HOMOGENIZATION OF ℓ_2-ADVERSARIAL TRAINING IN HIGH-DIMENSIONS: EXACT DYNAMICS UNDER STOCHASTIC GRADIENT DESCENT

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop a framework for analyzing the learning dynamics of ℓ_2-adversarial training of single-index models on binary mixtures of Gaussians in the high-dimensional limit under streaming stochastic gradient descent (SGD). We derive deterministic equivalents for a broad class of statistics of the SGD iterates, including the adversarial risk and distance to adversarial optimality, in terms of the solution to a system of ODEs. We use them to study two idealized learning rate schedules: the Polyak stepsize and exact line search. In the case of ℓ_2-adversarial least squares with a single class, we show that, unlike noiseless standard least squares, no constant learning rate guarantees monotone descent of SGD towards a minimizer of the adversarial risk. Moreover, we identify anisotropic covariance and a mismatch in ridge parameters as the two sources of suboptimality of exact line search relative to the Polyak stepsize. We also introduce a stochastic differential equation (SDE), called adversarial homogenized SGD, that captures the evolution of statistics of the iterates of SGD. For ℓ_2-adversarial least squares, using this SDE, we show the evolution of the risk is equivalent, up to a dimension-free error, to that of SGD on standard least squares with an adaptive learning rate and adaptive ℓ_2-regularization. When the dynamics converge, the limiting adversarial risk and SGD iterate are determined by a fixed-point equation, with the limiting iterate being the solution of a ridge regression problem whose regularization parameter is the limiting effective regularization of SGD.

1 INTRODUCTION

State-of-the-art machine learning models have been shown to be vulnerable to small input perturbations, better known as adversarial examples (Szegedy et al., 2014; Goodfellow et al., 2015; Fawzi et al., 2018; Ilyas et al., 2019; Yuan et al., 2019). These perturbations, often imperceptible to the human eye, are designed to maximize the model's error and result in large drops in performance.

Adversarial training has emerged as one of the most effective and persistent defenses against such perturbations (Goodfellow et al., 2015; Madry et al., 2018; Bai et al., 2021). It consists of training on adversarial examples in order to increase the robustness of the model to unseen adversarial examples. It replaces the standard risk with a robust risk in which each sample is evaluated against its worst-case perturbation. When minimizing this objective with stochastic gradient descent (SGD), the update depends not only on the current iterate and the newly sampled data point, but also on the adversarial perturbation of this data point, thereby modifying the training dynamics of SGD. This dependence manifests empirically in the higher sensitivity of adversarial training to optimization hyperparameters, including the learning rate schedule, training time, weight decay, and early stopping (Rice et al., 2020; Pang et al., 2021; Dabouei et al., 2022).

In this work, we analyze the dynamics of adversarial training of generalized linear models under ℓ_2-adversarial perturbations and SGD. We extend the frameworks of Collins-Woodfin et al. (2024a) and Collins-Woodfin & Seroussi (2025) to account for adversarial training of single-index models. The ℓ_2-adversarial objective function takes the form of the following min-max optimization problem

$$\min_{X\in\mathbb{R}^d}\left\{\mathcal{R}_\lambda^{\mathrm{adv}}(X)\overset{\mathrm{def}}{=}\mathbb{E}_{a,I,\epsilon}\left[\max_{\|\Delta a\|\le\delta} f_I(X^\top(a+\Delta a),\epsilon)\right]+\frac{\lambda}{2}\|X\|^2\right\}, \tag{1}$$

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

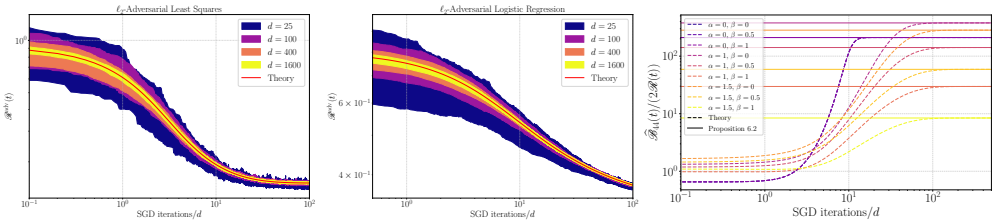

Figure 1: **Left, middle: concentration of the ℓ_2-adversarial risk** for noiseless ℓ_2-adversarial least squares with a single class $a \sim \mathcal{N}(0, K)$ (left) and noiseless hard label logistic regression on a binary Gaussian mixture with distinct means and common covariance (middle). As d grows, the adversarial risk concentrates around the deterministic limit (red) given by the ODEs (26), as predicted by Theorem 2.1. **Right: convergence of** $q(t) \overset{\text{def}}{=} \sqrt{\widehat{\mathscr{B}}_{44}(t)/(2\mathscr{R}(t))}$ for several power-law setups (see (486)), to the limit q_∞ predicted by the fixed-point equation of Theorem 3.2. See Figure 6 for further evidence and Appendix J for simulation details.

where $\|\cdot\|$ denotes the Euclidean norm[1], $\delta > 0$ is the adversarial budget, $\lambda \geq 0$ is a regularization parameter, $(a, I, \epsilon) \sim \mathcal{D} \subset \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}$, $X \in \mathbb{R}^d$ and $\{f_i : \mathbb{R}^2 \to \mathbb{R}\}_{i=1}^2$ satisfy the regularity and growth assumptions formulated below. Since $\{\langle \Delta a, X\rangle : \|\Delta a\| \leq \delta\} = [-\delta\|X\|, \delta\|X\|]$ by the Cauchy–Schwarz inequality, the inner maximization reduces to a scalar one and (1) becomes

$$\mathcal{R}_\lambda^{\text{adv}}(X) \overset{\text{def}}{=} \mathbb{E}_{a,I,\epsilon}\left[\max_{|s|\leq 1} f_I(X^\top a + \delta s\|X\|, \epsilon)\right] + \frac{\lambda}{2}\|X\|^2, \quad \mathcal{R}^{\text{adv}}(X) \overset{\text{def}}{=} \mathcal{R}_0^{\text{adv}}(X). \tag{2}$$

We suppose $a \in \mathbb{R}^d$ is distributed according to a Gaussian mixture with 2 classes. We minimize (2) using streaming SGD and explicitly solve the maximization problem at each step. At each SGD update, a new data triplet (a, I, ϵ) is sampled such that a belongs to class I and $\epsilon \sim \mathcal{N}(0, \eta^2)$ is independent label noise. Our main goal is not to study the optimal method for computing adversarial examples, but rather to provide a framework for analyzing how adversarial objective affects SGD.

Main contributions. In the large learning rate regime and when d is proportional to the number of samples, we derive a deterministic equivalent (det. equiv.) which consists of a system of d-dimensional coupled ODEs which allows us to predict the evolution of key statistics of the iterates of SGD under ℓ_2-adversarial training for general means and commuting covariances. For example, as the dimension d increases, the learning curves of $\mathcal{R}^{\text{Adv}}(X_k)$ concentrate around the det. equiv. $\mathscr{R}^{\text{Adv}}(t)$ with probability better than any inverse power of d (see Figure 1 and Theorem 2.1).

We introduce a SDE called *Adversarial Homogenized SGD* which captures the dynamics of SGD as $d \to \infty$, even for large learning rate (see Section 2). For data sampled from a single class $a \sim \mathcal{N}(0, K)$, using this SDE we show that the evolution of the adversarial risk $\mathcal{R}^{\text{Adv}}$ of ℓ_2-adversarial least squares is equivalent up to a dimension-free error to the evolution of $\mathcal{R}^{\text{Adv}}$ for SGD on standard least squares with an adaptive learning rate and adaptive ℓ_2-regularization (Section 3). We further use this framework to study two aspects of ℓ_2-adversarial training

- *Relationship to ridge regression* (Section 3). For ℓ_2-adversarial least squares on a single class $a \sim \mathcal{N}(0, K)$, the ratio $\|X_k\|/\sqrt{2\mathcal{R}(X_k)}$ governs the dynamics. Whenever it converges as $k \to \infty$, its limit solves a fixed point equation which predicts the limits of the det. equivs. of $\mathcal{R}(X_k)$, $\|X_k\|^2$ and $\mathcal{R}^{\text{Adv}}(X_k)$ and identifies the limiting iterate X_∞ as the solution of ridge regression whose parameter is SGD's limiting effective regularization.
- *Adversarial training mitigates the poor performance of greedy learning rates under strong anisotropy* (Section 4). Using the det. equivs., we study the *Polyak stepsize* (Polyak, 1987) and *exact line search*. For ℓ_2-adversarial least squares on a single class $a \sim \mathcal{N}(0, K)$, no constant learning rate guarantees monotone descent towards $X^{\star,\text{Adv}} \in \operatorname{argmin}_X \mathcal{R}^{\text{adv}}(X)$, unlike standard noiseless least squares. Moreover, at a given state of the det. equivs., the suboptimality of exact line search relative to the Polyak stepsize is driven by anisotropy and a mismatch in ridge parameters, which larger values of δ mitigate.

[1]We will also use $\|\cdot\|$ to denote the Frobenius norm of matrix and $\|\cdot\|_{\text{op}}$ for the operator norm.

Related Work. Adversarial examples and adversarial training have been studied extensively, in particular for deep networks (Szegedy et al., 2014; Goodfellow et al., 2015; Madry et al., 2018; Rice et al., 2020; Bai et al., 2021). Theoretical work has mostly focused on static properties, namely the generalization and regularization effects of adversarial training and the robust versus standard accuracy trade-off, for linear models (Raghunathan et al., 2020; Xing et al., 2021; Ribeiro et al., 2023; Ribeiro & Schön, 2023), with exact asymptotics in the proportional regime under Gaussian data (Javanmard et al., 2020; Hassani & Javanmard, 2024; Vilucchio et al., 2024; Dohmatob & Scetbon, 2024) and Gaussian mixtures (Dobriban et al., 2022; Javanmard & Soltanolkotabi, 2022; Taheri et al., 2021). By contrast, the theoretical understanding of the optimization dynamics of adversarial training is far more limited. From an algorithmic viewpoint, Scetbon & Dohmatob (2023) show that early-stopped gradient descent on the standard least squares objective is optimally robust to ℓ_2-attacks up to an absolute constant. Closest to our work, Gu et al. (2023) use a stochastic modified equation (Li et al., 2018) to approximate the min-max dynamics of SGD by an SDE in the vanishing learning rate limit. Our diffusion approximation does not rely on this and holds as $d \to \infty$ for any bounded learning rate schedule, including large ones and beyond the convergence threshold.

Our analysis builds on the long line of work deriving deterministic limits for stochastic optimization algorithms on isotropic Gaussian data (Biehl & Riegler, 1994; Biehl & Schwarze, 1995; Saad & Solla, 1995a;b). These results were subsequently extended to other models under isotropic data (Goldt et al., 2019; Ben Arous et al., 2022; Arnaboldi et al., 2023), to non-identity covariance (Collins-Woodfin & Paquette, 2024; Collins-Woodfin et al., 2024a; Goldt et al., 2020; 2022) and to Gaussian mixtures (Refinetti et al., 2021; Mignacco et al., 2021; Collins-Woodfin & Seroussi, 2025; Arous et al., 2026). We extend Collins-Woodfin et al. (2024a); Collins-Woodfin & Seroussi (2025) to the ℓ_2-adversarial objective. See Appendix A for an extended discussion of related work.

1.1 Model Setup

We study the setting of streaming SGD characterized by a sequence of independent samples $\{a_k, y_k\}$. Here y_k is the target which is either a function of data triplet (a_k, I_k, ϵ_k) or a hard label. We formalize this in the following assumptions analogously to Collins-Woodfin et al. (2024a); Collins-Woodfin & Seroussi (2025).

Assumption 1 (Data and classes)**.** *For $i = 1, 2$, conditioned on belonging to class i with probability p_i, $a \mid i \sim \mathcal{N}(\mu_i, K_i)$ with covariance $K_i \in \mathbb{R}^{d\times d}$ with operator norm $\|K_i\|_{\mathrm{op}} \le C$ for some constant $C > 0$ (independent of d). Moreover, K_1 and K_2 commute. We also scale the means $\|\mu_i\|^2 \le C$ and write $\mu \overset{\mathrm{def}}{=} \mu_1 \oplus \mu_2 \in \mathbb{R}^{d\times 2}$ where $\oplus$ denotes concatenation. Finally, $p_2 = 1 - p_1$.*

We also place the following assumption on the labels.

Assumption 2 (Target model and label noise)**.** *The targets satisfy one the following models:*

a) (Hard Label) $y \mid i = i$ with $i \in \{-1, 1\}$.

b) (Soft Label) For a fixed vector $X^\star$ and a link function $g : \mathbb{R} \to \mathbb{R}$ such that $y \mid i = g(X^{\star\top} a; \epsilon)$ for $a \mid i \sim \mathcal{N}(\mu_i, K_i)$. The label noise $\epsilon \sim \mathcal{N}(0, \eta^2)$, independent of (a, i).

Since the data a is sampled from a Gaussian mixture, the adversarial risk (2) decomposes as follows

$$\mathcal{R}_\lambda^{\mathrm{adv}}(X) = p_1 \cdot \mathbb{E}_{\,v,\epsilon}[\Psi_1(X; \sqrt{K_1}v + \mu_1, \epsilon)] + p_2 \mathbb{E}_{\,v,\epsilon}[\Psi_2(X; \sqrt{K_2}v + \mu_2, \epsilon)] + \frac{\lambda}{2}\|X\|^2, \quad (3)$$

where for $i = 1, 2$ we define $\Psi_i(X; \sqrt{K_i}v + \mu_i, \epsilon) \overset{\mathrm{def}}{=} \max_{|s|\le 1} f_i(X^\top(\sqrt{K_i}v + \mu_i) + \delta s\|X\|, \epsilon)$ and $v \sim \mathcal{N}(0, I_d)$. As we are minimizing (2), we consider $f_i(X^\top a, \epsilon)$ as a function of its first variable $X^\top a$ with parameter ϵ and write $f_i(X^\top a, \epsilon) = f_i(X^\top a; \epsilon)$. Given the presence of the $\max$ in (2), we must first formalize when the objective is differentiable.

Assumption 3 (pseudo-Lipschitz f)**.** *The function $f_i : \mathbb{R}^3 \to \mathbb{R}$ is α-pseudo-Lipschitz with constant $L(f) > 0$ for $i = 1, 2$.* [2]

[2]For normed vector spaces $\mathcal{A}, \mathcal{B}$ with norms $\|\cdot\|_{\mathcal{A}}$ and $\|\cdot\|_{\mathcal{B}}$, respectively, and for $\alpha \ge 0$, we say a function $F : \mathcal{A} \to \mathcal{B}$ is *α-pseudo-Lipschitz* with constant L if for any $A, \widehat{A} \in \mathcal{A}$, we have

$$\|F(A) - F(\widehat{A})\|_{\mathcal{B}} \le L\|A - \widehat{A}\|_{\mathcal{A}}(1 + \|A\|_{\mathcal{A}}^{\alpha} + \|\widehat{A}\|_{\mathcal{A}}^{\alpha}).$$

We denote by f_i' the (almost everywhere) derivative of f_i with respect to its first variable. Because of the maximum over s, the differentiability of Ψ_i in X is governed by Danskin (1967)'s theorem: for $X \neq 0$, Ψ_i is differentiable whenever all maximizing s yield the same gradient of $f_i(X^\top a + \delta s \|X\|; \epsilon)$, and its gradient is then this common value. At $X = 0$, differentiability fails since $\|X\|$ is not differentiable there, so we require $X_0 \neq 0$. Our analysis therefore follows SGD up to a stopping time before which these conditions hold (see Assumption 6). We also impose an assumption on $X^\star$.

Assumption 4 (Initialization and signal)**.** *For some constant C independent of d, the initialization $X_0 \in \mathbb{R}^d \setminus \{0\}$ and signal $X^\star \in \mathbb{R}^d$ satisfy $\max\{\|X_0\|, \|X^\star\|\} \leq C$.*

Since the minimizer of $\mathcal{R}^{\mathrm{Adv}}$ need not coincide with $X^\star$, we also track a minimizer of $\mathcal{R}^{\mathrm{Adv}}$.

Assumption 5 (Adversarial minimizer)**.** *There exists $X^{\star,\mathrm{Adv}} \in \operatorname{argmin}_X \mathcal{R}^{\mathrm{adv}}(X)$ such that $\|X^{\star,\mathrm{Adv}}\| \leq C$ for some $C > 0$ independent of d.*

We build upon the unified notation of Collins-Woodfin & Seroussi (2025) for hard and soft labels:

$$\hat{X} \overset{\mathrm{def}}{=} \begin{cases} X \oplus X^\star \in \mathbb{R}^{d\times 2} & \text{(soft labels)} \\ X \oplus 0 \in \mathbb{R}^{d\times 2} & \text{(hard labels)} \end{cases} \quad W = \hat{X} \oplus \mu \in \mathbb{R}^{d\times 4} \text{ and } \widehat{W} = W \oplus X^{\star,\mathrm{Adv}} \in \mathbb{R}^{d\times 5}.$$

For $i = 1, 2$, we introduce the following block matrices

$$\widehat{B}_i(W) \overset{\mathrm{def}}{=} \begin{bmatrix} B_i(W) & m_i(W) & 0 \\ m_i(W)^\top & \|\mu_i\|^2 & 0 \\ 0 & 0 & \|X\|^2 \end{bmatrix}, \quad B_i(W) \overset{\mathrm{def}}{=} \hat{X}^\top K_i \hat{X}, \quad m_i(W) \overset{\mathrm{def}}{=} \hat{X}^\top \mu_i. \tag{4}$$

Here $m_i(W)$ and $B_i(W)$ denote the mean vector and covariance matrix of the random vector $\hat{X}^\top a_i$ for $a_i \sim \mathcal{N}(\mu_i, K_i)$. Since $\mathcal{R}^{\mathrm{adv}}(X)$ is an expectation over the Gaussian random vector $\hat{X}^\top a_i$ for $a_i \sim \mathcal{N}(\mu_i, K_i)$, there exists some well-behaved functions $h_i : \mathbb{R}^{4\times 4} \to \mathbb{R}$ such that

$$\mathcal{R}^{\mathrm{adv}}(X) \overset{\mathrm{def}}{=} p_1 h_1(\widehat{B}_1) + p_2 h_2(\widehat{B}_2).$$

The next assumption requires the $\mathcal{R}^{\mathrm{Adv}}$ to be a smooth function of $\widehat{B}_i(W)$ so we can commute differentiation and expectation. Moreover, it defines the set where Danskin (1967)'s Theorem holds.

Assumption 6 (Risk representation)**.** *There exists an open set $\mathcal{U} \subset \mathbb{R}^{4\times 4}$ such that $(\widehat{B}_1(W_0), \widehat{B}_2(W_0)) \in \mathcal{U}^2$. Provided $(\widehat{B}_2(W), \widehat{B}_1(W)) \in \mathcal{U}^2$, there exists functions $h_i : \mathbb{R}^{4\times 4} \to \mathbb{R}$ such that $\mathcal{R}^{\mathrm{adv}}(X) \overset{\mathrm{def}}{=} p_1 h_1(\widehat{B}_1) + p_2 h_2(\widehat{B}_2)$ is differentiable and satisfies*

$$\nabla_X \mathcal{R}^{\mathrm{adv}}(X) = p_1 \mathbb{E}_{v,\epsilon}[\nabla_X \Psi_1(X; (\sqrt{K_1} v + \mu_1), \epsilon)] + p_2 \mathbb{E}_{v,\epsilon}[\nabla_X \Psi_2(X; (\sqrt{K_2} v + \mu_2), \epsilon)].$$

Furthermore, h_1 and h_2 are continuously differentiable on $\mathcal{U}$ and their derivatives ∇h_1 and ∇h_2 are α-pseudo-Lipschitz on $\mathcal{U}$ with constant $L(h) > 0$.

We also place an assumption requiring the well-behavior of the Fisher information matrices of the gradients as they controls the size of the stochastic gradient noise.

Assumption 7. *For $i = 1, 2$, let $\bar{s}_i = \operatorname{argmax}_{|s|\leq 1} f_i(X^\top((\sqrt{K_i} v + \mu_i)) + \delta s \|X\|, \epsilon)$ such that*

$$\mathcal{I}_i(\widehat{B}_i) \overset{\mathrm{def}}{=} \mathbb{E}_{v,\epsilon}[(f_i'(X^\top((\sqrt{K_i} v + \mu_i)) + \bar{s}_i \delta \|X\|; \epsilon))^2],$$

where the functions $\mathcal{I}_i : \mathbb{R}^{4\times 4} \to \mathbb{R}$ are α-pseudo-Lipschitz on $\mathcal{U}$ with constant $L(\mathcal{I}) > 0$.

A large class of regression problems fit these assumptions, notably phase retrieval and generalized linear models such as such as least squares and binary logistic regression (see Appendix I.1 I.2 and Appendix D in Collins-Woodfin et al. (2024a) for more examples).

1.2 Algorithmic Setup

We consider *online/streaming* SGD with a deterministic learning rate γ_k given by the update

$$X_{k+1} = X_k - \frac{\gamma_k}{d}\left(f'_{I_{k+1}}(g_{k,I_{k+1}}; \epsilon_{k+1})(a_{k+1,I_{k+1}} + \delta s_{k,I_{k+1}} \frac{X_k}{\|X_k\|}) + \lambda X_k\right) \tag{5}$$

where $a_k = \sqrt{K_{I_k}} v_k + \mu_{I_k}$ for $v_k \sim \mathcal{N}(0, I_d)$, $I_k \in \{1, 2\}$ is randomly sampled and we define $r_{k,I_{k+1}} = \hat{X}_k^\top a_{k+1,I_{k+1}} \in \mathbb{R}^2$ such that

$$\begin{aligned} g_{k,I_{k+1}} &= r_{k,I_{k+1}} + [\delta s_{k,I_{k+1}} \|X_k\| \quad 0]^\top, \\ s_{k,I_{k+1}} &= \operatorname{argmax}_{|s|\leq 1} f_{I_{k+1}}(X_k^\top a_{k+1,I_{k+1}} + \delta s \|X_k\|; \epsilon_{k+1}). \end{aligned} \tag{6}$$

Note that the initialization X_0 and $X^\star$ may be random. In this case, the following results are stated while conditioning on their values. Finally, we place an assumption on the learning rate schedule.

Assumption 8 (Bounded learning rate)**.** *There is a $\bar{\gamma} < \infty$ and a continuous function $\gamma : \mathbb{R}^+ \to \mathbb{R}^+$ which is bounded by $\bar{\gamma}$ such that $\gamma_k = \gamma(k/d)$.*

2 Deterministic dynamics for SGD under ℓ_2-adversarial training

Intuition behind the det. equivs. By Assumptions 6 and 7, $\mathcal{R}^{\mathrm{Adv}}(X)$ and $\mathcal{I}_i$ are determined by $\widehat{B}_i(W)$, and statistics such as $\|X - X^{\star,\mathrm{Adv}}\|^2$ by $X^{\star,\mathrm{Adv}}$ as well. All of these quantities are entries of $\widehat{W}^\top q(K_1, K_2)\widehat{W}$ for polynomials q and Cauchy's integral formula reduces this to a single object, the resolvent $R(z_i; K_i) \overset{\mathrm{def}}{=} (K_i - z_i I_d)^{-1}$ with $z_i \in \mathbb{C} \setminus \operatorname{spec}(K_i)$. Indeed, for any contour Γ enclosing the eigenvalues of K_1 and K_2 with a fixed positive margin, we have

$$\widehat{W}^\top q(K_1, K_2)\widehat{W} = \oint q(z)\widehat{W}^\top \mathcal{R}(\mathcal{K}; z)\widehat{W}\, \mathrm{D}z, \text{ and define } \oint f(z)\, \mathrm{D}z \overset{\mathrm{def}}{=} \frac{-1}{4\pi^2} \oint_{\Gamma^2} f(z)\, \mathrm{d}z_1\, \mathrm{d}z_2,$$

for $z = (z_1, z_2)$, $f : \mathbb{C}^2 \to \mathbb{C}$ and $\mathcal{R}(z; \mathcal{K}) \overset{\mathrm{def}}{=} R(z_1; K_1)R(z_2; K_2)$. Hence the deterministic dynamics of these quantities follow from those of

$$Z^{\mathrm{Adv}}(\widehat{W}, z) \overset{\mathrm{def}}{=} \widehat{W}^\top \mathcal{R}(z; \mathcal{K})\widehat{W}. \tag{7}$$

Since SGD is a discrete time process, we introduce the following time parametrization to compare it to its continuous time limit where $\lfloor \cdot \rfloor$ denotes the floor function:

$$k \text{ iterations of SGD} = \lfloor td \rfloor, \quad \text{for } t \in \mathbb{R}^+ \text{ the continuous time parameter.}$$

High-dimensional diffusion approximation. We also consider a continuous time counterpart of SGD, *adversarial homogenized SGD* (AdvHSGD), defined by the stochastic differential equation

$$\mathrm{d}\mathfrak{X}_t = -\gamma(t)\nabla \mathcal{R}_\lambda^{\mathrm{adv}}(\mathfrak{X}_t)\, \mathrm{d}t + \gamma(t)\sqrt{\frac{1}{d}\sum_{i=1}^{2} p_i \mathcal{I}_i(\widehat{B}(\mathcal{W}_t))(K_i + \mu_i \mu_i^\top)}\, \mathrm{d}B_t, \ \mathfrak{X}_0 = X_0, \tag{8}$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion in $\mathbb{R}^d$ and $\widehat{\mathcal{W}}_t$ is defined analogously to $\widehat{W}$. AdvHSGD matches the limiting dynamics of SGD rather than its moments: by Assumption 6, its drift is the expectation of a SGD update, while its diffusion coefficient is not the covariance of the SGD increments but differs from it only by terms whose effect on the dynamics vanishes as $d \to \infty$.

Deterministic limit dynamics. We now make the notion of det. equiv. rigorous (see Theorem 2.1). Let $\Gamma \overset{\mathrm{def}}{=} \{w \in \mathbb{C} : |w| = \max(1, 2\|K_2\|_{\mathrm{op}}, 2\|K_1\|_{\mathrm{op}})\}$. Then for any $z \in \Gamma^2$, let $\mathcal{Z}^{\mathrm{Adv}}(t, z) \in \mathbb{C}^{5\times 5}$ denote the solution to the integro-differential equation

$$\mathrm{d}\mathcal{Z}^{\mathrm{Adv}}(t, \cdot) = \mathcal{F}^{\mathrm{Adv}}(z, \mathcal{Z}^{\mathrm{Adv}}(t, \cdot))\, \mathrm{d}t, \text{ with initial condition } \mathcal{Z}^{\mathrm{Adv}}(0, z) = Z^{\mathrm{Adv}}(\widehat{W}_0, z), \tag{9}$$

where $\mathcal{F}^{\mathrm{Adv}}$ is defined in (32) and involves contour integrals over Γ.[3] Each entry of $\mathcal{Z}^{\mathrm{Adv}}(t, z)$ is the det. equiv. of the corresponding entry of $Z^{\mathrm{Adv}}(\widehat{W}, z)$, and only the entries involving X evolve in time. We denote by $\mathcal{S}(t, z)$ and $\mathcal{M}(t, z)$ the blocks of $\mathcal{Z}^{\mathrm{Adv}}(t, z)$ corresponding to

$$\hat{X}^\top \mathcal{R}(z; \mathcal{K})\hat{X} \quad \text{and} \quad \hat{X}^\top \mathcal{R}(z; \mathcal{K})\mu$$

[3] For numerical simulations of (9), we use the system of ODEs (26) in Appendix B from which we can recover a solution $\mathcal{Z}^{\mathrm{Adv}}(t, z)$ of (9). See Remark B.2 for more details.

and by $\mathcal{M}_i(t,z)$ the ith column of $\mathcal{M}(t,z)$. The det. equiv. $\widehat{\mathscr{B}}_i(t)$ of $\widehat{B}_i(W)$ is then obtained from (4) by replacing $B_i(W)$, $m_i(W)$ and $\|X\|^2$ respectively by

$$\mathscr{B}_i(t) \stackrel{\text{def}}{=} \oint z_i\, \mathcal{S}(t,z)\,\mathrm{D}z, \quad \mathfrak{m}_i(t) \stackrel{\text{def}}{=} \oint \mathcal{M}_i(t,z)\,\mathrm{D}z \quad \text{and} \quad \oint \mathcal{S}_{11}(t,z)\,\mathrm{D}z.$$

The det. equivs. of $\mathcal{R}^{\mathrm{Adv}}(X_{\lfloor td\rfloor})$, $\mathcal{I}_i(\widehat{B}_i(W_{\lfloor td\rfloor}))$, $\|X_{\lfloor td\rfloor}\|^2$ and $N(W_{\lfloor td\rfloor}) \stackrel{\text{def}}{=} \|W_{\lfloor td\rfloor}\|^2$ are[4]

$$\begin{gathered}\mathscr{R}^{\mathrm{Adv}}(t) \stackrel{\text{def}}{=} p_1 h_1(\widehat{\mathscr{B}}_1(t)) + p_2 h_2(\widehat{\mathscr{B}}_2(t)), \quad \mathscr{I}_i(t) \stackrel{\text{def}}{=} \mathcal{I}_i(\widehat{\mathscr{B}}_i(t)), \quad \widehat{\mathscr{B}}_{44}(t) \\ \text{and} \quad \mathscr{N}(t) \stackrel{\text{def}}{=} \oint \mathrm{Tr}(\mathcal{Z}^{\mathrm{Adv}}(t,z))\,\mathrm{D}z - \|X^{\star,\mathrm{Adv}}\|^2.\end{gathered} \tag{10}$$

These quantities also define the stopping time in Theorem 2.1. Moreover, the statistics above belong to the following class, where for $\mathcal{R}^{\mathrm{Adv}}$ and $\mathcal{I}_i$ this holds on $\mathcal{U}^2$.

Assumption 9. *Fix $\ell \geq 1$. The statistic $\varphi : \mathbb{R}^d \to \mathbb{R}$ satisfies $\varphi(X) = g(\{\widehat{W}^\top q_j(K_1,K_2)\widehat{W}\}_{j=1}^{\ell})$, where $g : \mathbb{R}^{\ell\times 5\times 5} \to \mathbb{R}$ is α-pseudo-Lipschitz and $\{q_j\}_{j=1}^{\ell}$ are polynomials.*

For a statistic φ satisfying Assumption 9, its det. equiv. is given by

$$\phi(t) \stackrel{\text{def}}{=} g\Big(\Big\{ \oint q_j(z)\, \mathcal{Z}^{\mathrm{Adv}}(t,z)\,\mathrm{D}z \Big\}_{j=1}^{\ell}\Big).$$

For any $\omega > 0$, define $\mathcal{U}_\omega \stackrel{\text{def}}{=} \{B \in \mathcal{U} : \inf_{V\notin\mathcal{U}} \|B - V\| \geq \omega\}$. We now state our main result.

Theorem 2.1. *Suppose that the assumptions stated above hold and let φ satisfy Assumption 9. Let Θ_M be the first time $(\widehat{\mathscr{B}}_1(t), \widehat{\mathscr{B}}_2(t))$ exits $\mathcal{U}_\omega^2$ or that $\mathscr{N}(t) \geq M$. Then for any $T, M > 0$ and any $\varepsilon \in (0, \frac{1}{2})$, there exists a constant $C > 0$ (independent of d) such that w.o.p.*[5]

$$\sup_{0\leq t\leq T\wedge\Theta_M} \Big(|\varphi(X_{\lfloor td\rfloor}) - \phi(t)| + |\varphi(\mathfrak{X}_t) - \phi(t)|\Big) \leq Cd^{-\varepsilon} \quad w.o.p.$$

Theorem 2.1 implies that statistics of SGD satisfying Assumption 9 can be studied without running the algorithm, either by solving (9) or by applying Itô calculus to AdvHSGD. Properties obtained through either analysis then transfer back to SGD for sufficiently large d. Assumption 5 is only needed for statistics involving $X^{\star,\mathrm{Adv}}$. For statistics of the form $g(\{W^\top q_j(K_1,K_2)W\}_{j=1}^{\ell})$, Theorem 2.1 holds without it (see Theorem C.3). See Figure 1 for an illustration and Appendix B for a complete overview of the theoretical results in this framework.

Intuition for the proof. By Cauchy's integral formula, $\varphi(X_{\lfloor td\rfloor})$, $\varphi(\mathfrak{X}_t)$ and $\phi(t)$ are the same pseudo-Lipschitz function g of contour integrals of $Z^{\mathrm{Adv}}(\widehat{W}_{\lfloor td\rfloor},\cdot)$, $Z^{\mathrm{Adv}}(\widehat{\mathcal{W}}_t,\cdot)$ and $\mathcal{Z}^{\mathrm{Adv}}(t,\cdot)$, so it suffices to compare these. Since the entries of Z^{Adv} are quadratic, we expand $Z^{\mathrm{Adv}}(\widehat{W}_{\lfloor td\rfloor}, z)$ through a Doob decomposition and $Z^{\mathrm{Adv}}(\widehat{\mathcal{W}}_t, z)$ through Itô's formula. As K_1 and K_2 commute, both expansions close in terms of Z^{Adv} with drift $\mathcal{F}^{\mathrm{Adv}}$, up to martingale terms and, for SGD, error terms, which all vanish w.o.p. as $d \to \infty$ for each fixed $z \in \Gamma^2$. A net argument makes this uniform over Γ^2, so both processes approximately solve (9) and stability of approximate solutions gives closeness to $\mathcal{Z}^{\mathrm{Adv}}$. Finally, Θ_M keeps $\mathcal{Z}^{\mathrm{Adv}}$ bounded and $(\widehat{\mathscr{B}}_1(t), \widehat{\mathscr{B}}_2(t))$ in $\mathcal{U}^2$, where h_1 and h_2 have the regularity required by Assumption 6. The first condition can be dropped when f is α-pseudo-Lipschitz with $\alpha \leq 1$ (Proposition B.1, Corollary B.4) and the second when $\mathcal{U} = \mathbb{R}^{4\times 4}$.

Example 2.1 (Least squares)**.** We consider least squares with centered $a \sim \mathcal{N}(0, K)$ and targets $(X^\star)^\top a + \epsilon$. See Appendix I.1 for more details. The least-squares risk is given by

$$\mathcal{R}(X) = \frac{1}{2}(X - X^\star)^\top K (X - X^\star) + \frac{\eta^2}{2} \text{ and } \mathcal{R}^{\mathrm{adv}}(X) = \mathcal{R}(X) + \delta\sqrt{\tfrac{2}{\pi}}\|X\|\sqrt{2\mathcal{R}(X)} + \tfrac{\delta^2}{2}\|X\|^2.$$

Dropping the subscript i, we write $\mathscr{R}(t)$ and $\widehat{\mathscr{B}}_{44}(t)$ for the det. equivs. of $\mathcal{R}(X_{\lfloor td\rfloor})$ and $\|X_{\lfloor td\rfloor}\|^2$, from which $\mathscr{R}^{\mathrm{Adv}}(t)$ follows by substitution. Since h is not differentiable where $\mathcal{R}(X) = 0$ or $X = 0$, $\mathcal{U}$ is the set where both are positive, and there $\mathcal{I} = 2h$, so $\mathscr{I}(t) = 2\mathscr{R}^{\mathrm{Adv}}(t)$.

[4] Assumption 6 entails that the integro-differential equation (9) has locally unique solutions whilst $(\widehat{\mathscr{B}}_1(t), \widehat{\mathscr{B}}_2(t)) \in \mathcal{U}^2$ and $\mathscr{N}(t)$ does not blow up.

[5] We say an event holds *with overwhelming probability* (w.o.p.) if there exists a function $\omega : \mathbb{N} \to \mathbb{R}$ with $\omega(d)/\log d \to \infty$ such that the event holds with probability at least $1 - e^{-\omega(d)}$.

Remark 2.1. *Theorem 2.1 extends Collins-Woodfin et al. (2024a); Collins-Woodfin & Seroussi (2025) to ℓ_2-adversarial training. Since* (1) *is a min-max problem, each SGD update depends on the adversarial perturbation of the sampled data point, which scales with $\|X_k\|$ and is selected through the argmax $s_{k,I_{k+1}}$. The det. equiv. must then account for both, which leads to more martingale and error terms to control. For soft labels, $X^{\star,\mathrm{Adv}}$ and $X^\star$ need not coincide, so we also extend the theory of approximate solutions to track the overlap between X_k and $X^{\star,\mathrm{Adv}}$. These differences already appear for least squares: the det. equivs. above satisfy coupled nonlinear Volterra equations, whereas setting $\delta = 0$ recovers the linear convolution Volterra equation of Collins-Woodfin et al. (2024a), Collins-Woodfin & Paquette (2024) and Paquette et al. (2022).*

3 ℓ_2-Adversarial Training v.s. ℓ_2-Regularization

We consider ℓ_2-adversarial least squares (Example 2.1). In this setting, AdvHSGD (8) simplifies to

$$\mathrm{d}\mathcal{X}_t = -\gamma^{\mathrm{eff}}(t)\left(\nabla\mathcal{R}(\mathcal{X}_t) + \lambda^{\mathrm{eff}}(t)\mathcal{X}_t\right)\mathrm{d}t + \gamma(t)\sqrt{\tfrac{2}{d}\,\mathcal{R}^{\mathrm{Adv}}(\mathcal{X}_t)\,K}\;\mathrm{d}B_t, \quad q(X) \overset{\mathrm{def}}{=} \frac{\|X\|}{\sqrt{2\mathcal{R}(X)}},$$

where the *effective learning rate* and *effective regularization* induced by adversarial training are

$$\gamma^{\mathrm{eff}}(t) \overset{\mathrm{def}}{=} \gamma(t)\Big(1 + \delta\sqrt{\tfrac{2}{\pi}}\,q(\mathcal{X}_t)\Big) \quad \text{and} \quad \lambda^{\mathrm{eff}}(t) \overset{\mathrm{def}}{=} \frac{\delta^2 + \delta\sqrt{\frac{2}{\pi}}\,(q(\mathcal{X}_t))^{-1}}{1+\delta\sqrt{\frac{2}{\pi}}\,q(\mathcal{X}_t)}.$$

Homogenized SGD (HSGD) (Collins-Woodfin & Paquette, 2024; Paquette et al., 2022; Collins-Woodfin et al., 2024a) is the analogue of AdvHSGD for SGD on ℓ_2-regularized least squares with a fixed parameter $\lambda^{\mathrm{Reg}} > 0$. Let $\gamma^{\mathrm{Reg}}(t)$ and $\lambda^{\mathrm{Reg}}(t)$ be defined as $\gamma^{\mathrm{eff}}(t)$ and $\lambda^{\mathrm{eff}}(t)$ with $\mathcal{X}_t$ replaced by $\mathcal{X}_t^{\mathrm{Reg}}$. HSGD with learning rate $\gamma^{\mathrm{Reg}}(t)$ and regularization $\lambda^{\mathrm{Reg}}(t)$ is

$$\begin{aligned}\mathrm{d}\mathcal{X}_t^{\mathrm{Reg}} = &-\gamma^{\mathrm{Reg}}(t)\left(\nabla\mathcal{R}(\mathcal{X}_t^{\mathrm{Reg}}) + \lambda^{\mathrm{Reg}}(t)\mathcal{X}_t^{\mathrm{Reg}}\right)\mathrm{d}t \\ &+ \gamma(t)\sqrt{\tfrac{2}{d}\left(\mathcal{R}^{\mathrm{Adv}}(\mathcal{X}_t^{\mathrm{Reg}}) - \delta^2(\tfrac{1}{2} - \tfrac{1}{\pi})\|\mathcal{X}_t^{\mathrm{Reg}}\|^2\right)K}\,\mathrm{d}B_t.\end{aligned} \tag{11}$$

The drifts of (3) and (11) coincide and their diffusion coefficients differ only in the coefficient of $\delta^2\|\cdot\|^2$. The next theorem shows this discrepancy has a dimension-free effect on the dynamics.

Theorem 3.1. *Let X_k^{Reg} be SGD on ℓ_2-regularized least squares with adaptive γ_k^{Reg} and λ_k^{Reg} given by the functions defining γ^{Reg}, λ^{Reg} evaluated at X_k^{Reg}. Let $X_0^{\mathrm{Reg}} = X_0$. For $\omega, M > 0$, let Θ be the first time the det. equiv. of $\mathcal{R}$ or of $\|\cdot\|^2$ along either algorithm leaves $[\omega, M]$. Then for any $T > 0$ and $\varepsilon \in (0, \frac{1}{2})$ there exists $C = C(\|K\|_{\mathrm{op}}, \bar{\gamma}, T, \omega, M)$, independent of d, such that w.o.p.*

$$\sup_{0\le t\le T\wedge\Theta}\left(|\mathcal{R}(X_{\lfloor td\rfloor}) - \mathcal{R}(X_{\lfloor td\rfloor}^{\mathrm{Reg}})| + |\mathcal{R}^{\mathrm{Adv}}(X_{\lfloor td\rfloor}) - \mathcal{R}^{\mathrm{Adv}}(X_{\lfloor td\rfloor}^{\mathrm{Reg}})|\right) \le C(\delta^2\left(\tfrac{1}{2} - \tfrac{1}{\pi}\right) + d^{-\varepsilon}).$$

Remark 3.1. *Theorem 2.1 remains valid if Assumption 8 is replaced by the analogue of Collins-Woodfin et al. (2024b, Assumption 6), under which γ_k is an α-pseudo-Lipschitz function of the history of the iterates, gradients and adversarial risk, or of any statistic satisfying Assumption 9, with the pseudo-Lipschitz condition imposed on $\mathcal{U}$ only (they impose it globally); the same holds for an adaptive regularization λ_k. Combining our proof with that of Collins-Woodfin et al. (2024b, Theorem 2.1), the det. equivs. of $\gamma_{\lfloor td\rfloor}$ and $\lambda_{\lfloor td\rfloor}$ are functions of $\{\mathcal{Z}^{\mathrm{Adv}}(s,\cdot)\}_{0\le s\le t}$ up to Θ_M. Hence the det. equivs. computed with $\gamma^{\mathrm{Reg}}(t)$ and $\lambda^{\mathrm{Reg}}(t)$, or with $\gamma^{\mathrm{Polyak,Adv}}(t)$ and $\gamma^{\mathrm{line}}(t)$ of Section 4, remain valid for SGD, as shown in Figures 3 and 4. Figure 3 (left) confirms Theorem 3.1.*

Limit $t \to \infty$. Let $\gamma(t) \equiv \gamma$ and $\eta = 0$. The det. equiv. $q(t) \overset{\mathrm{def}}{=} \sqrt{\widehat{\mathscr{B}}_{44}(t)/(2\mathscr{R}(t))}$ of $q(X_{\lfloor td\rfloor})$ determines the det. equivs. $\tilde{\gamma}^{\mathrm{eff}}(t)$ and $\tilde{\lambda}^{\mathrm{eff}}(t)$ of the effective learning rate and regularization, and hence the dynamics. It solves a nonlinear ODE coupled with $\mathscr{R}(t)$ (Appendix G). Figures 1 (right) and 6 give numerical evidence that $q(t)$ converges, in which case its limit solves a fixed point equation from which the limits of the det. equivs. are recovered. We state a simplified version of the following Theorem given its lengthiness. See Theorem G.1 for more details.

Theorem 3.2. *Suppose $\delta > 0$, $\mathscr{R}(0), \widehat{\mathscr{B}}_{44}(0) > 0$, $(X^\star)^\top K X^\star > 0$, $\sup_{t\geq 0} \mathscr{R}^{\mathrm{Adv}}(t) < \infty$, $q(t) \to q_\infty \in (0, \infty)$ and that δ satisfies the condition of Theorem G.1. Then q_∞ is the unique admissible solution a fixed point equation, and the limits as $t \to \infty$ of $\mathscr{R}$, $\widehat{\mathscr{B}}_{44}$, $\mathscr{R}^{\mathrm{Adv}}$ and $\rho_j(t)$, the det. equiv. of $\langle X_{\lfloor td \rfloor}, \omega_j \rangle$, are explicit functions of q_∞. Moreover, $\tilde{\lambda}^{\mathrm{eff}}(q_\infty) \stackrel{\mathrm{def}}{=} \lim_{t\to\infty} \tilde{\lambda}^{\mathrm{eff}}(t)$,*

$$\rho_j(t) \underset{t\to\infty}{\to} \frac{\lambda_j}{\lambda_j + \tilde{\lambda}^{\mathrm{eff}}(q_\infty)} x_j^\star, \quad \textit{i.e.} \quad X_\infty \stackrel{\mathrm{def}}{=} \sum_{j=1}^{d} \rho_j(\infty)\, \omega_j = \left(K + \tilde{\lambda}^{\mathrm{eff}}(q_\infty) I_d\right)^{-1} K X^\star$$

is the solution to ridge regression with parameter $\tilde{\lambda}^{\mathrm{eff}}(q_\infty)$. Here ω_j is an eigenvector of K.

Even for $K = I_d$, the fixed point equation is extremely hard to interpret, so we let $\gamma \to 0$.

Theorem 3.3. *Under the assumptions of Theorem 3.2 with K positive definite, there exist thresholds $0 < \delta_- < \delta_+$, depending only on $X^\star$ and the eigenpairs of K, such that as $\gamma \to 0$, $\tilde{\lambda}^{\mathrm{eff}}(q_\infty) \to \tilde{\lambda}_\star$ and $X_\infty \to X^{\star,\mathrm{Adv}} = (K + \tilde{\lambda}_\star I_d)^{-1} K X^\star$, where $\tilde{\lambda}_\star = 0$ and $X^{\star,\mathrm{Adv}} = X^\star$ if $\delta \leq \delta_-$, $\tilde{\lambda}_\star \in (0, \infty)$ if $\delta_- < \delta < \delta_+$, and $\tilde{\lambda}_\star = \infty$ and $X^{\star,\mathrm{Adv}} = 0$ if $\delta \geq \delta_+$.*

Theorem 3.3 recovers $X^{\star,\mathrm{Adv}}$ in Proposition 1 of Xing et al. (2021). The thresholds, the fixed point equation solved by $\tilde{\lambda}_\star$ and the case of singular K are given in Theorem G.2. Thus δ interpolates between minimizing $\mathcal{R}$ and $\|\cdot\|^2$, and for finite γ the effective regularization $\tilde{\lambda}^{\mathrm{eff}}(q_\infty)$ of SGD differs from $\tilde{\lambda}_\star$. Xing et al. (2021) identify the minimizer of the population ℓ_2-adversarial least squares risk with a ridge solution. Theorem 3.1 extends this along the SGD trajectory, with a ridge parameter $\lambda^{\mathrm{eff}}(t)$ adapting through $q(X_k)$, and Theorem 3.2 at its endpoint. Both rest on $\mathcal{R}^{\mathrm{Adv}}$ being a function of $\mathcal{R}$ and $\|X\|$ alone, which already fails at the level of the objective for two-layer networks (Robin et al., 2026). Whether the dynamics of nonlinear models admit such a description is left for future work.

4 Idealized Exact Line Search and Polyak Stepsize

Given the sensitivity of adversarial training to the learning rate schedule (Rice et al., 2020; Pang et al., 2021; Dabouei et al., 2022), we study two classical idealized strategies: *exact line search*, which selects at each iteration the learning rate producing the largest decrease of the objective, and the *Polyak stepsize* (Polyak, 1987), which does so for the distance to optimality. Throughout this section, we take $\lambda = 0$. Following Collins-Woodfin et al. (2024b), we impose these requirements on the det. equivs.:

$$\gamma^{\mathrm{line}}(t) \in \mathrm{argmin}_\gamma\, \mathrm{d}\mathscr{R}^{\mathrm{Adv}}(t) \quad \text{and} \quad \gamma^{\mathrm{Polyak,Adv}}(t) \in \mathrm{argmin}_\gamma\, \mathrm{d}\mathscr{D}^{2,\mathrm{Adv}}(t), \tag{12}$$

where $\mathscr{D}^{2,\mathrm{Adv}}(t)$ is the det. equiv. of $\|X_{\lfloor td \rfloor} - X^{\star,\mathrm{Adv}}\|^2$. Although neither algorithm is implementable, as both require distributional knowledge of a, $X^\star$ and $X^{\star,\mathrm{Adv}}$, they provide a benchmark for adaptive schedules and reveal how the adversarial perturbation should enter them. See Remark 3.1 and Figure 4 in Appendix F for why Theorem 2.1 still applies to SGD run with these schedules. See Collins-Woodfin et al. (2024b) for an overview of their stochastic counterparts (Vaswani et al., 2019; Loizou et al., 2021).

Polyak stepsize. Define the *stability threshold* $\gamma^{\mathrm{Stable,Adv}}(t)$ as the largest learning rate for which $\mathrm{d}\mathscr{D}^{2,\mathrm{Adv}}(t) < 0$ then $\gamma^{\mathrm{Polyak,Adv}}(t) = \frac{1}{2}\gamma^{\mathrm{Stable,Adv}}(t)$. Translating this back to SGD using Theorem 2.1 and Remark 3.1, the threshold is the ratio of $\langle X_k - X^{\star,\mathrm{Adv}}, \nabla\mathcal{R}^{\mathrm{Adv}}(X_k)\rangle$ to the stochastic gradient noise and decomposes into the standard descent term towards $X^\star$, a correction for the shift of the minimizer to $X^{\star,\mathrm{Adv}}$, and a term due to the adversarial perturbation which vanishes when $\delta = 0$.

This hints that a practical adaptive schedule for adversarial training should thus account for both the shift of the objective and the perturbation itself. Moreover, in contrast with noiseless standard least squares, whose threshold is the constant $2d/\operatorname{Tr}(K)$ (Loizou et al., 2021; Paquette et al., 2021), the stability threshold $\gamma_k^{\mathrm{Stable,adv}}$ of ℓ_2-adversarial least-squares (see Example 2.1) vanishes as $X_k \to X^{\star,\mathrm{Adv}}$ because adversarial training does not interpolate. Hence, no constant learning rate guarantees monotone descent towards $X^{\star,\mathrm{Adv}}$. See Appendix F.1 for the general expressions.

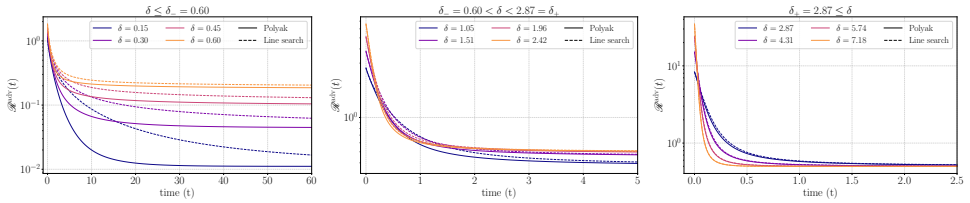

Figure 2: **Comparison for Exact Line Search and Polyak Stepsize under strong anisotropy** on noiseless ℓ_2-adversarial least squares for different values of δ in the three regimes of $X^{\star,\mathrm{Adv}}$ (See Theorem 3.3). The three plots illustrate the convergence of the ℓ_2-adversarial risk and how δ and $\tilde{\lambda}^{\mathrm{eff}}(t)$ mitigate the influence of strong anisotropy on the discrepancy between Polyak and exact line search. See Figure 5 for the case of weak anisotropy and Section J for simulation details.

Exact line search. Likewise, γ_k^{line} is half the largest learning rate for which $\mathrm{d}\mathscr{R}^{\mathrm{Adv}}(t) < 0$ (Appendix F.2). For ℓ_2-adversarial least squares (Example 2.1), Propositions F.1 and F.2 express both learning rates through ridge objectives and, at a given state of the ODEs, their ratio factorizes as

$$\frac{\gamma^{\mathrm{line}}(t)}{\gamma^{\mathrm{Polyak,Adv}}(t)} = \underbrace{\frac{\mathrm{Tr}(K)\big(\sum_{j=1}^d \alpha_j(t)\lambda_j + \tilde{\lambda}^{\mathrm{eff}}(t)\big)}{\mathrm{Tr}(K^2) + \tilde{\lambda}^{\mathrm{eff}}(t)\,\mathrm{Tr}(K)}}_{\text{spectral factor}} \cdot \underbrace{\frac{\Phi_{\tilde{\lambda}^{\mathrm{eff}}(t)}(t) - \Phi_{\min}(\tilde{\lambda}^{\mathrm{eff}}(t))}{\Phi_{\tilde{\lambda}^{\mathrm{Polyak}}(t)}(t) - \Phi_{\min}(\tilde{\lambda}_\star)}}_{\text{ridge factor}}, \tag{13}$$

where $\mathcal{R}$ is the least squares risk of Example 2.1, $\Phi_{\tilde{\lambda}}(t) \overset{\mathrm{def}}{=} \mathscr{R}(t) + \frac{\tilde{\lambda}}{2}\widehat{\mathscr{B}}_{44}(t)$ is the det. equiv. of the ridge objective $\mathcal{R}(X_{\lfloor td \rfloor}) + \frac{\tilde{\lambda}}{2}\|X_{\lfloor td \rfloor}\|^2$ and $\Phi_{\min}(\tilde{\lambda}) \overset{\mathrm{def}}{=} \min_{X\in\mathbb{R}^d}\left\{\mathcal{R}(X) + \frac{\tilde{\lambda}}{2}\|X\|^2\right\}$ is its minimum value. Here $\alpha_j(t) \ge 0$ with $\sum_{j=1}^d \alpha_j(t) = 1$ is the share of the excess $\mathcal{E}(t) \overset{\mathrm{def}}{=} \Phi_{\tilde{\lambda}^{\mathrm{eff}}(t)}(t) - \Phi_{\min}(\tilde{\lambda}^{\mathrm{eff}}(t))$ carried by the j-th eigendirection of K and

$\tilde{\lambda}^{\mathrm{Polyak}}(t) \overset{\mathrm{def}}{=} \tilde{\lambda}^{\mathrm{eff}}(t) + \big(\tilde{\lambda}_\star - \tilde{\lambda}^{\mathrm{eff}}(t)\big)\frac{\mathscr{A}^{\mathrm{Adv}}(t)}{\widehat{\mathscr{B}}_{44}(t)}$, with $\mathscr{A}^{\mathrm{Adv}}(t)$ the det. equiv. of $\langle X_{\lfloor td \rfloor}, X^{\star,\mathrm{Adv}}\rangle$,

where $\tilde{\lambda}_\star = \lim_{\gamma\to 0}\lim_{t\to\infty}\tilde{\lambda}^{\mathrm{eff}}(t)$ (Theorem 3.3). The ridge factor measures the mismatch between the ridge problem SGD currently follows and the one solved by $X^{\star,\mathrm{Adv}}$; it equals one when $\tilde{\lambda}^{\mathrm{eff}}(t) = \tilde{\lambda}_\star$.The spectral factor equals one if and only if $\sum_{j=1}^d \alpha_j(t)\lambda_j = \mathrm{Tr}(K^2)/\mathrm{Tr}(K)$. Under strong anisotropy (that is, $\mathrm{Tr}(K^2)/d \gg \mathrm{Tr}(K)/d$), it is far below one once the top eigendirections of K have been learned and the remaining excess $\mathcal{E}(t)$ is concentrated in the tail of the spectrum, and exact line search is then much more conservative than the Polyak stepsize, as in least squares (Collins-Woodfin et al., 2024b). Theorem 3.3 determines how much δ damps this discrepancy:

- $\delta \le \delta_-$: $\tilde{\lambda}^{\mathrm{eff}}(t) \to \tilde{\lambda}_\star = 0$ and anisotropy affects exact line search as in standard least squares;
- $\delta_- < \delta < \delta_+$: $\tilde{\lambda}^{\mathrm{eff}}(t) \to \tilde{\lambda}_\star \in (0, \infty)$ and the discrepancy is partially damped;
- $\delta \ge \delta_+$: $\tilde{\lambda}^{\mathrm{eff}}(t) \to \tilde{\lambda}_\star = \infty$, the spectral factor approaches one and only the ridge factor remains.

Figures 2 and 5 illustrate the three regimes under strong and weak anisotropy: larger δ mitigates the poorer performance of exact line search relative to the Polyak stepsize.

5 Limitations and Future Work

Our framework is limited to traces of order d and, as noted in Collins-Woodfin et al. (2024b), smaller traces call for different learning rate scalings. Extending it to multi-pass SGD, multi-index models with $\mathcal{O}(\log(d))$ number of classes similarly to Collins-Woodfin & Seroussi (2025) appears within reach. Beyond these extensions, we see two promising directions. First, we assumed that adversarial examples are computed exactly. Our framework could be adapted to describe the training dynamics when they are generated by approximate attacks such as the Fast Gradient Sign Method (Goodfellow et al., 2015), and to measure how far these dynamics are from the exact ones. Second, our decomposition of the Polyak stepsize hints that an adaptive learning rate schedule for adversarial training must account for both the shift of the objective and the adversarial perturbation. Designing a practical schedule on this principle and testing it on real data is a natural next step.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing, for proofreading, and to help search for and identify relevant literature. We have not used generative AI tools for any other task requiring or recommending disclosure. We have reviewed all AI-assisted work, including verifying every suggested reference against the original source, and take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

All assumptions are stated in Section 1.1 and complete proofs of every result are given in the appendix, whose organization is described in the outline at the start of the appendix. We also state the limitations of our framework in Section 5. All information needed to reproduce the experiments, including model parameters, data covariances, initializations and the number of runs for each figure, is provided in Appendix J. The simulations use synthetic Gaussian data and standard algorithms. We also plan to release the code on GitHub with instructions to reproduce the results.

REFERENCES

Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional and mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks. *arXiv preprint arXiv:2302.05882*, 2023.

Luca Arnaboldi, Florent Krzakala, Bruno Loureiro, and Ludovic Stephan. Escaping mediocrity: how two-layer networks learn hard generalized linear models with sgd, 2024. URL https://arxiv.org/abs/2305.18502.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for sgd: Effective dynamics and critical scaling, 2023. URL https://arxiv.org/abs/2206.04030.

Gerard Ben Arous, Reza Gheissari, Jiaoyang Huang, and Aukosh Jagannath. Local geometry of high-dimensional mixture models: Effective spectral theory and dynamical transitions, 2026. URL https://arxiv.org/abs/2502.15655.

Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 4312–4321. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/591. URL https://doi.org/10.24963/ijcai.2021/591. Survey Track.

Krishnakumar Balasubramanian, Promit Ghosal, and Ye He. High-dimensional scaling limits and fluctuations of online least-squares sgd with smooth covariance, 2024. URL https://arxiv.org/abs/2304.00707.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25349–25362, New York, 2022. Curran Associates, Inc.

Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport, 2019. URL https://arxiv.org/abs/1909.12272.

Michael Biehl and Peter Riegler. On-line learning with a perceptron. *Europhysics Letters*, 28(7): 525, 1994.

Michael Biehl and Holm Schwarze. Learning by on-line gradient descent. *Journal of Physics A: Mathematical and general*, 28(3):643, 1995.

Blake Bordelon and Cengiz Pehlevan. Learning curves for sgd on structured features, 2022. URL https://arxiv.org/abs/2106.02713.

Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data, 2026. URL https://arxiv.org/abs/2112.07572.

Kabir Aladin Chandrasekher, Ashwin Pananjady, and Christos Thrampoulidis. Sharp global convergence guarantees for iterative nonconvex optimization with random data. *Ann. Statist.*, 51 (1):179–210, 2023. ISSN 0090-5364,2168-8966. doi: 10.1214/22-aos2246. URL https://doi.org/10.1214/22-aos2246.

Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qZzy5urZw9.

Jacob Clarysse, Julia Hörrmann, and Fanny Yang. Why adversarial training can hurt robust accuracy, 2022. URL https://arxiv.org/abs/2203.02006.

Elizabeth Collins-Woodfin and Elliot Paquette. High-dimensional limit of one-pass SGD on least squares. *Electronic Communications in Probability*, 29:1–15, 2024. doi: 10.1214/23-ECP571.

Elizabeth Collins-Woodfin and Inbar Seroussi. Exact dynamics of multi-class stochastic gradient descent, 2025. URL https://arxiv.org/abs/2510.14074.

Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: an ode for sgd learning dynamics on glms and multi-index models. *Information and Inference: A Journal of the IMA*, 13(4):iaae028, 12 2024a. ISSN 2049-8772. doi: 10.1093/imaiai/iaae028. URL https://doi.org/10.1093/imaiai/iaae028.

Elizabeth Collins-Woodfin, Inbar Seroussi, Begoña García Malaxechebarría, Andrew Mackenzie, Elliot Paquette, and Courtney Paquette. The high line: Exact risk and learning rate curves of stochastic adaptive learning rate algorithms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=4VWnC5unAV.

Ali Dabouei, Fariborz Taherkhani, Sobhan Soleymani, and Nasser M. Nasrabadi. Revisiting outer optimization in adversarial training, 2022. URL https://arxiv.org/abs/2209.01199.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D. Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models, 2023. URL https://arxiv.org/abs/2305.10633.

Chen Dan, Yuting Wei, and Pradeep Ravikumar. Sharp statistical guarantees for adversarially robust gaussian classification, 2020. URL https://arxiv.org/abs/2006.16384.

Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborová, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. *arXiv preprint arXiv:2402.03220*, 2024.

John M. Danskin. The theory of max-min and its application to weapons allocation problems. 1967. URL https://api.semanticscholar.org/CorpusID:122915464.

Edgar Dobriban, Hamed Hassani, David Hong, and Alexander Robey. Provable tradeoffs in adversarially robust classification, 2022. URL https://arxiv.org/abs/2006.05161.

Elvis Dohmatob and Meyer Scetbon. Precise accuracy / robustness tradeoffs in regression: Case of general norms. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11198–11226. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/dohmatob24c.html.

Zhou Fan and Leda Wang. High-dimensional learning dynamics of multi-pass stochastic gradient descent in multi-index models, 2026. URL https://arxiv.org/abs/2601.21093.

Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers' robustness to adversarial perturbations. *Mach. Learn.*, 107(3):481–508, March 2018. ISSN 0885-6125. doi: 10.1007/s10994-017-5663-3. URL https://doi.org/10.1007/s10994-017-5663-3.

Cédric Gerbelot, Emanuele Troiani, Francesca Mignacco, Florent Krzakala, and Lenka Zdeborová. Rigorous dynamical mean-field theory for stochastic gradient descent methods. *SIAM Journal on Mathematics of Data Science*, 6(2):400–427, 2024. doi: 10.1137/23M1594388. URL https://doi.org/10.1137/23M1594388.

Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in neural information processing systems*, 32, 2019.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044, 2020.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pp. 426–471, New York, New York, USA, 2022. PMLR.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Haotian Gu, Xin Guo, and Xinyu Li. Adversarial training for gradient descent: Analysis through its continuous-time approximation, 2023. URL https://arxiv.org/abs/2105.08037.

Hamed Hassani and Adel Javanmard. The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression, 2024. URL https://arxiv.org/abs/2201.05149.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf.

Adel Javanmard and Mahdi Soltanolkotabi. Precise statistical analysis of classification accuracies for adversarial training, 2022. URL https://arxiv.org/abs/2010.11213.

Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression, 2020. URL https://arxiv.org/abs/2002.10477.

Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, Alan Yuille, Sangxia Huang, Yao Zhao, Yuzhe Zhao, Zhonglin Han, Junjiajia Long, Yerkebulan Berdibekov, Takuya Akiba, Seiya Tokui, and Motoki Abe. Adversarial attacks and defences competition. In Sergio Escalera and Markus Weimer (eds.), *The NIPS '17 Competition: Building Intelligent Systems*, pp. 195–231, Cham, 2018. Springer International Publishing. ISBN 978-3-319-94042-7.

Kiwon Lee, Andrew N. Cheng, Courtney Paquette, and Elliot Paquette. Trajectory of Mini-Batch Momentum: Batch Size Saturation and Convergence in High Dimensions. *To Appear in NeurIPS 2022*, art. arXiv:2206.01029, June 2022.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations, 2018. URL https://arxiv.org/abs/1811.01558.

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

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Begoña García Malaxechebarría, Courtney Paquette, Maryam Fazel, and Dmitriy Drusvyatskiy. High-dimensional limit of sgd for diagonal linear networks, 2026. URL https://arxiv.org/abs/2605.17177.

Noah Marshall, Ke Liang Xiao, Atish Agarwala, and Elliot Paquette. To clip or not to clip: the dynamics of sgd with gradient clipping in high-dimensions, 2024. URL https://arxiv.org/abs/2406.11733.

Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124008, December 2021. ISSN 1742-5468. doi: 10.1088/1742-5468/ac3a80. URL http://dx.doi.org/10.1088/1742-5468/ac3a80.

Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training, 2021. URL https://arxiv.org/abs/2010.00467.

C. Paquette, K. Lee, F. Pedregosa, and E. Paquette. SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality. In *Proceedings of Thirty Fourth Conference on Learning Theory (COLT)*, volume 134, pp. 3548–3626, 2021.

Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Homogenization of SGD in high-dimensions: Exact dynamics and generalization properties. *arXiv e-prints*, art. arXiv:2205.07069, May 2022.

Boris T Polyak. Introduction to optimization. 1987.

P.E. Protter. *Stochastic integration and differential equations*, volume 21 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin, 2005. doi: 10.1007/978-3-662-10061-5. URL https://doi.org/10.1007/978-3-662-10061-5.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy, 2020. URL https://arxiv.org/abs/2002.10716.

Maria Refinetti, Sebastian Goldt, Florent Krzakala, and Lenka Zdeborová. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed, 2021. URL https://arxiv.org/abs/2102.11742.

Antonio H. Ribeiro, Thomas B. Schön, Dave Zachariah, and Francis Bach. Efficient optimization algorithms for linear adversarial training. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1207–1215. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/ribeiro25a.html.

Antônio H. Ribeiro and Thomas B. Schön. Overparameterized linear regression under adversarial attacks. *IEEE Transactions on Signal Processing*, 71:601–614, 2023. doi: 10.1109/TSP.2023.3246228.

Antônio H. Ribeiro, Dave Zachariah, Francis Bach, and Thomas B. Schön. Regularization properties of adversarially-trained linear regression, 2023. URL https://arxiv.org/abs/2310.10807.

Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning, 2020. URL https://arxiv.org/abs/2002.11569.

David A. R. Robin, Rafael Pinot, and Yann Chevaleyre. Robustness cannot be reduced to regularization: Studying adversarial training beyond the linear case, 2026. URL https://arxiv.org/abs/2606.21488.

David Saad and Sara Solla. Dynamics of on-line gradient descent learning for multilayer neural networks. In *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995a.

David Saad and Sara A Solla. Exact solution for on-line learning in multilayer neural networks. *Physical Review Letters*, 74(21):4337, 1995b.

Meyer Scetbon and Elvis Dohmatob. Robust linear regression: Gradient-descent, early-stopping, and beyond, 2023. URL https://arxiv.org/abs/2301.13486.

Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free!, 2019. URL https://arxiv.org/abs/1904.12843.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, January 2014.

Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Asymptotic behavior of adversarial training in binary classification, 2021. URL https://arxiv.org/abs/2010.13275.

Kasimir Tanner, Matteo Vilucchio, Bruno Loureiro, and Florent Krzakala. A high dimensional statistical model for adversarial training: Geometry and trade-offs, 2024. URL https://arxiv.org/abs/2402.05674.

S. Vaswani, A. Mishkin, I. Laradji, M. Schmidt, G. Gidel, and S. Lacoste-Julien. Painless Stochastic Gradient: Interpolation, Line-Search, and Convergence Rates. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 3732–3745, 2019.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, Cambridge, UK, 2018. doi: 10.1017/9781108231596. URL https://doi.org/10.1017/9781108231596.

Matteo Vilucchio, Nikolaos Tsilivis, Bruno Loureiro, and Julia Kempe. On the geometry of regularization in adversarial training: High-dimensional asymptotics and generalization bounds, 2024. URL https://arxiv.org/abs/2410.16073.

Matteo Vilucchio, Lenka Zdeborová, and Bruno Loureiro. On the existence of consistent adversarial attacks in high-dimensional linear classification. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=PUIivg3GrO.

Chuang Wang, Hong Hu, and Yue Lu. A solvable high-dimensional model of GAN. In *Advances in Neural Information Processing Systems*, volume 32, New York, 2019. Curran Associates, Inc.

Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23549–23588. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wei22a.html.

Ke Liang Xiao, Noah Marshall, Atish Agarwala, and Elliot Paquette. Exact risk curves of signsgd in high-dimensions: Quantifying preconditioning and noise-compression effects, 2026. URL https://arxiv.org/abs/2411.12135.

Yue Xing, Ruizhi Zhang, and Guang Cheng. Adversarially robust estimate and risk analysis in linear regression. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 514–522. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/xing21c.html.

Yuki Yoshida and Masato Okada. Data-dependence of plateau phenomenon in learning with neural network—statistical mechanical analysis. In *Advances in Neural Information Processing Systems*, volume 32, New York, 2019. Curran Associates, Inc.

Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9):2805–2824, 2019. doi: 10.1109/TNNLS.2018.2886017.

Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data, 2019. URL https://arxiv.org/abs/1906.00555.

Notation. We say an event holds *with overwhelming probability* (w.o.p.) if there exists a function $\omega : \mathbb{N} \to \mathbb{R}$ with $\omega(d)/\log d \to \infty$ such that the event holds with probability at least $1 - e^{-\omega(d)}$. We let $\mathbb{1}_A(x)$ be the indicator function of the set A where it is 1 if $x \in A$ and 0 otherwise. For a matrix $A \in \mathbb{R}^{m\times d}$, we use $\|A\|$ to denote the Frobenius norm, $\|A\|_{\mathrm{op}}$ to denote the operator-2 norm. For normed vector spaces $\mathcal{A}$, $\mathcal{B}$ with norms $\|\cdot\|_{\mathcal{A}}$ and $\|\cdot\|_{\mathcal{B}}$, respectively, and for $\alpha \geq 0$, we say a function $F : \mathcal{A} \to \mathcal{B}$ is α-*pseudo-Lipschitz* with constant L if for any $A, \widehat{A} \in \mathcal{A}$, we have

$$\|F(A) - F(\widehat{A})\|_{\mathcal{B}} \leq L\|A - \widehat{A}\|_{\mathcal{A}}(1 + \|A\|_{\mathcal{A}}^{\alpha} + \|\widehat{A}\|_{\mathcal{A}}^{\alpha}).$$

We write $f(t) \asymp g(t)$ if there exist *absolute* constants $C, c > 0$ such that $c \cdot g(t) \leq f(t) \leq C \cdot g(t)$ for all t. If the constants depend on parameters, e.g., α, then we write $\asymp_{\alpha}$. We write $f(t) \sim g(t)$ as $t \to a$ if $\lim_{t\to a} \frac{f(t)}{g(t)} = 1$. Finally, the operator $\oplus$ denotes vector/matrix concatenation.

Outline of the paper. The remainder of the article is structured as follows:

1. Appendix A extends the discussion of related work.
2. Appendix B restates Section 2 in more detail. It provides the integro-differential equation (9), the det. equivs. of the resolvent statistics for SGD and AdvHSGD, their extension to statistics involving $X^{\star,\mathrm{Adv}}$, the general form of Theorem 2.1 and the conditions under which its stopping time can be removed.
3. Appendix C contains the proof Theorem 2.1. It shows through Doob decompositions that SGD and AdvHSGD are approximate solutions of the integro-differential equation (9), develops the stability theory of such approximate solutions together with the net argument over the contour, and transfers the resulting concentration to general statistics.
4. Appendix D consists of preliminary results which help simplify the proofs in Appendix C: derivatives of the special statistics, concentration and pseudo-Lipschitz estimates, the explicit Doob decompositions of SGD and AdvHSGD, and the bounds on the martingale and error terms.
5. Appendix E details the comparison of ℓ_2-adversarial least squares with ℓ_2-regularized least squares under adaptive learning rate and regularization. It derives the det. equivs. of HSGD with this schedule (Corollary E.1) and proves Theorem 3.1.
6. Appendix F derives the Polyak stepsize and exact line search for general losses, specializes them to ℓ_2-adversarial least squares and proves the ridge factorization (13) of their ratio (Propositions F.1 and F.2).
7. Appendix G analyzes the dynamics of ℓ_2-adversarial least squares: the coupled Volterra equations for $\mathscr{R}$ and $\widehat{\mathscr{B}}_{44}$, the ODE for $q(t)$ and its boundedness, and the proofs of Theorems 3.2 and 3.3 in their general form.
8. Appendix H solves the integro-differential equation in the eigenbasis of the covariances and derives the equivalent system of ODEs used in the numerical simulations.
9. Appendix I verifies the assumptions for least squares and binary logistic regression and computes the corresponding det. equivs.
10. Appendix J provides the parameters and details of the numerical simulations for each figure.

CONTENTS

1 Introduction **1**

1.1 Model Setup 3

1.2 Algorithmic Setup 4

2 Deterministic dynamics for SGD under ℓ_2-adversarial training **5**

3 ℓ_2-Adversarial Training v.s. ℓ_2-Regularization **7**

4 Idealized Exact Line Search and Polyak Stepsize **8**

5 Limitations and Future Work **9**

A Additional Related Work **18**

B Deterministic Dynamics for SGD under ℓ_2-Adversarial Training **19**

B.1 Adversarial Homogenized SGD 22

B.2 Extension to the Minimizer of the Adversarial Population Risk 23

B.3 Deterministic Dynamics for General Statistics 25

C SGD and AdvHSGD are Approximate Solutions **26**

C.1 The Doob Decomposition of SGD and AdvHSGD 26

C.2 Theory of Approximate Solutions to the Integro-Differential Equation 29

C.3 A Sufficient Condition to Remove the Stopping-time 38

C.4 Concentration Result for General Statistics 41

C.5 Concentration of General Statistics Involving the Minimizer of the ℓ_2-Adversarial Risk 42

D Preliminaries for the Proofs **46**

D.1 Derivative of Special Statistics 47

D.2 Concentration and Pseudo-Lipschitz 47

D.3 Doob Decomposition of Adversarial Homogenized SGD 48

D.4 Doob Decomposition for SGD 52

D.5 Error Bounds 55

D.5.1 Adversarial Homogenized SGD Martingale Error 56

D.5.2 Bounds on the Martingales $\mathcal{M}_k^{\mathrm{Grad}}$ and $\mathcal{M}_k^{\mathrm{Hess}}$ 58

D.5.3 Bounds on the Error Term $\mathcal{E}_k^{\mathrm{Hess}}$ of the Hessian Martingale Increments . . 67

E ℓ_2-Adversarial Training v.s. ℓ_2-Regularization **70**

F Idealized Exact Line Search and Polyak Stepsize **75**

F.1 Polyak Stepsize 75

F.1.1 General Polyak Stepsize 75

F.1.2 Polyak Stepsize For ℓ_2-adversarial least-squares 77

F.2 Line Search 79

F.2.1 General Line Search 79

F.3 Line search for ℓ_2-Adversarial Least Squares 79

F.4 Comparing the Polyak Stepsize and Line Search for ℓ_2-adversarial least-squares . . 80

G ℓ_2-Adversarial Linear Regression Analysis **85**

H Integro-Differential Equation Analysis **98**

I Analysis of Examples **105**

I.1 Least Squares 105

I.1.1 Supporting Results 109

I.2 Binary Logistic Regression 111

J Numerical Simulation Details **113**

A Additional Related Work

This section extends the discussion of related work. Given the breadth of the literatures on adversarial training and on high-dimensional dynamics of stochastic algorithms, it is necessarily non-exhaustive and highlights only a subset of related developments.

Adversarial training. Adversarial robustness and examples have been extensively studied in machine learning (Szegedy et al., 2014; Goodfellow et al., 2015; Madry et al., 2018; Kurakin et al., 2018; Yuan et al., 2019; Bai et al., 2021). Adversarial training has also been studied extensively in the case of deep neural networks (Goodfellow et al., 2015; Madry et al., 2018; Shafahi et al., 2019; Rice et al., 2020; Zhai et al., 2019; Chen et al., 2021). From a theoretical perspective, the generalization and regularization properties of adversarial training as well as the robust versus standard accuracy trade-off have been studied in linear models (Raghunathan et al., 2020; Dan et al., 2020; Clarysse et al., 2022; Ribeiro et al., 2023; 2025; Xing et al., 2021; Ribeiro & Schön, 2023) and two-layer neural networks (Robin et al., 2026). Under Gaussian data, exact asymptotics for these properties are derived in the proportional regime (Tanner et al., 2024; Vilucchio et al., 2024; Javanmard et al., 2020; Hassani & Javanmard, 2024; Dohmatob & Scetbon, 2024; Vilucchio et al., 2026). These properties were also investigated for Gaussian mixture models (Dobriban et al., 2022; Dan et al., 2020; Bhagoji et al., 2019; Javanmard & Soltanolkotabi, 2022; Taheri et al., 2021).

Deterministic dynamics of stochastic algorithms in high-dimensions. As mentioned in Collins-Woodfin et al. (2024b); Collins-Woodfin & Seroussi (2025), the theory of deterministic dynamics of stochastic algorithms on isotropic Gaussian data has a long history (Biehl & Riegler, 1994; Biehl & Schwarze, 1995; Saad & Solla, 1995a;b). These rigorous results were also extended to other models (Goldt et al., 2019; Wang et al., 2019; Arnaboldi et al., 2023; Damian et al., 2023; Arnaboldi et al., 2024; Dandi et al., 2024) under isotropic Gaussian data. Det. equivs. were also proven for multi-pass SGD with small mini-batches (Paquette et al., 2022) and with momentum (Lee et al., 2022). Other high-dimensional det. equivs. yielding different types of dynamics have also been studied (Mignacco et al., 2021; Gerbelot et al., 2024; Celentano et al., 2026; Fan & Wang, 2026; Chandrasekher et al., 2023; Bordelon & Pehlevan, 2022; Arous et al., 2023). Ben Arous et al. (2022); Arous et al. (2026); Refinetti et al. (2021) study the high-dimensional training dynamics of streaming SGD under GMM with isotropic covariance and Mignacco et al. (2021) study the analogous setting under multi-pass SGD.

In recent years, a line of work has made significant contributions to understanding the effect of non-identity covariance on the training dynamics of SGD under data from a single class (Collins-Woodfin & Paquette, 2024; Balasubramanian et al., 2024; Goldt et al., 2022; Yoshida & Okada, 2019; Goldt et al., 2020; Collins-Woodfin et al., 2024a; Xiao et al., 2026; Marshall et al., 2024; Malaxechebarría et al., 2026). As mentioned in Collins-Woodfin et al. (2024a), the non-identity covariance modifies the optimization landscape and affects convergence properties. Collins-Woodfin & Seroussi (2025) extend Collins-Woodfin et al. (2024a) to Gaussian mixture data with non-identity covariance and where the number of classes grows logarithmically with the input dimension d. Our work extends Collins-Woodfin et al. (2024a); Collins-Woodfin & Seroussi (2025) by accounting for the ℓ_2-adversarial objective under Gaussian mixture data with two classes and non-identity covariance.

B Deterministic Dynamics for SGD under ℓ_2-Adversarial Training

This section restates Section 2 in more detail and provides more context which may help interpret the results.

From Assumption 6 and Assumption 7, the adversarial risk $\mathcal{R}^{\mathrm{adv}}(X)$ and functions $\mathcal{I}_i$ are entirely determined by the matrices $\widehat{B}_i(W)$ which involve the matrix products $B_i = \hat{X}^\top K_i \hat{X}$, $\hat{X}^\top \mu_i$ and the scalar $X^\top X$. Note that these matrices are embedded into the following polynomials in K_i: $W^\top K_i W$ and $W^\top W$. Hence, if we are able to compute a det. equiv. for $W^\top q(K_i) W^\top$ for an arbitrary polynomial q, then we can study the deterministic dynamics of these statistics.

Through the use of Cauchy's integral formula, we can recover these statistics using the resolvents $R(z_i; K_i) \overset{\mathrm{def}}{=} (K_i - z_i I_d)^{-1}$ of K_i with $z_i \in \mathbb{C} \setminus \mathrm{spec}(K_i)$. For any polynomial q we have

$$W^\top q(K_i) W = -\frac{1}{2\pi \mathrm{i}} \oint_\Gamma q(z_i) W^\top R(z_i; K_i) W \, \mathrm{d}z_i,$$

where Γ is a contour enclosing the eigenvalues of K_i. Since we are interested in tracking simultaneously $W^\top q(K_1) W$ and $W^\top q(K_2) W$, we introduce $\mathcal{K} \overset{\mathrm{def}}{=} \{K_1, K_2\}$ and $\mathcal{R}(z; \mathcal{K}) \overset{\mathrm{def}}{=} R(z_1; K_1) R(z_2; K_2)$. We also introduce a fixed contour $\Gamma = \{w \in \mathbb{C} : |w| = \max\{1, 2\|\mathcal{K}\|_{\mathrm{op}}\}\}$ with $\|\mathcal{K}\|_{\mathrm{op}} \overset{\mathrm{def}}{=} \max\{\|K_1\|_{\mathrm{op}}, \|K_2\|_{\mathrm{op}}\}$. Note that by construction the distance between the spectra of K_1 and K_2 and the contour is always greater or equal than $\frac{1}{2}$.

For $z \in \Gamma^2$, we write $z = (z_1, z_2)$ and when integrating over z_1 and z_2 simultaneously, we write for any function $f : \mathbb{C}^2 \to \mathbb{C}$

$$\oint f(z) \mathrm{D}z \overset{\mathrm{def}}{=} \frac{-1}{4\pi^2} \oint_{\Gamma^2} f(z) \, \mathrm{d}z_1 \, \mathrm{d}z_2.$$

We introduce a norm over the contour Γ^2.

Definition B.1. *For a continuous function $H : \mathbb{C}^2 \to \mathbb{C}^{4\times 4}$, define*

$$\|H\|_\Gamma \overset{\mathrm{def}}{=} \sup_{z \in \Gamma^2} \|H(z)\|.$$

Similarly to Collins-Woodfin & Seroussi (2025), we will be interested in computing det. equivs. for the following matrices

$$S(W, z) \overset{\mathrm{def}}{=} \hat{X}^\top \mathcal{R}(z; \mathcal{K}) \hat{X} \quad \text{and} \quad M(W, z) \overset{\mathrm{def}}{=} \hat{X}^\top \mathcal{R}(z; \mathcal{K}) \mu \in \mathbb{C}^{2\times 2}, \tag{14}$$

which we regroup under

$$Z(W, z) \overset{\mathrm{def}}{=} W^\top \mathcal{R}(z; \mathcal{K}) W = \begin{bmatrix} S(W, z) & M(W, z) \\ M(W, z)^\top & \mu^\top \mathcal{R}(z; \mathcal{K}) \mu \end{bmatrix} \in \mathbb{C}^{4\times 4}. \tag{15}$$

Let $M_i(W, z)$ denote the ith column of $M(W, z)$ then from Cauchy's integral formula, we can recover $\widehat{B}_i(W)$ as follows

$$\widehat{B}_i(W) = \begin{bmatrix} \oint z_i S(W, z) \mathrm{D}z & \oint M_i(W, z) \mathrm{D}z & 0 \\ \oint M_i(W, z)^\top \mathrm{D}z & \|\mu_i\|^2 & 0 \\ 0 & 0 & \oint S_{11}(W, z) \mathrm{D}z. \end{bmatrix} \tag{16}$$

Given that we are studying the dynamics of SGD, the det. equiv. of $Z(W_k, z)$ will evolve with time. This det. equiv. which we denote $\mathcal{Z}(t, z)$ corresponds to the solution of the following integral-differential equation (17).

Integro-Differential Equation for $\mathcal{Z}(t, z)$. For any contour $\Gamma \subset \mathbb{C}$ enclosing the eigenvalues of K_1 and K_2, we have an expression for the derivative of $\mathcal{Z}$ for $z \in \Gamma^2$:

$$\mathrm{d}\mathcal{Z}(t, \cdot) = \mathcal{F}(z, \mathcal{Z}(t, \cdot)) \, \mathrm{d}t, \quad \text{with initial condition} \quad \mathcal{Z}(0, z) = Z(W_0, z), \tag{17}$$

where

$$\mathcal{Z}(t,z) \overset{\text{def}}{=} \begin{bmatrix} \mathcal{S}(t,z) & \mathcal{M}(t,z) \\ \mathcal{M}(t,z)^\top & \mu^\top \mathcal{R}(z;\mathcal{K})\mu \end{bmatrix}, \quad \text{and} \quad \mathcal{F}(z,\mathcal{Z}(t,\cdot)) \overset{\text{def}}{=} \begin{bmatrix} \mathcal{F}_{\mathcal{S}}(z,\mathcal{Z}(t,\cdot)) & \mathcal{F}_{\mathcal{M}}(z,\mathcal{Z}(t,\cdot)) \\ \mathcal{F}_{\mathcal{M}}(z,\mathcal{Z}(t,\cdot))^\top & 0_{2\times 2} \end{bmatrix}, \tag{18}$$

The maps $\mathcal{F}_{\mathcal{S}}$ and $\mathcal{F}_{\mathcal{M}}$ are defined as follows

$$\begin{aligned} \mathrm{d}\mathcal{M}(t,z) = -2\gamma(t)\sum_{i=1}^{2} p_i \Bigg(& H_{1,i}^\top(\widehat{\mathscr{B}}_i(t)) \left(z_i \mathcal{M}(t,z) - \frac{1}{2\pi\,\mathrm{i}} \oint_\Gamma \mathcal{M}(t,z)\,\mathrm{d}z_i \right) \\ & + H_{2,i}(\widehat{\mathscr{B}}_i(t)) \otimes \mu^\top \mathcal{R}(z;\mathcal{K})\mu_i + \left(\frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) \right) \mathrm{D}\,\mathcal{M}(t,z) \Bigg)\,\mathrm{d}t, \end{aligned} \tag{19}$$

and

$$\begin{aligned} \mathrm{d}\mathcal{S}(t,z) = -2\gamma(t)\sum_{i=1}^{2} p_i \Bigg(& \left(z_i \mathcal{S}(t,z) - \frac{1}{2\pi\,\mathrm{i}} \oint_\Gamma \mathcal{S}(t,z)\,\mathrm{d}z_i \right) H_{1,i}(\widehat{\mathscr{B}}_i(t)) \\ & + H_{1,i}^\top(\widehat{\mathscr{B}}_i(t)) \left(z_i \mathcal{S}(t,z) - \frac{1}{2\pi\,\mathrm{i}} \oint_\Gamma \mathcal{S}(t,z)\,\mathrm{d}z_i \right) \\ & + \mathcal{M}_i(t,z) \otimes H_{2,i}(\widehat{\mathscr{B}}_i(t)) + H_{2,i}(\widehat{\mathscr{B}}_i(t)) \otimes \mathcal{M}_i(t,z) \\ & + \left(\frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) \right) (\mathcal{S}(t,z)\,\mathrm{D} + \mathrm{D}\,\mathcal{S}(t,z)) \Bigg)\,\mathrm{d}t \\ & + \frac{\gamma(t)^2}{d} \sum_{i=1}^{2} p_i \mathcal{I}_i(\widehat{\mathscr{B}}_i(t))\,\mathrm{Tr}((K_i + \mu_i\mu_i^\top)\mathcal{R}(z;\mathcal{K}))\,\mathrm{D}\;\mathrm{d}t, \end{aligned} \tag{20}$$

$$\text{where} \quad H_{1,i}(\widehat{\mathscr{B}}_i) = \left[\begin{array}{c|c} \partial_{11} h_i(\widehat{\mathscr{B}}_i) & 0 \\ \hline \partial_{21} h_i(\widehat{\mathscr{B}}_i) & 0 \end{array} \right], \quad H_{2,i}(\widehat{\mathscr{B}}_i) = \left[\begin{array}{c} \partial_{13} h_i(\widehat{\mathscr{B}}_i) \\ \hline 0 \end{array} \right],$$

$$\mathrm{D} = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & 0 \end{array} \right] \quad \text{and} \quad \widehat{\mathscr{B}}_i(t) = \begin{bmatrix} \mathscr{B}_i(t) & \mathfrak{m}_i(t) & 0 \\ \mathfrak{m}_i(t)^\top & \|\mu_i\|^2 & 0 \\ 0 & 0 & \oint \mathcal{S}_{11}(t,z)\,\mathrm{D}z \end{bmatrix}. \tag{21}$$

Here $\mathscr{B}_i(t) = \oint z_i \mathcal{S}(t,z)\,\mathrm{D}z \in \mathbb{R}^{2\times 2}$, $\mathfrak{m}_i(t) = \oint \mathcal{M}_i(t)\,\mathrm{D}z$ where $\mathcal{M}_i(t) \in \mathbb{R}^2$ denotes the ith column of $\mathcal{M}$, associated with μ_i. For $i = 1, 2$, the functions $h_i : \mathbb{R}^{4\times 4} \to \mathbb{R}$ and $\mathcal{I}_i : \mathbb{R}^{4\times 4} \to \mathbb{R}$ are defined in Assumptions 6 and 7 respectively. Finally, here $W_0 = \hat{X}_0 \oplus \mu$ with $\hat{X}_0 = X_0 \oplus X^\star$ in the soft label setting and $\hat{X}_0 = X_0 \oplus 0$ in the hard label setting.

Remark B.1. *We note that, in the hard label setting, since $X^\star = 0$ and from the definition of $\mathcal{Z}(t,z)$, $\mathcal{S}(t,z)$ and $\mathcal{M}(t,z)$, the only non-zero terms in the integro-differential equation* (17) *are $\mathcal{S}_{11}(t,z)$ and the first row in $M(t,z)$.*

Note that the integral-differential equation in (17) corresponds to the adversarial version of the equations presented in Collins-Woodfin et al. (2024a) and (Collins-Woodfin & Seroussi, 2025). Hence, the two equations follow a very similar structure which differs due to the need to track the statistic $\|X_k\|^2$ as presented in (4) and due to different definitions for the functions h_i and $\mathcal{I}_i$ defined in Assumptions 6 and 7 respectively.

Since SGD is a discrete time process, we introduce the following time parametrization to compare it to its continuous time limit:

$$k \text{ iterations of SGD} = \lfloor td \rfloor, \quad \text{for } t \in \mathbb{R}^+ \text{ the continuous time parameter.}$$

Here $\lfloor \cdot \rfloor$ denotes the floor function. We will show that $\mathcal{Z}(t,z)$ approximates $Z(W_{\lfloor td \rfloor}, z)$ as $d \to \infty$ and formalize this in Theorem B.1. We thus denote $\mathcal{Z}(t,z)$ as the det. equiv. of $Z(W_{\lfloor td \rfloor}, z)$. Similarly, it follows from Theorem B.4 which we will introduce later that $\widehat{\mathscr{B}}_i(t)$, $\widehat{\mathscr{B}}_{i,44}(t)$, $\mathscr{B}_i(t)$

and $\mathfrak{m}_i(t)$ as defined in (21) respectively correspond to the det. equivs. of $\widehat{B}_i(W_{\lfloor td \rfloor})$, $\|X_{\lfloor td \rfloor}\|^2$, $B_i(W_{\lfloor td \rfloor})$ and $\hat{X}_{\lfloor td \rfloor}^\top \mu_i$.

Another statistic we can extract from $Z(\widehat{W}_{\lfloor td \rfloor}, z)$ and $\mathcal{Z}(t,z)$ through contour integration is the norm of the iterates $X_{\lfloor td \rfloor}$. We extend this to $W_{\lfloor td \rfloor}$ by introducing the following pair of statistics:

$$N(W_k) \stackrel{\text{def}}{=} \|W_k\|^2 = \|X_k\|^2 + \|X^\star\|^2 + \|\mu_1\|^2 + \|\mu_2\|^2 \quad \text{and} \quad \mathscr{N}(t) \stackrel{\text{def}}{=} \oint \operatorname{Tr}(\mathcal{Z}(t,z))\,\mathrm{D}z. \tag{22}$$

Here $\mathscr{N}(t)$ corresponds to the det. equiv. of $N(W_{\lfloor td \rfloor})$. We are now ready to present Theorem B.1.

Theorem B.1. *Suppose that the previous assumptions holds. Let ϑ_M be the first time either $(\widehat{\mathscr{B}}_1(t), \widehat{\mathscr{B}}_2(t))$ or $(\widehat{B}_1(W_{\lfloor td \rfloor}), \widehat{B}_2(W_{\lfloor td \rfloor}))$ exits $\mathcal{U}^2$ or that $\mathscr{N}(t) \geq M$ or $N(W_{\lfloor td \rfloor}) \geq M$. Then for any $T, M > 0$ and any $\varepsilon \in (0, \frac{1}{2})$, with overwhelming probability*

$$\sup_{0 \leq t \leq T \wedge \vartheta_M} \|Z(W_{\lfloor td \rfloor}, z) - \mathcal{Z}(t,z)\|_\Gamma \leq C d^{-\varepsilon},$$

for some constant C (independent of d) and where $\mathcal{Z}(t,z)$ is the solution to the integro-differential equation (17)*.*

The stopping time ϑ_M in Theorem B.1 encodes two conditions. The first, $\mathscr{N}(t) < M$ and $N(W_{\lfloor td \rfloor}) < M$, is a non-explosiveness condition: it keeps the iterates of SGD and their det. equiv. bounded, so that the constants in the α-pseudo-Lipschitz bounds of Assumptions 3 and 6 stay controlled. When f is α-pseudo-Lipschitz with $\alpha \leq 1$, a standard Grönwall argument shows that $\mathscr{N}(t)$ grows at most exponentially, and this condition can be removed (see Proposition B.1).

The second, that $(\widehat{\mathscr{B}}_1(t), \widehat{\mathscr{B}}_2(t))$ and $(\widehat{B}_1(W_{\lfloor td \rfloor}), \widehat{B}_2(W_{\lfloor td \rfloor}))$ remain in $\mathcal{U}^2$, keeps the trajectories in the region where Assumption 6 holds, namely where h_1 and h_2 are continuously differentiable with α-pseudo-Lipschitz derivatives. When $\mathcal{U} = \mathbb{R}^{4\times 4}$ this condition is automatically satisfied.

Since $\widehat{B}_i(W)$ and $N(W)$ are contour integrals of $Z(W,z)$ by (16) and (22), and $\widehat{\mathscr{B}}_i(t)$ and $\mathscr{N}(t)$ the same contour integrals of $\mathcal{Z}(t,z)$, Theorem B.1 transfers to them.

Corollary B.1. *Under the assumptions of Theorem B.1, for any $T, M > 0$ and $\varepsilon \in (0, \frac{1}{2})$, with overwhelming probability*

$$\sup_{0 \leq t \leq T \wedge \vartheta_M} \Big(\max_{i=1,2} \big\| \widehat{B}_i(W_{\lfloor td \rfloor}) - \widehat{\mathscr{B}}_i(t) \big\| + \big| N(W_{\lfloor td \rfloor}) - \mathscr{N}(t) \big| \Big) \leq C d^{-\varepsilon},$$

for some constant $C > 0$ independent of d.

When K_1 and K_2 are finite-dimensional matrices, to facilitate the numerical simulations of the integral-differential equation in (17) we can decompose the equation in terms of the eigenspaces of the covariance matrices K_1 and K_2 since they commute. Let $((\lambda_j^{(\ell)}, \omega_j) : 1 \leq j \leq d)$ be the eigenvalues and orthonormal eigenvectors of K_ℓ for $\ell = 1, 2$ and define the pairs of statistics

$$V_j(W_k) \stackrel{\text{def}}{=} \hat{X}_k^\top \omega_j \omega_j^\top \hat{X}_k = \frac{-1}{4\pi^2} \oint_{\Gamma_j^2} S(W_k, z)\,\mathrm{d}z_1\,\mathrm{d}z_2 \quad \text{and} \quad \mathscr{V}_j(t) \stackrel{\text{def}}{=} \begin{bmatrix} \mathscr{V}_{j,11}(t) & \mathscr{V}_{j,12}(t) \\ \mathscr{V}_{j,21}(t) & \mathscr{V}_{j,22}(t) \end{bmatrix},$$
$$m_{j,i}(W_k) \stackrel{\text{def}}{=} \hat{X}_k^\top \omega_j \omega_j^\top \mu_i = \frac{-1}{4\pi^2} \oint_{\Gamma_j^2} M_i(W_k, z)\,\mathrm{d}z_1\,\mathrm{d}z_2 \quad \text{and} \quad \mathfrak{m}_{j,i}(t) \stackrel{\text{def}}{=} \begin{bmatrix} \mathfrak{m}_{j,i,1}(t) \\ \mathfrak{m}_{j,i,2}(t) \end{bmatrix}, \tag{23}$$

where Γ_j is a contour enclosing solely the eigenvalues $\lambda_j^{(1)}$ and $\lambda_j^{(2)}$ defined analogously to Γ. Similarly to Collins-Woodfin et al. (2024a); Collins-Woodfin & Seroussi (2025), from the partial fractions decomposition of meromorphic functions, we have the representations for $i = 1, 2$

$$\mathcal{S}(t,z) = \sum_{j=1}^d \frac{\mathscr{V}_j(t)}{\prod_{\ell=1}^2 (\lambda_j^{(\ell)} - z_\ell)} \quad \text{and} \quad \mathcal{M}_i(t,z) = \sum_{j=1}^d \frac{\mathfrak{m}_{j,i}(t)}{\prod_{\ell=1}^2 (\lambda_j^{(\ell)} - z_\ell)}. \tag{24}$$

Summing over the eigenspaces of K_1 and K_2, we also have the representations for $\ell = 1, 2$

$$\mathscr{B}_\ell(t) = \sum_{j=1}^d \lambda_j^{(\ell)} \mathscr{V}_j(t), \quad \mathfrak{m}_\ell = \sum_{j=1}^d \mathfrak{m}_{j,\ell}(t) \quad \text{and} \quad \widehat{\mathscr{B}}_{\ell,44}(t) = \sum_{j=1}^d \mathscr{V}_{j,11}(t). \tag{25}$$

Given the relation (24), we obtain the d-dimensional set of coupled ODEs

$$\begin{aligned}
\frac{\mathrm{d}\mathscr{V}_j(t)}{\mathrm{d}t} &= -2\gamma(t)\sum_{i=1}^{2} p_i\Big(\lambda_j^{(i)}(\mathscr{V}_j(t)H_{1,i}(\widehat{\mathscr{B}}_i(t)) + H_{1,i}(\widehat{\mathscr{B}}_i(t))^\top\mathscr{V}_j(t)) + \mathfrak{m}_{j,i}(t)\otimes H_{2,i}(\widehat{\mathscr{B}}_i(t)) \\
&\quad + H_{2,i}(\widehat{\mathscr{B}}_i(t))\otimes\mathfrak{m}_{j,i}(t) + \left(\frac{\lambda}{2} + \partial_{44}h_i(\widehat{\mathscr{B}}_i(t))\right)(\mathscr{V}_j(t)\,\mathrm{D} + \mathrm{D}\,\mathscr{V}_j(t))\Big) \\
&\quad + \frac{\gamma(t)^2}{d}\left(\sum_{i=1}^{2} p_i\mathcal{I}_i(\widehat{\mathscr{B}}_i(t))(\lambda_j^{(i)} + (\omega_j^\top\mu_i)^2)\right)\mathrm{D}, \\
\frac{\mathrm{d}\mathfrak{m}_{j,\ell}(t)}{\mathrm{d}t} &= -2\gamma(t)\sum_{i=1}^{2} p_i\Big(\lambda_j^{(i)}H_{1,i}(\widehat{\mathscr{B}}_i(t))^\top\mathfrak{m}_{j,\ell}(t) + \langle\omega_j,\mu_\ell\rangle\langle\omega_j,\mu_i\rangle\cdot H_{2,i}(\widehat{\mathscr{B}}_i(t)) \\
&\quad + \left(\frac{\lambda}{2} + \partial_{44}h_i(\widehat{\mathscr{B}}_i(t))\right)\mathrm{D}\,\mathfrak{m}_{j,\ell}(t)\Big),
\end{aligned} \tag{26}$$

with initial condition $\mathscr{V}_j(0) = \hat{X}_0^\top\omega_j\omega_j^\top\hat{X}_0$ and $\mathfrak{m}_{j,\ell}(0) = \hat{X}_0^\top\omega_j\omega_j^\top\mu_\ell$. See Appendix H for more details on the derivation of the system of ODEs. Note that we show in Appendix H that $\mathscr{V}_j(t)$ are positive semidefinite matrices for $1 \le j \le d$ and $t \ge 0$. It follows that $\mathscr{B}_\ell(t)$ is positive semidefinite, $\widehat{\mathscr{B}}_{\ell,44}(t) \ge 0$ and $\mathscr{N}(t) \ge 0$ for all $t \ge 0$. Assumption 6 entails that the ODEs (26) have locally unique solutions whilst $(\widehat{\mathscr{B}}_1(t), \widehat{\mathscr{B}}_2(t)) \in \mathcal{U}^2$.

Remark B.2. *By Lemma 2 in Collins-Woodfin & Seroussi (2025), Theorem B.1 can be reformulated with $\mathcal{Z}(t,z)$ constructed with $\mathcal{S}(t,z)$ and $\mathcal{M}(t,z)$ defined as in* (24)*. We omit the proof as the argument is almost identical. Hence, we can recover a solution to* (17) *using a solution to the system of coupled ODEs* (26)*.*

B.1 Adversarial Homogenized SGD

An alternative way to compute the det. equiv. of $Z(W,z)$ as defined in (14) is through the use of Itô calculus. We introduce *adversarial homogenized SGD* (AdvHSGD)

$$\mathrm{d}\mathfrak{X}_t = -\gamma(t)\nabla\mathcal{R}_\lambda^{\mathrm{adv}}(\mathfrak{X}_t)\,\mathrm{d}t + \gamma(t)\sqrt{\frac{1}{d}\sum_{i=1}^{2} p_i\mathbb{E}_{\,v,\epsilon}[f_i'(\rho_{t,i};\epsilon)^2](K_i + \mu_i\mu_i^\top)}\,\mathrm{d}B_t, \tag{27}$$

where the initial condition is given by $\mathfrak{X}_0 = X_0$, $(B_t, t \ge 0)$ is a $\mathbb{R}^d$ standard Brownian motion and for $I \in \{1,2\}$ and $v \sim \mathcal{N}(0, I_d)$ we have

$$\begin{aligned}
\widehat{\mathfrak{X}}_t &\overset{\mathrm{def}}{=} \mathfrak{X}_t \oplus X^\star \in \mathbb{R}^{d\times 2}, \quad \mathcal{W}_t \overset{\mathrm{def}}{=} \widehat{\mathfrak{X}}_t \oplus \mu \in \mathbb{R}^{d\times 4}, \\
\rho_{t,I} &\overset{\mathrm{def}}{=} \widehat{\mathfrak{X}}_t^\top(\sqrt{K_I}v + \mu_I) + [\delta u_{t,I}\|\mathfrak{X}_t\| \quad 0]^\top, \\
\text{and} \quad u_{t,I} &= \mathrm{argmax}_{|u|\le 1}\, f_I(\widehat{\mathfrak{X}}_t^\top(\sqrt{K_I}v + \mu_I) + [\delta u\|\mathfrak{X}_t\| \quad 0]^\top, \epsilon).
\end{aligned} \tag{28}$$

In Appendix D.3, we perform a Doob decomposition elementwise on the function $Z(\mathcal{W}_t, z)$ and show (17) corresponds to the expression obtained by dropping the martingale terms which arise from Itô's lemma. Through the following theorem, we see that AdvHSGD concentrates around (17) with overwhelming probability.

Theorem B.2. *Suppose that the previous assumptions holds. For any $\omega > 0$, define*

$$\mathcal{U}_\omega \overset{\mathrm{def}}{=} \left\{B \in \mathcal{U} : \inf_{V\in\mathcal{U}^c}\|B - V\| \ge \omega\right\}. \tag{29}$$

Let Θ_M be the first time $(\widehat{\mathscr{B}}_1(t), \widehat{\mathscr{B}}_2(t))$ exits $\mathcal{U}_\omega^2$ or that $\mathscr{N}(t) \ge M$. Then for any $T, M > 0$ and any $\varepsilon \in (0, \frac{1}{2})$, with overwhelming probability

$$\sup_{0\le t\le T\wedge\Theta_M}\|Z(\mathcal{W}_t, z) - \mathcal{Z}(t,z)\|_\Gamma \le Cd^{-\varepsilon},$$

for some constant C (independent of d) and where $\mathcal{Z}(t,z)$ is the solution to the integro-differential equation (17) *or the system of coupled ODEs* (26)*.*

Theorem B.2 then leads to the following Corollary analogous to Corollary B.1.

Corollary B.2. *Under the assumptions of Theorem B.2, for any $T, M > 0$ and $\varepsilon \in (0, \frac{1}{2})$, with overwhelming probability*

$$\sup_{0\le t\le T\wedge\Theta_M}\Big(\max_{i=1,2}\big\|\widehat{B}_i(\mathcal{W}_t)-\widehat{\mathscr{B}}_i(t)\big\|+\big|N(\mathcal{W}_t)-\mathscr{N}(t)\big|\Big)\le Cd^{-\varepsilon},$$

for some constant $C > 0$ independent of d.

Notice that the stopping time in Theorem B.1 is formulated in terms of both $(\widehat{B}_1(W_{\lfloor td\rfloor}), \widehat{B}_2(W_{\lfloor td\rfloor}))$ and $(\widehat{\mathscr{B}}_1(t), \widehat{\mathscr{B}}_2(t))$ but the stopping time in Theorem B.2 in terms of $(\widehat{\mathscr{B}}_1(t), \widehat{\mathscr{B}}_2(t))$.

It turns out that only one of the two processes being compared needs to behave regularly. In particular, Theorem B.1 holds with the stopping time of Theorem B.2, which only requires $\mathcal{Z}(t, z)$ to remain bounded (i.e. $\mathscr{N}(t) < M$) and $(\widehat{\mathscr{B}}_1(t), \widehat{\mathscr{B}}_2(t))$ to remain in $\mathcal{U}_\omega^2$ (see Theorem C.2).

It is also possible to reverse the roles in Theorem B.1 and B.2 using the conditions $(\widehat{B}_1(W_{\lfloor td\rfloor}), \widehat{B}_2(W_{\lfloor td\rfloor})) \in \mathcal{U}_\omega \times \mathcal{U}_\omega$ and $N(W_{\lfloor td\rfloor}) < M$ or $(\widehat{B}_1(\mathcal{W}_{\lfloor td\rfloor}), \widehat{B}_2(\mathcal{W}_{\lfloor td\rfloor})) \in \mathcal{U}_\omega \times \mathcal{U}_\omega$ and $N(\mathcal{W}_t) < M$ respectively. For more details, see Section C.3, specifically Theorem C.2, Lemma C.1 and Remark C.2.

B.2 Extension to the Minimizer of the Adversarial Population Risk

The minimizer of the adversarial population risk is defined as follows.

Definition B.2. *Let $X^{\star,\mathrm{adv}}$ denote the adversarial ground truth. In other words*

$$X^{\star,\mathrm{adv}} \in \operatorname*{argmin}_X \left\{ \mathcal{R}^{\mathrm{adv}}(X) = \mathbb{E}_{a,I,\epsilon}[\max_{|s|\le 1} f_I(X^\top a + \delta s\|X\|, \epsilon)] \right\}. \tag{30}$$

In some cases such as least squares (Xing et al., 2021), there exists a threshold δ_0 such that for any $\delta < \delta_0$, $X^{\star,\mathrm{adv}} = X^\star$. We will assume that the norm of $X^{\star,\mathrm{Adv}}$ is bounded by a constant independent of dimension.

Assumption 10. *The adversarial ground truth $X^{\star,\mathrm{Adv}}$ satisfies $\|X^{\star,\mathrm{Adv}}\| \le C$ for some constant C independent of d.*

Given that we are also interested in statistics involving the adversarial ground truth $X^{\star,\mathrm{adv}}$ such as the descent to adversarial optimality $\|X - X^{\star,\mathrm{adv}}\|^2$, we introduce $\widehat{W} = W \oplus X^{\star,\mathrm{Adv}}$. To recover statistics of the form $\widehat{W}^\top q(K_1, K_2)\widehat{W}$, we must also keep track of the statistic $A_{\mathcal{K}}(\widehat{W}, z) \stackrel{\mathrm{def}}{=} X^\top \mathcal{R}(z; \mathcal{K}) X^{\star,\mathrm{adv}}$. Hence, analogously to Z defined in (15), we introduce Z^{Adv} as follows

$$Z^{\mathrm{Adv}}(\widehat{W}, z) \stackrel{\mathrm{def}}{=} \widehat{W}^\top \mathcal{R}(z; \mathcal{K})\widehat{W} = \begin{bmatrix} Z(W, z) & (\star)_0 \\ (\star)_0^\top & (X^{\star,\mathrm{Adv}})^\top \mathcal{R}(z; \mathcal{K})(X^{\star,\mathrm{Adv}}) \end{bmatrix}, \tag{31}$$

with $A_{\mathcal{K}}(\widehat{W}, z) \stackrel{\mathrm{def}}{=} X^\top \mathcal{R}(z; \mathcal{K}) X^{\star,\mathrm{Adv}}$ and where $(\star)_0$ corresponds to

$$(\star)_0 \stackrel{\mathrm{def}}{=} \begin{bmatrix} A_{\mathcal{K}}(\widehat{W}, z) & (X^\star)^\top \mathcal{R}(z; \mathcal{K}) X^{\star,\mathrm{Adv}} & \mu_1^\top \mathcal{R}(z; \mathcal{K}) X^{\star,\mathrm{Adv}} & \mu_2^\top \mathcal{R}(z; \mathcal{K}) X^{\star,\mathrm{Adv}} \end{bmatrix}^\top.$$

Let $\mathcal{A}_{\mathcal{K}}(t, z)$ denote the det. equiv. of $A_{\mathcal{K}}(\widehat{W}_{td}, z) = X_{td}^\top \mathcal{R}(z; \mathcal{K}) X^{\star,\mathrm{Adv}}$. The det. equiv. of $Z^(W_k, z)$ which we denote $\mathcal{Z}^{\mathrm{Adv}}(t, z)$ corresponds to the solution of the following integral-differential equation (32)

Integro-Differential Equation for $\mathcal{Z}^{\mathrm{Adv}}(t, z)$. For any contour $\Gamma \subset \mathbb{C}$ enclosing the eigenvalues of K_1 and K_2, we have an expression for the derivative of $\mathcal{Z}^{\mathrm{Adv}}$ for $z \in \Gamma^2$:

$$\mathrm{d}\mathcal{Z}^{\mathrm{Adv}}(t, \cdot) = \mathcal{F}^{\mathrm{Adv}}(z, \mathcal{Z}^{\mathrm{Adv}}(t, \cdot))\,\mathrm{d}t, \quad \text{with initial condition } \mathcal{Z}^{\mathrm{Adv}}(0, z) = Z^{\mathrm{Adv}}(\widehat{W}_0, z), \tag{32}$$

where

$$\mathcal{Z}^{\mathrm{Adv}}(t,z) \stackrel{\mathrm{def}}{=} \begin{bmatrix} \mathcal{Z}(t,z) & (\star)_2 \\ (\star)_2^\top & (X^{\star,\mathrm{Adv}})^\top \mathcal{R}(z;\mathcal{K})(X^{\star,\mathrm{Adv}}) \end{bmatrix},$$
$$(\star)_2 \stackrel{\mathrm{def}}{=} \begin{bmatrix} \mathcal{A}_{\mathcal{K}}(t,z) & (X^\star)^\top \mathcal{R}(z;\mathcal{K}) X^{\star,\mathrm{Adv}} & \mu_1^\top \mathcal{R}(z;\mathcal{K}) X^{\star,\mathrm{Adv}} & \mu_2^\top \mathcal{R}(z;\mathcal{K}) X^{\star,\mathrm{Adv}} \end{bmatrix}^\top,$$
$$\text{and} \quad \mathcal{F}^{\mathrm{Adv}}(z, \mathcal{Z}^{\mathrm{Adv}}(t,\cdot)) \stackrel{\mathrm{def}}{=} \begin{bmatrix} \mathcal{F}(z, \mathcal{Z}(t,\cdot)) & \mathcal{F}_\star(z, \mathcal{Z}^{\mathrm{Adv}}(t,\cdot)) \\ \mathcal{F}_\star(z, \mathcal{Z}^{\mathrm{Adv}}(t,\cdot))^\top & 0 \end{bmatrix},$$
$$\text{with} \quad \mathcal{F}_\star(z, \mathcal{Z}^{\mathrm{Adv}}(t,\cdot)) \stackrel{\mathrm{def}}{=} \begin{bmatrix} \mathcal{F}_{\mathcal{A}_{\mathcal{K}}}(z, \mathcal{Z}^{\mathrm{Adv}}(t,\cdot)) & 0 \end{bmatrix}^\top \in \mathbb{R}^5. \tag{33}$$

The map $\mathcal{F}$ is defined in (17) and

$$\mathrm{d}\mathcal{A}_{\mathcal{K}}(t,\cdot) = \mathcal{F}_{\mathcal{A}_{\mathcal{K}}}(z, \mathcal{Z}^{\mathrm{Adv}}(t,\cdot))\, \mathrm{d}t, \quad \text{with initial condition} \quad \mathcal{A}_{\mathcal{K}}(0,z) = A_{\mathcal{K}}(\widehat{W}_0, z), \tag{34}$$

where the map $\mathcal{F}_{\mathcal{A}_{\mathcal{K}}}$ is defined as follows

$$\mathcal{F}_{\mathcal{A}_{\mathcal{K}}}(z, \mathcal{Z}^{\mathrm{Adv}}(t,\cdot)) \stackrel{\mathrm{def}}{=} -2\gamma(t) \sum_{i=1}^{2} p_i \Bigg(\partial_{11} h_i(\widehat{\mathscr{B}}_i(t)) \left(z_i \mathcal{A}_{\mathcal{K}}(t,z) - \frac{1}{2\pi\,\mathrm{i}} \oint_\Gamma \mathcal{A}_{\mathcal{K}}(t,z)\, \mathrm{d}z_i \right)$$
$$+ \partial_{12} h_i(\widehat{\mathscr{B}}_i(t)) \left(z_i \langle X^\star \otimes X^{\star,\mathrm{Adv}}, \mathcal{R}(z;\mathcal{K}) \rangle - \frac{1}{2\pi\,\mathrm{i}} \oint_\Gamma \langle X^\star \otimes X^{\star,\mathrm{Adv}}, \mathcal{R}(z;\mathcal{K}) \rangle\, \mathrm{d}z_i \right)$$
$$+ \partial_{13} h_i(\widehat{\mathscr{B}}_i(t)) \langle \mu_i \otimes X^{\star,\mathrm{Adv}}, \mathcal{R}(z;\mathcal{K}) \rangle + \left(\frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) \right) \mathcal{A}_{\mathcal{K}}(t,z) \Bigg)\, \mathrm{d}t. \tag{35}$$

Here $\widehat{\mathscr{B}}_i(t)$ is defined in (17) and the functions $h_i : \mathbb{R}^{4\times 4} \to \mathbb{R}$ and $\mathcal{I}_i : \mathbb{R}^{4\times 4} \to \mathbb{R}$ are defined in Assumptions 6 and 7 respectively. Finally, here $\widehat{W}_0 = \hat{X}_0 \oplus \mu \oplus X^{\star,\mathrm{Adv}}$ where $\hat{X}_0 = X_0 \oplus X^\star$ in the soft label setting and $\hat{X}_0 = X_0 \oplus 0$ in the hard label setting.

The descent towards $X^{\star,\mathrm{Adv}}$ (i.e. $\|X - X^{\star,\mathrm{Adv}}$ in Appendix F.1) involves an expectation over the Gaussian vector $\widetilde{X}^\top a_i$ with $\widetilde{X} \stackrel{\mathrm{def}}{=} X \oplus X^\star \oplus X^{\star,\mathrm{Adv}} \in \mathbb{R}^{d\times 3}$ and $a_i \sim \mathcal{N}(\mu_i, K_i)$. Its covariance and mean, together with $\|X\|^2$, extend (4) to

$$\widehat{B}_i^{\mathrm{Adv}}(\widehat{W}) \stackrel{\mathrm{def}}{=} \begin{bmatrix} B_i^{\mathrm{Adv}}(\widehat{W}) & m_i^{\mathrm{Adv}}(\widehat{W}) & 0 \\ m_i^{\mathrm{Adv}}(\widehat{W})^\top & \|\mu_i\|^2 & 0 \\ 0 & 0 & \|X\|^2 \end{bmatrix} \in \mathbb{R}^{5\times 5},$$
$$B_i^{\mathrm{Adv}}(\widehat{W}) \stackrel{\mathrm{def}}{=} \widetilde{X}^\top K_i \widetilde{X} \quad \text{and} \quad m_i^{\mathrm{Adv}}(\widehat{W}) \stackrel{\mathrm{def}}{=} \widetilde{X}^\top \mu_i, \tag{36}$$

of which $\widehat{B}_i(W)$ is the principal submatrix on the rows and columns $\{1,2,4,5\}$. Since $\widetilde{X}^\top \mathcal{R}(z;\mathcal{K}) \widetilde{X}$ and $\widetilde{X}^\top \mathcal{R}(z;\mathcal{K}) \mu_i$ are submatrices of $Z^{\mathrm{Adv}}(\widehat{W}, z)$, Cauchy's integral formula gives $B_i^{\mathrm{Adv}}(\widehat{W}) = \oint z_i\, \widetilde{X}^\top \mathcal{R}(z;\mathcal{K}) \widetilde{X}\, \mathrm{D}z$ and $m_i^{\mathrm{Adv}}(\widehat{W}) = \oint \widetilde{X}^\top \mathcal{R}(z;\mathcal{K}) \mu_i\, \mathrm{D}z$, exactly as in (16). Replacing $Z^{\mathrm{Adv}}(\widehat{W}_{\lfloor td \rfloor}, z)$ by $\mathcal{Z}^{\mathrm{Adv}}(t,z)$ defines the det. equiv. $\widehat{\mathscr{B}}_i^{\mathrm{Adv}}(t)$. Once again, only the entries involving X evolve, such that

$$\mathscr{B}_i^{\mathrm{Adv}}(t) = \begin{bmatrix} \mathscr{B}_i(t) & \mathscr{A}_{K_i}(t) \\ \mathscr{A}_{K_i}(t) & (X^\star)^\top K_i X^{\star,\mathrm{Adv}} \quad (X^{\star,\mathrm{Adv}})^\top K_i X^{\star,\mathrm{Adv}} \in \mathbb{R}^{3\times 3} \end{bmatrix},$$
$$\text{and} \quad \mathfrak{m}_i^{\mathrm{Adv}}(t) = \mathfrak{m}_i(t) \oplus \mu_i^\top X^{\star,\mathrm{Adv}} \quad \text{with} \quad \mathscr{A}_{K_i}(t) = \oint z_i \mathcal{A}_{\mathcal{K}}(t,z)\, \mathrm{D}z \tag{37}$$

the det. equiv. of $X_{\lfloor td \rfloor}^\top K_i X^{\star,\mathrm{Adv}}$. Similarly to (24), when K_1 and K_2 are finite-dimensional matrices, we can decompose (34) in terms of the eigenspaces of the covariance matrices K_1 and K_2. Define the pair of statistics

$$U_j(\widehat{W}_k) \stackrel{\mathrm{def}}{=} X_k^\top \omega_j \omega_j^\top X^{\star,\mathrm{Adv}} = \frac{-1}{4\pi^2} \oint_{\Gamma_j^2} A_{\mathcal{K}}(\widehat{W}_k, z)\, \mathrm{d}z_1\, \mathrm{d}z_2 \quad \text{and its deterministic equiv. } \mathscr{U}_j(t), \tag{38}$$

where Γ_i is a contour enclosing solely the eigenvalues $\lambda_j^{(1)}$ and $\lambda_j^{(2)}$. From the partial fractions decomposition of meromorphic functions, we have the representation

$$\mathcal{A}_{\mathcal{K}}(t,z) = \sum_{j=1}^{d} \frac{\mathscr{U}_j(t)}{\prod_{\ell=1}^{2}(\lambda_j^{(\ell)} - z_\ell)}. \tag{39}$$

From Cauchy's integral formula, we have $\mathscr{A}_{K_i}(t) = \oint z_i \mathcal{A}_{\mathcal{K}}(t,z)\,\mathrm{D}z$ and summing over the eigenspaces of K_1 and K_2, for $i = 1,2$ we obtain $\mathscr{A}_{K_i}(t) = \sum_{j=1}\lambda_j^{(i)}\mathscr{U}_j(t)$.Thus, we obtain the coupled ODEs

$$\begin{aligned}\frac{\mathrm{d}\mathscr{U}_j(t)}{\mathrm{d}t} = -2\gamma(t)\sum_{i=1}^{2} p_i\Bigg(&\left(\lambda_j^{(i)}\partial_{11}h_i(\widehat{\mathscr{B}}_i(t)) + \left(\frac{\lambda}{2} + \partial_{44}h_i(\widehat{\mathscr{B}}_i(t))\right)\right)\mathscr{U}_j(t)\\ &+\left(\lambda_j\partial_{12}h_i(\widehat{\mathscr{B}}_i(t))\langle X^\star,\omega_j\rangle + \partial_{13}h_i(\widehat{\mathscr{B}}_i(t))\langle\mu_i,\omega_j\rangle\right)\langle X^{\star,\mathrm{Adv}},\omega_j\rangle\Bigg),\end{aligned} \tag{40}$$

with initial condition $\mathscr{U}_j(0) = X_0^\top\omega_j\omega_j^\top X^{\star,\mathrm{Adv}}$. As in Remark B.2, using an almost identical argument to Lemma 2 in Collins-Woodfin & Seroussi (2025), we can recover $\mathcal{Z}^{\mathrm{Adv}}(t,z)$ from a solution to the system of coupled ODEs (26) and (40).

We also obtain a parallel result to Theorems B.1 and B.2.

Theorem B.3. *Suppose that the previous assumptions holds. Let Θ_M be the first time $(\widehat{\mathscr{B}}_1(t),\widehat{\mathscr{B}}_2(t))$ exits $\mathcal{U}_\omega^2$ (see* (29)*) or that $\mathscr{N}(t) \geq M$. Then for any $T, M > 0$ and any $\varepsilon \in (0,\frac{1}{2})$, with overwhelming probability*

$$\sup_{0\leq t\leq T\wedge\Theta_M} \|Z^{\mathrm{Adv}}(\widehat{\mathcal{W}}_t,z) - \mathcal{Z}^{\mathrm{Adv}}(t,z)\|_\Gamma + \|Z^{\mathrm{Adv}}(\widehat{W}_{\lfloor td\rfloor},z) - \mathcal{Z}^{\mathrm{Adv}}(t,z)\|_\Gamma \leq Cd^{-\varepsilon},$$

for some constant $C > 0$ (independent of d) and where $\mathcal{Z}^{\mathrm{Adv}}(t,z)$ is the solution to the integro-differential equation (32) *or the system of coupled ODEs* (26) *and* (40)*.*

Once again, the following Corollary follows immediately from Cauchy's integral formula as in Corollaries B.1 and B.2.

Corollary B.3. *Under the assumptions of Theorem B.3, for any $T, M > 0$ and $\varepsilon \in (0,\frac{1}{2})$, with overwhelming probability*

$$\sup_{0\leq t\leq T\wedge\Theta_M}\left(\max_{i=1,2}\left\|\widehat{B}_i^{\mathrm{Adv}}(\widehat{W}_{\lfloor td\rfloor}) - \widehat{\mathscr{B}}_i^{\mathrm{Adv}}(t)\right\| + \max_{i=1,2}\left\|\widehat{B}_i^{\mathrm{Adv}}(\widehat{\mathcal{W}}_t) - \widehat{\mathscr{B}}_i^{\mathrm{Adv}}(t)\right\|\right) \leq Cd^{-\varepsilon},$$

for some constant $C > 0$ independent of d.

B.3 Deterministic Dynamics for General Statistics

As mentioned previously, we can recover statistics of the form $\widehat{W}^\top q(K_1,K_2)\widehat{W}$ for a polynomial q from Cauchy's integral formula. It is then natural to expect Theorems B.1, B.2 and B.3 extend to a more general set of polynomial functions of K_1 and K_2.

Assumption 11. *Fix $\ell \geq 1$. Suppose the statistic $\varphi : \mathbb{R}^d \to \mathbb{R}$ satisfies the composite structure*

$$\varphi(X) = g\left(\{\widehat{W}^\top q_j(K_1,K_2)\widehat{W}\}_{j=1}^{\ell}\right),$$

where $g : \mathbb{R}^{\ell\times 5\times 5} \to \mathbb{R}$ is α-pseudo-Lipschitz and $\{q_j\}_{j=1}^{\ell}$ are polynomials.

It is easy to see that $\widehat{B}_i(W)$ and h_i hence $\mathcal{R}_\lambda^{\mathrm{adv}}$ satisfy Assumption 11. Other examples of important statistics which satisfy this assumption are $\|X\|^2$, $\|X - X^\star\|^2$, $\|X - X^{\star,\mathrm{Adv}}\|^2$ and $\mathcal{I}(\widehat{B}_i(W))$.

For statistics satisfying Assumption 11, the det. equiv. is defined as follows

$$\phi(t) \overset{\mathrm{def}}{=} g\left(\left\{\oint q_j(z)\mathcal{Z}^{\mathrm{Adv}}(t,z)\,\mathrm{D}z\right\}_{j=1}^{\ell}\right). \tag{41}$$

We may finally present our main result.

Theorem B.4. *Suppose that the previous assumptions holds. Let Θ_M be the first time $(\widehat{\mathscr{B}}_1(t), \widehat{\mathscr{B}}_2(t))$ exits $\mathcal{U}_\omega^2$ (see* (29)*) or that $\mathscr{N}(t) \geq M$. For any function satisfying Assumption 11, for any $T, M > 0$ and any $\varepsilon \in (0, \frac{1}{2})$, with overwhelming probability*

$$\sup_{0 \leq t \leq T \wedge \Theta_M} |\varphi(\mathfrak{X}_t) - \varphi(X_{\lfloor td \rfloor})| + |\varphi(\mathfrak{X}_t) - \phi(t)| \leq C d^{-\varepsilon},$$

for some constant C (independent of d) and where $\phi(t)$ is defined in (41)*.*

Provided either $(\widehat{\mathscr{B}}_1(t), \widehat{\mathscr{B}}_2(t))$, $(\widehat{B}_1(\widehat{W}_{\lfloor td \rfloor}), \widehat{B}_2(\widehat{W}_{\lfloor td \rfloor}))$ or $(\widehat{B}_1(\widehat{\mathcal{W}}_t), \widehat{B}_2(\widehat{\mathcal{W}}_t))$ stays in $\mathcal{U}^2$, we formalize a condition under which it is possible to remove the stopping time Θ_M.

Proposition B.1 (Non-explosiveness)**.** *Suppose that the assumptions stated in this framework hold and suppose that the objective function f is α-pseudo-Lipschitz with $\alpha \leq 1$. Then there exists a constant C such that*

$$\mathscr{N}(t) \leq (1 + \mathscr{N}(0)) e^{Ct},$$

for $C = C(L(f), \eta, \delta, \|\mathcal{K}\|_{\mathrm{op}}, \max_{i=1,2} \|\mu_i\|, \lambda, \bar{\gamma})$ and all time t such that $(\widehat{\mathscr{B}}_1(t), \widehat{\mathscr{B}}_2(t))$ is in $\mathcal{U}^2$.

For more details, see Corollary C.3 and Proposition C.5. Hence, from Proposition B.1, we obtain a simplification of Theorem B.4.

Corollary B.4. *Suppose that the previous assumptions hold. Suppose $\mathcal{U} = \mathbb{R}^{4\times 4}$ and that f is α-pseudo-Lipschitz with $\alpha \leq 1$. For any function satisfying Assumption 11, for any $T > 0$ and any $\varepsilon \in (0, \frac{1}{2})$, with overwhelming probability*

$$\sup_{0 \leq t \leq T} |\varphi(\mathfrak{X}_t) - \varphi(X_{\lfloor td \rfloor})| + |\varphi(\mathfrak{X}_t) - \phi(t)| \leq C d^{-\varepsilon},$$

for some constant $C > 0$ (independent of d) and where $\phi(t)$ is defined in (41)*.*

Remark B.3. *Note that Section C.4 contains analogous results, Theorem C.3 and Corollary C.2, for statistics not involving $X^{\star,\mathrm{Adv}}$ in the case that Assumption 10 fails.*

Remark B.4. *Assumption 5 is only needed for statistics involving $X^{\star,\mathrm{Adv}}$. For statistics of the form $g(\{W^\top q_j(K_1, K_2) W\}_{j=1}^{\ell})$, Theorem 2.1 holds without it. See Section C.4, precisvely Theorem C.3 and Corollary C.2. Nevertheless, Assumption 5 is essential to the proof and can fail to hold. Consider hard label binary logistic regression with means $\mu_1 = -\mu_2 = \rho u$, where $\|u\| = 1$ and $\rho > \delta$, and covariances $K_1 = K_2 = (I_d - uu^\top) + \sigma_d^2 uu^\top$ with $\sigma_d \to 0$. Then Assumptions 1–4 hold uniformly in d, but $\|X^{\star,\mathrm{Adv}}\| \to \infty$ so Assumption 5fails. Indeed, as $\sigma_d \to 0$ the classes become separable along u with adversarial margin $\rho - \delta$, so the minimizer escapes to infinity.*

C SGD AND ADVHSGD ARE APPROXIMATE SOLUTIONS

Outline of the section. In this section, we will prove the main result Theorem B.4. The section is divided as follows. Section C.1 introduces the Doob decompositions of SGD and AdvHSGD and provides an overview of the necessary steps to prove SGD and AdvHSGD are approximate solutions to the integro-differential equation (17). Section C.2 introduces the theory of approximate solutions to the integro-differential equation (17). It also includes proofs on the stability of approximate solutions and the proofs that AdvHSGD and SGD are approximate solutions to (17). Section C.3 contains a rigorous justification for the removal of the stopping time in Theorems B.1 and B.2. Section C.4 contains analogous results to Theorems B.3 and B.4 whenever Assumption 10 fails. Section C.5 introduces the theory of approximate solutions to the integro-differential equation (32) and the rigorous justification for the removal of the stopping time in Theorems B.3 and B.4.

Finally, Appendix D contains the details of the derivations of the Doob decompositions for AdvHSGD and SGD and supporting results for the proofs that AdvHSGD and SGD are approximate solutions to (17). We choose to delegate these results to the Appendix as they are quite computational.

C.1 THE DOOB DECOMPOSITION OF SGD AND ADVHSGD

In order to effectively compare the discrete-time iterates of SGD to its continuous time limit, recall we introduce the following time parametrization

$$k \text{ iterations of SGD} = \lfloor td \rfloor, \quad \text{for } t \in \mathbb{R}^+ \text{ the continuous time parameter.}$$

For simplicity of notation, we denote $X_{td} \overset{\text{def}}{=} X_{\lfloor td \rfloor}$, $\hat{X}_{td} \overset{\text{def}}{=} \hat{X}_{\lfloor td \rfloor}$ and $W_{td} \overset{\text{def}}{=} W_{\lfloor td \rfloor}$. The goal of Section C is to show AdvHSGD and SGD approximately solve (17). Before diving into the theory of approximate solutions to integral-dfferential equations, we provide intuition behind the existence of the deterministic limits for statistics of the iterates of SGD and AdvHSGD.

Since the entries of $Z(W_k, z)$ are quadratic statistics of the entries of W_k, a natural starting point is to consider the Taylor expansion of a quadratic function of W_{k+1} centered at W_k. Formally, let a quadratic function denote a C^∞ function $\varphi : \mathbb{R}^d \to \mathbb{R}$ such that $(\nabla^{(j)}\varphi)(X) \equiv 0$ for all $X \in \mathbb{R}^d$ and $j \geq 3$. Here we define $\nabla^{(j)}\varphi \overset{\text{def}}{=} \nabla^{(j)}_X \varphi$.

Recall we defined $\hat{X}_k \overset{\text{def}}{=} X_k \oplus X^\star$, $W_k \overset{\text{def}}{=} \hat{X}_k \oplus \mu$, $a_{k+1,I_{k+1}} = \sqrt{K_{I_{k+1}}} v_{k+1} + \mu_{I_{k+1}}$ for $v_{k+1} \sim \mathcal{N}(0, I_d)$, $I_{k+1} \in \{1, 2\}$, $r_{k,I_{k+1}} = \hat{X}_k^\top a_{k+1,I_{k+1}}$ and

$$\begin{aligned} g_{k,I_{k+1}} &= r_{k,I_{k+1}} + [\delta s_{k,I_{k+1}} \|X_k\| \quad 0]^\top, \\ s_{k,I_{k+1}} &= \operatorname{argmax}_{|s| \leq 1} f_{I_{k+1}}(X_k^\top a_{k+1,I_{k+1}} + \delta s \|X_k\|; \epsilon_{k+1}). \end{aligned} \tag{42}$$

By Taylor expansion, setting $\Delta_k \overset{\text{def}}{=} f'_{I_{k+1}}(g_{k,I_{k+1}}, \epsilon_{k+1})(a_{k+1,I_{k+1}} + \delta s_{k,I_{k+1}} \frac{X_k}{\|X_k\|}) + \lambda X_k$, we have

$$\varphi(X_{k+1}) = \varphi(X_k) - \frac{\gamma_k}{d} \langle \nabla \varphi(X_k), \Delta_k \rangle + \frac{\gamma_k^2}{2d^2} \cdot \langle \nabla^2 \varphi(X_k), \Delta_k^{\otimes 2} \rangle \tag{43}$$

To derive the Doob decomposition, the idea is to iteratively condition on $r_{k,I_{k+1}}$, $\hat{X}_k$ and I_{k+1}. For this, we define the σ-algebras,

$$\mathcal{G}_{k,i} \overset{\text{def}}{=} \sigma(\{\hat{X}_j\}_{j=0}^k, \{r_i\}_{i=0}^k, \{I_j\}_{j=1}^k, \epsilon_{k+1}, I_{k+1} = i) \quad \text{and} \quad \mathcal{F}_k \overset{\text{def}}{=} \sigma(\{\hat{X}_j\}_{j=0}^k).$$

It is clear that $s_{k,I_{k+1}}$ is measurable with respect to $\mathcal{G}_{k,i}$ such that $s_{k,I_{k+1}} \mid \mathcal{G}_{k,i}$ corresponds to a deterministic maximization problem. This leads to the decomposition for SGD

$$\begin{aligned} \varphi(X_{k+1}) = \varphi(X_k) &- \frac{\gamma_k}{d} \langle \nabla \varphi(X_k), \nabla \mathcal{R}_\lambda^{\text{adv}}(X_k) \rangle \\ &+ \frac{\gamma_k^2}{2d^2} \sum_{i=1}^2 p_i \langle \nabla^2 \varphi(X_k), K_i + \mu_i \mu_i^\top \rangle \mathbb{E}\left[(f_i'(g_{k,i}, \epsilon_{k+1}))^2 \,|\, \mathcal{F}_k \right] \\ &+ \Delta \mathcal{M}_k^{\text{Grad}}(\varphi) + \Delta \mathcal{M}_k^{\text{Hess}}(\varphi) + \mathbb{E}\left[\mathcal{E}_k^{\text{Hess}}(\varphi) \,|\, \mathcal{F}_k \right] \end{aligned} \tag{44}$$

where $\mathcal{M}_k^{\text{Grad}}(\varphi)$ and $\mathcal{M}_k^{\text{Hess}}(\varphi)$ are martingale increments and $\mathcal{E}_k^{\text{Hess}}(\varphi)$ is an error term:

$$\begin{aligned} \Delta \mathcal{M}_k^{\text{Grad}}(\varphi) &= \frac{\gamma_k}{d} f'_{I_{k+1}}(g_{k,I_{k+1}}) \langle \nabla \varphi(X_k), a_{k+1,I_{k+1}} + \delta s_{k,I_{k+1}} \tfrac{X_k}{\|X_k\|} \rangle \\ &\quad - \frac{\gamma_k}{d} \sum_{i=1}^2 p_i \mathbb{E}\left[f_i'(g_{k,i}, \epsilon) \langle \nabla \varphi(X_k), a_{k+1,i} + \delta s_{k,i} \tfrac{X_k}{\|X_k\|} \rangle \,|\, \mathcal{F}_k \right], \\ \Delta \mathcal{M}_k^{\text{Hess}}(\varphi) &= \frac{\gamma_k^2}{2d^2} \left(\langle \nabla^2 \varphi(W_k), \Delta_k^{\otimes 2} \rangle - \mathbb{E}\left[\langle \nabla^2 \varphi(X_k), \Delta_k^{\otimes 2} \rangle \,|\, \mathcal{F}_k \right] \right), \\ \text{and} \quad \mathcal{E}_k^{\text{Hess}}(\varphi) &= \frac{\gamma_k^2}{2d^2} \sum_{i=1}^2 p_i \langle \nabla^2 \varphi(W_k), \Delta^{\otimes 2} - (K_i + \mu_i \mu_i^\top) f_i'(g_{k,i}, \epsilon_{k+1})^2 \rangle. \end{aligned} \tag{45}$$

From Assumption 8, recall we defined $\gamma_t = \gamma(t/d)$ from which it follows $\gamma_{\lfloor td \rfloor} = \gamma(\lfloor td \rfloor / d)$. Similarly, we define $\gamma_{td} = \gamma_{\lfloor td \rfloor}$ for simplicity of notation. With the following forward difference

$$(\Delta \varphi)(X_j) \overset{\text{def}}{=} \varphi(X_{j+1}) - \varphi(X_j),$$

we can use our continuous time equivalents to sum up (integrate) and obtain

$$\varphi(X_{td}) = \varphi(X_0) + \sum_{j=0}^{\lfloor td \rfloor - 1} (\Delta \varphi)(W_j) \overset{\text{def}}{=} \varphi(X_0) + \int_0^t d \cdot (\Delta \varphi)(X_{sd}) \, \mathrm{d}s + \xi_{td}(\varphi), \tag{46}$$

where $|\xi_{td}(\varphi)| = \left| \int_{(\lfloor td \rfloor - 1)/d}^{t} d \cdot \Delta\varphi(X_{sd}) \, \mathrm{d}s \right| \leq 2 \max_{0 \leq j \leq \lfloor td \rfloor} \{|\Delta\varphi(X_j)|\}$. With this, we obtain our Doob decomposition for SGD

$$\varphi(X_{td}) = \varphi(X_0) - \int_0^t \gamma_{sd} \langle \nabla\varphi(X_{sd}), \nabla\mathcal{R}_\lambda^{\mathrm{adv}}(X_{sd}) \rangle \, \mathrm{d}s \tag{47}$$

$$+ \frac{1}{2d} \int_0^t \gamma_{sd}^2 \sum_{i=1}^{2} p_i \mathbb{E}\left[f_i'(g_{sd,i}, \epsilon)^2 \,|\, \mathcal{F}_{sd} \right] \langle \nabla^2\varphi(X_{sd}), K_i + \mu_i \mu_i^\top \rangle \, \mathrm{d}s \tag{48}$$

$$+ \sum_{j=0}^{\lfloor td \rfloor - 1} \left(\Delta\mathcal{M}_j^{\mathrm{Grad}}(\varphi) + \Delta\mathcal{M}_j^{\mathrm{Hess}}(\varphi) + \mathbb{E}\left[\mathcal{E}_j^{\mathrm{Hess}}(\varphi) \,|\, \mathcal{F}_j \right] \right) + \xi_{td}(\varphi). \tag{49}$$

We refer to Section D.4 for more details on the derivation (44). Recall from (27) the definition of AdvHSGD

$$\mathrm{d}\mathcal{X}_t = -\gamma(t) \nabla\mathcal{R}_\lambda^{\mathrm{adv}}(\mathcal{X}_t) \, \mathrm{d}t + \gamma(t) \sqrt{\frac{1}{d} \sum_{i=1}^{2} p_i \mathbb{E}_{\, v,\epsilon}[f_i'(\rho_{t,i}, \epsilon)^2](K_i + \mu_i \mu_i^\top)} \, \mathrm{d}B_t, \tag{50}$$

where the initial condition is given by $\mathcal{X}_0 = X_0$, $(B_t, t \geq 0)$ is a $\mathbb{R}^d$ standard Brownian motion and for $I \in \{1, 2\}$ and $v \sim \mathcal{N}(0, I_d)$ we have

$$\begin{aligned} \widehat{\mathcal{X}}_t &\stackrel{\mathrm{def}}{=} \mathcal{X}_t \oplus X^\star \in \mathbb{R}^{d \times 2}, \quad \mathcal{W}_t \stackrel{\mathrm{def}}{=} \widehat{\mathcal{X}}_t \oplus \mu \in \mathbb{R}^{d \times 4}, \\ \rho_{t,I} &\stackrel{\mathrm{def}}{=} \widehat{\mathcal{X}}_t^\top (\sqrt{K_I} v + \mu_I) + \left[\delta u_{t,I} \|\mathcal{X}_t\| \quad 0 \right]^\top, \\ \text{and} \quad u_{t,I} &= \mathrm{argmax}_{|u| \leq 1} \, f(\widehat{\mathcal{X}}_t^\top (\sqrt{K_I} v + \mu_I) + \left[\delta u \|\mathcal{X}_t\| \quad 0 \right]^\top, \epsilon). \end{aligned} \tag{51}$$

By Itô's lemma (Protter, 2005, Thm. 33, Chapt. 2) applied to a quadratic statistic, we obtain

$$\begin{aligned} \mathrm{d}\varphi(\mathcal{X}_t) &= \langle \nabla\varphi(\mathcal{X}_t), \mathrm{d}\mathcal{X}_t \rangle + \frac{1}{2} \langle \nabla^2\varphi(\mathcal{X}_t), (\mathrm{d}\mathcal{X}_t)^{\otimes 2} \rangle \\ &= -\gamma(t) \langle \nabla\varphi(\mathcal{X}_t), \nabla\mathcal{R}_\lambda^{\mathrm{adv}}(\mathcal{X}_t) \rangle \, \mathrm{d}t \\ &\quad + \frac{\gamma^2(t)}{2d} \sum_{i=1}^{2} p_i \mathbb{E}_{\, v,\epsilon}[f_i'(\rho_{t,i}, \epsilon)^2] \, \mathrm{Tr}(\nabla^2\varphi(\mathcal{X}_t)(K_i + \mu_i \mu_i^\top)) \, \mathrm{d}t \\ &\quad + \mathrm{d}\mathcal{M}_t^{\mathrm{AdvHSGD}}(\varphi), \end{aligned}$$

$$\text{where} \quad \mathrm{d}\mathcal{M}_t^{\mathrm{AdvHSGD}}(\varphi) = \frac{\gamma(t)}{\sqrt{d}} \left\langle \sqrt{\sum_{i=1}^{2} p_i \mathbb{E}_{\, v,\epsilon}[f_i'(\rho_{t,i}, \epsilon)^2](K_i + \mu_i \mu_i^\top)}, \nabla\varphi(\mathcal{X}_t) \otimes \mathrm{d}B_t \right\rangle \tag{52}$$

corresponds to the martingale increment for AdvHSGD. By integrating, we obtain the Doob decomposition for $\varphi(\mathcal{X}_t)$

$$\varphi(\mathcal{X}_t) = \varphi(W_0) - \int_0^t \gamma(s) \langle \nabla\varphi(\mathcal{X}_s), \nabla\mathcal{R}_\lambda^{\mathrm{adv}}(\mathcal{X}_s) \rangle \, \mathrm{d}s + \int_0^t \mathrm{d}\mathcal{M}_s^{\mathrm{AdvHSGD}}(\varphi) \tag{53}$$

$$+ \frac{1}{2d} \int_0^t \gamma^2(s) \sum_{i=1}^{2} p_i \mathbb{E}_{\, v,\epsilon}[f_i'(\rho_{s,i}, \epsilon)^2] \, \mathrm{Tr}(\nabla^2\varphi(\mathcal{X}_s)(K_i + \mu_i \mu_i^\top)) \, \mathrm{d}s. \tag{54}$$

Equipped with these Doob decompositions, we must now prove that the term (49) and the martingale increment (54) become negligible as $d \to \infty$. The other terms (47), (48) and (53) survive the limit while also being well-approximated by our deterministic system of coupled ODEs.

In Section D.5, we show that the terms (49) and (54) vanish as $d \to \infty$. Having this in mind, we finally show in Section C that SGD and AdvHSGD on S are (ε, M, T)-approximate solutions.

In the following section, we introduce the theory of approximate solutions to the integro-differential equation and an equivalent formulation of our deterministic system of ODEs in terms of statistics of the resolvents of K_1 and K_2.

C.2 Theory of Approximate Solutions to the Integro-Differential Equation

In this section, we build upon the theory of approximate solutions to integro-differential equations presented in Section 4 of Collins-Woodfin et al. (2024a) and Section 5 in Collins-Woodfin & Seroussi (2025). Recall by Assumptions 6 and Assumption 7, we have for $i = 1, 2$

$$\mathcal{R}^{\mathrm{adv}}(X) \stackrel{\mathrm{def}}{=} p_1 h_1(\widehat{B}_1(W)) + p_2 h_2(\widehat{B}_2(W)),$$
$$\text{and} \quad \mathcal{I}_i(\widehat{B}_i(W)) \stackrel{\mathrm{def}}{=} \mathbb{E}_{v,\epsilon}[(f_i'(X^\top((\sqrt{K_i}v + \mu_i)) + \bar{s}_i\delta\|X\|;\epsilon))^2],$$

where $h_i : \mathbb{R}^{4\times 4} \to \mathbb{R}$ and $I_i : \mathbb{R}^{4\times 4} \to \mathbb{R}$ are α-pseudo-Lipschitz functions with constants $L(h), L(\mathcal{I})$ respectively and h_i are continuously differentiable provided $(\widehat{B}_1(W), \widehat{B}_2(W)) \in \mathcal{U}^2$. We also recall the notation we previously introduced

$$\widehat{B}_i(W) \stackrel{\mathrm{def}}{=} \begin{bmatrix} U_i & 0 \\ 0 & \|X\|^2 \end{bmatrix}, \quad U_i(W) \stackrel{\mathrm{def}}{=} \begin{bmatrix} B_i(W) & m_i(W) \\ m_i(W)^\top & \|\mu_i\|^2 \end{bmatrix}, \quad B_i(W) \stackrel{\mathrm{def}}{=} \hat{X}^\top K_i \hat{X} \tag{55}$$

and $m_i(W) \stackrel{\mathrm{def}}{=} \hat{X}^\top \mu_i$. Here W is a placeholder value for AdvHSGD or SGD. In this section, we are interested in approximate solutions to the integro-differential equation (17).

Fundamentally, the adversarial framework requires rederiving the results presented in Collins-Woodfin et al. (2024a); Collins-Woodfin & Seroussi (2025) to account for the new structure of the det. equivs. involving $\partial_{44}h$ and $\mathscr{B}_{44}(t)$. Hence, the proofs will build upon the two papers while adding the necessary technicalities. The general idea behind the framework presented in Collins-Woodfin et al. (2024a); Collins-Woodfin & Seroussi (2025) is that both $Z(\mathcal{W}_t, z)$ and $Z(W_{\lfloor td \rfloor}, z)$ which are respectively functions of AdvHSGD and SGD are approximate solutions to (17).

Recall we previously defined the contour $\Gamma = \{w \in \mathbb{C} : |w| = \max\{1, 2\|\mathcal{K}\|_{\mathrm{op}}\}\}$ with $\|\mathcal{K}\|_{\mathrm{op}} \stackrel{\mathrm{def}}{=} \max\{\|K_1\|_{\mathrm{op}}, \|K_2\|_{\mathrm{op}}\}$ such that the distance of any element of Γ to the spectrum of K_1 and K_2 is at least $\frac{1}{2}$. Recall for $z \in \Gamma^2$, we write $z = (z_1, z_2)$ and when integrating over z_1 and z_2 simultaneously, we write for any function $f : \mathbb{C}^2 \to \mathbb{C}$

$$\oint f(z)\mathrm{D}z \stackrel{\mathrm{def}}{=} \frac{-1}{4\pi^2} \oint_{\Gamma^2} f(z)\,\mathrm{d}z_1\,\mathrm{d}z_2.$$

Recall we introduced a norm over the contour Γ^2 for a continuous function $H : \mathbb{C}^2 \to \mathbb{C}^{4\times 4}$ in Definition B.1:

$$\|H\|_\Gamma \stackrel{\mathrm{def}}{=} \sup_{z\in\Gamma^2} \|H(z)\|.$$

Remark C.1. *The results and proofs in this section will be presented for the soft label setting. Note that proofs may be adapted to the hard label setting by considering $X^\star = 0$ and very similar arguments.*

Lemma 3 in Collins-Woodfin & Seroussi (2025) shows that up to constants depending on $\|\mathcal{K}\|_{\mathrm{op}}$, the norm $\|\cdot\|_\Gamma$ applied to the statistics $Z(W_{\lfloor td \rfloor}, \cdot)$ or $\mathcal{Z}(t, \cdot)$ is proportional to the norm of the iterates. We adapt this result slightly as our contour Γ differs from their contour Ω. We omit this proof however as it is almost identical. This result is analogous to Lemma 5 in Collins-Woodfin et al. (2024a).

Lemma C.1 (Lemma 3, (Collins-Woodfin & Seroussi, 2025)). *Recall $\mathscr{N}(t) \stackrel{\mathrm{def}}{=} \oint \mathrm{Tr}(\mathcal{Z}(t, z))\,\mathrm{D}z$. Then there exists a constant $C > 0$ depending on $\|\mathcal{K}\|_{\mathrm{op}}$ and such that*

$$C \le \frac{\|Z(\mathcal{W}_t, \cdot)\|_\Gamma}{\|\mathcal{W}_t\|^2}, \frac{\|Z(W_{td}, \cdot)\|_\Gamma}{\|W_{td}\|^2}, \frac{\|\mathcal{Z}(t, \cdot)\|_\Gamma}{\mathscr{N}(t)} \le 4.$$

Using the norm $\|\cdot\|_\Gamma$, we introduce a stopping time τ_M analogous to ϑ_M in Theorem B.1.

Definition C.1 (Stopping time τ_M). *For any $M > 0$ and a continuous function $\mathscr{Z} : [0, \infty] \times \mathbb{C}^2 \to \mathbb{R}^{4\times 4}$, define*

$$\tau_M(\mathscr{Z}) \stackrel{\mathrm{def}}{=} \inf\left\{t \ge 0 : \|\mathscr{Z}(t, \cdot)\|_\Gamma > M \quad \textit{or} \quad \widehat{\mathcal{B}}_1(t) \notin \mathcal{U} \quad \textit{or} \quad \widehat{\mathcal{B}}_2(t) \notin \mathcal{U}\right\},$$

where we define

$$\mathscr{Z}(t,\cdot) \stackrel{\text{def}}{=} \begin{bmatrix} \mathscr{S}(t,\cdot) & \mathscr{M}(t,\cdot) \\ \mathscr{M}(t,\cdot)^\top & \mu^\top \mathcal{R}(z;\mathcal{K})\mu \end{bmatrix}. \tag{56}$$

with

$$\widehat{\mathcal{B}}_i(t) \stackrel{\text{def}}{=} \begin{bmatrix} \mathcal{B}_i(t) & \mathcal{M}_i(t) & 0 \\ (\mathcal{M}_i(t))^\top & \|\mu_i\|^2 & 0 \\ 0 & 0 & \mathcal{V}_{11}(t) \end{bmatrix},$$

and

$$\mathcal{B}_i(t) \stackrel{\text{def}}{=} \oint z_i \mathscr{S}(t,z)\, \mathrm{D}z, \quad \mathcal{M}_i(t) \stackrel{\text{def}}{=} \oint \mathscr{M}_i(t,z)\, \mathrm{D}z, \quad \mathcal{V}(t) \stackrel{\text{def}}{=} \oint \mathscr{S}(t,z)\, \mathrm{D}z,$$

where $\mathscr{M}_i$ denotes the ith column of $\mathscr{M}$ associated with μ_i and $\mathcal{V}_{11}$ denotes the element in the position $(1,1)$ of $\mathcal{V}$

When clear from context, we omit the dependence $\tau_M = \tau_M(\mathscr{Z})$. The stopping time τ_M differs from ϑ_M introduced in Theorem B.1 as they respectively bound $\|\mathcal{Z}(t,\cdot)\|_\Gamma$ and $\mathscr{N}(t)$. Note that Lemma C.1 implies there exists constants $c, C > 0$ such that $\vartheta_{c\cdot M} \leq \tau_M \leq \vartheta_{C\cdot M}$ so the stopping times are interchangeable. We are now ready to introduce the notion of approximate solutions to integro-differential equations.

Definition C.2 ((ε, M, T)-approximate solution to the integro-differential equation)**.** *For $M, T, \varepsilon > 0$ and a continuous function $\mathscr{Z} : [0,\infty] \times \mathbb{C}^2 \to \mathbb{R}^{4\times 4}$, we denote $\mathscr{Z}$ an (ε, M, T)-approximate solution of* (17) *if*

$$\sup_{0\leq t\leq T\wedge\tau_M} \left\| \mathscr{Z}(t,\cdot) - \mathscr{Z}(0,\cdot) - \int_0^t \mathcal{F}(\cdot, \mathscr{Z}(s,\cdot))\, \mathrm{d}s \right\|_\Gamma \leq \varepsilon,$$

where $\mathscr{Z}(0,\cdot) = W_0^\top \mathcal{R}(\cdot;\mathcal{K}) W_0$ and $W_0 = X_0 \oplus X^\star \oplus \mu$.

It is clear from the definition C.2 and the definition of $\mathcal{Z}$ in (17) that $\mathcal{Z}$ is an approximate solution with $\varepsilon = 0$. In the next subsection, we will show $Z(W_{td})$ and $Z(\mathcal{W}_t)$ are (ε, M, T)-approximate solutions to (17). Before proving these results, we study the stability of the approximate solutions of (17) analogous to Proposition 10 in Collins-Woodfin et al. (2024a) and Corollary 1 in Collins-Woodfin & Seroussi (2025).

Proposition C.1 (Stability)**.** *For all (ε, M, T)-approximate solutions $\mathscr{Z}_1$ and $\mathscr{Z}_2$, there exists a constant $C = C(\bar{\gamma}, \lambda, M, \alpha, L(h), L(\mathcal{I}), \|\mathcal{K}\|_{\text{op}}, \max_{i=1,2}\|\mu_i\|, T) > 0$ such that*

$$\sup_{0\leq t\leq T\wedge\widehat{\tau}_M} \|\mathscr{Z}_1(t,\cdot) - \mathscr{Z}_2(t,\cdot)\|_\Gamma \leq C\varepsilon,$$

where $\widehat{\tau}_M = \min\{\tau_M(\mathscr{Z}_1), \tau_M(\mathscr{Z}_2)\}$.

Proof. Since $\mathscr{Z}_1$ and $\mathscr{Z}_2$ are (ε, M, T)-approximate solutions then for $i = 1, 2$ we can write

$$\mathscr{Z}_i(t,\cdot) = \mathscr{Z}_i(0,\cdot) + \int_0^t \mathcal{F}(\cdot, \mathscr{Z}_i(s,\cdot))\, \mathrm{d}s + \varepsilon_t(\mathscr{Z}_i), \tag{57}$$

where $\varepsilon_t(\mathscr{Z}_i)$ are error terms satisfying $\sup_{0\leq t\leq T\wedge\widehat{\tau}_M} \|\varepsilon_t(\mathscr{Z}_i)\|_\Gamma \leq \varepsilon$. By construction, we have $\widehat{\tau}_M \leq \tau_M(\mathscr{Z}_1), \tau_M(\mathscr{Z}_2)$ so we will work with $\widehat{\tau}_M$ throughout the proof. Since $\mathscr{Z}_1(0,\cdot) = \mathscr{Z}_2(0,\cdot)$, proving the result amounts to showing the function $\mathcal{F}$ is Lipschitz with respect to its second input:

$$\|\mathcal{F}(\cdot, \mathscr{Z}_1(t\wedge\widehat{\tau}_M,\cdot)) - \mathcal{F}(\cdot, \mathscr{Z}_2(t\wedge\widehat{\tau}_M,\cdot))\|_\Gamma \leq L(\mathcal{F})\|\mathscr{Z}_1(t\wedge\widehat{\tau}_M,\cdot) - \mathscr{Z}_2(t\wedge\widehat{\tau}_M,\cdot)\|_\Gamma.$$

From the definition of $\mathcal{F}$, we have that

$$\begin{aligned} &\|\mathcal{F}(\cdot, \mathscr{Z}_1(t\wedge\widehat{\tau}_M,\cdot)) - \mathcal{F}(\cdot, \mathscr{Z}_2(t\wedge\widehat{\tau}_M,\cdot))\|_\Gamma \\ &\qquad \leq \|\mathcal{F}_{\mathcal{S}}(\cdot, \mathscr{Z}_1(t\wedge\widehat{\tau}_M,\cdot)) - \mathcal{F}_{\mathcal{S}}(\cdot, \mathscr{Z}_2(t\wedge\widehat{\tau}_M,\cdot))\|_\Gamma \\ &\qquad\quad + 2\|\mathcal{F}_{\mathcal{M}}(\cdot, \mathscr{Z}_1(t\wedge\widehat{\tau}_M,\cdot)) - \mathcal{F}_{\mathcal{M}}(\cdot, \mathscr{Z}_2(t\wedge\widehat{\tau}_M,\cdot))\|_\Gamma. \end{aligned} \tag{58}$$

Hence, showing $\mathcal{F}$ is Lipschitz amounts to showing $\mathcal{F}_{\mathcal{S}}$ and $\mathcal{F}_{\mathcal{M}}$ are Lipschitz. To show this, we build upon the proofs of Proposition 10 in Collins-Woodfin et al. (2024a) and account for the contributions

of the adversarial terms. Similarly to the notation introduced in Definition C.1, we will represent the structure of $\mathscr{Z}_j$ for $j = 1, 2$ as follows

$$\mathscr{Z}_j(t,\cdot) \stackrel{\text{def}}{=} \begin{bmatrix} \mathscr{S}_j(t,\cdot) & \mathscr{M}_j(t,\cdot) \\ \mathscr{M}_j(t,\cdot)^\top & \mu^\top \mathcal{R}(z;\mathcal{K})\mu \end{bmatrix}. \tag{59}$$

For simplicity of notation, we define for $j = 1, 2$

$$\mathcal{B}_{i,j}(t) \stackrel{\text{def}}{=} \oint z_i \mathscr{S}_j(t,z)\,\mathrm{D}z, \quad \mathcal{M}_{i,j}(t) \stackrel{\text{def}}{=} \oint \mathscr{M}_{i,j}(t,z)\,\mathrm{D}z, \quad \mathcal{V}_j(t) \stackrel{\text{def}}{=} \oint \mathscr{S}_j(t,z)\,\mathrm{D}z,$$
$$\text{and} \quad \widehat{\mathcal{B}}_{i,j}(t) \stackrel{\text{def}}{=} \begin{bmatrix} \mathcal{B}_{i,j}(t) & \mathcal{M}_{i,j}(t) & 0 \\ \mathcal{M}_{i,j}(t) & \|\mu_i\|^2 & 0 \\ 0 & 0 & \mathcal{V}_{j,11}(t) \end{bmatrix}. \tag{60}$$

where $\mathscr{M}_{i,j}$ denotes the ith column of $\mathscr{M}_j$ associated with μ_i and $\mathcal{V}_{j,11}$ denotes the element in the position $(1,1)$ of $\mathcal{V}_j$. We will work with $s \le \widehat{\tau}_M$ and omit the stopping time in the next steps where we show $\mathcal{F}$ is Lipschitz. First, since $\oint |q(z_i)|\,\mathrm{D}|z|$ is uniformly bounded by a constant depending on $\|\mathcal{K}\|_{\text{op}}$, it follows from the definition of the Frobenius norm

$$\begin{gathered} \|\mathscr{S}_j(t,\cdot)\|_\Gamma \le M, \quad \|z_i\mathscr{S}_j(t,\cdot)\|_\Gamma \le C(M, \|\mathcal{K}\|_{\text{op}}), \\ \|\mathscr{M}_{i,j}(t,\cdot)\|_\Gamma \le \|\mathscr{M}_j(t,\cdot)\|_\Gamma \le M, \quad \|z_i\mathscr{M}_j(t,\cdot)\|_\Gamma \le C(M, \|\mathcal{K}\|_{\text{op}}), \\ \left\| \frac{-1}{2\pi\,\mathrm{i}} \oint_\Gamma \mathscr{M}_j(t,z)\,\mathrm{d}z_i \right\|_\Gamma \le C(M, \|\mathcal{K}\|_{\text{op}}) \quad \text{and} \quad \left\| \frac{-1}{2\pi\,\mathrm{i}} \oint_\Gamma \mathscr{S}_j(t,z)\,\mathrm{d}z_i \right\|_\Gamma \le C(M, \|\mathcal{K}\|_{\text{op}}). \end{gathered} \tag{61}$$

Similarly, we have the following inequalities which also follow from linearity and the definition of the Frobenius norm

$$\begin{gathered} \|\mathscr{S}_1(t,\cdot) - \mathscr{S}_2(t,\cdot)\|_\Gamma \le \|\mathscr{Z}_1(t,\cdot) - \mathscr{Z}_2(t,\cdot)\|_\Gamma, \\ \|\mathscr{M}_{i,1}(t,\cdot) - \mathscr{M}_{i,2}(t,\cdot)\|_\Gamma \le \|\mathscr{M}_1(t,\cdot) - \mathscr{M}_2(t,\cdot)\|_\Gamma \le \|\mathscr{Z}_1(t,\cdot) - \mathscr{Z}_2(t,\cdot)\|_\Gamma, \\ \left\| \frac{-1}{2\pi\,\mathrm{i}} \oint_\Gamma \mathscr{M}_1(t,z) - \mathscr{M}_2(t,z)\,\mathrm{d}z_i \right\|_\Gamma \le C(\|\mathcal{K}\|_{\text{op}}) \|\mathscr{Z}_1(t,\cdot) - \mathscr{Z}_2(t,\cdot)\|_\Gamma, \\ \left\| \frac{-1}{2\pi\,\mathrm{i}} \oint_\Gamma \mathscr{S}_1(t,z) - \mathscr{S}_2(t,z)\,\mathrm{d}z_i \right\|_\Gamma \le C(\|\mathcal{K}\|_{\text{op}}) \|\mathscr{Z}_1(t,\cdot) - \mathscr{Z}_2(t,\cdot)\|_\Gamma. \end{gathered} \tag{62}$$

Hence, we have

$$\|\mathcal{B}_{i,j}(s)\| \le C(\|\mathcal{K}\|_{\text{op}}) \|\mathscr{S}_j(\cdot,z)\|_\Gamma \le C(\|\mathcal{K}\|_{\text{op}}) \|\mathscr{Z}_j(\cdot,z)\|_\Gamma \le C(\|\mathcal{K}\|_{\text{op}}, M). \tag{63}$$

Similarly, we obtain for $C = C(\|\mathcal{K}\|_{\text{op}}, M)$

$$\begin{gathered} \|\mathcal{M}_{i,j}(\cdot)\| \le C, \quad \|\mu_i\| \le C, \quad \text{and} \quad \|\mathcal{V}_{j,11}(s)\| \le \|\mathcal{V}_j(s)\| \le C \\ \text{from which it follows that} \quad \|\widehat{\mathcal{B}}_{i,j}(s)\| \le C. \end{gathered} \tag{64}$$

From the definitions of $H_{1,i}$ and $H_{2,i}$ in (17) and Assumption 6, ∇h_i is α-pseudo-Lipschitz so using (64) it follows for $\ell = 1, 2$

$$\begin{aligned} \|H_{\ell,i}(\widehat{\mathcal{B}}_{i,1}(s)) - H_{\ell,i}(\widehat{\mathcal{B}}_{i,2}(s))\| &\le L(h)(1 + \|\widehat{\mathcal{B}}_{i,1}(s)\|^\alpha + \|\widehat{\mathcal{B}}_{i,2}(s)\|^\alpha)\|\widehat{\mathcal{B}}_{i,1}(s) - \widehat{\mathcal{B}}_{i,2}(s)\| \\ &\le C(M, \alpha, L(h), \|\mathcal{K}\|_{\text{op}})\|\widehat{\mathcal{B}}_{i,1}(s) - \widehat{\mathcal{B}}_{i,2}(s)\|. \end{aligned} \tag{65}$$

Once again since $\oint |q(z_i)|\,\mathrm{D}|z|$ is uniformly bounded and from (62), then we have

$$\begin{gathered} \|\mathcal{B}_{i,1}(s) - \mathcal{B}_{i,2}(s)\| \le C(\|\mathcal{K}\|_{\text{op}}) \|\mathscr{Z}_1(s,\cdot) - \mathscr{Z}_2(s,\cdot)\|_\Gamma, \\ \|\mathcal{M}_{i,1}(s) - \mathcal{M}_{i,2}(s)\| \le C(\|\mathcal{K}\|_{\text{op}}) \|\mathscr{Z}_1(s,\cdot) - \mathscr{Z}_2(s,\cdot)\|_\Gamma, \\ \|\mathcal{V}_{1,11}(s) - \mathcal{V}_{2,11}(s)\| \le \|\mathcal{V}_1(s) - \mathcal{V}_2(s)\| \le C(\|\mathcal{K}\|_{\text{op}}) \|\mathscr{Z}_1(s,\cdot) - \mathscr{Z}_2(s,\cdot)\|_\Gamma. \end{gathered} \tag{66}$$

Hence, we obtain

$$\|\widehat{\mathcal{B}}_{i,1}(s) - \widehat{\mathcal{B}}_{i,2}(s)\| \le C(\|\mathcal{K}\|_{\text{op}}) \|\mathscr{Z}_1(s,\cdot) - \mathscr{Z}_2(s,\cdot)\|_\Gamma. \tag{67}$$

Hence, inserting (67) into (65), we obtain

$$\|H_{\ell,i}(\widehat{\mathcal{B}}_{i,1}(s)) - H_{\ell,i}(\widehat{\mathcal{B}}_{i,2}(s))\| \le C(M, \alpha, L(h), \|\mathcal{K}\|_{\text{op}}) \|\mathscr{Z}_1(s,\cdot) - \mathscr{Z}_2(s,\cdot)\|_\Gamma. \tag{68}$$

From Assumption 6 and (64), we obtain the bound

$$\|H_{\ell,i}(\widehat{\mathcal{B}}_{i,j}(s))\| \le CL(h)(1+\|\widehat{\mathcal{B}}_{i,j}(s)\|)^{1+\alpha} \le C(\|\mathcal{K}\|_{\mathrm{op}}, M, \alpha, L(h)). \tag{69}$$

Similarly to (65) and (69), from Assumption 6 and (64), we have

$$\begin{aligned} |\partial_{44}h_i(\widehat{\mathcal{B}}_{i,1}(s)) - \partial_{44}h_i(\widehat{\mathcal{B}}_{i,2}(s))| &\le C(M,\alpha,L(h),\|\mathcal{K}\|_{\mathrm{op}})\|\mathscr{Z}_1(s,\cdot)-\mathscr{Z}_2(s,\cdot)\|_\Gamma, \\ |\frac{\lambda}{2} + \partial_{44}h_i(\widehat{\mathcal{B}}_{i,j}(s))| &\le C(\lambda,\|\mathcal{K}\|_{\mathrm{op}},M,\alpha,L(h)). \end{aligned} \tag{70}$$

It is also clear from (59) that $\|\mu^\top \mathcal{R}(z;\mathcal{K})\mu_i\| \le \|\mu^\top \mathcal{R}(z;\mathcal{K})\mu\| \le \|\mathscr{Z}_j\|_\Gamma \le M$. Thus, it follows from (61), (62), (68), (69) and (70) and summing over i that

$$\begin{aligned} &\|\mathcal{F}_{\mathcal{M}}(\cdot,\mathscr{Z}_1(t\wedge\widehat{\tau}_M,\cdot)) - \mathcal{F}_{\mathcal{M}}(\cdot,\mathscr{Z}_2(t\wedge\widehat{\tau}_M,\cdot))\|_\Gamma \\ &\qquad \le C(\lambda,\bar{\gamma},M,\alpha,L(h),\|\mathcal{K}\|_{\mathrm{op}})\|\mathscr{Z}_1(s)-\mathscr{Z}_2(s)\|. \end{aligned} \tag{71}$$

From Assumption 7, (64) and (67), we obtain

$$|\mathcal{I}_i(\widehat{\mathcal{B}}_{i,1}(s)) - \mathcal{I}_i(\widehat{\mathcal{B}}_{i,2}(s))| \le C(L(\mathcal{I}),\|\mathcal{K}\|_{\mathrm{op}},M,\alpha)\|\mathscr{Z}_1(s,\cdot)-\mathscr{Z}_2(s,\cdot)\|_\Gamma. \tag{72}$$

As shown in Lemma A.3 of Collins-Woodfin et al. (2024a), from the Neumann series expansion of the resolvent $\sup_{z_i\in\Gamma}\|R(z_i;K_i)\|_{\mathrm{op}} \le 2$ from which it follows that $\sup_{z\in\Gamma^2}\|\mathcal{R}(z;\mathcal{K})\|_{\mathrm{op}} \le 4$. Given the previous two bounds and using the identity $K_iR(z_i;K_i) = I_d + z_iR(z_i;K_i)$, it follows

$$\begin{aligned} &\sup_{z\in\Gamma^2}\left|\frac{\gamma_s^2}{d}\operatorname{Tr}((K_1+\mu_1\mu_1^\top)\mathcal{R}(z;\mathcal{K}))\left(\mathcal{I}_i(\widehat{\mathcal{B}}_{1,1}(s)) - \mathcal{I}_i(\widehat{\mathcal{B}}_{1,2}(s))\right)\right| \\ &\le \frac{C(\bar{\gamma})}{d}\sup_{z\in\Gamma^2}(|\operatorname{Tr}(R(z_2;K_2))| + |z_1||\operatorname{Tr}(\mathcal{R}(z;\mathcal{K}))| + |\mu_1^\top\mathcal{R}(z;\mathcal{K})\mu_1|)|\mathcal{I}_i(\widehat{\mathcal{B}}_{1,1}(s)) - \mathcal{I}_i(\widehat{\mathcal{B}}_{1,2}(s))| \\ &\le C(\bar{\gamma},L(\mathcal{I}),\|\mathcal{K}\|_{\mathrm{op}},\max_{i=1,2}\|\mu_i\|,M,\alpha)\|\mathscr{Z}_1(s,\cdot)-\mathscr{Z}_2(s,\cdot)\|_\Gamma. \end{aligned} \tag{73}$$

The same result holds with K_2, μ_2, $\widehat{\mathcal{B}}_{2,1}(s)$ and $\widehat{\mathcal{B}}_{2,1}(s)$. Thus, it follows from (61), (62), (68), (69), (70) and (73) and summing over i that

$$\begin{aligned} &\|\mathcal{F}_{\mathcal{S}}(\cdot,\mathscr{Z}_1(t\wedge\widehat{\tau}_M,\cdot)) - \mathcal{F}_{\mathcal{S}}(\cdot,\mathscr{Z}_2(t\wedge\widehat{\tau}_M,\cdot))\|_\Gamma \\ &\qquad \le C(\lambda,\bar{\gamma},M,\alpha,L(h),L(\mathcal{I}),\|\mathcal{K}\|_{\mathrm{op}})\|\mathscr{Z}_1(s)-\mathscr{Z}_2(s)\|. \end{aligned} \tag{74}$$

Hence, combining (71) and (74), we have shown $\mathcal{F}$ is $L(\mathcal{F})$-Lipschitz with respect to its second input where $L(\mathcal{F}) = C(\bar{\gamma},\lambda,M,\alpha,L(h),L(\mathcal{I}),\|\mathcal{K}\|_{\mathrm{op}},\max_{i=1,2}\|\mu_i\|)$. It thus follows

$$\begin{aligned} \sup_{0\le t\le T\wedge\widehat{\tau}_M}\|\mathscr{Z}_1(t,\cdot)-\mathscr{Z}_2(t,\cdot)\|_\Gamma &\le 2\varepsilon + \sup_{0\le t\le T\wedge\widehat{\tau}_M}\int_0^t \|\mathcal{F}(\cdot,\mathscr{Z}_1(s,\cdot)) - \mathcal{F}(\cdot,\mathscr{Z}_2(s,\cdot))\|_\Gamma\,\mathrm{d}s \\ &\le 2\varepsilon + \sup_{0\le t\le T}\int_0^t \|\mathcal{F}(\cdot,\mathscr{Z}_1(s\wedge\widehat{\tau}_M,\cdot)) - \mathcal{F}(\cdot,\mathscr{Z}_2(s\wedge\widehat{\tau}_M,\cdot))\|_\Gamma\,\mathrm{d}s \\ &\le 2\varepsilon + L(\mathcal{F})\int_0^t \|\mathscr{Z}_1(s\wedge\widehat{\tau}_M,\cdot)-\mathscr{Z}_2(s\wedge\widehat{\tau}_M,\cdot)\|_\Gamma\,\mathrm{d}s \\ &\le 2\varepsilon + L(\mathcal{F})\int_0^t \sup_{0\le u\le s\wedge\widehat{\tau}_M}\|\mathscr{Z}_1(u,\cdot)-\mathscr{Z}_2(u,\cdot)\|_\Gamma\,\mathrm{d}s. \end{aligned} \tag{75}$$

Applying Grönwall's inequality, we obtain

$$\sup_{0\le t\le T\wedge\widehat{\tau}_M}\|\mathscr{Z}_1(t,\cdot)-\mathscr{Z}_2(t,\cdot)\|_\Gamma \le 2\varepsilon e^{L(\mathcal{F})T}, \tag{76}$$

from which the result follows. □

Proposition C.1 can be extended to any function $\varphi(X)$ satisfying the following Assumption 12.

Assumption 12. *Fix $\ell \ge 1$. Suppose the statistic $\varphi : \mathbb{R}^d \to \mathbb{R}$ satisfies the composite structure*

$$\varphi(X) = g\left(\{W^\top q_j(K_1,K_2)W\}_{j=1}^\ell\right),$$

where $g : \mathbb{R}^{\ell\times 5\times 5} \to \mathbb{R}$ is α-pseudo-Lipschitz on $(\mathcal{U})^\ell$ and $\{q_j\}_{j=1}^\ell$ are polynomials.

This result is analogous to Proposition 11 in Collins-Woodfin et al. (2024a) and Proposition 7 in Collins-Woodfin & Seroussi (2025).

Proposition C.2. *Fix $\ell \geq 1$ and suppose $\varphi : \mathbb{R}^d \to \mathbb{R}$ satisfies Assumption 12. Let $\mathscr{Z}_1$ and $\mathscr{Z}_2$ be (ε, M, T)-approximate solutions. Then there exists a positive constant C such that*

$$\sup_{0\leq t\leq T\wedge\widehat{\tau}_M}\left|g\left(\left\{\oint q_j(z)\mathscr{Z}_1(t,z)\,\mathrm{D}z\right\}_{j=1}^{\ell}\right)-g\left(\left\{\oint q_j(z)\mathscr{Z}_2(t,z)\,\mathrm{D}z\right\}_{j=1}^{\ell}\right)\right|\leq C\varepsilon,$$

where $C = C(\bar{\gamma}, T, q, \lambda, M, \alpha, L(h), L(g), \|K\|_{\mathrm{op}})$ and $\widehat{\tau}_M = \min\{\tau_M(\mathscr{Z}_1), \tau_M(\mathscr{Z}_2)\}$.

Proof. The case $\ell = 1$ follows an almost identical argument to the proof Proposition 11 in Collins-Woodfin et al. (2024a). The proof $\ell > 1$ follows immediately from pseudo-Lipschitzness of g. □

As mentioned previously, in Section D.5, we show for each fixed $z \in \Gamma^2$ that the terms (49) and (54) vanish as $d \to \infty$ with overwhelming probability. From the definition C.2, we must show these results hold simultaneously for each $z \in \Gamma^2$. Since union bounds only hold over countable sets, we use a net argument similar to Collins-Woodfin et al. (2024a); Collins-Woodfin & Seroussi (2025).

Recall we defined the contour $\Gamma = \{w \in \mathbb{C} : |w| = \max(1, 2\|\mathcal{K}\|_{\mathrm{op}})\}$. For a fixed $\xi > 0$, we build a $d^{-\xi}$-mesh of $\Gamma_\xi \subset \Gamma$ such that for every $w \in \Gamma$, there exists a $\bar{w} \in \Gamma_\xi$ with $|w - \bar{w}| < d^{-\xi}$, with Γ_ξ having cardinality $|\Gamma_\xi| = C(|\Gamma|)d^{\xi}$. Here $|\Gamma|$ denotes the contour length and $|\Gamma_\xi|$ denotes the cardinality. We aso introduce the following resolvent Lemma from Collins-Woodfin & Seroussi (2025).

Lemma C.2 (Lemma 4 (Collins-Woodfin & Seroussi, 2025))**.** *For $z, \bar{z} \in \mathbb{C}^2 \setminus (\mathrm{spec}(K_1) \times \mathrm{spec}(K_2))$, the following formula holds*

$$\|\mathcal{R}(z;\mathcal{K}) - \mathcal{R}(\bar{z};\mathcal{K})\|_{\mathrm{op}} \leq \sum_{i=1}^{2}\left(\prod_{j=1}^{i}\|R(z_i;K_i)\|_{\mathrm{op}}\right)|z_i - \bar{z}_i|\left(\prod_{j=i}^{2}\|R(\bar{z}_i;K_i)\|_{\mathrm{op}}\right).$$

Note that the normal resolvent $R(z_i; K_i) = (K_i - z_i I_d)^{-1}$ satisfies a similar identity for $z_i, \bar{z}_i \in \mathbb{C} \setminus \mathrm{spec}(K_i)$

$$\|R(z_i;K_i) - R(\bar{z}_i;K_i)\|_{\mathrm{op}} \leq |z_i - \bar{z}_i|\|R(z_i;K_i)\|_{\mathrm{op}}\|R(\bar{z}_i;K_i)\|_{\mathrm{op}}. \tag{77}$$

We build upon Lemma A.1 in Collins-Woodfin et al. (2024a) and Lemma 5 in Collins-Woodfin & Seroussi (2025).

Lemma C.3 (Net argument)**.** *Fix $T, M > 0, \xi > 0$. Suppose Γ_ξ is a $d^{-\xi}$ mesh of Γ with $|\Gamma_\xi| = Cd^{\xi}$ for some $C > 0$. Let the function $Z(t, z) = Z(W_{td}, z)$ or $Z(t, z) = Z(\mathcal{W}_t, z)$ satisfy*

$$\sup_{0\leq t\leq(T\wedge\tau_M)}\|Z(t,\cdot) - Z(0,\cdot) - \int_0^t \mathcal{F}(\cdot, Z(s,\cdot))\,\mathrm{d}s\|_{\Gamma_\xi} \leq \varepsilon \tag{78}$$

with τ_M as in Definition C.1. Then Z is an $(\varepsilon + Cd^{-\xi}, M, T)$-approximate solution to the integro-differential equation, that is,

$$\sup_{0\leq t\leq(T\wedge\tau_M)}\|Z(t,\cdot) - Z(0,\cdot) - \int_0^t \mathcal{F}(\cdot, Z(s,\cdot))\,\mathrm{d}s\|_{\Gamma} \leq \varepsilon + Cd^{-\xi},$$

where $C = C(\bar{\gamma}, \alpha, L(\mathcal{I}), M, \|\mathcal{K}\|_{\mathrm{op}}, L(h), \lambda, \max_{i=1,2}\|\mu_i\|, T) > 0$.

Proof. We will prove the lemma for $Z(t, z) \stackrel{\mathrm{def}}{=} Z(\mathcal{W}_t, z)$ as the result follows for SGD with a similar argument. We will also work with the stopped process $Z(t, z) \stackrel{\mathrm{def}}{=} Z(t \wedge \tau_M, z)$ and omit the dependence on the stopping time for notational convenience. For a fixed $z \in \Gamma^2$ with $\bar{z} \in \Gamma_\xi^2$ such

that $|z_i - \bar{z}_i| < d^{-\xi}$ for $i = 1, 2$, we have

$$\begin{aligned}\big\|Z(t,z)-Z(0,z) - \int_0^t \mathcal{F}(z, Z(s,\cdot))\,\mathrm{d}s\big\| \\ \leq \|Z(t,z) - Z(t,\bar{z})\| + \|Z(0,z) - Z(0,\bar{z})\| \\ + \int_0^t \|\mathcal{F}(z, Z(s,\cdot)) - \mathcal{F}(\bar{z}, Z(s,\cdot))\|\,\mathrm{d}s \\ + \big\|Z(t,\bar{z}) - Z(0,\bar{z}) - \int_0^t \mathcal{F}(\bar{z}, Z(s,\cdot))\,\mathrm{d}s\big\|.\end{aligned} \tag{79}$$

We will bound each of these terms separately. The last term is bounded by ε by the assumption (78) stated in the lemma. To bound the other terms, we first introduce some useful identities. From Proposition C.1, we have

$$\sup_{z_i \in \Gamma} \|R(z_i, K_i)\|_{\mathrm{op}} \leq 2 \quad \text{and} \quad \sup_{z \in \Gamma^2} \|\mathcal{R}(z, \mathcal{K})\|_{\mathrm{op}} \leq 4. \tag{80}$$

For the first term, using Lemma C.2, (80) and recalling that we are working with the stopped process we obtain

$$\begin{aligned}\|Z(t,z) - Z(t,\bar{z})\| &\leq \|\mathcal{W}_t\|^2 \|\mathcal{R}(z;\mathcal{K}) - \mathcal{R}(\bar{z};\mathcal{K})\|_{\mathrm{op}} \\ &\leq C \left\| \oint Z(t,\tilde{z})\,\mathrm{D}\tilde{z} \right\| \sum_{i=1}^{2} |z_i - \bar{z}_i| \\ &\leq C d^{-\xi} \oint \|Z(t,\tilde{z})\|\,\mathrm{D}|\tilde{z}| \\ &\leq C(\|\mathcal{K}\|_{\mathrm{op}}, M) d^{-\xi},\end{aligned} \tag{81}$$

where we used Cauchy's integral formula and the arc length $|\Gamma| \leq C(\|\mathcal{K}\|_{\mathrm{op}})$. The same argument can be used to bound the second term in (79). It remains to bound the third term. We have for any $s \leq t \wedge \tau_M$

$$\begin{aligned}\|\mathcal{F}(z, Z(s,\cdot)) - \mathcal{F}(\bar{z}, Z(s,\cdot))\| \leq \|\mathcal{F}_{\mathcal{S}}(z, Z(s,\cdot)) - \mathcal{F}_{\mathcal{S}}(\bar{z}, Z(s,\cdot))\| \\ + 2\|\mathcal{F}_{\mathcal{M}}(z, Z(s,\cdot)) - \mathcal{F}_{\mathcal{M}}(\bar{z}, Z(s,\cdot))\|.\end{aligned} \tag{82}$$

We will use an analogous notation for $\widehat{\mathcal{B}}_i$ as defined in Definition C.1 in terms of $Z(t,z)$, $S(t,z)$, $M(t,z)$. Before bounding the terms on the right-hand side of (82), we derive some preliminary bounds. By Assumption 7 and the bound derived in (64), we have

$$|\mathcal{I}_i(\widehat{\mathcal{B}}_i(s))| \leq C L(\mathcal{I})(1 + \|\widehat{\mathcal{B}}_i(s)\|)^{1+\alpha} \leq C(\alpha, L(\mathcal{I}), M, \|\mathcal{K}\|_{\mathrm{op}}). \tag{83}$$

From (81), it follows that

$$\begin{aligned}\|z_i S(s,z) - \bar{z}_i S(s,\bar{z})\| \leq |z_i - \bar{z}_i|\|S(s,z)\| + |\bar{z}_i|\|S(s,z) - S(s,\bar{z})\| \leq C(\|\mathcal{K}\|_{\mathrm{op}}, M) d^{-\xi}, \\ \text{and similarly} \quad \|z_i M(s,z) - \bar{z}_i M(s,\bar{z})\| \leq C(\|\mathcal{K}\|_{\mathrm{op}}, M) d^{-\xi}.\end{aligned} \tag{84}$$

From (81), we also obtain

$$\begin{aligned}&\left\| \frac{-1}{2\pi\,\mathrm{i}} \oint_{\Gamma} S(t,(w,z_2)) - S(t,(w,\bar{z}_2))\,\mathrm{d}w \right\| \\ &\leq C(\|\mathcal{K}\|_{\mathrm{op}}) \sup_{w \in \Gamma} \|S(t,(w,z_2)) - S(t,(w,\bar{z}_2))\| \leq C(\|\mathcal{K}\|_{\mathrm{op}}, M) d^{-\xi}.\end{aligned} \tag{85}$$

It is clear that this hold if we integrate over the second position in $z = (z_1, z_2)$. We also bound the trace terms in (82). From the construction of Γ_ξ, we have from Lemma C.2 and the resolvent bounds in (80)

$$\begin{aligned}\frac{1}{d}|\operatorname{Tr}((K_i + \mu_i \mu_i^\top)(\mathcal{R}(z;\mathcal{K}) - \mathcal{R}(\bar{z};\mathcal{K}))| &\leq (\|\mathcal{K}\|_{\mathrm{op}} + \max_{j=1,2} \|\mu_j\|^2) \|\mathcal{R}(z;\mathcal{K}) - \mathcal{R}(\bar{z};\mathcal{K})\|_{\mathrm{op}} \\ &\leq C(\|\mathcal{K}\|_{\mathrm{op}}, \max_{j=1,2} \|\mu_j\|^2) d^{-\xi}.\end{aligned} \tag{86}$$

Given the definition of $\mathcal{F}_{\mathcal{S}}$ in (17), we have

$$\begin{aligned}\|\mathcal{F}_{\mathcal{S}}(z, Z(s,\cdot)) - \mathcal{F}_{\mathcal{S}}(\bar{z}, Z(s,\cdot))\| \leq \frac{\bar{\gamma}^2}{d}\sum_{i=1}^{2} p_i |\mathcal{I}_i(\widehat{\mathcal{B}}_i(s))|| \operatorname{Tr}((K_i + \mu_i\mu_i^\top)(\mathcal{R}(z;\mathcal{K}) - \mathcal{R}(\bar{z};\mathcal{K}))| \\ + 4\bar{\gamma}\sum_{i=1}^{2} p_i \|H_{2,i}(\widehat{\mathcal{B}}_i(s))\|\|M_i(s,z) - M_i(s,\bar{z})\| \\ + 4\bar{\gamma}\, p_1\|H_{1,1}(\widehat{\mathcal{B}}_1(s))\| \left\| \frac{-1}{2\pi\,\mathrm{i}} \oint_{\Gamma} (S(s,(w,z_2)) - S(s,(w,\bar{z}_2)))\,\mathrm{d}w \right\| \\ + 4\bar{\gamma}\, p_2\|H_{1,2}(\widehat{\mathcal{B}}_2(s))\| \left\| \frac{-1}{2\pi\,\mathrm{i}} \oint_{\Gamma} (S(s,(z_1,w)) - S(s,(\bar{z}_1,w)))\,\mathrm{d}w \right\| \\ + 4\bar{\gamma}\sum_{i=1}^{2} p_i \|H_{1,i}(\widehat{\mathcal{B}}_i(s))\|\|z_i S(s,z) - \bar{z}_i S(s,\bar{z})\| \\ + 8\bar{\gamma}\sum_{i=1}^{2} p_i |\lambda + \partial_{44} h_i(\widehat{\mathcal{B}}_i(s))|\|S(s,z) - S(s,\bar{z})\|,\end{aligned} \tag{87}$$

Hence, from (69), (70), (81), (83), (84), (85) and (86) we have

$$\|\mathcal{F}_{\mathcal{S}}(z, Z(s,\cdot)) - \mathcal{F}_{\mathcal{S}}(\bar{z}, Z(s,\cdot))\| \leq C(\bar{\gamma}, \lambda, \alpha, M, \|\mathcal{K}\|_{\mathrm{op}}, L(h), L(\mathcal{I}), \max_{j=1,2}\|\mu_i\|) d^{-\xi}. \tag{88}$$

Using a similar argument, we obtain for $\mathcal{F}_{\mathcal{M}}$

$$\|\mathcal{F}_{\mathcal{M}}(z, Z(s,\cdot)) - \mathcal{F}_{\mathcal{M}}(\bar{z}, Z(s,\cdot))\| \leq C(\bar{\gamma}, \lambda, \alpha, M, \|\mathcal{K}\|_{\mathrm{op}}, L(h), \max_{j=1,2}\|\mu_i\|) d^{-\xi}. \tag{89}$$

Combining (88), (89) and integrating from 0 to t, we obtain

$$\int_0^t \|\mathcal{F}(z, S(s,\cdot)) - \mathcal{F}(\bar{z}, S(s,\cdot))\|\,\mathrm{d}s \leq C(\bar{\gamma}, \lambda, \alpha, M, \|\mathcal{K}\|_{\mathrm{op}}, L(h), L(\mathcal{I}), \max_{j=1,2}\|\mu_i\|) d^{-\xi} t. \tag{90}$$

Taking the supremum over $z \in \Gamma^2$ and the supremum over $0 \leq s \leq t \wedge \tau_M$ in (79) and from the bounds (81), (90) and the assumption stated in (78), the result follows. □

As mentioned previously, in order to show SGD and AdvHSGD are approximates solutions to (17), we must control the error terms which arise in the Doob decompositions of AdvHSGD and SGD respectively. In Definition C.1 and Theorem B.1, we introduced the stopping times τ_M and ϑ_M. We formally introduce ϑ_M for fixed $M > 0$:

$$\begin{aligned}\vartheta_M &\stackrel{\mathrm{def}}{=} \inf\{t \geq 0 : \|W_{td}\|^2 > M \quad \text{or} \quad \widehat{B}_1(W_{td}) \notin \mathcal{U} \quad \text{or} \quad \widehat{B}_2(W_{td}) \notin \mathcal{U}\} \\ \text{or} \quad \vartheta_M &\stackrel{\mathrm{def}}{=} \inf\{t \geq 0 : \|\mathcal{W}_t\|^2 > M \quad \text{or} \quad \widehat{B}_1(\mathcal{W}_t) \notin \mathcal{U} \quad \text{or} \quad \widehat{B}_2(\mathcal{W}_t) \notin \mathcal{U}\},\end{aligned} \tag{91}$$

Here we overload the notation so that ϑ_M can be applied to either the iterates of SGD or AdvHSGD. Which stopping time is used will be clear from context. By Lemma C.1, there exists constants $c, C > 0$ such that $\vartheta_{c\cdot M} \leq \tau_M \leq \vartheta_{C\cdot M}$ implying the stopping times τ_M and ϑ_M are interchangeable. When clear from context, we will omit the subscript $\vartheta = \vartheta_M$. Hence, in the following proofs of Propositions C.3 and C.4, we will instead work with the stopped processes $W_{td}^{\vartheta} = W_{(t\wedge\vartheta)d}$ and $\mathcal{W}_t^{\vartheta} = \mathcal{W}_{t\wedge\vartheta}$. The following proof builds upon Proposition 12 in Collins-Woodfin et al. (2024a) and Proposition 5 in Collins-Woodfin & Seroussi (2025).

Proposition C.3 (SGD is an approximate solution)**.** *Fix $T, M > 0$ and $0 < \varepsilon < \frac{1}{2}$. Then $Z(W_{td}, z)$ is an $(Cd^{-\varepsilon}, M, T)$-approximate solution w.o.p, that is*

$$\sup_{0\leq t\leq (T\wedge\tau_M)} \|Z(W_{td},\cdot) - Z(W_0,\cdot) - \int_0^t \mathcal{F}(\cdot, Z(W_{sd},\cdot))\,\mathrm{d}s\|_{\Gamma} \leq Cd^{-\varepsilon} \quad w.o.p.$$

for some constant $C > 0$ independent of d.

Proof. Given that the elements of $Z(W_{td}, \cdot)$ are at most quadratic functions of W_{td}, applying (47), (48) and (49) position-wise, we obtain the following Doob decomposition for $Z(W_{td}, \cdot)$ for a fixed $z \in \Gamma^2$

$$\begin{aligned} Z(W_{td}, z) = W_0^\top \mathcal{R}(z; \mathcal{K}) W_0 + \int_0^t \mathcal{F}(z, Z(W_{sd}, \cdot))\, \mathrm{d}s \\ + \sum_{j=0}^{\lfloor td \rfloor - 1} \left(\Delta \mathcal{M}_j^{\text{Grad}}(Z) + \Delta \mathcal{M}_j^{\text{Hess}}(Z) + \mathbb{E}\left[\mathcal{E}_j^{\text{Hess}}(Z) \,|\, \mathcal{F}_j \right] \right) + \xi_{td}(Z). \end{aligned} \tag{92}$$

See Appendix D.4 for a detailed derivation on the Doob decomposition for an individual quadratic statistic of W_{td} and Appendix D.3 for a detailed derivation of (92) but replacing $\mathcal{W}_t$ by W_{td} and ρ_t by g_{td}. Taking the supremum from $0 \le t \le T \wedge \tau_M$, we then have

$$\begin{aligned} \sup_{0 \le t \le T \wedge \tau_M} \| Z(W_{td}, \cdot) - Z(W_0, \cdot) - \int_0^t \mathcal{F}(\cdot, Z(W_{sd}, \cdot))\, \mathrm{d}s \| \\ \le \sup_{0 \le t \le T \wedge \vartheta_{CM}} \| \mathcal{M}_{td}^{\text{Grad}}(Z(\cdot, z)) \| + \sup_{0 \le t \le T \wedge \vartheta_{CM}} \| \mathcal{M}_{td}^{\text{Hess}}(Z(\cdot, z)) \| \\ + \sup_{0 \le t \le T \wedge \vartheta_{CM}} \| \sum_{j=0}^{\lfloor td \rfloor - 1} \mathbb{E}\left[\mathcal{E}_j^{\text{Hess}}(Z) \,|\, \mathcal{F}_j \right] \| + \sup_{0 \le t \le T \wedge \tau_M} \| \xi_{td}(Z) \|. \end{aligned} \tag{93}$$

where we used the fact that $\tau_M \le \vartheta_{CM}$ for some constant $C > 0$. We defer the proof the bounds on $\mathcal{M}_{td}^{\text{Grad}}$ and $\mathcal{M}_{td}^{\text{Hess}}$ to Proposition D.2 and Proposition D.3 in Section D.5.1. We also defer the proof the bound on the Hessian error term to Proposition D.4 in Section D.5.3.

Fix a constant $\beta > 0$ and $\Gamma_\beta \subset \Gamma$ such that there exists a $\bar{z} \in \Gamma_\xi^2$ with $|z_i - \bar{z}_i| < d^{-\beta}$ for $i = 1, 2$ and the cardinality $|\Gamma_\beta| = C(\|\mathcal{K}\|_{\text{op}}) d^\beta$. From Proposition D.2 and Proposition D.3 and union bounds since the cardinality of Γ_β is polynomial in d, we have that for any $\zeta > 0$

$$\sup_{z \in \Gamma_\beta} \sup_{0 \le t \le T} \| \mathcal{M}_{(t \wedge \vartheta_{CM})d}^{\text{Grad}}(Z(\cdot, z)) \| < d^{-\frac{1}{2} + \zeta}, \quad w.o.p. \tag{94}$$

and

$$\sup_{z \in \Gamma_\beta} \sup_{0 \le t \le T} \| \mathcal{M}_{(t \wedge \vartheta_{CM})d}^{\text{Hess}}(Z(\cdot, z)) \| < d^{-\frac{1}{2} + \zeta}, \quad w.o.p. \tag{95}$$

From Proposition D.4, the triangle inequality and a union bound over Γ_β, we also have

$$\sup_{z \in \Gamma_\beta} \sup_{0 \le t \le T} \sum_{j=0}^{\lfloor (t \wedge \vartheta_{CM})d \rfloor - 1} \| \mathbb{E}\left[\mathcal{E}_j^{\text{Hess}}(Z) \,|\, \mathcal{F}_j \right] \| \le C(L(f))^2 d^{-1/2 + \zeta}. \tag{96}$$

It remains to bound the last error term ξ_{td}. We have

$$\begin{aligned} \sup_{0 \le t \le T \wedge \tau_M} \| \xi_{td}(Z) \|_\Gamma \le \frac{2}{d} \sup_{0 \le t \le T \wedge \tau_M} \| \Delta(Z_{td}) \|_\Gamma = \frac{2}{d} \sup_{0 \le t \le T \wedge \tau_M} \| \mathcal{F}(\cdot, Z(W_{td}, \cdot)) \|_\Gamma \\ \le \frac{2}{d} \sup_{0 \le t \le T \wedge \tau_M} (\| \mathcal{F}_{\mathcal{S}}(\cdot, Z(W_{td}, \cdot)) \|_\Gamma + 2 \| \mathcal{F}_{\mathcal{M}}(\cdot, Z(W_{td}, \cdot)) \|_\Gamma). \end{aligned} \tag{97}$$

We will use an analogous notation for $\widehat{\mathcal{B}}_i$ as defined in (60) in terms of $Z(t,z)$, $S(t,z)$, $M(t,z)$ and drop the subscript j as we work with a single approximate solution. For $0 \le t \le T \wedge \tau_M$, we have

$$\begin{aligned}\|\mathcal{F}_{\mathcal{S}}(\cdot, Z(W_{td},\cdot))\|_\Gamma &\le C(\bar\gamma, \|\mathcal{K}\|_{\mathrm{op}}, M)\sum_{i=1}^2 p_i \|H_{1,i}(\widehat{\mathcal{B}}_i(W_{td}))\|(1+\sup_{z\in\Gamma^2}\|z_i S(W_{td},z)\|)\\ &+ C(\bar\gamma)\sum_{i=1}^2 p_i \|H_{2,i}(\widehat{\mathcal{B}}_i(W_{td}))\|\|M_i(W_{td},\cdot)\|_\Gamma\\ &+ \frac{C(\bar\gamma)}{d}\sum_{i=1}^2 p_i \sup_{z\in\Gamma}|\operatorname{Tr}((K_i+\mu_i\mu_i^\top)\mathcal{R}(z;\mathcal{K}))|\cdot|\mathcal{I}_i(\widehat{\mathcal{B}}_i(W_{td}))| \qquad (98)\\ &+ C(\bar\gamma)\sum_{i=1}^2 p_i \|H_{1,i}(\widehat{\mathcal{B}}_i(W_{td}))\|\sup_{z\in\Gamma^2}\|z_i Z(W_{td},z)\|\\ &+ C(\bar\gamma)\sum_{i=1}^2 p_i |\lambda+\partial_{44}h_i(\widehat{\mathcal{B}}_i(W_{td}))|\|S(W_{td},\cdot)\|_\Gamma.\end{aligned}$$

Using a similar argument as (86) from (80), we have

$$\frac{1}{d}|\operatorname{Tr}((K_i+\mu_i\mu_i^\top)\mathcal{R}(z;\mathcal{K}))| \le (\|\mathcal{K}\|_{\mathrm{op}}+\max_{j=1,2}\|\mu_j\|^2)\|\mathcal{R}(z;\mathcal{K})\|_{\mathrm{op}} \le C(\|\mathcal{K}\|_{\mathrm{op}}, \max_{j=1,2}\|\mu_j\|^2). \qquad (99)$$

From Assumption 6, Assumption 7, since $\sup_{z\in\Gamma^2}\|z_i Z(W_{td},\cdot)\|_\Gamma \le C(\|\mathcal{K}\|_{\mathrm{op}}, M)$ and from (69), (70), (83) and (99), it follows that

$$\|\mathcal{F}_{\mathcal{S}}(\cdot, Z(W_{td},\cdot))\|_\Gamma \le C(\bar\gamma,\lambda,\alpha,M,\|\mathcal{K}\|_{\mathrm{op}}, L(h), L(\mathcal{I}), \max_{j=1,2}\|\mu_i\|). \qquad (100)$$

Using a similar argument, we obtain

$$\|\mathcal{F}_{\mathcal{M}}(\cdot, Z(W_{td},\cdot))\|_\Gamma \le C(\bar\gamma,\lambda,\alpha,M,\|\mathcal{K}\|_{\mathrm{op}}, L(h), \max_{j=1,2}\|\mu_i\|). \qquad (101)$$

Hence, there exist a positive constant $C = C(\bar\gamma,\lambda,\alpha,M,\|\mathcal{K}\|_{\mathrm{op}}, L(h), L(\mathcal{I}), \max_{j=1,2}\|\mu_i\|)$ such that

$$\|\mathcal{F}(\cdot, Z(W_{td},\cdot))\|_\Gamma \le C$$

from which it follows

$$\sup_{0\le t\le T\wedge\tau_M}\|\xi_{td}(S)\|_\Gamma \le Cd^{-1} \quad a.s. \qquad (102)$$

Combining all these bounds, we then obtain from a union over Γ_β^2

$$\sup_{0\le t\le T\wedge\tau_M}\|Z(W_{td},\cdot) - Z(W_0,\cdot) - \int_0^t \mathcal{F}(\cdot, Z(W_{sd},\cdot))\,\mathrm{d}s\|_{\Gamma_\beta} \le Cd^{\zeta-\frac{1}{2}} \quad w.o.p. \qquad (103)$$

The result follows from the net argument presented in Lemma C.3 with $\zeta = \frac{1}{2} - \beta$ for any $\beta \in (0, \frac{1}{2})$. □

The proof Proposition C.4 follows a similar argument to the proof Proposition C.3 and builds upon Proposition 13 in Collins-Woodfin et al. (2024a).

Proposition C.4 (AdvHSGD is an approximate solution)**.** *Fix $T, M > 0$ and $0 < \varepsilon < \frac{1}{2}$. Then $Z(\mathcal{W}_t, z)$ is an $(Cd^{-\varepsilon}, M, T)$-approximate solution w.o.p, that is*

$$\sup_{0\le t\le(T\wedge\tau_M)}\|Z(\mathcal{W}_t,\cdot) - Z(\mathcal{W}_0,\cdot) - \int_0^t \mathcal{F}(\cdot, Z(\mathcal{W}_s,\cdot))\,\mathrm{d}s\|_\Gamma \le Cd^{-\varepsilon} \quad w.o.p.$$

for some constant $C > 0$.

Proof. Given that the elements of $Z(\mathcal{W}_t,\cdot)$ are quadratics, from Itô's lemma and (53), we obtain the Doob decomposition for $Z(\mathcal{W}_t,\cdot)$

$$Z(\mathcal{W}_t, z) = W_0^\top \mathcal{R}(z;\mathcal{K})W_0 + \int_0^t \mathcal{F}(z, Z(\mathcal{W}_s,\cdot))\,\mathrm{d}s + \int_0^t \mathrm{d}\mathcal{M}_s^{\mathrm{AdvHSGD}}(Z(\mathcal{W}_s,\cdot)). \qquad (104)$$

See Appendix D.3 for a detailed derivation. Showing $Z(\mathcal{W}_t, \cdot)$ is an approximate solution is equivalent to bounding the martingale term. It then follows that for any $z \in \Gamma^2$

$$\begin{aligned}\sup_{0\le t\le (T\wedge\tau_M)} \|Z(\mathcal{W}_t, z) - Z(W_0, z) - \int_0^t \mathcal{F}(\cdot, Z(\mathcal{W}_s, \cdot))\, \mathrm{d}s\| \\ \le \sup_{0\le t\le \vartheta_{CM}\wedge T} \|\mathcal{M}_s^{\mathrm{AdvHSGD}}(Z(\mathcal{W}_s, z))\|\end{aligned} \tag{105}$$

where we used the fact that $\tau_M \le \vartheta_{CM}$ for some constant $C > 0$. We defer the proof the bound on $\mathcal{M}_s^{\mathrm{AdvHSGD}}$ to Proposition D.1 in Section D.5.1. Fix a constant $\xi > 0$ and $\Gamma_\xi \subset \Gamma$ such that there exists a $\bar{z} \in \Gamma_\xi^2$ such that $|z_i - \bar{z}_i| < d^{-\xi}$ for $i = 1, 2$ and the cardinality $|\Gamma_\xi| = C(\|\mathcal{K}\|_{\mathrm{op}})d^{\xi}$. From Proposition D.1 and a union bound since the cardinality of Γ_ξ is polynomial in d, we have that for any $\zeta > 0$

$$\sup_{z\in\Gamma_\xi}\sup_{0\le t\le T} \|\mathcal{M}_{t\wedge\vartheta}^{\mathrm{AdvHSGD}}(Z(\cdot, z))\| \le CL(f)\ d^{\zeta/2-1/2} \quad w.o.p. \tag{106}$$

It then follows that

$$\sup_{0\le t\le (T\wedge\tau_M)} \|Z(\mathcal{W}_t, \cdot) - Z(W_0, \cdot) - \int_0^t \mathcal{F}(\cdot, Z(\mathcal{W}_s, \cdot))\, \mathrm{d}s\|_{\Gamma_\xi} \le CL(f)\ d^{\zeta/2-1/2} \quad w.o.p. \tag{107}$$

The result follows from the net argument presented in Lemma C.3 with $\zeta = 1 - 2\xi$ for any $\xi \in (0, \frac{1}{2})$. □

Through Proposition C.1 on the stability of approximate solutions, we finally formalize the relationship between SGD, AdvHSGD and $\mathcal{Z}(t, \cdot)$ the solution to (17).

Theorem C.1. *For all $\varepsilon \in (0, \frac{1}{2})$ such that for any $T, M > 0$ and d sufficiently large, with overwhelming probability*

$$\begin{aligned}&\sup_{0\le t\le T\wedge\widehat{\tau}_M(Z(W,\cdot),\mathcal{Z})} \|Z(W_{td}, \cdot) - \mathcal{Z}(t, \cdot)\|_\Gamma \le Cd^{-\varepsilon}, \\ &\sup_{0\le t\le T\wedge\widehat{\tau}_M(Z(\mathcal{W},\cdot),\mathcal{Z})} \|Z(\mathcal{W}_t, \cdot) - \mathcal{Z}(t, \cdot)\|_\Gamma \le Cd^{-\varepsilon}, \\ \textit{and}\quad &\sup_{0\le t\le T\wedge\widehat{\tau}_M(Z(W,\cdot),Z(\mathcal{W},\cdot))} \|Z(W_{td}, \cdot) - Z(\mathcal{W}_t, \cdot)\|_\Gamma \le Cd^{-\varepsilon},\end{aligned} \tag{108}$$

where $C > 0$ is a constant independent of d, where $\mathcal{Z}(t, z)$ is the deterministic solution to the integro-differential equation (17) *and*

$$\widehat{\tau}_M(\mathscr{Z}_1, \mathscr{Z}_2) = \min\{\tau_M(\mathscr{Z}_1), \tau_M(\mathscr{Z}_2)\}.$$

Proof. The proof follows exactly the same steps as Theorem 14 in Collins-Woodfin et al. (2024a). Similarly, we provide the details for the case $\mathscr{Z}_1(t, \cdot) = Z(\mathcal{W}_t, \cdot)$ and $\mathscr{Z}_2(t, \cdot) = Z(W_{td}, \cdot)$ as the argument is analogous in both cases $\mathscr{Z}_1(t, \cdot) = Z(\mathcal{W}_t, \cdot)$ and $\mathscr{Z}_2(t, \cdot) = \mathcal{Z}(t, \cdot)$ and $\mathscr{Z}_1(t, \cdot) = Z(W_{td}, \cdot)$ and $\mathscr{Z}_2(t, \cdot) = \mathcal{Z}(t, \cdot)$. For notational convenience, we write $\widehat{\tau}_M = \widehat{\tau}_M(\mathscr{Z}_1, \mathscr{Z}_2)$.

By Proposition C.3 and Proposition C.4, for some $\varepsilon \in (0, \frac{1}{2})$ we have shown $Z(W_{td}, \cdot)$ and $Z(\mathcal{W}_t, \cdot)$ are respectively $(Cd^{-\varepsilon}, M, T)$-approximate solutions with overwhelming probability. (Recall by definition C.2 the deterministic solution $\mathcal{Z}(t, \cdot)$ is an $(0, M, T)$.) Hence, by the stability result on solutions of (17) proven in Proposition C.1, it follows that

$$\sup_{0\le t\le T\wedge\widehat{\tau}_M} \|Z(W_{td}, \cdot) - Z(\mathcal{W}_t, \cdot)\|_\Gamma \le Cd^{-\varepsilon}, w.o.p. \tag{109}$$

which concludes the proof. □

C.3 A Sufficient Condition to Remove the Stopping-time

In the previous section, we showed in theorem C.1 that $Z(\mathcal{W}_t, \cdot)$ and $Z(W_{td}, \cdot)$ both concentrate around the deterministic solution $\mathcal{Z}(t, z)$ of the integro-differential equation (17) under the conditions that both processes behave sufficiently regularly. We now introduce a two stopping times

which we will use to show that only one of SGD or AdvHSGD must behave sufficiently regularly is enough for theorem C.1 to hold. Let $\mathscr{Z}(t,\cdot)$ be an (ε, M, T)-approximate solution and define the stopping time for $M, \omega > 0$

$$\tau_{M,\omega} \stackrel{\text{def}}{=} \inf\Big\{t \geq 0 : \|\mathscr{Z}(t,\cdot)\|_\Gamma > M \quad \text{or} \quad \inf_{V\in\mathcal{U}^c} \|\widehat{\mathcal{B}}_1(t) - V\| \leq \omega \\ \text{or} \quad \inf_{V\in\mathcal{U}^c} \|\widehat{\mathcal{B}}_2(t) - V\| \leq \omega\Big\}, \tag{110}$$

where $\widehat{\mathcal{B}}_i(t)$ is defined in Definition C.1 and $\mathcal{U}^c$ is the complement of $\mathcal{U}$ as defined in Assumption 6. It is clear that $\tau_{M,0} = \tau_M$ as defined in Definition C.1. The next theorem will replace $\widehat{\tau}_M(\mathscr{Z}_1, \mathscr{Z}_2)$ in Theorem C.1 by the following stopping time which requires only one of the approximate solutions to stay bounded

$$\Theta_{M,\omega}(\mathscr{Z}_1, \mathscr{Z}_2) \stackrel{\text{def}}{=} \max\{\inf\{t \geq 0 : \|\mathscr{Z}_j(t,\cdot)\|_\Gamma > M\} \quad \text{for} \quad j = 1,2\} \\ \wedge \max\left\{\inf\{t \geq 0 : \inf_{V\in\mathcal{U}^c} \|\widehat{\mathcal{B}}_i(\mathscr{X}_j(t,\cdot)) - V\| \leq \omega\} \quad \text{for} \quad i = 1,2 \quad j = 1,2\right\} \tag{111}$$

where we write $\widehat{\mathcal{B}}_i(\mathscr{X}_j(t,\cdot))$ to emphasize the dependence on $\mathscr{X}_j$.

Theorem C.2. *There exists an $\varepsilon \in (0, \frac{1}{2})$ such that for any $T, M, \omega > 0$ and d sufficiently large, with overwhelming probability*

$$\sup_{0\leq t\leq T\wedge\Theta_{M,\omega}(Z(W,\cdot),\mathcal{Z})} \|Z(W_{td},\cdot) - \mathcal{Z}(t,\cdot)\|_\Gamma \leq Cd^{-\varepsilon}, \\ \sup_{0\leq t\leq T\wedge\Theta_{M,\omega}(Z(\mathcal{W},\cdot),\mathcal{Z})} \|Z(\mathcal{W}_t,\cdot) - \mathcal{Z}(t,\cdot)\|_\Gamma \leq Cd^{-\varepsilon}, \\ \textit{and} \quad \sup_{0\leq t\leq T\wedge\Theta_{M,\omega}(Z(W,\cdot),Z(\mathcal{W},\cdot))} \|Z(W_{td},\cdot) - Z(\mathcal{W}_t,\cdot)\|_\Gamma \leq Cd^{-\varepsilon}, \tag{112}$$

where $C > 0$ is a constant independent of d, where $\mathcal{Z}(t,z)$ is the deterministic solution to the integro-differential equation (17)*.*

Proof. Fix a $\omega > 0$ and for two approximate solutions $\mathscr{Z}_1$ and $\mathscr{Z}_2$, define the stopping time

$$\widehat{\tau}^{\mathscr{Z}_1,\mathscr{Z}_2}_{M+1,0} \stackrel{\text{def}}{=} \min\{\tau_{M+1,0}(\mathscr{Z}_1), \tau_{M+1,0}(\mathscr{Z}_2)\}. \tag{113}$$

Similarly to the proof theorem C.1, the proof follows similar steps in all three cases $\mathscr{Z}_1(t,\cdot) = Z(\mathcal{W}_t,\cdot)$ and $\mathscr{Z}_2(t,\cdot) = Z(W_{td},\cdot)$, $\mathscr{Z}_1(t,\cdot) = Z(\mathcal{W}_t,\cdot)$ and $\mathscr{Z}_2(t,\cdot) = \mathcal{Z}(t,\cdot)$, $\mathscr{Z}_1(t,\cdot) = Z(W_{td},\cdot)$ and $\mathscr{Z}_2(t,\cdot) = \mathcal{Z}(t,\cdot)$. We omit the details of the proof as it follows almost identical steps as the proof Theorem 15 in Collins-Woodfin et al. (2024a) after noting that for any $0 \leq t \leq T \wedge \widehat{\tau}_{M+1,0}$ and $i = 1,2$

$$\|\widehat{\mathcal{B}}_i(\mathscr{Z}_1(t,\cdot)) - \widehat{\mathcal{B}}_i(\mathscr{Z}_2(t,\cdot))\| \leq C(\|\mathcal{K}\|_{\text{op}})\|\mathscr{Z}_1(t,\cdot) - \mathscr{Z}_2(t,\cdot)\|_\Gamma \leq C(\|\mathcal{K}\|_{\text{op}})d^{-\varepsilon} \quad w.o.p. \tag{114}$$

which follows from Theorem C.1 and (67). □

It immediately follows from Theorem C.2 that SGD and AdvHSGD concentrate around $\mathcal{Z}(t,z)$ provided that AdvHSGD or $\mathcal{Z}(t,z)$ stay bounded. Analogously to Corollary 4 in Collins-Woodfin et al. (2024a) and , we formalize this in the following Corollary. We omit the proof as it is identical.

Corollary C.1 (Bounded $\mathscr{N}$ and concentration)**.** *Recall we defined the statistic $\mathscr{N}(t) = \oint \mathrm{Tr}(\mathcal{Z}(t,z))\,\mathrm{D}z$. Suppose for a fixed $T, \omega > 0$ that*

$$\sup_{0\leq t\leq T} \mathscr{N}(t) \leq M \quad \textit{and} \quad \inf_{i\in\{1,2\}} \inf_{0\leq t\leq T} \inf_{V\in\mathcal{U}^c} \|\widehat{\mathscr{B}}_i(t) - V\| > \omega \quad \textit{hold w.o.p.} \tag{115}$$

where $M > 0$ is independent of d. Then there exists $\varepsilon \in (0, \frac{1}{2})$ so that for d sufficiently large, with overwhelming probability,

$$\sup_{0\leq t\leq T} \|Z(\mathcal{W}_t,\cdot) - \mathcal{Z}(t,\cdot)\|_\Gamma \leq Cd^{-\varepsilon} \quad \textit{and} \quad \sup_{0\leq t\leq T} \|Z(W_{td},\cdot) - \mathcal{Z}(t,\cdot)\|_\Gamma \leq Cd^{-\varepsilon}, \tag{116}$$

where $C > 0$ is a constant independent of d. Hence, by the triangle inequality, it follows that

$$\sup_{0\leq t\leq T} \|Z(W_{td},\cdot) - Z(\mathcal{W}_t,\cdot)\|_\Gamma \leq 2Cd^{-\varepsilon}. \tag{117}$$

Remark C.2. *Intuitively, Corollary C.1 is written in terms of a condition on the size of $\mathscr{N}(t)$ instead of $\|\mathcal{Z}(t,\cdot)\|_\Gamma$ as per Lemma C.1, there exists constants $c, C > 0$ (independent of d) such that $c\mathscr{N}(t) \le \|\mathcal{Z}(t,\cdot)\|_\Gamma \le C\mathscr{N}(t)$. Hence, we obtain the inequality for a fixed $M > 0$*

$$\inf\{t \ge 0 : \|\mathcal{Z}(t,\cdot)\|_\Gamma > CM\} \ge \inf\{t \ge 0 : \mathscr{N}(t) > M\}, \tag{118}$$

from which it is not hard to see that Theorem C.2 holds. Corollary C.1 also holds if we replace $(\mathscr{N}(t), \widehat{\mathscr{B}}_1(t), \widehat{\mathscr{B}}_2(t))$ in (115) *with the equivalents for either SGD or AdvHSGD by Lemma C.1.*

Theorem B.1 and Theorem B.2 follow from direct applications of Theorem C.2 and Lemma C.1 as explained in Remark C.2. We also provide a sufficient condition for Corollary C.1 to hold.

Proposition C.5 (Non-explosiveness)**.** *Suppose that the assumptions stated in this framework hold then the function $\mathscr{N}(t)$ is the solution to the following ordinary differential equation*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{N}(t) = -2\gamma(t)\Bigg(\sum_{i=1}^{2} p_i\, \mathbb{E}_{v,\epsilon}[f_i'(q_{t,i},\epsilon)x_{t,i}] + \lambda(\mathscr{N}(t) - \|X^\star\|^2 - \mathrm{Tr}(\mu^\top\mu))\Bigg) \\ + \frac{\gamma(t)^2}{d}\sum_{i=1}^{2} p_i\, \mathrm{Tr}(K_i + \mu_i\mu_i^\top)\, \mathbb{E}_{v,\epsilon}[f_i'(q_{t,i},\epsilon)^2],\end{aligned} \tag{119}$$

where

$$\begin{aligned} q_{t,i} &= x_{t,i} \oplus x_i^\star = \sqrt{\mathscr{B}_i(t)}v + \mathfrak{m}_i(t) + \begin{bmatrix}\delta u_{t,i}\sqrt{\widehat{\mathscr{B}}_{i,44}(t)} & 0\end{bmatrix}^\top, \\ u_{t,i} &= \mathrm{argmax}_{|u|\le 1}\, f_i(\sqrt{\mathscr{B}_i(t)}v + \mathfrak{m}_i(t) + \begin{bmatrix}\delta u\sqrt{\widehat{\mathscr{B}}_{i,44}(t)} & 0\end{bmatrix}^\top, \epsilon),\end{aligned} \tag{120}$$

for $v \sim \mathcal{N}(0, I_2)$ and where $\mathscr{B}_i(t)$ is defined in (17)*. Suppose further that the objective function f is α-pseudo-Lipschitz with $\alpha \le 1$. Then there exists $C = C(\bar\gamma, L(f), \eta, \delta, \|\mathcal{K}\|_{\mathrm{op}}, \lambda, \max_{i=1,2}\|\mu_i\|^2)$ such that*

$$\mathscr{N}(t) \le (1 + \mathscr{N}(0))e^{Ct},$$

for all time t such that $(\widehat{\mathscr{B}}_1(t), \widehat{\mathscr{B}}_2(t))$ is in $\mathcal{U}^2$.

Proof. To derive (119), we first apply Itô's lemma to the statistic $\varphi(\mathcal{X}_t) \overset{\mathrm{def}}{=} \|\mathcal{X}_t\|^2 + \|X^\star\|^2 + \|\mu_1\|^2 + \|\mu_2\|^2$ and obtain

$$\begin{aligned}\mathrm{d}\varphi(\mathcal{X}_t) = -2\gamma(t)\langle\mathcal{X}_t, \nabla\mathcal{R}_\lambda^{\mathrm{adv}}(\mathcal{W}_t)\rangle\,\mathrm{d}t + \frac{\gamma(t)^2}{d}\sum_{i=1}^{2} p_i\,\mathrm{Tr}(K_i + \mu_i\mu_i^\top)\,\mathbb{E}_{a,\epsilon}[f_i'(\rho_{t,i},\epsilon)^2]\,\mathrm{d}t \\ + \mathrm{d}\mathcal{M}_t^{\mathrm{AdvHSGD}}(\varphi),\end{aligned} \tag{121}$$

where $\mathcal{M}_t^{\mathrm{AdvHSGD}}(\varphi)$ is a martingale. By Assumption 7, we know there exist functions $\mathcal{I}_i$ for $i = 1, 2$ such that $\mathcal{I}_i(\widehat{B}_i(\mathcal{W}_t)) = \mathbb{E}_{a,\epsilon}[f_i'(\rho_{t,i},\epsilon)^2]$. We also have that

$$\langle\mathcal{X}_t, \nabla\mathcal{R}_\lambda^{\mathrm{adv}}(\mathcal{W}_t)\rangle = \sum_{i=1}^{2} p_i(\mathbb{E}_{v,\epsilon}[f_i'(\rho_{t,i},\epsilon)\langle\mathcal{X}_t, \sqrt{K_i}v + \mu_i\rangle + \delta u_{t,i}\|\mathcal{X}_t\|] + \lambda\|\mathcal{X}_t\|^2. \tag{122}$$

This is a Gaussian expectation hence there exists some functions A such that

$$A(\widehat{B}_i(\mathcal{W}_t)) \overset{\mathrm{def}}{=} \mathbb{E}_{v,\epsilon}[f_i'(\rho_{t,i},\epsilon)\langle\mathcal{X}_t, \sqrt{K_i}v + \mu_i\rangle + \delta u_{t,i}\|\mathcal{X}_t\|]. \tag{123}$$

We then obtain the representation

$$\begin{aligned}\mathrm{d}\varphi(\mathcal{X}_t) = -2\gamma(t)\Bigg(\sum_{i=1}^{2} p_i A(\widehat{B}_i(\mathcal{W}_t)) + \lambda(\varphi(\mathcal{X}_t) - \|X^\star\|^2 - \|\mu_1\|^2 - \|\mu_2\|^2)\Bigg)\,\mathrm{d}t \\ + \frac{\gamma(t)^2}{d}\sum_{i=1}^{2} p_i\,\mathrm{Tr}(K_i + \mu_i\mu_i^\top)\,\mathbb{E}_{v,\epsilon}[f_i'(q_{t,i},\epsilon)^2]\,\mathrm{d}t + \mathrm{d}\mathcal{M}_t^{\mathrm{AdvHSGD}}(\varphi).\end{aligned} \tag{124}$$

Using Cauchy's integral formula, we also have the representation

$$\varphi(\mathcal{X}_t) = \oint (S_{11}(\mathcal{W}_t, z) + S_{22}(\mathcal{W}_t, z))\,\mathrm{D}z + \|\mu_1\|^2 + \|\mu_2\|^2. \tag{125}$$

From (177), we also have

$$\mathrm{d}\varphi(\mathcal{X}_t) = \frac{-1}{2\pi i}\oint \left(\mathcal{F}_{\mathcal{S},11}(z, Z(\mathcal{W}_t, z)) + \mathcal{F}_{\mathcal{S},22}(z, Z(\mathcal{W}_t, z))\right)\mathrm{D}z + \mathrm{d}\mathcal{M}_t^{\mathrm{AdvHSGD}}(\varphi) \tag{126}$$

for some martingale $\mathcal{M}_t^{\mathrm{AdvHSGD}}$. Given that the semi-martingale decomposition of an Itô process is unique and that

$$\mathscr{N}(t) = \oint \mathrm{Tr}(\mathcal{Z}(t,z))\,\mathrm{D}z = \oint \left(\mathcal{S}_{11}(t,z) + \mathcal{S}_{22}(t,z)\right)\mathrm{D}z + \|\mu_1\|^2 + \|\mu_2\|^2 \tag{127}$$

then the terms with finite quadratic variation are equal from which it follows that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{N}(t) = -2\gamma(t)\Bigg(\sum_{i=1}^{2} p_i A(\widehat{\mathscr{B}}_i(t)) + \lambda(\mathscr{N}(t) - \|X^\star\|^2 - \|\mu_1\|^2 - \|\mu_2\|^2)\Bigg) \\ + \frac{\gamma(t)^2}{d}\sum_{i=1}^{2} p_i\,\mathrm{Tr}(K_i + \mu_i\mu_i^\top)\,\mathbb{E}_{v,\epsilon}[f_i'(q_{t,i},\epsilon)^2]\,\mathrm{d}t.\end{aligned} \tag{128}$$

This completes the derivation of (119). From Assumption 3, the function f is α-pseudo-Lipschitz with $\alpha \leq 1$. Hence, from Lemma D.3 we obtain the almost sure bound

$$|f_i'(q_{t,i},\epsilon)| \leq C(L(f))(1 + \|q_{t,i}\| + |\epsilon|) \quad a.s. \tag{129}$$

Squaring and taking the expectation, we obtain

$$\mathbb{E}_{v,\epsilon}[|f_i'(q_{t,i},\epsilon)|^2] \leq C(L(f), \eta, \delta, \max_{i=1,2}\|\mu_i\|^2, \|\mathcal{K}\|_{\mathrm{op}})(1 + \mathscr{N}(t)) \tag{130}$$

since $\widehat{\mathscr{B}}_{i,44}(t) = \mathscr{N}(t) - \|X^\star\|^2 - \mathrm{Tr}(\mu^\top\mu)$, since by Cauchy–Schwarz and Lemma C.1

$$\begin{aligned}\mathrm{Tr}(\mathscr{B}_i(t)) \leq C\|\mathscr{B}_i(t)\| \leq C(\|\mathcal{K}\|_{\mathrm{op}})\|\mathcal{S}(t,\cdot)\|_\Gamma \leq C(\|\mathcal{K}\|_{\mathrm{op}})\mathscr{N}(t), \\ \text{and from (443)} \quad \|\mathfrak{m}_i\|^2 \leq \|\mu_i\|^2\mathscr{N}(t).\end{aligned} \tag{131}$$

Similarly, we obtain from Cauchy–Schwarz

$$\left|\mathbb{E}_{v,\epsilon}[f_i'(q_{t,i},\epsilon)x_{t,i}]\right|^2 \leq C(L(f), \eta, \delta, \max_{i=1,2}\|\mu_i\|^2, \|\mathcal{K}\|_{\mathrm{op}})(1 + \mathscr{N}(t))^2. \tag{132}$$

Combining these inequalities, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{N}(t) \leq C(1 + \mathscr{N}(t)). \tag{133}$$

where $C = C(\bar{\gamma}, L(f), \eta, \delta, \|\mathcal{K}\|_{\mathrm{op}}, \lambda, \max_{i=1,2}\|\mu_i\|^2)$. The result follows from an application of Grönwall's inequality $(1 + \mathscr{N}(t)) \leq (1 + \mathscr{N}(0))e^{Ct}$. □

C.4 Concentration Result for General Statistics

In this section, we prove a result for general statistics satisfying Assumption 12. For statistics satisfying Assumption 12, the det. equiv. is defined as follows

$$\phi(t) \overset{\mathrm{def}}{=} g\left(\left\{\oint q_j(z_1,z_2)\mathcal{Z}(t,z)\,\mathrm{D}z\right\}_{j=1}^{\ell}\right). \tag{134}$$

We are now ready to prove our result for the concentration of $\varphi(\mathcal{X}_t)$ and $\varphi(W_{td})$ around $\phi(t)$.

Theorem C.3 (Concentration of any statistic)**.** *Suppose that the assumptions stated in this framework hold. Then there exists an $\varepsilon \in (0, \frac{1}{2})$ such that for any $T, M, \omega > 0$ and d sufficiently large, with overwhelming probability*

$$\begin{aligned}&\sup_{0\leq t\leq T\wedge\Theta_{M,\omega}(Z(W,\cdot),\mathcal{Z})} |\varphi(W_{td}) - \phi(t)| \leq Cd^{-\varepsilon}, \\ &\sup_{0\leq t\leq T\wedge\Theta_{M,\omega}(Z(\mathcal{W},\cdot),\mathcal{Z})} |\varphi(\mathcal{W}_t) - \phi(t)| \leq Cd^{-\varepsilon}, \\ \textit{and}\quad &\sup_{0\leq t\leq T\wedge\Theta_{M,\omega}(Z(W,\cdot),Z(\mathcal{W},\cdot))} |\varphi(W_{td}) - \varphi(\mathcal{W}_t)| \leq Cd^{-\varepsilon},\end{aligned} \tag{135}$$

where $C > 0$ is a constant independent of d, where ϕ is defined in (134) *and where the stopping time $\Theta_M(\mathscr{Z}_1, \mathscr{Z}_2) \equiv \Theta_{M,0}(\mathscr{Z}_1, \mathscr{Z}_2)$ is defined in* (111)*.*

Proof. This proof builds upon the proof Theorem 15 and Theorem 16 in Collins-Woodfin et al. (2024a). Similarly to the proof Theorem C.2, we define the stopping time

$$\widehat{\tau}_{M+1,\eta}^{\mathscr{Z}_1,\mathscr{Z}_2} \stackrel{\text{def}}{=} \min\{\tau_{M+1,\eta}(\mathscr{Z}_1), \tau_{M+1,\eta}(\mathscr{Z}_2)\}, \tag{136}$$

and consider the case $\mathscr{Z}_1(t,\cdot) = Z(\mathcal{W}_t,\cdot)$ and $\mathscr{Z}_2(t,\cdot) = Z(W_{td},\cdot)$ as the other cases follow analogously. Fix $T, M > 0$. By Proposition C.4 and Proposition C.3, there exists a $\varepsilon \in (0, \frac{1}{2})$ and constant $C > 0$ such that $Z(\mathcal{W}_t, z)$ and $Z(W_{td}, z)$ are $(Cd^{-\varepsilon}, M+1, T)$-approximate solutions. (Recall by definition C.2 the deterministic solution $\mathcal{Z}(t,\cdot)$ is an $(0, M, T)$.) By Cauchy's integral formula, we have for each $1 \le j \le \ell$

$$\mathcal{W}_t^\top q_j(K_1,K_2)\mathcal{W}_t = \oint q_j(z)Z(\mathcal{W}_t,z)\,\mathrm{D}z \quad \text{and} \quad W_{td}^\top q_j(K_1,K_2)W_{td} = \oint q_j(z)Z(W_{td},z)\,\mathrm{D}z. \tag{137}$$

Hence from Proposition C.2, it follows that

$$\sup_{0\le t\le T\wedge\widehat{\tau}_{M+1,0}} |\varphi(\mathcal{W}_t) - \varphi(W_{td})| \le Cd^{-\varepsilon}, \quad w.o.p. \tag{138}$$

□

Corollary C.2. *Suppose that the assumptions stated in this framework hold and the assumptions in Corollary C.1 hold. Then there exists an $\varepsilon \in (0, \frac{1}{2})$ such that for any $T > 0$ and sufficiently large d, with overwhelming probability,*

$$\sup_{0\le t\le T} |\varphi(\mathcal{X}_t) - \phi(t)| \le Cd^{-\varepsilon} \quad \textit{and} \qquad \sup_{0\le t\le T} |\varphi(X_{td}) - \phi(t)| \le Cd^{-\varepsilon}. \tag{139}$$

Hence, by the triangle inequality, it follows that

$$\sup_{0\le t\le T} |\varphi(X_{td}) - \varphi(\mathcal{X}_t)| \le 2Cd^{-\varepsilon}. \tag{140}$$

C.5 Concentration of General Statistics Involving the Minimizer of the ℓ_2-Adversarial Risk

Recall from Definition B.2 the minimizer of the adversarial population risk is defined as follows

Definition C.3. *Let $X^{\star,\mathrm{adv}}$ denote the adversarial ground truth. In other words, we have*

$$X^{\star,\mathrm{adv}} \in \operatorname*{argmin}_X \left\{ \mathcal{R}^{\mathrm{adv}}(X) = \mathbb{E}_{a,i,\epsilon}[\max_{|s|\le 1} f_i(X^\top a + \delta s\|X\|,\epsilon)] \right\}. \tag{141}$$

To study the distance to adversarial optimality, we must compute the det. equiv. of the statistic $\|X - X^{\star,\mathrm{adv}}\|^2$. We cannot use a similar argument as presented in the proof Proposition C.5 given that the first expectation in the drift contribution from Itô's lemma applied to $\|\mathcal{X}_t - X^{\star,\mathrm{adv}}\|^2$, involves an expectation over the jointly Gaussian scalars $\mathcal{X}_t^\top a_i$, $(X^\star)^\top a$ and $(X^{\star,\mathrm{adv}})^\top a_i$. Consequently, we must compute the det. equiv. for the mean and covariance matrix of the gaussian vector $\widetilde{X}_t^\top a_i$ for $\widetilde{X}_t \stackrel{\text{def}}{=} X_t \oplus X^\star \oplus X^{\star,\mathrm{Adv}}$. Note that the results in this section hold for the soft label setting and hard label setting with $X^\star = 0$.

Recall from (31) the definition of Z^{Adv}:

$$Z^{\mathrm{Adv}}(\widehat{W},z) \stackrel{\text{def}}{=} \widehat{W}^\top \mathcal{R}(z;\mathcal{K})\widehat{W} = \begin{bmatrix} Z(W,z) & (\star)_0 \\ (\star)_0^\top & (X^{\star,\mathrm{Adv}})^\top \mathcal{R}(z;\mathcal{K})(X^{\star,\mathrm{Adv}}) \end{bmatrix}. \tag{142}$$

with $A_{\mathcal{K}}(\widehat{W},z) \stackrel{\text{def}}{=} X^\top \mathcal{R}(z;\mathcal{K})X^{\star,\mathrm{Adv}}$ such that $(\star)_0$ corresponds to the following term

$$(\star)_0 \stackrel{\text{def}}{=} \begin{bmatrix} A_{\mathcal{K}}(\widehat{W},z) & (X^\star)^\top \mathcal{R}(z;\mathcal{K})X^{\star,\mathrm{Adv}} & \mu_1^\top \mathcal{R}(z;\mathcal{K})X^{\star,\mathrm{Adv}} & \mu_2^\top \mathcal{R}(z;\mathcal{K})X^{\star,\mathrm{Adv}} \end{bmatrix}^\top.$$

We also define $\widehat{B}_i^{\mathrm{Adv}}(W)$ similarly to $\widehat{B}_i(W)$

$$\widehat{\mathcal{B}}_i^{\mathrm{Adv}}(\widehat{W}) \stackrel{\text{def}}{=} \begin{bmatrix} \widehat{B}_i(W) & (\star)_1 \\ (\star)_1^\top & \|X^{\star,\mathrm{Adv}}\|^2 \end{bmatrix} \in \mathbb{R}^{5\times 5}, \tag{143}$$

with $(\star)_1$ corresponding to the following term

$$(\star)_1 \stackrel{\text{def}}{=} \left[\oint z_i A_{\mathcal{K}}(\widehat{W}, z)\, \mathrm{D}z \quad (X^\star)^\top (X^{\star,\mathrm{Adv}}) \quad \mu_i^\top X^{\star,\mathrm{Adv}} \quad 0 \right]^\top.$$

From the structure of $Z^{\mathrm{Adv}}(\widehat{W})$, it is clear that computing its det. equiv. amounts to computing the det. equiv. of $A_{\mathcal{K}}(\widehat{W}, z)$.

Given that the argument is extremely similar to the steps presented in Section C.1 to Section C.4, we will only provide a brief sketch of the proof, highlighting the differences whenever they arise. We omit the steps of the proof for AdvHSGD and focus on SGD as the argument is extremely similar.

Define $\widehat{W}_k = W_k \oplus X^{\star,\mathrm{Adv}}$ for SGD and $\widehat{\mathcal{W}}_t$ analogously for AdvHSGD. From the Doob decompositions presented in Section C.1, the doob decomposition of $\varphi(\widehat{W}_{td}) = X_{td}^\top \mathcal{R}(z;\mathcal{K}) X^{\star,\mathrm{Adv}}$ is given by

$$\varphi(\widehat{W}_{td}) = \varphi(\widehat{W}_0) - \int_0^t \gamma_{sd} \langle \nabla \varphi(\widehat{W}_{sd}), \nabla \mathcal{R}_\lambda^{\mathrm{adv}}(X_{sd}) \rangle \, \mathrm{d}s + \sum_{j=0}^{\lfloor td \rfloor - 1} \Delta \mathcal{M}_j^{\mathrm{Grad}}(\varphi) + \xi_{td}(\varphi). \quad (144)$$

We redefine the σ-algebras as follows

$$\mathcal{G}_{k,i} \stackrel{\text{def}}{=} \sigma(\{\hat{X}_j\}_{j=0}^k, \{r_i\}_{i=0}^k, \epsilon_{k+1}, \{I_j\}_{j=1}^k, I_{k+1} = i, X^{\star,\mathrm{Adv}}) \text{ and } \mathcal{F}_k \stackrel{\text{def}}{=} \sigma(\{\hat{X}_j\}_{j=0}^k, X^{\star,\mathrm{Adv}}).$$

The following Lemma is analogous to Lemma C.1 and follows an argument almost identical to Lemma 3 in Collins-Woodfin & Seroussi (2025) and Lemma 5 in Collins-Woodfin et al. (2024a).

Lemma C.4. *Recall $\mathscr{N}(t) \stackrel{\text{def}}{=} \oint \mathrm{Tr}(\mathcal{Z}(t,z))\, \mathrm{D}z$. Then there exists a constant $C > 0$ depending on $\|\mathcal{K}\|_{\mathrm{op}}$ and such that*

$$C \leq \frac{\|Z^{\mathrm{Adv}}(\widehat{W}_{td}, \cdot)\|_\Gamma}{\|\widehat{W}_{td}\|^2}, \frac{\|Z^{\mathrm{Adv}}(\widehat{\mathcal{W}}_t, \cdot)\|_\Gamma}{\|\widehat{\mathcal{W}}_t\|^2}, \frac{\|\mathcal{Z}^{\mathrm{Adv}}(t, z \cdot\|_\Gamma}{\mathscr{N}(t) + \|X^{\star,\mathrm{Adv}}\|^2} \leq 4.$$

We introduce a stopping time ψ_M analogous to τ_M in Definition C.1.

Definition C.4 (Stopping time ψ_M)**.** *Define the open set $\mathcal{U}^{\mathrm{Adv}}$ such that $(\widehat{B}_1^{\mathrm{Adv}}(W_0), \widehat{B}_2^{\mathrm{Adv}}(W_0)) \in \mathcal{U}^{\mathrm{Adv}} \times \mathcal{U}^{\mathrm{Adv}}$ and*

$$\mathcal{U}^{\mathrm{Adv}} = \left\{ \begin{bmatrix} B & u_1 \\ u_2^\top & w \end{bmatrix} : \quad B \in \mathcal{U}, \quad u_1, u_2 \in \mathbb{R}^4 \quad \textit{and} \quad w \in \mathbb{R} \right\}. \quad (145)$$

Here $\mathcal{U}^{\mathrm{Adv}}$ is the extension of $\mathcal{U}$ to $\mathbb{R}^{5\times 5}$. For any $M > 0$ and a continuous function $\mathscr{Z}^{\mathrm{Adv}} : [0,\infty] \times \mathbb{C}^2 \to \mathbb{R}^{5\times 5}$, define

$$\psi_M(\mathscr{Z}^{\mathrm{Adv}}) \stackrel{\text{def}}{=} \inf \left\{ t \geq 0 : \|\mathscr{Z}^{\mathrm{Adv}}(t,\cdot)\|_\Gamma > M \quad \textit{or} \quad \widehat{\mathcal{B}}_1^{\mathrm{Adv}}(t) \notin \mathcal{U}^{\mathrm{Adv}} \quad \textit{or} \quad \widehat{\mathcal{B}}_2^{\mathrm{Adv}}(t) \notin \mathcal{U}^{\mathrm{Adv}} \right\},$$

where we define

$$\mathscr{Z}^{\mathrm{Adv}}(t,\cdot) \stackrel{\text{def}}{=} \begin{bmatrix} \mathscr{Z}(t,\cdot) & (\star)_3 \\ (\star)_3^\top & (X^{\star,\mathrm{Adv}})^\top \mathcal{R}(z;\mathcal{K})(X^{\star,\mathrm{Adv}}) \end{bmatrix}. \quad (146)$$

with $(\star)_3$ corresponding to the following term

$$(\star)_3 \stackrel{\text{def}}{=} \left[\mathscr{O}(t,\cdot) \quad (X^\star)^\top \mathcal{R}(z;\mathcal{K}) X^{\star,\mathrm{Adv}} \quad \mu_1^\top \mathcal{R}(z;\mathcal{K}) X^{\star,\mathrm{Adv}} \quad \mu_2^\top \mathcal{R}(z;\mathcal{K}) X^{\star,\mathrm{Adv}} \right]^\top$$

and

$$\widehat{\mathcal{B}}_i^{\mathrm{Adv}}(t) \stackrel{\text{def}}{=} \begin{bmatrix} \widehat{\mathcal{B}}_i(t) & (\star)_4 \\ (\star)_4^\top & \|X^{\star,\mathrm{Adv}}\|^2 \end{bmatrix} \in \mathbb{R}^{5\times 5}, \quad (147)$$

with

$$(\star)_4 \stackrel{\text{def}}{=} \left[\oint z_i \mathscr{O}(t,z)\, \mathrm{D}z \quad (X^\star)^\top (X^{\star,\mathrm{Adv}}) \quad \mu_i^\top X^{\star,\mathrm{Adv}} \quad 0 \right]^\top$$

where $\mathscr{Z}(t,\cdot)$ is defined in Definition C.1. Here $\mathscr{O}(t,\cdot)$ is either $A_{\mathcal{K}}(\widehat{W}_{td},\cdot)$, $A_{\mathcal{K}}(\widehat{\mathcal{W}}_t,\cdot)$ or $\mathcal{A}_{\mathcal{K}}(t,\cdot)$.

When clear from context, we omit the dependence $\psi_M = \psi_M(\mathscr{Z}^{\mathrm{Adv}})$. We are now ready to introduce the notion of approximate solutions to integro-differential equations to (32).

Definition C.5 ((ε, M, T)-approximate solution to the integro-differential equation)**.** *For $M, T, \varepsilon > 0$ and a continuous function $\mathscr{Z}^{\mathrm{Adv}} : [0,\infty] \times \mathbb{C}^2 \to \mathbb{R}^{5\times 5}$, we denote $\mathscr{Z}^{\mathrm{Adv}}$ an (ε, M, T)-approximate solution of* (32) *if*

$$\sup_{0\le t\le T\wedge\psi_M} \left\| \mathscr{Z}^{\mathrm{Adv}}(t,\cdot) - \mathscr{Z}^{\mathrm{Adv}}(0,\cdot) - \int_0^t \mathcal{F}^{\mathrm{Adv}}(z, \mathscr{Z}^{\mathrm{Adv}}(s,\cdot))\,\mathrm{d}s \right\|_{\Gamma} \le \varepsilon, \tag{148}$$

where $\mathscr{Z}^{\mathrm{Adv}}(0,\cdot) = \widehat{W}_0^\top \mathcal{R}(\cdot;\mathcal{K})\widehat{W}_0$.

It is clear from the definition C.2 and the definition of $\mathcal{Z}^{\mathrm{Adv}}$ in (32) that $\mathcal{Z}^{\mathrm{Adv}}$ is an approximate solution with $\varepsilon = 0$. In the next subsection, we will show $\mathcal{Z}^{\mathrm{Adv}}(\widehat{W}_{td}, z)$ and $Z^{\mathrm{Adv}}(\widehat{\mathcal{W}}_t, z)$ are (ε, M, T)-approximate solutions to (148)

As we have already shown in Propositions C.1 and C.2 that approximate solutions to (17) satisfy stability properties, showing stability of approximate solutions as defined above in Definition C.5 amounts to showing stability for the term $\mathscr{O}(t,\cdot)$ in $\mathscr{Z}^{\mathrm{Adv}}(t,\cdot)$. As the proofs follow an almost identical argument to the proofs of Propositions C.1 and C.2, we state the results and omit them.

Proposition C.6 (Stability)**.** *For all (ε, M, T)-approximate solutions $\mathscr{Z}_1^{\mathrm{Adv}}$ and $\mathscr{Z}_2^{\mathrm{Adv}}$, there exists a constant $C = C(\bar\gamma, \lambda, M, \alpha, L(h), L(\mathcal{I}), \|\mathcal{K}\|_{\mathrm{op}}, \max_{i=1,2}\|\mu_i\|, T) > 0$ such that*

$$\sup_{0\le t\le T\wedge\widehat{\psi}_M} \|\mathscr{Z}_1^{\mathrm{Adv}}(t,\cdot) - \mathscr{Z}_2^{\mathrm{Adv}}(t,\cdot)\|_{\Gamma} \le C\varepsilon,$$

where $\widehat{\psi}_M = \min\{\psi_M(\mathscr{Z}_1^{\mathrm{Adv}}), \psi_M(\mathscr{Z}_2^{\mathrm{Adv}})\}$.

Proposition C.1 can be extended to any function $\varphi(X)$ satisfying Assumption 11.

Proposition C.7. *Fix $\ell \ge 1$ and suppose $\varphi : \mathbb{R}^d \to \mathbb{R}$ satisfies Assumption 11. Let $\mathscr{Z}_1^{\mathrm{Adv}}$ and $\mathscr{Z}_2^{\mathrm{Adv}}$ be (ε, M, T)-approximate solutions. Then there exists a positive constant $C = C(\bar\gamma, \lambda, M, \alpha, L(h), L(g), \|K\|_{\mathrm{op}})$ such that*

$$\sup_{0\le t\le T\wedge\widehat{\psi}_M} \left| g\left(\left\{ \oint q_j(z)\mathscr{Z}_1^{\mathrm{Adv}}(t,z)\,\mathrm{D}z \right\}_{j=1}^{\ell} \right) - g\left(\left\{ \oint q_j(z)\mathscr{Z}_2^{\mathrm{Adv}}(t,z)\,\mathrm{D}z \right\}_{j=1}^{\ell} \right) \right| \le C\varepsilon,$$

where $\widehat{\psi}_M = \min\{\psi_M(\mathscr{Z}_1^{\mathrm{Adv}}), \psi_M(\mathscr{Z}_2^{\mathrm{Adv}})\}$.

We require a net argument as in Lemma C.3. We omit the proof once again.

Lemma C.5 (Net argument)**.** *Fix $T, M > 0, \xi > 0$. Suppose Γ_ξ is a $d^{-\xi}$ mesh of Γ with $|\Gamma_\xi| = Cd^\xi$ for some $C > 0$. Let the function $Z^{\mathrm{Adv}}(t,z) = Z^{\mathrm{Adv}}(\widehat{W}_{td}, z)$ or $Z^{\mathrm{Adv}}(t,z) = Z^{\mathrm{Adv}}(\widehat{\mathcal{W}}_t, z)$ satisfy*

$$\sup_{0\le t\le (T\wedge\tau_M)} \|Z^{\mathrm{Adv}}(t,\cdot) - Z^{\mathrm{Adv}}(0,\cdot) - \int_0^t \mathcal{F}^{\mathrm{Adv}}(\cdot, Z^{\mathrm{Adv}}(s,\cdot))\,\mathrm{d}s\|_{\Gamma_\xi} \le \varepsilon \tag{149}$$

with ψ_M as in Definition C.4. Then Z^{Adv} is an $(\varepsilon + Cd^{-\xi}, M, T)$-approximate solution to the integro-differential equation (32)*, that is,*

$$\sup_{0\le t\le (T\wedge\tau_M)} \|Z^{\mathrm{Adv}}(t,\cdot) - Z^{\mathrm{Adv}}(0,\cdot) - \int_0^t \mathcal{F}^{\mathrm{Adv}}(\cdot, Z^{\mathrm{Adv}}(s,\cdot))\,\mathrm{d}s\|_{\Gamma} \le \varepsilon + Cd^{-\xi},$$

where $C = C(\bar\gamma, \alpha, L(\mathcal{I}), M, \|\mathcal{K}\|_{\mathrm{op}}, L(h), \lambda, \max_{i=1,2}\|\mu_i\|, T) > 0$ and

Remark C.3. *The set $\mathcal{U}$ enters our assumptions only through Assumption 6, which concerns h_1 and h_2, and these are functions of the 4×4 statistic $\widehat{B}_i(W)$ alone. Since the upper-diagonal 4×4 block of $\widehat{B}_i^{\mathrm{Adv}}(\widehat{W})$ is $\widehat{B}_i(W)$ by* (143)*, we have $\widehat{B}_i^{\mathrm{Adv}}(\widehat{W}) \in \mathcal{U}^{\mathrm{Adv}}$ if and only if $\widehat{B}_i(W) \in \mathcal{U}$. Moreover, $\widehat{W} = W \oplus X^{\star,\mathrm{Adv}}$ gives*

$$\|\widehat{W}\|^2 = \|W\|^2 + \|X^{\star,\mathrm{Adv}}\|^2,$$

so that $\|\widehat{W}\|^2 > M$ if and only if $\|W\|^2 > M - \|X^{\star,\mathrm{Adv}}\|^2$. Define $\vartheta_M^{\mathrm{Adv}}$ analogously to ϑ_M in (91) *with $\widehat{W}$, $\widehat{B}_1^{\mathrm{Adv}}(\widehat{W})$, $\widehat{B}_2^{\mathrm{Adv}}(\widehat{W})$ and $\mathcal{U}^{\mathrm{Adv}}$ instead. Combining the two previous observations,*

we have $\vartheta_M^{\mathrm{Adv}} = \vartheta_{M-\|X^{\star,\mathrm{Adv}}\|^2}$. *From Assumption 10, we have* $\|X^{\star,\mathrm{Adv}}\| \leq C_\star$ *for* $C_\star > 0$ *independent of* d. *Since* ϑ_M *is nondecreasing in* M, *it follows for all* $M > 0$ *that*

$$\vartheta_M \leq \vartheta_{M+C_\star^2}^{\mathrm{Adv}} \leq \vartheta_{M+C_\star^2}.$$

The stopping times are thus interchangeable, which allows us to relate ψ_M *to* τ_M. *Indeed, Lemma C.4 provides constants* $c, C > 0$ *such that* $\vartheta_{c\cdot M}^{\mathrm{Adv}} \leq \psi_M \leq \vartheta_{C\cdot M}^{\mathrm{Adv}}$. *Thus, for any* $M > 0$ *we have*

$$\psi_{M/C} \leq \vartheta_M \leq \psi_{(M+C_\star^2)/c}.$$

Since ϑ_M *and* τ_M *are interchangeable by Lemma C.1, then there exists constants* $c', C' > 0$ *such that for all* $M > 0$

$$\psi_{c'M/C} \leq \tau_M \leq \psi_{(C'M+C_\star^2)/c}$$

The same equivalence applies to the stopping times of Section C.3. Indeed, by (145) *a matrix* $V \in (\mathcal{U}^{\mathrm{Adv}})^c$ *is exactly a matrix whose* 4×4 *block* B *satisfies* $B \in \mathcal{U}^c$. *This leads to the bound*

$$\|\widehat{\mathcal{B}}_i^{\mathrm{Adv}}(t) - V\| \;\geq\; \|\widehat{\mathcal{B}}_i(t) - B\| \;\geq\; \inf_{B' \in \mathcal{U}^c} \|\widehat{\mathcal{B}}_i(t) - B'\|.$$

Conversely, given $B \in \mathcal{U}^c$, *the matrix* V *obtained from* $\widehat{\mathcal{B}}_i^{\mathrm{Adv}}(t)$ *by replacing its* 4×4 *block by* B *lies in* $(\mathcal{U}^{\mathrm{Adv}})^c$ *and satisfies* $\|\widehat{\mathcal{B}}_i^{\mathrm{Adv}}(t) - V\| = \|\widehat{\mathcal{B}}_i(t) - B\|$ *from which it follows that*

$$\inf_{V \in (\mathcal{U}^{\mathrm{Adv}})^c} \|\widehat{\mathcal{B}}_i^{\mathrm{Adv}}(t) - V\| = \inf_{B \in \mathcal{U}^c} \|\widehat{\mathcal{B}}_i(t) - B\|.$$

Hence, we will state the results in terms of the stopping times τ_M *and* Θ_M *defined previously.*

The proof the following Theorem C.4 follows basically the same argument as the proof Theorem C.2. Indeed, Propositions D.2 and D.1 were proven for an arbitrary quadratic statistic of SGD and AdvHSGD respectively. Hence, using the same arguments as the proof Propositions C.3 and C.4, it follows that for fixed $T, M > 0$ and $0 < \varepsilon < \frac{1}{2}$, $Z^{\mathrm{Adv}}(\widehat{W}_{td}, \cdot)$ and $Z^{\mathrm{Adv}}(\widehat{\mathcal{W}}_t, \cdot)$ are $(Cd^{-\varepsilon}, M, T)$ approximate solutions. Finally, the theorem follows from the stability proven in Proposition C.6.

Theorem C.4. *There exists an* $\varepsilon \in (0, \frac{1}{2})$ *such that for any* $T, M > 0$ *and* d *sufficiently large, with overwhelming probability*

$$\begin{aligned} &\sup_{0 \leq t \leq T \wedge \widehat{\tau}_M(Z(W), \mathcal{Z})} \|Z^{\mathrm{Adv}}(\widehat{W}_{td}, \cdot) - \mathcal{Z}^{\mathrm{Adv}}(t, \cdot)\|_\Gamma \leq Cd^{-\varepsilon}, \\ &\sup_{0 \leq t \leq T \wedge \widehat{\tau}_M(Z(\mathcal{W}), \mathcal{Z})} \|Z^{\mathrm{Adv}}(\mathcal{W}_t, \cdot) - \mathcal{Z}^{\mathrm{Adv}}(t, \cdot)\|_\Gamma \leq Cd^{-\varepsilon}, \\ \textit{and} \quad &\sup_{0 \leq t \leq T \wedge \widehat{\tau}_M(Z(W), Z(\mathcal{W}))} \|Z^{\mathrm{Adv}}(\widehat{W}_{td}, \cdot) - Z^{\mathrm{Adv}}(\widehat{\mathcal{W}}_t, \cdot)\|_\Gamma \leq Cd^{-\varepsilon}, \end{aligned} \tag{150}$$

where $C > 0$ *is a constant independent of* d, *where* $\mathcal{Z}^{\mathrm{Adv}}(t, z)$ *is the deterministic solution to the integro-differential equation* (34) *and*

$$\widehat{\tau}_M(\mathscr{Z}_1, \mathscr{Z}_2) = \min\{\tau_M(\mathscr{Z}_1), \tau_M(\mathscr{Z}_2)\}.$$

As in Section C.3, the next Theorem replaces of the stopping time $\widehat{\tau}_{M,\omega}$ with $\Theta_{M,\omega}$. We omit its proof as it uses the same argument as the proof Theorem C.2 but applied to $\widehat{\mathcal{B}}_i^{\mathrm{Adv}}(\mathscr{X}_j^{\mathrm{Adv}})$ instead.

Theorem C.5. *There exists an* $\varepsilon \in (0, \frac{1}{2})$ *such that for any* $T, M, \omega > 0$ *and* d *sufficiently large, with overwhelming probability*

$$\begin{aligned} &\sup_{0 \leq t \leq T \wedge \Theta_{M,\omega}(Z(W,\cdot), \mathcal{Z})} \|Z^{\mathrm{Adv}}(\widehat{W}_{td}, \cdot) - \mathcal{Z}^{\mathrm{Adv}}(t, \cdot)\|_\Gamma \leq Cd^{-\varepsilon}, \\ &\sup_{0 \leq t \leq T \wedge \Theta_{M,\omega}(Z(\mathcal{W},\cdot), \mathcal{Z})} \|Z^{\mathrm{Adv}}(\widehat{\mathcal{W}}_t, \cdot) - \mathcal{Z}^{\mathrm{Adv}}(t, \cdot)\|_\Gamma \leq Cd^{-\varepsilon}, \\ \textit{and} \quad &\sup_{0 \leq t \leq T \wedge \Theta_{M,\omega}(Z(W,\cdot), Z(\mathcal{W},\cdot))} \|Z^{\mathrm{Adv}}(\widehat{W}_{td}, \cdot) - Z^{\mathrm{Adv}}(\widehat{\mathcal{W}}_t, \cdot)\|_\Gamma \leq Cd^{-\varepsilon}, \end{aligned} \tag{151}$$

where $C > 0$ *is a constant independent of* d, *where* $\mathcal{Z}^{\mathrm{Adv}}(t, z)$ *is the deterministic solution to the integro-differential equation* (32).

Hence, it follows from Theorem C.5 that SGD and AdvHSGD concentrate around $\mathcal{Z}^{\mathrm{Adv}}(t,z)$ provided that AdvHSGD or $\mathcal{Z}^{\mathrm{Adv}}(t,z)$ stay bounded. Analogously to Corollary C.1, we formalize this in the following corollary.

Corollary C.3. *Suppose for a fixed $T,\omega>0$ that*

$$\sup_{0\le t\le T}\mathscr{N}(t)\le M \quad \textit{and} \quad \inf_{i\in\{1,2\}}\inf_{0\le t\le T}\inf_{V\in\mathcal{U}^c}\|\widehat{\mathscr{B}}_i(t)-V\|>\omega \quad \textit{hold w.o.p.} \tag{152}$$

where $M>0$ is independent of d. Then there exists $\varepsilon\in(0,\frac{1}{2})$ so that for d sufficiently large, with overwhelming probability,

$$\sup_{0\le t\le T}\|Z^{\mathrm{Adv}}(\widehat{\mathcal{W}}_t,\cdot)-\mathcal{Z}^{\mathrm{Adv}}(t,\cdot)\|_\Gamma\le Cd^{-\varepsilon} \textit{ and } \sup_{0\le t\le T}\|Z^{\mathrm{Adv}}(\widehat{W}_{td},\cdot)-\mathcal{Z}^{\mathrm{Adv}}(t,\cdot)\|_\Gamma\le Cd^{-\varepsilon}, \tag{153}$$

where $C>0$ is a constant independent of d. Hence, by the triangle inequality, it follows that

$$\sup_{0\le t\le T}\|Z^{\mathrm{Adv}}(\widehat{W}_{td},\cdot)-Z^{\mathrm{Adv}}(\widehat{\mathcal{W}}_t,\cdot)\|_\Gamma\le 2Cd^{-\varepsilon}. \tag{154}$$

Finally, using the same argument as the proof Theorem C.3 which we also omit, we obtain a general concentration result for statistics satisfying Assumption 11, Recall the det. equiv. of $\varphi(\mathfrak{X}_t)$ and $\varphi(X_{td})$ for φ satisfying Assumption 11 is defined as

$$\phi^{\mathrm{Adv}}(t)=g\left(\left\{\oint q_j(z)\mathcal{Z}^{\mathrm{Adv}}(t,z)\,\mathrm{D}z\right\}_{j=1}^{\ell}\right). \tag{155}$$

Theorem C.6 (Concentration of any statistic)**.** *Suppose that the assumptions stated in this framework hold. Then there exists an $\varepsilon\in(0,\frac{1}{2})$ such that for any $T,M,\omega>0$ and d sufficiently large, with overwhelming probability*

$$\begin{aligned}
&\sup_{0\le t\le T\wedge\Theta^{\mathrm{Adv}}_{M,\omega}(Z(W,\cdot),\mathcal{Z})}|\varphi(X_{td})-\phi^{\mathrm{Adv}}(t)|\le Cd^{-\varepsilon},\\
&\sup_{0\le t\le T\wedge\Theta_{M,\omega}(Z(\mathcal{W},\cdot),\mathcal{Z})}|\varphi(\mathfrak{X}_t)-\phi^{\mathrm{Adv}}(t)|\le Cd^{-\varepsilon},\\
\textit{and}\quad &\sup_{0\le t\le T\wedge\Theta_{M,\omega}(Z(W,\cdot),Z(\mathcal{W},\cdot))}|\varphi(X_{td})-\varphi(\mathfrak{X}_t)|\le Cd^{-\varepsilon},
\end{aligned} \tag{156}$$

where $C>0$ is a constant independent of d, where ϕ^{Adv} is defined in (155)*.*

Corollary C.4. *Suppose that the assumptions stated in this framework hold and the assumptions in Corollary C.3 hold. Then there exists an $\varepsilon\in(0,\frac{1}{2})$ such that for any $T>0$ and sufficiently large d, with overwhelming probability,*

$$\sup_{0\le t\le T}|\varphi(\mathfrak{X}_t)-\phi^{\mathrm{Adv}}(t)|\le Cd^{-\varepsilon} \quad \textit{and} \quad \sup_{0\le t\le T}|\varphi(X_{td})-\phi^{\mathrm{Adv}}(t)|\le Cd^{-\varepsilon}. \tag{157}$$

Hence, by the triangle inequality, it follows that

$$\sup_{0\le t\le T}|\varphi(X_{td})-\varphi(\mathfrak{X}_t)|\le 2Cd^{-\varepsilon}. \tag{158}$$

D Preliminaries for the Proofs

The results in this section build upon (Collins-Woodfin et al., 2024a; Collins-Woodfin & Seroussi, 2025) in order to fit the current framework. We refer the reader to Section 3 of Collins-Woodfin et al. (2024a) for a thorough overview to contractions over abtract vector spaces and tensor differentiation. To facilitate understanding of the proofs and streamline notation, we will include results from these papers and use similar notation. Recall we defined $r_{k,i}=\hat{X}_k^\top a_{k+1,i}$ such that

$$\begin{aligned}
g_{k,i}&=r_{k,i}+\begin{bmatrix}\delta s_{k,i}\|X_k\| & 0\end{bmatrix}^\top,\\
s_{k,i}&=\mathrm{argmax}_{|s|\le 1}\,f_i(X_k^\top a_{k+1,i}+\delta s\|X_k\|;\epsilon_{k+1}),
\end{aligned} \tag{159}$$

where $a_{k,I_k}=\sqrt{K_{I_k}}v_k+\mu_{I_k}$ for $v_k\sim\mathcal{N}(0,I_d)$.

D.1 Derivative of Special Statistics

In this section, we compute the derivative of the adversarial risk function $\mathcal{R}^{\mathrm{adv}}$. From Assumption 3, the functions f_i α-pseudo-Lipschitz for $i = 1, 2$. Hence, on any bounded set, the functions f_i and and $\Psi_i : \mathbb{R}^d \to \mathbb{R}$ defined in (3) are Lipschitz so differentiable a.e. by Radamacher's theorem. Thus, the derivatives f_i' and $\nabla_X \Psi_i$, exist a.e. In section D.2, we derive growth bounds on f_i'. We will now to compute the derivative of the risk $\mathcal{R}^{\mathrm{adv}}(X)$. We omit the proof as it follows a similar argument to Lemma 2 in Collins-Woodfin et al. (2024a).

Lemma D.1 (Derivatives of the statistic, $\mathcal{R}^{\mathrm{adv}}$)**.** *Suppose the risk is $\mathcal{R}^{\mathrm{adv}}(X)$ satisfies Assumption 6. Then, one has*

$$\nabla \mathcal{R}^{\mathrm{adv}}(X) = 2 \sum_{i=1}^{2} p_i \Big(\partial_{11} h_i(\widehat{B}_i(W)) \cdot K_i X + \partial_{12} h_i(\widehat{B}_i(W)) \cdot K_i X^\star + \partial_{13} h_i(\widehat{B}_i(W)) \cdot \mu_i + \partial_{44} h_i(\widehat{B}_i(W)) \cdot X \Big)$$

where $h : \mathbb{R}^{4\times 4} \to \mathbb{R}$ with gradient

$$\nabla h_i = \left[\begin{array}{c|c|c|c} \partial_{11} h_i & \partial_{12} h_i & \partial_{13} h_i & 0 \\ \hline \partial_{21} h_i & \partial_{22} h_i & \partial_{23} h_i & 0 \\ \hline \partial_{31} h_i & \partial_{32} h_i & \partial_{33} h_i & 0 \\ \hline 0 & 0 & 0 & \partial_{44} h_i \end{array}\right] \in \mathbb{R}^{4\times 4}.$$

D.2 Concentration and Pseudo-Lipschitz

In order to simplify the proofs in section D.5, we will make use of the subgaussian norm $\|\cdot\|_{\psi_2}$ (see e.g., (Vershynin, 2018) for more details):

$$\|X\|_{\psi_2} \asymp \inf\{V > 0 : \forall\, t > 0 \ \Pr(|X| > t) \le 2e^{-t^2/V^2}\}. \tag{160}$$

We will also adapt some concentration results for pseudo-Lipschitz functions from Collins-Woodfin et al. (2024a) to our adversarial framework.

The proofs in section D.5 will involve statistics whose gradients are α-pseudo-Lipschitz. Before introducing bounds on the norm of the gradient of α-pseudo-Lipschitz functions, we define the following projection operator onto the ball of radius β in the vector space V, $\mathrm{Proj}_\beta : V \to V$:

$$\mathrm{Proj}_\beta(x) \stackrel{\mathrm{def}}{=} \begin{cases} x & \text{if } \|x\| \le \beta \\ \beta \frac{x}{\|x\|} & \text{otherwise.} \end{cases} \tag{161}$$

It can be shown that taking compositions of projections with α-pseudo-Lipschitz functions creates Lipschitz functions.

Lemma D.2 (Lemma 3, (Collins-Woodfin et al., 2024a))**.** *Suppose f is α-pseudo-Lipschitz with constant L. Then the composition $f \circ \mathrm{Proj}_\beta$ is Lipschitz with constant $L(1 + 2\beta^\alpha)$.*

By Assumption 3 f α-pseudo-Lipschitz. We build upon Lemma 3.4 from Collins-Woodfin et al. (2024a) to obtain some bounds on the growth of moments of f' in terms of $\hat{X} = X \oplus X^\star$. We omit the proof as it follows an almost identical argument.

Lemma D.3 (Growth of f')**.** *Suppose a function $f : \mathbb{R}^3 \to \mathbb{R}$ is α-pseudo-Lipschitz with constant $L(f)$. Suppose the noise $\epsilon \sim \mathcal{N}(0, \eta^2)$ is independent of $v \sim \mathcal{N}(0, K)$. Then for $p > 0$ and any $g \in \mathbb{R}^2$,*

$$|f'(g)|^p \le C(\alpha, p)(L(f))^p (1 + \|g\| + |\epsilon|)^{\max\{1, \alpha p\}}, \tag{162}$$

Moreover, for $\mu \in \mathbb{R}^d$, if $g = \hat{X}^\top v + \hat{X}^\top \mu + [\delta s \|X\| \quad 0]^\top$ with $\hat{X} = X \oplus X^\star$ and $|s| \le 1$, there is a growth rate on $f'(g)$ and sub-Gaussian norm on g in terms of $\hat{X}$, X and η,

$$\begin{aligned} &\mathbb{E}_{v,\epsilon}[|f'(g)|^p] \le C(\alpha, p)(L(f))^p \big(1 + (\|K\|_{\mathrm{op}}^{1/2} + \|\mu\|)\|\hat{X}\| + \delta\|X\| + \eta\big)^{\max\{1,\alpha p\}} \\ &\quad \textit{and} \qquad \|(1 + \|g\| + |\epsilon|)\|_{\psi_2} \le C(1 + (\|K\|_{\mathrm{op}}^{1/2} + \|\mu\|)\|\hat{X}\| + \delta\|X\| + \eta). \end{aligned} \tag{163}$$

D.3 Doob Decomposition of Adversarial Homogenized SGD

The arguments in this section are built upon the results presented in Appendix A of Collins-Woodfin et al. (2024a) and Section 5 in Collins-Woodfin & Seroussi (2025). We present the steps for the soft label setting as the hard label setting follows similarly. We are interested in the behavior of AdvHSGD under the following statistics:

$$S(W,z) \stackrel{\text{def}}{=} \hat{X}^\top \mathcal{R}(z;\mathcal{K})\hat{X} \in \mathbb{C}^{2\times 2}, \quad M(W,z) \stackrel{\text{def}}{=} \hat{X}^\top \mathcal{R}(z;\mathcal{K})\mu \in \mathbb{C}^{2\times 2}, \quad \text{for} \quad z \in \Gamma^2, \tag{164}$$

where we recall $\mathcal{K} \stackrel{\text{def}}{=} \{K_1, K_2\}$, $\Gamma \stackrel{\text{def}}{=} \{w \in \mathbb{C} : |w| = \max(1, 2\|\mathcal{K}\|_{\text{op}})\}$ with $\|\mathcal{K}\|_{\text{op}} \stackrel{\text{def}}{=} \max_{i=1,2} \|K_i\|_{\text{op}}$ and $\mathcal{R}(z;\mathcal{K}) \stackrel{\text{def}}{=} R(z_1;K_1)R(z_2;K_2)$ with $z_i \in \Gamma$ for $i = 1, 2$. Recall from (27) the definition of AdvHSGD

$$\mathrm{d}\mathcal{X}_t = -\gamma(t)\nabla\mathcal{R}_\lambda^{\text{adv}}(\mathcal{X}_t)\,\mathrm{d}t + \gamma(t)\sqrt{\frac{1}{d}\sum_{i=1}^{2} p_i \mathbb{E}_{v,\epsilon}[f_i'(\rho_{t,i},\epsilon)^2](K_i + \mu_i\mu_i^\top)}\,\mathrm{d}B_t, \tag{165}$$

where the initial condition is given by $\mathcal{X}_0 = X_0$, $(B_t, t \geq 0)$ is a $\mathbb{R}^d$ standard Brownian motion and for $I \in \{1, 2\}$ and $v \sim \mathcal{N}(0, I_d)$ we have

$$\begin{aligned} \widehat{\mathcal{X}}_t &\stackrel{\text{def}}{=} \mathcal{X}_t \oplus X^\star \in \mathbb{R}^{d\times 2}, \quad \mathcal{W}_t \stackrel{\text{def}}{=} \widehat{\mathcal{X}}_t \oplus \mu \in \mathbb{R}^{d\times 4}, \\ \rho_{t,I} &\stackrel{\text{def}}{=} \widehat{\mathcal{X}}_t^\top(\sqrt{K_I}v + \mu_I) + [\delta u_{t,I}\|\mathcal{X}_t\| \quad 0]^\top, \\ \text{and} \quad u_{t,I} &= \text{argmax}_{|u|\leq 1} f(\widehat{\mathcal{X}}_t^\top(\sqrt{K_I}v + \mu_I) + [\delta u\|\mathcal{X}_t\| \quad 0]^\top, \epsilon). \end{aligned} \tag{166}$$

From (164), it is immediate that $S(\mathcal{W}_t, z)$ is a bilinear form. Hence, we decompose it into block form

$$S(\mathcal{W}_t,z) = \left[\begin{array}{c|c} \mathcal{X}_t^\top\mathcal{R}(z;\mathcal{K})\mathcal{X}_t & \mathcal{X}_t^\top\mathcal{R}(z;\mathcal{K})X^\star \\ \hline (X^{\star\top})\mathcal{R}(z;\mathcal{K})\mathcal{X}_t & (X^\star)^\top\mathcal{R}(z;\mathcal{K})X^\star \end{array}\right] = \left[\begin{array}{c|c} S_{11}(\mathcal{W}_t,z) & S_{12}(\mathcal{W}_t,z) \\ \hline S_{21}(\mathcal{W}_t,z) & S_{22}(\mathcal{W}_t,z) \end{array}\right].$$

With this notation established, we derive an explicit equation for $S(\mathcal{W}_t, z)$ in terms of its martingale increment to facilitate the proof Proposition C.4 using Itô calculus.

Itô calculus applied to $S(\mathcal{W}_t, z)$. From Assumption 6, recall that the adversarial expected risk $\mathcal{R}^{\text{adv}}$ can be decomposed as the sum of two function compositions

$$\mathcal{R}^{\text{adv}}(\mathcal{X}_t) = \sum_{i=1}^{2} p_i \cdot (h_i \circ \widehat{B}_i(\widehat{\mathcal{W}}_t)),$$

for some functions $h_i : \mathbb{R}^{4\times 4} \to \mathbb{R}$ for $i = 1, 2$ and

$$\widehat{B}_i(\mathcal{W}_t) \stackrel{\text{def}}{=} \begin{bmatrix} U_i(\mathcal{W}_t) & 0 \\ 0 & \|\mathcal{X}_t\|^2 \end{bmatrix}, \quad U_i(\mathcal{W}_t) \stackrel{\text{def}}{=} \begin{bmatrix} B_i(\mathcal{W}_t) & m_i(\mathcal{W}_t) \\ m_i(\mathcal{W}_t)^\top & \|\mu_i\|^2 \end{bmatrix}, \quad B_i(\mathcal{W}_t) \stackrel{\text{def}}{=} \widehat{\mathcal{X}}_t^\top K_i \widehat{\mathcal{X}}_t \tag{167}$$

and $m_i(\mathcal{W}_t) \stackrel{\text{def}}{=} \widehat{\mathcal{X}}_t^\top \mu_i$ Recall from Lemma D.1 the gradient of the adversarial risk

$$\begin{aligned} \nabla\mathcal{R}^{\text{adv}}(X) = 2\sum_{i=1}^{2} p_i \Big(&\partial_{11}h_i(\widehat{B}_i(W)) \cdot K_i X + \partial_{12}h_i(\widehat{B}_i(W)) \cdot K_i X^\star \\ &+ \partial_{13}h_i(\widehat{B}_i(W)) \cdot \mu_i + \partial_{44}h_i(\widehat{B}_i(W)) \cdot X \Big) \end{aligned}$$

For simplicity of notation, we will omit the composition of the partial derivatives with $\widehat{B}_i(\mathcal{W}_t)$ in the next steps of the proof. Since $\mathrm{d}\widehat{\mathfrak{X}}_t = \mathrm{d}\mathfrak{X}_t \oplus 0_d$ and using the product rule for Itô derivatives,

$$
\begin{aligned}
\mathrm{d}S &= \langle \mathrm{d}\widehat{\mathfrak{X}}_t \otimes \widehat{\mathfrak{X}}_t, \mathcal{R}(z;\mathcal{K})\rangle_{\mathbb{R}^{d\times d}} + \langle \widehat{\mathfrak{X}}_t \otimes \mathrm{d}\widehat{\mathfrak{X}}_t, \mathcal{R}(z;\mathcal{K})\rangle_{\mathbb{R}^{d\times d}} + \langle \mathrm{d}\widehat{\mathfrak{X}}_t \otimes \mathrm{d}\widehat{\mathfrak{X}}_t, \mathcal{R}(z;\mathcal{K})\rangle_{\mathbb{R}^{d\times d}} \\
&= \langle (\mathrm{d}\mathfrak{X}_t \oplus 0_d) \otimes \widehat{\mathfrak{X}}_t, \mathcal{R}(z;\mathcal{K})\rangle_{\mathbb{R}^{d\times d}} + \langle \widehat{\mathfrak{X}}_t \otimes (\mathrm{d}\mathfrak{X}_t \oplus 0_d), \mathcal{R}(z;\mathcal{K})\rangle_{\mathbb{R}^{d\times d}} \\
&\quad + \langle (\mathrm{d}\mathfrak{X}_t \oplus 0_d) \otimes (\mathrm{d}\mathfrak{X}_t \oplus 0_d), \mathcal{R}(z;\mathcal{K})\rangle_{\mathbb{R}^{d\times d}} \\
&= -\gamma(t) \cdot \langle ((\nabla \mathcal{R}^{\mathrm{adv}}(\mathfrak{X}_t) + \lambda \mathfrak{X}_t) \oplus 0_d) \otimes \widehat{\mathfrak{X}}_t, \mathcal{R}(z;\mathcal{K})\rangle_{\mathbb{R}^{d\times d}}\, \mathrm{d}t \\
&\quad - \gamma(t) \cdot \langle \widehat{\mathfrak{X}}_t \otimes ((\nabla \mathcal{R}^{\mathrm{adv}}(\mathfrak{X}_t) + \lambda \mathfrak{X}_t) \oplus 0_d), \mathcal{R}(z;\mathcal{K})\rangle_{\mathbb{R}^{d\times d}}\, \mathrm{d}t \\
&\quad + \begin{bmatrix} \frac{\gamma_t^2}{d} \sum_{i=1}^2 p_i \mathbb{E}_{\,v,\epsilon}[f_i'(\rho_{t,i},\epsilon)^2] \operatorname{Tr}((K_i + \mu_i\mu_i^\top)\mathcal{R}(z;\mathcal{K}))\, \mathrm{d}t & 0 \\ 0 & 0 \end{bmatrix} + \mathrm{d}\mathcal{M}_t^{\mathrm{AdvHSGD}}(S(\mathcal{W}_t, z)),
\end{aligned}
\tag{168}
$$

where we define the martingale increment of $S(\mathcal{W}_t, z)$ elementwise such that for $1 \le j, \ell \le 2$

$$
\begin{aligned}
\big(\mathrm{d}\mathcal{M}_t^{\mathrm{AdvHSGD}}(S(\mathcal{W}_t, z))\big)_{j\ell} &\overset{\mathrm{def}}{=} \mathrm{d}\mathcal{M}_t^{\mathrm{AdvHSGD}}(S_{j\ell}(\mathcal{W}_t, z)) \\
&\overset{\mathrm{def}}{=} \gamma(t) \left\langle \sqrt{\frac{1}{d}\sum_{i=1}^2 p_i \mathbb{E}_{\,v,\epsilon}[f_i'(\rho_{t,i},\epsilon)^2](K_i + \mu_i\mu_i^\top)}, \nabla_X(S_{j\ell}(\mathcal{W}_t, z)) \otimes \mathrm{d}B_t \right\rangle.
\end{aligned}
$$

Similarly, we define the martingale term elementwise

$$
\big(\mathcal{M}_t^{\mathrm{AdvHSGD}}(S(\cdot, z))\big)_{j\ell} = \int_0^t \mathrm{d}\mathcal{M}_s^{\mathrm{AdvHSGD}}(S_{j\ell}(\cdot, z)).
$$

Plugging in $\nabla\mathcal{R}^{\mathrm{adv}}$ into the first term of (168) and Developing, we have

$$
\begin{aligned}
&\langle ((\nabla \mathcal{R}^{\mathrm{adv}}(\mathfrak{X}_t) + \lambda \mathfrak{X}_t) \oplus 0_d) \otimes \widehat{\mathfrak{X}}_t, \mathcal{R}(z;\mathcal{K})\rangle_{\mathbb{R}^{d\times d}}\, \mathrm{d}t \\
&= 2\sum_{i=1}^2 p_i \Big(\langle (\partial_{11} h_i \cdot K_i \mathfrak{X}_t \oplus 0_d) \otimes \widehat{\mathfrak{X}}_t, \mathcal{R}(z;\mathcal{K})\rangle_{\mathbb{R}^{d\times d}} \\
&\qquad + \langle (\partial_{12} h_i \cdot K_i X^\star \oplus 0_d) \otimes \widehat{\mathfrak{X}}_t, \mathcal{R}(z;\mathcal{K})\rangle_{\mathbb{R}^{d\times d}} + \langle (\partial_{13} h_i \cdot \mu_i \oplus 0_d) \otimes \widehat{\mathfrak{X}}_t, \mathcal{R}(z;\mathcal{K})\rangle_{\mathbb{R}^{d\times d}} \\
&\qquad + \langle (\partial_{44} h_i \cdot \mathfrak{X}_t \oplus 0_d) \otimes \widehat{\mathfrak{X}}_t, \mathcal{R}(z;\mathcal{K})\rangle_{\mathbb{R}^{d\times d}} \Big)\, \mathrm{d}t + \lambda \langle (\mathfrak{X}_t \oplus 0_d) \otimes \widehat{\mathfrak{X}}_t, \mathcal{R}(z;\mathcal{K})\rangle_{\mathbb{R}^{d\times d}}\, \mathrm{d}t
\end{aligned}
$$

Expanding the terms with $\widehat{\mathfrak{X}}_t = \mathfrak{X}_t \oplus X^\star$ then we obtain

$$
-\gamma(t)\cdot\langle ((\nabla \mathcal{R}^{\mathrm{adv}}(\mathfrak{X}_t) + \lambda \mathfrak{X}_t) \oplus 0_d) \otimes \widehat{\mathfrak{X}}_t, \mathcal{R}(z;\mathcal{K})\rangle_{\mathbb{R}^{d\times d}}\, \mathrm{d}t = \left[\begin{array}{c|c} A_1 + \tilde{A}_1 & E + \tilde{E} \\ \hline 0 & 0 \end{array}\right],
$$

$$
\begin{aligned}
\text{where} \quad A_1 &= -2\gamma(t) \sum_{i=1}^2 p_i \Big(\partial_{11} h_i \cdot \langle \mathfrak{X}_t \otimes \mathfrak{X}_t, K_i \mathcal{R}(z;\mathcal{K})\rangle + \partial_{12} h_i \cdot \langle X^\star \otimes \mathfrak{X}_t, K_i \mathcal{R}(z;\mathcal{K})\rangle \\
&\qquad + \partial_{13} h_i \cdot \langle \mu_i \otimes \mathfrak{X}_t, \mathcal{R}(z;\mathcal{K})\rangle + \partial_{44} h_i \cdot \langle \mathfrak{X}_t \otimes \mathfrak{X}_t, \mathcal{R}(z;\mathcal{K})\rangle \Big)\, \mathrm{d}t, \\
\tilde{A}_1 &= -\gamma(t) \cdot \lambda \cdot \langle \mathfrak{X}_t \otimes \mathfrak{X}_t, \mathcal{R}(z;\mathcal{K})\rangle\, \mathrm{d}t, \\
E &= -2\gamma(t) \sum_{i=1}^2 p_i \Big(\partial_{11} h_i \cdot \langle \mathfrak{X}_t \otimes X^\star, K_i \mathcal{R}(z;\mathcal{K})\rangle + \partial_{12} h_i \cdot \langle X^\star \otimes X^\star, K_i \mathcal{R}(z;\mathcal{K})\rangle \\
&\qquad + \partial_{13} h_i \cdot \langle \mu_i \otimes X^\star, \mathcal{R}(z;\mathcal{K})\rangle + \partial_{44} h_i \cdot \langle \mathfrak{X}_t \otimes X^\star, \mathcal{R}(z;\mathcal{K})\rangle \Big)\, \mathrm{d}t, \\
\text{and} \quad \tilde{E} &= -\gamma(t) \cdot \lambda \cdot \langle \mathfrak{X}_t \otimes X^\star, \mathcal{R}(z;\mathcal{K})\rangle\, \mathrm{d}t.
\end{aligned}
\tag{169}
$$

Similarly for the second term in (168),

$$-\gamma(t)\cdot\langle\widehat{\mathfrak{X}}_t\otimes((\nabla\mathcal{R}^{\mathrm{adv}}(\mathfrak{X}_t)+\lambda\mathfrak{X}_t)\oplus 0_d),\mathcal{R}(z;\mathcal{K})\rangle_{\mathbb{R}^{d\times d}}\,\mathrm{d}t=\left[\begin{array}{c|c}A_2+\tilde{A}_2 & 0\\ \hline C+\tilde{C} & 0\end{array}\right],$$

$$\begin{aligned}
\text{where}\quad A_2 &= -2\gamma(t)\sum_{i=1}^{2}p_i\Big(\partial_{11}h_i\cdot\langle\mathfrak{X}_t\otimes\mathfrak{X}_t,K_i\mathcal{R}(z;\mathcal{K})\rangle+\partial_{12}h_i\cdot\langle\mathfrak{X}_t\otimes X^\star,K_i\mathcal{R}(z;\mathcal{K})\rangle\\
&\quad+\partial_{13}h_i\cdot\langle\mathfrak{X}_t\otimes\mu_i,\mathcal{R}(z;\mathcal{K})\rangle+\partial_{44}h_i\cdot\langle\mathfrak{X}_t\otimes\mathfrak{X}_t,\mathcal{R}(z;\mathcal{K})\rangle\Big)\,\mathrm{d}t,\\
\tilde{A}_2 &= -\gamma(t)\cdot\lambda\cdot\langle\mathfrak{X}_t\otimes\mathfrak{X}_t,\mathcal{R}(z;\mathcal{K})\rangle\,\mathrm{d}t,\\
C &= -2\gamma(t)\sum_{i=1}^{2}p_i\Big(\partial_{11}h_i\cdot\langle X^\star\otimes\mathfrak{X}_t,K_i\mathcal{R}(z;\mathcal{K})\rangle+\partial_{12}h_i\cdot\langle X^\star\otimes X^\star,K_i\mathcal{R}(z;\mathcal{K})\rangle\\
&\quad+\partial_{13}h_i\cdot\langle X^\star\otimes\mu_i,\mathcal{R}(z;\mathcal{K})\rangle+\partial_{44}h_i\cdot\langle X^\star\otimes\mathfrak{X}_t,\mathcal{R}(z;\mathcal{K})\rangle\Big)\,\mathrm{d}t,\\
\text{and}\quad \tilde{C} &= -\gamma(t)\cdot\lambda\cdot\langle X^\star\otimes\mathfrak{X}_t,\mathcal{R}(z;\mathcal{K})\rangle\,\mathrm{d}t.
\end{aligned}\tag{170}$$

Finally, denote the last term in (168) as

$$A_3=\frac{\gamma_t^2}{d}\sum_{i=1}^{2}p_i\mathbb{E}_{\,v,\epsilon}[f_i'(\rho_{t,i},\epsilon)^2]\,\mathrm{Tr}((K_i+\mu_i\mu_i^\top)\mathcal{R}(z;\mathcal{K}))\,\mathrm{d}t.$$

It follows then that

$$\begin{aligned}
(\mathrm{d}S)(\mathcal{W}_t,z)&=\left[\begin{array}{c|c}\mathrm{d}S_{11} & \mathrm{d}S_{12}\\ \hline \mathrm{d}S_{21} & \mathrm{d}S_{22}\end{array}\right]\\
&=\left[\begin{array}{c|c}A_1+\tilde{A}_1+A_2+\tilde{A}_2+A_3 & E+\tilde{E}\\ \hline C+\tilde{C} & 0\end{array}\right]+\mathrm{d}\mathcal{M}_t^{\mathrm{AdvHSGD}}(S(\mathcal{W}_t,z)).
\end{aligned}$$

Through a simple computation, we observe that

$$K_1\mathcal{R}(z;\mathcal{K})=R(z_2;K_2)+z_1\mathcal{R}(z;\mathcal{K}).$$

Similarly, we have $K_2\mathcal{R}(z;\mathcal{K})=R(z_1;K_1)+z_2\mathcal{R}(z;\mathcal{K})$. From Cauchy's integral formula, we also note that

$$\widehat{\mathfrak{X}}_t^\top R(z_1;K_1)\widehat{\mathfrak{X}}_t=\frac{-1}{2\pi\,\mathrm{i}}\oint_\Gamma S(\mathcal{W}_t,z)\,\mathrm{d}z_2.$$

Similar results hold for z_2. Define $M_i(\mathcal{W}_t,z)\overset{\mathrm{def}}{=}\widehat{\mathfrak{X}}_t^\top\mathcal{R}(z;\mathcal{K})\mu_i$ for $i=1,2$. Using the above identity, the quantities $A_1,A_2,E,C,\tilde{A}_1,\tilde{A}_2,\tilde{E}$, and $\tilde{C}$ from (169) and (170) can be reexpressed in terms of $S(\mathcal{W}_t,z)=\langle\widehat{\mathfrak{X}}_t\otimes\widehat{\mathfrak{X}}_t,\mathcal{R}(z;\mathcal{K})\rangle$. It follows that

$$\begin{aligned}
\mathrm{d}S(\mathcal{W}_t,z)&=-2\gamma(t)\sum_{i=1}^{2}p_i\Bigg(\left(z_iS(\mathcal{W}_t,z)-\frac{1}{2\pi\,\mathrm{i}}\oint_\Gamma S(\mathcal{W}_t,z)\,\mathrm{d}z_i\right)H_{1,i}(\widehat{B}_i(\mathcal{W}_t))\\
&\quad+H_{1,i}^\top(\widehat{B}_i(\mathcal{W}_t))\left(z_iS(\mathcal{W}_t,z)-\frac{1}{2\pi\,\mathrm{i}}\oint_\Gamma S(\mathcal{W}_t,z)\,\mathrm{d}z_i\right)\\
&\quad+M_i(\mathcal{W}_t,z)\otimes H_{2,i}(\widehat{B}_i(\mathcal{W}_t))+H_{2,i}(\widehat{B}_i(\mathcal{W}_t))\otimes M_i(\mathcal{W}_t,z)\\
&\quad+\left(\frac{\lambda}{2}+\partial_{44}h_i(\widehat{B}_i(\mathcal{W}_t))\right)(S(\mathcal{W}_t,z)\,\mathrm{D}+\mathrm{D}\,S(\mathcal{W}_t,z))\Bigg)\,\mathrm{d}t\\
&\quad+\frac{\gamma_t^2}{d}\sum_{i=1}^{2}p_i\mathbb{E}_{\,v,\epsilon}[f_i'(\rho_{t,i},\epsilon)^2]\,\mathrm{Tr}((K_i+\mu_i\mu_i^\top)\mathcal{R}(z;\mathcal{K}))\,\mathrm{d}t\,\mathrm{D}\,\mathrm{d}t\\
&\quad+\mathrm{d}\mathcal{M}_t^{\mathrm{AdvHSGD}}(S(\mathcal{W}_t,z)),
\end{aligned}\tag{171}$$

$$\text{where}\quad H_{1,i}(\widehat{B}_i)=\left[\begin{array}{c|c}\partial_{11}h_i(\widehat{B}_i) & 0\\ \hline \partial_{21}h_i(\widehat{B}_i) & 0\end{array}\right],\quad H_{2,i}(\widehat{B}_i)=\left[\begin{array}{c}\partial_{13}h_i(\widehat{B}_i)\\ \hline 0\end{array}\right],\quad \mathrm{D}=\left[\begin{array}{c|c}1 & 0\\ \hline 0 & 0\end{array}\right],$$

$$\text{and initialized with } S(0,z)=\hat{X}_0^\top\mathcal{R}(z;\mathcal{K})\hat{X}_0.$$

Now, recall by Assumption 7, there exists a function $\mathcal{I}_i : \mathbb{R}^{4\times 4} \to \mathbb{R}$ such that

$$\mathbb{E}_{v,\epsilon}[f_i'(\rho_{t,i},\epsilon)^2] = \mathcal{I}_i(\widehat{B}_i(\mathcal{W}_t)).$$

From the Cauchy integral formula, we have

$$\widehat{B}_i(W) = \begin{bmatrix} \oint z_i S(W,z)\ \mathrm{D}z & \oint M_i(W,z)\ \mathrm{D}z & 0 \\ \oint M_i(W,z)^\top\ \mathrm{D}z & \oint \mu_i^\top \mathcal{R}(z;\mathcal{K})\mu_i\ \mathrm{D}z & 0 \\ 0 & 0 & \oint S_{11}(W,z)\ \mathrm{D}z \end{bmatrix}. \tag{172}$$

where $M_i(W,z)$ denotes the ith columns of $M(W,z)$. We must now derive the Doob decomposition of $M(\mathcal{W}_t,z)$. We first decompose it into block form

$$M(\mathcal{W}_t,z) = \left[\begin{array}{c|c} \mathcal{X}_t^\top \mathcal{R}(z;\mathcal{K})\mu_1 & \mathcal{X}_t^\top \mathcal{R}(z;\mathcal{K})\mu_2 \\ \hline X^{\star\top} \mathcal{R}(z;\mathcal{K})\mu_1 & X^{\star\top} \mathcal{R}(z;\mathcal{K})\mu_2 \end{array}\right] = \left[\begin{array}{c|c} M_{11}(\mathcal{W}_t,z) & M_{12}(\mathcal{W}_t,z) \\ \hline M_{21}(\mathcal{W}_t,z) & M_{22}(\mathcal{W}_t,z) \end{array}\right] \in \mathbb{R}^{2\times 2}.$$

From Itô's lemma, we have

$$\begin{aligned} \mathrm{d}M(\mathcal{W}_t,z) &= (\mathrm{d}\widehat{\mathcal{X}}_t)^\top \mathcal{R}(z;\mathcal{K})\mu \\ &= -\gamma(t)\langle((\nabla \mathcal{R}_\lambda^{\mathrm{adv}}(\mathcal{X}_t) + \lambda \mathcal{X}_t) \oplus 0_d) \otimes \mu, \mathcal{R}(z;\mathcal{K})\rangle_{\mathbb{R}^{d\times d}}\, \mathrm{d}t + \mathrm{d}\mathcal{M}_t^{\mathrm{AdvHSGD}}(M(\mathcal{W}_t,z)), \end{aligned} \tag{173}$$

where for $1 \le j,\ell \le 2$

$$\begin{aligned} &\mathrm{d}\mathcal{M}_t^{\mathrm{AdvHSGD}}(M_{j\ell}(\mathcal{W}_t,z)) \\ &\quad \stackrel{\mathrm{def}}{=} \gamma(t) \left\langle \sqrt{\frac{1}{d}\sum_{i=1}^2 p_i \mathbb{E}_{v,\epsilon}[f_i'(\rho_{t,i},\epsilon)^2](K_i + \mu_i\mu_i^\top)}, \nabla_X (M_{j\ell}(\mathcal{W}_t,z)) \otimes \mathrm{d}B_t \right\rangle. \end{aligned} \tag{174}$$

Using a similar argument as we did in (169), we obtain

$$-\gamma(t)\cdot\langle((\nabla \mathcal{R}^{\mathrm{adv}}(\mathcal{X}_t) + \lambda \mathcal{X}_t) \oplus 0_d) \otimes \mu, \mathcal{R}(z;\mathcal{K})\rangle_{\mathbb{R}^{d\times d}}\ \mathrm{d}t = \left[\begin{array}{c|c} O_1 + \tilde{O}_1 & P + \tilde{P} \\ \hline 0 & 0 \end{array}\right],$$

$$\begin{aligned} \text{where}\quad O_1 &= -2\gamma(t)\sum_{i=1}^2 p_i \Big(\partial_{11} h_i \cdot \langle \mathcal{X}_t \otimes \mu_1, K_i \mathcal{R}(z;\mathcal{K})\rangle + \partial_{12} h_i \cdot \langle X^\star \otimes \mu_1, K_i \mathcal{R}(z;\mathcal{K})\rangle \\ &\qquad + \partial_{13} h_i \cdot \langle \mu_i \otimes \mu_1, \mathcal{R}(z;\mathcal{K})\rangle + \partial_{44} h_i \cdot \langle \mathcal{X}_t \otimes \mu_1, \mathcal{R}(z;\mathcal{K})\rangle \Big)\, \mathrm{d}t, \\ \tilde{O}_1 &= -\gamma(t)\cdot\lambda\cdot\langle \mathcal{X}_t \otimes \mathcal{X}_t, \mathcal{R}(z;\mathcal{K})\rangle\ \mathrm{d}t, \\ P &= -2\gamma(t)\sum_{i=1}^2 p_i \Big(\partial_{11} h_i \cdot \langle \mathcal{X}_t \otimes \mu_2, K_i \mathcal{R}(z;\mathcal{K})\rangle + \partial_{12} h_i \cdot \langle X^\star \otimes \mu_2, K_i \mathcal{R}(z;\mathcal{K})\rangle \\ &\qquad + \partial_{13} h_i \cdot \langle \mu_i \otimes \mu_2, \mathcal{R}(z;\mathcal{K})\rangle + \partial_{44} h_i \cdot \langle \mathcal{X}_t \otimes \mu_2, \mathcal{R}(z;\mathcal{K})\rangle \Big)\, \mathrm{d}t, \\ \text{and}\quad \tilde{P} &= -\gamma(t)\cdot\lambda\cdot\langle \mathcal{X}_t \otimes X^\star, \mathcal{R}(z;\mathcal{K})\rangle\ \mathrm{d}t. \end{aligned} \tag{175}$$

It then follows from the resolvent identities presented above that

$$\begin{aligned} \mathrm{d}M(\mathcal{W}_t,z) &= -2\gamma(t)\sum_{i=1}^2 p_i \Big(H_{1,i}^\top(\widehat{B}_i(\mathcal{W}_t)) \Big(z_i M(\mathcal{W}_t,z) - \frac{1}{2\pi\,\mathrm{i}} \oint_\Gamma M(\mathcal{W}_t,z)\,\mathrm{d}z_i \Big) \\ &+ H_{2,i}(\widehat{B}_i(\mathcal{W}_t)) \otimes \mu^\top \mathcal{R}(z;\mathcal{K})\mu_i + \Big(\frac{\lambda}{2} + \partial_{44} h_i(\widehat{B}_i(\mathcal{W}_t)) \Big)\, \mathrm{D}\, M(\mathcal{W}_t,z) \Big)\, \mathrm{d}t \\ &+ \mathrm{d}\mathcal{M}_t^{\mathrm{AdvHSGD}}(M(\mathcal{W}_t,z)). \end{aligned} \tag{176}$$

From the Doob decompositions (171) and (176), we obtain the representation

$$\mathrm{d}Z(\mathcal{W}_t,\cdot) = \mathcal{F}(z, Z(\mathcal{W}_t,\cdot))\ \mathrm{d}t + \mathrm{d}\mathcal{M}_t^{\mathrm{AdvHSGD}}(Z(\mathcal{W}_t,\cdot)), \tag{177}$$

where

$$\mathcal{F}(z, Z(\mathcal{W}_t, \cdot)) \stackrel{\text{def}}{=} \begin{bmatrix} \mathcal{F}_{\mathcal{S}}(z, Z(\mathcal{W}_t, \cdot)) & \mathcal{F}_{\mathcal{M}}(z, Z(\mathcal{W}_t, \cdot)) \\ \mathcal{F}_{\mathcal{M}}(z, Z(\mathcal{W}_t, \cdot))^\top & 0_{2\times 2} \end{bmatrix}, \tag{178}$$

and

$$Z(\mathcal{W}_0, \cdot) \stackrel{\text{def}}{=} \begin{bmatrix} S(\mathcal{W}_0, \cdot) & M(\mathcal{W}_0, \cdot) \\ M(\mathcal{W}_0, \cdot)^\top & \mu^\top \mathcal{R}(z; \mathcal{K})\mu \end{bmatrix} = \begin{bmatrix} \hat{X}_0^\top \mathcal{R}(z; \mathcal{K})\hat{X}_0 & \hat{X}_0^\top \mathcal{R}(z; \mathcal{K})\mu \\ \mu^\top \mathcal{R}(z; \mathcal{K})\hat{X}_0 & \mu^\top \mathcal{R}(z; \mathcal{K})\mu \end{bmatrix}. \tag{179}$$

The Doob decomposition in the hard label setting can be recovered using similar arguments or simply be setting $X^\star = 0$.

D.4 Doob Decomposition for SGD

We provide explicit details behind the derivation of the Doob decomposition of SGD for an arbitrary quadratic statistic presented in Section C.1. In order to streamline notation, we write

$$f_i(g_{k,i}) \stackrel{\text{def}}{=} f_i(g_{k,i}; \epsilon_{k+1}) \quad \text{and} \quad f_i'(g_{k,i}) \stackrel{\text{def}}{=} f_i'(g_{k,i}; \epsilon_{k+1}). \tag{180}$$

The arguments in this section build upon Appendix A in Collins-Woodfin et al. (2024a) and Section 5.2.2 in Collins-Woodfin & Seroussi (2025).

Recall we defined $a_k = \sqrt{K_{I_k}} v_k + \mu_{I_k}$ for $v_k \sim \mathcal{N}(0, I_d)$, $I_k \in \{1, 2\}$ is randomly sampled, $r_{k,I_{k+1}} = \hat{X}_k^\top a_{k+1,I_{k+1}}$ and

$$\begin{aligned} g_{k,I_{k+1}} &= r_{k,I_{k+1}} + \begin{bmatrix} \delta s_{k,I_{k+1}} \|X_k\| & 0 \end{bmatrix}^\top, \\ s_{k,I_{k+1}} &= \operatorname{argmax}_{|s|\le 1} f_{I_{k+1}}(r_{k,I_{k+1}} + \delta s \|X_k\|; \epsilon_{k+1}). \end{aligned} \tag{181}$$

By Taylor expansion, setting $\Delta_k \stackrel{\text{def}}{=} f'_{I_{k+1}}(g_{k,I_{k+1}})(a_{k+1,I_{k+1}} + \delta s_{k,I_{k+1}} \frac{X_k}{\|X_k\|}) + \lambda X_k$,

$$\varphi(X_{k+1}) = \varphi(X_k - \frac{\gamma_k}{d}\Delta_k) = \varphi(X_k) - \frac{\gamma_k}{d}\langle \nabla\varphi(X_k), \Delta_k\rangle + \frac{\gamma_k^2}{2d^2} \cdot \langle \nabla^2\varphi(X_k), \Delta_k^{\otimes 2}\rangle \tag{182}$$

To derive the Doob decomposition, the idea is to iteratively condition on r_k, I_{k+1} and $\hat{X}_k$. For this, we define the σ-algebras,

$$\mathcal{G}_{k,i} \stackrel{\text{def}}{=} \sigma(\{\hat{X}_j\}_{j=0}^k, \{r_i\}_{i=0}^k, \{I_j\}_{j=1}^k, , \epsilon_{k+1}, I_{k+1} = i) \quad \text{and} \quad \mathcal{F}_k \stackrel{\text{def}}{=} \sigma(\{\hat{X}_j\}_{j=0}^k).$$

It is clear that $s_{k,I_{k+1}}$ is measurable with respect to $\mathcal{G}_{k,i}$ such that $s_{k,I_{k+1}} \mid \mathcal{G}_{k,i}$ corresponds to a deterministic maximization problem.

Gradient term in Taylor expansion. First, we start by decomposing the gradient term in (43),

$$\frac{\gamma_k}{d}\langle \nabla\varphi(X_k), f'_{I_{k+1}}(g_{k,I_{k+1}})(a_{k+1,I_{k+1}} + \delta s_{k,I_{k+1}} \frac{X_k}{\|X_k\|}) + \lambda X_k\rangle. \tag{183}$$

We define a martingale increment associated with it as

$$\begin{aligned} \Delta\mathcal{M}_k^{\text{Grad}}(\varphi) \stackrel{\text{def}}{=}& \frac{\gamma_k}{d} f'_{I_{k+1}}(g_{k,I_{k+1}})\langle \nabla\varphi(\hat{X}_k), a_{k+1,I_{k+1}} + \delta s_{k,I_{k+1}} \frac{X_k}{\|X_k\|}\rangle \\ &- \frac{\gamma_k}{d}\mathbb{E}_{a,I,\epsilon}\big[f'_{I_{k+1}}(g_{k,I_{k+1}})\langle \nabla\varphi(\hat{X}_k), a_{k+1,I_{k+1}} + \delta s_{k,I_{k+1}} \frac{X_k}{\|X_k\|}\rangle \mid \mathcal{F}_k\big]. \end{aligned} \tag{184}$$

From Danskin (1967) theorem, Assumption 6 and interchanging the derivative and expectation, it follows that

$$\mathbb{E}_{a,I,\epsilon}\big[f'_{I_{k+1}}(g_{k,I_{k+1}})\langle \nabla\varphi(\hat{X}_k), a_{k+1,I_{k+1}} + \delta s_{k,I_{k+1}} \frac{X_k}{\|X_k\|}\rangle \mid \mathcal{F}_k\big] = \langle \nabla\varphi(\hat{X}_k), \nabla\mathcal{R}^{\text{adv}}(X_k)\rangle.$$

Hence, the doob decomposition of the gradient term in (183) is given by

$$\begin{aligned} &\frac{\gamma_k}{d}\langle \nabla\varphi(X_k), f'_{I_{k+1}}(g_{k,I_{k+1}})(a_{k+1,I_{k+1}} + \delta s_{k,I_{k+1}} \frac{X_k}{\|X_k\|}) + \lambda X_k\rangle \\ &\quad = \frac{\gamma_k}{d}\langle \nabla\varphi(\hat{X}_k), \nabla\mathcal{R}^{\text{adv}}(X_k) + \lambda X_k\rangle + \Delta\mathcal{M}_k^{\text{Grad}}, \end{aligned} \tag{185}$$

where

$$\begin{aligned} \Delta\mathcal{M}_k^{\text{Grad}}(\varphi) \stackrel{\text{def}}{=}& \frac{\gamma_k}{d} f'_{I_{k+1}}(g_{k,I_{k+1}})\langle \nabla\varphi(\hat{X}_k), a_{k+1,I_{k+1}} + \delta s_{k,I_{k+1}} \frac{X_k}{\|X_k\|}\rangle \\ &- \frac{\gamma_k}{d}\mathbb{E}_{a,I,\epsilon}\big[f'_{I_{k+1}}(g_{k,I_{k+1}})\langle \nabla\varphi(\hat{X}_k), a_{k+1,I_{k+1}} + \delta s_{k,I_{k+1}} \frac{X_k}{\|X_k\|}\rangle \mid \mathcal{F}_k\big]. \end{aligned} \tag{186}$$

Hessian term in the Taylor expansion. Next, we continue by computing the Doob decomposition of the Hessian term in the Taylor expansion (43). Developing the outer product, we have

$$\begin{aligned}
&\frac{\gamma_k^2}{2d^2}\langle\nabla^2\varphi(X_k),\big(f'_{I_{k+1}}(g_{k,I_{k+1}})(a_{k+1,I_{k+1}}+\delta s_{k,I_{k+1}}\tfrac{X_k}{\|X_k\|})+\lambda X_k\big)^{\otimes 2}\rangle\\
&\quad=\frac{\gamma_k^2}{2d^2}(f'_{I_{k+1}}(g_{k,I_{k+1}}))^2\langle\nabla^2\varphi(\hat{X}_k),a_{k+1,I_{k+1}}^{\otimes 2}\rangle\\
&\qquad+\frac{\gamma_k^2}{d^2}\delta s_{k,I_{k+1}}(f'_{I_{k+1}}(g_{k,I_{k+1}}))^2\langle\nabla^2\varphi(\hat{X}_k),a_{k+1,I_{k+1}}\otimes\tfrac{X_k}{\|X_k\|}\rangle\\
&\qquad+\frac{\gamma_k^2}{2d^2}\delta^2 s_{k,I_{k+1}}^2(f'_{I_{k+1}}(g_{k,I_{k+1}}))^2\langle\nabla^2\varphi(\hat{X}_k),\tfrac{X_k^{\otimes 2}}{\|X_k\|^2}\rangle+\frac{\gamma_k^2}{2d^2}\langle\nabla^2\varphi(X_k),(\lambda X_k)^{\otimes 2}\rangle\\
&\qquad+\frac{\gamma_k^2}{d^2}f'_{I_{k+1}}(g_{k,I_{k+1}})\langle\nabla^2\varphi(X_k),(a_{k+1,I_{k+1}}+\delta s_{k,I_{k+1}}\tfrac{X_k}{\|X_k\|})\otimes\lambda X_k\rangle.
\end{aligned}\tag{187}$$

The martingale increment associated with the Hessian is given by

$$\Delta\mathcal{M}_k^{\text{Hess}}(\varphi)\overset{\text{def}}{=}\frac{\gamma_k^2}{2d^2}\Big(\langle\nabla^2\varphi(X_k),\Delta_k^{\otimes 2}\rangle-\mathbb{E}\left[\langle\nabla^2\varphi(X_k),\Delta_k^{\otimes 2}\rangle\,|\,\mathcal{F}_k\right]\Big).\tag{188}$$

In order to explicitly characterize the martingale increment, we will compute the conditional expectation of (187) given $\mathcal{F}_k$. We first condition on $\mathcal{G}_k$ and use the following Lemma D.4 to simplify the computation of the conditional expectation of the first term in (187).

Lemma D.4 (Conditioning, Lemma A2 (Collins-Woodfin et al., 2024a)). *Suppose $v\in\mathbb{R}^d$ is distributed $N(0,I_d)$ and $U\in\mathbb{R}^{d\times 2}$ has orthonormal columns. Then*

$$v\,|\,\langle U,v\rangle_{\mathbb{R}^d}\sim v-UU^\top v+UU^\top v,\tag{189}$$

where $v-UU^\top v\sim\mathcal{N}(0,I_d-UU^\top)$ and $UU^\top v\sim\mathcal{N}(0,UU^\top)$ with $v-UU^\top v$ independent of $UU^\top v$.

A simple computation yields

$$\begin{aligned}
&\mathbb{E}\left[(f_i'(g_{k,i}))^2\langle\nabla^2\varphi(\hat{X}_k),a_{k+1,i}^{\otimes 2}\rangle\,|\,\mathcal{G}_{k,i}\right]\\
&\quad=\mathbb{E}\left[(f_i'(g_{k,i}))^2\langle\nabla^2\varphi(X_k),(a_{k+1,i}-\mathbb{E}\left[a_{k+1,i}\,|\,\mathcal{G}_{k,i}\right])^{\otimes 2}\rangle\,|\,\mathcal{G}_{k,i}\right]\\
&\qquad+\mathbb{E}_{\epsilon_{k+1}}[(f_i'(g_{k,i}))^2]\cdot\langle\nabla^2\varphi(X_k),\mathbb{E}\left[a_{k+1,i}\,|\,\mathcal{G}_{k,i}\right]^{\otimes 2}\rangle.
\end{aligned}\tag{190}$$

We will compute these conditional expectations using Lemma D.4. By Assumption 1, we write $a_{k+1,i}=\mu_i+\sqrt{K_i}v_{k+1}$ for $v_{k+1}\sim N(0,I_d)$. It then follows

$$v_{k+1}\mid r_{k,i},\hat{X}_k\overset{d}{=}v_{k+1}\mid\hat{X}_k^\top(\mu_i+\sqrt{K_i}v_{k+1}),\hat{X}_k\overset{d}{=}v_{k+1}\mid\hat{X}_k^\top\sqrt{K_i}v_{k+1}.$$

In order to apply Lemma D.4, we apply the QR decomposition to $\sqrt{K}\hat{X}_k=Q_{k,i}R_{k,i}$ such that $R_{k,i}\in\mathbb{R}^{\ell_i\times\ell_i}$ is upper-triangular and invertible, $Q_{k,i}\in\mathbb{R}^{d\times\ell_i}$ has orthonormal columns and where $\ell_i=\operatorname{rank}(\sqrt{K_i}W_k)\le 2$.

Set $\Pi_{k,i}=Q_{k,i}Q_{k,i}^\top$. By Lemma D.4 and since $R_{k,i}$ is invertible, in distribution we have

$$\begin{aligned}
v_{k+1}\,|\,\hat{X}_k^\top\sqrt{K_i}v_{k+1}&\overset{\text{d}}{=}v_{k+1}\,|\,R_{k,i}^\top Q_{k,i}^\top v_{k+1}\overset{\text{d}}{=}v_{k+1}\,|\,Q_{k,i}^\top v_{k+1}\\
&\overset{\text{d}}{=}\big(v_{k+1}-\Pi_{k,i}v_{k+1}\big)+\Pi_{k,i}v_{k+1}.
\end{aligned}$$

Hence, it follows that

$$a_{k+1,i}\mid r_{k,i},\hat{X}_k\overset{d}{=}\sqrt{K_i}(v_{k+1}-\Pi_{k,i}v_{k+1})+\sqrt{K_i}\Pi_{k,i}v_{k+1}+\mu_i\tag{191}$$

By Lemma D.4, we have $(I_d-\Pi_{k,i})v_{k+1}\sim N(0,I_d-\Pi_{k,i})$ and $\Pi_{k,i}v_{k+1}\sim N(0,\Pi_{k,i})$ such that $(I_d-\Pi_{k,i})v_{k+1}$ independent of $\Pi_{k,i}v_{k+1}$ from which we obtain

$$\mathbb{E}\left[a_{k+1,i}\,|\,\mathcal{G}_{k,i}\right]=\sqrt{K_i}\Pi_{k,i}v_{k+1}+\mu_i,\quad\text{where } v_{k+1}\sim N(0,I_d).\tag{192}$$

Furthermore, it follows from a simple computation that

$$\mathbb{E}\left[(a_{k+1,i}-\mathbb{E}\left[a_{k+1,i}\,|\,\mathcal{G}_{k,i}\right])^{\otimes 2}|\,\mathcal{G}_k\right]=\sqrt{K_i}(I_d-\Pi_{k,i})\sqrt{K_i}.\tag{193}$$

Putting this all together, we obtain the decomposition of the leading order term of (190)

$$\mathbb{E}\left[(f_i'(g_{k,i}))^2\langle\nabla^2\varphi(\hat{X}_k), a_{k+1,i}^{\otimes 2}\rangle \,|\, \mathcal{G}_{k,i}\right] = \mathbb{E}_{\,\epsilon_{k+1}}[(f_i'(g_{k,i}))^2]\cdot\Big(\langle\nabla^2\varphi(X_k), K_i\rangle - \langle\nabla^2\varphi(X_k), \sqrt{K_i}\Pi_{k,i}\sqrt{K_i}\rangle + \langle\nabla^2\varphi(\hat{X}_k), \big(\sqrt{K_i}\Pi_{k,i}v_{k+1}+\mu_i\big)^{\otimes 2}\rangle\Big). \tag{194}$$

With this in mind, we obtain

$$\frac{\gamma_k^2}{2d^2}\mathbb{E}\left[(f_i'(g_{k,i}))^2\langle\nabla^2\varphi(X_k), a_{k+1}^{\otimes 2}\rangle \,|\, \mathcal{F}_k\right] = \frac{\gamma_k^2}{2d^2}\sum_{i=1}^{2} p_i\mathbb{E}\left[(f_i'(g_{k,i}))^2 \,|\, \mathcal{F}_k\right]\langle\nabla^2\varphi(X_k), K_i+\mu_i\mu_i^\top\rangle + \mathbb{E}\left[\mathcal{E}_{k,1}^{\text{Hess}} \,|\, \mathcal{F}_k\right]. \tag{195}$$

where

$$\mathcal{E}_{k,1}^{\text{Hess}}(\varphi) \stackrel{\text{def}}{=} \frac{\gamma_k^2}{2d^2}\sum_{i=1}^{2} p_i\mathbb{E}_{\,\epsilon_{k+1}}[(f_i'(g_{k,i}))^2]\cdot\Big(-\langle\nabla^2\varphi(X_k), \sqrt{K_i}\Pi_{k,i}\sqrt{K_i}\rangle + \langle\nabla^2\varphi(\hat{X}_k), \big(\sqrt{K_i}\Pi_{k,i}v_{k+1}\big)^{\otimes 2}\rangle + 2\langle\nabla^2\varphi(\hat{X}_k), \sqrt{K_i}\Pi_{k,i}v_{k+1}\mu_i^\top\rangle\Big).$$

We will show in Section D.5.3 that $\mathcal{E}_{k,1}^{\text{Hess}}(\varphi)$ is of lower order and will disappear as $d\to\infty$. For the second term in (187), by (192) and a similar argument as above, we obtain

$$\mathbb{E}\left[\delta s_{k,I_{k+1}}(f'_{I_{k+1}}(g_{k,I_{k+1}}))^2\langle\nabla^2\varphi(\hat{X}_k), a_{k+1,I_{k+1}}\otimes\tfrac{X_k}{\|X_k\|}\rangle \,|\, \mathcal{F}_k\right] = \sum_{i=1}^{2} p_i\mathbb{E}\left[\delta s_{k,i}(f_i'(g_{k,i}))^2\langle\nabla^2\varphi(\hat{X}_k), (\sqrt{K_i}\Pi_{k,i}v_{k+1}+\mu_i)\otimes\tfrac{X_k}{\|X_k\|}\rangle \,|\, \mathcal{F}_k\right], \tag{196}$$

with $v_{k+1}\sim N(0, I_d)$. Due to the factor $\frac{1}{d^2}$, this lower order term will disappear as $d\to\infty$ (see Section D.5.3). Hence, we define

$$\mathcal{E}_{k,2}^{\text{Hess}}(\varphi) \stackrel{\text{def}}{=} \frac{\gamma_k^2}{d^2}\sum_{i=1}^{2} p_i\delta s_{k,i}(f_i'(g_{k,i}))^2\langle\nabla^2\varphi(\hat{X}_k), (\sqrt{K_i}\Pi_{k,i}v_{k+1}+\mu_i)\otimes\tfrac{X_k}{\|X_k\|}\rangle \tag{197}$$

For the third term and fourth terms in (187), we leave them as is and define

$$\mathcal{E}_{k,3}^{\text{Hess}}(\varphi) \stackrel{\text{def}}{=} \frac{\gamma_k^2}{2d^2}\sum_{i=1}^{2} p_i\delta^2 s_{k,i}^2(f_i'(g_{k,i}))^2\langle\nabla^2\varphi(\hat{X}_k), \tfrac{X_k^{\otimes 2}}{\|X_k\|^2}\rangle,$$

$$\mathcal{E}_{k,4}^{\text{Hess}}(\varphi) \stackrel{\text{def}}{=} \frac{\gamma_k^2}{2d^2}\langle\nabla^2\varphi(X_k), (\lambda X_k)^{\otimes 2}\rangle.$$

For the last term in (187), by (192) and a similar argument as above, we obtain

$$\mathbb{E}\left[f'_{I_{k+1}}(g_{k,I_{k+1}})\langle\nabla^2\varphi(X_k), (a_{k+1,I_{k+1}}+\delta s_{k,I_{k+1}}\tfrac{X_k}{\|X_k\|})\otimes\lambda X_k\rangle \,|\, \mathcal{F}_k\right] = \sum_{i=1}^{2} p_i\mathbb{E}\left[f_i'(g_{k,i})\langle\nabla^2\varphi(X_k), (\sqrt{K_i}\Pi_{k,i}v_{k+1}+\mu_i+\delta s_{k,i}\tfrac{X_k}{\|X_k\|})\otimes\lambda X_k\rangle \,|\, \mathcal{F}_k\right]. \tag{198}$$

Finally, due to the $\frac{1}{d^2}$, this term will be of lower order (see Section D.5.3), and thus, we define

$$\mathcal{E}_{k,5}^{\text{Hess}}(\varphi) \stackrel{\text{def}}{=} \frac{\gamma_k^2}{d^2}\sum_{i=1}^{2} p_i f_i'(g_{k,i})\langle\nabla^2\varphi(X_k), (\sqrt{K_i}\Pi_{k,i}v_{k+1}+\mu_i+\delta s_{k,i}\tfrac{X_k}{\|X_k\|})\otimes\lambda X_k\rangle. \tag{199}$$

We have thus identified the martingale increments of a *single* update of SGD, that is

$$\varphi(\hat{X}_{k+1}) = \varphi(X_k) - \frac{\gamma_k}{d}\langle\nabla\varphi(X_k), \nabla\mathcal{R}_\lambda^{\text{adv}}(X_k)\rangle + \frac{\gamma_k^2}{2d^2}\sum_{i=1}^{2} p_i\langle\nabla^2\varphi(X_k), K_i+\mu_i\mu_i^\top\rangle\mathbb{E}\left[(f_i'(g_{k,i}))^2 \,|\, \mathcal{F}_k\right] + \Delta\mathcal{M}_k^{\text{Grad}}(\varphi) + \Delta\mathcal{M}_k^{\text{Hess}}(\varphi) + \mathbb{E}\left[\mathcal{E}_k^{\text{Hess}}(\varphi) \,|\, \mathcal{F}_k\right] \tag{200}$$

where the error terms look like

$$\begin{aligned}
\Delta\mathcal{M}_k^{\text{Grad}}(\varphi) &= \frac{\gamma_k}{d} f'_{I_{k+1}}(g_{k,I_{k+1}})\langle\nabla\varphi(\hat{X}_k), a_{k+1,I_{k+1}} + \delta s_{k,I_{k+1}}\tfrac{X_k}{\|X_k\|}\rangle \\
&\quad - \frac{\gamma_k}{d}\mathbb{E}_{a,I,\epsilon}\big[f'_{I_{k+1}}(g_{k,I_{k+1}})\langle\nabla\varphi(\hat{X}_k), a_{k+1,I_{k+1}} + \delta s_{k,I_{k+1}}\tfrac{X_k}{\|X_k\|}\rangle \,|\, \mathcal{F}_k\big], \\
\Delta\mathcal{M}_k^{\text{Hess}}(\varphi) &= \frac{\gamma_k^2}{2d^2}\Big(\langle\nabla^2\varphi(X_k), \Delta_k^{\otimes 2}\rangle - \mathbb{E}\left[\langle\nabla^2\varphi(X_k), \Delta_k^{\otimes 2}\rangle \,|\, \mathcal{F}_k\right]\Big), \\
\text{and}\quad \mathcal{E}_k^{\text{Hess}}(\varphi) &= \mathcal{E}_{k,1}^{\text{Hess}}(\varphi) + \mathcal{E}_{k,2}^{\text{Hess}}(\varphi) + \mathcal{E}_{k,3}^{\text{Hess}}(\varphi) + \mathcal{E}_{k,4}^{\text{Hess}}(\varphi) + \mathcal{E}_{k,5}^{\text{Hess}}(\varphi) \\
&= \frac{\gamma_k^2}{d^2}\sum_{i=1}^{2} p_i \mathbb{E}_{\epsilon_{k+1}}[(f_i'(g_{k,i}))^2]\cdot\Big(-\langle\nabla^2\varphi(X_k), \sqrt{K_i}\Pi_{k,i}\sqrt{K_i}\rangle \\
&\quad + \langle\nabla^2\varphi(\hat{X}_k), \big(\sqrt{K_i}\Pi_{k,i}v_{k+1}\big)^{\otimes 2}\rangle + 2\langle\nabla^2\varphi(\hat{X}_k), \sqrt{K_i}\Pi_{k,i}v_{k+1}\mu_i^\top\rangle\Big) \\
&\quad + \frac{\gamma_k^2}{d^2}\sum_{i=1}^{2} p_i\delta s_{k,i}(f_i'(g_{k,i}))^2\langle\nabla^2\varphi(\hat{X}_k), (\sqrt{K_i}\Pi_{k,i}v_{k+1} + \mu_i)\otimes\tfrac{X_k}{\|X_k\|}\rangle \\
&\quad + \frac{\gamma_k^2}{2d^2}\sum_{i=1}^{2} p_i\delta^2 s_{k,i}^2(f_i'(g_{k,i}))^2\langle\nabla^2\varphi(\hat{X}_k), \tfrac{X_k^{\otimes 2}}{\|X_k\|^2}\rangle \\
&\quad + \frac{\gamma_k^2}{2d^2}\langle\nabla^2\varphi(X_k), (\lambda X_k)^{\otimes 2}\rangle \\
&\quad + \frac{\gamma_k^2}{d^2}\sum_{i=1}^{2} p_i f_i'(g_{k,i})\langle\nabla^2\varphi(X_k), (\sqrt{K_i}\Pi_{k,i}v_{k+1} + \mu_i + \delta s_{k,i}\tfrac{X_k}{\|X_k\|})\otimes\lambda X_k\rangle.
\end{aligned} \tag{201}$$

The Doob decomposition in the hard label setting can be recovered using similar arguments or simply be setting $X^\star = 0$.

D.5 Error Bounds

The results in this section build upon Appendix A in Collins-Woodfin et al. (2024a) and Section 5.5 in Collins-Woodfin & Seroussi (2025). Recall, we defined $W = \hat{X} \oplus \mu$. We are studying the statistics

$$\begin{aligned}
&S(W,\cdot) = \hat{X}^\top\mathcal{R}(\cdot;\mathcal{K})\hat{X}, \quad M(W,\cdot) = \hat{X}^\top\mathcal{R}(\cdot;\mathcal{K})\mu, \\
\text{and}\quad &Z(W,\cdot) \overset{\text{def}}{=} W^\top\mathcal{R}(\cdot;\mathcal{K})W = \begin{bmatrix} S(W,\cdot) & M(W,\cdot) \\ M(W,\cdot)^\top & \mu^\top\mathcal{R}(\cdot;\mathcal{K})\mu \end{bmatrix}.
\end{aligned} \tag{202}$$

Recall we defined the fixed contour

$$\Gamma = \{w\in\mathbb{C} : |w| = \max\{1, 2\|\mathcal{K}\|_{\text{op}}\}\} \quad \text{and} \quad \|\mathcal{K}\|_{\text{op}} \overset{\text{def}}{=} \max\{\|K_1\|_{\text{op}}, \|K_1\|_{\text{op}}\}.$$

Note that by construction the distance between the spectra of K_1 and K_2 and the contour is always greater or equal than $\frac{1}{2}$. Recall for $z\in\Gamma^2$, we write $z = (z_1, z_2)$ and when integrating over all z_1 simultaneously, we write for any function $f : \mathbb{C}^2 \to \mathbb{C}$

$$\oint f(z)\mathrm{D}z \overset{\text{def}}{=} \frac{-1}{4\pi^2}\oint_{\Gamma^2} f(z)\,\mathrm{d}z_1\,\mathrm{d}z_2.$$

Recall we also defined the norm $\|\cdot\|_\Gamma$ for a continuous function $H : \mathbb{C}^2 \to \mathbb{R}^{4\times 4}$:

$$\|H\|_\Gamma = \max_{z\in\Gamma^2}\|H(z)\|.$$

In the next subsections, we will control the error terms which arise in the Doob decompositions of adversarial homogenized SGD and SGD respectively. This results will then be used in the proofs of SGD and AdvHSGD being approximate solutions to the integro-differential equation (17). Before diving into the proofs, we introduce Lemma 3 from Collins-Woodfin & Seroussi (2025) on the norm of the derivatives of $Z(W,\cdot)$.

Lemma D.5. *There exist constants $c, C > 0$ such that*

$$c\|W\|^2 \le \|Z(W,\cdot)\|_\Gamma \le C\|W\|^2, \quad \|\nabla_X Z(W,\cdot)\|_\Gamma \le C\|W\|,$$
$$\sup_{z\in\Gamma^2} \|\nabla_X^2 Z(W,z)\|_{\mathrm{op}} \le C \quad \textit{and} \quad \|\nabla_X^2 Z(W,\cdot)\|_\Gamma \le C\sqrt{d}.$$

Lemma D.5 differs from Lemma 3 from Collins-Woodfin & Seroussi (2025) and is analogous to Lemma A3 in Collins-Woodfin et al. (2024a) as we are studying single-index models (i.e. $\ell = 1, \ell^\star = 1, \ell^+ = 4$ in Collins-Woodfin & Seroussi (2025)).

To control the martingale terms, we will work with the stopped processes $W_{td}^\vartheta \stackrel{\mathrm{def}}{=} W_{t\wedge\vartheta d}$ and $\mathcal{W}_t^\vartheta \stackrel{\mathrm{def}}{=} \mathcal{W}_{t\wedge\vartheta}$ where we recall the stopping time ϑ for fixed $M > 0$ which we defined in (91):

$$\begin{aligned} \vartheta_M &\stackrel{\mathrm{def}}{=} \inf\{t \ge 0 : \|W_{td}\|^2 > M \quad \text{or} \quad \widehat{\mathcal{B}}_1(t) \notin \mathcal{U} \quad \text{or} \quad \widehat{\mathcal{B}}_2(t) \notin \mathcal{U}\} \\ \text{or} \quad \vartheta_M &\stackrel{\mathrm{def}}{=} \inf\{t \ge 0 : \|\mathcal{W}_t\|^2 > M \quad \text{or} \quad \widehat{\mathcal{B}}_1(t) \notin \mathcal{U} \quad \text{or} \quad \widehat{\mathcal{B}}_2(t) \notin \mathcal{U}\}, \end{aligned} \tag{203}$$

Here we overload the notation so that ϑ_M can be applied to either the iterates of SGD or AdvHSGD. When clear from context, we drop the superscript ϑ for notational simplicity. We will first control the errors with the stopped processes and later derive a sufficient condition under which we can remove the stopping time.

Throughout the proofs in the following subsections, the constant C may depend on the following various bounded quantities: $\bar{\gamma}$, T, λ, δ, α, $\|K\|_{\mathrm{op}}$, and M. However, C does not depend on d. The value of C may also change from line to line in these proofs and may differ from the one presented in Lemma D.5.

D.5.1 Adversarial Homogenized SGD Martingale Error

Given that the elements of $Z(\mathcal{W}_t, z)$ are at most quadratics in $\mathcal{X}_t$, it suffices to control the martingale increment that arises in the Doob decomposition of AdvHSGD for an arbitrary quadratic statistic $\varphi : \mathbb{R}^d \to \mathbb{R}$:

$$\begin{aligned} \mathcal{M}_t^{\mathrm{AdvHSGD}}(\varphi) &\stackrel{\mathrm{def}}{=} \int_0^t \mathrm{d}\mathcal{M}_s^{\mathrm{AdvHSGD}}(\varphi) \\ &= \frac{1}{\sqrt{d}} \int_0^t \gamma(s) \left\langle \sqrt{\sum_{i=1}^2 p_i \mathbb{E}_{v,\epsilon}[f_i'(\rho_{s,i},\epsilon)^2](K_i + \mu_i\mu_i^\top)}, \nabla\varphi(\mathcal{X}_s) \otimes \mathrm{d}B_s \right\rangle. \end{aligned} \tag{204}$$

To show this martingale increment vanishes as $d \to \infty$, we will control its fluctuations by showing elementwise, its quadratic variation as defined below vanishes. Given a partition $\pi = \{t_0, .., t_n\}$ of $[0, t]$ such that $0 = t_0 < t_1 < \cdots < t_n = t$ then the π-quadratic variation of $\mathcal{M}_t^{\mathrm{AdvHSGD}}(Z_{j\ell}(\cdot, z))$ is defined as

$$Q_\pi(\mathcal{M}_t^{\mathrm{AdvHSGD}}(Z_{j\ell}(\cdot,z))) = \sum_{i=0}^{n-1} (\mathcal{M}_{t_{i+1}}^{\mathrm{AdvHSGD}}(Z_{j\ell}(\cdot,z)) - \mathcal{M}_{t_i}^{\mathrm{AdvHSGD}}(Z_{j\ell}(\cdot,z)))^2.$$

If there exists process denoted $\langle \mathcal{M}^{\mathrm{AdvHSGD}}(Z_{j\ell}(\cdot,z))\rangle_t$ such that for all sequence of partitions π_ℓ with $\max_{0\le i\le n_\ell - 1} |t_{i+1}^{(\ell)} - t_i^{(\ell)}| \underset{k\to\infty}{\to} 0$, we have

$$Q_{\pi_\ell}(\mathcal{M}_t^{\mathrm{AdvHSGD}}(Z_{j\ell}(\cdot,z))) \underset{\ell\to\infty}{\to} \langle \mathcal{M}_t^{\mathrm{AdvHSGD}}(Z_{j\ell}(\cdot,z))\rangle_t,$$

then $\langle \mathcal{M}^{\mathrm{AdvHSGD}}(Z_{j\ell}(\cdot,z))\rangle$ is called the *quadratic variation* of $\langle \mathcal{M}^{\mathrm{AdvHSGD}}(Z_{j\ell}(\cdot,z))\rangle$ (see (Protter, 2005, Chapter 1) for details). We build upon Proposition A.1 of Collins-Woodfin et al. (2024a). We provide the proof for the soft label setting as the proof for hard labels follows the same argument with $X^\star = 0$.

Proposition D.1 (AdvHSGD small.)**.** *For $i = 1, 2$, suppose $f_i : \mathbb{R}^3 \to \mathbb{R}$ is an α-pseudo-Lipschitz function with constant $L(f)$ (see Assumption 3). Let the statistic $Z : \mathbb{R}^d \to \mathbb{R}^{4\times 4}$ be defined as in*

(202). *For any $T > 0$, $\zeta > 0$ and fix $z \in \Gamma^2$, there is some constant C such that, with overwhelming probability,*

$$\sup_{0\le t\le T} \|\mathcal{M}^{\mathrm{AdvHSGD}}_{t\wedge\vartheta}(S(\cdot,z))\| \le CL(f)\ d^{\zeta/2-1/2}. \tag{205}$$

Proof. Let $\varphi(\mathcal{X}_t) \overset{\mathrm{def}}{=} Z_{j\ell}(\mathcal{W}_t, z)$ be the $j\ell$-coordinate of Z for a fixed $z \in \Gamma^2$. Recall the martingale increment elementwise is given by,

$$\mathrm{d}\mathcal{M}^{\mathrm{AdvHSGD}}_t(\varphi) = \frac{\gamma(t)}{\sqrt{d}} \left\langle \sqrt{\sum_{i=1}^{2} p_i \mathbb{E}_{\,v,\epsilon}[f_i'(\rho_{t,i},\epsilon)^2](K_i + \mu_i\mu_i^\top)}, \nabla_X \varphi(\mathcal{X}_t) \otimes \mathrm{d}B_t \right\rangle \tag{206}$$

The quadratic variation of $\mathcal{M}^{\mathrm{AdvHSGD}}_t$ is

$$\begin{aligned}
&\langle \mathcal{M}^{\mathrm{AdvHSGD}}(\varphi)\rangle_t \\
&= \frac{1}{d}\int_0^t \gamma^2(s) \left\| \left\langle \sqrt{\sum_{i=1}^{2} p_i \mathbb{E}_{\,v,\epsilon}[f_i'(\rho_{s,i},\epsilon)^2](K_i + \mu_i\mu_i^\top)}, \nabla_X \varphi(\mathcal{X}_s) \right\rangle_{\mathbb{R}^d} \right\|^2 \mathrm{d}s.
\end{aligned} \tag{207}$$

Our goal is to show $\sup_{0\le t\le T} \langle \mathcal{M}^{\mathrm{AdvHSGD}}(\varphi)\rangle_{t\wedge\vartheta}$ is small. Since $\gamma : \mathbb{R}^+ \to \mathbb{R}^+$ is a bounded function by Assumption 8, we must then show the $\|\cdot\|^2$ is also small. We have

$$\begin{aligned}
&\left\| \left\langle \sqrt{\sum_{i=1}^{2} p_i \mathbb{E}_{\,v,\epsilon}[f_i'(\rho^\vartheta_{s,i},\epsilon)^2](K_i + \mu_i\mu_i^\top)}, \nabla_X \varphi(\mathcal{X}^\vartheta_s) \right\rangle_{\mathbb{R}^d} \right\|^2 \\
&\le \left\| \sqrt{\sum_{i=1}^{2} p_i \mathbb{E}_{\,v,\epsilon}[f_i'(\rho^\vartheta_{s,i},\epsilon)^2](K_i + \mu_i\mu_i^\top)} \right\|_{\mathrm{op}}^2 \|\nabla_X \varphi(\mathcal{X}^\vartheta_s)\|^2 \\
&\le \sum_{i=1}^{2} p_i \mathbb{E}_{\,v,\epsilon}[f_i'(\rho^\vartheta_{s,i},\epsilon)^2] \left\|(K_i + \mu_i\mu_i^\top)\right\|_{\mathrm{op}} \|\nabla_X \varphi(\mathcal{X}^\vartheta_s)\|^2 \\
&\le \sum_{i=1}^{2} p_i \mathbb{E}_{\,v,\epsilon}[f_i'(\rho^\vartheta_{s,i},\epsilon)^2](\|K_i\|_{\mathrm{op}} + \|\mu_i\|^2)\|\nabla_X \varphi(\mathcal{X}^\vartheta_s)\|^2.
\end{aligned} \tag{208}$$

By Lemma D.5, we have a bound on

$$\|\nabla_X \varphi(\mathcal{X}^\vartheta_s)\| = \|\nabla_X Z_{j\ell}(\mathcal{W}^\vartheta_s, z)\| \le \|\nabla_X Z_{j\ell}(\mathcal{W}^\vartheta_s, \cdot)\|_\Gamma \le C\|\mathcal{W}^\vartheta_s\|.$$

From Assumption 1 for $i = 1, 2$, we have $\|K_i\|_{\mathrm{op}} \le C$, $p_i \le 1$ and $\|\mu_i\|^2 \le C$. From Lemma D.3, we have a bound on $\mathbb{E}_{\,v,\epsilon}[f_i'(\rho^\vartheta_{s,i},\epsilon)^2]$, and since $\|\mathcal{X}^\vartheta_t\| \le \|\widehat{\mathcal{X}}^\vartheta_t\| \le \|\widehat{\mathcal{W}}^\vartheta_t\|$ then

$$\begin{aligned}
&\left\| \left\langle \sqrt{\sum_{i=1}^{2} p_i \mathbb{E}_{\,v,\epsilon}[f_i'(\rho^\vartheta_{s,i},\epsilon)^2](K_i + \mu_i\mu_i^\top)}, \nabla_X \varphi(\mathcal{X}^\vartheta_s) \right\rangle_{\mathbb{R}^d} \right\|^2 \le C\|\mathcal{W}^\vartheta_s\|^2 \sum_{i=1}^{2} \mathbb{E}_{\,v,\epsilon}[f_i'(\rho^\vartheta_{s,i},\epsilon)^2] \\
&\le C(L(f))^2\|\mathcal{W}^\vartheta_s\|^2 \sum_{i=1}^{2} \big(1 + (\|K_i\|_{\mathrm{op}}^{1/2} + \|\mu_i\|)\|\widehat{\mathcal{W}}^\nu_t\| + \delta\|\mathcal{X}^\nu_t\| + \eta\big)^{\max\{1,2\alpha\}} \\
&\le C(L(f))^2 M\big(1 + (1+\delta)\sqrt{M} + \eta\big)^{\max\{1,2\alpha\}}
\end{aligned} \tag{209}$$

Thus, from (207) and (209), it follows

$$\sup_{0\le t\le T} \langle \mathcal{M}^{\mathrm{AdvHSGD}}(\varphi)\rangle_{t\wedge\vartheta} \le C(L(f))^2 d^{-1}. \tag{210}$$

From Doob's maximal inequality for non-negative super-martingales and the exponential super-martingale

$$Z^a_t = e^{a\mathcal{M}^{\mathrm{AdvHSGD}}_{t\wedge\vartheta}(\varphi) - \frac{a^2}{2}\langle \mathcal{M}^{\mathrm{AdvHSGD}}(\varphi)\rangle_{t\wedge\vartheta}}, \quad \text{for} \quad a, b > 0,$$

if $\sup_{0\le t\le T}\langle \mathcal{M}^{\mathrm{AdvHSGD}}(\varphi)\rangle_{t\wedge\vartheta}\le b$ a.s, then it is easy to show

$$\Pr(\sup_{0\le t\le T}|\mathcal{M}^{\mathrm{AdvHSGD}}_{t\wedge\vartheta}(\varphi)|>p)\le 2\exp(-p^2/2b).$$

Absorbing $\bar\gamma^2$ into C and setting $p=\sqrt{2C}L(f)d^{\zeta/2-1/2}$ for any $\zeta>0$, then

$$\Pr(\sup_{0\le t\le T}|\mathcal{M}^{\mathrm{AdvHSGD}}_{t\wedge\vartheta}(\varphi)|>\sqrt{C}L(f)d^{\zeta/2-1/2})\le C\exp(-d^\zeta).$$

The result immediately follows after noting that the number of $j\ell$ coordinates is 16 which is independent of d. Indeed, from a union bound we have

$$\begin{aligned}\Pr(\sup_{1\le i,j\le 4}\sup_{0\le t\le T}|\mathcal{M}^{\mathrm{AdvHSGD}}_{t\wedge\vartheta}(\varphi)|<p)&=1-\Pr(\sup_{1\le i,j\le 4}\sup_{0\le t\le T}|\mathcal{M}^{\mathrm{AdvHSGD}}_{t\wedge\vartheta}(\varphi)|>p)\\&=1-\Pr(\cup_{1\le i,j\le 4}\sup_{0\le t\le T}|\mathcal{M}^{\mathrm{AdvHSGD}}_{t\wedge\vartheta}(\varphi)|>p)\\&\ge 1-C\exp(-d^\zeta).\end{aligned}\tag{211}$$

Noting that

$$\sup_{0\le t\le T}\|\mathcal{M}^{\mathrm{AdvHSGD}}_{t\wedge\vartheta}(S(\cdot,z))\|\le 4\sup_{1\le i,j\le 2}\sup_{0\le t\le T}|\mathcal{M}^{\mathrm{AdvHSGD}}_{t\wedge\vartheta}(\varphi)|,\tag{212}$$

concludes the proof. □

D.5.2 Bounds on the Martingales $\mathcal{M}_k^{\mathrm{Grad}}$ and $\mathcal{M}_k^{\mathrm{Hess}}$

Given that the elements of $Z(W_k,z)$ are at most quadratics in X_k, it suffices to control the martingale increment that arise in the Doob decomposition of SGD for an arbitrary quadratic statistic $\varphi:\mathbb{R}^d\to\mathbb{R}$:

$$\begin{aligned}\Delta\mathcal{M}^{\mathrm{Grad}}_k(\varphi)&=\frac{\gamma_k}{d}f'_{I_{k+1}}(g_{k,I_{k+1}})\langle\nabla\varphi(\hat X_k),a_{k+1,I_{k+1}}+\delta s_{k,I_{k+1}}\tfrac{X_k}{\|X_k\|}\rangle\\&\quad-\frac{\gamma_k}{d}\sum_{i=1}^2 p_i\mathbb{E}_{a,\epsilon}\big[f'_i(g_{k,i})\langle\nabla\varphi(\hat X_k),a_{k+1,i}+\delta s_{k,i}\tfrac{X_k}{\|X_k\|}\rangle\,|\,\mathcal{F}_k\big],\\\Delta\mathcal{M}^{\mathrm{Hess}}_k(\varphi)&=\frac{\gamma_k^2}{2d^2}\Big(\langle\nabla^2\varphi(X_k),\Delta_k^{\otimes 2}\rangle-\mathbb{E}\left[\langle\nabla^2\varphi(X_k),\Delta_k^{\otimes 2}\rangle\,|\,\mathcal{F}_k\right]\Big),\end{aligned}$$

where we recall $a_{k,I_k}=\sqrt{K_{I_k}}v_k+\mu_{I_k}$ for $v_k\sim\mathcal{N}(0,I_d)$, $I_k\in\{1,2\}$, $r_{k,I_{k+1}}=\hat X_k^\top a_{k+1,I_{k+1}}$ and

$$\begin{aligned}g_{k,I_{k+1}}&=r_{k,I_{k+1}}+[\delta s_{k,I_{k+1}}\|X_k\|\quad 0]^\top,\\s_{k,I_{k+1}}&=\mathrm{argmax}_{|s|\le 1}f_{I_{k+1}}(X_k^\top a_{k+1,I_{k+1}}+\delta s\|X_k\|;\epsilon_{k+1}).\end{aligned}\tag{213}$$

We also define

$$\mathcal{M}^{\mathrm{Grad}}_k(\varphi)=\sum_{j=0}^{k-1}\Delta\mathcal{M}^{\mathrm{Grad}}_j(\varphi)\quad\text{and}\quad\mathcal{M}^{\mathrm{Hess}}_k(\varphi)=\sum_{j=0}^{k-1}\Delta\mathcal{M}^{\mathrm{Hess}}_j(\varphi)$$

Once again, we provide the proofs for the soft label setting as the proofs for hard labels follow the same arguments with $X^\star=0$. In the following proof Proposition D.2, we build upon Proposition A.2 of Collins-Woodfin et al. (2024a) and Lemma 9 in Collins-Woodfin & Seroussi (2025).

Proposition D.2 (Gradient martingale)**.** *For $i=1,2$, suppose $f_i:\mathbb{R}^3\to\mathbb{R}$ is an α-pseudo-Lipschitz function with constant $L(f)$ (see Assumption 3). Let the statistic $Z:\mathbb{R}^d\to\mathbb{R}^{4\times 4}$ be defined as in* (202)*. Then, for any $\zeta>0$ and $T>0$, and with overwhelming probability,*

$$\sup_{0\le t\le T}\|\mathcal{M}^{\mathrm{Grad}}_{\lfloor(t\wedge\vartheta)d\rfloor}(Z(\cdot,z))\|<d^{-\frac12+\zeta}.\tag{214}$$

Proof. Let $\varphi\overset{\mathrm{def}}{=}Z_{j\ell}(W,z)$ be the $j\ell$-coordinate of Z for a fixed $z\in\Gamma^2$. Similarly as in the proof Proposition D.1, we will be working on the stopped process $\mathcal{M}^{\mathrm{Grad}}_{(t\wedge\vartheta)d}(\varphi)$. To simplify notation, we

omit the dependence on the stopping time ϑ and on φ such that $\mathcal{M}_k^{\mathrm{Grad}} \overset{\mathrm{def}}{=} \mathcal{M}_{(k/d\wedge\vartheta)d}^{\mathrm{Grad}}(\varphi)$. We also omit the dependence on X in $\nabla\varphi = \nabla_X\varphi$.

Define the projected martingale increment

$$\begin{aligned}\Delta\mathcal{M}_k^{\mathrm{Grad},\beta} &= \frac{\gamma_k}{d} f'_{I_{k+1}}(\mathrm{Proj}_\beta(g_{k,I_{k+1}})) \cdot \big(\mathrm{Proj}_\beta(\langle\nabla\varphi(\hat{X}_k), \sqrt{K_i}v_{k+1}\rangle) + \langle\nabla\varphi(\hat{X}_k), \mu_i\rangle \\ &+ \delta s_{k,I_{k+1}}\langle\nabla\varphi(\hat{X}_k), \tfrac{X_k}{\|X_k\|}\rangle\big) - \frac{\gamma_k}{d}\sum_{i=1}^{2} p_i\mathbb{E}\left[f_i'(\mathrm{Proj}_\beta(g_{k,i}))\cdot(\mathrm{Proj}_\beta(\langle\nabla\varphi(\hat{X}_k), \sqrt{K_i}v_{k+1}\rangle)\right. \\ &\left. + \langle\nabla\varphi(\hat{X}_k), \mu_i\rangle + \delta s_{k,I_{k+1}}\langle\nabla\varphi(\hat{X}_k), \tfrac{X_k}{\|X_k\|}\rangle) \,|\, \mathcal{F}_k\right]\end{aligned} \tag{215}$$

For some $\zeta > 0$, set $\beta = d^\zeta$. We will later optimize the value of ζ. Our goal will be to bound $\Delta\mathcal{M}_k^{\mathrm{Grad},\beta}$ and then the difference $|\Delta\mathcal{M}_k^{\mathrm{Grad}} - \Delta\mathcal{M}_k^{\mathrm{Grad},\beta}|$.

To bound the difference, we will make use of subgaussian tail bounds of the projected quantities taking advantage of the fact that $a_{k,i} = \sqrt{K_i}v_{k+1} + \mu_i$ for v_{k+1} and v_{k+1} is subgaussian. Denoting $\widehat{r}_{k,i} \overset{\mathrm{def}}{=} \hat{X}_k^\top\sqrt{K_i}v_{k+1}$ then $\widehat{r}_{k,i}$ is $\|\sqrt{K}\|_{\mathrm{op}}\|\hat{X}_k\|$-subgaussian and $\langle\sqrt{K}\nabla\varphi(X_k), v_{k+1}\rangle$ is $\|\sqrt{K}\|_{\mathrm{op}}\|\nabla\varphi(X_k)\|$-subgaussian.

Given that $|s_{k,i}| \le 1$ and noting that $(g_{k,i}, \epsilon_{k+1}) = (\widehat{r}_{k,i}, \epsilon_{k+1}) + (\hat{X}_k^\top\mu_i, 0) + (\delta s_{k,i}\|X_k\|, 0_2) \in \mathbb{R}^3$, we have by the triangle inequality and Cauchy–Schwarz

$$\|(g_{k,i}, \epsilon_{k+1})\| \le \|(\hat{X}_k^\top\sqrt{K_i}v_{k+1}, \epsilon_{k+1})\| + \|\hat{X}^\top\mu_i\| + \delta\|X_k\|. \tag{216}$$

From Lemma D.5, we have

$$\|\nabla\varphi(X_k)\| = \|\nabla Z_{j\ell}(W_k, z)\| \le \sup_{z\in\Gamma}\|\nabla_X Z_{j\ell}(W_k, z)\| \le \|\nabla_X Z(W_k, z)\|_\Gamma \le C\|W_k\|. \tag{217}$$

Now, By Assumption 1, ϵ_{k+1} is η-subgaussian. Since η is a constant, $\|\sqrt{K_i}\|_{\mathrm{op}} \le C$, by (217) and we are working with the stopped processes, then for large enough C it follows

$$\|(\widehat{r}_{k,i}, \epsilon_{k+1})\|_{\psi_2} \le C, \quad \text{and} \quad \|\langle\sqrt{K_i}\nabla\varphi(X_k), v_{k+1}\rangle\|_{\psi_2} \le C. \tag{218}$$

Here we omit the dependence of the constant C on i as we could always redefine it as $C = \max\{C_1, C_2\}$. We will proceed with the same logic throughout the rest of the proof.

Hence, using (216) and for the stopped processes $\delta\|X_k\| \le \delta\|W_k\| \le CM$, $\|\hat{X}_k^\top\mu_i\| \le \|W_k\|\|\mu_i\| \le CM$. We then may apply the following argument to bound the tails of the projected quantities for $\beta > CM$

$$\mathbb{P}(\|(g_{k,i}, \epsilon_{k+1})\| > \beta) \le \mathbb{P}(\|(r_k, \epsilon_{k+1})\| > \beta - CM) \le 2\exp\left(-\frac{(\beta - CM)^2}{2C}\right). \tag{219}$$

Since we set $\beta = d^\zeta$ for $\zeta > 0$, the condition $\beta > CM$ holds for sufficiently large d hence it follows that

$$\mathbb{P}(\|(g_{k,i}, \epsilon_{k+1})\| > \beta) \le \exp(-\Omega(\beta^2)), \tag{220}$$

This tail bounds will be useful when we bound the difference between $\mathcal{M}_k^{\mathrm{Grad}}$ and $\mathcal{M}_k^{\mathrm{Grad},\beta}$.

From the definition of the projection and the bounds on the absolute value of f_i' in Lemma D.2 and Lemma D.3, we have

$$\begin{aligned}|f_i'(\mathrm{Proj}_\beta(g_{k,i}, \epsilon_{k+1}))| &\le CL(f)\beta^{\max(1,\alpha)}, \\ |\mathrm{Proj}_\beta(\langle\nabla\varphi(X_k), \sqrt{K_i}v_{k+1}\rangle)| &\le \beta.\end{aligned} \tag{221}$$

It then follows from Cauchy–Schwarz

$$|f_i'(\mathrm{Proj}_\beta(g_{k,i}, \epsilon_{k+1}))\cdot\mathrm{Proj}_\beta(\langle\nabla\varphi(X_k), \sqrt{K_i}v_{k+1}\rangle)| \le CL(f)\beta^{2+\alpha}. \tag{222}$$

Since we are working with the stopped process and $|s_{k,i}| \le 1$, we also have from Cauchy–Schwarz and (217)

$$\begin{aligned}|f_i'(\mathrm{Proj}_\beta(g_{k,i}, \epsilon_{k+1}))\cdot\langle\nabla\varphi(\hat{X}_k), \mu_i\rangle| &\le CML(f)\beta^{1+\alpha}, \\ |f_i'(\mathrm{Proj}_\beta(g_{k,i}, \epsilon_{k+1}))\cdot\delta s_{k,I_{k+1}}\langle\nabla\varphi(\hat{X}_k), \tfrac{X_k}{\|X_k\|}\rangle| &\le CML(f)\beta^{1+\alpha}.\end{aligned} \tag{223}$$

Since this is an almost sure bound, it also holds under expectation from which it follows that

$$|\Delta\mathcal{M}_k^{\mathrm{Grad},\beta}| \leq \frac{2\bar{\gamma}}{d} CL(f)\beta^{2+\alpha}, \quad a.s. \tag{224}$$

By Azuma's inequality then for $0 \leq k \leq Td$

$$\mathbb{P}\left(|\mathcal{M}_k^{\mathrm{Grad},\beta}| > t\right) \leq 2\exp\left(\frac{-t^2}{2\sum_{i=1}^k Cd^{-2}\beta^{2(2+\alpha)}}\right) \leq 2\exp\left(\frac{-t^2}{TCd^{-1}\beta^{2(2+\alpha)}}\right) \tag{225}$$

Thus, with overwhelming probability, for every $0 \leq k \leq Td$

$$|\mathcal{M}_k^{\mathrm{Grad},\beta}| < d^{-1/2}\beta^{3+\alpha}. \tag{226}$$

Taking a union bound from 0 to Td does not modify the overwhelming probability statement. Hence, with overwhelming probability

$$\sup_{0\leq k\leq Td} |\mathcal{M}_k^{\mathrm{Grad},\beta}| < d^{-1/2}\beta^{3+\alpha}. \tag{227}$$

It remains to bound the difference between $\mathcal{M}_k^{\mathrm{Grad}}$ and $\mathcal{M}_k^{\mathrm{Grad},\beta}$. For notational convenience, we write

$$\begin{aligned} F_{k,i} &\stackrel{\mathrm{def}}{=} \frac{\gamma_k}{d} f_i'(g_{k,i})\langle\nabla\varphi(\hat{X}_k), \sqrt{K_i}v_{k+1} + \mu_i + \delta s_{k,i}\tfrac{X_k}{\|X_k\|}\rangle, \\ F_{k,i}^\beta &\stackrel{\mathrm{def}}{=} \frac{\gamma_k}{d} f_i'(\mathrm{Proj}_\beta(g_{k,i}))\cdot\Big(\mathrm{Proj}_\beta(\langle\nabla\varphi(\hat{X}_k), \sqrt{K_i}v_{k+1}\rangle) \\ &\qquad + \langle\nabla\varphi(\hat{X}_k), \mu_i\rangle + \delta s_{k,i}\langle\nabla\varphi(\hat{X}_k), \tfrac{X_k}{\|X_k\|}\rangle\Big). \end{aligned} \tag{228}$$

Our goal is to show the quantity

$$|(F_{k,i} - \mathbb{E}[F_{k,i}]) - (F_{k,i}^\beta - \mathbb{E}[F_{k,i}^\beta])| \leq |F_{k,i} - F_{k,i}^\beta| + |\,\mathbb{E}[F_{k,i} - F_{k,i}^\beta]|, \tag{229}$$

vanishes as $d \to \infty$. Notice that

$$\begin{aligned} &|F_{k,i} - F_{k,i}^\beta| \\ &\leq \frac{\gamma_k}{d}|f_i'(g_{k,i})\langle\nabla\varphi(\hat{X}_k), \sqrt{K}v_{k+1}\rangle - f_i'(\mathrm{Proj}_\beta(g_{k,i}))\cdot\mathrm{Proj}_\beta(\langle\nabla\varphi(\hat{X}_k), \sqrt{K}v_{k+1}\rangle)| \\ &+ \frac{\gamma_k}{d}|\langle\nabla\varphi(\hat{X}_k), \mu_i\rangle|\cdot|f_i'(g_{k,i}) - f_i'(\mathrm{Proj}_\beta(g_{k,i}))| \\ &+ \frac{\gamma_k}{d}\delta|\langle\nabla\varphi(\hat{X}_k), \tfrac{X_k}{\|X_k\|}\rangle|\cdot|f_i'(g_{k,i}) - f_i'(\mathrm{Proj}_\beta(g_{k,i}))| \end{aligned}$$

Thus, using the subgaussian tail bounds we previously derived in (220) and (220), we obtain

$$\begin{aligned} \mathbb{P}(|F_{k,I_{k+1}} - F_{k,I_{k+1}}^\beta| > 0) &= \mathbb{P}(\{\|(g_{k,I_{k+1}}, \epsilon_{k+1})\| > \beta\} \cup \{|\langle\nabla\varphi(\hat{X}_k), \sqrt{K_{I_{k+1}}}v_{k+1}\rangle| > \beta\}) \\ &\leq \sum_{i=1}^2 p_i(\mathbb{P}(\|(g_{k,i}, \epsilon_{k+1})\| > \beta) + \mathbb{P}(|\langle\nabla\varphi(\hat{X}_k), \sqrt{K_i}v_{k+1}\rangle| > \beta)) \\ &\leq 4\exp\left(-\Omega(\beta^2)\right). \end{aligned} \tag{230}$$

Since we defined $\beta = d^\zeta$ for some $\zeta > 0$, then (230) implies $F_k - F_k^\beta = 0$ with overwhelming probability. It then follows from a union bound that

$$\sum_{k=1}^{Td} |F_{k,I_{k+1}} - F_{k,I_{k+1}}^\beta| = 0, \tag{231}$$

with overwhelming probability. We must now bound the difference in expectation. In the following expectations, we omit the conditional expectation on $\mathcal{F}_k$ for notational simplicity. We have by Cauchy–Schwarz

$$\begin{aligned} |\,\mathbb{E}[F_{k,i} - F_{k,i}^\beta]| &= |\,\mathbb{E}[(F_{k,i} - F_{k,i}^\beta)\mathbb{1}_{|F_{k,i}-F_{k,i}^\beta|>0}]| \\ &\leq |\,\mathbb{E}[F_{k,i}\mathbb{1}_{|F_{k,i}-F_{k,i}^\beta|>0}]| + |\,\mathbb{E}[F_{k,i}^\beta\mathbb{1}_{|F_{k,i}-F_{k,i}^\beta|>0}]|. \end{aligned} \tag{232}$$

Using a similar argument as in (A.78)-(A.82) in Collins-Woodfin et al. (2024a), we will show both of the terms in the upper bound are exponentially small with high-probability. First, using (222) and (223) we have

$$\begin{aligned} |\,\mathbb{E}[F_{k,i}^{\beta}\mathbb{1}_{|F_{k,i}-F_{k,i}^{\beta}|>0}]| &\leq \frac{\bar{\gamma}}{d}CL(f)\beta^{2+\alpha}\mathbb{P}\left(|F_{k,i}-F_{k,i}^{\beta}|>0\right) \\ &\leq \frac{\bar{\gamma}}{d}CL(f)\beta^{2+\alpha}\exp\left(-\Omega(\beta^2)\right). \end{aligned} \tag{233}$$

Then for the second term we have

$$\begin{aligned} |\mathbb{E}\left[F_{k,i}\cdot\mathbb{1}_{|F_{k,i}-F_{k,i}^{\beta}|>0}\right]| &\leq \mathbb{E}\left[|F_{k,i}|\cdot\mathbb{1}_{E_1}\right]+\mathbb{E}\left[|F_{k,i}|\cdot\mathbb{1}_{E_2}\right]+\mathbb{E}\left[|F_{k,i}|\cdot\mathbb{1}_{E_3}\right], \\ \text{where } E_1 &\stackrel{\text{def}}{=}\{\|(g_{k,i},\epsilon_{k+1})\|\leq\beta\}\cap\{|\langle\nabla\varphi(\hat{X}_k),\sqrt{K_i}v_{k+1}\rangle|>\beta\}, \\ E_2 &\stackrel{\text{def}}{=}\{\|(g_{k,i},\epsilon_{k+1})\|>\beta\}\cap\{|\langle\nabla\varphi(\hat{X}_k),\sqrt{K_i}v_{k+1}\rangle|\leq\beta\}, \\ E_3 &\stackrel{\text{def}}{=}\{\|(g_{k,i},\epsilon_{k+1})\|>\beta\}\cap\{|\langle\nabla\varphi(\hat{X}_k),\sqrt{K_i}v_{k+1}\rangle|>\beta\}. \end{aligned} \tag{234}$$

From the bounds in (221) and (223), we obtain for the term involving E_1

$$\begin{aligned} \mathbb{E}\left[|F_{k,i}|\cdot\mathbb{1}_{E_1}\right] \leq{}& \frac{\bar{\gamma}}{d}CML(f)\beta^{1+\alpha}\mathbb{P}(|\langle\nabla\varphi(\hat{X}_k),\sqrt{K_i}v_{k+1}\rangle|>\beta) \\ &+\frac{\bar{\gamma}}{d}CL(f)\beta^{1+\alpha}\,\mathbb{E}\left[|\langle\nabla\varphi(\hat{X}_k),\sqrt{K_i}v_{k+1}\rangle|\mathbb{1}_{|\langle\nabla\varphi(\hat{X}_k),\sqrt{K_i}v_{k+1}\rangle|>\beta}\right]. \end{aligned} \tag{235}$$

We will show $\mathbb{E}\left[|\langle\nabla\varphi(\hat{X}_k),\sqrt{K_i}v_{k+1}\rangle|\mathbb{1}_{|\langle\nabla\varphi(\hat{X}_k),\sqrt{K_i}v_{k+1}\rangle|>\beta}\right]$ is exponentially small. Using (220), we have for some constant $c>0$

$$\begin{aligned} &\mathbb{E}\left[\langle\nabla\varphi(\hat{X}_k),\sqrt{K_i}v_{k+1}\rangle|\mathbb{1}_{|\langle\nabla\varphi(\hat{X}_k),\sqrt{K_i}v_{k+1}\rangle|>\beta}\right] \\ &\quad=\beta\mathbb{P}(|\langle\nabla\varphi(\hat{X}_k),\sqrt{K_i}v_{k+1}\rangle|>\beta)+\int_{\beta}^{\infty}\mathbb{P}(|\langle\nabla\varphi(\hat{X}_k),\sqrt{K_i}v_{k+1}\rangle|>t)\,\mathrm{d}t \\ &\quad\leq 2\beta\exp\left(-\Omega(\beta^2)\right)+\int_{\beta}^{\infty}2\exp\left(-\frac{t^2}{2C}\right)\mathrm{d}t \\ &\quad=2\beta\exp\left(-\Omega(\beta^2)\right)+2\sqrt{2\pi C}\,\mathbb{P}\left(\mathcal{N}(0,1)>\frac{\beta}{\sqrt{C}}\right) \\ &\quad\leq C\beta\exp\left(-\Omega(\beta^2)\right), \end{aligned} \tag{236}$$

as Gaussian random variables are subgaussian. Since $\beta=d^{\zeta}$ for some $\zeta>0$ then the bound is exponentially small in d. Thus, we obtain

$$\mathbb{E}\left[|F_{k,i}|\cdot\mathbb{1}_{E_1}\right]\leq\frac{\bar{\gamma}}{d}L(f)C\beta^{2+\alpha}\exp\left(-\Omega(\beta^2)\right). \tag{237}$$

Since we are working with the stopped processes, from Cauchy–Schwarz, (220), (221) and the moments bound in Lemma D.3, we have for the term involving E_2

$$\begin{aligned} \mathbb{E}\left[|F_{k,i}|\cdot\mathbb{1}_{E_2}\right] &\leq \frac{\bar{\gamma}}{d}CM\beta\,\mathbb{E}\left[|f_i'(g_{k,i},\epsilon_{k+1})|\mathbb{1}_{\|(g_{k,i},\epsilon_{k+1})\|>\beta}\right] \\ &\leq \frac{\bar{\gamma}}{d}CM\beta\left(\mathbb{E}\left[|f_i'(g_{k,i},\epsilon_{k+1})|^2\right]\right)^{1/2}\mathbb{P}(\|(g_{k,i},\epsilon_{k+1})\|>\beta)^{1/2} \\ &\leq \frac{\bar{\gamma}}{d}CML(f)\beta\left(1+(\|K_i\|_{\mathrm{op}}^{1/2}+\|\mu_i\|)\|\hat{X}_k\|+\delta\|X_k\|+\eta\right)^{1+\alpha}\exp(-\Omega(\beta^2)) \end{aligned} \tag{238}$$

Since we are working with the stopped processes and from Assumption 1, $\|\hat{X}_k\|$, $\|W_k\|$, $\|K\|_{\mathrm{op}}$, η^2, δ are bounded by constants (independent of d) so we obtain

$$\mathbb{E}\left[|F_{k,i}|\cdot\mathbb{1}_{E_2}\right]\leq\frac{\bar{\gamma}}{d}CL(f)\beta\exp(-\Omega(\beta^2)) \tag{239}$$

For the term involving E_3, we use Cauchy–Schwarz to obtain

$$\begin{aligned}
&\mathbb{E}\left[|F_{k,i}| \cdot \mathbb{1}_{E_3}\right] \\
&\quad \leq \frac{\bar{\gamma}}{d}(\mathbb{E}\left[|\langle \nabla\varphi(\hat{X}_k), \sqrt{K_i} v_{k+1}\rangle|^2 \mathbb{1}_{\|(g_{k,i},\epsilon_{k+1})\|>\beta}\right])^{1/2}(\mathbb{E}\left[|f_i'(g_{k,i},\epsilon_{k+1})|^2 \mathbb{1}_{\|(g_{k,i},\epsilon_{k+1})\|>\beta}\right])^{1/2} \\
&\quad + \frac{\bar{\gamma}}{d} CM\, \mathbb{E}\left[|f_i'(g_{k,i},\epsilon_{k+1})| \mathbb{1}_{\|(g_{k,i},\epsilon_{k+1})\|>\beta}\right].
\end{aligned} \tag{240}$$

Using similar steps as for E_1 and E_2, this is also exponentially small. Putting this altogether, we obtain

$$|\mathbb{E}\left[F_{k,i} \cdot \mathbb{1}_{|F_{k,i} - F^{\beta}_{k,i}|>0}\right]| \leq C d^{-1} \text{poly}(\beta) \exp(-\Omega(\beta^2)) \tag{241}$$

For simplicity of notation, let $\mathcal{E} = \{\sum_{k=1}^{Td} |F_{k,I_{k+1}} - F^{\beta}_{k,I_{k+1}}| = 0\}$ then with $\beta = d^{\zeta}$ we have

$$\begin{aligned}
&\mathbb{P}(\sup_{0\leq k\leq Td} |\mathcal{M}_k^{\text{Grad}}| > t) \\
&\leq \mathbb{P}(\sup_{0\leq k\leq Td} |\mathcal{M}_k^{\text{Grad},\beta}| > t - \sum_{k=1}^{Td} |F_{k,I_{k+1}} - F^{\beta}_{k,I_{k+1}}| - \sum_{k=1}^{Td}\sum_{i=1}^{2} p_i |\,\mathbb{E}[F_{k,i} - F^{\beta}_{k,i}]|) \\
&\leq \mathbb{P}(\sup_{0\leq k\leq Td} |\mathcal{M}_k^{\text{Grad},\beta}| > t - \sum_{k=1}^{Td}\sum_{i=1}^{2} p_i |\,\mathbb{E}[F_{k,i} - F^{\beta}_{k,i}]|,\ \mathcal{E}) + \mathbb{P}(\mathcal{E}^c) \\
&\leq \mathbb{P}(\sup_{0\leq k\leq Td} |\mathcal{M}_k^{\text{Grad},\beta}| > t - \text{poly}(d^{\zeta}) \exp(-\Omega(d^{2\zeta}))) + \exp(-\Omega(d^{2\zeta})).
\end{aligned} \tag{242}$$

Setting $t \overset{\text{def}}{=} d^{-\frac{1}{2}+(3+\alpha)\zeta}$ then for sufficiently large d, $t - \text{poly}(d^{\zeta}) \exp(-\Omega(d^{2\zeta}))) \geq \frac{t}{2}$. Thus, it follows from Azuma's inequality and a union bound as done in (227) that, with overwhelming probability

$$\sup_{0\leq k\leq Td} |\mathcal{M}_k^{\text{Grad}}| < d^{-\frac{1}{2}+(3+\alpha)\zeta}. \tag{243}$$

Hence for any arbitrarily small value of $\zeta < \frac{1}{2(3+\alpha)}$, the result follows. □

In the following proof Proposition D.3, we build upon Proposition A.3 of Collins-Woodfin et al. (2024a) and Lemma 10 in Collins-Woodfin & Seroussi (2025).

Proposition D.3 (Hessian martingale)**.** *For $i = 1, 2$, suppose $f_i : \mathbb{R}^3 \to \mathbb{R}$ is an α-pseudo-Lipschitz function with constant $L(f)$ (see Assumption 3). Let the statistic $Z : \mathbb{R}^d \to \mathbb{R}^{4\times 4}$ be defined as in* (202)*. Then, for any $\zeta > 0$ and $T > 0$, and with overwhelming probability,*

$$\sup_{0\leq t\leq T} \|\mathcal{M}^{\text{Hess}}_{\lfloor (t\wedge\vartheta)d\rfloor}(Z(\cdot, z))\| < d^{-\frac{1}{2}+\zeta}. \tag{244}$$

Proof. Similarly to in the proof Proposition D.2, we denote $\varphi(\hat{X}_k) \overset{\text{def}}{=} Z_{j\ell}(W, z)$ as the $j\ell$-coordinate of Z for a fixed $z \in \Gamma^2$. We will also work with the stopped process $\mathcal{M}^{\text{Hess}}_{(t\wedge\vartheta)d}(\varphi)$ and, for notational simplicity, omit the dependence on the stopping time ϑ and on φ such that $\mathcal{M}^{\text{Hess}}_k \overset{\text{def}}{=} \mathcal{M}^{\text{Hess}}_{(k/d\wedge\vartheta)d}(\varphi)$. We also omit the dependence on X in $\nabla^2\varphi = \nabla^2_X\varphi$.

For some $\zeta > 0$, set $\beta = d^{\zeta}$. We will later optimize the value of ζ similarly to the proof Proposition D.2. Recall the hessian martingale increment

$$\Delta\mathcal{M}^{\text{Hess}}_k = \frac{\gamma_k^2}{2d^2}\left(\langle \nabla^2\varphi(X_k), \Delta_k^{\otimes 2}\rangle - \mathbb{E}\left[\langle \nabla^2\varphi(X_k), \Delta_k^{\otimes 2}\rangle \,|\, \mathcal{F}_k\right]\right), \tag{245}$$

where $\Delta_k = f'_{I_{k+1}}(g_{k,I_{k+1}})(a_{k+1,I_{k+1}} + \delta s_{k,I_{k+1}} \frac{X_k}{\|X_k\|}) + \lambda X_k$. We first split this term into nine separate increments

$$\Delta\mathcal{M}^{\text{Hess}}_k = \sum_{q=1}^{9} \Delta\mathcal{M}^{\text{Hess},q}_k, \tag{246}$$

where

$$
\begin{aligned}
\Delta\mathcal{M}_k^{\text{Hess},1} &= \frac{\gamma_k^2}{2d^2}\Bigg((f'_{I_{k+1}}(g_{k,I_{k+1}}))^2 \langle \nabla^2\varphi(X_k), (\sqrt{K_{I_{k+1}}}v_{k+1})^{\otimes 2}\rangle \\
&\qquad - \sum_{i=1}^{2} p_i \mathbb{E}\left[(f'_i(g_{k,i}))^2 \langle \nabla^2\varphi(X_k), (\sqrt{K_i}v_{k+1})^{\otimes 2}\rangle \mid \mathcal{F}_k\right]\Bigg) \\
\Delta\mathcal{M}_k^{\text{Hess},2} &= \frac{\gamma_k^2}{d^2}\Bigg((f'_{I_{k+1}}(g_{k,I_{k+1}}))^2 \langle \nabla^2\varphi(X_k), \sqrt{K_{I_{k+1}}}v_{k+1}\otimes \mu_{I_{k+1}}\rangle \\
&\qquad - \sum_{i=1}^{2} p_i \mathbb{E}\left[(f'_i(g_{k,i}))^2 \langle \nabla^2\varphi(X_k), \sqrt{K_{I_{k+1}}}v_{k+1}\otimes \mu_{I_{k+1}}\rangle \mid \mathcal{F}_k\right]\Bigg) \\
\Delta\mathcal{M}_k^{\text{Hess},3} &= \frac{\gamma_k^2}{2d^2}\Bigg((f'_{I_{k+1}}(g_{k,I_{k+1}}))^2 \langle \nabla^2\varphi(X_k), \mu_{I_{k+1}}^{\otimes 2}\rangle \\
&\qquad - \sum_{i=1}^{2} p_i \mathbb{E}\left[(f'_i(g_{k,i}))^2 \langle \nabla^2\varphi(X_k), \mu_i^{\otimes 2}\rangle \mid \mathcal{F}_k\right]\Bigg) \\
\Delta\mathcal{M}_k^{\text{Hess},4} &= \frac{\gamma_k^2}{d^2}\Bigg(\delta s_{k,I_{k+1}} (f'_{I_{k+1}}(g_{k,I_{k+1}}))^2 \langle \nabla^2\varphi(X_k), \mu_{I_{k+1}}\otimes \frac{X_k}{\|X_k\|}\rangle \\
&\qquad - \sum_{i=1}^{2} p_i \mathbb{E}\left[\delta s_{k,i} (f'_i(g_{k,i}))^2 \langle \nabla^2\varphi(X_k), \mu_i\otimes \frac{X_k}{\|X_k\|}\rangle \mid \mathcal{F}_k\right]\Bigg) \\
\Delta\mathcal{M}_k^{\text{Hess},5} &= \frac{\gamma_k^2}{d^2}\Bigg(\delta s_{k,I_{k+1}} (f'_{I_{k+1}}(g_{k,I_{k+1}}))^2 \langle \nabla^2\varphi(X_k), \sqrt{K_{I_{k+1}}}v_{k+1}\otimes \frac{X_k}{\|X_k\|}\rangle \\
&\qquad - \sum_{i=1}^{2} p_i \mathbb{E}\left[\delta s_{k,i} (f'_i(g_{k,i}))^2 \langle \nabla^2\varphi(X_k), \sqrt{K_i}v_{k+1}\otimes \frac{X_k}{\|X_k\|}\rangle \mid \mathcal{F}_k\right]\Bigg) \\
\Delta\mathcal{M}_k^{\text{Hess},6} &= \frac{\gamma_k^2}{2d^2}\delta^2 \langle \nabla^2\varphi(X_k), \frac{X_k^{\otimes 2}}{\|X_k\|^2}\rangle \Bigg((s_{k,I_{k+1}})^2 (f'_{I_{k+1}}(g_{k,I_{k+1}}))^2 \\
&\qquad - \sum_{i=1}^{2} p_i \mathbb{E}\left[(s_{k,i}^2 (f'_i(g_{k,i}))^2 \mid \mathcal{F}_k\right]\Bigg) \\
\Delta\mathcal{M}_k^{\text{Hess},7} &= \frac{\gamma_k^2}{d^2}\Bigg(f'_{I_{k+1}}(g_{k,I_{k+1}}) \langle \nabla^2\varphi(X_k), \sqrt{K_{I_{k+1}}}v_{k+1}\otimes \lambda X_k\rangle \\
&\qquad - \sum_{i=1}^{2} p_i \mathbb{E}\left[f'_i(g_{k,i}) \langle \nabla^2\varphi(X_k), \sqrt{K_i}v_{k+1}\otimes \lambda X_k\rangle \mid \mathcal{F}_k\right]\Bigg) \\
\Delta\mathcal{M}_k^{\text{Hess},8} &= \frac{\gamma_k^2}{d^2}\Bigg(f'_{I_{k+1}}(g_{k,I_{k+1}}) \langle \nabla^2\varphi(X_k), \mu_{I_{k+1}}\otimes \lambda X_k\rangle \\
&\qquad - \sum_{i=1}^{2} p_i \mathbb{E}\left[f'_i(g_{k,i}) \langle \nabla^2\varphi(X_k), \mu_i\otimes \lambda X_k\rangle \mid \mathcal{F}_k\right]\Bigg) \\
\Delta\mathcal{M}_k^{\text{Hess},9} &= \frac{\gamma_k^2}{d^2}\delta\lambda \langle \nabla^2\varphi(X_k), \frac{X_k^{\otimes 2}}{\|X_k\|}\rangle \Bigg(s_{k,I_{k+1}} f'_{I_{k+1}}(g_{k,I_{k+1}}) \\
&\qquad - \sum_{i=1}^{2} p_i \mathbb{E}\left[s_{k,i} f'_i(g_{k,i}) \mid \mathcal{F}_k\right]\Bigg)
\end{aligned}
\tag{247}
$$

As done in the proof Proposition D.2, we introduce the projected increments

$$
\begin{aligned}
\Delta\mathcal{M}_k^{\text{Hess},1,\beta} &= \frac{\gamma_k^2}{2d^2}\Big((f'_{I_{k+1}}(\text{Proj}_\beta(g_{k,I_{k+1}})))^2 \,\text{Proj}_{d\beta}(\langle\nabla^2\varphi(X_k), (\sqrt{K_{I_{k+1}}}v_{k+1})^{\otimes 2}\rangle) \\
&\qquad - \sum_{i=1}^2 p_i \mathbb{E}\left[(f'_i(\text{Proj}_\beta(g_{k,i})))^2 \,\text{Proj}_{d\beta}(\langle\nabla^2\varphi(X_k), (\sqrt{K_i}v_{k+1})^{\otimes 2}\rangle) \,|\, \mathcal{F}_k\right]\Big) \\
\Delta\mathcal{M}_k^{\text{Hess},2,\beta} &= \frac{\gamma_k^2}{d^2}\Big((f'_{I_{k+1}}(\text{Proj}_\beta(g_{k,I_{k+1}})))^2 \,\text{Proj}_\beta(\langle\nabla^2\varphi(X_k), \sqrt{K_{I_{k+1}}}v_{k+1}\otimes\mu_{I_{k+1}}\rangle) \\
&\qquad - \sum_{i=1}^2 p_i \mathbb{E}\left[(f'_i(\text{Proj}_\beta(g_{k,i})))^2 \,\text{Proj}_\beta(\langle\nabla^2\varphi(X_k), \sqrt{K_{I_{k+1}}}v_{k+1}\otimes\mu_{I_{k+1}}\rangle) \,|\, \mathcal{F}_k\right]\Big) \\
\Delta\mathcal{M}_k^{\text{Hess},3,\beta} &= \frac{\gamma_k^2}{2d^2}\Big((f'_{I_{k+1}}(\text{Proj}_\beta(g_{k,I_{k+1}})))^2 \langle\nabla^2\varphi(X_k), \mu_{I_{k+1}}^{\otimes 2}\rangle \\
&\qquad - \sum_{i=1}^2 p_i \mathbb{E}\left[(f'_i(\text{Proj}_\beta(g_{k,i})))^2 \langle\nabla^2\varphi(X_k), \mu_i^{\otimes 2}\rangle \,|\, \mathcal{F}_k\right]\Big) \\
\Delta\mathcal{M}_k^{\text{Hess},4,\beta} &= \frac{\gamma_k^2}{d^2}\Big(\delta s_{k,I_{k+1}} (f'_{I_{k+1}}(\text{Proj}_\beta(g_{k,I_{k+1}})))^2 \langle\nabla^2\varphi(X_k), \mu_{I_{k+1}}\otimes\frac{X_k}{\|X_k\|}\rangle \\
&\qquad - \sum_{i=1}^2 p_i \mathbb{E}\left[\delta s_{k,i} (f'_i(\text{Proj}_\beta(g_{k,i})))^2 \langle\nabla^2\varphi(X_k), \mu_i\otimes\frac{X_k}{\|X_k\|}\rangle \,|\, \mathcal{F}_k\right]\Big) \\
\Delta\mathcal{M}_k^{\text{Hess},5,\beta} &= \frac{\gamma_k^2}{d^2}\Big(\delta s_{k,I_{k+1}} (f'_{I_{k+1}}(\text{Proj}_\beta(g_{k,I_{k+1}})))^2 \,\text{Proj}_\beta(\langle\nabla^2\varphi(X_k), \sqrt{K_{I_{k+1}}}v_{k+1}\otimes\frac{X_k}{\|X_k\|}\rangle) \\
&\qquad - \sum_{i=1}^2 p_i \mathbb{E}\left[\delta s_{k,i} (f'_i(\text{Proj}_\beta(g_{k,i})))^2 \,\text{Proj}_\beta(\langle\nabla^2\varphi(X_k), \sqrt{K_i}v_{k+1}\otimes\frac{X_k}{\|X_k\|}\rangle) \,|\, \mathcal{F}_k\right]\Big) \\
\Delta\mathcal{M}_k^{\text{Hess},6,\beta} &= \frac{\gamma_k^2}{2d^2}\delta^2\langle\nabla^2\varphi(X_k), \frac{X_k^{\otimes 2}}{\|X_k\|^2}\rangle\Big(s_{k,I_{k+1}}^2 (f'_{I_{k+1}}(\text{Proj}_\beta(g_{k,I_{k+1}})))^2 \\
&\qquad - \sum_{i=1}^2 p_i \mathbb{E}\left[(s_{k,i}^2 (f'_i(\text{Proj}_\beta(g_{k,i})))^2 \,|\, \mathcal{F}_k\right]\Big) \\
\Delta\mathcal{M}_k^{\text{Hess},7,\beta} &= \frac{\gamma_k^2}{d^2}\Big(f'_{I_{k+1}}(\text{Proj}_\beta(g_{k,I_{k+1}})) \,\text{Proj}_\beta(\langle\nabla^2\varphi(X_k), \sqrt{K_{I_{k+1}}}v_{k+1}\otimes\lambda X_k\rangle) \\
&\qquad - \sum_{i=1}^2 p_i \mathbb{E}\left[f'_i(\text{Proj}_\beta(g_{k,i})) \,\text{Proj}_\beta(\langle\nabla^2\varphi(X_k), \sqrt{K_i}v_{k+1}\otimes\lambda X_k\rangle) \,|\, \mathcal{F}_k\right]\Big) \\
\Delta\mathcal{M}_k^{\text{Hess},8,\beta} &= \frac{\gamma_k^2}{d^2}\lambda\Big(f'_{I_{k+1}}(\text{Proj}_\beta(g_{k,I_{k+1}}))\langle\nabla^2\varphi(X_k), \mu_{I_{k+1}}\otimes X_k\rangle \\
&\qquad - \sum_{i=1}^2 p_i \mathbb{E}\left[f'_i(\text{Proj}_\beta(g_{k,i}))\langle\nabla^2\varphi(X_k), \mu_i\otimes X_k\rangle \,|\, \mathcal{F}_k\right]\Big) \\
\Delta\mathcal{M}_k^{\text{Hess},9,\beta} &= \frac{\gamma_k^2}{d^2}\delta\lambda\langle\nabla^2\varphi(X_k), \frac{X_k^{\otimes 2}}{\|X_k\|}\rangle\Big(s_{k,I_{k+1}} f'_{I_{k+1}}(\text{Proj}_\beta(g_{k,I_{k+1}})) \\
&\qquad - \sum_{i=1}^2 p_i \mathbb{E}\left[s_{k,i} f'_i(\text{Proj}_\beta(g_{k,i})) \,|\, \mathcal{F}_k\right]\Big)
\end{aligned}
\tag{248}
$$

Note that the projection radius of $\Delta\mathcal{M}_k^{\text{Hess},1,\beta}$ differs from the projection radius β due to the order of the fluctuations of $\langle\nabla^2\varphi(X_k), (\sqrt{K_i}v_{k+1})^{\otimes 2}\rangle$. We begin by bounding the terms not involving v_{k+1} which are $\mathcal{M}_k^{\text{Hess},3}$, $\mathcal{M}_k^{\text{Hess},4}$, $\mathcal{M}_k^{\text{Hess},6}$, $\mathcal{M}_k^{\text{Hess},8}$, and $\mathcal{M}_k^{\text{Hess},9}$. From Lemma D.5 and a similar

reasoning as (217), we have

$$|\langle\nabla^2\varphi(X_k),\mu_i^{\otimes 2}\rangle|\le C\|\mu_i\|^2,\quad |\langle\nabla^2\varphi(X_k),\mu_i\otimes\frac{X_k}{\|X_k\|}\rangle|\le C\|\mu_i\|,$$
$$|\langle\nabla^2\varphi(X_k),\mu_i\otimes X_k\rangle|\le C\|\mu_i\|\|X_k\|,\quad |\langle\nabla^2\varphi(X_k),\frac{X_k^{\otimes 2}}{\|X_k\|}\rangle|\le C\|X_k\|, \tag{249}$$
$$\text{and}\quad |\langle\nabla^2\varphi(X_k),\frac{X_k^{\otimes 2}}{\|X_k\|^2}\rangle|\le C.$$

Here we omit the dependence of the constant C on i as we could always redefine it as $C=\max\{C_1,C_2\}$. We will proceed with the same logic throughout the rest of the proof.

Since we are working with the stopped processes and $\|\mu_i\|^2\le C$, it follows

$$|\langle\nabla^2\varphi(X_k),\mu_i^{\otimes 2}\rangle|\le C,\quad |\langle\nabla^2\varphi(X_k),\mu_i\otimes\frac{X_k}{\|X_k\|}\rangle|\le C,$$
$$|\langle\nabla^2\varphi(X_k),\mu_i\otimes X_k\rangle|\le CM,\quad |\langle\nabla^2\varphi(X_k),\frac{X_k^{\otimes 2}}{\|X_k\|}\rangle|\le CM, \tag{250}$$
$$\text{and}\quad |\langle\nabla^2\varphi(X_k),\frac{X_k^{\otimes 2}}{\|X_k\|^2}\rangle|\le C$$

Using the bounds on the absolute value of f' in Lemma D.2 and Lemma D.3, we have

$$\begin{aligned}|f_i'(\mathrm{Proj}_\beta(g_{k,i}))|&\le CL(f)\beta^{1+\alpha},\\ |(f_i'(\mathrm{Proj}_\beta(g_{k,i})))^2|&\le (CL(f)\beta^{1+\alpha})^2.\end{aligned} \tag{251}$$

Since $|s_k|\le 1$, it immediately follows

$$\begin{aligned}|s_kf_i'(\mathrm{Proj}_\beta(g_{k,i}))|&\le CL(f)\beta^{\max(1+\alpha)},\\ |s_k^2(f_i'(\mathrm{Proj}_\beta(g_{k,i})))^2|&\le (CL(f)\beta^{1+\alpha})^2.\end{aligned} \tag{252}$$

These bounds holding almost surely imply they also hold under expectation from which we obtain

$$|\Delta\mathcal{M}_k^{\mathrm{Hess},3,\beta}|\le C\frac{\bar\gamma^2}{d^2}L(f)^2\beta^{2+2\alpha},\quad |\Delta\mathcal{M}_k^{\mathrm{Hess},4,\beta}|\le C\frac{\bar\gamma^2\delta}{d^2}L(f)^2\beta^{2+2\alpha},$$
$$|\Delta\mathcal{M}_k^{\mathrm{Hess},6,\beta}|\le C\frac{\bar\gamma^2\delta^2}{d^2}L(f)^2\beta^{2+2\alpha},\quad |\Delta\mathcal{M}_k^{\mathrm{Hess},8,\beta}|\le 2C\frac{\bar\gamma^2\lambda}{d^2}\lambda L(f)\beta^{1+\alpha}, \tag{253}$$
$$\text{and}\quad |\Delta\mathcal{M}_k^{\mathrm{Hess},9,\beta}|\le C\frac{\bar\gamma^2}{d^2}\lambda\delta L(f)\beta^{1+\alpha}.$$

Using a similar argument involving Azuma's inequality as in the proof Proposition D.2, we obtain with overwhelming probability

$$\sup_{0\le k\le Td}|\mathcal{M}_k^{\mathrm{Hess},3,\beta}|<d^{-3/2}\beta^{3+2\alpha},\quad \sup_{0\le k\le Td}|\mathcal{M}_k^{\mathrm{Hess},4,\beta}|<d^{-3/2}\beta^{3+2\alpha},$$
$$\sup_{0\le k\le Td}|\mathcal{M}_k^{\mathrm{Hess},6,\beta}|<d^{-3/2}\beta^{3+2\alpha},\quad \sup_{0\le k\le Td}|\mathcal{M}_k^{\mathrm{Hess},8,\beta}|<d^{-3/2}\beta^{2+\alpha}, \tag{254}$$
$$\text{and}\quad \sup_{0\le k\le Td}|\mathcal{M}_k^{\mathrm{Hess},9,\beta}|<d^{-3/2}\beta^{2+\alpha}.$$

Using similar steps as in the proof Proposition D.2, it then follows that the differences $|\mathcal{M}_k^{\mathrm{Hess},q}-\mathcal{M}_k^{\mathrm{Hess},q,\beta}|$ for $q\in\{3,4,6,8,9\}$ are exponentially small with overwhelming probability.

We now continue by bounding the terms linear in v_{k+1} which are $\mathcal{M}_k^{\mathrm{Hess},2,\beta}$, $\mathcal{M}_k^{\mathrm{Hess},5,\beta}$ and $\mathcal{M}_k^{\mathrm{Hess},7,\beta}$. The argument is once again extremely similar to the proof Proposition D.2. Setting $\widehat{r}_{k,i}\overset{\mathrm{def}}{=}\hat{X}_k\sqrt{K_i}v_{k+1}$ and using a similar argument as (218) and from Lemma D.5, for large enough constant C we have

$$\|(\widehat{r}_{k,i},\epsilon_{k+1})\|_{\psi_2}\le C,\quad \|\langle\sqrt{K_i}\nabla^2\varphi(X_k)\mu_i,v_{k+1}\rangle\|_{\psi_2}\le C$$
$$\|\langle\sqrt{K_i}\nabla^2\varphi(X_k)\frac{X_k}{\|X_k\|},v_{k+1}\rangle\|_{\psi_2}\le C,\quad\text{and}\quad \|\lambda\langle\sqrt{K_i}\nabla^2\varphi(X_k)X_k,v_{k+1}\rangle\|_{\psi_2}\le C. \tag{255}$$

By (251), (252), the definition of the projection operator and using a similar argument as in (222) in the proof Proposition D.2, we have

$$|\Delta\mathcal{M}_k^{\mathrm{Hess},2,\beta}| \le \frac{\bar{\gamma}^2}{d^2} C(L(f))^2\beta^{3+2\alpha} \quad , |\Delta\mathcal{M}_k^{\mathrm{Hess},5,\beta}| \le \delta\frac{\bar{\gamma}^2}{d^2} CL(f)\beta^{3+2\alpha}, \\ \text{and} \quad \Delta\mathcal{M}_k^{\mathrm{Hess},7,\beta}| \le \frac{\bar{\gamma}^2}{d^2} CL(f)\beta^{2+\alpha}. \tag{256}$$

from which it follows

$$\sup_{0\le k\le Td} |\mathcal{M}_k^{\mathrm{Hess},2,\beta}| < d^{-3/2}\beta^{4+2\alpha}, \quad \sup_{0\le k\le Td} |\mathcal{M}_k^{\mathrm{Hess},5,\beta}| < d^{-3/2}\beta^{4+2\alpha}, \\ \text{and} \quad \sup_{0\le k\le Td} |\mathcal{M}_k^{\mathrm{Hess},7,\beta}| < d^{-3/2}\beta^{3+\alpha}. \tag{257}$$

From the subgaussian norm bounds (255) and bounds on f_i' from Lemma D.3, we can use an almost identical argument as in Proposition D.2 to show the differences $|\mathcal{M}_k^{\mathrm{Hess},q} - \mathcal{M}_k^{\mathrm{Hess},q,\beta}|$ are exponentially small with overwhelming probability for $q = 2, 5, 7$.

Finally, it remains to bound the term $\mathcal{M}_k^{\mathrm{Hess},1,\beta}$ which is quadratic in v_{k+1}. The argument will once again be similar to the proof Proposition D.2 and the steps above. We have

$$\langle\nabla^2\varphi(X_k), (\sqrt{K_i}v_{k+1})^{\otimes 2}\rangle = \langle\sqrt{K_i}\nabla^2\varphi(X_k)\sqrt{K_i}, v_{k+1}^{\otimes 2}\rangle. \tag{258}$$

It then follows from Hanson-Wright's inequality

$$\mathbb{P}\left(|\langle\sqrt{K_i}\nabla^2\varphi(X_k)\sqrt{K_i}, v_{k+1}^{\otimes 2}\rangle - \mathrm{Tr}(\nabla^2\varphi(X_k)K_i)| > t\right) \\ \le 2\exp\left(-C\min\left\{\frac{t^2}{\|\sqrt{K_i}\nabla^2\varphi(X_k)\sqrt{K_i}\|^2}, \frac{t}{\|\sqrt{K_i}\nabla^2\varphi(X_k)\sqrt{K_i}\|_{\mathrm{op}}}\right\}\right). \tag{259}$$

From Lemma D.5 and since $\|K_i\|_{\mathrm{op}} \le C$ then

$$\|\sqrt{K_i}\nabla^2\varphi(X_k)\sqrt{K_i}\|_{\mathrm{op}} \le \|\sqrt{K_i}\|_{\mathrm{op}}^2\|\nabla^2\varphi(X_k)\|_{\mathrm{op}} \le C. \tag{260}$$

Combining this with $\|\cdot\| \le \sqrt{d}\|\cdot\|_{\mathrm{op}}$ then it follows

$$\mathbb{P}\left(|\langle\sqrt{K_i}\nabla^2\varphi(X_k)\sqrt{K_i}, v_{k+1}^{\otimes 2}\rangle - \mathrm{Tr}(\nabla^2\varphi(X_k)K_i)| > t\right) \le 2\exp\left(-\frac{\min(t^2d^{-1}, t)}{C}\right). \tag{261}$$

From Cauchy–Schwarz and Lemma D.5, $|\mathrm{Tr}(\nabla^2\varphi(X_k)K)| \le d\|\nabla^2\varphi(X_k)\|_{\mathrm{op}}\|K\|_{\mathrm{op}} \le Cd$. Hence, we obtain

$$\mathbb{P}\left(\left|\langle\sqrt{K_i}\nabla^2\varphi(X_k)\sqrt{K_i}, v_{k+1}^{\otimes 2}\rangle\right| > t\right) \le 2\exp\left(-\frac{\min((t-Cd)^2d^{-1}, t-Cd)}{C}\right) \tag{262}$$

Hence, for any $\zeta > 0$, setting $t = d^{1+\zeta}$ then with overwhelming probability

$$\left|\langle\sqrt{K_i}\nabla^2\varphi(X_k)\sqrt{K_i}, v_{k+1}^{\otimes 2}\rangle\right| < d^{1+\zeta}. \tag{263}$$

Now, from (251) and $\mathrm{Proj}_{d\beta}(\langle\nabla^2\varphi(X_k), (\sqrt{K_i}v_{k+1})^{\otimes 2}\rangle) \le d\beta$, we obtain the almost sure bound

$$|(f_i'(\mathrm{Proj}_\beta(g_{k,i})))^2 \cdot \mathrm{Proj}_{d\beta}(\langle\nabla^2\varphi(X_k), (\sqrt{K_i}v_{k+1})^{\otimes 2}\rangle)| \le dC(L(f))^2\beta^{3+2\alpha} \tag{264}$$

Since an almost sure bound holds under expectation, it follows

$$|\mathcal{M}_k^{\mathrm{Hess},1,\beta}| \le \frac{\bar{\gamma}^2}{d} C(L(f))^2\beta^{3+2\alpha}. \tag{265}$$

By Azuma's inequality then for $0 \le k \le Td$

$$\mathbb{P}\left(|\mathcal{M}_k^{\mathrm{Hess},1,\beta}| > t\right) \le 2\exp\left(\frac{-t^2}{2\sum_{i=1}^k Cd^{-2}\beta^{2(3+2\alpha)}}\right) \le 2\exp\left(\frac{-t^2}{TCd^{-1}\beta^{2(3+2\alpha)}}\right) \tag{266}$$

Thus, with overwhelming probability, for every $0 \leq k \leq Td$

$$|\mathcal{M}_k^{\mathrm{Hess},1,\beta}| < d^{-1/2}\beta^{4+2\alpha}. \tag{267}$$

Since taking a union bound from 0 to Td does not modify the overwhelming probability statement, then with overwhelming probability

$$\sup_{0\leq k\leq Td} |\mathcal{M}_k^{\mathrm{Hess},1,\beta}| < d^{-1/2}\beta^{4+2\alpha}. \tag{268}$$

It remains to bound the difference between $\mathcal{M}_k^{\mathrm{Hess},1,}$ and $\mathcal{M}_k^{\mathrm{Hess},1,\beta}$. Once again, the argument is extremely similar to the one used in the proof Proposition D.2. For notational convenience, we write

$$\begin{aligned} F_{k,i} &\stackrel{\mathrm{def}}{=} \frac{\gamma_k^2}{2d^2}(f_i'(g_{k,i}))^2\langle\nabla^2\varphi(X_k), (\sqrt{K_i}v_{k+1})^{\otimes 2}\rangle, \\ F_{k,i}^{\beta} &\stackrel{\mathrm{def}}{=} \frac{\gamma_k^2}{2d^2}(f_i'(\mathrm{Proj}_{\beta}(g_{k,i})))^2 \cdot \mathrm{Proj}_{d\beta}(\langle\nabla^2\varphi(X_k), (\sqrt{K_i}v_{k+1})^{\otimes 2}\rangle). \end{aligned} \tag{269}$$

We will show the following expression is

$$|(F_{k,i} - \mathbb{E}[F_{k,i}]) - (F_{k,i}^{\beta} - \mathbb{E}[F_{k,i}^{\beta}])| \leq |F_{k,i} - F_{k,i}^{\beta}| + |\,\mathbb{E}[F_{k,i} - F_{k,i}^{\beta}]|, \tag{270}$$

exponentially small in d. Using the subgaussian tail bound (220) we previously derived in the proof Proposition D.2 and (263), then

$$\begin{aligned} \mathbb{P}(|F_{k,i} - F_{k,i}^{\beta}| > 0) &\leq \sum_{i=1} p_i(\mathbb{P}(\|(g_{k,i}, \epsilon_{k+1})\| > \beta) + \mathbb{P}(|\langle\nabla^2\varphi(X_k), (\sqrt{K_i}v_{k+1})^{\otimes 2}\rangle| > d\beta)) \\ &\leq 2\exp\left(-\Omega(\beta^2)\right) + 2\exp\left(-\frac{d\min((\beta - C)^2, \beta - C)}{C}\right) \end{aligned} \tag{271}$$

Since we defined $\beta = d^{\zeta}$ for some $\zeta > 0$, then for sufficiently large d, we have $\beta - C \geq \frac{\beta}{2}$. Hence, for sufficiently large d we have

$$\mathbb{P}(|F_{k,i} - F_{k,i}^{\beta}| > 0) \leq C\exp\left(-\Omega(d\min(\beta^2, \beta))\right). \tag{272}$$

which implies $F_{k,i} - F_{k,i}^{\beta} = 0$ with overwhelming probability. It then follows from a union bound that

$$\sum_{k=1}^{Td} |F_{k,I_{k+1}} - F_{k,I_{k+1}}^{\beta}| = 0, \tag{273}$$

with overwhelming probability. We omit the proof that $|\,\mathbb{E}[F_{k,i} - F_{k,i}^{\beta}]|$ is exponentially small in d, since the steps are almost identical to the ones in the proof Proposition D.2 using the growth bound on f', (220) and (263). Thus, the difference $|\mathcal{M}_k^{\mathrm{Hess},1} - \mathcal{M}_k^{\mathrm{Hess},1,\beta}|$ is exponentially small with overwhelming probability.

Finally, combining (254), (257) and (268) with the bounds we derived for the differences $|\mathcal{M}_k^{\mathrm{Hess},q} - \mathcal{M}_k^{\mathrm{Hess},q,\beta}|$ which hold with overwhelming probability for $q = 1, .., 9$, with a similar argument as the end of the proof Proposition D.2, we obtain

$$\sup_{0\leq k\leq Td} |\mathcal{M}_k^{\mathrm{Hess}}| < d^{-\frac{1}{2}+(4+2\alpha)\zeta}. \tag{274}$$

Hence for any arbitrarily small value of $\zeta < \frac{1}{2(4+2\alpha)}$, the result follows. □

D.5.3 Bounds on the Error Term $\mathcal{E}_k^{\mathrm{Hess}}$ of the Hessian Martingale Increments

In this section, we will bound the lower order error terms arising from the Hessian martingale increments defined in (201). We build upon Proposition A.4 in Collins-Woodfin et al. (2024a) and Lemma 11 in Collins-Woodfin & Seroussi (2025). Once again, we provide the proof for the soft label setting as the proof for hard labels follows the same argument with $X^\star = 0$.

Proposition D.4 (Hessian error term). *For $i = 1, 2$, suppose $f_i : \mathbb{R}^3 \to \mathbb{R}$ is an α-pseudo-Lipschitz function with constant $L(f)$ (see Assumption 3). Let the statistic $Z : \mathbb{R}^d \to \mathbb{R}^{4\times 4}$ be defined as in* (202)*. Then, for any $T > 0$, with overwhelming probability*

$$\sup_{0\le t\le T} \sum_{k=0}^{\lfloor (t\wedge\vartheta)d\rfloor - 1} |\mathbb{E}\left[\mathcal{E}_k^{\mathrm{Hess}}(Z(\cdot, z)) \,|\, \mathcal{F}_k\right]| \le C(L(f))^2 d^{-1/2}. \tag{275}$$

Proof. Similarly to in the proofs of Propositions D.2 and D.3, we denote $\varphi \overset{\text{def}}{=} Z_{j\ell}(W, z)$ as the $j\ell$-coordinate of Z for a fixed $z \in \Gamma^2$. Throughout this proof, we will be working with the stopped versions of the stochastic processes involved. For notational simplicity, we omit the dependence on the stopping time ϑ and on φ such that $\mathcal{E}_k^{\mathrm{Hess}} \overset{\text{def}}{=} \mathcal{E}^{\mathrm{Hess}}_{(k/d\wedge\vartheta)d}(\varphi)$. As in the proofs of Propositions D.2 and D.3, for constant $C > 0$, we omit the dependence on i as we could redefine $C = \max\{C_1, C_2\}$. We also write $\nabla\varphi(X_k) \overset{\text{def}}{=} \nabla_X \varphi(X_k)$.

Recall we defined $\Pi_{k,i} = Q_{k,i}Q_{k,i}^\top$ where $Q_{k,i} \in \mathbb{R}^{d\times \ell_i}$ a matrix with orthonormal columns with $\ell_i = \mathrm{rank}(\sqrt{K_i}\hat{X}_k) \le 2$. It is then easy to see that $\|\Pi_{k,i}\|^2 \le 2$. Before proceeding with the proof, we introduce the definition of the nuclear norm

$$\|A\|_* \overset{\text{def}}{=} \sup_{\|B\|_{\mathrm{op}}=1} \langle A, B\rangle. \tag{276}$$

Since $Q_{k,i}$ is a matrix with orthonormal columns, note that $\|\Pi_{k,i}\|_* \le 2$. Recall from (201) the definition of $\mathcal{E}_k^{\mathrm{Hess}}$

$$\begin{aligned}
\mathcal{E}_k^{\mathrm{Hess}}(\varphi) = \frac{\gamma_k^2}{d^2}\sum_{i=1}^2 p_i (f_i'(g_{k,i}))^2 \cdot \Bigg(&- \langle \nabla^2\varphi(X_k), \sqrt{K_i}\Pi_{k,i}\sqrt{K_i}\rangle \\
&+ \langle \nabla^2\varphi(\hat{X}_k), \left(\sqrt{K_i}\Pi_{k,i}v_{k+1}\right)^{\otimes 2}\rangle + 2\langle \nabla^2\varphi(\hat{X}_k), \sqrt{K_i}\Pi_{k,i}v_{k+1}\mu_i^\top\rangle \Bigg) \\
&+ \frac{\gamma_k^2}{d^2}\sum_{i=1}^2 p_i \delta s_{k,i}(f_i'(g_{k,i}))^2 \langle \nabla^2\varphi(\hat{X}_k), (\sqrt{K_i}\Pi_{k,i}v_{k+1} + \mu_i)\otimes \tfrac{X_k}{\|X_k\|}\rangle \\
&+ \frac{\gamma_k^2}{2d^2}\sum_{i=1}^2 p_i \delta^2 s_{k,i}^2 (f_i'(g_{k,i}))^2 \langle \nabla^2\varphi(\hat{X}_k), \tfrac{X_k^{\otimes 2}}{\|X_k\|^2}\rangle \\
&+ \frac{\gamma_k^2}{2d^2}\langle \nabla^2\varphi(X_k), (\lambda X_k)^{\otimes 2}\rangle \\
&+ \frac{\gamma_k^2}{d^2}\sum_{i=1}^2 p_i f_i'(g_{k,i}) \langle \nabla^2\varphi(X_k), (\sqrt{K_i}\Pi_{k,i}v_{k+1} + \mu_i + \delta s_{k,i}\tfrac{X_k}{\|X_k\|})\otimes \lambda X_k\rangle.
\end{aligned} \tag{277}$$

Starting with the first term, we have

$$\begin{aligned}
|(f_i'(g_{k,i}))^2 \cdot \langle \nabla^2\varphi(X_k), \sqrt{K_i}\Pi_k\sqrt{K_i}\rangle| &\le |f_i'(g_{k,i})|^2 \|\nabla^2\varphi(X_k)\| \|\sqrt{K_i}\Pi_k\sqrt{K_i}\| \\
&\le |f_i'(g_{k,i})|^2 \|\nabla^2\varphi(X_k)\| \|K_i\|_{\mathrm{op}} \|\Pi_k\|_* \\
&\le 2|f_i'(g_{k,i})|^2 \|\nabla^2\varphi(X_k)\| \|K_i\|_{\mathrm{op}}.
\end{aligned} \tag{278}$$

From Lemma D.5, we have

$$\|\nabla^2\varphi(X_k)\| = \|\nabla^2 Z_{j\ell}(W_k, z)\| \le \sup_{z\in\Gamma} \|\nabla^2 Z_{j\ell}(W_k, z)\| \le \|\nabla^2 Z(W_k, z)\|_\Gamma \le C\sqrt{d}. \tag{279}$$

By Assumption 1, $\|K_i\|_{\mathrm{op}} \le C$ and from Lemma D.3, we have $\mathbb{E}[|f_i'(g_{k,i})|^2 \mid \mathcal{F}_k] \le CL(f)^2(1 + (\|K_i\|_{\mathrm{op}}^{1/2} + \|\mu_i\|)\|\hat{X}_k\| + \delta\|X_k\| + \eta)^{\max(1,2\alpha)}$. Since we are working with the stopped processes, it then follows

$$\begin{aligned}
&\mathbb{E}[|f_i'(g_{k,i})^2 \cdot \langle \nabla^2\varphi(X_k), \sqrt{K_i}\Pi_k\sqrt{K_i}\rangle| \mid \mathcal{F}_k] \\
&\qquad \le C\sqrt{d}L(f)^2(1 + (\|K_i\|_{\mathrm{op}}^{1/2} + \|\mu_i\|)\|\hat{X}_k\| + \delta\|X_k\| + \eta)^{\max(1,2\alpha)} = \mathcal{O}(d^{1/2}).
\end{aligned} \tag{280}$$

For the second term, we obtain

$$|f_i'(g_{k,i})^2\langle\nabla^2\varphi(X_k),\left(\sqrt{K_i}\Pi_{k,i}v_{k+1}\right)^{\otimes 2}\rangle| \le |f_i'(g_{k,i})|^2\|\nabla^2\varphi(X_k)\|\|\sqrt{K_i}\Pi_{k,i}v_{k+1}\|^2 \le |f_i'(g_{k,i})|^2\|\nabla^2\varphi(X_k)\|\|K_i\|_{\mathrm{op}}\|\Pi_{k,i}v_{k+1}\|^2 \tag{281}$$

Once again, from Lemma D.3, we have $\mathbb{E}[|f_i'(g_{k,i})|^4 \mid \mathcal{F}_k] \le CL(f)^4(1+(\|K\|_{\mathrm{op}}^{1/2}+\|\mu_i\|)\|\hat{X}_k\|+\delta\|X_k\|+\eta)^{\max(1,4\alpha)}$. Now, since $Q_{k,i}$ has $\ell \le 2$ orthonormal columns then $\Pi_{k,i}$ is a projection matrix and since $Q_{k,i}v_{k+1} \sim \mathcal{N}(0, I_\ell)$ it follows $\mathbb{E}[\|\Pi_{k,i}v_{k+1}\|^4] = \mathbb{E}[(\|Q_kv_{k+1}\|^2)^2] = \mathbb{E}[(\chi_\ell^2)^2] \le 8$. Hence, by (279) and (281)

$$\mathbb{E}[|f_i'(g_{k,i})^2\langle\nabla^2\varphi(X_k),\left(\sqrt{K_i}\Pi_{k,i}v_{k+1}\right)^{\otimes 2}\rangle| \mid \mathcal{F}_k] \le C\sqrt{d}L(f)^2(1+(\|K_i\|_{\mathrm{op}}^{1/2}+\|\mu_i\|)\|\hat{X}_k\|+\delta\|X_k\|+\eta)^{\max(\frac{1}{2},2\alpha)} = \mathcal{O}(d^{1/2}). \tag{282}$$

Using a similar argument, we obtain for the third term

$$\mathbb{E}[|f_i'(g_{k,i})^2\cdot\langle\nabla^2\varphi(\hat{X}_k),\sqrt{K_i}\Pi_{k,i}v_{k+1}\mu_i^\top\rangle| \mid \mathcal{F}_k] = \mathcal{O}(d^{1/2}). \tag{283}$$

Similarly, since $|s_k| \le 1$ for the fourth term

$$|\delta s_{k,i}(f_i'(g_{k,i}))^2\langle\nabla^2\varphi(\hat{X}_k),(\sqrt{K_i}\Pi_{k,i}v_{k+1}+\mu_i)\otimes\tfrac{X_k}{\|X_k\|}\rangle| \le |f_i'(g_{k,i})|^2\|\nabla^2\varphi(X_k)\|(\|K\|_{\mathrm{op}}^{1/2}\|\Pi_kv_{k+1}\|+\|\mu_i\|) \tag{284}$$

Using a similar argument to what we just did, from Cauchy–Schwarz, Lemma D.3, (279), (284) and since $\mathbb{E}[\|\Pi_kv_{k+1}\|^2] = \mathbb{E}[\chi_\ell^2] \le 2$ we obtain

$$\mathbb{E}[|s_k^2f_i'(g_{k,i})^2\cdot\langle\nabla^2\varphi(X_k),\sqrt{K_i}\Pi_kv_{k+1}\otimes\tfrac{X_k}{\|X_k\|}\rangle| \mid \mathcal{F}_k] \le C\sqrt{d}L(f)^2(1+(\|K_i\|_{\mathrm{op}}^{1/2}+\|\mu_i\|)\|\hat{X}_k\|+\delta\|X_k\|+\eta)^{\max(\frac{1}{2},2\alpha)} = \mathcal{O}(d^{1/2}). \tag{285}$$

For the fifth term, since $|s_{k,i}| \le 1$ we have

$$|s_{k,i}^2(f_i'(g_{k,i}))^2\langle\nabla^2\varphi(\hat{X}_k),\tfrac{X_k^{\otimes 2}}{\|X_k\|^2}\rangle| \le |f_i'(g_{k,i})|^2\|\nabla^2\varphi(X_k)\|. \tag{286}$$

Hence, it follows from Lemma D.3 and (279)

$$\mathbb{E}[|s_{k,i}^2(f_i'(g_{k,i}))^2\langle\nabla^2\varphi(\hat{X}_k),\tfrac{X_k^{\otimes 2}}{\|X_k\|^2}\rangle| \mid \mathcal{F}_k] \le C\sqrt{d}L(f)^2(1+(\|K_i\|_{\mathrm{op}}^{1/2}+\|\mu_i\|)\|\hat{X}_k\|+\delta\|X_k\|+\eta)^{\max(1,2\alpha)} = \mathcal{O}(d^{1/2}). \tag{287}$$

For the firth term, from (279) and recalling that we are working with the stopped processes we have

$$\mathbb{E}[|\langle\nabla^2\varphi(X_k),(\lambda X_k)^{\otimes 2}\rangle| \mid \mathcal{F}_k] \le C\sqrt{d}. \tag{288}$$

Finally, for the last term, since $|s_k| \le 1$ we have

$$|f_i'(g_{k,i})\langle\nabla^2\varphi(X_k),(\sqrt{K_i}\Pi_{k,i}v_{k+1}+\mu_i+\delta s_{k,i}\tfrac{X_k}{\|X_k\|})\otimes\lambda X_k\rangle| \le \lambda|f_i'(g_{k,i})|\|\nabla^2\varphi(X_k)\|\|X_k\|(\|K_i\|_{\mathrm{op}}^{1/2}\|\Pi_{k,i}v_{k+1}\|+\|\mu_i\|+\delta) \tag{289}$$

Thus, using a similar argument as above, from Lemma D.3, (279) and Cauchy–Schwarz we have

$$\mathbb{E}[|f_i'(g_{k,i})\cdot\langle\nabla^2\varphi(X_k),(\sqrt{K_i}\Pi_kv_{k+1}+\delta s_k\tfrac{X_k}{\|X_k\|})\otimes\lambda X_k\rangle| \mid \mathcal{F}_k] \le C\sqrt{d}L(f)(1+(\|K_i\|_{\mathrm{op}}^{1/2}+\|\mu_i\|)\|\hat{X}_k\|+\delta\|X_k\|+\eta)^{\max(\frac{1}{2},\alpha)} = \mathcal{O}(d^{1/2}). \tag{290}$$

Since $\frac{\gamma_k^2}{d^2} \le \frac{\bar{\gamma}^2}{d^2}$ for all $k \ge 0$ and these bounds hold uniformly in i, the result follows from summing over i, from Jensen's inequality, the triangle inequality and noting that the summation contributes a factor Td. □

E ℓ_2-ADVERSARIAL TRAINING V.S. ℓ_2-REGULARIZATION

In this section, we provide a detailed description of the relationship between ℓ_2-adversarial least squares and ℓ_2-regularized least squares. From (B.1), AdvHSGD for ℓ_2-adversarial least squares with $\lambda = 0$ on a single data class $a \sim N(0, K)$ (i.e. $p_1 = 1$, $K_1 = K$, $\mu_1 = 0$) with soft labels is the solution to the SDE

$$\mathrm{d}\mathcal{X}_t = -\gamma^{\mathrm{eff}}(t)\left(\nabla\mathcal{R}(\mathcal{X}_t) + \lambda^{\mathrm{eff}}(t)\mathcal{X}_t\right)\mathrm{d}t + \gamma(t)\sqrt{\frac{2}{d}\mathcal{R}^{\mathrm{Adv}}(\mathcal{X}_t)K}\,\mathrm{d}B_t, \tag{291}$$

where we define

$$\gamma^{\mathrm{eff}}(t) = \gamma(t)\left(1 + \delta\sqrt{\frac{2}{\pi}}\frac{\|\mathcal{X}_t\|}{\sqrt{2\mathcal{R}(\mathcal{X}_t)}}\right) \quad \text{and} \quad \lambda^{\mathrm{eff}}(t) = \frac{\delta^2 + \delta\sqrt{\frac{2}{\pi}}\frac{\sqrt{2\mathcal{R}(\mathcal{X}_t)}}{\|\mathcal{X}_t\|}}{1 + \delta\sqrt{\frac{2}{\pi}}\frac{\|\mathcal{X}_t\|}{\sqrt{2\mathcal{R}(\mathcal{X}_t)}}}.$$

Here $\gamma^{\mathrm{eff}}(t)$ and $\lambda^{\mathrm{eff}}(t)$ correspond to the effective learning rate and regularization induced by adversarial training.

Effective learning rate γ^{eff} and regularization λ^{eff} The schedules of γ^{eff} and λ^{eff} are entirely determined by the path of the ratio $q(\mathcal{X}_t) \overset{\mathrm{def}}{=} \left(\|\mathcal{X}_t\|^2/(2\mathcal{R}(\mathcal{X}_t))\right)^{1/2}$. The effective learning rate is a strictly increasing function of $q(\mathcal{X}_t)$, whereas the effective regularization parameter is a strictly decreasing function of $q(\mathcal{X}_t)$. We may reexpress AdvHSGD as follows

$$\mathrm{d}\mathcal{X}_t = -\gamma^{\mathrm{eff}}(t)\left(\nabla\mathcal{R}(\mathcal{X}_t) + \lambda^{\mathrm{eff}}(t)\mathcal{X}_t\right)\mathrm{d}t + \gamma(t)\sqrt{\frac{2}{d}G(q(\mathcal{X}_t))\mathcal{R}(\mathcal{X}_t)K}\,\mathrm{d}B_t, \tag{292}$$

where $G(q) \overset{\mathrm{def}}{=} 1 + 2\delta\sqrt{\frac{2}{\pi}}q + \delta^2q^2$. Consider $\gamma(t) \equiv \gamma$ for some $\gamma > 0$. When $q(\mathcal{X}_t)$ is large (i.e. $\mathcal{R}(\mathcal{X}_t) \ll \|\mathcal{X}_t\|^2$), we have

$$\gamma^{\mathrm{eff}}(t)\lambda^{\mathrm{eff}}(t) = \gamma\left(\delta^2 + \delta\sqrt{\frac{2}{\pi}}\frac{1}{q(\mathcal{X}_t)}\right) \approx \gamma\delta^2,$$

and $\lambda^{\mathrm{eff}}(t) \to 0$ as $q(\mathcal{X}_t) \to \infty$. Hence, when $q(\mathcal{X}_t)$ is large, $\gamma^{\mathrm{eff}}(t)$ grows linearly in $q(\mathcal{X}_t)$ which implies the descent term of (292) favors the least squares component relative to the ℓ_2-regularization component and acts similarly to least squares with learning rate $\gamma^{\mathrm{eff}}(t)$. When $q(\mathcal{X}_t) \ll 1$ (i.e. $\|\mathcal{X}_t\|^2 \ll \mathcal{R}(\mathcal{X}_t)$), the effective learning rate $\gamma^{\mathrm{eff}}(t) \approx \gamma$ and

$$\gamma^{\mathrm{eff}}(t)\lambda^{\mathrm{eff}}(t) = \gamma\left(\delta^2 + \delta\sqrt{\frac{2}{\pi}}\frac{1}{q(\mathcal{X}_t)}\right) \approx \gamma\delta\sqrt{\frac{2}{\pi}}\frac{1}{q(\mathcal{X}_t)}.$$

Thus, for small $q(\mathcal{X}_t) \ll 1$, adversarial training favors the ℓ_2-regularization component and acts similarly to least squares with constant learning rate γ and large regularization parameter. This is further emphasized by the following decomposition of $\nabla\mathcal{R}^{\mathrm{Adv}}$:

$$\nabla\mathcal{R}^{\mathrm{adv}}(\mathcal{X}_t) = \nabla\mathcal{R}(\mathcal{X}_t) + \delta^2\mathcal{X}_t + \delta\sqrt{\frac{2}{\pi}}\left(q(\mathcal{X}_t)\nabla\mathcal{R}(\mathcal{X}_t) + \frac{1}{q(\mathcal{X}_t)}\mathcal{X}_t\right). \tag{293}$$

Furthermore, $q(\mathcal{X}_t) \ll 1$ implies $G(q(\mathcal{X}_t)) \approx 1$ from which it follows that the diffusion coefficient of (292) acts similarly to the diffusion coefficient of Homogenized SGD (Collins-Woodfin et al., 2024a; Collins-Woodfin & Paquette, 2024; Paquette et al., 2022):

$$\gamma(t)\sqrt{\frac{2}{d}G(q(\mathcal{X}_t))\mathcal{R}(\mathcal{X}_t)K}\,\mathrm{d}B_t \approx \gamma(t)\sqrt{\frac{2}{d}\mathcal{R}(\mathcal{X}_t)K}\,\mathrm{d}B_t. \tag{294}$$

Relation to ridge regularization It has been shown that, under suitable assumptions on the data and adversarial radius δ, the minimizer of ℓ_2-adversarial least squares can coincide with the ridge regression solution (Ribeiro et al., 2023; Xing et al., 2021). We emphasize this connection by reexpressing the drift term of AdvHSGD as follows

$$-\gamma(t)\nabla\mathcal{R}^{\mathrm{adv}}(\mathcal{X}_t) = -\gamma^{\mathrm{eff}}(t)\left(\left(K + \lambda^{\mathrm{eff}}(t)I_d\right)\mathcal{X}_t - KX^\star\right). \tag{295}$$

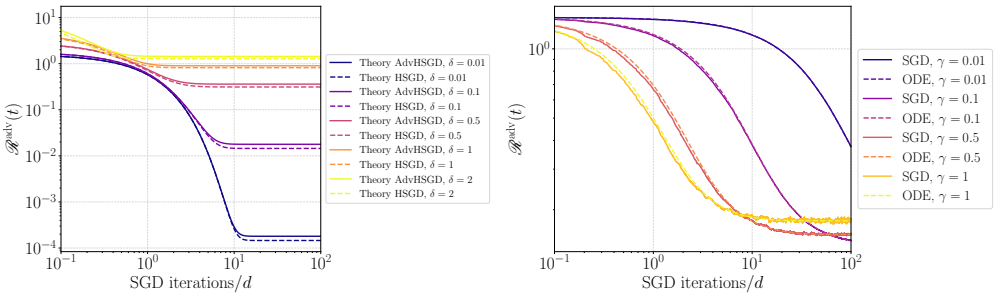

Figure 3: **Comparison of ℓ_2-adversarial least squares and ℓ_2-regularized least squares with adaptive learning rate and regularization.** The left plot compares the paths of the det. equivs. of $\mathcal{R}^{\mathrm{Adv}}$ for AdvHSGD and HSGD and confirms Theorem 3.1. The right plot compares the path of $\mathcal{R}^{\mathrm{Adv}}(X_k)$ for SGD with adaptive learning rate $\gamma^{\mathrm{Reg}}(t)$ and regularization $\lambda^{\mathrm{Reg}}(t)$ versus the det. equiv. computed from (301) and (302) and shows close agreement between them for a variety of constant $\gamma(t) \equiv \gamma$. The ℓ_2-adversarial risk is computed using $\mathcal{R}^{\mathrm{Adv}}(X_k) = \mathcal{R}(X_k) + \delta\sqrt{\frac{2}{\pi}}\sqrt{2\mathcal{R}(X_k)}\|X_k\| + \frac{\delta^2}{2}\|X_k\|^2$. See Appendix J for simulation details.

Conditional on the parameters $\gamma^{\mathrm{eff}}(t)$ and $\lambda^{\mathrm{eff}}(t)$, the drift is that of ridge regression with adaptive learning rate and regularization. This connection becomes explicit by projecting onto the eigenbasis of K. Recall (λ_j, ω_j) the eigenvalue-eigenvectors pairs of K. Then by Itô's formula

$$\mathrm{d}\langle \mathcal{X}_t, \omega_j\rangle = -\gamma^{\mathrm{eff}}(t)\left(\left(\lambda_j + \lambda^{\mathrm{eff}}(t)\right)\langle \mathcal{X}_t, \omega_j\rangle - \lambda_j\langle X^\star, \omega_j\rangle\right)\mathrm{d}t + \mathrm{d}\mathcal{M}_j(\mathcal{X}_t), \tag{296}$$

where $\mathrm{d}\mathcal{M}_j(\mathcal{X}_t)$ is a martingale term associated with eigenvector ω_j which vanishes as $d \to \infty$ (see Theorem B.4). Let $\rho_j(t)$, $\tilde{\lambda}^{\mathrm{eff}}(t)$ $q(t)$ denote the det. equivs. of $\varphi(\mathcal{X}_t)$, $\lambda^{\mathrm{eff}}(t)$ and $q(\mathcal{X}_t)$ in the sense of Theorem B.4. In Theorem G.1, assuming $q(t) \to q_\infty$, we show that as $t \to \infty$

$$\rho_j(t) \to \frac{\lambda_j}{\lambda_j + \tilde{\lambda}^{\mathrm{eff}}(q_\infty)}\langle X^\star, \omega_j\rangle. \tag{297}$$

Hence, whenever $q(t) \underset{t\to\infty}{\longrightarrow} q_\infty$ and as $d \to \infty$, SGD converges to the ridge optimum associated with $\tilde{\lambda}^{\mathrm{eff}}(q_\infty)$:

$$X_\infty = \left(K + \tilde{\lambda}^{\mathrm{eff}}(q_\infty)\right)^{-1} KX^\star.$$

In particular, directions with $\lambda_j \gg \tilde{\lambda}^{\mathrm{eff}}(q_\infty)$ are nearly identical to $X^\star$, whereas directions with $\lambda_j \ll \tilde{\lambda}^{\mathrm{eff}}(q_\infty)$ are strongly attenuated.

Relation to Homogenized SGD As originally shown in Collins-Woodfin & Paquette (2024), Homogenized SGD (HSGD) is a similar SDE to AdvHSGD for ℓ_2-regularized least squares with parameter $\lambda > 0$ and is the solution to the following SDE

$$\mathrm{d}\mathcal{X}_t^{\mathrm{Reg}} = -\gamma(t)\left(\nabla\mathcal{R}(\mathcal{X}_t^{\mathrm{Reg}}) + \lambda\mathcal{X}_t^{\mathrm{Reg}}\right)\mathrm{d}t + \gamma(t)\sqrt{\frac{2}{d}\mathcal{R}(\mathcal{X}_t^{\mathrm{Reg}})K}\,\mathrm{d}B_t. \tag{298}$$

Now, introduce the adaptive learning rate and regularization parameters

$$\gamma^{\mathrm{Reg}}(t) = \gamma(t)\left(1 + \delta\sqrt{\frac{2}{\pi}}\frac{\|\mathcal{X}_t^{\mathrm{Reg}}\|}{\sqrt{2\mathcal{R}(\mathcal{X}_t^{\mathrm{Reg}})}}\right) \quad \text{and} \quad \lambda^{\mathrm{Reg}}(t) = \frac{\delta^2 + \delta\sqrt{\frac{2}{\pi}}\frac{\sqrt{2\mathcal{R}(\mathcal{X}_t^{\mathrm{Reg}})}}{\|\mathcal{X}_t^{\mathrm{Reg}}\|}}{1 + \delta\sqrt{\frac{2}{\pi}}\frac{\|\mathcal{X}_t^{\mathrm{Reg}}\|}{\sqrt{2\mathcal{R}(\mathcal{X}_t^{\mathrm{Reg}})}}}. \tag{299}$$

Then the SDE for ℓ_2-regularized least squares amounts to

$$\begin{aligned}\mathrm{d}\mathcal{X}_t^{\mathrm{Reg}} = &-\gamma^{\mathrm{Reg}}(t)\left(\nabla\mathcal{R}(\mathcal{X}_t^{\mathrm{Reg}}) + \lambda^{\mathrm{Reg}}(t)\mathcal{X}_t^{\mathrm{Reg}}\right)\mathrm{d}t \\ &+ \gamma(t)\sqrt{\frac{2}{d}\left(\mathcal{R}(\mathcal{X}_t^{\mathrm{Reg}}) + \delta\sqrt{\frac{2}{\pi}}\|\mathcal{X}_t^{\mathrm{Reg}}\|\sqrt{2\mathcal{R}(\mathcal{X}_t^{\mathrm{Reg}})} + \frac{\delta^2}{\pi}\|\mathcal{X}_t^{\mathrm{Reg}}\|^2\right)K}\,\mathrm{d}B_t.\end{aligned} \tag{300}$$

Let $\mathcal{R}(t)$ and $\Phi(t)$ be the solutions of the coupled Volterra equations

$$\begin{aligned}\mathcal{R}(t) = \frac{1}{2}\left\langle K, \left(e^{-K\tilde{\Gamma}^{\mathrm{Reg}}(t)-\Lambda^{\mathrm{Reg}}(t)}\left(X_0 + \left(\int_0^t K\mathfrak{g}(u)e^{K\tilde{\Gamma}^{\mathrm{Reg}}(u)+\Lambda^{\mathrm{Reg}}(u)}\,\mathrm{d}u\right)X^\star\right) - X^\star\right)^{\otimes 2}\right\rangle \\ + \frac{\eta^2}{2} + \frac{1}{d}\int_0^t \gamma^2(s)\left(\mathcal{R}(s) + \delta\sqrt{\frac{2}{\pi}}\sqrt{2\mathcal{R}(s)\Phi(s)} + \frac{\delta^2}{\pi}\Phi(s)\right) \\ \times \mathrm{Tr}\left(K^2 e^{-2K\left(\tilde{\Gamma}^{\mathrm{Reg}}(t)-\tilde{\Gamma}^{\mathrm{Reg}}(s)\right)-2\left(\Lambda^{\mathrm{Reg}}(t)-\Lambda^{\mathrm{Reg}}(s)\right)}\right)\mathrm{d}s.\end{aligned} \tag{301}$$

and

$$\begin{aligned}\Phi(t) = \left\|e^{-K\tilde{\Gamma}^{\mathrm{Reg}}(t)-\Lambda^{\mathrm{Reg}}(t)}\left(X_0 + \left(\int_0^t K\mathfrak{g}(u)e^{K\tilde{\Gamma}^{\mathrm{Reg}}(u)+\Lambda^{\mathrm{Reg}}(u)}\,\mathrm{d}u\right)X^\star\right)\right\|^2 \\ + \frac{2}{d}\int_0^t \gamma^2(s)\left(\mathcal{R}(s) + \delta\sqrt{\frac{2}{\pi}}\sqrt{2\mathcal{R}(s)\Phi(s)} + \frac{\delta^2}{\pi}\Phi(s)\right) \\ \times \mathrm{Tr}\left(Ke^{-2K\left(\tilde{\Gamma}^{\mathrm{Reg}}(t)-\tilde{\Gamma}^{\mathrm{Reg}}(s)\right)-2\left(\Lambda^{\mathrm{Reg}}(t)-\Lambda^{\mathrm{Reg}}(s)\right)}\right)\mathrm{d}s.\end{aligned} \tag{302}$$

where

$$\mathfrak{g}(t) = \gamma(t)\left(1 + \delta\sqrt{\frac{2}{\pi}}\sqrt{\frac{\Phi(t)}{2\mathcal{R}(t)}}\right) \quad \text{and} \quad \mathfrak{l}(t) = \frac{\delta^2 + \delta\sqrt{\frac{2}{\pi}}\sqrt{\frac{2\mathcal{R}(t)}{\Phi(t)}}}{1 + \delta\sqrt{\frac{2}{\pi}}\sqrt{\frac{\Phi(t)}{2\mathcal{R}(t)}}}. \tag{303}$$

and

$$\tilde{\Gamma}^{\mathrm{Reg}}(t) = \int_0^t \mathfrak{g}(s)\,\mathrm{d}s \quad \text{and} \quad \Lambda^{\mathrm{Reg}}(t) = \int_0^t \mathfrak{g}(s)\mathfrak{l}(s)\,\mathrm{d}s. \tag{304}$$

The following consequence of Theorem 2.1 and Remark 3.1 identifies $\mathcal{R}$ and Φ as the det. equivs. of the least squares risk and squared norm along (300), and along SGD on ℓ_2-regularized least squares run with the same schedule.

Corollary E.1. *Let X_k^{Reg} be the iterates of SGD* (5) *on ℓ_2-regularized least squares with $a \sim \mathcal{N}(0, K)$, where γ_k and λ_k are the functions defining γ^{Reg} and λ^{Reg} evaluated at X_k^{Reg}, and let $\mathcal{X}_t^{\mathrm{Reg}}$ solve* (300)*. For $\omega, M > 0$, let Θ^{Reg} be the first time $\mathcal{R}(t)$ or $\Phi(t)$ leaves $[\omega, M]$. Then for any $T > 0$ and $\varepsilon \in (0, \frac{1}{2})$ there exists $C = C(\|K\|_{\mathrm{op}}, \bar{\gamma}, T, \omega, M)$, independent of d, such that w.o.p.*

$$\sup_{0\le t\le T\wedge\Theta^{\mathrm{Reg}}}\left(\left|\mathcal{R}(X_{\lfloor td\rfloor}^{\mathrm{Reg}}) - \mathcal{R}(t)\right| + \left|\|X_{\lfloor td\rfloor}^{\mathrm{Reg}}\|^2 - \Phi(t)\right| + \left|\mathcal{R}(\mathcal{X}_t^{\mathrm{Reg}}) - \mathcal{R}(t)\right| + \left|\|\mathcal{X}_t^{\mathrm{Reg}}\|^2 - \Phi(t)\right|\right) \le Cd^{-\varepsilon}.$$

Proof. With $\delta = 0$, a single class and soft labels, (2) is ℓ_2-regularized least squares and AdvHSGD (8) with the schedule $(\gamma^{\mathrm{Reg}}, \lambda^{\mathrm{Reg}})$ is (300). For $\delta = 0$, h is smooth on $\mathbb{R}^{4\times 4}$, so the set $\mathcal{U}$, where $\mathcal{R}(X)$ and $\|X\|^2$ are positive, is dictated by the schedule alone: γ^{Reg} and λ^{Reg} are functions of $\mathcal{R}(X)$ and $\|X\|^2$) which is pseudo-Lipschitz on $\mathcal{U}$ but not at its boundary. Hence Theorem 2.1 applies by Remark 3.1, and the det. equivs. of $\gamma_{\lfloor td\rfloor}$ and $\lambda_{\lfloor td\rfloor}$ are $\mathfrak{g}(t)$ and $\mathfrak{l}(t)$ in (303). The det. equivs. of $\mathcal{R}$ and $\|X\|^2$ then follow from (26) with $\delta = 0$ and schedule $(\mathfrak{g}, \mathfrak{l})$, exactly as for AdvHSGD in Appendix I.1, which gives (301) and (302). Finally, on $[0, \Theta^{\mathrm{Reg}}]$ the det. equiv. of $\widehat{B}$ lies in $\mathcal{U}_\omega$, since its distance to the complement of $\mathcal{U}$ is $\min(\mathcal{R}(t), \Phi(t)) \ge \omega$, and $\mathscr{N}(t) = \Phi(t) + \|X^\star\|^2 \le M + C$ by Assumption 4, so $\Theta^{\mathrm{Reg}} \le \Theta_{M+C}$ in the notation of Theorem 2.1. □

We are now ready to prove Theorem 3.1.

Proof. For notational simplicity, we write $D(t) \stackrel{\mathrm{def}}{=} |\mathscr{R}(t) - \mathcal{R}(t)| + \left|\widehat{\mathscr{B}}_{44}(t) - \Phi(t)\right|$ and define

$$q_{\mathrm{Adv}}(t) \stackrel{\mathrm{def}}{=} \sqrt{\frac{\widehat{\mathscr{B}}_{44}(t)}{2\mathscr{R}(t)}} \quad \text{and} \quad q_{\mathrm{Reg}}(t) \stackrel{\mathrm{def}}{=} \sqrt{\frac{\Phi(t)}{2\mathcal{R}(t)}}.$$

Throughout this proof, we work with $t \in [0, T \wedge \Theta]$, so that from Θ, we have $\omega \leq \mathscr{R}(t), \mathcal{R}(t), \widehat{\mathscr{B}}_{44}(t), \Phi(t) \leq M$. This is analogous to the stopping time of Theorem 2.1. Hence

$$q_- \stackrel{\text{def}}{=} \sqrt{\tfrac{\omega}{2M}} \leq q_{\text{Adv}}(t), q_{\text{Reg}}(t) \leq \sqrt{\tfrac{M}{2\omega}} \stackrel{\text{def}}{=} q_+.$$

for all $t \in [0, T]$. We also define for notational simplicity

$$A_{\text{Adv}}(t) \stackrel{\text{def}}{=} 1 + \delta\sqrt{\frac{2}{\pi}} q_{\text{Adv}}(t), \quad B_{\text{Adv}}(t) \stackrel{\text{def}}{=} \delta^2 + \delta\sqrt{\frac{2}{\pi}} \frac{1}{q_{\text{Adv}}(t)}$$
$$A_{\text{Reg}}(t) \stackrel{\text{def}}{=} 1 + \delta\sqrt{\frac{2}{\pi}} q_{\text{Reg}}(t) \quad \text{and} \quad B_{\text{Reg}}(t) \stackrel{\text{def}}{=} \delta^2 + \delta\sqrt{\frac{2}{\pi}} \frac{1}{q_{\text{Reg}}(t)}.$$

We first show that $\tilde{\Gamma}$, $\tilde{\Gamma}^{\text{Reg}}$, Λ and Λ^{Reg} are Lipschitz. Since the map $(R, N) \mapsto \sqrt{\frac{N}{2R}}$ is smooth on compact subsets of $(0, \infty) \times (0, \infty)$, the assumptions imply

$$|q_{\text{Adv}}(t) - q_{\text{Reg}}(t)| \leq C \left(|\mathscr{R}(t) - \mathcal{R}(t)| + |\widehat{\mathscr{B}}_{44}(t) - \Phi(t)| \right) = CD(t).$$

Since $q_{\text{Adv}}, q_{\text{Reg}} \in [q_-, q_+]$, the map $q \mapsto q^{-1}$ is also Lipschitz on the relevant interval. Hence

$$|A_{\text{Adv}}(t) - A_{\text{Reg}}(t)| + |B_{\text{Adv}}(t) - B_{\text{Reg}}(t)| \leq CD(t). \tag{305}$$

Therefore

$$\left| \tilde{\Gamma}(t) - \tilde{\Gamma}^{\text{Reg}}(t) \right| \leq C(\bar{\gamma}) \int_0^t D(s)\, \mathrm{d}s \quad \text{and} \quad \left| \Lambda(t) - \Lambda^{\text{Reg}}(t) \right| \leq C(\bar{\gamma}) \int_0^t D(s)\, \mathrm{d}s. \tag{306}$$

More generally, for $0 \leq s \leq t$,

$$\begin{aligned} \left| \left(\tilde{\Gamma}(t) - \tilde{\Gamma}(s) \right) - \left(\tilde{\Gamma}^{\text{Reg}}(t) - \tilde{\Gamma}^{\text{Reg}}(s) \right) \right| + \left| (\Lambda(t) - \Lambda(s)) - \left(\Lambda^{\text{Reg}}(t) - \Lambda^{\text{Reg}}(s) \right) \right| \\ \leq C(\bar{\gamma}) \int_s^t D(u)\, \mathrm{d}u. \end{aligned} \tag{307}$$

For $p = 1, 2$, define the kernels

$$\mathcal{K}_p^{\text{Adv}}(t, s) \stackrel{\text{def}}{=} \text{Tr} \left(K^p e^{-2K(\tilde{\Gamma}(t) - \tilde{\Gamma}(s)) - 2(\Lambda(t) - \Lambda(s))} \right),$$

and

$$\mathcal{K}_p^{\text{Reg}}(t, s) \stackrel{\text{def}}{=} \text{Tr} \left(K^p e^{-2K(\tilde{\Gamma}^{\text{Reg}}(t) - \tilde{\Gamma}^{\text{Reg}}(s)) - 2(\Lambda^{\text{Reg}}(t) - \Lambda^{\text{Reg}}(s))} \right).$$

From the eigendecomposition of K, we obtain the bounds

$$\mathcal{K}_p^{\text{Adv}}(t, s) \leq \text{Tr}(K^p), \qquad \mathcal{K}_p^{\text{Reg}}(t, s) \leq \text{Tr}(K^p),$$

which imply

$$\frac{1}{d} \mathcal{K}_p^{\text{Adv}}(t, s) + \frac{1}{d} \mathcal{K}_p^{\text{Reg}}(t, s) \leq C(\|K\|_{\text{op}}). \tag{308}$$

Moreover, since $|e^{-x} - e^{-y}| \leq |x - y|$ for all $x, y \geq 0$ and using (307), we obtain

$$\frac{1}{d} \left| \mathcal{K}_p^{\text{Adv}}(t, s) - \mathcal{K}_p^{\text{Reg}}(t, s) \right| \leq C(\|K\|_{\text{op}}, \bar{\gamma}) \int_s^t D(u)\, \mathrm{d}u, \quad \text{for} \quad p = 1, 2. \tag{309}$$

Now, let $\mathcal{R}_{\text{Reg}}(t) \stackrel{\text{def}}{=} \mathcal{R}(t) + \delta\sqrt{\frac{2}{\pi}} \sqrt{2\mathcal{R}(t)\Phi(t)} + \frac{\delta^2}{\pi}\Phi(t)$. Then we obtain adding and subtracting $\frac{\delta^2}{\pi}\widehat{\mathscr{B}}_{44}(t)$

$$\begin{aligned} \mathscr{R}^{\text{Adv}}(t) - \mathcal{R}_{\text{Reg}}(t) = \mathscr{R}(t) - \mathcal{R}(t) + \delta\sqrt{\frac{2}{\pi}} \left(\sqrt{2\mathscr{R}(t)\widehat{\mathscr{B}}_{44}(t)} - \sqrt{2\mathcal{R}(t)\Phi(t)} \right) \\ + \frac{\delta^2}{\pi} \left(\widehat{\mathscr{B}}_{44}(t) - \Phi(t) \right) + \delta^2 \left(\frac{1}{2} - \frac{1}{\pi} \right) \widehat{\mathscr{B}}_{44}(t) \end{aligned}$$

Since the map $(R, N) \mapsto \sqrt{2RN}$ is smooth on compact subsets of $(0,\infty)\times(0,\infty)$, the boundedness assumption implies

$$\left|\sqrt{2\mathscr{R}(t)\widehat{\mathscr{B}}_{44}(t)} - \sqrt{2\mathcal{R}(t)\Phi(t)}\right| \le CD(t).$$

Therefore

$$\left|\mathscr{R}^{\mathrm{Adv}}(t) - \mathcal{R}_{\mathrm{Reg}}(t)\right| \le CD(t) + \delta^2\left(\frac{1}{2} - \frac{1}{\pi}\right)\widehat{\mathscr{B}}_{44}(t). \tag{310}$$

For simplicity of notation, we define the following terms which appear in the first term of the det. equivs.:

$$m_{\mathrm{Adv}}(t) \stackrel{\mathrm{def}}{=} e^{-K\tilde{\Gamma}(t)-\Lambda(t)}\left(X_0 + \left(\int_0^t K\gamma(u)A_{\mathrm{Adv}}(u)e^{K\tilde{\Gamma}_Z(u)+\Lambda_Z(u)}\,\mathrm{d}u\right)X^\star\right),$$
$$m_{\mathrm{Reg}}(t) \stackrel{\mathrm{def}}{=} e^{-K\tilde{\Gamma}^{\mathrm{Reg}}(t)-\Lambda^{\mathrm{Reg}}(t)}\left(X_0 + \left(\int_0^t K\gamma(u)A_{\mathrm{Reg}}(u)e^{K\tilde{\Gamma}^{\mathrm{Reg}}(u)+\Lambda^{\mathrm{Reg}}(u)}\,\mathrm{d}u\right)X^\star\right).$$

The corresponding deterministic first terms are

$$G_{\mathscr{R}}(t) \stackrel{\mathrm{def}}{=} \frac{1}{2}\left\langle K, (m_{\mathrm{Adv}}(t) - X^\star)^{\otimes 2}\right\rangle + \frac{\eta^2}{2}, \quad G_{\mathscr{B}}(t) \stackrel{\mathrm{def}}{=} \|m_{\mathrm{Adv}}(t)\|^2,$$
$$G_{\mathcal{R}}(t) \stackrel{\mathrm{def}}{=} \frac{1}{2}\left\langle K, (m_{\mathrm{Reg}}(t) - X^\star)^{\otimes 2}\right\rangle + \frac{\eta^2}{2}, \quad G_{\Phi}(t) \stackrel{\mathrm{def}}{=} \|m_{\mathrm{Reg}}(t)\|^2.$$

We now compare m_{Adv} and m_{Reg} using their explicit representations. For conciseness, let $P_{\mathrm{Adv}}(t) \stackrel{\mathrm{def}}{=} e^{-K\tilde{\Gamma}(t)-\Lambda(t)}$. and $P_{\mathrm{Reg}}(t) \stackrel{\mathrm{def}}{=} e^{-K\tilde{\Gamma}^{\mathrm{Reg}}(t)-\Lambda^{\mathrm{Reg}}(t)}$ then we have

$$m_{\mathrm{Adv}}(t) - m_{\mathrm{Reg}}(t) = (P_{\mathrm{Adv}}(t) - P_{\mathrm{Reg}}(t))\left(X_0 + \int_0^t K\gamma(u)A_{\mathrm{Adv}}(u)(P_{\mathrm{Adv}}(u))^{-1}X^\star\,\mathrm{d}u\right)$$
$$+ P_{\mathrm{Reg}}(t)\int_0^t K\gamma(u)\left(A_{\mathrm{Adv}}(u)(P_{\mathrm{Adv}}(u))^{-1} - A_{\mathrm{Reg}}(u)(P_{\mathrm{Reg}}(u))^{-1}\right)X^\star\,\mathrm{d}u.$$

Once again from $|e^{-x} - e^{-y}| \le |x - y|$ for all $x, y \ge 0$, we obtain

$$\|P_{\mathrm{Adv}}(t) - P_{\mathrm{Reg}}(t)\|_{\mathrm{op}} \le C(\|K\|_{\mathrm{op}}, \bar{\gamma})\int_0^t D(s)\,\mathrm{d}s.$$

Similarly, we obtain $\|P_{\mathrm{Adv}}(u)^{-1} - P_{\mathrm{Reg}}(u)^{-1}\|_{\mathrm{op}} \le C(\|K\|_{\mathrm{op}}, \bar{\gamma})\int_0^u D(s)\,\mathrm{d}s$. Since $\|(P_{\mathrm{Adv}}(t))^{-1}\|_{\mathrm{op}}, \|(P_{\mathrm{Reg}}(t))^{-1}\|_{\mathrm{op}} \le C(T, \|K\|_{\mathrm{op}}, \bar{\gamma})$ and from (305), it follows that

$$\|m_{\mathrm{Adv}}(t) - m_{\mathrm{Reg}}(t)\| \le C(T, \|K\|_{\mathrm{op}}, \bar{\gamma})\int_0^t D(s)\,\mathrm{d}s.$$

Thus, we obtain the following bounds on the first terms

$$\begin{aligned} |G_{\mathscr{R}}(t) - G_{\mathcal{R}}(t)| &\le C(T, \|K\|_{\mathrm{op}}, \bar{\gamma})\int_0^t D(s)\,\mathrm{d}s \\ \text{and} \quad |G_{\mathscr{B}}(t) - G_{\Phi}(t)| &\le C(T, \|K\|_{\mathrm{op}}, \bar{\gamma})\int_0^t D(s)\,\mathrm{d}s. \end{aligned} \tag{311}$$

For the difference in risks, we have

$$\mathscr{R}(t) - \mathcal{R}(t) = G_{\mathscr{R}}(t) - G_{\mathcal{R}}(t) + \frac{1}{d}\int_0^t \gamma^2(s)\left[\mathcal{F}_{\mathrm{Adv}}(s)\mathcal{K}_2^{\mathrm{Adv}}(t,s) - \mathcal{R}_{\mathrm{Reg}}(s)\mathcal{K}_2^{\mathrm{Reg}}(t,s)\right]\mathrm{d}s.$$

Thus, since the boundedness assumption implies $|\mathcal{R}_{\mathrm{Reg}}(t)| \le C$ and by (308), (309), (310) and (311) we obtain

$$|\mathscr{R}(t) - \mathcal{R}(t)| \le C(T, \|K\|_{\mathrm{op}}, \bar{\gamma})\int_0^t D(s)\,\mathrm{d}s + C(\|K\|_{\mathrm{op}}, \bar{\gamma})\delta^2\left(\frac{1}{2} - \frac{1}{\pi}\right)\int_0^t \widehat{\mathscr{B}}_{44}(s)\,\mathrm{d}s. \tag{312}$$

Using a similar argument, we obtain

$$\left|\widehat{\mathscr{B}}_{44}(t) - \Phi(t)\right| \leq C(T, \|K\|_{\mathrm{op}}, \bar{\gamma}) \int_0^t D(s)\,\mathrm{d}s + C(\|K\|_{\mathrm{op}}, \bar{\gamma})\delta^2 \left(\frac{1}{2} - \frac{1}{\pi}\right) \int_0^t \widehat{\mathscr{B}}_{44}(t)\,\mathrm{d}s. \tag{313}$$

Adding (312) and (313), we obtain

$$D(t) \leq C(T, \|K\|_{\mathrm{op}}, \bar{\gamma}) \int_0^t D(s)\,\mathrm{d}s + C(\|K\|_{\mathrm{op}}, \bar{\gamma}, T)\delta^2 \left(\frac{1}{2} - \frac{1}{\pi}\right) \int_0^t \widehat{\mathscr{B}}_{44}(t)\,\mathrm{d}s. \tag{314}$$

An application of Grönwall's inequality, Theorem 2.1, Corollary E.1, the definition of $\mathcal{R}^{\mathrm{Adv}}$ and taking the supremum over $[0, T \wedge \Theta]$ completes the proof. □

F Idealized Exact Line Search and Polyak Stepsize

F.1 Polyak Stepsize

F.1.1 General Polyak Stepsize

In this section, we determine the learning rate threshold for which the functions $\mathcal{D}^2(X) \stackrel{\mathrm{def}}{=} \|X - X^\star\|^2$ and $\mathcal{D}^{2,\mathrm{Adv}}(X) \stackrel{\mathrm{def}}{=} \|X - X^{\star,\mathrm{Adv}}\|^2$ are decreasing when $d \to \infty$ using their det. equivs. We will work with the regularization parameter $\lambda = 0$ for simplicity. Our analysis will build upon the system of ODEs (26). The det. equiv. for the distance to least squares optimality $\|X - X^\star\|^2$ is given by

$$\mathscr{D}^2(t) = \sum_{i=1}^{d} V_{j,11}(t) - 2V_{j,12}(t) + V_{j,22}(t). \tag{315}$$

Using a similar argument as in the proof Proposition C.5, we can show that $\mathscr{D}^2(t)$ solves the following ODE.

Corollary F.1. *$\mathscr{D}^2(t)$ solves the differential equation*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}^2(t) = -2\gamma(t) \sum_{i=1}^{2} p_i Q(\widehat{\mathscr{B}}_i(t)) + \frac{\gamma(t)^2}{d} \sum_{i=1}^{2} p_i \,\mathrm{Tr}(K_i + \mu_i\mu_i^\top)\,\mathbb{E}_{v,\epsilon}[f_i'(q_{t,i}, \epsilon)^2], \tag{316}$$

where

$$Q(\widehat{\mathscr{B}}_i(t)) = \mathbb{E}_{v,\epsilon}[f_i'(q_{t,i}, \epsilon)(x_{t,i} - x_i^\star)] - \delta \frac{\mathscr{A}(t)}{\sqrt{\widehat{\mathscr{B}}_{i,44}(t)}}\,\mathbb{E}_{v,\epsilon}[u_{t,i} f_i'(q_{t,i}, \epsilon)], \tag{317}$$

for

$$\begin{aligned} q_{t,i} &= x_{t,i} \oplus x_i^\star = \sqrt{\mathscr{B}_i(t)}v + \mathfrak{m}_i(t) + \begin{bmatrix} \delta u_{t,i}\sqrt{\widehat{\mathscr{B}}_{i,44}(t)} & 0 \end{bmatrix}^\top, \\ u_{t,i} &= \mathrm{argmax}_{|u| \leq 1}\, f_i(\sqrt{\mathscr{B}_i(t)}v + \mathfrak{m}_i(t) + \begin{bmatrix} \delta u\sqrt{\widehat{\mathscr{B}}_{i,44}(t)} & 0 \end{bmatrix}^\top, \epsilon), \end{aligned} \tag{318}$$

for $v \sim \mathcal{N}(0, I_2)$. Here $\mathscr{A}(t)$ is the det. equiv. of $X_{\lfloor td \rfloor}^\top X^\star$ and solves the following ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{A}(t) = -\gamma(t) \sum_{i=1}^{2} p_i \left(\mathbb{E}_{v,\epsilon}[f_i'(q_{t,i}, \epsilon)x_i^\star] + \delta \frac{\mathscr{A}(t)}{\sqrt{\widehat{\mathscr{B}}_{i,44}(t)}}\,\mathbb{E}_{v,\epsilon}[u_{t,i} f_i'(q_{t,i}, \epsilon)] \right). \tag{319}$$

Given (316), the exact threshold for monotone decrease of $\|X - X^\star\|^2$ for SGD (i.e. $\mathrm{d}\mathscr{D}^2(t) < 0$) which we call $\gamma^{\mathrm{Stable},\star}$ is given by

$$\gamma(t) < \gamma_{\lfloor td \rfloor}^{\mathrm{Stable},\star} = \frac{\sum_{i=1}^{2} p_i Q(\widehat{\mathscr{B}}_i(t))}{\sum_{i=1}^{2} p_i \frac{\mathrm{Tr}(K_i + \mu_i\mu_i^\top)}{2d}\,\mathbb{E}_{v,\epsilon}[f_i'(q_{t,i}, \epsilon)^2]}. \tag{320}$$

Given that adversarial training is an adaptive algorithm which depends on the sampled data point at each step, it follows that $\gamma_{\lfloor td \rfloor}^{\mathrm{Stable},\star}$ is adaptive. Translating this quantity back to SGD, we obtain the threshold for stability with respect to descent optimality towards $X^\star$

$$\gamma_k^{\mathrm{Stable},\star} = \frac{\langle X_k - X^\star, \nabla \mathcal{R}^{\mathrm{Adv}}(X_k)\rangle}{\sum_{i=1}^2 p_i \frac{\mathrm{Tr}(K_i+\mu_i\mu_i^\top)}{2d}\,\mathbb{E}_{v,\epsilon}[f_i'(g_{k,i},\epsilon_{k+1})^2]}. \tag{321}$$

This stability threshold corresponds to the adversarial version of the results presented in Section 3 and Appendix E of Collins-Woodfin et al. (2024b) and should be compared to the greedy learning rate $\gamma^{\mathrm{Polyak},\star}(t)$ that maximizes the decrease of $\mathscr{D}^2(t)$ at each iteration: $\gamma^{\mathrm{Polyak},\star}(t) \in \mathrm{argmin}_\gamma\, \mathrm{d}\mathscr{D}^2(t)$. Analogously to Collins-Woodfin et al. (2024b), we call this learning rate the *Polyak stepsize*. Solving for γ_k and $\gamma(t)$ respectively in (320) and (321), we obtain the following closed forms for the Polyak learning rate

$$\gamma_k^{\mathrm{Polyak},*} = \tfrac{1}{2}\gamma_k^{\mathrm{Stable},*} \quad \text{with det. equiv.} \quad \gamma^{\mathrm{Polyak},*}(t) = \tfrac{1}{2}\gamma^{\mathrm{Stable},*}(t). \tag{322}$$

Setting $\delta = 0$ recovers the results in Collins-Woodfin et al. (2024b). Hence, ℓ_2-adversarial training preserves the ratio of the Polyak learning rate to stable learning rate with respect to $X^\star$. We also obtain a threshold for descent of $\|X - X^{\star,\mathrm{Adv}}\|^2$ whose det. equiv. is given by

$$\mathscr{D}^{2,\mathrm{Adv}}(t) = \sum_{i=1}^d V_{j,11}(t) - 2\mathscr{U}_j(t) + \|X^{\star,\mathrm{Adv}}\|^2, \tag{323}$$

where $\mathscr{U}_j(t)$ is the det. equiv. of $X_{\lfloor td \rfloor}^\top \omega_j \omega_j^\top X^{\star,\mathrm{Adv}}$. Using a similar argument as in the proof Proposition C.5, we can show that $\mathscr{D}^{2,\mathrm{Adv}}(t)$ solves the following ODE.

Corollary F.2. *$\mathscr{D}^{2,\mathrm{Adv}}(t)$ solves the differential equation*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}^{2,\mathrm{Adv}}(t) = -2\gamma(t)\sum_{i=1}^2 p_i Q^{\mathrm{Adv}}(\widehat{\mathscr{B}}_i(t)) + \frac{\gamma(t)^2}{d}\sum_{i=1}^2 p_i\,\mathrm{Tr}(K_i+\mu_i\mu_i^\top)\,\mathbb{E}_{v,\epsilon}[f_i'(q_{t,i},\epsilon)^2], \tag{324}$$

where

$$Q^{\mathrm{Adv}}(\widehat{\mathscr{B}}_i^{\mathrm{Adv}}(t)) = \mathbb{E}_{w,\epsilon}[f_i'(q_{t,i},\epsilon)(x_{t,i} - x_i^{\star,\mathrm{Adv}})] - \delta\frac{\mathscr{A}^{\mathrm{Adv}}(t)}{\sqrt{\widehat{\mathscr{B}}_{i,44}(t)}}\,\mathbb{E}_{v,\epsilon}[u_{t,i}f_i'(q_{t,i},\epsilon)], \tag{325}$$

for

$$\begin{aligned} q_{t,i} &= \sqrt{\mathscr{B}_i(t)}v + \mathfrak{m}_i(t) + \begin{bmatrix}\delta u_{t,i}\sqrt{\widehat{\mathscr{B}}_{i,44}(t)} & 0\end{bmatrix}^\top, \\ u_{t,i} &= \mathrm{argmax}_{|u|\le 1}\, f_i(\sqrt{\mathscr{B}_i(t)}v + \mathfrak{m}_i(t) + \begin{bmatrix}\delta u\sqrt{\widehat{\mathscr{B}}_{i,44}(t)} & 0\end{bmatrix}^\top, \epsilon), \end{aligned} \tag{326}$$

for $w = v \oplus y \sim \mathcal{N}(0, I_3)$ for $v \in \mathbb{R}^2$ and $y \in \mathbb{R}$, where

$$x_{t,i} \oplus x_i^\star \oplus x_i^{\star,\mathrm{Adv}} = \sqrt{\mathscr{B}_i^{\mathrm{Adv}}(t)}w + \mathfrak{m}_i(t) \oplus (\mu_i^\top X^{\star,\mathrm{Adv}}) + \begin{bmatrix}\delta u_{t,i}\sqrt{\widehat{\mathscr{B}}_{i,44}(t)} & 0 & 0\end{bmatrix}^\top. \tag{327}$$

Here $\mathscr{B}_i^{\mathrm{Adv}}(t)$ is defined in (37) *and $\mathscr{A}^{\mathrm{Adv}}(t)$ is the det. equiv. of $X_{\lfloor td \rfloor}^\top X^{\star,\mathrm{Adv}}$ which solves the following ODE*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{A}^{\mathrm{Adv}}(t) = -\gamma(t)\sum_{i=1}^2 p_i\left(\mathbb{E}_{w,\epsilon}[f_i'(q_{t,i},\epsilon)x^{\star,\mathrm{Adv}}] + \delta\frac{\mathscr{A}^{\mathrm{Adv}}(t)}{\sqrt{\widehat{\mathscr{B}}_{i,44}(t)}}\,\mathbb{E}_{v,\epsilon}[u_{t,i}f_i'(q_{t,i},\epsilon)]\right). \tag{328}$$

Given (324), the exact threshold for monotone decrease of $\|X - X^{\star,\mathrm{Adv}}\|^2$ for SGD (i.e. $\mathrm{d}\mathscr{D}^{2,\mathrm{Adv}}(t) < 0$) is given by

$$\gamma(t) < \gamma_{\lfloor td \rfloor}^{\mathrm{Stable,Adv}} = \frac{\sum_{i=1}^2 p_i Q^{\mathrm{Adv}}(\widehat{\mathscr{B}}_i^{\mathrm{Adv}}(t))}{\sum_{i=1}^2 p_i \frac{\mathrm{Tr}(K_i+\mu_i\mu_i^\top)}{2d}\,\mathbb{E}_{v,\epsilon}[f_i'(q_{t,i},\epsilon)^2]}. \tag{329}$$

Translating this quantity back to SGD, we obtain the threshold for stability for descent towards $X^{\star,\mathrm{Adv}}$:

$$\begin{aligned}\gamma_k^{\mathrm{Stable,Adv}} &= \frac{\langle X_k - X^{\star,\mathrm{Adv}}, \nabla\mathcal{R}^{\mathrm{Adv}}(X_k)\rangle}{\sum_{i=1}^2 p_i \frac{\mathrm{Tr}(K_i+\mu_i\mu_i^\top)}{2d}\,\mathbb{E}_{v,\epsilon}[f_i'(g_{k,i},\epsilon_{k+1})^2]} \\ &= \frac{\sum_{i=1}^2 p_i\,\mathbb{E}_{a,\epsilon}\left[f_i'(g_{k,i},\epsilon)(X_k - X^\star)^\top a_{k+1,i}\right]}{\sum_{i=1}^2 p_i \frac{\mathrm{Tr}(K_i+\mu_i\mu_i^\top)}{2d}\,\mathbb{E}_{v,\epsilon}[f_i'(g_{k,i},\epsilon_{k+1})^2]} \\ &\quad + \frac{\sum_{i=1}^2 p_i\,\mathbb{E}_{a,\epsilon}\left[f_i'(g_{k,i},\epsilon)(X^\star - X^{\star,\mathrm{Adv}})^\top a_{k+1,i}\right]}{\sum_{i=1}^2 p_i \frac{\mathrm{Tr}(K_i+\mu_i\mu_i^\top)}{2d}\,\mathbb{E}_{v,\epsilon}[f_i'(g_{k,i},\epsilon_{k+1})^2]} \\ &\quad + \frac{\delta\left(\|X_k\| - \langle X^{\star,\mathrm{Adv}}, \frac{X_k}{\|X_k\|}\rangle\right)\sum_{i=1}^2 p_i\,\mathbb{E}_{a,\epsilon}\left[f_i'(g_{k,i},\epsilon)s_{k,i}\right]}{\sum_{i=1}^2 p_i \frac{\mathrm{Tr}(K_i+\mu_i\mu_i^\top)}{2d}\,\mathbb{E}_{v,\epsilon}[f_i'(g_{k,i},\epsilon_{k+1})^2]}.\end{aligned} \tag{330}$$

This stability threshold admits a natural decomposition into three contributions. The first term is the standard descent contribution towards $X^\star$. The second term is the correction for the shift in ground truth between $X^\star$ and $X^{\star,\mathrm{Adv}}$. The third term is the contribution from the adversarial perturbation of the input.

F.1.2 Polyak Stepsize For ℓ_2-adversarial least-squares

A natural question is how the stability threshold differs for SGD on the population risk $\mathcal{R}(X_k)$ versus the adversarial risk $\mathcal{R}^{\mathrm{Adv}}(X_k)$. We focus on the case of SGD ran on the least squares loss in the soft label setting with anisotropic centered Gaussian data $a \sim \mathcal{N}(0, K)$ which corresponds to the following minimization problem

$$\min_{X\in\mathbb{R}^d}\left\{\mathcal{R}(X) = \tfrac{1}{2}\mathbb{E}_{a,\epsilon}[((X-X^\star)^\top a - \epsilon)^2] = \tfrac{1}{2}\eta^2 + \tfrac{1}{2}(X-X^\star)^\top K(X-X^\star)\right\}. \tag{331}$$

Translating the relevant det. equivs. to their non-deterministic counterparts, the Polyak stepsize for (331) which we call $\mathfrak{g}^{\mathrm{Polyak}}$ corresponds to Collins-Woodfin et al. (2024b)

$$\mathfrak{g}_k^{\mathrm{Polyak}} = \frac{2\mathcal{R}(X_k) - \eta^2}{\frac{2\,\mathrm{Tr}(K)}{d}\mathcal{R}(X_k)} \text{ and on noiseless least squares,} \quad \mathfrak{g}_k^{\mathrm{Polyak}} = \frac{d}{\mathrm{Tr}(K)}. \tag{332}$$

In the case of ℓ_2-adversarial least squares, we define γ_k^{eff} and λ_k^{eff} as the effective learning rate and effective regularization

$$\gamma_k^{\mathrm{eff}} = \gamma_k\left(1 + \delta\sqrt{\frac{2}{\pi}}\frac{\|X_k\|}{\sqrt{2\mathcal{R}(X_k)}}\right) \quad \text{and} \quad \lambda_k^{\mathrm{eff}} = \frac{\delta^2 + \delta\sqrt{\frac{2}{\pi}}\frac{\sqrt{2\mathcal{R}(X_k)}}{\|X_k\|}}{1 + \delta\sqrt{\frac{2}{\pi}}\frac{\|X_k\|}{\sqrt{2\mathcal{R}(X_k)}}}, \tag{333}$$

with γ_k corresponding to the learning rate defined in Assumption 8. See Section E for more details on γ_k^{eff} and λ_k^{eff}. Let $\tilde{\gamma}^{\mathrm{eff}}(t)$ and $\tilde{\lambda}^{\mathrm{eff}}(t)$ denote the det. equivs. of γ_k^{eff} and λ_k^{eff} written in terms of $\mathscr{R}(t)$ and $\widehat{\mathscr{B}}_{44}(t)$. Then the descent to least squares optimality for ℓ_2-adversarial least squares solves the following ODE:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}^2(t) = -2\tilde{\gamma}^{\mathrm{eff}}(t)\left(2\mathscr{R}(t) - \eta^2 + \tilde{\lambda}^{\mathrm{eff}}(t)(\widehat{\mathscr{B}}_{44}(t) - \mathscr{A}(t))\right) + \frac{2\gamma^2(t)}{d}\,\mathrm{Tr}(K)\mathscr{R}^{\mathrm{Adv}}(t). \tag{334}$$

where $\mathscr{R}^{\mathrm{Adv}}(t)$ is the det. equiv. of $\mathcal{R}^{\mathrm{Adv}}(X_k)$. Noting the relation

$$\mathscr{R}^{\mathrm{adv}}(t) = \frac{\tilde{\gamma}^{\mathrm{eff}}(t)}{\gamma(t)}\left(\mathscr{R}(t) + \frac{\tilde{\lambda}^{\mathrm{eff}}(t)}{2}\widehat{\mathscr{B}}_{44}(t)\right), \tag{335}$$

then translating these quantities back to SGD, the Polyak stepsize with respect to $X^\star$ is naturally adaptive and given by

$$\gamma_k^{\mathrm{Polyak,ls}} = \frac{2\mathcal{R}(X_k) - \eta^2 + \lambda_k^{\mathrm{eff}}(\|X_k\|^2 - \langle X_k, X^\star\rangle)}{\frac{\mathrm{Tr}(K)}{d}(2\mathcal{R}(X_k) + \lambda_k^{\mathrm{eff}}\|X_k\|^2)}. \tag{336}$$

In the noiseless setting, this simplifies to

$$\begin{aligned}\gamma_k^{\text{Polyak,ls}} &= \frac{d}{\operatorname{Tr}(K)}\left(1 - \lambda_k^{\text{eff}}\frac{\langle X_k, X^\star\rangle}{2\mathcal{R}(X_k) + \lambda_k^{\text{eff}}\|X_k\|^2}\right) \\ &= \frac{d}{\operatorname{Tr}(K)}\left(1 - \left(\delta^2 + \delta\sqrt{\frac{2}{\pi}}\frac{\sqrt{2\mathcal{R}(X_k)}}{\|X_k\|}\right)\frac{\langle X_k, X^\star\rangle}{2\mathcal{R}^{\text{Adv}}(X_k)}\right).\end{aligned} \tag{337}$$

In the case of ℓ_2-adversarial least squares, the descent to adversarial optimality solves the following ODE:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}^{2,\text{Adv}}(t) = -2\tilde{\gamma}^{\text{eff}}(t)\Bigg(&2\mathscr{R}(t) - \eta^2 + \tilde{\lambda}^{\text{eff}}(t)(\widehat{\mathscr{B}}_{44}(t) - \mathscr{A}^{\text{Adv}}(t)) \\ &+ \sum_{j=1}^{d}\lambda_j\left(\mathscr{V}_{j,21}(t) - \mathscr{V}_{j,22}(t) - \mathscr{U}_j(t) + \langle X^\star, \omega_j\rangle\langle X^{\star,\text{Adv}}, \omega_j\rangle\right)\Bigg) \\ &+ \frac{2\gamma^2(t)}{d}\operatorname{Tr}(K)\mathscr{R}^{\text{Adv}}(t).\end{aligned} \tag{338}$$

Whenever $X^{\star,\text{Adv}} = X^\star$, we recover (334). Using a similar argument as (336), the Polyak stepsize $\gamma_k^{\text{Polyak,adv}}$ with respect to $X^{\star,\text{Adv}}$ is naturally adaptive and given by

$$\gamma_k^{\text{Polyak,adv}} = \frac{2\mathcal{R}(X_k) - \eta^2 - (X_k - X^\star)^\top K(X^{\star,\text{Adv}} - X^\star) + \lambda_k^{\text{eff}}(\|X_k\|^2 - \langle X_k, X^{\star,\text{Adv}}\rangle)}{\frac{\operatorname{Tr}(K)}{d}\left(2\mathcal{R}(X_k) + \lambda_k^{\text{eff}}\|X_k\|^2\right)}. \tag{339}$$

Adversarial training is designed to improve robust performance by minimizing the worst-case loss over admissible perturbations, but this can come at the expense of standard risk performance. This standard risk versus robust risk trade-off has been extensively studied empirically and theoretically.

This brings us to the question: can this trade-off be controlled through the learning rate scheduling of adversarial training itself. In the context of our framework, this can be phrased as: given values of δ, $X^\star$ and K, does there exist a learning rate schedule which minimizes the distance to optimality towards $X^\star$? The thresholds (321) and (330) are pointwise stability thresholds along the SGD trajectory. In particular, descent towards $X^\star$ at iteration k is possible precisely when

$$\langle X_k - X^\star, \nabla\mathcal{R}^{\text{Adv}}(X_k)\rangle > 0. \tag{340}$$

This condition may hold along a particular path even when $X^\star$ is not the minimizer of $\mathcal{R}^{\text{Adv}}$ which implies it is possible for adversarial training to move towards $X^\star$ over a region of state space despite the adversarial optimum satisfying $X^{\star,\text{Adv}} \neq X^\star$. In order for (340) to hold globally on $\mathcal{U}$ (see Assumption 6), this requires $\nabla\mathcal{R}^{\text{Adv}}(X^\star) = 0$ as shown in the following Lemma.

Lemma F.1. *Suppose $\mathcal{R}^{\text{Adv}}$ is continuously differentiable at $X^\star$ then*

$$\langle X_k - X^\star, \nabla\mathcal{R}^{\text{Adv}}(X_k)\rangle > 0 \;\forall\; X \neq X^\star \implies \nabla\mathcal{R}^{\text{Adv}}(X^\star) = 0.$$

Proof. We proceed by proof by contradiction. Suppose $\nabla\mathcal{R}^{\text{Adv}}(X^\star) \neq 0$ and for some $\varepsilon > 0$, define $X_\varepsilon = X^\star + \varepsilon u$ for $u = -\frac{\nabla\mathcal{R}^{\text{Adv}}(X^\star)}{\|\nabla\mathcal{R}^{\text{Adv}}(X^\star)\|}$. Then we have $\|X_\varepsilon - X^\star\| = \varepsilon$.

Since $\mathcal{R}^{\text{Adv}}$ is continuously differentiable at $X^\star$ then $u^\top\nabla\mathcal{R}^{\text{Adv}}(X_\varepsilon) \to u^\top\nabla\mathcal{R}^{\text{Adv}}(X^\star)$ as $\varepsilon \to 0$. Now, $u^\top\nabla\mathcal{R}^{\text{Adv}}(X^\star) = -\|\nabla\mathcal{R}^{\text{Adv}}(X^\star)\| < 0$. By continuity, for sufficiently small ε we have

$$\|\mathcal{R}^{\text{Adv}}(X_\varepsilon) - \mathcal{R}^{\text{Adv}}(X^\star)\| < \frac{1}{2}\|\mathcal{R}^{\text{Adv}}(X_\star)\|.$$

Hence, we obtain

$$\begin{aligned}u^\top\nabla\mathcal{R}^{\text{Adv}}(X_\varepsilon) &= u^\top\mathcal{R}^{\text{Adv}}(X^\star) + u^\top(\nabla\mathcal{R}^{\text{Adv}}(X_\varepsilon) - \mathcal{R}^{\text{Adv}}(X^\star)) \\ &\leq -\frac{1}{2}\|\nabla\mathcal{R}^{\text{Adv}}(X^\star)\| < 0.\end{aligned}$$

Thus, it follows that $(X_\varepsilon - X^\star)^\top\nabla\mathcal{R}^{\text{Adv}}(X_\varepsilon) = \varepsilon u^\top\mathcal{R}^{\text{Adv}}(X_\varepsilon) < 0$ which is a contradiction. □

Sufficient conditions for $\langle X_k - X^{\star,\mathrm{Adv}}, \nabla\mathcal{R}^{\mathrm{Adv}}(X_k)\rangle > 0$ for all $X \neq X^{\star,\mathrm{Adv}}$ are either $\mathcal{R}^{\mathrm{Adv}}$ is convex and has a unique minimizer or $\mathcal{R}^{\mathrm{Adv}}$ is υ-strongly convex for some $\upsilon > 0$.

A consequence of Lemma F.1 is that if $X^{\star,\mathrm{Adv}}$ is the unique minimizer fo $\mathcal{R}^{\mathrm{Adv}}$ then $\langle X_k - X^{\star}, \nabla\mathcal{R}^{\mathrm{Adv}}(X_k)\rangle > 0$ holding globally requires $X^{\star,\mathrm{Adv}} = X^{\star}$. As proven in Xing et al. (2021), in the case of ℓ_2-adversarial least squares with $K \succ 0$, $X^{\star,\mathrm{Adv}} = X^{\star}$ whenever

$$\delta < \sqrt{\frac{2}{\pi}} \frac{\|X^{\star}\|}{(X^{\star})^{\top} K^{-1} X^{\star}}. \tag{341}$$

By Lemma I.1, for least squares $\mathcal{R}^{\mathrm{Adv}}$ is υ-strongly convex with unique minimizer $X^{\star,\mathrm{Adv}}$ which entails that for any iterate $X_k \neq X^{\star,\mathrm{Adv}}$ of SGD, there exists a learning rate for which descent towards $X^{\star,\mathrm{Adv}}$ is possible.

F.2 Line Search

F.2.1 General Line Search

Similarly to the Polyak stepsize F.1, we determine the learning rate threshold for which the adversarial risk $\mathcal{R}^{\mathrm{Adv}}(X)$ is decreasing at each iteration when $d \to \infty$ using its det. equiv. We will work with the regularization parameter $\lambda = 0$ for simplicity. Given that the expressions in this section are very involved for mixture of Gaussians, we work with a single covariance K, a single mean $\mu = 0$ and regularization parameter $\lambda = 0$. We thus drop the subscripts i on $\widehat{B}_i(W)$ for example. Similarly to Equation (90) in Collins-Woodfin et al. (2024b), from Lemma D.1 and Assumptions 6 and 7, there exists functions m and v such that

$$\begin{aligned} \|\nabla\mathcal{R}^{\mathrm{Adv}}(X)\|^2 &= m\left(\widehat{B}(W), \hat{X}^{\top} K^2 \hat{X}\right), \\ \text{and} \quad \mathrm{Tr}\left(\nabla^2\mathcal{R}^{\mathrm{Adv}}(X) K\right) &= v\left(\widehat{B}(W), \hat{X}^{\top} K^3 \hat{X}, \hat{X}^{\top} K^2 \hat{X}, K\right). \end{aligned} \tag{342}$$

From Theorem C.6 and since m and v satisfy Assumption 12 then define $\mathscr{M}(t)$ and $\upsilon(t)$ as their respective det. equivs. It follows using a similar argument as in the proof Proposition C.5 that $\mathscr{R}^{\mathrm{Adv}}$ satisfies the following ODE

$$\mathrm{d}\mathscr{R}^{\mathrm{Adv}}(t) = -\gamma(t)\mathscr{M}(t)\,\mathrm{d}t + \frac{\gamma^2(t)}{d}\upsilon(t)\mathcal{I}(\widehat{\mathscr{B}}(t))\,\mathrm{d}t. \tag{343}$$

Similarly to the section on the Polyak stepsize, we get a maximal learning rate threshold which guarantees monotone decrease of $\mathscr{R}^{\mathrm{Adv}}(t)$ for a given time t (i.e. $\mathrm{d}\mathscr{R}^{\mathrm{Adv}}(t) < 0$). Translating relevant det. equivs. to SGD, we obtain

$$\gamma_k^{\mathrm{Stable},\mathscr{R}^{\mathrm{Adv}}} = \frac{\|\nabla\mathcal{R}^{\mathrm{Adv}}(X_k)\|^2}{\frac{\mathrm{Tr}(\nabla^2\mathcal{R}^{\mathrm{Adv}}(X_k)K)}{d}\mathcal{I}(\widehat{B}(W_k))} \quad \text{and det. equiv.} \quad \gamma^{\mathrm{Stable},\mathscr{R}^{\mathrm{Adv}}}(t) = \frac{\mathscr{M}(t)}{\frac{1}{d}\upsilon(t)\mathcal{I}(\widehat{\mathscr{B}}(t))}. \tag{344}$$

Analogously to Collins-Woodfin et al. (2024b), optimizing over γ in the ODE (343) and translating relevant det. equivs. back to SGD, we obtain the learning rate schedule of *exact line search*

$$\gamma_k^{\mathrm{line}} = \frac{\gamma_k^{\mathrm{Stable},\mathscr{R}^{\mathrm{Adv}}}}{2} \quad \text{and det. equiv.} \quad \gamma^{\mathrm{line}}(t) = \frac{\gamma^{\mathrm{Stable},\mathscr{R}^{\mathrm{Adv}}}}{2}. \tag{345}$$

F.3 Line search for ℓ_2-Adversarial Least Squares

We study the learning rate schedule produced by exact line search on ℓ_2-adversarial least squares. See (331) in Section F.1 for the problem setup. In this case, the det. equiv. $\mathscr{R}^{\mathrm{Adv}}$ satisfies the following ODE

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathscr{R}^{\mathrm{Adv}}(t) = -\gamma\left(1 + \delta\sqrt{\frac{2}{\pi}}q(t)\right)^2 \sum_{j=1}^{d}\Big(&\big(\lambda_j + \tilde{\lambda}^{\mathrm{eff}}(t)\big)^2 \mathscr{V}_{j,11}(t) \\ &- 2\lambda_j\big(\lambda_j + \tilde{\lambda}^{\mathrm{eff}}(t)\big)\mathscr{V}_{j,12}(t) + \lambda_j^2\mathscr{V}_{j,22}(t)\Big) \\ &+ \frac{\gamma^2}{d}\left(1 + \delta\sqrt{\frac{2}{\pi}}\,q(t)\right)\mathscr{R}^{\mathrm{Adv}}(t)\left(\mathrm{Tr}(K^2) + \tilde{\lambda}^{\mathrm{eff}}(t)\,\mathrm{Tr}(K)\right). \end{aligned} \tag{346}$$

Hence, for ℓ_2-adversarial least squares we obtain

$$\gamma^{\text{line}}(t) = \frac{\sum_{j=1}^{d}\Big(\big(\lambda_j + \tilde{\lambda}^{\text{eff}}(t)\big)^2 \mathscr{V}_{j,11}(t) - 2\lambda_j\big(\lambda_j + \tilde{\lambda}^{\text{eff}}(t)\big)\mathscr{V}_{j,12}(t) + \lambda_j^2 \mathscr{V}_{j,22}(t)\Big)}{\frac{2}{d}\left(\mathscr{R}(t) + \frac{\tilde{\lambda}^{\text{eff}}(t)}{2}\widehat{\mathscr{B}}_{44}(t)\right)\Big(\operatorname{Tr}(K^2) + \tilde{\lambda}^{\text{eff}}(t)\operatorname{Tr}(K)\Big)}. \tag{347}$$

F.4 Comparing the Polyak Stepsize and Line Search for ℓ_2-adversarial least-squares

In this section, we provide an extended version of Section 4. We consider the idealized algorithms: *exact line search* and *Polyak stepsize*. See (Collins-Woodfin et al., 2024b) for a discussion of stochastic line search and Polyak-type stepsizes. Let $\mathscr{D}^{2,\text{Adv}}(t)$ be the det. equiv. of $\|X_{\lfloor td \rfloor} - X^{\star,\text{Adv}}\|^2$ where we recall $X^{\star,\text{Adv}} = \operatorname{argmin}_X \mathcal{R}^{\text{Adv}}(X)$. In deterministic optimization, the Polyak stepsize is the learning rate schedule chosen for maximum decrease of the distance to optimality at the following iteration. Exact line search is the learning rate strategy chosen for maximum decrease of the objective function at the next iteration.

Analogously to Collins-Woodfin et al. (2024b), given that $\mathcal{R}^{\text{Adv}}(X_{\lfloor td \rfloor})$ and $\|X_{\lfloor td \rfloor} - X^{\star,\text{Adv}}\|^2$ are close to $\mathscr{R}^{\text{Adv}}(t)$ and $\mathscr{D}^{2,\text{Adv}}(t)$ when $d \to \infty$ (See Theorem B.4), we can ask this to hold for the det. equivs. We restrict the results in this section to ℓ_2-adversarial least squares (see start of Section E) as the general expressions derived are too cumbersome. See Appendix F.1 and F.2 for general results and the proofs. Given that we do not have access to $X^\star$ nor $X^{\star,\text{Adv}}$ nor distributional knowledge of a in practice, these are idealized algorithms. Nevertheless, they serve as a foundation for developing more practical algorithms.

Polyak Stepsize. The stability threshold for descent towards $X^{\star,\text{Adv}}$ corresponds to the largest learning rate such that $\mathrm{d}\mathscr{D}^{2,\text{Adv}}(t) < 0$, which we call $\gamma_k^{\text{Stable,adv}}$. For ℓ_2-adversarial least squares, translating back det. equivs. to SGD, this corresponds to

$$\gamma_k^{\text{Stable,adv}} = \frac{2\mathcal{R}(X_k) - \eta^2 - (X_k - X^\star)^\top K(X^{\star,\text{Adv}} - X^\star) + \lambda_k^{\text{eff}}(\|X_k\|^2 - \langle X_k, X^{\star,\text{Adv}}\rangle)}{\frac{\operatorname{Tr}(K)}{2d}\left(2\mathcal{R}(X_k) + \lambda_k^{\text{eff}}\|X_k\|^2\right)}. \tag{348}$$

By Lemma I.1, for any $\delta > 0$, $\mathcal{R}^{\text{Adv}}$ is υ-strongly convex with unique minimizer $X^{\star,\text{Adv}}$ which implies that for any X and combination of δ and K, there exists a learning rate sequence for which descent towards $X^{\star,\text{Adv}}$ is possible. However, at the optimum $X = X^{\star,\text{Adv}}$, we have $\gamma_k^{\text{Polyak,adv}} = 0$ from which it follows that no constant learning rate schedule guarantees global monotone descent towards $X^{\star,\text{Adv}}$. This differs from standard noiseless least squares, where the stability threshold for descent towards $X^\star$ is the constant $\gamma_k^{\text{Stable}} = \frac{2d}{\operatorname{Tr}(K)}$ (Loizou et al., 2021; Paquette et al., 2021).

The numerator of stability threshold above admits a natural decomposition into three contributions:

1. $2\mathcal{R}(X_k) - \eta^2$ is the standard descent contribution towards $X^\star$
2. $(X_k - X^\star)^\top K(X^{\star,\text{Adv}} - X^\star)$ is the correction for the shift between $X^\star$ and $X^{\star,\text{Adv}}$
3. $\lambda_k^{\text{eff}}(\|X_k\|^2 - \langle X_k, X^{\star,\text{Adv}}\rangle)$ arises from the adversarial perturbation of the input

This also holds for our general results and hints that a practical learning rate schedule for ℓ_2-adversarial training must take into consideration the difference in objective between the standard loss and adversarial loss and must take into consideration the adversarial perturbation of the input.

The Polyak stepsize corresponds to $\gamma^{\text{Polyak,adv}}(t) \in \operatorname{argmin}_\gamma \mathrm{d}\mathscr{D}^{2,\text{Adv}}(t)$. In the case of ℓ_2-adversarial least squares, this corresponds to $\gamma_k^{\text{Polyak,adv}} = \frac{1}{2}\gamma_k^{\text{Stable,adv}}$.

Exact Line Search. Inspired by Collins-Woodfin et al. (2024b), we denote the greedy learning rate $\gamma^{\text{line}}(t) \in \operatorname{argmin}_\gamma \mathrm{d}\mathscr{R}^{\text{Adv}}(t)$ which produces the largest decreases in adversarial risk at the next iteartion as *exact line search*. For ℓ_2-adversarial least squares, translating back det. equivs. to SGD, this corresponds to

$$\gamma_k^{\text{line}} = \frac{\|\nabla\mathcal{R}^{\text{Adv}}(X_k)\|^2}{\frac{2}{d}\left(1 + \delta\sqrt{\frac{2}{\pi}}\frac{\|X_k\|}{\sqrt{2\mathcal{R}(X_k)}}\right)\mathcal{R}^{\text{Adv}}(X_k)\left(\operatorname{Tr}(K^2) + \lambda_k^{\text{eff}}\operatorname{Tr}(K)\right)}. \tag{349}$$

Here λ_k^{eff} is the effective regularization of SGD as defined in Section E. It turns out $\gamma^{\text{line}}(t)$, the det. equiv. of (349), can be expressed in terms of a ridge regression objective with fixed parameter $\tilde{\lambda}^{\text{eff}}(t)$, the det. equiv. of λ_k^{eff}.

Proposition F.1. *Consider ℓ_2-adversarial least squares and suppose K has $0 < n \leq d$ distinct eigenvalues $\lambda_n < \cdots < \lambda_1$ each with algebraic multiplicity $\#\lambda_j$. Let $J_+ \stackrel{\text{def}}{=} \{j : \lambda_j > 0\}$. Then the learning rate schedule of exact line search for all $t \geq 0$ where $\mathscr{E}(t) > 0$ is given by*

$$\gamma^{\text{line}}(t) = \frac{\sum_{j=1}^{d} \alpha_j(t)(\lambda_j + \tilde{\lambda}^{\text{eff}}(t))}{\sum_{j=1}^{n} \frac{\#\lambda_j}{d} \lambda_j(\lambda_j + \tilde{\lambda}^{\text{eff}}(t))} \left(1 - \frac{\Phi_{\min}(\tilde{\lambda}^{\text{eff}}(t))}{\Phi_{\tilde{\lambda}^{\text{eff}}(t)}(t)}\right). \tag{350}$$

Here $\Phi_{\tilde{\lambda}^{\text{eff}}(t)}(t) \stackrel{\text{def}}{=} \mathscr{R}(t) + \frac{\tilde{\lambda}^{\text{eff}}(t)}{2}\widehat{\mathscr{B}}_{44}(t)$ is the ridge regression objective with fixed parameter $\tilde{\lambda}^{\text{eff}}(t)$ and $\Phi_{\min}(\tilde{\lambda}^{\text{eff}}(t)) = \frac{1}{2}\sum_{j\in J_+} \lambda_j \frac{\tilde{\lambda}^{\text{eff}}(t)}{\lambda_j + \tilde{\lambda}^{\text{eff}}(t)} \mathscr{V}_{j,22}(t) + \frac{\eta^2}{2}$ is its minimum. Here where

$$\mathcal{E}_j(t) \stackrel{\text{def}}{=} \frac{1}{2}\left(\left(\lambda_j + \tilde{\lambda}^{\text{eff}}(t)\right)\mathscr{V}_{j,11}(t) - 2\lambda_j \mathscr{V}_{j,12}(t) + \frac{\lambda_j^2}{\lambda_j + \tilde{\lambda}^{\text{eff}}(t)}\mathscr{V}_{j,22}(t)\right),$$

denotes the λ_j eigenspace's contribution to the excess risk $\mathcal{E}(t) \stackrel{\text{def}}{=} \Phi_{\tilde{\lambda}^{\text{eff}}(t)}(t) - \Phi_{\min}(\tilde{\lambda}^{\text{eff}}(t))$ such that $\mathcal{E}(t) = \sum_{j=1}^{n} \mathcal{E}_j(t)$ and, $0 \leq \alpha_j \stackrel{\text{def}}{=} \frac{\mathscr{E}_j}{\mathscr{E}(t)} \leq 1$ for $1 \leq j \leq n$. In the isotropic setting, the learning rate schedule further simplifies to

$$\gamma^{\text{line}}(t) = 1 - \frac{\Phi_{\min}(\tilde{\lambda}^{\text{eff}}(t))}{\Phi_{\tilde{\lambda}^{\text{eff}}(t)}(t)}.$$

Note that when $\mathcal{E}(t) = 0$, we have reached the optimum.

Proof. For this given covariance setup, we have

$$\frac{1}{d}\left(\operatorname{Tr}(K^2) + \tilde{\lambda}^{\text{eff}}(t)\operatorname{Tr}(K)\right) = \sum_{j=1}^{n} \frac{\#\lambda_j}{d}\lambda_j(\lambda_j + \tilde{\lambda}^{\text{eff}}(t)).$$

Given the relation of ℓ_2-adversarial least squares to ridge regression presented in Section E, we define the ridge objective function $\Phi_{\tilde{\lambda}^{\text{eff}}(t)}(t) = \mathscr{R}(t) + \frac{\tilde{\lambda}^{\text{eff}}(t)}{2}\widehat{\mathscr{B}}_{44}(t)$. For a fixed $\tilde{\lambda}^{\text{eff}}(t)$ and $J_+ \stackrel{\text{def}}{=} \{j : \lambda_j > 0\}$, the minimum value this objective is

$$\Phi_{\min}(\tilde{\lambda}^{\text{eff}}(t)) = \frac{1}{2}\sum_{j\in J_+} \lambda_j \frac{\tilde{\lambda}^{\text{eff}}(t)}{\lambda_j + \tilde{\lambda}^{\text{eff}}(t)}\mathscr{V}_{j,22}(t) + \frac{\eta^2}{2} \tag{351}$$

Define the excess risk $\mathcal{E}(t) \stackrel{\text{def}}{=} \Phi_{\tilde{\lambda}^{\text{eff}}(t)}(t) - \Phi_{\min}(\tilde{\lambda}^{\text{eff}}(t))$ then we obtain

$$\mathcal{E}(t) = \frac{1}{2}\sum_{j\in J_+}\left(\left(\lambda_j + \tilde{\lambda}^{\text{eff}}(t)\right)\mathscr{V}_{j,11}(t) - 2\lambda_j\mathscr{V}_{j,12}(t) + \frac{\lambda_j^2}{\lambda_j + \tilde{\lambda}^{\text{eff}}(t)}\mathscr{V}_{j,22}(t)\right).$$

It is thus easy to see that $\mathscr{E}(t) = \sum_{j\in J_+} \mathscr{E}_j(t)$. Denote the numerator of (347) as $\mathscr{H}(t)$ then it satisfies

$$\mathscr{H}(t) = 2\sum_{j\in J_+}\left(\lambda_j + \tilde{\lambda}^{\text{eff}}(t)\right)\mathcal{E}_j(t).$$

It thus follows that

$$\gamma^{\text{line}}(t) = \frac{\sum_{j\in J_+}\left(\lambda_j + \tilde{\lambda}^{\text{eff}}(t)\right)\mathcal{E}_j(t)}{\left(\mathscr{R}(t) + \frac{\tilde{\lambda}^{\text{eff}}(t)}{2}\widehat{\mathscr{B}}_{44}(t)\right)\sum_{j=1}^{n}\frac{\#\lambda_j}{d}\lambda_j(\lambda_j + \tilde{\lambda}^{\text{eff}}(t))},$$

where $\mathcal{E}_j(t) = 0$ for $\lambda_j = 0$. Whenever $\mathcal{E}(t) > 0$, define

$$\alpha_j(t) \stackrel{\text{def}}{=} \frac{\mathcal{E}_j(t)}{\mathcal{E}(t)} \quad \text{then} \quad 0 \leq \alpha(t) \leq 1,$$

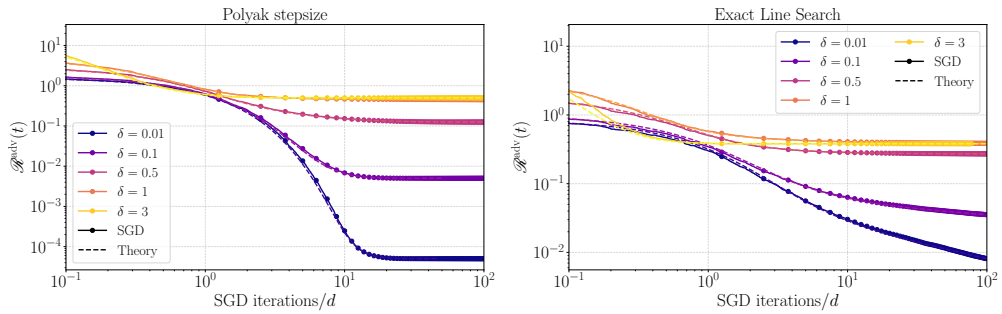

Figure 4: **SGD with exact line search γ_k^{line} or Polyak stepsize $\gamma_k^{\mathrm{Polyak,adv}}$** matches closely the path of our system of ODEs (26) with deterministic learning rates schedules $\gamma(t) = \gamma^{\mathrm{line}}(t)$ and $\gamma(t) = \gamma^{\mathrm{Polyak,adv}}(t)$ for $\mathcal{R}^{\mathrm{Adv}}(X_k)$ on noiseless ℓ_2-adversarial least squares. See Appendix J for simulation details.

then

$$\sum_{j=1}^{d}\left(\lambda_j + \tilde{\lambda}^{\mathrm{eff}}(t)\right)\mathcal{E}_j(t) = \mathcal{E}(t)\sum_{j=1}^{d}\alpha_j\left(\lambda_j + \tilde{\lambda}^{\mathrm{eff}}(t)\right),$$

from which it follows that

$$\gamma^{\mathrm{line}}(t) = \frac{\sum_{j=1}^{d}\alpha_j\left(\lambda_j + \tilde{\lambda}^{\mathrm{eff}}(t)\right)}{\sum_{j=1}^{n}\frac{\#\lambda_j}{d}\lambda_j(\lambda_j + \tilde{\lambda}^{\mathrm{eff}}(t))}\frac{\mathcal{E}(t)}{\mathscr{R}(t) + \frac{\tilde{\lambda}^{\mathrm{eff}}(t)}{2}\widehat{\mathscr{B}}_{44}(t)}.$$

The result follows from the definition of $\Phi_{\tilde{\lambda}^{\mathrm{eff}}(t)}(t)$ which implies

$$\frac{\mathcal{E}(t)}{\mathscr{R}(t) + \frac{\tilde{\lambda}^{\mathrm{eff}}(t)}{2}\widehat{\mathscr{B}}_{44}(t)} = 1 - \frac{\Phi_{\min}(\tilde{\lambda}^{\mathrm{eff}}(t))}{\Phi_{\tilde{\lambda}^{\mathrm{eff}}(t)}(t)}.$$

The bounds follow from the fact that $0 \leq \alpha(t) \leq 1$ and the isotropic result follows from simply $\lambda_1 = .. = \lambda_n$. □

It turns out the det. equiv. of the Polyak stepsize (349) can also be expressed in terms of a ridge regression objective similarly to Proposition F.1.

Proposition F.2. *Consider ℓ_2-adversarial least squares and assume K has an arbitrary number $0 < n \leq d$ distinct eigenvalues. For $1 \leq j \leq n$, define*

$$s_j^\star \overset{\mathrm{def}}{=} \begin{cases} \frac{\lambda_j}{\lambda_j + \tilde{\lambda}_\star}, & 0 < \lambda_j + \tilde{\lambda}_\star < \infty \\ 0 & \lambda_j = \tilde{\lambda}_\star = 0 \text{ or } \tilde{\lambda}_\star = \infty, \end{cases} \qquad \vartheta_j^\star(t) \overset{\mathrm{def}}{=} \begin{cases} \left(\tilde{\lambda}_\star - \tilde{\lambda}^{\mathrm{eff}}(t)\right)s_j^\star & \text{if} \quad \tilde{\lambda}_\star < \infty \\ \lambda_j & \text{if} \quad \tilde{\lambda}_\star = \infty \end{cases}.$$

Then the det. equiv. of the Polyak stepsize (348) *for all $t \geq 0$ where $\widehat{\mathscr{B}}_{44}(t) > 0$ is given by*

$$\gamma^{\mathrm{Polyak,Adv}}(t) = \frac{\Phi_{\tilde{\lambda}_\star^{\mathrm{Polyak}}(t)}(t) - \Phi_{\min}(\tilde{\lambda}_\star)}{\frac{\mathrm{Tr}(K)}{d}\left(\Phi_{\tilde{\lambda}^{\mathrm{eff}}(t)}(t)\right)}, \quad \text{for} \quad \tilde{\lambda}^{\mathrm{Polyak}}(t) \overset{\mathrm{def}}{=} \tilde{\lambda}^{\mathrm{eff}}(t) + \sum_{j=1}^{d}\vartheta_j^\star(t)\frac{\mathscr{V}_{j,12}(t)}{\widehat{\mathscr{B}}_{44}(t)}. \tag{352}$$

Here $\Phi_{\tilde{\lambda}_\star^{\mathrm{Polyak}}(t)}(t) \overset{\mathrm{def}}{=} \mathscr{R}(t) + \frac{\tilde{\lambda}_\star^{\mathrm{Polyak}}(t)}{2}\widehat{\mathscr{B}}_{44}(t)$ is the ridge regression objective with fixed parameter $\tilde{\lambda}_\star^{\mathrm{Polyak}}(t)$ and $\Phi_{\min}(\tilde{\lambda}_\star^{\mathrm{Polyak}}(t))$ is its minimum defined in Proposition F.1. When $\tilde{\lambda}_\star = \infty$, we note that

$$\Phi_{\min}(\tilde{\lambda}_\star) \overset{\mathrm{def}}{=} \lim_{\lambda\to\infty}\Phi_{\min}(\lambda) = \frac{\eta^2}{2} + \frac{1}{2}(X^\star)^\top K X^\star. \tag{353}$$

Proof. From (Xing et al., 2021, Proposition 1) and Theorem G.2, the minimizer of $\mathcal{R}^{\mathrm{Adv}}$ is given by $X^{\star,\mathrm{Adv}} = \sum_{j\in J_+}\frac{\lambda_j}{\lambda_j + \tilde{\lambda}_\star}\langle X^\star, \omega_j\rangle\omega_j$ for $J_+ \overset{\mathrm{def}}{=} \{j : \lambda_j > 0\}$ where $\tilde{\lambda}_\star$ is defined in Theorem G.2.

Hence we have for $1 \leq j \leq J_+$

$$\langle X^{\star,\mathrm{Adv}}, \omega_j \rangle = \frac{\lambda_j}{\lambda_j + \tilde{\lambda}_\star} \langle X^\star, \omega_j \rangle.$$

such that $X^{\star,\mathrm{Adv}} = 0$ when $\tilde{\lambda}_\star = 0$. Through this representation, we can express the det. equiv. of (339) as follows

$$\gamma^{\mathrm{Polyak,Adv}}(t) = \frac{\mathcal{F}^{\mathrm{Polyak}}(t)}{\frac{2\,\mathrm{Tr}(K)}{d}\left(\Phi_{\tilde{\lambda}^{\mathrm{eff}}(t)}(t)\right)},$$

where

$$\mathcal{F}^{\mathrm{Polyak}}(t) \stackrel{\mathrm{def}}{=} \sum_{j=1}^{d} \left(\left(\lambda_j + \tilde{\lambda}^{\mathrm{eff}}(t)\right) \mathscr{V}_{j,11}(t) - \left(\lambda_j + s_j^\star \left(\lambda_j + \tilde{\lambda}^{\mathrm{eff}}(t)\right)\right) \mathscr{V}_{j,12}(t) + \lambda_j s_j^\star \mathscr{V}_{j,22}(t) \right).$$

From a simple computation, (351) and (353), we obtain

$$\sum_{j=1}^{d} \lambda_j s_j^\star \mathscr{V}_{j,22}(t) = \eta^2 + \sum_{j=1}^{d} \lambda_j \mathscr{V}_{j,22}(t) - 2\Phi_{\min}(\tilde{\lambda}_\star).$$

which leads to $\mathcal{F}^{\mathrm{Polyak}}(t) = S(t) - 2\Phi_{\min}(\tilde{\lambda}_\star)$ where we define

$$S(t) \stackrel{\mathrm{def}}{=} \eta^2 + \sum_{j=1}^{d} \left(\left(\lambda_j + \tilde{\lambda}^{\mathrm{eff}}(t)\right) \mathscr{V}_{j,11}(t) - \left(\lambda_j + s_j^\star \left(\lambda_j + \tilde{\lambda}^{\mathrm{eff}}(t)\right)\right) \mathscr{V}_{j,12}(t) + \lambda_j \mathscr{V}_{j,22}(t) \right).$$

Our goal is to rewrite $S(t)$ in terms of $\Phi_{\tilde{\lambda}^{\mathrm{Polyak}}(t)}(t)$ for some explicit $\tilde{\lambda}^{\mathrm{Polyak}}(t)$. It is easy to see from the definition of $\mathscr{R}(t)$ and the identity

$$\lambda_j - s_j^\star \left(\lambda_j + \tilde{\lambda}^{\mathrm{eff}}(t)\right) = \vartheta_j^\star(t),$$

that

$$S(t) - 2\mathscr{R}(t) = \widetilde{\lambda}^{\mathrm{eff}}(t) \widehat{\mathscr{B}}_{44}(t) + \sum_{j=1}^{d} \vartheta_j^\star(t) \mathscr{V}_{j,12}(t).$$

Assuming $\widehat{\mathscr{B}}_{44}(t) > 0$, define

$$\tilde{\lambda}^{\mathrm{Polyak}}(t) \stackrel{\mathrm{def}}{=} \widetilde{\lambda}^{\mathrm{eff}}(t) + \left(\tilde{\lambda}_\star - \widetilde{\lambda}^{\mathrm{eff}}(t)\right) \sum_{j=1}^{d} s_j^\star \frac{\mathscr{V}_{j,12}(t)}{\widehat{\mathscr{B}}_{44}(t)},$$

then it is clear that $S(t) = 2\Phi_{\tilde{\lambda}^{\mathrm{Polyak}}(t)}(t)$ from which the results follows. □

Note that Theorems B.4, C.3 and C.6 are formulated in terms of the set $\mathcal{U}$ in which $\widehat{\mathscr{B}}_{44}(t) > 0$. Hence, one can view Proposition F.2 as remaining valid for all time these theorems hold.

Comparison of Exact Line Search and Polyak stepsize. We now compare the det. equivs. (348) and (349). The representations (350) and (352) allow us to compare exact line search and the Polyak stepsize locally, at a fixed state of the limiting ODE. We use this comparison to identify the factors that can cause exact line search to be suboptimal relative to the Polyak stepsize. Given a fixed state of the coupled ODEs (461), the ratio of $\gamma^{\mathrm{line}}(t)/\gamma^{\mathrm{Polyak,Adv}}(t)$ decomposes into a factor driven by anisotropy and a factor driven by the mismatch in $\tilde{\lambda}^{\mathrm{eff}}(t)$ and $\tilde{\lambda}_\star^{\mathrm{Polyak}}(t)$:

$$\boxed{\frac{\gamma^{\mathrm{line}}(t)}{\gamma^{\mathrm{Polyak,Adv}}(t)} = \underbrace{\frac{\frac{\mathrm{Tr}(K)}{d}\sum_{r=1}^{d} \alpha_r(t)\left(\lambda_r + \tilde{\lambda}^{\mathrm{eff}}(t)\right)}{\sum_{r=1}^{n} \frac{\#\lambda_r}{d}\lambda_r\left(\lambda_r + \tilde{\lambda}^{\mathrm{eff}}(t)\right)}}_{Spectral\,factor} \cdot \underbrace{\frac{\Phi_{\tilde{\lambda}^{\mathrm{eff}}(t)}(t) - \Phi_{\min}\left(\tilde{\lambda}^{\mathrm{eff}}(t)\right)}{\Phi_{\tilde{\lambda}_\star^{\mathrm{Polyak}}(t)}(t) - \Phi_{\min}(\tilde{\lambda}_\star)}}_{Ridge\,factor}.} \tag{354}$$

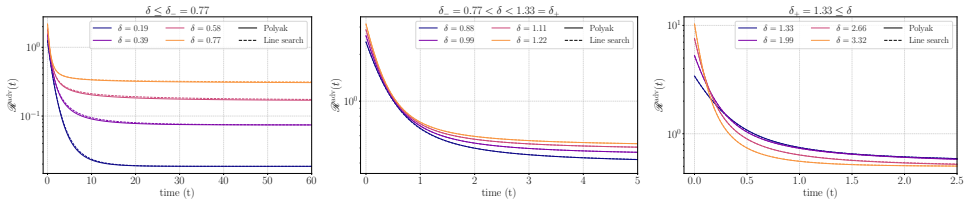

Figure 5: **Comparison between Exact Line Search and Polyak Stepsize under weak anisotropy** on noiseless ℓ_2-adversarial least squares for different values of δ in the three regimes of $X^{\star,\mathrm{Adv}}$ (See Theorem 3.3). The three plots illustrate the convergence of the ℓ_2-adversarial risk and that, under weak anisotropy, exact line search and the Polyak stepsize perform similarly. See Figure 2 for results in the case of strong anisotropy. See (354) for more details and Section J for simulation details.

From a simple computation, the first term in (354) is equal to one if and only if

$$\sum_{r=1}^{d} \alpha_r(t)\lambda_r = \frac{\mathrm{Tr}(K^2)}{\mathrm{Tr}(K)}. \tag{355}$$

One natural way for equality in (355) to hold is that the excess risk weights $\alpha_j(t)$ match each eigenspace's contribution to the trace. If K is nearly isotropic then this condition holds approximately so the spectral factor is nearly one and the mismatch is mainly driven by the ridge factor.

Conversely, in the presence of strong anisotropy, the spectral factor can be far from one. When the remaining ridge excess $\mathcal{E}(t) = \Phi_{\tilde{\lambda}^{\mathrm{eff}}(t)}(t) - \Phi_{\min}(\tilde{\lambda}^{\mathrm{eff}}(t))$ is concentrated in directions associated with small eigenvalues, we have $\sum_{r=1}^{d} \alpha_r(t)\lambda_r < \frac{\mathrm{Tr}(K^2)}{\mathrm{Tr}(K)}$ and, for a fixed ODE state, exact line search chooses a smaller stepsize than Polyak.

When the excess is sufficiently concentrated in directions associated with large eigenvalues, the opposite can occur. Thus anisotropy can cause greedy line search to deviate substantially from the Polyak stepsize. Under the normalization $\frac{1}{d}\mathrm{Tr}(K) = 1$, the spectral factor in (354) takes the form

$$\frac{\sum_{r=1}^{d} \alpha_r(t)\lambda_r + \tilde{\lambda}^{\mathrm{eff}}(t)}{\frac{1}{d}\mathrm{Tr}(K^2) + \tilde{\lambda}^{\mathrm{eff}}(t).}$$

Thus the discrepancy in the spectral factor is governed by how much $\sum_{j=1}^{d} \alpha_j(t)\lambda_j$ differs from $\frac{1}{d}\mathrm{Tr}(K^2)$ and by $\tilde{\lambda}^{\mathrm{eff}}(t)$. Strong anisotropy increases $\frac{1}{d}\mathrm{Tr}(K^2)$, while concentration of the remaining ridge excess directions associated with small eigenvalues decreases $\sum_{j=1}^{d} \alpha_j(t)\lambda_j$. Both effects make the line search stepsize smaller relative to the Polyak stepsize. In contrast, a large effective regularization parameter dampens this discrepancy. We emphasize this for K with the eigenvalues generated by Collins-Woodfin et al. (2024b)

$$\lambda_i = \frac{d}{\sum_{j=1}^{d}\left(\frac{j}{d+1}\right)^{-2/s}}\left(\frac{i}{d+1}\right)^{-2/s}, \qquad i = 1, \dots, d, \qquad s > 2. \tag{356}$$

By construction, $\frac{1}{d}\mathrm{Tr}(K) = 1$. As $s \downarrow 2$, the spectrum becomes more spread out and $\frac{1}{d}\mathrm{Tr}(K^2)$ increases. Consequently, when $\tilde{\lambda}^{\mathrm{eff}}(t)$ is small and the remaining ridge excess is not concentrated in high-curvature eigenspaces, the spectral factor can be far below one.

For a fixed ODE state, the three adversarial regimes (see Theorem G.2 and (Xing et al., 2021, Proposition 1)) determine how strongly this anisotropy can affect the same-state comparison. If $\delta \le \delta_-$, then $\tilde{\lambda}^{\mathrm{eff}}(t) \to \tilde{\lambda}_\star = 0$, so the damping from the ridge term disappears near the optimum. This is the regime most analogous to standard least squares as strong anisotropy can make exact line search substantially more conservative than Polyak (Collins-Woodfin et al., 2024b). If $\delta_- < \delta < \delta_+$, then $\tilde{\lambda}^{\mathrm{eff}}(t) \to \tilde{\lambda}_\star \in (0, \infty)$, so the ridge term partially damps the spectral discrepancy.

If $\delta \geq \delta_+$, then $\tilde{\lambda}^{\text{eff}}(t) \to \infty$ as $t \to \infty$ so the spectral factor approaches one. In this large δ regime, anisotropy no longer creates a large mismatch between line search and Polyak and any remaining discrepancy is driven by the ridge factor. See Figure 5 and Figure 2 for numerical evidence illustrating the impact of δ and anisotropy on the convergence of the ℓ_2-adversarial risk.

G ℓ_2-Adversarial Linear Regression Analysis

In this section, we will analyze the training dynamics of ℓ_2-adversarial least squares with regularization parameter $\lambda = 0$, with $\gamma_k = \gamma(t) \equiv \gamma$ and under the assumption that the data $a \sim \mathcal{N}(0, K)$. Given that this is equivalent to $p_1 = 1$ and $p_2 = 0$ with $\mu_1 = 0$, we drop the subscript i in this section. In this setting, the adversarial risk simplifies to

$$\mathcal{R}^{\text{adv}}(X) = \mathcal{R}(X) + \delta\sqrt{\frac{2}{\pi}}\|X\|\sqrt{2\mathcal{R}(X)} + \frac{\delta^2}{2}\|X\|^2 \tag{357}$$

From (446) and $\mathscr{R}^{\text{adv}}(t) = h \circ (\widehat{\mathscr{B}}(t))$, the det. equiv. of $\mathcal{R}^{\text{adv}}$ is given by

$$\mathscr{R}^{\text{adv}}(t) = \mathscr{R}(t) + \delta\sqrt{\frac{2}{\pi}}\sqrt{2\widehat{\mathscr{B}}_{44}(t)\mathscr{R}(t)} + \frac{\delta^2}{2}\widehat{\mathscr{B}}_{44}(t). \tag{358}$$

It is clear that we lose differentiability at $\mathscr{R}(t) = \frac{1}{2}\left(\mathscr{B}_{11}(t) - \mathscr{B}_{12} - \mathscr{B}_{21} + \mathscr{B}_{22}\right) + \frac{\eta^2}{2} = 0$ and $\widehat{\mathscr{B}}_{44}(t) = 0$. To apply the theory in this paper, we work on the set $\mathcal{U}$ defined as follows

$$\mathcal{U} \stackrel{\text{def}}{=} \left\{\widehat{B} : \widehat{B}_{11} - \widehat{B}_{12} - \widehat{B}_{21} + \widehat{B}_{22} + \eta > 0, \quad \widehat{B}_{44} > 0\right\} \tag{359}$$

As in Section E, we introduce the effective learning rate $\tilde{\gamma}^{\text{eff}}(t)$ and regularization parameter $\tilde{\lambda}^{\text{eff}}(t)$ whose dynamics are determined from the det. equiv. $q(t)$ of $q(X_k) \stackrel{\text{def}}{=} \left(\|X_k\|^2/(2\mathcal{R}(X_k))\right)^{1/2}$. Their respective det. equivs. are given by

$$q(t) \stackrel{\text{def}}{=} \sqrt{\frac{\widehat{\mathscr{B}}_{44}(t)}{2\mathscr{R}(t)}}, \quad \tilde{\gamma}^{\text{eff}}(t) = \gamma\left(1 + \delta\sqrt{\frac{2}{\pi}}q(t)\right) \quad \text{and} \quad \tilde{\lambda}^{\text{eff}}(t) = \frac{\delta^2 + \delta\sqrt{\frac{2}{\pi}}\frac{1}{q(t)}}{1 + \delta\sqrt{\frac{2}{\pi}}q(t)}. \tag{360}$$

We also introduce the following integral terms

$$\tilde{\Gamma}(t) = \int_0^t \tilde{\gamma}^{\text{eff}}(s)\,\mathrm{d}s \quad \text{and} \quad \Lambda(t) = \int_0^t \tilde{\gamma}^{\text{eff}}(s)\tilde{\lambda}^{\text{eff}}(s)\,\mathrm{d}s. \tag{361}$$

As derived in detail in Appendix I.1, $\mathscr{R}(t)$ and $\widehat{\mathscr{B}}_{44}(t)$ solve a system of coupled nonlinear Volterra equations

$$\begin{aligned}\mathscr{R}(t) = \frac{1}{2}\langle K, (e^{-K\tilde{\Gamma}(t)-\Lambda(t)}\left(X_0 + \left(\int_0^t K\tilde{\gamma}^{\text{eff}}(s)e^{K\tilde{\Gamma}(s)+\Lambda(s)}\,\mathrm{d}s\right)X^\star\right) - X^\star)^{\otimes 2}\rangle + \frac{\eta^2}{2} \\ + \frac{\gamma}{d}\int_0^t \tilde{\gamma}^{\text{eff}}(s)\left(\mathscr{R}(s) + \frac{\tilde{\lambda}^{\text{eff}}(s)}{2}\widehat{\mathscr{B}}_{44}(s)\right)\mathrm{Tr}\big(K^2 e^{-2K(\tilde{\Gamma}(t)-\tilde{\Gamma}(s))-2(\Lambda(t)-\Lambda(s))}\big)\,\mathrm{d}s.\end{aligned} \tag{362}$$

and

$$\begin{aligned}\widehat{\mathscr{B}}_{44}(t) = \left\|e^{-K\tilde{\Gamma}(t)-\Lambda(t)}\left(X_0 + \left(\int_0^t K\tilde{\gamma}^{\text{eff}}(s)e^{K\tilde{\Gamma}(s)+\Lambda(s)}\,\mathrm{d}s\right)X^\star\right)\right\|^2 \\ + \frac{2\gamma}{d}\int_0^t \tilde{\gamma}^{\text{eff}}(s)\left(\mathscr{R}(s) + \frac{\tilde{\lambda}^{\text{eff}}(s)}{2}\widehat{\mathscr{B}}_{44}(s)\right)\mathrm{Tr}\big(Ke^{-2K(\tilde{\Gamma}(t)-\tilde{\Gamma}(s))-2(\Lambda(t)-\Lambda(s))}\big)\,\mathrm{d}s.\end{aligned} \tag{363}$$

See Appendix I.1 for more details. Setting $\delta = 0$, then $\mathscr{R}(t)$ satisfies the linear convolution Volterra equation derived in Collins-Woodfin et al. (2024a); Collins-Woodfin & Paquette (2024); Paquette

et al. (2022). The central object of the dynamics is $q(t)$ which satisfies the following nonlinear ODE:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(q(t))^2 = \frac{\gamma}{\mathscr{R}(t)}(q(t))^2\left(1+\delta\sqrt{\frac{2}{\pi}}q(t)\right)\sum_{j=1}^{d}\lambda_j^2\left(\mathscr{V}_{j,11}(t)-2\mathscr{V}_{j,12}(t)+\mathscr{V}_{j,22}(t)\right)\\ +\frac{\gamma}{\mathscr{R}(t)}(\delta^2(q(t))^2-1)\sum_{j=1}^{d}\lambda_j\left(\mathscr{V}_{j,11}(t)-\mathscr{V}_{j,12}(t)\right)-2\gamma q(t)\left(\delta^2 q(t)+\delta\sqrt{\frac{2}{\pi}}\right)\\ +\frac{\gamma^2}{d}\left(1+2\delta\sqrt{\frac{2}{\pi}}q(t)+\delta^2(q(t))^2\right)(\mathrm{Tr}(K)-\mathrm{Tr}(K^2)(q(t))^2).\end{aligned} \tag{364}$$

We start by showing $q(t)$ is bounded above and below using a barrier argument conditionally on $\mathscr{R}(t)$ remaining positive.

Proposition G.1. *For any $T>0$, if $\mathscr{R}(t)>0$ for all $t\in[0,T]$ then for all $t\in[0,T]$*

$$\min(q(0),q_L)\le q(t)\le\max(q(0),q_U), \tag{365}$$

where q_L and q_U are defined as follows

$$\begin{aligned}q_L &= \min\left(\frac{1}{\delta},\sqrt{\frac{\mathrm{Tr}(K)}{2\,\mathrm{Tr}(K^2)}},\frac{\gamma\,\mathrm{Tr}(K)}{16d\left(\|K\|_{\mathrm{op}}^{1/2}+\delta\sqrt{\frac{2}{\pi}}\right)},\sqrt{\frac{\gamma\,\mathrm{Tr}(K)}{16d\,(2\|K\|_{\mathrm{op}}+\delta^2)}}\right)\\ q_U &= \max\left(1,\frac{1}{\delta},\sqrt{\frac{2\,\mathrm{Tr}(K)}{\mathrm{Tr}(K^2)}},\frac{8d\left(\|K\|_{\mathrm{op}}\left(1+\delta\sqrt{\frac{2}{\pi}}\right)+\delta^2(1+\|K\|_{\mathrm{op}}^{1/2})\right)}{\gamma\delta^2\,\mathrm{Tr}(K^2)}\right).\end{aligned} \tag{366}$$

Suppose further that there exists a $c>0$ such that $\frac{\mathrm{Tr}(K^2)}{d}\ge c$ uniformly in d and $0<c_0\le q(0)\le C_0<\infty$ where $c_0,C_0>0$ are independent of d. Then we obtain the following bounds uniform in d:

$$\begin{aligned}q_L &= \min\left(\frac{1}{\delta},\frac{1}{\sqrt{2\|K\|_{\mathrm{op}}}},\frac{c\gamma}{16\|K\|_{\mathrm{op}}\left(\|K\|_{\mathrm{op}}^{1/2}+\delta\sqrt{\frac{2}{\pi}}\right)},\sqrt{\frac{c\gamma}{16\|K\|_{\mathrm{op}}\,(2\|K\|_{\mathrm{op}}+\delta^2)}}\right),\\ q_U &= \max\left(1,\frac{1}{\delta},\sqrt{\frac{2\|K\|_{\mathrm{op}}}{c}},\frac{8\left(\|K\|_{\mathrm{op}}\left(1+\delta\sqrt{\frac{2}{\pi}}\right)+\delta^2(1+\|K\|_{\mathrm{op}}^{1/2})\right)}{\gamma\delta^2 c}\right).\end{aligned} \tag{367}$$

Proof. For simplicity of notation, we define the following notation

$$\begin{aligned}\mathrm{term}_1 &\stackrel{\mathrm{def}}{=} \frac{\gamma}{\mathscr{R}(t)}(q(t))^2\left(1+\delta\sqrt{\frac{2}{\pi}}q(t)\right)\sum_{j=1}^{d}\lambda_j^2\left(\mathscr{V}_{j,11}(t)-2\mathscr{V}_{j,12}(t)+\mathscr{V}_{j,22}(t)\right),\\ \mathrm{term}_2 &\stackrel{\mathrm{def}}{=} \frac{\gamma}{\mathscr{R}(t)}(\delta^2(q(t))^2-1)\sum_{j=1}^{d}\lambda_j\left(\mathscr{V}_{j,11}(t)-\mathscr{V}_{j,12}(t)\right),\\ \mathrm{term}_3 &\stackrel{\mathrm{def}}{=} -2\gamma q(t)\left(\delta^2 q(t)+\delta\sqrt{\frac{2}{\pi}}\right),\\ \mathrm{term}_4 &\stackrel{\mathrm{def}}{=} \frac{\gamma^2}{d}\left(1+2\delta\sqrt{\frac{2}{\pi}}q(t)+\delta^2(q(t))^2\right)(\mathrm{Tr}(K)-\mathrm{Tr}(K^2)(q(t))^2).\end{aligned} \tag{368}$$

We will refer to each term using ordinals. Suppose $0<q(t)\le q_L$. Since $\mathscr{V}_j(t)$ is psd then $\mathscr{V}_{j,11}(t)-2\mathscr{V}_{j,12}(t)+\mathscr{V}_{j,22}(t)\ge 0$ which implies the first term (i.e. term_1) is nonnegative. Since $q(t)\le q_L$ then

$$q(t)\le\sqrt{\frac{\mathrm{Tr}(K)}{2\,\mathrm{Tr}(K^2)}},$$

and thus

$$\operatorname{Tr}(K) - (q(t))^2 \operatorname{Tr}(K^2) \geq \frac{1}{2} \operatorname{Tr}(K). \tag{369}$$

Hence, we obtain

$$\text{term}_4 \geq \frac{\gamma^2}{2d} \operatorname{Tr}(K). \tag{370}$$

Since $q(t) \leq q_L$ then $q(t) \leq \frac{1}{\delta}$ which implies $|\delta^2 (q(t))^2 - 1| \leq 1$. From the second bound in Lemma I.2 then we obtain

$$\text{term}_2 \geq -\gamma \left(4\|K\|_{\text{op}} (q(t))^2 + 2\|K\|_{\text{op}}^{1/2} q(t) \right). \tag{371}$$

Hence, for all $0 \leq q(t) \leq q_L$ we have

$$\frac{\mathrm{d}}{\mathrm{d}t} (q(t))^2 \geq \frac{\gamma^2}{2d} \operatorname{Tr}(K) - 2\gamma \left(2\|K\|_{\text{op}} + \delta^2 \right) (q(t))^2 - 2\gamma \left(\|K\|_{\text{op}}^{1/2} + \delta \sqrt{\frac{2}{\pi}} \right) q(t). \tag{372}$$

Since $q(t) \leq q_L$ then we have

$$2\gamma \left(2\|K\|_{\text{op}} + \delta^2 \right) (q(t))^2 \leq \frac{\gamma^2}{8d} \operatorname{Tr}(K),$$

$$\text{and} \quad 2\gamma \left(\|K\|_{\text{op}}^{1/2} + \delta \sqrt{\frac{2}{\pi}} \right) q(t) \leq \frac{\gamma^2}{8d} \operatorname{Tr}(K).$$

Thus, whenever $0 \leq q(t) \leq q_L$ then

$$\frac{\mathrm{d}}{\mathrm{d}t} (q(t))^2 \geq \frac{\gamma^2}{4d} \operatorname{Tr}(K) \geq 0. \tag{373}$$

Now, let $q(t) \geq q_U$ then $q(t) \geq 1$, $q(t) \geq \frac{1}{\delta}$ and

$$q(t) \geq \sqrt{\frac{2 \operatorname{Tr}(K)}{\operatorname{Tr}(K^2)}},$$

from which it follows that

$$\operatorname{Tr}(K) - (q(t))^2 \operatorname{Tr}(K^2) \leq -\frac{1}{2} (q(t))^2 \operatorname{Tr}(K^2). \tag{374}$$

Thus, we obtain the bound for the fourth term

$$\text{term}_4 \leq -\frac{\gamma^2 \delta^2}{2d} (q(t))^4 \operatorname{Tr}(K^2). \tag{375}$$

Since $\lambda_j^2 \leq \|K\|_{\text{op}} \lambda_j$ then

$$\sum_{j=1}^{d} \lambda_j^2 \left(\mathscr{V}_{j,11}(t) - 2\mathscr{V}_{j,12}(t) + \mathscr{V}_{j,22}(t) \right) \leq 2\|K\|_{\text{op}} \mathscr{R}(t).$$

Using this and since $q(t) \geq 1$ then for the first term we obtain

$$\text{term}_1 \leq 2\gamma \|K\|_{\text{op}} \left(1 + \delta \sqrt{\frac{2}{\pi}} \right) (q(t))^3. \tag{376}$$

Since $q(t) \geq \frac{1}{\delta}$ then $\delta^2 (q(t))^2 - 1 \geq 0$. Using this, $q(t) \geq 1$ and the first bound in Lemma I.2, we obtain

$$\text{term}_2 \leq 2\gamma \delta^2 \left(\|K\|_{\text{op}}^{1/2} + 1 \right) (q(t))^3. \tag{377}$$

Finally, the third term is nonpositive so we drop it to obtain the upper bound

$$\frac{\mathrm{d}}{\mathrm{d}t} (q(t))^2 \leq 2\gamma \left(\|K\|_{\text{op}} \left(1 + \delta \sqrt{\frac{2}{\pi}} \right) + \delta^2 \left(\|K\|_{\text{op}}^{1/2} + 1 \right) \right) (q(t))^3 - \frac{\gamma^2 \delta^2}{2d} (q(t))^4 \operatorname{Tr}(K^2). \tag{378}$$

Since $q(t) \geq q_U$ then we have

$$2\gamma \left(\|K\|_{\mathrm{op}} \left(1 + \delta \sqrt{\frac{2}{\pi}} \right) + \delta^2 \left(\|K\|_{\mathrm{op}}^{1/2} + 1 \right) \right) (q(t))^3 \leq \frac{\gamma^2 \delta^2}{4d} (q(t))^4 \operatorname{Tr}(K^2),$$

from which it follows that for all $q(t) \geq q_U$

$$\frac{\mathrm{d}}{\mathrm{d}t} (q(t))^2 \leq -\frac{\gamma^2 \delta^2}{4d} (q(t))^4 \operatorname{Tr}(K^2) < 0. \tag{379}$$

Hence, by a first exit argument, we obtain the first bound. To obtain the bound uniform in d, we note that given there exists a $c > 0$ such that $\frac{1}{d} \operatorname{Tr}(K^2) \geq c > 0$ then

$$\operatorname{Tr}(K) = \sum_{j=1}^{d} \lambda_j \geq \frac{1}{\|K\|_{\mathrm{op}}} \sum_{j=1}^{d} \lambda_j^2 = \frac{\operatorname{Tr}(K^2)}{\|K\|_{\mathrm{op}}}.$$

Hence, $\frac{1}{d} \operatorname{Tr}(K) \geq \frac{c}{\|K\|_{\mathrm{op}}}$ from which it follows that

$$\frac{\operatorname{Tr}(K)}{\operatorname{Tr}(K^2)} \geq \frac{1}{\|K\|_{\mathrm{op}}} \quad \text{and} \quad \frac{\operatorname{Tr}(K)}{\operatorname{Tr}(K^2)} \leq \frac{\|K\|_{\mathrm{op}}}{c}. \tag{380}$$

□

Assuming that $X_0 \neq X^\star$ then we show that $q(t)$ is bounded above and below for all time $t > 0$. Note that this condition is not necessary if $\eta > 0$ as then $\mathscr{R}(t) \geq \frac{\eta^2}{2}$ for all $t \geq 0$.

Proposition G.2. *Suppose $\mathscr{R}(0), \widehat{\mathscr{B}}_{44}(0) > 0$ then for all $t \geq 0$, $\mathscr{R}(t) > 0$ from which it follows $q(t)$ is uniformly bounded by constants for all time $t \geq 0$*

$$\min(q(0), q_L) \leq q(t) \leq \max(q(0), q_U), \tag{381}$$

where q_L and q_U are constants which depend on $\operatorname{Tr}(K)$, $\operatorname{Tr}(K^2)$, γ and δ and are defined in Proposition G.1. Suppose further that there exists a $c > 0$ such that $\frac{\operatorname{Tr}(K^2)}{d} \geq c$ uniformly in d and $0 < c_0 \leq q(0) \leq C_0 < \infty$ where $c_0, C_0 > 0$ are independent of d then q_L and q_U are uniform bounds with respect to d.

Proof. Without loss of generality, we assume $q_L \leq q(0) \leq q_U$. We will proceed by contradiction. Given the definition of $\mathscr{R}^{\mathrm{Adv}}$, if $\mathscr{R}(t) > 0$ then $\mathscr{R}^{\mathrm{Adv}}(t) > 0$. Suppose there exists a $\tau > 0$ such that $\mathscr{R}(\tau) = 0$. Formally, τ is defined

$$\tau = \inf\{t > 0 : \mathscr{R}(t) = 0\} < \infty. \tag{382}$$

By assumption, we have $\mathscr{R}(0) > 0$ and since $\mathscr{R}(t)$ is continuous then $\mathscr{R}(t) > 0$ for every $t \in [0, \tau)$. By Proposition G.1 then $q(t)$ is well-defined on $[0, \tau)$ such that $0 < q_L \leq q(t) \leq q_U < \infty$. On $[0, \tau)$, we have

$$\tilde{\Gamma}'(t) = \gamma \left(1 + \delta \sqrt{\frac{2}{\pi}} q(t) \right) \leq \gamma \left(1 + \delta \sqrt{\frac{2}{\pi}} q_U \right).$$

Thus, we obtain

$$\tilde{\Gamma}(t) \leq \gamma \left(1 + \delta \sqrt{\frac{2}{\pi}} q_U \right) \tau,$$

and given $\tilde{\Gamma}(t)$ is nondecreasing then $\tilde{\Gamma}(\tau^-) = \lim_{t \uparrow \tau} \tilde{\Gamma}(t) < \infty$. Similarly, we have

$$\Lambda'(t) = \gamma \left(\delta^2 + \delta \sqrt{\frac{2}{\pi}} \frac{1}{q(t)} \right) \leq \gamma \left(\delta^2 + \delta \sqrt{\frac{2}{\pi}} \frac{1}{q_L} \right),$$

from which it follows as $\Lambda(t)$ is nondecreasing that $\Lambda(\tau^-) = \lim_{t \uparrow \tau} \Lambda(t) < \infty$. Set $0 < \varepsilon < \tau$. Then for every $s \in [0, \varepsilon]$, we have $\mathscr{R}(s) > 0$. Since $\mathscr{R}$ is continuous then it attains a minimum on $[0, \varepsilon]$ which we denote $m_\varepsilon = \min_{0 \leq s \leq \varepsilon} \mathscr{R}(s)$. Now, for any $\varepsilon \leq t < \tau$ we have

$$0 \leq \tilde{\Gamma}(t) - \tilde{\Gamma}(s) \leq \tilde{\Gamma}(\tau^-) - \tilde{\Gamma}(0) = \tilde{\Gamma}(\tau^-),$$
$$0 \leq \Lambda(t) - \Lambda(s) \leq \Lambda(\tau^-) - \Lambda(0) = \Lambda(\tau^-),$$

from which it follows that

$$e^{-2\lambda_j(\tilde{\Gamma}(t)-\tilde{\Gamma}(s))-2(\Lambda(t)-\Lambda(s))} \geq e^{-2(\|K\|_{\mathrm{op}}\tilde{\Gamma}(\tau^-)+\Lambda(\tau^-))}.$$

Thus, we have $\mathrm{Tr}\big(K^2 e^{-2K(\tilde{\Gamma}(t)-\tilde{\Gamma}(s))-2(\Lambda(t)-\Lambda(s))}\big) \geq e^{-2(\|K\|_{\mathrm{op}}\tilde{\Gamma}(\tau^-)+\Lambda(\tau^-))}\,\mathrm{Tr}(K^2) > 0$ which implies from (456)

$$\begin{aligned}
\mathscr{R}(t) &\geq \frac{\gamma^2}{d}\int_0^t \mathscr{R}(s)\mathrm{Tr}\big(K^2 e^{-2K(\tilde{\Gamma}(t)-\tilde{\Gamma}(s))-2(\Lambda(t)-\Lambda(s))}\big)\,\mathrm{d}s \\
&\geq \frac{\gamma^2}{d}\int_0^{\varepsilon} m_\varepsilon e^{-2(\|K\|_{\mathrm{op}}\tilde{\Gamma}(\tau^-)+\Lambda(\tau^-))}\,\mathrm{Tr}(K^2)\,\mathrm{d}s \\
&\geq \frac{\gamma^2}{d}\,\mathrm{Tr}(K^2) m_\varepsilon e^{-2(\|K\|_{\mathrm{op}}\tilde{\Gamma}(\tau^-)+\Lambda(\tau^-))}\varepsilon > 0.
\end{aligned} \tag{383}$$

Hence, $\mathscr{R}(t) > 0$ for all $t \in [\varepsilon, \tau)$ which implies since $\mathscr{R}(t)$ is continuous that $\mathscr{R}(\tau) = \lim_{t\uparrow\tau}\mathscr{R}(t) \geq \frac{\gamma^2}{d}\,\mathrm{Tr}(K^2) m_\varepsilon e^{-2(\|K\|_{\mathrm{op}}\tilde{\Gamma}(\tau^-)+\Lambda(\tau^-))}\varepsilon > 0$ which is a contradiction. Therefore, $\mathscr{R}(t) > 0$ for every $t > 0$. Hence, Proposition G.1 is valid on $[0, \infty)$. □

Before the proof Theorem G.1, we introduce the following notation to simplify the derivations:

$$\begin{gathered}
\rho_j(t) \stackrel{\mathrm{def}}{=} e^{-\lambda_j\tilde{\Gamma}(t)-\Lambda(t)}\left(x_{0,j} + \lambda_j x_j^\star \int_0^t \tilde{\gamma}^{\mathrm{eff}}(s) e^{\lambda_j\tilde{\Gamma}(s)+\Lambda(s)}\,\mathrm{d}s\right), \\
\text{where} \quad x_{0,j} \stackrel{\mathrm{def}}{=} X_0^\top \omega_j \quad \text{and} \quad x_j^\star \stackrel{\mathrm{def}}{=} (X^\star)^\top \omega_j.
\end{gathered} \tag{384}$$

Define $\ell(u) \stackrel{\mathrm{def}}{=} \Lambda(\tilde{\Gamma}^{-1}(u))$ then $\ell'(u) = \tilde{\lambda}^{\mathrm{eff}}(\tilde{\Gamma}^{-1}(u))$. We now express $\rho_j(t)$ as a function of $\tilde{\Gamma}$ using the time change $u = \tilde{\Gamma}(s)$:

$$\rho_j(u) \stackrel{\mathrm{def}}{=} e^{-\lambda_j u-\ell(u)}\left(x_{0,j} + \lambda_j x_j^\star \int_0^u e^{\lambda_j s+\ell(s)}\,\mathrm{d}s\right). \tag{385}$$

We will work with each definition of ρ_j depending on the context. Given that $\frac{\mathrm{d}}{\mathrm{d}s}e^{\lambda_j s+\ell(s)} = (\lambda_j + \ell'(s))e^{\lambda_j s+\ell(s)}$ then

$$\int_0^u \lambda_j e^{\lambda_j s+\ell(s)}\,\mathrm{d}s = e^{\lambda_j u+\ell(u)} - 1 - \int_0^u \ell'(s) e^{\lambda_j s+\ell(s)}\,\mathrm{d}s.$$

From this identity, we also obtain the representation

$$\rho_j(u) \stackrel{\mathrm{def}}{=} x_j^\star + e^{-\lambda_j u-\ell(u)}\left(x_{0,j} - x_j^\star\right) - x_j^\star \int_0^u \ell'(s) e^{-\lambda_j(u-s)-(\ell(u)-\ell(s))}\,\mathrm{d}s. \tag{386}$$

Using these identities, we define the following functions

$$\begin{aligned}
F(u) &\stackrel{\mathrm{def}}{=} \frac{1}{2}\sum_{j=1}^d \lambda_j(\rho_j(u) - x_j^\star)^2 \\
&= \frac{1}{2}\sum_{j=1}^d \lambda_j(e^{-\lambda_j u-\ell(u)}\left(x_{0,j} - x_j^\star\right) - x_j^\star \int_0^u \ell'(s) e^{-\lambda_j(u-s)-(\ell(u)-\ell(s))}\,\mathrm{d}s)^2,
\end{aligned} \tag{387}$$

and

$$L(u) \stackrel{\mathrm{def}}{=} \sum_{j=1}^d (\rho_j(u))^2 = \sum_{j=1}^d e^{-2(\lambda_j u+\ell(u))}\left(x_{0,j} + \lambda_j x_j^\star \int_0^u e^{\lambda_j s+\ell(s)}\,\mathrm{d}s\right)^2. \tag{388}$$

We also introduce the following notation

$$\begin{gathered}
M_j(t) \stackrel{\mathrm{def}}{=} \gamma^2 \int_0^t \mathscr{R}^{\mathrm{adv}}(s)\lambda_j^2 e^{-2\lambda_j(\tilde{\Gamma}(t)-\tilde{\Gamma}(s))-2(\Lambda(t)-\Lambda(s))}\,\mathrm{d}s, \\
\text{and} \quad N_j(t) \stackrel{\mathrm{def}}{=} 2\gamma^2 \int_0^t \mathscr{R}^{\mathrm{adv}}(s)\lambda_j e^{-2\lambda_j(\tilde{\Gamma}(t)-\tilde{\Gamma}(s))-2(\Lambda(t)-\Lambda(s))}\,\mathrm{d}s.
\end{gathered}$$

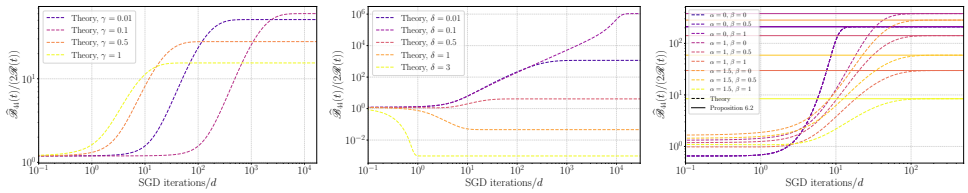

Figure 6: **Numerical evidence that** $q(t) \stackrel{\text{def}}{=} \sqrt{\frac{\widehat{\mathscr{B}}_{44}(t)}{2\mathscr{R}(t)}}$ **converges** for noiseless ℓ_2-adversarial least squares. The first (left) and second (middle) plots provide evidence that $q(t)$ converges for a variety of constant learning rates γ and δ. Here we fix either parameter and vary the other according to the values presented in the legends. The third plot provides evidence that $q(t)$ converges for a variety of $X^\star$ and covariances K under different power law setups (see (486)) and validates that the fixed point equation in Theorem G.1 predicts the limiting value q_∞. See Appendix J for simulation details.

This leads to

$$\mathscr{R}(t) = F(\tilde{\Gamma}(t)) + \frac{1}{d}\sum_{j=1}^{d} M_j(t), \quad \text{and} \quad \widehat{\mathscr{B}}_{44}(t) = L(\tilde{\Gamma}(t)) + \frac{1}{d}\sum_{j=1}^{d} N_j(t).$$

We are now ready to prove Theorem G.1.

Theorem G.1. *Suppose* $\mathscr{R}(0), \widehat{\mathscr{B}}_{44}(0) > 0$, $(X^\star)^\top K X^\star > 0$, $\sup_{t\geq 0} \mathscr{R}^{\text{Adv}}(t) < \infty$ *and* $q(t) \to q_\infty \in (0,\infty)$, $\delta > 0$. *If, for all* $q \in [\min(q(0), q_L), \max(q(0), q_U)]$, *the following condition is satisfied*

$$\frac{B_d(q)}{A_d(q)} \geq \max\left(\frac{P_\delta(q)}{Q_\delta(q)}, 0\right) \stackrel{\text{def}}{=} \psi_\delta(q), \tag{389}$$

where

$$A_d(q) \stackrel{\text{def}}{=} \frac{1}{d}\sum_{j=1}^{d} \frac{\lambda_j}{\delta^2\left(\delta q + \sqrt{\frac{2}{\pi}} + r_j\delta q(1+\sqrt{\frac{2}{\pi}}\delta q)\right)^2},$$

$$B_d(q) \stackrel{\text{def}}{=} \frac{1}{d}\sum_{j=1}^{d} \frac{\lambda_j^2}{\delta^4\left(\delta q + \sqrt{\frac{2}{\pi}} + r_j\delta q(1+\sqrt{\frac{2}{\pi}}\delta q)\right)^2}$$

$$P_\delta(q) \stackrel{\text{def}}{=} \left(2 - \frac{6}{\pi}\right)(\delta q)^4 + \left(3\sqrt{\frac{2}{\pi}} - 7\left(\frac{2}{\pi}\right)^{3/2}\right)(\delta q)^3$$
$$- \left(4\left(\frac{2}{\pi}\right)^2 + \frac{4}{\pi}\right)(\delta q)^2 - \left(3\left(\frac{2}{\pi}\right)^{3/2} + \sqrt{\frac{2}{\pi}}\right)(\delta q) - \frac{2}{\pi},$$

and $\quad Q_\delta(q) \stackrel{\text{def}}{=} \frac{2}{\pi}(\delta q)^6 + \left(7\left(\frac{2}{\pi}\right)^{3/2} - \sqrt{\frac{2}{\pi}}\right)(\delta q)^5 + \left(8\left(\frac{2}{\pi}\right)^2 + \frac{16}{\pi} - 2\right)(\delta q)^4$
$$+ \left(15\left(\frac{2}{\pi}\right)^{3/2} + \sqrt{\frac{2}{\pi}}\right)(\delta q)^3 + \frac{18}{\pi}(\delta q)^2 + 2\sqrt{\frac{2}{\pi}}\delta q.$$

Then q_∞ *is the unique fixed point satisfying the fixed point equation*

$$q_\infty^2 = (1 - G(q_\infty)C_{\mathscr{R}}(q_\infty))\frac{L_\infty(q_\infty)}{2F_\infty(q_\infty)} + \frac{1}{2}G(q_\infty)C_{\mathscr{B}}(q_\infty) \tag{390}$$

and the admissibility condition $G(q_\infty)C_{\mathscr{R}}(q_\infty) < 1$ *for* $G(q) = 1 + 2\delta\sqrt{\frac{2}{\pi}}q + \delta^2 q^2$ *and* $C_{\mathscr{R}}(q) = \frac{\gamma}{2d}\sum_{j=1}^{d}\frac{\lambda_j^2}{D_j(q)}$ *with* $D_j(q) = \lambda_j(1 + \delta\sqrt{\frac{2}{\pi}}q) + \delta^2 + \delta\sqrt{\frac{2}{\pi}}\frac{1}{q}$. *The limiting values of* $\mathscr{R}$, $\widehat{\mathscr{B}}_{44}$ *and* $\mathscr{R}^{\text{Adv}}$ *are given by*

$$\mathscr{R}(\infty) = \frac{F_\infty(q_\infty)}{1 - G(q_\infty)C_{\mathscr{R}}(q_\infty)}, \quad \widehat{\mathscr{B}}_{44}(\infty) = 2q_\infty^2\mathscr{R}(\infty), \quad \text{and} \quad \mathscr{R}^{\text{Adv}}(\infty) = G(q_\infty)\mathscr{R}(\infty),$$

where $C_{\mathscr{B}}(q) = \frac{\gamma}{d}\sum_{j=1}^{d}\frac{\lambda_j}{D_j(q)}$, $F_\infty(q) = \frac{1}{2}\sum_{j=1}^{d}\lambda_j\left(\frac{\tilde\lambda^{\rm eff}(q)}{\lambda_j+\tilde\lambda^{\rm eff}(q)}\right)^2(x_j^\star)^2$, $L(\infty) = \sum_{j=1}^{d}\left(\frac{\lambda_j}{\lambda_j+\tilde\lambda^{\rm eff}(\infty)}\right)^2(x_j^\star)^2$, *and* $\tilde\lambda^{\rm eff}(q) = \frac{\delta^2+\delta\sqrt{\frac{2}{\pi}}\frac{1}{q}}{1+\delta\sqrt{\frac{2}{\pi}}q}$. *In the case of isotropic covariance, the condition in* (389) *simplifies to*

$$\delta^2 \geq \max\left(\frac{P_\delta(q)}{Q_\delta(q)}, 0\right).$$

Furthermore, let $r_\star \overset{\rm def}{=} \sup_{q>0}\psi_\delta(q)$ *with* $r_\star \approx 8.35\times10^{-5}$. *Then a less precise sufficient condition than* (389) *is*

$$\delta \leq \sqrt{\frac{\lambda_{\min}^+}{r_\star}},$$

which leads to $\delta \leq r_\star^{-1/2}$ *for isotropic covariance.*

Proof. Recall we have

$$\tilde\Gamma'(t) = \tilde\gamma^{\rm eff}(t) = \gamma\left(1+\delta\sqrt{\frac{2}{\pi}}q(t)\right) \quad \text{and} \quad \Lambda'(t) = \tilde\gamma^{\rm eff}(t)\tilde\lambda^{\rm eff}(t) = \gamma\left(\delta^2+\delta\sqrt{\frac{2}{\pi}}\frac{1}{q(t)}\right). \tag{391}$$

Recall $\ell(u) = \Lambda(\tilde\Gamma^{-1}(u))$ then $\ell'(u) = \tilde\lambda^{\rm eff}(\tilde\Gamma^{-1}(u))$. Given that $\tilde\Gamma(t) \geq \gamma t$ we have $\tilde\Gamma^{-1}(u) \to \infty$ when $u \to \infty$ which implies $q(\tilde\Gamma^{-1}(u)) \to q_\infty$ and

$$\ell'(u) \to \tilde\lambda^{\rm eff}(\infty) = \frac{\delta^2+\delta\sqrt{\frac{2}{\pi}}\frac{1}{q_\infty}}{1+\delta\sqrt{\frac{2}{\pi}}q_\infty}.$$

We now compute the limits of $F(u)$ and $L(u)$. From the definition of $\rho_j(u)$, it satisfies the following ODE

$$\frac{\mathrm{d}}{\mathrm{d}u}\rho_j(u) = -(\lambda_j+\ell'(u))\rho_j(u) + \lambda_j x_j^\star.$$

We will show that $\rho_j(u) \to \rho_j(\infty) \overset{\rm def}{=} \frac{\lambda_j}{\lambda_j+\tilde\lambda^{\rm eff}(\infty)}x_j^\star$. Let $y_j(u) \overset{\rm def}{=} \rho_j(u) - \rho_j(\infty)$ then $y_j'(u) = \rho_j'(u)$ from which we obtain after substituting $\rho_j(u) = y_j(u)+\rho_j(\infty)$ and $\lambda_j x_j^\star = (\lambda_j+\tilde\lambda^{\rm eff}(\infty))\rho_j(\infty)$

$$y_j'(u) = -(\lambda_j+\ell'(u))y_j(u) + (\tilde\lambda^{\rm eff}(\infty)-\ell'(u))\rho_j(\infty) \tag{392}$$

This ODE takes the form $y'(u) = -A(u)y(u)+B(u)$ with $A_j(u) = (\lambda_j+\ell'(u))$ and $B_j(u) = (\tilde\lambda^{\rm eff}(\infty)-\ell'(u))\rho_j(\infty)$. Since $\ell'(u) \to \tilde\lambda^{\rm eff}(\infty)$ and $\lambda_j+\tilde\lambda^{\rm eff}(\infty) > 0$, for large enough u, $A_j(u)$ is bounded below by a positive constant while $B_j(u) \to 0$ as $u \to \infty$. Thus, from variation of constants, we obtain $y_j(u) \to 0$ hence $\rho_j(u) \to \frac{\lambda_j}{\lambda_j+\tilde\lambda^{\rm eff}(\infty)}x_j^\star$.

From the definitions of F and L, we obtain

$$F(\infty) = \frac{1}{2}\sum_{j=1}^{d}\lambda_j\left(\frac{\tilde\lambda^{\rm eff}(\infty)}{\lambda_j+\tilde\lambda^{\rm eff}(\infty)}\right)^2(x_j^\star)^2 \quad \text{and} \quad L(\infty) = \sum_{j=1}^{d}\left(\frac{\lambda_j}{\lambda_j+\tilde\lambda^{\rm eff}(\infty)}\right)^2(x_j^\star)^2.$$

Now, the adversarial risk can be written as $\mathscr{R}^{\rm Adv}(t) = G(q(t))\mathscr{R}(t)$ for $G(q) = 1+2\delta\sqrt{\frac{2}{\pi}}q+\delta^2q^2$. Since $\sup_{t\geq0}\mathscr{R}^{\rm Adv}(t) < C < \infty$ for some constant $C > 0$ then we have the uniform bounds

$$\begin{aligned} 0 \leq M_j(t) &\leq \gamma^2\lambda_j^2 C\int_0^\infty e^{-2\gamma(\lambda_j+\delta^2)u}\,\mathrm{d}u = \frac{C\gamma\lambda_j^2}{2(\lambda_j+\delta^2)}, \\ 0 \leq N_j(t) &\leq 2\gamma^2\lambda_j C\int_0^\infty e^{-2\gamma(\lambda_j+\delta^2)u}\,\mathrm{d}u = \frac{C\gamma\lambda_j}{\lambda_j+\delta^2}. \end{aligned} \tag{393}$$

Proposition G.2 also implies there exists $q_-, q_+ > 0$ such that $q_- \leq q(t) \leq q_+$. Differentiating $M_j(t)$ and $N_j(t)$, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}M_j(t) = -2\gamma D_j(q(t))M_j(t) + \gamma^2 \mathscr{R}^{\mathrm{Adv}}(t)\lambda_j^2,$$
$$\frac{\mathrm{d}}{\mathrm{d}t}N_j(t) = -2\gamma D_j(q(t))N_j(t) + 2\gamma^2 \mathscr{R}^{\mathrm{Adv}}(t)\lambda_j,$$

where we define $D_j(q) \stackrel{\mathrm{def}}{=} \lambda_j \left(1 + \delta\sqrt{\frac{2}{\pi}}q\right) + \delta^2 + \delta\sqrt{\frac{2}{\pi}}\frac{1}{q}$. Since

$$\mathscr{R}^{\mathrm{Adv}}(t) = G(q(t))\mathscr{R}(t) = G(q(t))\left(F(\tilde{\Gamma}(t)) + \frac{1}{d}\sum_{j=1}^{d} M_j(t)\right)$$

then

$$\frac{\mathrm{d}}{\mathrm{d}t}M_j(t) = -2\gamma D_j(q(t))M_j(t) + \gamma^2\lambda_j^2 G(q(t))\left(F(\tilde{\Gamma}(t)) + \frac{1}{d}\sum_{k=1}^{d} M_k(t)\right) \tag{394}$$

Now, suppose $M_j(u) \to M_j(\infty)$ then the stationary equation would be

$$0 = -2\gamma D_j(q_\infty)M_j(\infty) + \gamma^2\lambda_j^2 G(q_\infty)\left(F(\infty) + \frac{1}{d}\sum_{j=1}^{d} M_j(\infty)\right). \tag{395}$$

Let $S(t) = \frac{1}{d}\sum_{j=1}^{d} M_j(t)$ then $M_j(\infty) = \frac{\gamma\lambda_j^2}{2D_j(q_\infty)}G(q_\infty)\left(F(\infty) + S(\infty)\right)$. Summing over j and multiplying by $\frac{1}{d}$, we obtain

$$S(\infty) = G(q_\infty)C_{\mathscr{R}}(q_\infty)(F(\infty) + S(\infty)). \tag{396}$$

Thus, if $G(q_\infty)C_{\mathscr{R}}(q_\infty) < 1$ then

$$S(\infty) = \frac{G(q_\infty)C_{\mathscr{R}}(q_\infty)}{1 - G(q_\infty)C_{\mathscr{R}}(q_\infty)}F(\infty). \tag{397}$$

Given that $\mathscr{R}(t) = F(\tilde{\Gamma}(t)) + S(t)$ then the candidate limit for $\mathscr{R}$ is

$$\mathscr{R}(\infty) = F(\infty) + S(\infty) = \frac{F(\infty)}{1 - G(q_\infty)C_{\mathscr{R}}(q_\infty)}. \tag{398}$$

We will now show $M_j(t) \to M_j(\infty)$ as $t \to \infty$. Let $y_j(t) = M_j(t) - M_j(\infty)$. Subtracting (395) from (394), we obtain

$$y_j'(t) = -2\gamma D_j(q_\infty)y_j(t) + \gamma^2\lambda_j^2 G(q_\infty)\sum_{k=1}^{d} y_k(t) + \mathrm{err}_j(t),$$

where

$$\mathrm{err}_j(t) = \gamma^2\lambda_j^2 G(q(\infty))(F(\tilde{\Gamma}(t)) - F(\infty)) + \gamma^2\lambda_j^2\left(G(q(t)) - G(q(\infty))\right)\mathscr{R}(t)$$
$$- 2\gamma\left(\lambda_j\delta\sqrt{\frac{2}{\pi}}(q(t) - q_\infty) + \delta\sqrt{\frac{2}{\pi}}\left(\frac{1}{q(t)} - \frac{1}{q_\infty}\right)\right)M_j(t).$$

Since $\mathscr{R}(t) \leq \sup_{t\geq 0}\mathscr{R}^{\mathrm{Adv}}(t) < \infty$, $M_j(t)$ are uniformly bounded, $F(\tilde{\Gamma}(t)) \to F(\infty)$ and $q(t) \to q_\infty$ then the error term $\mathrm{err}_j(t) \to 0$ as $t \to \infty$. Now, the vector $y(t)$ composed of $y_j(t)$ satisfies the following ODE

$$y'(t) = A_\infty y(t) + \mathrm{err}(t),$$

where A_∞ is defined such that

$$[A_\infty]_{jk} = -2\gamma D_j(q_\infty)\mathbb{1}_{jk} + \gamma^2\lambda_j^2 G(q_\infty),$$

where $\mathbb{1}_{jk} = 1$ when $j = k$ and 0 if not. Hence, we have $A_\infty = -Q + u\mathbf{1}^\top$ where $\mathbf{1}$ is a vector of 1, D is diagonal such that

$$Q_{jj} = 2\gamma D_j(q_\infty) \quad \text{and} \quad u_j = \gamma^2\lambda_j^2 G(q_\infty).$$

The condition $G(q_\infty)C_{\mathscr{R}}(q_\infty) < 1$ is equivalent to $\mathbf{1}^\top Q^{-1}u < 1$ as $\mathbf{1}^\top Q^{-1}u = G(q_\infty)C_{\mathscr{R}}(q_\infty)$. We will now show that A_∞ is a Hurwitz-stable matrix. Let υ be an eigenvalue of A_∞ then $\det(A_\infty - \upsilon I_d) = 0$.

We proceed by contradiction. Suppose there is an eigenvalue υ with $\Re[\upsilon] \geq 0$. Then $-Q - \upsilon I_d$ is invertible which implies $\det(A_\infty - \upsilon I_d) = \det(-Q - \upsilon I_d)(1 - 1^\top (Q + \upsilon I_d)^{-1}u)$.

Now, every diagonal entry of $-Q - \upsilon I_d$ has negative real part hence $\det(-Q - \upsilon I_d) \neq 0$. Since υ is an eigenvalue then this enforces that $1 = 1^\top (Q + \upsilon I_d)^{-1}u$. Since $\Re[\upsilon] \geq 0$ and $Q_{jj} > 0$ then taking absolute values, we obtain

$$1 \leq \sum_{j=1}^{d} \frac{u_j}{|\upsilon + Q_{jj}|} \leq \sum_{j=1}^{d} \frac{u_j}{Q_{jj}} = \mathbf{1}^\top Q^{-1}u < 1,$$

which is a contradiction. Hence A_∞ is Hurwitz and it follows that there exists $C > 0$ and $\kappa > 0$ such that $\|e^{A_\infty t}\| \leq Ce^{-\kappa t}$. Since $\mathrm{err}(t) \to 0$ then by variation of constants then $y(t) \to 0$ which implies $M_j(t) \to M_j(\infty)$.

Thus, it follows that $\mathscr{R}(t) \to \mathscr{R}(\infty)$ and since $q(t) \to q_\infty$ then $\mathscr{R}^{\mathrm{Adv}}(t) \to \mathscr{R}^{\mathrm{Adv}}(\infty) = \mathscr{R}(\infty)G(q_\infty)$ and $\widehat{\mathscr{B}}_{44}(t) \to \widehat{\mathscr{B}}_{44}(\infty) = 2q_\infty^2 \mathscr{R}(\infty)$. We could also use a similar argument to show that $\widehat{\mathscr{B}}_{44}(t) \to \widehat{\mathscr{B}}_{44}(\infty)$ as $t \to \infty$ with

$$\widehat{\mathscr{B}}_{44}(\infty) = L(\infty) + G(q_\infty)C_{\mathscr{B}}(q_\infty)\mathscr{R}(\infty) \quad \text{where} \quad C_{\mathscr{B}}(q) = \frac{\gamma}{d}\sum_{j=1}^{d} \frac{\lambda_j}{D_j(q)}. \tag{399}$$

Finally, we have $q_\infty^2 = \frac{\widehat{\mathscr{B}}_{44}(\infty)}{2\mathscr{R}(\infty)}$. Plugging (398) and (399) into this expression, we obtain

$$q_\infty^2 = (1 - G(q_\infty)C_{\mathscr{R}}(q_\infty))\frac{L_\infty(q_\infty)}{2F_\infty(q_\infty)} + \frac{1}{2}G(q_\infty)C_{\mathscr{B}}(q_\infty), \tag{400}$$

where we emphasize the dependence on q_∞ by writing $L_\infty(q_\infty) \overset{\mathrm{def}}{=} L(\infty)$ and $F_\infty(q_\infty) \overset{\mathrm{def}}{=} F(\infty)$. We now prove uniqueness of the admissible fixed point. We have

$$\frac{\mathrm{d}}{\mathrm{d}q}\big(q\tilde{\lambda}^{\mathrm{eff}}(q)\big) = \frac{\delta^2 - \delta^2\frac{2}{\pi}}{(1 + \delta\sqrt{\frac{2}{\pi}}q)^2} = \frac{\delta^2(1 - \frac{2}{\pi})}{\left(1 + \delta\sqrt{\frac{2}{\pi}}q\right)^2} > 0.$$

Thus $q \mapsto (q\tilde{\lambda}^{\mathrm{eff}}(q))^2$ is strictly increasing. Multiplying the fixed point equation by $\tilde{\lambda}^{\mathrm{eff}}(q)^2$, it is sufficient to show that

$$\Theta(q) \overset{\mathrm{def}}{=} \tilde{\lambda}^{\mathrm{eff}}(q)^2 \left[(1 - G(q)C_{\mathscr{R}}(q))\frac{L_\infty(q)}{2F_\infty(q)} + \frac{1}{2}G(q)C_{\mathscr{B}}(q)\right]$$

is strictly decreasing on $\mathcal{I} = \{q > 0 : G(q)C_{\mathscr{R}}(q) < 1\}$ as the fixed point equation is equivalent to $(q\tilde{\lambda}^{\mathrm{eff}}(q))^2 = \Theta(q)$. First, we have

$$\frac{\mathrm{d}}{\mathrm{d}q}(\tilde{\lambda}^{\mathrm{eff}}(q)) = -\frac{\delta\sqrt{\frac{2}{\pi}}\left(\frac{1}{q^2} + 2\delta\sqrt{\frac{2}{\pi}}\frac{1}{q} + \delta^2\right)}{\left(1 + \delta\sqrt{\frac{2}{\pi}}q\right)^2} < 0.$$

Next, define the functions

$$A(r) = \sum_{j=1}^{d} \lambda_j w_j(r), \quad \text{with} \quad w_j(r) = \frac{\frac{\lambda_j (x_j^\star)^2}{(\lambda_j + r)^2}}{\sum_{k=1}^{d} \frac{\lambda_k (x_k^\star)^2}{(\lambda_k + r)^2}},$$

such that $\tilde{\lambda}^{\mathrm{eff}}(q)^2 \frac{L_\infty(q)}{2F_\infty(q)} = A(\tilde{\lambda}^{\mathrm{eff}}(q))$. Differentiating we obtain

$$\frac{\mathrm{d}}{\mathrm{d}r}A(r) = 2\sum_{1 \leq i < j \leq d} w_i(r)w_j(r)\frac{(\lambda_i - \lambda_j)^2}{(\lambda_i + r)(\lambda_j + r)}.$$

Apart from some degenerate cases such as $X^\star = c\omega_j$ with $\lambda_j = 0$ or $K = 0$ then $\frac{\mathrm{d}}{\mathrm{d}r}A(r) \geq 0$. From the chain rule then we obtain

$$\frac{\mathrm{d}}{\mathrm{d}q}A(\tilde{\lambda}^{\mathrm{eff}}(q)) = A'(\tilde{\lambda}^{\mathrm{eff}}(q))\frac{\mathrm{d}}{\mathrm{d}q}(\tilde{\lambda}^{\mathrm{eff}}(q)) \leq 0.$$

Finally, $q \mapsto G(q)C_{\mathscr{R}}(q)$ is strictly increasing, because for each j,

$$\frac{\mathrm{d}}{\mathrm{d}q}\left(\frac{G(q)}{D_j(q)}\right) = \frac{\delta}{q^2 D_j(q)^2}\left(2\delta^3 q^3 + \delta^2\lambda_j q^4\sqrt{\frac{2}{\pi}} + 5\delta^2 q^2\sqrt{\frac{2}{\pi}}\right.$$
$$\left. +2\delta\lambda_j q^3 + \delta q\frac{8}{\pi} + \lambda_j q^2\sqrt{\frac{2}{\pi}} + \sqrt{\frac{2}{\pi}}\right) > 0.$$

Therefore $1 - G(q)C_{\mathscr{R}}(q)$ is strictly decreasing and positive on $\mathcal{I}$ from which it follows that the first term in $\Theta(q)$ is nonincreasing on $\mathcal{I}$. For the second term, we have

$$\frac{(\tilde{\lambda}^{\mathrm{eff}}(q))^2}{2}G(q)C_{\mathscr{B}} = \frac{\gamma}{2d}\sum_{j=1}^{d} T_{\lambda_j,\delta}(q),$$

where we define

$$T_{\lambda,\delta}(q) = \lambda\frac{\left(\delta^2 + \delta\sqrt{\frac{2}{\pi}}\frac{1}{q}\right)^2\left(1 + 2\delta\sqrt{\frac{2}{\pi}}q + \delta^2 q^2\right)}{\lambda\left(1 + \delta\sqrt{\frac{2}{\pi}}q\right)^3 + \left(\delta^2 + \delta\sqrt{\frac{2}{\pi}}\frac{1}{q}\right)\left(1 + \delta\sqrt{\frac{2}{\pi}}q\right)^2}.$$

Our goal is to find a sufficient condition on δ for which $\frac{d}{dq}\sum_{j=1}^{d} T_{\lambda_j,\delta}(q) \leq 0$. For a single eigenvalue, define $r \stackrel{\mathrm{def}}{=} \frac{\lambda}{\delta^2}$. and note that

$$\delta^2 + \delta\sqrt{\frac{2}{\pi}}\frac{1}{q} = \delta^2\frac{\delta q + \sqrt{\frac{2}{\pi}}}{\delta q}.$$

Substituting this and $\lambda = \delta^2 r$ into $T_{\lambda,\delta}$, we obtain after a simple computation

$$T_{\lambda,\delta}(q) = \delta^3\frac{r\left(\delta q + \sqrt{\frac{2}{\pi}}\right)^2\left(1 + 2\sqrt{\frac{2}{\pi}}\delta q + \delta^2 q^2\right)}{q\left(1 + \delta\sqrt{\frac{2}{\pi}}q\right)^2\left(\delta q r\left(1 + \delta\sqrt{\frac{2}{\pi}}q\right) + \delta q + \sqrt{\frac{2}{\pi}}\right)}.$$

Differentiating with respect to q, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}q}T_{\lambda,\delta}(q) = \delta^3\frac{r\left(\delta q + \sqrt{\frac{2}{\pi}}\right)\{P_\delta(q) - rQ_\delta(q)\}}{q^2\left(1 + \delta\sqrt{\frac{2}{\pi}}q\right)^3\left(\delta q r\left(1 + \delta\sqrt{\frac{2}{\pi}}q\right) + \delta q + \sqrt{\frac{2}{\pi}}\right)^2}.$$

Since all coefficients of $Q_\delta(q)$ are positive then we have $Q_\delta(q) > 0$ for all $q > 0$. For each eigenvalue, define $r_j \stackrel{\mathrm{def}}{=} \frac{\lambda_j}{\delta^2}$ then

$$\frac{\gamma}{2d}\frac{\mathrm{d}}{\mathrm{d}q}\sum_{j=1}^{d} T_{\lambda_j,\delta}(q) = \frac{\delta^3\gamma}{2d}\frac{\delta q + \sqrt{\frac{2}{\pi}}}{q^2\left(1 + \delta\sqrt{\frac{2}{\pi}}q\right)^3}\sum_{j=1}^{d}\frac{r_j\{P_\delta(q) - r_j Q_\delta(q)\}}{\left(\delta q r_j\left(1 + \delta\sqrt{\frac{2}{\pi}}q\right) + \delta q + \sqrt{\frac{2}{\pi}}\right)^2}.$$

The sign of the summed derivative is determined by the sign of the difference $P_\delta(q)A_d(q) - Q_\delta(q)B_d(q)$ where we defined the terms

$$A_d(q) \stackrel{\mathrm{def}}{=} \frac{1}{d}\sum_{j=1}^{d}\frac{r_j}{\left(\delta q + \sqrt{\frac{2}{\pi}} + r_j\delta q(1 + \sqrt{\frac{2}{\pi}}\delta q)\right)^2},$$

$$\text{and} \quad B_d(q) \stackrel{\mathrm{def}}{=} \frac{1}{d}\sum_{j=1}^{d}\frac{r_j^2}{\left(\delta q + \sqrt{\frac{2}{\pi}} + r_j\delta q(1 + \sqrt{\frac{2}{\pi}}\delta q)\right)^2}.$$

Since $Q_\delta(q) > 0$ and ignoring the degenerate case $K = 0$, the condition $P_\delta(q)A_d(q) - Q_\delta(q)B_d(q) \leq 0$ is equivalent to

$$\frac{B_d(q)}{A_d(q)} \geq \frac{P_\delta(q)}{Q_\delta(q)},$$

whenever $P_\delta(q) > 0$ and whenever $P_\delta(q) \leq 0$ then $P_\delta(q)A_d(q) - Q_\delta(q)B_d(q) < 0$. Thus, a sufficient condition for the second term in $\Theta(q)$ to be nonincreasing is for all $q \in [q_-, q_+]$

$$\frac{B_d(q)}{A_d(q)} \geq \max\left(\frac{P_\delta(q)}{Q_\delta(q)}, 0\right) \stackrel{\text{def}}{=} \psi_\delta(q).$$

We now derive a sufficient conditions on δ for which this holds depending on the assumptions on the eigenvalues of K. From the definitions of P_δ and Q_δ, it is clear that they only depend on q through δq. Define the quantity $r_\star \stackrel{\text{def}}{=} \sup_{q>0} \psi_\delta(q)$. Then $r_\star$ is independent of the value of δ. Indeed, we have

$$r_\star \stackrel{\text{def}}{=} \sup_{q>0} \psi_\delta(q) = \sup_{q>0} \max\left(\frac{P_\delta(q)}{Q_\delta(q)}, 0\right) = \sup_{q>0} \max\left(\frac{\widehat{P}(\delta q)}{\widehat{Q}(\delta q)}, 0\right) = \sup_{s>0} \max\left(\frac{\widehat{P}(s)}{\widehat{Q}(s)}, 0\right)$$

where $\widehat{P}(\delta q) = P_\delta(q)$, $\widehat{Q}(\delta q) = Q_\delta(q)$, $s = \delta q$ and since for $\delta > 0$ then $q > 0 \iff s > 0$. Hence a simpler sufficient condition which guarantees monotonicity is for all $q \in [q_-, q_+]$

$$\frac{B_d(q)}{A_d(q)} \geq r_\star.$$

Numerically, we obtain $r_\star \approx 8.35 \times 10^{-5}$. We rigorously show $0 < r_\star < \infty$. Define $\widehat{\psi}$ such that $\widehat{\psi}(\delta q) = \psi_\delta(q)$ We have $\widehat{P}(0) = 0$. The leading coefficient of $\widehat{P}$ is greater than zero which implies that $\widehat{P}(s) \to \infty$ as $s \to \infty$ and thus there exists s_0 such that $\widehat{P}(s_0) > 0$. It follows that $r_\star > 0$. Now, as $s \to \infty$ we have $\widehat{P}(s) \sim \left(2 - \frac{6}{\pi}\right) s^4$ while $\widehat{Q} \sim \frac{2}{\pi} s^6$ from which it follows that

$$\frac{\widehat{P}(s)}{\widehat{Q}(s)} \sim \frac{2 - \frac{6}{\pi}}{\frac{2}{\pi}} \frac{1}{s^2} \to 0 \quad \text{as} \quad s \to \infty.$$

Since $\widehat{P}(0) < 0$ then by continuity $\widehat{P}(s) < 0$ for sufficiently small $s > 0$ which implies $\widehat{\psi}(s) = 0$ near 0. Since $\widehat{\psi}(s) \to 0$ as $s \to \infty$ and $\widehat{\psi}$ is continuous on $(0, \infty)$, then $r_\star < \infty$.

From the definitions of $B_d(q)$ and $A_d(q)$, it is easy to see that if $\lambda_j = 0$ then its contribution to $B_d(q)$ and $A_d(q)$ is zero. Let $J_+ \stackrel{\text{def}}{=} \{j : \lambda_j > 0\}$ and ignoring the degenerate case $K = 0$ then we have

$$\frac{B_d(q)}{A_d(q)} = \sum_{j \in J_+} r_j w_j,$$

where we define the weights $\sum_{j \in J_+} w_j = 1$ as follows

$$w_j \stackrel{\text{def}}{=} \frac{\frac{r_j}{\left(\delta q + \sqrt{\frac{2}{\pi}} + r_j \delta q \left(1 + \sqrt{\frac{2}{\pi}} \delta q\right)\right)^2}}{\sum_{k \in J_+}^{d} \frac{r_k}{\left(\delta q + \sqrt{\frac{2}{\pi}} + r_k \delta q \left(1 + \sqrt{\frac{2}{\pi}} \delta q\right)\right)^2}}.$$

Consequently, $B_d(q)/A_d(q)$ is a weighted average which implies

$$\frac{\lambda_{\min}^+}{\delta^2} = \min_{j \in J_+} r_j \leq \frac{B_d(q)}{A_d(q)},$$

where $\lambda_{\min}^+$ is the smallest non-zero eigenvalue. Thus, a sufficient condition is

$$\delta \leq \sqrt{\frac{\lambda_{\min}^+}{r_\star}}.$$

The condition for isotropic covariance then follows immediately. □

Remark G.1. *We note that we could predict the limiting values of a larger set of statistics of $q(t)$ (e.g. distance to optimality) than just $\mathscr{R}$, $\widehat{\mathscr{B}}_{44}$ and $\mathscr{R}^{\mathrm{Adv}}$. See Figure 6 for numerical evidence that $q(t)$ converges and the fixed point equation* (390) *predicts its limit.*

The fixed point equation (390) is hard to interpret. Even for the simpler setting of $K = I_d$, (390) remains a polynomial of degree 6 in q. To provide insights on the limiting values q_∞ and $\tilde{\lambda}(\infty)$, we derive q_∞ as $\gamma \to 0$. Note that we recover Proposition 1 in Xing et al. (2021) in the limit $\gamma \to 0$.

Theorem G.2. *Suppose the assumptions of Theorem G.1 hold. Define $J_+ \overset{\text{def}}{=} \{j : \lambda_j > 0\}$ and define the two thresholds*

$$\delta_- = \sqrt{\frac{2}{\pi}} \sqrt{\frac{\sum_{j\in J_+} (x_j^\star)^2}{\sum_{j\in J_+} \frac{(x_j^\star)^2}{\lambda_j}}} \quad \textit{and} \quad \delta_+ = \sqrt{\frac{\pi}{2}} \sqrt{\frac{\sum_{j\in J_+} \lambda_j^2 (x_j^\star)^2}{\sum_{j\in J_+} \lambda_j (x_j^\star)^2}}. \tag{401}$$

For $\gamma > 0$, let $q_\gamma(\infty)$ be an admissible solution of the fixed point equation of Theorem G.1. Then as $\gamma \to 0$

$$\begin{aligned} &\textit{if } \delta \le \delta_- \textit{ then} \quad q_\gamma(\infty) \to \infty \quad \textit{and} \quad \tilde{\lambda}(q_\gamma(\infty)) \to \tilde{\lambda}_\star = 0, \\ &\textit{if } \delta_- < \delta < \delta_+ \textit{ then} \quad q_\gamma(\infty) \to q_\star \in (0,\infty) \quad \textit{and} \quad \tilde{\lambda}(q_\gamma(\infty)) \to \tilde{\lambda}_\star \in (0,\infty), \\ &\textit{and if } \delta \ge \delta_+ \textit{ then} \quad q_\gamma(\infty) \to 0 \quad \textit{and} \quad \tilde{\lambda}(q_\gamma(\infty)) \to \tilde{\lambda}_\star = \infty. \end{aligned} \tag{402}$$

If $\delta_- < \delta < \delta_+$, then $(q_\star, \tilde{\lambda}_\star)$ is the unique solution of

$$q_\star = \sqrt{\frac{\sum_{j\in J_+} \frac{\lambda_j^2}{(\lambda_j+\tilde{\lambda}_\star)^2} (x_j^\star)^2}{\sum_{j\in J_+} \lambda_j \frac{\tilde{\lambda}_\star^2}{(\lambda_j+\tilde{\lambda}_\star)^2} (x_j^\star)^2}} \quad \textit{and} \quad \tilde{\lambda}_\star = \frac{\delta^2 + \delta\sqrt{\frac{2}{\pi}} \frac{1}{q_\star}}{1 + \delta\sqrt{\frac{2}{\pi}} q_\star}. \tag{403}$$

If $K = I_d$ then $\delta_- = \sqrt{\frac{2}{\pi}}$, $\delta_+ = \sqrt{\frac{\pi}{2}}$ and for $\delta_- < \delta < \delta_+$

$$q_\star = \frac{1 - \delta\sqrt{\frac{2}{\pi}}}{\delta\left(\delta - \sqrt{\frac{2}{\pi}}\right)} \quad \textit{and} \quad \tilde{\lambda}_\star = \frac{\delta^2 - \delta\sqrt{\frac{2}{\pi}}}{1 - \delta\sqrt{\frac{2}{\pi}}}. \tag{404}$$

Proof. We omit the proof $\delta_- < \delta_+$ as it follows from two simple applications of Cauchy–Schwarz. Since K is symmetric, we have $\mathbb{R} = \mathrm{Range}(K) \oplus \ker(K)$. Hence, we can decompose $X = X_+ + X^0$ for $X_+ \in \mathrm{Range}(K)$ and $X^0 \in \ker(K)$. Similarly, we can decompose $X^\star = X_+^\star + X^{\star,0}$ and $X^{\star,\mathrm{Adv}} = X_+^{\star,\mathrm{Adv}} + X^{\star,\mathrm{Adv},0}$.

We now show by contradiction that $X^{\star,\mathrm{Adv},0} = 0$. Assume that $X^{\star,\mathrm{Adv},0} \neq 0$. From the above decompositions, we have that$(X - X^\star)^\top K (X - X^\star) = (X_+ - X_+^\star)^\top K (X_+ - X_+^\star)$ and thus $\mathcal{R}(X) = \mathcal{R}(X_+)$.

Since $\mathrm{Range}(K) \perp \ker(K)$ then $\|X\|^2 = \|X_+\|^2 + \|X^0\|^2 \ge \|X_+\|^2$. Therefore, we have

$$\mathcal{R}^{\mathrm{Adv}}(X) - \mathcal{R}^{\mathrm{Adv}}(X_+) = \delta\sqrt{\frac{2}{\pi}} \sqrt{2\mathcal{R}(X_+)} \left(\|X\| - \|X_+\|\right) + \frac{\delta^2}{2} \|X^0\|^2.$$

Since $X^{\star,\mathrm{Adv},0} \neq 0$ then $\|X^{\star,\mathrm{Adv},0}\| > 0$ and $\|X^{\star,\mathrm{Adv}}\|^2 > \|X_+^{\star,\mathrm{Adv}}\|^2$ which implies $\mathcal{R}^{\mathrm{Adv}}(X^{\star,\mathrm{Adv}}) - \mathcal{R}^{\mathrm{Adv}}(X_+^{\star,\mathrm{Adv}}) > 0$. Thus, we have reached a contradiction as $X^{\star,\mathrm{Adv}}$ is a minimizer of $\mathcal{R}^{\mathrm{Adv}}$ from which it follows that $X^{\star,\mathrm{Adv}} \in \mathrm{Range}(K)$.

On $\mathrm{Range}(K)$, the operator $K|_{\mathrm{Range}(K)}$ is positive definite. The preceding argument shows that minimizing $\mathcal{R}^{\mathrm{Adv}}$ over $\mathbb{R}^d$ is equivalent to minimizing it over $\mathrm{Range}(K)$:

$$\arg\min_{X\in\mathbb{R}^d} \mathcal{R}^{\mathrm{Adv}}(X) = \arg\min_{X\in\mathrm{Range}(K)} \mathcal{R}^{\mathrm{Adv}}(X)$$

Plugging the definition of $q_\star$ into $\tilde{\lambda}_\star$ in (405) recovers Equation 4 in Xing et al. (2021) restricted to $\mathrm{Range}(K)$. Thus, we apply Proposition 1 in Xing et al. (2021) to the positive definite covariance matrix $K|_{\mathrm{Range}(K)}$ and obtain the minimizer

$$X^{\star,\mathrm{Adv}} = \sum_{j\in J_+} \frac{\lambda_j}{\lambda_j + \tilde{\lambda}_\star} x_j^\star \omega_j,$$

from which we obtain the limiting values as $\gamma \to 0$ in (402). It remains to prove that $q_\gamma(\infty) \to q_\star$. We start with the regime $\delta_- < \delta < \delta_+$. From Theorem G.1 we have

$$D_j(q) = \lambda_j(1 + \delta\sqrt{\frac{2}{\pi}}q) + \delta^2 + \delta\sqrt{\frac{2}{\pi}}\frac{1}{q} \geq \delta^2 > 0.$$

For all $q \in [a,b] \subset (0,\infty)$, this implies that $C_{\mathscr{R}}(q) = \mathcal{O}(\gamma)$ and $C_{\mathscr{B}}(q) = \mathcal{O}(\gamma)$. The functions $G(q)$, $F_\infty(q)$, $L_\infty(q)$ are continuous on $[a,b]$ and do not depend on γ. Moreover, $F_\infty(q) > 0$ on $[a,b]$ from which it follows that

$$\sup_{q\in[a,b]} \left| (1 - G(q)C_{\mathscr{R}}(q)) \frac{L_\infty(q)}{2F_\infty(q)} + \frac{1}{2}G(q)C_{\mathscr{B}}(q) - \frac{L_\infty(q)}{2F_\infty(q)} \right| \underset{\gamma\to 0}{\longrightarrow} 0.$$

Hence, for a subsequence $\gamma_n \downarrow 0$ such that $q_{\gamma_n} \to q_0 \in (0,\infty)$, then

$$q_0^2 = \frac{L_\infty(q_0)}{2F_\infty(q_0)}. \tag{405}$$

Thus every finite positive accumulation point of q_γ must be a solution of the fixed point equation (405). In the regime $\delta_- < \delta < \delta_+$, the solution to (405) is equivalent to the solution to Equation 4 in Xing et al. (2021). Moreover, Xing et al. (2021) show $\tilde{\lambda}_\star$ is the unique stationary fixed point satisfying Equation 4. They also prove that Equation 4 has no solution in $(0,\infty)$ whenever $\delta \leq \delta_-$ or $\delta_+ \leq \delta$. Therefore, outside the middle regime, $q_\gamma(\infty)$ cannot remain in any γ-independent compact subset of $(0,\infty)$.

With this in mind, we prove the limits in the boundary regimes. First suppose, towards contradiction, that along some sequence $\gamma_n \downarrow 0$, $q_{\gamma_n}(\infty) \to 0$. For $0 < q \leq 1$, $D_j(q) \geq \delta\sqrt{\frac{2}{\pi}}\frac{1}{q}$. Therefore

$$C_{\mathscr{R}}(q) \leq \sqrt{\frac{\pi}{2}}\frac{\gamma q}{2d\delta}\sum_{j=1}^d \lambda_j^2 \quad \text{and} \quad C_{\mathscr{B}}(q) \leq \sqrt{\frac{\pi}{2}}\frac{\gamma q}{d\delta}\sum_{j=1}^d \lambda_j.$$

Since $G(q) \to 1$ as $q \downarrow 0$, this implies $G(q_{\gamma_n})C_{\mathscr{R}}(q_{\gamma_n}) \to 0$ and $G(q_{\gamma_n})C_{\mathscr{B}}(q_{\gamma_n}) \to 0$. Using the fixed point equation and the fact that $C_{\mathscr{B}}(q) \geq 0$, we get

$$(q_{\gamma_n}(\infty))^2 \geq (1 - G(q_{\gamma_n}(\infty))C_{\mathscr{R}}(q_{\gamma_n}(\infty))) \frac{L_\infty(q_{\gamma_n}(\infty))}{2F_\infty(q_{\gamma_n}(\infty))}.$$

Dividing by $(q_{\gamma_n}(\infty))^2$ gives

$$1 \geq (1 - G(q_{\gamma_n}(\infty))C_{\mathscr{R}}(q_{\gamma_n}(\infty))) \frac{L_\infty(q_{\gamma_n}(\infty))}{2F_\infty(q_{\gamma_n}(\infty))} \frac{1}{q_{\gamma_n}^2(\infty)}.$$

Since $\tilde{\lambda}(q) =\sim \delta\sqrt{\frac{2}{\pi}}\frac{1}{q}$ as $q \downarrow 0$ then

$$F_\infty(q) \to \frac{1}{2}\sum_{j\in J_+} \lambda_j (x_j^\star)^2 \quad \text{and} \quad L_\infty(q) \sim \frac{\pi}{2\delta^2}q^2 \sum_{j\in J_+} \lambda_j^2 (x_j^\star)^2,$$

from which it follows that

$$\frac{L_\infty(q_{\gamma_n}(\infty))}{2F_\infty(q_{\gamma_n}(\infty))} \frac{1}{q_{\gamma_n}^2(\infty)} \longrightarrow \left(\frac{\delta_+}{\delta}\right)^2.$$

Thus, we obtain $1 \geq \left(\frac{\delta_+}{\delta}\right)^2$ which implies whenever $\delta < \delta_+$ and then q_γ cannot converge to 0. A similar argument shows that along some subsequence $\gamma_n \downarrow 0$, when $\delta > \delta_+$, q_{γ_n} cannot converge to ∞.

If $\delta \leq \delta_-$, then $\delta < \delta_+$, so no subsequence of $q_\gamma(\infty)$ can converge to 0 nor converge to any point in $(0, \infty)$ as the fixed point equation (405) does not admit a solution in this regime. This implies $q_\gamma \to \infty$ when $\gamma \to 0$ and thus $\tilde{\lambda}(q_\gamma) \to 0$. If $\delta_- < \delta < \delta_+$, then q_γ cannot converge to 0 nor ∞. Hence every sequence $\gamma_n \downarrow 0$ has a subsequence along which $q_{\gamma_n} \to q_0 \in (0, \infty)$ which solves the fixed point equation (405). Since $q_\star$ is the unique solution to (405), then every subsequential limit equals $q_\star$ from which it follows that $q_\gamma \to q_\star$ and by continuity then $\tilde{\lambda}^{\text{eff}}(q_\gamma) \to \tilde{\lambda}^{\text{eff}}(q_\star)$.

Finally, if $\delta \geq \delta_+$, then $\delta > \delta_-$, so no subsequence of $q_\gamma(\infty)$ can converge to ∞ nor to a point in $(0, \infty)$ as the fixed point equation (405) does not admit a solution in this regime. This implies $q_\gamma \to 0$ when $\gamma \to 0$ and thus $\tilde{\lambda}(q_\gamma) \to \infty$.

The limits for the isotropic setting follow from a simple computation. □

With $J_+ \stackrel{\text{def}}{=} \{j : \lambda_j > 0\}$ as in Theorem G.2, then from Xing et al. (2021, Proposition 1) and Theorem G.2, we see that $X^{\star,\text{Adv}}$ admits the following representation depending on δ

$$X^{\star,\text{Adv}} = \begin{cases} \sum_{j \in J_+} \frac{\lambda_j}{\lambda_j + \tilde{\lambda}_\star} \langle X^\star, \omega_j \rangle \omega_j & \text{if} \quad \delta < \delta_+ \\ 0 & \text{if} \quad \delta \geq \delta_+ \end{cases} \tag{406}$$

Thus, depending on the value of δ, we can view $X^{\star,\text{Adv}}$ as a compromise between minimizing the least squares risk $\mathcal{R}$ and $\|\cdot\|^2$. For $\delta \leq \delta_-$ we have $\mathcal{R}(X^{\star,\text{Adv}}) = 0$ and for $\delta \geq \delta_+$ we have $\|X^{\star,\text{Adv}}\|^2 = 0$.

H Integro-Differential Equation Analysis

In this section, we provide alternative characterization for the solution to the integro-differential equation (17) which we recall below.

Integro-Differential Equation for $\mathcal{Z}(t,z)$. For any contour $\Gamma \subset \mathbb{C}$ enclosing the eigenvalues of K_1 and K_2, we have an expression for the derivative of $\mathcal{Z}$ for $z \in \Gamma^2$:

$$\mathrm{d}\mathcal{Z}(t,\cdot) = \mathcal{F}(z, \mathcal{Z}(t,\cdot))\ \mathrm{d}t, \quad \text{with initial condition} \quad \mathcal{Z}(0,z) = Z(W_0, z), \tag{407}$$

where

$$\begin{gathered} \mathcal{Z}(t,z) \stackrel{\text{def}}{=} \begin{bmatrix} \mathcal{S}(t,z) & \mathcal{M}(t,z) \\ \mathcal{M}(t,z)^\top & \mu^\top \mathcal{R}(z;\mathcal{K})\mu \end{bmatrix}, \\ \text{and} \quad \mathcal{F}(z, \mathcal{Z}(t,\cdot)) \stackrel{\text{def}}{=} \begin{bmatrix} \mathcal{F}_\mathcal{S}(z, \mathcal{Z}(t,\cdot)) & \mathcal{F}_\mathcal{M}(z, \mathcal{Z}(t,\cdot)) \\ \mathcal{F}_\mathcal{M}(z, \mathcal{Z}(t,\cdot))^\top & 0_{2\times 2} \end{bmatrix}, \end{gathered} \tag{408}$$

The maps $\mathcal{F}_\mathcal{S}$ and $\mathcal{F}_\mathcal{M}$ are defined as follows

$$\begin{aligned} \mathrm{d}\mathcal{M}(t,z) = -2\gamma(t) \sum_{i=1}^{2} p_i \Bigg(& H_{1,i}^\top(\widehat{\mathscr{B}}_i(t)) \left(z_i \mathcal{M}(t,z) - \frac{1}{2\pi\,\mathrm{i}} \oint_\Gamma \mathcal{M}(t,z)\,\mathrm{d}z_i \right) \\ & + H_{2,i}(\widehat{\mathscr{B}}_i(t)) \otimes \mu^\top \mathcal{R}(z;\mathcal{K})\mu_i + \left(\frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) \right) \mathrm{D}\,\mathcal{M}(t,z) \Bigg)\,\mathrm{d}t, \end{aligned} \tag{409}$$

and

$$
\begin{aligned}
\mathrm{d}\mathcal{S}(t,z) = -2\gamma(t)\sum_{i=1}^{2} p_i \Bigg(& \left(z_i \mathcal{S}(t,z) - \frac{1}{2\pi\,\mathrm{i}} \oint_{\Gamma} \mathcal{S}(t,z)\,\mathrm{d}z_i \right) H_{1,i}(\widehat{\mathscr{B}}_i(t)) \\
& + H_{1,i}^{\top}(\widehat{\mathscr{B}}_i(t)) \left(z_i \mathcal{S}(t,z) - \frac{1}{2\pi\,\mathrm{i}} \oint_{\Gamma} \mathcal{S}(t,z)\,\mathrm{d}z_i \right) \\
& + \mathcal{M}_i(t,z) \otimes H_{2,i}(\widehat{\mathscr{B}}_i(t)) + H_{2,i}(\widehat{\mathscr{B}}_i(t)) \otimes \mathcal{M}_i(t,z) \\
& + \left(\frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) \right) (\mathcal{S}(t,z)\,\mathrm{D} + \mathrm{D}\,\mathcal{S}(t,z)) \Bigg)\,\mathrm{d}t \\
& + \frac{\gamma(t)^2}{d} \sum_{i=1}^{2} p_i \mathcal{I}_i(\widehat{\mathscr{B}}_i(t))\,\mathrm{Tr}((K_i + \mu_i\mu_i^{\top})\mathcal{R}(z;\mathcal{K}))\,\mathrm{D}\;\mathrm{d}t,
\end{aligned}
\tag{410}
$$

where $H_{1,i}(\widehat{\mathscr{B}}_i) = \left[\begin{array}{c|c} \partial_{11} h_i(\widehat{\mathscr{B}}_i) & 0 \\ \hline \partial_{21} h_i(\widehat{\mathscr{B}}_i) & 0 \end{array}\right]$, $H_{2,i}(B) = \left[\begin{array}{c} \partial_{13} h_i(\widehat{\mathscr{B}}_i) \\ \hline 0 \end{array}\right]$, $\mathrm{D} = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & 0 \end{array}\right]$

$$
\widehat{\mathscr{B}}_i(t) = \begin{bmatrix} \mathscr{B}_i(t) & \mathfrak{m}_i(t) & 0 \\ \mathfrak{m}_i(t)^{\top} & \oint \mu_i^{\top} \mathcal{R}(z;\mathcal{K}) \mu_i\,\mathrm{D}z & 0 \\ 0 & 0 & \oint \mathcal{S}_{11}(t,z)\,\mathrm{D}z \end{bmatrix}. \tag{411}
$$

Here $\mathscr{B}_i(t) = \oint z_i \mathcal{S}(t,z)\,\mathrm{D}z \in \mathbb{R}^{2\times 2}$, $\mathfrak{m}_i(t) = \oint \mathcal{M}_i(t)\,\mathrm{D}z$ where $\mathcal{M}_i(t) \in \mathbb{R}^2$ denotes the ith column of $\mathcal{M}$, associated with μ_i.

We proceed using the spectral decomposition of the resolvent $\mathcal{R}(z;\mathcal{K})$ similarly to Lemma 2 in Collins-Woodfin & Seroussi (2025) as follows:

$$
\mathcal{R}(z;\mathcal{K}) = \sum_{j=1}^{d} \frac{1}{\lambda_j^{(1)} - z_1} \omega_j \omega_j^{\top} \sum_{\ell=1}^{d} \frac{1}{\lambda_\ell^{(2)} - z_2} \omega_\ell \omega_\ell^{\top} = \sum_{j=1}^{d} \frac{1}{\prod_{\ell=1}^{2} (\lambda_j^{(\ell)} - z_\ell)} \omega_j \omega_j^{\top} \tag{412}
$$

We will now assume the following decompositions of $\mathcal{S}(t,z)$ and $\mathcal{M}(t,z)$ for some functions $\mathscr{V}_j$ and $\mathfrak{m}_{j,\ell}$ which we will determine later.

$$
\mathcal{S}(t,z) = \sum_{j=1}^{d} \frac{\mathscr{V}_j(t)}{\prod_{\ell=1}^{2} (\lambda_j^{(\ell)} - z_\ell)} \quad \text{and} \quad \mathcal{M}_i(t,z) = \sum_{j=1}^{d} \frac{\mathfrak{m}_{j,i}(t)}{\prod_{\ell=1}^{2} (\lambda_j^{(\ell)} - z_\ell)} \tag{413}
$$

Using (407), we determine $\mathscr{V}_j$ and $\mathfrak{m}_{j,\ell}$ are defined as follows

$$
\begin{aligned}
\frac{\mathrm{d}\mathscr{V}_j(t)}{\mathrm{d}t} = -2\gamma(t) \sum_{i=1}^{2} p_i \Bigg(& \lambda_j^{(i)} (\mathscr{V}_j(t) H_{1,i}(\widehat{\mathscr{B}}_i(t)) + H_{1,i}(\widehat{\mathscr{B}}_i(t))^{\top} \mathscr{V}_j(t)) + \mathfrak{m}_{j,i}(t) \otimes H_{2,i}(\widehat{\mathscr{B}}_i(t)) \\
& + H_{2,i}(\widehat{\mathscr{B}}_i(t)) \otimes \mathfrak{m}_{j,i}(t) + \left(\frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) \right) (\mathscr{V}_j(t)\,\mathrm{D} + \mathrm{D}\,\mathscr{V}_j(t)) \Bigg) \\
& + \frac{\gamma(t)^2}{d} \left(\sum_{i=1}^{2} p_i \mathcal{I}_i(\widehat{\mathscr{B}}_i(t)) (\lambda_j^{(i)} + (\omega_j^{\top} \mu_i)^2) \right) \mathrm{D}, \\
\frac{\mathrm{d}\mathfrak{m}_{j,\ell}(t)}{\mathrm{d}t} = -2\gamma(t) \sum_{i=1}^{2} p_i \Bigg(& \lambda_j^{(i)} H_{1,i}(\widehat{\mathscr{B}}_i(t))^{\top} \mathfrak{m}_{j,\ell}(t) + \langle \omega_j, \mu_\ell \rangle \langle \omega_j, \mu_i \rangle \cdot H_{2,i}(\widehat{\mathscr{B}}_i(t)) \\
& + \left(\frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) \right) \mathrm{D}\,\mathfrak{m}_{j,\ell}(t) \Bigg),
\end{aligned}
\tag{414}
$$

with $\mathscr{V}_j(0) = \hat{X}_0^{\top} \omega_j \omega_j^{\top} \hat{X}_0$ and $\mathfrak{m}_{j,\ell}(0) = \hat{X}_0^{\top} \omega_j \omega_j^{\top} \mu_\ell$. It immediately follows from Lemma 2 in Collins-Woodfin & Seroussi (2025) that $\mathcal{S}(t,z)$ and $\mathcal{M}_\ell(t,z)$ defined in (413) with $\mathscr{V}_j$ and $\mathfrak{m}_{j,\ell}$ satisfying (414) are solutions to (407). We omit the proof as the argument is identical. We will now

use a similar argument to Appendix A in Collins-Woodfin et al. (2024a) to determine the explicit dynamics of $\widehat{\mathscr{B}}_i(t))$ for $i = 1, 2$. Let Φ be the fundamental matrix of the ODE:

$$\dot{\Phi}_j = 2\gamma(t) \sum_{i=1}^{2} p_i \left(\lambda_j^{(i)} H_{1,i}(\widehat{\mathscr{B}}_i(t)) + \left(\frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) \right) \mathrm{D} \right) \Phi_j, \quad \Phi_j(0) = I_2. \tag{415}$$

Define the following terms:

$$\begin{aligned} U_{0,\mathscr{V}_j}(t) \stackrel{\text{def}}{=} & -2\gamma(t) \sum_{i=1}^{2} p_i \left(\mathfrak{m}_{j,i}(t) \otimes H_{2,i}(\widehat{\mathscr{B}}_i(t)) + H_{2,i}(\widehat{\mathscr{B}}_i(t)) \otimes \mathfrak{m}_{j,i}(t) \right) \\ & + \frac{\gamma(t)^2}{d} \left(\sum_{i=1}^{2} p_i \mathcal{I}_i(\widehat{\mathscr{B}}_i(t))(\lambda_j^{(i)} + (\omega_j^\top \mu_i)^2) \right) \mathrm{D}, \end{aligned}$$

and

$$U_{0,\mathfrak{m}_{j,\ell}}(t) = -2\gamma(t) \sum_{i=1}^{2} p_i \langle \omega_j, \mu_\ell \rangle \langle \omega_j, \mu_i \rangle \cdot H_{2,i}(\widehat{\mathscr{B}}_i(t)).$$

Then it follows that

$$\begin{aligned} (\Phi_j^\top \dot{\mathfrak{m}}_{j,\ell}) &= \dot{\Phi}_j^{\ \top} \mathfrak{m}_{j,\ell} + \Phi_j^\top \dot{\mathfrak{m}}_{j,\ell} \\ &= \Phi_j^\top \left(2\gamma(t) \sum_{i=1}^{2} p_i \left(\lambda_j^{(i)} H_{1,i}(\widehat{\mathscr{B}}_i(t))^\top + \left(\frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) \right) \mathrm{D} \right) \right) \mathfrak{m}_{j,\ell} \\ &\quad + \Phi_j^\top \left(U_{0,\mathfrak{m}_{j,\ell}} - 2\gamma(t) \sum_{i=1}^{2} p_i \left(\lambda_j^{(i)} H_{1,i}(\widehat{\mathscr{B}}_i(t))^\top + \left(\frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) \right) \mathrm{D} \right) \mathfrak{m}_{j,\ell} \right) \\ &= \Phi_j^\top U_{0,\mathfrak{m}_{j,\ell}}. \end{aligned}$$

Similarly, we obtain

$$\begin{aligned} (\Phi_j^\top \dot{\mathscr{V}}_j \Phi_j) &= \dot{\Phi}_j^{\ \top} \mathscr{V}_j \Phi_j + \Phi_j^\top \dot{\mathscr{V}}_j \Phi_j + \Phi_j^\top \mathscr{V}_j \dot{\Phi}_j \\ &= \Phi_j^\top \left(2\gamma(t) \sum_{i=1}^{2} p_i \left(\lambda_j^{(i)} H_{1,i}(\widehat{\mathscr{B}}_i(t))^\top + \left(\frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) \right) \mathrm{D} \right) \right) \mathscr{V}_j \Phi_j \\ &\quad + \Phi_j^\top \left(U_{0,\mathscr{V}_j} - \mathscr{V}_j \left(2\gamma(t) \sum_{i=1}^{2} p_i \left(\lambda_j^{(i)} H_{1,i}(\widehat{\mathscr{B}}_i(t)) + \left(\frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) \right) \mathrm{D} \right) \right) \right. \\ &\quad \left. - \left(2\gamma(t) \sum_{i=1}^{2} p_i \left(\lambda_j^{(i)} H_{1,i}(\widehat{\mathscr{B}}_i(t))^\top + \left(\frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) \right) \mathrm{D} \right) \right) \mathscr{V}_j \right) \Phi_j \\ &\quad + \Phi_j^\top \mathscr{V}_j \left(2\gamma(t) \sum_{i=1}^{2} p_i \left(\lambda_j^{(i)} H_{1,i}(\widehat{\mathscr{B}}_i(t)) + \left(\frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) \right) \mathrm{D} \right) \right) \Phi_j \\ &= \Phi_j^\top U_{0,\mathscr{V}_j} \Phi_j. \end{aligned} \tag{416}$$

These ODEs are solvable.

Resolvent formula. We have

$$\mathfrak{m}_{j,\ell}(t) = \Phi_j^{-\top}(t) \mathfrak{m}_{j,\ell}(0) + \int_0^t \Phi_j^{-\top}(t) \Phi_j^\top(s) U_{0,\mathfrak{m}_{j,\ell}}(s) \, \mathrm{d}s \tag{417}$$

and

$$\mathscr{V}_j(t) = \Phi_j^{-\top}(t) \mathscr{V}_j(0) \Phi_j^{-1}(t) + \int_0^t \Phi_j^{-\top}(t) \Phi_j^\top(s) U_{0,\mathscr{V}_j}(s) \Phi_j(s) \Phi_j^{-1}(t) \, \mathrm{d}s \tag{418}$$

$$\text{where} \quad \mathscr{V}_j(0) = \hat{X}_0^\top \omega_j \omega_j^\top \hat{X}_0, \quad \mathfrak{m}_{j,\ell}(0) = \hat{X}_0^\top \omega_j \omega_j^\top \mu_\ell,$$

$$\Phi_j(t) \text{ is the solution to } \dot{\Phi}_j = 2\gamma(t) \sum_{i=1}^{2} p_i \left(\lambda_j^{(i)} H_{1,i}(\widehat{\mathscr{B}}_i(t)) + \left(\frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) \right) \mathrm{D} \right) \Phi_j$$

$$\text{with} \quad \Phi(0, z) = I_2, \quad U_{0,\mathscr{V}_j}(t) = -2\gamma(t) \sum_{i=1}^{2} p_i \Big(\mathfrak{m}_{j,i}(t) \otimes H_{2,i}(\widehat{\mathscr{B}}_i(t))$$

$$+ H_{2,i}(\widehat{\mathscr{B}}_i(t)) \otimes \mathfrak{m}_{j,i}(t) \Big) + \frac{\gamma(t)^2}{d} \left(\sum_{i=1}^{2} p_i \mathcal{I}_i(\widehat{\mathscr{B}}_i(t)) (\lambda_j^{(i)} + (\omega_j^\top \mu_i)^2) \right) \mathrm{D}$$

$$\text{and} \quad U_{0,\mathfrak{m}_{j,\ell}}(t) = -2\gamma(t) \sum_{i=1}^{2} p_i \langle \omega_j, \mu_\ell \rangle \langle \omega_j, \mu_i \rangle \cdot H_{2,i}(\widehat{\mathscr{B}}_i(t)).$$

This alternative characterization requires to solve the ODE,

$$\dot{\Phi}_j = 2\gamma(t) \sum_{i=1}^{2} p_i \left(\lambda_j^{(i)} H_{1,i}(\widehat{\mathscr{B}}_i(t)) + \left(\frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) \right) \mathrm{D} \right) \Phi_j, \quad \Phi_j(0) = I_2. \tag{419}$$

We must now solve for $\mathfrak{m}_{j,\ell}(t)$ and $\mathscr{V}_j(t)$ in (417) and (418) respectively. Most of the work towards this is to solve the matrix ODE (419) which we express in matrix form:

$$\dot{\Phi}_j = 2\gamma(t) \sum_{i=1}^{2} p_i \begin{bmatrix} \lambda_j^{(i)} \partial_{11} h_i(\widehat{\mathscr{B}}_i(t)) + \frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) & 0 \\ \lambda_j^{(i)} \partial_{21} h_i(\widehat{\mathscr{B}}_i(t)) & 0 \end{bmatrix} \Phi, \quad \Phi(0) = I_2. \tag{420}$$

Hence, this reduces to the following system of first-order linear ODEs:

$$\begin{aligned}
\dot{\Phi}_{j,11} &= 2\gamma(t) \sum_{i=1}^{2} p_i (\lambda_j^{(i)} \partial_{11} h_i(\widehat{\mathscr{B}}_i(t)) + \tfrac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t))) \Phi_{j,11} \quad \Phi_{j,11}(0) = 1 \\
\dot{\Phi}_{j,21} &= 2\gamma(t) (\sum_{i=1}^{2} p_i \lambda_j^{(i)} \partial_{21} h_i(\widehat{\mathscr{B}}_i(t))) \Phi_{j,11}, \quad \Phi_{j,21}(0) = 0, \\
\dot{\Phi}_{j,12} &= 2\gamma(t) \sum_{i=1}^{2} p_i (\lambda_j^{(i)} \partial_{11} h_i(\widehat{\mathscr{B}}_i(t)) + \tfrac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t))) \Phi_{j,12}, \quad \Phi_{j,12}(0) = 0, \\
\dot{\Phi}_{j,22} &= 2\gamma(t) (\sum_{i=1}^{2} p_i \lambda_j^{(i)} \partial_{21} h_i(\widehat{\mathscr{B}}_i(t))) \Phi_{j,12}, \quad \Phi_{j,22}(0) = 1.
\end{aligned} \tag{421}$$

Using integration factors to solve the ODEs of $\Phi_{j,11}$ and $\Phi_{j,12}$, we obtain

$$\Phi_{j,11}(t) = \exp\left(\int_0^t 2\gamma(s) \sum_{i=1}^{2} p_i \left(\lambda_j^{(i)} \partial_{11} h_i(\widehat{\mathscr{B}}_i(s)) + \tfrac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(s)) \right) \mathrm{d}s \right),$$

$$\Phi_{j,21}(t) = \int_0^t 2\gamma(s) \left(\sum_{i=1}^{2} p_i \lambda_j^{(i)} \partial_{21} h_i(\widehat{\mathscr{B}}_i(s)) \right) \Phi_{j,11}(s) \, \mathrm{d}s, \quad \Phi_{j,22}(t) = 1, \quad \text{and} \quad \Phi_{j,12}(t) = 0. \tag{422}$$

Given that $\Phi_j \in \mathbb{R}^{2\times 2}$ then its inverse is given by

$$\Phi_j^{-1}(t) = \frac{1}{\Phi_{j,11}(t)} \begin{bmatrix} 1 & 0 \\ -\Phi_{j,21}(t) & \Phi_{j,11}(t) \end{bmatrix}. \tag{423}$$

It remains to compute the expressions in (417) and (418). We have

$$\begin{aligned}
&\Phi_j^{-\top}(t)\mathscr{V}_j(0)\Phi_j^{-1}(t) \\
&= \frac{1}{\Phi_{j,11}^2(t)} \begin{bmatrix} X_0^\top \omega_j \omega_j^\top X_0 & \Phi_{j,11}(t) X_0^\top \omega_j \omega_j^\top X^\star \\ -2\Phi_{j,21}(t) X_0^\top \omega_j \omega_j^\top X^\star + \Phi_{j,21}^2(t)(X^\star)^\top \omega_j \omega_j^\top X^\star & -\Phi_{j,11}(t)\Phi_{j,21}(t)(X^\star)^\top \omega_j \omega_j^\top X^\star \\ \Phi_{j,11}(t) X_0^\top \omega_j \omega_j^\top X^\star & \\ -\Phi_{j,21}(t)\Phi_{j,11}(t)(X^\star)^\top \omega_j \omega_j^\top X^\star & \Phi_{11}^2(t)(X^\star)^\top \omega_j \omega_j^\top X^\star \end{bmatrix}.
\end{aligned} \tag{424}$$

Now, it is easy to see that

$$\Phi_j(s)\Phi_j^{-1}(t) = \begin{bmatrix} \frac{\Phi_{j,11}(s)}{\Phi_{j,11}(t)} & 0 \\ \frac{\Phi_{j,21}(s) - \Phi_{j,21}(t)}{\Phi_{j,11}(t)} & 1 \end{bmatrix}, \tag{425}$$

from which it follows that

$$\begin{aligned}
&\frac{\gamma(t)^2}{d}\left(\sum_{i=1}^2 p_i \mathcal{I}_i(\widehat{\mathscr{B}}_i(t))(\lambda_j^{(i)} + (\omega_j^\top \mu_i)^2)\right)\Phi_j^{-\top}(t)\Phi_j^\top(s) D\Phi_j(s)\Phi_j^{-1}(t) \\
&= \frac{\gamma(t)^2}{d}\frac{\Phi_{j,11}^2(s)}{\Phi_{j,11}^2(t)}\left(\sum_{i=1}^2 p_i \mathcal{I}_i(\widehat{\mathscr{B}}_i(t))(\lambda_j^{(i)} + (\omega_j^\top \mu_i)^2)\right) D,
\end{aligned}$$

and

$$\begin{aligned}
&\Phi_j^{-\top}(t)\Phi_j^\top(s)(\mathfrak{m}_{j,i}(s) \otimes H_{2,i}(\widehat{\mathscr{B}}_i(s)))\Phi_j(s)\Phi_j^{-1}(t) \\
&= \partial_{13} h_i(\widehat{\mathscr{B}}_i(s))\frac{\Phi_{j,11}(s)}{\Phi_{j,11}(t)} \begin{bmatrix} \frac{\Phi_{j,11}(s)}{\Phi_{j,11}(t)}\mathfrak{m}_{j,i,1}(s) + \frac{\Phi_{j,21}(s)-\Phi_{j,21}(t)}{\Phi_{j,11}(t)}\mathfrak{m}_{j,i,2}(s) & 0 \\ \mathfrak{m}_{j,i,2}(s) & 0 \end{bmatrix},
\end{aligned} \tag{426}$$

where $\mathfrak{m}_{j,i,1}(t)$ and $\mathfrak{m}_{j,i,2}(t)$ denote the two elements composing $\mathfrak{m}_{j,i}(t) \in \mathbb{R}^2$. Similarly, we obtain

$$\begin{aligned}
&\Phi_j^{-\top}(t)\Phi_j^\top(s)(H_{2,i}(\widehat{\mathscr{B}}_i(s)) \otimes \mathfrak{m}_{j,i}(s))\Phi_j(s)\Phi_j^{-1}(t) \\
&= \partial_{13} h_i(\widehat{\mathscr{B}}_i(s))\frac{\Phi_{j,11}(s)}{\Phi_{j,11}(t)} \begin{bmatrix} \frac{\Phi_{j,11}(s)}{\Phi_{j,11}(t)}\mathfrak{m}_{j,i,1}(s) + \frac{\Phi_{j,21}(s)-\Phi_{j,21}(t)}{\Phi_{j,11}(t)}\mathfrak{m}_{j,i,2}(s) & \mathfrak{m}_{j,i,2}(s) \\ 0 & 0 \end{bmatrix}
\end{aligned}$$

We obtain

$$\begin{aligned}
\mathscr{V}_j(t) &= \frac{1}{\Phi_{j,11}^2(t)} \begin{bmatrix} X_0^\top \omega_j \omega_j^\top X_0 & \Phi_{j,11}(t) X_0^\top \omega_j \omega_j^\top X^\star \\ -2\Phi_{j,21}(t) X_0^\top \omega_j \omega_j^\top X^\star + \Phi_{j,21}^2(t)(X^\star)^\top \omega_j \omega_j^\top X^\star & -\Phi_{j,11}(t)\Phi_{j,21}(t)(X^\star)^\top \omega_j \omega_j^\top X^\star \\ \Phi_{j,11}(t) X_0^\top \omega_j \omega_j^\top X^\star & \\ -\Phi_{j,21}(t)\Phi_{j,11}(t)(X^\star)^\top \omega_j \omega_j^\top X^\star & \Phi_{11}^2(t)(X^\star)^\top \omega_j \omega_j^\top X^\star \end{bmatrix} \\
&- \int_0^t 2\gamma(s)\sum_{i=1}^2 p_i \partial_{13} h_i(\widehat{\mathscr{B}}_i(s))\frac{\Phi_{j,11}(s)}{\Phi_{j,11}(t)} \\
&\quad \times \begin{bmatrix} 2\left(\frac{\Phi_{j,11}(s)}{\Phi_{j,11}(t)}\mathfrak{m}_{j,i,1}(s) + \frac{\Phi_{j,21}(s)-\Phi_{j,21}(t)}{\Phi_{j,11}(t)}\mathfrak{m}_{j,i,2}(s)\right) & \mathfrak{m}_{j,i,2}(s) \\ \mathfrak{m}_{j,i,2}(s) & 0 \end{bmatrix} \mathrm{d}s \\
&+ \int_0^t \frac{\gamma(s)^2}{d}\frac{\Phi_{j,11}^2(s)}{\Phi_{j,11}^2(t)}\left(\sum_{i=1}^2 p_i \mathcal{I}_i(\widehat{\mathscr{B}}_i(s))(\lambda_j^{(i)} + (\omega_j^\top \mu_i)^2)\right) \mathrm{D}\, \mathrm{d}s.
\end{aligned}$$

We also have

$$\Phi_j^{-\top}(t,z)\mathfrak{m}_{j,\ell}(0) = \frac{1}{\Phi_{j,11}(t)} \begin{bmatrix} X_0^\top \omega_j \omega_j^\top \mu_\ell \\ -\Phi_{j,21}(t)(X^\star)^\top \omega_j \omega_j^\top \mu_\ell \\ \Phi_{j,11}(t)(X^\star)^\top \omega_j \omega_j^\top \mu_\ell \end{bmatrix}. \tag{427}$$

We also have

$$\begin{aligned}
&\Phi_j^{-\top}(t)\Phi_j^\top(s)\Big(-2\gamma(s)\sum_{i=1}^2 p_i \langle \omega_j, \mu_\ell\rangle\langle \omega_j, \mu_i\rangle \cdot H_{2,i}(\widehat{\mathscr{B}}_i(s))\Big) \\
&= -2\gamma(s)\sum_{i=1}^2 p_i \frac{\Phi_{j,11}(s)}{\Phi_{j,11}(t)}\langle \omega_j, \mu_\ell\rangle\langle \omega_j, \mu_i\rangle \cdot H_{2,i}(\widehat{\mathscr{B}}_i(s)).
\end{aligned} \tag{428}$$

Hence, we obtain

$$\mathfrak{m}_{j,\ell}(t) = \frac{1}{\Phi_{j,11}(t)} \begin{bmatrix} X_0^\top \omega_j \omega_j^\top \mu_\ell \\ -\Phi_{j,21}(t)(X^\star)^\top \omega_j \omega_j^\top \mu_\ell \\ \\ \Phi_{j,11}(t)(X^\star)^\top \omega_j \omega_j^\top \mu_\ell \end{bmatrix} - \int_0^t 2\gamma(s) \sum_{i=1}^2 p_i \frac{\Phi_{j,11}(s)}{\Phi_{j,11}(t)} \langle \omega_j, \mu_\ell \rangle \langle \omega_j, \mu_i \rangle \cdot H_{2,i}(\widehat{\mathscr{B}}_i(s))\, \mathrm{d}s. \tag{429}$$

The result (431) now immediately follows from summing over j from 1 to d. Indeed, we have for $\ell = 1, 2$

$$\mathscr{B}_\ell(t) = \sum_{j=1}^d \lambda_j^{(\ell)} \mathscr{V}_j(t), \quad \mathfrak{m}_\ell = \sum_{j=1}^d \mathfrak{m}_{j,\ell}(t) \quad \text{and} \quad \widehat{\mathscr{B}}_{\ell,44}(t) = \sum_{j=1}^d \mathscr{V}_{j,11}(t). \tag{430}$$

Evolution of $\widehat{\mathscr{B}}_\ell(t)$ for $\ell = 1, 2$.

$$\widehat{\mathscr{B}}_\ell(t) = \begin{bmatrix} \mathscr{B}_\ell(t) & \mathfrak{m}_\ell(t) & 0 \\ \mathfrak{m}_\ell(t)^\top & \|\mu_\ell\|^2 & 0 \\ 0 & 0 & \widehat{\mathscr{B}}_{\ell,44}(t) \end{bmatrix}, \quad \mathscr{B}_\ell(t) = \begin{bmatrix} \mathscr{B}_{\ell,11}(t) & \mathscr{B}_{\ell,12}(t) \\ \mathscr{B}_{\ell,21}(t) & \mathscr{B}_{\ell,22}(t) \end{bmatrix}$$

where

$$\begin{aligned}
\mathscr{B}_{\ell,11}(t) &= X_0^\top \tfrac{K_\ell}{\Phi_{11}^2(t,\mathcal{K})} X_0 - 2X_0^\top \tfrac{K_\ell \Phi_{21}(t,\mathcal{K})}{\Phi_{11}^2(t,\mathcal{K})} X^\star + (X^\star)^\top \tfrac{K_\ell \Phi_{21}^2(t,\mathcal{K})}{\Phi_{11}^2(t,\mathcal{K})} X^\star \\
&\quad - 4\int_0^t \gamma(s) \sum_{i=1}^2 p_i \partial_{13} h_i(\widehat{\mathscr{B}}_i(s)) \sum_{j=1}^d \lambda_j^{(\ell)} \frac{\Phi_{j,11}(s)}{\Phi_{j,11}(t)} \Bigg(\frac{\Phi_{j,11}(s)}{\Phi_{j,11}(t)} \mathfrak{m}_{j,i,1}(s) \\
&\qquad + \frac{\Phi_{j,21}(s) - \Phi_{j,21}(t)}{\Phi_{j,11}(t)} \mathfrak{m}_{j,i,2}(s) \Bigg) \mathrm{d}s \\
&\quad + \int_0^t \frac{\gamma(s)^2}{d} \sum_{i=1}^2 p_i \mathcal{I}_i(\widehat{\mathscr{B}}_i(s)) \operatorname{Tr}((K_i + \mu_i \mu_i^\top) K_\ell \tfrac{\Phi_{11}^2(s,\mathcal{K})}{\Phi_{11}^2(t,\mathcal{K})})\ \mathrm{d}s, \\
\mathscr{B}_{\ell,12}(t) &= X_0^\top \tfrac{K_\ell}{\Phi_{11}(t,\mathcal{K})} X^\star - (X^\star)^\top \tfrac{K_\ell \Phi_{21}(t,\mathcal{K})}{\Phi_{11}(t,\mathcal{K})} X^\star \\
&\quad - \int_0^t 2\gamma(s) \sum_{i=1}^2 p_i \partial_{13} h_i(\widehat{\mathscr{B}}_i(s)) \sum_{j=1}^d \lambda_j^{(\ell)} \frac{\Phi_{j,11}(s)}{\Phi_{j,11}(t)} \mathfrak{m}_{j,i,2}(s)\, \mathrm{d}s, \\
\mathfrak{m}_{\ell,1}(t) &= X_0^\top \Phi_{11}^{-1}(t,\mathcal{K}) \mu_\ell - (X^\star)^\top \tfrac{\Phi_{21}(t,\mathcal{K})}{\Phi_{11}(t,\mathcal{K})} \mu_\ell \\
&\quad - 2\int_0^t \gamma(s) \sum_{i=1}^2 p_i \partial_{13} h_i(\widehat{\mathscr{B}}_i(s)) \cdot \mu_i^\top \tfrac{\Phi_{11}(s,\mathcal{K})}{\Phi_{11}(t,\mathcal{K})} \mu_\ell\ \mathrm{d}s, \\
\mathscr{B}_{\ell,44}(t) &= X_0^\top \Phi_{11}^{-2}(t,\mathcal{K}) X_0 - 2X_0^\top \tfrac{\Phi_{21}(t,\mathcal{K})}{\Phi_{11}^2(t,\mathcal{K})} X^\star + (X^\star)^\top \tfrac{\Phi_{21}^2(t,\mathcal{K})}{\Phi_{11}^2(t,\mathcal{K})} X^\star \\
&\quad - 4\int_0^t \gamma(s) \sum_{i=1}^2 p_i \partial_{13} h_i(\widehat{\mathscr{B}}_i(s)) \sum_{j=1}^d \frac{\Phi_{j,11}(s)}{\Phi_{j,11}(t)} \Bigg(\frac{\Phi_{j,11}(s)}{\Phi_{j,11}(t)} \mathfrak{m}_{j,i,1}(s) \\
&\qquad + \frac{\Phi_{j,21}(s) - \Phi_{j,21}(t)}{\Phi_{j,11}(t)} \mathfrak{m}_{j,i,2}(s) \Bigg) \mathrm{d}s \\
&\quad + \int_0^t \frac{\gamma(s)^2}{d} \sum_{i=1}^2 p_i \mathcal{I}_i(\widehat{\mathscr{B}}_i(s)) \operatorname{Tr}((K_i + \mu_i \mu_i^\top) \tfrac{\Phi_{11}^2(s,\mathcal{K})}{\Phi_{11}^2(t,\mathcal{K})})\ \mathrm{d}s, \\
\mathscr{B}_{\ell,21}(t) &= \mathscr{B}_{\ell,12}(t), \quad \mathscr{B}_{\ell,22}(t) = (X^\star)^\top K_\ell X^\star \quad \text{and} \quad \mathfrak{m}_{\ell,2} = (X^\star)^\top \mu_\ell
\end{aligned} \tag{431}$$

The function $\Phi_{11}(t,\mathcal{K})$ and $\Phi_{21}(t,\mathcal{K})$, by solving a differential equation, are given by

$$\Phi_{11}(t,\mathcal{K}) = \exp\left(\int_0^t 2\gamma(s)\sum_{i=1}^2 p_i\Big(K_i\partial_{11}h_i(\widehat{\mathscr{B}}_i(s)) + \frac{\lambda}{2} + \partial_{44}h_i(\widehat{\mathscr{B}}_i(s))\Big)\,\mathrm{d}s\right),$$
$$\text{and}\quad \Phi_{21}(t,\mathcal{K}) = \int_0^t 2\gamma(s)\Big(\sum_{i=1}^2 p_i K_i\partial_{21}h_i(\widehat{\mathscr{B}}_i(s))\Big)\Phi_{11}(s,\mathcal{K})\,\mathrm{d}s. \tag{432}$$

Let $\mathscr{R}^{\mathrm{adv}}(t)$ denote the det. equiv. of $\mathcal{R}^{\mathrm{adv}}$. Given that $\mathcal{R}^{\mathrm{adv}}(X) = \sum_{i=1}^2 p_i \cdot h_i(\widehat{B}_i(W))$ then we obtain

$$\mathscr{R}^{\mathrm{adv}}(t) = \sum_{i=1}^2 p_i \cdot h_i(\widehat{\mathscr{B}}_i(t)).$$

Analogously to the spectral decomposition in (413), we may use a similar argument to decompose $\mathcal{Z}(t,z)$ as follows

$$\mathcal{Z}(t,z) = \sum_{j=1}^d \frac{\mathcal{Z}_j(t)}{\prod_{\ell=1}^2(\lambda_j^{(\ell)} - z_\ell)}, \tag{433}$$

where

$$\mathcal{Z}_j(t) \overset{\text{def}}{=} \begin{bmatrix} \mathscr{V}_j(t) & \mathfrak{M}_j(t) \\ \mathfrak{M}_j(t)^\top & \mu^\top\omega_j\omega_j^\top\mu \end{bmatrix} \quad \text{with} \quad \mathfrak{M}_j(t) \overset{\text{def}}{=} [\mathfrak{m}_{j,1} \quad \mathfrak{m}_{j,2}]. \tag{434}$$

Using (407), we obtain the differential equation

$$\frac{\mathrm{d}\mathcal{Z}_j(t)}{\mathrm{d}t} = -\mathcal{Z}_j(t)\mathcal{B}_j(t) - \mathcal{B}_j(t)^\top\mathcal{Z}_j(t) + \mathcal{G}_j(t), \tag{435}$$

where we define for e_i the standard basis vectors of $\mathbb{R}^2$

$$\mathcal{B}_j(t) \overset{\text{def}}{=} \begin{bmatrix} \mathcal{A}_j(t) & 0 \\ \mathcal{C}(t) & 0 \end{bmatrix}, \quad \mathcal{C}(t) = 2\gamma(t)\sum_{i=1}^2 p_i\,\mathrm{e}_i \otimes H_{2,i}(\widehat{\mathscr{B}}_i(t)),$$
$$\mathcal{A}_j(t) \overset{\text{def}}{=} 2\gamma(t)\sum_{i=1}^2 p_i\Big(\lambda_j^{(i)}H_{1,i}(\widehat{\mathscr{B}}_i(t)) + \Big(\frac{\lambda}{2} + \partial_{44}h_i(\widehat{\mathscr{B}}_i(t))\Big)\,\mathrm{D}\Big), \tag{436}$$
$$\text{and}\quad \mathcal{G}_j(t) \overset{\text{def}}{=} \begin{bmatrix} \frac{\gamma(t)^2}{d}\left(\sum_{i=1}^2 p_i\mathcal{I}_i(\widehat{\mathscr{B}}_i(t))(\lambda_j^{(i)} + (\omega_j^\top\mu_i)^2)\right)\mathrm{D} & 0 \\ 0 & 0 \end{bmatrix}.$$

Analogously to (415), we define the fundamental ODE

$$\dot{\Xi}_j = \mathcal{B}_j(t)\Xi_j, \quad \Xi_j(0) = I_4. \tag{437}$$

Given that $\mathcal{B}_j(t)$ is a block lower triangular matrix then we obtain from the initial conditions

$$\Xi_j(t) = \begin{bmatrix} \Phi_j(t) & 0 \\ \Upsilon_j(t) & I_2 \end{bmatrix}, \quad \text{with} \quad \Upsilon_j(t) = \int_0^t \mathcal{C}_j(s)\Phi_j(s)\,\mathrm{d}s, \tag{438}$$

and where Φ_j is defined in (422). Using a similar argument as (416), we obtain

$$\mathcal{Z}_j(t) = \Xi_j(t)^{-\top}\mathcal{Z}_j(0)\Xi_j(t)^{-1} + \int_0^t \Xi_j(t)^{-\top}\Xi_j(s)^\top\mathcal{G}_j(s)\Xi_j(s)\Xi_j(t)^{-1}\,\mathrm{d}s. \tag{439}$$

Now, setting $x_j \overset{\text{def}}{=} \hat{X}_0^\top\omega_j \in \mathbb{R}^2$ and $q_j \overset{\text{def}}{=} \mu^\top\omega_j \in \mathbb{R}^2$ then we have

$$\mathscr{V}_j(0) = x_jx_j^\top, \mathcal{M}_j(t) = x_jq_j^\top \quad \text{and} \quad \mu^\top\omega_j\omega_j^\top\mu = q_jq_j^\top. \tag{440}$$

It thus follows that

$$\mathcal{Z}_j(0) = \begin{bmatrix} x_j \\ q_j \end{bmatrix}\begin{bmatrix} x_j \\ q_j \end{bmatrix}^\top \succeq 0. \tag{441}$$

By Assumption 7, it is clear that for $i = 1, 2$, $I_i(\mathscr{B}(s)) \geq 0$ for $0 \leq s \leq t$ and every $t \geq 0$. Since $\mathcal{Z}_j(0)$ are positive semidefinite (psd) and congruence preserves positive semidefiniteness, it follows that $\mathcal{Z}_j(t)$ is psd for every $1 \leq j \leq d$ and $t \geq 0$.

Since the diagonal block matrices of a psd matrix are psd, it follows that $\mathscr{V}_j(t)$ is psd for all $1 \leq j \leq d$ and $t \geq 0$ and thus $\mathscr{B}_\ell(t)$ is psd for all $t \geq 0$ and $\ell = 1, 2$. Since the diagonal entries of a psd matrix are nonnegative, it follows that $\mathscr{V}_{j,11}(t) \geq 0$ for all $1 \leq j \leq d$ and $t \geq 0$ and thus $\mathscr{B}_{\ell,44}(t) \geq 0$ for every $t \geq 0$ and $\ell = 1, 2$.

Now, define $\mathcal{Z}(t) = \oint \mathcal{Z}(t) \, \mathrm{D}z$ then by (433) we have $\mathcal{Z}(t) = \sum_{j=1}^{d} \mathcal{Z}_j(t)$ which implies $\mathcal{Z}(t)$ is psd for all $t \geq 0$. Every principal submatrix of $\mathcal{Z}(t)$ is then psd which implies the following two matrices obtained by removing the third column and row to obtain $\mathcal{Z}_2(t)$ and by removing the fourth column and row to obtain $\mathcal{Z}_1(t)$ are psd:

$$\mathcal{Z}_i(t) \stackrel{\text{def}}{=} \begin{bmatrix} \sum_{j=1}^{d} \mathscr{V}_j(t) & \mathfrak{m}_i \\ \mathfrak{m}_i^\top & \|\mu_i\|^2 \end{bmatrix} \quad \text{for} \quad i = 1, 2. \tag{442}$$

When $\mu_1 \neq 0$ and $\mu_2 \neq 0$ then from the Schur complement, we obtain the following inequality for $i = 1, 2$

$$\|\mathfrak{m}_i\|^2 \leq \|\mu_i\|^2 \operatorname{Tr}(\sum_{j=1}^{d} \mathscr{V}_j(t)) \leq \|\mu_i\|^2 \mathscr{N}(t). \tag{443}$$

We close off this section by providing the hard label version of the system of ODEs (26) which simplifies to

$$\begin{aligned} \frac{\mathrm{d}\mathscr{V}_{j,11}(t)}{\mathrm{d}t} &= -4\gamma(t) \sum_{i=1}^{2} p_i \Bigg(\left(\lambda_j^{(i)} \partial_{11} h_i(\widehat{\mathscr{B}}_i(t)) + \frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) \right) \mathscr{V}_{j,11}(t) \\ &\quad + \partial_{13} h_i(\widehat{\mathscr{B}}_i(t)) \mathfrak{m}_{j,i,1}(t) \Bigg) + \frac{\gamma(t)^2}{d} \sum_{i=1}^{2} p_i \mathcal{I}_i(\widehat{\mathscr{B}}_i(t)) (\lambda_j^{(i)} + (\omega_j^\top \mu_i)^2), \\ \frac{\mathrm{d}\mathfrak{m}_{j,\ell,1}(t)}{\mathrm{d}t} &= -2\gamma(t) \sum_{i=1}^{2} p_i \Bigg(\left(\lambda_j^{(i)} \partial_{11} h_i(\widehat{\mathscr{B}}_i(t)) + \frac{\lambda}{2} + \partial_{44} h_i(\widehat{\mathscr{B}}_i(t)) \right) \mathfrak{m}_{j,\ell,1}(t) \\ &\quad + \partial_{13} h_i(\widehat{\mathscr{B}}_i(t)) \langle \omega_j, \mu_\ell \rangle \langle \omega_j, \mu_i \rangle \Bigg), \end{aligned} \tag{444}$$

with initial condition $\mathscr{V}_{j,11}(0) = X_0^\top \omega_j \omega_j^\top X_0$ and $\mathfrak{m}_{j,\ell,1}(0) = X_0^\top \omega_j \omega_j^\top \mu_\ell$.

I Analysis of Examples

In this section, we derive the functions h_i, ∇h_i, f_i, f_i', $\mathbb{E}_{a,\epsilon}[f_i'^2]$ for two examples fitting our framework: least squares with soft labels and binary logistic regression with hard labels. See Section 2 and Appendix D in Collins-Woodfin et al. (2024a) for more examples.

I.1 Least Squares

We study the dynamics of the ℓ_2-adversarial least squares with regularization parameter $\lambda = 0$ under the assumption that the data $a \sim \mathcal{N}(0, K)$. Given that this is equivalent to $p_1 = 1$ and $p_2 = 0$ with $\mu_1 = 0$, we drop the subscript i in this section. The classical least squares setting consists of minimizing $X \in \mathbb{R}^d$ over the risk function

$$\mathcal{R}(X) = \frac{1}{2} \mathbb{E}_{a,\epsilon} \left[\left(X^\top a - ((X^\star)^\top a + \epsilon) \right)^2 \right] = \frac{1}{2} (X - X^\star)^\top K (X - X^\star) + \frac{\eta^2}{2},$$

with respect to targets $y = X^{\star\top} a + \epsilon$. It is easy to identify the functions f as

$$f(X^\top a, \epsilon) = \frac{1}{2} \left(X^\top a - ((X^\star)^\top a + \epsilon) \right)^2. \tag{445}$$

The adversarial formulation consists of the following risk

$$\begin{aligned}
\mathcal{R}^{\text{adv}}(X) &\stackrel{\text{def}}{=} \frac{1}{2}\mathbb{E}_{a,\epsilon}\left[\max_{|s|\leq 1}\left((X^\top a - ((X^\star)^\top a + \epsilon)) + \delta s\|X\|\right)^2\right] \\
&= \frac{1}{2}\mathbb{E}_{a,\epsilon}\left[\max_{s\in\{-1,1\}}\left(((X - X^\star)^\top a - \epsilon) + \delta s\|X\|\right)^2\right] \\
&= \frac{1}{2}\mathbb{E}_{a,\epsilon}\left[\left(\left|(X - X^\star)^\top a - \epsilon\right| + \delta\|X\|\right)^2\right] \\
&= \frac{1}{2}\left(\mathbb{E}_{a,\epsilon}\left[\left((X - X^\star)^\top a - \epsilon\right)^2\right] + 2\delta\|X\|\,\mathbb{E}_{a,\epsilon}\left[\left|(X - X^\star)^\top a - \epsilon\right|\right] + \frac{\delta^2}{2}\|X\|^2\right) \\
&= \mathcal{R}(X) + \delta\sqrt{\frac{2}{\pi}}\|X\|\sqrt{2\mathcal{R}(X)} + \frac{\delta^2}{2}\|X\|^2
\end{aligned} \tag{446}$$

since $u \mapsto u^2$ is convex. The function $h : \mathbb{R}^{4\times 4} \to \mathbb{R}$ must satisfy $h(\widehat{B}(W)) = \mathcal{R}^{\text{adv}}(X)$. Similarly to Section B.1 in Collins-Woodfin et al. (2024a), we identify the following terms:

$$z_{11} = X^\top KX, \quad z_{12} = X^\top KX^\star \quad z_{21} = X^{\star\top} KX,$$
$$z_{22} = X^{\star\top} KX^\star, \quad z_{44} = \|X\|^2$$

Hence, we have the identifications

$$\mathcal{R}(X) = \frac{1}{2}\left(z_{11} - z_{12} - z_{21} + z_{22}\right) + \frac{\eta^2}{2}.$$

which leads to

$$h\left(\begin{bmatrix} z_{11} & z_{12} & 0 & 0 \\ z_{21} & z_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & z_{44} \end{bmatrix}\right) = \frac{1}{2}\left(z_{11} - z_{12} - z_{21} + z_{22}\right) + \frac{\eta^2}{2} + \delta\sqrt{\frac{2}{\pi}}\sqrt{z_{44}\left(\left(z_{11} - z_{12} - z_{21} + z_{22}\right) + \eta^2\right)} + \frac{\delta^2}{2}z_{44}.$$

Consequently, it is clear that

$$\mathscr{R}(t) = \frac{1}{2}\left(\mathscr{B}_{11}(t) - \mathscr{B}_{12}(t) - \mathscr{B}_{21}(t) + \mathscr{B}_{22}(t)\right) + \frac{\eta^2}{2}, \tag{447}$$

where we denote $\mathscr{R}(t)$ as the det. equiv. of $\mathcal{R}(X)$. Similarly, from the definition of z_{44}, it is clear that $\widehat{\mathscr{B}}_{44}(t)$ is the det. equiv. of $\|X\|^2$. From (446) and $\mathscr{R}^{\text{adv}}(t) = h \circ (\widehat{\mathscr{B}}(t))$, we obtain

$$\mathscr{R}^{\text{adv}}(t) = \mathscr{R}(t) + \delta\sqrt{\frac{2}{\pi}}\sqrt{2\widehat{\mathscr{B}}_{44}(t)\mathscr{R}(t)} + \frac{\delta^2}{2}\widehat{\mathscr{B}}_{44}(t). \tag{448}$$

Continuing, we then see that

$$\nabla h(B) = \left[\begin{array}{c|c|c|c} \partial_{11}h(B) & \partial_{12}h(B) & 0 & 0 \\ \hline \partial_{21}h(B) & \partial_{22}h(B) & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & \partial_{44}h(B) \end{array}\right],$$

with

$$\partial_{ij}h(B) = \frac{1}{2} + \frac{\delta}{2}\sqrt{\frac{2}{\pi}}\sqrt{\frac{z_{44}}{z_{11} - z_{12} - z_{21} + z_{22} + \eta^2}}, \text{when} \quad (i,j) \in \{(1,1),(2,2)\},$$

$$\partial_{ij}h(B) = -\frac{1}{2} - \frac{\delta}{2}\sqrt{\frac{2}{\pi}}\sqrt{\frac{z_{44}}{z_{11} - z_{12} - z_{21} + z_{22} + \eta^2}}, \text{when} \quad (i,j) \in \{(1,2),(2,1)\}$$

and $\partial_{44}h(B) = \frac{\delta^2}{2} + \frac{\delta}{2}\sqrt{\frac{2}{\pi}}\sqrt{\frac{z_{11} - z_{12} - z_{21} + z_{22} + \eta^2}{z_{44}}}.$

It immediately follows that

$$H_1(\widehat{\mathscr{B}}(t)) = \frac{1}{2}\begin{bmatrix} 1+\delta\sqrt{\frac{2}{\pi}}\sqrt{\frac{\widehat{\mathscr{B}}_{44}(t)}{2\mathscr{R}(t)}} & 0 \\ -1-\delta\sqrt{\frac{2}{\pi}}\sqrt{\frac{\widehat{\mathscr{B}}_{44}(t)}{2\mathscr{R}(t)}} & 0 \end{bmatrix}, \quad H_2(\widehat{\mathscr{B}}(t)) = \frac{1}{2}\begin{bmatrix} 0 \\ 0 \end{bmatrix},$$
$$\text{and} \quad \partial_{44}h(\widehat{\mathscr{B}}(t)) = \frac{\delta^2}{2} + \frac{\delta}{2}\sqrt{\frac{2}{\pi}}\sqrt{\frac{2\mathscr{R}(t)}{\widehat{\mathscr{B}}_{44}(t)}}$$

It is clear that we lose differentiability at $\mathscr{R}(t) = \frac{1}{2}\left(\mathscr{B}_{11}(t) - \mathscr{B}_{12} - \mathscr{B}_{21} + \mathscr{B}_{22}\right) + \frac{\eta^2}{2} = 0$ and $\widehat{\mathscr{B}}_{44}(t) = 0$. . To apply the theory in this paper, we work on the set $\mathcal{U}$ defined as follows

$$\mathcal{U} \stackrel{\text{def}}{=} \left\{\widehat{B} : \widehat{B}_{11} - \widehat{B}_{12} - \widehat{B}_{21} + \widehat{B}_{22} + \eta > 0, \quad \widehat{B}_{44} > 0\right\} \tag{449}$$

Recall $\bar{s} = \operatorname{argmax}_{s\in\{-1,1\}}\left(X^\top a + \delta s\|X\| - ((X^\star)^\top a + \epsilon)\right)^2$. From (445), considering f as a map $r \mapsto \frac{1}{2}(r - (X^{\star\top}a + \epsilon))^2$ we see that the almost everywhere derivative of f with respect to $X^\top a$ is given by

$$f'(x) = x - (X^{\star\top}a - \epsilon)$$

So it follows from our initial computation of $\mathcal{R}^{\text{adv}}$ that

$$\begin{aligned}\mathbb{E}_{a,\epsilon}[f'((X^\top a + \bar{s}\delta\|X\|, \epsilon))^2] &= \mathbb{E}_{a,\epsilon}[(X^\top a + \bar{s}\delta\|X\| - (X^{\star\top}a - \epsilon))^2] \\ &= \mathbb{E}_{a,\epsilon}[\max_{s\in\{-1,1\}}(((X - X^\star)^\top a - \epsilon) + s\delta\|X\|)^2] \\ &= 2\mathcal{R}^{\text{adv}}(X),\end{aligned}$$

which leads to

$$\mathcal{I}(\widehat{\mathscr{B}}(t)) = 2\mathscr{R}(t) + 2\delta\sqrt{\frac{2}{\pi}}\sqrt{2\widehat{\mathscr{B}}_{44}(t)\mathscr{R}(t)} + \delta^2\widehat{\mathscr{B}}_{44}(t) = 2\mathscr{R}^{\text{adv}}(t).$$

We now have all the components to explicitly compute Φ as defined in (432) and $\widehat{\mathscr{B}}(t)$ as defined in (431). We have

$$\begin{aligned}\Phi_{11}(t,K) = \exp\Bigg(\int_0^t \gamma(s)\Bigg(&\left(1+\delta\sqrt{\frac{2}{\pi}}\sqrt{\frac{\widehat{\mathscr{B}}_{44}(s)}{2\mathscr{R}(s)}}\right)K \\ &+ \left(\delta^2 + \delta\sqrt{\frac{2}{\pi}}\sqrt{\frac{2\mathscr{R}(s)}{\widehat{\mathscr{B}}_{44}(s)}}\right)I_d\Bigg)\,\mathrm{d}s\Bigg), \qquad (450)\\ \text{and } \Phi_{21}(t,K) = -\int_0^t \gamma(s)&\left(1+\delta\sqrt{\frac{2}{\pi}}\sqrt{\frac{\widehat{\mathscr{B}}_{44}(s)}{2\mathscr{R}(s)}}\right)K\Phi_{11}(s,K)\,\mathrm{d}s.\end{aligned}$$

We define the following adaptive learning rate and regularization parameter respectively

$$\tilde{\gamma}^{\text{eff}}(t) = \gamma(t)\left(1+\delta\sqrt{\frac{2}{\pi}}\sqrt{\frac{\widehat{\mathscr{B}}_{44}(t)}{2\mathscr{R}(t)}}\right) \quad \text{and} \quad \tilde{\lambda}^{\text{eff}}(t) = \frac{\delta^2 + \delta\sqrt{\frac{2}{\pi}}\sqrt{\frac{2\mathscr{R}(t)}{\widehat{\mathscr{B}}_{44}(t)}}}{1+\delta\sqrt{\frac{2}{\pi}}\sqrt{\frac{\widehat{\mathscr{B}}_{44}(t)}{2\mathscr{R}(t)}}}. \tag{451}$$

Define the following integral terms

$$\tilde{\Gamma}(t) = \int_0^t \tilde{\gamma}^{\text{eff}}(s)\,\mathrm{d}s \quad \text{and} \quad \Lambda(t) = \int_0^t \tilde{\gamma}^{\text{eff}}(s)\tilde{\lambda}^{\text{eff}}(s)\,\mathrm{d}s. \tag{452}$$

Then plugging these terms into (450), we obtain

$$\Phi_{11}(t,K) = e^{K\tilde{\Gamma}(t)+\Lambda(t)}, \quad \text{and} \quad \Phi_{21}(t,K) \quad = -\int_0^t K\tilde{\gamma}^{\text{eff}}(s)e^{K\tilde{\Gamma}(s)+\Lambda(s)}\,\mathrm{d}s. \tag{453}$$

We can also view $\Phi_{21}(t,K)$ the following way

$$\Phi_{21}(t,K) = 1 - \Phi_{11}(t,K) + \int_0^t \tilde{\gamma}^{\text{eff}}(s)\tilde{\lambda}^{\text{eff}}(s)\Phi_{11}(s,K)\,\mathrm{d}s.$$

From (431), the first three terms of $\mathscr{B}_{11}(t)$ and $\widehat{\mathscr{B}}_{44}(t)$ form a perfect square from which it follows

$$\begin{aligned}
\mathscr{B}_{11}(t) &= \langle K, \left(\Phi_{11}^{-1}(t,K)\left(X_0 - \Phi_{21}(t,K)X^\star\right)\right)^{\otimes 2}\rangle \\
&\qquad + \frac{2}{d}\int_0^t \gamma^2(s)\mathscr{R}^{\text{adv}}(s)\text{Tr}\big(K^2 e^{-2K(\tilde{\Gamma}(t)-\tilde{\Gamma}(s))-2(\Lambda(t)-\Lambda(s))}\big)\,\mathrm{d}s, \\
\widehat{\mathscr{B}}_{44}(t) &= \|\Phi_{11}^{-1}(t,K)(X_0 - \Phi_{21}(t,K)X^\star)\|^2 \\
&\qquad + \frac{2}{d}\int_0^t \gamma^2(s)\mathscr{R}^{\text{adv}}(s)\text{Tr}\big(K e^{-2K(\tilde{\Gamma}(t)-\tilde{\Gamma}(s))-2(\Lambda(t)-\Lambda(s))}\big)\,\mathrm{d}s,
\end{aligned} \tag{454}$$

Since

$$\begin{aligned}
\mathscr{B}_{12}(t) = \mathscr{B}_{21}(t) &= X_0^\top \tfrac{K}{\Phi_{11}(t,K)}X^\star - (X^\star)^\top \tfrac{K\Phi_{21}(t,K)}{\Phi_{11}(t,K)}X^\star \\
&= \left(X_0 - \Phi_{21}(t,K)X^\star\right)^\top \Phi_{11}^{-1}(t,K)KX^\star
\end{aligned}$$

From this, it follows that

$$\begin{aligned}
&\mathscr{B}_{11}(t) - \mathscr{B}_{12}(t) - \mathscr{B}_{21}(t) + \mathscr{B}_{22}(t) \\
&= \langle K, (e^{-K\tilde{\Gamma}(t)-\Lambda(t)}\left(X_0 + \left(\int_0^t K\tilde{\gamma}^{\text{eff}}(s)e^{K\tilde{\Gamma}(s)+\Lambda(s)}\,\mathrm{d}s\right)X^\star\right) - X^\star)^{\otimes 2}\rangle \\
&\quad + \frac{2}{d}\int_0^t \gamma^2(s)\mathscr{R}^{\text{adv}}(s)\text{Tr}\big(K^2 e^{-2K(\tilde{\Gamma}(t)-\tilde{\Gamma}(s))-2(\Lambda(t)-\Lambda(s))}\big)\,\mathrm{d}s.
\end{aligned} \tag{455}$$

Hence, we obtain

$$\begin{aligned}
\mathscr{R}(t) &= \frac{1}{2}\langle K, (e^{-K\tilde{\Gamma}(t)-\Lambda(t)}\left(X_0 + \left(\int_0^t K\tilde{\gamma}^{\text{eff}}(s)e^{K\tilde{\Gamma}(s)+\Lambda(s)}\,\mathrm{d}s\right)X^\star\right) - X^\star)^{\otimes 2}\rangle + \frac{\eta^2}{2} \\
&\quad + \frac{1}{d}\int_0^t \gamma^2(s)\mathscr{R}^{\text{adv}}(s)\text{Tr}\big(K^2 e^{-2K(\tilde{\Gamma}(t)-\tilde{\Gamma}(s))-2(\Lambda(t)-\Lambda(s))}\big)\,\mathrm{d}s.
\end{aligned} \tag{456}$$

Plugging the values of Φ_{11} and Φ_{21} in (454), we obtain

$$\begin{aligned}
\widehat{\mathscr{B}}_{44}(t) &= \left\|e^{-K\tilde{\Gamma}(t)-\Lambda(t)}\left(X_0 + \left(\int_0^t K\tilde{\gamma}^{\text{eff}}(s)e^{K\tilde{\Gamma}(s)+\Lambda(s)}\,\mathrm{d}s\right)X^\star\right)\right\|^2 \\
&\qquad + \frac{2}{d}\int_0^t \gamma^2(s)\mathscr{R}^{\text{adv}}(s)\text{Tr}\big(K e^{-2K(\tilde{\Gamma}(t)-\tilde{\Gamma}(s))-2(\Lambda(t)-\Lambda(s))}\big)\,\mathrm{d}s.
\end{aligned} \tag{457}$$

Finally, combining (456) and (457), we recover $\mathscr{R}^{\text{adv}}(t)$ using (448). Note that we may express $\mathscr{R}^{\text{adv}}(t)$ in terms of $\tilde{\gamma}^{\text{eff}}(t)$ and $\tilde{\lambda}^{\text{eff}}(t)$ as follows

$$\mathscr{R}^{\text{adv}}(t) = \frac{\tilde{\gamma}^{\text{eff}}(t)}{\gamma(t)}\left(\mathscr{R}(t) + \frac{\tilde{\lambda}^{\text{eff}}(t)}{2}\widehat{\mathscr{B}}_{44}(t)\right). \tag{458}$$

Using this reformulation, we obtain

$$\begin{aligned}
\mathscr{R}(t) &= \frac{1}{2}\langle K, (e^{-K\tilde{\Gamma}(t)-\Lambda(t)}\left(X_0 + \left(\int_0^t K\tilde{\gamma}^{\text{eff}}(s)e^{K\tilde{\Gamma}(s)+\Lambda(s)}\,\mathrm{d}s\right)X^\star\right) - X^\star)^{\otimes 2}\rangle + \frac{\eta^2}{2} \\
&\quad + \frac{1}{d}\int_0^t \gamma(s)\tilde{\gamma}^{\text{eff}}(s)\left(\mathscr{R}(s) + \frac{\tilde{\lambda}^{\text{eff}}(s)}{2}\widehat{\mathscr{B}}_{44}(s)\right)\text{Tr}\big(K^2 e^{-2K(\tilde{\Gamma}(t)-\tilde{\Gamma}(s))-2(\Lambda(t)-\Lambda(s))}\big)\,\mathrm{d}s.
\end{aligned} \tag{459}$$

and

$$\widehat{\mathscr{B}}_{44}(t) = \left\| e^{-K\tilde{\Gamma}(t)-\Lambda(t)} \left(X_0 + \left(\int_0^t K\tilde{\gamma}^{\mathrm{eff}}(s) e^{K\tilde{\Gamma}(s)+\Lambda(s)} \, \mathrm{d}s \right) X^\star \right) \right\|^2 + \frac{2}{d} \int_0^t \gamma(s)\tilde{\gamma}^{\mathrm{eff}}(s) \left(\mathscr{R}(s) + \frac{\tilde{\lambda}^{\mathrm{eff}}(s)}{2} \widehat{\mathscr{B}}_{44}(s) \right) \mathrm{Tr}\big(K e^{-2K(\tilde{\Gamma}(t)-\tilde{\Gamma}(s))-2(\Lambda(t)-\Lambda(s))} \big) \, \mathrm{d}s. \tag{460}$$

The system of ODEs (26) in this setting simplifies to

$$\begin{aligned} \frac{\mathrm{d}\mathscr{V}_{j,11}(t)}{\mathrm{d}t} &= -2\gamma(t) \left(\lambda_j \left(1 + \delta\sqrt{\frac{2}{\pi}} \sqrt{\frac{\widehat{\mathscr{B}}_{44}(t)}{2\mathscr{R}(t)}} \right) + \delta^2 + \delta\sqrt{\frac{2}{\pi}} \sqrt{\frac{2\mathscr{R}(t)}{\widehat{\mathscr{B}}_{44}(t)}} \right) \mathscr{V}_{j,11}(t) \\ &\quad + 2\gamma(t)\lambda_j \left(1 + \delta\sqrt{\frac{2}{\pi}} \sqrt{\frac{\widehat{\mathscr{B}}_{44}(t)}{2\mathscr{R}(t)}} \right) \mathscr{V}_{j,21}(t) + 2\lambda_j \frac{\gamma^2(t)}{d} \mathscr{R}^{\mathrm{adv}}(t) \\ \frac{\mathrm{d}\mathscr{V}_{j,21}(t)}{\mathrm{d}t} &= -\gamma(t) \left(\lambda_j \left(1 + \delta\sqrt{\frac{2}{\pi}} \sqrt{\frac{\widehat{\mathscr{B}}_{44}(t)}{2\mathscr{R}(t)}} \right) + \delta^2 + \delta\sqrt{\frac{2}{\pi}} \sqrt{\frac{2\mathscr{R}(t)}{\widehat{\mathscr{B}}_{44}(t)}} \right) \mathscr{V}_{j,21}(t) \\ &\quad + \gamma(t)\lambda_j \left(1 + \delta\sqrt{\frac{2}{\pi}} \sqrt{\frac{\widehat{\mathscr{B}}_{44}(t)}{2\mathscr{R}(t)}} \right) \mathscr{V}_{j,22}(t), \end{aligned} \tag{461}$$

where we recall $\mathscr{V}_{j,22}(t) = X^{\star\top} \omega_j \omega_j^\top X^\star$. Hence, from (451) and (458), we obtain the simplified representation

$$\begin{aligned} \frac{\mathrm{d}\mathscr{V}_{j,11}(t)}{\mathrm{d}t} &= -2\tilde{\gamma}^{\mathrm{eff}}(t) \left(\lambda_j + \tilde{\lambda}^{\mathrm{eff}}(t) \right) \mathscr{V}_{j,11}(t) \\ &\quad + 2\tilde{\gamma}^{\mathrm{eff}}(t)\lambda_j \mathscr{V}_{j,21}(t) + 2\lambda_j \frac{\gamma(t)\tilde{\gamma}^{\mathrm{eff}}(t)}{d} \left(\mathscr{R}(t) + \frac{\tilde{\lambda}^{\mathrm{eff}}(t)}{2} \widehat{\mathscr{B}}_{44}(t) \right) \\ \frac{\mathrm{d}\mathscr{V}_{j,21}(t)}{\mathrm{d}t} &= -\tilde{\gamma}^{\mathrm{eff}}(t) \left(\lambda_j + \tilde{\lambda}^{\mathrm{eff}}(t) \right) \mathscr{V}_{j,21}(t) + \tilde{\gamma}^{\mathrm{eff}}(t)\lambda_j \mathscr{V}_{j,22}(t). \end{aligned} \tag{462}$$

I.1.1 Supporting Results

Lemma I.1. *As per Definition B.2, recall* $X^{\star,\mathrm{Adv}} = \operatorname{argmin}_X \mathcal{R}^{\mathrm{Adv}}(X)$. *The* ℓ_2*-adversarial risk* $\mathcal{R}^{\mathrm{Adv}}(X)$ *is strongly convex with parameter* $\delta^2(1 - \frac{2}{\pi})$ *and* $X^{\star,\mathrm{Adv}}$ *is its unique minimizer.*

Proof. Recall

$$\mathcal{R}^{\mathrm{adv}}(X) = \mathcal{R}(X) + \delta\sqrt{\frac{2}{\pi}} \|X\| \sqrt{2\mathcal{R}(X)} + \frac{\delta^2}{2} \|X\|^2. \tag{463}$$

Completing the square, we can rewrite this as

$$\mathcal{R}^{\mathrm{adv}}(X) = \frac{1}{2} \left(\sqrt{2\mathcal{R}(X)} + \delta\sqrt{\frac{2}{\pi}} \|X\| \right)^2 + \frac{\delta^2}{2} \left(1 - \frac{2}{\pi} \right) \|X\|^2. \tag{464}$$

The map $X \mapsto \frac{1}{2} \left(\sqrt{2\mathcal{R}(X)} + \delta\sqrt{\frac{2}{\pi}} \|X\| \right)^2$ is convex. For all $\delta > 0$, the second term is $\delta^2(1 - \frac{2}{\pi})$-strongly convex. Hence, the result follows. □

Lemma I.2. *Recall* $q(t) = \sqrt{\frac{\widehat{\mathscr{B}}_{44}(t)}{2\mathscr{R}(t)}}$. *For any* $T > 0$, *suppose* $\mathscr{R}(t) > 0$ *for all* $t \in [0, T]$ *then for all* $t \in [0, T]$

$$\frac{\left| \sum_{j=1}^d \lambda_j \left(\mathscr{V}_{j,11}(t) - \mathscr{V}_{j,12}(t) \right) \right|}{\mathscr{R}(t)} \leq 2\|K\|_{\mathrm{op}}^{1/2} q(t) + 2, \tag{465}$$

and

$$\frac{\left| \sum_{j=1}^d \lambda_j \left(\mathscr{V}_{j,11}(t) - \mathscr{V}_{j,12}(t) \right) \right|}{\mathscr{R}(t)} \leq 4\|K\|_{\mathrm{op}} (q(t))^2 + 2\|K\|_{\mathrm{op}}^{1/2} q(t). \tag{466}$$

Proof. For simplicity of notation, define $\mathcal{E}_j(t) = \mathscr{V}_{j,11}(t) - 2\mathscr{V}_{j,12}(t) + \mathscr{V}_{j,22}(t)$ then $2\mathscr{R}(t) = \sum_{j=1}^d \lambda_j \mathcal{E}_j(t) + \eta^2$. Since $\mathscr{V}_j(t)$ is psd then its determinant is nonnegative which implies $(\mathscr{V}_{j,12}(t))^2 \leq \mathscr{V}_{j,11}(t)\mathscr{V}_{j,22}(t)$. Using this inequality, it follows that

$$\mathcal{E}_j(t) \geq \left(\sqrt{\mathscr{V}_{j,11}(t)} - \sqrt{\mathscr{V}_{j,22}(t)}\right)^2. \tag{467}$$

We also obtain the following bounds using Cauchy–Schwarz

$$\begin{aligned} 2\mathscr{R}(t) &\geq \left(\sqrt{\sum_{j=1}^d \lambda_j \mathscr{V}_{j,11}(t)} - \sqrt{\sum_{j=1}^d \lambda_j \mathscr{V}_{j,22}(t)}\right)^2, \\ \text{and} \quad \left|\sum_{j=1}^d \lambda_j \mathscr{V}_{j,12}(t)\right| &\leq \sqrt{\sum_{j=1}^d \lambda_j \mathscr{V}_{j,11}(t)}\sqrt{\sum_{j=1}^d \lambda_j \mathscr{V}_{j,22}(t)}. \end{aligned} \tag{468}$$

We start by deriving the first bound. We may rewrite the difference $\mathscr{V}_{j,11}(t) - \mathscr{V}_{j,12}(t)$ as follows

$$\mathscr{V}_{j,11}(t) - \mathscr{V}_{j,12}(t) = \frac{1}{2}\mathcal{E}_j(t) + \frac{1}{2}\left(\mathscr{V}_{j,11}(t) - \mathscr{V}_{j,22}(t)\right). \tag{469}$$

Summing over λ_j, it follows

$$\sum_{j=1}^d \lambda_j \left(\mathscr{V}_{j,11}(t) - \mathscr{V}_{j,12}(t)\right) \leq \mathscr{R}(t) + \frac{1}{2}\sum_{j=1}^d \lambda_j \left(\mathscr{V}_{j,11}(t) - \mathscr{V}_{j,22}(t)\right). \tag{470}$$

Since $\mathscr{V}_{j,11}(t), \mathscr{V}_{j,22}(t) \geq 0$ and using (467) then

$$\begin{aligned} |\mathscr{V}_{j,11}(t) - \mathscr{V}_{j,22}(t)| &= \left|\sqrt{\mathscr{V}_{j,11}(t)} - \sqrt{\mathscr{V}_{j,22}(t)}\right| \left(\sqrt{\mathscr{V}_{j,11}(t)} + \sqrt{\mathscr{V}_{j,22}(t)}\right) \\ &\leq \sqrt{\mathcal{E}_j(t)} \left(\sqrt{\mathscr{V}_{j,11}(t)} + \sqrt{\mathscr{V}_{j,22}(t)}\right). \end{aligned}$$

Summing over λ_j and from Cauchy–Schwarz, it follows

$$\left|\sum_{j=1}^d \lambda_j \left(\mathscr{V}_{j,11}(t) - \mathscr{V}_{j,22}(t)\right)\right| \leq \sqrt{2\mathscr{R}(t)} \left(\sqrt{\sum_{j=1}^d \lambda_j \mathscr{V}_{j,11}(t)} + \sqrt{\sum_{j=1}^d \lambda_j \mathscr{V}_{j,22}(t)}\right). \tag{471}$$

Now, using the definition of $q(t)$ we obtain the simple bound

$$\sum_{j=1}^d \lambda_j \mathscr{V}_{j,11}(t) \leq \|K\|_{\mathrm{op}} \widehat{\mathscr{B}}_{44}(t) = 2\|K\|_{\mathrm{op}} \mathscr{R}(t)(q(t))^2. \tag{472}$$

Using this bound and (468), it follows that

$$\sqrt{\sum_{j=1}^d \lambda_j \mathscr{V}_{j,22}(t)} \leq \sqrt{\sum_{j=1}^d \lambda_j \mathscr{V}_{j,11}(t)} + \sqrt{2\mathscr{R}(t)} \leq \sqrt{2\mathscr{R}(t)} \left(1 + \|K\|_{\mathrm{op}}^{1/2} q(t)\right). \tag{473}$$

Thus, from these previous bounds, we obtain

$$\frac{1}{2}\left|\sum_{j=1}^d \lambda_j \left(\mathscr{V}_{j,11}(t) - \mathscr{V}_{j,22}(t)\right)\right| \leq \mathscr{R}(t)\left(2\|K\|_{\mathrm{op}}^{1/2} q(t) + 1\right). \tag{474}$$

Dividing by $\mathscr{R}(t)$ obtains the first bound. For the second bound, from (468), (472) and (468) we have

$$\begin{aligned} \frac{\left|\sum_{j=1}^d \lambda_j \left(\mathscr{V}_{j,11}(t) - \mathscr{V}_{j,12}(t)\right)\right|}{\mathscr{R}(t)} &\leq \frac{\left|\sum_{j=1}^d \lambda_j \mathscr{V}_{j,11}(t)\right| + \sqrt{\sum_{j=1}^d \lambda_j \mathscr{V}_{j,11}(t)}\sqrt{\sum_{j=1}^d \lambda_j \mathscr{V}_{j,22}(t)}}{\mathscr{R}(t)} \\ &\leq 4\|K\|_{\mathrm{op}}(q(t))^2 + 2\|K\|_{\mathrm{op}}^{1/2} q(t). \end{aligned}$$

□

I.2 Binary Logistic Regression

We consider a binary logistic regression problem with $\epsilon = 0$ where we are trying to classify two classes. Setting the hard labels $y \in \{-1, 1\}$, the standard logistic regression objective function (cross-entropy) can be rewritten as

$$f_I(X^\top a_I) = \log(1 + e^{-y_I \cdot X^\top a_I}). \tag{475}$$

Since the map $t \mapsto \log(1 + e^{-t})$ is convex and strictly decreasing, the inner-maximization problem in the adversarial risk simplifies to Ribeiro et al. (2023)

$$\begin{aligned}
\mathcal{R}^{\text{adv}}(X) &= \mathbb{E}_{a,I}\left[\max_{|s|\leq 1} \log(1 + \exp(-y_I \cdot (X^\top a_I + \delta s \|X\|)))\right] \\
&= \mathbb{E}_{a,I}\left[\log(1 + \exp(-y_I \cdot (X^\top a_I) + \delta\|X\|))\right] \\
&= \sum_{i=1}^{2} p_i \mathbb{E}_v\left[\log(1 + \exp((-1)^i \cdot (\sqrt{X^\top K_i X} v + X^\top \mu_i) + \delta\|X\|))\right]
\end{aligned} \tag{476}$$

where $v \sim \mathcal{N}(0, 1)$. To simplify the derivation, we take $\mu_1 = \mu$ and $\mu_2 = -\mu$. We note that this is equivalent to any other choices of means for binary logistic regression by recentering as follows $\tilde{a} = a - c$ with $c = \frac{\mu_1 + \mu_2}{2}$ such that

$$\mu_1 - c = \frac{\mu_1 - \mu_2}{2} \overset{\text{def}}{=} \mu \quad \text{and} \quad \mu_2 - c = -\frac{\mu_1 - \mu_2}{2} \overset{\text{def}}{=} -\mu.$$

We are studying the hard label setting, hence consider $X^\star = 0$. We identify the following terms

$$z_{11} = X^\top K_i X, \quad z_{13} = X^\top \mu_i, \quad z_{31} = z_{13} \quad \text{and} \quad z_{44} = \|X\|^2,$$

which leads to

$$h_i\left(\begin{bmatrix} z_{11} & 0 & z_{13} & 0 \\ 0 & 0 & 0 & 0 \\ z_{31} & 0 & 0 & 0 \\ 0 & 0 & 0 & z_{44} \end{bmatrix}\right) = \mathbb{E}_v\left[\log\left(1 + \exp\left((-1)^i\left(\sqrt{z_{11}} v + \frac{z_{13} + z_{31}}{2}\right) + \delta\sqrt{z_{44}}\right)\right)\right]$$

Here we include both z_{13} and z_{31} to stay consistent with the gradient derived in Lemma D.1. It follows from (476) that

$$\mathcal{R}^{\text{adv}}(X) = \sum_{i=1}^{2} p_i h_i(\widehat{B}_i(W)), \tag{477}$$

with

$$h_i(\widehat{B}_i(W)) \overset{\text{def}}{=} \mathbb{E}_v\left[\log\left(1 + \exp\left((-1)^i\left(\sqrt{B_{i,11}(W)} v + \widehat{B}_{i,13}(W)\right) + \delta\sqrt{\widehat{B}_{1,44}(W)}\right)\right)\right].$$

since $\mu_2 = -\mu_1$ and $\widehat{B}_{1,44}(W) = \widehat{B}_{2,44}(W)$. Hence, we obtain the following det. equivs. for $\mathcal{R}$ and $\mathcal{R}^{\text{adv}}$

$$\begin{aligned}
\mathscr{R}(t) &= \sum_{i=1}^{2} p_i \mathbb{E}_v\left[\log\left(1 + \exp\left((-1)^i \sqrt{\mathscr{B}_{i,11}(t)} v - \mathfrak{m}(t)\right)\right)\right], \\
\mathscr{R}^{\text{adv}}(t) &= \sum_{i=1}^{2} p_i \mathbb{E}_v\left[\log\left(1 + \exp\left((-1)^i \sqrt{\mathscr{B}_{i,11}(t)} v - \mathfrak{m}(t) + \delta\sqrt{\widehat{\mathscr{B}}_{1,44}(t)}\right)\right)\right].
\end{aligned} \tag{478}$$

Here $\mathfrak{m}_1 = -\mathfrak{m}_2$ so we drop the subscript in $\mathfrak{m}$ as we only need to track one of the means. Analogously to Collins-Woodfin & Seroussi (2025), we define the variables $w_{ij}(k)$ for $i = 1, 2$ and $j = 1, 2$ to denote the robust class probabilities with $r_i(k) = \sqrt{X_k^\top K_i X_k} v + (-1)^{i-1} X_k^\top \mu$ at iteration k of SGD

$$\begin{aligned}
w_{11}(k) &= \frac{e^{r_1(k) - \delta\|X\|}}{1 + e^{r_1(k) - \delta\|X\|}}, \quad w_{12}(k) = \frac{1}{1 + e^{r_1(k) - \delta\|X\|}}, \\
w_{21}(k) &= \frac{e^{r_2(k) - \delta\|X\|}}{1 + e^{r_2(k) - \delta\|X\|}} \quad \text{and} \quad w_{22}(k) = \frac{1}{1 + e^{r_2(k) - \delta\|X\|}}.
\end{aligned}$$

Hence, their det. equivs. are given by

$$\begin{aligned}\mathfrak{u}_{12}(t) &\stackrel{\text{def}}{=} \left(1+e^{\sqrt{\mathscr{B}_{1,11}(t)}v+\mathfrak{m}(t)-\delta\sqrt{\widehat{\mathscr{B}}_{1,44}(t)}}\right)^{-1}, \quad \mathfrak{u}_{11}(t) \stackrel{\text{def}}{=} 1-\mathfrak{u}_{12}(t)\\ \mathfrak{u}_{22}(t) &\stackrel{\text{def}}{=} \left(1+e^{\sqrt{\mathscr{B}_{2,11}(t)}v-\mathfrak{m}(t)-\delta\sqrt{\widehat{\mathscr{B}}_{1,44}(t)}}\right)^{-1}, \quad \mathfrak{u}_{21}(t) \stackrel{\text{def}}{=} 1-\mathfrak{u}_{22}(t)\end{aligned} \tag{479}$$

We now compute $\partial_{11}h_i$ (i.e. $H_{1,t}$), $\partial_{13}h_i$ (i.e. $H_{2,t}$) and $\partial_{44}h_i$. From the chain rule, we have

$$\begin{aligned}\partial_{11}h_i(\widehat{B}_i) &= \frac{(-1)^i}{2\sqrt{B_{i,11}}}\mathbb{E}_v\left[v\cdot\sigma\left((-1)^i\sqrt{B_{i,11}}v-\widehat{B}_{1,13}+\delta\sqrt{\widehat{B}_{1,44}}\right)\right],\\ \partial_{13}h_i(\widehat{B}_i) &= \frac{(-1)^i}{2}\mathbb{E}_v\left[\sigma\left((-1)^i\sqrt{B_{i,11}}v-\widehat{B}_{1,13}+\delta\sqrt{\widehat{B}_{1,44}}\right)\right],\\ \partial_{44}h_i(\widehat{B}_i) &= \frac{\delta}{2\sqrt{\widehat{B}_{1,44}}}\mathbb{E}_v\left[\sigma\left((-1)^i\sqrt{B_{i,11}}v-\widehat{B}_{1,13}+\delta\sqrt{\widehat{B}_{1,44}}\right)\right].\end{aligned} \tag{480}$$

where $\sigma(x) \stackrel{\text{def}}{=} \frac{1}{1+e^{-x}}$ denotes the sigmoid function and $\partial_{13}h_i(\widehat{B}_i) = \partial_{31}h_i(\widehat{B}_i)$. From Stein's Lemma and since $\sigma'(x) = \sigma(x)(1-\sigma(x))$, we obtain

$$\begin{aligned}\partial_{11}h_i(\widehat{B}_i) = \frac{1}{2}\mathbb{E}_v\Bigg[&\sigma\left((-1)^i\sqrt{B_{i,11}}v-\widehat{B}_{1,13}+\delta\sqrt{\widehat{B}_{1,44}}\right)\\ &\times\left(1-\sigma\left((-1)^i\sqrt{B_{i,11}}v-\widehat{B}_{1,13}+\delta\sqrt{\widehat{B}_{1,44}}\right)\right)\Bigg].\end{aligned}$$

Using the notation in (479), we obtain the relations

$$\begin{aligned}&\partial_{11}h_i(\widehat{\mathscr{B}}_i(t)) = \frac{1}{2}\mathbb{E}_v\left[\mathfrak{u}_{i1}(t)\mathfrak{u}_{i2}(t)\right], \quad \partial_{13}h_1(\widehat{\mathscr{B}}_1(t)) = -\frac{1}{2}\mathbb{E}_v\left[\mathfrak{u}_{12}(t)\right],\\ &\partial_{13}h_2(\widehat{\mathscr{B}}_2(t)) = \frac{1}{2}\mathbb{E}_v\left[\mathfrak{u}_{21}(t)\right], \quad \partial_{44}h_1(\widehat{\mathscr{B}}_1(t)) = \frac{\delta}{2\sqrt{\widehat{\mathscr{B}}_{1,44}(t)}}\mathbb{E}_v\left[\mathfrak{u}_{12}(t)\right]\\ &\qquad\text{and}\quad \partial_{44}h_2(\widehat{\mathscr{B}}_2(t)) = \frac{\delta}{2\sqrt{\widehat{\mathscr{B}}_{1,44}(t)}}\mathbb{E}_v\left[\mathfrak{u}_{21}(t)\right].\end{aligned} \tag{481}$$

As a function of the input $X^\top a$, we have

$$f_i'(x) = (-1)^i\sigma((-1)^i\cdot x). \tag{482}$$

Hence, we obtain

$$\mathcal{I}_i(\widehat{B}_i) = \mathbb{E}_v\left[\sigma^2\left((-1)^i\sqrt{X^\top K_i X}v - X^\top\mu + \delta\|X\|\right)\right], \tag{483}$$

and its det. equivs.

$$\mathcal{I}_1(\widehat{\mathscr{B}}_1(t)) = \mathbb{E}_v[(\mathfrak{u}_{12}(t))^2] \quad\text{and}\quad \mathcal{I}_2(\widehat{\mathscr{B}}_2(t)) = \mathbb{E}_v[(\mathfrak{u}_{21}(t))^2]. \tag{484}$$

Using this notation, the system of ODEs (26) is given by

$$\begin{aligned}\frac{\mathrm{d}\mathscr{V}_{j,11}(t)}{\mathrm{d}t} &= -2\gamma(t)\Bigg(\sum_{i=1}^2 p_i\lambda_j^{(i)}\mathbb{E}_v\left[\mathfrak{u}_{i1}(t)\mathfrak{u}_{i2}(t)\right]\\ &\quad+\frac{\delta}{\sqrt{\widehat{\mathscr{B}}_{1,44}(t)}}\left(p_1\mathbb{E}_v\left[\mathfrak{u}_{12}(t)\right]+p_2\mathbb{E}_v\left[\mathfrak{u}_{21}(t)\right]\right)\Bigg)\mathscr{V}_{j,11}(t)\\ &\quad+2\gamma(t)\left(p_1\mathbb{E}_v\left[\mathfrak{u}_{12}(t)\right]+p_2\mathbb{E}_v\left[\mathfrak{u}_{21}(t)\right]\right)\mathfrak{m}_{j,1,1}(t)\\ &\quad+\frac{\gamma(t)^2}{d}\sum_{i=1}^2 p_i\,\mathbb{E}_v[(\mathfrak{u}_{i(\sim i)}(t))^2](\lambda_j^{(i)}+(\omega_j^\top\mu)^2),\\ \frac{\mathrm{d}\mathfrak{m}_{j,1,1}}{\mathrm{d}t} &= -\gamma(t)\Bigg(\sum_{i=1}^2 p_i\lambda_j^{(i)}\mathbb{E}_v\left[\mathfrak{u}_{i1}(t)\mathfrak{u}_{i2}(t)\right]\\ &\quad+\frac{\delta}{\sqrt{\widehat{\mathscr{B}}_{1,44}(t)}}\left(p_1\mathbb{E}_v\left[\mathfrak{u}_{12}(t)\right]+p_2\mathbb{E}_v\left[\mathfrak{u}_{21}(t)\right]\right)\Bigg)\mathfrak{m}_{j,1,1}(t)\\ &\quad+\gamma(t)\left(p_1\mathbb{E}_v\left[\mathfrak{u}_{12}(t)\right]+p_2\,\mathbb{E}\left[\mathfrak{u}_{21}(t)\right]\right)(\mu^\top\omega_j)^2,\end{aligned} \tag{485}$$

with initial condition $\mathscr{V}_{j,11}(0) = X_0^\top \omega_j \omega_j^\top X_0$, $\mathfrak{m}_{j,1,1}(0) = X_0^\top \omega_j \omega_j^\top \mu$ and where $\sim i$ denotes not i such that $\mathfrak{u}_{i(\sim i)} = \mathfrak{u}_{12}$ for $i = 1$ and $\mathfrak{u}_{i(\sim i)} = \mathfrak{u}_{21}$ for $i = 2$. Note that if $K_1 = K_2$ then we have $\mathfrak{u}_{12}(t) \overset{d}{=} \mathfrak{u}_{21}(t)$ and $\mathfrak{u}_{11}(t) \overset{d}{=} \mathfrak{u}_{22}(t)$.

J Numerical Simulation Details

For the numerical simulations, power law covariance refers to the following setup. For a covariance matrix K, let (λ_j, ω_j) denote the eigenvalue-eigenvector pairs. Suppose $\alpha \geq 0$, $\beta \geq 0$. Define $x_{0,j} \overset{\text{def}}{=} \langle X_0, \omega_j \rangle$ and $x_j^\star \overset{\text{def}}{=} \langle X^\star, \omega_j \rangle$ such that

$$\lambda_j = \left(\frac{j}{d}\right)^{\alpha} \quad \text{and} \quad (x_j^\star)^2 = \frac{1}{d}\lambda_j^{-\beta}. \tag{486}$$

Remark J.1. *This power law parametrization is inspired by the parametrizations presented in Collins-Woodfin et al. (2024b); Collins-Woodfin & Seroussi (2025). The parameter β_{HL} of Collins-Woodfin et al. (2024b) satisfies the following relation with our parameter α:*

$$\alpha = \frac{1}{1 - \beta_{\text{HL}}}.$$

Hence for large d, the eigenvalues $\{\lambda_j\}_{j=1}^d$ follow a power law distribution with density function $f(\lambda) = \frac{1}{\alpha}\lambda^{\frac{1}{\alpha}-1}$ and support $(0, 1]$. $\alpha = 0$ corresponds to isotropic covariance and $\alpha > 1$ corresponds to power law setup of Collins-Woodfin et al. (2024b) with $0 < \beta_{\text{HL}} < 1$. See Remark 4 in Collins-Woodfin & Seroussi (2025) for more details on the role of α or Section 6 in Wei et al. (2022) for a discussion on power laws.

Figure 1: Left, middle: concentration of the ℓ_2-adversarial risk for noiseless ℓ_2-adversarial least squares with a single class $a \sim \mathcal{N}(0, K)$ (left) and noiseless binary logistic regression with hard labels on a binary Gaussian mixture with distinct means and common covariance (middle). For logistic regression, see Appendix I.2. For least squares, 30 runs of SGD with constant learning rate $\gamma = 0.3$ and $\delta = 0.3$ under power law covariance (see (486)) with $\alpha = 1.25$, $\beta = 0.75$, $\mu = 0$, $p_1 = 1$ and $X_0 \sim \mathcal{N}(0, I_d/d)$. For logistic regression, 30 runs of SGD with constant learning rate $\gamma = 0.3$ and $\delta = 0.1$ under power law covariance (see (486)) with $\alpha = 1.25$ and $K_1 = K_2$. In the hard label setup, we impose a power law assumption on the mean (Collins-Woodfin & Seroussi, 2025) instead of the ground truth such that $(\langle \mu, \omega_j \rangle)^2 = \frac{1}{d}\left(\frac{j}{d}\right)^{\beta}$ with $\beta = 0.75$. We also set $\mu_1 = -\mu_2$, $p_1 = p_2 = 0.5$ and $X_0 \sim \mathcal{N}(0, I_d/d)$. The shaded regions represent the max and min values at each step across the 30 SGD runs. As dimension d increases, in both plots the adversarial risk concentrates around the deterministic limit (red) described by the system of ODEs (26) as predicted by Theorem 2.1. **Right: convergence of** $q(t)$ to the limit predicted by the fixed-point equation of Theorem 3.2 under several power-law setups. This panel is the right plot of Figure 6; see its description below for the simulation details.

Figure 2: Comparison for Exact Line Search and Polyak Stepsize under strong anisotropy on noiseless ℓ_2-adversarial least squares for different values of δ in the three regimes of $X^{\star,\text{Adv}}$ (Theorem 3.3). The three plots illustrate the convergence of the ℓ_2-adversarial risk and how δ and $\tilde{\lambda}^{\text{eff}}(t)$ mitigate the impact of strong anisotropy on the discrepancy between the Polyak stepsize and exact line search. Across all the plots, we fix $X_0 \sim \mathcal{N}(0, \frac{1}{d}I_d)$, $X^\star = \mathbf{1}_d/\sqrt{d}$ where $\mathbf{1}$ is a vector of ones of size d, $d = 800$, $\eta = 0$, $\mu = 0$, $p_1 = 1$ and K satisfying the covariance setup presented in (356) with $s = 2.8$ which produces $\frac{1}{d}\operatorname{Tr}(K^2) \approx 5.25$. See (354) for more details.

Figure 3: Comparison of ℓ_2-adversarial least squares and ℓ_2-regularized least squares with adaptive learning rate and regularization. The left plot compares the paths of the det. equivs. of $\mathcal{R}^{\text{Adv}}$ for AdvHSGD and HSGD and confirms Theorem 3.1. The right plot compares the path of $\mathcal{R}^{\text{Adv}}(X_k)$ for SGD with adaptive learning rate $\gamma^{\text{Reg}}(t)$ and regularization $\lambda^{\text{Reg}}(t)$ versus the det. equiv. computed from (301) and (302) and shows close agreement between them for a variety of constant $\gamma(t) \equiv \gamma$. The ℓ_2-adversarial risk is computed using $\mathcal{R}^{\text{Adv}}(X_k) = \mathcal{R}(X_k) + \delta\sqrt{\frac{2}{\pi}}\sqrt{2\mathcal{R}(X_k)}\|X_k\| + \frac{\delta^2}{2}\|X_k\|^2$. For the left plot, we fixed the learning rate $\gamma = 1$, the

noise variance $\eta^2 = 0$, $K = I_d$, $\mu = 0$, $p_1 = 1$, $X_0 \sim \mathcal{N}(0, 4I_d/d)$ and $X^\star \sim \mathcal{N}(0, I_d/d)$ across all runs and only vary δ. For the right plot, we fix $\delta = 0.3$, $\eta^2 = 0$, power law covariance (See (486)) with $\alpha = 1.25$, $\beta = 0.75$, $\mu = 0$, $p_1 = 1$ and $X_0 \sim \mathcal{N}(0, 4I_d/d)$. All experiments are done for $d = 800$.

Figure 4: SGD with exact line search γ_k^{line} or Polyak stepsize $\gamma_k^{\text{Polyak,adv}}$ matches closely the path of our system of ODEs (26) with deterministic learning rates schedules $\gamma(t) = \gamma^{\text{line}}(t)$ and $\gamma(t) = \gamma^{\text{Polyak,adv}}(t)$ for $\mathcal{R}^{\text{Adv}}(X_k)$ on noiseless ℓ_2-adversarial least squares. For the plot with Polyak stepsize (left), we set $d = 800$, $\eta = 0$, $X_0 \sim \mathcal{N}(0, 4I_d/d)$, $X^\star \sim \mathcal{N}(0, I_d/d)$, $\mu = 0$, $p_1 = 1$, $K_1 = K_2 = I_d$. For the plot with exact line search, we set $d = 800$, $\eta = 0$, $X_0 \sim \mathcal{N}(0, 4I_d/d)$, $\mu = 0$, $p_1 = 1$, K and $X^\star$ satisfying a power law setup (see (486)) with $\alpha = 1.25$ and $\beta = 0.75$.

Figure 5: Comparison between Exact Line Search and Polyak Stepsize under weak anisotropy on noiseless ℓ_2-adversarial least squares for different values of δ in the three regimes of $X^{\star,\text{Adv}}$ (See Theorem 3.3). The three plots illustrate the convergence of the ℓ_2-adversarial risk and that, under weak anisotropy, exact line search and the Polyak stepsize perform similarly. See Figure 2 for results in the case of strong anisotropy. Across all the plots, we fix $X_0 \sim \mathcal{N}(0, \frac{1}{d} I_d)$, $X^\star = \mathbf{1}_d/\sqrt{d}$ where $\mathbf{1}$ is a vector of ones of size d, $d = 800$, $\eta = 0$, $\mu = 0$, $p_1 = 1$ and K satisfying the covariance setup presented in (356) with $s = 7.5$ which produces $\frac{1}{d}\operatorname{Tr}(K^2) \approx 1.12$. See (354) for more details.

Figure 6: Numerical evidence that $q(t) \stackrel{\text{def}}{=} \sqrt{\frac{\widehat{\mathscr{B}}_{44}(t)}{2\mathscr{R}(t)}}$ converges for noiseless ℓ_2-adversarial least squares. The first and second plots provide evidence that $q(t)$ converges for a variety of constant learning rates γ (left) and δ (middle). In either plot, we fix one of the parameters and vary the other. Here we fix the parameters $d = 800$, $\eta = 0$, $X_0 \sim \mathcal{N}(0, 4I_d/d)$, K and $X^\star$ satisfy a power law relationship (see (486)) with $\alpha = 1.25$ and $\beta = 0.75$. We set $\delta = 0.3$ when we vary γ and fix $\gamma = 0.3$ when we vary δ. The third plot (right) provides evidence that $q(t)$ converges for a variety of $X^\star$ and covariances K under different power law setups (see (486)) and validates that the fixed point equation in Theorem G.1 predicts the limiting value q_∞. For this plot, we use the same setup with $\delta = \gamma = 0.3$ fixed and vary α and β according to the values given in the legend.