_\delta) \geq 1-\delta$ and k was arbitrary, this completes the proof. ■

B.3.5. THEOREM 4.11

Proof Let

$$\Delta_\delta \overset{\text{def}}{=} P_0 + \Gamma_\delta, \qquad \alpha \overset{\text{def}}{=} \frac{p}{3p-2}, \qquad r \overset{\text{def}}{=} 1-\alpha = \frac{2(p-1)}{3p-2},$$

$$a \overset{\text{def}}{=} \frac{p-1}{3p-2} = \frac{r}{2}, \qquad R_\kappa \overset{\text{def}}{=} 128\kappa^2, \qquad H_{\kappa,\ell,p} \overset{\text{def}}{=} \sigma^2 B_{\kappa,\ell}^{(2-p)/p} A_{\kappa,\ell}^{2(p-1)/p}.$$

Set

$$\bar\eta \overset{\text{def}}{=} c_\eta\left(\frac{\Delta_\delta}{H_{\kappa,\ell,p}}\right)^\alpha, \qquad \bar\tau \overset{\text{def}}{=} \left(\frac{\sigma^p B_{\kappa,\ell}}{2A_{\kappa,\ell}\bar\eta}\right)^{1/p}.$$

Let $c_\tau, c_B > 0$ be sufficiently large universal constants, and define

$$T_0 \overset{\text{def}}{=} \left\lceil 1 \vee (R_\kappa\ell\,\bar\eta)^{1/\alpha} \vee \left(\frac{c_\tau\sqrt{\kappa\ell\Delta_\delta}}{\bar\tau}\right)^{3p-2} \vee \left(c_B\sqrt{R_\kappa}\,\bar\eta\bar\tau\sqrt{\kappa^2\ell\Delta_\delta}\right)^{1/a}$$

$$\vee \left(c_B R_\kappa^2 \ell\, \bar\eta^2 \bar\tau^2\right)^{1/(2a)} \vee \left(c_B (1+R_\kappa)\sigma^p \bar\eta \bar\tau^{2-p}\right)^{1/r}\Big\rceil .$$

Then the theorem's schedule is equivalently $\eta_x = \bar\eta T^{-\alpha}, \tau = \bar\tau T^{1/(3p-2)}$. Let $M \overset{\text{def}}{=} C_M \Delta_\delta$, where $C_M > 0$ is a sufficiently large universal constant. We first show that, under the balanced schedule and $T \geq T_0$, the assumptions needed to apply Theorem B.2 are automatically satisfied.

The smoothness-cap term in T_0 gives

$$\eta_x = \bar\eta T^{-\alpha} \leq \frac{1}{R_\kappa \ell} = \frac{1}{128\kappa^2 \ell}.$$

The clipping-bootstrap term in T_0 gives

$$\tau^2 = \bar\tau^2 T^{2/(3p-2)} \geq c_\tau^2 \kappa \ell \Delta_\delta.$$

Choosing c_τ sufficiently large, this implies

$$\tau^2 \geq 64\kappa\ell M.$$

Assume for a moment that $P_t \leq M$. Then $a_t \leq P_t$ and $b_t \leq 2P_t$. By Lemmas A.4 and A.3,

$$\|\nabla g(x_t)\|^2 \leq 4\kappa\ell M, \qquad \|\delta_t\|^2 \leq 4\kappa\ell M,$$

where $\delta_t = \nabla_x f(x_t, y_t) - \nabla g(x_t)$. Hence

$$\|\nabla_x f(x_t, y_t)\| \leq \|\nabla g(x_t)\| + \|\delta_t\| \leq 4\sqrt{\kappa\ell M} \leq \tau/2.$$

Also, since $y \mapsto g(x_t) - f(x_t, y)$ is nonnegative and ℓ-smooth,

$$\|\nabla_y f(x_t, y_t)\|^2 \leq 2\ell b_t \leq 4\ell M,$$

and hence

$$\|\nabla_y f(x_t, y_t)\| \leq 2\sqrt{\ell M} \leq \tau/4.$$

Finally,

$$\ell \eta_x \tau \leq \tau/4$$

because $\eta_x \leq 1/(128\kappa^2\ell)$ and $\kappa \geq 1$. Therefore the predictable premise of Theorem B.2 holds whenever $P_t \leq M$. Next, under $P_t \leq M$, the predictable envelope in Theorem B.2 satisfies

$$\bar G_t \leq C_G \kappa^2 \ell\, \eta_x M + C_G \kappa^2 \ell^2 \eta_x^3 \tau^2$$

for a universal constant C_G. Therefore the three terms defining B_t are controlled by

$$C\sqrt{R_\kappa}\, \eta_x \tau \sqrt{\kappa^2 \ell M}, \quad C R_\kappa^2 \ell\, \eta_x^2 \tau^2, \quad C(1+R_\kappa)\sigma^p \eta_x \tau^{2-p}.$$

By the last three terms in T_0, these quantities are all at most 1 after increasing c_B if necessary. Consequently, $B_t \leq 2$ and $z_t \geq 1/2$ on any interval on which $P_t \leq M$. Now define

$$K \overset{\text{def}}{=} \inf\{k \in \{1, \dots, T\} : P_k > M\},$$

with $K \overset{\text{def}}{=} T+1$ if the set is empty. Work on the event $\mathcal{E}_\delta$ from Theorem B.2. Suppose for contradiction that $K \le T$. Then $P_t \le M$ for all $t < K$, so the premise of Theorem B.2 holds up to time K, and $z_t \ge 1/2$ for $t < K$. Applying Theorem B.2 with $k = K$ and dropping the nonnegative progress term gives

$$z_{K-1} P_K \le 2P_0 + 2\Gamma_\delta + C\sigma^{2p} B_{\kappa,\ell} T \eta_x \tau^{2-2p} + C\sigma^p A_{\kappa,\ell} T \eta_x^2 \tau^{2-p}.$$

The balanced identity $\tau^p = \frac{\sigma^p B_{\kappa,\ell}}{2A_{\kappa,\ell}\eta_x}$ gives

$$\sigma^{2p} B_{\kappa,\ell} T \eta_x \tau^{2-2p} + \sigma^p A_{\kappa,\ell} T \eta_x^2 \tau^{2-p} \le C H_{\kappa,\ell,p} T \eta_x^{(3p-2)/p}.$$

Since

$$\eta_x = c_\eta \left(\frac{\Delta_\delta}{H_{\kappa,\ell,p} T} \right)^{p/(3p-2)},$$

the right-hand side is at most $C\Delta_\delta$. Thus, after choosing C_M sufficiently large,

$$z_{K-1} P_K \le M/2.$$

Since $z_{K-1} \ge 1/2$, this implies $P_K \le M$, a contradiction. Hence $K = T+1$. Therefore $P_t \le M$, Theorem B.2's premise holds, and $z_t \ge 1/2$ for all $t < T$. Apply Theorem B.2 with $k = T$. Since

$$G_t \ge \frac{1}{8} \eta_x \left\| \nabla g(x_t) \right\|^2,$$

we get

$$\frac{1}{8} \sum_{t=0}^{T-1} z_t \eta_x \left\| \nabla g(x_t) \right\|^2 \le C\Delta_\delta$$

using the balanced identity once more. Since $\sum_{t=0}^{T-1} z_t \eta_x \ge T\eta_x/2$, we obtain

$$\min_{0 \le t < T} \left\| \nabla g(x_t) \right\|^2 \le C \frac{\Delta_\delta}{T\eta_x}.$$

Substituting $\eta_x = c_\eta \left(\frac{\Delta_\delta}{H_{\kappa,\ell,p} T} \right)^\alpha$ gives us

$$\frac{\Delta_\delta}{T\eta_x} = C H_{\kappa,\ell,p}^{p/(3p-2)} \left(\frac{\Delta_\delta}{T} \right)^r = C\sigma^{\frac{2p}{3p-2}} B_{\kappa,\ell}^{\frac{2-p}{3p-2}} A_{\kappa,\ell}^{\frac{2(p-1)}{3p-2}} \left(\frac{\Delta_\delta}{T} \right)^r.$$

■

B.3.6. THEOREM 4.12

Proof Let $\Delta_\delta \overset{\text{def}}{=} P_0 + \Gamma_\delta$, $\alpha \overset{\text{def}}{=} p/(3p-2)$, $r \overset{\text{def}}{=} 2(p-1)/(3p-2)$, $a \overset{\text{def}}{=} (p-1)/(3p-2)$, $R_\kappa \overset{\text{def}}{=} 128\kappa^2$, and $H_{\kappa,\ell,p} \overset{\text{def}}{=} \sigma^2 B_{\kappa,\ell}^{(2-p)/p} A_{\kappa,\ell}^{2(p-1)/p}$. Set

$$\bar{\eta} \overset{\text{def}}{=} c_\eta \left(\frac{\Delta_\delta}{H_{\kappa,\ell,p}} \right)^\alpha, \qquad \bar{\tau} \overset{\text{def}}{=} \left(\frac{\sigma^p B_{\kappa,\ell}}{2A_{\kappa,\ell}\bar{\eta}} \right)^{1/p}.$$

Let $c_\tau, c_B > 0$ be sufficiently large universal constants, and define

$$s_0 \stackrel{\text{def}}{=} \left\lceil 1 \vee (R_\kappa \ell\, \bar\eta)^{1/\alpha} \vee \left(\frac{c_\tau \sqrt{\kappa\ell\Delta_\delta}}{\bar\tau}\right)^{3p-2} \vee \left(c_B \sqrt{R_\kappa}\, \bar\eta\bar\tau \sqrt{\kappa^2\ell\Delta_\delta}\right)^{1/a} \right.$$
$$\left. \vee \left(c_B R_\kappa^2 \ell\, \bar\eta^2\bar\tau^2\right)^{1/(2a)} \vee \left(c_B(1+R_\kappa)\sigma^p \bar\eta\bar\tau^{2-p}\right)^{1/r} \right\rceil.$$

The theorem's schedule is equivalently $\eta_{x,t} = \bar\eta(t+s_0)^{-\alpha}$ and $\tau_t = \bar\tau(t+s_0)^{1/(3p-2)}$. Fix an arbitrary integer $T \geq \max\{2, s_0\}$. Let $M_k \stackrel{\text{def}}{=} C_M\Delta_\delta(1 + \sum_{t=0}^{k-1}(t+s_0)^{-1})$ for $k \geq 1$, and set $M_0 \stackrel{\text{def}}{=} C_M\Delta_\delta$, where $C_M > 0$ is a sufficiently large universal constant. Under the shifted schedule,

$$\eta_{x,t}\tau_t^{2-2p} = \bar\eta\bar\tau^{2-2p}(t+s_0)^{-1}, \quad \eta_{x,t}^2\tau_t^{2-p} = \bar\eta^2\bar\tau^{2-p}(t+s_0)^{-1}.$$

The balanced identity $\tau_t^p = \sigma^p B_{\kappa,\ell}/(2A_{\kappa,\ell}\eta_{x,t})$ implies

$$\sigma^{2p} B_{\kappa,\ell}\eta_{x,t}\tau_t^{2-2p} + \sigma^p A_{\kappa,\ell}\eta_{x,t}^2\tau_t^{2-p} \leq C\Delta_\delta(t+s_0)^{-1}$$

after substituting the definitions of $\bar\eta$ and $\bar\tau$. Hence M_k dominates the deterministic upper envelope generated by the right-hand side of Theorem B.2.

We first verify the validity and normalization conditions. The first term in s_0 gives $\eta_{x,t} \leq \eta_{x,0} = \bar\eta s_0^{-\alpha} \leq (128\kappa^2\ell)^{-1}$. The second term in s_0 gives $\tau_0^2 \geq c_\tau^2\kappa\ell\Delta_\delta$. Moreover, for every $t \geq 0$,

$$\tau_t^2 = \bar\tau^2(t+s_0)^{2/(3p-2)} \geq c\,\bar\tau^2 s_0^{2/(3p-2)}\left(1 + \log\left(1 + \frac{t}{s_0}\right)\right).$$

The elementary inequality used here is that, for $\beta \in [1/2, 2]$, $(1+u)^\beta \geq c_\beta(1+\log(1+u))$ for all $u \geq 0$, and the constants can be chosen uniformly for $p \in (1, 2]$. Thus, for c_τ sufficiently large, $\tau_t^2 \geq 64\kappa\ell M_{t+1}$ for all $t \geq 0$. If $P_t \leq M_{t+1}$, then $a_t \leq M_{t+1}$ and $b_t \leq 2M_{t+1}$. Lemmas A.4 and A.3 give

$$\|\nabla_x f(x_t, y_t)\| \leq 4\sqrt{\kappa\ell M_{t+1}} \leq \tau_t/2, \quad \|\nabla_y f(x_t, y_t)\| \leq 2\sqrt{\ell M_{t+1}} \leq \tau_t/4.$$

Also $\ell\eta_{x,t}\tau_t \leq \tau_t/4$ because $\eta_{x,t} \leq 1/(128\kappa^2\ell)$ and $\kappa \geq 1$. Therefore the predictable premise of Theorem B.2 holds at time t whenever $P_t \leq M_{t+1}$. Under $P_t \leq M_{t+1}$, the predictable envelope satisfies

$$\bar G_t \leq C_G\kappa^2\ell\,\eta_{x,t}M_{t+1} + C_G\kappa^2\ell^2\eta_{x,t}^3\tau_t^2.$$

Consequently, the terms controlling B_t are bounded by

$$C\sqrt{R_\kappa}\,\eta_{x,t}\tau_t\sqrt{\kappa^2\ell M_{t+1}}, \quad CR_\kappa^2\ell\,\eta_{x,t}^2\tau_t^2, \quad C(1+R_\kappa)\sigma^p\eta_{x,t}\tau_t^{2-p}.$$

Using the shifted schedule,

$$\eta_{x,t}\tau_t = \bar\eta\bar\tau(t+s_0)^{-a}, \quad \eta_{x,t}^2\tau_t^2 = \bar\eta^2\bar\tau^2(t+s_0)^{-2a}, \quad \eta_{x,t}\tau_t^{2-p} = \bar\eta\bar\tau^{2-p}(t+s_0)^{-r}.$$

The last three terms in s_0, together with the same polynomial-dominates-log inequality, ensure these quantities are all at most a sufficiently small universal constant. Thus $B_t \leq 2$ and $z_t \geq 1/2$ whenever $P_i \leq M_{i+1}$ for all $i \leq t$.

Now bootstrap the potential. Let $K \stackrel{\text{def}}{=} \inf\{k \in \{1,\dots,T\} : P_k > M_k\}$, with $K = T+1$ if the set is empty. Work on the event $\mathcal{E}_\delta$ from Theorem B.2. Suppose $K \le T$. Then, for every $t < K$, either $t = 0$, in which case $P_0 \le \Delta_\delta \le M_1$, or $P_t \le M_t \le M_{t+1}$. Therefore Theorem B.2's premise holds for all $t < K$, and $z_t \ge 1/2$ for $t < K$.

Applying Theorem B.2 with $k = K$, dropping the nonnegative progress term, and using $z_t \le 1$,

$$z_{K-1} P_K \le 2P_0 + 2\Gamma_\delta + C\sigma^{2p} B_{\kappa,\ell} \sum_{t=0}^{K-1} \eta_{x,t} \tau_t^{2-2p} + C\sigma^p A_{\kappa,\ell} \sum_{t=0}^{K-1} \eta_{x,t}^2 \tau_t^{2-p}.$$

By the balanced identity and the definition of M_K, the right-hand side is at most $M_K/2$ for C_M sufficiently large. Since $z_{K-1} \ge 1/2$, this implies $P_K \le M_K$, contradicting the definition of K. Hence $K = T+1$. Thus $P_t \le M_t$, the premise of Theorem B.2 holds, and $z_t \ge 1/2$ for all $t < T$.

Apply Theorem B.2 with $k = T$. Since $G_t \ge \frac{1}{8}\eta_{x,t} \|\nabla g(x_t)\|^2$, we obtain

$$\frac{1}{8} \sum_{t=0}^{T-1} z_t \eta_{x,t} \|\nabla g(x_t)\|^2 \le C\Delta_\delta \left(1 + \sum_{t=0}^{T-1} \frac{1}{t+s_0}\right).$$

Since $z_t \ge 1/2$,

$$\sum_{t=0}^{T-1} z_t \eta_{x,t} \ge \frac{1}{2}\bar{\eta} \sum_{t=0}^{T-1} (t+s_0)^{-\alpha} \ge c\bar{\eta} T^{1-\alpha} = c\bar{\eta} T^r.$$

The second inequality uses $T \ge s_0$. Also, $\sum_{t=0}^{T-1} (t+s_0)^{-1} \le 1 + \log T$. Therefore

$$\min_{0 \le t < T} \|\nabla g(x_t)\|^2 \le C \frac{\Delta_\delta (1 + \log T)}{\bar{\eta} T^r}.$$

Substituting $\bar{\eta} = c_\eta (\Delta_\delta / H_{\kappa,\ell,p})^\alpha$ gives

$$\frac{\Delta_\delta}{\bar{\eta}} = C H_{\kappa,\ell,p}^{p/(3p-2)} \Delta_\delta^r, \quad H_{\kappa,\ell,p}^{p/(3p-2)} = \sigma^{\frac{2p}{3p-2}} B_{\kappa,\ell}^{\frac{2-p}{3p-2}} A_{\kappa,\ell}^{\frac{2(p-1)}{3p-2}}.$$

This proves the claimed logarithmic unknown-horizon rate for the fixed horizon T. Since $T \ge \max\{2, s_0\}$ was arbitrary and the event $\mathcal{E}_\delta$ does not depend on T, the claim follows. ■

Appendix C. Proofs of Nonconvex–Concave Games

C.1. SGDA in Subgaussian NC-C Games

C.1.1. LEMMA 5.1

Proof Let

$$u(x) := \operatorname{prox}_{\lambda f}(x) = \arg\min_u \left\{ f(u) + \frac{1}{2\lambda} \|u - x\|^2 \right\}.$$

Since $f(u) + \frac{\rho}{2}\|u\|^2$ is convex and $\lambda < 1/\rho$, the objective defining $u(x)$ is $(1/\lambda - \rho)$-strongly convex; hence $u(x)$ exists and is unique. The optimality condition gives

$$p(x) := \frac{x - u(x)}{\lambda} \in \partial f(u(x)).$$

For x, x' write $u = u(x)$, $u' = u(x')$, $p = p(x)$, and $p' = p(x')$. Weak convexity is equivalent to hypomonotonicity of the subdifferential:

$$\langle p - p', u - u' \rangle \geq -\rho \left\| u - u' \right\|^2 .$$

Using $u - u' = x - x' - \lambda(p - p')$, the last inequality implies

$$\langle p - p', x - x' \rangle \geq \lambda \left\| p - p' \right\|^2 - \rho \left\| x - x' - \lambda(p - p') \right\|^2 \geq -\frac{\rho}{1 - \rho\lambda} \left\| x - x' \right\|^2 ,$$

where the final inequality follows by minimizing the preceding quadratic lower bound over $p - p'$. Also, the same optimality relation gives the standard Lipschitz bound

$$\left\| u(x) - u(x') \right\| \leq \frac{1}{1 - \rho\lambda} \left\| x - x' \right\| .$$

The usual envelope argument, using $u(x)$ and $u(x')$ as comparison points in the two proximal problems and then sending $x' \to x$, yields continuous differentiability and

$$\nabla f_\lambda(x) = p(x) = \frac{1}{\lambda} \left(x - \operatorname{prox}_{\lambda f}(x) \right) .$$

Using $u(x)$ as a feasible point in the definition of $f_\lambda(x')$ gives

$$f_\lambda(x') \leq f_\lambda(x) + \langle \nabla f_\lambda(x), x' - x \rangle + \frac{1}{2\lambda} \left\| x' - x \right\|^2 .$$

Finally, the hypomonotonicity bound for $\nabla f_\lambda = p$ and the fundamental theorem of calculus along the segment $x + \theta(x' - x)$ give

$$f_\lambda(x') \geq f_\lambda(x) + \langle \nabla f_\lambda(x), x' - x \rangle - \frac{\rho}{2(1 - \rho\lambda)} \left\| x' - x \right\|^2 .$$

These two quadratic inequalities imply

$$\left\| \nabla f_\lambda(x') - \nabla f_\lambda(x) \right\| \leq \max \left\{ \frac{1}{\lambda}, \frac{\rho}{1 - \rho\lambda} \right\} \left\| x' - x \right\| ,$$

so if $\lambda \leq 1/(2\rho)$, then $\rho/(1 - \rho\lambda) \leq 1/\lambda$, and f_λ is $1/\lambda$-smooth. ■

C.1.2. LEMMA 5.2

Proof Fix $y \in \mathcal{Y}$ and set $h_y(x) := f(x, y)$. By Assumption 3.1, h_y is ℓ-smooth in x, so

$$h_y(x') \geq h_y(x) + \langle \nabla h_y(x), x' - x \rangle - \frac{\ell}{2} \left\| x' - x \right\|^2 .$$

Adding $\frac{\ell}{2} \left\| x' \right\|^2$ and using $\left\| x' \right\|^2 = \left\| x \right\|^2 + 2 \langle x, x' - x \rangle + \left\| x' - x \right\|^2$ gives

$$h_y(x') + \frac{\ell}{2} \left\| x' \right\|^2 \geq h_y(x) + \frac{\ell}{2} \left\| x \right\|^2 + \langle \nabla h_y(x) + \ell x, x' - x \rangle .$$

Thus $x \mapsto f(x,y) + \frac{\ell}{2}\|x\|^2$ is convex for each y. Taking the pointwise maximum over $y \in \mathcal{Y}$ preserves convexity, so

$$g(x) + \frac{\ell}{2}\|x\|^2 = \max_{y\in\mathcal{Y}} \left\{ f(x,y) + \frac{\ell}{2}\|x\|^2 \right\}$$

is convex, and g is ℓ-weakly convex.

Apply Lemma 5.1 to g with $\rho = \ell$ and $\lambda = 1/(2\ell)$. Then $\Phi = g_{1/(2\ell)}$ is continuously differentiable,

$$\nabla\Phi(x) = 2\ell\left(x - \mathrm{prox}_{g/(2\ell)}(x)\right),$$

and, since $\lambda = 1/(2\ell) \le 1/(2\rho)$, Φ is 2ℓ-smooth. Finally, $\Phi(x) \le g(x)$ by choosing $u = x$ in the envelope definition, while

$$\Phi(x) = \inf_u \left\{ g(u) + \ell\|u - x\|^2 \right\} \ge \inf_u g(u).$$

Taking infima in x gives $\inf_x \Phi(x) = \inf_x g(x)$. ■

C.1.3. LEMMA 5.3

Proof Recall $\Phi(\cdot) = g_{1/(2\ell)}(\cdot)$ and define

$$u_t \in \mathrm{prox}_{g/(2\ell)}(x_t) := \arg\min_{u\in\mathbb{R}^{d_x}} \left\{ g(u) + \ell\|u - x_t\|^2 \right\}.$$

By Lemma 5.2, Φ is 2ℓ-smooth and

$$\nabla\Phi(x_t) = 2\ell(x_t - u_t), \qquad \Phi(x_t) = g(u_t) + \ell\|u_t - x_t\|^2. \tag{42}$$

Moreover, the SGDA x-update can be written as

$$x_{t+1} = x_t - \eta_{x,t}\big(\nabla_x f(x_t, y_t) + \xi_t^x\big).$$

Step 1: Smoothness descent for Φ. By 2ℓ-smoothness of Φ,

$$\Phi(x_{t+1}) \le \Phi(x_t) + \langle \nabla\Phi(x_t), x_{t+1} - x_t\rangle + \ell\|x_{t+1} - x_t\|^2.$$

Substituting $x_{t+1} - x_t = -\eta_{x,t}(\nabla_x f(x_t, y_t) + \xi_t^x)$ gives

$$\begin{aligned}\Phi(x_{t+1}) &\le \Phi(x_t) - \eta_{x,t}\langle \nabla\Phi(x_t), \nabla_x f(x_t, y_t) + \xi_t^x\rangle + \ell\eta_{x,t}^2\|\nabla_x f(x_t, y_t) + \xi_t^x\|^2 \\ &= \Phi(x_t) - \eta_{x,t}\langle \nabla\Phi(x_t), \nabla_x f(x_t, y_t)\rangle - \eta_{x,t}\langle \nabla\Phi(x_t), \xi_t^x\rangle + \ell\eta_{x,t}^2\|\nabla_x f(x_t, y_t) + \xi_t^x\|^2.\end{aligned} \tag{43}$$

Subtracting $\Phi^\star$ from both sides yields the same inequality for $a_{t+1} = \Phi(x_{t+1}) - \Phi^\star$.

Step 2: Lower bound $\langle \nabla\Phi(x_t), \nabla_x f(x_t, y_t)\rangle$. Fix t and consider the function $x \mapsto f(x, y_t)$. By Assumption 3.1, it is ℓ-smooth, hence for any u,

$$f(u, y_t) \geq f(x_t, y_t) + \langle \nabla_x f(x_t, y_t), u - x_t\rangle - \frac{\ell}{2}\|u - x_t\|^2.$$

Applying this with $u = u_t$ and rearranging gives

$$\langle \nabla_x f(x_t, y_t), x_t - u_t\rangle \geq f(x_t, y_t) - f(u_t, y_t) - \frac{\ell}{2}\|x_t - u_t\|^2.$$

Multiplying by 2ℓ and using $\nabla\Phi(x_t) = 2\ell(x_t - u_t)$,

$$\langle \nabla\Phi(x_t), \nabla_x f(x_t, y_t)\rangle \geq 2\ell\big(f(x_t, y_t) - f(u_t, y_t)\big) - \ell^2\|x_t - u_t\|^2. \tag{44}$$

Since $\ell^2\|x_t - u_t\|^2 = \frac{1}{4}\|\nabla\Phi(x_t)\|^2$ by (42), we obtain

$$-\eta_{x,t}\langle \nabla\Phi(x_t), \nabla_x f(x_t, y_t)\rangle \leq -2\ell\eta_{x,t}\big(f(x_t, y_t) - f(u_t, y_t)\big) + \frac{\eta_{x,t}}{4}\|\nabla\Phi(x_t)\|^2.$$

Using $f(u_t, y_t) \leq g(u_t)$ and (42),

$$\begin{aligned}
-2\ell\eta_{x,t}\big(f(x_t, y_t) - f(u_t, y_t)\big) &\leq -2\ell\eta_{x,t} f(x_t, y_t) + 2\ell\eta_{x,t} g(u_t) \\
&= -2\ell\eta_{x,t} f(x_t, y_t) + 2\ell\eta_{x,t}\Phi(x_t) - 2\ell^2\eta_{x,t}\|x_t - u_t\|^2 \\
&= 2\ell\eta_{x,t}\big(\Phi(x_t) - f(x_t, y_t)\big) - \frac{\eta_{x,t}}{2}\|\nabla\Phi(x_t)\|^2.
\end{aligned}$$

Therefore,

$$-\eta_{x,t}\langle \nabla\Phi(x_t), \nabla_x f(x_t, y_t)\rangle \leq -\frac{\eta_{x,t}}{4}\|\nabla\Phi(x_t)\|^2 + 2\ell\eta_{x,t}\big(\Phi(x_t) - f(x_t, y_t)\big). \tag{45}$$

Finally, since $\Phi(x_t) \leq g(x_t) = \max_{y\in\mathcal{Y}} f(x_t, y)$, we have $\Phi(x_t) - f(x_t, y_t) \leq g(x_t) - f(x_t, y_t) = b_t$, hence

$$-\eta_{x,t}\langle \nabla\Phi(x_t), \nabla_x f(x_t, y_t)\rangle \leq -\frac{\eta_{x,t}}{4}\|\nabla\Phi(x_t)\|^2 + 2\ell\eta_{x,t} b_t.$$

Step 3: Bound the quadratic term. Using $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$,

$$\ell\eta_{x,t}^2\|\nabla_x f(x_t, y_t) + \xi_t^x\|^2 \leq 2\ell\eta_{x,t}^2\|\nabla_x f(x_t, y_t)\|^2 + 2\ell\eta_{x,t}^2\|\xi_t^x\|^2.$$

Assumption 3.3 states that $x \mapsto f(x, y)$ is L-Lipschitz for each fixed y; since $f(\cdot, y_t)$ is differentiable, this implies $\|\nabla_x f(x_t, y_t)\| \leq L$. Therefore

$$\ell\eta_{x,t}^2\|\nabla_x f(x_t, y_t) + \xi_t^x\|^2 \leq 2\ell L^2\eta_{x,t}^2 + 2\ell\eta_{x,t}^2\|\xi_t^x\|^2.$$

Step 4: Combine. Plugging the last two displays into (43) yields:

$$a_{t+1} \leq a_t - \frac{\eta_{x,t}}{4}\|\nabla\Phi(x_t)\|^2 + 2\ell\eta_{x,t} b_t - \eta_{x,t}\langle \nabla\Phi(x_t), \xi_t^x\rangle + 2\ell\eta_{x,t}^2 L^2 + 2\ell\eta_{x,t}^2\|\xi_t^x\|^2. \tag{46}$$

■

C.1.4. LEMMA 5.4

Proof Recall $b_{t+1} = g(x_{t+1}) - f(x_{t+1}, y_{t+1})$. We first relate b_{t+1} to the gap against the fixed comparator $y_s^\star$.

Step 1: Reduce b_{t+1} to a fixed-comparator gap. Since $x \mapsto f(x, y)$ is L-Lipschitz for each y, the outer function $g(x) = \max_{y \in \mathcal{Y}} f(x, y)$ is also L-Lipschitz:

$$g(x) - g(x') \le \max_{y \in \mathcal{Y}} \big(f(x, y) - f(x', y)\big) \le L\|x - x'\|,$$

and similarly $g(x') - g(x) \le L\|x - x'\|$. Therefore,

$$g(x_{t+1}) \le g(x_s) + L\|x_{t+1} - x_s\| = f(x_s, y_s^\star) + L\|x_{t+1} - x_s\|.$$

Also, by L-Lipschitzness of $x \mapsto f(x, y_s^\star)$,

$$f(x_s, y_s^\star) \le f(x_{t+1}, y_s^\star) + L\|x_{t+1} - x_s\|.$$

Combining the two inequalities gives

$$b_{t+1} = g(x_{t+1}) - f(x_{t+1}, y_{t+1}) \le 2L\|x_{t+1} - x_s\| + \big(f(x_{t+1}, y_s^\star) - f(x_{t+1}, y_{t+1})\big). \tag{47}$$

Step 2: One-step ascent bound in y. Define the concave function (in y)

$$h(y) := f(x_{t+1}, y), \qquad y \in \mathcal{Y},$$

which is ℓ-smooth by Assumption 3.1. The y-update is

$$y_{t+1} = \Pi_{\mathcal{Y}}\Big(y_t + \eta_{y,t}\big(\nabla h(y_t) + \xi_t^y\big)\Big).$$

Let $g_t := \nabla h(y_t) + \xi_t^y$ so that $y_{t+1} = \Pi_{\mathcal{Y}}(y_t + \eta_{y,t} g_t)$.

(i) Concavity + smoothness. Concavity implies

$$h(y_s^\star) \le h(y_t) + \langle \nabla h(y_t), y_s^\star - y_t \rangle, \tag{48}$$

and ℓ-smoothness implies (equivalently, smoothness of $-h$)

$$h(y_t) - h(y_{t+1}) \le \langle \nabla h(y_t), y_t - y_{t+1} \rangle + \frac{\ell}{2}\|y_{t+1} - y_t\|^2. \tag{49}$$

Adding (48) and (49) yields

$$h(y_s^\star) - h(y_{t+1}) \le \langle \nabla h(y_t), y_s^\star - y_{t+1} \rangle + \frac{\ell}{2}\|y_{t+1} - y_t\|^2. \tag{50}$$

(ii) Insert g_t and apply the projection inequality. Since $\nabla h(y_t) = g_t - \xi_t^y$, we can rewrite the inner product term as

$$\langle \nabla h(y_t), y_s^\star - y_{t+1} \rangle = \langle g_t, y_s^\star - y_{t+1} \rangle + \langle \xi_t^y, y_{t+1} - y_s^\star \rangle.$$

For Euclidean projection, the following inequality holds: if $y_{t+1} = \Pi_{\mathcal{Y}}(y_t + \eta g_t)$, then for any $y \in \mathcal{Y}$,

$$\langle g_t, y - y_{t+1} \rangle \leq \frac{\|y_t - y\|^2 - \|y_{t+1} - y\|^2 - \|y_{t+1} - y_t\|^2}{2\eta}. \tag{51}$$

Applying (51) with $y = y_s^\star$ and $\eta = \eta_{y,t}$ gives

$$\langle g_t, y_s^\star - y_{t+1} \rangle \leq \frac{\|y_t - y_s^\star\|^2 - \|y_{t+1} - y_s^\star\|^2 - \|y_{t+1} - y_t\|^2}{2\eta_{y,t}}.$$

Thus, from (50),

$$\begin{aligned} h(y_s^\star) - h(y_{t+1}) &\leq \frac{\|y_t - y_s^\star\|^2 - \|y_{t+1} - y_s^\star\|^2}{2\eta_{y,t}} - \frac{\|y_{t+1} - y_t\|^2}{2\eta_{y,t}} \\ &\quad + \langle \xi_t^y, y_{t+1} - y_s^\star \rangle + \frac{\ell}{2}\|y_{t+1} - y_t\|^2 \qquad (52) \\ &\leq \frac{\|y_t - y_s^\star\|^2 - \|y_{t+1} - y_s^\star\|^2}{2\eta_{y,t}} + \langle \xi_t^y, y_t - y_s^* \rangle + \frac{\eta_{y,t}}{2(1 - \ell\eta_{y,t})} \|\xi_t^y\|^2 \qquad (53) \\ &\leq \frac{\|y_t - y_s^\star\|^2 - \|y_{t+1} - y_s^\star\|^2}{2\eta_{y,t}} + \langle \xi_t^y, y_t - y_s^* \rangle + \eta_{y,t} \|\xi_t^y\|^2 \qquad (54) \end{aligned}$$

where we use Young's inequality with a parameter $\frac{\eta_{y,t}}{1-\ell\eta_{y,t}}$ for the second inequality, and $\eta_{y,t} \leq \frac{1}{2\ell}$ for the last inequality.

Step 3: Conclude. Since $h(y) = f(x_{t+1}, y)$, inequality (54) is exactly a bound on $f(x_{t+1}, y_s^\star) - f(x_{t+1}, y_{t+1})$. Substituting (54) into (47) yields:

$$b_{t+1} \leq 2L \|x_{t+1} - x_s\| + \frac{\|y_t - y_s^*\|^2 - \|y_{t+1} - y_s^*\|^2}{2\eta_{y,t}} + \langle \xi_t^y, y_t - y_s^* \rangle + \eta_{y,t} \|\xi_t^y\|^2. \tag{55}$$

■

C.1.5. LEMMA 5.5

Proof We start from Lemma 5.3 and rearrange it as:

$$\frac{\eta_{x,t}}{4}\|\nabla\Phi(x_t)\|^2 \leq (a_t - a_{t+1}) + 2\ell\eta_{x,t}b_t - \eta_{x,t}\langle \nabla\Phi(x_t), \xi_t^x \rangle + 2\ell\eta_{x,t}^2 L^2 + 2\ell\eta_{x,t}^2 \|\xi_t^x\|^2. \tag{56}$$

Since $\lambda_t = 2\ell\eta_{x,t}$, we have $2\ell\eta_{x,t}b_t = \lambda_t b_t$. Therefore,

$$(a_t - a_{t+1}) + \lambda_t b_t = (a_t + \lambda_t b_t) - a_{t+1} = P_t - a_{t+1}.$$

Moreover, $P_{t+1} = a_{t+1} + \lambda_{t+1}b_{t+1}$ implies

$$P_t - a_{t+1} = (P_t - P_{t+1}) + \lambda_{t+1}b_{t+1}.$$

Substituting these identities into (56) yields

$$\frac{\eta_{x,t}}{4}\|\nabla\Phi(x_t)\|^2 \leq (P_t - P_{t+1}) + \lambda_{t+1}b_{t+1} - \eta_{x,t}\langle \nabla\Phi(x_t), \xi_t^x \rangle + 2\ell\eta_{x,t}^2 L^2 + 2\ell\eta_{x,t}^2 \|\xi_t^x\|^2. \tag{57}$$

Finally, apply Lemma 5.4 to bound b_{t+1}:

$$b_{t+1} \le 2L\|x_{t+1} - x_s\| + \frac{\|y_t - y_s^\star\|^2 - \|y_{t+1} - y_s^\star\|^2}{2\eta_{y,t}} + \langle y_t - y_s^\star, \xi_t^y\rangle + \eta_{y,t}\,\|\xi_t^y\|^2 .$$

Multiplying by λ_{t+1} and plugging into (57) gives the desired result. ■

C.1.6. THEOREM C.1

Theorem C.1 (Convergence of SGDA in Subgaussian NC-C games) *Suppose Assumptions 3.1 and 3.3 to 3.5 hold. Fix $T \ge 1$ and a deterministic partition $0 = s_0 < s_1 < \cdots < s_M = T$. Let $B_m := s_{m+1} - s_m$ and let $s(t) = s_m$ whenever $s_m \le t < s_{m+1}$. For each block start, choose an $\mathcal{F}_{s_m}$-measurable maximizer $y_{s_m}^\star \in \arg\max_{y\in\mathcal{Y}} f(x_{s_m}, y)$. Assume deterministic steps satisfy $0 < \eta_{x,t} \le \frac{1}{8\ell}, 0 < \eta_{y,t} \le \frac{1}{2\ell}, \lambda_t := 2\ell\eta_{x,t}$, and both $(\eta_{x,t})_t$ and $(\lambda_{t+1}/\eta_{y,t})_t$ are nonincreasing. Define*

$$S_1(T) := \sum_{t=0}^{T-1} \eta_{x,t}, \quad S_2(T) := \sum_{t=0}^{T-1} \eta_{x,t}^2, \quad S_2'(T) := \sum_{t=0}^{T-1} \eta_{x,t}\eta_{y,t},$$

$$S_{2,B}(T) := \sum_{m=0}^{M-1} B_m \sum_{t=s_m}^{s_{m+1}-1} \eta_{x,t}^2, \qquad S_r(T) := \sum_{m=0}^{M-1} \frac{\eta_{x,s_m}}{\eta_{y,s_m}},$$

$$\eta_{\max} := \max_{0\le t<T} \eta_{x,t}, \qquad \Gamma_\delta := \max\{1, \log(c/\delta)\}.$$

Then with probability at least $1 - \delta$,

$$\min_{0\le t<T} \|\nabla\Phi(x_t)\|^2 \le \frac{C}{S_1(T)}\Big[P_0 + \ell D^2 S_r(T) + \ell\big(L^2 + \sigma^2\Gamma_\delta\big)\big(S_2(T) + S_{2,B}(T)\big) + \ell D\sigma\sqrt{S_2(T)\Gamma_\delta} + \ell\sigma^2\Gamma_\delta S_2'(T) + \sigma^2\eta_{\max}\Gamma_\delta\Big].$$

Proof We start from Lemma 5.5, introduce an augmented potential that telescopes the within-block y-distance term, apply Lemma 4.4 only to the self-bounding x-noise, and then postprocess the pathwise remainder.

Step 1: start from Lemma 5.5 Fix any $t \in \{0, 1, \ldots, T-1\}$ and let $s := s(t)$. Lemma 5.5 gives, almost surely,

$$\frac{\eta_{x,t}}{4}\|\nabla\Phi(x_t)\|^2 \le P_t - P_{t+1} + 2L^2\ell\,\eta_{x,t}^2 + \lambda_{t+1}\Bigg(2L\|x_{t+1} - x_s\| + \frac{\|y_t - y_s^*\|^2 - \|y_{t+1} - y_s^*\|^2}{2\eta_{y,t}} + \langle y_t - y_s^*, \xi_t^y\rangle + \eta_{y,t}\|\xi_t^y\|^2\Bigg) - \eta_{x,t}\langle\nabla\Phi(x_t), \xi_t^x\rangle + 2\ell\,\eta_{x,t}^2\|\xi_t^x\|^2. \tag{58}$$

Step 2: introduce the y-drift augmented potential. Define

$$w_t := \frac{\lambda_{t+1}}{2\eta_{y,t}}, \qquad \widehat{P}_t := P_t + w_t\|y_t - y^*_{s(t)}\|^2.$$

For the terminal potential, set $s(T) := s_{M-1}$ only to make $\widehat{P}_T$ well-defined. Note that $(w_t)_{t\ge 0}$ is nonincreasing because $(\lambda_{t+1}/\eta_{y,t})_{t\ge 0}$ is nonincreasing.

Rewrite the drift term in (58) as

$$\lambda_{t+1}\cdot\frac{\|y_t - y^*_s\|^2 - \|y_{t+1} - y^*_s\|^2}{2\eta_{y,t}} = w_t\big(\|y_t - y^*_s\|^2 - \|y_{t+1} - y^*_s\|^2\big).$$

We now compare the telescoping structure with the augmented potential:

$$\widehat{P}_t - \widehat{P}_{t+1} = (P_t - P_{t+1}) + w_t\|y_t - y^*_s\|^2 - w_{t+1}\|y_{t+1} - y^*_{s(t+1)}\|^2. \tag{59}$$

Subtracting (59) from the desired telescoping expression gives

$$\begin{aligned}&(P_t - P_{t+1}) + w_t\big(\|y_t - y^*_s\|^2 - \|y_{t+1} - y^*_s\|^2\big) - (\widehat{P}_t - \widehat{P}_{t+1})\\ &\quad = -w_t\|y_{t+1} - y^*_s\|^2 + w_{t+1}\|y_{t+1} - y^*_{s(t+1)}\|^2.\end{aligned}$$

If $s(t+1) = s(t)$, this residual equals $-(w_t - w_{t+1})\|y_{t+1} - y^\star_s\|^2 \le 0$. If $t+1 < T$ and $s(t+1) \ne s(t)$, the residual is at most $D^2 w_{t+1}$ since $y_{t+1}, y^*_{s(t+1)} \in \mathcal{Y}$. Therefore, for all $t < T$,

$$(P_t - P_{t+1}) + w_t\big(\|y_t - y^*_s\|^2 - \|y_{t+1} - y^*_s\|^2\big) \;\le\; \widehat{P}_t - \widehat{P}_{t+1} + D^2 w_{t+1}\mathbf{1}_{\{t+1<T,\, s(t+1)\ne s(t)\}}. \tag{60}$$

Plugging (60) into (58) yields, for all t,

$$\begin{aligned}\frac{\eta_{x,t}}{4}\|\nabla\Phi(x_t)\|^2 \le\; &(\widehat{P}_t - \widehat{P}_{t+1}) + 2L^2\ell\,\eta_{x,t}^2 + 2L\lambda_{t+1}\|x_{t+1} - x_{s(t)}\|\\ &+ D^2 w_{t+1}\mathbf{1}_{\{t+1<T,\, s(t+1)\ne s(t)\}} + \lambda_{t+1}\langle y_t - y^*_{s(t)}, \xi^y_t\rangle + \lambda_{t+1}\eta_{y,t}\|\xi^y_t\|^2\\ &- \eta_{x,t}\langle\nabla\Phi(x_t), \xi^x_t\rangle + 2\ell\,\eta_{x,t}^2\|\xi^x_t\|^2.\end{aligned} \tag{61}$$

Step 3: fit into Lemma 4.4 (self-bounding subgaussian martingales). Define

$$G_t := \frac{\eta_{x,t}}{4}\|\nabla\Phi(x_t)\|^2, \qquad \mathcal{P}_t := \widehat{P}_t, \qquad c^x_t := -\nabla\Phi(x_t), \qquad d^x_t := 2\ell.$$

Then the x-noise part in (61) equals

$$\eta_{x,t}\langle c^x_t, \xi^x_t\rangle + d^x_t\eta_{x,t}^2\|\xi^x_t\|^2.$$

Let the remaining terms be

$$\begin{aligned}r_t := \;&2L^2\ell\,\eta_{x,t}^2 + 2L\lambda_{t+1}\|x_{t+1} - x_{s(t)}\| + D^2 w_{t+1}\mathbf{1}_{\{t+1<T,\, s(t+1)\ne s(t)\}}\\ &+ \lambda_{t+1}\langle y_t - y^*_{s(t)}, \xi^y_t\rangle + \lambda_{t+1}\eta_{y,t}\|\xi^y_t\|^2.\end{aligned} \tag{62}$$

Then (61) becomes

$$G_t \le \mathcal{P}_t - \mathcal{P}_{t+1} + \eta_{x,t}\langle c^x_t, \xi^x_t\rangle + d^x_t\eta_{x,t}^2\|\xi^x_t\|^2 + r_t. \tag{63}$$

Moreover, the self-bounding condition in Lemma 4.4 holds with $C_1 = 4$ since

$$\eta_{x,t}\|c^x_t\|^2 = \eta_{x,t}\|\nabla\Phi(x_t)\|^2 = 4G_t.$$

Step 4: apply Lemma 4.4 to control the x-noise terms. Apply Lemma 4.4 to (63) with $C_1 = 4$, $d_t^x = 2\ell$, $c_t^y = 0$, $d_t^y = 0$, and $\eta_{y,t}^{\text{Lemma}} = 0$. Then for any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$,

$$\sum_{t=0}^{T-1} G_t \le 2\widehat{P}_0 + 2\sum_{t=0}^{T-1} r_t + C\sigma^2\eta_{\max}\Gamma_{\delta_1} + C\ell\sigma^2 S_2(T)\Gamma_{\delta_1}, \tag{64}$$

where $\Gamma_{\delta_1} := \max\{1, \log(c/\delta_1)\}$ and $\eta_{\max} = \max_{t\in[0,T)} \eta_{x,t}$.

Step 5: postprocess. It remains to upper bound $\widehat{P}_0$ and $\sum r_t$ using $S_2, S_2', S_{2,B}, S_r$.

(a) Bounding $\widehat{P}_0$ and the boundary term by $\ell D^2 S_r(T)$. Since $y_0, y^*_{s(0)} \in \mathcal{Y}$, $\|y_0 - y^*_{s(0)}\| \le D$, hence

$$\widehat{P}_0 = P_0 + w_0\|y_0 - y^*_{s(0)}\|^2 \le P_0 + D^2 w_0.$$

Moreover, $w_0 = \lambda_1/(2\eta_{y,0}) = \ell\,\eta_{x,1}/\eta_{y,0} \le \ell\,\eta_{x,0}/\eta_{y,0} \le \ell S_r(T)$, so

$$\widehat{P}_0 \le P_0 + \ell D^2 S_r(T). \tag{65}$$

Similarly,

$$\sum_{t=0}^{T-1} D^2 w_{t+1}\mathbf{1}_{\{t+1<T,\, s(t+1)\neq s(t)\}} = \sum_{m=1}^{M-1} D^2 w_{s_m} \le \ell D^2 \sum_{m=1}^{M-1} \frac{\eta_{x,s_m}}{\eta_{y,s_m}} \le \ell D^2 S_r(T).$$

Thus, the boundary contribution is also $\lesssim \ell D^2 S_r(T)$.

(b) Deterministic smooth/Lipschitz term.

$$\sum_{t=0}^{T-1} 2L^2\ell\,\eta_{x,t}^2 = 2\ell L^2 S_2(T). \tag{66}$$

(c) Blockwise x-drift term into $S_{2,B}(T)$. We bound $\sum_t 2L\lambda_{t+1}\|x_{t+1} - x_{s(t)}\|$. First, since $f(\cdot, y)$ is L-Lipschitz and differentiable, $\|\nabla_x f(x,y)\| \le L$ for all (x,y). Using the SGDA update $x_{t+1} = x_t - \eta_{x,t}(\nabla_x f(x_t, y_t) + \xi_t^x)$ gives

$$\|x_{t+1} - x_t\| \le \eta_{x,t}(L + \|\xi_t^x\|). \tag{67}$$

Fix a block m with start $s = s_m$. For $t \in \{s, \dots, s_{m+1} - 1\}$,

$$\|x_{t+1} - x_s\| \le \sum_{i=s}^{t} \|x_{i+1} - x_i\| \le \sum_{i=s}^{t} \eta_{x,i}(L + \|\xi_i^x\|).$$

Also $\lambda_{t+1} = 2\ell\eta_{x,t+1} \le 2\ell\eta_{x,t}$ since $(\eta_{x,t})$ is nonincreasing. Hence,

$$\begin{aligned}\sum_{t=s}^{s_{m+1}-1} 2L\lambda_{t+1}\|x_{t+1} - x_s\| &\le \sum_{t=s}^{s_{m+1}-1} 4\ell L\,\eta_{x,t}\sum_{i=s}^{t}\eta_{x,i}(L + \|\xi_i^x\|) \\ &= 4\ell L \sum_{i=s}^{s_{m+1}-1} \eta_{x,i}(L + \|\xi_i^x\|) \sum_{t=i}^{s_{m+1}-1} \eta_{x,t}.\end{aligned} \tag{68}$$

Since $\eta_{x,t} \le \eta_{x,i}$ for $t \ge i$, we have $\sum_{t=i}^{s_{m+1}-1} \eta_{x,t} \le B_m \eta_{x,i}$. Plugging this into (68) yields

$$\sum_{t=s}^{s_{m+1}-1} 2L\lambda_{t+1}\|x_{t+1} - x_s\| \le 4\ell L B_m \sum_{i=s}^{s_{m+1}-1} \eta_{x,i}^2 (L + \|\xi_i^x\|).$$

Using $L(L + \|\xi\|) \le \frac{3}{2}L^2 + \frac{1}{2}\|\xi\|^2$,

$$4\ell L B_m\, \eta_{x,i}^2 (L + \|\xi_i^x\|) \le 6\ell L^2 B_m\, \eta_{x,i}^2 + 2\ell B_m\, \eta_{x,i}^2 \|\xi_i^x\|^2.$$

Summing over blocks,

$$\sum_{t=0}^{T-1} 2L\lambda_{t+1}\|x_{t+1} - x_{s(t)}\| \le 6\ell L^2 S_{2,B}(T) + 2\ell \sum_{m=0}^{M-1} B_m \sum_{t=s_m}^{s_{m+1}-1} \eta_{x,t}^2 \|\xi_t^x\|^2. \tag{69}$$

By Theorem A.9 with weights $a_i = B_m \eta_{x,i}^2$ for $i \in [s_m, s_{m+1} - 1]$, with probability at least $1 - \delta_2$,

$$\sum_{m=0}^{M-1} B_m \sum_{t=s_m}^{s_{m+1}-1} \eta_{x,t}^2 \|\xi_t^x\|^2 \;\le\; C\sigma^2 S_{2,B}(T) \log(c/\delta_2). \tag{70}$$

Combining (69)–(70) gives

$$\sum_{t=0}^{T-1} 2L\lambda_{t+1}\|x_{t+1} - x_{s(t)}\| \le C\ell\big(L^2 + \sigma^2 \log(c/\delta_2)\big)\, S_{2,B}(T). \tag{71}$$

(d) y-noise inner product: a $\sqrt{S_2 \log}$ bound. Define $Z_t := \lambda_{t+1}\langle y_t - y^*_{s(t)}, \xi_t^y\rangle$. Conditioned on $\mathcal{F}_{t+1/2}$, $y_t - y^*_{s(t)}$ is measurable, so Theorem A.8 implies that Z_t is conditionally subgaussian up to a universal constant, with scale at most $\sigma D \lambda_{t+1}$. By standard subgaussian martingale concentration, for any $\delta_3 \in (0,1)$, with probability at least $1 - \delta_3$,

$$\sum_{t=0}^{T-1} \lambda_{t+1}\langle y_t - y^*_{s(t)}, \xi_t^y\rangle \le C\sigma D \sqrt{\log(1/\delta_3) \sum_{t=0}^{T-1} \lambda_{t+1}^2}. \tag{72}$$

Since $\lambda_{t+1} = 2\ell\eta_{x,t+1} \le 2\ell\eta_{x,t}$,

$$\sum_{t=0}^{T-1} \lambda_{t+1}^2 \le 4\ell^2 \sum_{t=0}^{T-1} \eta_{x,t}^2 = 4\ell^2 S_2(T).$$

Plugging into (72) gives

$$\sum_{t=0}^{T-1} \lambda_{t+1}\langle y_t - y^*_{s(t)}, \xi_t^y\rangle \le C\ell D\sigma \sqrt{S_2(T) \log(1/\delta_3)}. \tag{73}$$

(e) y-noise quadratic term: an $S_2'(T)\log$ bound. Since $\lambda_{t+1} \le 2\ell\eta_{x,t}$,

$$\sum_{t=0}^{T-1} \lambda_{t+1}\eta_{y,t} \le 2\ell \sum_{t=0}^{T-1} \eta_{x,t}\eta_{y,t} = 2\ell S_2'(T).$$

By Theorem A.9 applied to ξ_t^y with weights $\alpha_t = \lambda_{t+1}\eta_{y,t}$, with probability at least $1-\delta_4$,

$$\sum_{t=0}^{T-1} \lambda_{t+1}\eta_{y,t}\|\xi_t^y\|^2 \le C\sigma^2 \log(c/\delta_4) \sum_{t=0}^{T-1} \lambda_{t+1}\eta_{y,t} \le C\ell\sigma^2 S_2'(T)\log(c/\delta_4). \tag{74}$$

(f) Collecting the bounds for $\sum r_t$. Combining (66), (71), the boundary estimate in (a), and (73)–(74), we obtain that with probability at least $1-(\delta_2+\delta_3+\delta_4)$,

$$\begin{aligned}\sum_{t=0}^{T-1} r_t \le{}& C\ell D^2 S_r(T) + C\ell\big(L^2 + \sigma^2\log(c/\delta_2)\big)\big(S_{2,B}(T) + S_2(T)\big) \\ &+ C\ell D\sigma\sqrt{S_2(T)\log(1/\delta_3)} + C\ell\sigma^2 S_2'(T)\log(c/\delta_4).\end{aligned} \tag{75}$$

Step 6: divide by $S_1(T)$ to get $\min_t \|\nabla\Phi(x_t)\|^2$. By definition $G_t = \frac{\eta_{x,t}}{4}\|\nabla\Phi(x_t)\|^2$,

$$\sum_{t=0}^{T-1} G_t = \frac{1}{4}\sum_{t=0}^{T-1} \eta_{x,t}\|\nabla\Phi(x_t)\|^2 \ge \frac{1}{4} S_1(T)\cdot \min_{t\in[0,T)} \|\nabla\Phi(x_t)\|^2.$$

Hence

$$\min_{t\in[0,T)} \|\nabla\Phi(x_t)\|^2 \le \frac{4}{S_1(T)} \sum_{t=0}^{T-1} G_t. \tag{76}$$

Now combine (64), (65), and (75). Set $\delta_1=\delta_2=\delta_3=\delta_4=\delta/4$ and apply a union bound to obtain an event of probability at least $1-\delta$ on which

$$\begin{aligned}\sum_{t=0}^{T-1} G_t \le C\Bigg(& P_0 + \ell D^2 S_r(T) + \big(\ell L^2 + \ell\sigma^2\log(c/\delta)\big)\big(S_{2,B}(T)+S_2(T)\big) \\ & + \ell D\sigma\sqrt{S_2(T)\log(c/\delta)} + \ell\sigma^2 S_2'(T)\log(c/\delta) + \sigma^2\eta_{\max}\log(c/\delta)\Bigg),\end{aligned} \tag{77}$$

Finally apply (76) to conclude the theorem.

■

C.1.7. THEOREM 5.6

Proof If $\bar{P}_0 = 0$, then $a_0 = 0$ and $b_0 = 0$. Hence $\Phi(x_0) = \Phi^\star$, and since Φ is 2ℓ-smooth and minimized at x_0, $\nabla\Phi(x_0) = 0$. Thus the conclusion is trivial. Assume $\bar{P}_0 > 0$. Let

$$\Gamma_\delta := \max\{1, \log(c/\delta)\}, \quad Q_\delta := L^2 + \sigma^2\Gamma_\delta,$$

$$c_x := \frac{\bar{P}_0^{3/4}}{\big(\ell^3 D^2\sigma^2\Gamma_\delta Q_\delta\big)^{1/4}}, \quad c_y := \frac{\bar{P}_0^{1/4} D^{1/2} Q_\delta^{1/4}}{\ell^{1/4}\sigma^{3/2}\Gamma_\delta^{3/4}}, \quad c_B := \left(\frac{\ell D^2\sigma^2\Gamma_\delta}{\bar{P}_0 Q_\delta}\right)^{1/2},$$

$$T_0 := \left\lceil 2 \vee (8\ell c_x)^{4/3} \vee (2\ell c_y)^4 \vee 4c_B^2 \vee c_B^{-2} \right\rceil.$$

The proof uses a block partition only as an analysis device.

Step 1: finite fixed-step specialization of Theorem C.1. For arbitrary constant steps $\eta_{x,t} \equiv \eta_x$ and $\eta_{y,t} \equiv \eta_y$ obeying $\eta_x \le 1/(8\ell)$ and $\eta_y \le 1/(2\ell)$, and for any uniform proof partition with block length $B \in \{1, \dots, T\}$, Theorem C.1 gives, with probability at least $1 - \delta$,

$$\min_{0 \le t < T} \|\nabla \Phi(x_t)\|^2 \le C \left[\frac{P_0}{T\eta_x} + \frac{\ell D^2}{B\eta_y} + \ell Q_\delta \eta_x (B+1) + \ell \sigma^2 \Gamma_\delta \eta_y + \frac{\ell D \sigma \sqrt{\Gamma_\delta}}{\sqrt{T}} + \frac{\sigma^2 \Gamma_\delta}{T} \right]. \tag{C1-fixed}$$

Indeed, for this partition,

$$S_1 = T\eta_x, \qquad S_2 = T\eta_x^2, \qquad S_2' = T\eta_x\eta_y, \qquad \eta_{\max} = \eta_x,$$

while $S_r \le C(T/B)(\eta_x/\eta_y)$ and $S_{2,B} \le BT\eta_x^2$. Substituting these estimates into Theorem C.1 yields (C1-fixed).

Step 2: choose the hidden proof block length. Set

$$B := \left\lceil c_B \sqrt{T} \right\rceil.$$

Step 3: verify the step caps. Since $T \ge T_0$,

$$T \ge (8\ell c_x)^{4/3} \quad \Longrightarrow \quad \eta_x = c_x T^{-3/4} \le \frac{1}{8\ell},$$

and

$$T \ge (2\ell c_y)^4 \quad \Longrightarrow \quad \eta_y = c_y T^{-1/4} \le \frac{1}{2\ell}.$$

Step 4: verify the hidden block feasibility. Since $T \ge c_B^{-2}$, we have $c_B\sqrt{T} \ge 1$, and therefore

$$B = \left\lceil c_B \sqrt{T} \right\rceil \le 2c_B \sqrt{T}.$$

Also $T \ge 4c_B^2$ implies $2c_B\sqrt{T} \le T$. Thus $1 \le B \le T$.

Step 5: substitute the theorem's choices. By $P_0 \le \bar{P}_0$, (C1-fixed) gives

$$\min_{0 \le t < T} \|\nabla \Phi(x_t)\|^2 \le C \left[\frac{\bar{P}_0}{T\eta_x} + \frac{\ell D^2}{B\eta_y} + \ell Q_\delta \eta_x (B+1) + \ell \sigma^2 \Gamma_\delta \eta_y + \frac{\ell D \sigma \sqrt{\Gamma_\delta}}{\sqrt{T}} + \frac{\sigma^2 \Gamma_\delta}{T} \right].$$

Substitute $\eta_x = c_x T^{-3/4}$, $\eta_y = c_y T^{-1/4}$, and $B = \lceil c_B \sqrt{T} \rceil$. The block bounds from Step 4 give $B \ge c_B\sqrt{T}$ and $B \le 2c_B\sqrt{T}$. Therefore

$$\frac{\bar{P}_0}{T\eta_x} = \frac{\bar{P}_0}{c_x T^{1/4}}, \qquad \ell Q_\delta \eta_x B \le 2\ell Q_\delta c_x c_B T^{-1/4},$$

$$\frac{\ell D^2}{B\eta_y} \le \frac{\ell D^2}{c_B c_y T^{1/4}}, \qquad \ell \sigma^2 \Gamma_\delta \eta_y = \ell \sigma^2 \Gamma_\delta c_y T^{-1/4}.$$

With the definitions of c_x, c_y, c_B, each of the four displayed terms is bounded by a universal constant times

$$\left(\frac{\ell^3 D^2 \sigma^2 \Gamma_\delta Q_\delta \bar{P}_0}{T} \right)^{1/4}.$$

The remaining $+1$ in $B+1$ contributes

$$\ell Q_\delta \eta_x = \ell Q_\delta c_x T^{-3/4},$$

which is bounded by the same balanced term after enlarging the universal constant, since $B \geq 1$ implies $\eta_x \leq \eta_x B$. Keeping the two residual terms in (C1-fixed) gives

$$\min_{0\leq t<T} \|\nabla\Phi(x_t)\|^2 \leq C\left[\left(\frac{\ell^3 D^2\sigma^2\Gamma_\delta Q_\delta \bar{P}_0}{T}\right)^{1/4} + \frac{\ell D\sigma\sqrt{\Gamma_\delta}}{\sqrt{T}} + \frac{\sigma^2\Gamma_\delta}{T}\right].$$

■

C.1.8. THEOREM 5.7

Proof If $\bar{P}_0 = 0$, then $a_0 = b_0 = 0$, and the same argument as in the proof of Theorem 5.6 gives $\nabla\Phi(x_0) = 0$. Assume $\bar{P}_0 > 0$. Let

$$\Gamma_\delta := \max\{1, \log(c/\delta)\}, \qquad Q_\delta := L^2 + \sigma^2\Gamma_\delta,$$

$$A := \bar{P}_0, \qquad B := \ell Q_\delta, \qquad C_0 := \ell D^2, \qquad D_0 := \ell\sigma^2\Gamma_\delta.$$

Define

$$M_\delta := (ABC_0D_0)^{1/4} = \left(\ell^3 D^2\sigma^2\Gamma_\delta Q_\delta \bar{P}_0\right)^{1/4},$$

$$c_x := \frac{A^{3/4}}{(BC_0D_0)^{1/4}}, \qquad c_y := \frac{A^{1/4}(BC_0)^{1/4}}{D_0^{3/4}},$$

$$c_B := \left(\frac{C_0D_0}{AB}\right)^{1/2} = \left(\frac{\ell D^2\sigma^2\Gamma_\delta}{\bar{P}_0 Q_\delta}\right)^{1/2}.$$

Construct the deterministic shift

$$s_0 := \left\lceil 1 \vee (8\ell c_x)^{4/3} \vee (2\ell c_y)^4 \vee c_B^2 \vee c_B^{-2} \vee \left(\frac{\sigma^2\Gamma_\delta}{M_\delta}\right)^{4/3} \right\rceil.$$

The theorem's schedules are exactly

$$\eta_{x,t} = c_x(t+s_0)^{-3/4}, \qquad \eta_{y,t} = c_y(t+s_0)^{-1/4}.$$

Step 1: verify the step caps, monotonicity, and initial potential. By the definition of s_0,

$$s_0 \geq (8\ell c_x)^{4/3} \quad\Longrightarrow\quad \eta_{x,0} = c_x s_0^{-3/4} \leq \frac{1}{8\ell},$$

$$s_0 \geq (2\ell c_y)^4 \quad\Longrightarrow\quad \eta_{y,0} = c_y s_0^{-1/4} \leq \frac{1}{2\ell}.$$

Since both schedules are decreasing, these caps hold for every $t \geq 0$. Also $(\eta_{x,t})_t$ is nonincreasing. For the second monotonicity condition in Theorem C.1, write, for $u = t + s_0 \geq 1$,

$$\frac{\lambda_{t+1}}{\eta_{y,t}} = 2\ell\frac{c_x}{c_y}\frac{(t+s_0)^{1/4}}{(t+1+s_0)^{3/4}} = 2\ell\frac{c_x}{c_y}u^{1/4}(u+1)^{-3/4}.$$

The map $u \mapsto u^{1/4}(u+1)^{-3/4}$ is nonincreasing on $[1,\infty)$ because

$$\frac{d}{du}\left(\frac{1}{4}\log u - \frac{3}{4}\log(u+1)\right) = \frac{1-2u}{4u(u+1)} \le 0.$$

Finally, $\eta_{x,0} \le 1/(8\ell)$ gives

$$P_0 = \Phi(x_0) - \Phi^\star + 2\ell\eta_{x,0}\big(g(x_0) - f(x_0,y_0)\big) \le \bar{P}_0 = A.$$

Step 2: define the hidden adaptive proof partition. Fix an arbitrary horizon $T \ge \max\{2, s_0\}$ and set

$$\Lambda_T := 1 + \log(1 + T/s_0).$$

Define the deterministic partition recursively by $\tau_0 := 0$ and

$$\tau_{m+1} := \min\left\{T,\ \tau_m + \left\lceil c_B\sqrt{\tau_m + s_0}\right\rceil\right\}$$

until $\tau_M = T$, and let

$$B_m := \tau_{m+1} - \tau_m.$$

This partition is only used in the proof; it is not part of SGDA. For every nonempty block,

$$B_m \le \left\lceil c_B\sqrt{\tau_m + s_0}\right\rceil \le c_B\sqrt{\tau_m + s_0} + 1 \le 2c_B\sqrt{\tau_m + s_0},$$

because $c_B\sqrt{\tau_m + s_0} \ge c_B\sqrt{s_0} \ge 1$ by $s_0 \ge c_B^{-2}$. For every full nonterminal block, the recursion also gives

$$B_m \ge c_B\sqrt{\tau_m + s_0}.$$

Step 3: evaluate the sums in Theorem C.1. The ordinary sums satisfy

$$S_1(T) = c_x\sum_{t=0}^{T-1}(t+s_0)^{-3/4} \ge c\,c_x T^{1/4}.$$

Also,

$$S_2(T) = c_x^2\sum_{t=0}^{T-1}(t+s_0)^{-3/2} \le C c_x^2 s_0^{-1/2},$$

$$S_2'(T) = c_x c_y\sum_{t=0}^{T-1}(t+s_0)^{-1} \le C c_x c_y \Lambda_T, \qquad \eta_{\max} = c_x s_0^{-3/4}.$$

For the block-comparator sum, put $u_m := \tau_m + s_0$. Then

$$S_r(T) = \sum_{m=0}^{M-1}\frac{\eta_{x,\tau_m}}{\eta_{y,\tau_m}} = \frac{c_x}{c_y}\sum_{m=0}^{M-1} u_m^{-1/2}.$$

For every full nonterminal block,

$$\frac{c_B}{\sqrt{u_m}} \le \frac{B_m}{u_m}.$$

The upper block bound and $s_0 \geq c_B^2$ imply $B_m/u_m \leq 2$. Hence

$$\frac{B_m}{u_m} \leq C \log\left(\frac{u_{m+1}}{u_m}\right) = C \log\left(\frac{\tau_{m+1}+s_0}{\tau_m+s_0}\right).$$

Summing over the full nonterminal blocks telescopes, while the final block contributes at most $s_0^{-1/2} \leq c_B^{-1}$. Therefore

$$\sum_{m=0}^{M-1} u_m^{-1/2} \leq C c_B^{-1} \Lambda_T, \qquad S_r(T) \leq C \frac{c_x}{c_y c_B} \Lambda_T.$$

For the block-drift sum, monotonicity of $\eta_{x,t}$ within each block gives

$$\begin{aligned} S_{2,B}(T) &= \sum_{m=0}^{M-1} B_m \sum_{t=\tau_m}^{\tau_{m+1}-1} \eta_{x,t}^2 \\ &\leq \sum_{m=0}^{M-1} B_m^2 c_x^2 (\tau_m + s_0)^{-3/2} \\ &\leq C c_x^2 c_B^2 \sum_{m=0}^{M-1} (\tau_m + s_0)^{-1/2} \\ &\leq C c_x^2 c_B \Lambda_T. \end{aligned}$$

Step 4: apply Theorem C.1. Apply Theorem C.1 with the deterministic partition above and with block starts $s_m = \tau_m$. Using $P_0 \leq A$ and the preceding sum estimates, with probability at least $1-\delta$,

$$\begin{aligned} \min_{0\leq t<T} \|\nabla \Phi(x_t)\|^2 \leq \frac{C}{T^{1/4}} \Bigg[&\frac{A}{c_x} + \frac{\ell D^2}{c_y c_B} \Lambda_T + \ell Q_\delta c_x c_B \Lambda_T \\ &+ \ell \sigma^2 \Gamma_\delta c_y \Lambda_T + \ell Q_\delta c_x s_0^{-1/2} \\ &+ \ell D \sigma \sqrt{\Gamma_\delta} s_0^{-1/4} + \sigma^2 \Gamma_\delta s_0^{-3/4} \Bigg]. \end{aligned} \tag{*}$$

Indeed, the terms arise as follows:

$$\frac{P_0}{S_1(T)} \leq \frac{C}{T^{1/4}} \frac{A}{c_x},$$

$$\frac{\ell D^2 S_r(T)}{S_1(T)} \leq \frac{C}{T^{1/4}} \frac{\ell D^2}{c_y c_B} \Lambda_T,$$

$$\frac{\ell Q_\delta S_{2,B}(T)}{S_1(T)} \leq \frac{C}{T^{1/4}} \ell Q_\delta c_x c_B \Lambda_T,$$

$$\frac{\ell Q_\delta S_2(T)}{S_1(T)} \leq \frac{C}{T^{1/4}} \ell Q_\delta c_x s_0^{-1/2},$$

$$\frac{\ell D \sigma \sqrt{S_2(T) \Gamma_\delta}}{S_1(T)} \leq \frac{C}{T^{1/4}} \ell D \sigma \sqrt{\Gamma_\delta} s_0^{-1/4},$$

$$\frac{\ell\sigma^2\Gamma_\delta S_2'(T)}{S_1(T)} \le \frac{C}{T^{1/4}}\ell\sigma^2\Gamma_\delta c_y\Lambda_T,$$
$$\frac{\sigma^2\eta_{\max}\Gamma_\delta}{S_1(T)} \le \frac{C}{T^{1/4}}\sigma^2\Gamma_\delta s_0^{-3/4}.$$

Step 5: balance the constants. The definitions of c_x, c_y, c_B, M_δ give the four identities

$$\frac{A}{c_x} = M_\delta, \qquad \frac{\ell D^2}{c_y c_B} = M_\delta, \qquad \ell Q_\delta c_x c_B = M_\delta, \qquad \ell\sigma^2\Gamma_\delta c_y = M_\delta.$$

The remaining terms are lower order by the construction of s_0. First,

$$s_0 \ge c_B^{-2} \quad \Longrightarrow \quad s_0^{-1/2} \le c_B,$$

so

$$\ell Q_\delta c_x s_0^{-1/2} \le \ell Q_\delta c_x c_B = M_\delta.$$

Second,

$$s_0 \ge c_B^2 \quad \Longrightarrow \quad s_0^{-1/4} \le c_B^{-1/2},$$

and

$$\ell D\sigma\sqrt{\Gamma_\delta}\, c_B^{-1/2} = M_\delta,$$

so

$$\ell D\sigma\sqrt{\Gamma_\delta} s_0^{-1/4} \le M_\delta.$$

Third,

$$s_0 \ge \left(\frac{\sigma^2\Gamma_\delta}{M_\delta}\right)^{4/3} \quad \Longrightarrow \quad \sigma^2\Gamma_\delta s_0^{-3/4} \le M_\delta.$$

Substituting these bounds into $(*)$ and using $\Lambda_T \ge 1$ gives

$$\min_{0\le t<T} \|\nabla\Phi(x_t)\|^2 \le \frac{C\Lambda_T}{T^{1/4}} M_\delta.$$

Since $T \ge s_0 \ge 1$,

$$\Lambda_T = 1 + \log(1 + T/s_0) \le C(1 + \log T),$$

and the definition of M_δ gives the claimed bound. Since $T \ge \max\{2, s_0\}$ was arbitrary, the theorem follows. ■

C.2. SGDA$_{\text{Clip}}$ in Heavy-Tailed NC-C Games

C.2.1. LEMMA 5.8

Proof We start from Lemma 5.5 and replace the SGDA noise vectors by

$$\widetilde{\xi}_t^x + \beta_t^x, \qquad \widetilde{\xi}_t^y + \beta_t^y.$$

The block dual-gap term inherited from Lemma 5.4 is

$$2L\left\|x_{t+1} - x_s\right\|,$$

so the clipped version must keep the same shifted iterate. The direct substitution gives

$$\frac{\eta_{x,t}}{4}\left\|\nabla\Phi(x_t)\right\|^2 \le P_t - P_{t+1} + 2L^2\ell\eta_{x,t}^2 - \eta_{x,t}\left\langle\nabla\Phi(x_t), \widetilde{\xi}_t^x + \beta_t^x\right\rangle + 2\ell\eta_{x,t}^2\left\|\widetilde{\xi}_t^x + \beta_t^x\right\|^2 + \lambda_{t+1}\left[2L\left\|x_{t+1} - x_s\right\| + \frac{\left\|y_t - y_s^\star\right\|^2 - \left\|y_{t+1} - y_s^\star\right\|^2}{2\eta_{y,t}} + \left\langle y_t - y_s^\star, \widetilde{\xi}_t^y + \beta_t^y\right\rangle + \eta_{y,t}\left\|\widetilde{\xi}_t^y + \beta_t^y\right\|^2\right].$$

For the x-bias inner product,

$$-\eta_{x,t}\left\langle\nabla\Phi(x_t), \beta_t^x\right\rangle \le \frac{\eta_{x,t}}{8}\left\|\nabla\Phi(x_t)\right\|^2 + 2\eta_{x,t}\left\|\beta_t^x\right\|^2.$$

Also,

$$2\ell\eta_{x,t}^2\left\|\widetilde{\xi}_t^x + \beta_t^x\right\|^2 \le 4\ell\eta_{x,t}^2\left\|\widetilde{\xi}_t^x\right\|^2 + 4\ell\eta_{x,t}^2\left\|\beta_t^x\right\|^2 \le 4\ell\eta_{x,t}^2\left\|\widetilde{\xi}_t^x\right\|^2 + 2\eta_{x,t}\left\|\beta_t^x\right\|^2,$$

where the last inequality uses $\eta_{x,t} \le 1/(2\ell)$. Finally,

$$\left\|\widetilde{\xi}_t^y + \beta_t^y\right\|^2 \le 2\left\|\widetilde{\xi}_t^y\right\|^2 + 2\left\|\beta_t^y\right\|^2.$$

Substituting these estimates and moving the absorbed $(\eta_{x,t}/8)\left\|\nabla\Phi(x_t)\right\|^2$ term to the left proves the lemma. ■

C.2.2. THEOREM C.2

Theorem C.2 (Convergence of SGDA$_{\text{Clip}}$ in heavy-tailed NC-C games) *Suppose Assumptions 3.1, 3.3, 3.4 and 3.6 hold with $p \in (1, 2]$. Fix a horizon $T \ge 1$ and a deterministic partition $0 = s_0 < s_1 < \cdots < s_M = T$. Let $s(t) = s_m$ whenever $s_m \le t < s_{m+1}$, and choose $y_{s_m}^\star \in \arg\max_{y\in\mathcal{Y}} f(x_{s_m}, y)$. Assume $\eta_{x,t} \le \frac{1}{8\ell}, \eta_{y,t} \le \frac{1}{2\ell}, \tau_{x,t} \ge 2L, \tau_{y,t} \ge 2\ell D$, and let $\lambda_t \overset{def}{=} 2\ell\eta_{x,t}, w_t \overset{def}{=} \frac{\lambda_{t+1}}{2\eta_{y,t}}, u_t \overset{def}{=} D^2\lambda_{t+1}w_t = \frac{D^2\lambda_{t+1}^2}{2\eta_{y,t}}$, and assume w_t is nonincreasing. Define $G_t \overset{def}{=} \frac{\eta_{x,t}}{8}\left\|\nabla\Phi(x_t)\right\|^2 + u_t$ and $\bar{G}_t \overset{def}{=} \frac{\eta_{x,t}}{8}L^2 + u_t$. Let $C_0 > 0$ be a sufficiently large constant and define:*

$$B_t := \max\left\{1, \max_{0\le i\le t}\{B_i^x, B_i^y, B_i^v\}\right\}, \qquad z_t := B_t^{-1},$$

where

$$B_i^x := 2\tau_{x,i}\sqrt{C_0\eta_{x,i}\bar{G}_i} + 64\ell\eta_{x,i}^2\tau_{x,i}^2,$$

$$B_i^y := 2\tau_{y,i}\sqrt{C_0\eta_{y,i}\bar{G}_i} + 64w_i\eta_{y,i}^2\tau_{y,i}^2,$$

$$B_i^v := 96C_0\sigma^p\left(\eta_{x,i}\tau_{x,i}^{2-p} + \eta_{y,i}\tau_{y,i}^{2-p}\right).$$

For $k \le T$, define

$$S_d(k) := D^2 w_0 z_0 + \sum_{t=0}^{k-1} z_t \left[2L^2 \ell \eta_{x,t}^2 + 2L\lambda_{t+1} \left\| x_{t+1} - x_{s(t)} \right\| + D^2 w_t \mathbb{1}\{s(t+1) \neq s(t)\} + u_t \right],$$

$$S_{0,\tau}(k) := \sum_{t=0}^{k-1} z_t \lambda_{t+1} \tau_{y,t}^{1-p},$$

$$S_{1,\tau}(k) := \sum_{t=0}^{k-1} z_t \left[\eta_{x,t} \tau_{x,t}^{2-2p} + \lambda_{t+1} \eta_{y,t} \tau_{y,t}^{2-2p} \right],$$

$$S_{2,\tau}(k) := \sum_{t=0}^{k-1} z_t \left[\ell \eta_{x,t}^2 \tau_{x,t}^{2-p} + \lambda_{t+1} \eta_{y,t} \tau_{y,t}^{2-p} \right].$$

Then, with probability at least $1-\delta$, for all $1 \le k \le T$,

$$z_{k-1} P_k + \sum_{t=0}^{k-1} z_t G_t \le C \left[z_0 P_0 + S_d(k) + D\sigma^p S_{0,\tau}(k) + \sigma^{2p} S_{1,\tau}(k) + \sigma^p S_{2,\tau}(k) + \Gamma_\delta \right],$$

where $C, c > 0$ are universal numerical constants and $\Gamma_\delta := \max\{1, \log(c/\delta)\}$.

Proof Define the augmented potential

$$\widetilde{P}_t := P_t + w_t \left\| y_t - y^\star_{s(t)} \right\|^2.$$

From Lemma 5.8, for $s = s(t)$,

$$\begin{aligned}
\frac{\eta_{x,t}}{8} \left\| \nabla \Phi(x_t) \right\|^2 \le{}& P_t - P_{t+1} + 2L^2 \ell \eta_{x,t}^2 + 2L\lambda_{t+1} \left\| x_{t+1} - x_{s(t)} \right\| \\
&+ \lambda_{t+1} \frac{\left\| y_t - y^\star_{s(t)} \right\|^2 - \left\| y_{t+1} - y^\star_{s(t)} \right\|^2}{2\eta_{y,t}} - \eta_{x,t} \left\langle \nabla \Phi(x_t), \widetilde{\xi}_t^x \right\rangle + 4\ell \eta_{x,t}^2 \left\| \widetilde{\xi}_t^x \right\|^2 \\
&+ \lambda_{t+1} \left\langle y_t - y^\star_{s(t)}, \widetilde{\xi}_t^y \right\rangle + 2\lambda_{t+1} \eta_{y,t} \left\| \widetilde{\xi}_t^y \right\|^2 + 4\eta_{x,t} \left\| \beta_t^x \right\|^2 \\
&+ \lambda_{t+1} \left\langle y_t - y^\star_{s(t)}, \beta_t^y \right\rangle + 2\lambda_{t+1} \eta_{y,t} \left\| \beta_t^y \right\|^2.
\end{aligned}$$

Because $w_t = \lambda_{t+1}/(2\eta_{y,t})$, the distance difference equals

$$w_t \left(\left\| y_t - y^\star_{s(t)} \right\|^2 - \left\| y_{t+1} - y^\star_{s(t)} \right\|^2 \right).$$

Using nonincreasing w_t and the blockwise definition of $s(t)$,

$$w_t \left(\left\| y_t - y^\star_{s(t)} \right\|^2 - \left\| y_{t+1} - y^\star_{s(t)} \right\|^2 \right) \le \widetilde{P}_t - \widetilde{P}_{t+1} + D^2 w_t \mathbb{1}\{s(t+1) \neq s(t)\}.$$

Add u_t to both sides and set

$$c_t^x := -\nabla\Phi(x_t), \qquad c_t^y := \frac{\lambda_{t+1}}{\eta_{y,t}}(y_t - y^\star_{s(t)}), \qquad d_t^x := 4\ell, \qquad d_t^y := 4w_t.$$

Since $d_t^y \eta_{y,t}^2 = 2\lambda_{t+1}\eta_{y,t}$, we obtain

$$G_t \le \widetilde{P}_t - \widetilde{P}_{t+1} + \eta_{x,t}\left\langle c_t^x, \widetilde{\xi}_t^x\right\rangle + d_t^x\eta_{x,t}^2\left\|\widetilde{\xi}_t^x\right\|^2 + \eta_{y,t}\left\langle c_t^y, \widetilde{\xi}_t^y\right\rangle + d_t^y\eta_{y,t}^2\left\|\widetilde{\xi}_t^y\right\|^2 + r_t,$$

where the nonnegative remainder is

$$\begin{aligned} r_t :=& 2L^2\ell\eta_{x,t}^2 + 2L\lambda_{t+1}\left\|x_{t+1} - x_{s(t)}\right\| + D^2 w_t \mathbb{1}\{s(t+1)\neq s(t)\} + u_t \\ &+ 4\eta_{x,t}\left\|\beta_t^x\right\|^2 + D\lambda_{t+1}\left\|\beta_t^y\right\| + 2\lambda_{t+1}\eta_{y,t}\left\|\beta_t^y\right\|^2. \end{aligned}$$

This is the only place where the y-bias inner product is handled: it is paid linearly via $\lambda_{t+1}\left\langle y_t - y^\star_{s(t)}, \beta_t^y\right\rangle \le D\lambda_{t+1}\left\|\beta_t^y\right\|$. The self-bounding property is immediate:

$$\eta_{x,t}\left\|c_t^x\right\|^2 = \eta_{x,t}\left\|\nabla\Phi(x_t)\right\|^2 \le 8G_t,$$

$$\eta_{y,t}\left\|c_t^y\right\|^2 = \frac{\lambda_{t+1}^2}{\eta_{y,t}}\left\|y_t - y^\star_{s(t)}\right\|^2 \le \frac{D^2\lambda_{t+1}^2}{\eta_{y,t}} = 2u_t \le 2G_t.$$

Thus $\eta_{x,t}\left\|c_t^x\right\|^2 + \eta_{y,t}\left\|c_t^y\right\|^2 \le 10G_t$.

The deterministic low-signal conditions in the theorem imply Lemma 4.8 applies at every iterate: $\|\nabla_x f(x_t, y_t)\| \le L \le \tau_{x,t}/2$ and $\|\nabla_y f(x_{t+1}, y_t)\| \le \ell D \le \tau_{y,t}/2$. Hence, almost surely,

$$\left\|\widetilde{\xi}_t^x\right\| \le 2\tau_{x,t}, \qquad \left\|\widetilde{\xi}_t^y\right\| \le 2\tau_{y,t},$$

$$\mathbb{E}[\left\|\widetilde{\xi}_t^x\right\|^2 \mid \mathcal{F}_t] \le 16\sigma^p\tau_{x,t}^{2-p}, \qquad \mathbb{E}[\left\|\widetilde{\xi}_t^y\right\|^2 \mid \mathcal{F}_{t+1/2}] \le 16\sigma^p\tau_{y,t}^{2-p},$$

$$\left\|\beta_t^x\right\| \le 4\sigma^p\tau_{x,t}^{1-p}, \qquad \left\|\beta_t^y\right\| \le 4\sigma^p\tau_{y,t}^{1-p}.$$

Apply Lemma 4.10 to the process $\widetilde{P}_t$ with the deterministic envelope $\bar{G}_t$, $C_1 = 10$, $d_t^x = 4\ell$, and $d_t^y = 4w_t$. The definitions of B_t in the theorem dominate the corresponding normalization terms in that lemma. With probability at least $1-\delta$, simultaneously for all $k \le T$,

$$z_{k-1}\widetilde{P}_k + \sum_{t=0}^{k-1} z_t G_t \le 2z_0\widetilde{P}_0 + 2\sum_{t=0}^{k-1} z_t r_t + 4\sum_{t=0}^{k-1} z_t D_t + 2\Gamma_\delta,$$

where $D_t \le C\sigma^p\left(\ell\eta_{x,t}^2\tau_{x,t}^{2-p} + \lambda_{t+1}\eta_{y,t}\tau_{y,t}^{2-p}\right)$. Moreover, $\widetilde{P}_k \ge P_k$ and $\widetilde{P}_0 \le P_0 + D^2 w_0$. The squared clipping biases give

$$4\eta_{x,t}\left\|\beta_t^x\right\|^2 \le C\sigma^{2p}\eta_{x,t}\tau_{x,t}^{2-2p}, \qquad 2\lambda_{t+1}\eta_{y,t}\left\|\beta_t^y\right\|^2 \le C\sigma^{2p}\lambda_{t+1}\eta_{y,t}\tau_{y,t}^{2-2p},$$

while the linear y-bias gives

$$D\lambda_{t+1}\left\|\beta_t^y\right\| \le 4D\sigma^p\lambda_{t+1}\tau_{y,t}^{1-p}.$$

Substituting these bounds, absorbing numerical constants, and using the definitions of S_d, $S_{0,\tau}$, $S_{1,\tau}$, and $S_{2,\tau}$ proves the theorem. ∎

C.2.3. THEOREM 5.9

Proof Let $a := (2p-1)/(3p-2)$, $c := 1/(3p-2)$, $q := p/(3p-2)$, and $r := 1-a = (p-1)/(3p-2)$. Write

$$\Gamma_\delta := \max\{1, \log(c_0/\delta)\}, \qquad \Delta_\delta := \bar{P}_0 + \Gamma_\delta, \qquad H_\delta := L^2 + \sigma^p \Gamma_\delta, \qquad A_\delta := \ell^3 D^2 \sigma^p \Gamma_\delta H_\delta,$$

and set $Y := \ell D \sigma^p$ and $\tau_0 := \{Y/(A_\delta \Delta_\delta)^{1/4}\}^{4/(3p-2)}$, where $c_0 > 0$ is universal. Write the fixed schedules as $\eta_x = c_x T^{-a}$, $\eta_y = c_y T^{-c}$, and $\tau = \tau_0 T^c$, with

$$c_x := c_{x,p} \frac{\Delta_\delta^{3/4}}{A_\delta^{1/4} \tau_0^{(2-p)/4}}, \qquad c_y := c_{y,p} \frac{(A_\delta \Delta_\delta)^{1/4}}{\ell \sigma^p \Gamma_\delta \, \tau_0^{3(2-p)/4}}.$$

Let $\rho := (D^2/(H_\delta c_x c_y))^{1/2}$ and $B := \lceil \rho T^q \rceil$, and use the deterministic proof partition $0 = s_0 < s_1 < \cdots < s_M = T$ into consecutive blocks of length B, except possibly the last block. The partition is only an analysis device.

Choose T_0 large enough so that $1 \le B \le T$, $\eta_x \le (8\ell)^{-1}$, $\eta_y \le (2\ell)^{-1}$, and $\tau \ge 2\max\{L, \ell D\}$ for all $T \ge T_0$. We also choose T_0 large enough and $c_{x,p}, c_{y,p}$ small enough so that the normalization factors in Theorem C.2 satisfy $B_t \le 2$, hence $z_t \ge 1/2$. The quantities being estimated here are precisely the deterministic envelope $\bar{G}_t$ and the three self-normalizer coefficients B_t^x, B_t^y, B_t^v:

$$\begin{aligned}
\bar{G}_t &= \frac{\eta_x L^2}{8} + \frac{D^2 \lambda_{t+1}^2}{2\eta_y} \lesssim \eta_x L^2 + \ell^2 D^2 \frac{\eta_x^2}{\eta_y}, \\
B_t^x &\lesssim \tau \left(\eta_x L + \ell D \, \eta_x^{3/2} \eta_y^{-1/2} \right) + \ell \eta_x^2 \tau^2, \\
B_t^y &\lesssim \tau \left(L(\eta_x \eta_y)^{1/2} + \ell D \, \eta_x \right) + \ell \eta_x \eta_y \tau^2, \\
B_t^v &\lesssim \sigma^p (\eta_x \tau^{2-p} + \eta_y \tau^{2-p}).
\end{aligned}$$

Under the displayed schedules, the factors appearing in these four lines obey

$$\begin{aligned}
\tau \eta_x &= O(T^{-2r}), & \tau \eta_x^{3/2} \eta_y^{-1/2} &= O(T^{-3r}), & \eta_x^2 \tau^2 &= O(T^{-4r}), \\
\tau (\eta_x \eta_y)^{1/2} &= O(T^{-r}), & \eta_x \eta_y \tau^2 &= O(T^{-2r}), & \eta_x \tau^{2-p} + \eta_y \tau^{2-p} &= O(T^{-r}).
\end{aligned}$$

Thus all self-normalizer coefficients are controlled by the threshold T_0 and the p-dependent constants.

Apply Theorem C.2 with confidence $\delta/2$, and apply Lemma A.11 with confidence $\delta/2$ to the same deterministic partition. Enlarging the universal constant inside Γ_δ if necessary, both events use the same Γ_δ. On their intersection, whose probability is at least $1-\delta$,

$$\begin{aligned}
&\sum_{t=0}^{T-1} z_t G_t \le C \left[\Delta_\delta + S_d(T) + D\sigma^p S_{0,\tau}(T) + \sigma^{2p} S_{1,\tau}(T) + \sigma^p S_{2,\tau}(T) \right], \\
&\sum_{t<T} z_t G_t \ge \frac{1}{8} \left(\sum_{t<T} z_t \eta_x \right) \min_{t<T} \|\nabla \Phi(x_t)\|^2 \ge \frac{c_x}{16} T^r \min_{t<T} \|\nabla \Phi(x_t)\|^2 .
\end{aligned}$$

The lower bound uses $z_t \ge 1/2$ and $T\eta_x = c_x T^r$.

We next bound the terms on the right-hand side of the master inequality. The block-boundary term, the smoothness term, and the u_t-term satisfy

$$\sum_{\text{blocks}} D^2 w_{s_m} \lesssim D^2 \frac{T}{B} \frac{\ell \eta_x}{\eta_y} \lesssim \frac{\ell D^2 c_x}{\rho c_y},$$
$$\sum_{t<T} 2L^2 \ell \eta_x^2 = O(T^{1-2a}) = o(1),$$
$$\sum_{t<T} u_t \lesssim \ell^2 D^2 T \frac{\eta_x^2}{\eta_y} = O(T^{1-2a+c}) = o(1).$$

Here $1 - q - a + c = 0$ gives the boundary estimate, while $1 - 2a + c = -r < 0$ makes the last two terms lower order; increasing T_0 absorbs them.

It remains to control the block drift. For a block $I_m = [s_m, s_{m+1})$, write $M_{m,t} := \sum_{i=s_m}^{t} \eta_x \widetilde{\xi}_i^x$. The clipped update, the NC-C Lipschitz bound $\|\nabla_x f(x_i, y_i)\| \le L$, and the clipping-bias bound $\|\beta_i^x\| \le 4\sigma^p \tau^{1-p}$ give, for $t \in I_m$,

$$\|x_{t+1} - x_{s_m}\| \le \sum_{i=s_m}^{t} \eta_x \left(L + 4\sigma^p \tau^{1-p}\right) + \|M_{m,t}\|.$$

Therefore, for $\mathcal{D}_T := \sum_{t<T} 2L z_t \lambda_{t+1} \|x_{t+1} - x_{s(t)}\|$,

$$\mathcal{D}_T \le C\ell L^2 B T \eta_x^2 + C\ell L \sigma^p B T \eta_x^2 \tau^{1-p} + \sum_{m=0}^{M-1} \alpha_m \max_{t \in I_m} \|M_{m,t}\|,$$
$$\alpha_m := 2L \sum_{t \in I_m} z_t \lambda_{t+1} \le CL\ell B \eta_x.$$

The first deterministic block-drift term is $\lesssim \ell L^2 \rho c_x^2$ because $1 + q - 2a = 0$, and the deterministic clipping-bias contribution has the extra factor $T^{c(1-p)} = T^{-r}$.

For the weighted-block event, instantiate Lemma A.11 with $\zeta_i := \eta_x \widetilde{\xi}_i^x$, $R_i := 2\eta_x \tau$, and $V_m := 16\sigma^p B \eta_x^2 \tau^{2-p}$. Then $K_m \lesssim \eta_x \sqrt{B\sigma^p \tau^{2-p}} + \eta_x \tau$, and

$$\sum_{m=0}^{M-1} \alpha_m \max_{t \in I_m} \|M_{m,t}\| \le C \left[\sum_{m=0}^{M-1} \alpha_m K_m + \Gamma_\delta \max_{0 \le m < M} \alpha_m K_m \right],$$
$$\sum_{m=0}^{M-1} \alpha_m K_m \lesssim L\ell T \eta_x^2 \left(\sqrt{B\sigma^p \tau^{2-p}} + \tau \right).$$

The first summand in the last line scales as T^{-r} because $1 - 2a + \{q + c(2-p)\}/2 = -r$, and the second also scales as T^{-r} because $1 - 2a + c = -r$. The maximal term is lower order after increasing T_0, so

$$\mathcal{D}_T \lesssim C_p \ell H_\delta \rho c_x^2.$$

The remaining clipped-noise terms are deterministic under the fixed schedules:

$$D\sigma^p S_{0,\tau}(T) \lesssim D\sigma^p \sum_{t<T} \lambda_{t+1} \tau^{1-p} \lesssim Y c_x \tau_0^{1-p},$$
$$\sigma^{2p} S_{1,\tau}(T) \lesssim \sigma^{2p} \left[T\eta_x \tau^{2-2p} + T\eta_x \eta_y \tau^{2-2p} \right],$$
$$\sigma^p S_{2,\tau}(T) \lesssim \ell\sigma^p \left[T\eta_x^2 \tau^{2-p} + T\eta_x \eta_y \tau^{2-p} \right] \lesssim \ell\sigma^p \Gamma_\delta c_x c_y \tau_0^{2-p}.$$

The first line uses $1-a+c(1-p)=0$; the $S_{1,\tau}$ terms and the first $S_{2,\tau}$ term are lower order, while $T\eta_x\eta_y\tau^{2-p} = c_xc_y\tau_0^{2-p}$ and $\Gamma_\delta \geq 1$ give the last bound.

Combining the master estimate with the deterministic bounds gives

$$\sum_{t<T} z_t G_t \leq C_p \left[\Delta_\delta + \frac{\ell D^2 c_x}{\rho c_y} + \ell H_\delta \rho c_x^2 + \ell\sigma^p\Gamma_\delta c_x c_y \tau_0^{2-p} + Y c_x \tau_0^{1-p}\right],$$

and division by the lower bound $c_x T^r/16$ yields

$$\min_{0\leq t<T} \|\nabla\Phi(x_t)\|^2 \leq \frac{C_p}{T^r}\left[\frac{\Delta_\delta}{c_x} + \frac{\ell D^2}{\rho c_y} + \ell H_\delta \rho c_x + \ell\sigma^p\Gamma_\delta c_y\tau_0^{2-p} + Y\tau_0^{1-p}\right].$$

It remains only to evaluate the deterministic choice of constants. The chosen value of ρ balances the two block-boundary terms:

$$\frac{\ell D^2}{\rho c_y} + \ell H_\delta \rho c_x = 2\ell D\sqrt{\frac{H_\delta c_x}{c_y}}.$$

Balancing this expression with $\ell\sigma^p\Gamma_\delta c_y\tau_0^{2-p}$ gives

$$c_y \asymp_p \left(\frac{\ell D\sqrt{H_\delta c_x}}{\ell\sigma^p\Gamma_\delta\,\tau_0^{2-p}}\right)^{2/3}, \qquad \frac{\ell D^2}{\rho c_y} + \ell H_\delta\rho c_x + \ell\sigma^p\Gamma_\delta c_y\tau_0^{2-p} \lesssim_p (A_\delta c_x \tau_0^{2-p})^{1/3}.$$

Balancing this term with Δ_δ/c_x gives

$$c_x \asymp_p \frac{\Delta_\delta^{3/4}}{A_\delta^{1/4}\tau_0^{(2-p)/4}}, \qquad \frac{\Delta_\delta}{c_x} + (A_\delta c_x\tau_0^{2-p})^{1/3} \lesssim_p (A_\delta\Delta_\delta)^{1/4}\tau_0^{(2-p)/4}.$$

The corresponding c_y is exactly the stated $c_{y,p}(A_\delta\Delta_\delta)^{1/4}/(\ell\sigma^p\Gamma_\delta\tau_0^{3(2-p)/4})$, after adjusting the p-dependent constant. Finally, the chosen τ_0 balances the last two terms:

$$(A_\delta\Delta_\delta)^{1/4}\tau_0^{(2-p)/4} + Y\tau_0^{1-p} \lesssim_p (A_\delta\Delta_\delta)^r Y^{(2-p)/(3p-2)}.$$

Substituting into the preceding stationarity bound and using $r=(p-1)/(3p-2)$ gives

$$\min_{0\leq t<T}\|\nabla\Phi(x_t)\|^2 \leq C_p\left(\frac{A_\delta\Delta_\delta\, Y^{(2-p)/(p-1)}}{T}\right)^r,$$

which is the claimed rate after expanding $A_\delta = \ell^3 D^2\sigma^p\Gamma_\delta(L^2+\sigma^p\Gamma_\delta)$, $\Delta_\delta = \bar{P}_0 + \Gamma_\delta$, and $Y = \ell D\sigma^p$. ∎

C.2.4. THEOREM 5.10

Proof Let $a := (2p-1)/(3p-2)$, $c := 1/(3p-2)$, $q := p/(3p-2)$, and $r := 1-a = (p-1)/(3p-2)$. Write

$$\Gamma_\delta := \max\{1, \log(c_0/\delta)\}, \qquad \Delta_\delta := \bar{P}_0 + \Gamma_\delta, \qquad H_\delta := L^2 + \sigma^p \Gamma_\delta, \qquad A_\delta := \ell^3 D^2 \sigma^p \Gamma_\delta H_\delta,$$

and set $Y := \ell D \sigma^p$ and $\tau_0 := \{Y/(A_\delta \Delta_\delta)^{1/4}\}^{4/(3p-2)}$, where $c_0 > 0$ is universal. Write the shifted schedules as $\eta_{x,t} = c_x(t+s_0)^{-a}$, $\eta_{y,t} = c_y(t+s_0)^{-c}$, and $\tau_t = \tau_0(t+s_0)^c$, with

$$c_x := c_{x,p} \frac{\Delta_\delta^{3/4}}{A_\delta^{1/4} \tau_0^{(2-p)/4}}, \qquad c_y := c_{y,p} \frac{(A_\delta \Delta_\delta)^{1/4}}{\ell \sigma^p \Gamma_\delta \, \tau_0^{3(2-p)/4}}.$$

Set $\rho := (D^2/(H_\delta c_x c_y))^{1/2}$. Choose s_0 large enough that, for all $t \geq 0$, we have $\eta_{x,t} \leq (8\ell)^{-1}$, $\eta_{y,t} \leq (2\ell)^{-1}$, $\tau_t \geq 2\max\{L, \ell D\}$. We also choose s_0 large enough, and $c_{x,p}, c_{y,p}$ small enough, so that the normalization factors in Theorem C.2 satisfy $B_t \leq 2$, hence $z_t \geq 1/2$. The quantities being estimated are the deterministic envelope $\bar{G}_t$ and the self-normalizer coefficients B_t^x, B_t^y, B_t^v:

$$\begin{aligned}
\bar{G}_t &= \frac{\eta_{x,t} L^2}{8} + \frac{D^2 \lambda_{t+1}^2}{2\eta_{y,t}} \lesssim \eta_{x,t} L^2 + \ell^2 D^2 \frac{\eta_{x,t}^2}{\eta_{y,t}}, \\
B_t^x &\lesssim \tau_t \left(\eta_{x,t} L + \ell D \, \eta_{x,t}^{3/2} \eta_{y,t}^{-1/2} \right) + \ell \eta_{x,t}^2 \tau_t^2, \\
B_t^y &\lesssim \tau_t \left(L(\eta_{x,t}\eta_{y,t})^{1/2} + \ell D \, \eta_{x,t} \right) + \ell \eta_{x,t} \eta_{y,t} \tau_t^2, \\
B_t^v &\lesssim \sigma^p (\eta_{x,t} \tau_t^{2-p} + \eta_{y,t} \tau_t^{2-p}).
\end{aligned}$$

Here we used $\eta_{x,t+1} \leq \eta_{x,t}$ in the bound on $\bar{G}_t$. Under the shifted schedules, these factors decay as

$$\begin{aligned}
\tau_t \eta_{x,t} &= O((t+s_0)^{-2r}), \quad \tau_t \eta_{x,t}^{3/2} \eta_{y,t}^{-1/2} = O((t+s_0)^{-3r}), \quad \eta_{x,t}^2 \tau_t^2 = O((t+s_0)^{-4r}), \\
\tau_t (\eta_{x,t}\eta_{y,t})^{1/2} &= O((t+s_0)^{-r}), \quad \eta_{x,t}\eta_{y,t}\tau_t^2 = O((t+s_0)^{-2r}), \quad \eta_{x,t}\tau_t^{2-p} = O((t+s_0)^{-3r}), \\
\eta_{y,t}\tau_t^{2-p} &= O((t+s_0)^{-r}).
\end{aligned}$$

Thus the normalization conditions hold after increasing s_0 and shrinking the p-dependent constants if needed.

Also choose s_0 so that w_t is nonincreasing. Since $w_t = \lambda_{t+1}/(2\eta_{y,t}) = \ell(c_x/c_y)(t+1+s_0)^{-a}(t+s_0)^c$, it suffices to impose $s_0 \geq \lceil c/(a-c) \rceil = \lceil 1/(2(p-1)) \rceil$. Indeed, for $u = t + s_0$,

$$\frac{d}{du} \log\big(u^c (u+1)^{-a}\big) = \frac{c}{u} - \frac{a}{u+1} = \frac{c - (a-c)u}{u(u+1)} \leq 0$$

whenever $u \geq c/(a-c)$.

Define the deterministic proof partition by $r_0 = 0$ and

$$r_{j+1} = \min\left\{T, \, r_j + \lceil \rho (r_j + s_0)^q \rceil \right\}.$$

Let $I_j = [r_j, r_{j+1})$ and $B_j := r_{j+1} - r_j$. This partition is only an analysis device.

Apply Theorem C.2 with confidence $\delta/2$, and apply Lemma A.11 with confidence $\delta/2$ to the same deterministic partition. On their intersection event, whose probability is at least $1-\delta$,

$$\begin{aligned}
\sum_{t=0}^{T-1} z_t G_t &\le C\left[\Delta_\delta + S_d(T) + D\sigma^p S_{0,\tau}(T) + \sigma^{2p} S_{1,\tau}(T) + \sigma^p S_{2,\tau}(T)\right], \\
\sum_{t<T} z_t G_t &\ge \frac{1}{8}\sum_{t<T} z_t \eta_{x,t} \min_{t<T} \|\nabla\Phi(x_t)\|^2 \\
&\ge c_p c_x \left(\sum_{t<T}(t+s_0)^{-a}\right) \min_{t<T} \|\nabla\Phi(x_t)\|^2 \\
&\ge c_p c_x T^r \min_{t<T} \|\nabla\Phi(x_t)\|^2 .
\end{aligned}$$

The last line uses $z_t \ge 1/2$ and the integral lower bound $\sum_{t<T}(t+s_0)^{-a} \gtrsim_p T^{1-a}$, valid because $T \ge s_0$.

We now estimate the deterministic terms. By increasing s_0 once more, the adaptive blocks satisfy $r_{j+1}+s_0 \le 2(r_j+s_0)$, and the ceiling error in the block count is absorbed by $1+\log T$ for all $T \ge s_0$. Hence the usual integral comparison for this partition gives the block-boundary, smoothness, and u_t estimates

$$\begin{aligned}
\sum_j D^2 w_{r_j} &\lesssim \ell D^2 \frac{c_x}{c_y} \sum_j (r_j+s_0)^{-(a-c)} \\
&\lesssim \ell D^2 \frac{c_x}{c_y}\left(\frac{1}{\rho}\int_{s_0}^{T+s_0} u^{-q-(a-c)}\,du + 1\right) \\
&\lesssim \frac{\ell D^2 c_x}{\rho c_y}(1+\log T), \\
\sum_{t<T} \eta_{x,t}^2 &\lesssim c_x^2 \sum_t (t+s_0)^{-2a}, \\
\sum_{t<T} \frac{\eta_{x,t}^2}{\eta_{y,t}} &\lesssim \frac{c_x^2}{c_y} \sum_t (t+s_0)^{-2a+c}.
\end{aligned}$$

The boundary estimate uses the monotonicity of w_t and $q+(a-c)=1$; the last two sums are finite uniformly after the shift because $2a>1$ and $2a-c>1$.

For $t \in I_j$, the clipped update, the NC-C bound $\|\nabla_x f(x_i,y_i)\| \le L$, and $\|\beta_i^x\| \le 4\sigma^p \tau_i^{1-p}$ imply

$$\left\|x_{t+1} - x_{r_j}\right\| \le \sum_{i=r_j}^{t} \eta_{x,i}\left(L + 4\sigma^p \tau_i^{1-p}\right) + \left\|\sum_{i=r_j}^{t} \eta_{x,i}\widetilde{\xi}_i^x\right\|.$$

Consequently, the deterministic part of the block drift satisfies

$$\begin{aligned}
C\ell L^2 \sum_j B_j \sum_{t\in I_j} \eta_{x,t}^2 &\lesssim C\ell L^2 \rho c_x^2 \int_{s_0}^{T+s_0} u^{q-2a}\,du \\
&\lesssim \ell L^2 \rho c_x^2 (1+\log T).
\end{aligned}$$

Here $q - 2a = -1$. The deterministic clipping-bias part has the additional factor τ_t^{1-p}, whose exponent makes the corresponding integral summable after the shift.

For the martingale part, instantiate the weighted-block event with $\zeta_i := \eta_{x,i}\widetilde{\xi}_i^x$, $R_i := 2\eta_{x,i}\tau_i$, and $V_j := 16\sigma^p \sum_{i\in I_j} \eta_{x,i}^2 \tau_i^{2-p}$. Use block weights $\alpha_j := 2L \sum_{t\in I_j} z_t\lambda_{t+1}$; for $u_j := r_j + s_0$,

$$\begin{aligned}
\alpha_j &\lesssim L\ell c_x \rho u_j^{q-a},\\
K_j &\lesssim c_x\sqrt{\rho\sigma^p}\tau_0^{(2-p)/2} u_j^{-a+(q+c(2-p))/2} + c_x\tau_0 u_j^{-a+c}.
\end{aligned}$$

Since $q + c(2-p) = 2c$, both powers in K_j are $u_j^{-a+c} = u_j^{-2r}$. Converting the block sum with density $(\rho u^q)^{-1}\,du$ gives exponent $(q - a - 2r) - q = -a - 2r < -1$. Thus $\sum_j \alpha_j K_j$ and $\Gamma_\delta \max_j \alpha_j K_j$ are, after increasing s_0, dominated by the same logarithmic scale, and the whole block-drift term obeys

$$\sum_{t<T} 2Lz_t\lambda_{t+1} \left\|x_{t+1} - x_{s(t)}\right\| \le C_p \ell H_\delta \rho c_x^2 (1 + \log T).$$

The remaining clipped-noise terms are deterministic under the shifted schedules:

$$\begin{aligned}
D\sigma^p S_{0,\tau}(T) &\lesssim Y c_x \tau_0^{1-p} \sum_{t<T} (t+s_0)^{-a+c(1-p)} \lesssim Y c_x \tau_0^{1-p}(1+\log T),\\
\sigma^{2p} S_{1,\tau}(T) &\lesssim \sigma^{2p} \sum_{t<T} \left[\eta_{x,t}\tau_t^{2-2p} + \eta_{x,t}\eta_{y,t}\tau_t^{2-2p}\right],\\
\sigma^p S_{2,\tau}(T) &\lesssim \ell\sigma^p \sum_{t<T} \left[\eta_{x,t}^2\tau_t^{2-p} + \eta_{x,t}\eta_{y,t}\tau_t^{2-p}\right].
\end{aligned}$$

The $S_{0,\tau}$ line is logarithmic because $-a + c(1-p) = -1$. The $S_{1,\tau}$ terms and the first $S_{2,\tau}$ term are summable after the shift, whereas the second $S_{2,\tau}$ term is logarithmic because $-a - c + c(2-p) = -1$; therefore

$$\sigma^p S_{2,\tau}(T) \lesssim \ell\sigma^p \Gamma_\delta c_x c_y \tau_0^{2-p}(1+\log T).$$

Combining the bounds gives

$$\sum_{t<T} z_t G_t \le C_p(1+\log T)\left[\Delta_\delta + \frac{\ell D^2 c_x}{\rho c_y} + \ell H_\delta \rho c_x^2 + \ell\sigma^p\Gamma_\delta c_x c_y \tau_0^{2-p} + Y c_x \tau_0^{1-p}\right].$$

Dividing by the lower bound $c_p c_x T^r$ yields

$$\min_{0\le t<T} \|\nabla\Phi(x_t)\|^2 \le \frac{C_p(1+\log T)}{T^r}\left[\frac{\Delta_\delta}{c_x} + \frac{\ell D^2}{\rho c_y} + \ell H_\delta \rho c_x + \ell\sigma^p\Gamma_\delta c_y \tau_0^{2-p} + Y\tau_0^{1-p}\right].$$

The bracket is the same deterministic bracket optimized in the proof of Theorem 5.9. With the same choices of ρ, c_x, c_y, τ_0, it is bounded by $C_p(A_\delta\Delta_\delta)^r Y^{(2-p)/(3p-2)}$. Therefore

$$\min_{0\le t<T} \|\nabla\Phi(x_t)\|^2 \le C_p \frac{1+\log T}{T^r}(A_\delta\Delta_\delta)^r Y^{(2-p)/(3p-2)},$$

which is the claimed rate after expanding A_δ, Δ_δ, and $Y = \ell D\sigma^p$. ■