
Local Convergence of Gradient Methods for Min-Max Games: Partial Curvature Generically Suffices

Guillaume Wang

Institute of Mathematics
École polytechnique fédérale de Lausanne
Station Z, CH-1015 Lausanne
guillaume.wang@epfl.ch

Lénaïc Chizat

Institute of Mathematics
École polytechnique fédérale de Lausanne
Station Z, CH-1015 Lausanne
lenaic.chizat@epfl.ch

Abstract

We study the convergence to local Nash equilibria of gradient methods for two-player zero-sum differentiable games. It is well-known that such dynamics converge locally when $S \succ 0$ and may diverge when $S = 0$, where $S \succeq 0$ is the symmetric part of the Jacobian at equilibrium that accounts for the “potential” component of the game. We show that these dynamics also converge as soon as S is nonzero (*partial curvature*) and the eigenvectors of the antisymmetric part A are in general position with respect to the kernel of S . We then study the convergence rates when $S \ll A$ and prove that they typically depend on the *average* of the eigenvalues of S , instead of the minimum as an analogy with minimization problems would suggest. To illustrate our results, we consider the problem of computing mixed Nash equilibria of continuous games. We show that, thanks to partial curvature, conic particle methods – which optimize over both weights and supports of the mixed strategies – generically converge faster than fixed-support methods. For min-max games, it is thus beneficial to add degrees of freedom “with curvature”: this can be interpreted as yet another benefit of over-parameterization.

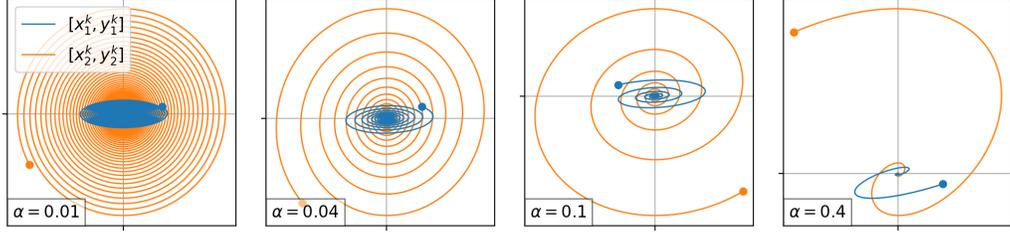
1 Introduction

Min-max optimization is notoriously subtler than minimization, even in convex-concave settings. While many of the proof techniques for minimization have a natural equivalent in the min-max world, some common intuitions fail to transfer. The picture is clear for strongly convex-strongly concave (SC-SC) min-max games: all the classical gradient methods converge exponentially (for small enough step-sizes) with worst-case convergence rates dependent on the strong convexity and strong concavity parameters μ_x, μ_y . But, most famously perhaps, for bilinear min-max games the last iterate of simultaneous Gradient Descent-Ascent (GDA) diverges, while the Proximal Point (PP) method converges, and alternating GDA and the continuous-time limit of all of these algorithms – Gradient Flow (GF) – exhibit a cycling behavior.

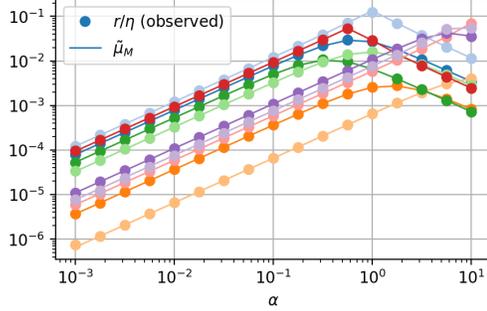
Based on those two extreme cases, it still seems that part of our intuition from minimization, where the last-iterate convergence rate is indeed determined by the strong convexity parameter, is preserved. In this paper, we argue that this intuition is in fact overly pessimistic, and that gradient methods behave in general more favorably in the min-max setting than in the minimization setting.

Let us first look at a basic example. Consider a bilinear min-max game regularized by a quadratic term *only in one scalar variable*:

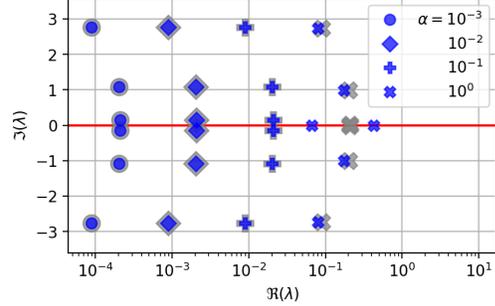
$$\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^d} x^\top P y + \frac{\alpha}{2} x_1^2.$$



(a) Local convergence of the GF iterates for a fixed draw of $P \in \mathbb{R}^{2 \times 2}$ and various values of α . Only the final phase of the dynamics is shown, so here we see the iterates evolve along the “dominant” eigenspace of M (the subspace along which convergence is the slowest).



(b) Observed and predicted ($\tilde{\mu}_M$) normalized convergence rate r/η of GDA with a small step-size η , i.e., $\|x^k\| + \|y^k\| = \Theta((1-r)^k)$, vs. regularization strength α (the higher the faster). Each color represents one draw of $P \in \mathbb{R}^{2 \times 2}$.



(c) Spectrum of the Jacobian M for a fixed draw of $P \in \mathbb{R}^{3 \times 3}$ (green) and its approximation by Equation (3.1) (gray)

Figure 1: Convergence of gradient methods on a random bilinear game regularized by $\frac{\alpha}{2}x_1^2$ for small step-sizes. The fact that $\tilde{\mu}_M$, and so r/η , scale linearly with α (for small α) is explained by Prop. 3.1.

Here there is no strong convexity in either player (so $\mu_x = \mu_y = 0$), yet as observed on Fig. 1a, Fig. 1b, when $P \in \mathbb{R}^{d \times d}$ is random with independent standard normal entries, GDA (with a small step-size) typically converges at an exponential rate that scales linearly with α .¹

As we will see, this phenomenon is a consequence of the existing theory for general smooth min-max games

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} f(x, y) \quad (1.1)$$

with a local Nash equilibrium (NE), or local saddle point, $z^* = (x^*, y^*)$. It is well-known from the dynamical systems literature that GF converges locally to z^* if

$$\tilde{\mu}_M := \min_{\lambda \in \text{Sp}(M)} \Re(\lambda) > 0, \quad \text{with } M = \begin{bmatrix} \nabla_{xx}^2 f & \nabla_{xy}^2 f \\ -\nabla_{xy}^2 f^\top & -\nabla_{yy}^2 f \end{bmatrix} (z^*) \in \mathbb{R}^{(n+m) \times (n+m)}$$

the Jacobian of the skewed gradient field at z^* , and where $\text{Sp}(\cdot)$ denotes the spectrum, i.e., the set of eigenvalues. For our starting basic example, this quantity² can be visualized in Fig. 1c. Moreover, $\tilde{\mu}_M$ also characterizes the convergence behavior of gradient methods (GDA, PP, etc.) in the leading order in the step-size η : they all converge to z^* with the rate $\eta \tilde{\mu}_M + O(\eta^2)$, see App. A for a review.³

While the condition $\tilde{\mu}_M > 0$ is very general and tight, it is not obvious how to control or interpret it. Our purpose with this paper is to explore what this condition entails and to emphasize some of its surprising consequences.

¹Julia code for the numerical experiments is available at https://github.com/guillaumew16/local_cvgce_minmax.

²In numerical analysis and stability theory, the quantity $-\tilde{\mu}_M$ is known as the *spectral abscissa* of $-M$.

³More precisely, gradient methods converge locally (for small enough step-sizes) if $\min_{\lambda \in \text{Sp}(M)} \Re(\lambda) > 0$, and only if $\min_{\lambda \in \text{Sp}(M) \setminus \{0\}} \Re(\lambda) \geq 0$. Throughout this paper we will assume for simplicity M invertible so that those two quantities are equal.

The rest of the paper is organized as follows. In [Sec. 2](#) we show that, generically, $\tilde{\mu}_M > 0$ as long as the problem has *partial curvature*, i.e. the diagonal blocks of M are non-zero. In [Sec. 3](#), we study more precisely the case of games where the “interaction” component dominates the “potential” component, i.e., when $\nabla_{xx}^2 f(z^*), \nabla_{yy}^2 f(z^*) \ll \nabla_{xy}^2 f(z^*)$. For such games, in a certain random setting, we show that $\tilde{\mu}_M$ scales as the *average* of the eigenvalues of the potential part, instead of the minimum as an analogy with minimization problems would suggest. In [Sec. 4](#) we consider the computation of mixed Nash equilibria of continuous games using particle methods, a setting where the convergence under partial curvature has a striking consequence. By optimizing over both the weights and supports of the mixed strategies, one obtains dynamics that generically converge faster than fixed-support methods, even when the latter are using the optimal supports.

For ease of exposition, in most of the paper we focus on the local convergence of GF. We discuss the convergence of discrete time algorithms in [Sec. 3.2](#).

1.1 Related work

Throughout this paper, by “convergence” we mean convergence of the last iterate to a (local) NE, while a different line of work also considers convergence of the averaged iterate [[Nem04](#)], and another considers convergence to any critical point [[ALW21](#)]. Also to prevent possible confusion, let us mention a related but distinct line of work [[DH19](#); [Doa22](#)] that considers using different step-sizes to update the x and the y variables. The resulting two-timescale dynamics can be analyzed globally assuming a min-max analog of the Polyak-Łojasiewicz inequality, using quite different considerations than the ones developed in this paper.

Analysis of GDA and PP for SC-SC, bilinear or convex-SC games. The fact that PP converges for SC-SC and bilinear min-max games is well-established since at least [[Roc76](#)]; for a modern reformulation, see e.g. [[MOP20](#)]. The convergence of simultaneous GDA for SC-SC games and its divergence for bilinear games is also a classical fact, see e.g. [[LS19](#); [Lu22](#)]. The cycling behavior of alternating GDA for bilinear games is proved in [[BGP20](#), Theorem 4].

As shown more recently in [[NK17](#), Appendix G] and [[ZWL22](#), Theorem 6], convexity in x , strong concavity in y and non-degeneracy of the interaction component are in fact sufficient to ensure $\tilde{\mu}_M > 0$, and so local convergence of GDA. The setting of the former work corresponds to $\nabla_{xx}^2 f(z^*) = 0$, $-\nabla_{yy}^2 f(z^*) \succ 0$ and $\nabla_{xy}^2 f(z^*)$ full-row-rank, extended in the latter work to $\nabla_{xx}^2 f(z^*) \succeq 0$. Those two works also provide bounds on $\tilde{\mu}_M$ for those specific cases in terms of the least eigenvalues of $-\nabla_{yy}^2 f(z^*)$ and of $\nabla_{xy}^2 f(z^*) \nabla_{xy}^2 f(z^*)^\top$.

Spectral analysis-based convergence analyses. Our approach to analyzing min-max algorithms is to directly study the properties of the update operator (e.g., of $T(z) = z - \begin{pmatrix} \nabla_x f \\ -\nabla_y f \end{pmatrix}(z)$ for simultaneous GDA). Specifically, by a classical result on discrete-time dynamical systems, the local exponential convergence is characterized by the spectral radius of that operator’s Jacobian. This approach is used for example in [[Gid+19](#)] to analyze the local convergence behavior of alternating GDA with negative momentum, and in [[AML20](#)] to derive tight convergence bounds – dependent explicitly on η and $\text{Sp}(M)$ – for simultaneous GDA and multi-step Extra-Gradient.

Average-case analysis of the (local) convergence rate. In [[PS20](#); [DPS21](#)], the authors analyze the convergence of gradient methods for affine operator problems – i.e., for finding z such that $M(z - z^*) = 0$ – when M is a random normal matrix with a known spectral distribution. They derive average-case optimal methods for this setting. Our analysis in [Sec. 3](#) also has an average-case flavor, but our random model is different: we assume that the symmetric S and antisymmetric parts A of M are independent and that $S \ll A$. In particular we do not require M to be normal (i.e., S and A do not commute a priori, which is the typical case in min-max optimization).

Hypo-coercivity. In the context of partial differential equations (PDEs), the phenomenon that a linear PDE $\partial_t u_t = -Lu_t$ may exhibit linear convergence to 0 even when the generator L is not coercive, is called hypo-coercivity and is studied in detail in [[Vil09](#)]. This is precisely the infinite-dimensional analog of the phenomenon studied in the present paper: still denoting by S, A the symmetric resp. antisymmetric parts of M , coercivity of L corresponds to $S \succ 0$, while hypo-coercivity corresponds to

$\tilde{\mu}_M > 0$. Specifically, [Vil09] shows how to construct Lyapunov functions to establish hypocoercivity and convergence rates of certain PDEs, by exploiting properties of the iterated commutators of S and A . By contrast our focus is on the finite-dimensional case, where it is easier and more natural to directly study the spectrum of M .

1.2 Notation

For any matrix $T \in \mathbb{R}^{d \times d}$ or $\mathbb{C}^{d \times d}$, denote by $\text{Sp}(T) \subset \mathbb{C}$ its spectrum, i.e., the set of its eigenvalues, and by $\rho(T) = \max_{\lambda \in \text{Sp}(T)} |\lambda|$ its spectral radius. Recall that the spectral radius is distinct from the operator norm, which is the largest singular value, although they coincide for Hermitian matrices. Denote eigenspaces as $E_\lambda(T) = \{z \in \mathbb{C}^d; Tz = \lambda z\}$ and let $\text{Eigvecs}(T) = \bigcup_{\lambda \in \text{Sp}(T)} E_\lambda(T)$ the set of all (complex) eigenvectors. $\|\cdot\|$ denotes Euclidean or Hermitian norm. For a collection of square matrices (resp. scalars) $(C_k)_k$, $\text{Diag}((C_k))$ denotes the (block-)diagonal matrix with blocks (resp. coefficients) $(C_k)_k$.

2 Characterization of local convergence

Consider the general smooth min-max game of Equation (1.1) and assume M is invertible. We decompose M into its symmetric and antisymmetric parts as

$$M = \begin{bmatrix} \nabla_{xx}^2 f & \nabla_{xy}^2 f \\ -\nabla_{xy}^2 f^\top & -\nabla_{yy}^2 f \end{bmatrix} (z^*) =: \begin{bmatrix} Q & P \\ -P^\top & R \end{bmatrix}, \quad S := \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix}, \quad A := \begin{bmatrix} 0 & P \\ -P^\top & 0 \end{bmatrix}.$$

By the second-order optimality condition in the definition of NE, $Q \succeq 0$ and $R \succeq 0$.

Following [Let+19], S and A can be thought of intuitively as the ‘‘potential’’ resp. ‘‘interaction’’ (or Hamiltonian) components of the two-player zero-sum game (1.1). Indeed, consider the quadratic game

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} \left\{ \frac{1}{2} \begin{pmatrix} x \\ y \end{pmatrix}^\top \nabla^2 f(z^*) \begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{2} x^\top Q x - \frac{1}{2} y^\top R y + x^\top P y \right\},$$

which is essentially sufficient to understand the local behavior of GF for (1.1) around z^* . Then we can interpret the terms in Q and R as quadratic potentials to be optimized independently by each player, with the bilinear term in P capturing all of the interaction between the players.

Partial curvature generically suffices. Let us recall that $\tilde{\mu}_M = \min_{\lambda \in \text{Sp}(M)} \Re(\lambda)$ governs the local convergence of GF around the local NE z^* . It is a general fact that this quantity is nonnegative (the proof is included below). In the following theorem, we give necessary and sufficient conditions for it to be positive, in terms of geometric conditions on Q , R and P .

Theorem 2.1. *The following conditions are equivalent:*

- (i) $\tilde{\mu}_M > 0$.
- (ii) $\text{Eigvecs}(A) \cap \text{Ker } S = \{0\}$.
- (iii) For any eigenvector x of PP^\top (i.e., left-singular vector of P), $x \notin \text{Ker } Q$ or $P^\top x \notin \text{Ker } R$.
- (iv) For any eigenvector y of $P^\top P$ (i.e., right-singular vector of P), $Py \notin \text{Ker } Q$ or $y \notin \text{Ker } R$.

As a consequence of **Thm. 2.1**, the condition $\tilde{\mu}_M > 0$ holds generically in the following sense: for any fixed $S \neq 0$, the set of matrices P such that $\tilde{\mu}_M > 0$ is dense and open in $\mathbb{R}^{n \times m}$. In particular, this property holds with probability 1 if P is drawn from an absolutely continuous distribution and is independent from Q and R , as in the experiment of **Fig. 1**.

Proof. Let $\lambda \in \text{Sp}(M)$ and $z \in \mathbb{C}^{n+m}$ non-zero such that $Mz = (S + A)z = \lambda z$. Since S and A are real and symmetric resp. antisymmetric,

$$\begin{aligned} \overline{\bar{z}^\top S z} &= z^\top S \bar{z} = (z^\top S \bar{z})^\top = \bar{z}^\top S z, \text{ so } \bar{z}^\top S z \in \mathbb{R} \\ \text{and } \overline{\bar{z}^\top A z} &= z^\top A \bar{z} = (z^\top A \bar{z})^\top = -\bar{z}^\top A z, \text{ so } \bar{z}^\top A z \in i\mathbb{R}. \end{aligned}$$

So by taking the real part in $\bar{z}^\top (S + A)z = \lambda \|z\|^2$,

$$\Re(\lambda) \|z\|^2 = \bar{z}^\top S z = \Re(z)^\top S \Re(z) + \Im(z)^\top S \Im(z) \geq 0$$

since $S = \text{Diag}(Q, R) \succeq 0$. This shows that $\tilde{\mu}_M \geq 0$.

Now let us show the equivalence of the conditions.

(ii) \implies (i): By contraposition, suppose there exists $\lambda \in \text{Sp}(M)$ with $\Re(\lambda) = 0$, and let z non-zero such that $Mz = \lambda z$. Then $\bar{z}^\top Sz = \Re(\lambda) \|z\|^2 = 0$, so $z \in \text{Ker } S$. So $Mz = Az = \lambda z$, and $z \in (E_\lambda(A) \cap \text{Ker } S) \setminus \{0\}$.

(i) \implies (ii): By contraposition, suppose there exists $\lambda \in \text{Sp}(A)$ and a non-zero $z \in E_\lambda(A) \cap \text{Ker } S$. Since A is antisymmetric, then $\Re(\lambda) = 0$. On the other hand, $Mz = (S + A)z = Az = \lambda z$, i.e., $\lambda \in \text{Sp}(M)$. So $\tilde{\mu}_M \leq \Re(\lambda) = 0$.

(i) \implies (iii), (iv): By contraposition, suppose there exists an eigenvector x of PP^\top such that $x \in \text{Ker } Q$ and $P^\top x \in \text{Ker } R$, and denote $\sigma \in \mathbb{R}$ such that $PP^\top x = \sigma^2 x$ (since $PP^\top \succeq 0$). Then

$$M \begin{pmatrix} i\sigma x \\ P^\top x \end{pmatrix} = \begin{bmatrix} Q & P \\ -P^\top & R \end{bmatrix} \begin{pmatrix} i\sigma x \\ P^\top x \end{pmatrix} = \begin{pmatrix} \sigma^2 x \\ -i\sigma P^\top x \end{pmatrix} = -i\sigma \begin{pmatrix} i\sigma x \\ P^\top x \end{pmatrix}$$

and so $-i\sigma \in \text{Sp}(M)$. This shows (i) \implies (iii), and (i) \implies (iv) follows analogously.

(iii), (iv) \implies (ii): By contraposition, suppose there exists $\lambda = i\sigma \in \text{Sp}(A)$ and a non-zero $z = (x, y) \in E_\lambda(A) \cap \text{Ker } S$. Expanding the blocks in $Az = \lambda z$,

$$\begin{cases} Py = i\sigma x \\ -P^\top x = i\sigma y \end{cases} \quad \text{and so} \quad \begin{cases} P^\top Py = \sigma^2 y \\ PP^\top x = \sigma^2 x. \end{cases}$$

Moreover, this implies that $x = 0 \iff y = 0$, and since $z \neq 0$, then both $x \neq 0$ and $y \neq 0$. So x is an eigenvector of PP^\top and, since $Sz = 0$, $x \in \text{Ker } Q$ and $P^\top x = -i\sigma y \in \text{Ker } R$, which contradicts (iii). Likewise, y is an eigenvector of $P^\top P$ and $y \in \text{Ker } R$ and $Py = i\sigma x \in \text{Ker } Q$, which contradicts (iv). \square

Geometric interpretation using real vectors. We draw the attention of the reader to the fact that (ii) involves complex eigenvectors. For a rephrasing in terms of real objects, note that if $z \in E_{i\sigma}(A)$ with $\sigma \in \mathbb{R}$, then $A [\Re(z) \quad \Im(z)] = [\Re(z) \quad \Im(z)] \begin{bmatrix} 0 & \sigma \\ -\sigma & 0 \end{bmatrix}$; geometrically, $F_{i\sigma}(A) = \text{span}(\Re(z), \Im(z))$ is a ‘‘rotation plane’’ of the GF for the bilinear game $\min_x \max_y x^\top Py$, in the sense that the projection of GF on $F_{i\sigma}(A)$ is a circular motion with constant speed σ . Condition (ii) expresses that for each such $F_{i\sigma}(A)$, there exists an eigenspace $E_\mu(S)$ of S (for a $\mu > 0$) that is not orthogonal to it. This causes the GF for $\min_x \max_y \frac{1}{2}x^\top Qx - \frac{1}{2}y^\top Ry + x^\top Py$ projected on $F_{i\sigma}(A)$ to spiral down to 0 instead of cycling around it.

One may naturally wonder whether a notion of non-orthogonality between the potential (S) and interaction components (A) can be used to bound the convergence quantitatively; this is developed in the next section in the particular case $S \ll A$.

3 Convergence rate when interaction dominates ($S \ll A$)

Let us now discuss the case of games with a small symmetric part, that is, whose Jacobian at optimum is $M_\alpha = A + \alpha S$ for some symmetric $S = \begin{bmatrix} Q & 0 \\ 0^\top & R \end{bmatrix}$, antisymmetric $A = \begin{bmatrix} 0 & P \\ -P^\top & 0 \end{bmatrix}$ and some small $\alpha > 0$. In this section we assume $n = m$ (the general case is technically more challenging as Prop. 3.1 requires A to have distinct eigenvalues, which requires $|n - m| \leq 1$).

3.1 Convergence rate of Gradient Flow

As discussed previously, the normalized local exponential convergence rate r/η of gradient methods in the asymptotic regime $\eta \rightarrow 0$ – or equivalently, the convergence rate of GF – for a game with Jacobian at optimum M_α is equal to $\tilde{\mu}_{M_\alpha}$. We can estimate this quantity using the standard formula for the asymptotic expansion of the eigenvalues of a perturbed matrix, which takes an interesting form in our context.

Proposition 3.1. *Suppose that P is full-rank and has distinct singular values, and let $P = U\Sigma V^\top = \sum_{j=1}^n \sigma_j u_j v_j^\top$ be its singular value decomposition. Then it holds*

$$\tilde{\mu}_{M_\alpha} = \frac{1}{2}\alpha \left(\min_{1 \leq j \leq n} u_j^\top Q u_j + v_j^\top R v_j \right) + O(\alpha^3).$$

This expansion explains in particular why the normalized convergence rate $r/\eta \sim \tilde{\mu}_{M_\alpha}$ is approximately proportional to α in Fig. 1. Interestingly, the error term is $O(\alpha^3)$, which suggests that the approximation can be reasonably accurate even for quite large values of α , as illustrated in Fig. 1c.

Proof. Here $M_0 = A$ has distinct eigenvalues $\{is\sigma_j, s \in \{-1, 1\}, 1 \leq j \leq n\}$, with unit-norm eigenvectors $A \begin{pmatrix} -isu_j/\sqrt{2} \\ v_j/\sqrt{2} \end{pmatrix} = is\sigma_j \begin{pmatrix} -isu_j/\sqrt{2} \\ v_j/\sqrt{2} \end{pmatrix}$. By the calculation of the eigenvalue derivatives from [Tao08], we obtain the following expansion for $\text{Sp}(M_\alpha)$:

$$\text{Sp}(M_\alpha) = \left\{ is\sigma_j + \frac{1}{2}\alpha (u_j^\top Q u_j + v_j^\top R v_j) + \frac{1}{4}\alpha^2 \sum_{(s', j') \neq (s, j)} \frac{1}{is\sigma_j - is'\sigma_{j'}} (ss' u_{j'}^\top Q u_j + v_{j'}^\top R v_j)^2 + O(\alpha^3), s \in \{-1, 1\}, 1 \leq j \leq n \right\}. \quad (3.1)$$

Now the zeroth- and second-order terms are all in $i\mathbb{R}$, hence the announced expansion for $\tilde{\mu}_{M_\alpha}$. \square

Estimate of the leading term under a probabilistic model. Assuming the singular vectors $(u_1, \dots, u_n), (v_1, \dots, v_n)$ of P are distributed uniformly at random – which is the case for example if P has i.i.d. Gaussian entries by rotational invariance⁴ –, the leading term in the expansion of $\tilde{\mu}_{M_\alpha}$ can be estimated in expectation as follows. The proof is placed in App. B, where we also include a high-probability version of the estimate as Prop. B.5.

Proposition 3.2. *Suppose Q, R are fixed and U, V are independently distributed uniformly on the set of $n \times n$ orthonormal matrices. Then*

$$\frac{\text{Tr}(S)}{n} \left(1 - 2 \frac{\|S\|_F}{\text{Tr}(S)} \sqrt{\log n} \right) \leq \mathbb{E} \left[\min_{1 \leq j \leq n} u_j^\top Q u_j + v_j^\top R v_j \right] \leq \frac{\text{Tr}(S)}{n}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. In particular, provided that $\frac{\text{Tr}(S)}{\|S\|_F} \geq 2\sqrt{\log n}(1+c)$ for some fixed $c > 0$, we have $\mathbb{E} [\min_{1 \leq j \leq n} u_j^\top Q u_j + v_j^\top R v_j] \asymp \frac{\text{Tr}(S)}{n}$ as $n \rightarrow \infty$.

Note that $\frac{\text{Tr}(S)}{\|S\|_F}$, which always lies in the interval $[1, \sqrt{2n}]$, is a measure of the effective sparsity of the spectrum of S (larger meaning less sparse), so the condition $\frac{\text{Tr}(S)}{\|S\|_F} \geq 2\sqrt{\log n}(1+c)$ means the spectrum of S is well spread-out. So the proposition shows that **the exponential convergence rate depends on the average of the eigenvalues of S** , when $\alpha S \ll A$ and the spectrum of S is well spread-out. This fact should be contrasted with the case of minimization, where the convergence rate scales as the *minimum* eigenvalue of the Hessian.

Interestingly, when the spectrum of S is sparse, the typical behavior of the leading term in the expansion of $\tilde{\mu}_{M_\alpha}$, $\min_j u_j^\top Q u_j + v_j^\top R v_j$, is quite different. In this case, that quantity depends on the geometric mean of the non-zero eigenvalues of S , rather than the arithmetic mean, as formalized in the following proposition. The proof is placed in App. B, along with a high-probability version of the estimate in Prop. B.9.

Proposition 3.3. *Suppose Q, R are fixed and U, V are independently distributed uniformly on the set of $n \times n$ orthonormal matrices. Let $s_1 \geq \dots \geq s_r > 0 = s_{r+1} = \dots = s_{2n}$ the eigenvalues of S .*

⁴If $P = U\Sigma V^\top$ has i.i.d. Gaussian entries, then P has the same law as $\tilde{U}^\top P \tilde{V}$, for any $\tilde{U}, \tilde{V} \in \mathcal{O}_n$ the set of $n \times n$ orthonormal matrices. So (U, V) has the same law as $(\tilde{U}^\top U, \tilde{V}^\top V)$. This shows that (U, V) is distributed according to the Haar measure on the product group $\mathcal{O}_n \times \mathcal{O}_n$, that is, U and V are independently distributed uniformly on \mathcal{O}_n .

Table 1: Expansions of $\rho(\nabla T(z^*))$ in α and η for classical gradient methods

Algorithm	$T(z)$	$\nabla T(z^*)$	$\rho(\nabla T(z^*))^2 = \max_{j \leq 2n} [\dots] + O(\eta\alpha^3 + \eta^2\alpha^2)$
Sim-GDA	$z - \eta g(z)$	$I - \eta M$	$1 - 2\alpha\eta (\bar{w}_j^\top S w_j) + \eta^2 \sigma_j^2$
Alt-GDA	see text	$I - \eta(I - \frac{\eta}{2}A)M + O(\eta^3)$	$1 - 2\alpha\eta (\bar{w}_j^\top S w_j) + O(\eta^3)$
EG	$z - \eta g(z - \eta g(z))$	$I - \eta(I - \eta M)M$	$1 - 2\alpha\eta (\bar{w}_j^\top S w_j) - \eta^2 \sigma_j^2 + O(\eta^3)$

Then

$$\mathbb{E} \left[\min_{j \leq n} u_j^\top Q u_j + v_j^\top R v_j \right] \geq \max_{S \subset \{1, \dots, r\}} \frac{1}{e} \frac{|S|}{n} n^{-\frac{2}{|S|}} \left[\prod_{l \in S} s_l \right]^{\frac{1}{|S|}}.$$

In particular, $\mathbb{E} \left[\min_{j \leq n} u_j^\top Q u_j + v_j^\top R v_j \right] \gtrsim n^{-\frac{2}{r}-1}$ when $n \rightarrow \infty$ and r and $s \in \mathbb{R}^r$ are fixed.

Numerically, the quantity $\min_{j \leq n} u_j^\top Q u_j + v_j^\top R v_j$ indeed scales as $n^{-\frac{2}{r}-1}$ under the conditions of the proposition, suggesting that our lower estimate could be tight.

3.2 Convergence rate of discrete-time algorithms

[Prop. 3.1](#) gave an expansion of the normalized convergence rate r/η of gradient methods (for a game with Jacobian at optimum M_α), non-asymptotically in α and in the asymptotic limit $\eta \rightarrow 0$. In this subsection, we give expansions of the convergence rate r that are non-asymptotic in α and η .

The algorithms we will consider can be written in the form $z^{k+1} = T(z^k)$ with the update operator T dependent only on ∇f and on step-size η , and satisfying $z^* = T(z^*)$. It is well-known that local convergence of such methods is determined by $\rho(\nabla T(z^*))$, where ∇T is the Jacobian of T and $\rho(\cdot)$ denotes spectral radius. Namely, if $\rho(\nabla T(z^*)) < 1$ then the iterates converge locally with $\|z^k - z^*\| = O((\rho(\nabla T(z^*)) + \varepsilon)^k)$, with $\varepsilon > 0$ an arbitrarily small constant [[Ber97](#), Proposition 4.4.1].

In [Table 1](#) we give an expansion for $\rho(\nabla T(z^*))$ for three classical gradient methods: simultaneous GDA (Sim-GDA), alternating GDA (Alt-GDA), and Extra-Gradient (EG). Let us clarify immediately that throughout this paper, statements made about ‘‘GDA’’ without further specification apply to both Sim-GDA and Alt-GDA. In the second column we denoted by $g(z) = \begin{pmatrix} \nabla_x f \\ -\nabla_y f \end{pmatrix}(z)$ the skewed gradient field of the game, and in the third column we wrote M instead of M_α for concision. In the fourth column, $\pm i\sigma_j$ denotes the eigenvalues of A assumed distinct and w_j are the associated eigenvectors – equivalently, the singular value decomposition of P is $P = \sum_{j=1}^n \sigma_j u_j v_j^\top$ and for each $j \leq n$, $\bar{w}_j^\top S w_j = \bar{w}_{j+n}^\top S w_{j+n} = \frac{1}{2} (u_j^\top Q u_j + v_j^\top R v_j)$. We refer to [App. D](#) for the derivation of this table and for explicit bounds on the ‘‘ $O(\cdot)$ ’’ terms.

Informally, as one can directly see from the fourth column, Sim-GDA requires a very small step-size for the first term in η to overcome the terms $+\eta^2 \sigma_j^2$, while for EG those terms actually appear with a favorable sign, and Alt-GDA neither benefits nor suffers from those terms. We also see that Alt-GDA is quite faithful to GF, in that their normalized convergence rates coincide up to terms of order $\eta^3 + \alpha^4$. All of these insights are in line with common intuition in the min-max optimization literature [[BGP20](#); [Lu22](#)], as well as with our numerical experiments for the next section, [Fig. 2](#).

A symmetrized formulation of Alt-GDA. In order to derive the rate for Alt-GDA, we used the following symmetrized formulation of it: we let $(x^0, y^{1/2}) \in \mathbb{R}^d \times \mathbb{R}^d$ and

$$\begin{cases} \forall k \in \mathbb{N}, & x^{k+1} = x^k - \eta \nabla_x f(x^k, y^{k+1/2}) \\ \forall k \in \mathbb{N} + 1/2, & y^{k+1} = y^k + \eta \nabla_y f(x^{k+1/2}, y^k) \end{cases} \quad \text{and} \quad \begin{cases} \forall k \in \mathbb{N}, & x^{k+1/2} = \frac{x^{k+1} + x^k}{2} \\ \forall k \in \mathbb{N} + 1/2, & y^{k+1/2} = \frac{y^{k+1} + y^k}{2}. \end{cases}$$

That is, x gets updated with the gradient rule at integer time-steps, y at half-integer time-steps, and we define x^k, y^k at non-updating time-steps as the average of the preceding and following updating time-steps. Assuming $\eta \leq \|\nabla_{xx}^2 f\|_\infty^{-1} \wedge \|\nabla_{yy}^2 f\|_\infty^{-1}$, we show in [App. C](#) that z^{k+1} is indeed entirely

determined by $z^k = (x^k, y^k)$ for each $k \in \mathbb{N}$, and that the associated update operator T satisfies

$$T(z) = z - \eta g(z) + \frac{\eta^2}{2} A(z)g(z) + O(\eta^3 \|g(z)\|) \quad \text{where} \quad A(z) = \frac{\nabla g(z) - \nabla g(z)^\top}{2}.$$

For comparison, the usual formulation of Alt-GDA considers as the iterates $(\tilde{x}^k, \tilde{y}^k) = (x^k, y^{k+1/2})$. We emphasize that $(\tilde{x}^k, \tilde{y}^k)$ and (x^k, y^k) have the same convergence rate if they converge exponentially, as one can check directly from the definition.

4 Illustration: sparse mixed Nash equilibria of continuous games

In this section we apply the above considerations to a particular class of min-max problems, which is of its own interest in game theory. Namely we consider the classical problem of finding the mixed Nash equilibria (MNE) of two-player zero-sum games, that is, given strategy spaces \mathcal{X} , \mathcal{Y} and a payoff function $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, solving the min-max problem

$$\min_{\mu \in \mathcal{P}(\mathcal{X})} \max_{\nu \in \mathcal{P}(\mathcal{Y})} \{F(\mu, \nu) = \mathbb{E}_{x \sim \mu, y \sim \nu} [f(x, y)]\}.$$

Here $\mathcal{P}(\mathcal{X})$ denotes the space of probability measures – representing mixed strategies – over \mathcal{X} . Let us focus on cases where \mathcal{X} and \mathcal{Y} are continuous sets, say, $\mathcal{X} = \mathcal{Y} = \mathbb{T}^1$ the one-dimensional Euclidean torus,⁵ and $f : \mathbb{T}^1 \times \mathbb{T}^1 \rightarrow \mathbb{R}$ is smooth. The above min-max problem is then infinite-dimensional, and algorithms to solve it explicitly must be based on reparameterized formulations. More specifically, suppose that the MNE (μ^*, ν^*) is unique and “sparse”, i.e., has finite supports: $\text{supp}(\mu^*) = \{x_I^*, 1 \leq I \leq N\}$, $\text{supp}(\nu^*) = \{y_J^*, 1 \leq J \leq M\}$ (this is the case for example if f is a sum of separable functions [SOP08]).

In this setting there are two natural reparameterizations and associated algorithms:

1. If the optimal support points $\{x_I^*\}_I$, $\{y_J^*\}_J$ are known, then we may reparameterize by $\mu = \sum_{I=1}^N a_I \delta_{x_I^*}$, $\nu = \sum_{J=1}^M b_J \delta_{y_J^*}$ and optimize over the a_I, b_J . The problem reduces to a constrained bilinear game

$$\min_{a \in \Delta_N} \max_{b \in \Delta_M} \left\{ F_1(a, b) = \sum_{I=1}^N \sum_{J=1}^M a_I b_J f(x_I^*, y_J^*) =: a^\top P b \right\}$$

where Δ_N denotes the standard simplex. A classical approach is then to apply Mirror Prox (MP) with entropy link function [Nem04] (MP is the Bregman-geometry analog of EG).

2. If only the number of optimal support points is known, then we may reparameterize by $\mu = \sum_{I=1}^N a_I \delta_{x_I}$, $\nu = \sum_{J=1}^M b_J \delta_{y_J}$ and optimize over both the weights (a_I, b_J) and the support points (x_I, y_J) . The problem reduces to

$$\min_{(a,x) \in \Delta_N \times (\mathbb{T}^1)^N} \max_{(b,y) \in \Delta_M \times (\mathbb{T}^1)^M} \left\{ F_2(a, x, b, y) = \sum_{I=1}^N \sum_{J=1}^M a_I b_J f(x_I, y_J) \right\}.$$

A possible approach is to iteratively update (simultaneously) the a, b using MP steps with step-size η and the x, y using EG steps with step-size $\gamma\eta$, for some parameter $\gamma > 0$. This algorithm is called Conic Particle Mirror Prox in [WC22], but for concision we will refer to it simply as “EG” in this section. Note that the main challenge in that reference is to deal with the case where N and M are unknown, but here we assume they are known for the sake of simplicity.

In Fig. 2, we show the dependency on η and γ of the local convergence rate of MP and EG, as well as that of the analogs of Sim-GDA and Alt-GDA and GF for the problem $\min_{(a,x)} \max_{(b,y)} F_2(a, x, b, y)$ in order to illustrate the insights from Sec. 3.2. We used a randomly generated payoff function f , and the convergence is measured by the iterates’ ℓ_2 distance to the solution; see Sec. F.1 for details.

MP (the algorithm based on the reparameterization F_1) converges to the MNE for any small enough η with an exponential rate at least proportional to η^2 [WLZL21], and this scaling is tight numerically

⁵The choice of $\mathcal{X} = \mathcal{Y} = \mathbb{T}^1$ is made for simplicity of exposition. The discussion below extends straightforwardly to toruses of any dimension, and could be extended to \mathcal{X}, \mathcal{Y} Riemannian manifolds without boundaries at the cost of more technical notation.

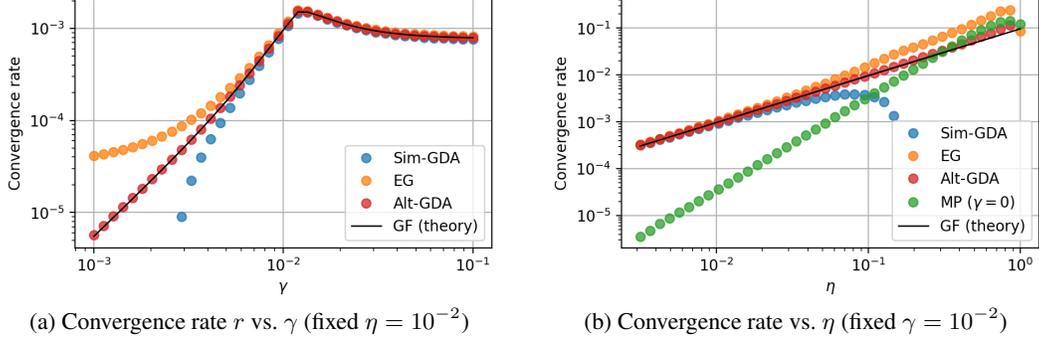


Figure 2: Observed local convergence rates r (i.e., $\|z^k - z^*\| = \Theta((1-r)^k)$) for various conic particle methods using step-size η for the weights (a, b) and $\gamma\eta$ for the positions (x, y) and for a fixed random draw of f . (left) Fixing $\eta = 10^{-2}$, we observe a rate for GF (we plot $\eta \cdot \tilde{\mu}_\gamma$) scaling as γ^2 as predicted in Prop. 4.1. Interestingly, Alt-GDA has exactly the same behavior, while the behavior of Sim-GDA and EG differ for small γ , due to the corrective terms shown in Table 1. (right) Fixing $\gamma = 10^{-2}$, we observe a rate scaling as η for EG and its variants, and as η^2 for MP; the convergence of EG is mostly faster than MP.

(green dots in Fig. 2b). Its explicit variant (Mirror Descent-Ascent) and its continuous-time flow are known to diverge for any η [BP18] [MPP18].

EG is known to converge locally to the MNE for any small enough η (and any $\gamma > 0$) with an exponential rate at least proportional to η^2 , despite the non-convexity of F_2 , under some non-degeneracy assumptions [WC22, Sec. 3.1]. However numerically the convergence rate of EG typically scales as η , not η^2 (orange dots in Fig. 2b) – a fact which we will explain below. The explicit variant of EG and its continuous-time flow have not previously been analyzed; the discussion below will give a characterization of when they converge locally.

Note that, at least for the particular f and γ used for Fig. 2b, EG converges locally faster than MP for the same η (in number of iterations, the per-iteration costs differing only by a constant factor), even though the former does not use the knowledge of the $\{x_I^*\}_I, \{y_J^*\}_J$! In other words, even when the optimal support points are known, it is beneficial to use the overparameterized formulation F_2 where we also vary the support points.

Overparameterization induces partial curvature. Let us inspect the “Jacobians” at optimum for the two dynamics, MP vs. EG with parameter γ . Due to the simplex constraints and the non-Euclidean nature of the updates for a, b , the relevant matrices are M_{MP} and M_γ defined below, in the sense that the exponential convergence rate of each algorithm is $\tilde{\mu}_M \eta + O(\eta^2)$ (see Sec. F.2, F.3). Namely, omitting half of the antisymmetric off-diagonals for readability,

$$M_{\text{MP}} = \begin{bmatrix} \mathbf{0} & \Pi_a D_a P D_b \Pi_b^\top \\ -(*)^\top & \mathbf{0} \end{bmatrix}$$

where $D_a = \text{Diag}(\sqrt{a^*})$ (square roots being taken component-wise) and $\Pi_a \in \mathbb{R}^{(N-1) \times N}$ is any matrix such that $\Pi_a \Pi_a^\top = I_{N-1}$ and $\Pi_a^\top \Pi_a = I_N - \sqrt{a^*} \sqrt{a^*}^\top$, and likewise for D_b and $\Pi_b \in \mathbb{R}^{(M-1) \times M}$; and for EG,

$$M_\gamma = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \Pi_a D_a P D_b \Pi_b^\top & \sqrt{\gamma} \Pi_a D_a [\partial_y P] D_b \\ \mathbf{0} & \gamma \text{Diag}(\partial_{xx}^2 P b^*) & \sqrt{\gamma} D_a [\partial_x P] D_b \Pi_b^\top & \gamma D_a [\partial_{xy}^2 P] D_b \\ -(*)^\top & -(*)^\top & \mathbf{0} & \mathbf{0} \\ -(*)^\top & -(*)^\top & \mathbf{0} & -\gamma \text{Diag}(\partial_{yy}^2 P^\top a^*) \end{bmatrix}$$

where $[\partial_x P]_{IJ} = \partial_x f(x_I^*, y_J^*)$, and likewise for $\partial_y P, \partial_{xx}^2 P, \partial_{yy}^2 P, \partial_{xy}^2 P$.

For MP, it is clear that the equivalent conditions of Thm. 2.1 are violated for any payoff function f , and so $\tilde{\mu}_{M_{\text{MP}}} = 0$. For EG, depending on f and γ , they may or may not be violated. For all of the random payoff functions we considered in our experiments, we observed that $\tilde{\mu}_{M_\gamma} > 0$, suggesting that the conditions hold generically. They are violated for certain f ’s however, as shown in Sec. F.4, so that the scaling in η^2 of the convergence rate proved in [WC22] is tight in the worst case.

More precisely as we show in the next proposition, assuming that the blocks of M_γ are in general position, we expect $\tilde{\mu}_{M_\gamma}$ to scale as γ^2 , which is indeed observed in the numerical experiment reported in Fig. 2a. The proof, placed in Sec. F.5, relies on the same tools as Prop. 3.1, that is, on the asymptotic expansions of the eigenvalues of perturbed matrices.

Proposition 4.1. *Let S_2 symmetric and A_0, A_1, A_2 antisymmetric real matrices of the form*

$$S_2 = \left[\begin{array}{c|c} \mathbf{0} & \\ \hline * & \\ \hline & \mathbf{0} \\ & * \end{array} \right], A_0 = \left[\begin{array}{c|c} * & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \\ \hline * & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{array} \right], A_1 = \left[\begin{array}{c|c} \mathbf{0} & * \\ \hline & * \\ \hline \mathbf{0} & * \\ * & \mathbf{0} \end{array} \right], A_2 = \left[\begin{array}{c|c} \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & * \\ \hline \mathbf{0} & \mathbf{0} \\ \mathbf{0} & * \end{array} \right]$$

and $M_\gamma = \gamma S_2 + A_0 + \sqrt{\gamma} A_1 + \gamma A_2$ for all $\gamma > 0$. Then $\tilde{\mu}_{M_\gamma} = O(\gamma^2)$ as $\gamma \rightarrow 0$.

5 Conclusion

We have investigated the local convergence of gradient methods for min-max games and found that they converge generically under partial curvature. In more specific settings, we have obtained quantitative estimates of the local convergence rate which exhibit the *average* of the eigenvalues of S as the driving quantity for typical problems. For the computation of mixed Nash equilibria of continuous games, this leads to a behavior of conic particle gradient methods that is more favorable than that described by the worst-case bounds.

More generally, our analysis leads to the following insights: (i) worst-case bounds might be looser in min-max optimization (compared to minimization) as they fail to capture the interplay between interaction and potential parts; (ii) the addition of new degrees of freedom with curvature typically accelerates the local convergence, as we illustrated in Sec. 4.

We note that the phenomenon described in this paper is fundamentally a consequence of the fact that the skewed gradient field’s Jacobian at optimum has a positive-semidefinite symmetric part. This property is satisfied at local Nash equilibria of min-max games, i.e., of two-player zero-sum differentiable games, but is not true for general differentiable games.

References

- [ALW21] Jacob Abernethy, Kevin A Lai, and Andre Wibisono. “Last-iterate convergence rates for min-max optimization: Convergence of hamiltonian gradient descent and consensus optimization”. In: *Algorithmic Learning Theory*. PMLR. 2021, pp. 3–47.
- [AW20] Ehsan Amid and Manfred KK Warmuth. “Reparameterizing mirror descent as gradient descent”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 8430–8439.
- [AMLG20] Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. “A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 2863–2873.
- [BGP20] James P Bailey, Gauthier Gidel, and Georgios Piliouras. “Finite regret and cycles with fixed step-size via alternating gradient descent-ascent”. In: *Conference on Learning Theory*. PMLR. 2020, pp. 391–407.
- [BP18] James P Bailey and Georgios Piliouras. “Multiplicative weights update in zero-sum games”. In: *Proceedings of the 2018 ACM Conference on Economics and Computation*. 2018, pp. 321–338.
- [Ber97] Dimitri P Bertsekas. “Nonlinear programming”. In: *Journal of the Operational Research Society* 48.3 (1997), pp. 334–334.
- [Doa22] Thinh Doan. “Convergence rates of two-time-scale gradient descent-ascent dynamics for solving nonconvex min-max problems”. In: *Learning for Dynamics and Control Conference*. PMLR. 2022, pp. 192–206.
- [DPS21] Carles Domingo-Enrich, Fabian Pedregosa, and Damien Scieur. “Average-case Acceleration for Bilinear Games and Normal Matrices”. In: *International Conference on Learning Representations*. 2021.

- [DH19] Simon S Du and Wei Hu. “Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 196–205.
- [Gid+19] Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Rémi Le Priol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. “Negative momentum for improved game dynamics”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 1802–1811.
- [GLWM22] Benjamin Grimmer, Haihao Lu, Pratik Worah, and Vahab Mirrokni. “Limiting behaviors of nonconvex-nonconcave minimax optimization via continuous-time systems”. In: *International Conference on Algorithmic Learning Theory*. PMLR. 2022, pp. 465–487.
- [GWS21] Suriya Gunasekar, Blake Woodworth, and Nathan Srebro. “Mirrorless mirror descent: A natural derivation of mirror descent”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 2305–2313.
- [LM00] Beatrice Laurent and Pascal Massart. “Adaptive estimation of a quadratic functional by model selection”. In: *Annals of Statistics* (2000), pp. 1302–1338.
- [Let+19] Alistair Letcher, David Balduzzi, Sébastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. “Differentiable game mechanics”. In: *The Journal of Machine Learning Research* 20.1 (2019), pp. 3032–3071.
- [LS19] Tengyuan Liang and James Stokes. “Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 907–915.
- [Lu22] Haihao Lu. “An $O(s^r)$ -resolution ODE framework for understanding discrete-time algorithms and applications to the linear convergence of minimax problems”. In: *Mathematical Programming* 194.1 (2022), pp. 1061–1112.
- [MPP18] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. “Cycles in adversarial regularized learning”. In: *Proceedings of the twenty-ninth annual ACM-SIAM symposium on discrete algorithms*. SIAM. 2018, pp. 2703–2717.
- [MOP20] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. “A unified analysis of extragradient and optimistic gradient methods for saddle point problems: Proximal point approach”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 1497–1507.
- [NK17] Vaishnavh Nagarajan and J Zico Kolter. “Gradient descent GAN optimization is locally stable”. In: *Advances in neural information processing systems* 30 (2017).
- [Nem04] Arkadi Nemirovski. “Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems”. In: *SIAM Journal on Optimization* 15.1 (2004), pp. 229–251.
- [PS20] Fabian Pedregosa and Damien Scieur. “Acceleration through spectral density estimation”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7553–7562.
- [Per13] Lawrence Perko. *Differential equations and dynamical systems*. Vol. 7. Springer Science & Business Media, 2013.
- [PP+08] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. “The matrix cookbook”. In: *Technical University of Denmark* 7.15 (2008), p. 510.
- [Roc76] R Tyrrell Rockafellar. “Monotone operators and the proximal point algorithm”. In: *SIAM journal on control and optimization* 14.5 (1976), pp. 877–898.
- [SOP08] Noah D Stein, Asuman Ozdaglar, and Pablo A Parrilo. “Separable and low-rank continuous games”. In: *International Journal of Game Theory* 37.4 (2008), pp. 475–504.
- [Tao08] Terence Tao. *When are eigenvalues stable*. Oct. 28, 2008. URL: <https://terrytao.wordpress.com/2008/10/28/when-are-eigenvalues-stable>.
- [VTM07] Nico Van der Aa, H Ter Morsche, and R Mattheij. “Computation of eigenvalue and eigenvector derivatives for a general complex-valued eigensystem”. In: *The electronic journal of linear algebra* 16 (2007), pp. 300–314.
- [Vil09] Cédric Villani. *Hypocoercivity*. Vol. 202. 950. American Mathematical Society, 2009.

- [Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge university press, 2019.
- [WC22] Guillaume Wang and Lénaïc Chizat. “An Exponentially Converging Particle Method for the Mixed Nash Equilibrium of Continuous Games”. In: *arXiv preprint arXiv:2211.01280* (2022).
- [WLZL21] Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. “Linear Last-iterate Convergence in Constrained Saddle-point Optimization”. In: *International Conference on Learning Representations*. 2021.
- [ZWLG22] Guodong Zhang, Yuanhao Wang, Laurent Lessard, and Roger B Grosse. “Near-optimal local convergence of alternating gradient descent-ascent for minimax optimization”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 7659–7679.

A Local convergence rates of classical gradient methods

Consider a game $\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} f(x, y)$ with local NE z^* and let $M = \begin{bmatrix} \nabla_{xx}^2 f & \nabla_{xy}^2 f \\ -\nabla_{xy}^2 f^\top & -\nabla_{yy}^2 f \end{bmatrix}(z^*)$, assumed invertible for simplicity. Gradient-based iterative methods can be written in the form $z^{k+1} = T(z^k)$ with the update operator T dependent only on ∇f and on step-size η , and satisfying $z^* = T(z^*)$. It is well-known that local convergence is determined by $\rho(\nabla T(z^*))$ where ∇T is the Jacobian of T . Namely, if $\rho(\nabla T(z^*)) < 1$ then the iterates converge locally with $\|z^k - z^*\| = O(\rho(\nabla T(z^*)) + \varepsilon)^k$ with $\varepsilon > 0$ an arbitrarily small constant [Ber97, Proposition 4.4.1].

Table 2 summarizes the local convergence rates of three classical gradient methods, up to third order in η . The last two columns of the table are valid for $\eta \leq \eta_{\max}$ (given by the second column), and the $O(\cdot)$'s hide only universal constants. In the last column, we denoted $\tilde{\mu} = \tilde{\mu}_M$ for concision, and $L = \rho(M)$. “ k -EG” stands for k -step Extra-Gradient. The table was extracted from the proofs of [AMLG20, Appendix E].

As shown in the fourth column, the quantity $\tilde{\mu}_M$ appears naturally in the bounds on the local convergence rate $1 - \rho(\nabla T(z^*))$, with all methods benefitting from $\tilde{\mu}_M$ being larger. Moreover, the rate depends linearly on $\tilde{\mu}_M$ in the asymptotic regime of small η , as can be seen in the third column, since $1 - \rho(\nabla T(z^*))^2 \sim 2\eta \min_{\lambda \in \text{Sp}(M)} \Re(\lambda) = 2\eta \tilde{\mu}_M$.⁶ This asymptotic equivalence corresponds to the fact that GF, the continuous-time flow of all of the classical gradient methods, has local exponential convergence rate equal to $\tilde{\mu}_M$. (Of course using a large η , or following a different continuous-time flow than GF, may lead to faster convergence, but it may require knowledge of problem parameters or the use of more complex adaptive schemes.)

Interestingly, as can be seen in the third column, EG may have a slower convergence rate than GDA if $|\Im(\lambda)| < \Re(\lambda)$ for $\lambda = \arg \min_{\text{Sp}(M)} \Re(\cdot)$, when η is small.

Table 2: Moduli of eigenvalues of $\nabla T(z^*)$ for classical gradient methods

Algorithm	η_{\max}	$\{ \nu ^2, \nu \in \text{Sp}(\nabla T(z^*))\}$	Upper bound on $\rho(\nabla T(z^*))^2$
Sim-GDA	∞	$\{1 - 2\eta \Re(\lambda) + \eta^2 \lambda ^2, \lambda \in \text{Sp}(M)\}$	$1 - 2\eta \tilde{\mu} + \eta^2 L^2$
k -EG ($k \geq 2$)	$(1-c)/L$ ($\forall c > 0$)	$\{1 - 2\eta \Re(\lambda) - \eta^2 (\lambda ^2 - 4\Re(\lambda)^2) + O(\frac{1}{1-c}\eta^3 \lambda ^3), \lambda \in \text{Sp}(M)\}$	$1 - \frac{1}{(2-c)^2} \max(2\eta \tilde{\mu}, \eta^2 L^2)$
PP	∞	$\{\frac{1}{1+2\eta \Re(\lambda) + \eta^2 \lambda ^2}, \lambda \in \text{Sp}(M)\}$	$1 - \max(\frac{2\eta \tilde{\mu}}{1+2\eta \tilde{\mu}}, \frac{\eta^2 L^2}{1+\eta^2 L^2})$

B Details for Sec. 3.1

B.1 Proof of Prop. 3.2

For ease of reference, we restate the proposition below.

Proposition 3.2. *Suppose Q, R are fixed and U, V are independently distributed uniformly on the set of $n \times n$ orthonormal matrices. Then*

$$\frac{\text{Tr}(S)}{n} \left(1 - 2 \frac{\|S\|_F}{\text{Tr}(S)} \sqrt{\log n}\right) \leq \mathbb{E} \left[\min_{1 \leq j \leq n} u_j^\top Q u_j + v_j^\top R v_j \right] \leq \frac{\text{Tr}(S)}{n}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. In particular, provided that $\frac{\text{Tr}(S)}{\|S\|_F} \geq 2\sqrt{\log n}(1+c)$ for some fixed $c > 0$, we have $\mathbb{E}[\min_{1 \leq j \leq n} u_j^\top Q u_j + v_j^\top R v_j] \asymp \frac{\text{Tr}(S)}{n}$ as $n \rightarrow \infty$.

⁶ $1 - \rho(\nabla T(z^*))^2$ is the exponential convergence rate for $\|z^k - z^*\|^2$, and $1 - \rho(\nabla T(z^*))$ is the one for $\|z^k - z^*\|$. Hence those two quantities differ by a factor 2 at first order in η . Equivalently, the additional factor 2 comes from the fact that $\rho^2 = [1 - (1 - \rho)]^2 \approx 1 - 2(1 - \rho)$ when ρ is close to 1.

For the sake of concision, let $W = \begin{bmatrix} U \\ V \end{bmatrix}$ and $w_j = W_{\bullet j} = \begin{pmatrix} u_j \\ v_j \end{pmatrix}$, and (in this appendix only) $\mu := \min_{j \leq n} u_j^\top Q u_j + v_j^\top R v_j = \min_{j \leq n} w_j^\top S w_j$ the quantity which we want to estimate. Furthermore, denote $s_1 \geq \dots \geq s_r > 0 = s_{r+1} = \dots = s_{2n}$ the eigenvalues of S , and assume S is diagonal w.l.o.g. Also let, for all $j \leq n$, $N_1, \dots, N_n, M_1, \dots, M_n \sim \chi_n^2$ i.i.d. and independent of U and V , and pose $a_j = \begin{pmatrix} \sqrt{N_j} u_j \\ \sqrt{M_j} v_j \end{pmatrix}$.

Note that for each j , $a_j \sim \mathcal{N}(0, I_{2n})$. Indeed, $\sqrt{N_j} u_j \sim \mathcal{N}(0, I_n)$ since it is isotropic and its norm has the correct distribution, and likewise for $\sqrt{M_j} v_j$; so $a_j \sim \mathcal{N}(0, I_{2n})$ as the concatenation of two independent standard Gaussian vectors.

Lemma B.1. *We have $\frac{1}{n} \mathbb{E} [\min_{j \leq n} a_j^\top S a_j] \leq \mathbb{E} \mu \leq \frac{\text{Tr}(S)}{n}$.*

Proof. Let $J := \arg \min_{j \leq n} w_j^\top S w_j$. Then $a_J^\top S a_J = N_J u_J^\top Q u_J + M_J v_J^\top R v_J$. Now J is a deterministic function of the $(w_j)_{j \leq n}$, and $(N_j, M_j)_{j \leq n}$ are i.i.d. and independent of $(w_j)_{j \leq n}$, so we have the conditional expectation $\mathbb{E}[N_J | (w_j)_{j \leq n}] = \mathbb{E} N_1 = \mathbb{E} \chi_n^2 = n$ and likewise for M_J . Thus,

$$\mathbb{E} [a_J^\top S a_J | (w_j)_{j \leq n}] = \mathbb{E} [N_J | (w_j)_{j \leq n}] u_J^\top Q u_J + \mathbb{E} [M_J | (w_j)_{j \leq n}] v_J^\top R v_J = n \mu,$$

and the first inequality follows by taking total expectations.

For the second inequality, $\mathbb{E} \mu \leq \mathbb{E} u_1^\top Q u_1 + v_1^\top R v_1$. Now letting $a_1^{(1)} = \sqrt{N_1} u_1 \sim \mathcal{N}(0, I_n)$, $N_1 u_1^\top Q u_1 = (a_1^{(1)})^\top Q a_1^{(1)}$, so taking expectations on both sides, $n \mathbb{E} u_1^\top Q u_1 = \text{Tr}(Q)$. So $\mathbb{E} \mu \leq \frac{\text{Tr}(Q) + \text{Tr}(R)}{n} = \frac{\text{Tr}(S)}{n}$. \square

Let for concision $\zeta_j = a_j^\top S a_j$ for each $j \leq n$.

Lemma B.2. *The moment-generating function of $\zeta_1 (\sim \zeta_2 \sim \dots \sim \zeta_n)$ is*

$$\mathbb{E} e^{t \zeta_1} = \prod_{l=1}^r (1 - 2s_l t)^{-1/2} \quad \text{for all } t < \frac{1}{2 \max_l s_l}.$$

Proof. Since $\zeta_1 = \sum_{l=1}^{2n} s_l a_1[l]^2$ and $a_1[l]^2 \sim \chi^2$ i.i.d., we have $\mathbb{E} e^{t a_1[l]^2} = (1 - 2t)^{-1/2}$ for all $t < \frac{1}{2}$ and $\mathbb{E} e^{t \zeta_1} = \prod_{l=1}^r \mathbb{E} e^{t s_l a_1[l]^2} = \prod_{l=1}^r (1 - 2s_l t)^{-1/2}$ for all $t < \frac{1}{2 \max_l s_l}$. \square

We now lower-bound the expectation of $\min_{j \leq n} \zeta_j$ using a Chernoff bound-type argument, which we note does not require independence.

Lemma B.3. *We have*

$$\mathbb{E} \min_{j \leq n} \zeta_j \geq \text{Tr}(S) \left(1 - 2 \frac{\|S\|_F}{\text{Tr}(S)} \sqrt{\log n} \right)$$

where $\|\cdot\|_F$ denotes Frobenius norm, i.e., ℓ_2 -norm of the vector of eigenvalues.

Proof. By Jensen's inequality and monotonicity of $\exp(\cdot)$,

$$\forall t > 0, \quad \exp \left(t \mathbb{E} \max_{j \leq n} -\zeta_j \right) \leq \mathbb{E} \exp \left(t \max_{j \leq n} -\zeta_j \right) = \mathbb{E} \max_{j \leq n} \exp(-t \zeta_j).$$

So, taking log and optimizing the bound,

$$\begin{aligned} \mathbb{E} \max_{j \leq n} (-\zeta_j) &\leq \inf_{t > 0} \frac{1}{t} \log \mathbb{E} \max_{j \leq n} \exp(t(-\zeta_j)) \\ &\leq \inf_{t > 0} \frac{1}{t} \log \left[n \mathbb{E} e^{t(-\zeta_1)} \right] = \inf_{t > 0} \frac{1}{t} \log \left[n \prod_{l=1}^r (1 + 2s_l t)^{-1/2} \right] \\ &= \inf_{t > 0} \frac{1}{t} \left(\log(n) - \frac{1}{2} \sum_{l=1}^r \log(1 + 2s_l t) \right) =: \inf_{t > 0} g(t). \end{aligned} \quad (\text{B.1})$$

By calculating we find that $g'(t) > 0 \iff \sum_{l=1}^r \left[\log(1 + 2s_l t) - \frac{2s_l t}{1+2s_l t} \right] > 2 \log n$. With the case $r \gg \log n$ in mind, let us evaluate at $t^{(1)}$ defined by $\sum_{l=1}^r 2(s_l t)^2 = 2 \log n$, i.e., $t^{(1)} = \sqrt{\log n} / \|S\|_F$. (This choice is obtained by Taylor expansion for $t \rightarrow 0$ of the condition $g'(t) = 0$.) Using that $\log(1 + y) \geq y - \frac{1}{2}y^2$ for $y \geq 0$, we get

$$\begin{aligned} g(t^{(1)}) &= \frac{\|S\|_F}{\sqrt{\log n}} \left(\log n - \frac{1}{2} \sum_{l=1}^r \log \left(1 + 2s_l \frac{\sqrt{\log n}}{\|S\|_F} \right) \right) \\ &\leq \frac{\|S\|_F}{\sqrt{\log n}} \left(\log n - \frac{1}{2} \sum_{l=1}^r \left(2s_l \frac{\sqrt{\log n}}{\|S\|_F} - \frac{1}{2} 4s_l^2 \frac{\log n}{\|S\|_F^2} \right) \right) \\ &= \frac{\|S\|_F}{\sqrt{\log n}} \left(\log n - \frac{\text{Tr}(S) \sqrt{\log n}}{\|S\|_F} + \log n \right) \\ &= -\text{Tr}(S) + 2 \|S\|_F \sqrt{\log n}. \end{aligned}$$

Thus, $\mathbb{E} \min_{j \leq n} \zeta_j \geq -g(t^{(1)}) \geq \text{Tr}(S) - 2 \|S\|_F \sqrt{\log n} = \text{Tr}(S) \left(1 - 2 \frac{\|S\|_F}{\text{Tr}(S)} \sqrt{\log n} \right)$. \square

The upper bound of [Prop. 3.2](#) is shown in [Lemma B.1](#), and the lower bound follows immediately from substituting [Lemma B.3](#) into [Lemma B.1](#).

B.2 Proof of [Prop. 3.3](#)

For ease of reference, we restate the proposition below.

Proposition 3.3. *Suppose Q, R are fixed and U, V are independently distributed uniformly on the set of $n \times n$ orthonormal matrices. Let $s_1 \geq \dots \geq s_r > 0 = s_{r+1} = \dots = s_{2n}$ the eigenvalues of S . Then*

$$\mathbb{E} \left[\min_{j \leq n} u_j^\top Q u_j + v_j^\top R v_j \right] \geq \max_{\mathcal{S} \subset \{1, \dots, r\}} \frac{1}{e} \frac{|\mathcal{S}|}{n} n^{-\frac{2}{|\mathcal{S}|}} \left[\prod_{l \in \mathcal{S}} s_l \right]^{\frac{1}{|\mathcal{S}|}}.$$

In particular, $\mathbb{E} \left[\min_{j \leq n} u_j^\top Q u_j + v_j^\top R v_j \right] \gtrsim n^{-\frac{2}{r}-1}$ when $n \rightarrow \infty$ and r and $s \in \mathbb{R}^r$ are fixed.

We are still exactly in the same setting as for [Prop. 3.2](#), so all the lemmas of [Sec. B.1](#) apply. We also reuse notations from that subsection: $\mu = \min_{j \leq n} u_j^\top Q u_j + v_j^\top R v_j$, $a_j = \begin{pmatrix} \sqrt{N_j} u_j \\ \sqrt{M_j} v_j \end{pmatrix}$ where $N_1, \dots, N_n, M_1, \dots, M_n \sim \chi_n^2$ i.i.d. and independent of U and V , and $\zeta_j = a_j^\top S a_j$ for $j \leq n$. We also assume S diagonal w.l.o.g.

Recall from [Lemma B.1](#) that $\mathbb{E} \mu \geq \frac{1}{n} \mathbb{E} \min_{j \leq n} \zeta_j$. The proposition then follows immediately from the following lemma.

Lemma B.4. *Let any $\mathcal{S} \subset \{1, \dots, r\}$, denote $r_{\mathcal{S}} = |\mathcal{S}|$ and $G_{\mathcal{S}} = \left[\prod_{l \in \mathcal{S}} s_l \right]^{1/r_{\mathcal{S}}}$. Then we have $\mathbb{E} \min_{j \leq n} \zeta_j \geq \frac{1}{e} r_{\mathcal{S}} n^{-\frac{2}{r_{\mathcal{S}}}} G_{\mathcal{S}}$.*

Proof. Recall from [\(B.1\)](#) that

$$\mathbb{E} \max_{j \leq n} (-\zeta_j) \leq \inf_{t > 0} \frac{1}{t} \left(\log(n) - \frac{1}{2} \sum_{l=1}^r \log(1 + 2s_l t) \right) =: \inf_{t > 0} g(t)$$

and that $g'(t) > 0 \iff \sum_{l=1}^r \left[\log(1 + 2s_l t) - \frac{2s_l t}{1+2s_l t} \right] > 2 \log n$.

With the case $r \ll \log n$ in mind, let us evaluate at $t^{(2)}$ defined by $\sum_{l=1}^r [\log(1 + 2s_l t) - 1] = 2 \log n$, i.e., $\prod_{l=1}^r (1 + 2s_l t) = e^{2 \log n + r} = n^2 e^r$ – more precisely, let $t^{(2)}$ be the smallest positive such t (since there may be several solutions to that polynomial equation). Then $g(t^{(2)}) = \frac{1}{t^{(2)}} (-r/2)$.

Let us further upper-bound $t^{(2)}$ by some $u > 0$. Since $t^{(2)}$ is defined as the smallest positive root of the polynomial $P(t) = \prod_{l=1}^r (1 + 2s_l t) - n^2 e^r$, and since $P(0) < 0$, it suffices to pick

any $u > 0$ that $P(u) \geq 0$. Since $P(t) \geq [\prod_{l=1}^r 2s_l] t^r - n^2 e^r$ for any $t > 0$,⁷ we may choose $u = (n^2 e^r)^{\frac{1}{r}} / [\prod_{l=1}^r 2s_l]^{\frac{1}{r}}$.

Thus we have

$$\begin{aligned} -\mathbb{E} \min_{j \leq n} \zeta_j &\leq g(t^{(2)}) = \frac{1}{t^{(2)}}(-r/2) \leq \frac{1}{u}(-r/2) = \frac{[\prod_{l=1}^r 2s_l]^{1/r}}{(n^2 e^r)^{1/r}}(-r/2) \\ \mathbb{E} \min_{j \leq n} \zeta_j &\geq r \frac{[\prod_{l=1}^r s_l]^{1/r}}{e \cdot n^{2/r}}. \end{aligned}$$

This shows the lemma in the case $\mathcal{S} = \{1, \dots, r\}$.

For the case of arbitrary $\mathcal{S} \subset \{1, \dots, r\}$, drop some terms in the Chernoff bound-type inequality; that is, upper-bound $-\sum_{l \notin \mathcal{S}} \log(1 + 2s_l t)$ by zero:

$$-\mathbb{E} \min_{j \leq n} \zeta_j \leq \inf_{t > 0} \frac{1}{t} \left(\log(n) - \frac{1}{2} \sum_{l=1}^r \log(1 + 2s_l t) \right) \leq \inf_{t > 0} \frac{1}{t} \underbrace{\left(\log(n) - \frac{1}{2} \sum_{l \in \mathcal{S}} \log(1 + 2s_l t) \right)}_{=: g_{\mathcal{S}}(t)}.$$

Thereafter, going through exactly the same considerations as above, we obtain the inequality where we restricted attention to the components $l \in \mathcal{S}$. \square

B.3 High-probability bounds for the spread-out spectrum case

In this subsection we provide a high-probability counterpart to the expectation estimate from [Prop. 3.2](#), where we showed $\mathbb{E} [\min_{j \leq n} u_j^\top Q u_j + v_j^\top R v_j] \sim \frac{\text{Tr}(S)}{n}$ when $\frac{\text{Tr}(S)}{\|S\|_F} \geq \Omega(\sqrt{\log n})$.

Proposition B.5. *Suppose Q, R are fixed and U, V are independently distributed uniformly on the set of $n \times n$ orthonormal matrices. Then, denoting $\mu = \min_{j \leq n} u_j^\top Q u_j + v_j^\top R v_j$,*

$$\begin{aligned} \forall 0 \leq \gamma \leq 1, \quad \mathbb{P} \left(\mu \geq \frac{\text{Tr}(S)}{2n} (1 - \gamma) \right) &\geq 1 - n e^{-\frac{\text{Tr}(S)^2}{4\|S\|_F^2} \gamma^2} - 2e^{-\frac{n}{8}} \\ \text{and } \forall \gamma \geq 0, \quad \mathbb{P} \left(\mu \leq \frac{4 \text{Tr}(S)}{n} (1 + \gamma) \right) &\geq \begin{cases} 1 - e^{-\frac{\text{Tr}(S)^2}{8\|S\|_F^2} \gamma^2} - 2e^{-\frac{n}{8}} & \text{if } \gamma \leq \frac{\|S\|_F^2}{\text{Tr}(S)\|S\|} \\ 1 - e^{-\frac{\text{Tr}(S)}{8\|S\|} \gamma} - 2e^{-\frac{n}{8}} & \text{otherwise} \end{cases} \end{aligned}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\|\cdot\|$ denotes the operator norm.

The remainder of this subsection is dedicated to proving the above proposition. Let as in [Sec. B.1](#)

$a_j = \begin{pmatrix} \sqrt{N_j} u_j \\ \sqrt{M_j} v_j \end{pmatrix}$ where $N_1, \dots, N_n, M_1, \dots, M_n \sim \chi_n^2$ i.i.d. and independent of U and V , and $\zeta_j = a_j^\top S a_j$ for $j \leq n$. Also denote $s_1 \geq \dots \geq s_r > 0 = s_{r+1} = \dots = s_{2n}$ the eigenvalues of S , and assume S diagonal w.l.o.g.

Lemma B.6. *For each j , $\mathbb{P} \left(\frac{1}{2n} \leq \frac{1}{N_j} \right) \geq 1 - e^{-\frac{n}{8}}$ and $\mathbb{P} \left(\frac{1}{N_j} \leq \frac{4}{n} \right) \geq 1 - e^{-\frac{n}{8}}$.*

Proof. Since $N_j \sim \chi_n^2$, we have the classical concentration bounds [[LM00](#), Equations (4.3), (4.4)]

$$\forall t > 0, \quad \mathbb{P} \left(N_j - n \leq 2\sqrt{nt} + 2t \right) \geq 1 - e^{-t} \quad \text{and} \quad \mathbb{P} \left(N_j - n \geq -2\sqrt{nt} \right) \geq 1 - e^{-t}.$$

Evaluating the first inequality at $t = \frac{n}{8}$ yields $N_j \leq n + \frac{n}{\sqrt{2}} + \frac{n}{4} \leq 2n$ with probability $\geq 1 - e^{-\frac{n}{8}}$. Evaluating the second inequality at $t = \frac{n}{8}$ yields $N_j \geq n - \frac{n}{\sqrt{2}} \geq \frac{n}{4}$ with probability $\geq 1 - e^{-\frac{n}{8}}$. \square

⁷And this approximation of $P(t)$ intuitively makes sense to do since our choice of $t^{(2)}$ was guided by the Ansatz that $\frac{2s_l t}{1+2s_l t} \approx 1$, i.e. $s_l t \gg 1$, for all $l \leq r$.

Letting the random variable $J = \arg \min_{j \leq n} u_j^\top Q u_j + v_j^\top R v_j$ that is only dependent on U and V , so independent of the N_j, M_j , we have

$$\begin{aligned} \mu &= \frac{1}{N_J} \cdot \sqrt{N_J} u_J^\top Q \sqrt{N_J} u_J + \frac{1}{M_J} \cdot \sqrt{M_J} v_J^\top R \sqrt{M_J} v_J \geq \left(\frac{1}{N_J} \wedge \frac{1}{M_J} \right) a_J^\top S a_J \\ &\geq \left(\frac{1}{N_J} \wedge \frac{1}{M_J} \right) \min_{j \leq n} a_j^\top S a_j \end{aligned}$$

and so by union bound, using that $N_J \sim N_1$ and $M_J \sim M_1$ by independence,

$$\forall c \geq 0, \mathbb{P}(\mu \geq c) \geq \mathbb{P}\left(\min_{j \leq n} a_j^\top S a_j \geq 2nc\right) - 2e^{-\frac{n}{8}}. \quad (\text{B.2})$$

Let for concision $\zeta_j = a_j^\top S a_j$ for each $j \leq n$. Recall from [Lemma B.2](#) that the moment-generating function of $\zeta_1 (\sim \zeta_2 \sim \dots \sim \zeta_n)$ is

$$\mathbb{E}e^{t\zeta_1} = \prod_{l=1}^r (1 - 2s_l t)^{-1/2} \quad \text{for all } t < \frac{1}{2 \max_l s_l}.$$

We can now use union bound and Chernoff's bound to lower-bound $\min_{j \leq n} \zeta_j$ with high probability. The derivation is essentially an instantiation of the general subexponential tail bound [[Wai19](#), Sec. 2.1.3] using our precise knowledge of the moment-generating function of the ζ_j .

Lemma B.7. *For any $0 \leq \gamma \leq 1$, we have*

$$\mathbb{P}\left(\min_{j \leq n} \zeta_j \geq \frac{\text{Tr}(S)}{2n}(1 - \gamma)\right) \geq 1 - ne^{-\frac{\text{Tr}(S)^2}{4\|S\|_F^2} \gamma^2}$$

where $\|\cdot\|_F$ denotes Frobenius norm, i.e., ℓ_2 -norm of the vector of eigenvalues.

Proof. By union bound, since the ζ_j are identically distributed, $\mathbb{P}(\min_j \zeta_j \leq x) \leq n\mathbb{P}(\zeta_1 \leq x)$. By Markov's inequality,

$$\begin{aligned} \mathbb{P}(\zeta_1 \leq x) &= \mathbb{P}(-\zeta_1 \geq -x) \leq \inf_{t>0} \frac{1}{e^{-tx}} \mathbb{E}e^{-t\zeta_1} = \inf_{t>0} e^{tx} \prod_{l=1}^r (1 + 2s_l t)^{-1/2} \\ \log \mathbb{P}(\zeta_1 \leq x) &\leq \inf_{t>0} tx - \frac{1}{2} \sum_{l=1}^r \log(1 + 2s_l t) =: \inf_{t>0} g_x(t). \end{aligned}$$

By calculating we find that $g'_x(t) > 0 \iff 2x > \sum_{l=1}^r \frac{2s_l t}{1+2s_l t}$. With the case $x \approx \mathbb{E}\zeta_1 = \text{Tr}(S) \ll r$ in mind, let us evaluate at $t_x^{(1)}$ defined by $2x = \sum_{l=1}^r 2s_l(1 - 2s_l t) = 2\text{Tr}(S) - 4\|S\|_F^2 t$, i.e., $t_x^{(1)} = \frac{\text{Tr}(S) - x}{2\|S\|_F^2}$. Assume henceforth that $x \leq \text{Tr}(S)$ so that $t_x^{(1)} \geq 0$. Using that $\log(1 + y) \geq y - \frac{1}{2}y^2$ for $y \geq 0$, we get

$$\begin{aligned} g_x(t_x^{(1)}) &= \frac{\text{Tr}(S) - x}{2\|S\|_F^2} x - \frac{1}{2} \sum_{l=1}^r \log\left(1 + 2s_l \frac{\text{Tr}(S) - x}{2\|S\|_F^2}\right) \\ &\leq \frac{\text{Tr}(S)x - x^2}{2\|S\|_F^2} - \frac{1}{2} \sum_{l=1}^r 2s_l \frac{\text{Tr}(S) - x}{2\|S\|_F^2} + \frac{1}{4} \sum_{l=1}^r 4s_l^2 \left(\frac{\text{Tr}(S) - x}{2\|S\|_F^2}\right)^2 \\ &= \frac{\text{Tr}(S)x - x^2}{2\|S\|_F^2} - \text{Tr}(S) \frac{\text{Tr}(S) - x}{2\|S\|_F^2} + \|S\|_F^2 \left(\frac{\text{Tr}(S) - x}{2\|S\|_F^2}\right)^2 \\ &= \frac{1}{2\|S\|_F^2} \left[\text{Tr}(S)x - x^2 - \text{Tr}(S)^2 + \text{Tr}(S)x + \frac{1}{2}(\text{Tr}(S) - x)^2 \right] \\ &= -\frac{1}{4\|S\|_F^2} [\text{Tr}(S) - x]^2. \end{aligned}$$

Hence we have shown

$$\forall x \leq \text{Tr}(S), \mathbb{P}\left(\min_j \zeta_j \leq x\right) \leq n\mathbb{P}(\zeta_1 \leq x) \leq n \exp\left(-\frac{1}{4\|S\|_F^2}[\text{Tr}(S) - x]^2\right)$$

and the announced bound follows by a change of variables $\gamma = 1 - \frac{x}{\text{Tr}(S)}$. \square

Combining (B.2) with Lemma B.7 yields the high-probability lower bound

$$\forall 0 \leq \gamma \leq 1, \quad \mathbb{P}\left(\mu \geq \frac{\text{Tr}(S)}{2n}(1 - \gamma)\right) \geq 1 - ne^{-\frac{\text{Tr}(S)^2}{4\|S\|_F^2}\gamma^2} - 2e^{-\frac{n}{8}}$$

which is exactly the lower bound of Prop. B.5.

For a high-probability upper bound on $\mu = \min_{j \leq n} w_j^\top S w_j$, in the regime of interest in this subsection it is sufficient to start from

$$\mu \leq w_1^\top S w_1 = \frac{1}{N_1} \cdot \sqrt{N_1} u_1^\top Q \sqrt{N_1} u_1 + \frac{1}{M_1} \cdot \sqrt{M_1} v_1^\top R \sqrt{M_1} v_1 \leq \left(\frac{1}{N_1} \vee \frac{1}{M_1}\right) a_1^\top S a_1$$

so that, by union bound and Lemma B.6,

$$\forall C \geq 0, \mathbb{P}(\mu \leq C) \geq \mathbb{P}\left(a_1^\top S a_1 \leq \frac{n}{4}C\right) - 2e^{-\frac{n}{8}}. \quad (\text{B.3})$$

Hence it is sufficient to upper-bound $\zeta_1 = a_1^\top S a_1$ with high probability.

Lemma B.8. *For any $\varepsilon \geq 0$, we have*

$$\mathbb{P}(\zeta_1 \leq \text{Tr}(S) + \varepsilon) \geq \begin{cases} 1 - e^{-\frac{\varepsilon^2}{8\|S\|_F^2}} & \text{if } 0 \leq \varepsilon \leq \frac{\|S\|_F^2}{\max_l s_l} \\ 1 - e^{-\frac{\varepsilon}{8 \max_l s_l}} & \text{otherwise.} \end{cases}$$

Proof. $\zeta_1 = \sum_{l=1}^r s_l a_1[l]^2$ is subexponential with parameters $\left(\sqrt{\sum_{l=1}^r (2 \cdot s_l)^2}, \max_{l \leq r} 4 \cdot s_l\right)$ as a linear combinations of independent χ^2 random variables $a_j[l]^2$ which are subexponential with parameters $(2, 4)$ [Wai19, Sec. 2.1.3]. The announced subexponential tail bound follows by a direct application of [Wai19, Prop. 2.9]. \square

Combining (B.3) with Lemma B.8 yields the high-probability upper bound

$$\forall \varepsilon \geq 0, \quad \mathbb{P}\left(\mu \leq 4 \frac{\text{Tr}(S) + \varepsilon}{n}\right) \geq \begin{cases} 1 - e^{-\frac{\varepsilon^2}{8\|S\|_F^2}} - 2e^{-\frac{n}{8}} & \text{if } \varepsilon \leq \frac{\|S\|_F^2}{\max_l s_l} \\ 1 - e^{-\frac{\varepsilon}{8 \max_l s_l}} - 2e^{-\frac{n}{8}} & \text{otherwise} \end{cases}$$

$$\text{and so } \forall \gamma \geq 0, \quad \mathbb{P}\left(\mu \leq \frac{4 \text{Tr}(S)}{n}(1 + \gamma)\right) \geq \begin{cases} 1 - e^{-\frac{\text{Tr}(S)^2}{8\|S\|_F^2}\gamma^2} - 2e^{-\frac{n}{8}} & \text{if } \gamma \leq \frac{\|S\|_F^2}{\text{Tr}(S)[\max_l s_l]} \\ 1 - e^{-\frac{\text{Tr}(S)\gamma}{8 \max_l s_l}} - 2e^{-\frac{n}{8}} & \text{otherwise} \end{cases}$$

by the change of variables $\text{Tr}(S)\gamma = \varepsilon$, which is exactly the upper bound of Prop. B.5.

B.4 High-probability bounds for the sparse spectrum case

In this subsection we provide a high-probability counterpart to the expectation lower estimate from Prop. 3.3, where we showed $\mathbb{E}[\min_{j \leq n} u_j^\top Q u_j + v_j^\top R v_j] \gtrsim n^{-\frac{2}{r}-1}$ when $n \rightarrow \infty$ with $r = \text{rank}(S)$ and $s \in \mathbb{R}^r$ fixed.

Proposition B.9. *Suppose Q, R are fixed and U, V are independently distributed uniformly on the set of $n \times n$ orthonormal matrices. Let $s_1 \geq \dots \geq s_r > 0 = s_{r+1} = \dots = s_{2n}$ the eigenvalues of S . Then denoting $\mu = \min_{j \leq n} u_j^\top Q u_j + v_j^\top R v_j$, for any $\mathcal{S} \subset \{1, \dots, r\}$,*

$$\forall 0 \leq \gamma \leq 1, \quad \mathbb{P}\left(\mu \geq \frac{1}{2e} \frac{|\mathcal{S}|}{n} n^{-\frac{2}{|\mathcal{S}|}} \left[\prod_{l \in \mathcal{S}} s_l\right]^{\frac{1}{|\mathcal{S}|}} \cdot (1 - \gamma)\right) \geq 1 - e^{-\frac{|\mathcal{S}|}{2}\gamma} - 2e^{-\frac{n}{8}}.$$

The remainder of this subsection is dedicated to proving the above proposition. As we are exactly in the same setting as for [Prop. B.5](#), all the lemmas of [Sec. B.3](#) apply. We also reuse notations from that subsection: $a_j = \left(\frac{\sqrt{N_j} u_j}{\sqrt{M_j} v_j} \right)$ where $N_1, \dots, N_n, M_1, \dots, M_n \sim \chi_n^2$ i.i.d. and independent of U and V , and $\zeta_j = a_j^\top S a_j$ for $j \leq n$. We also assume S diagonal w.l.o.g.

Given [\(B.2\)](#), in order to prove the proposition it suffices to derive a different lower tail bound for $\min_{j \leq n} \zeta_j$ that is more adapted to the sparse spectrum case.

Lemma B.10. *Let any $S \subset \{1, \dots, r\}$, $r_S = |S|$ and $G_S = [\prod_{l \in S} s_l]^{1/r_S}$. Then*

$$\forall \gamma \in \mathbb{R}, \quad \mathbb{P} \left(\min_{j \leq n} \zeta_j \geq \frac{r_S}{e} n^{-\frac{2}{r_S}} \left[\prod_{l \in S} s_l \right]^{\frac{1}{r_S}} \cdot (1 - \gamma) \right) \geq 1 - e^{-\frac{r_S}{2} \gamma}.$$

Proof. Let any $x \in \mathbb{R}$. By Markov's inequality,

$$\mathbb{P} \left(\min_j \zeta_j \leq x \right) = \mathbb{P} \left(\max_j (-\zeta_j) \geq -x \right) \leq \inf_{t>0} \frac{1}{e^{-tx}} \mathbb{E} e^{t \max_j (-\zeta_j)} = \inf_{t>0} e^{tx} \cdot \mathbb{E} \max_j e^{t(-\zeta_j)}.$$

Now for any $\alpha \geq 1$, by monotonicity and concavity of $y \mapsto y^{1/\alpha}$,

$$\begin{aligned} \mathbb{E} \max_{j \leq n} e^{t(-\zeta_j)} &= \mathbb{E} \left[\left(\max_{j \leq n} e^{\alpha t(-\zeta_j)} \right)^{1/\alpha} \right] \leq \mathbb{E} \left[\left(\sum_{j \leq n} e^{\alpha t(-\zeta_j)} \right)^{1/\alpha} \right] \leq \left[\mathbb{E} \sum_{j \leq n} e^{\alpha t(-\zeta_j)} \right]^{1/\alpha} \\ &= \left[n \mathbb{E} e^{\alpha t(-\zeta_1)} \right]^{1/\alpha}. \end{aligned}$$

Hence, taking the infimum over α and substituting into the above inequality,⁸

$$\begin{aligned} \mathbb{P} \left(\min_j \zeta_j \leq x \right) &\leq \inf_{t>0} e^{tx} \cdot \inf_{\alpha \geq 1} \left[n \mathbb{E} e^{\alpha t(-\zeta_1)} \right]^{1/\alpha} \\ \log \mathbb{P} \left(\min_j \zeta_j \leq x \right) &\leq \inf_{t>0} tx + \inf_{\alpha \geq 1} \frac{1}{\alpha} \left(\log n + \log \mathbb{E} e^{\alpha t(-\zeta_1)} \right) \\ &= \inf_{t>0} t \left[x + \inf_{\beta \geq t} \frac{1}{\beta} \left(\log n + \log \mathbb{E} e^{\beta(-\zeta_1)} \right) \right]. \end{aligned}$$

Let, as in the proof of [Lemma B.4](#),

$$g(\beta) := \frac{1}{\beta} \left(\log n + \log \mathbb{E} e^{-\beta \zeta_1} \right) = \frac{1}{\beta} \left(\log n - \frac{1}{2} \sum_{l=1}^r \log(1 + 2s_l \beta) \right) \quad \text{and} \quad u := \frac{e}{2} n^{\frac{2}{r}} \left[\prod_{l=1}^r s_l \right]^{-\frac{1}{r}}.$$

One can check, by lower-bounding $\log(1 + 2s_l u)$ by $\log(2s_l u)$ for each term, that $g(u) \leq \frac{1}{u}(-r/2)$. Hence,

$$\begin{aligned} \log \mathbb{P} \left(\min_j \zeta_j \leq x \right) &\leq \inf_{t>0} t \left[x + \inf_{\beta \geq t} g(\beta) \right] \leq \inf_{0 < t \leq u} t(x + g(u)) \\ &\leq u(x + g(u)) \leq ux - \frac{r}{2} = \frac{r}{2} \cdot \left(\frac{u}{r/2} x - 1 \right). \end{aligned}$$

Hence we have shown that for all $x \in \mathbb{R}$,

$$\mathbb{P} \left(\min_j \zeta_j \leq x \right) \leq \exp \left(-\frac{r}{2} \left(1 - \frac{u}{r/2} x \right) \right).$$

⁸In fact one can check, by inverting \inf_t and \inf_α and optimizing first over t , that the infimum over the joint (t, α) is always attained at some $(t^*, 1)$; that is, the true optimum is always at $\alpha = 1$. What introducing the extra degree of freedom α affords us is a way to obtain a tractable upper bound, whereas upper-bounding the \inf_t with $\alpha = 1$ fixed seems more difficult.

The inequality of the lemma with $\mathcal{S} = \{1, \dots, r\}$ follows by the change of variable $\gamma = 1 - \frac{u}{r/2}x$ and substituting the expression of u .

The inequality for arbitrary $\mathcal{S} \subset \{1, \dots, r\}$ follows by exactly the same considerations as above applied to

$$g_{\mathcal{S}}(\beta) = \frac{1}{\beta} \left(\log n - \frac{1}{2} \sum_{l \in \mathcal{S}} \log(1 + 2s_l \beta) \right) \quad \text{and} \quad u_{\mathcal{S}} = \frac{e}{2} n^{\frac{2}{r_{\mathcal{S}}}} \left[\prod_{l \in \mathcal{S}} s_l \right]^{-\frac{1}{r_{\mathcal{S}}}}$$

instead of g and u , noting that $g(\beta) \leq g_{\mathcal{S}}(\beta)$ for any $\beta > 0$. \square

Combining (B.2) with Lemma B.10 yields the high-probability lower bound

$$\forall \gamma \in \mathbb{R}, \quad \mathbb{P} \left(\mu \geq \frac{1}{2n} \frac{|\mathcal{S}|}{e} n^{-\frac{2}{|\mathcal{S}|}} \left[\prod_{l \in \mathcal{S}} s_l \right]^{\frac{1}{|\mathcal{S}|}} \cdot (1 - \gamma) \right) \geq 1 - e^{-\frac{|\mathcal{S}|}{2} \gamma} - 2e^{-\frac{n}{8}}$$

for any $\mathcal{S} \subset \{1, \dots, r\}$, which is exactly the inequality of Prop. B.5.

C A symmetric expression for Alt-GDA

This appendix contains results stated in Sec. 3.2 and used in Sec. D.3 and that may be of independent interest.

As announced in Sec. 3.2, we analyze a symmetrized formulation of Alt-GDA, whose definition we recall here for ease of reference: let $(x^0, y^{1/2}) \in \mathbb{R}^d \times \mathbb{R}^d$ and

$$\begin{cases} \forall k \in \mathbb{N}, & x^{k+1} = x^k - \eta \nabla_x f(x^k, y^{k+1/2}) \\ \forall k \in \mathbb{N} + 1/2, & y^{k+1} = y^k + \eta \nabla_y f(x^{k+1/2}, y^k) \end{cases} \quad \text{and} \quad \begin{cases} \forall k \in \mathbb{N}, & x^{k+1/2} = \frac{x^{k+1} + x^k}{2} \\ \forall k \in \mathbb{N} + 1/2, & y^{k+1/2} = \frac{y^{k+1} + y^k}{2}. \end{cases} \quad (\text{C.1})$$

Note that this symmetrized formulation is indeed equivalent to standard one (used e.g. in [Gid+19; BGP20; ZWLG22; GLWM22]), via the correspondence $\forall k \in \mathbb{N}$, $\tilde{x}^k = x^k$ and $\tilde{y}^k = y^{k+1/2}$. We emphasize that, as one can easily check from the definition, the symmetrized iterates $(z^k)_k = (x^k, y^k)_k$ converge if and only if the classical ones $(\tilde{z}^k)_k = (x^k, y^{k+1/2})_{k \in \mathbb{N}}$ do, and if they converge exponentially, then they have the same convergence rate.

The following proposition shows that the evolution of the symmetrized Alt-GDA iterates can be expressed approximately in terms only of the (skewed) gradient field $g(z)$.

Proposition C.1. *Assume $\eta \leq \|\nabla_{xx}^2 f\|_{\infty}^{-1} \wedge \|\nabla_{yy}^2 f\|_{\infty}^{-1}$. Then the symmetrized Alt-GDA iterates defined by (C.1) satisfy*

$$\forall k \in \mathbb{N} \cup (\mathbb{N} + 1/2), \quad z^{k+1} = z^k - \eta g(z^k) + \frac{\eta^2}{2} A(z^k) g(z^k) + O(\eta^3 \|g(z^k)\|)$$

where $A(z) = \begin{bmatrix} 0 & \nabla_{xy}^2 f \\ -\nabla_{xy}^2 f^{\top} & 0 \end{bmatrix} (z)$. More precisely the “ $O(\cdot)$ ” term is a vector with norm at most $10\eta^3 \|g(z^k)\| \left(\|\nabla^3 f\|_{\infty} (1 + \|\nabla_{xy}^2 f\|_{\infty}^2 \eta^2) + \|\nabla^2 f\|_{\infty} \|g(z^k)\| \right)$.

Remark C.1 (High-resolution ODE for the symmetrized iterates). The symmetrized update operator with the $O(\eta^3)$ term neglected, $z \mapsto z - \eta g(z) + \frac{\eta^2}{2} A(z) g(z)$, leads to a different high-resolution ODE than the one derived by [GLWM22, Eq. (12)] for the standard formulation. The one derived in that reference is

$$\frac{d\tilde{z}}{dt} = -g(\tilde{z}) - \frac{\eta}{2} \begin{bmatrix} Q & P \\ P^{\top} & -R \end{bmatrix} (\tilde{z}) g(\tilde{z}) \quad \text{where} \quad \begin{bmatrix} Q & P \\ -P^{\top} & R \end{bmatrix} (z) := \nabla g(z)$$

while the high-resolution ODE corresponding to the symmetrized update $z^{k+1} \approx T(z^k)$ can be shown – simply by applying the high-resolution ODE for explicit Euler steps [Lu22, Eq. (20)] to the vector field $-g(z) + \frac{\eta}{2} A(z) g(z)$ – to be

$$\frac{dz}{dt} = -g(z) - \frac{\eta}{2} M(z) g(z) + \frac{\eta}{2} A(z) g(z) = -g(z) - \frac{\eta}{2} \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} (z) g(z).$$

There is no contradiction here, as the classical Alt-GDA iterates \tilde{z}^k are not identical pointwise to the symmetrized ones z^k ; instead the correspondence involves a time-shift of $+1/2$ for one of the variables. Informally, the former high-resolution ODE tracks the evolution of the Alt-GDA iterates using a piecewise-constant (forward in time) interpolation scheme, with the arbitrary choice of measuring at time-steps where y was updated more recently than x , and the latter tracks the same iterates but using a piecewise-linear (trapezoidal) interpolation scheme.

Proof. As usual in this paper we let for concision $z^k = (x^k, y^k)$ for all $k \in \mathbb{N} \cup (\mathbb{N} + 1/2)$. Denote $L_{xx} = \|\nabla_{xx}^2 f\|_\infty$, $L_{yy} = \|\nabla_{yy}^2 f\|_\infty$, $L_{xy} = \|\nabla_{xy}^2 f\|_\infty$, $L_2 = L_{xx} \vee L_{yy} \vee L_{xy}$ and $L_3 = \|\nabla^3 f\|_\infty$. Throughout this proof we will write ε or ε' to denote elements of $[-1, 1]$ or of the unit ball of \mathbb{R}^n or \mathbb{R}^m or \mathbb{R}^{n+m} , and that may change from line to line.

Let $k \in \mathbb{N}$. By Taylor expansion with remainder in Lagrange form,

$$\begin{aligned} x^{k+1} - x^k &= -\eta \nabla_x f(x^k, y^{k+1/2}) \\ &= -\eta \left[\nabla_x f(z^k) + \nabla_{xy}^2 f(z^k) \cdot (y^{k+1/2} - y^k) + \varepsilon \left(\frac{L_3}{2} \|y^{k+1/2} - y^k\|^2 \right) \right]. \end{aligned}$$

Now

$$\begin{aligned} y^{k+1/2} - y^k &= y^{k+1/2} - \frac{y^{k+1/2} + y^{k-1/2}}{2} = \frac{y^{k+1/2} - y^{k-1/2}}{2} \\ \text{and } y^{k+1/2} - y^{k-1/2} &= \eta \nabla_y f(x^k, y^{k-1/2}) = \eta \nabla_y f(z^k) + \varepsilon \left(\eta L_{yy} \|y^k - y^{k-1/2}\| \right) \\ &= \eta \nabla_y f(z^k) + \varepsilon \left(\eta L_{yy} \frac{1}{2} \|y^{k+1/2} - y^{k-1/2}\| \right) \\ \text{since } y^{k-1/2} - y^k &= y^{k-1/2} - \frac{y^{k+1/2} + y^{k-1/2}}{2} = -\frac{y^{k+1/2} - y^{k-1/2}}{2} = -(y^{k+1/2} - y^k). \end{aligned}$$

Moreover, the above computation also shows that

$$\begin{aligned} \|y^{k+1/2} - y^{k-1/2}\| &\leq \eta \|\nabla_y f(z^k)\| + \eta L_{yy} \frac{1}{2} \|y^{k+1/2} - y^{k-1/2}\| \\ \text{and so } \|y^{k+1/2} - y^{k-1/2}\| &\leq \frac{1}{1 - \eta L_{yy}/2} \eta \|\nabla_y f(z^k)\| \end{aligned} \quad (\text{C.2})$$

for any $\eta \leq 2/L_{yy}$. In particular, since we assume $\eta \leq 1/L_{yy}$ then $\|y^{k+1/2} - y^{k-1/2}\| \leq 2\eta \|\nabla_y f(z^k)\|$. Thus, using this last fact to express the error terms in terms of $\|g(z^k)\| = \|\nabla f(z^k)\|$, by substituting into the Taylor expansion of $x^{k+1} - x^k$ we obtain

$$\begin{aligned} \forall k \in \mathbb{N}, x^{k+1} - x^k &= -\eta \nabla_x f(z^k) - \frac{1}{2} \eta^2 \nabla_{xy}^2 f(z^k) \cdot \nabla_y f(z^k) \\ &\quad + \varepsilon \left(\eta^2 L_{xy} L_{yy} \frac{1}{2} \|y^{k+1/2} - y^{k-1/2}\| \right) + \varepsilon' \left(\eta \frac{L_3}{8} \|y^{k+1/2} - y^{k-1/2}\|^2 \right) \\ &= -\eta \nabla_x f(z^k) - \frac{1}{2} \eta^2 \nabla_{xy}^2 f(z^k) \cdot \nabla_y f(z^k) + \eta^3 \varepsilon \left(L_{xy} L_{yy} \|g(z^k)\| + \frac{L_3}{2} \|g(z^k)\|^2 \right). \end{aligned} \quad (\text{C.3})$$

By the symmetric calculations, we have

$$\begin{aligned} \forall k \in \mathbb{N} + 1/2, y^{k+1} - y^k &= \eta \nabla_y f(z^k) - \frac{1}{2} \eta^2 \nabla_{yx}^2 f(z^k) \nabla_x f(z^k) \\ &\quad + \eta^3 \varepsilon \left(L_{xy} L_{yy} \|g(z^k)\| + \frac{L_3}{2} \|g(z^k)\|^2 \right). \end{aligned} \quad (\text{C.4})$$

Now let us compute an expansion for the increments between non-updating time-steps. Let $k \in \mathbb{N}$ and let us compute $y^{k+1} - y^k$:

$$\begin{aligned}
y^{k+1} - y^k &= \frac{y^{k+\frac{3}{2}} + y^{k+1/2}}{2} - \frac{y^{k+1/2} + y^{k-1/2}}{2} = \frac{1}{2} \left(y^{k+\frac{3}{2}} - y^{k-1/2} \right) \\
&= \frac{1}{2} \left(\eta \nabla_y f \left(x^{k+1}, y^{k+1/2} \right) + \eta \nabla_y f \left(x^k, y^{k-1/2} \right) \right) \\
&= \frac{1}{2} \eta \left[2 \nabla_y f(z^k) + \nabla_{yx}^2 f(z^k) \cdot (x^{k+1} - x^k) + \nabla_{yy}^2 f(z^k) \cdot (y^{k+1/2} - y^k) + \nabla_{yy}^2 f(z^k) \cdot (y^{k-1/2} - y^k) \right] \\
&\quad + \varepsilon \left(\eta \frac{L_3}{2} \|x^{k+1} - x^k\|^2 + \eta \frac{L_3}{2} \|y^{k+1/2} - y^k\|^2 \right) + \varepsilon' \left(\eta \frac{L_3}{2} \|y^k - y^{k-1/2}\|^2 \right) \\
&= \eta \nabla_y f(z^k) + \frac{1}{2} \eta \nabla_{yx}^2 f(z^k) \cdot (x^{k+1} - x^k) + \frac{1}{2} \eta \nabla_{yy}^2 f(z^k) \cdot \underbrace{(y^{k+1/2} + y^{k-1/2} - 2y^k)}_{=0} \\
&\quad + \varepsilon \left(\eta \frac{L_3}{2} \|x^{k+1} - x^k\|^2 + \eta \frac{L_3}{4} \|y^{k+1/2} - y^{k-1/2}\|^2 \right) \\
&= \eta \nabla_y f(z^k) + \frac{1}{2} \eta \nabla_{yx}^2 f(z^k) \cdot (x^{k+1} - x^k) + \varepsilon \left(\eta^3 L_3 (2 + L_{xy}^2 \eta^2) \|g(z^k)\|^2 \right).
\end{aligned}$$

Here in order to bound the error term in $\|x^{k+1} - x^k\|$ we used that

$$\begin{aligned}
\|x^{k+1} - x^k\| &= \eta \left\| \nabla_x f(x^k, y^{k+1/2}) \right\| \leq \eta \left\| \nabla_x f(z^k) \right\| + \eta L_{xy} \|y^{k+1/2} - y^k\| \\
&= \eta \left\| \nabla_x f(z^k) \right\| + \frac{1}{2} \eta L_{xy} \|y^{k+1/2} - y^{k-1/2}\| \leq \eta (1 + L_{xy} \eta) \|g(z^k)\| \\
\|x^{k+1} - x^k\| &\leq \eta^2 (2 + 2L_{xy}^2 \eta^2) \|g(z^k)\|^2
\end{aligned}$$

since $\|y^{k+1/2} - y^{k-1/2}\| \leq 2\eta \|\nabla_y f(z^k)\|$ as shown previously (C.2). Hence, substituting the expansion of $x^{k+1} - x^k$ from the previous paragraph, we have

$$\begin{aligned}
\forall k \in \mathbb{N}, y^{k+1} - y^k &= \eta \nabla_y f(z^k) + \frac{1}{2} \eta \nabla_{yx}^2 f(z^k) \cdot \left[-\eta \nabla_x f(z^k) - \frac{1}{2} \eta^2 \nabla_{xy}^2 f(z^k) \cdot \nabla_y f(z^k) \right] \\
&\quad + \eta^3 \varepsilon \left(L_3 (2 + L_{xy}^2 \eta^2) \|g(z^k)\|^2 + L_{xy} L_{yy} \|g(z^k)\| + \frac{L_3}{2} \|g(z^k)\|^2 \right) \\
&= \eta \nabla_y f(z^k) - \frac{1}{2} \eta^2 \nabla_{yx}^2 f(z^k) \cdot \nabla_x f(z^k) + \eta^3 \varepsilon \left(L_3 \left(\frac{5}{2} + L_{xy}^2 \eta^2 \right) \|g(z^k)\|^2 + \frac{5}{4} L_2^2 \|g(z^k)\| \right). \tag{C.5}
\end{aligned}$$

By the symmetric calculations, we have

$$\begin{aligned}
\forall k \in \mathbb{N} + 1/2, x^{k+1} - x^k &= -\eta \nabla_x f(z^k) - \frac{1}{2} \eta^2 \nabla_{xy}^2 f(z^k) \cdot \nabla_y f(z^k) \\
&\quad + \eta^3 \varepsilon \left(L_3 \left(\frac{5}{2} + L_{xy}^2 \eta^2 \right) \|g(z^k)\|^2 + \frac{5}{4} L_2^2 \|g(z^k)\| \right). \tag{C.6}
\end{aligned}$$

The announced expansion for $z^{k+1} - z^k$, both at time-steps $k \in \mathbb{N}$ and $k \in \mathbb{N} + 1/2$, follows immediately from (C.3), (C.4), (C.5) and (C.6). \square

The above Prop. C.1 is already a good indication of the fact that local convergence of Alt-GDA can be analyzed via the spectrum of $\nabla \left[\text{id} - \eta g + \frac{\eta^2}{2} A \cdot g \right] (z^*) = I - \eta M + \frac{\eta^2}{2} AM$. In fact, since the error term in the above proposition is $O(\eta^3 \|g(z^k)\|)$, one can follow the potential-based approach developed in [Lu22, Sec. 4] (via continuous-time) or [WC22, Sec. H] (for a special case), to lower-bound the decrease of $\|g(z^k)\|$ or $\|z^k - z^*\|$ at each iteration. However that approach may require η to be smaller than actually necessary. The following proposition offers a more direct path to linking convergence of Alt-GDA to spectral properties of the aforementioned matrix.

Proposition C.2. Assume $\eta \leq \|\nabla_{xx}^2 f\|_\infty^{-1} \wedge \|\nabla_{yy}^2 f\|_\infty^{-1}$. Consider $z^k = (x^k, y^k)_{k \in \mathbb{N} \cup (\mathbb{N}+1/2)}$ the symmetrized Alt-GDA iterates defined by (C.1).

For any $k \in \mathbb{N} \cup (\mathbb{N}+1/2)$, $z^{k+1/2}$ is entirely determined by z^k . More precisely, there exist well-defined operators $T_{xy}^{1/2}, T_{yx}^{1/2}$ such that
$$\begin{cases} \forall k \in \mathbb{N}, & z^{k+1/2} = T_{xy}^{1/2}(z^k) \\ \forall k \in \mathbb{N} + 1/2, & z^{k+1/2} = T_{yx}^{1/2}(z^k). \end{cases}$$

Moreover, for z^* such that $g(z^*) = 0$, it holds $T_{xy}^{1/2}(z^*) = T_{yx}^{1/2}(z^*) = z^*$ and

$$\nabla \left[T_{xy}^{1/2} \circ T_{yx}^{1/2} \right] (z^*), \quad \nabla \left[T_{yx}^{1/2} \circ T_{xy}^{1/2} \right] (z^*) = I - \eta M + \frac{\eta^2}{2} AM + O(\eta^3).$$

More precisely the “ $O(\cdot)$ ” term is a matrix with operator norm – and so spectral radius – at most $2\eta^3 \|A\| (\|A\| \vee \|S\|)^2$, where $\|\cdot\|$ denotes operator norm.

Thanks to the above proposition, one can analyze the local convergence of the integer-time-step iterates $(z^k)_{k \in \mathbb{N}}$, say, by applying the usual analysis to its well-defined update operator $T_{yx}^{1/2} \circ T_{xy}^{1/2}$; that is to say, by bounding the spectral radius of its Jacobian at z^* .

For comparison, the update operator \tilde{T}_{xy} for the standard formulation is given by $\tilde{z}^{k+1} = \begin{pmatrix} \tilde{x}^k - \nabla_x f(\tilde{x}^k, \tilde{y}^k) \\ \tilde{y}^k + \nabla_y f(\tilde{x}^{k+1}, \tilde{y}^k) \end{pmatrix}$. Denoting $M = \begin{bmatrix} Q & P \\ -P^\top & R \end{bmatrix}$ as usual the Jacobian of g at a fixed point z^* , the Jacobian of \tilde{T}_{xy} at z^* writes [ZWL22, Sec. A.3]

$$\nabla \tilde{T}_{xy}(z^*) = \begin{bmatrix} I & 0 \\ \eta P^\top & I - \eta R \end{bmatrix} \begin{bmatrix} I - \eta Q & -\eta P \\ 0 & I \end{bmatrix} = \begin{bmatrix} I - \eta Q & -\eta P \\ \eta P^\top - \eta^2 P^\top Q & I - \eta R - \eta^2 P^\top P \end{bmatrix}$$

which cannot be written only in terms of the symmetric and antisymmetric parts of M . Interestingly, in order to heuristically obtain an expression for a “Jacobian” that is symmetric in the x and y players, a natural idea is to simply consider the average

$$\frac{\nabla T_{xy}(z^*) + \nabla T_{yx}(z^*)}{2} = \begin{bmatrix} I - \eta Q - \frac{\eta^2}{2} P P^\top & -\eta P + \frac{\eta^2}{2} P R \\ \eta P^\top - \frac{\eta^2}{2} P^\top Q & I - \eta R - \frac{\eta^2}{2} P^\top P \end{bmatrix} = I - \eta M + \frac{\eta^2}{2} AM,$$

which yields the same matrix as in Prop. C.2 (ignoring the $O(\eta^3)$ term).

Proof. Denote $L_{xx} = \|\nabla_{xx}^2 f\|_\infty$, $L_{yy} = \|\nabla_{yy}^2 f\|_\infty$, $L_{xy} = \|\nabla_{xy}^2 f\|_\infty$, $L_2 = L_{xx} \vee L_{yy} \vee L_{xy}$ and $L_3 = \|\nabla^3 f\|_\infty$.

Let $k \in \mathbb{N}$. We have by definition

$$\begin{cases} x^{k+1/2} = \frac{x^k - \eta \nabla_x f(x^k, y^{k+1/2}) + x^k}{2} \\ y^{k+1/2} = y^{k-1/2} + \eta \nabla_y f(x^k, y^{k-1/2}) \\ y^{k-1/2} = y^k - (y^{k+1/2} - y^k) \end{cases} \quad \text{and so} \quad \begin{cases} x^{k+1/2} = x^k - \frac{1}{2} \eta \nabla_x f(x^k, y^{k+1/2}) \\ y^{k+1/2} = 2y^k - y^{k+1/2} + \eta \nabla_y f(x^k, 2y^k - y^{k+1/2}) \end{cases}$$

i.e.,
$$\begin{cases} x^{k+1/2} = x^k - \frac{1}{2} \eta \nabla_x f(x^k, y^{k+1/2}) \\ y^{k+1/2} = y^k + \frac{1}{2} \eta \nabla_y f(x^k, 2y^k - y^{k+1/2}). \end{cases}$$

It remains to check that the second equation determines $y^{k+1/2}$ completely for any given x^k and y^k . And indeed, thanks to our assumption that $\frac{\eta}{2} L_{yy} < 1$, the mapping $y \mapsto y^k + \frac{1}{2} \eta \nabla_y f(x^k, 2y^k - y)$ is a contraction, so in particular it has a unique fixed point. This shows that $z^{k+1/2}$ is entirely determined by z^k for $k \in \mathbb{N}$, and the case $k \in \mathbb{N} + 1/2$ is similar.

Moreover the above also shows that $T_{xy}^{1/2}$ is characterized by

$$(x^{k+1/2}, y^{k+1/2}) = T_{xy}^{1/2}(x^k, y^k) \iff \begin{cases} x^{k+1/2} = x^k - \frac{1}{2} \eta \nabla_x f(x^k, y^{k+1/2}) \\ y^{k+1/2} = y^k + \frac{1}{2} \eta \nabla_y f(x^k, 2y^k - y^{k+1/2}) \end{cases}$$

and symmetrically for $T_{yx}^{1/2}$. So by substituting, we indeed find that $T_{xy}^{1/2}(x^*, y^*) = T_{yx}^{1/2}(x^*, y^*) = (x^*, y^*)$. In order to compute the Jacobian, let any small $\delta = (\delta_x, \delta_y)$, and consider $T_{xy}^{1/2}(z^* + \delta) -$

$T_{xy}^{1/2}(z^*) =: \Delta = (\Delta_x, \Delta_y)$. Then, denoting y' the unique solution of $y' = y^* + \delta_y + \frac{1}{2}\eta\nabla_y f(x^* + \delta_x, 2y^* + 2\delta_y - y')$, we have

$$\begin{aligned} y' - y^* &= \delta_y + \frac{\eta}{2} [\nabla_y f(x^* + \delta_x, 2y^* + 2\delta_y - y') - \nabla_y f(z^*)] \\ \|y' - y^*\| &\leq \|\delta_y\| + \frac{\eta}{2} [L_{xy} \|\delta_x\| + L_{yy} (2\|\delta_y\| + \|y^* - y'\|)] \\ \|y' - y^*\| &\leq \frac{1}{1 - \eta L_{yy}/2} \cdot \left(\|\delta_y\| + \frac{\eta}{2} [L_{xy} \|\delta_x\| + L_{yy} 2\|\delta_y\|] \right) = O(\|\delta_x\| + \|\delta_y\|) \end{aligned}$$

since we assume that $\eta L_{yy}/2 \leq 1/2$, and so

$$\begin{aligned} y' - y^* &= \delta_y + \frac{\eta}{2} \left[\nabla_{yx}^2 f(z^*) \delta_x + \nabla_{yy}^2 f(z^*) (2\delta_y + y^* - y') \right] + O(\|\delta_x\|^2 + \|\delta_y\|^2) \\ y' - y^* &= \left[I + \frac{\eta}{2} \nabla_{yy}^2 f(z^*) \right]^{-1} \left(\delta_y + \frac{\eta}{2} [\nabla_{yx}^2 f(z^*) \delta_x + 2\nabla_{yy}^2 f(z^*) \delta_y] \right) + O(\|\delta_x\|^2 + \|\delta_y\|^2). \end{aligned}$$

This directly gives an expansion for $\Delta_y = y' - y^*$, and we have

$$\begin{aligned} \Delta_x &= \delta_x - \frac{1}{2}\eta [\nabla_x f(x^* + \delta_x, y') - \nabla_x f(z^*)] \\ &= \delta_x - \frac{1}{2}\eta [\nabla_{xx}^2 f(z^*) \delta_x + \nabla_{xy}^2 f(z^*) (y' - y^*)] + O(\|\delta_x\|^2 + \|\delta_y\|^2) \\ &= \delta_x - \frac{\eta}{2} \nabla_{xx}^2 f(z^*) \delta_x - \frac{\eta}{2} \nabla_{xy}^2 f(z^*) \left[I + \frac{\eta}{2} \nabla_{yy}^2 f(z^*) \right]^{-1} \left(\delta_y + \frac{\eta}{2} [\nabla_{yx}^2 f(z^*) \delta_x + 2\nabla_{yy}^2 f(z^*) \delta_y] \right) \\ &\quad + O(\|\delta_x\|^2 + \|\delta_y\|^2) \\ &= \left\{ I - \frac{\eta}{2} \nabla_{xx}^2 f(z^*) - \frac{\eta^2}{4} \nabla_{xy}^2 f(z^*) \left[I + \frac{\eta}{2} \nabla_{yy}^2 f(z^*) \right]^{-1} \nabla_{yx}^2 f(z^*) \right\} \delta_x \\ &\quad + \left\{ -\frac{\eta}{2} \nabla_{xy}^2 f(z^*) \left[I + \frac{\eta}{2} \nabla_{yy}^2 f(z^*) \right]^{-1} (I + \eta \nabla_{yy}^2 f(z^*)) \right\} \delta_y + O(\|\delta_x\|^2 + \|\delta_y\|^2). \end{aligned}$$

Writing for concision $\begin{bmatrix} \nabla_{xx}^2 f & \nabla_{xy}^2 f \\ \nabla_{yx}^2 f & \nabla_{yy}^2 f \end{bmatrix} (z^*) = \begin{bmatrix} Q & P \\ P^\top & -R \end{bmatrix}$ as usual in this paper, the above expansions write

$$\begin{aligned} \Delta_x &= \left\{ I - \frac{\eta}{2} Q - \frac{\eta^2}{4} P \left[I - \frac{\eta}{2} R \right]^{-1} P^\top \right\} \delta_x + \left\{ -\frac{\eta}{2} P \left[I - \frac{\eta}{2} R \right]^{-1} (I - \eta R) \right\} \delta_y + O(\|\delta_x\|^2 + \|\delta_y\|^2) \\ \Delta_y &= \left\{ \left[I - \frac{\eta}{2} R \right]^{-1} \frac{\eta}{2} P^\top \right\} \delta_x + \left\{ \left[I - \frac{\eta}{2} R \right]^{-1} (I - \eta R) \right\} \delta_y + O(\|\delta_x\|^2 + \|\delta_y\|^2) \end{aligned}$$

which implies by definition of the Jacobian that

$$\nabla T_{xy}^{1/2}(z^*) = \begin{bmatrix} I - \frac{\eta}{2} Q - \frac{\eta^2}{4} P \left[I - \frac{\eta}{2} R \right]^{-1} P^\top & -\frac{\eta}{2} P \left[I - \frac{\eta}{2} R \right]^{-1} (I - \eta R) \\ \left[I - \frac{\eta}{2} R \right]^{-1} \frac{\eta}{2} P^\top & \left[I - \frac{\eta}{2} R \right]^{-1} (I - \eta R) \end{bmatrix}.$$

By symmetry, the Jacobian for $T_{yx}^{1/2}$ can be obtained from the above expression by swapping η for $-\eta$, Q for $-R$, P^\top for P , and swapping the lines resp. columns of the block matrix, yielding

$$\nabla T_{yx}^{1/2}(z^*) = \begin{bmatrix} \left[I - \frac{\eta}{2} Q \right]^{-1} (I - \eta Q) & -\left[I - \frac{\eta}{2} Q \right]^{-1} \frac{\eta}{2} P \\ \frac{\eta}{2} P^\top \left[I - \frac{\eta}{2} Q \right]^{-1} (I - \eta Q) & I - \frac{\eta}{2} R - \frac{\eta^2}{4} P^\top \left[I - \frac{\eta}{2} Q \right]^{-1} P \end{bmatrix}.$$

By straightforward but tedious computations, and noting that $\left[I - \frac{\eta}{2} R \right]^{-1}$ commutes with R , we arrive at the following expression for $\nabla \left[T_{xy}^{1/2} \circ T_{yx}^{1/2} \right] (z^*) = \nabla T_{xy}^{1/2}(z^*) \cdot \nabla T_{yx}^{1/2}(z^*)$ (since z^* is a fixed point of both operators):

$$\begin{aligned} &\nabla \left[T_{xy}^{1/2} \circ T_{yx}^{1/2} \right] (z^*) \\ &= \begin{bmatrix} I - \eta Q - \frac{\eta^2}{2} P P^\top \left[I - \frac{\eta}{2} Q \right]^{-1} (I - \eta Q) & -\eta P + \frac{\eta^2}{2} P R + \frac{\eta^3}{4} P P^\top \left[I - \frac{\eta}{2} Q \right]^{-1} P \\ \eta P^\top \left[I - \frac{\eta}{2} Q \right]^{-1} (I - \eta Q) & I - \eta R - \frac{\eta^2}{2} P^\top \left[I - \frac{\eta}{2} Q \right]^{-1} P \end{bmatrix} \\ &= \begin{bmatrix} I - \eta Q - \frac{\eta^2}{2} P P^\top & -\eta P + \frac{\eta^2}{2} P R \\ \eta P^\top - \frac{\eta^2}{2} P^\top Q & I - \eta R - \frac{\eta^2}{2} P^\top P \end{bmatrix} + \mathbf{E} = I - \eta M + \frac{\eta^2}{2} A M + \mathbf{E} \end{aligned}$$

where the error term \mathbf{E} is

$$\mathbf{E} = \begin{bmatrix} -\frac{\eta^2}{2} P P^\top \left\{ \left[I - \frac{\eta}{2} Q \right]^{-1} (I - \eta Q) - I \right\} & \frac{\eta^3}{4} P P^\top \left[I - \frac{\eta}{2} Q \right]^{-1} P \\ \eta P^\top \left\{ \left[I - \frac{\eta}{2} Q \right]^{-1} (I - \eta Q) - \left(I - \frac{\eta}{2} Q \right) \right\} & -\frac{\eta^2}{2} P^\top \left\{ \left[I - \frac{\eta}{2} Q \right]^{-1} - I \right\} P \end{bmatrix} = O(\eta^3)$$

thanks to our assumption that $\frac{\eta}{2} L_{xx} < 1$. More precisely, by bounding the operator norm of each block and summing the bounds, and using that $\left\| \left[I - \frac{\eta}{2} Q \right]^{-1} \right\| \leq \sum_{k=0}^{\infty} \left\| \frac{\eta}{2} Q \right\|^k = \frac{1}{1 - \eta \|Q\|/2} \leq \frac{1}{1 - \eta L_{xx}/2} \leq 2$, one can check that

$$\begin{aligned} \left\| R \right\| &\leq \left\| \left[I - \frac{\eta}{2} Q \right]^{-1} \right\| \left(\frac{\eta^2}{2} \left\| P \right\|^2 \cdot \frac{\eta}{2} \left\| Q \right\| + \frac{\eta^3}{4} \left\| P \right\|^3 + \eta \left\| P \right\| \cdot \frac{\eta^2}{4} \left\| Q \right\|^2 + \frac{\eta^2}{2} \left\| P \right\|^2 \cdot \frac{\eta}{2} \left\| Q \right\| \right) \\ &\leq \frac{1}{1 - \eta L_{xx}/2} \cdot \eta^3 \left\| P \right\| (\left\| P \right\| \vee \left\| Q \right\|)^2 \leq 2\eta^3 \left\| P \right\| (\left\| P \right\| \vee \left\| Q \right\|)^2. \end{aligned}$$

Note that by definition of $A = \begin{bmatrix} 0 & P \\ -P^\top & 0 \end{bmatrix}$ and $S = \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix}$, we have $\left\| A \right\| = \left\| P \right\|$ and $\left\| S \right\| = \left\| Q \right\| \vee \left\| R \right\|$.

This proves the claimed expansion for $\nabla \left[T_{xy}^{1/2} \circ T_{yx}^{1/2} \right] (z^*)$, and the expansion for $\nabla \left[T_{yx}^{1/2} \circ T_{xy}^{1/2} \right] (z^*)$ follows by symmetry. \square

D Details for Sec. 3.2

In this section we prove the expansions of the convergence rates of discrete-time algorithms (with non-asymptotic bounds in η and α) reported in Table 1. That is, we derive approximate expressions for $\rho(\nabla T(z^*))$ the spectral radius of the update operator's Jacobian at optimum, for various update rules (GDA, EG, etc.), when the skewed gradient field's Jacobian M has a small symmetric part.

We will repeatedly make use of the following estimate for the spectrum of a perturbed normal matrix. The expansion itself is a special case of [Tao08, Eqs. (5) and (7)] for M_0 normal, explicitly pointed out in that reference. Deriving the explicit bound on the error term involves rather tedious calculations however, so the proof is deferred to Sec. D.5.

Proposition D.1. *Let $M_0, M_1, M_2 \in \mathbb{R}^{d \times d}$ or $\mathbb{C}^{d \times d}$ and $M_\alpha = M_0 + \alpha M_1 + \frac{\alpha^2}{2} M_2$. Assume M_0 is normal and has distinct eigenvalues; denote its eigenvalue decomposition as $M_0 w_j = \lambda_j w_j$ with $(w_j)_j$ unitary, and let $\gamma_0 = \min_{k \neq j} |\lambda_k - \lambda_j|$. Then for all α such that $\left\| \alpha M_1 + \frac{\alpha^2}{2} M_2 \right\| \leq \frac{\gamma_0}{4\sqrt{2d}}$,*

$$\text{Sp}(M_\alpha) = \left\{ \lambda_j + \alpha \bar{w}_j^\top M_1 w_j + \frac{\alpha^2}{2} \bar{w}_j^\top M_2 w_j + \alpha^2 \sum_{k \neq j} \frac{(\bar{w}_j^\top M_1 w_k)(\bar{w}_k^\top M_1 w_j)}{\lambda_j - \lambda_k} + \mathbf{r}_j, 1 \leq j \leq d \right\}$$

where $\|\cdot\|$ denotes operator norm and for each j , $|\mathbf{r}_j| \leq \alpha^3 \cdot 8d\gamma_0^{-1} \left\| M_1 \right\| \left(\left\| M_2 \right\| + 4d\gamma_0^{-1} \left\| M_1 \right\|^2 \right)$.

In the remainder of this section, consider S real symmetric and A real antisymmetric in $\mathbb{R}^{d \times d}$, and let $M = M_\alpha = \alpha S + A$. Assume A has simple eigenvalues and denote its eigenvalue decomposition as $A w_j = i\sigma_j w_j$ with $\sigma_j \in \mathbb{R}$. This corresponds exactly to Sec. 3 with the correspondence $d := n + n$ and $\text{Sp}(A) = \{i\sigma_j, j \leq d\} := \{is\tilde{\sigma}_j, s \in \{-1, 1\}, j \leq n\}$ with $(\tilde{\sigma}_j)_{j \leq n}$ the singular values of P . Furthermore, let for concision $\gamma_A = \min_{k \neq j} |\sigma_k - \sigma_j|$ and $L_S = \|S\|$, $L_A = \|A\|$.

Throughout the derivations of this section, we will write $\varepsilon, \varepsilon', \varepsilon_j$ or ε'_j to denote elements of $[-1, 1]$ or of the unit ball of \mathbb{R}^d , and that may change from line to line. Likewise, $\zeta, \zeta', \zeta_j, \zeta'_j$ will denote elements of $\{z \in \mathbb{C}; |z| \leq 1\}$ or of the unit ball of \mathbb{C}^d , and that may change from line to line.

D.1 Simultaneous GDA

The Sim-GDA update rule is $z^{k+1} = T(z^k) = z^k - \eta g(z^k)$, so the update operator's Jacobian at optimum is

$$\nabla T(z^*) = I - \eta M = (I - \eta A) - \alpha \eta S.$$

Observe that $I - \eta A$ is normal with eigenvalue/eigenvector pairs $(1 - i\eta\sigma_j, w_j)$. Hence let us apply [Prop. D.1](#) with $M_0 = I - \eta A$ and $M_1 = -\eta S$ (and $M_2 = 0$). We get that

$$\forall \alpha \leq \frac{1}{\eta L_S} \frac{\eta \gamma_A}{4\sqrt{2d}} = \frac{\gamma_A}{L_S \cdot 4\sqrt{2d}},$$

$$\text{Sp}(\nabla T(z^*)) = \left\{ 1 - \eta i \sigma_j - \alpha \eta \bar{w}_j^\top S w_j + \alpha^2 \sum_{k \neq j} \frac{\eta^2 |\bar{w}_j^\top S w_k|^2}{-i\eta\sigma_j + i\eta\sigma_k} + \alpha^3 \eta \cdot \zeta_j (32d^2 \gamma_A^{-2} L_S^3), 1 \leq j \leq d \right\}.$$

It seems unlikely that the term in α^2 will ever lead to friendly expressions, so we will bound it uniformly; let us nonetheless note that it is pure imaginary. Namely, since

$$\left| \alpha^2 \sum_{k \neq j} \frac{\eta^2 |\bar{w}_j^\top S w_k|^2}{-i\eta\sigma_j + i\eta\sigma_k} \right| \leq \alpha^2 \eta \sum_{k \neq j} \frac{|\bar{w}_k^\top S w_j|^2}{|\sigma_k - \sigma_j|} \leq \alpha^2 \eta \gamma_A^{-1} \underbrace{\sum_k |\bar{w}_k^\top S w_j|^2}_{=\|S w_j\|^2 \leq L_S^2}, \quad (\text{D.1})$$

then we may write that term as $i\alpha^2 \eta \cdot \varepsilon_j (\gamma_A^{-1} L_S^2)$. For the spectral radius we get

$$\begin{aligned} \rho(\nabla T(z^*))^2 &= \max_{\lambda \in \text{Sp}(\nabla T(z^*))} |\lambda|^2 \\ &= \max_{j \leq d} \left| 1 - \alpha \eta \bar{w}_j^\top S w_j + \alpha^3 \eta \cdot \varepsilon_j' 32d^2 \gamma_A^{-2} L_S^3 \right|^2 + \left| \eta \sigma_j + \alpha^2 \eta \cdot \varepsilon_j \gamma_A^{-1} L_S^2 + \alpha^3 \eta \cdot \varepsilon_j'' 32d^2 \gamma_A^{-2} L_S^3 \right|^2 \\ &= \max_{j \leq d} 1 - 2\alpha \eta (\bar{w}_j^\top S w_j) + \eta^2 \sigma_j^2 + O(\alpha^3 \eta + \alpha^2 \eta^2) \end{aligned}$$

and more precisely one can check that the $O(\cdot)$ term is absolutely bounded by $\alpha^3 \eta \cdot 128d^2 \gamma_A^{-2} L_S^3 (1 + 5\eta L_A) + \alpha^2 \eta^2 \cdot 2\gamma_A^{-1} L_S^2 L_A$, for all $\alpha \leq \frac{\gamma_A}{L_S \cdot 4\sqrt{2d}}$.

D.2 Proximal Point

The PP update rule is $z^{k+1} = T(z^k)$ with $T^{-1}(z) = z + \eta g(z)$ so the update operator's Jacobian at optimum is $\nabla T(z^*) = (I + \eta M)^{-1}$. So $\text{Sp}(\nabla T(z^*)) = \{\lambda^{-1}, \lambda \in \text{Sp}(I + \eta M)\}$ and so, by the exact same calculations as for Sim-GDA with η replaced by $-\eta$,

$$\rho(\nabla T(z^*))^2 = \left[\min_{j \leq d} 1 + 2\alpha \eta (\bar{w}_j^\top S w_j) + \eta^2 \sigma_j^2 + O(\alpha^3 \eta + \alpha^2 \eta^2) \right]^{-1}$$

for all $\alpha \leq \frac{\gamma_A}{L_S \cdot 4\sqrt{2d}}$, and we have the same absolute bound on the $O(\cdot)$ term as for Sim-GDA.

D.3 Alternating GDA

As we show in [App. C](#), for any $\eta \leq \|\nabla_{xx}^2 f\|_\infty^{-1} \wedge \|\nabla_{yy}^2 f\|_\infty^{-1}$, the (symmetrized) Alt-GDA update operator's Jacobian at optimum is

$$\begin{aligned} \nabla T(z^*) &= \nabla \bar{T}(z^*) + \mathbf{E} \quad \text{for some } \mathbf{E} \in \mathbb{R}^{d \times d} \text{ with } \|\mathbf{E}\| \leq 2\eta^3 L_A (L_A \vee \alpha L_S)^2, \\ \text{where } \nabla \bar{T}(z^*) &= I - \eta M + \frac{\eta^2}{2} A M = \left(I - \eta A + \frac{\eta^2}{2} A^2 \right) + \alpha \left(-\eta S + \frac{\eta^2}{2} A S \right). \end{aligned}$$

Observe that $I - \eta A + \frac{\eta^2}{2} A^2$ is normal with eigenvalue/eigenvector pairs $(1 - i\eta\sigma_j - \frac{\eta^2}{2}\sigma_j^2, w_j)$.

Hence let us apply [Prop. D.1](#) with $M_0 = I - \eta A + \frac{\eta^2}{2} A^2$ and $M_1 = -\eta S + \frac{\eta^2}{2} A S$ (and $M_2 = 0$). Further observe that

$$\bar{w}_j^\top A S w_k = \bar{w}_j^\top \left(\sum_l i\sigma_l w_l \bar{w}_l^\top \right) S w_k = i\sigma_j \cdot \bar{w}_j^\top S w_k.$$

So we get that, for all $\alpha \leq \frac{\gamma_A}{4\sqrt{2d} \cdot L_s (1 + \frac{\eta}{2} L_A)}$,

$$\begin{aligned}
& \text{Sp}(\nabla \bar{T}(z^*)) \\
&= \left\{ 1 - \eta i \sigma_j - \frac{\eta^2}{2} \sigma_j^2 + \alpha \bar{w}_j^\top \left(-\eta S + \frac{\eta^2}{2} AS \right) w_j + \alpha^2 \sum_{k \neq j} \frac{\eta^2 (\bar{w}_j^\top (S - \frac{\eta}{2} AS) w_k) (\bar{w}_k^\top (S - \frac{\eta}{2} AS) w_j)}{-i\eta \sigma_j + i\eta \sigma_k + \frac{\eta^2}{2} (-\sigma_j^2 + \sigma_k^2)} \right. \\
&\quad \left. + \alpha^3 \eta \cdot \zeta_j \left(32d^2 \gamma_A^{-2} L_S^3 \left(1 + \frac{\eta}{2} L_A \right)^3 \right), 1 \leq j \leq d \right\} \\
&= \left\{ 1 - \eta i \sigma_j - \frac{\eta^2}{2} \sigma_j^2 - \alpha \eta (\bar{w}_j^\top S w_j) + \alpha \frac{\eta^2}{2} i \sigma_j \cdot (\bar{w}_j^\top S w_j) + i \cdot \alpha^2 \eta \cdot \varepsilon_j (\gamma_A^{-1} L_S^2) \right. \\
&\quad \left. + \alpha^3 \eta \cdot \zeta_j \left(32d^2 \gamma_A^{-2} L_S^3 \left(1 + \frac{\eta}{2} L_A \right)^3 \right) + \alpha^2 \eta^3 \cdot \zeta'_j (\gamma_A^{-1} L_S^2 L_A^2 / 4), 1 \leq j \leq d \right\}.
\end{aligned}$$

Here for the second equality, we computed the coefficient for the term in α^2 as

$$\begin{aligned}
& \sum_{k \neq j} \frac{\eta^2 (\bar{w}_j^\top (S - \frac{\eta}{2} AS) w_k) (\bar{w}_k^\top (S - \frac{\eta}{2} AS) w_j)}{-i\eta \sigma_j + i\eta \sigma_k + \frac{\eta^2}{2} (-\sigma_j^2 + \sigma_k^2)} \\
&= \sum_{k \neq j} \eta \frac{(\bar{w}_j^\top S w_k - \frac{\eta}{2} i \sigma_j \bar{w}_j^\top S w_k) (\bar{w}_k^\top S w_j - \frac{\eta}{2} i \sigma_k \bar{w}_k^\top S w_j)}{i(\sigma_k - \sigma_j) + \frac{\eta}{2} (\sigma_k^2 - \sigma_j^2)} = \sum_{k \neq j} \eta |\bar{w}_j^\top S w_k|^2 \frac{(1 - \frac{\eta}{2} i \sigma_j) (1 - \frac{\eta}{2} i \sigma_k)}{i(\sigma_k - \sigma_j) (1 - i \frac{\eta}{2} (\sigma_k + \sigma_j))} \\
&= \sum_{k \neq j} \frac{\eta |\bar{w}_j^\top S w_k|^2}{i(\sigma_k - \sigma_j)} \left(1 + \frac{-\frac{\eta^2}{4} \sigma_j \sigma_k}{1 - i \frac{\eta}{2} (\sigma_k + \sigma_j)} \right) = i\eta \cdot \varepsilon_j (\gamma_A^{-1} L_S^2) + \eta^3 \cdot \zeta'_j (\gamma_A^{-1} L_S^2 \cdot L_A^2 / 4)
\end{aligned}$$

where the last equality follows from the same bound as in (D.1). For the spectral radius we get

$$\begin{aligned}
\rho(\nabla \bar{T}(z^*))^2 &= \max_{j \leq d} \left| 1 - \frac{\eta^2}{2} \sigma_j^2 - \alpha \eta (\bar{w}_j^\top S w_j) \right|^2 + \left| -\eta \sigma_j + \alpha \frac{\eta^2}{2} \sigma_j \cdot (\bar{w}_j^\top S w_j) + O(\alpha^2 \eta) \right|^2 \\
&\quad + O(\alpha^3 \eta + \alpha^2 \eta^3) \\
&= \max_{j \leq d} 1 + \frac{\eta^4}{4} \sigma_j^4 - \eta^2 \sigma_j^2 - 2\alpha \eta (\bar{w}_j^\top S w_j) + \alpha \eta^3 \sigma_j^2 (\bar{w}_j^\top S w_j) \\
&\quad + \eta^2 \sigma_j^2 - \alpha \eta^3 \sigma_j^2 (\bar{w}_j^\top S w_j) + O(\alpha^3 \eta + \alpha^2 \eta^2) \\
&= \max_{j \leq d} 1 - 2\alpha \eta (\bar{w}_j^\top S w_j) + \frac{\eta^4}{4} \sigma_j^4 + O(\alpha^3 \eta + \alpha^2 \eta^2).
\end{aligned}$$

The two cancellations that occur in the last line (of $\pm \eta^2 \sigma_j^2$ and of $\pm \alpha \eta^3 \sigma_j^2 (\bar{w}_j^\top S w_j)$) are consistent with the intuition that Alt-GDA is a good symplectic integrator. More precisely, one can check that the final $O(\cdot)$ term is absolutely bounded by $\alpha^3 \eta \cdot 512d^2 \gamma_A^{-2} L_S^3 (1 + \eta L_A)^6 + \alpha^2 \eta^2 \cdot 4d \gamma_A^{-1} L_S^2 L_A (1 + \eta L_A)^4$, for all $\alpha \leq \frac{\gamma_A}{4\sqrt{2d} \cdot L_s (1 + \frac{\eta}{2} L_A)}$.

Finally, by smoothness of the eigenvalues of perturbed matrices (Lemma D.2),

$$\rho(\nabla T(z^*))^2 = \rho(\nabla \bar{T}(z^*) + \mathbf{E})^2 = \rho(\nabla \bar{T}(z^*))^2 + O(\eta^3).$$

One could further derive explicit bounds on the $O(\eta^3)$ term in that last expression by controlling the distance to normality of $\nabla \bar{T}(z^*)$ – since it is continuous in α and normal for $\alpha = 0$ – and adapting the proof of Lemma D.5 to matrices that are close to normal.

D.4 Extra-Gradient

The EG update rule is $z^{k+1} = T(z^k) = z^k - \eta g(z^k - \eta g(z^k))$ so the update operator's Jacobian at optimum is

$$\begin{aligned}
\nabla T(z^*) &= I - \eta M (I - \eta M) = I - \eta A - \alpha \eta S + \eta^2 M^2 \\
&= (I - \eta A + \eta^2 A^2) + \alpha (-\eta S + \eta^2 (AS + SA)) + \frac{1}{2} \alpha^2 (2\eta^2 S^2).
\end{aligned}$$

Similarly as in the previous subsection, observe that $I - \eta A + \eta^2 A^2$ is normal with eigenvalue/eigenvector pairs $(1 - i\eta\sigma_j - \eta^2\sigma_j^2, w_j)$, and that

$$\bar{w}_j^\top (AS + SA)w_k = i(\sigma_j + \sigma_k) \cdot \bar{w}_j^\top Sw_k.$$

Let us apply [Prop. D.1](#) to $M_0 = I - \eta A + \eta^2 A^2$, $M_1 = -\eta S + \eta^2(AS + SA)$ and $M_2 = 2\eta^2 S^2$. The proposition yields a bound for all α such that $\alpha \left\| S - \eta(AS + SA) + \alpha\eta S^2 \right\| \leq \frac{\gamma_A}{4\sqrt{2d}}$, for which a simpler sufficient condition is $\alpha \leq 1 \wedge \frac{\gamma_A}{4\sqrt{2d} \cdot L_S(1 + \eta L_S(1 + 2L_A))}$. Using that $\|M_1\| \leq \eta L_S(1 + 2\eta L_A)$ to upper-bound the term in α^3 , we get that for all such α ,

$$\begin{aligned} \text{Sp}(\nabla T(z^*)) &= \left\{ 1 - i\eta\sigma_j - \eta^2\sigma_j^2 + \alpha(-\eta\bar{w}_j^\top Sw_j + \eta^2 2i\sigma_j\bar{w}_j^\top Sw_j) + \alpha^2\eta^2\bar{w}_j^\top S^2w_j \right. \\ &\quad + \alpha^2 \sum_{k \neq j} \frac{\eta^2\bar{w}_j^\top (S - \eta(AS + SA))w_k \cdot \bar{w}_k^\top (S - \eta(AS + SA))w_j}{-i\eta\sigma_j + i\eta\sigma_k + \eta^2(-\sigma_j^2 + \sigma_k^2)} \\ &\quad \left. + \alpha^3\eta \cdot \zeta_j \cdot 8d\gamma_A^{-1}L_S(1 + 2\eta L_A) (2\eta L_S^2 + 4d\gamma_A^{-1}L_S^2(1 + 2\eta L_A)^2), 1 \leq j \leq d \right\} \\ &= \left\{ 1 - i\eta\sigma_j - \eta^2\sigma_j^2 - \alpha\eta(\bar{w}_j^\top Sw_j) + 2i\alpha\eta^2\sigma_j(\bar{w}_j^\top Sw_j) \right. \\ &\quad + i \cdot \alpha^2\eta \cdot \varepsilon_j(\gamma_A^{-1}L_S^2) + \alpha^2\eta^2 \cdot \varepsilon'_j(2\gamma_A^{-1}L_S^2L_A) + \alpha^2\eta^2\bar{w}_j^\top S^2w_j \\ &\quad \left. + \alpha^3\eta \cdot \zeta_j(512d^2\gamma_A^{-2}L_S^3(1 + \eta L_A)^3), 1 \leq j \leq d \right\}. \end{aligned}$$

Here for the second equality we computed the terms in α^2 as

$$\begin{aligned} &\alpha^2 \sum_{k \neq j} \frac{\eta^2\bar{w}_j^\top (S - \eta(AS + SA))w_k \cdot \bar{w}_k^\top (S - \eta(AS + SA))w_j}{-i\eta\sigma_j + i\eta\sigma_k + \eta^2(-\sigma_j^2 + \sigma_k^2)} \\ &= \alpha^2 \sum_{k \neq j} \eta \frac{(\bar{w}_j^\top Sw_k)(1 - i\eta(\sigma_j + \sigma_k)) \cdot (\bar{w}_k^\top Sw_j)(1 - i\eta(\sigma_j + \sigma_k))}{i(\sigma_k - \sigma_j)(1 - i\eta(\sigma_k + \sigma_j))} \\ &= \alpha^2 \sum_{k \neq j} \frac{\eta |\bar{w}_j^\top Sw_k|^2}{i(\sigma_k - \sigma_j)} (1 - i\eta(\sigma_k + \sigma_j)) = i \cdot \alpha^2\eta \cdot \varepsilon_j(\gamma_A^{-1}L_S^2) + \alpha^2\eta^2 \cdot \varepsilon'_j(\gamma_A^{-1}L_S^2 \cdot 2L_A) \end{aligned}$$

where the last equality follows from the same bound as in [\(D.1\)](#). For the spectral radius we get

$$\begin{aligned} \rho(\nabla T(z^*))^2 &= \max_{j \leq d} \left| 1 - \eta^2\sigma_j^2 - \alpha\eta(\bar{w}_j^\top Sw_j) + O(\alpha^2\eta^2) \right|^2 + \left| -\eta\sigma_j + 2\alpha\eta^2\sigma_j(\bar{w}_j^\top Sw_j) + O(\alpha^2\eta) \right|^2 \\ &\quad + O(\alpha^3\eta) \\ &= \max_{j \leq d} \left(1 + \eta^4\sigma_j^4 - 2\eta^2\sigma_j^2 - 2\alpha\eta(\bar{w}_j^\top Sw_j) + 2\alpha\eta^3\sigma_j^2(\bar{w}_j^\top Sw_j) \right. \\ &\quad \left. + \eta^2\sigma_j^2 - 4\alpha\eta^3\sigma_j^2(\bar{w}_j^\top Sw_j) + O(\alpha^3\eta + \alpha^2\eta^2) \right) \\ &= \max_{j \leq d} \left(1 - 2\alpha\eta(\bar{w}_j^\top Sw_j) - \eta^2\sigma_j^2 - 2\alpha\eta^3\sigma_j^2(\bar{w}_j^\top Sw_j) + \eta^4\sigma_j^4 + O(\alpha^3\eta + \alpha^2\eta^2) \right) \end{aligned}$$

and more precisely one can check that the final $O(\cdot)$ term is absolutely bounded by $\alpha^3\eta \cdot 2^{12}d^2\gamma_A^{-2}L_S^3(1 + \eta L_A)^7 + \alpha^2\eta^2 \cdot 15d\gamma_A^{-1}L_S^2L_A(1 + \eta L_A)^2$.

D.5 Proof of [Prop. D.1](#)

The following lemma summarizes the eigenvalue derivative formulas up to order 2 for perturbed matrices with distinct eigenvalues.

Lemma D.2 ([\[Tao08\]](#)). *Let $M_\alpha \in \mathbb{C}^{d \times d}$ for $\alpha \in (-1, 1)$ a smooth curve of matrices such that M_0 has distinct eigenvalues. Then there exists an open interval $I \ni 0$ such that M_α has distinct eigenvalues*

for all $\alpha \in I$; in particular they are diagonalizable. Denote their eigenvalue decompositions as $M_\alpha = \sum_{k=1}^d \lambda_k(\alpha) v_k(\alpha) \bar{w}_k(\alpha)^\top$; that is, the eigenvalues of M_α are $(\lambda_k(\alpha))_k$, the associated eigenvectors are $(v_k(\alpha))_k$, and $(w_k(\alpha))_k$ is a dual basis to the basis of eigenvectors, i.e., $\bar{w}_k(\alpha)^\top v_j(\alpha) = \mathbb{1}_{j=k}$.

The eigenvalues of M_α are smooth over I and their first two derivatives at any $\alpha \in I$ are given by, using $\dot{\bullet}$ to denote differentiation w.r.t. α and leaving the dependency on α implicit,

$$\dot{\lambda}_k = w_k^* \dot{M} v_k \quad \text{and} \quad \ddot{\lambda}_k = w_k^* \ddot{M} v_k + 2 \sum_{j \neq k} \frac{(w_k^* \dot{M} v_j)(w_j^* \dot{M} v_k)}{\lambda_k - \lambda_j}$$

where we denoted $w_k^* = \bar{w}_k^\top$. Furthermore the eigenvectors v_k and dual basis vectors w_k can also be chosen smooth and their derivatives at any $\alpha \in I$ are given by

$$\dot{v}_k = \sum_{j \neq k} \frac{w_j^* \dot{M} v_k}{\lambda_k - \lambda_j} v_j + c_k v_k \quad \text{and} \quad \dot{w}_k^* = \sum_{j \neq k} \frac{w_k^* \dot{M} v_j}{\lambda_k - \lambda_j} w_j^* - c_k w_k^*$$

for some scalars c_k that reflect the normalization of the eigenvectors.

By applying a Taylor expansion with remainder in Lagrange form to the eigenvalues $\lambda_k(\alpha)$ of the matrix $M_\alpha = M_0 + \alpha M_1 + \frac{\alpha^2}{2} M_2$ of [Prop. D.1](#) – since the eigenvalues are smooth by the above lemma –, we have that for all α in some neighborhood of zero,

$$\lambda_k(\alpha) = \lambda_k(0) + \alpha \dot{\lambda}_k(0) + \frac{\alpha^2}{2} \ddot{\lambda}_k(0) + \frac{\alpha^3}{6} \ddot{\lambda}_k(\xi) \quad \text{for some } 0 < \xi < \alpha. \quad (\text{D.2})$$

By substituting the expressions from [Lemma D.2](#) for the first two eigenvalue derivatives, we already get the terms in α and α^2 in [Prop. D.1](#). In order to control the last term in α^3 , we need to compute the third eigenvalue derivatives $\ddot{\lambda}_k(\xi)$. Note that M_ξ is never assumed normal for any $\xi > 0$, which is why we do not use normality in [Lemma D.2](#) nor in [Lemma D.3](#) below.

Lemma D.3. *Under the conditions of [Lemma D.2](#),*

$$\begin{aligned} \ddot{\lambda}_k &= w_k^* \ddot{M} v_k + 3 \sum_{j \neq k} \frac{1}{\lambda_k - \lambda_j} \left[(w_k^* \dot{M} v_j)(w_j^* \ddot{M} v_k) + (w_k^* \ddot{M} v_j)(w_j^* \dot{M} v_k) \right] \\ &\quad + 6 \sum_{j, l \neq k} \frac{(w_k^* \dot{M} v_j)(w_j^* \dot{M} v_l)(w_l^* \dot{M} v_k)}{(\lambda_k - \lambda_j)(\lambda_k - \lambda_l)} - 6 \sum_{j \neq k} \frac{(w_k^* \dot{M} v_j)(w_j^* \dot{M} v_k)}{(\lambda_k - \lambda_j)^2} w_k^* \dot{M} v_k. \end{aligned}$$

In particular,

$$\max_k \|\ddot{\lambda}_k\| \leq \chi \left(\|\ddot{M}\| + 6d\gamma^{-1}\chi^2 \|\dot{M}\| \|\ddot{M}\| + 6d^2\gamma^{-2}\chi^3 \|\dot{M}\|^3 \right)$$

where $\gamma = \min_{j \neq k} |\lambda_j - \lambda_k|$, $\chi = \max_k \|v_k\| \|w_k\|$ and $\|\cdot\|$ denotes operator norm.

Proof. By differentiating the identity $M v_k = \lambda_k v_k$, we have that

$$\begin{aligned} \dot{M} v_k + M \dot{v}_k &= \dot{\lambda}_k v_k + \lambda_k \dot{v}_k \\ \ddot{M} v_k + 2\dot{M} \dot{v}_k + M \ddot{v}_k &= \ddot{\lambda}_k v_k + 2\dot{\lambda}_k \dot{v}_k + \lambda_k \ddot{v}_k \\ \ddot{\lambda}_k v_k + 3\dot{\lambda}_k \dot{v}_k + 3\dot{\lambda}_k \ddot{v}_k + \lambda_k \ddot{v}_k &= \ddot{\lambda}_k v_k + 3\dot{\lambda}_k \dot{v}_k + 3\dot{\lambda}_k \ddot{v}_k + \lambda_k \ddot{v}_k. \end{aligned}$$

Also note that $w_k^* M = \lambda_k w_k^*$. By multiplying the third identity by w_k^* on the left, we have that

$$\begin{aligned} w_k^* \ddot{M} v_k + 3w_k^* \ddot{M} \dot{v}_k + 3w_k^* \dot{M} \ddot{v}_k + w_k^* M \ddot{v}_k &= \ddot{\lambda}_k + 3\dot{\lambda}_k w_k^* \dot{v}_k + 3\dot{\lambda}_k w_k^* \ddot{v}_k + \lambda_k w_k^* \ddot{v}_k \\ w_k^* \ddot{M} v_k + 3w_k^* \ddot{M} \dot{v}_k + 3w_k^* \dot{M} \ddot{v}_k &= \ddot{\lambda}_k + 3\dot{\lambda}_k w_k^* \dot{v}_k + 3\dot{\lambda}_k w_k^* \ddot{v}_k \end{aligned}$$

since $(w_k^* M) \ddot{v}_k = \lambda_k w_k^* \ddot{v}_k$. Now let us compute \ddot{v}_k . By multiplying the identity for the second derivatives by w_j^* on the left for any $j \neq k$, we have that

$$\begin{aligned} w_j^* \ddot{M} v_k + 2w_j^* \dot{M} \dot{v}_k + \underbrace{w_j^* M \ddot{v}_k}_{=\lambda_j w_j^*} &= \ddot{\lambda}_k w_j^* v_k + 2\dot{\lambda}_k w_j^* \dot{v}_k + \lambda_k w_j^* \ddot{v}_k \\ w_j^* \ddot{v}_k &= \frac{1}{\lambda_j - \lambda_k} \left(2\dot{\lambda}_k w_j^* \dot{v}_k - w_j^* \ddot{M} v_k - 2w_j^* \dot{M} \dot{v}_k \right). \end{aligned}$$

Hence we can compute

$$\begin{aligned} w_k^* \dot{M} \ddot{v}_k - \dot{\lambda}_k w_k^* \ddot{v}_k &= w_k^* \dot{M} \sum_j (w_j^* \ddot{v}_k) v_j - \dot{\lambda}_k w_k^* \ddot{v}_k \\ &= \sum_{j \neq k} \left(w_k^* \dot{M} v_j \right) \frac{1}{\lambda_j - \lambda_k} \left(2\dot{\lambda}_k w_j^* \ddot{v}_k - w_j^* \ddot{M} v_k - 2w_j^* \dot{M} \dot{v}_k \right) + \underbrace{(w_k^* \dot{M} v_k)(w_k^* \ddot{v}_k) - \dot{\lambda}_k w_k^* \ddot{v}_k}_{=0} \end{aligned}$$

since $\dot{\lambda}_k = w_k^* \dot{M} v_k$ by [Lemma D.2](#). Substituting back into the identity for the third derivative left-multiplied by w_k^* , we find that

$$\ddot{\lambda}_k = w_k^* \ddot{M} v_k + 3w_k^* \ddot{M} \dot{v}_k - 3\ddot{\lambda}_k w_k^* \dot{v}_k + 3 \sum_{j \neq k} \frac{w_k^* \dot{M} v_j}{\lambda_j - \lambda_k} \left(2\dot{\lambda}_k w_j^* \dot{v}_k - w_j^* \ddot{M} v_k - 2w_j^* \dot{M} \dot{v}_k \right).$$

The claimed expression for $\ddot{\lambda}_k$ will follow by substituting the expressions for \dot{v}_k , $\dot{\lambda}_k$ and $\ddot{\lambda}_k$ from [Lemma D.2](#) and simplifying. Namely, letting for concision $\delta_{jk} = \lambda_j - \lambda_k$ and $M_{jk} = w_j^* \dot{M} v_k$, $\dot{M}_{jk} = w_j^* \dot{M} v_k$ and similarly for \ddot{M}_{jk} , \ddot{M}_{jk} for any $j, k \leq d$, we have

$$\dot{\lambda}_k = \dot{M}_{kk}, \quad \dot{v}_k = \sum_{j \neq k} \frac{\dot{M}_{jk}}{\delta_{kj}} v_j + c_k v_k, \quad \ddot{\lambda}_k = \ddot{M}_{kk} + 2 \sum_{j \neq k} \frac{\dot{M}_{kj} \dot{M}_{jk}}{\delta_{kj}}$$

and in particular $w_j^* \dot{v}_k = \frac{\dot{M}_{jk}}{\delta_{kj}}$ for all $j \neq k$ and so

$$\begin{aligned} \ddot{\lambda}_k &= \ddot{M}_{kk} + 3 \sum_{j \neq k} \frac{\ddot{M}_{kj} \dot{M}_{jk}}{\delta_{kj}} + 3\ddot{M}_{kk} c_k - 3 \left(\ddot{M}_{kk} + 2 \sum_{j \neq k} \frac{\dot{M}_{kj} \dot{M}_{jk}}{\delta_{kj}} \right) c_k \\ &\quad - 3 \sum_{j \neq k} \frac{\dot{M}_{kj} \ddot{M}_{jk}}{\delta_{jk}} + 6 \sum_{j \neq k} \frac{\dot{M}_{kj}}{\delta_{jk}} \left(\dot{M}_{kk} \frac{\dot{M}_{jk}}{\delta_{kj}} - \sum_{l \neq k} \frac{\dot{M}_{jl} \dot{M}_{lk}}{\delta_{kl}} - \dot{M}_{jk} c_k \right) \\ &= \ddot{M}_{kk} + 3 \sum_{j \neq k} \frac{\ddot{M}_{kj} \dot{M}_{jk} + \dot{M}_{kj} \ddot{M}_{jk}}{\delta_{kj}} + 6 \sum_{j \neq k} \frac{\dot{M}_{kj}}{\delta_{jk}} \left(\dot{M}_{kk} \frac{\dot{M}_{jk}}{\delta_{kj}} - \sum_{l \neq k} \frac{\dot{M}_{jl} \dot{M}_{lk}}{\delta_{kl}} \right). \end{aligned}$$

In order to simplify the last term, we simply write it as

$$\sum_{j \neq k} \frac{\dot{M}_{kj}}{\delta_{jk}} \left(\dot{M}_{kk} \frac{\dot{M}_{jk}}{\delta_{kj}} - \sum_{l \neq k} \frac{\dot{M}_{jl} \dot{M}_{lk}}{\delta_{kl}} \right) = \sum_{j, l \neq k} \frac{\dot{M}_{kj} \dot{M}_{jl} \dot{M}_{lk}}{\delta_{kj} \delta_{kl}} - \sum_{j \neq k} \frac{\dot{M}_{kj} \dot{M}_{jk}}{\delta_{kj}^2} \dot{M}_{kk}.$$

By substituting, we obtain the announced expression for $\ddot{\lambda}_k$. \square

In order to bound the term in $\ddot{\lambda}_k(\xi)$ in [\(D.2\)](#), we want to control

$$\gamma(\alpha) := \min_{j \neq k} |\lambda_j(\alpha) - \lambda_k(\alpha)| \quad \text{and} \quad \chi(\alpha) := \max_k \|v_k(\alpha)\| \|w_k(\alpha)\|$$

the uniform eigengap resp. maximal eigenvalue condition number, uniformly in α for α in a neighborhood of zero. Now assuming M_0 has distinct eigenvalues then $\gamma(0) > 0$, and assuming M_0 is normal – as is the case for [Prop. D.1](#) – then $\chi(0) = 1$. So we want to bound the perturbation of the eigenvalues and eigenvectors uniformly.

For the eigenvalues, thanks to the assumption that M_0 is normal, we easily get the following bound.

Lemma D.4. *Under the conditions of [Lemma D.2](#), if additionally M_0 is normal, then it holds*

$$\max_k |\lambda_k(\alpha) - \lambda_k(0)| \leq \|M_\alpha - M_0\|$$

for all $\alpha \in I'$ the maximal open interval containing 0 such that $\forall \alpha \in I', \|M_\alpha - M_0\| < \frac{1}{2} \gamma(0)$.

In particular, we have for all $\alpha \in I'$

$$\gamma(\alpha) \geq \gamma(0) - 2\|M_\alpha - M_0\| > 0.$$

Note that $I' \subset I$ the interval from [Lemma D.2](#), since $\gamma(\alpha) > 0$ if and only if M_α has distinct eigenvalues.

Proof. Let λ any eigenvalue of M_α and v such that $M_\alpha v = \lambda v$ and $\|v\| = 1$. Since M_0 is normal, it is orthonormally diagonalizable, so we may write it as $M_0 = U_0 \Lambda_0 \bar{U}_0^\top$ with U_0 unitary and $\Lambda_0 = \text{Diag}((\lambda_k(0)))$. Then, letting $\tilde{v} = \bar{U}_0^\top v$ and $\Delta = M_\alpha - M_0$ and $\tilde{\Delta} = \bar{U}_0^\top \Delta U_0$,

$$M_\alpha v = (M_0 + \Delta)v = U_0^\top (\Lambda_0 + \tilde{\Delta}) \bar{U}_0^\top v = \lambda v, \text{ i.e., } (\Lambda_0 + \tilde{\Delta})\tilde{v} = \lambda \tilde{v}$$

and so $\min_j |\lambda - \lambda_{0j}|^2 \leq \sum_j |\tilde{v}[j]|^2 |\lambda - \lambda_{0j}|^2 = \|(\Lambda_0 + \tilde{\Delta})\tilde{v}\|^2 \leq \|\tilde{\Delta}\tilde{v}\|^2 \leq \|\Delta\|^2$

since U_0 is unitary and $\|v\| = 1$.

This shows that for any $\lambda \in \text{Sp}(M_\alpha)$, there exists a $\lambda_k(0) \in \text{Sp}(M_0)$ which is close to it, i.e.,

$$\forall k, \exists j \text{ s.t. } |\lambda_k(\alpha) - \lambda_j(0)| \leq \|M_\alpha - M_0\|.$$

A fortiori, we can ensure that the smooth parametrization of the eigenvalues $(\lambda_k(\alpha))_k$ satisfies the inequality announced in the lemma, by restraining α to some I' small enough so that $\arg \min_j |\lambda_k(\alpha) - \lambda_j(0)| = k$ for all $\alpha \in I'$. More explicitly, this can be achieved by choosing I' such that $\sup_{\alpha \in I'} \|M_\alpha - M_0\| < \frac{1}{2} \min_{j \neq k} |\lambda_j(0) - \lambda_k(0)| = \frac{1}{2} \gamma(0)$. Hence the choice of I' announced in the lemma. \square

For the eigenvectors, also using the assumption that M_0 is normal, we get the following bound.

Lemma D.5. *Under the conditions of [Lemma D.2](#), assume additionally that M_0 is normal, and choose the normalization of the eigenvectors $(v_k(\alpha))_k$ such that $\|v_k(\alpha)\| = 1$ and $v_k^*(0) \cdot v_k(\alpha) \in \mathbb{R}_+$ for all k . Then*

$$\forall k, \|v_k(\alpha) - v_k(0)\| \leq \frac{2\sqrt{2}\|M_\alpha - M_0\|}{\gamma(0)}$$

and $\|w_k^*(\alpha) - v_k^*(0)\| \leq \sqrt{d} \|w_k(\alpha)\| \frac{2\sqrt{2}\|M_\alpha - M_0\|}{\gamma(0)}$

for all $\alpha \in I'$ the interval from [Lemma D.4](#).

In particular, we have

$$\chi(\alpha) \leq \frac{\gamma(0)}{\gamma(0) - 2\sqrt{2d}\|M_\alpha - M_0\|}$$

for all $\alpha \in I''$ the maximal open interval containing 0 such that $\forall \alpha \in I'', \|M_\alpha - M_0\| < \frac{1}{2\sqrt{2d}}\gamma(0)$.

Proof. We will write for concision $v_k = v_k(\alpha)$ and $v_{0k} = v_k(0) = w_k(0)$ since M_0 is normal, and $\lambda_k = \lambda_k(\alpha)$, $\lambda_{0k} = \lambda_k(0)$. Fix $\alpha \in I'$ and let $\Delta = M_\alpha - M_0$; by [Lemma D.4](#) we have $\max_k |\lambda_k - \lambda_{0k}| \leq \|\Delta\| \leq \frac{1}{2}\gamma(0)$.

Fix k . Subtracting $M_0 v_{0k} = \lambda_{0k} v_{0k}$ from $(M_0 + \Delta)v_k = \lambda_k v_k$, we have

$$M_0(v_k - v_{0k}) + \Delta v_k = \lambda_k v_k - \lambda_{0k} v_{0k} = \lambda_{0k}(v_k - v_{0k}) + (\lambda_k - \lambda_{0k})v_{0k}$$

$$(M_0 - \lambda_{0k}I)(v_k - v_{0k}) = -\Delta v_k + (\lambda_k - \lambda_{0k})v_{0k}$$

$$\sum_{j \neq k} |\lambda_{0j} - \lambda_{0k}|^2 |v_{0j}^* v_k|^2 = \|(M_0 - \lambda_{0k}I)(v_k - v_{0k})\|^2 \leq \|\Delta\|^2 (\|v_k\| + 1)^2 = 4\|\Delta\|^2$$

where $v_{0j}^* := \bar{v}_{0j}^\top$, using the unitary basis $(v_{0j})_j$ to compute the norm on the left-hand side. The left-hand side is further lower-bounded by $\gamma(0)^2 \sum_{j \neq k} |v_{0j}^* v_k|^2 = \gamma(0)^2 (1 - |v_{0k}^* v_k|^2)$, so

$$1 - |v_{0k}^* v_k|^2 \leq \frac{4\|\Delta\|^2}{\gamma(0)^2}.$$

Consequently, by our choice of normalization: $v_k^*(0)v_k(\alpha) \in \mathbb{R}_+$, we have

$$\|v_k - v_{0k}\|^2 = 2 - 2\Re(v_{0k}^*v_k) = 2 - 2|v_{0k}^*v_k| \leq 2 \left(1 - \sqrt{1 - \frac{4\|\Delta\|^2}{\gamma(0)^2}} \right) \leq \frac{8\|\Delta\|^2}{\gamma(0)^2}$$

using that $\forall y \in [0, 1]$, $1 - \sqrt{1 - y^2} \leq y^2$. This shows the first inequality of the lemma.

Let us now show the control on $w_k := w_k^*(\alpha) - v_k$ – the second inequality of the lemma. Using the unitary basis $(v_{0j})_j$ to compute the norm, we have $\|w_k^* - v_{0k}^*\|^2 = \sum_{j \neq k} |w_k^*v_{0j}|^2 + |w_k^*v_{0k} - 1|^2$ and since by definition $w_k^*v_j = \mathbb{1}_{j=k}$, we can bound each term as

$$\forall j \neq k, |w_k^*v_{0j}| \leq |w_k^*v_j| + |w_k^*(v_{0j} - v_j)| \leq 0 + \|w_k^*\| \frac{2\sqrt{2}\|\Delta\|}{\gamma(0)}$$

$$\text{and } |w_k^*v_{0k} - 1| = |w_k^*v_k - 1 + w_k^*(v_{0k} - v_k)| \leq \|w_k^*\| \|v_{0k} - v_k\| \leq \|w_k^*\| \frac{2\sqrt{2}\|\Delta\|}{\gamma(0)}.$$

So in total, $\|w_k^* - v_{0k}^*\|^2 \leq d \cdot \|w_k^*\|^2 \frac{8\|\Delta\|^2}{\gamma(0)^2}$, as announced.

The second part of the lemma follows by noting that $\|v_k\| \|w_k\| = 1 \cdot \|w_k\|$ is bounded by

$$\begin{aligned} \|w_k\| &\leq \|v_{0k}\| + \|w_k - v_{0k}\| \leq 1 + \sqrt{d} \|w_k\| \frac{2\sqrt{2}\|\Delta\|}{\gamma(0)} \\ \implies \|w_k\| &\leq \frac{1}{1 - \sqrt{d} \cdot \frac{2\sqrt{2}\|\Delta\|}{\gamma(0)}} = \frac{\gamma(0)}{\gamma(0) - 2\sqrt{2}\sqrt{d}\|\Delta\|} \end{aligned}$$

for all α such that the denominator in the second line is positive, as announced. \square

By combining the three above lemmas, we have that under the conditions of [Lemma D.2](#), if additionally M_0 is normal, then for all α in a neighborhood of zero such that $\|M_\alpha - M_0\| \leq \frac{\gamma(0)}{4\sqrt{2d}}$, one can check that $\gamma(\alpha) \geq \frac{1}{2}\gamma(0)$ and $\chi(\alpha) \leq 2$ and so

$$\frac{1}{6} \max_k |\lambda_k^{\dots}| \leq \frac{1}{3} \|\ddot{M}\| + 8d\gamma(0)^{-1} \|\dot{M}\| \cdot \|\ddot{M}\| + 32d^2\gamma(0)^{-2} \|\dot{M}\|^3.$$

In particular, for a matrix $M_\alpha = M_0 + \alpha M_1 + \frac{\alpha^2}{2} M_2$ with M_0 normal, the right-hand side translates exactly to the bound on $|r_j|$ announced in [Prop. D.1](#). So that proposition follows directly from [\(D.2\)](#) and the above discussion.

E Local convergence of equality-constrained Mirror Flow

This appendix contains results used in [App. F](#).

Let \mathcal{X}, \mathcal{Y} convex subsets of \mathbb{R}^n resp. \mathbb{R}^m and $\phi_x : \mathcal{X} \rightarrow \mathbb{R}$, $\phi_y : \mathcal{Y} \rightarrow \mathbb{R}$ strictly convex and differentiable. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\phi : \mathcal{Z} \rightarrow \mathbb{R}$ with $\phi(x, y) = \phi_x(x) + \phi_y(y)$. Consider a twice continuously differentiable min-max objective $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, denote $g(z) = \text{Diag}(I_n, -I_m) \nabla f(z)$ and $M(z) = \nabla g(z)$. Throughout this appendix we make the following assumption.

Assumption 1. *The constraint set is defined by equalities: $\mathcal{Z} = \{z \in \mathbb{R}^{n+m}; Az = b\} = z_b + \text{Ker } A$, where z_b is any solution of $Az_b = b$. Furthermore, ϕ is strictly convex and three times differentiable.*

Definition E.1 ([\[AW20, Proposition 1\]](#)). For an initial point $z^0 \in \mathcal{Z}$, the mirror flow (MF) with link function ϕ is the unique curve $z(t)$ such that $z(0) = z^0$ and

$$\frac{dz}{dt} = -\Phi_z^{-1} P_z g(z) =: -g_{\text{eff}}(z),$$

where Φ_z denotes the Hessian of ϕ at z and $P_z := I - A^\top [A\Phi_z^{-1}A^\top]^{-1} A\Phi_z^{-1}$.

One can check that $P_z^2 = P_z$ and that $P_z \Phi_z = \Phi_z P_z^\top$. Furthermore, $P_z A^\top = 0$, i.e., $\text{Im } P_z^\top \subset \text{Ker } A$, and in fact $\text{Im } P_z^\top = \text{Ker } A$ since $(\text{Im } P_z^\top)^\perp = \text{Ker } P_z = \text{Im}(I - P_z) \subset \text{Im } A^\top = (\text{Ker } A)^\perp$. In particular the MF preserves the constraint set, since $\frac{d}{dt}(Az - b) = -AP^\top \Phi_z^{-1}g(z) = 0$. As a consequence, g_{eff} can be seen as an operator from the affine space $\mathcal{Z} = z_b + \text{Ker } A$ to itself.

Lemma E.1. *The Jacobian of $g_{\text{eff}} : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^{n+m}$ at a local NE z^* is equal to $\Phi_{z^*}^{-1} P_{z^*} \cdot M(z^*)$. Furthermore, the Jacobian of $g_{\text{eff}} : \mathcal{Z} \rightarrow \mathcal{Z}$ (seen as an operator between affine spaces) at z^* is*

$$M_{\text{eff}}(z^*) = \Phi_{z^*}^{-1} P_{z^*} \cdot M(z^*) P_{z^*}^\top$$

which is a linear operator from $\text{Ker } A$ to itself.

Proof. In this proof, write for concision $P = P_z$ and $\Phi^{-1} = \Phi_z^{-1}$. Using Einstein's summation notation, pose

$$P_i^j = P = I - A^\top [A\Phi^{-1}A^\top]^{-1} A\Phi^{-1} \quad \text{and} \quad Q^{ij} = Q := \Phi^{-1}P.$$

In particular $g_{\text{eff}}(z) = Qg$. Now

$$\nabla\{Qg\}^i_j = \frac{\partial[Qg]^i}{\partial z^j} = (Q\nabla g)^i_j + \frac{\partial Q^{ik}}{\partial z^j} g_k. \quad (\text{E.1})$$

Using formula (59) from [PP+08]: $\frac{\partial Y^{-1}}{\partial x} = -Y^{-1} \frac{\partial Y}{\partial x} Y^{-1}$, we have that

$$\begin{aligned} \frac{\partial(\Phi^{-1})^{il}}{\partial z^j} &= -(\Phi^{-1})^{ia} K_{abj} (\Phi^{-1})^{bl} \\ \frac{\partial P_l^k}{\partial z^j} &= -(A^\top [A\Phi^{-1}A^\top]^{-1} A)_l^s K_{stj} (\Phi^{-1}A^\top [A\Phi^{-1}A^\top]^{-1} A\Phi^{-1})^{tk} \\ &\quad - (A^\top)_{la} \frac{\partial\{[A\Phi^{-1}A^\top]^{-1}\}^{ab}}{\partial z^j} (A\Phi^{-1})_b^k \end{aligned}$$

where $K = \nabla^3 \phi(z)$, and after calculating and simplifying,

$$\frac{\partial P_l^k}{\partial z^j} = (I - P)_l^s K_{stj} (\Phi^{-1}P)^{tk}.$$

So

$$\begin{aligned} \frac{\partial Q^{ik}}{\partial z^j} &= (\Phi^{-1})^{il} \frac{\partial P_l^k}{\partial z^j} + \frac{\partial(\Phi^{-1})^{il}}{\partial z^j} P_l^k \\ &= -(\Phi^{-1}P)^{is} K_{stj} (\Phi^{-1}P)^{tk} = -Q^{is} K_{stj} Q^{tk} \end{aligned}$$

$$\text{and finally } \frac{\partial Q^{ik}}{\partial z^j} g_k = -Q^{is} K_{stj} (Qg)^t.$$

Now $g_{\text{eff}}(z^*) = 0$, so $Qg = 0$ at z^* . The first part of the lemma follows by substituting into (E.1).

To check the second part of the lemma, simply use that $\Phi_{z^*}^{-1} P_{z^*} = P_{z^*}^\top \Phi_{z^*}^{-1}$ and that $P_{z^*}^\top$ is a projector onto the kernel of A , as remarked under Definition E.1. \square

The next lemma is a special case of the stable manifold theorem [Per13, Section 2.7].

Lemma E.2. *Let $\mathcal{Z} = z_b + T\mathcal{Z}$ an affine subspace of \mathbb{R}^d . Let a continuously differentiable operator $g : \mathcal{Z} \rightarrow \mathcal{Z}$ and denote its Jacobian as $\nabla g : \mathcal{Z} \rightarrow T\mathcal{Z} \times T\mathcal{Z}$. Suppose there exists z^* such that $g(z^*) = 0$ and $\min_{\lambda \in \text{Sp}(\nabla g(z^*))} \Re(\lambda) > 0$.*

Then for any $\varepsilon > 0$, there exists $C > 0$ and a relative neighborhood of z^ such that, for any z^0 in that neighborhood, the flow $\frac{dz}{dt} = -g(z)$ converges exponentially to z^* with*

$$\|z(t) - z^*\|^2 \leq C \|z^0 - z^*\|^2 \exp\left(-2 \left[\min_{\lambda \in \text{Sp}(\nabla g(z^*))} \Re(\lambda) - \varepsilon \right] t\right).$$

Proposition E.3. *If $\widetilde{M} := \Phi_{z^*}^{-\frac{1}{2}} P_{z^*} \cdot M(z^*) \cdot P_{z^*}^\top \Phi_{z^*}^{-\frac{1}{2}}$ is invertible and satisfies the equivalent conditions of Thm. 2.1, then MF converges locally exponentially to z^* at a rate $\widetilde{\mu}_{\widetilde{M}}$.*

Proof. \widetilde{M} is similar to $M_{\text{eff}}(z^*)$. In particular they have the same eigenvalues. So the proposition follows immediately from the above two lemmas. \square

Remark E.1. The matrix \widetilde{M} can be interpreted as

$$\widetilde{M} = \underbrace{\Phi_{z^*}^{-\frac{1}{2}} P_{z^*} \Phi_{z^*}^{\frac{1}{2}}}_{=: R_{z^*}} \cdot \Phi_{z^*}^{-\frac{1}{2}} M(z^*) \Phi_{z^*}^{-\frac{1}{2}} \cdot \underbrace{\Phi_{z^*}^{\frac{1}{2}} P_{z^*}^\top \Phi_{z^*}^{-\frac{1}{2}}}_{=: R_{z^*}^\top}.$$

Note that R_{z^*} is an orthogonal projection. Furthermore, the central factor consists in a transformation of $M(z^*)$ that is compatible with the block structure:

- If we write $M(z^*) = S + A$ with S symmetric and A antisymmetric, then $\Phi_{z^*}^{-\frac{1}{2}} M(z^*) \Phi_{z^*}^{-\frac{1}{2}} = \Phi_{z^*}^{-\frac{1}{2}} S \Phi_{z^*}^{-\frac{1}{2}} + \Phi_{z^*}^{-\frac{1}{2}} A \Phi_{z^*}^{-\frac{1}{2}}$ and the first term is symmetric and the second term is antisymmetric.
- If we write $M(z^*) = \begin{bmatrix} Q & P \\ -P^\top & R \end{bmatrix}$ with Q, R symmetric, then
$$\Phi_{z^*}^{-\frac{1}{2}} M(z^*) \Phi_{z^*}^{-\frac{1}{2}} = \begin{bmatrix} \Phi_{x^*}^{-\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \Phi_{y^*}^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} Q & P \\ -P^\top & R \end{bmatrix} \begin{bmatrix} \Phi_{x^*}^{-\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \Phi_{y^*}^{-\frac{1}{2}} \end{bmatrix} = \begin{bmatrix} \Phi_{x^*}^{-\frac{1}{2}} Q \Phi_{x^*}^{-\frac{1}{2}} & \Phi_{x^*}^{-\frac{1}{2}} P \Phi_{y^*}^{-\frac{1}{2}} \\ -\left(\Phi_{x^*}^{-\frac{1}{2}} P \Phi_{y^*}^{-\frac{1}{2}}\right)^\top & \Phi_{y^*}^{-\frac{1}{2}} R \Phi_{y^*}^{-\frac{1}{2}} \end{bmatrix}$$
is of the same form, where Φ_{x^*} is the Hessian of ϕ_x at x^* and similarly for Φ_{y^*} .

We state an analogous result for Mirror Descent-Ascent (MDA) and its variants, Mirror Prox (MP) and Bregman PP. Its proof, omitted for brevity, involves exactly the same ideas as for MF.

Proposition E.4. *Under the assumptions of Prop. E.3, if additionally ϕ is strongly convex, then MDA, MP and Bregman PP converge locally exponentially to z^* for any small enough step-size at a rate $\eta \tilde{\mu}_{\widetilde{M}} + O(\eta^2)$.*

F Details for Sec. 4

In this appendix we provide details for Fig. 2, derive the first-order term (in η) in the local convergence rate of MP and EG, and discuss a counter-example. For our purpose it is sufficient to consider only the continuous-time flows of those algorithms.

F.1 Details for Fig. 2

In the numerical experiment reported in Fig. 2, we used the same random payoff functions as in [WC22, Section 4]:

$$f(x, y) = \Re \sum_{-K \leq k \leq K} \sum_{-L \leq l \leq L} c_{kl} e^{2\pi i(kx + ly)} \quad \text{for } (c_{kl}) \in \mathbb{C}^{(2K+1) \times (2L+1)}, \quad (\text{F.1})$$

with $K = L = 2$ and $\Re(c_{kl}), \Im(c_{kl})$ drawn independently from the standard Gaussian distribution. We measured the distance between the solution $z^* = (a^*, x^*, b^*, y^*)$ and the iterates z^k by

$$\|z - z^*\|^2 := \|a - a^*\|^2 + \|b - b^*\|^2 + \|x - x^*\|^2 + \|y - y^*\|^2.$$

For z in a neighborhood of z^* , $\|z - z^*\|$ is an upper bound on the duality gap of the original MNE problem up to a constant dependent only on f [WC22, Proposition 3.1 and Claim C.3].

F.2 First-order term in the local convergence rate of MP

Denote the joint variable as $z = (a, b) \in \Delta_N \times \Delta_M$ and let z^* such that $(\mu^*, \nu^*) = (\sum_I a_I^* \delta_{x_I^*}, \sum_J b_J^* \delta_{y_J^*})$ is the MNE. The continuous-time flow of MP is MF (Definition E.1) with $\phi(z) = \sum_I z_I \log z_I$, $A = \begin{bmatrix} \mathbf{1}_N^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_M^\top \end{bmatrix} \in \mathbb{R}^{2 \times (N+M)}$ and $b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. So by Prop. E.4, the first-order term in the local convergence rate of MP is $\eta \tilde{\mu}_{M_{\text{MP}}}$ for some M_{MP} which we now compute.

Following the notations of App. E, we have $\Phi_{z^*} = \text{Diag} \left(\left(\frac{1}{a_I^*} \right)_I, \left(\frac{1}{b_J^*} \right)_J \right)$, and by straightforward calculations $P_{z^*} = \text{Diag} (I_N - \mathbf{1}_N (a^*)^\top, I_M - \mathbf{1}_M (b^*)^\top)$, so that $R_{z^*} = \Phi_{z^*}^{-1/2} P_{z^*} \Phi_{z^*}^{1/2} =$

$\text{Diag}\left(I_N - \sqrt{a^*}\sqrt{a^{*\top}}, I_M - \sqrt{b^*}\sqrt{b^{*\top}}\right)$. In order to project out the constraints explicitly, let Π_a any matrix in $\mathbb{R}^{(N-1)\times N}$ with orthonormal rows, i.e., $\Pi_a\Pi_a^\top = I_{N-1}$, and such that $\Pi_a^\top\Pi_a = I_N - \sqrt{a^*}\sqrt{a^{*\top}}$, and likewise for $\Pi_b \in \mathbb{R}^{(M-1)\times M}$. For example the rows of Π_a may be obtained by completing the unit vector $\sqrt{a^*}$ into an orthonormal basis and removing $\sqrt{a^*}$ from the basis. Then R_{z^*} as a linear projector from \mathbb{R}^{N+M} to $\{\sqrt{a^*}\}^\perp \times \{\sqrt{b^*}\}^\perp \simeq \mathbb{R}^{N+M-2}$ can be written as the matrix $R_{z^*} = \text{Diag}(\Pi_a, \Pi_b)$.⁹ So \widetilde{M} as a linear operator over $\{\sqrt{a^*}\}^\perp \times \{\sqrt{b^*}\}^\perp$ can be written as the matrix, denoting $D_a = \text{Diag}(\sqrt{a^*})$ and $D_b = \text{Diag}(\sqrt{b^*})$ for concision,

$$\begin{aligned} M_{\text{MP}} &:= R_{z^*}\Phi_{z^*}^{-1/2}M(z^*)\Phi_{z^*}^{-1/2}R_{z^*}^\top \\ &= \text{Diag}(\Pi_a, \Pi_b)\text{Diag}\left(\sqrt{a^*}, \sqrt{b^*}\right)\begin{bmatrix} \mathbf{0} & P \\ -P^\top & \mathbf{0} \end{bmatrix}\text{Diag}\left(\sqrt{a^*}, \sqrt{b^*}\right)\text{Diag}(\Pi_a^\top, \Pi_b^\top) \\ &= \begin{bmatrix} \mathbf{0} & \Pi_a D_a P D_b \Pi_b^\top \\ -\Pi_b D_b P^\top D_a \Pi_a^\top & \mathbf{0} \end{bmatrix}. \end{aligned}$$

F.3 First-order term in the local convergence rate of EG (conic particle methods)

As in the main text, we write ‘‘EG’’ to refer to the Conic Particle Mirror Prox algorithm of [WC22].

Denote the joint variable as $z = (a, x, b, y) \in \Delta_N \times (\mathbb{T}^1)^N \times \Delta_M \times (\mathbb{T}^1)^M$ and let any z^* such that $(\mu^*, \nu^*) = (\sum_I a_I^* \delta_{x_I^*}, \sum_J b_J^* \delta_{y_J^*})$ is the MNE. The flow is given by the system of ODEs

$$\begin{aligned} \frac{da}{dt} &= -\Phi_a^{-1}P_a\nabla_a F(z) & \text{and} & & \frac{db}{dt} &= \Phi_b^{-1}P_b\nabla_b F(z) & (\text{F.2}) \\ \frac{dx}{dt} &= -\gamma\text{Diag}\left(\frac{1}{a_I}\right)\nabla_x F(z) & & & \frac{dy}{dt} &= \gamma\text{Diag}\left(\frac{1}{b_J}\right)\nabla_y F(z). \end{aligned}$$

Here $F(z) = \sum_I \sum_J a_I b_J f(x_I, y_J)$, γ is a constant parameter, $\Phi_a = \text{Diag}\left(\frac{1}{a_I}\right)$ and $P_a = I_N - \mathbf{1}_N a^\top$, and Φ_b and P_b are defined similarly.

In general, the flow (F.2) does not match the structure of Mirror Flow because of the factor ‘‘ $\text{Diag}\left(\frac{1}{a_I}\right)$ ’’ in the equation for $\frac{dx}{dt}$ [GWS21, Section 2.4]. However, by adapting the reasoning of App. E – namely only Lemma E.1 needs to be adapted – it is easy to show that the statement of Prop. E.3 holds also for this dynamics. By the same adaptation one can show that the statement of Prop. E.4 holds also for EG. Hence the first-order term in the local convergence rate of EG is $\eta\tilde{\mu}_{M_\gamma}$ for some M_γ which we now compute.

The statement of Prop. E.3 applies to $\widetilde{M} = R_{z^*} \cdot \Phi_{z^*}^{-1/2}M(z^*)\Phi_{z^*}^{-1/2} \cdot R_{z^*}^\top$ where $\Phi_{z^*} = \text{Diag}\left(\left(\frac{1}{a_I^*}\right)_I, \frac{1}{\gamma}a^*, \left(\frac{1}{b_J^*}\right)_J, \frac{1}{\gamma}b^*\right)$ and $P_{z^*} = \text{Diag}(P_a, I_N, P_b, I_M)$ so that $R_{z^*} = \Phi_{z^*}^{-1/2}P_{z^*}\Phi_{z^*}^{1/2} = \text{Diag}\left(I_N - \sqrt{a^*}\sqrt{a^{*\top}}, I_N, I_M - \sqrt{b^*}\sqrt{b^{*\top}}, I_M\right)$. In order to project out the constraints explicitly, let $\Pi_a \in \mathbb{R}^{(N-1)\times N}$, $\Pi_b \in \mathbb{R}^{(M-1)\times M}$ the same matrices as in the previous subsection. Then R_{z^*} as a linear projector from \mathbb{R}^{2N+2M} to $\{\sqrt{a^*}\}^\perp \times \mathbb{R}^N \times \{\sqrt{b^*}\}^\perp \times \mathbb{R}^M \simeq \mathbb{R}^{2N+2M-2}$ can be written as the matrix $R_{z^*} = \text{Diag}(\Pi_a, I_N, \Pi_b, I_M)$. Moreover by [WC22, Claim C.2], dropping superscript *’s for concision only in this equation,

$$M(z^*) = \begin{bmatrix} \mathbf{0} & \mathbf{0} & P & \partial_y P \text{Diag}(b) \\ \mathbf{0} & \text{Diag}(a)\text{Diag}(\partial_{xx}^2 P b) & \text{Diag}(a)\partial_x P & \text{Diag}(a)\partial_{xy}^2 P \text{Diag}(b) \\ -P^\top & -(\text{Diag}(a)\partial_x P)^\top & \mathbf{0} & \mathbf{0} \\ -(\partial_y P \text{Diag}(b))^\top & -(\text{Diag}(a)\partial_{xy}^2 P \text{Diag}(b))^\top & \mathbf{0} & -\text{Diag}(b)\text{Diag}(\partial_{yy}^2 P^\top a) \end{bmatrix}$$

where $[\partial_x P]_{IJ} = \partial_x f(x_I^*, y_J^*)$, and likewise for $\partial_y P$, $\partial_{xx}^2 P$, $\partial_{yy}^2 P$, $\partial_{xy}^2 P$. So, finally, $\widetilde{M} = R_{z^*}\Phi_{z^*}^{-1/2}M(z^*)\Phi_{z^*}^{-1/2}R_{z^*}^\top$ as a linear operator over $\{\sqrt{a^*}\}^\perp \times \mathbb{R}^N \times \{\sqrt{b^*}\}^\perp \times \mathbb{R}^M$ can be written

⁹The expression ‘‘ $\text{Diag}(\Pi_a, \Pi_b)$ ’’, as well as ‘‘ $\text{Diag}(\Pi_a, I_N, \Pi_b, I_M)$ ’’ in Sec. F.3, constitutes a slight abuse of our notation ‘‘Diag’’, since Π_a and Π_b are not square matrices. To remove any ambiguity: by $\text{Diag}(\Pi_a, \Pi_b)$ we mean the matrix $\begin{bmatrix} \Pi_a & \mathbf{0}_{(N-1)\times M} \\ \mathbf{0}_{(M-1)\times N} & \Pi_b \end{bmatrix}$, and similarly for $\text{Diag}(\Pi_a, I_N, \Pi_b, I_M)$.

as the matrix, denoting $D_a = \text{Diag}(\sqrt{a^*})$ and $D_b = \text{Diag}(\sqrt{b^*})$ for concision,

$$M_\gamma := \begin{bmatrix} \mathbf{0} & \mathbf{0} & \Pi_a D_a P D_b \Pi_b^\top & \sqrt{\gamma} \Pi_a D_a [\partial_y P] D_b \\ \mathbf{0} & \gamma \text{Diag}(\partial_{xx}^2 P b^*) & \sqrt{\gamma} D_a [\partial_x P] D_b \Pi_b^\top & \gamma D_a [\partial_{xy}^2 P] D_b \\ -(*)^\top & -(*)^\top & \mathbf{0} & \mathbf{0} \\ -(*)^\top & -(*)^\top & \mathbf{0} & -\gamma \text{Diag}(\partial_{yy}^2 P^\top a^*) \end{bmatrix}.$$

F.4 An example where the first-order term in the local convergence rate of EG is zero

Consider the payoff function defined by (F.1) with $c_{20} = c_{02} = -i$, $c_{11} = 2$ and $c_{kl} = 0$ otherwise, i.e., $f(x, y) = \sin(4\pi x) + \sin(4\pi y) + 2 \cos(2\pi x + 2\pi y)$. As shown in [WC22, Example 4.1], the MNE is unique and given by $a^* = b^* = (\frac{1}{2}, \frac{1}{2})$, $x^* = (\frac{3}{8}, \frac{7}{8})$, and $y^* = (\frac{1}{8}, \frac{5}{8})$. So we can compute M_γ explicitly in this case: we find

$$P = \begin{pmatrix} -2 & 2 \\ 2 & -2 \end{pmatrix}, \partial_x P = \partial_y P = 0, \partial_{xx}^2 P \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} = \partial_{yy}^2 P \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} = \begin{pmatrix} 16\pi^2 & \\ & 16\pi^2 \end{pmatrix}, \partial_{xy} P = \begin{pmatrix} 8\pi^2 & -8\pi^2 \\ -8\pi^2 & 8\pi^2 \end{pmatrix}$$

and $D_a = D_b = \frac{1}{\sqrt{2}} I$, and so (for a certain choice of Π_a and Π_b , each of which is anyway determined up to a sign)

$$M_\gamma = \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & -2 & 0 & 0 \\ 0 & \gamma(4\pi)^2 & 0 & 0 & \gamma(2\pi)^2 & -\gamma(2\pi)^2 \\ 0 & 0 & \gamma(4\pi)^2 & 0 & -\gamma(2\pi)^2 & \gamma(2\pi)^2 \\ \hline 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\gamma(2\pi)^2 & \gamma(2\pi)^2 & 0 & \gamma(4\pi)^2 & 0 \\ 0 & \gamma(2\pi)^2 & -\gamma(2\pi)^2 & 0 & 0 & \gamma(4\pi)^2 \end{array} \right].$$

This matrix clearly does not satisfy condition (iii) of Thm. 2.1, so $\tilde{\mu}_{M_\gamma} = 0$.

F.5 Proof of Prop. 4.1

For ease of reference, we restate the proposition below.

Proposition 4.1. *Let S_2 symmetric and A_0, A_1, A_2 antisymmetric real matrices of the form*

$$S_2 = \left[\begin{array}{c|c} \mathbf{0} & \\ \hline * & \mathbf{0} \\ \hline & * \end{array} \right], A_0 = \left[\begin{array}{c|c} * & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \\ \hline * & \mathbf{0} \end{array} \right], A_1 = \left[\begin{array}{c|c} \mathbf{0} & * \\ \hline * & \mathbf{0} \\ \hline & \end{array} \right], A_2 = \left[\begin{array}{c|c} \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & * \\ \hline & \end{array} \right]$$

and $M_\gamma = \gamma S_2 + A_0 + \sqrt{\gamma} A_1 + \gamma A_2$ for all $\gamma > 0$. Then $\tilde{\mu}_{M_\gamma} = O(\gamma^2)$ as $\gamma \rightarrow 0$.

Pose $\alpha = \sqrt{\gamma}$, $M_0 = A_0 + A_\varepsilon$, $M_1 = A_1$ and $M_2 = 2(A_2 + S_2)$, where A_ε is any antisymmetric matrix such that $\|A_\varepsilon\| \leq \varepsilon$ and M_0 has distinct eigenvalues. We will prove the proposition by applying the spectral expansions of Sec. D.5 to the matrix curve $M_\alpha = M_0 + \alpha M_1 + \frac{\alpha^2}{2} M_2$.

Adopting the notations of that section, we have the expansion for the eigenvalues $\lambda_k(\alpha)$ of M_α :

$$\lambda_k(\alpha) = \lambda_k(0) + \alpha \dot{\lambda}_k(0) + \frac{\alpha^2}{2} \ddot{\lambda}_k(0) + \frac{\alpha^3}{3!} \dddot{\lambda}_k(0) + \frac{\alpha^4}{4!} \dots \lambda_k(0) + O(\alpha^5),$$

with $\{\lambda_k(0)\}_k = \text{Sp}(M_0) \subset i\mathbb{R}$ by antisymmetry, $\dot{\lambda}_k(0) = v_{0k}^* A_1 v_{0k} \in i\mathbb{R}$ – where v_{0k} are the eigenvectors of M_0 – by normality of M_0 and antisymmetry of A_1 , and

$$\ddot{\lambda}_k(0) = 2v_{0k}^* (A_2 + S_2) v_{0k} + 2 \sum_{j \neq k} \underbrace{\frac{(v_{0k}^* A_1 v_{0j})(v_{0j}^* A_1 v_{0k})}{\lambda_k(0) - \lambda_j(0)}}_{\in i\mathbb{R}}$$

since $v_{0k}^* A_1 v_{0j} = -(v_{0j}^* A_1 v_{0k})^*$, and

$$\begin{aligned} \ddot{\lambda}_k(0) &= 3 \sum_{j \neq k} \frac{1}{\lambda_k(0) - \lambda_j(0)} [(v_{0k}^* A_1 v_{0j})(v_{0j}^* M_2 v_{0k}) + (v_{0k}^* M_2 v_{0j})(v_{0j}^* A_1 v_{0k})] \\ &+ 6 \underbrace{\sum_{j, l \neq k} \frac{(v_{0k}^* A_1 v_{0j})(v_{0j}^* A_1 v_{0l})(v_{0l}^* A_1 v_{0k})}{(\lambda_k(0) - \lambda_j(0))(\lambda_k(0) - \lambda_l(0))}}_{\in i\mathbb{R}} - 6 \sum_{j \neq k} \underbrace{\frac{(v_{0k}^* A_1 v_{0j})(v_{0j}^* A_1 v_{0k})}{(\lambda_k(0) - \lambda_j(0))^2}}_{\in i\mathbb{R}} v_{0k}^* A_1 v_{0k} \end{aligned}$$

as one can check by computing the convex conjugate of each of the underbraced expressions.

Now, let \mathcal{K} the set of indices corresponding to the non-zero eigenvalues of A_0 , and $\mathcal{K}^c = \{1, \dots, d\} \setminus \mathcal{K}$. Note that the eigenvectors of A_0 are of the form

$$\forall k \in \mathcal{K}, \tilde{v}_k = \begin{pmatrix} * \\ 0 \\ * \\ 0 \end{pmatrix} \quad \text{and} \quad \forall h \in \mathcal{K}^c, \tilde{v}_h = \begin{pmatrix} 0 \\ * \\ 0 \\ * \end{pmatrix}.$$

Furthermore, note that

$$\begin{aligned} A_1 \begin{pmatrix} * \\ 0 \\ * \\ 0 \end{pmatrix} &= \begin{pmatrix} 0 \\ * \\ 0 \\ * \end{pmatrix}, & S_2 \begin{pmatrix} * \\ 0 \\ * \\ 0 \end{pmatrix} &= \mathbf{0}, & A_2 \begin{pmatrix} * \\ 0 \\ * \\ 0 \end{pmatrix} &= \mathbf{0}, \\ A_1 \begin{pmatrix} 0 \\ * \\ 0 \\ * \end{pmatrix} &= \begin{pmatrix} * \\ 0 \\ * \\ 0 \end{pmatrix}, & A_2 \begin{pmatrix} 0 \\ * \\ 0 \\ * \end{pmatrix} &= \begin{pmatrix} 0 \\ * \\ 0 \\ * \end{pmatrix}, & A_2 \begin{pmatrix} 0 \\ * \\ 0 \\ * \end{pmatrix} &= \begin{pmatrix} 0 \\ * \\ 0 \\ * \end{pmatrix}. \end{aligned}$$

Using this structure, one can check that for all $k \in \mathcal{K}$, $\Re \tilde{\lambda}_k(0) = o_\varepsilon(1)$ and $\Re \ddot{\lambda}_k(0) = o_\varepsilon(1)$.

Thus, the spectrum of M_α consists of eigenvalues (corresponding to indices $h \in \mathcal{K}^c$) with $\Re \tilde{\lambda}_h(0)$ non-zero a priori, in which case $\Re \lambda_h(\alpha) = \Theta(\alpha^2) = \Theta(\gamma)$, and of eigenvalues (corresponding to indices $k \in \mathcal{K}$) with $\Re \lambda_k(\alpha) = O(\alpha^4) + o_\varepsilon(1) = O(\gamma^2) + o_\varepsilon(1)$. By letting $\varepsilon \rightarrow 0$ and using that eigenvalues are continuous, this shows that $\tilde{\mu}_{M_\gamma} = O(\gamma^2)$ as $\gamma \rightarrow 0$.