OJMO

Open Journal of Mathematical Optimization

Guillaume Wang & Lénaïc Chizat

An Exponentially Converging Particle Method for the Mixed Nash Equilibrium of Continuous Games

Volume 6 (2025), article no. 1 (66 pages)

https://doi.org/10.5802/ojmo.37

Article submitted on February 1, 2024, revised on August 28, 2024,
accepted on September 16, 2024.

CENTRE MERSENNE

Open Journal of Mathematical Optimization is a member of the
Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/
e-ISSN: 2777-5860

An Exponentially Converging Particle Method for the Mixed Nash Equilibrium of Continuous Games

Guillaume Wang
Institute of Mathematics, École polytechnique fédérale de Lausanne (EPFL), Station Z, CH-1015 Lausanne

Lénaïc Chizat
Institute of Mathematics, École polytechnique fédérale de Lausanne (EPFL), Station Z, CH-1015 Lausanne

Abstract

We consider the problem of computing mixed Nash equilibria of two-player zero-sum games with continuous sets of pure strategies and with first-order access to the payoff function. This problem arises for example in game-theory-inspired machine learning applications, such as distributionally-robust learning. In those applications, the strategy sets are high-dimensional and thus methods based on discretisation cannot tractably return high-accuracy solutions. In this paper, we introduce and analyze a particle-based method that enjoys guaranteed local convergence for this problem. This method consists in parametrizing the mixed strategies as atomic measures and applying proximal point updates to both the atoms' weights and positions. It can be interpreted as an implicit time discretization of the "interacting" Wasserstein–Fisher–Rao gradient flow.

We prove that, under non-degeneracy assumptions, this method converges at an exponential rate to the exact mixed Nash equilibrium from any initialization satisfying a natural notion of closeness to optimality. We illustrate our results with numerical experiments and discuss applications to max-margin and distributionally-robust classification using two-layer neural networks, where our method has a natural interpretation as a simultaneous training of the network's weights and of the adversarial distribution.

Digital Object Identifier 10.5802/ojmo.37

1 Introduction

Consider the min-max, or saddle-point, optimization problem

$$\min_{\mu\in\mathcal{P}(\mathcal{X})}\max_{\nu\in\mathcal{P}(\mathcal{Y})}\int_{\mathcal{X}}\int_{\mathcal{Y}} f(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y) \tag{1}$$

where $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ are the sets of probability distributions over the sets of *pure strategies* $\mathcal{X}$ and $\mathcal{Y}$, and $f : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ is the *payoff function*. In the language of game theory, $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ are the sets of *mixed strategies* and solutions (μ^*,ν^*) of (1) are the *mixed Nash equilibria* (MNEs) of the two-player zero-sum game $(f,\mathcal{X},\mathcal{Y})$. The conditions for the existence of a MNE are well-known since the 1950s. In particular, by Glicksberg's theorem [14] (later generalized as Sion's minimax theorem [33]) a MNE always exists if $\mathcal{X}$ and $\mathcal{Y}$ are finite, or if $\mathcal{X}$ and $\mathcal{Y}$ are compact and f is continuous.

Many methods have been proposed to compute MNEs given zeroth-order access to f, including in noisy, online or decentralized settings [6, 25]. Those methods are typically derived and studied for finite games (i.e., with finite strategy sets). When $\mathcal{X}$ and $\mathcal{Y}$ are continuous (say, differentiable manifolds) and we additionally have access to the gradients of f, it is still possible to reduce the game to a finite one by discretization, but this not only wastes the gradient information, it may also incur a prohibitively high discretization cost when $\mathcal{X}$ and $\mathcal{Y}$ have high dimension. In this paper, we thus study how to efficiently compute a MNE of $(f,\mathcal{X},\mathcal{Y})$ to a high accuracy, when the strategy sets $\mathcal{X},\mathcal{Y}$ are continuous and given first-order access to f.

Particle methods

In this setting, a possible strategy is to parametrize the unknowns via

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}, \qquad \nu = \sum_{j=1}^{m} b_j \delta_{y_j},$$

and to use gradient methods to solve the reparametrized min-max problem

$$\min_{\substack{a\in\Delta_n \\ x\in\mathcal{X}^n}} \max_{\substack{b\in\Delta_m \\ y\in\mathcal{Y}^m}} \left\{ F_{n,m}\left((a,x),(b,y)\right) \;\coloneqq\; \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j f(x_i,y_j) \right\}, \tag{2}$$

where $\Delta_n \coloneqq \left\{a \in \mathbb{R}_+^n ; \sum_{i=1}^n a_i = 1\right\}$ is the n-simplex. This approach takes its inspiration from the recent guarantees obtained for "weighted particle methods" for convex minimization on the space of measures. In particular, adapting the Conic[1] Particle Gradient Descent (CP-GD) method [8] to the constrained min-max context, leads to the *Conic Particle Mirror Descent-Ascent* (CP-MDA) method which defines a sequence of iterates $(a^k, x^k, b^k, y^k)_{k\geq 0}$ by the update rule

$$\begin{cases} a_i^{k+1} \propto a_i^k e^{-\eta \frac{\partial}{\partial a_i} F_{n,m}(a^k,x^k,b^k,y^k)} \\ x_i^{k+1} = x_i^k - \sigma \frac{1}{a_i^k} \frac{\partial}{\partial x_i} F_{n,m}(a^k,x^k,b^k,y^k) \end{cases} \qquad \begin{cases} b_j^{k+1} \propto b_j^k e^{\eta \frac{\partial}{\partial b_j} F_{n,m}(a^k,x^k,b^k,y^k)} \\ y_j^{k+1} = y_j^k + \sigma \frac{1}{b_j^k} \frac{\partial}{\partial y_j} F_{n,m}(a^k,x^k,b^k,y^k). \end{cases} \tag{3}$$

Here $\eta, \sigma > 0$ are step-sizes to be chosen and a^k and b^k are normalized to sum to 1 at each step. (The equations above are for the case where $\mathcal{X}$, $\mathcal{Y}$ are Euclidean and without boundaries; in general a retraction step is needed for the update of x^k, y^k.) In the limit where $\eta, \sigma \to 0$, we obtain a continuous-time dynamics studied by [12] under the name "interacting Wasserstein–Fisher–Rao gradient flow", which they show admits a mean-field limiting dynamics when $n, m \to \infty$ and the initial iterates are randomly independently sampled.

Convergence to mixed Nash equilibria

The reparametrized saddle-point objective $F_{n,m}$ is finite-dimensional, but is unfortunately not convex-concave in general, and there is no known convergence guarantee for CP-MDA. In fact, taking $\sigma = 0$, we recover the Mirror Descent-Ascent algorithm on the finite game $(f, \{x_i^0\}_i, \{y_j^0\}_j)$ (a.k.a. multiplicative weight updates), and it is known that the Bregman divergence of the iterates to the MNE is then non-decreasing [3]! For finite games, this non-convergence issue can be resolved by considering instead the implicit version of the same algorithm, or other methods which can be interpreted as tractable approximations of it [10, 29].

In this paper, we propose the implicit version of CP-MDA, which we call the *Conic Particle Proximal Point* algorithm (CP-PP).

- We show that, if (1) admits a unique and non-degenerate sparse saddle point (μ^*, ν^*), and if CP-PP is initialized close enough to optimality, then it converges to (μ^*, ν^*) at an exponential rate. Note that one can always find such an initialization by sampling sufficiently many particles, setting $\sigma = 0$ and taking the averaged iterate in an initial warm-up phase, see Section 2.3. The convergence is established both for the Nikaido–Isoda error (the natural measure of optimality for min-max problems) and for the Wasserstein–Fisher–Rao distance to (μ^*, ν^*).
- While CP-PP itself is not directly implementable, we also prove in a simplified setting that a computationally efficient approximation of CP-PP, the *Conic Particle Mirror Prox* algorithm (CP-MP), also converges to (μ^*, ν^*) under the same conditions and with the same rate. We observe experimentally that its convergence behavior is the same in the general setting.
- We illustrate our work with numerical experiments, including examples of applications to max-$\mathcal{F}_1$-margin and distributionally-robust classification with two-layer neural networks. We observe experimentally that the explicit method CP-MDA does not always converge (although it does in some cases), so that using an implicit time discretization of the dynamics, like CP-PP, is necessary for convergence in general.

[1] The term "conic" refers to the particular geometry on the space of couples (a_i, x_i) that leads to multiplicative updates on a_i and additive updates on x_i in (3).

1.1 Related work

Infinite-dimensional Mirror Descent-Ascent

For finite strategy sets $\mathcal{X}$ and $\mathcal{Y}$, the min-max problem (1) is finite-dimensional, and the classical Mirror Descent-Ascent algorithm can be applied to obtain convergence of the averaged iterate to a MNE [4]. In [16], the authors use sampling by Langevin dynamics to formulate an implementable version of Mirror Descent-Ascent for continuous strategy sets, which coincides with the bona fide infinite-dimensional Mirror Descent-Ascent algorithm in expectation.

Computing approximate MNEs via regularization

[12], [24] and [23] propose and analyze methods that solve an entropy-regularized variant of (1):

$$\min_{\mu\in\mathcal{P}(\mathcal{X})}\max_{\nu\in\mathcal{P}(\mathcal{Y})}\int_{\mathcal{X}}\int_{\mathcal{Y}} f(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y)+\beta^{-1}H(\mu)-\beta^{-1}H(\nu)$$

where $H(\mu)=\int_{\mathcal{X}}\log\frac{\mathrm{d}\mu}{\mathrm{d}x}(x)\,\mathrm{d}\mu(x)$ denotes negative differential entropy (and $H(\mu)=+\infty$ if μ is not absolutely continuous with respect to Lebesgue measure), and β is a fixed regularization parameter. The methods analyzed in these papers correspond to the continuous-time dynamics called "entropy-regularized interacting Wasserstein gradient flow" in [12]. Qualitative convergence properties are shown in [12], [24] proves convergence to an approximate MNE in the quasi-static regime (i.e., when the step-size used to update μ is infinitely smaller than the one for ν), and [23] proves convergence in a regime with finite timescale separation. The continuous-time guarantees of the aforementioned works can be translated to discrete-time algorithms, thanks to the general framework developed in [18] and [19].

Last-iterate convergence of proximal point methods for min-max optimization

In the optimization and learning community, there has been much interest in general convex-concave min-max problems $\min_x\max_y G(x,y)$. Some works focus on ergodic convergence, that is, convergence of the averaged iterate: $\big(\frac{1}{T}\sum_{k=1}^{T}x^k,\frac{1}{T}\sum_{k=1}^{T}y^k\big)$. For instance, the Mirror Prox and Proximal Point methods were introduced in [30] to attain $O(1/T)$ ergodic convergence for convex-concave $C^{1,1}$ functions, instead of $O(1/\sqrt{T})$ using Mirror Descent-Ascent [4, 5].

Recent works showed that, while Mirror Descent-Ascent may not in fact converge in the last-iterate sense [27], Proximal Point and related methods (e.g. Mirror Prox, Optimistic Mirror Descent-Ascent) do exhibit last-iterate convergence [21]. In the special case of finding MNEs of finite two-player zero-sum games, convergence rates for the last iterate have been derived by [10] and by [37] for Optimistic Mirror Descent-Ascent, under the assumption that the MNE is unique.

It should be noted that, when using particle methods for the problem (1), the averaged iterate $\big(\frac{1}{T}\sum_{k=1}^{T}\mu^k,\frac{1}{T}\sum_{k=1}^{T}\nu^k\big)$ consists of $(n+m)T$ atoms in general. This means that averaging in measure space would result in unacceptably large memory requirements in cases where the domain $\mathcal{X}$ or $\mathcal{Y}$ is large, such as for mixtures of GANs [12]. Another option is to take the average of the (a_i^k,x_i^k) directly, but it would not a priori improve upon the last iterate because $F_{n,m}$ is not convex-concave.

There also exists a growing literature on nonconvex-nonconcave min-max optimization, which focuses on the problem of finding local saddle points or even just stationary points of the gradient descent-ascent flow [11, 13]. Our result are stronger than what one could expect to achieve with techniques from that literature: We are able to find the solution of (1), which gives a global Nash equilibrium of $F_{n,m}$, instead of simply stationary points.

The remainder of this paper is structured as follows. In Section 2, we state the problem, describe the algorithm, and present the main result. In Section 3, we prove the main result. In Section 4, we discuss examples of applications and present numerical experiments. In Section 5 we conclude and proof details are deferred to the appendix.

2 Main result

2.1 Problem setting: Computing MNEs of continuous games

Preliminaries

Let us first recall the general convex-concave min-max optimization framework. For convex sets M, N and a convex-concave function $F : M \times N \to \mathbb{R}$, a *saddle point* or *solution of the min-max problem*

$$\min_{\mu\in M}\max_{\nu\in N} F(\mu,\nu)$$

is any pair (μ^*,ν^*) such that[2]

$$\forall\,\mu\in M,\ \forall\,\nu\in N,\quad F(\mu^*,\nu)\le F(\mu^*,\nu^*)\le F(\mu,\nu^*).$$

The existence of saddle points is guaranteed for example by Sion's minimax theorem when M and N are compact and F is continuous. The goodness of a pair $(\widehat{\mu},\widehat{\nu})$ can be quantified by its *Nikaido–Isoda error* (NI error), a.k.a. duality gap, defined as

$$\mathrm{NI}(\widehat{\mu},\widehat{\nu}*) = \max_{\mu,\nu} F(\widehat{\mu},\nu) - F(\mu,\widehat{\nu}).$$

Indeed, it is easily seen that $\mathrm{NI}(\widehat{\mu},\widehat{\nu})\ge 0$ with equality if and only if $(\widehat{\mu},\widehat{\nu})$ is a saddle point.

As an example, the problem of finding the MNE of a two-player zero-sum game with finite strategy sets $\mathcal{X}=[n]\coloneqq\{1,\dots,n\}$ and $\mathcal{Y}=[m]$ and payoff function $f(i,j)$ can be written as

$$\min_{a\in\Delta_n}\max_{b\in\Delta_m}\left\{F(a,b)=\sum_{i=1}^{n}\sum_{j=1}^{m}a_ib_jf(i,j)=a^\top Mb\right\}$$

where $M_{ij}=f(i,j)$ and $\Delta_n=\left\{a\in\mathbb{R}^n_+;\sum_i a_i=1\right\}\simeq\mathcal{P}([n])$. Since Δ_n and Δ_m are convex compact and $F(a,b)=a^\top Mb$ is convex-concave and continuous, Sion's minimax theorem applies so saddle points exist.

Problem setting and assumptions

The min-max problem we are concerned with in this paper is that of finding a MNE of the continuous game $(f,\mathcal{X},\mathcal{Y})$, as defined in (1):

$$\min_{\mu\in\mathcal{P}(\mathcal{X})}\max_{\nu\in\mathcal{P}(\mathcal{Y})}\left\{F(\mu,\nu)\coloneqq\int_{\mathcal{X}}\int_{\mathcal{Y}}f(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y)\right\},$$

with the following assumptions.

▶ **Assumptions.**

1. The strategy sets are the d_x- resp. d_y-dimensional tori $\mathcal{X}=\mathbb{T}^{d_x}$, $\mathcal{Y}=\mathbb{T}^{d_y}$.
2. The payoff function $f:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ is $C^{2,1}$, i.e., it has Lipschitz-continuous second-order differentials.
3. The MNE (μ^*,ν^*) of (1) is unique.
4. The MNE (μ^*,ν^*) of (1) is sparse, that is,

$$\mathrm{supp}(\mu^*)=\{x_I^*, I\in[n^*]\}\quad\text{and}\quad\mathrm{supp}(\nu^*)=\{y_J^*, J\in[m^*]\}$$

for some $n^*,m^*<\infty$.

Note that points 1 and 2 imply the existence of a MNE by Glicksberg's theorem. Also note that point 4 is guaranteed to hold if the game is *separable*, that is, if f can be written as a finite sum of the form $f(x,y)=\sum_{kl}c_{kl}g_k(x)h_l(y)$ for some $c_{kl}\in\mathbb{R}$ and $g_k:\mathcal{X}\to\mathbb{R}, h_l:\mathcal{Y}\to\mathbb{R}$ continuous [34, Cor. 2.10]. Moreover, point 1 could be replaced by assuming only that $\mathcal{X}$ and $\mathcal{Y}$ are compact Riemannian manifolds without boundaries; our analysis could be generalized to this setting (following [8]) at the expense of more technical

[2] When F is not convex-concave, several notions of min-max solutions can be considered [11, §2.1], but we will never need such considerations in this article. To fix ideas, we will use the strongest such notion (that of global saddle point) and call *solution of a nonconvex-nonconcave problem* $\min_x\max_y G(x,y)$ any (x^*,y^*) satisfying $G(x^*,y)\le G(x,y^*)$ for all x,y; but we emphasize that this choice has no impact on any of our discussion.

notation. Point 3 is crucial to our analysis, but also to all known last-iterate convergence analyses for MNEs in finite dimension [10, 37].

Before stating the rest of our assumptions, let us remark a useful fact about the structure of the problem. By definition the MNE (μ^*, ν^*) is characterized by

$$\forall\, \mu \in \mathcal{P}(\mathcal{X}),\ \forall\, \nu \in \mathcal{P}(\mathcal{Y}),\ F(\mu^*, \nu) \le \rho \le F(\mu, \nu^*) \quad \text{where } \rho \coloneqq F(\mu^*, \nu^*),$$

i.e.,

$$\forall\, \mu \in \mathcal{P}(\mathcal{X}),\ \int_{\mathcal{X}} \underbrace{\left(\int_{\mathcal{Y}} f(x,y)\,\mathrm{d}\nu^*(y)\right)}_{=:\ (F\nu^*)(x)} \mathrm{d}\mu(x) \ge \rho \quad \text{and} \quad \forall\, \nu \in \mathcal{P}(\mathcal{Y}),\ \int_{\mathcal{Y}} \underbrace{\left(\int_{\mathcal{X}} f(x,y)\,\mathrm{d}\mu^*(x)\right)}_{=:\ ((\mu^*)^\top F)(y)} \mathrm{d}\nu(y) \le \rho.$$

The function $F\nu \in \mathcal{C}(\mathcal{X})$ defined by this equation is the first variation of F with respect to μ at any (μ, ν) [32]; note that it is independent of μ thanks to bilinearity of F. Since $\min_{\mu \in \mathcal{P}(\mathcal{X})} \int g\,\mathrm{d}\mu = \min_{\mathcal{X}} g$ for any $g \in \mathcal{C}(\mathcal{X})$, the above inequalities are equivalent to

$$\forall\, x \in \mathcal{X},\ (F\nu^*)(x) \ge \rho \quad \text{and} \quad \forall\, y \in \mathcal{Y},\ ((\mu^*)^\top F)(y) \le \rho. \tag{4}$$

As a partial converse, we also have

$$\forall\, I \in [n^*],\ (F\nu^*)(x_I^*) = \rho \quad \text{and} \quad \forall\, J \in [m^*],\ ((\mu^*)^\top F)(y_J^*) = \rho \tag{5}$$

since if $(F\nu^*)(x_I^*) > \rho$ for some I then we would have $F(\mu^*, \nu^*) = \int_{\mathcal{X}} (F\nu^*)(x)\,\mathrm{d}\mu^*(x) > \rho$.

Our second set of assumptions requires the inequalities (4) to be strict wherever possible, and even "strong" locally.

▶ **Assumptions** (Non-degeneracy)**.**

5. The first variations at optimum, $F\nu^* \in \mathcal{C}(\mathcal{X})$ resp. $(\mu^*)^\top F \in \mathcal{C}(\mathcal{Y})$, are equal to $\rho \coloneqq F(\mu^*, \nu^*)$ only at the $\{x_I^*, I \in [n^*]\}$ resp. $\{y_J^*, J \in [m^*]\}$.
6. The local kernels are non-degenerate, that is,

$$\forall\, I \in [n^*],\ \nabla^2 (F\nu^*)(x_I^*) \succ 0 \quad \text{and} \quad \forall\, J \in [m^*],\ \nabla^2 ((\mu^*)^\top F)(y_J^*) \prec 0.$$

These non-degeneracy assumptions are analogous to the ones made in [8] for minimization in the space of measures. As for minimization, they generally cannot be checked a priori; but they can be checked a posteriori after computing (μ^*, ν^*) by computing the Hessians of $F\nu^*$ on $\mathrm{supp}(\mu^*)$ and of $(\mu^*)^\top F$ on $\mathrm{supp}(\nu^*)$, or simply visually in the one-dimensional case as in Section 4.1 (Figure 3). Even though we do not have any rigorous result in this direction, informally we expect the non-degeneracy assumptions to be generic in some sense. For example they turned out to be satisfied in all of our experiments with random payoff functions of the form of Section 4.1, with any dimension d_x, d_y. Example 13 provides a case where all the assumptions can be checked analytically, including the uniqueness of the MNE.

Our work provides the first convergence guarantee for the sparse continuous-game MNE setting (Assumptions 1, 2, 4). On the other hand, Assumptions 3, 5, 6 are admittedly difficult to check a priori for any given $(f, \mathcal{X}, \mathcal{Y})$. But they are the minimal ones for our convergence analysis to go through, and we expect that they are very difficult to relax.

2.2 The Conic Particle Proximal Point algorithm

In order to solve the saddle-point problem (1), we reparametrize the problem via $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j}$ and we use a particle gradient algorithm to tackle the reparametrized problem (2): $\min_{a,x} \max_{b,y} F_{n,m}((a,x),(b,y))$. Specifically, we analyze the Conic Particle Proximal Point (CP-PP) algorithm given by the update rule

$$\begin{aligned}((a^{k+1}, x^{k+1}), (b^{k+1}, y^{k+1})) = \underset{\substack{a \in \Delta_n \\ x \in \mathcal{X}^n}}{\arg\min} \underset{\substack{b \in \Delta_m \\ y \in \mathcal{Y}^m}}{\arg\max}\ & F_{n,m}((a,x),(b,y)) + \frac{1}{\eta} D(a, a^k) + \frac{1}{2\sigma} \sum_{i=1}^n a_i^k \left\| x_i - x_i^k \right\|^2 \\ & - \frac{1}{\eta} D(b, b^k) - \frac{1}{2\sigma} \sum_{j=1}^m b_j^k \left\| y_j - y_j^k \right\|^2 \end{aligned} \tag{6}$$

Algorithm 1: Conic Particle Proximal Point, implementable variant

Input: $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, $n, m \in \mathbb{N}^*$, $\eta, \sigma > 0$, $T, L \in \mathbb{N}^*$
Initialize $z^0 = (a^0, x^0, b^0, y^0) \in \Delta_n \times \mathcal{X}^n \times \Delta_m \times \mathcal{Y}^m$;
for $k = 0, \dots, T-1$ **do**
  $\widetilde{z}^0 \leftarrow z^k$;
  for $l = 0, \dots, L-1$ **do**
    $\forall\, i,\ \widetilde{a}_i^{l+1} \leftarrow \widetilde{a}_i^l e^{-\eta \frac{\partial}{\partial a_i} F_{n,m}(\widetilde{z}^l)}/Z$ with Z such that $\widetilde{a}^{l+1} \in \Delta_n$;
    $\forall\, i,\ \widetilde{x}_i^{l+1} \leftarrow \widetilde{x}_i^l - \sigma \frac{1}{a_i^k} \frac{\partial}{\partial x_i} F_{n,m}(\widetilde{z}^l)$;
    similarly for $\widetilde{b}^{l+1}$ and $\widetilde{y}^{l+1}$
  $z^{k+1} \leftarrow \widetilde{z}^L$
return $\mu^T = \sum_{i=1}^n a_i^T \delta_{x_i^T}$, $\nu^T = \sum_{j=1}^m b_i^T \delta_{y_j^T}$

where $\eta, \sigma > 0$ are constant step-sizes to be chosen and $D(w, \widehat{w}) = \sum_i w_i \log \frac{w_i}{\widehat{w}_i}$ denotes Kullback–Leibler divergence a.k.a. relative entropy. Interestingly, the function $D\left((a,x),(\widehat{a},\widehat{x})\right) = D(a,\widehat{a}) + \frac{\eta}{2\sigma}\sum_i \widehat{a}_i \left\|x_i - \widehat{x}_i\right\|^2$ is technically not a Bregman divergence (it does not satisfy the last point of Lemma 47), due to the "cross-terms" in the second term.

The following lemma, proved in Appendix B, justifies that the CP-PP update is well-defined.

▶ **Lemma 1.** *Under Assumptions 1-2, there exist $\eta_0, \sigma_0 > 0$ (dependent only on $(f, \mathcal{X}, \mathcal{Y})$) such that if $\eta \le \eta_0$ and $\sigma \le \sigma_0$, then the objective function in (6) is convex-concave over a ball centered at $((a^k, x^k), (b^k, y^k))$, and it has a saddle point in the interior of that ball.*

Implementable variant: Conic Particle Mirror-Prox (CP-MP)

Note that every CP-PP update requires solving a min-max optimization problem (6) exactly. In practice, in the spirit of [30], one may obtain an approximate solution to (6) by running an inner loop where $F_{n,m}((a,x),(b,y))$ is replaced by its first-order approximation at the current point, starting from $(\widetilde{a}^0, \widetilde{x}^0, \widetilde{b}^0, \widetilde{y}^0) = (a^k, x^k, b^k, y^k)$:

$$
\begin{aligned}
&(\widetilde{a}^{l+1}, \widetilde{x}^{l+1}, \widetilde{b}^{l+1}, \widetilde{y}^{l+1}) \\
&= \operatorname*{arg\,min}_{\substack{a \in \Delta_n \\ x \in \mathcal{X}^n}} \operatorname*{arg\,max}_{\substack{b \in \Delta_m \\ y \in \mathcal{Y}^m}} \left\langle \nabla F_{n,m}\left(\widetilde{a}^l, \widetilde{x}^l, \widetilde{b}^l, \widetilde{y}^l\right), (a,x,b,y) \right\rangle + \frac{1}{\eta} D(a, a^k) + \frac{1}{2\sigma} \sum_{i=1}^n a_i^k \left\|x_i - x_i^k\right\|^2 \\
&\qquad\qquad - \frac{1}{\eta} D(b, b^k) - \frac{1}{2\sigma} \sum_{j=1}^m b_j^k \left\|y_j - y_j^k\right\|^2
\end{aligned}
$$

and letting $(a^{k+1}, x^{k+1}, b^{k+1}, y^{k+1}) = (\widetilde{a}^L, \widetilde{x}^L, \widetilde{b}^L, \widetilde{y}^L)$, where $L \ge 1$ is the number of times we run the inner loop at each k. Each iteration of the inner loop decomposes into four independent mirror descent updates. Pseudocode for this implementable variant of CP-PP is given in Algorithm 1, where to lighten the notation we use the shorthand $z^k = (a^k, x^k, b^k, y^k)$.

One can check that $L = 1$ corresponds to the CP-MDA algorithm described in the introduction, and that for $L = \infty$ we recover CP-PP. When $L = 2$, we refer to this method as Conic Particle Mirror Prox (CP-MP). Similarly as in [30], one can expect that $L = 2$ actually suffices to obtain the same convergence behavior as $L = \infty$; this is confirmed in numerical experiments, and proved in a simplified setting (Proposition 6). This behavior can be explained by the fact that Proximal Point and Mirror Prox updates coincide up to order-3 terms in the step-size (see Lemma 49 and Lemma 50).

Relation to Wasserstein–Fisher–Rao (WFR) geometry

The Wasserstein–Fisher–Rao distance, a.k.a. Hellinger–Kantorovich distance [9, 20, 22] is a distance on the set of non-negative measures which metrizes narrow convergence and combines features of the Fisher–Rao and of the Wasserstein distances. This last fact is perhaps easiest to see in its dynamical formulation [22, Thm. 8.18]:

$$
\mathrm{WFR}_2^2(\mu_0, \mu_1) = \inf_{(\mu_t, v_t, r_t) \in \mathcal{A}(\mu_0, \mu_1)} \int_0^1 \int_{\mathcal{X}} \left(\frac{\eta}{2\sigma} \left\|v_t(x)\right\|^2 + \frac{1}{2} r_t(x)^2 \right) \mathrm{d}\mu_t(x)\, \mathrm{d}t \tag{7}
$$

where $\mathcal{A}(\mu_0, \mu_1)$ is the set of triples $(\mu_t, v_t, r_t)_{0\le t\le 1}$ such that $(\mu_t)_{t\in[0,1]}$ is a weakly continuous curve in $\mathcal{M}_+(\mathcal{X})$ the set of non-negative measures with endpoints μ_0 and μ_1, $v_t \in L^2_{\mu_t}(\mathcal{X})^{d_x}$, $r_t \in L^2_{\mu_t}(\mathcal{X})$, and satisfying the continuity equation with source $\partial_t\mu_t + \nabla\cdot(\mu_t v_t) = \mu_t r_t$ in the sense of distributions. (See [9, 22] for equivalent static formulations). Here the scalars η and σ trade off the Fisher–Rao ("local growth and destruction of mass") and Wasserstein ("movement of mass") components of the distance.

The CP-MDA, CP-PP and CP-MP algorithms are time-discretizations of the interacting WFR gradient flow [12], which is the finite-particle version of the WFR gradient flow in measure space [8, Prop. 2.1]. Because of this connection, the CP-PP iterates' WFR distance to the MNE (or rather a simpler proxy of it that is sufficient for our purpose) will play a central role in our convergence analysis.

2.3 Main convergence result

Our main result is that the CP-PP algorithm (6) converges locally at an exponential rate. Its proof follows from Proposition 7, Proposition 8 and Theorem 9, as shown in Appendix I.

▶ **Theorem 2.** *Fix any $\Gamma_0 \ge 1$. Under Assumptions 1-6, there exist $\eta_0, \sigma_0 > 0$ such that for all $\eta \le \eta_0, \sigma \le \sigma_0$ with $\Gamma_0^{-1} \le \frac{\sigma}{\eta} \le \Gamma_0$, there exists $C, C', r_0, \kappa > 0$ such that if $\mathrm{NI}(\mu^0, \nu^0) \le r_0$, then the CP-PP iterates $(\mu^k, \nu^k) = \left(\sum_{i=1}^n a_i^k \delta_{x_i^k}, \sum_{j=1}^m b_j^k \delta_{y_j^k}\right)$ satisfy for all $k \in \mathbb{N}$*

$$\mathrm{NI}(\mu^k, \nu^k) \le C(1-\kappa)^k$$

and $\quad \mathrm{WFR}_2^2(\mu^k, \mu^*) + \mathrm{WFR}_2^2(\nu^k, \nu^*) \le C'(1-\kappa)^k.$

In particular, since WFR *metrizes narrow convergence, μ^k, ν^k converge narrowly to μ^*, ν^*, that is $\forall\, \phi \in \mathcal{C}(\mathcal{X})$, $\int \phi \,\mathrm{d}\mu^k \to \int \phi \,\mathrm{d}\mu^*$ and $\forall\, \psi \in \mathcal{C}(\mathcal{Y})$, $\int \psi \,\mathrm{d}\nu^k \to \int \psi \,\mathrm{d}\nu^*$.*

The dependency of C, C', r_0, κ on the problem data $(f, \mathcal{X}, \mathcal{Y})$ appears quite subtle, unfortunately. It can be traced back to Lemma 18 establishing an "error bound" type inequality which relies on uniqueness of the MNE, making it difficult to quantify. The known analyses for finite-game MNEs [10, 37] face the same drawback.

Dependency of the constants on the step-sizes η, σ

The quantities C, C', r_0, κ appearing in the theorem depend on the step-sizes. Our proof technique requires to take them at most of order $r_0 \lesssim \eta^{17/4}$, $\kappa \lesssim \eta^2$, and $C \gtrsim \eta^{-1/5} r_0^{2/5}$, $C' \gtrsim \eta^{-2/5} r_0^{4/5}$ (supposing $\eta \asymp \sigma$), as one can check from the proof in Appendix I and the statements of Proposition 7, Proposition 8 and Theorem 9. So our approach only applies for discrete-time algorithms, and would not allow to show convergence of the continuous-time flow associated to CP-PP.

Indeed if we formally let $\eta, \sigma \to 0$, the localness requirement on the initial iterate $\mathrm{NI}(\mu^0, \nu^0) \le r_0 \xrightarrow{\eta,\sigma\to 0} 0$ reduces to requiring (μ^0, ν^0) to already be the MNE. Even ignoring the localness requirement, the bound $\left(1 - \Theta(\eta^2)\right)^k$ becomes constant when $k = \lfloor \frac{t}{\eta} \rfloor$ for a fixed t and $\eta \asymp \sigma \to 0$; so the bound does not vanish as t increases.

Interestingly, experimentally we do observe convergence of the continuous-time flow in some (but not all) cases; see Section 4.1. It is worth mentioning that for $\mathcal{X}$ and $\mathcal{Y}$ finite, the Fisher–Rao gradient flow does not converge [27], so the fact that we sometimes observe convergence of the flow may be specific to conic particle gradient methods.

Necessity of taking $\eta, \sigma > 0$ (comparison with pure Fisher–Rao and pure Wasserstein particle methods)

- If $\sigma = 0$, the position variables (x^k, y^k) of CP-PP stay constant throughout the algorithm, and only the weights (a^k, b^k) vary. This corresponds to the Fisher–Rao gradient dynamics.
  - If the initial measure variables (μ^0, ν^0) were supported on a large number of points $\{\widehat{x}_1, \dots, \widehat{x}_n\}$ resp. $\{\widehat{y}_1, \dots, \widehat{y}_m\}$ covering $\mathcal{X}$ resp. $\mathcal{Y}$ uniformly, one may expect CP-PP to converge locally exponentially to (μ^∞, ν^∞) a MNE of the finite game $(f, \{\widehat{x}_i\}_i, \{\widehat{y}_j\}_j)$.[3] By continuity of f, the "price" of the discretization with respect to the original game can then be bounded by $\mathrm{NI}(\mu^\infty, \nu^\infty) = O\left(n^{-1/d_x} + m^{-1/d_y}\right)$. That is,

[3] In fact there is no guarantee so far that the last iterate of CP-PP with $\sigma = 0$ will converge locally to a MNE of the finite game $\left(f, \{\widehat{x}_i\}_i, \{\widehat{y}_j\}_j\right)$ (although the averaged iterate is guaranteed to converge globally to one by [30]). Indeed, while [26, App. D] shows qualitative convergence without a rate, known quantitative last-iterate convergence guarantees for finite games [10, 37] require the MNE to be unique, which may not be the case a priori for $\left(f, \{\widehat{x}_i\}_i, \{\widehat{y}_j\}_j\right)$.

in order to achieve a NI error of $O(\varepsilon)$, it is sufficient to let $n = (1/\varepsilon)^{d_x}, m = (1/\varepsilon)^{d_y}$, and to run CP-PP to convergence with $\sigma = 0$; however the computational complexity of each iteration would then be $\Theta(nm)$, which can be prohibitively costly if d_x, d_y are large.
  - Also note that if (μ^0, ν^0) are supported on the entire space $\mathcal{X}$ resp. $\mathcal{Y}$, we can only expect convergence of CP-PP to the exact MNE at a rate $\asymp 1/k$ in the worst case, and not at an exponential rate [7] (a worst-case example can be constructed by using a similar idea as in Proposition 5.5, Setting II of that paper).
- If $\eta = 0$, the weight variables of CP-PP stay fixed, and only the positions vary. This corresponds to the interacting Wasserstein gradient dynamics [12].
  - This dynamics has a degraded behavior because of it has fewer degrees of freedom. For instance, for (μ^0, ν^0) supported on finitely many point and $a_i^0 = \frac{1}{n}, b_j^0 = \frac{1}{m}$, the CP-PP iterates cannot converge to (μ^*, ν^*) exactly, unless the solution weights a_I^*, b_J^* happen to all lie in $\frac{1}{n}\mathbb{Z}$ resp. $\frac{1}{m}\mathbb{Z}$.
  - Even allowing for (μ^0, ν^0) supported on the entire space, we are not aware of any convergence guarantee to (μ^*, ν^*) for this dynamics, in continuous time or otherwise.

Agnosticity to the numbers of particles n, m

The numbers of particles n, m used in the CP-PP algorithm do not appear in the theorem, nor are they hidden in the constants. In particular the convergence rate does not deteriorate with large n, m while the per-iteration cost is linear in $n + m$. Even the condition that $n \geq n^*$ and $m \geq m^*$ (the sparsities of (μ^*, ν^*)) does not appear explicitly, but it is implied by the localness condition $\mathrm{NI}(\mu^0, \nu^0) \leq r_0$. That is, r_0 is defined such that, if $n < n^*$ or $m < m^*$, then there simply do not exist atomic measures μ^0, ν^0 with n resp. m atoms that achieve NI error less than r_0.

The fact that our result is agnostic to such overparametrization should be viewed as a strength. Indeed the convergence guarantee does not deteriorate with large n, m, and on the other hand allowing ourselves to use arbitrary $(n, m) \neq (n^*, m^*)$ enables simpler warm-up procedures. In terms of the application to classification with two-layer neural networks presented in Section 4, this agnosticity means that our results apply regardless of the number of hidden neurons, as long as it is not too small.

A possible two-phase procedure

While our proposed algorithm is only shown to be locally convergent, we would like to stress that it is the only known one that can provably converge to the actual solution (μ^*, ν^*). This is in contrast to any algorithm relying on discretization of the strategy spaces $\mathcal{X}, \mathcal{Y}$, since these algorithms can only ever output measures (μ^k, ν^k) whose support does not even match the optimal one (unless one is extremely lucky when choosing the discretization). So one way to take advantage of our proposed algorithm and performance guarantee, is to use it as a second "high-accuracy" phase, preceded by a warm-up phase with global convergence guarantees.

A simple such warm-up procedure is to fix $\varepsilon_{\text{warm-up}} \leq r_0$, to discretize $\mathcal{X}$ and $\mathcal{Y}$ by $n = (1/\varepsilon_{\text{warm-up}})^{d_x}$, $m = (1/\varepsilon_{\text{warm-up}})^{d_y}$ grid-points $\{\widehat{x}_i\}_i, \{\widehat{y}_j\}_j$, to run Fisher–Rao Proximal Point (i.e, CP-PP with $\sigma = 0$) for $T_{\text{warm-up}} = 1/\varepsilon_{\text{warm-up}}$ iterations and take the average of the iterates. Indeed this ensures a NI error of $O(\varepsilon_{\text{warm-up}})$ for the discretized game [30], and the NI error for the discretized game is $O(\varepsilon_{\text{warm-up}})$-close to the NI error for the original game by the choice of n, m. The per-iteration complexity of the first phase is $\Theta(nm)$ and that of the second phase is $\Theta(n + m) = \Theta((1/\varepsilon_{\text{warm-up}})^{d_x \vee d_y})$, which could still be large. But note that *with the same per-iteration complexity*, one can then exponentially fast achieve NI error less than ε, for any $\varepsilon < \varepsilon_{\text{warm-up}}$. This "high-accuracy" regime is where our method improves upon classical discretization-based algorithms.

Note however that, unfortunately, we cannot control the size r_0 of the neighborhood where our local result applies. Even worse: even if the quantity r_0 was known, this would still not suffice to certify efficiently that neighborhood is reached, as the NI error is difficult to compute and even to upper-bound. Thus we are at present unable to deduce a provably globally convergent algorithm from our work. (If convergence rates are not desired, then it suffices to use the two-phase procedure proposed above along with some form of the doubling trick to choose $\varepsilon_{\text{warm-up}}$.)

3 Convergence proof

Throughout this section, we make the Assumptions 1–6 described in Section 2.1. Furthermore, to lighten notation, we denote by $z^k = (a^k, x^k, b^k, y^k)$ the iterates of the CP-PP algorithm. We denote the sparse unique MNE as

$$\mu^* = \sum_{I \in [n^*]} a_I^* \delta_{x_I^*} \quad \text{and} \quad \nu^* = \sum_{J \in [m^*]} b_J^* \delta_{y_J^*} \quad \text{with } a_I^*, b_J^* > 0.$$

The variational inequality characterizing CP-PP

From Lemma 1, we know that z^{k+1} is well-defined as the saddle point of a convex-concave function in the interior of its domain. Just by writing out the first-order optimality conditions in the argmin/argmax, we see that the CP-PP update (6) is characterized by the variational formula

$$\begin{aligned}\forall\, z = (a,x,b,y),\ \eta\, \widetilde{\mathrm{gap}}(z; z^{k+1}) \le \sum_i (a_i - a_i^{k+1}) \log \frac{a_i^{k+1}}{a_i^k} + \sum_j (b_j - b_j^{k+1}) \log \frac{b_j^{k+1}}{b_j^k} \\ + \frac{\eta}{\sigma} \sum_i a_i^k \left\langle x_i^{k+1} - x_i^k, x_i - x_i^{k+1} \right\rangle + \frac{\eta}{\sigma} \sum_j b_j^k \left\langle y_j^{k+1} - y_j^k, y_j - y_j^{k+1} \right\rangle \end{aligned} \tag{8}$$

(both sides are linear in $z - z^{k+1}$), where we introduce, for all $z = (a,x,b,y)$ and $\widehat{z} = (\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y})$,

$$\widetilde{\mathrm{gap}}(z; \widehat{z}) \coloneqq \left\langle \begin{pmatrix} \nabla_a \\ \nabla_x \\ -\nabla_b \\ -\nabla_y \end{pmatrix} F_{n,m}(\widehat{z}), \begin{pmatrix} \widehat{a} - a \\ \widehat{x} - x \\ \widehat{b} - b \\ \widehat{y} - y \end{pmatrix} \right\rangle \quad \text{and} \quad \mathrm{gap}(z; \widehat{z}) \coloneqq F_{n,m}((\widehat{a}, \widehat{x}), (b, y)) - F_{n,m}((a, x), (\widehat{b}, \widehat{y})).$$

One can check that z^k is in the (relative) interior of $(\Delta_n \times \mathcal{X}^n) \times (\Delta_m \times \mathcal{Y}^m)$ for all k (provided z^0 is), so (8) holds in fact with an equality (but we continue to write inequalities to show the generality of our arguments).

The significance of the quantity $\widetilde{\mathrm{gap}}(z; \widehat{z})$ comes from the fact that, if $F_{n,m}$ was convex-concave, $\widehat{z}$ would be a saddle point if and only if $\forall\, z, \widetilde{\mathrm{gap}}(z; \widehat{z}) \le 0$, as a solution of the Stampacchia variational inequality [11, §2.1]. Furthermore, $\widetilde{\mathrm{gap}}(z; \widehat{z})$ can also be interpreted as a first-order approximation of $\mathrm{gap}(z; \widehat{z})$, whose significance is that $\mathrm{NI}\big(\sum_i \widehat{a}_i \delta_{\hat{x}_i}, \sum_j \widehat{b}_j \delta_{\hat{y}_j}\big) = \max_z \mathrm{gap}(z; \widehat{z})$.

3.1 Exact-parametrization case

In this subsection, we present a short proof of our result in the simpler case where we additionally assume that the number of particles (n, m) is exactly equal to the sparsity of the solution (n^*, m^*). Relabel the solution particles (a_I^*, x_I^*) resp. (b_J^*, y_J^*) arbitrarily so that they are indexed by $i \in [n] = [n^*]$ resp. $j \in [m] = [m^*]$.

The convergence analysis relies on the Lyapunov function $V(a,x,b,y) = V(a,x) + V(b,y)$ where

$$V(a,x) = \underbrace{D(a^*, a)}_{=: V_{\mathrm{wei}}(a,x)} + \frac{\eta}{\sigma} \underbrace{\frac{1}{2} \sum_{i=1}^n a_i \big\| x_i^* - x_i \big\|^2}_{=: V_{\mathrm{pos}}(a,x)} \tag{9}$$

and similarly for $V(b,y)$. For ease of presentation, also let $V_1(a,x) = V_{\mathrm{wei}}(a,x) + V_{\mathrm{pos}}(a,x)$, and similarly for $V_1(b,y)$ and $V_1(a,x,b,y)$. Note that we always have $(1 \wedge \sigma/\eta) V \le V_1 \le (1 \vee \sigma/\eta) V$.

Note that $V(a,x,b,y) \ge 0$ and that equality holds if and only if $(a,x,b,y) = (a^*, x^*, b^*, y^*)$. We can also relate this quantity to the NI error as follows; in particular V is arbitrarily small for NI small, and vice-versa. The proof can be found in Appendix D.4.

▶ **Proposition 3.** *Assume that $n = n^*, m = m^*$ and define V_1 as in (9). There exist a constant $C > 0$ dependent only on $(f, \mathcal{X}, \mathcal{Y})$ such that, for any $z = (a,x,b,y)$, denoting $\mu = \sum_i a_i \delta_{x_i}$ and $\nu = \sum_j b_j \delta_{y_j}$,*

$$\mathrm{NI}(\mu, \nu) \le C \sqrt{V_1(z)}.$$

Moreover, there exist $C', r > 0$ dependent only on $(f, \mathcal{X}, \mathcal{Y})$ such that if $\mathrm{NI}(\mu,\nu) \le r$, then up to permuting the labels of the solution particles (so that, for each $i \in [n]$, x_i is in a neighborhood of x_i^*, and not necessarily of $x_{i'}^*$ for $i' \ne i$),

$$C' V_1(z)^{5/4} \le \mathrm{NI}(\mu, \nu).$$

Our choice of Lyapunov function is essentially a proxy for the squared WFR distance (7) of μ to μ^* and of ν to ν^*, as shown in [8, Lem. D.1]. In our notation:

▶ **Proposition 4** ([8, Lem. D.1]). *Assume that $n = n^*$ and define V as in (9). There exist constants $C, r > 0$ (dependent only on μ^*) such that for any (a, x) with $V(a, x) \le r$, denoting $\mu = \sum_i a_i \delta_{x_i}$,*

$$\mathrm{WFR}_2^2(\mu, \mu^*) \le 2V(a, x)\left(1 + C\frac{\eta}{\sigma}\right).$$

The main result of this subsection is that the CP-PP algorithm converges locally at an exponential rate, as measured by the Lyapunov function. Convergence measured by NI error and by WFR distance (Theorem 2 with the additional exact-parametrization assumption) could be shown by combining Theorem 5 with Proposition 3 and Proposition 4.

▶ **Theorem 5.** *Assume that $n = n^*, m = m^*$ and define V as in (9). Fix any $\Gamma_0 \ge 1$. There exist η_0, σ_0 such that for all $\eta \le \eta_0, \sigma \le \sigma_0$ with $\Gamma_0^{-1} \le \frac{\sigma}{\eta} \le \Gamma_0$, there exists $r_0 > 0$ such that if $V(z^0) \le r_0$, then the CP-PP iterates z^k satisfy*

$$\forall\, k,\ V(z^k) \le V(z^0)\,(1 - \kappa)^k$$

for some constant $\kappa > 0$.

More precisely, one can check from the last step of the proof that the rate κ and the localness level r_0 can at most be chosen equal to η^2 times a constant (dependent on $(f, \mathcal{X}, \mathcal{Y})$ and Γ_0).

Proof. Evaluate (8) at $z = (a^*, x^*, b^*, y^*)$. By the Bregman three-point identity on $h : a \mapsto a \log a - a + 1$ (so that the Kullback–Leibler divergence $D(\cdot, \cdot)$ is equal to the Bregman divergence of h summed component-wise) and on $x \mapsto \frac{1}{2}\|x\|^2$, we can rewrite the obtained inequality as

$$\begin{aligned}\eta\, \widetilde{\mathrm{gap}}(z^*; z^{k+1}) \le V(z^k) - V(z^{k+1}) \\ \underbrace{- \left(D(a^{k+1}, a^k) + D(b^{k+1}, b^k) + \frac{\eta}{2\sigma}\sum_i a_i^k \left\|x_i^{k+1} - x_i^k\right\|^2 + \frac{\eta}{2\sigma}\sum_j b_j^k \left\|y_j^{k+1} - y_j^k\right\|^2 \right)}_{=:\ D(k+1,k)} \\ \underbrace{+ \frac{\eta}{2\sigma}\sum_i (a_i^{k+1} - a_i^k)\left\|x_i^* - x_i^{k+1}\right\|^2 + \frac{\eta}{2\sigma}\sum_j (b_j^{k+1} - b_j^k)\left\|y_j^* - y_j^{k+1}\right\|^2}_{=:\ [\mathrm{err}]}. \end{aligned} \tag{10}$$

Now one can show that, if $\eta \le \eta_0, \sigma \le \sigma_0$ and $V(z^k) \le r_0$ for some η_0, σ_0, r_0 dependent only on $(f, \mathcal{X}, \mathcal{Y})$ and Γ_0, then both $\left[\min_i a_i^k \wedge \min_j b_j^k\right]$ and $\left[\min_i a_i^{k+1} \wedge \min_j b_j^{k+1}\right]$ are lower-bounded by a fixed positive constant (Lemma 31), and so

- (Lemma 33) The left-hand side is lower-bounded as $\eta\, \widetilde{\mathrm{gap}}(z^*; z^{k+1}) \ge \eta \frac{\sigma_{\min}}{2} V_{\mathrm{pos}}(z) + O\left(\eta V(z^{k+1})^{3/2}\right)$ for some $\sigma_{\min} > 0$ dependent only on $(f, \mathcal{X}, \mathcal{Y})$. This inequality is a consequence of the "quadratic growth" and "star-convexity" properties discussed in Section 3.3.1, resp. Section 3.3.3.
- (Lemma 38) There exists a constant $C > 0$ dependent only on $(f, \mathcal{X}, \mathcal{Y})$ and Γ_0 such that $D(k+1, k) \ge C\eta^2 V(z^{k+1}) + O\left(\eta V(z^{k+1})^2\right)$. This inequality follows from an "error bound"-type result discussed in Section 3.3.2.
- (Lemma 36) The terms on the third line, that arise due to the fact that the divergence used in the update (6) is not a Bregman divergence, are bounded as $[\mathrm{err}] = O\left(\eta V(z^{k+1})^{3/2}\right)$.

In each of the bounds above, the $O(\,\cdot\,)$ hides a constant dependent only on $(f, \mathcal{X}, \mathcal{Y})$ and Γ_0. By plugging these bounds back into (10), we obtain

$$V(z^{k+1}) \le V(z^k) - C\eta^2 V(z^{k+1}) + O\left(\eta V(z^{k+1})^{3/2}\right).$$

In particular, since $V(z^{k+1})$ is bounded by a constant (Lemma 31), for small enough η_0 and σ_0 we have that $V(z^{k+1}) \le 2V(z^k)$. By rearranging we get

$$V(z^{k+1}) \le \frac{V(z^k)}{1 + \eta^2\left[C - O\left(\frac{\sqrt{V(z^{k+1})}}{\eta}\right)\right]}.$$

Hence, for appropriately small choices of r_0, we have $V(z^k) \le r_0 \implies V(z^{k+1}) \le 2r_0 \implies V(z^{k+1}) \le V(z^k)(1-\kappa)$ for some $\kappa > 0$. The final result follows by induction. ◀

3.1.1 Convergence of CP-MP

In the exact-parametrization case it is relatively easy to extend our convergence result for CP-PP to CP-MP (the implementable variant of CP-PP, Algorithm 1, with $L = 2$). Namely CP-MP converges under the same conditions and with the same rate as CP-PP.

▶ **Proposition 6.** *The statement of Theorem 5 also holds for z^k being the iterates of CP-MP.*

The proof of the proposition, in Appendix H, essentially relies on the convergence result for CP-PP and on the fact that the CP-MP and CP-PP updates coincide up to order-3 terms in the step-size. In particular we derive order-2 expressions for the Mirror Prox and Proximal Point updates under quite general assumptions (Lemma 49 and Lemma 50), which may be of independent interest.

3.2 General case

In general, the sparsity of the solution (n^*, m^*) is not known in advance, and $n \neq n^*, m \neq m^*$. Contrary to the exact-parametrization case where the choice of Lyapunov function was relatively straightforward, here it must be carefully designed, due to overparametrization. Indeed, the variables (a, x, b, y) and the solution (a^*, x^*, b^*, y^*) live in different spaces: $a \in \Delta_n \neq \Delta_{n^*} \ni a^*$, so we cannot just evaluate the algorithm's characterizing inequality (8) at the solution.

We define a Lyapunov function $V(a, x, b, y) = V(a, x) + V(b, y)$ by the following construction, generalizing [8, Eq. (20)]. See Figure 1a for an illustration.

- Fix $(\varphi_I)_{I \in [0,n^*]}$ a partition of unity of $\mathcal{X}$ centered at the $(x_I^*)_I$, i.e.,
  - Each φ_I is a measurable function $\mathcal{X} \to \mathbb{R}$;
  - $\forall\, I \in [n^*],\ \varphi_I \ge 0$ and $\varphi_0 = 1 - \sum_{I \in \mathcal{I}^*} \varphi_I \ge 0$ over $\mathcal{X}$;
  - $\forall\, I \in [n^*],\ \varphi_I(x_I^*) = 1$.
- For any $a \in \Delta_n, x \in \mathcal{X}^n$, define the *aggregated weights*, the *aggregated positions* and the *local covariance matrices* of $\mu = \sum_i a_i \delta_{x_i}$ as

$$\forall\, I \in [0, n^*],\ \bar{a}_I = \int_{\mathcal{X}} \varphi_I \,\mathrm{d}\mu \qquad \text{and} \qquad \forall\, I \in [n^*],\ \bar{x}_I = \int_{\mathcal{X}} x\ \frac{\varphi_I(x)}{\bar{a}_I} \,\mathrm{d}\mu(x)$$
$$\Sigma_I = \int_{\mathcal{X}} (x - \bar{x}_I)(x - \bar{x}_I)^\top\ \frac{\varphi_I(x)}{\bar{a}_I} \,\mathrm{d}\mu(x).$$

  I.e., in discrete notation,

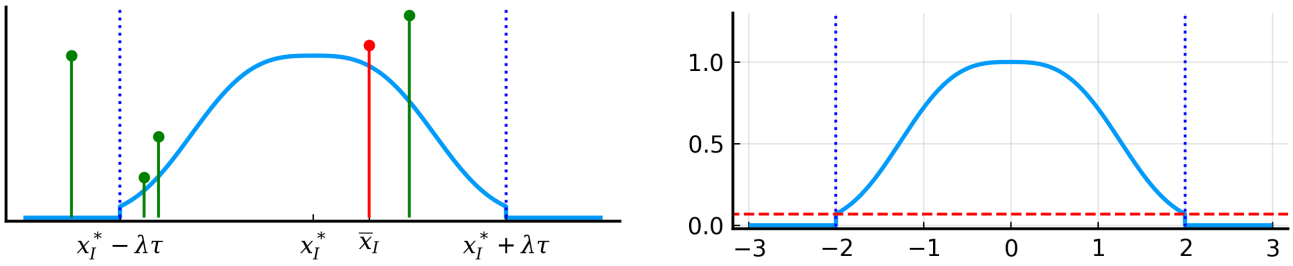
$$\forall\, I \in [n^*],\ \bar{a}_I = \sum_i \varphi_{Ii} a_i \qquad \bar{x}_I = \sum_i \frac{\varphi_{Ii} a_i}{\bar{a}_I} x_i \qquad \Sigma_I = \sum_i \frac{\varphi_{Ii} a_i}{\bar{a}_I} (x_i - \bar{x}_I)(x_i - \bar{x}_I)^\top \tag{11}$$

  where $\varphi_{Ii} = \varphi_I(x_i)$, and $\bar{a}_0 = 1 - \sum_I \bar{a}_I$ is the *stray weight*.
- Let, for any $a \in \Delta_n, x \in \mathcal{X}^n$,

$$V(a, x) = \underbrace{D(a^*, \bar{a})}_{=: V_{\mathrm{wei}}(a,x)} + \frac{\eta}{\sigma} \underbrace{\frac{1}{2} \sum_I \bar{a}_I \left(\|x_I^* - \bar{x}_I\|^2 + \mathrm{Tr}(\Sigma_I) \right)}_{=: V_{\mathrm{pos}}(a,x)}. \tag{12}$$

Similarly, fix $(\psi_J)_{J \in [0,m^*]}$ a partition of unity of $\mathcal{Y}$ centered at the y_J^*, similarly define $\bar{b} \in \Delta_{m^*}$ and $\bar{y} \in \mathcal{Y}^{m^*}$ for any $b \in \Delta_m, y \in \mathcal{Y}^m$, and similarly define $V(b, y)$. For ease of presentation, also let $V_1(a, x) = V_{\mathrm{wei}}(a, x) + V_{\mathrm{pos}}(a, x)$, and similarly for $V_1(b, y)$ and $V_1(a, x, b, y)$. Note that we always have $(1 \wedge \sigma/\eta)V \le V_1 \le (1 \vee \sigma/\eta)V$.

The Lyapunov function V depends on the choice of partitions of unity $(\varphi_I)_I$ and $(\psi_J)_J$. They can be freely designed so as to make the proof go through, as long as they satisfy the conditions announced above (non-negative, sum to 1, $\varphi_I(x_I^*) = 1$). For example, our analysis for the exact-parametrization case was equivalent to choosing as φ_I the indicator function of a small ball centered at x_I^* (Claim 16). Proving convergence in the general case requires a much subtler choice; specifically, the partitions of unity we use for the proof of our main result are



(a) Green stems represent the particles: abscissa indicates x_i and stem height indicates a_i

(b) $y = \exp\left(-\frac{|x|^3}{3}\right)$ if $|x| \le 2$, $y = 0$ otherwise

Figure 1 Illustration of the construction defining $V(a, x)$

defined as

$$\varphi_I(x) = \begin{cases} \exp\left(-\frac{\|x - x_I^*\|^3}{3\tau^3}\right) & \text{if } \left\|x - x_I^*\right\| \le \lambda\tau \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

for some bandwidth resp. cut-off parameters $\tau, \lambda > 0$ chosen as functions of η and σ. The cut-off parameter λ is used to ensure that $\sum_{I' \neq I} \varphi_{I'}(x_I^*) = 0$ and so $\varphi_0(x_I^*) \ge 0$. See Figure 1b for an illustration in one dimension with $\tau = 1$, $\lambda = 2$.

Note that $V(a, x, b, y) \ge 0$ with equality if and only if $(\bar{a}, \bar{x}, \bar{b}, \bar{y}) = (a^*, x^*, b^*, y^*)$ and $\Sigma_I, \Sigma_J = 0$ for all I, J, i.e., if and only if $(\mu, \nu) = (\mu^*, \nu^*)$. Beyond this equivalence, similarly as for the exact-parametrization case, we can show the following relation between V and NI error. The proof of the following proposition, as well as a more quantitative version of it, can be found in Appendix D.[4]

▶ **Proposition 7.** *Define V_1 as in (12) with the partitions of unity $(\varphi_I)_I$ and $(\psi_J)_J$ as in (13). Suppose that $\lambda\tau$ is less than some constant dependent on $(f, \mathcal{X}, \mathcal{Y})$. There exists a constant $C > 0$ dependent only on $(f, \mathcal{X}, \mathcal{Y})$ such that, for any $z = (a, x, b, y)$, denoting $\mu = \sum_i a_i \delta_{x_i}$ and $\nu = \sum_j b_j \delta_{y_j}$,*

$$\mathrm{NI}(\mu, \nu) \le C\sqrt{V_1(z)}.$$

Moreover, there exist $C', r > 0$ dependent on $(f, \mathcal{X}, \mathcal{Y})$, λ and τ such that, if $\mathrm{NI}(\mu, \nu) \le r$, then

$$C' V_1(z)^{5/4} \le \mathrm{NI}(\mu, \nu).$$

More precisely if λ, τ are chosen as functions of η, σ as in (23) and $\Gamma_0^{-1} \le \frac{\sigma}{\eta} \le \Gamma_0$ for some $\Gamma_0 \ge 1$, then r and C' can be chosen as $\sqrt{\sigma}$ times constants dependent only on $(f, \mathcal{X}, \mathcal{Y})$ and Γ_0.

The Lyapunov function V is by design essentially a proxy for squared WFR distance (7) of μ to μ^* and of ν to ν^*. Indeed we followed the same construction as for [8, Lem. D.1], with the nuance that φ_I and ψ_J are not necessarily indicator functions. A simple modification of their proof shows that, in our notation:[5]

▶ **Proposition 8** (Modification of [8, Lem. D.1])**.** *Define V as in (12) with the partitions of unity $(\varphi_I)_I$ and $(\psi_J)_J$ as in (13). There exist constants $C, r > 0$ (dependent only on μ^*) such that for any (a, x) with $V(a, x) \le r$, denoting $\mu = \sum_i a_i \delta_{x_i}$,*

$$\mathrm{WFR}_2^2(\mu, \mu^*) \le 2V(a, x)\left(1 + C\frac{\eta}{\sigma}(\lambda\tau)^2\right).$$

The main result of this subsection, proved in Appendix E, is that the CP-PP algorithm converges locally at an exponential rate, as measured by the Lyapunov function. Convergence measured by NI error and by WFR distance (Theorem 2) follows by combining Theorem 9 with Proposition 7 and Proposition 8 as shown in Appendix I.

---

[4] The reason why we need to split Proposition 7 into two parts is that the inequality $V^\alpha \lesssim \mathrm{NI}$ (for any exponent α) cannot be true for all (a, x, b, y). Indeed, NI is bounded, but V may be infinite due to the terms $D(a^*, \bar{a})$ if μ has no mass near one of the x_I^*'s.

[5] Namely the modification to bring to the proof of [8, Lem. D.1] is (in the notations of that paper) to use the transport plan that sends (r, θ) to $\left(\frac{r}{\bar{r}_I} r_I^*, \theta_I^*\right)$ with probability $\varphi_I(\theta)$ and to $(0, \theta)$ with probability $\varphi_0(\theta)$. The present work uses Kullback–Leibler divergence while [8] uses squared Hellinger distance, but the difference can be controlled similarly as in Lemma 46, thanks to Lemma 19. The factor $\frac{\eta}{\sigma}$ can be thought of simply as a linear rescaling of $\|\cdot\|_{\mathcal{X}}^2$.

▶ **Theorem 9.** *Define $(\varphi_I)_I, (\psi_J)_J$ as in (13) and define V as in (12). Choose λ, τ as functions of η, σ as in (23). Fix any $\Gamma_0 \geq 1$. There exist η_0, σ_0 such that for all $\eta \leq \eta_0, \sigma \leq \sigma_0$ with $\Gamma_0^{-1} \leq \frac{\sigma}{\eta} \leq \Gamma_0$, there exists $r_0 > 0$ such that if $V(z^0) \leq r_0$, then the CP-PP iterates satisfy*

$$\forall\, k,\ V(z^k) \leq V(z^0)(1-\kappa)^k$$

for some constant $\kappa > 0$.

More precisely, one can check from the last step of the proof (Appendix E.6) that the rate κ can at most be chosen equal to η^2 times a constant, and that the localness level r_0 can at most be chosen equal to η^3 times a constant (dependent on $(f, \mathcal{X}, \mathcal{Y})$ and Γ_0).

The full proof of the theorem can be found in Appendix E. It has the same general structure as for the exact-parametrization case, but needs to deal with the following difficulties:

- The variables (a, x, b, y) and the solution (a^*, x^*, b^*, y^*) live in different spaces so we cannot just evaluate the characterizing inequality (8) at the solution particles. Instead we identify a notion of "proxy solution particles" $(a^{(*)}, x^{(*)}, b^{(*)}, y^{(*)}) \in \Delta_n \times \mathcal{X}^n \times \Delta_m \times \mathcal{Y}^m$, namely

$$x_i^{(*)} \coloneqq x_i^{k+1} + \sum_{I \in [n^*]} \varphi_{Ii}^{k+1}(x_I^* - x_i^{k+1}) \quad \text{and} \quad a_i^{(*)} \coloneqq \sum_{I \in [n^*]} a_I^* \frac{\varphi_{Ii}^{k+1} a_i^{k+1}}{\bar{a}_I^{k+1}} \tag{14}$$

  where $\varphi_{Ii}^{k+1} = \varphi_I(x_i^{k+1})$, and similarly for $b^{(*)}, y^{(*)}$. We show that evaluating (8) at $(a^{(*)}, x^{(*)}, b^{(*)}, y^{(*)})$ makes the Lyapunov function emerge naturally, yielding a general-case equivalent of (10).
- Compared to the exact-parametrization case, several additional error terms appear, which are much more delicate to control. For example, we need to bound $\frac{1}{2}\sum_I \sum_i a_i^k \left(\varphi_{Ii}^{k+1} - \varphi_{Ii}^k\right) \left\|x_I^* - x_i^k\right\|^2$, a term which we did not appear in (10). This requires some technical work, and it is here that we benefit from choosing the partitions of unity as (13); the choice of the parameters λ and τ also requires care.
- The stray-weight and variance terms of the Lyapunov function do not appear in the error-bound-type inequality of Section 3.3.2, so we do need to combine that result with the quadratic growth and star-convexity properties, contrary to the exact-parametrization case.

3.3 A crucial proof ingredient: lower growth properties

For unconstrained min-max optimization of a smooth convex-concave objective $G(x, y)$, the proof of convergence of the (Euclidean) Proximal Point method essentially reduces to three steps:

1. Letting $\mathrm{gap}(z; \widehat{z}) = G(\widehat{x}, y) - G(x, \widehat{y})$ and $\widetilde{\mathrm{gap}}(z; \widehat{z}) = \left\langle \begin{pmatrix} \nabla_x \\ -\nabla_y \end{pmatrix} G(\widehat{z}), \widehat{z} - z \right\rangle$, notice that

   $$\widetilde{\mathrm{gap}}(z; \widehat{z}) = \mathrm{gap}(z; \widehat{z}) + D_{G(\cdot, \hat{y})}(x, \widehat{x}) - D_{G(\hat{x}, \cdot)}(y, \widehat{y}) \geq \mathrm{gap}(z; \widehat{z})$$

   where D denotes a Bregman divergence, by convexity-concavity.
2. The Proximal Point update is characterized by the variational inequality, analogous to (8),

   $$\forall\, z,\ \eta\, \widetilde{\mathrm{gap}}(z; z^{k+1}) \leq \left\langle z - z^{k+1}, z^{k+1} - z^k \right\rangle = \frac{1}{2}\left\|z - z^k\right\|^2 - \frac{1}{2}\left\|z - z^{k+1}\right\|^2 - \frac{1}{2}\left\|z^{k+1} - z^k\right\|^2.$$

3. In particular evaluating at the saddle point z^* assumed unique for simplicity,

   $$\begin{aligned} \frac{1}{2}\left\|z^* - z^{k+1}\right\|^2 \leq \frac{1}{2}\left\|z^* - z^k\right\|^2 - \eta\, \mathrm{gap}(z^*; z^{k+1}) \\ - \eta \left[D_{G(\cdot, y^{k+1})}(x^*, x^{k+1}) - D_{G(x^{k+1}, \cdot)}(y^*, y^{k+1})\right] - \frac{1}{2}\left\|z^{k+1} - z^k\right\|^2. \end{aligned}$$

   Depending on the properties of G, we lower-bound one of the three terms appearing with a negative sign on the right-hand side. For example,
   - If G satisfies (a min-max analog of) the quadratic growth property [15, Def. 5.1], i.e., if there exists $C > 0$ such that

   $$\forall\, z,\ \mathrm{gap}(z^*; z) \geq \frac{C}{2}\left\|z^* - z\right\|^2,$$

   then we can directly conclude to exponential decrease of the Lyapunov function $V(z) = \frac{1}{2}\left\|z^* - z\right\|^2$ with a rate at least $\frac{C}{2}\eta$.

- If G satisfies the error bound property with constant $C' > 0$ [17, 35], i.e., if

$$\forall\, z,\ \left\| \begin{pmatrix} \nabla_x G(z) \\ -\nabla_y G(z) \end{pmatrix} \right\| \geq C' \|z - z^*\|,$$

  then since $z^{k+1} - z^k = \eta \begin{pmatrix} -\nabla_x G(z^{k+1}) \\ \nabla_y G(z^{k+1}) \end{pmatrix}$, we can conclude to exponential convergence with a rate at least $(C'\eta)^2$.
- We note that convexity-concavity of G is not essential. (Suppose existence and uniqueness of the saddle point z^* is guaranteed by some other property that convexity-concavity.) Indeed, to lower-bound the second term on the right-hand side, it suffices to have G (μ'-strongly) star-convex-concave [15, Def. 5.1], that is

$$\forall\, z,\ D_{G(\cdot,y)}(x^*, x) - D_{G(x,\cdot)}(y^*, y) \geq \frac{\mu'}{2} \|z^* - z\|^2.$$

  In total, if G is (μ'-strongly) star-convex-concave, satisfies quadratic growth with constant C, and error bound with constant C', then we can conclude to exponential convergence with a rate at least $\frac{C+\mu'}{2}\eta + (C'\eta)^2$.

In our case (constrained min-max optimization of the overparametrized objective $F_{n,m}$ using the divergence $D\left((a,x),(\widehat{a},\widehat{x})\right) = D(a,\widehat{a}) + \frac{\eta}{2\sigma}\sum_i \widehat{a}_i \|x_i - \widehat{x}_i\|^2$ which is non-Euclidean and not even a Bregman divergence) the analysis is significantly more technical, but it involves the same basic ingredients.

3.3.1 "Quadratic growth" with respect to the position and stray weight variables

We establish a quadratic growth property for $F_{n,m}$ involving only some of the desired terms in the lower bound. Analogously to the analysis of [8] for minimization, the proof relies on the non-degeneracy Assumptions 5–6; a crucial difference however, is that we do not have quadratic growth in the weight variables. To be precise, compared to the assumption (A5) of [8], our non-degeneracy assumption concerns only the so-called local kernels H_I, H_J, and the min-max analog of the global kernel (K in that paper's notations) is necessarily zero due to the bilinearity of $F(\mu,\nu)$.

The precise statement of our result is given in Appendix C.1; here we state a simplified version to give the intuition.

▶ **Lemma 10** ("Quadratic growth", simplified). *There exist constants $r, C > 0$ only dependent on $(f, \mathcal{X}, \mathcal{Y})$ (and on λ, τ in the general case) such that, for any $z = (a, x, b, y)$ with $V_1(z) \leq r$, then*

$$F(\mu, \nu^*) - F(\mu^*, \nu) \geq C\left(V_{\mathrm{pos}}(z) + \bar{a}_0 + \bar{b}_0\right)$$

where $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j}$ (and $\bar{a}_0 = \bar{b}_0 = 0$ by definition in the exact-parametrization case).

Note that $\|\bar{a} - a^*\|_1$ does not appear in this inequality, but it will appear in the error-bound-type inequality discussed in the next paragraph. Conversely $\bar{a}_0$ and $\sum_I \bar{a}_I \operatorname{Tr}(\Sigma_I)$ appear in the inequality of this paragraph but not of the next.

3.3.2 "Error bound" with respect to the weight and aggregated position variables

It is well-known that the error bound property holds for strongly-convex-strongly-concave and smooth min-max objectives, or for bilinear objectives constrained to a product of polytopes [35]. In our case, the reparametrized objective $F_{n,m}(a,x,b,y)$ is bilinear in the weight components (a,b), and intuitively it possesses some local strong convexity-concavity in the position components (x,y) thanks to Assumption 6. But these two facts are not enough to directly show an error bound inequality, because the constant C for the components (a,b) may depend arbitrarily badly on (x,y) a priori. Instead we use an argument, inspired by [37, Lem. 14], that also exploits the Assumption 3 of uniqueness of the MNE.

Again, the precise statement of our result is deferred to the appendix Appendix C.2; here we state an informal version to give the intuition. We also refer to the second paragraph of that appendix for an interpretation of the quantity appearing on the left-hand side.

▶ **Lemma 11** ("Error bound", informal). *Consider any* $\widehat{z} = (\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y}) \in \Delta_n \times \mathcal{X}^n \times \Delta_m \times \mathcal{Y}^m$. *For any* $(A_I)_{I\in[n^*]}, (B_J)_{J\in[m^*]}$ *(and* $A_0 = B_0 = 0$*) and* $(X_I)_{I\in[n^*]}, (Y_J)_{J\in[m^*]}$, *define "proxy particles" from* (A, X, B, Y) *and* $(\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y})$ *analogously to* (14), *and denote them by* z. *There exist* $r, C > 0$ *only dependent on* $(f, \mathcal{X}, \mathcal{Y})$ *such that if* $V_1(\widehat{z}) \le r$, *then*

$$\max_{A,X,B,Y} \widetilde{\mathrm{gap}}(z;\widehat{z}) \ge C\sqrt{\sum_I d_h(a_I^*, \widehat{a}_I) + \sum_J d_h(b_J^*, \widehat{b}_J) + \sum_I \widehat{a}_I \left\|\widehat{x}_I - x_I^*\right\|^2 + \sum_J \widehat{b}_J \left\|\widehat{y}_J - y_J^*\right\|^2} + O(V_1(\widehat{z})).$$

3.3.3 Local "star-convexity-concavity" (strong with respect to the position variables)

Note that $\widetilde{\mathrm{gap}}(z;\widehat{z}) - \mathrm{gap}(z;\widehat{z}) = D_{F_{n,m}(\cdot,\hat{y})}(x,\widehat{x}) - D_{F_{n,m}(\hat{x},\cdot)}(y,\widehat{y})$ where D denotes a Bregman divergence. Intuitively, $F_{n,m}$ is bilinear in the weight variables and, in a neighborhood of the MNE, it possesses some local strong convexity-concavity with respect to the position variables thanks to Assumption 6. And indeed, by Taylor expansions, one can obtain lower-bounds on $\widetilde{\mathrm{gap}}(z;\widehat{z}) - \mathrm{gap}(z;\widehat{z})$ consisting of positive terms and of error terms in $V(z)$, $V(\widehat{z})$. Note however that (for the general case) due to the overparametrization, there are many ways to write Taylor expansions.

Specifically, we will use the following bound. Again the precise statement of the result is deferred to the appendix Appendix C.3; here we state an informal version to give the intuition.

▶ **Lemma 12** ("Local star-convexity-concavity", informal). *Consider any* $\widehat{z} = (\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y}) \in \Delta_n \times \mathcal{X}^n \times \Delta_m \times \mathcal{Y}^m$, *and let* $z^{(*)} = (a^{(*)}, x^{(*)}, b^{(*)}, y^{(*)})$ *the "proxy solution particles" defined as in* (14). *Denote* $\widehat{\mu} = \sum_{i=1}^n \widehat{a}_i \delta_{\hat{x}_i}$ *and* $\widehat{\nu} = \sum_{j=1}^m \widehat{b}_j \delta_{\hat{y}_j}$. *There exist* $r, C > 0$ *only dependent on* $(f, \mathcal{X}, \mathcal{Y})$ *such that if* $V_1(\widehat{z}) \le r$, *then for an appropriate choice of the partitions of unity* φ_I, ψ_J,

$$\widetilde{\mathrm{gap}}(z^{(*)};\widehat{z}) \ge F(\widehat{\mu}, \nu^*) - F(\mu^*, \widehat{\nu}) + CV_{\mathrm{pos}}(\widehat{z}) + O\left(V_1(\widehat{z})^{3/2}\right).$$

4 Numerical experiments

In this section, we illustrate the CP-PP algorithm and its convergence properties on simple examples of applications. As discussed in Section 2.2, in experiments we actually run the CP-MP algorithm since the CP-PP update cannot be computed exactly, but based on Lemmas 49 and 50 we strongly expect the same convergence behavior for these two algorithms, as proved in the exact-parametrization case in Proposition 6.

Julia code to reproduce the experiments is publicly available online at `https://github.com/guillaumew16/particle-MNE`.

4.1 Payoff drawn from a Gaussian process

We start by an application of our method on a toy example where the payoff function is drawn from a Gaussian process. More precisely we apply CP-MP on the function $f : \mathbb{T}^{d_x} \times \mathbb{T}^{d_y} \to \mathbb{R}$ defined by

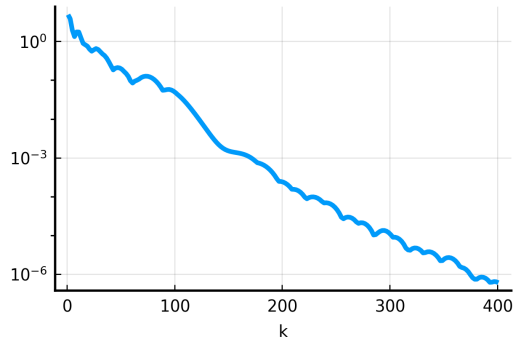
$$f(x,y) = \Re \sum_{|k|\le K} \sum_{|l|\le L} c_{k,l} e^{2\pi i(\langle k,x\rangle + \langle l,y\rangle)}$$

where the $c_{k,l} \in \mathbb{C}$ are drawn randomly, namely $\Re[c_{k,l}], \Im[c_{k,l}]$ are drawn i.i.d. from the standard normal distribution. The orders K and L control the smoothness of the function. Remark that the game is separable, i.e., f can be written as a finite sum of the form $f(x,y) = \sum_{k,l} c'_{k,l} g_k(x) h_l(y)$ for some $c'_{k,l} \in \mathbb{R}$ and continuous g_k, h_l, since $f(x,y) = \sum_{k,l} |c_{k,l}| \cos\left(2\pi \langle k, x\rangle + 2\pi \langle l, y\rangle + \arg(c_{k,l})\right)$ and $\cos(a+b) = \cos a \cos b - \sin a \sin b$; so we are guaranteed that a sparse MNE exists [34, Cor. 2.10].

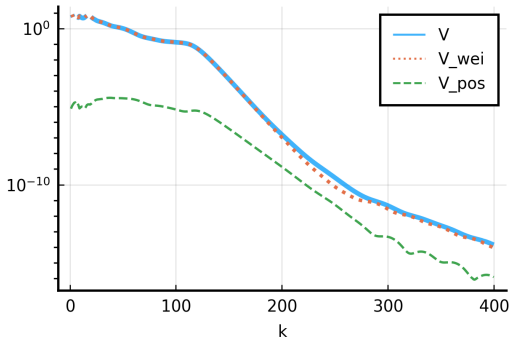
We illustrate the behavior of the CP-MP algorithm on such a payoff function with $d_x = d_y = 1$, $K = L = 3$ and $n = m = 15$. A contour plot of f is contained in Figure 3a.

In Figure 2, we plot the NI errors and Lyapunov potentials of the iterates for $\eta = 0.04$ and $\sigma = 0.001$, up to $T = 400$. Those values decrease exponentially as expected from our upper bounds. In order to compute the NI errors, we computed $\max_\nu F(\mu^k, \nu) = \max_\nu \int_{\mathcal{Y}} ((\mu^k)^\top F)\,\mathrm{d}\nu = \max_{y\in\mathcal{Y}} ((\mu^k)^\top F)(y)$ simply by discretization of $\mathcal{Y} = \mathbb{T}^1$. In order to compute the Lyapunov potentials $V(z^k)$ defined in (12), we used an estimation of the $(a_I^*, x_I^*)_I, (b_J^*, y_J^*)_J$ obtained by clustering the particles of μ^{2T}, ν^{2T}.

In Figure 3, we plot (a smoothed version of) the measures (μ^T, ν^T) as well as the first variations $(F\nu^T)(x)$ and $((\mu^T)^\top F)(y)$ at the last iterate. On all three subfigures, green lines indicate the support of μ^T and ν^T. The

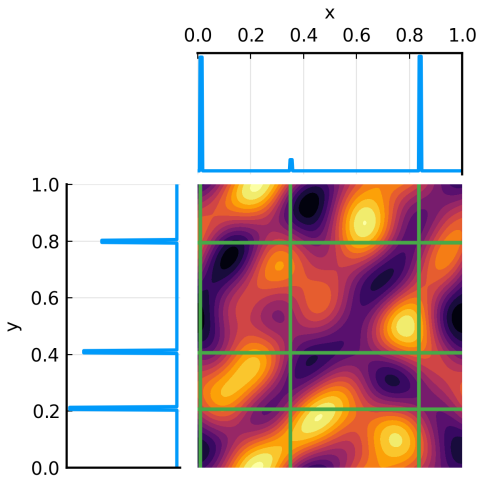


(a) NI errors $\mathrm{NI}(\mu^k, \nu^k)$ (log-linear scale)

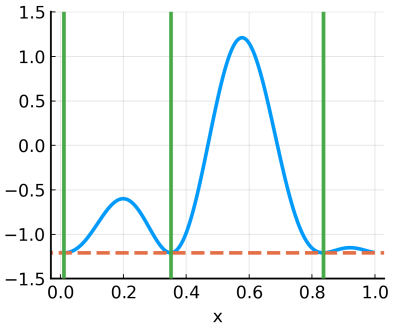


(b) Lyapunov potentials $V(z^k)$ (log-linear scale)

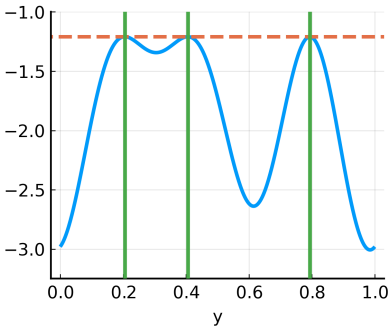
Figure 2 Optimality metrics of CP-MP iterates for Gaussian process payoff



(a) The measures smoothed by convolution with $h = \mathbb{1}_{[0,0.01]}$.
Middle: contour plot of $f(x, y)$.
Top: $(\mu^T * h)(x)$. Left: $(\nu^T * h)(y)$



(b) First variation $(F\nu^T)(x)$ at the last iterate.
Orange dashed line: payoff at equilibrium $F(\mu^*, \nu^*)$



(c) First variation $((\mu^T)^\top F)(y)$ at the last iterate.
Orange dashed line: payoff at equilibrium $F(\mu^*, \nu^*)$

Figure 3 Smoothed measures and first variations at the last iterate

iterates converge to a sparse measure (here $n^* = m^* = 3$), as expected. The first variations visibly satisfy the inequalities (4), which characterize the MNE, as well as the non-degeneracy Assumptions 5–6.

Convergence of the continuous-time flow

Interestingly, for payoff functions of this form, we observe experimentally that the continuous-time flow corresponding to CP-MP typically also converges to the MNE. This behavior is not captured by our upper bound. Indeed, we observe that the slopes of the lines in the log-linear plots of Figure 2 scale as η instead of η^2. Moreover, experimentally, the explicit time-discretization CP-MDA (i.e. Algorithm 1 with $L = 1$) also converges exponentially to the solution in this setting.

This phenomenon is specific to CP-MP, as it does not arise for Mirror Prox in finite games.[6] We emphasize that it is not always the case that the continuous-time flow of CP-MP converges, as shown experimentally for the synthetic example below.

The mechanism behind this phenomenon is explained in depth in the follow-up work [36] (subsequent to the completion of this work), where in particular the conditions for convergence of the continuous-time flow are described precisely, in the exact-parametrization case.

[6] For finite games with a unique MNE, Mirror Descent-Ascent diverges, unless the MNE consists of two Dirac deltas (i.e. there exists a pure-strategy Nash equilibrium) [3]. Moreover, in all our experiments with random payoff matrices, we observed that the convergence rate of Mirror Prox scaled as η^2.

▶ **Example 13** (The continuous-time flow may not converge)**.** Take $d_x = d_y = 1$ and $c_{2,0} = -i$, $c_{0,2} = -i$, $c_{1,1} = 2$, $c_{k,l} = 0$ otherwise, i.e.,

$$f(x,y) = \sin(4\pi x) \ + \ \sin(4\pi y) \ + \ 2\cos(2\pi x + 2\pi y).$$

For this payoff function, the MNE is unique and equal to $(\mu^*, \nu^*) = \left(\frac{1}{2}\delta_{\frac{3}{8}} + \frac{1}{2}\delta_{\frac{7}{8}}, \frac{1}{2}\delta_{\frac{1}{8}} + \frac{1}{2}\delta_{\frac{5}{8}}\right)$. Indeed,

1. Using that $-1 \leq \sin \leq 1$ one can check that this (μ^*, ν^*) is a MNE, and so the value at optimum is $\rho = 0$.
2. Suppose by contradiction that there exists (μ', ν') a MNE with $\mathbb{E}_{\nu'}[\sin(4\pi y) - 1] < 0$, and pose $x_+ = \frac{3}{8}, x_- = \frac{7}{8}$. Using that $\cos(2\pi x_+ + 2\pi y) + \cos(2\pi x_- + 2\pi y) = 0$ for all y, we have either $\mathbb{E}_{\delta_{x_+}, \nu'}[2\cos(2\pi x + 2\pi y)] \leq 0$ or $\mathbb{E}_{\delta_{x_-}, \nu'}[2\cos(2\pi x + 2\pi y)] \leq 0$. So $F(\delta_{x_+}, \nu') < 0 = \rho$ or $F(\delta_{x_-}, \nu') < \rho$, contradicting optimality of ν'.
3. By the previous point and the symmetric argument for μ', any MNE (μ', ν') must satisfy $\mathbb{E}_{\mu'}[\sin(4\pi x) + 1] = 0$ and $\mathbb{E}_{\nu'}[\sin(4\pi y) - 1] = 0$, i.e., must be of the form $\mu' = a\delta_{\frac{3}{8}} + (1-a)\delta_{\frac{7}{8}}$, $\nu' = b\delta_{\frac{1}{8}} + (1-b)\delta_{\frac{5}{8}}$. By explicit calculations, one can show that necessarily $a = b = \frac{1}{2}$.

On the other hand, experimentally we observe that CP-MDA does not converge, while CP-MP converges with an exponential rate that scales as η^2.

4.2 Max-$\mathcal{F}_1$-margin classification with two-layer neural networks

A well-known machine learning task which uses the min-max framework is max-margin classification. In particular when using a two-layer neural network as the classifier, the training task is exactly of the form (1). Indeed, a two-layer network with non-decreasing positive-homogeneous activation σ (without bias terms) can be represented as a signed measure $\nu_\pm$ on the space of normalized hidden neurons $\Theta = \mathbb{S}^{d-1} = \left\{\theta \in \mathbb{R}^d; \left\|\theta\right\|_2 = 1\right\}$ via

$$\mathrm{NN}(x; \nu_\pm) = \int_\Theta \sigma(\theta^\top x) \, \mathrm{d}\nu_\pm(\theta),$$

or equivalently as a non-negative measure ν on the space $\Theta_+ \sqcup \Theta_-$, the disjoint union of two copies of Θ, via

$$\mathrm{NN}(x; \nu) = \int_{\Theta_+} \sigma(\theta_+^\top x) \, \mathrm{d}\nu(\theta_+) - \int_{\Theta_-} \sigma(\theta_-^\top x) \, \mathrm{d}\nu(\theta_-).$$

Two-layer networks with bias terms can be represented in the same way, by appending a constant component 1 to the input vector x and taking $\Theta = \mathbb{S}^d$ instead of $\mathbb{S}^{d-1}$. One can define the $\mathcal{F}_1$ norm of a function $f : \mathbb{R}^d \to \mathbb{R}$ as the infimum of $\nu(\Theta_+ \sqcup \Theta_-)$ over all ν such that $f = \mathrm{NN}(\,\cdot\,; \nu)$. Balls for this norm admit advantageous estimation/approximation trade-offs in a supervised learning task [2].

Consider a supervised classification task with covariates $x \in \mathbb{R}^d$ and labels $y \in \{-1, 1\}$. Given N observations $(x_i, y_i)_{1 \leq i \leq N}$, the max-$\mathcal{F}_1$-margin classification task consists in finding ν that maximizes the following problem

$$\begin{aligned}
&\max_{\substack{\nu \in \mathcal{M}(\Theta_+ \sqcup \Theta_-) \\ \nu(\Theta_+ \sqcup \Theta_-) = 1}} \ \min_{1 \leq i \leq N} \ y_i \ \mathrm{NN}(x_i; \nu) \\
&\equiv \max_{\nu \in \mathcal{P}(\Theta_+ \sqcup \Theta_-)} \ \min_{a \in \mathcal{P}([N])} \ \sum_{i=1}^N a_i y_i \left(\int_{\Theta_+} \sigma(\theta_+^\top x_i) \, \mathrm{d}\nu(\theta_+) - \int_{\Theta_-} \sigma(\theta_-^\top x_i) \, \mathrm{d}\nu(\theta_-) \right).
\end{aligned}$$

This problem is indeed an instance of (1) for $\mathcal{X} = [N]$, $\mathcal{Y} = \Theta_+ \sqcup \Theta_-$, and $f(i, \theta) = \begin{cases} y_i \sigma(\theta^\top x_i) & \text{if } \theta \in \Theta_+ \\ -y_i \sigma(\theta^\top x_i) & \text{if } \theta \in \Theta_- \end{cases}$.

Numerical results

As detailed above, the max-$\mathcal{F}_1$-margin classification problem can be written as

$$\max_{\nu \in \mathcal{P}(\Theta_+ \sqcup \Theta_-)} \ \min_{a \in \mathcal{P}([N])} \ \sum_{i=1}^N \int_{\Theta_+ \sqcup \Theta_-} a_i f(i, \theta) \, \mathrm{d}\nu(\theta).$$

It is straightforward to adapt the CP-MP algorithm to this setting. Namely, choose $m' = 2m$ a number of neurons, reparametrize by $\nu = \sum_{j=1}^{2m} b_j \delta_{\theta_j}$ for $\theta_1, \dots, \theta_m \in \Theta_+$ and $\theta_{m+1}, \dots, \theta_{2m} \in \Theta_-$, and consider the reparametrized problem

$$\min_{a \in \Delta_N} \ \max_{\substack{b \in \Delta_{2m} \\ \theta \in (\mathbb{S}^{d-1})^{2m}}} \left\{ \sum_{i=1}^N \sum_{j=1}^{2m} a_i b_j \cdot \left(\mathbf{1}_{[j \leq m]} - \mathbf{1}_{[j > m]} \right) \cdot y_i \sigma(\theta_j^\top x_i) \ =: \ F_{n,m}(a, (b, \theta)) \right\}.$$

We can then apply Algorithm 1 with $y = \theta$ and with x_i kept constant equal to i for all i.

In Figure 4 we present the results of an experiment with $N = 5$ samples, two positively labeled and three negatively labeled, and $2m = 2 * 50$ neurons and activation $\sigma(s) = \max(0, s)^3$. The dimensionality of the problem is $d = 3$ with each sample having 1 as the last coordinate, meaning that the last component of θ acts as a bias term. Our analysis does not cover this case, strictly speaking, since one strategy space is discrete, and there is no guarantee that the MNE is unique; yet the experimental results indicate a similar behavior as for continuous games.

- Figure 4a shows that the NI error, here $\mathrm{NI}(a^k, \nu^k) = \max_\theta \left| \sum_{i=1}^N a_i y_i \sigma(\theta^\top x_i) \right| - \min_i y_i \mathrm{NN}(x_i; \nu^k)$, decreases exponentially to 0.
- In particular the margin is non-negative at optimum so all points are classified correctly, as expected from the universality of two-layer neural networks [31]. This can also be seen from the decision regions shown in Figure 4b.
- The solution found (ν^T) turns out to be sparse, as shown by the plots in Figure 4c, where blue dots correspond to positively weighted neurons and red dots to negatively weighted neurons, and the distance from the origin represents the associated magnitude b_j^T. A green sphere of radius $\frac{1}{m}$ was added for scale.
- While they are not represented in the figure so as not to overload it, the variables a are also of interest as a measure of each sample's importance. For example in this experiment, a_i^T is close to zero for the topmost sample ($x_i \approx (0, 2)$ and $y_i = +1$), and non-zero for all other samples. In particular, removing the topmost sample from the dataset does not modify the learned network.

The activation function $\sigma(s) = \max(0, s)^3$ chosen for this experiment has locally Lipschitz-continuous second derivative, so our results' smoothness assumption on the payoff is verified. Interestingly, when using the ReLU activation $\sigma(s) = \max(0, s)$ for the same toy dataset, we observe that the NI error first decreases at an exponential rate and then oscillates around a value of about 10^{-3}, even for large m. For $\sigma(s) = \max(0, s)^2$, in all our experiments we observed that the NI error vanishes exponentially.

4.3 Distributionally-robust classification with two-layer neural networks

Consider again a supervised classification task with covariates $x \in \mathbb{R}^d$ and labels $y \in \{-1, 1\}$. Consider a dataset of N observations $(\widehat{x}_k, \widehat{y}_k)_{1 \le k \le N}$ and let $\widehat{\mu} = \frac{1}{N} \sum_{k=1}^N \delta_{(\hat{x}_k, \hat{y}_k)}$ the empirical distribution. Let W_∞ denote the L^∞-Wasserstein distance on $\mathcal{P}(\mathbb{R}^d \times \{-1, 1\})$, defined by

$$W_\infty(\mu, \mu') = \inf_{\gamma \in \Pi(\mu, \mu')} \max_{\substack{((x,y),(x',y')) \\ \in \mathrm{supp}(\gamma)}} d((x, y), (x', y')) \quad \text{where} \quad d((x, y), (x', y')) = \begin{cases} \left\| x - x' \right\|_2 & \text{if } y = y' \\ +\infty & \text{otherwise} \end{cases}$$

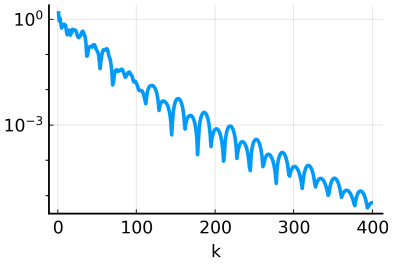
where $\Pi(\mu, \mu')$ is the set of couplings of μ and μ'. Fix a "robustness level" $r > 0$. In the spirit of [28], the distributionally-robust classification task with respect to W_∞, using two-layer neural networks $\mathrm{NN}(\,\cdot\,; \nu)$, is to find ν that maximizes

$$\begin{aligned} &\max_{\substack{\nu \in \mathcal{M}(\Theta_+ \sqcup \Theta_-) \\ \nu(\Theta_+ \sqcup \Theta_-) = 1}} \min_{\substack{\mu \in \mathcal{P}(\mathbb{R}^d \times \{-1,1\}) \\ W_\infty(\mu, \hat{\mu}) \le r}} \int_{\mathbb{R}^d \times \{-1,1\}} y \ \mathrm{NN}(x; \nu) \, \mathrm{d}\mu(x, y) \\ &\equiv \max_{\nu \in \mathcal{P}(\Theta_+ \sqcup \Theta_-)} \min_{\substack{\mu \in \mathcal{P}(\mathbb{R}^d \times \{-1,1\}) \\ W_\infty(\mu, \hat{\mu}) \le r}} \int_{\mathbb{R}^d \times \{-1,1\}} \int_{\Theta_+ \sqcup \Theta_-} f\left((x, y), \theta\right) \, \mathrm{d}\nu(\theta) \, \mathrm{d}\mu(x, y) \end{aligned}$$

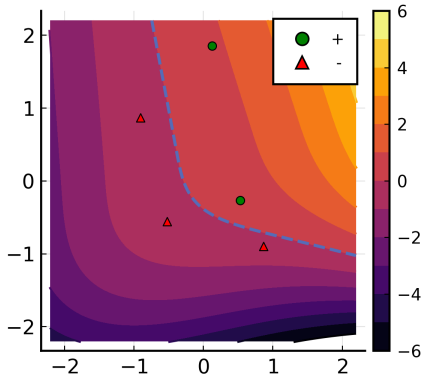
with $f\left((x, y), \theta\right) = \begin{cases} y\sigma(\theta^\top x) & \text{if } \theta \in \Theta_+ \\ -y\sigma(\theta^\top x) & \text{if } \theta \in \Theta_- \end{cases}$ a "payoff" function over $\left(\mathbb{R}^d \times \{-1, 1\}\right) \times (\Theta_+ \sqcup \Theta_-)$. More concretely, since $\widehat{\mu} = \frac{1}{N} \sum_{k=1}^N \delta_{\hat{x}_k, \hat{y}_k}$, then $W_\infty(\mu, \widehat{\mu}) \le r$ means that

$$\mathrm{supp}(\mu) \subset \{(x, y);\ \exists\, k, d\left((x, y), (\widehat{x}_k, \widehat{y}_k)\right) \le r\} = \bigcup_{k \in [N]} (\widehat{x}_k + r\mathbb{B}) \times \{\widehat{y}_k\}$$

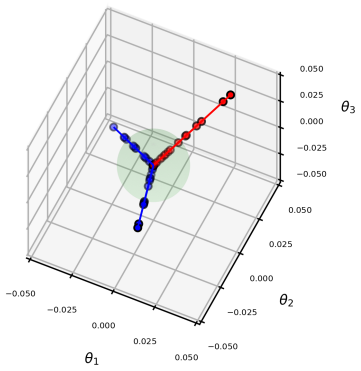
where $\mathbb{B}$ denotes the unit Euclidean ball. In the language of adversarial robustness, the inner minimization means that we train the model $\mathrm{NN}(\,\cdot\,; \nu)$ to correctly classify potential adversarial examples that are within a distance of r from an instance present in the dataset.



(a) NI errors $\mathrm{NI}(a^k, \nu^k)$ (log-linear scale)

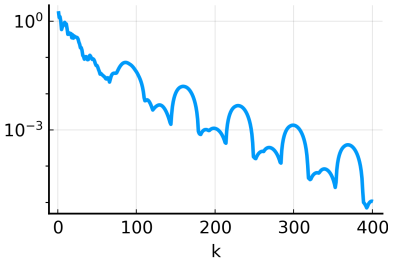


(b) Soft decision regions (logits) at the last iteration. The dashed blue line indicates the decision boundary.

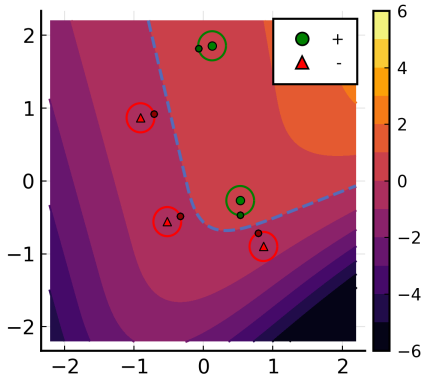


(c) Neurons at the last iteration

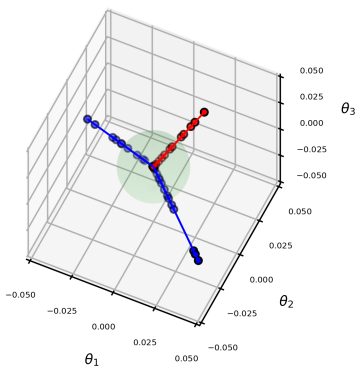
Figure 4 Results for the max-$\mathcal{F}_1$-margin classification experiment



(a) NI errors $\mathrm{NI}(\mu^k, \nu^k)$ (log-linear scale)

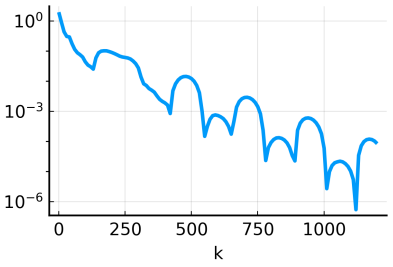


(b) Decision regions at the last iteration. See text for further description.

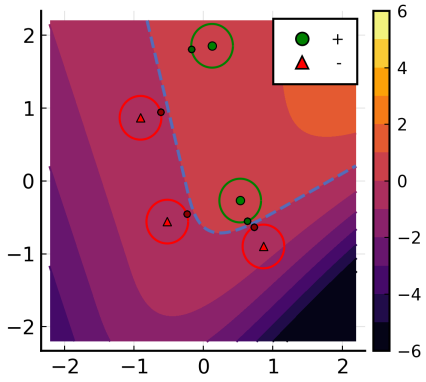


(c) Neurons at the last iteration

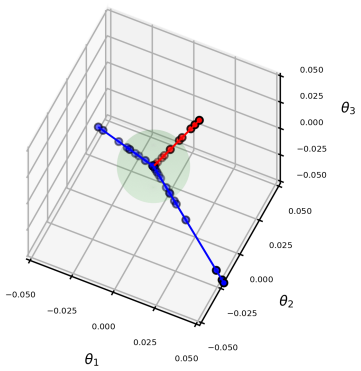
Figure 5 Results for the distributionally-robust classification experiment with $r = 0.2$



(a) NI errors $\mathrm{NI}(\mu^k, \nu^k)$ (log-linear scale)



(b) Decision regions at the last iteration. See text for further description.



(c) Neurons at the last iteration

Figure 6 Results for the distributionally-robust classification experiment with $r = 0.3$

Numerical results

We showed how the task of distributionally-robust classification can be rewritten as

$$\max_{\nu \in \mathcal{P}(\Theta_+ \sqcup \Theta_-)} \min_{\mu \in \mathcal{P}\left(\bigcup_{k \in [N]} (\hat{x}_k + r\mathbb{B}) \times \{\hat{y}_k\}\right)} \int_{\mathbb{R}^d \times \{-1,1\}} \int_{\Theta_+ \sqcup \Theta_-} f\left((x,y),\theta\right) \mathrm{d}\nu(\theta) \mathrm{d}\mu(x,y).$$

Let us adapt the CP-MP algorithm to this setting.

- For the classifier (ν), similarly to the previous example, choose $m' = 2m$ a number of neurons and let $\nu = \sum_{j=1}^{2m} b_j \delta_{\theta_j}$ with $\theta_1, \dots, \theta_m \in \Theta_+$ and $\theta_1, \dots, \theta_m \in \Theta_-$.
- For the adversary (μ), choose $n' = Nn$ a number of particles (n per sample), and let $\mu = \sum_{k=1}^{N} \sum_{i=1}^{n} a_{ki} \delta_{(\hat{x}_k + u_{ki}, \hat{y}_k)}$ with $\|u_{ki}\|_2 \leq r$. To deal with the constraint on the u_{ki}'s, we project those variables back to $r\mathbb{B}$ after each gradient step.

We obtain the reparametrized problem

$$\min_{\substack{a \in \Delta_{Nn} \\ u \in (r\mathbb{B})^{Nn}}} \max_{\substack{b \in \Delta_{2m} \\ \theta \in (\mathbb{S}^{d-1})^{2m}}} \left\{ \sum_{k=1}^{N} \sum_{i=1}^{n} \sum_{j=1}^{2m} a_{ki} b_j \cdot \left(\mathbf{1}_{[j \leq m]} - \mathbf{1}_{[j > m]} \right) \cdot \widehat{y}_k \sigma\left(\theta_j^\top (\widehat{x}_k + u_{ki}) \right) =: F_{n,m}((a,u),(b,\theta)) \right\}$$

on which we can apply Algorithm 1 with $x = u$, $y = \theta$, modified with a projection step to ensure $u \in r\mathbb{B}$.

In Figure 5 (resp. Figure 6), we show the results of experiments with the same dataset and the same network architecture as in the previous subsection, with robustness level $r = 0.2$ (resp. $r = 0.3$), and using $n = 10$ particles per datapoint. The bias terms are taken into account, i.e., each u_{ki} has 0 as the last coordinate.

- In both experiments, similar to the previous subsection, the NI error decreases exponentially to 0 (Figure 5a, Figure 6a).
- In particular the robust margin $\min_k \min_{x \in \hat{x}_k + r\mathbb{B}} \widehat{y}_k \mathrm{NN}(x; \nu^T)$ is non-negative at optimum. In other words, the disks of radius r around the sample covariates are classified correctly, as can be seen in Figure 5b, Figure 6b, where the disks' boundaries are shown by green and red circles. In those figures we also represented the adversary's support points $(\widehat{x}_k + u_{ki}^T)$ by slightly darker marks. We see that they are concentrated on the points of the disks that are closest to the decision boundary (dashed blue line).
- Just like in the max-$\mathcal{F}_1$-margin experiment of the previous subsection, the learned network (ν^T) is sparse, as shown in Figure 5c, Figure 6c. In fact, max-$\mathcal{F}_1$-margin can be seen as an instance of distributionally-robust classification with level $r = 0$, and increasing r seems to perturb the learned neurons in a continuous way.
- Again, the variables a are not represented in the figures to avoid overloading them. In both experiments, $\sum_{i=1}^{n} a_{ki}^T$ is close to zero for the topmost sample ($\widehat{x}_k \approx (0, 2)$ and $\widehat{y}_k = +1$) and non-zero for all other samples, just like in the max-$\mathcal{F}_1$-margin experiment.

5 Conclusion

In this paper, we showed that weighted particle methods can be successfully used to compute the MNE of continuous games. Specifically, we prove local exponential convergence of Conic Particle Proximal Point (CP-PP) under non-degeneracy assumptions. This algorithm is easily implementable as a descent-ascent method on a reparametrized finite-dimensional (but nonconvex-nonconcave) objective, and corresponds to an implicit time-discretization of the Wasserstein–Fisher–Rao gradient flow. Applied to max-margin and distributionally-robust classification, our result indicates (and our numerical experiments confirm) that training the classifier and the adversary simultaneously is sufficient for convergence, with no need for timescale separation nor for any reformulation as in [28].

An interesting question for further research would be to relax the assumption that the step-sizes for the weight (η) and position variables (σ) are of the same order, as this would allow a direct comparison with the convergence behavior of pure Fisher–Rao or pure Wasserstein gradient methods. Another open direction is to adapt our algorithm and analysis to the case where only noisy access to the payoff function or its derivatives is available. Finally, it could be interesting to extend our study of distributionally-robust classification (Section 4.3) to regression tasks, or to classification using the logistic loss.

Acknowledgments

We would like to thank Praneeth Netrapalli for insightful discussions.

References

1 Ehsan Amid and Manfred K. K. Warmuth. Reparameterizing mirror descent as gradient descent. *Adv. Neural Inf. Process. Syst.*, 33:8430–8439, 2020.

2 Francis Bach. Breaking the curse of dimensionality with convex neural networks. *J. Mach. Learn. Theory*, 18(1):629–681, 2017.

3 James P. Bailey and Georgios Piliouras. Multiplicative weights update in zero-sum games. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 321–338, 2018.

4 Sébastien Bubeck. ORF523: Mirror Descent, part II/II | I'm a bandit, 2013. `https://blogs.princeton.edu/imabandit/2013/04/18/orf523-mirror-descent-part-iiii/`.

5 Sébastien Bubeck. ORF523: Mirror Prox | I'm a bandit, 2013. `https://blogs.princeton.edu/imabandit/2013/04/23/orf523-mirror-prox/`.

6 Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. *Adv. Neural Inf. Process. Syst.*, 34:27952–27964, 2021.

7 Lénaïc Chizat. Convergence rates of gradient methods for convex optimization in the space of measures. `https://arxiv.org/abs/2105.08368`, 2021.

8 Lénaïc Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Math. Program.*, 194(1):487–532, 2022.

9 Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *J. Funct. Anal.*, 274(11):3090–3123, 2018.

10 Constantinos Daskalakis and Ioannis Panageas. Last-Iterate Convergence: Zero-Sum Games and Constrained Min-Max Optimization. In *10th innovations in theoretical computer science conference, ITCS 2019, January 10–12, 2019, San Diego, CA, USA*, volume 124 of *LIPIcs – Leibniz International Proceedings in Informatics*. Leibniz Zentrum für Informatik, 2019. article no. 27 (18 pages).

11 Jelena Diakonikolas, Constantinos Daskalakis, and Michael I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2746–2754. PMLR, 2021.

12 Carles Domingo-Enrich, Samy Jelassi, Arthur Mensch, Grant Rotskoff, and Joan Bruna. A mean-field analysis of two-player zero-sum games. *Adv. Neural Inf. Process. Syst.*, 33:20215–20226, 2020.

13 Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A Variational Inequality Perspective on Generative Adversarial Networks. In *International Conference on Learning Representations*, 2018.

14 Irving L. Glicksberg. A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points. *Proc. Am. Math. Soc.*, 3(1):170–174, 1952.

15 Charles Guille-Escuret, Manuela Girotti, Baptiste Goujaud, and Ioannis Mitliagkas. A study of condition numbers for first-order optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1261–1269. PMLR, 2021.

16 Ya-Ping Hsieh, Chen Liu, and Volkan Cevher. Finding mixed Nash equilibria of generative adversarial networks. In *International Conference on Machine Learning*, pages 2810–2819. PMLR, 2019.

17 Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. *Adv. Neural Inf. Process. Syst.*, 33:16223–16234, 2020.

18 Mohammad Reza Karimi Jaghargh, Ya-Ping Hsieh, and Andreas Krause. A Dynamical System View of Langevin-Based Non-Convex Sampling. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pages 41051–41075, 2023.

19 Mohammad Reza Karimi Jaghargh, Ya-Ping Hsieh, and Andreas Krause. Stochastic Approximation Algorithms for Systems of Interacting Particles. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pages 55826–88847, 2023.

20 Stanislav Kondratyev, Léonard Monsaingeon, and Dmitry Vorotnikov. A new optimal transport distance on the space of finite Radon measures. *Adv. Differ. Equ.*, 21(11/12):1117–1164, 2016.

21 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*, pages 907–915. PMLR, 2019.

22 Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Invent. Math.*, 211(3):969–1117, 2018.

23 Yulong Lu. Two-Scale Gradient Descent Ascent Dynamics Finds Mixed Nash Equilibria of Continuous Games: A Mean-Field Perspective. `https://arxiv.org/abs/2212.08791`, 2022.

24 Chao Ma and Lexing Ying. Provably convergent quasistatic dynamics for mean-field two-player zero-sum games. In *International Conference on Learning Representations*, 2021.

25 Panayotis Mertikopoulos, Ya-Ping Hsieh, and Volkan Cevher. Learning in games from a stochastic approximation viewpoint. `https://arxiv.org/abs/2206.03922`, 2022.

26 Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *International Conference on Learning Representations*, 2018.

27 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*, pages 2703–2717.

Association for Computing Machinery, 2018.

28 Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Program.*, 171(1):115–166, 2018.

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

30 Arkadi Nemirovski. Prox-method with rate of convergence O (1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. Optim.*, 15(1):229–251, 2004.

31 Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numer.*, 8:143–195, 1999.

32 Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bull. Math. Sci.*, 7(1):87–154, 2017.

33 Maurice Sion. On general minimax theorems. *Pac. J. Math.*, 8(1):171–176, 1958.

34 Noah D. Stein, Asuman Ozdaglar, and Pablo A. Parrilo. Separable and low-rank continuous games. *Int. J. Game Theory*, 37(4):475–504, 2008.

35 Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *J. Comput. Appl. Math.*, 60(1-2):237–252, 1995.

36 Guillaume Wang and Lénaïc Chizat. Local convergence of gradient methods for min-max games: partial curvature generically suffices. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pages 60841–60852, 2023.

37 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*, 2020.

Appendix

The appendix is structured as follows.

- In Appendix A we introduce notations used throughout the proofs in the appendix.
- In Appendix B we prove Lemma 1 stating that the CP-PP update is well-defined.
- In Appendix C we formalize and prove the lower growth properties discussed in Section 3.3.
  This section, and in particular the "steepness" result Claim 22, represents the crux of our analysis. In particular much of Appendix D will rely on the same proof ideas, and Appendix E will make crucial use of the results from this section.
- In Appendix D we prove the bounds of Proposition 3 and Proposition 7 relating NI error and our Lyapunov potential.
- In Appendix E we present the complete convergence analysis of CP-PP for the general case, proving Theorem 9.
  It is instructive to see how the aggregated weights and positions naturally appear in the derivations, so that the steps of the proof almost perfectly match the ones for the exact-parametrization case (proof of Theorem 5). The manipulations required to deal with the additional error terms are purely technical however, and they are deferred to the last subsection.
- Appendix F collects elementary auxiliary facts used in some of the above sections.
- Appendix G contains some delayed calculatory proofs for the above sections.
- In Appendix H we prove Proposition 6 stating that (in the exact-parametrization case) CP-MP has the same convergence behavior as CP-PP.
  The proof consists in deriving generically applicable approximate expressions for the Mirror Prox and Proximal Point updates, up to order-3 terms in the step-size. In particular we show and exploit the fact that the error terms are also proportional to the projected gradient norm.
- In Appendix I we show in detail how our main result Theorem 2 follows from combining Proposition 7, Proposition 8 and Theorem 9.

A Notations used in the proofs

In this section we collect notations used throughout the proof. Most of them are natural, except perhaps our use of the $O(\,\cdot\,)$ notation (last paragraph).

Relative entropy

Let $h : \mathbb{R}_+ \to \mathbb{R}$ defined by $h(s) = s \log s - s + 1$. h is convex and its Bregman divergence is

$$d_h(s, s') = \begin{cases} +\infty & \text{if } s' = 0, s > 0 \\ s \log \frac{s}{s'} - s + s' & \text{otherwise.} \end{cases}$$

The Kullback Leibler (KL) divergence between w and $\widehat{w} \in [0,1]^n$ is given by $D(w, \widehat{w}) = \sum_i d_h(w_i, w'_i)$.

Indexing

- We use $I \in [n^*]$ resp. $J \in [m^*]$ to index the "true" particles, i.e., the unique MNE (μ^*, ν^*) is

  $$\mu^* = \sum_{I \in [n^*]} a^*_I \delta_{x^*_I} \quad (a^*_I > 0) \qquad\qquad \nu^* = \sum_{J \in [m^*]} b^*_J \delta_{y^*_J} \quad (b^*_J > 0).$$

  We use $i \in [n]$ resp. $j \in [m]$ to index the particles used by the algorithm.
- Let by convention $a^*_0 = b^*_0 = 0$. In particular we can write that

  $$\forall\, \bar{a} \in \Delta_{[0,n^*]},\ D(a^*, \bar{a}) = \sum_{I \in [0,n^*]} a^*_I \log \frac{a^*_I}{\bar{a}_I}.$$

  Unless specified, summations over I refer to $I \in [n^*]$ (excluding index 0). To lift any ambiguity, $\left\|\bar{a} - a^*\right\|_1$ refers to ℓ_1-norm for $\Delta_{[n^*]}$: $\left\|\bar{a} - a^*\right\|_1 = \sum_{I \in [n^*]} |\bar{a} - a^*|$, even when $\bar{a} \in \Delta_{[0,n^*]}$.
- For $\bar{a} \in \Delta_{[0,n^*]}$, $\Delta\bar{a}_I = \bar{a}_I - a^*_I$, and for $\bar{x} \in \mathcal{X}^{n^*}$, $\Delta\bar{x}_I = \bar{x}_I - x^*_I$.

- Generically denote the joint weight resp. position variables by $w = \binom{a}{b} \in \Delta_n \times \Delta_m$, resp. $p = \binom{x}{y} \in \mathcal{X}^n \times \mathcal{Y}^m$. Summations over w_i will implicitly be over $[n] \sqcup [m]$, that is, $\sum_i f(w_i) = \sum_{i=1}^n f(a_i) + \sum_{j=1}^m f(b_j)$. Likewise for $\overline{w} = \binom{\bar{a}}{\bar{b}} \in \Delta_{[n^*]} \times \Delta_{[m^*]}$ and $\bar{p} = \binom{\bar{x}}{\bar{y}} \in \mathcal{X}^{n^*} \times \mathcal{Y}^{m^*}$, for which summations will implicitly be over $[n^*] \sqcup [m^*]$ (excluding the two indices 0). Finally, $\overline{w}_0 = \bar{a}_0 + \bar{b}_0$.

Local payoff matrices

- Generically let, for $(\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y}) \in (\Delta_n \times \mathcal{X}^n) \times (\Delta_m \times \mathcal{Y}^m)$,

$$\widehat{M}_{ij} = f(\widehat{x}_i, \widehat{y}_j) \qquad (^\wedge M^*)_{iJ} = f(\widehat{x}_i, y^*_J)$$
$$(^*M^\wedge)_{Ij} = f(x^*_I, \widehat{y}_j) \qquad M^*_{IJ} = f(x^*_I, y^*_J)$$

  as well as $\partial_x \widehat{M}_{ij} = \partial_x f(\widehat{x}_i, \widehat{y}_j)$, $\partial_x M^*_{IJ} = \partial_x f(x^*_I, y^*_J)$, etc., and similarly for ∂_y, ∂^2_{xx}, ∂^2_{xy}, and ∂^2_{yy}. For example, we have the Taylor expansion for all i, j, I

$$\widehat{M}_{ij} = (^*M^\wedge)_{Ij} + (\widehat{x}_i - x^*_I)^\top \partial_x (^*M^\wedge)_{Ij} + O(\|\widehat{x}_i - x^*_I\|^2).$$

- Let

$$\forall\, I \in [n^*],\ H_I = \sum_J \partial^2_{xx} M^*_{IJ} b^*_J \quad \text{resp.} \quad \forall\, J \in [m^*],\ H_J = -\sum_I a^*_I \partial^2_{yy} M^*_{IJ}$$

  the local kernels; that is, $H_I = \partial^2_{xx}(F\nu^*)(x^*_I)$ and $H_J = -\partial^2_{yy}((\mu^*)^\top F)(y^*_J)$. Let H_x the $n^* d_x$ by $n^* d_x$ block-diagonal matrix with blocks $(H_I)_{I \in [n^*]}$, and likewise let H_y the block-diagonal matrix with blocks $(H_J)_{J \in [m^*]}$.
- We will use the usual matrix-vector product notation, so that for example $H_I = \partial^2_{xx} M^*_{I\bullet} b^*$. In addition we introduce the following shorthands: For $\bar{a}, \bar{x}, \bar{b}, \bar{y} = (\Delta_{n^*} \times \mathcal{X}^{n^*}) \times (\Delta_{m^*} \times \mathcal{Y}^{m^*})$,
  - Even though $\bar{x}$ is a vector in $(\mathbb{R}^{d_x})^{n^*}$ and $\bar{a}$ is a vector in $\mathbb{R}^{n^*}$, we denote by $\bar{a} \odot \bar{x}$ the vector in $(\mathbb{R}^{d_x})^{n^*}$ such that $(\bar{a} \odot \bar{x})_I = \bar{a}_I \bar{x}_I$.
  - We denote by $\bar{a} H_x$ the block-diagonal matrix such that $[\bar{a} H_x]_{II'} = \mathbf{1}_{I=I'} \bar{a}_I H_I$. Likewise,
  - $\bar{a} \partial_x M^*$ is the matrix such that $[\bar{a} \partial_x M^*]_{IJ} = \bar{a}_I \partial_x M^*_{IJ}$, and
  - $\partial_y M^* \bar{b}$ is the matrix such that $[\partial_y M^* \bar{b}]_{IJ} = \bar{b}_J \partial_y M^*_{IJ}$, and
  - $\bar{a} \partial^2_{xy} M^* \bar{b}$ is the matrix such that $[\bar{a} \partial^2_{xy} M^* \bar{b}]_{IJ} = \bar{a}_I \bar{b}_J \partial^2_{xy} M^*_{IJ}$.

  Finally, we use id to denote the identity matrix, and its size will be clear from context.

Norms and dual norms on $\mathcal{X}$ and $\mathcal{Y}$

We assume that $\mathcal{X}$ and $\mathcal{Y}$ are the d_x- resp. d_y-dimensional tori, that is, $\mathcal{X} = \mathbb{T}^{d_x} = (\mathbb{R}/\mathbb{Z})^{d_x}$ and

$$\forall\, x, x' \in \mathcal{X},\ \|x - x'\|_{\mathcal{X}} = \inf_{k \in \mathbb{Z}^{d_x}} \|x - x' + k\|_2.$$

In particular $\mathcal{X}$ is a compact Riemannian manifold, the tangent space at any point is isometric to $\mathbb{R}^{d_x}$, and the norm of a tangent vector is

$$\forall\, v \in T_x\mathcal{X},\ \|v\|_x = \|v\|_2.$$

The same considerations apply for $\mathcal{Y} = \mathbb{T}^{d_y}$.

To lighten notation, we use $\|\cdot\|$ to denote the norm over $\mathcal{X}$ or $\mathcal{Y}$ or $T_x\mathcal{X}$ or $T_y\mathcal{Y}$; which one is meant in each situation will be clear from context.

The quantities arising from the Assumptions 1–6

- Since $\mathcal{X}$ and $\mathcal{Y}$ are compact Riemannian manifolds, let $R = \mathrm{diameter}(\mathcal{X}) \vee \mathrm{diameter}(\mathcal{Y}) < \infty$.
- Since f has bounded differentials of order up to 3, let ∂_x, ∂_y the partial derivative operators and denote the smoothness constants of f as

$$L_0 = \sup_{\mathcal{X}\times\mathcal{Y}} f - \inf_{\mathcal{X}\times\mathcal{Y}} f, \qquad L_1 = \sup_{\mathcal{X}\times\mathcal{Y}} \|\partial_x f\| \vee \|\partial_y f\|, \qquad L_2 = \sup_{\mathcal{X}\times\mathcal{Y}} \|\partial^2_{xx} f\| \vee \|\partial^2_{xy} f\| \vee \|\partial^2_{yy} f\|$$

  and L_3 such that $\partial^2_{xx} f$, $\partial^2_{xy} f$, $\partial^2_{xx} f$ are L_3-Lipschitz-continuous, and $L = L_0 \vee L_1 \vee L_2 \vee L_3$.
- By definition of MNE, the local kernels $H_I, H_J \succeq 0$, and by non-degeneracy assumption, $H_I, H_J \succ 0$. Denote $\sigma_{\min} = (\min_I \sigma_{\min}(H_I)) \wedge (\min_J \sigma_{\min}(H_J)) > 0$ the least eigenvalue.

Shorthands for partitions of unity

We recall the following notations, already introduced in our construction of the Lyapunov function in Section 3.2.

- For each $I \in [n^*]$, $\varphi_I : \mathcal{X} \to \mathbb{R}$ is the function defined in (13), and $\varphi_0 = 1 - \sum_I \varphi_I$.
- Generically denote, for any $a \in \Delta_n, x \in \mathcal{X}^n$: $\forall\, I \in [0, n^*], \forall\, i \in [n],\ \varphi_{Ii} = \varphi_I(x_i)$, and

$$\forall\, I \in [n^*],\ \bar{a}_I = \sum_i \varphi_{Ii} a_i \qquad \bar{x}_I = \sum_i \frac{\varphi_{Ii} a_i}{\bar{a}_I} x_i \qquad \Sigma_I = \sum_i \frac{\varphi_{Ii} a_i}{\bar{a}_I} (x_i - \bar{x}_I)(x_i - \bar{x}_I)^\top$$

  as well as $\bar{a}_0 = 1 - \sum_I \bar{a}_I$. We refer to $\bar{a}_I$ as the aggregated weights, to $\bar{x}_I$ as the aggregated positions, to Σ_I as the local covariance matrices, and to $\bar{a}_0$ as the stray weight. For example, the iterates at k have aggregated weights $\bar{a}_I^k = \sum_i \varphi_{Ii}^k a_i^k$.
- For any $J \in [0, m^*]$, then $\psi_J : \mathcal{Y} \to \mathbb{R}$ is defined similarly. For any $b \in \Delta_m$ and $y \in \mathcal{Y}^m$, then $\bar{b} \in \Delta_{m^*}$ and $\bar{y} \in \mathcal{Y}^{m^*}$ are defined similarly.

  In addition, we let $\varepsilon = e^{-\lambda^3/3}$ be the value of φ_I and ψ_J at the cut-off.

What we hide in the $O(\,\cdot\,)$'s

- Fix an arbitrary constant $\Gamma_0 \geq 1$ and restrict attention to choices of step-sizes such that $\eta, \sigma \leq 1000$ and $\Gamma_0^{-1} \leq \frac{\sigma}{\eta} \leq \Gamma_0$.
- Let

$$c = \frac{a^*_{\min} \wedge b^*_{\min}}{4} = \frac{(\min_I a_I^*) \wedge (\min_J b_J^*)}{4}.$$

  We will justify in Lemma 31 that, locally, this quantity is a uniform lower bound on the iterates' aggregated weights: $\min_I \bar{a}_I^k, \min_J \bar{b}_J^k, \min_I \bar{a}_I^{k+1}, \min_J \bar{b}_J^{k+1} \geq c$.
- We will use $O(\,\cdot\,)$ and $\lesssim$ and $\asymp$ to hide only constants dependent on $(f, \mathcal{X}, \mathcal{Y})$ (such as $c, R, L, n^*, m^* \dots$) and on Γ_0. That is,
  - $a = O(b)$ means that there exists a constant C only dependent on those quantities, such that $|a| \leq C\,|b|$.
  - $a \lesssim b$ means that there exists a constant C only dependent on those quantities, such that $a \leq Cb$.
  - $a \asymp b$ means that $a \lesssim b$ and $a \gtrsim b$.

  For example, we have $\eta, \sigma = O(1)$ and $\eta \asymp \sigma$.
- Likewise, by "for η sufficiently small" we mean that a property holds for all $\eta \leq \eta_0$ for some η_0 dependent only on those quantities.

B Proof of Lemma 1

Fix k and denote the objective function in (6) as

$$G((a,x),(b,y)) \coloneqq F_{n,m}((a,x),(b,y)) + \frac{1}{\eta} D(a, a^k) + \frac{1}{2\sigma} \sum_{i=1}^n a_i^k \left\|x_i - x_i^k\right\|^2 - \frac{1}{\eta} D(b, b^k) - \frac{1}{2\sigma} \sum_{j=1}^m b_j^k \left\|y_j - y_j^k\right\|^2.$$

Recall that L_1 denotes the Lipschitz constant of f and L_2 its smoothness constant. We prove a quantitative version of Lemma 1.

▶ **Lemma 14.** *G is convex-concave over $(A^k \times \mathcal{X}^n) \times (B^k \times \mathcal{Y}^n)$ where*

$$A^k = \left\{a \in \Delta_n; \forall\, i,\ c_1 a_i^k \leq a_i \leq c_2 a_i^k\right\}$$
$$\text{and} \quad B^k = \left\{b \in \Delta_m; \forall\, j,\ c_1 b_j^k \leq b_j \leq c_2 b_j^k\right\}$$

for any c_1, c_2 such that

$$c_1 \leq 1 \leq c_2 \quad \text{and} \quad \frac{c_2}{c_1} \leq \frac{2}{\eta L_1} \quad \text{and} \quad c_2 \leq \frac{1}{(L_1 + L_2)\sigma}.$$

Such c_1, c_2 exist if and only if $\eta \leq \frac{2}{L_1}$ and $\sigma \leq \frac{1}{L_1 + L_2}$. In particular, if $\eta \leq \frac{1}{L_1}$ and $\sigma \leq \frac{1}{2(L_1+L_2)}$, then we can take $c_1 = 0.75$ and $c_2 = 1.5$. Furthermore, let $((a^*, x^*), (b^*, y^*))$ denote a saddle point of G over $(A^k \times \mathcal{X}^n) \times (B^k \times \mathcal{Y}^n)$. If $\eta \leq \frac{1}{L_1} \frac{c_1}{c_2}$, then we have $D(a^*, a^k) + D(b^*, b^k) \leq O(\eta)$. In particular, for η small enough, a^* resp. b^* belong to the interior of A^k resp. B^k.

Proof. Fix $(\widehat{b},\widehat{y}) \in B^k \times \mathcal{Y}^m$ and let us show that $G(\cdot,(\widehat{b},\widehat{y}))$ is convex over $A^k \times \mathcal{X}^n$. For this, it suffices to show that its Bregman divergence is non-negative, i.e., that

$$\begin{aligned}
&\forall\, (a,x),(\widehat{a},\widehat{x}) \in A^k \times \mathcal{X}^n,\\
&D_{G(\cdot,(\hat{b},\hat{y}))}((a,x),(\widehat{a},\widehat{x})) \coloneqq G((a,x),(\widehat{b},\widehat{y})) - G((\widehat{a},\widehat{x}),(\widehat{b},\widehat{y})) - \left\langle \begin{pmatrix}\nabla_a\\ \nabla_x\end{pmatrix} G((\widehat{a},\widehat{x}),(\widehat{b},\widehat{y})), \begin{pmatrix}a-\widehat{a}\\ x-\widehat{x}\end{pmatrix} \right\rangle \geq 0.
\end{aligned}$$

By straightforward calculations summarized in Lemma 47, this quantity is equal to

$$D_{G(\cdot,(\hat{b},\hat{y}))}((a,x),(\widehat{a},\widehat{x})) = D_{F_{n,m}(\cdot,(\hat{b},\hat{y}))}((a,x),(\widehat{a},\widehat{x})) + \frac{1}{\eta}D(a,\widehat{a}) + \frac{1}{2\sigma}\sum_i a_i^k \|x_i - \widehat{x}_i\|^2.$$

Let us now estimate the term in $D_{F_{n,m}(\cdot,(\hat{b},\hat{y}))}$. Using the shorthands for the local payoff matrices,

$$\begin{aligned}
&D_{F_{n,m}(\cdot,(\hat{b},\hat{y}))}((a,x),(\widehat{a},\widehat{x}))\\
&= F_{n,m}((a,x),(\widehat{b},\widehat{y})) - F_{n,m}((\widehat{a},\widehat{x}),(\widehat{b},\widehat{y})) - \left\langle \nabla_{(a,x)} F_{n,m}((\widehat{a},\widehat{x}),(\widehat{b},\widehat{y})), (a,x)-(\widehat{a},\widehat{x}) \right\rangle\\
&= a^\top (M^\wedge)\widehat{b} - \widehat{a}^\top \widehat{M}\widehat{b} - (a-\widehat{a})^\top \widehat{M}\widehat{b} - \sum_{i,j} \widehat{a}_i (x_i - \widehat{x}_i)\cdot [\partial_x \widehat{M}]_{i,j}\widehat{b}_j\\
&= a^\top \left((M^\wedge) - \widehat{M}\right)\widehat{b} - \widehat{a}^\top \left[\mathrm{Diag}(x-\widehat{x})\partial_x \widehat{M}\right]\widehat{b}\\
&= \underbrace{\widehat{a}^\top \left((M^\wedge) - \widehat{M} - \left[\mathrm{Diag}(x-\widehat{x})\partial_x\widehat{M}\right]\right)\widehat{b}}_{=:S_1} + \underbrace{(a-\widehat{a})^\top \left((M^\wedge) - \widehat{M}\right)\widehat{b}}_{=:S_2}.
\end{aligned}$$

For the first term: For all i,j, by L_2-smoothness of $f(\cdot,\widehat{y}_j)$,

$$\begin{aligned}
\left|\left((M^\wedge) - \widehat{M} - \left[\mathrm{Diag}(x-\widehat{x})\partial_x\widehat{M}\right]\right)_{i,j}\right| &= |f(x_i,\widehat{y}_j) - f(\widehat{x}_i,\widehat{y}_j) - (x_i-\widehat{x}_i)\partial_x f(\widehat{x}_i,\widehat{y}_j)|\\
&\leq \frac{L_2}{2}\|x_i-\widehat{x}_i\|^2\\
\text{so}\quad |S_1| &\leq \frac{L_2}{2}\sum_i \widehat{a}_i \|x_i-\widehat{x}_i\|^2.
\end{aligned}$$

For the second term: For all i,j, by L_1-Lipschitz-continuity of $f(\cdot,\widehat{y}_j)$,

$$\begin{aligned}
\left|\left((M^\wedge) - \widehat{M}\right)_{i,j}\right| &= |f(x_i,\widehat{y}_j) - f(\widehat{x}_i,\widehat{y}_j)| \leq L_1\|x_i-\widehat{x}_i\|\\
\text{so}\quad |S_2| &\leq L_1 \sum_i |a_i-\widehat{a}_i|\,\|x_i-\widehat{x}_i\| = L_1 \sum_i \frac{|a_i-\widehat{a}_i|}{\sqrt{\widehat{a}_i}}\cdot\sqrt{\widehat{a}_i}\|x_i-\widehat{x}_i\|\\
&\leq \frac{L_1}{2}\left(\sum_i \frac{(a_i-\widehat{a}_i)^2}{\widehat{a}_i} + \sum_i \widehat{a}_i\|x_i-\widehat{x}_i\|^2\right).
\end{aligned}$$

Thus we have that for all $(a,x),(\widehat{a},\widehat{x})$,

$$\begin{aligned}
&D_{G(\cdot,(\hat{b},\hat{y}))}((a,x),(\widehat{a},\widehat{x}))\\
&\geq -\frac{L_1}{2}\sum_i \frac{(a_i-\widehat{a}_i)^2}{\widehat{a}_i} - \frac{L_1+L_2}{2}\sum_i \widehat{a}_i\|x_i-\widehat{x}_i\|^2 + \frac{1}{\eta}D(a,\widehat{a}) + \frac{1}{2\sigma}\sum_i a_i^k\|x_i-\widehat{x}_i\|^2\\
&= \underbrace{\frac{1}{\eta}D(a,\widehat{a}) - \frac{L_1}{2}\chi^2(a,\widehat{a})}\ + \underbrace{\sum_i \left(\frac{1}{2\sigma}a_i^k - \frac{L_1+L_2}{2}\widehat{a}_i\right)\|x_i-\widehat{x}_i\|^2}.
\end{aligned}$$

- By Lemma 46, if $\max_i \frac{a_i}{\widehat{a}_i} \leq \frac{2}{\eta L_1}$, then $\chi^2(a,\widehat{a}) \leq \frac{2}{\eta L_1}D(a,\widehat{a})$, and so the first underbrace is non-negative.
- If furthermore $\frac{1}{2\sigma}a_i^k - \frac{L_1+L_2}{2}\widehat{a}_i \geq 0$ for all i, then the second underbrace is non-negative.

Both of these conditions can be ensured by imposing $a,\widehat{a} \in A^k$ with c_1,c_2 as defined in the lemma, since $c_1 a_i^k \leq a_i,\widehat{a}_i \leq c_2 a_i^k \implies \frac{a_i}{\widehat{a}_i} \leq \frac{c_2}{c_1}$. The conditions on η,σ and the possible choices of c_1,c_2 are straightforward

to check. Finally, fix some admissible c_1, c_2 and let us prove the last part of the lemma. Let $((a^*, x^*), (b^*, y^*))$ denote a saddle point of G over $(A^k \times \mathcal{X}^n) \times (B^k \times \mathcal{Y}^n)$, i.e., such that

$$\forall\, ((a,x),(b,y)) \in (A^k \times \mathcal{X}^n) \times (B^k \times \mathcal{Y}^n),\ G((a^*,x^*),(b,y)) \le G((a^*,x^*),(b^*,y^*)) \le G((a,x),(b^*,y^*)).$$

We have shown above that for all $(a,x), (\widehat{a},\widehat{x}) \in A^k \times \mathcal{X}^n$ and $(b,y), (\widehat{b},\widehat{y}) \in B^k \times \mathcal{Y}^n$,

$$\begin{aligned} D_{G(\cdot,(\hat{b},\hat{y}))}((a,x),(\widehat{a},\widehat{x})) &\ge \frac{1}{\eta} D(a,\widehat{a}) - \frac{L_1}{2}\chi^2(a,\widehat{a}) + \sum_i \left(\frac{1}{2\sigma} a_i^k - \frac{L_1+L_2}{2}\widehat{a}_i\right) \left\|x_i - \widehat{x}_i\right\|^2 \\ &\ge \left(\frac{1}{\eta} - \frac{L_1}{2}\frac{c_2}{c_1}\right) D(a,\widehat{a}) \end{aligned}$$

and symmetrically,

$$D_{G((\hat{a},\hat{x}),\cdot)}((b,y),(\widehat{b},\widehat{y})) \le -\left(\frac{1}{\eta} - \frac{L_1}{2}\frac{c_2}{c_1}\right) D(b,\widehat{b}).$$

In particular, for $(a,x) = (a^*,x^*)$, $(\widehat{a},\widehat{x}) = (a^k,x^k)$ and $(b,y) = (b^*,y^*)$, $(\widehat{b},\widehat{y}) = (b^k,y^k)$, the difference of the left-hand sides reads

$$\begin{aligned} &D_{G(\cdot,(b^k,y^k))}((a^*,x^*),(a^k,x^k)) - D_{G((a^k,x^k),\cdot)}((b^*,y^*),(b^k,y^k)) \\ &= G((a^*,x^*),(b^k,y^k)) - G((a^k,x^k),(b^k,y^k)) - \left\langle \begin{pmatrix} \nabla_a \\ \nabla_x \end{pmatrix} G(a^k,x^k,b^k,y^k), \begin{pmatrix} a^*-a^k \\ x^*-x^k \end{pmatrix} \right\rangle \\ &\quad - \left(G((a^k,x^k),(b^*,y^*)) - G((a^k,x^k),(b^k,y^k)) - \left\langle \begin{pmatrix} \nabla_b \\ \nabla_y \end{pmatrix} G(a^k,x^k,b^k,y^k), \begin{pmatrix} b^*-b^k \\ y^*-y^k \end{pmatrix} \right\rangle \right) \\ &= G((a^*,x^*),(b^k,y^k)) - G((a^k,x^k),(b^*,y^*)) - \left\langle \begin{pmatrix} \nabla_a \\ \nabla_x \\ -\nabla_b \\ -\nabla_y \end{pmatrix} G(a^k,x^k,b^k,y^k), \begin{pmatrix} a^*-a^k \\ x^*-x^k \\ b^*-b^k \\ y^*-y^k \end{pmatrix} \right\rangle. \end{aligned}$$

Now $G((a^*,x^*),(b^k,y^k)) - G((a^k,x^k),(b^*,y^*)) \le 0$ by definition of the saddle point, so we have

$$\left(\frac{1}{\eta} - \frac{L_1}{2}\frac{c_2}{c_1}\right)\left(D(a^*,a^k) + D(b^*,b^k)\right) \le -\left\langle \begin{pmatrix} \nabla_a \\ \nabla_x \\ -\nabla_b \\ -\nabla_y \end{pmatrix} G(a^k,x^k,b^k,y^k), \begin{pmatrix} a^*-a^k \\ x^*-x^k \\ b^*-b^k \\ y^*-y^k \end{pmatrix} \right\rangle.$$

Since $\nabla_a D(a,a^k)\big|_{a=a^k} = 0$ and $\nabla_{x_i} \frac{1}{2}\left\|x_i - x_i^k\right\|^2\Big|_{x_i=x_i^k} = 0$, the right-hand side is equal to

$$-\left\langle \begin{pmatrix} \nabla_a \\ \nabla_x \\ -\nabla_b \\ -\nabla_y \end{pmatrix} F_{n,m}(a^k,x^k,b^k,y^k), \begin{pmatrix} a^*-a^k \\ x^*-x^k \\ b^*-b^k \\ y^*-y^k \end{pmatrix} \right\rangle.$$

By proceeding similarly as for our bound of $D_{F_{n,m}(\cdot,(\hat{b},\hat{y}))}$, one can show that it is bounded by $C := 2(L_0 \vee L_1)(1+R)$ (in particular the bound does not depend on n, m). Thus we have as announced

$$D(a^*,a^k) + D(b^*,b^k) \le \frac{C}{\left(\frac{1}{\eta} - \frac{L_1}{2}\frac{c_2}{c_1}\right)} = O(\eta). \qquad \blacktriangleleft$$

C Proof of the lower growth properties

Here we give the proof of the lower growth properties, which are crucial ingredients in our convergence analysis of CP-PP, and which were discussed in Section 3.3.

C.1 Proof of "quadratic growth"

▶ **Lemma 15** ("Quadratic growth", general case)**.** *Consider the Lyapunov function V as in* (12) *with the partitions of unity $(\varphi_I)_I$ and $(\psi_J)_J$ as in* (13)*. Then for any $(\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y}) \in \Delta_n \times \mathcal{X}^n \times \Delta_m \times \mathcal{Y}^m$,*

$$F(\widehat{\mu}, \nu^*) - F(\mu^*, \widehat{\nu}) \geq \max\left\{ \left[\frac{\sigma_{\min}}{2} \wedge \frac{2\xi}{(\lambda\tau)^2}\right] V_{\mathrm{pos}}(\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y}), \quad \left[\frac{\sigma_{\min}}{4}\frac{3(\lambda\tau)^2}{\lambda^3} \wedge \xi\right](\widehat{\bar{a}}_0 + \widehat{\bar{b}}_0) \right\}$$

for some constant $\xi > 0$ only dependent on $(f, \mathcal{X}, \mathcal{Y})$.

Proof. We have

$$F(\widehat{\mu}, \nu^*) - F(\mu^*, \widehat{\nu}) = \langle \widehat{\mu}, F\nu^* - \rho\rangle_{\mathcal{C}(\mathcal{X})} + \left\langle \rho - (\mu^*)^\top F, \widehat{\nu}\right\rangle_{\mathcal{C}(\mathcal{Y})}.$$

Focus on the first term. By the non-degeneracy Assumptions 5 and 6, $F\nu^*$ grows quadratically as $\frac{1}{2}\sigma_{\min}(H_I)$ on the neighborhood of x_I^* for each I, and is lower-bounded by a constant everywhere else. In symbols, there exists a constant $\xi > 0$ such that

$$\forall\, x \in \mathcal{X},\ (F\nu^*)(x) - \rho \geq \left(\frac{\sigma_{\min}}{4} \min_I \left\|x - x_I^*\right\|^2\right) \wedge \xi. \tag{15}$$

This directly implies a lower bound in terms of the position variables. Indeed,

$$\forall\, x \in \mathrm{supp}(\varphi_I) = B_{x_I^*, \lambda\tau},\ (F\nu^*)(x) - \rho \geq \left[\frac{\sigma_{\min}}{4} \wedge \frac{\xi}{(\lambda\tau)^2}\right] \left\|x - x_I^*\right\|^2,$$

and $(F\nu^*)(x) - \rho \geq 0$ on all of $\mathcal{X}$, so by decomposing $\widehat{\mu} = \sum_I \varphi_I \widehat{\mu} + \varphi_0 \widehat{\mu}$,

$$\langle \widehat{\mu}, F\nu^* - \rho\rangle_{\mathcal{C}(\mathcal{X})} \geq \left[\frac{\sigma_{\min}}{4} \wedge \frac{\xi}{(\lambda\tau)^2}\right] \cdot \underbrace{\sum_{I \in [n^*]} \sum_{i=1}^{n} \widehat{\varphi}_{Ii} \widehat{a}_i \left\|\widehat{x}_i - x_I^*\right\|^2}_{= 2V_{\mathrm{pos}}(\hat{a}, \hat{x})}.$$

We can also get a lower bound in terms of the "stray weights" $\widehat{\bar{a}}_0$ and $\widehat{\bar{b}}_0$. Indeed,

$$\forall\, r \leq \lambda,\ 1 - e^{-\frac{r^3}{3}} \leq \frac{r^3}{3} \leq \frac{\lambda}{3} \cdot r^2$$

so that for each I,

$$\forall\, x \in B_{x_I^*, \lambda\tau},\ \varphi_0(x) = 1 - \exp\left(-\frac{\left\|x - x_I^*\right\|^3}{3\tau^3}\right) \leq \frac{\lambda}{3} \cdot \frac{\left\|x - x_I^*\right\|^2}{\tau^2}$$

$$\text{and so} \quad \forall\, x \in \mathcal{X},\ \varphi_0(x) \leq \left(\frac{\lambda}{3\tau^2} \min_I \left\|x - x_I^*\right\|^2\right) \wedge 1.$$

Hence, (15) implies

$$\begin{aligned}
\forall\, x \in \mathcal{X},\ (F\nu^* - \rho)(x) \geq \left(\frac{\sigma_{\min}}{4} \min_I \left\|x - x_I^*\right\|^2\right) \wedge \xi &\geq \left[\frac{\sigma_{\min}}{4}\frac{3\tau^2}{\lambda} \wedge \xi\right] \cdot \left[\left(\frac{\lambda}{3\tau^2} \min_I \left\|x - x_I^*\right\|^2\right) \wedge 1\right] \\
&\geq \left[\frac{\sigma_{\min}}{4}\frac{3\tau^2}{\lambda} \wedge \xi\right] \cdot \varphi_0(x) \\
\text{and so finally} \qquad \langle \widehat{\mu}, F\nu^* - \rho\rangle_{\mathcal{C}(\mathcal{X})} &\geq \left[\frac{\sigma_{\min}}{4}\frac{3\tau^2}{\lambda} \wedge \xi\right] \cdot \underbrace{\int_{\mathcal{X}} \varphi_0 \,\mathrm{d}\widehat{\mu}}_{= \widehat{\bar{a}}_0}.
\end{aligned}$$

◀

For the exact-parametrization case, we can actually reuse the result for the general case, using that our two Lyapunov functions "coincide locally". We summarize that fact in the following easily checked claim, which will also be useful in Sections C.2 and D.4.

▷ Claim 16. Consider the exact-parametrization case. Denote
- $d_* = \min_{i' \neq i''} \|x^*_{i'} - x^*_{i''}\| \wedge \min_{j' \neq j''} \|y^*_{j'} - y^*_{j''}\|$. Fix any $\underline{d}_* \leq d_*$.
- V_e the Lyapunov function designed for the exact-parametrization case (9).
- V_g the Lyapunov function designed for the general case (12) with φ_I being the indicator function of $B_{x^*_I, \underline{d}_*/4}$ (well-defined since those balls do not intersect).
- $V_{\lambda,\tau}$ the Lyapunov function designed for the general case (12) with the parametrized partitions of unity $(\varphi_I)_I$ and $(\psi_J)_J$ from (13).

There exists $r > 0$ such that if $V_e(\widehat{z}) \leq r$, then $\widehat{a}_i \geq \frac{a^*_i}{2}$ and $\|\widehat{x}_i - x^*_i\| \leq \underline{d}_*/4$ for all $i \in [n]$. That is, for all i, $\widehat{x}_i \in B_{x^*_i, \underline{d}_*/4}$. For such $\widehat{z}$, we have $V_e(\widehat{z}) = V_g(\widehat{z})$. Moreover, V_g is the point-wise limit of $V_{\lambda,\tau}$ when $\lambda\tau$ is held constant equal to $\underline{d}_*/4$ and $\tau \to \infty$.

The result for the exact-parametrization case now follows immediately from the above lemma and claim:

▶ **Lemma 17** ("Quadratic growth", exact-parametrization case)**.** *Assume $n = n^*, m = m^*$, and consider the Lyapunov function V defined in (9). There exist constants $r, C > 0$ only dependent on $(f, \mathcal{X}, \mathcal{Y})$ such that, for any $\widehat{z} = (\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y})$ with $V(\widehat{z}) \leq r$, then*

$$F(\widehat{\mu}, \nu^*) - F(\mu^*, \widehat{\nu}) \geq C V_{\mathrm{pos}}(\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y}).$$

C.2 Proof of "error bound"

Our error-bound-type result is contained in the following lemma. It is stated for the general case, i.e., V_1 refers to the Lyapunov function defined in (12). But since we make no assumption on the partitions of unity $(\varphi_I)_I, (\psi_J)_J$, the conclusion of the lemma ("if $V_1(\widehat{z}) \leq r$ then we have this error bound inequality") is also true for the exact-parametrization case with V_1 referring to the Lyapunov function from (9), as one can deduce a posteriori thanks to Claim 16.

▶ **Lemma 18.** *Consider any $\widehat{z} = (\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y}) \in \Delta_n \times \mathcal{X}^n \times \Delta_m \times \mathcal{Y}^m$. For any $(A_I)_{I \in [n^*]}, (B_J)_{J \in [m^*]}$ (and $A_0 = B_0 = 0$) and $(X_I)_{I \in [n^*]}, (Y_J)_{J \in [m^*]}$, let analogously to (14) the "proxy particles"*

$$x_i = \widehat{x}_i + \sum_I \widehat{\varphi}_{Ii}(X_I - \widehat{x}_i) \quad \text{and} \quad a_i = \sum_I A_I \frac{\widehat{\varphi}_{Ii}\widehat{a}_i}{\widehat{\widehat{a}}_I} \tag{16}$$

and similarly for b, y, and $z = (a, x, b, y)$. There exist $r, C > 0$ only dependent on $(f, \mathcal{X}, \mathcal{Y})$ such that if $V_1(\widehat{z}) \leq r$, then

$$\max_{A,X,B,Y} \widetilde{\mathrm{gap}}(z; \widehat{z}) \geq C \sqrt{\sum_I d_h(w^*_I, \widehat{\widehat{w}}_I) + \sum_I \widehat{\widehat{w}}_I \|\Delta \widehat{\widehat{p}}_I\|^2} + O\left(V_1(\widehat{z})\right).$$

Informally, $\max_{A,X,B,Y} \widetilde{\mathrm{gap}}(z; \widehat{z})$ can be interpreted as a lower bound (up to a constant) on

$$\max_{\|\delta z\|_* \leq 1} \left\langle \begin{pmatrix} \nabla_a \\ \nabla_x \\ -\nabla_b \\ -\nabla_y \end{pmatrix} F_{n,m}(\widehat{z}), \begin{pmatrix} \delta a \\ \delta x \\ \delta b \\ \delta y \end{pmatrix} \right\rangle = \|\nabla F_{n,m}(\widehat{z})\|.$$

Note that in the latter expression, δx has n degrees of freedom, whereas in $\max_{A,X,B,Y} \widetilde{\mathrm{gap}}(z; \widehat{z})$, X has only n^* degrees of freedom (and similarly for A, B and Y). Thus, $\max_{A,X,B,Y} \widetilde{\mathrm{gap}}(z; \widehat{z})$ represents a "norm" of $\nabla F_{n,m}(\widehat{z})$ (which justifies why we refer to Lemma 18 as an error bound), for a notion of norm that is adapted to the geometry of the algorithm and of the problem at hand.

The remainder of this subsection is dedicated to proving the above lemma. To lighten notation, we continue to leave the dependence of $z = (a, x, b, y)$ on A, X, B and Y implicit.

Let us start by showing that locally (i.e. for z's with small enough Lyapunov potential), we have a constant lower bound on the aggregated weights $\bar{a}_I, \bar{b}_J$. This fact will be used repeatedly throughout this appendix and the next ones.

▶ **Lemma 19.** *There exists $r > 0$ (only dependent on a^*, b^*) such that if $V_{\mathrm{wei}}(a, x, b, y) \leq r$, then*

$$\left(\min_{I \neq 0} \bar{a}_I\right) \wedge \left(\min_{J \neq 0} \bar{b}_J\right) \geq \frac{a^*_{\min} \wedge b^*_{\min}}{2}.$$

Proof. $h : [0,1] \to \mathbb{R}, s \mapsto s\log s - s + 1$ is 1-strongly convex (just bound h''), so for any I,

$$(a_I^* - \bar{a}_I)^2 \le 2d_h(a_I^*, \bar{a}_I) \le 2V_{\mathrm{wei}}(a,x) \le 2r.$$

Similarly, for any J, $(b_J^* - \bar{b}_J)^2 \le 2r$. So by choosing r small enough, we can ensure that $\min_I \bar{a}_I \ge \frac{\min_I a_I^*}{2}$ and $\min_J \bar{b}_J \ge \frac{\min_J b_J^*}{2}$, and the lemma follows by combining these two inequalities. ◀

C.2.1 Step 1: Reduce to a bilinear functional in the aggregated weights and positions

▷ Claim 20. For any $\widehat{z}$ and A, X, B, Y, $\widetilde{\mathrm{gap}}(z;\widehat{z})$ is approximately given by

$$\begin{aligned}\widetilde{\mathrm{gap}}^{(\mathrm{loc}*)}(z;\widehat{z}) = &-\Delta A^\top M^* \Delta\widehat{\bar{b}} \;+\; \Delta\widehat{\bar{a}}^\top M^* \Delta B \\ &+ \Delta\widehat{\bar{a}}^\top \partial_y M^*(\Delta Y \odot b^*) \;-\; (a^* \odot \Delta X)^\top \partial_x M^* \Delta\widehat{\bar{b}} \\ &- \Delta A^\top \partial_y M^*(\Delta\widehat{\bar{y}} \odot b^*) \;+\; (a^* \odot \Delta\widehat{\bar{x}})^\top \partial_x M^* \Delta B \\ &- \sum_I a_I^* \left\langle \Delta X_I, \Delta\widehat{\bar{x}}_I \right\rangle_{H_I} - \sum_J b_J^* \left\langle \Delta Y_I, \Delta\widehat{\bar{y}}_J \right\rangle_{H_J} \\ &- (a^* \odot \Delta X)^\top \partial_{xy}^2 M^*(\Delta\widehat{\bar{y}} \odot b^*) \;+\; (a^* \odot \Delta\widehat{\bar{x}})^\top \partial_{xy}^2 M^*(\Delta Y \odot b^*)\end{aligned}$$

and more precisely

$$\widetilde{\mathrm{gap}}(z;\widehat{z}) = \widetilde{\mathrm{gap}}^{(\mathrm{loc}*)}(z;\widehat{z}) + O\left((\min_I \widehat{\bar{w}}_I)^{-1} V_1(\widehat{z})\right).$$

The proof consists of simple but tedious calculations, which we defer to Appendix G.1. Essentially we just write out the expression of $\widetilde{\mathrm{gap}}(z;\widehat{z})$ and do Taylor expansions of $f(\widehat{x}_i, \widehat{y}_j)$ around (x_I^*, y_J^*).

We see that $\widetilde{\mathrm{gap}}^{(\mathrm{loc}*)}(z;\widehat{z})$ has a bilinear structure; in matrix form, denoting $[a^* H_x]_{II'} = \mathbb{1}_{I=I'} a_I^* H_I$ and $[a^* \partial_x M^*]_{IJ} = a_I^* \partial_x M_{IJ}^*$, etc.,

$$\widetilde{\mathrm{gap}}^{(\mathrm{loc}*)}(z;\widehat{z}) = -\begin{pmatrix}\Delta A \\ \Delta X \\ \Delta B \\ \Delta Y\end{pmatrix}^\top \begin{bmatrix} 0 & 0 & M^* & \partial_y M^* b^* \\ 0 & a^* H_x & a^* \partial_x M^* & a^* \partial_{xy}^2 M^* b^* \\ -(M^*)^\top & -(a^* \partial_x M^*)^\top & 0 & 0 \\ -(\partial_y M^* b^*)^\top & -(a^* \partial_{xy}^2 M^* b^*)^\top & 0 & b^* H_y \end{bmatrix} \begin{pmatrix}\Delta\widehat{\bar{a}} \\ \Delta\widehat{\bar{x}} \\ \Delta\widehat{\bar{b}} \\ \Delta\widehat{\bar{y}}\end{pmatrix}.$$

Here the first component $\Delta A / \Delta\widehat{\bar{a}}$ is a vector in $\mathbb{R}^{n^*}$, and the second component $\Delta X / \Delta\widehat{\bar{x}}$ is in $\mathcal{X}^{n^*}$. Note that on the right, the first component $\Delta\widehat{\bar{a}}$ does not sum to 0, but to $-\widehat{\bar{a}}_0$. For concision and for clarity of the argument to follow, introduce some notation for the remainder of this section:

- Let $\mathcal{H}$ denote the above block matrix.
- Let

$$\mathcal{Z} = \left\{ \begin{pmatrix}\alpha \\ \bar{x} \\ \beta \\ \bar{y}\end{pmatrix} \in [0,1]^{n^*} \times \mathcal{X}^{n^*} \times [0,1]^{m^*} \times \mathcal{Y}^{m^*} \;;\; \sum_I \alpha_I \le 1, \sum_J \beta_J \le 1 \right\},$$

$$Z = \begin{pmatrix}A \\ X \\ B \\ Y\end{pmatrix} \quad \text{and} \quad \widehat{Z} = \begin{pmatrix}\widehat{\bar{a}} \\ \widehat{\bar{x}} \\ \widehat{\bar{b}} \\ \widehat{\bar{y}}\end{pmatrix} \quad \text{and} \quad Z^* = \begin{pmatrix}a^* \\ x^* \\ b^* \\ y^*\end{pmatrix} \in \mathcal{Z}.$$

- For any $\widetilde{Z} \in \mathcal{Z}$, denote $\Delta\widetilde{Z} = \widetilde{Z} - Z^*$ and let

$$\left\|\Delta\widetilde{Z}\right\|^2 = \left\|\widetilde{\alpha} - a^*\right\|_1^2 + \max_I \left\|\widetilde{\bar{x}}_I - x_I^*\right\|^2 + \left\|\widetilde{\beta} - b^*\right\|_1^2 + \max_J \left\|\widetilde{\bar{y}}_J - y_J^*\right\|^2. \tag{17}$$

Clearly $\|\cdot\|$ defines a norm on $\mathcal{Z} - Z^*$.

So far we showed that, for all $\widehat{z}$ with $V(\widehat{z})$ less than some r so that Lemma 19 applies,

$$\max_{W,P} \widetilde{\mathrm{gap}}(z;\widehat{z}) = \max_{\substack{A\in\Delta_{n^*}, X\in\mathcal{X}^{n^*} \\ B\in\Delta_{m^*}, Y\in\mathcal{Y}^{m^*}}} -\begin{pmatrix}\Delta A \\ \Delta X \\ \Delta B \\ \Delta Y\end{pmatrix}^\top \mathcal{H} \begin{pmatrix}\Delta\widehat{\bar{a}} \\ \Delta\widehat{\bar{x}} \\ \Delta\widehat{\bar{b}} \\ \Delta\widehat{\bar{y}}\end{pmatrix} + O\left(V_1(\widehat{z})\right).$$

To complete the proof of the lemma, it suffices to prove that there exist $r, C > 0$ only dependent on $(f, \mathcal{X}, \mathcal{Y})$ such that if $V_1(\widehat{z}) \le r$, then

$$\max_{\substack{A\in\Delta_{n^*}, X\in\mathcal{X}^{n^*}\\ B\in\Delta_{m^*}, Y\in\mathcal{Y}^{m^*}}} -\Delta Z^\top \mathcal{H} \Delta\widehat{Z} \ge C\sqrt{\sum_I d_h(w_I^*, \widehat{\widehat{w}}_I) + \sum_I \widehat{\widehat{w}}_I \left\|\Delta\widehat{\widehat{p}}_I\right\|^2}.$$

We will do so by working with the aggregated variables $\Delta Z, \Delta\widehat{Z}$. So let us first clarify the relation between the norm on $\mathcal{Z} - Z^*$ and the desired divergence. Namely, we show that $\left\|\Delta\widehat{Z}\right\|^2$ is equivalent to $V_1(\widehat{z}) - \widehat{\widehat{w}}_0$.

▷ Claim 21. Suppose $\min_I \widehat{\widehat{a}}_I, \min_J \widehat{\widehat{b}}_J \ge c$ for some constant $c > 0$. Then

$$2c\sum_I d_h(w_I^*, \widehat{\widehat{w}}_I) + \sum_I \widehat{\widehat{w}}_I \left\|\Delta\widehat{\widehat{p}}_I\right\|^2 \;\le\; \left\|\Delta\widehat{Z}\right\|^2 \;\le\; 2(n^* \wedge m^*)\sum_I d_h(w_I^*, \widehat{\widehat{w}}_I) + \frac{1}{c}\sum_I \widehat{\widehat{w}}_I \left\|\Delta\widehat{\widehat{p}}_I\right\|^2.$$

Proof. For the first inequality, for the weight part: $h : x \mapsto x\log x - x + 1$ is $\frac{1}{c}$-smooth over $[c, 1]$ (since $1 \le h''(x) = \frac{1}{x} \le \frac{1}{c}$), so $\forall\, s, s' \in [c, 1],\ d_h(s, s') \le \frac{1}{2c}\left|s - s'\right|^2$. So

$$\sum_I d_h(a_I^*, \widehat{\widehat{a}}_I) \le \frac{1}{2c}\sum_I (a_I^* - \widehat{\widehat{a}}_I)^2 \le \frac{1}{2c}\left(\sum_I \left|a_I^* - \widehat{\widehat{a}}_I\right|\right)^2.$$

For the position part: just write $\sum_I \widehat{\widehat{a}}_I \left\|\Delta\widehat{\widehat{x}}_I\right\|^2 \le \max_I \left\|\Delta\widehat{\widehat{x}}_I\right\|^2$. For the second inequality, for the weight part: h is 1-strongly concave over $[c, 1]$, so $\forall\, s, s' \in [c, 1],\ d_h(s, s') \ge \frac{1}{2}\left|s - s'\right|^2$. So

$$\sum_I d_h(a_I^*, \widehat{\widehat{a}}_I) \ge \frac{1}{2}\sum_I (a_I^* - \widehat{\widehat{a}}_I)^2 \ge \frac{1}{2n^*}\left(\sum_I \left|a_I^* - \widehat{\widehat{a}}_I\right|\right)^2.$$

For the position part: since $c \le \widehat{\widehat{a}}_I$, just write $\max_I \left\|\Delta\widehat{\widehat{x}}_I\right\|^2 \le \sum_I \frac{\widehat{\widehat{a}}_I}{c}\left\|\Delta\widehat{\widehat{x}}_I\right\|^2$. ◀

C.2.2 Step 2: "Steepness" of the reduced game

The following claim is the crucial point of our analysis. It extends [37, Lem. 14] to the case of continuous instead of finite games, using the Wasserstein–Fisher–Rao instead of the Fisher–Rao geometry. In particular, the proof crucially relies on the assumption that the MNE is unique.

▷ Claim 22. For all $\widehat{Z} \in \mathcal{Z} \setminus \{Z^*\}$, $\max_{\substack{A\in\Delta_{n^*}, X\in\mathcal{X}^{n^*}\\ B\in\Delta_{m^*}, Y\in\mathcal{Y}^{m^*}}} -\Delta Z^\top \mathcal{H} \Delta\widehat{Z} > 0$.

Proof. Let $\widehat{Z} = \begin{pmatrix}\widehat{\alpha} & \widehat{\widehat{x}} & \widehat{\beta} & \widehat{\widehat{y}}\end{pmatrix} \in \mathcal{Z}\setminus\{Z^*\}$. Suppose by contradiction $\max_{\substack{A\in\Delta_{n^*}, X\in\mathcal{X}^{n^*}\\ B\in\Delta_{m^*}, Y\in\mathcal{Y}^{m^*}}} -\Delta Z^\top \mathcal{H} \Delta\widehat{Z} \le 0$. Since the set $\left\{\Delta Z = \begin{pmatrix}\Delta A\\ \Delta X\\ \Delta B\\ \Delta Y\end{pmatrix};\, A \in \Delta_{n^*}, X \in \mathcal{X}^{n^*}, B \in \Delta_{m^*}, Y \in \mathcal{Y}^{m^*}\right\}$ contains a (relative) neighborhood of zero since a^* is in the interior of Δ_{n^*}, clearly the inequality to contradict is equivalent to

$$\forall\, A \in \Delta_{n^*}, X \in \mathcal{X}^{n^*}, B \in \Delta_{m^*}, Y \in \mathcal{Y}^{m^*},\ \Delta Z^\top \mathcal{H} \Delta\widehat{Z} = 0. \tag{18}$$

Denote $\widehat{\alpha}_0 = 1 - \sum_I \widehat{\alpha}_I$ and $\widehat{\beta}_0 = 1 - \sum_J \widehat{\beta}_J$. Pose for some $\lambda \ne 0$ to be specified

$$A = a^* + \lambda(\Delta\widehat{\alpha} + \widehat{\alpha}_0 a^*) \quad \text{and} \quad B = b^* + \lambda(\Delta\widehat{\beta} + \widehat{\beta}_0 b^*).$$

It is straightforward to check that $\sum_I A_I = 1$ and $\sum_J B_J = 1$. Moreover since a^* lies in the interior of Δ_{n^*}, $|\lambda|$ can be chosen small enough so that $A_I \ge 0$, and so $A \in \Delta_{n^*}$ (and likewise $B \in \Delta_{m^*}$). Further pose

$$X = x^* + \lambda\Delta\widehat{\widehat{x}} \quad \text{and} \quad Y = y^* + \lambda\Delta\widehat{\widehat{y}}.$$

Evaluating (18) at this (A, X, B, Y) yields

$$\begin{aligned}
0 &= \lambda \begin{pmatrix} \Delta\widehat{\alpha} + \widehat{\alpha}_0 a^* \\ \Delta\widehat{\widetilde{x}} \\ \Delta\widehat{\beta} + \widehat{\beta}_0 b^* \\ \Delta\widehat{\widetilde{y}} \end{pmatrix}^\top \begin{bmatrix} 0 & 0 & M^* & \partial_y M^* b^* \\ 0 & a^* H_x & a^* \partial_x M^* & a^* \partial^2_{xy} M^* b^* \\ -(M^*)^\top & -(a^* \partial_x M^*)^\top & 0 & 0 \\ -(\partial_y M^* b^*)^\top & -(a^* \partial^2_{xy} M^* b^*)^\top & 0 & b^* H_y \end{bmatrix} \begin{pmatrix} \Delta\widehat{\alpha} \\ \Delta\widehat{\widetilde{x}} \\ \Delta\widehat{\beta} \\ \Delta\widehat{\widetilde{y}} \end{pmatrix} \\
&= \Delta\widehat{Z}^\top \mathcal{H} \Delta\widehat{Z} \;+\; \widehat{\alpha}_0 \cdot \underbrace{a^* M^*}_{=\rho \mathbf{1}^\top} \Delta\widehat{\beta} - \widehat{\beta}_0 \cdot \Delta\widehat{\alpha} \underbrace{M^* b^*}_{=\rho \mathbf{1}} \\
&= a^* \Delta\widehat{\widetilde{x}}^\top H_x \Delta\widehat{\widetilde{x}} + b^* \Delta\widehat{\widetilde{y}}^\top H_y \Delta\widehat{\widetilde{y}} \;+\; \widehat{\alpha}_0 \cdot \rho(-\widehat{\beta}_0) - \widehat{\beta}_0 \cdot \rho(-\widehat{\alpha}_0) \\
&= \sum_I a^*_I \big\| \Delta\widehat{\widetilde{x}}_I \big\|^2_{H_I} + \sum_J b^*_J \big\| \Delta\widehat{\widetilde{y}}_J \big\|^2_{H_J}.
\end{aligned}$$

Since $H_I, H_J \succ 0$, this implies that $\Delta\widehat{\widetilde{x}} = 0$ and $\Delta\widehat{\widetilde{y}} = 0$. So the inequality to contradict reduces to

$$\forall\, A \in \Delta_{n^*}, X \in \mathcal{X}^{n^*}, B \in \Delta_{m^*}, Y \in \mathcal{Y}^{m^*},\; \begin{pmatrix} \Delta A \\ \Delta X \\ \Delta B \\ \Delta Y \end{pmatrix}^\top \begin{bmatrix} 0 & M^* \\ 0 & a^* \partial_x M^* \\ -(M^*)^\top & 0 \\ -(\partial_y M^* b^*)^\top & 0 \end{bmatrix} \begin{pmatrix} \Delta\widehat{\alpha} \\ \Delta\widehat{\beta} \end{pmatrix} = 0. \tag{19}$$

Since $\Delta\widehat{Z} \neq 0$ and $\Delta\widehat{\widetilde{x}}, \Delta\widehat{\widetilde{y}} = 0$, then w.l.o.g. $\Delta\widehat{\alpha} \neq 0$. We want to show that there exists $\theta > 0$ such that, denoting

$$\widehat{\alpha}^\theta = a^* + \theta\,(\Delta\widehat{\alpha} + \widehat{\alpha}_0 a^*) \quad \text{and} \quad \widehat{\mu}^\theta = \sum_I \widehat{\alpha}^\theta_I \delta_{x^*_I} = \mu^* + \theta \sum_I (\Delta\widehat{\alpha}_I + \widehat{\alpha}_0 a^*_I)\, \delta_{x^*_I},$$

$(\widehat{\mu}^\theta, \nu^*)$ is a MNE, which will contradict uniqueness of the MNE (μ^*, ν^*). Equivalently, we want to show that the first variation $(\widehat{\mu}^\theta)^\top F$ is everywhere upper-bounded by ρ:

$$\forall\, y \in \mathcal{Y},\; ((\widehat{\mu}^\theta)^\top F)(y) \leq \rho.$$

First remark that:

1. By the non-degeneracy Assumption 6, there exists $\tau > 0$ such that

$$\forall\, J,\; \forall\, y,\; \big\| y - y^*_J \big\| =: \big\| \delta y \big\| \leq \tau \implies ((\mu^*)^\top F)(y) \leq \rho - \frac{1}{4} \sigma_{\min} \big\| \delta y \big\|^2.$$

2. By (19) evaluated at $A = a^*, X = x^*, Y = y^*$ and $B = e_J$,

$$\begin{aligned}
&\forall\, J,\; [\Delta\widehat{\alpha} M^*]_J - \Delta\widehat{\alpha} M^* b^* = 0 \\
&\qquad [\Delta\widehat{\alpha} M^*]_J = \Delta\widehat{\alpha} M^* b^* = \rho(-\widehat{\alpha}_0).
\end{aligned}$$

3. By (19) evaluated at $A = a^*, X = x^*, B = b^*, Y_{J'} = y^*_{J'}$ for $J' \neq J$ and Y_J arbitrary,

$$\forall\, J,\; [\Delta\widehat{\alpha} \partial_y M^*]_J = 0. \tag{20}$$

Now,

- Fix $J \in [m^*]$. Let us show that there exists $\theta_0 > 0$ such that for all $\theta \leq \theta_0$, we have $\forall\, y \in B_{y^*_J, \tau},\; ((\widehat{\mu}^\theta)^\top F)(y) \leq \rho$. Indeed for all $y \in B_{y^*_J, \tau}$ and letting $\delta y = y - y^*_J$,

$$\begin{aligned}
((\widehat{\mu}^\theta)^\top F)(y) &= ((\mu^*)^\top F)(y) + \theta \sum_I (\Delta\widehat{\alpha}_I + \widehat{\alpha}_0 a^*_I)\, f(x^*_I, y) \\
&\leq \rho - \frac{1}{4} \sigma_{\min} \big\| \delta y \big\|^2 + \theta \sum_I (\Delta\widehat{\alpha}_I + \widehat{\alpha}_0 a^*_I) \Big[M^*_{IJ} + \partial_y M^*_{IJ} \cdot \delta y + O(\big\| \delta y \big\|^2) \Big]
\end{aligned}$$

by point 1. Now

$$\begin{aligned}
\sum_I (\Delta\widehat{\alpha}_I + \widehat{\alpha}_0 a^*_I)\, M^*_{IJ} &= [\Delta\widehat{\alpha} M^*]_J + \widehat{\alpha}_0 [a^* M^*]_J \\
&= \rho(-\widehat{\alpha}_0) + \widehat{\alpha}_0 \rho = 0
\end{aligned}$$

by point 2 and

$$\sum_I (\Delta\widehat{\alpha}_I + \widehat{\alpha}_0 a_I^*)\, \partial_y M_{IJ}^* = 0$$

by point 3. So

$$\begin{aligned}((\widehat{\mu}^\theta)^\top F)(y) &\le \rho - \frac{1}{4}\sigma_{\min}\|\delta y\|^2 + \theta\left[0 + 0\cdot\delta y + O(\|\delta y\|^2)\right] \\ &\le \rho - \left(\frac{1}{4}\sigma_{\min} + O(\theta)\right)\|\delta y\|^2.\end{aligned}$$

So clearly we can choose such a θ_0.

- By the non-degeneracy Assumption 5 there exists $\xi_\tau > 0$ such that for any $y \in \mathcal{Y} \setminus \left(\cup_j B_{y_j^*,\tau}\right)$, $((\mu^*)^\top F)(y) \le \rho - \xi_\tau$. So for all such y,

  $$((\widehat{\mu}^\theta)^\top F)(y) \le \rho - \xi_\tau + O\left(\theta(\|\Delta\widehat{\alpha}\| + \widehat{\alpha}_0)\right).$$

  So we can indeed choose $0 < \theta \le \theta_0$ that satisfies the requirement. ◀

C.2.3 Step 3: Leverage homogeneity

▷ Claim 23. There exist a constant $C > 0$ (dependent only on $(f, \mathcal{X}, \mathcal{Y})$) such that

$$\forall\, \widehat{Z} \in \mathcal{Z},\quad \max_{\substack{A\in\Delta_{n^*}, X\in\mathcal{X}^{n^*}\\ B\in\Delta_{m^*}, Y\in\mathcal{Y}^{m^*}}} -\Delta Z^\top \mathcal{H}\Delta\widehat{Z} \ge C\|\Delta\widehat{Z}\|.$$

Proof. Let for concision $g(\Delta\widehat{Z}) = \max_{\substack{A\in\Delta_{n^*}, X\in\mathcal{X}^{n^*}\\ B\in\Delta_{m^*}, Y\in\mathcal{Y}^{m^*}}} -\Delta Z^\top \mathcal{H}\Delta\widehat{Z}$. Note that g is continuous and positive-homogeneous. Since a^* resp. b^* lies in the interior of its domain, it is not hard to check that there exists $r > 0$ (only dependent on a^*, b^*) such that, for any $\widehat{Z} \in \mathcal{Z}$, then $Z^* + r\frac{\Delta\widehat{Z}}{\|\Delta\widehat{Z}\|} \in \mathcal{Z}$. In other words, for any $\widehat{Z} \in \mathcal{Z}$, we can write $\Delta\widehat{Z} = \frac{\|\Delta\widehat{Z}\|}{r}\Delta\widetilde{Z}$ for some $\widetilde{Z} \in \mathcal{S}_{Z^*,r} \coloneqq \left\{\widetilde{Z} \in \mathcal{Z}; \|\Delta\widetilde{Z}\| = r\right\}$. Now by Claim 22, $g(\Delta\widetilde{Z}) > 0$ for all $\widetilde{Z} \in \mathcal{S}_{Z^*,r}$. Since $\mathcal{S}_{Z^*,r}$ is a compact set and g is continuous, we have

$$\forall\, \widetilde{Z} \in \mathcal{Z} \quad \text{s.t.} \quad \|\Delta\widetilde{Z}\| = r,\ g(\Delta\widetilde{Z}) \ge \inf_{\mathcal{S}_{Z^*,r}} g > 0$$

so by positive-homogeneity, $$\forall\, \widehat{Z} \in \mathcal{Z},\ g(\Delta\widehat{Z}) \ge \left(\inf_{\mathcal{S}_{Z^*,r}} g\right)\frac{\|\Delta\widehat{Z}\|}{r} =: C\|\Delta\widehat{Z}\|.$$ ◀

Lemma 18 follows by using Claim 21 to further lower-bound the result of Claim 23, and by substituting into Claim 20.

C.3 Proof of "local star-convexity-concavity"

▶ **Lemma 24.** *Consider the Lyapunov function V_1 as in* (12) *with the partitions of unity $(\varphi_I)_I$ and $(\psi_J)_J$ as in* (13).[7] *Consider any $\widehat{z} = (\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y}) \in \Delta_n \times \mathcal{X}^n \times \Delta_m \times \mathcal{Y}^m$, and let $z^{(*)} = (a^{(*)}, x^{(*)}, b^{(*)}, y^{(*)})$ "proxy solution particles" similarly as in* (14)*:*

$$x_i^{(*)} \coloneqq \widehat{x}_i + \sum_{I\in[n^*]} \widehat{\varphi}_{Ii}(x_I^* - \widehat{x}_i) \quad \textit{and} \quad a_i^{(*)} \coloneqq \sum_{I\in[n^*]} a_I^* \frac{\widehat{\varphi}_{Ii}\widehat{a}_i}{\widehat{\widehat{a}}_I},$$

and similarly for $a^{(*)}, b^{(*)}$. Suppose $\left(\min_I \widehat{\widehat{a}}_I\right) \wedge \left(\min_J \widehat{\widehat{b}}_J\right) \ge c$ for some $c > 0$. Then, denoting $\widehat{\mu} = \sum_{i=1}^n \widehat{a}_i\delta_{\hat{x}_i}$ and $\widehat{\nu} = \sum_{j=1}^m \widehat{b}_j\delta_{\hat{y}_j}$,

$$\begin{aligned}\widetilde{\mathrm{gap}}(z^{(*)}; \widehat{z}) &= \widehat{a}^\top({}^\wedge M^*)b^* - (a^*)^\top({}^* M^\wedge)\widehat{b} + \frac{1}{2}\sum_I\sum_i \widehat{\varphi}_{Ii}\widehat{a}_i\|\widehat{x}_i - x_I^*\|_{H_I}^2 + \frac{1}{2}\sum_J\sum_j \widehat{\psi}_{Jj}\widehat{b}_j\|\widehat{y}_j - y_J^*\|_{H_J}^2 \\ &\quad + O\left(c^{-1}V_1(\widehat{z})^{3/2}\right) + \boldsymbol{R} \qquad\qquad \textit{where} \quad |\boldsymbol{R}| \le 2L_3c^{-1}\cdot\lambda\tau\cdot V_{\mathrm{pos}}(\widehat{z}) \\ &\ge F(\widehat{\mu}, \nu^*) - F(\mu^*, \widehat{\nu}) + V_{\mathrm{pos}}(\widehat{z})\left(\sigma_{\min} - 2L_3c^{-1}\cdot\lambda\tau\right) + O\left(c^{-1}V_1(\widehat{z})^{3/2}\right).\end{aligned}$$

[7] For this lemma, the precise choice of the partitions of unity $(\varphi_I)_I$ and $(\psi_J)_J$ does not matter, only the fact that $\mathrm{supp}(\varphi_I) \subset B_{x_I^*,\lambda\tau}$ and $\mathrm{supp}(\psi_J) \subset B_{y_J^*,\lambda\tau}$.

Recall from Lemma 19 that we can always ensure $\left(\min_I \widehat{a}_I\right) \wedge \left(\min_J \widehat{\widehat{b}}_J\right) \geq c$ for some constant $c > 0$ by assuming $V_1(\widehat{z}) \leq r$ for some constant $r > 0$. In order to recover the informal statement of Section 3.3.3, note that if in addition $\lambda\tau \leq \frac{c\ \sigma_{\min}}{4L_3}$, then

$$\widetilde{\mathrm{gap}}(z^{(*)}; \widehat{z}) \geq F(\widehat{\mu}, \nu^*) - F(\mu^*, \widehat{\nu}) + \frac{\sigma_{\min}}{2} V_{\mathrm{pos}}(\widehat{z}) + O\left(V_1(\widehat{z})^{3/2}\right).$$

The proof proceeds by Taylor expansions to estimate $\widetilde{\mathrm{gap}}(z^{(*)}; \widehat{z})$. This involves rather tedious calculations, so we defer it to Appendix G.2. In a nutshell, we do Taylor expansions of $f(\widehat{x}_i, \widehat{y}_j)$ around $(x_I^*, \widehat{y}_j)$ or $(\widehat{x}_i, y_J^*)$. In order to make $\widehat{a}^\top ({}^\wedge M^*) b^*$ and $(a^*)^\top ({}^* M^\wedge) \widehat{b}$ appear, at first we only expand the side with $z^{(*)} - \widehat{z}$ (e.g., when estimating the terms $\left\langle \nabla_a F_{n,m}(\widehat{z}), \widehat{a} - a^{(*)} \right\rangle$ and $\left\langle \nabla_x F_{n,m}(\widehat{z}), \widehat{x} - x^{(*)} \right\rangle$, start by expanding only with respect to x and keeping $\widehat{y}, \widehat{b}$ as is). Once the expansion of $\widetilde{\mathrm{gap}}(z^{(*)}, \widehat{z})$ (the equality) is proved, the lower bound follow straightforwardly.

D Proof of the relation between Lyapunov function and NI error

In this section, we show that the Lyapunov function can be used as a proxy for the NI error. We present the proof for the general case (Proposition 7), and describe in Appendix D.4 the necessary adaptations to prove the proposition for the exact-parametrization case (Proposition 3).

As announced in the main text, here is a more quantitative version of Proposition 7.

▶ **Proposition 25.** *Define V_1 as in (12) with the partitions of unity $(\varphi_I)_I$ and $(\psi_J)_J$ as in (13). Suppose that $\lambda\tau \leq \frac{\sigma_{\min}}{2L_3}$. There exist constants C_1, C_2 dependent on $(f, \mathcal{X}, \mathcal{Y})$ and K dependent on λ, τ such that, for any $\widehat{z} = (\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y}) \in \Delta_n \times \mathcal{X}^n \times \Delta_m \times \mathcal{Y}^m$, denoting $\widehat{\mu} = \sum_i \widehat{a}_i \delta_{\widehat{x}_i}$ and $\widehat{\nu} = \sum_j \widehat{b}_j \delta_{\widehat{y}_j}$,*

$$C_1 K \left[\left(\min_I \widehat{a}_I \wedge \min_J \widehat{\widehat{b}}_J\right) V_1(\widehat{z})\right]^{5/4} \leq \mathrm{NI}(\widehat{\mu}, \widehat{\nu}) \leq C_2 \sqrt{V_1(\widehat{z})}.$$

Moreover, there exists $r > 0$ dependent only on $(f, \mathcal{X}, \mathcal{Y})$ such that, if $\mathrm{NI}(\widehat{\mu}, \widehat{\nu}) \leq rK$, then $\min_I \widehat{\widehat{w}}_I \geq c = \frac{w_{\min}^*}{4}$, and so

$$C_1 c^{5/4}\ K\ V_1(\widehat{z})^{5/4} \leq \mathrm{NI}(\widehat{\mu}, \widehat{\nu}).$$

The expression of K can be found in (21). In particular, if λ, τ are chosen as in the proof of convergence for the general case (23), then $K \asymp \sqrt{\sigma}$.

The rest of this section is dedicated to proving the above proposition, with the exception of the last subsection where we deal with the exact-parametrization case.

D.1 Proof of the first inequality: $\mathrm{NI} \lesssim \sqrt{V_1}$

By bilinearity of $F(\mu, \nu)$, for any $\widehat{\mu} = \sum_{i=1}^n \widehat{a}_i \delta_{\widehat{x}_i}$, $\widehat{\nu} = \sum_{j=1}^m \widehat{b}_j \delta_{\widehat{y}_j}$,

$$\begin{aligned}\mathrm{NI}(\widehat{\mu}, \widehat{\nu}) = \max_{\mu,\nu} F(\widehat{\mu}, \nu) - F(\mu, \widehat{\nu}) &= \max_{\mu,\nu} \int_{\mathcal{Y}} \sum_i \widehat{a}_i f(\widehat{x}_i, \cdot)\, \mathrm{d}\nu - \int_{\mathcal{X}} \sum_j \widehat{b}_j f(\cdot, \widehat{y}_j)\, \mathrm{d}\mu \\ &= \max_{x,y} \sum_i \widehat{a}_i f(\widehat{x}_i, y) - \sum_j \widehat{b}_j f(x, \widehat{y}_j).\end{aligned}$$

Now, denoting $\widehat{\varphi}_{Ii} = \varphi_I(\widehat{x}_i)$, for any $y \in \mathcal{Y}$

$$\begin{aligned}\sum_i \widehat{a}_i f(\widehat{x}_i, y) &= \sum_I \sum_i \widehat{\varphi}_{Ii} \widehat{a}_i f(\widehat{x}_i, y) + \sum_i \widehat{\varphi}_{0i} \widehat{a}_i f(\widehat{x}_i, y) \\ &= \sum_I \sum_i \widehat{\varphi}_{Ii} \widehat{a}_i \left(f(x_I^*, y) + O(\left\|\widehat{x}_i - x_I^*\right\|)\right) + O(\widehat{\widehat{a}}_0) \\ &= \sum_I \widehat{\widehat{a}}_I f(x_I^*, y) + O\left(\sum_I \sum_i \widehat{\varphi}_{Ii} \widehat{a}_i \left\|\widehat{x}_i - x_I^*\right\|\right) + O(\widehat{\widehat{a}}_0) \\ &\leq \rho + O\left(\left\|\Delta\widehat{\widehat{a}}\right\|_1\right) + O\left(\sqrt{V_{\mathrm{pos}}(\widehat{a}, \widehat{x})}\right) + O(\widehat{\widehat{a}}_0) = \rho + O\left(\sqrt{V_1(\widehat{a}, \widehat{x})}\right)\end{aligned}$$

where we used Jensen's inequality on $s \mapsto s^2$ and (36) to bound $\sum_{I,i} \widehat{\varphi}_{Ii}\widehat{a}_i\|\widehat{x}_i - x_I^*\|$. Similarly, for any $x \in \mathcal{X}$, $\sum_j \widehat{b}_j f(x, \widehat{y}_j) \geq \rho + O\left(\sqrt{V_1(\widehat{b}, \widehat{y})}\right)$. Hence

$$\mathrm{NI}(\widehat{\mu}, \widehat{\nu}) = \max_{x,y} \sum_i \widehat{a}_i f(\widehat{x}_i, y) - \sum_j \widehat{b}_j f(x, \widehat{y}_j) \lesssim \sqrt{V_1(\widehat{z})}.$$

This shows the first inequality in Proposition 25.

D.2 Proof of the second inequality: $\mathrm{NI} \gtrsim [(\min_I \bar{a}_I \wedge \min_J \bar{b}_J)V_1]^{5/4}$

Lower bound on "gap" to solution (μ^*, ν^*)

Lemma 15 directly implies a lower bound on $\mathrm{NI}(\widehat{\mu}, \widehat{\nu}) = \max_{\mu,\nu} F(\widehat{\mu}, \nu) - F(\mu, \widehat{\nu}) \geq F(\widehat{\mu}, \nu^*) - F(\mu^*, \widehat{\nu})$. For concision, within this section, denote K the constant appearing in that lower bound:

$$\mathrm{NI}(\widehat{\mu}, \widehat{\nu}) \geq K\left(\widehat{\overline{w}}_0 + V_{\mathrm{pos}}(\widehat{z})\right) \quad \text{where} \quad K = \frac{1}{2}\left(\frac{\sigma_{\min}}{2} \wedge \frac{2\xi}{(\lambda\tau)^2} \wedge \frac{\sigma_{\min}}{4}\frac{3(\lambda\tau)^2}{\lambda^3} \wedge \xi\right) \tag{21}$$

and where $\xi > 0$ is a constant only dependent on $(f, \mathcal{X}, \mathcal{Y})$. It remains to lower-bound $\mathrm{NI}(\widehat{\mu}, \widehat{\nu})$ by $\left\|\Delta\widehat{\overline{w}}\right\|_1$ with some exponent.

Lower bound on maximum "gap" to perturbations of the solution

In the remainder of this section, we adopt again the notations of Appendix C.2 (17) for the set $\mathcal{Z}$ and the norm on $\mathcal{Z} - Z^*$.

▷ Claim 26. Suppose $\lambda\tau \leq \frac{\sigma_{\min}}{4L_3}$. For any $A \in \Delta_{n^*}, X \in \mathcal{X}^{n^*}, B \in \Delta_{m^*}, Y \in \mathcal{Y}^{m^*}$,

$$\begin{aligned} &F_{n,m^*}(\widehat{a}, \widehat{x}, B, Y) - F_{n^*,m}(A, X, \widehat{b}, \widehat{y}) \\ &\geq -\begin{pmatrix} \Delta A \\ \Delta X \\ \Delta B \\ \Delta Y \end{pmatrix}^\top \begin{bmatrix} 0 & 0 & M^* & \partial_y M^* b^* \\ 0 & a^*\frac{\sigma_{\min}}{2}\,\mathrm{id} & a^*\partial_x M^* & a^*\partial^2_{xy} M^* b^* \\ -(M^*)^\top & -(a^*\partial_x M^*)^\top & 0 & 0 \\ -(\partial_y M^* b^*)^\top & -(a^*\partial^2_{xy} M^* b^*)^\top & 0 & b^*\frac{\sigma_{\min}}{2}\,\mathrm{id} \end{bmatrix} \begin{pmatrix} \Delta\widehat{a} \\ \Delta\widehat{x} \\ \Delta\widehat{b} \\ \Delta\widehat{y} \end{pmatrix} \\ &\qquad + O\left(\left\|\Delta\widehat{Z}\right\|^3 + \left(\widehat{\overline{w}}_0 + V_{\mathrm{pos}}(\widehat{z})\right)^2 + \|\delta z\|^2\right) \end{aligned}$$

where we denote $\left[a^*\frac{\sigma_{\min}}{2}\,\mathrm{id}\right]_{II'} = \mathbb{1}_{I=I'}a_I^*\frac{\sigma_{\min}}{2}\,\mathrm{id}_{\mathcal{X}}$ for each I, I'.

The proof of this claim follows from simple but tedious calculations, which we defer to Appendix G.3. Essentially, we do Taylor expansions of f around (x_I^*, y_J^*), rearrange the terms so as to get an expression of order 2 in $\Delta\widehat{Z}$ and of order 1 in ΔZ, and check that the remaining terms are non-negative or negligible.

Denote $\mathcal{H}$ the block matrix in the above claim. We reuse the result of Claim 23 (actually it was proved for a slightly different $\mathcal{H}$, with $\frac{\sigma_{\min}}{2}\,\mathrm{id}$ replaced by H_x resp. H_y, but the proof can be very easily adapted): There exist a constant $C > 0$ (dependent only on $(f, \mathcal{X}, \mathcal{Y})$) such that

$$\forall\, \widehat{Z} \in \mathcal{Z}, \quad \max_{\substack{A\in\Delta_{n^*}, X\in\mathcal{X}^{n^*} \\ B\in\Delta_{m^*}, Y\in\mathcal{Y}^{m^*}}} -\Delta Z^\top \mathcal{H}\Delta\widehat{Z} \geq C\left\|\Delta\widehat{Z}\right\|.$$

We further refine that result slightly by also exploiting the positive-homogeneity with respect to ΔZ instead of just $\Delta\widehat{Z}$.

▷ Claim 27. There exist constants $q_1, C > 0$ dependent only on $(f, \mathcal{X}, \mathcal{Y})$ such that for any $q \leq q_1$,

$$\forall\, \widehat{Z} \in \mathcal{Z}, \quad \max_{\|\Delta Z\| \leq q} -\Delta Z^\top \mathcal{H}\Delta\widehat{Z} \geq Cq\left\|\Delta\widehat{Z}\right\|.$$

Proof. Let $\widetilde{\mathcal{Z}} = \Delta_{n^*} \times \mathcal{X}^{n^*} \times \Delta_{m^*} \times \mathcal{Y}^{m^*}$ and $\Delta\widetilde{\mathcal{Z}} = \widetilde{\mathcal{Z}} - Z^*$, and

$$D = \left\{ \begin{pmatrix} \Delta A \\ \Delta X \\ \Delta B \\ \Delta Y \end{pmatrix} \in \mathbb{R}^{n^*} \times \mathcal{X}^{n^*} \times \mathbb{R}^{m^*} \times \mathcal{Y}^{m^*} \;;\; \sum_I \Delta A_I = \sum_J \Delta B_J = 0 \text{ and } \|\delta z\| \le 1 \right\}.$$

Since a^*, x^*, b^*, y^* lie in the (relative) interior of their domains, then there exist $q_1, q_2 > 0$ such that $q_1 D \subset \Delta\widetilde{\mathcal{Z}} \subset q_2 D$. So, using Claim 23 for the second inequality,

$$\max_{v \in q_2 D} -v^\top \mathcal{H} \Delta\widehat{Z} \ge \max_{v \in \Delta\widetilde{\mathcal{Z}}} -v^\top \mathcal{H} \Delta\widehat{Z} = \max_{Z \in \widetilde{\mathcal{Z}}} -\Delta Z^\top \mathcal{H} \Delta\widehat{Z} \ge C\big\|\Delta\widehat{Z}\big\|$$
$$\max_{v \in D} -v^\top \mathcal{H} \Delta\widehat{Z} = \frac{1}{q_2} \max_{v \in q_2 D} -v^\top \mathcal{H} \Delta\widehat{Z} \ge \frac{1}{q_2} C\big\|\Delta\widehat{Z}\big\|$$

and so, for any $q \le q_1$, since $qD \subset q_1 D \subset \Delta\widetilde{\mathcal{Z}}$,

$$\max_{Z \in \widetilde{\mathcal{Z}} \text{ s.t. } \|\Delta Z\| \le q} -\Delta Z^\top \mathcal{H} \Delta\widehat{Z} = \max_{v \in qD \cap \Delta\widetilde{\mathcal{Z}}} -v^\top \mathcal{H} \Delta\widehat{Z} = \max_{v \in qD} -v^\top \mathcal{H} \Delta\widehat{Z}$$
$$= q \max_{v \in D} -v^\top \mathcal{H} \Delta\widehat{Z} \ge \frac{q}{q_2} C\big\|\Delta\widehat{Z}\big\|. \qquad \blacktriangleleft$$

With this we can prove the following lower bound on $\mathrm{NI}(\widehat{\mu}, \widehat{\nu})$:

▷ Claim 28. There exists a constant C dependent only on $(f, \mathcal{X}, \mathcal{Y})$ such that, for any $\beta \ge 1$,

$$\max_{A,X,B,Y} F_{n,m^*}(\widehat{a}, \widehat{x}, B, Y) - F_{n^*,m}(A, X, \widehat{b}, \widehat{y}) \ge C\big\|\Delta\widehat{Z}\big\|^{1+\beta} + O\left(\big\|\Delta\widehat{Z}\big\|^{2\beta} + \big\|\Delta\widehat{Z}\big\|^3 + \left(\widehat{\overline{w}}_0 + V_{\mathrm{pos}}(\widehat{z})\right)^2\right).$$

In particular for $\beta = 3/2$, we have that

$$\mathrm{NI}(\widehat{\mu}, \widehat{\nu}) \ge C\big\|\Delta\widehat{Z}\big\|^{5/2} + O\left(\big\|\Delta\widehat{Z}\big\|^3 + \left(\widehat{\overline{w}}_0 + V_{\mathrm{pos}}(\widehat{z})\right)^2\right),$$

and there exists $r > 0$ such that

$$\forall\ (\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y}) \text{ s.t. } \big\|\Delta\widehat{Z}\big\| \le r,\ \mathrm{NI}(\widehat{\mu}, \widehat{\nu}) \ge \frac{C}{2}\big\|\Delta\widehat{Z}\big\|^{5/2} + O\left(\left(\widehat{\overline{w}}_0 + V_{\mathrm{pos}}(\widehat{z})\right)^2\right).$$

Proof. Note that for any $(\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y})$, $\big\|\Delta\widehat{Z}\big\| \le 4 + 2R =: R'$. To prove the claim, for any fixed $(\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y})$, simply apply Claim 27 with $q = q_1 \frac{\|\Delta\widehat{Z}\|^\beta}{(R')^\beta}$ and substitute into Claim 26. The second part of the claim follows straightforwardly. For the third part, let C_1 denote the constant hidden in the $O(\,\cdot\,)$ and pick r such that

$$C\big\|\Delta\widehat{Z}\big\|^{5/2} - C_1\big\|\Delta\widehat{Z}\big\|^3 = \big\|\Delta\widehat{Z}\big\|^{5/2}\left(C - C_1\big\|\Delta\widehat{Z}\big\|^{1/2}\right) \ge \big\|\Delta\widehat{Z}\big\|^{5/2}\frac{C}{2}$$

for any $\big\|\Delta\widehat{Z}\big\| \le r$. $\blacktriangleleft$

This gives the desired bound on a neighborhood of (μ^*, ν^*). Outside of that neighborhood, we can simply use that NI is non-zero and continuous and has a compact domain.

▷ Claim 29. For any $r > 0$, there exists $C' > 0$ such that $\forall\ (\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y})$ s.t. $\big\|\Delta\widehat{Z}\big\| > r,\ \mathrm{NI}(\widehat{\mu}, \widehat{\nu}) \ge C'$.

Proof. It suffices to show that $\inf_{\mathcal{S}_{n,m,Z^*,r}} \mathrm{NI}(\widehat{\mu}, \widehat{\nu}) > 0$ where $\mathcal{S}_{n,m,Z^*,r} = \left\{(\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y}); \big\|\Delta\widehat{Z}\big\| \ge r\right\}$.

The set $\mathcal{S}_{n,m,Z^*,r}$ is closed as a preimage by the continuous function $(\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y}) \mapsto \big\|\Delta\widehat{Z}\big\|$, and compact as a closed subset of the compact $\Delta_n \times \mathcal{X}^n \times \Delta_m \times \mathcal{Y}^m$. Furthermore, the mapping

$$(\widehat{a}, \widehat{x}, \widehat{b}, \widehat{y}) \mapsto \mathrm{NI}(\widehat{\mu}, \widehat{\nu}) = \max_y \sum_i \widehat{a}_i f(\widehat{x}_i, y) - \min_x \sum_j \widehat{b}_j f(x, \widehat{y}_j)$$

is continuous, since each term is continuous as the maximum (or minimum) of uniformly continuous functions. It just remains to check that NI is non-zero on $\mathcal{S}_{n,m,Z^*,r}$, which follows from uniqueness of the MNE since $\mathrm{NI}(\widehat{\mu}, \widehat{\nu}) = 0 \implies (\widehat{\mu}, \widehat{\nu}) = (\mu^*, \nu^*) \implies \widehat{Z} = Z^*$. $\blacktriangleleft$

Putting together the two above claims, and using that $\|\Delta\widehat{Z}\|$ is anyway bounded by $4+2R=O(1)$, we have shown that there exists a constant $C>0$ dependent only on $(f,\mathcal{X},\mathcal{Y})$ such that

$$\mathrm{NI}(\widehat{\mu},\widehat{\nu}) \geq C\|\Delta\widehat{Z}\|^{5/2} + O\left(\left(\widehat{\overline{w}}_0 + V_{\mathrm{pos}}(\widehat{z})\right)^2\right) \tag{22}$$

for any $(\widehat{a},\widehat{x},\widehat{b},\widehat{y})$.

Conclusion

In the first paragraph we have shown (K is given by Eq. (21))

$$\mathrm{NI}(\widehat{\mu},\widehat{\nu}) \geq F(\widehat{\mu},\nu^*) - F(\mu^*,\widehat{\nu}) \geq K\left(\widehat{\overline{w}}_0 + V_{\mathrm{pos}}(\widehat{z})\right).$$

In the second paragraph we have shown (Eq. (22))

$$\mathrm{NI}(\widehat{\mu},\widehat{\nu}) \geq \max_{A,X,B,Y} F(\widehat{a},\widehat{x},B,Y) - F(A,X,\widehat{b},\widehat{y}) \geq C\|\Delta\widehat{Z}\|^{5/2} - C_1\left(\widehat{\overline{w}}_0 + V_{\mathrm{pos}}(\widehat{z})\right)^2$$

for some $C, C_1 > 0$ only dependent on $(f,\mathcal{X},\mathcal{Y})$; further, using that $\|\Delta\widehat{Z}\|^2 \geq 2(\min_I \widehat{\overline{w}}_I)\sum_I d_h(w_I^*,\widehat{\overline{w}}_I) + \sum_I \widehat{\overline{w}}_I\|\Delta\widehat{\overline{p}}_I\|^2$ (Claim 21), we have

$$\begin{aligned}\mathrm{NI}(\widehat{\mu},\widehat{\nu}) &\geq C\left(2(\min_I \widehat{\overline{w}}_I)\sum_I d_h(w_I^*,\widehat{\overline{w}}_I) + \sum_I \widehat{\overline{w}}_I\|\Delta\widehat{\overline{p}}\|^2\right)^{5/4} - C_1\left(\left(\widehat{\overline{w}}_0 + V_{\mathrm{pos}}(\widehat{z})\right)^2\right)\\ &\geq C\left(2(\min_I \widehat{\overline{w}}_I)\sum_I d_h(w_I^*,\widehat{\overline{w}}_I) + \sum_I \widehat{\overline{w}}_I\|\Delta\widehat{\overline{p}}\|^2\right)^{5/4} - \underbrace{C_1(2+R)}_{=:C_2}\left(\widehat{\overline{w}}_0 + V_{\mathrm{pos}}(\widehat{z})\right).\end{aligned}$$

Taking a convex combination of these two inequalities with ratio $\frac{C_2}{C_2+\frac{K}{2}}$, and since $K=O(1)$ by definition, we get

$$\mathrm{NI}(\widehat{\mu},\widehat{\nu}) \geq C_3 K\left((\min_I \widehat{\overline{w}}_I)\cdot V_1(\widehat{a},\widehat{x},\widehat{b},\widehat{y})\right)^{5/4}$$

for some $C_3>0$ only dependent on $(f,\mathcal{X},\mathcal{Y})$. This concludes the proof of the second inequality of the proposition.

D.3 Proof of the second part of the proposition

We showed in Eq. (22) that there exist constants $C, C_1 > 0$ such that

$$\mathrm{NI}(\widehat{\mu},\widehat{\nu}) \geq C\|\Delta\widehat{Z}\|^{5/2} - C_1\left(\widehat{\overline{w}}_0 + V_{\mathrm{pos}}(\widehat{z})\right)^2 \quad \text{i.e.,} \quad \|\Delta\widehat{Z}\|^{5/2} \leq \frac{1}{C}\mathrm{NI}(\widehat{\mu},\widehat{\nu}) + \frac{C_1}{C}\left(\widehat{\overline{w}}_0 + V_{\mathrm{pos}}(\widehat{z})\right)^2.$$

Now by Eq. (21), we have

$$\mathrm{NI}(\widehat{\mu},\widehat{\nu}) \geq K\left(\widehat{\overline{w}}_0 + V_{\mathrm{pos}}(\widehat{z})\right) \quad \text{i.e.,} \quad \widehat{\overline{w}}_0 + V_{\mathrm{pos}}(\widehat{z}) \leq \frac{1}{K}\mathrm{NI}(\widehat{\mu},\widehat{\nu}).$$

Let $r \leq r_1 \wedge r_2$ where r_1, r_2 are defined by $\frac{K}{C}r_1 = \frac{C_1}{C}r_2^2 = \frac{1}{2}\left(\frac{w^*_{\min}}{10}\right)^{5/2}$. Then, for any $(\widehat{a},\widehat{x},\widehat{b},\widehat{y})$ such that $\mathrm{NI}(\widehat{\mu},\widehat{\nu}) \leq rK$,

$$\frac{1}{C}\mathrm{NI}(\widehat{\mu},\widehat{\nu}) \leq \frac{1}{C}r_1 K = \frac{1}{2}\left(\frac{w^*_{\min}}{10}\right)^{5/2} \qquad \text{and} \qquad \frac{C_1}{C}\left(\widehat{\overline{w}}_0 + V_{\mathrm{pos}}(\widehat{z})\right)^2 \leq \frac{C_1}{C}r_2^2 = \frac{1}{2}\left(\frac{w^*_{\min}}{10}\right)^{5/2},$$

and so $\|\Delta\widehat{Z}\|^{5/2} \leq \left(\frac{w^*_{\min}}{10}\right)^{5/2}$, and in particular $\min_I \widehat{a}_I, \min_J \widehat{b}_J \geq \frac{w^*_{\min}}{4}$. Note that by definition, $K=O(1)$, and so r can be chosen independent of λ,τ and only dependent on $(f,\mathcal{X},\mathcal{Y})$.

This concludes the proof of Proposition 25, and so of Proposition 7.

D.4 Proof for the exact-parametrization case (Proposition 3)

The first part of Proposition 3 (the upper bound on NI) follows from exactly the same computations as in the general case.

The second part of the proposition follows from the same considerations as for the general case; only Eq. (21) and Claim 26 need to be adapted. For the former, simply use Lemma 17 instead of Lemma 15. For the latter, the same bound as in the general case holds; indeed this can be deduced from the general case using Claim 16, by holding $\lambda\tau$ constant and letting $\tau \to \infty$.

E Proof of convergence in the general case

In this section we prove Theorem 9.

Choice of the partitions of unity's parameters

Our specific choice for the parameters λ, τ appearing in the definition of $(\varphi_I)_I$, Eq. (13), will not come into play until later in the proof, but to fix ideas we give their expressions right away. We choose

$$\lambda^3 = \frac{1}{\sqrt{\sigma}} \quad \text{and} \quad \lambda\tau = \min\left\{\sqrt{\frac{1}{2}\frac{\sigma}{\eta}},\ \frac{c\,\sigma_{\min}}{4L_3},\ \frac{\min_{I,I'}\left\|x_I^* - x_{I'}^*\right\| \wedge \min_{J,J'}\left\|y_J^* - y_{J'}^*\right\|}{4}\right\} \asymp 1. \tag{23}$$

Intuitively, in terms of the illustration Figure 1a, the cut-off abscissa $\lambda\tau$ should be thought of as $\Theta(1)$, and the blue curve as being "spiky" with a scale of $\tau = \Theta(\sigma^{1/6})$, when η, σ are small.

Note that $\varepsilon = e^{-\lambda^3/3} = e^{-1/(3\sqrt{\sigma})}$ (the value of φ_I at the cut-off) is exponentially small for $\sigma \asymp \eta$ small, so that $\varepsilon \cdot poly(\eta, \sigma, \lambda, \tau)$ is arbitrary small for η, σ small enough, where $poly(\dots)$ is any polynomial expression of the arguments. Essentially, any term where ε appears can be neglected (will be compensated by other terms), for η and σ small enough.

E.1 Making the Lyapunov function appear in the characterizing inequality

We start from the characterizing inequality (8); for reference it reads

$$\begin{aligned}\forall\, z,\ \eta\,\widetilde{\mathrm{gap}}(z; z^{k+1}) \le \sum_i (a_i - a_i^{k+1}) \log\frac{a_i^{k+1}}{a_i^k} + \sum_j (b_j - b_j^{k+1}) \log\frac{b_j^{k+1}}{b_j^k} \\ + \frac{\eta}{\sigma}\sum_i a_i^k \left\langle x_i^{k+1} - x_i^k, x_i - x_i^{k+1}\right\rangle + \frac{\eta}{\sigma}\sum_j b_j^k \left\langle y_j^{k+1} - y_j^k, y_j - y_j^{k+1}\right\rangle\end{aligned}$$

where

$$\widetilde{\mathrm{gap}}(z; \widehat{z}) = \left\langle \begin{pmatrix} \nabla_a \\ \nabla_x \\ -\nabla_b \\ -\nabla_y \end{pmatrix} F_{n,m}(\widehat{z}), \begin{pmatrix} \widehat{a} - a \\ \widehat{x} - x \\ \widehat{b} - b \\ \widehat{y} - y \end{pmatrix} \right\rangle.$$

As announced in Section 3.2, (14), our first step is to evaluate it at "proxy solution particles" $(a^{(*)}, x^{(*)}, b^{(*)}, y^{(*)}) \in \Delta_n \times \mathcal{X}^n \times \Delta_m \times \mathcal{Y}^m$ given by

$$x_i^{(*)} = x_i^{k+1} + \sum_{I \in [n^*]} \varphi_{Ii}^{k+1}(x_I^* - x_i^{k+1}) \quad \text{and} \quad a_i^{(*)} = \sum_{I \in [n^*]} a_I^* \frac{\varphi_{Ii}^{k+1} a_i^{k+1}}{\bar{a}_I^{k+1}}$$

and similarly for $b^{(*)}, y^{(*)}$.

Position terms

The term $\frac{\eta}{\sigma}\sum_i a_i^k \langle x_i^{k+1} - x_i^k, x_i^{(*)} - x_i^{k+1}\rangle$ on the right-hand side of (8) becomes, by Pythagorean identity and Eq. (36),

$$\begin{aligned}
&\sum_I \sum_i a_i^k\ \varphi_{Ii}^{k+1} \left\langle x_i^{k+1} - x_i^k, x_I^* - x_i^{k+1}\right\rangle = \frac{1}{2}\sum_{I,i} a_i^k\ \varphi_{Ii}^{k+1} \left(\left\|x_I^* - x_i^k\right\|^2 - \left\|x_I^* - x_i^{k+1}\right\|^2 - \left\|x_i^{k+1} - x_i^k\right\|^2\right) \\
&= \left[\frac{1}{2}\sum_{I,i} a_i^k\ \varphi_{Ii}^{k+1}\left\|x_I^* - x_i^k\right\|^2\right] - \left[\frac{1}{2}\sum_{I,i} a_i^k\ \varphi_{Ii}^{k+1}\left\|x_I^* - x_i^{k+1}\right\|^2\right] - \left[\frac{1}{2}\sum_{I,i} a_i^k\ \varphi_{Ii}^{k+1}\left\|x_i^{k+1} - x_i^k\right\|^2\right] \\
&= \left[V_{\mathrm{pos}}(a^k, x^k) + \frac{1}{2}\sum_{I,i} a_i^k \left(\varphi_{Ii}^{k+1} - \varphi_{Ii}^k\right)\left\|x_I^* - x_i^k\right\|^2\right] \\
&\quad - \left[V_{\mathrm{pos}}(a^{k+1}, x^{k+1}) + \frac{1}{2}\sum_{I,i}\left(a_i^k - a_i^{k+1}\right)\ \varphi_{Ii}^{k+1}\left\|x_I^* - x_i^{k+1}\right\|^2\right] - \left[\frac{1}{2}\sum_i a_i^k\ \left(1 - \varphi_{0i}^{k+1}\right)\left\|x_i^{k+1} - x_i^k\right\|^2\right].
\end{aligned}$$

Weight terms

For all $I \in [0, n^*]$, let $u_i^{k+1,I} = \frac{\varphi_{Ii}^{k+1} a_i^{k+1}}{\bar{a}_I^{k+1}}$, so that $u^{k+1,I} \in \Delta_n$ for each I. Then since $a_i^{(*)} = \sum_I a_I^* u_i^{k+1,I}$,

$$\begin{aligned}
\sum_i a_i^{(*)} \log\frac{a_i^{k+1}}{a_i^k} &= \sum_I \sum_i a_I^* u_i^{k+1,I} \log\frac{a_i^{k+1}}{a_i^k} \\
&= \sum_I a_I^* \sum_i u_i^{k+1,I} \log\frac{\bar{a}_I^{k+1}}{\bar{a}_I^k} \cdot \frac{a_i^{k+1}/\bar{a}_I^{k+1}}{a_i^k/\bar{a}_I^k} \\
&= \sum_I a_I^* \left(\sum_i u_i^{k+1,I}\right) \log\frac{\bar{a}_I^{k+1}}{\bar{a}_I^k} + \sum_I a_I^* \sum_i u_i^{k+1,I} \log\frac{a_i^{k+1}/\bar{a}_I^{k+1}}{a_i^k/\bar{a}_I^k} \\
&= D(a^*, \bar{a}^k) - D(a^*, \bar{a}^{k+1}) + \sum_I a_I^* \sum_i u_i^{k+1,I} \log\frac{a_i^{k+1}/\bar{a}_I^{k+1}}{a_i^k/\bar{a}_I^k}
\end{aligned}$$

and, since $a_i^{k+1} = \sum_{I\in[0,n^*]} \bar{a}_I^{k+1} u_i^{k+1,I}$, by the same calculation with a_I^* replaced by $\bar{a}_I^{k+1}$

$$\begin{aligned}
\sum_i a_i^{k+1} \log\frac{a_i^{k+1}}{a_i^k} &= \sum_{I\in[0,n^*]} \sum_i \bar{a}_I^{k+1} u_i^{k+1,I} \log\frac{a_i^{k+1}}{a_i^k} \\
&= D(\bar{a}^{k+1}, \bar{a}^k) + \sum_{I\in[0,n^*]} \bar{a}_I^{k+1} \sum_i u_i^{k+1,I} \log\frac{a_i^{k+1}/\bar{a}_I^{k+1}}{a_i^k/\bar{a}_I^k},
\end{aligned}$$

and the term for $I = 0$ in this last sum is equal to

$$\begin{aligned}
\sum_i \varphi_{0i}^{k+1} a_i^{k+1}\left[\log\frac{a_i^{k+1}}{a_i^k} - \log\frac{\bar{a}_0^{k+1}}{\bar{a}_0^k}\right] &= \sum_i \varphi_{0i}^{k+1}\left[d_h(a_i^{k+1}, a_i^k) + a_i^{k+1} - a_i^k\right] - \left[d_h(\bar{a}_0^{k+1}, \bar{a}_0^k) + \bar{a}_0^{k+1} - \bar{a}_0^k\right] \\
&= \sum_i \varphi_{0i}^{k+1} d_h(a_i^{k+1}, a_i^k) - d_h(\bar{a}_0^{k+1}, \bar{a}_0^k) - \sum_i \left(\varphi_{0i}^{k+1} - \varphi_{0i}^k\right) a_i^k.
\end{aligned}$$

So the weight term on the right-hand side of (8) becomes

$$\begin{aligned}
\sum_i (a_i^{(*)} - a_i^{k+1}) \log\frac{a_i^{k+1}}{a_i^k} &= V_{\mathrm{wei}}(a^k, x^k) - V_{\mathrm{wei}}(a^{k+1}, x^{k+1}) - \sum_I d_h(\bar{a}_I^{k+1}, \bar{a}_I^k) \\
&\quad + \sum_I (a_I^* - \bar{a}_I^{k+1}) \sum_i u_i^{k+1,I} \log\frac{a_i^{k+1}/\bar{a}_I^{k+1}}{a_i^k/\bar{a}_I^k} \\
&\quad - \sum_i \varphi_{0i}^{k+1} d_h(a_i^{k+1}, a_i^k)\ +\ \sum_i (\varphi_{0i}^{k+1} - \varphi_{0i}^k) a_i^k. \qquad (24)
\end{aligned}$$

All in all, evaluating the characterizing inequality (8) at $(a^{(*)}, x^{(*)}, b^{(*)}, y^{(*)}) = z^{(*)}$ yields

$$\begin{aligned}
\eta\,\widetilde{\mathrm{gap}}(z^{(*)}; z^{k+1}) \le\ & V(z^k) - V(z^{k+1}) \qquad (25)\\
& - \sum_I d_h(\overline{w}_I^{k+1}, \overline{w}_I^k) - \frac{\eta}{2\sigma}\sum_i w_i^k(1-\varphi_{0i}^{k+1})\big\|p_i^{k+1} - p_i^k\big\|^2\\
& + \sum_I (w_I^* - \overline{w}_I^{k+1}) \sum_i \frac{\varphi_{Ii}^{k+1} w_i^{k+1}}{\overline{w}_I^{k+1}} \log \frac{w_i^{k+1}/\overline{w}_I^{k+1}}{w_i^k/\overline{w}_I^k} \qquad (\text{err1})\\
& - \sum_i \varphi_{0i}^{k+1} d_h(w_i^{k+1}, w_i^k)\\
& + \sum_i (\varphi_{0i}^{k+1} - \varphi_{0i}^k) w_i^k + \frac{\eta}{2\sigma}\sum_I\sum_i w_i^k\left(\varphi_{Ii}^{k+1} - \varphi_{Ii}^k\right)\big\|p_I^* - p_i^k\big\|^2 \qquad (\text{err2})\\
& + \frac{\eta}{2\sigma}\sum_I\sum_i (w_i^{k+1} - w_i^k)\varphi_{Ii}^{k+1}\big\|p_I^* - p_i^{k+1}\big\|^2 \qquad (\text{err3})
\end{aligned}$$

where we let for concision $w^k = \binom{a^k}{b^k} \in \Delta_n \times \Delta_m$ and $p^k = \binom{x^k}{y^k} \in \mathcal{X}^n \times \mathcal{Y}^m$, and similarly for $\overline{w}^k, w^* \in \Delta_{[0,n^*]} \times \Delta_{[0,m^*]}$ and $\overline{p}^k, p^* \in \mathcal{X}^{n^*} \times \mathcal{Y}^{m^*}$.

The left-hand side looks like "gap from MNE to iterates" so morally non-negative, which we will show and quantify. The second line consists of minus "divergence from $(k+1)$ to k" terms which we will lower-bound. The last four lines consist of error terms which we will control.

E.2 Preliminary lemmas

As this phrase is used many times in the proof, let us emphasize again that by "for η, σ small enough" we always mean that a property holds for all $\eta \le \eta_0, \sigma \le \sigma_0$ for some η_0, σ_0 only dependent on $(f, \mathcal{X}, \mathcal{Y})$ and Γ_0 (the same constants that may be hidden in $O(\,\cdot\,)$'s).

Next, we state some useful elementary facts about the algorithm. The following equations are clear from the update rule (6). Alternatively, they can be seen as a consequence of (8) (holding with equality) applied to $(a^{k+1} + \delta a, x^{k+1}, b^{k+1}, y^{k+1})$ for all $\delta a \in \{\mathbf{1}_n\}^\top$, resp. $(a^{k+1}, x^{k+1} + \delta x^{(i_0)}, b^{k+1}, y^{k+1})$ where $\delta x_i^{(i_0)} = \mathbb{1}_{i=i_0}$.

$$\begin{aligned}
& a_i^{k+1} = a_i^k\ e^{-\eta[(M^{k+1}b^{k+1})_i - \rho]}/Z \quad \text{where} \quad Z = \sum_{i'} a_{i'}^k\ e^{-\eta[(M^{k+1}b^{k+1})_{i'} - \rho]}\\
& \text{and} \quad x_i^{k+1} = x_i^k - \sigma\frac{a_i^{k+1}}{a_i^k}\partial_x M_{i\bullet}^{k+1} b^{k+1}.
\end{aligned} \qquad (26)$$

We formalize the trivial fact that iterates move by no more than the step-size at each time-step.

▶ **Lemma 30.** *For all k,*

$$\begin{aligned}
&\forall\, i,\ \left|a_i^{k+1} - a_i^k\right| \le \min\{a_i^k, a_i^{k+1}\}(e^{2\eta L_0} - 1)\\
&\quad \left\|a^{k+1} - a^k\right\|_1 \le e^{2\eta L_0} - 1 = O(\eta)
\end{aligned}$$

and in particular $a_i^{k+1} = a_i^k(1 + O(\eta))$, and similarly for b, and

$$\forall\, i,\ \left\|x_i^{k+1} - x_i^k\right\| \le \sigma e^{2\eta L_0} L_1 = O(\sigma)$$

and similarly for y.

Proof. From (26), we have

$$\begin{aligned}
&\forall\, i, && a_i^{k+1} = a_i^k \frac{e^{-\eta(M^{k+1}b^{k+1})_i}}{\sum_{i'} a_{i'}^k e^{-\eta(M^{k+1}b^{k+1})_{i'}}}\\
&\text{and so} && e^{-2\eta L_0} = \frac{e^{-\eta L_0}}{\sum_{i'} a_{i'}^k e^{\eta L_0}} \le \frac{a_i^{k+1}}{a_i^k} \le \frac{e^{\eta L_0}}{\sum_{i'} a_{i'}^k e^{-\eta L_0}} = e^{2\eta L_0}
\end{aligned}$$

from which the first result follows. Furthermore, also from (26), $\left\|x_i^{k+1} - x_i^k\right\| = \sigma\frac{a_i^{k+1}}{a_i^k}\left\|\partial_x M_{i\bullet}^{k+1} b^{k+1}\right\| \le \sigma\ e^{2\eta L_0} L_1$. ◀

Recall from Lemma 19 that locally (i.e. if $V(z^k)$ is small enough), we have a constant lower bound on the iterates' aggregated weights $\bar{a}_I^k, \bar{b}_I^k$. That is, there exists $r > 0$ (dependent only on a^*, b^*) such that $V(z) \le r \implies (\min_{I\neq 0} \bar{a}_I) \wedge \left(\min_{J\neq 0} \bar{b}_J\right) \ge \frac{a^*_{\min}\wedge b^*_{\min}}{2}$. Thanks to the considerations above, we can show that locally, the aggregated weights are lower-bounded by a constant for iterates both at k and at $k+1$.

▶ **Lemma 31.** *There exists $r > 0$ (only dependent on a^*, b^*) such that if $V(z^k) \le r$, then for small enough η, σ,*

$$\left(\min_{I\neq 0} \bar{a}_I^k\right) \wedge \left(\min_{I\neq 0} \bar{a}_I^{k+1}\right) \wedge \left(\min_{J\neq 0} \bar{b}_J^k\right) \wedge \left(\min_{J\neq 0} \bar{b}_J^{k+1}\right) \ge \frac{a^*_{\min}\wedge b^*_{\min}}{4} =: c.$$

The proof is conceptually simple but annoyingly technical due to the fact that, to compare $\bar{a}_I^k = \sum_i \varphi_I(x_i^k) a_i^k$ and $\bar{a}_I^{k+1} = \sum_i \varphi_I(x_i^{k+1}) a_i^{k+1}$, we also need to control the variation between $\varphi_I(x_i^k)$ and $\varphi_I(x_i^{k+1})$. Namely we have the following bound, which will also be useful elsewhere in this appendix.

▶ **Lemma 32.** *For any $I \in [n^*]$,*

$$\left|\sum_i (\varphi_{Ii}^{k+1} - \varphi_{Ii}^k) a_i^k\right| \lesssim \varepsilon\left(\bar{a}_0^k + \bar{a}_0^{k+1}\right) + \sqrt{\sigma}\min\left\{1, V(z^{k+1})\right\} \lesssim \varepsilon \bar{a}_0^k + \sqrt{\sigma} V(z^{k+1}).$$

The proof of Lemma 32 is quite technical and is deferred to Appendix E.7. In particular it relies on our specific choice of partitions of unity (13) and of λ, τ (23).

Proof of Lemma 31. Let r the constant from Lemma 19, so that $\eta V(z) \le r \implies \min_I \bar{a}_I, \min_J \bar{b}_J \ge \frac{a^*_{\min}\wedge b^*_{\min}}{2} = 2c$. This immediately ensures that $\min_I \bar{a}_I^k, \min_I \bar{b}_I^k \ge 2c$. For any $I \in [n^*]$,

$$\begin{aligned}
\bar{a}_I^{k+1} - \bar{a}_I^k = \sum_i \varphi_{Ii}^{k+1} a_i^{k+1} - \varphi_{Ii}^k a_i^k &= \sum_i \varphi_{Ii}^{k+1}(a_i^{k+1} - a_i^k) + \sum_i (\varphi_{Ii}^{k+1} - \varphi_{Ii}^k) a_i^k \\
&\ge -O(\eta)\bar{a}_I^{k+1} + \sum_i (\varphi_{Ii}^{k+1} - \varphi_{Ii}^k) a_i^k \\
\bar{a}_I^{k+1} &\ge (1 - O(\eta))\left[\bar{a}_I^k + \sum_i (\varphi_{Ii}^{k+1} - \varphi_{Ii}^k) a_i^k\right].
\end{aligned}$$

Now $\bar{a}_I^k \ge c$, and by Lemma 32, $\left|\sum_i (\varphi_{Ii}^{k+1} - \varphi_{Ii}^k) a_i^k\right| \lesssim \varepsilon + \sqrt{\sigma}$. So for η, σ small enough we indeed have $\bar{a}_I^{k+1} \ge c$, and similarly $\bar{b}_J^{k+1} \ge c$ for all $J \in [m^*]$. ◀

In the remainder of this section except Appendix E.6, we assume that the conditions of this lemma are satisfied. As a first useful consequence, we have that both $V(a^k, x^k)$ and $V(a^{k+1}, x^{k+1})$ are uniformly bounded. Indeed, $V(z^k) \le r$ and $V_{\mathrm{pos}}(a, x) = \sum_{I,i} \varphi_{Ii} a_i \left\|x_i - x_I^*\right\|^2 \le R^2 = O(1)$ for any (a, x) anyway, and

$$V_{\mathrm{wei}}(a^{k+1}, x^{k+1}) = \bar{a}_0^{k+1} + \sum_I d_h(a_I^*, \bar{a}_I^{k+1}) \le 1 + \sum_I \frac{1}{c}(a_I^* - \bar{a}_I^{k+1})^2 \le 1 + \frac{n^*}{c} = O(1) \tag{27}$$

by $\frac{1}{c}$-smoothness of h over $[c, 1]$. A second useful consequence is that, by Claim 21, for $(a, x) = (a^k, x^k)$ or (a^{k+1}, x^{k+1}),

$$\bar{a}_0 + \left\|\bar{a} - a^*\right\|_1^2 \asymp V_{\mathrm{wei}}(a, x) \quad \text{and} \quad \max_I \left\|\bar{x}_I - x_I^*\right\|^2 \lesssim V_{\mathrm{pos}}(a, x).$$

E.3 Lower-bounding $\widetilde{\mathrm{gap}}(z^{(*)}; z^{k+1})$

Let us give a quantitative lower bound on the term on the left-hand side of (25): $\widetilde{\mathrm{gap}}(z^{(*)}; z^{k+1})$. It is here that we make use of the "quadratic growth" and "star-convexity-concavity" properties discussed in Sections 3.3.1 and 3.3.3.

▶ **Lemma 33.** *There exists a constant $\xi > 0$ only dependent on $(f, \mathcal{X}, \mathcal{Y})$ such that*

$$\widetilde{\mathrm{gap}}(z^{(*)}; z^{k+1}) \ge \left[\frac{\sigma_{\min}}{4}\frac{3(\lambda\tau)^3}{\lambda^3} \wedge \xi\right]\left(\bar{a}_0^{k+1} + \bar{b}_0^{k+1}\right) + \frac{\sigma_{\min}}{2} V_{\mathrm{pos}}(z^{k+1}) + O\left(V(z^{k+1})^{3/2}\right).$$

Proof. As a direct consequence of Lemma 24, since $\left(\min_I \bar{a}_I^{k+1}\right) \wedge \left(\min_J \bar{b}_J^{k+1}\right) \geq c$, then denoting $\mu^{k+1} = \sum_{i=1}^n a_i^{k+1} \delta_{x_i^{k+1}}$ and $\nu^{k+1} = \sum_{j=1}^m b_j^{k+1} \delta_{y_j^{k+1}}$,

$$\widetilde{\mathrm{gap}}(z^{(*)}; z^{k+1}) \geq F(\mu^{k+1}, \nu^*) - F(\mu^*, \nu^{k+1}) + V_{\mathrm{pos}}(z^{k+1}) \underbrace{\left(\sigma_{\min} - 2L_3 c^{-1} \cdot \lambda\tau\right)}_{} + O\left(V(z^{k+1})^{3/2}\right).$$

Note that by our choice of $\lambda\tau \leq \frac{c\sigma_{\min}}{4L_3}$, we have $\sigma_{\min} - 2L_3 c^{-1} \cdot \lambda\tau \geq \frac{\sigma_{\min}}{2}$. Furthermore, as a direct consequence of Lemma 15, we have

$$F(\mu^{k+1}, \nu^*) - F(\mu^*, \nu^{k+1}) \geq \left[\frac{\sigma_{\min}}{4} \frac{3(\lambda\tau)^2}{\lambda^3} \wedge \xi\right] \left(\bar{a}_0^{k+1} + \bar{b}_0^{k+1}\right)$$

for some constant $\xi > 0$ only dependent on $(f, \mathcal{X}, \mathcal{Y})$. The lemma follows by combining the two inequalities. ◀

E.4 Controlling the error terms

The proofs for the lemmas in this subsection are all technical, relying on our specific choice of partitions of unity $(\varphi_I)_I$ as well as of parameters λ, τ. We defer the proofs to Appendix E.7.

▶ **Lemma 34** (Bound for (err1))**.** *For η, σ small enough,*

$$\forall\, I, \qquad \sum_i \frac{\varphi_{Ii}^{k+1} w_i^{k+1}}{\overline{w}_I^{k+1}} \log \frac{w_i^{k+1}/\overline{w}_I^{k+1}}{w_i^k/\overline{w}_I^k} = O\left(\varepsilon \overline{w}_0^k + \sqrt{\eta} V(z^{k+1})\right).$$

In particular, $$\sum_I (w_I^* - \overline{w}_I^{k+1}) \sum_i \frac{\varphi_{Ii}^{k+1} w_i^{k+1}}{\overline{w}_I^{k+1}} \log \frac{w_i^{k+1}/\overline{w}_I^{k+1}}{w_i^k/\overline{w}_I^k} \lesssim \sqrt{V(z^{k+1})} \cdot \left[\varepsilon \overline{w}_0^k + \sqrt{\eta} V(z^{k+1})\right].$$

▶ **Lemma 35** (Bound for (err2))**.** *For η, σ small enough,*

$$\sum_i (\varphi_{0i}^{k+1} - \varphi_{0i}^k) w_i^k + \frac{\eta}{2\sigma} \sum_{I,i} w_i^k \left(\varphi_{Ii}^{k+1} - \varphi_{Ii}^k\right) \left\|p_I^* - p_i^k\right\|^2 \lesssim \varepsilon \overline{w}_0^{k+1} + \eta^{3/2} V_{\mathrm{pos}}(z^{k+1}) + \sqrt{\eta} V(z^{k+1})^{3/2}.$$

▶ **Lemma 36** (Bound for (err3))**.** *For η, σ small enough,*

$$\frac{\eta}{2\sigma} \sum_I \sum_i (w_i^{k+1} - w_i^k) \varphi_{Ii}^{k+1} \left\|p_I^* - p_i^{k+1}\right\|^2 \lesssim \eta V(z^{k+1})^{3/2}.$$

$V(z^k)$ does not grow too fast

At this point, we have all we need to show the following.

▶ **Lemma 37.** *For η, σ small enough, $V(z^{k+1}) \leq 2V(z^k)$.*

Proof. Starting from (25), upper-bound the second line by 0, lower-bound the left-hand side using Lemma 33 and bound the error terms using Lemmas 34, 35 and 36. Simplify the obtained inequality using that $\lambda\tau \asymp 1$, $\lambda^3 = \frac{1}{\sqrt{\sigma}}$, $\eta \asymp \sigma$, and $V(z^{k+1}) = O(1)$ as noted above (27). Rearranging, we get

$$V(z^{k+1}) - V(z^k) \leq O(\varepsilon) V(z^k) + O\left(\sqrt{\eta} \cdot V(z^{k+1})\right)$$
$$V(z^{k+1}) \leq (1 + O(\sqrt{\eta}))(1 + O(\varepsilon)) V(z^k)$$

and so $V(z^{k+1}) \leq 2V(z^k)$ for η, σ small enough, as announced. ◀

E.5 Lower-bounding the "divergence from $(k+1)$ to k" terms

In this subsection, we lower-bound the quantity

$$D(k+1, k) \coloneqq \sum_I d_h(\overline{w}_I^{k+1}, \overline{w}_I^k) + \frac{\eta}{2\sigma} \sum_i w_i^k (1 - \varphi_{0i}^{k+1}) \left\|p_i^{k+1} - p_i^k\right\|^2$$

appearing with a negative sign on the right-hand side of (25). The bound relies on the "error bound"-type property discussed in Section 3.3.2.

▶ **Lemma 38.** *There exist constants $r, C > 0$ only dependent on $(f, \mathcal{X}, \mathcal{Y})$ and Γ_0 such that, for $V(z^k) \le r$ and small enough η, σ, then $D(k+1,k)$ is lower-bounded by*

$$D(k+1,k) \ge C\eta^2 \left(\sum_I d_h(w_I^*, \overline{w}_I^{k+1}) + \sum_I \overline{w}_I^{k+1} \left\| \Delta \overline{p}_I^{k+1} \right\|^2 \right) + O\left(\left(\varepsilon \overline{w}_0^k\right)^2 + \eta V(z^{k+1})^2 \right).$$

The remainder of this subsection is dedicated to proving this lemma.

For any $(A_I)_{I\in[n^*]}, (B_J)_{J\in[m^*]}$ (and $A_0 = B_0 = 0$) and $(X_I)_{I\in[n^*]}, (Y_J)_{J\in[m^*]}$, define "proxy particles" as in (16):

$$x_i = x_i^{k+1} + \sum_I \varphi_{Ii}^{k+1}(X_I - x_i^{k+1}) \quad \text{and} \quad a_i = \sum_I A_I \frac{\varphi_{Ii}^{k+1} a_i^{k+1}}{\overline{a}_I^{k+1}},$$

similarly for b and y and let $z = (a, x, b, y)$, leaving the dependence on A, X, B and Y implicit to lighten notation. For concision, let as usual $w = \binom{a}{b}$, $p = \binom{x}{y}$, $W = \binom{A}{B}$, and $P = \binom{X}{Y}$. We proceed by upper- and lower-bounding the gradient-norm-like quantity $\max_{A,X,B,Y} \widetilde{\mathrm{gap}}(z; z^{k+1})$.

Upper bound on the gradient-norm-like quantity

▷ Claim 39. There exists $C > 0$ only dependent on $(f, \mathcal{X}, \mathcal{Y})$ and Γ_0 such that

$$\max_{A,X,B,Y} \widetilde{\mathrm{gap}}(z; z^{k+1}) \le \frac{C}{\eta} \sqrt{D(k+1,k)} \; + \; \frac{1}{\eta} O\left(\varepsilon \overline{w}_0^k + \sqrt{\eta} V(z^{k+1})\right).$$

Proof. Evaluate (8) at the proxy particles $z = (a, x, b, y)$:

$$\forall\, A, X, B, Y,\ \eta\, \widetilde{\mathrm{gap}}(z; z^{k+1}) \le \sum_i (w_i - w_i^{k+1}) \log \frac{w_i^{k+1}}{w_i^k} + \frac{\eta}{\sigma} \sum_i w_i^k \left\langle p_i^{k+1} - p_i^k, p_i - p_i^{k+1} \right\rangle.$$

On the right-hand side, we get for the position terms

$$\begin{aligned}
\sum_i w_i^k \left\langle p_i^{k+1} - p_i^k, p_i - p_i^{k+1} \right\rangle &= \sum_I \sum_i w_i^k \varphi_{Ii}^{k+1} \Big\langle p^{k+1} - p_i^k, \underbrace{P_I - p_i^{k+1}}_{\|\cdot\| \le R} \Big\rangle \\
&\le R \sum_I \sum_i w_i^k \varphi_{Ii}^{k+1} \left\| p_i^{k+1} - p_i^k \right\| = R \sum_i w_i^k (1 - \varphi_{0i}^{k+1}) \left\| p_i^{k+1} - p_i^k \right\| \\
&\le R \underbrace{\sqrt{\sum_i w_i^k (1 - \varphi_{0i}^{k+1})}}_{\le \sqrt{2}} \sqrt{\sum_i w_i^k (1 - \varphi_{0i}^{k+1}) \left\| p_i^{k+1} - p_i^k \right\|^2}.
\end{aligned}$$

In the last inequality, we used Cauchy–Schwarz inequality. For the weight terms, by the same calculation as for (24) with w^* replaced by W,

$$\begin{aligned}
\sum_i (w_i - w_i^{k+1}) \log \frac{w_i^{k+1}}{w_i^k} &= \sum_I W_I \log \frac{\overline{w}_I^{k+1}}{\overline{w}_I^k} - \sum_I d_h(\overline{w}_I^{k+1}, \overline{w}_I^k) \\
&\quad + \sum_I (W_I - \overline{w}_I^{k+1}) \sum_i \frac{\varphi_{Ii}^{k+1} a_i^{k+1}}{\overline{a}_I^{k+1}} \log \frac{w_i^{k+1} / \overline{w}_I^{k+1}}{w_i^k / \overline{w}_I^k} \\
&\quad - \sum_i \varphi_{0i}^{k+1} d_h(w_i^{k+1}, w_i^k) + \sum_i (\varphi_{0i}^{k+1} - \varphi_{0i}^k) w_i^k.
\end{aligned}$$

The second and fourth terms are non-positive, the third term is bounded by $O\big(\varepsilon \overline{w}_0^k + \sqrt{\eta} V(z^{k+1})\big)$ by Lemma 34, and so is the last term by Lemma 32. Further upper-bound the first term, using that $\overline{w}_I^k \ge c$ (Lemma 31), by

$$\begin{aligned}
\sum_I W_I \log \frac{\overline{w}_I^{k+1}}{\overline{w}_I^k} &\le \sum_I \frac{W_I}{\overline{w}_I^k} \left(\overline{w}_I^{k+1} - \overline{w}_I^k\right) \le \frac{1}{c} \sum_I \left| \overline{w}_I^{k+1} - \overline{w}_I^k \right| \\
&\le \frac{\sqrt{n^* + m^*}}{c} \sqrt{\sum_I \left| \overline{w}_I^{k+1} - \overline{w}_I^k \right|^2} \le \frac{\sqrt{2(n^* + m^*)}}{c} \sqrt{\sum_I d_h(\overline{w}_I^{k+1}, \overline{w}_I^k)}
\end{aligned}$$

since h is 1-strongly convex over $[0,1]$. Thus,

$$\sum_i (w_i - w_i^{k+1}) \log \frac{w_i^{k+1}}{w_i^k} \leq \frac{\sqrt{2(n^*+m^*)}}{c} \sqrt{\sum_I d_h(\overline{w}_I^{k+1}, \overline{w}_I^k)} + O\left(\varepsilon \overline{w}_0^k + \sqrt{\eta} V(z^{k+1})\right).$$

By putting the two parts together and using that $\sqrt{A} + \sqrt{B} \leq \sqrt{2}\sqrt{A+B}$, we obtain

$$\eta \max_{A,X,B,Y} \widetilde{\mathrm{gap}}(z; z^{k+1}) \leq 2 \left(\frac{\sqrt{n^*+m^*}}{c} \vee R\sqrt{\frac{\eta}{\sigma}} \right) \sqrt{D(k+1,k)} \; + \; O\left(\varepsilon \overline{w}_0^k + \sqrt{\eta} V(z^{k+1})\right)$$

and the claim follows directly. ◀

Lower bound on the gradient-norm-like quantity

As a direct application of Lemma 18, there exist $r', C' > 0$ only dependent on $(f, \mathcal{X}, \mathcal{Y})$ and Γ_0 such that if $V(z^{k+1}) \leq r'$, then

$$\max_{A,X,B,Y} \widetilde{\mathrm{gap}}(z; z^{k+1}) \geq C' \sqrt{\sum_I d_h(w_I^*, \overline{w}_I^{k+1}) + \sum_I \overline{w}_I^{k+1} \left\| \Delta \bar{p}_I^{k+1} \right\|^2} + O\left(V(z^{k+1})\right).$$

Note that thanks to Lemma 37, we can indeed assume $V(z^{k+1}) \leq r'$ by choosing r small enough in the statement of Lemma 38.

Putting the two bounds together

All in all, we showed that (assuming $V(z^k) \leq r$ for some r small enough)

$$\frac{C}{\eta}\sqrt{D(k+1,k)} + \frac{1}{\eta} O\big(\varepsilon \overline{w}_0^k + \sqrt{\eta} V(z^{k+1})\big) \geq C' \sqrt{\sum_I d_h(w_I^*, \overline{w}_I^{k+1}) + \sum_I \overline{w}_I^{k+1} \left\| \Delta \bar{p}_I^{k+1} \right\|^2} + O\big(V(z^{k+1})\big)$$

for some C, C' only dependent on $(f, \mathcal{X}, \mathcal{Y})$ and Γ_0. Rearranging and taking squares,

$$D(k+1,k) \geq \left(\frac{C'}{C}\right)^2 \eta^2 \left(\sum_I d_h(w_I^*, \overline{w}_I^{k+1}) + \sum_I \overline{w}_I^{k+1} \left\| \Delta \bar{p}_I^{k+1} \right\|^2 \right) + O\big((\varepsilon \overline{w}_0^k)^2 + \eta V(z^{k+1})^2\big).$$

This concludes the proof of Lemma 38.

E.6 Proof conclusion

It just remains to put everything together by substituting the terms by their lower bound in (25). Our choice of λ, τ and our assumption that $\sigma \asymp \eta$ simplify things considerably. The only subtlety is that some of the terms in the upper bound of Lemma 34 and Lemma 35 need to be compensated by the lower bound of Lemma 33, which can be done by assuming η, σ small enough.

In the remainder of this subsection, $C_1, C_2, \cdots > 0$ will denote constants dependent only on $(f, \mathcal{X}, \mathcal{Y})$ and Γ_0 (the same things as what we hide in $O(\,\cdot\,)$).

Putting all the bounds together

Assume that $V(z^k) \leq r$ and that r, η, σ are small enough so that all of the lemmas apply. Just substitute the terms in (25) by their bounds:

$$V(z^{k+1}) - V(z^k) \leq -\eta\, \widetilde{\mathrm{gap}}(z^{(*)}; z^{k+1}) - D(k+1,k) + (\mathrm{err1}) + (\mathrm{err2}) + (\mathrm{err3}).$$

By Lemma 33, there exist C_1, C_2 such that

$$\eta\, \widetilde{\mathrm{gap}}(z^{(*)}; z^{k+1}) \geq \eta \cdot C_1 \sqrt{\sigma} \overline{w}_0^{k+1} + \eta \cdot C_2 V_{\mathrm{pos}}(z^{k+1}) + O\left(\eta V(z^{k+1})^{3/2}\right).$$

By Lemma 38, there exists C_3 such that

$$D(k+1,k) \geq C_3 \eta^2 \left(\sum_I d_h(w_I^*, \overline{w}_I^{k+1}) + \sum_I \overline{w}_I^{k+1} \left\| \Delta \bar{p}_I^{k+1} \right\|^2 \right) + O\left(\left(\varepsilon \overline{w}_0^k\right)^2 + \eta V(z^{k+1})^2 \right).$$

By Lemma 34,

$$\begin{aligned}(\text{err1}) &\lesssim \sqrt{\eta} V(z^{k+1})^{3/2} + \varepsilon \overline{w}_0^k \cdot \sqrt{V(z^{k+1})} \\ &\lesssim \sqrt{\eta} V(z^{k+1})^{3/2} + \varepsilon \left(\overline{w}_0^k\right)^2 + \varepsilon V(z^{k+1}).\end{aligned}$$

By Lemma 35,

$$(\text{err2}) \lesssim \varepsilon \overline{w}_0^{k+1} + \eta^{3/2} V_{\text{pos}}(z^{k+1}) + \sqrt{\eta} V(z^{k+1})^{3/2}.$$

By Lemma 36,

$$(\text{err3}) \lesssim \eta V(z^{k+1})^{3/2}.$$

All in all, since $\overline{w}_0^k = O(V(z^k))$ by Eq. (37), we get

$$\begin{aligned}V(z^{k+1}) - V(z^k) \leq &- \left(C_1 \eta^{3/2} - O(\varepsilon)\right) \overline{w}_0^{k+1} - (C_2 - O(\sqrt{\eta}))\, \eta V_{\text{pos}}(z^{k+1}) \\ &- C_3 \eta^2 \left(\sum_I d_h(w_I^*, \overline{w}_I^{k+1}) + \sum_I \overline{w}_I^{k+1} \left\| \Delta \overline{p}_I^{k+1} \right\|^2 \right) \\ &+ O\left(\sqrt{\eta} V(z^{k+1})^{3/2}\right) + O\left(\varepsilon V(z^{k+1})\right) + O\left(\varepsilon V(z^k)^2\right).\end{aligned}$$

To be explicit, this means that there exists a constant $M > 0$ dependent only on $(f, \mathcal{X}, \mathcal{Y})$ and Γ_0 such that

$$\begin{aligned}V(z^{k+1}) - V(z^k) \leq &- \left(C_1 \eta^{3/2} - M\varepsilon\right) \overline{w}_0^{k+1} - (C_2 - M\sqrt{\eta})\, \eta V_{\text{pos}}(z^{k+1}) \\ &- C_3 \eta^2 \left(\sum_I d_h(w_I^*, \overline{w}_I^{k+1}) + \sum_I \overline{w}_I^{k+1} \left\| \Delta \overline{p}_I^{k+1} \right\|^2 \right) \\ &+ M\left(\sqrt{\eta} V(z^{k+1})^{3/2} + \varepsilon V(z^{k+1}) + \varepsilon V(z^k)^2\right).\end{aligned}$$

Since $\varepsilon = e^{-\lambda^3/3} = e^{-\frac{1}{3\sqrt{\sigma}}}$ and $\sigma \asymp \eta$, then for small enough η, σ, we have $C_1 \eta^{3/2} - M\varepsilon \geq \frac{C_1}{2} \eta^{3/2}$. Furthermore, for small enough η, we have $C_2 - M\sqrt{\eta} \geq \frac{C_2}{2}$. Thus, there exists $C_4 > 0$ such that

$$V(z^{k+1}) - V(z^k) \leq -C_4 \eta^2 V(z^{k+1}) + M\left(\sqrt{\eta} V(z^{k+1})^{3/2} + \varepsilon V(z^{k+1}) + \varepsilon V(z^k)^2\right).$$

Moreover for small enough η, σ, we have $C_4 \eta^2 - M\varepsilon \geq \frac{C_4}{2} \eta^2$. Then,

$$\begin{aligned}&V(z^{k+1}) - V(z^k) \leq -(C_4/2)\eta^2 V(z^{k+1}) + M\left(\sqrt{\eta} V(z^{k+1})^{3/2} + \varepsilon V(z^k)^2\right) \\ &V(z^{k+1}) \left[1 + (C_4/2)\eta^2 - M\sqrt{\eta}\sqrt{V(z^{k+1})}\right] \leq V(z^k) \left[1 + M\varepsilon V(z^k)\right].\end{aligned}$$

Sufficient decrease of the Lyapunov function

For a fixed $r_0 > 0$ to be chosen (small enough so that all of the lemmas apply), assume that $V(z^k) \leq r_0$. By Lemma 37, we can assume η, σ small enough such that $V(z^{k+1}) \leq 2r_0$. We then have that

$$\begin{aligned}V(z^{k+1}) \left[1 + (C_4/2)\eta^2 - M\sqrt{2\eta r_0}\right] &\leq V(z^k) \left[1 + M\varepsilon r_0\right] \\ \frac{V(z^{k+1})}{V(z^k)} &\leq \frac{1 + M\varepsilon r_0}{1 + (C_4/2)\eta^2 - M\sqrt{2\eta r_0}} =: 1 - \kappa.\end{aligned} \tag{28}$$

Clearly r_0 can be chosen (dependent on η) such that the right-hand side is strictly less than 1, i.e., $\kappa > 0$. By induction, if $V(z^0) \leq r_0$, then for all k, $V(z^{k+1}) \leq r_0$ and

$$\begin{aligned}\forall\, k,\ \frac{V(z^{k+1})}{V(z^k)} &\leq 1 - \kappa \\ V(z^k) &\leq V(z^0)(1-\kappa)^k.\end{aligned}$$

This concludes the proof of Theorem 9.

▶ Remark 40. More precisely, for the right-hand side of (28) to be less than 1, r_0 needs to be chosen less than η^3 times a constant (dependent on $(f, \mathcal{X}, \mathcal{Y})$ and Γ_0). The rate κ can be seen to be of order η^2, for any admissible choice of r_0.

E.7 Delayed technical proofs

In some of the proofs of this subsection we use the expressions and a priori bounds for V_{pos} and V_{wei} from Appendix F.1 without explicit mention.

E.7.1 Auxiliary claims

As a consequence of the fact that $\left\|x_i^{k+1} - x_i^k\right\| = O(\sigma)$ (Lemma 30), we can meaningfully classify the particles according to which $\mathrm{supp}(\varphi_I)$ they belong to, both at k and at $k+1$. For a fixed k, denote

$$\forall\, I \in [n^*],\ \mathcal{N}(I) = \left\{i;\quad x_i^{k+1} \in B_{x_I^*,\lambda\tau} \text{ or } x_i^k \in B_{x_I^*,\lambda\tau}\right\} \quad \text{and} \quad \mathcal{N}(0) = [n] \setminus (\cup_I \mathcal{N}(I))\,.$$

Since $x_i^{k+1} - x_i^k = O(\sigma)$, for σ chosen small enough compared to $\min_{I\neq I'}\left\|x_I^* - x_{I'}^*\right\|$ we have that $x_i^k \in B_{x_I^*,\lambda\tau} \implies \forall\, I' \neq I,\ x_i^k, x_i^{k+1} \notin B_{x_{I'}^*,\lambda\tau}$, and so the $\mathcal{N}(I)$ are pair-wise disjoint. In other words, $\bigsqcup_{I\in[0,n^*]} \mathcal{N}(I)$ then forms a partition of $[n]$; and similarly for the $(y_j)_{j\in[m]}$. *In the remainder of this section, we assume σ small enough so that this is the case.*

Let

$$\forall\, x \in \mathcal{X},\ \widetilde{\varphi}_I(x) = \exp\left(-\frac{\left\|x - x_I^*\right\|^3}{3\tau^3}\right)$$

so that $\varphi_I(x)$ coincides with $\widetilde{\varphi}_I(x)$ if and only if $\left\|x - x_I^*\right\| \leq \lambda\tau$.

▷ Claim 41. For any $I \in [n^*]$,

$$\sum_i (\varphi_{Ii}^{k+1} - \varphi_{Ii}^k) a_i^k = \sum_{i\in\mathcal{N}(I)} (\widetilde{\varphi}_{Ii}^{k+1} - \widetilde{\varphi}_{Ii}^k) a_i^k + O\left(\varepsilon(\bar{a}_0^{k+1} + \bar{a}_0^k)\right).$$

Proof. Since $\varphi_I(x)$ coincides with $\widetilde{\varphi}_I(x)$ if and only if $\left\|x - x_I^*\right\| \leq \lambda\tau$,
- if $i \notin \mathcal{N}(I)$, i.e., if both $x_i^{k+1}, x_i^k \notin B_{x_I^*,\lambda\tau}$, then $\varphi_{Ii}^{k+1} - \varphi_{Ii}^k = 0$;
- if $i \in \mathcal{N}(I)$,

$$\begin{aligned}\left|(\varphi_{Ii}^{k+1} - \varphi_{Ii}^k) - (\widetilde{\varphi}_{Ii}^{k+1} - \widetilde{\varphi}_{Ii}^k)\right| &\leq \left|\varphi_{Ii}^{k+1} - \widetilde{\varphi}_{Ii}^{k+1}\right| + \left|\varphi_{Ii}^k - \widetilde{\varphi}_{Ii}^k\right| \\ &\leq \varepsilon \cdot \mathbb{1}\left[x_i^{k+1} \notin B_{x_I^*,\lambda\tau} \wedge x_i^k \in B_{x_I^*,\lambda\tau}\right] + \varepsilon \cdot \mathbb{1}\left[x_i^{k+1} \in B_{x_I^*,\lambda\tau} \wedge x_i^k \notin B_{x_I^*,\lambda\tau}\right].\end{aligned}$$

Further note that, by definition,

$$\sum_I \sum_{i\in\mathcal{N}(I)} \varepsilon \mathbb{1}\left[x_i^{k+1} \notin B_{x_I^*,\lambda\tau} \ \wedge\ x_i^k \in B_{x_I^*,\lambda\tau}\right] a_i^k \leq \varepsilon \sum_i \varphi_0(x_i^{k+1}) a_i^k \leq \varepsilon(1 + O(\eta))\bar{a}_0^{k+1}$$

$$\text{and} \quad \sum_I \sum_{i\in\mathcal{N}(I)} \varepsilon \mathbb{1}\left[x_i^{k+1} \in B_{x_I^*,\lambda\tau} \ \wedge\ x_i^k \notin B_{x_I^*,\lambda\tau}\right] a_i^k \leq \varepsilon \sum_i \varphi_0(x_i^k) a_i^k = \varepsilon \bar{a}_0^k.$$

Thus

$$\left|\sum_i (\varphi_{Ii}^{k+1} - \varphi_{Ii}^k) a_i^k - \sum_{i\in\mathcal{N}(I)} (\widetilde{\varphi}_{Ii}^{k+1} - \widetilde{\varphi}_{Ii}^k) a_i^k\right| \leq \varepsilon\left((1 + O(\eta))\bar{a}_0^{k+1} + \bar{a}_0^k\right)$$

and hence the announced estimate. ◀

The following claim follows from a Taylor expansion of $x \mapsto \left\|x - x_I^*\right\|^3$.

▷ Claim 42. For any i, I, we have

$$\begin{aligned}\left\|x_i^{k+1} - x_I^*\right\|^3 - \left\|x_i^k - x_I^*\right\|^3 &= 3\left\langle x_i^{k+1} - x_i^k, x_i^{k+1} - x_I^*\right\rangle \left\|x_i^{k+1} - x_I^*\right\| \\ &\quad + O\left(\left\|x_i^{k+1} - x_i^k\right\|^2 \left\|x_i^{k+1} - x_I^*\right\| + \left\|x_i^{k+1} - x_i^k\right\|^3\right).\end{aligned}$$

Proof. The first and second derivatives of $\|\cdot - x_I^*\|^3$ are given, up to translation, by

$$(\nabla\|\cdot\|^3)(x) = 3\|x\|x \qquad \text{and} \qquad 0 \preceq (\nabla^2\|\cdot\|^3)(x) = 3\|x\|\,\mathrm{id} + 3\|x\|\frac{xx^\top}{\|x\|^2} \preceq 6\|x\|\,\mathrm{id}\,.$$

By Taylor expansion of $\|\cdot - x_I^*\|^3$ centered at x_i^{k+1} with remainder in Lagrange form, there exists $\theta \in [0,1]$ such that

$$\left\|x_i^{k+1} - x_I^*\right\|^3 - \left\|x_i^k - x_I^*\right\|^3 = 3\left\langle x_i^{k+1} - x_i^k, x_i^{k+1} - x_I^*\right\rangle \left\|x_i^{k+1} - x_I^*\right\| - \boldsymbol{R}$$

where

$$\begin{aligned}
\boldsymbol{R} &= \frac{1}{2}(x_i^{k+1} - x_i^k)^\top \left[\left(\nabla^2 \|\cdot - x_I^*\|^3\right)\left(\theta x_i^{k+1} + (1-\theta)x_i^k\right)\right](x_i^{k+1} - x_i^k) \\
2\,|\boldsymbol{R}| &\le 6\left\|x_i^{k+1} - x_i^k\right\|^2 \left\|\theta(x_i^{k+1} - x_I^*) + (1-\theta)(x_i^k - x_I^*)\right\| \\
&\le 6\left\|x_i^{k+1} - x_i^k\right\|^2 \Big(\left\|x_i^{k+1} - x_I^*\right\| + \underbrace{\left\|x_i^k - x_I^*\right\|}_{\le \|x_i^{k+1} - x_I^*\| + \|x_i^{k+1} - x_i^k\|}\Big) \\
&\le 12\left\|x_i^{k+1} - x_i^k\right\|^2 \left\|x_i^{k+1} - x_I^*\right\| + 6\left\|x_i^{k+1} - x_i^k\right\|^3.
\end{aligned}$$

◀

We will repeatedly use the following Taylor expansions of the local payoff matrices.

▷ Claim 43. For any i, I,

$$(M^{k+1}b^{k+1})_i = \rho + \frac{1}{2}\left\|x_i^{k+1} - x_I^*\right\|_{H_I}^2 + O\left(\left\|x_i^{k+1} - x_I^*\right\|^3\right) + O\left(\sqrt{V(b^{k+1}, y^{k+1})}\right) \tag{29}$$

and more precisely if $\left\|x_i^{k+1} - x_I^*\right\| \le \frac{\sigma_{\min}}{2L_3}$, then

$$(M^{k+1}b^{k+1})_i \ge \rho + \frac{1}{4}\sigma_{\min}\left\|x_i^{k+1} - x_I^*\right\|^2 + O\left(\sqrt{V(b^{k+1}, y^{k+1})}\right). \tag{30}$$

Furthermore,

$$\partial_x M_{i\bullet}^{k+1} b^{k+1} = (x_i^{k+1} - x_I^*)^\top H_I + O\left(\left\|x_i^{k+1} - x_I^*\right\|^2\right) + O\left(\min\left\{1, \sqrt{V(b^{k+1}, y^{k+1})}\right\}\right) \tag{31}$$

and more precisely if $\left\|x_i^{k+1} - x_I^*\right\| \le \frac{\sigma_{\min}}{2L_3}$, then

$$(x_i^{k+1} - x_I^*)^\top \partial_x M_{i\bullet}^{k+1} b^{k+1} \ge \frac{\sigma_{\min}}{2}\left\|x_i^{k+1} - x_I^*\right\|^2 + O\left(\left\|x_i^{k+1} - x_I^*\right\|\sqrt{V(b^{k+1}, y^{k+1})}\right). \tag{32}$$

Note that our choice of λ, τ implies $\lambda\tau \le \frac{\sigma_{\min}}{4L_3}$, and that for all $i \in \mathcal{N}(I)$, $\left\|x_i^{k+1} - x_I^*\right\| \le \lambda\tau + O(\sigma)$. *In the remainder of this section, we assume σ small enough so that $\left\|x_i^{k+1} - x_I^*\right\| \le \frac{\sigma_{\min}}{2L_3}$ holds for all $i \in \mathcal{N}(I)$* and similarly for the y_j, y_J^*.

Proof. To lighten notation in the calculations, denote $\widehat{x} = x^{k+1}$, $\widehat{y} = y^{k+1}$ and $\widehat{b} = b^{k+1}$. By Taylor expansion, for all i, I,

$$\begin{aligned}
(M^{k+1}b^{k+1})_i = (\widehat{M}\widehat{b})_i &= \sum_J \sum_j \widehat{\psi}_{Jj}\left[(^\wedge M^*)_{iJ} + \widehat{M}_{ij} - (^\wedge M^*)_{iJ}\right]\widehat{b}_j + \sum_j \widehat{\psi}_{0j}\widehat{M}_{ij}\widehat{b}_j \\
&= \sum_J (^\wedge M^*)_{iJ}\widehat{\bar{b}}_J + \sum_J \sum_j \widehat{\psi}_{Jj} O(\left\|\widehat{y}_j - y_J^*\right\|)\widehat{b}_j + O(\widehat{\bar{b}}_0) \\
&= \sum_J \left[M_{IJ}^* + (\widehat{x}_i - x_I^*)^\top \partial_x M_{IJ}^* + \frac{1}{2}((\widehat{x}_i - x_I^*)^2)^\top \partial_{xx}^2 M_{IJ}^* + O\left(\left\|\widehat{x}_i - x_I^*\right\|^3\right)\right]\widehat{\bar{b}}_J + O\left(\sqrt{V_{\mathrm{pos}}(\widehat{b}, \widehat{y})}\right) + O(\widehat{\bar{b}}_0) \\
&= \rho + \sum_J M_{IJ}^* \Delta\widehat{\bar{b}}_J + \sum_J (\widehat{x}_i - x_I^*)^\top \partial_x M_{IJ}^* \Delta\widehat{\bar{b}}_J + \frac{1}{2}\left\|\widehat{x}_i - x_I^*\right\|_{H_I}^2 + O\left(\left\|\widehat{x}_i - x_I^*\right\|^3\right) + O\left(\sqrt{V(\widehat{b}, \widehat{y})}\right) \\
&= \rho + \frac{1}{2}\left\|\widehat{x}_i - x_I^*\right\|_{H_I}^2 + O\left(\left\|\widehat{x}_i - x_I^*\right\|^3\right) + O\left(\sqrt{V(\widehat{b}, \widehat{y})}\right).
\end{aligned}$$

More precisely,

$$\begin{aligned}
(\widehat{M}\widehat{b})_i &\ge \rho + \frac{1}{2}\left\|\widehat{x}_i - x_I^*\right\|_{H_I}^2 - \frac{L_3}{2}\left\|\widehat{x}_i - x_I^*\right\|^3 + O\left(\sqrt{V(\widehat{b}, \widehat{y})}\right) \\
&\ge \rho + \frac{1}{4}\sigma_{\min}\left\|\widehat{x}_i - x_I^*\right\|^2 + O\left(\sqrt{V(\widehat{b}, \widehat{y})}\right) \quad \text{if} \quad \left\|\widehat{x}_i - x_I^*\right\| \le \frac{\sigma_{\min}}{2L_3}.
\end{aligned}$$

Also by Taylor expansion, for all i, I,

$$\begin{aligned}
\partial_x M^{k+1}_{i\bullet} b^{k+1} &= \partial_x \widehat{M}_{i\bullet}\widehat{b} = \sum_J \sum_j \widehat{\psi}_{Jj}\left[\partial_x(\hat{}M^*)_{iJ} + \partial_x \widehat{M}_{ij} - \partial_x(\hat{}M^*)_{iJ}\right]\widehat{b}_j + \sum_j \widehat{\psi}_{0j}\partial_x\widehat{M}_{ij}\widehat{b}_j \\
&= \sum_J \partial_x(\hat{}M^*)_{iJ}\widehat{\bar{b}}_J + \sum_J\sum_j \widehat{\psi}_{Jj} O(\|\widehat{y}_j - y^*_J\|)\widehat{b}_j + O(\widehat{\bar{b}}_0) \\
&= \sum_J \left[\partial_x M^*_{IJ} + (\widehat{x}_i - x^*_I)^\top \partial^2_{xx} M^*_{IJ} + O(\|\widehat{x}_i - x^*_I\|^2)\right]\widehat{\bar{b}}_J + O\left(\sqrt{V_{\mathrm{pos}}(\widehat{b},\widehat{y})}\right) + O(\widehat{\bar{b}}_0) \\
&= \partial_x M^*_{I\bullet}\Delta\widehat{\bar{b}} + \sum_J (\widehat{x}_i - x^*_I)^\top \partial^2_{xx} M^*_{IJ}\widehat{\bar{b}}_J + O(\|\widehat{x}_i - x^*_I\|^2) + O\left(\min\left\{1, \sqrt{V(\widehat{b},\widehat{y})}\right\}\right) \\
&= (\widehat{x}_i - x^*_I)^\top H_I + O(\|\widehat{x}_i - x^*_I\|^2) + O\left(\min\left\{1, \sqrt{V(\widehat{b},\widehat{y})}\right\}\right).
\end{aligned}$$

On lines 4 and 5, the fact that the last error term is $O(1)$ can be checked by noting that $\widehat{\bar{b}}_0, \|\Delta\widehat{\bar{b}}\| \le 2$ and $V_{\mathrm{pos}}(\widehat{b},\widehat{y}) = \sum_J \sum_j \widehat{\psi}_{Jj}\widehat{b}_j\|\widehat{y}_j - y^*_J\|^2 \le R^2$. More precisely, for any δx,

$$\left\langle \delta x, \partial_x \widehat{M}_{i\bullet}\widehat{b}\right\rangle \ge \langle \delta x, H_I(\widehat{x}_i - x^*_I)\rangle - L_3\|\delta x\|\|\widehat{x}_i - x^*_I\|^2 + O\left(\|\delta x\|\sqrt{V(\widehat{b},\widehat{y})}\right).$$

So if $\|\widehat{x}_i - x^*_I\| \le \frac{\sigma_{\min}}{2L_3}$, then

$$\begin{aligned}
(\widehat{x}_i - x^*_I)^\top \partial_x \widehat{M}_{i\bullet}\widehat{b} &\ge \sigma_{\min}\|\widehat{x}_i - x^*_I\|^2 - L_3\|\widehat{x}_i - x^*_I\|^3 + O\left(\|\widehat{x}_i - x^*_I\|\sqrt{V(\widehat{b},\widehat{y})}\right) \\
&\ge \frac{\sigma_{\min}}{2}\|\widehat{x}_i - x^*_I\|^2 + O\left(\|\widehat{x}_i - x^*_I\|\sqrt{V(\widehat{b},\widehat{y})}\right).
\end{aligned}$$

◀

E.7.2 Proof of Lemma 32

Note that the proof does not make use of the fact that $V(z^{k+1}) = O(1)$ (inequality (27)), so as to avoid circular reasoning since we showed that fact as a consequence of Lemma 32.

Proof. Fix $I \in [n^*]$. We showed in Claim 41 that

$$\sum_i (\varphi^{k+1}_{Ii} - \varphi^k_{Ii})a^k_i = \sum_{i\in\mathcal{N}(I)} (\widetilde{\varphi}^{k+1}_{Ii} - \widetilde{\varphi}^k_{Ii})a^k_i + O\left(\varepsilon(\bar{a}^{k+1}_0 + \bar{a}^k_0)\right)$$

and it remains to upper- and lower-bound the first term.

First note that

$$\widetilde{\varphi}^k_{Ii} - \widetilde{\varphi}^{k+1}_{Ii} = \widetilde{\varphi}^k_{Ii}\cdot\left(1 - \frac{\widetilde{\varphi}^{k+1}_{Ii}}{\widetilde{\varphi}^k_{Ii}}\right) \le \widetilde{\varphi}^k_{Ii}\cdot\left(\log\widetilde{\varphi}^k_{Ii} - \log\widetilde{\varphi}^{k+1}_{Ii}\right) = \widetilde{\varphi}^k_{Ii}\cdot\frac{1}{3\tau^3}\left(\|x^{k+1}_i - x^*_I\|^3 - \|x^k_i - x^*_I\|^3\right) \quad (33)$$

and that

$$\widetilde{\varphi}^k_{Ii} - \widetilde{\varphi}^{k+1}_{Ii} = \widetilde{\varphi}^{k+1}_{Ii}\cdot\left(\frac{\widetilde{\varphi}^k_{Ii}}{\widetilde{\varphi}^{k+1}_{Ii}} - 1\right) \ge \widetilde{\varphi}^{k+1}_{Ii}\cdot\left(\log\widetilde{\varphi}^k_{Ii} - \log\widetilde{\varphi}^{k+1}_{Ii}\right) = \widetilde{\varphi}^{k+1}_{Ii}\cdot\frac{1}{3\tau^3}\left(\|x^{k+1}_i - x^*_I\|^3 - \|x^k_i - x^*_I\|^3\right).$$

Furthermore, by Claim 42,

$$\begin{aligned}
\left|\|x^{k+1}_i - x^*_I\|^3 - \|x^k_i - x^*_I\|^3\right| &\le 3\|x^{k+1}_i - x^k_i\|\|x^{k+1}_i - x^*_I\|^2 + O\left(\|x^{k+1}_i - x^k_i\|^2\|x^{k+1}_i - x^*_I\| + \|x^{k+1}_i - x^k_i\|^3\right) \\
&\lesssim \|x^{k+1}_i - x^k_i\|\|x^{k+1}_i - x^*_I\| + \|x^{k+1}_i - x^k_i\|^2
\end{aligned}$$

and by the update equation (26) and the expansion (31), since $a^k_i = (1 + O(\eta))a^{k+1}_i \asymp a^{k+1}_i$ by Lemma 30,

$$\|x^{k+1}_i - x^k_i\| = \sigma\frac{a^{k+1}_i}{a^k_i}\|\partial_x M^{k+1}_{i\bullet}b^{k+1}\| \lesssim \sigma\left(\|x^{k+1}_i - x^*_I\| \;+\; 1 \wedge \sqrt{V(b^{k+1}, y^{k+1})}\right),$$

and so

$$\begin{aligned}
&\left|\left\|x_i^{k+1}-x_I^*\right\|^3-\left\|x_i^k-x_I^*\right\|^3\right| \\
&\lesssim \sigma\left\|x_i^{k+1}-x_I^*\right\|^2+\sigma\left[1\wedge\sqrt{V(b^{k+1},y^{k+1})}\right]\left\|x_i^{k+1}-x_I^*\right\|+\sigma^2\left[1\wedge V(b^{k+1},y^{k+1})\right] \\
&\lesssim \sigma\left\|x_i^{k+1}-x_I^*\right\|^2+\sigma\left[1\wedge V(b^{k+1},y^{k+1})\right]
\end{aligned}$$

where we used that $ab\le\frac{a^2}{2}+\frac{b^2}{2}$ to bound the second term of the first line. So

$$\begin{aligned}
\sum_{i\in\mathcal{N}(I)}(\widetilde{\varphi}_{Ii}^k-\widetilde{\varphi}_{Ii}^{k+1})a_i^k &\le \frac{1}{3\tau^3}\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^k a_i^k\cdot\left(\left\|x_i^{k+1}-x_I^*\right\|^3-\left\|x_i^k-x_I^*\right\|^3\right) \\
&\lesssim \sigma\frac{\lambda^3}{(\lambda\tau)^3}\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^k a_i^{k+1}\left(\left\|x_i^{k+1}-x_I^*\right\|^2+\left[1\wedge V(b^{k+1},y^{k+1})\right]\right) \\
&\lesssim \sqrt{\sigma}\left(\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^k a_i^{k+1}\left\|x_i^{k+1}-x_I^*\right\|^2+\left[1\wedge V(b^{k+1},y^{k+1})\right]\right)
\end{aligned}$$

where the last line follows from our choice of $\lambda^3=\frac{1}{\sqrt{\sigma}}$ and $\lambda\tau\asymp 1$. Similarly, on the other side,

$$\sum_{i\in\mathcal{N}(I)}(\widetilde{\varphi}_{Ii}^{k+1}-\widetilde{\varphi}_{Ii}^k)a_i^k\lesssim\sqrt{\sigma}\left(\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^{k+1}a_i^{k+1}\left\|x_i^{k+1}-x_I^*\right\|^2+\left[1\wedge V(b^{k+1},y^{k+1})\right]\right)$$

Finally, it remains to bound $\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^k a_i^{k+1}\left\|x_i^{k+1}-x_I^*\right\|^2$ as well as $\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^{k+1}a_i^{k+1}\left\|x_i^{k+1}-x_I^*\right\|^2$, in terms of $\sum_i\varphi_{Ii}^{k+1}a_i^{k+1}\left\|x_i^{k+1}-x_I^*\right\|^2=O\left(\sigma V_{\mathrm{pos}}(a^{k+1},x^{k+1})\right)$. This is done in the following Claim 44.

By putting everything together, we obtain that $\left|\sum_i(\varphi_{Ii}^{k+1}-\varphi_{Ii}^k)a_i^k\right|\lesssim\varepsilon\left(\bar{a}_0^k+\bar{a}_0^{k+1}\right)+\sqrt{\sigma}\left[1\wedge V(z^{k+1})\right]$, which is the first inequality of the lemma. The second inequality of the lemma follows by noting that $\varepsilon\bar{a}_0^{k+1}=O\left(\sqrt{\sigma}V(z^{k+1})\right)$, since $\varepsilon=e^{-1/(3\sqrt{\sigma})}=O(\sqrt{\sigma})$. ◀

▷ Claim 44. For small enough η and σ, for any $I\in[n^*]$,

$$\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^k a_i^{k+1}\left\|x_i^{k+1}-x_I^*\right\|^2\lesssim\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^{k+1}a_i^{k+1}\left\|x_i^{k+1}-x_I^*\right\|^2\lesssim\varepsilon\bar{a}_0^{k+1}+\sum_{i\in\mathcal{N}(I)}\varphi_{Ii}^{k+1}a_i^{k+1}\left\|x_i^{k+1}-x_I^*\right\|^2.$$

Proof. By the same reasoning as in the proof of Claim 41, one can show that

$$\sum_{i\in\mathcal{N}(I)}\left(\widetilde{\varphi}_{Ii}^{k+1}-\varphi_{Ii}^{k+1}\right)a_i^{k+1}\left\|x_i^{k+1}-x_I^*\right\|^2\le\varepsilon\bar{a}_0^{k+1}R^2.$$

Hence the second inequality. For the first inequality: As we saw in (33),

$$\sum_{i\in\mathcal{N}(I)}(\widetilde{\varphi}_{Ii}^k-\widetilde{\varphi}_{Ii}^{k+1})a_i^{k+1}\left\|x_i^{k+1}-x_I^*\right\|^2\le\frac{1}{3\tau^3}\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^k a_i^{k+1}\left\|x_i^{k+1}-x_I^*\right\|^2\left(\left\|x_i^{k+1}-x_I^*\right\|^3-\left\|x_i^k-x_I^*\right\|^3\right),$$

and by Claim 42 and Lemma 30, $\left\|x_i^{k+1}-x_I^*\right\|^3-\left\|x_i^k-x_I^*\right\|^3\lesssim\left\|x_i^{k+1}-x_i^k\right\|=O(\sigma)$. So, by our choice of $\lambda^3=\frac{1}{\sqrt{\sigma}}$ and $\lambda\tau\asymp 1$,

$$\sum_{i\in\mathcal{N}(I)}\left(\widetilde{\varphi}_{Ii}^k-\widetilde{\varphi}_{Ii}^{k+1}\right)a_i^{k+1}\left\|x_i^{k+1}-x_I^*\right\|^2\lesssim\underbrace{\sigma\frac{\lambda^3}{(\lambda\tau)^3}}_{\asymp\sqrt{\sigma}}\cdot\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^k a_i^{k+1}\left\|x_i^{k+1}-x_I^*\right\|^2.$$

Thus,

$$\begin{aligned}
(1-O(\sqrt{\sigma}))\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^k a_i^{k+1}\left\|x_i^{k+1}-x_I^*\right\|^2&\le\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^{k+1}a_i^{k+1}\left\|x_i^{k+1}-x_I^*\right\|^2 \\
\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^k a_i^{k+1}\left\|x_i^{k+1}-x_I^*\right\|^2&\le(1+O(\sqrt{\sigma}))\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^{k+1}a_i^{k+1}\left\|x_i^{k+1}-x_I^*\right\|^2.
\end{aligned}$$

◀

E.7.3 Proof of Lemma 34 (bound on (err1))

▷ Claim 45. For any $I \in [n^*]$,

$$\sum_i \frac{\varphi_{Ii}^{k+1} a_i^{k+1}}{\bar{a}_I^{k+1}} \log \frac{a_i^{k+1}/\bar{a}_I^{k+1}}{a_i^k/\bar{a}_I^k} = O\left(\sum_i (\varphi_{Ii}^{k+1} - \varphi_{Ii}^k) a_i^k\right) + \eta^2 O\left(V(z^{k+1})\right).$$

Lemma 34 follows straightforwardly from the claim and from Lemma 32.

Proof of the claim. Fix $I \in [n^*]$. Let $u_i^{k+1,I} = \frac{\varphi_{Ii}^{k+1} a_i^{k+1}}{\bar{a}_I^{k+1}}$, and we want to bound

$$\sum_i u_i^{k+1,I} \log \frac{a_i^{k+1}/\bar{a}_I^{k+1}}{a_i^k/\bar{a}_I^k} = \sum_i u_i^{k+1,I} \log \frac{Z a_i^{k+1}}{a_i^k} \; + \; \log \frac{\bar{a}_I^k}{Z \bar{a}_I^{k+1}}.$$

By (26), we have $\log \frac{a_i^{k+1}}{a_i^k} = -\eta[(M^{k+1}b^{k+1})_i - \rho] - \log Z$ where $Z = \sum_{i'} a_{i'}^k \; e^{-\eta[(M^{k+1}b^{k+1})_{i'} - \rho]}$, so

$$\sum_i u_i^{k+1,I} \log \frac{Z a_i^{k+1}}{a_i^k} = \sum_i u_i^{k+1,I} (-\eta)[(M^{k+1}b^{k+1})_i - \rho]$$

and

$$\begin{aligned}
\log \frac{\bar{a}_I^k}{Z \bar{a}_I^{k+1}} &= \log \frac{\sum_i \varphi_{Ii}^k a_i^k}{Z \sum_i \varphi_{Ii}^{k+1} a_i^{k+1}} \\
&= \log \frac{\sum_i \varphi_{Ii}^k a_i^k}{\sum_i \varphi_{Ii}^{k+1} a_i^k} + \log \frac{\sum_i \varphi_{Ii}^{k+1} a_i^k}{Z \sum_i \varphi_{Ii}^{k+1} a_i^{k+1}} \\
&= \log \frac{\sum_i \varphi_{Ii}^k a_i^k}{\sum_i \varphi_{Ii}^{k+1} a_i^k} + \log \frac{\sum_i \varphi_{Ii}^{k+1} a_i^{k+1} e^{\eta[(M^{k+1}b^{k+1})_i - \rho]}}{\sum_i \varphi_{Ii}^{k+1} a_i^{k+1}} \\
&= \log \frac{\sum_i \varphi_{Ii}^k a_i^k}{\sum_i \varphi_{Ii}^{k+1} a_i^k} + \log \sum_i u_i^{k+1,I} e^{\eta[(M^{k+1}b^{k+1})_i - \rho]}.
\end{aligned}$$

Now by Jensen inequality on concavity of log, since $\sum_i u_i^{k+1,I} = 1$,

$$\log \left[\sum_i u_i^{k+1,I} e^{\eta[(M^{k+1}b^{k+1})_i - \rho]}\right] + \sum_i u_i^{k+1,I} (-\eta)[(M^{k+1}b^{k+1})_i - \rho] \geq 0.$$

Furthermore, since $\log x \leq x - 1$ and $e^x = 1 + x + O(x^2)$ and using (29),

$$\begin{aligned}
&\log \left[\sum_i u_i^{k+1,I} e^{\eta[(M^{k+1}b^{k+1})_i - \rho]}\right] + \sum_i u_i^{k+1,I} (-\eta)[(M^{k+1}b^{k+1})_i - \rho] \\
&\leq \sum_i u_i^{k+1,I} \left(e^{\eta[(M^{k+1}b^{k+1})_i - \rho]} - 1 - \eta[(M^{k+1}b^{k+1})_i - \rho]\right) \\
&= \sum_i u_i^{k+1,I} O\left(\eta^2 [(M^{k+1}b^{k+1})_i - \rho]^2\right) \\
&= \eta^2 \sum_i \frac{\varphi_{Ii}^{k+1} a_i^{k+1}}{\bar{a}_I^{k+1}} O\left(\left\|x_i^{k+1} - x_I^*\right\|^2 + V(b^{k+1}, y^{k+1})\right) \\
&\lesssim \eta^2 O\left(V(z^{k+1})\right).
\end{aligned}$$

Thus,

$$\sum_i u_i^{k+1,I} \log \frac{a_i^{k+1}/\bar{a}_I^{k+1}}{a_i^k/\bar{a}_I^k} = \log \frac{\sum_i \varphi_{Ii}^k a_i^k}{\sum_i \varphi_{Ii}^{k+1} a_i^k} + \eta^2 O\left(V(z^{k+1})\right).$$

To upper- and lower-bound the first term, note that by Lemma 31

$$\sum_i \varphi_{Ii}^k a_i^k = \bar{a}_I^k \geq c \quad \text{and} \quad \sum_i \varphi_{Ii}^{k+1} a_i^k \geq \sum_i \varphi_{Ii}^{k+1} a_i^{k+1} (1 - O(\eta)) = \bar{a}_I^{k+1} (1 - O(\eta)) \geq c(1 - O(\eta)).$$

So just by bounding the derivative of log we have

$$\left|\log\sum_i \varphi_{Ii}^k a_i^k - \log\sum_i \varphi_{Ii}^{k+1} a_i^k\right| \le \underbrace{\frac{1}{c(1-O(\eta))}}_{=O(1)}\left|\sum_i(\varphi_{Ii}^{k+1}-\varphi_{Ii}^k)a_i^k\right|. \qquad \blacktriangleleft$$

E.7.4 Proof of Lemma 35 (bound on (err2))

Proof. Focus on the a terms. The quantity we want to upper-bound is

$$\begin{aligned}&\sum_i(\varphi_{0i}^{k+1}-\varphi_{0i}^k)a_i^k + \frac{\eta}{2\sigma}\sum_I\sum_i a_i^k\big(\varphi_{Ii}^{k+1}-\varphi_{Ii}^k\big)\big\|x_I^*-x_i^k\big\|^2\\ &= \sum_I\sum_i\big(\varphi_{Ii}^k-\varphi_{Ii}^{k+1}\big)\,a_i^k\left[1-\frac{\eta}{2\sigma}\big\|x_I^*-x_i^k\big\|^2\right].\end{aligned}$$

Fix $I\in[n^*]$. Note that the sum is only over indices $i\in\mathcal{N}(I)$ (otherwise $\varphi_{Ii}^{k+1}=\varphi_{Ii}^k=0$) and that we have for all such i

$$\begin{aligned}\big\|x_I^*-x_i^k\big\|^2 &\le (\lambda\tau+O(\sigma))^2 \le 2(\lambda\tau)^2+O(\sigma^2)\le\frac{\sigma}{\eta}+O(\sigma^2)\\ \frac{\eta}{2\sigma}\big\|x_I^*-x_i^k\big\|^2 &\le \frac12+O(\eta\sigma)\end{aligned}$$

due to our choice of λ,τ that ensures that $(\lambda\tau)^2\le\frac12\frac{\sigma}{\eta}$. So for small enough η,σ, we have for all $i\in\mathcal{N}(I)$ that $0\le 1-\frac{\eta}{2\sigma}\big\|x_I^*-x_i^k\big\|^2\le 1$. Following a similar reasoning as for Claim 41, but taking into account that we know the sign of the objects involved and that we are only interested in an upper bound, one can check that

$$\sum_i(\varphi_{Ii}^k-\varphi_{Ii}^{k+1})a_i^k\left[1-\frac{\eta}{\sigma}\big\|x_I^*-x_i^k\big\|^2\right]\le\sum_{i\in\mathcal{N}(I)}\big(\widetilde{\varphi}_{Ii}^k-\widetilde{\varphi}_{Ii}^{k+1}\big)a_i^k\left[1-\frac{\eta}{\sigma}\big\|x_I^*-x_i^k\big\|^2\right]+\varepsilon(1+O(\eta))\bar{a}_0^{k+1}$$

(note that $\bar{a}_0^k$ does not appear on the right-hand side). It remains to bound the first term. As we already saw in the proof of Lemma 32 (Eq. (33)), we have

$$\widetilde{\varphi}_{Ii}^k-\widetilde{\varphi}_{Ii}^{k+1}=\widetilde{\varphi}_{Ii}^k\cdot\left(1-\frac{\widetilde{\varphi}_{Ii}^{k+1}}{\widetilde{\varphi}_{Ii}^k}\right)\le\widetilde{\varphi}_{Ii}^k\cdot\big(\log\widetilde{\varphi}_{Ii}^k-\log\widetilde{\varphi}_{Ii}^{k+1}\big)=\widetilde{\varphi}_{Ii}^k\cdot\frac{1}{3\tau^3}\left(\big\|x_i^{k+1}-x_I^*\big\|^3-\big\|x_i^k-x_I^*\big\|^3\right)$$

and by Claim 42, we have the Taylor expansion

$$\begin{aligned}\big\|x_i^{k+1}-x_I^*\big\|^3-\big\|x_i^k-x_I^*\big\|^3 &= 3\left\langle x_i^{k+1}-x_i^k,x_i^{k+1}-x_I^*\right\rangle\big\|x_i^{k+1}-x_I^*\big\|\\ &\quad+O\left(\big\|x_i^{k+1}-x_i^k\big\|^2\big\|x_i^{k+1}-x_I^*\big\|+\big\|x_i^{k+1}-x_i^k\big\|^3\right).\end{aligned}$$

Now by the update equation (26) and the expansion (32), for any $i\in\mathcal{N}(I)$,

$$\begin{aligned}\left\langle x_i^{k+1}-x_i^k,x_i^{k+1}-x_I^*\right\rangle &= \left\langle-\sigma\frac{a_i^{k+1}}{a_i^k}\partial_x M_{i\bullet}^{k+1}b^{k+1},x_i^{k+1}-x_I^*\right\rangle = -\sigma\frac{a_i^{k+1}}{a_i^k}(x_i^{k+1}-x_I^*)^\top\partial_x M_{i\bullet}^{k+1}b^{k+1}\\ &\le -\sigma\frac{a_i^{k+1}}{a_i^k}\frac{\sigma_{\min}}{2}\big\|x_i^{k+1}-x_I^*\big\|^2+\sigma\frac{a_i^{k+1}}{a_i^k}O\left(\big\|x_i^{k+1}-x_I^*\big\|\sqrt{V(b^{k+1},y^{k+1})}\right)\\ &\le \sigma(1+O(\eta))O\left(\big\|x_i^{k+1}-x_I^*\big\|\sqrt{V(b^{k+1},y^{k+1})}\right)\end{aligned}$$

and so the terms arising from the order-1 terms in the Taylor expansion of $\big\|x_i^{k+1}-x_I^*\big\|^3-\big\|x_i^k-x_I^*\big\|^3$ are upper-bounded by

$$\begin{aligned}&\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^k a_i^k\left[1-\frac{\eta}{\sigma}\big\|x_I^*-x_i^k\big\|^2\right]\cdot\frac{1}{\tau^3}\left\langle x_i^{k+1}-x_i^k,x_i^{k+1}-x_I^*\right\rangle\big\|x_i^{k+1}-x_I^*\big\|\\ &\lesssim\frac{\sigma}{\tau^3}\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^k a_i^{k+1}\left[1-\frac{\eta}{\sigma}\big\|x_I^*-x_i^k\big\|^2\right]\cdot\big\|x_i^{k+1}-x_I^*\big\|^2\cdot\sqrt{V(b^{k+1},y^{k+1})}\\ &\lesssim\sigma\frac{\lambda^3}{(\lambda\tau)^3}\left(\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^k a_i^{k+1}\big\|x_i^{k+1}-x_I^*\big\|^2\right)\sqrt{V(b^{k+1},y^{k+1})}.\end{aligned}\tag{34}$$

Here in the last line we just bounded $\left|1-\frac{\eta}{\sigma}\|x_I^*-x_i^k\|^2\right|$ by 1. Further, by the update equation (26) and the expansion (31),

$$\|x_i^{k+1}-x_i^k\| = \sigma\frac{a_i^{k+1}}{a_i^k}\|\partial_x M_{i\bullet}^{k+1}b^{k+1}\| \lesssim \sigma(1+O(\eta))\left(\|x_i^{k+1}-x_I^*\|+\sqrt{V(b^{k+1},y^{k+1})}\right)$$

so the terms arising from higher-order terms (the $O(\,\cdot\,)$ terms) in the Taylor expansion of $\|x_i^{k+1}-x_I^*\|^3-\|x_i^k-x_I^*\|^3$ are upper-bounded by

$$\begin{aligned}
&\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^k a_i^k\left[1-\frac{\eta}{\sigma}\|x_I^*-x_i^k\|^2\right]\cdot\frac{1}{\tau^3}O\left(\|x_i^{k+1}-x_i^k\|^2\|x_i^{k+1}-x_I^*\|+\|x_i^{k+1}-x_i^k\|^3\right)\\
&\lesssim\frac{1}{\tau^3}\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^k a_i^{k+1}\cdot\left(\sigma^2\|x_i^{k+1}-x_I^*\|^3+\sigma^2\cdot V(b^{k+1},y^{k+1})\cdot\|x_i^{k+1}-x_I^*\|+\sigma^3\left[V(b^{k+1},y^{k+1})\right]^{3/2}\right)\\
&\lesssim\frac{1}{\tau^3}\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^k a_i^{k+1}\cdot\left(\sigma^2\|x_i^{k+1}-x_I^*\|^3+\sigma^2\left[V(b^{k+1},y^{k+1})\right]^{3/2}\right)\\
&\lesssim\sigma^2\frac{\lambda^3}{(\lambda\tau)^2}\left(\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^k a_i^{k+1}\|x_i^{k+1}-x_I^*\|^2\right)+\sigma^2\frac{\lambda^3}{(\lambda\tau)^3}\left[V(b^{k+1},y^{k+1})\right]^{3/2}
\end{aligned}\tag{35}$$

where in the second line we just bounded $\left|1-\frac{\eta}{\sigma}\|x_I^*-x_i^k\|^2\right|$ by 1 again, the third line follows from Young's inequality, and the last line uses that $\|x_i^{k+1}-x_I^*\|=O(\lambda\tau+\sigma)=O(\lambda\tau)$ for $i\in\mathcal{N}(I)$.

Putting everything together, by summing (34) and (35) and by using Claim 44 to bound $\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^k a_i^{k+1}\|x_i^{k+1}-x_I^*\|^2$, we get

$$\begin{aligned}
&\sum_I\sum_i\left(\varphi_{Ii}^k-\varphi_{Ii}^{k+1}\right)a_i^k\left[1-\frac{\eta}{2\sigma}\|x_I^*-x_i^k\|^2\right]\\
&\lesssim\varepsilon(\bar{a}_0^{k+1}+\bar{a}_0^k)+\sigma^2\frac{\lambda^3}{(\lambda\tau)^3}\left[V(b^{k+1},y^{k+1})\right]^{3/2}\\
&\quad+\sigma\frac{\lambda^3}{(\lambda\tau)^3}\left(\sum_I\sum_{i\in\mathcal{N}(I)}\widetilde{\varphi}_{Ii}^k a_i^{k+1}\|x_i^{k+1}-x_I^*\|^2\right)\left(\sqrt{V(b^{k+1},y^{k+1})}+\sigma\lambda\tau\right)\\
&\lesssim\varepsilon(\bar{a}_0^{k+1}+\bar{a}_0^k)+\sigma^2\frac{\lambda^3}{(\lambda\tau)^3}\left[V(b^{k+1},y^{k+1})\right]^{3/2}\\
&\quad+\sigma\frac{\lambda^3}{(\lambda\tau)^3}\left(\varepsilon\bar{a}_0^{k+1}+V_{\mathrm{pos}}(a^{k+1},x^{k+1})\right)\left(\sqrt{V(b^{k+1},y^{k+1})}+\sigma\lambda\tau\right).
\end{aligned}$$

Finally, we use that $\lambda\tau\asymp 1$ and that $\lambda^3=\frac{1}{\sqrt{\sigma}}$ and that $\eta\asymp\sigma$ to simplify the bound, and we obtain as announced that the above is upper-bounded up to a constant factor by

$$\varepsilon(\bar{a}_0^{k+1}+\bar{a}_0^k)+\sqrt{\sigma}\cdot\sigma V_{\mathrm{pos}}(a^{k+1},x^{k+1})+\sqrt{\sigma}V(z^{k+1})^{3/2}.$$ ◀

E.7.5 Proof of Lemma 36 (bound on (err3))

Proof. Focus on the a terms. We want to bound $\frac{\eta}{2\sigma}\sum_I\sum_i(a_i^{k+1}-a_i^k)\varphi_{Ii}^{k+1}\|x_I^*-x_i^{k+1}\|^2$. By the update equation (26),

$$a_i^{k+1}=a_i^k\ e^{-\eta[(M^{k+1}b^{k+1})_i-\rho]}/Z\quad\text{where}\quad Z=\sum_{i'}a_{i'}^k\ e^{-\eta[(M^{k+1}b^{k+1})_{i'}-\rho]}$$

$$\begin{aligned}
\text{i.e.}\quad a_i^{k+1}-a_i^k&=a_i^{k+1}\left[1-e^{\eta[(M^{k+1}b^{k+1})_i-\rho]}Z\right]\\
&=a_i^{k+1}\left[1-e^{\eta[(M^{k+1}b^{k+1})_i-\rho]}+e^{\eta[(M^{k+1}b^{k+1})_i-\rho]}(1-Z)\right].
\end{aligned}$$

So

$$\begin{aligned}
\sum_I\sum_i(a_i^{k+1}-a_i^k)\varphi_{Ii}^{k+1}\|x_I^*-x_i^{k+1}\|^2&=\sum_I\sum_i\varphi_{Ii}^{k+1}a_i^{k+1}\|x_I^*-x_i^{k+1}\|^2\left(1-e^{\eta[(M^{k+1}b^{k+1})_i-\rho]}\right)\\
&\quad+\sum_I\sum_i\varphi_{Ii}^{k+1}a_i^{k+1}\|x_I^*-x_i^{k+1}\|^2e^{\eta[(M^{k+1}b^{k+1})_i-\rho]}\ (1-Z).
\end{aligned}$$

For the first term, since $1-e^x \le -x$, then by the expansion (30),

$$1-e^{\eta[(M^{k+1}b^{k+1})_i-\rho]} \le -\eta[(M^{k+1}b^{k+1})_i-\rho] \le -\eta\left[\frac{1}{4}\sigma_{\min}\left\|x_i^{k+1}-x_I^*\right\|^2 + O\left(\sqrt{V(b^{k+1},y^{k+1})}\right)\right]$$
$$\lesssim \eta\sqrt{V(b^{k+1},y^{k+1})}.$$

So we get

$$\sum_I\sum_i \varphi_{Ii}^{k+1}a_i^{k+1}\left\|x_I^*-x_i^{k+1}\right\|^2\left(1-e^{\eta[(M^{k+1}b^{k+1})_i-\rho]}\right) \lesssim \sum_I\sum_i \varphi_{Ii}^{k+1}a_i^{k+1}\left\|x_I^*-x_i^{k+1}\right\|^2\cdot\eta\sqrt{V(b^{k+1},y^{k+1})}$$
$$\lesssim V_{\mathrm{pos}}(a^{k+1},x^{k+1})\cdot\eta\sqrt{V(b^{k+1},y^{k+1})}.$$

For the second term, write Z as

$$\forall\, i\in[n], a_i^{k+1} = a_i^k e^{-\eta[(M^{k+1}b^{k+1})_i-\rho]}/Z \quad\Longrightarrow\quad 1/Z = \sum_i a_i^{k+1}e^{\eta[(M^{k+1}b^{k+1})_i-\rho]}.$$

So using the expansion (29),

$$\begin{aligned}\frac{1}{Z}-1 &= \sum_i a_i^{k+1}\left(e^{\eta[(M^{k+1}b^{k+1})_i-\rho]}-1\right) = \sum_i a_i^{k+1}O\left(\eta[(M^{k+1}b^{k+1})_i-\rho]\right)\\ &= \sum_I\sum_i \varphi_{Ii}^{k+1}a_i^{k+1}O\left(\eta[(M^{k+1}b^{k+1})_i-\rho]\right) + \sum_i \varphi_{0i}^{k+1}a_i^{k+1}O\left(\eta[(M^{k+1}b^{k+1})_i-\rho]\right)\\ &= \eta\sum_I\sum_i \varphi_{Ii}^{k+1}a_i^{k+1}O\left(\left\|x_i^{k+1}-x_I^*\right\|^2+\sqrt{V(b^{k+1},y^{k+1})}\right) + O(\eta\bar a_0^{k+1})\\ &= \eta O\left(\sigma V_{\mathrm{pos}}(a^{k+1},x^{k+1})+\sqrt{V(b^{k+1},y^{k+1})}+\bar a_0^{k+1}\right) = \eta O\left(\sqrt{V(z^{k+1})}\right).\end{aligned}$$

So, since $e^{\eta[(M^{k+1}b^{k+1})_i-\rho]} \le e^{2\eta_0 L_0}$ and $Z=\sum_{i'} a_{i'}^k\, e^{-\eta[(M^{k+1}b^{k+1})_{i'}-\rho]} \le e^{2\eta_0L_0} = O(1)$, we get

$$\sum_I\sum_i \varphi_{Ii}^{k+1}a_i^{k+1}\left\|x_I^*-x_i^{k+1}\right\|^2 e^{\eta[(M^{k+1}b^{k+1})_i-\rho]}\,(1-Z) \le \sum_I\sum_i \varphi_{Ii}^{k+1}a_i^{k+1}\left\|x_I^*-x_i^{k+1}\right\|^2\cdot e^{4\eta_0L_0}\left|\frac{1}{Z}-1\right|$$
$$\lesssim V_{\mathrm{pos}}(a^{k+1},x^{k+1})\left|\frac{1}{Z}-1\right|$$
$$\lesssim V_{\mathrm{pos}}(a^{k+1},x^{k+1})\cdot\eta\sqrt{V(z^{k+1})}.$$

Putting the two bounds together gives the announced inequality. ◀

F Auxiliary lemmas

▶ **Lemma 46** (KL- vs. χ^2-divergence comparison). *For any $a,\widehat a\in\Delta_n$, denoting $D(a,\widehat a)$ the KL-divergence and $\chi^2(a,\widehat a)=\sum_i \frac{(\hat a_i-a_i)^2}{\hat a_i}$ the χ^2-divergence,*

$$D(a,\widehat a)\le \log\left(1+\chi^2(a,\widehat a)\right)\le\chi^2(a,\widehat a) \qquad \textit{and} \qquad D(a,\widehat a)\ge\left(\max_i\frac{a_i}{\widehat a_i}\right)^{-1}\chi^2(a,\widehat a).$$

Proof. For the first inequality, use Jensen's inequality on log, and that $\log(1+x)\le x$ for all x. For the second inequality, recall that the f-divergence of a w.r.t. $\widehat a$ is defined as $\sum_{i=1}^n \widehat a_i f\left(\frac{a_i}{\widehat a_i}\right)$. The KL-divergence $D(a,\widehat a)$ is the f-divergence for $f_D(t)=t\log(t)$ and the χ^2-divergence $\chi^2(a,\widehat a)$ is the f-divergence for $f_{\chi^2}(t)=t^2-t$. Note that for any $c>0$,

$$\forall\, t\le c,\quad f_D(t)\ge\frac{1}{c}f_{\chi^2}(t).$$

The claimed inequality follows by evaluating at $t=\frac{a_i}{\widehat a_i}$ and taking the sum weighted by $\widehat a_i$. ◀

▶ **Lemma 47.** *Let D_Φ denote the Bregman divergence associated to an arbitrary differentiable function $\Phi : X \to \mathbb{R}$, that is, $D_\Phi(x,y) = \Phi(x) - \Phi(y) - \langle \nabla\Phi(y), x-y\rangle$.*

- *D_Φ is non-negative if and only if Φ is convex, and D_Φ is zero if and only if Φ is linear.*
- *If Φ is convex, then $D_\Phi(x,y)$ is convex in x (but not in y in general).*
- *D_Φ is linear in Φ, and $D_{D_\Phi(\cdot,z)}(x,y) = D_\Phi(x,y)$.*
- *We have the three-point identity (or Pythagorean identity)*

$$\forall\, x,y,z \in X,\ D_\Phi(x,z) = D_\Phi(x,y) + D_\Phi(y,z) - \langle \nabla\Phi(z) - \nabla\Phi(y), x-y\rangle\,.$$

In particular,

$$\forall\, x,y,z \in X,\ \nabla_y D_\Phi(y,z)^\top (x-y) = D_\Phi(x,z) - D_\Phi(x,y) - D_\Phi(y,z).$$

- *For any fixed y_0 and y_1, $x \mapsto D_\Phi(x,y_1) - D_\Phi(x,y_0)$ is affine in x.*

The following lemma is just a rewriting of the facts remarked in Section 2.1 about the structure of the problem.

▶ **Lemma 48.** *Under the Assumptions 1-6, letting $\rho = F(\mu^*,\nu^*)$,*

$$\begin{aligned} &\forall\, x \in \mathcal{X},\ (F\nu^*)(x) \geq \rho && \forall\, y \in \mathcal{Y},\ ((\mu^*)^\top F)(y) \leq \rho \\ \text{and}\quad &\forall\, x \in \operatorname{supp}(\mu^*),\ (F\nu^*)(x) = \rho && \forall\, y \in \operatorname{supp}(\nu^*),\ ((\mu^*)^\top F)(y) = \rho \end{aligned}$$

and we have the first- and second-order conditions

$$\begin{aligned} &\forall\, x \in \operatorname{supp}(\mu^*),\ \partial_x(F\nu^*)(x) = 0 && \forall\, y \in \operatorname{supp}(\nu^*),\ \partial_y((\mu^*)^\top F)(y) = 0 \\ &\qquad \text{and}\quad \partial^2_{xx}(F\nu^*)(x) \succ 0 && \partial^2_{yy}((\mu^*)^\top F)(y) \prec 0. \end{aligned}$$

Using the shorthand notations detailed in Appendix A, this means that

$$\begin{aligned} &\forall\, I \in [n^*],\ M^*_{I\bullet} b^* = \rho && \forall\, J \in [m^*],\ (a^*)^\top M^*_{\bullet J} = \rho \\ &\quad \text{and}\quad \partial_x M^*_{I\bullet} b^* = 0 && (a^*)^\top \partial_y M^*_{\bullet J} = 0 \\ &\quad \text{and}\quad \partial^2_{xx} M^*_{I\bullet} b^* =: H_I \succ 0 && (a^*)^\top \partial^2_{yy} M^*_{\bullet J} =: -H_J \prec 0. \end{aligned}$$

F.1 Useful expressions and a priori bounds for V_{wei} and V_{pos}

In many technical proofs, it will be helpful to keep in mind the following decomposition of the Lyapunov function:

- By Pythagorean identity, we have the bias-variance decomposition

$$\begin{aligned} 2V_{\mathrm{pos}}(a,x) &= \sum_I \bar{a}_I \left(\left\|x^*_I - \bar{x}_I\right\|^2 + \operatorname{Tr}(\Sigma_I)\right) = \sum_I \sum_i a_i \varphi_{Ii} \left(\left\|x^*_I - \bar{x}_I\right\|^2 + \left\|x_i - \bar{x}_I\right\|^2 \right) \\ &= \sum_I \sum_i a_i \varphi_{Ii} \left(\left\|x^*_I - x_i\right\|^2 + 2\left\langle x^*_I - \bar{x}_I, x_i - \bar{x}_I\right\rangle \right) \\ &= \sum_I \sum_i a_i \varphi_{Ii} \left\|x^*_I - x_i\right\|^2. \end{aligned} \tag{36}$$

  In particular, note that by Jensen's inequality, $\sum_I \left\|(\bar{a}\odot\Delta\bar{x})_I\right\| = \sum_I \left\|\sum_i \varphi_{Ii} a_i (x_i - x^*_I)\right\| \leq \sum_{I,i} \varphi_{Ii} a_i \left\|x_i - x^*_I\right\| \leq \sqrt{\sum_{I,i} \varphi_{Ii} a_i \left\|x_i - x^*_I\right\|^2} = \sqrt{2V_{\mathrm{pos}}(a,x)}$.
- The stray weights $\bar{a}_0, \bar{b}_0$ play a special role. Indeed since $a^*_0 = 0$, then $d_h(a^*_0, \bar{a}_0) = \bar{a}_0$, so

$$V_{\mathrm{wei}}(a,x) = D(a^*, \bar{a}) = \sum_I d_h(a^*_I, \bar{a}_I) + \bar{a}_0. \tag{37}$$

  Now $d_h(s,s') \geq 0$. So V_{wei} can be viewed as a sum of two terms, both positive: The first one measures the (unnormalized entropic Bregman) distance between a^* and $(\bar{a}_I)_{I\in\mathcal{I}^*}$ in $\mathbb{R}_+^{\mathcal{I}^*}$, and the second one accounts for the stray weights $\bar{a}_0$. In particular, $\bar{a}_0 = O(V_{\mathrm{wei}}(a,x))$, while for $I \neq 0$ we only have $|\bar{a}_I - a^*_I| \leq \left\|\bar{a} - a^*\right\|_1 = O(\sqrt{V_{\mathrm{wei}}(a,x)})$ (by Pinsker's inequality or by 1-strong convexity of h).

G Calculations

In this section we present the simple but tedious calculations that constitute the proofs of Claim 20, Lemma 24 and Claim 26. They all consist in writing Taylor expansions of f around (x_I^*, y_J^*) or $(\widehat{x}_i, y_J^*)$ or $(x_I^*, \widehat{y}_j)$, and applying the facts collected in Lemma 48 and Appendix F.1, which we will use without explicit mention throughout this subsection.

We will also repeatedly use that as a consequence of (36) and (37),

$$\begin{aligned}\sum_I \left\|(\Delta\bar{a}\odot\Delta\bar{x})_I\right\| &= \sum_I |\Delta\bar{a}_I^*| \left\|\Delta\bar{x}_I\right\| = \sum_I \frac{1}{\bar{a}_I} |\Delta\bar{a}_I^*| \left\|\bar{a}_I\Delta\bar{x}_I\right\| \\ &\le \left(\min_I \bar{a}_I\right)^{-1} \left\|\Delta\bar{a}^*\right\|_1 \max_I \left\|\bar{a}_I\Delta\bar{x}_I\right\| \\ &= O\left(\left(\min_I \bar{a}_I\right)^{-1} \sqrt{V_{\mathrm{wei}}(a,x)}\sqrt{V_{\mathrm{pos}}(a,x)}\right) = O\left(\left(\min_I \bar{a}_I\right)^{-1} V_1(a,x)\right).\end{aligned}$$

G.1 Proof of Claim 20

Proof of Claim 20. Let us compute separately the four terms of $\widetilde{\mathrm{gap}}(z;\widehat{z}) = \left\langle \begin{pmatrix} \nabla_a \\ \nabla_x \\ -\nabla_b \\ -\nabla_y \end{pmatrix} F_{n,m}(\widehat{z}), \begin{pmatrix} \hat{a}-a \\ \hat{x}-x \\ \hat{b}-b \\ \hat{y}-y \end{pmatrix} \right\rangle$, where for ease of reference we recall that $x_i \coloneqq \widehat{x}_i + \sum_I \widehat{\varphi}_{Ii}(X_I - \widehat{x}_i)$ and $a_i \coloneqq \sum_I A_I \frac{\widehat{\varphi}_{Ii}\widehat{a}_i}{\widehat{a}_I}$. Focus on the $\nabla_a F_{n,m}$ term (and the $-\nabla_b F_{n,m}$ term is dealt with analogously). By definition $\langle \delta a, \nabla_a F_{n,m}(\widehat{z})\rangle = (\delta a)^\top \widehat{M}\widehat{b}$, so

$$\langle \nabla_a F_{n,m}(\widehat{z}), \widehat{a} - a\rangle = \widehat{a}^\top \widehat{M}\widehat{b} - a^\top \widehat{M}\widehat{b}$$

and by Taylor expansions of f around (x_I^*, y_J^*),

$$\begin{aligned}
&- a^\top \widehat{M}\widehat{b} = -\sum_I A_I \sum_{ij} \frac{\widehat{\varphi}_{Ii}\widehat{a}_i}{\widehat{a}_I} \widehat{M}_{ij}\widehat{b}_j \\
&= -\sum_{IJ} \frac{A_I}{\widehat{a}_I} \sum_{ij} \widehat{\varphi}_{Ii}\widehat{\psi}_{Jj}\cdot\widehat{a}_i \Big[M_{IJ}^* + (\widehat{x}_i - x_I^*)^\top \partial_x M_{IJ}^* + \partial_y M_{IJ}^*(\widehat{y}_j - y_J^*) \\
&\qquad\qquad + O(\|\widehat{x}_i - x_I^*\|^2 + \|\widehat{y}_j - y_J^*\|^2)\Big]\widehat{b}_j + O\left(\widehat{\bar{b}}_0\right) \\
&= -A^\top M^* \widehat{\bar{b}} - (A\odot(\widehat{x} - x^*))^\top \partial_x M^* \widehat{\bar{b}} - A^\top \partial_y M^*((\widehat{y} - y^*)\odot\widehat{\bar{b}}) + O\left(\widehat{\bar{b}}_0 + (\min_I \widehat{a}_I)^{-1} V_{\mathrm{pos}}(\widehat{a},\widehat{x}) + V_{\mathrm{pos}}(\widehat{b},\widehat{y})\right) \\
&= -A^\top M^* \widehat{\bar{b}} - (A\odot\Delta\widehat{x})^\top \partial_x M^* \Delta\widehat{\bar{b}} - \Delta A^\top \partial_y M^*(\Delta\widehat{y}\odot\widehat{\bar{b}}) + O\left(\left(\min_I \widehat{w}_I\right)^{-1} V_1(\widehat{z})\right) \\
&= -A^\top M^* \widehat{\bar{b}} - \Delta A^\top \partial_y M^*(\Delta\widehat{y}\odot b^*) + O\left(\left(\min_I \widehat{w}_I\right)^{-1} V_1(\widehat{z})\right).
\end{aligned}$$

For the $\nabla_x F_{n,m}$ term (and the $-\nabla_y F_{n,m}$ term is dealt with analogously), by definition $\langle \delta x, \nabla_x F_{n,m}(\widehat{z})\rangle = (\widehat{a}\odot\delta x)^\top \partial_x \widehat{M}\widehat{b}$, so

$$\begin{aligned}
&\langle \nabla_x F_{n,m}(\widehat{z}), \widehat{x} - x\rangle = \sum_i \widehat{a}_i (\widehat{x}_i - x_i)^\top \partial_x \widehat{M}_{i\bullet}\widehat{b} = \sum_I \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i (\widehat{x}_i - X_I)^\top \partial_x \widehat{M}_{i\bullet}\widehat{b} \\
&= \sum_{IJ}\sum_{ij} \widehat{\varphi}_{Ii}\widehat{\psi}_{Jj}\cdot\widehat{a}_i(\widehat{x}_i - X_I)^\top \Big[\partial_x M_{IJ}^* + (\widehat{x}_i - x_I^*)^\top \partial_{xx}^2 M_{IJ}^* + \partial_{xy}^2 M_{IJ}^*(\widehat{y}_j - y_J^*) \\
&\qquad\qquad + O\left(\left\|\widehat{x}_i - x_I^*\right\|^2 + \left\|\widehat{y}_j - y_J^*\right\|^2\right)\Big]\widehat{b}_j + O\left(\widehat{\bar{b}}_0\right) \\
&= (\widehat{a}\odot(\widehat{x} - X))^\top \partial_x M^* \widehat{\bar{b}} + \sum_{I,i} \widehat{\varphi}_{Ii}\widehat{a}_i (\widehat{x}_i - X_I, \widehat{x}_i - x_I^*)^\top \partial_{xx}^2 M_{I\bullet}^* \widehat{\bar{b}} + (\widehat{a}\odot(\widehat{x} - X))^\top \partial_{xy}^2 M^*((\widehat{y} - y^*)\odot\widehat{\bar{b}}) \\
&\quad + O\left(\widehat{\bar{b}}_0 + V_{\mathrm{pos}}(\widehat{z})\right) \\
&= (\widehat{a}\odot(\widehat{x} - X))^\top \partial_x M^* \Delta\widehat{\bar{b}} + \sum_{I,i} \widehat{\varphi}_{Ii}\widehat{a}_i \langle \widehat{x}_i - X_I, \widehat{x}_i - x_I^*\rangle_{H_I} + (\widehat{a}\odot(\widehat{x} - X))^\top \partial_{xy}^2 M^*((\widehat{y} - y^*)\odot\widehat{\bar{b}}) \\
&\quad + O\left(V_1(\widehat{z})\right) \\
&= -(\widehat{a}\odot\Delta X)^\top \partial_x M^* \Delta\widehat{\bar{b}} + \sum_{I,i} \widehat{\varphi}_{Ii}\widehat{a}_i \langle \widehat{x}_i - X_I, \widehat{x}_i - x_I^*\rangle_{H_I} - (\widehat{a}\odot\Delta X)\partial_{xy}^2 M^*(\Delta\widehat{y}\odot\widehat{\bar{b}}) + O\left(V_1(\widehat{z})\right)
\end{aligned}$$

$$= -(a^* \odot \Delta X)^\top \partial_x M^* \Delta\widehat{\bar{b}} + \sum_{I,i} \widehat{\varphi}_{Ii}\widehat{a}_i \langle \widehat{x}_i - X_I, \widehat{x}_i - x_I^* \rangle_{H_I} - (a^* \odot \Delta X)^\top \partial_{xy}^2 M^* (\Delta\widehat{\bar{y}} \odot b^*)$$
$$+ O\left(\left(\min_I \widehat{\bar{w}}_I \right)^{-1} V_1(\widehat{z}) \right).$$

Further transform the second term as

$$\sum_{I,i} \widehat{\varphi}_{Ii}\widehat{a}_i \langle \widehat{x}_i - X_I, \widehat{x}_i - x_I^* \rangle_{H_I} = \sum_{I,i} \widehat{\varphi}_{Ii}\widehat{a}_i \left\| \widehat{x}_i - x_I^* \right\|_{H_I}^2 + \sum_{I,i} \widehat{\varphi}_{Ii}\widehat{a}_i \langle x_I^* - X_I, \widehat{x}_i - x_I^* \rangle_{H_I}$$
$$= O\left(V_{\mathrm{pos}}(\widehat{a}, \widehat{x}) \right) - \sum_I \widehat{\bar{a}}_I \left\langle \Delta X_I, \Delta\widehat{\bar{x}}_I \right\rangle_{H_I}.$$

Putting everything together, and using that

$$-A^\top M^* \widehat{\bar{b}} + \widehat{\bar{a}}^\top M^* B = -A^\top M^* \Delta\widehat{\bar{b}} + \Delta\widehat{\bar{a}}^\top M^* B = -\Delta A^\top M^* \Delta\widehat{\bar{b}} + \Delta\widehat{\bar{a}}^\top M^* \Delta B + O\left(\widehat{\bar{a}}_0 + \widehat{\bar{b}}_0 \right),$$

we get the desired estimate. ◀

G.2 Proof of Lemma 24

Proof of Lemma 24. Let us compute separately the four terms of $\widehat{\mathrm{gap}}(z^{(*)}; \widehat{z}) = \left\langle \begin{pmatrix} \nabla_a \\ \nabla_x \\ -\nabla_b \\ -\nabla_y \end{pmatrix} F_{n,m}(\widehat{z}), \begin{pmatrix} \hat{a} - a^{(*)} \\ \hat{x} - x^{(*)} \\ \hat{b} - b^{(*)} \\ \hat{y} - y^{(*)} \end{pmatrix} \right\rangle$, where for ease of reference we recall that $x_i^{(*)} \coloneqq \widehat{x}_i + \sum_I \widehat{\varphi}_{Ii}(x_I^* - \widehat{x}_i)$ and $a_i^{(*)} \coloneqq \sum_I a_I^* \frac{\widehat{\varphi}_{Ii}\widehat{a}_i}{\widehat{\bar{a}}_I}$.

In the calculations below, we write $\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}_{Iij} \in [-1, 1]$ or $\in B_{0,1}$ to denote quantities possibly dependent on summation indices, and that may change from line to line. This is done in order to track error terms with more precision than using $O(\,\cdot\,)$'s. Focus on the $\nabla_a F_{n,m}$ term (and the $-\nabla_b F_{n,m}$ term is dealt with analogously). By definition $\langle \delta a, \nabla_a F_{n,m}(\widehat{z}) \rangle = (\delta a)^\top \widehat{M}\widehat{b}$, so

$$\langle \nabla_a F_{n,m}(\widehat{z}), \widehat{a} - a \rangle = \widehat{a}^\top \widehat{M}\widehat{b} - (a^{(*)})^\top \widehat{M}\widehat{b},$$

and by Taylor expansions of f around $(x_I^*, \widehat{y}_j)$,

$$- (a^{(*)})^\top \widehat{M}\widehat{b} = -\sum_I a_I^* \sum_{ij} \frac{\widehat{\varphi}_{Ii}\widehat{a}_i}{\widehat{\bar{a}}_I} \widehat{M}_{ij}\widehat{b}_j$$
$$= -\sum_I a_I^* \sum_{ij} \frac{\widehat{\varphi}_{Ii}\widehat{a}_i}{\widehat{\bar{a}}_I} \left[({}^*M^\wedge)_{Ij} + (\widehat{x}_i - x_I^*)^\top \partial_x ({}^*M^\wedge)_{Ij} + \frac{1}{2}((\widehat{x}_i - x_I^*)^2)^\top \partial_{xx}^2 ({}^*M^\wedge)_{Ij} + L_3 \boldsymbol{\varepsilon}_{Iij} \left\| \widehat{x}_i - x_I^* \right\|^3 \right] \widehat{b}_j$$
$$= -(a^*)^\top ({}^*M^\wedge)\widehat{b} - (a^* \odot \Delta\widehat{\bar{x}})^\top \partial_x ({}^*M^\wedge)\widehat{b} - \frac{1}{2}\sum_I a_I^* \sum_i \frac{\widehat{\varphi}_{Ii}\widehat{a}_i}{\widehat{\bar{a}}_I} ((\widehat{x}_i - x_I^*)^2)^\top \partial_{xx}^2 ({}^*M^\wedge)_{I\bullet}\widehat{b}$$
$$+ L_3 \boldsymbol{\varepsilon} \sum_I a_I^* \sum_i \frac{\widehat{\varphi}_{Ii}\widehat{a}_i}{\widehat{\bar{a}}_I} \left\| \widehat{x}_i - x_I^* \right\|^3.$$

Now for any I, J, j, $\partial_{xx}^2 ({}^*M^\wedge)_{Ij} = \partial_{xx}^2 M_{IJ}^* + O\left(\left\| \widehat{y}_j - y_J^* \right\| \right)$, so

$$\sum_I a_I^* \sum_i \frac{\widehat{\varphi}_{Ii}\widehat{a}_i}{\widehat{\bar{a}}_I} ((\widehat{x}_i - x_I^*)^2)^\top \partial_{xx}^2 ({}^*M^\wedge)_{I\bullet}\widehat{b}$$
$$= \sum_{IJ} \frac{a_I^*}{\widehat{\bar{a}}_I} \sum_{ij} \widehat{\varphi}_{Ii}\psi_{Jj} \cdot \widehat{a}_i ((\widehat{x}_i - x_I^*)^2)^\top \partial_{xx}^2 ({}^*M^\wedge)_{I\bullet}\widehat{b}_j + \sum_I \frac{a_I^*}{\widehat{\bar{a}}_I} \sum_{ij} \widehat{\varphi}_{Ii}\psi_{0j} \cdot \widehat{a}_i ((\widehat{x}_i - x_I^*)^2)^\top \partial_{xx}^2 ({}^*M^\wedge)_{I\bullet}\widehat{b}_j$$
$$= \sum_{IJ} \frac{a_I^*}{\widehat{\bar{a}}_I} \sum_{ij} \widehat{\varphi}_{Ii}\psi_{Jj} \cdot \widehat{a}_i ((\widehat{x}_i - x_I^*)^2)^\top \left[\partial_{xx}^2 M_{IJ}^* + O\left(\left\| \widehat{y}_j - y_J^* \right\| \right) \right] \widehat{b}_j + O\left(\sum_I \frac{a_I^*}{\widehat{\bar{a}}_I} \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i \left\| \widehat{x}_i - x_I^* \right\|^2 \cdot \widehat{\bar{b}}_0 \right)$$
$$= \sum_{IJ} \frac{a_I^*}{\widehat{\bar{a}}_I} \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i ((\widehat{x}_i - x_I^*)^2)^\top \partial_{xx}^2 M_{IJ}^* \widehat{\bar{b}}_J + O\left(\sum_I \frac{a_I^*}{\widehat{\bar{a}}_I} \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i \left\| \widehat{x}_i - x_I^* \right\|^2 \left(\sqrt{V_{\mathrm{pos}}(\widehat{b}, \widehat{y})} + \widehat{\bar{b}}_0 \right) \right)$$
$$= \sum_I \frac{a_I^*}{\widehat{\bar{a}}_I} \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i \left\| \widehat{x}_i - x_I^* \right\|_{H_I}^2 + O\left(\left(\min_I \widehat{\bar{a}}_I \right)^{-1} V_{\mathrm{pos}}(\widehat{a}, \widehat{x}) \left(\sqrt{V_{\mathrm{pos}}(\widehat{b}, \widehat{y})} + \left\| \Delta\widehat{\bar{b}} \right\|_1 \right) \right).$$

For the $\nabla_x F_{n,m}$ term (and the $-\nabla_y F_{n,m}$ term is dealt with analogously), by definition $\langle \delta x, \nabla_x F_{n,m}(\widehat{z})\rangle = (\widehat{a}\odot\delta x)^\top \partial_x \widehat{M}\widehat{b}$, so

$$\begin{aligned}
&\left\langle \nabla_x F_{n,m}(\widehat{z}), \widehat{x} - x^{(*)} \right\rangle = \sum_i \widehat{a}_i (\widehat{x}_i - x_i^{(*)})^\top \partial_x \widehat{M}_{i\bullet}\widehat{b} = \sum_I \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i (\widehat{x}_i - x_I^*)^\top \partial_x \widehat{M}_{i\bullet}\widehat{b} \\
&= \sum_I \sum_{ij} \widehat{\varphi}_{Ii}\widehat{a}_i (\widehat{x}_i - x_I^*)^\top \left[\partial_x ({}^*M^\wedge)_{Ij} + (\widehat{x}_i - x_I^*)^\top \partial^2_{xx} ({}^*M^\wedge)_{Ij} + L_3\boldsymbol{\varepsilon}_{Iij}\left\|\widehat{x}_i - x_I^*\right\|^2\right]\widehat{b}_j \\
&= (\widehat{\widehat{a}}\odot\Delta\widehat{\widehat{x}})^\top \partial_x ({}^*M^\wedge)\widehat{b} + \sum_I \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i ((\widehat{x}_i - x_I^*)^2)^\top \partial^2_{xx} ({}^*M^\wedge)_{I\bullet}\widehat{b} + L_3\boldsymbol{\varepsilon} \sum_I \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i \left\|\widehat{x}_i - x_I^*\right\|^3.
\end{aligned}$$

Now by the same calculation as previously,

$$\sum_I \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i ((\widehat{x}_i - x_I^*)^2)^\top \partial^2_{xx} ({}^*M^\wedge)_{I\bullet}\widehat{b} = \sum_I \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i \left\|\widehat{x}_i - x_I^*\right\|^2_{H_I} + O\left(V_{\mathrm{pos}}(\widehat{a},\widehat{x})\sqrt{V_1(\widehat{b},\widehat{y})}\right).$$

Putting everything together we get

$$\begin{aligned}
\widetilde{\mathrm{gap}}(z;\widehat{z}) = {}& \widehat{a}^\top ({}^\wedge M^*) b^* - (a^*)^\top ({}^*M^\wedge)\widehat{b} \\
&+ ((\widehat{\widehat{a}} - a^*)\odot\Delta\widehat{\widehat{x}})^\top \partial_x ({}^*M^\wedge)\widehat{b} - \widehat{a}^\top \partial_y ({}^\wedge M^*)(\Delta\widehat{\widehat{y}}\odot(\widehat{\widehat{b}} - b^*)) \\
&+ \sum_I \left(1 - \frac{1}{2}\frac{a_I^*}{\widehat{\widehat{a}}_I}\right) \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i \left\|\widehat{x}_i - x_I^*\right\|^2_{H_I} + \sum_J \left(1 - \frac{1}{2}\frac{b_J^*}{\widehat{\widehat{b}}_J}\right) \sum_j \widehat{\psi}_{Jj}\widehat{b}_j \left\|\widehat{y}_j - y_J^*\right\|^2_{H_J} \\
&+ L_3\boldsymbol{\varepsilon} \sum_I \left(1 + \frac{w_I^*}{\widehat{\widehat{w}}_I}\right) \sum_i \widehat{\varphi}_{Ii}\widehat{w}_i \left\|\widehat{p}_i - p_I^*\right\|^3 + O\left(\left(\min_I \widehat{\widehat{a}}_I\right)^{-1} V_1(\widehat{z})^{3/2}\right).
\end{aligned}$$

Now,

- The terms on the second line are negligible, as

$$\begin{aligned}
\partial_x ({}^*M^\wedge)_{I\bullet}\widehat{b} &= \sum_J \sum_j \widehat{\psi}_{Jj}\left[\partial_x M^*_{IJ} + O(\left\|\widehat{y}_j - y_J^*\right\|)\right]\widehat{b}_j + O\left(\widehat{\widehat{b}}_0\right) \\
&= \partial_x M^*_{I\bullet}\Delta\widehat{\widehat{b}} + O\left(\sqrt{V_{\mathrm{pos}}(\widehat{b},\widehat{y})}\right) + O\left(\widehat{\widehat{b}}_0\right) = O\left(\sqrt{V_1(\widehat{b},\widehat{y})}\right)
\end{aligned}$$

  so $\quad ((\widehat{\widehat{a}} - a^*)\odot\Delta\widehat{\widehat{x}})^\top \partial_x ({}^*M^\wedge)\widehat{b} = O\left(\left(\min_I \widehat{\widehat{a}}_I\right)^{-1} V_1(\widehat{a},\widehat{x})\cdot\sqrt{V_1(\widehat{b},\widehat{y})}\right).$

- Part of the terms on the third line turn out to be negligible, as

$$\begin{aligned}
\sum_I \left(\frac{1}{2} - \frac{1}{2}\frac{a_I^*}{\widehat{\widehat{a}}_I}\right) \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i \left\|\widehat{x}_i - x_I^*\right\|^2_{H_I} &= \frac{1}{2}\sum_I \frac{\widehat{\widehat{a}}_I - a_I^*}{\widehat{\widehat{a}}_I} \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i \left\|\widehat{x}_i - x_I^*\right\|^2 \\
&= O\left(\left(\min_I \widehat{\widehat{a}}_I\right)^{-1} \left\|\widehat{\widehat{a}} - a^*\right\|_1 V_{\mathrm{pos}}(\widehat{a},\widehat{x})\right).
\end{aligned}$$

- On the last line, the term in $L_3\boldsymbol{\varepsilon}$ is absolutely bounded by

$$L_3 \cdot 1 \cdot \sum_I 2\left(\min_I \widehat{\widehat{w}}_I\right)^{-1} \widehat{\varphi}_{Ii}\widehat{w}_i \left\|\widehat{p}_i - p_I^*\right\|^2 \cdot \lambda\tau \le 2L_3\left(\min_I \widehat{\widehat{w}}_I\right)^{-1} \cdot V_{\mathrm{pos}}(\widehat{z})\cdot\lambda\tau.$$

So as announced,

$$\begin{aligned}
\widetilde{\mathrm{gap}}(z;\widehat{z}) = {}& \widehat{a}^\top ({}^\wedge M^*) b^* - (a^*)^\top ({}^*M^\wedge)\widehat{b} + \frac{1}{2}\sum_I \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i \left\|\widehat{x}_i - x_I^*\right\|^2_{H_I} + \frac{1}{2}\sum_J \sum_j \widehat{\psi}_{Jj}\widehat{b}_j \left\|\widehat{y}_j - y_J^*\right\|^2_{H_J} \\
&+ O\left(\left(\min_I \widehat{\widehat{w}}_I\right)^{-1} V_1(\widehat{z})^{3/2}\right) + \boldsymbol{\varepsilon}\cdot\left[2L_3\left(\min_I \widehat{\widehat{w}}_I\right)^{-1}\cdot V_{\mathrm{pos}}(\widehat{z})\cdot\lambda\tau\right].
\end{aligned}$$

◀

G.3 Proof of Claim 26

Proof of Claim 26. Let any $\widehat{z} = (\widehat{a},\widehat{x},\widehat{b},\widehat{y}) \in \Delta_n \times \mathcal{X}^n \times \Delta_m \times \mathcal{Y}^m$ and $Z = (A,X,B,Y) \in \Delta_{n^*} \times \mathcal{X}^{n^*} \times \Delta_{m^*} \times \mathcal{Y}^{m^*}$, and denote $\widehat{Z} = (\widehat{a},\widehat{\widehat{a}},\widehat{\widehat{b}},\widehat{\widehat{y}})$. Recall that, as defined in (17), for any $\widetilde{X} \in \mathcal{X}^{n^*}$, $\left\|\Delta\widetilde{X}\right\| = \max_I \left\|\Delta\widetilde{X}_I\right\|$, and for any $\widetilde{Z} = (\widetilde{A},\widetilde{X},\widetilde{B},\widetilde{Y})$, $\left\|\Delta\widetilde{Z}\right\|^2 = \left\|\Delta\widetilde{A}\right\|_1^2 + \left\|\Delta\widetilde{X}\right\|^2 + \left\|\Delta\widetilde{B}\right\|_1^2 + \left\|\Delta\widetilde{Y}\right\|^2$.

In the calculations below, we write $\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}_{Ii}, \boldsymbol{\varepsilon}_{IJi} \in [-1,1]$ to denote quantities possibly dependent on summation indices, and that may change from line to line. This is done in order to track error terms with more precision than using $O(\,\cdot\,)$'s. By Taylor expansions of f around (x_I^*, y_J^*), we have

$$
\begin{aligned}
&F_{n,m^*}(\widehat{a},\widehat{x},B,Y) = \sum_{i=1}^{n}\sum_{J\in[m^*]} \widehat{a}_i f(\widehat{x}_i, Y_J) B_J \\
&= \sum_{IJ}\sum_i \widehat{\varphi}_{Ii}\widehat{a}_i B_J \Bigg[M^*_{IJ} + (\widehat{x}_i - x_I^*)^\top \partial_x M^*_{IJ} + \partial_y M^*_{IJ}(Y_J - y_J^*) + (\widehat{x}_i - x_I^*)^\top \partial^2_{xy} M^*_{IJ}(Y_J - y_J^*) \\
&\qquad + \frac{1}{2}((\widehat{x}_i - x_I^*)^2)^\top \partial^2_{xx} M^*_{IJ} + L_3 \boldsymbol{\varepsilon}_{IJi} \big\|\widehat{x}_i - x_I^*\big\|^3 + O\left(\big\|Y_J - y_J^*\big\|^2\right) \Bigg] \\
&\quad + \sum_J \sum_i \widehat{\varphi}_{0i}\widehat{a}_i B_J \left[(^\wedge M^*)_{iJ} + O\left(\big\|Y_J - y_J^*\big\|\right)\right] \\
&= \widehat{\bar{a}}^\top M^* B + (\widehat{\bar{a}} \odot \Delta\widehat{\bar{x}})^\top \partial_x M^* B + \widehat{\bar{a}}^\top \partial_y M^*(\Delta Y \odot B) + (\widehat{\bar{a}} \odot \Delta\widehat{\bar{x}})^\top \partial^2_{xy} M^*(\Delta Y \odot B) \\
&\quad + \frac{1}{2}\sum_I \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i \big\|\widehat{x}_i - x_I^*\big\|^2_{H_I} + \sum_{IJ}\sum_i \widehat{\varphi}_{Ii}\widehat{a}_i \Delta B_J O\left(\big\|\widehat{x}_i - x_I^*\big\|^2\right) + \sum_I \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i L_3 \boldsymbol{\varepsilon}_{Ii}\big\|\widehat{x}_i - x_I^*\big\|^3 \\
&\quad + (\widehat{\varphi}_0 \odot \widehat{a})^\top (^\wedge M^*) B + O\left(\widehat{\bar{a}}_0 \big\|\Delta Y\big\|\right) + O\left(\big\|\Delta Y\big\|^2\right) \\
&= \widehat{\bar{a}}^\top M^* B + (\widehat{\varphi}_0 \odot \widehat{a})^\top (^\wedge M^*) B \\
&\quad + (\widehat{\bar{a}} \odot \Delta\widehat{\bar{x}})^\top \partial_x M^* B + \widehat{\bar{a}}^\top \partial_y M^*(\Delta Y \odot B) + (\widehat{\bar{a}} \odot \Delta\widehat{\bar{x}})^\top \partial^2_{xy} M^*(\Delta Y \odot B) \\
&\quad + \frac{1}{2}\sum_I \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i \left(\big\|\widehat{x}_i - x_I^*\big\|^2_{H_I} + 2L_3 \boldsymbol{\varepsilon}_{Ii}\big\|\widehat{x}_i - x_I^*\big\|^3\right) \\
&\quad + O\left(\widehat{\bar{a}}_0 \big\|\Delta Y\big\|\right) + O\left(\big\|\Delta Y\big\|^2\right) + O\left(\big\|\Delta B\big\|_1 \cdot V_{\mathrm{pos}}(\widehat{a},\widehat{x})\right).
\end{aligned}
$$

- The first line can be rewritten as

$$
\begin{aligned}
\widehat{\bar{a}}^\top M^* B + (\widehat{\varphi}_0 \odot \widehat{a})^\top (^\wedge M^*) B &= (1 - \widehat{\bar{a}}_0)\rho + \widehat{\bar{a}}^\top M^* \Delta B + (\widehat{\varphi}_0 \odot \widehat{a})^\top (^\wedge M^*) B \\
&= (1 - \widehat{\bar{a}}_0)\rho + \Delta\widehat{\bar{a}}^\top M^* \Delta B + (\widehat{\varphi}_0 \odot \widehat{a})^\top (^\wedge M^*) B \\
&= \rho + \Delta\widehat{\bar{a}}^\top M^* \Delta B + (\widehat{\varphi}_0 \odot \widehat{a})^\top \left[(^\wedge M^*) B - \rho \mathbf{1}\right] \\
&= \rho + \Delta\widehat{\bar{a}}^\top M^* \Delta B + (\widehat{\varphi}_0 \odot \widehat{a})^\top \underbrace{\left[(^\wedge M^*) b^* - \rho \mathbf{1}\right]}_{\geq 0} + O\left(\widehat{\bar{a}}_0 \big\|\Delta B\big\|_1\right).
\end{aligned}
$$

- On the second line, it is not hard to check that

$$
(\widehat{\bar{a}} \odot \Delta\widehat{\bar{x}})^\top \partial_x M^* B = (\widehat{\bar{a}} \odot \Delta\widehat{\bar{x}})^\top \partial_x M^* \Delta B = (a^* \odot \Delta\widehat{\bar{x}})^\top \partial_x M^* \Delta B + O\left(\big\|\Delta\widehat{\bar{a}}\big\|_1 \big\|\Delta\widehat{\bar{x}}\big\| \big\|\Delta B\big\|_1\right)
$$

and likewise for the other terms, and so

$$
\begin{aligned}
&(\widehat{\bar{a}} \odot \Delta\widehat{\bar{x}})^\top \partial_x M^* B + \widehat{\bar{a}}^\top \partial_y M^*(\Delta Y \odot B) + (\widehat{\bar{a}} \odot \Delta\widehat{\bar{x}})^\top \partial^2_{xy} M^*(\Delta Y \odot B) \\
&= (a^* \odot \Delta\widehat{\bar{x}})^\top \partial_x M^* \Delta B + \Delta\widehat{\bar{a}}^\top \partial_y M^*(\Delta Y \odot b^*) + (a^* \odot \Delta\widehat{\bar{x}})^\top \partial^2_{xy} M^*(\Delta Y \odot b^*) + O\left(\big\|\Delta\widehat{Z}\big\|^3 + \|\delta z\|^3\right).
\end{aligned}
$$

- We can lower-bound the first term on the third line as

$$
\frac{1}{2}\sum_I \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i \left(\big\|\widehat{x}_i - x_I^*\big\|^2_{H_I} + 2L_3 \boldsymbol{\varepsilon}_{Ii}\big\|\widehat{x}_i - x_I^*\big\|^3\right) \geq \frac{1}{2}\sum_I \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i \big\|\widehat{x}_i - x_I^*\big\|^2 \underbrace{(\sigma_{\min} - 2L_3 \boldsymbol{\varepsilon}_{Ii} \cdot \lambda\tau)}_{\geq \frac{\sigma_{\min}}{2}}
$$

by our assumption that $\lambda\tau \leq \frac{\sigma_{\min}}{4L_3}$. Further, we can decompose this lower bound as

$$
\sum_I \sum_i \widehat{\varphi}_{Ii}\widehat{a}_i \big\|\widehat{x}_i - x_I^*\big\|^2 = \sum_I \widehat{\bar{a}}_I \big\|\widehat{\bar{x}}_I - x_I^*\big\|^2 + \underbrace{\sum_I \widehat{\bar{a}}_I \operatorname{Tr}(\widehat{\Sigma}_I)}_{\geq 0} \geq \sum_I a_I^* \big\|\widehat{\bar{x}}_I - x_I^*\big\|^2 + O\left(\big\|\Delta\widehat{\bar{a}}\big\|_1 \big\|\Delta\widehat{\bar{x}}\big\|^2\right)
$$

and transform the remaining quadratic term, using that $2\langle a,b\rangle \le 2\|a\|\|b\| \le \|a\|^2 + \|b\|^2$, as

$$\sum_I a_I^* \|\widehat{\bar{x}}_I - x_I^*\|^2 \ge \sum_I a_I^* \left(-2\left\langle \Delta\widehat{\bar{x}}_I, \Delta X_I\right\rangle - \|\Delta X_I\|^2\right) = -2\sum_I a_I^* \left\langle \Delta\widehat{\bar{x}}_I, \Delta X_I\right\rangle + O\left(\|\Delta X\|^2\right).$$

Thus, combining the above bounds, we have

$$\begin{aligned} F_{n,m^*}(\widehat{a},\widehat{x},B,Y) \ge \rho &+ \Delta\widehat{\bar{a}}^\top M^* \Delta B \\ &+ (a^*\odot\Delta\widehat{\bar{x}})^\top \partial_x M^* \Delta B + \Delta\widehat{\bar{a}}^\top \partial_y M^*(\Delta Y \odot b^*) + (a^*\odot\Delta\widehat{\bar{x}})^\top \partial^2_{xy} M^*(\Delta Y\odot b^*) \\ &- \frac{\sigma_{\min}}{2}\sum_I a_I^* \left\langle \Delta\widehat{\bar{x}}_I, \Delta X_I\right\rangle \quad + O\left(\left\|\Delta\widehat{Z}\right\|^3 + \|\delta z\|^2 + \|\delta z\|\left(\widehat{\bar{a}}_0 + V_{\rm pos}(\widehat{a},\widehat{x})\right)\right). \end{aligned}$$

One can derive the analogous upper bound for $F_{n^*,m}(A,X,\widehat{b},\widehat{y})$. Combining the two, we obtain

$$\begin{aligned} &F_{n,m^*}(\widehat{a},\widehat{x},B,Y) - F_{n^*,m}(A,X,\widehat{b},\widehat{y}) \\ &\ge -\begin{pmatrix}\Delta A\\ \Delta X\\ \Delta B\\ \Delta Y\end{pmatrix}^\top \begin{bmatrix} 0 & 0 & M^* & \partial_y M^* b^* \\ 0 & a^*\frac{\sigma_{\min}}{2}\,\mathrm{id} & a^*\partial_x M^* & a^*\partial^2_{xy} M^* b^* \\ -(M^*)^\top & -(a^*\partial_x M^*)^\top & 0 & 0 \\ -(\partial_y M^* b^*)^\top & -(a^*\partial^2_{xy} M^* b^*)^\top & 0 & b^*\frac{\sigma_{\min}}{2}\,\mathrm{id}\end{bmatrix}\begin{pmatrix}\Delta\widehat{\bar{a}}\\ \Delta\widehat{\bar{x}}\\ \Delta\widehat{\bar{b}}\\ \Delta\widehat{\bar{y}}\end{pmatrix} \\ &\quad + O\left(\left\|\Delta\widehat{Z}\right\|^3 + \|\delta z\|\left(\widehat{\bar{w}}_0 + \sigma V_{\rm pos}(\widehat{z})\right) + \|\delta z\|^2\right). \end{aligned}$$

It just remains to note that $\|\delta z\|\left(\widehat{\bar{w}}_0 + \sigma V_{\rm pos}(\widehat{z})\right) \le \frac{1}{2}\|\delta z\|^2 + \frac{1}{2}\left(\widehat{\bar{w}}_0 + \sigma V_{\rm pos}(\widehat{z})\right)^2$. ◀

H Proof of convergence of CP-MP in the exact-parametrization case

In this section we prove Proposition 6, which states that CP-MP converges under the same conditions and with the same rate as CP-PP. The proof essentially combines the convergence result for CP-PP with the general fact that the Mirror Prox and Proximal Point updates coincide up to order-3 terms, a consequence of the two following lemmas.

▶ **Lemma 49.** *Let $A \in \mathbb{R}^{m\times d}$ and $b\in\mathbb{R}^m$ for some $m<d$, with A having full rank, and denote $\mathcal{Z} = \{z\in\mathbb{R}^d; Az=b\}$. Define the semi-norm*

$$\forall\, v\in\mathbb{R}^d,\ \|v\|_{*\mathcal{Z}} = \max_{\substack{\|\delta\|\le 1\\ A\delta=0}} \langle\delta, v\rangle$$

where $\|\delta\|$ is the usual Euclidean norm. Consider some function $F:\mathbb{R}^d\to\mathbb{R}$ with Lipschitz-continuous second-order differentials. Let $D:\mathcal{Z}\times\mathcal{Z}\to\mathbb{R}_+$ such that for any z^0 in some subset $\mathcal{Z}_0\subset\mathcal{Z}$,

- *$D^0(z) \coloneqq D(z,z^0)$ is strongly convex and smooth over $z\in\mathcal{Z}$, and the constants do not depend on z^0.*
- *There exist $H = H_{ij} \succ 0$ and $K = K_{ijk}$ an order-3 symmetric tensor (that depend on z^0) such that $\nabla D^0(z) = H(z-z^0) + \frac{K}{2}(z-z^0)^2 + O\left(\|z-z^0\|^3\right)$. That is, using Einstein's summation notation,*

$$[\nabla D^0(z)]_i = H_{ij}(z-z^0)^j + \frac{1}{2}K_{ijk}(z-z^0)^j(z-z^0)^k + O\left(\|z-z^0\|^3\right).$$

To be clear, we assume that $\sigma_{\min}(H)^{-1}$ and the norms of H and K, as well as the constant hidden in the $O(\cdot)$ in the above equation, are all bounded by a constant that does not depend on z^0. Consider the Mirror Prox (MP) update $z^{k+1}_{\rm MP} = \mathrm{MP}(z^k;\eta)$ defined by

$$\begin{aligned} \widehat{z} &= \operatorname*{arg\,min}_{Az=b} \langle\nabla F(z^k), z\rangle + \frac{1}{\eta}D(z,z^k) \\ z^{k+1}_{\rm MP} &= \operatorname*{arg\,min}_{Az=b} \langle\nabla F(\widehat{z}), z\rangle + \frac{1}{\eta}D(z,z^k). \end{aligned}$$

Then, if $z^k\in\mathcal{Z}_0$ (and for η small enough),

$$\begin{aligned} z^{k+1}_{\rm MP} - z^k = -\eta H^{-1}P\nabla F(z^k) + \eta^2 H^{-1}P\cdot&\left(\nabla^2 F(z^k)H^{-1}P\nabla F(z^k) - \frac{1}{2}K\left[H^{-1}P\nabla F(z^k)\right]^2\right) \\ &+ O\left(\eta^3\|\nabla F(z^k)\|^2_{*\mathcal{Z}}\right) \end{aligned}$$

where $O(\cdot)$ hides only the aforementioned constant and the smoothness constants of F, and where

$$P = I - A^\top \left[AH^{-1}A^\top\right]^{-1} AH^{-1}.$$

▶ **Lemma 50.** *Under the same conditions and using the same notations as in the previous lemma, the Proximal Point (PP) update $z_{\mathrm{PP}}^{k+1} = \mathrm{PP}(z^k;\eta)$ defined by*

$$z_{\mathrm{PP}}^{k+1} = \operatorname*{arg\,min}_{Az=b} F(z) + \frac{1}{\eta} D(z, z^k)$$

satisfies

$$z_{\mathrm{PP}}^{k+1} - z^k = -\eta H^{-1} P \nabla F(z^k) + \eta^2 H^{-1} P \cdot \left(\nabla^2 F(z^k) H^{-1} P \nabla F(z^k) - \frac{1}{2} K \left[H^{-1} P \nabla F(z^k) \right]^2 \right) + O\left(\eta^3 \left\| \nabla F(z^k) \right\|_{*\mathcal{Z}}^3\right).$$

That is, $z_{\mathrm{PP}}^{k+1} - z_{\mathrm{MP}}^{k+1} = O\left(\eta^3 \left\| \nabla F(z^k) \right\|_{*\mathcal{Z}}^3\right)$.

▶ Remark 51. One can check that $P^2 = P$ and that $H^{-1}P$ is symmetric, i.e., P is a projection which is orthogonal for $\langle \cdot, \cdot \rangle_{H^{-1}}$. Furthermore, $PA^\top = 0$, i.e., $P^\top$ projects onto the kernel of A, and so the semi-norm $\|Pv\| = \max_{\|\delta\| \le 1} \langle \delta, Pv \rangle$ is dominated by $\|v\|_{*\mathcal{Z}}$ (the operator norm of $P^\top$ being bounded by a constant).

▶ Remark 52. In the Euclidean case where $D(\cdot,\cdot) = \frac{1}{2}\|\cdot - \cdot\|^2$, we recover the formulas from [29, Prop. 2]. In the case where the divergence function $D(\cdot,\cdot)$ is a Bregman divergence, we provide a finer (order-2) expansion than [1, Prop. 1] (which was order-1).

In the next subsection we show how to prove Proposition 6 using (a min-max version of) the two above lemmas, and in the two following subsections we prove Lemma 49 and Lemma 50 respectively.

H.1 Proof of Proposition 6

In this subsection we assume the exact-parametrization setting, i.e. $n = n^*, m = m^*$, and we use the notations introduced in Appendix A.

Preliminaries

For all $z, \widehat{z} \in \Delta_n \times \mathcal{X}^n \times \Delta_m \times \mathcal{Y}^m$, denote

$$D\left((a,x),(\widehat{a},\widehat{x})\right) = D(a,\widehat{a}) + \frac{\eta}{2\sigma} \sum_i \widehat{a}_i \|x_i - \widehat{x}_i\|^2,$$

similarly for $D\big((b,y),(\widehat{b},\widehat{y})\big)$, and $D(z,\widehat{z}) = D\big((a,x),(\widehat{a},\widehat{x})\big) + D\big((b,y),(\widehat{b},\widehat{y})\big)$; in particular we have $V(\widehat{z}) = D(z^*,\widehat{z})$. Also let

$$\|z - \widehat{z}\|^2 = \|a - \widehat{a}\|_1^2 + \max_i \|x_i - \widehat{x}_i\|^2 + \|b - \widehat{b}\|_1^2 + \max_j \|y_j - \widehat{y}_j\|^2$$

and recall from Claim 21 that divergence and squared norm are equivalent in the sense that, if $a_i, \widehat{a}_i, b_j, \widehat{b}_j \ge c = \Theta(1)$ for all i, j, then $\|z - \widehat{z}\|^2 \lesssim D(z,\widehat{z}) \lesssim \|z - \widehat{z}\|^2$.

Furthermore, denote $g(z) = \begin{pmatrix} \nabla_a \\ \nabla_x \\ -\nabla_b \\ -\nabla_y \end{pmatrix} F_{n,m}(z)$ and define the semi-norm

$$\|v\|_{*\mathcal{Z}} = \max_{\substack{\delta z \in \mathbb{R}^n \times \mathcal{X}^n \times \mathbb{R}^m \times \mathcal{Y}^m \\ \|\delta z\| \le 1 \\ \mathbf{1}^\top \delta a = \mathbf{1}^\top \delta b = 0}} \langle \delta z, v \rangle .$$

By definition, z^* is a stationary point of the vector flow $g(z)$ under the constraint $z \in \Delta_n \times \mathcal{X}^n \times \Delta_m \times \mathcal{Y}^m$, and z^* belongs to the relative interior of that domain. So by smoothness of $F_{n,m}$,

$$\|g(z)\|_{*\mathcal{Z}} = \|g(z) - g(z^*)\|_{*\mathcal{Z}} \lesssim \|z - z^*\|.$$

Comparing the CP-MP and CP-PP updates

Starting from z^k, the CP-MP update $\widetilde{z}^{k+1}$ is given by

$$\widehat{z} = \operatorname*{arg\,min}_{\substack{a\in\Delta_n \\ x\in\mathcal{X}^n}} \operatorname*{arg\,max}_{\substack{b\in\Delta_m \\ y\in\mathcal{Y}^m}} \ \langle g(z^k), z\rangle + \frac{1}{\eta}\left[D\left((a,x),(a^k,x^k)\right) - D\left((b,y),(b^k,y^k)\right)\right]$$

$$\widetilde{z}^{k+1} = \operatorname*{arg\,min}_{\substack{a\in\Delta_n \\ x\in\mathcal{X}^n}} \operatorname*{arg\,max}_{\substack{b\in\Delta_m \\ y\in\mathcal{Y}^m}} \ \langle g(\widehat{z}), z\rangle + \frac{1}{\eta}\left[D\left((a,x),(a^k,x^k)\right) - D\left((b,y),(b^k,y^k)\right)\right]$$

and the CP-PP update z^{k+1} by

$$z^{k+1} = \operatorname*{arg\,min}_{\substack{a\in\Delta_n \\ x\in\mathcal{X}^n}} \operatorname*{arg\,max}_{\substack{b\in\Delta_m \\ y\in\mathcal{Y}^m}} \ F_{n,m}(z) + \frac{1}{\eta}\left[D\left((a,x),(a^k,x^k)\right) - D\left((b,y),(b^k,y^k)\right)\right].$$

It is not hard to adapt the proofs of Lemma 49 and Lemma 50 to cover min-max updates of these forms, as $D(a,\widehat{a}) = +\infty$ for a on the relative boundary of Δ_n so that the constraints reduce to $\mathbf{1}^\top a = \mathbf{1}^\top b = 1$. Furthermore, it is not hard to show that $\|\Delta\widetilde{z}^{k+1}\|, \|\Delta z^{k+1}\| \lesssim \|\Delta z^k\| + \eta$, so in particular by choosing r_0 and η small enough, we may assume $z^k, z^{k+1}, \widetilde{z}^{k+1} \in \mathcal{Z}_0 = \{z;\ \min_i a_i, \min_j b_j \geq c\}$, and the assumptions of Lemma 49 and Lemma 50 on $D(\,\cdot\,, z^k)$ are satisfied. Thus we have

$$\|\widetilde{z}^{k+1} - z^{k+1}\| \lesssim \eta^3 \|g(z^k)\|_{*\mathcal{Z}} \lesssim \eta^3 \|\Delta z^k\|.$$

Let us convert the above bound on $\|\widetilde{z}^{k+1} - z^{k+1}\|$ into a bound on $|V(z^{k+1}) - V(\widetilde{z}^{k+1})|$. Since we can assume $z^{k+1}, \widetilde{z}^{k+1} \in \mathcal{Z}_0$, by using that $h : s \mapsto s\log s - s + 1$ is c-smooth over $[c,1]$ one easily checks that, denoting $w = (a,b)$ and $p = (x,y)$,

$$\begin{aligned} D(w^*, w^{k+1}) - D(w^*, \widetilde{w}^{k+1}) &= D(\widetilde{w}^{k+1}, w^{k+1}) - \sum_i \left(h'(w_i^{k+1}) - h'(\widetilde{w}_i^{k+1})\right)\left(w_i^* - \widetilde{w}_i^{k+1}\right) \\ &= O\left(\|w^{k+1} - \widetilde{w}^{k+1}\|^2 + \|w^{k+1} - \widetilde{w}^{k+1}\|\|\Delta\widetilde{w}^{k+1}\|\right) \end{aligned}$$

by Bregman three-point identity; and similarly, for each i

$$\|p_i^* - p_i^{k+1}\|^2 - \|p_i^* - \widetilde{p}_i^{k+1}\|^2 = O\left(\|p_i^{k+1} - \widetilde{p}_i^{k+1}\|^2 + \|p_i^{k+1} - \widetilde{p}_i^{k+1}\|\|\Delta\widetilde{p}_i^{k+1}\|\right).$$

Now it is not hard to show (in fact this is just (38) below) that $\|\widetilde{z}^{k+1} - z^k\| \lesssim \eta\|\nabla g(z^k)\|$ and so $\|\Delta\widetilde{z}^{k+1}\| \lesssim \|\Delta z^k\|$. So

$$|V(z^{k+1}) - V(\widetilde{z}^{k+1})| = |D(z^*, z^{k+1}) - D(z^*, \widetilde{z}^{k+1})| \lesssim \eta^3\|\Delta z^k\|^2 \lesssim \eta^3 D(z^*, z^k) = \eta^3 V(z^k).$$

Proof conclusion

In the proof of Theorem 5 we showed that

$$V(z^{k+1}) \leq V(z^k) - (C/2)\eta^2 V(z^{k+1})$$

for some C dependent only on $(f, \mathcal{X}, \mathcal{Y})$ and Γ_0, for η, σ small enough and r_0 small enough (depending on η, σ). So

$$V(\widetilde{z}^{k+1}) \leq V(z^k) - (C/2)\eta^2 V(\widetilde{z}^{k+1}) + O\left(\eta^3 V(z^k)\right),$$

and we can conclude to the local exponential convergence of the sequence of CP-MP iterates in exactly the same way as for Theorem 5.

H.2 Proof of Lemma 49

To lighten notation and since we focus on a single iteration, instead of "$z_{\mathrm{MP}}^{k+1} = \mathrm{MP}(z^k;\eta)$" we will consider $\mathrm{MP}(z^0;\eta) = z^2$ with

$$z^1 = \operatorname*{arg\,min}_{Az=b} \left\langle \nabla F(z^0), z\right\rangle + \frac{1}{\eta} D(z, z^0) \tag{U1}$$

$$z^2 = \operatorname*{arg\,min}_{Az=b} \left\langle \nabla F(z^1), z\right\rangle + \frac{1}{\eta} D(z, z^0). \tag{U2}$$

The goal is to get an order-2 expansion for $\delta z \coloneqq z^2 - z^0$.

Also to lighten notation, we will write $\|\nabla F(z^0)\|$ for $\|\nabla F(z^0)\|_{*\mathcal{Z}}$ (and similarly for $\nabla F(z^1)$).

First estimates

By Lagrangian duality, there exist $\lambda^1, \lambda^2 \in \mathbb{R}^m$ such that

$$\nabla F(z^0) + \frac{1}{\eta}\nabla D^0(z^1) - A^\top \lambda^1 = 0 \tag{S1}$$

$$\text{and} \quad \nabla F(z^1) + \frac{1}{\eta}\nabla D^0(z^2) - A^\top \lambda^2 = 0. \tag{S2}$$

As a first consequence, since $(z^1 - z^0)^\top A^\top = 0$, we get that

$$(z^1 - z^0)^\top \left[\nabla F(z^0) + \frac{1}{\eta}\nabla D^0(z^1)\right] = 0$$

$$\frac{\mu}{2}\|z^1 - z^0\|^2 \le (z^1 - z^0)^\top \left(\nabla D^0(z^1) - \nabla D^0(z^0)\right) \le \eta\|z^1 - z^0\|\|\nabla F(z^0)\|$$

$$\|z^1 - z^0\| \lesssim \eta\|\nabla F(z^0)\|$$

and also consequently $\|\nabla F(z^1)\| \le \|\nabla F(z^0)\| + O(\|z^1 - z^0\|) \lesssim \|\nabla F(z^0)\|$. Similarly since $(z^2 - z^0)^\top A^\top = 0$,

$$\|z^2 - z^0\| \lesssim \eta\|\nabla F(z^1)\| \lesssim \eta\|\nabla F(z^0)\|. \tag{38}$$

An order-1 expansion for the first update (U1)

Next we want to get an explicit approximate expression for $\nabla F(z^1)$ only in terms of z^0, based on the expansion

$$\nabla F(z^1) = \nabla F(z^0) + \nabla^2 F(z^0)(z^1 - z^0) + O\left(\|z^1 - z^0\|)^2\right).$$

For this we want to get an explicit approximate expression for $z^1 - z^0$.

From (S1) and an order-1 expansion of ∇D^0, we have

$$\eta\nabla F(z^0) + H(z^1 - z^0) - \eta A^\top \lambda^1 = O(\|z^1 - z^0\|^2)$$

$$z^1 - z^0 = \eta H^{-1}\left(-\nabla F(z^0) + A^\top \lambda^1\right) + O(\|z^1 - z^0\|^2)$$

and this will get us an expression of $z^1 - z^0$ of the correct order for this paragraph's purpose. It remains to identify $A^\top \lambda^1$. An approximate expression of it can be obtained simply by

$$\frac{1}{\eta}A(z^1 - z^0) = 0 = AH^{-1}\left(-\nabla F(z^0) + A^\top \lambda^1\right) + \frac{1}{\eta}O(\|z^1 - z^0\|^2)$$

$$AH^{-1}A^\top \lambda^1 = AH^{-1}\nabla F(z^0) + \frac{1}{\eta}O(\|z^1 - z^0\|^2)$$

$$\lambda^1 = \left[AH^{-1}A^\top\right]^{-1} AH^{-1}\nabla F(z^0) + \frac{1}{\eta}O(\|z^1 - z^0\|^2)$$

since $AH^{-1}A^\top$ is invertible as an $m \times m$ product of full-rank matrices. Thus we get

$$z^1 - z^0 = -\eta H^{-1}\underbrace{\left(I - A^\top\left[AH^{-1}A^\top\right]^{-1} AH^{-1}\right)}_{=P}\nabla F(z^0) + O(\|z^1 - z^0\|^2). \tag{P}$$

To recap, we showed that

$$\nabla F(z^1) = \nabla F(z^0) - \eta\nabla^2 F(z^0)H^{-1}P\nabla F(z^0) + O(\|z^1 - z^0\|^2). \tag{39}$$

An order-2 expansion of δz (the second update (U2))

Recall that we denote $\delta z = z^2 - z^0$. From (S2) and an order-1 expansion of ∇D^0, by exactly the same calculations as in the previous paragraph,

$$\lambda^2 = \left[AH^{-1}A^\top\right]^{-1} AH^{-1}\nabla F(z^1) + \frac{1}{\eta}O(\|\delta z\|^2)$$

$$\text{and} \quad \delta z = -\eta H^{-1}P\nabla F(z^1) + O(\|\delta z\|^2). \tag{40}$$

However this is not precise enough for our goal, as the error is order-2 in η.

From (S2) and an order-2 expansion of ∇D^0, we have

$$\eta\nabla F(z^1) + H_{ij}\delta z^j + \frac{1}{2}K_{ijk}\delta z^j\delta z^k - \eta A^\top\lambda^2 = O(\|\delta z\|^3)$$
$$\left(H_{ij} + \frac{1}{2}K_{ijk}\delta z^k\right)\delta z^j = -\eta\nabla F(z^1) + \eta A^\top\lambda^2 + O(\|\delta z\|^3)$$

where unmarked vectors are implicitly indexed by subscript i. Denoting for concision

$$G_{ij} \coloneqq \frac{1}{2}K_{ijk}\delta z^k \qquad \text{and} \qquad v \coloneqq H^{-1}\left(\nabla F(z^1) - A^\top\lambda^2\right)$$

(note that G is symmetric since K is), the above equation writes

$$(H+G)\delta z = -\eta H v + O(\|\delta z\|^3)$$
$$(I + H^{-1}G)\delta z = -\eta v + O(\|\delta z\|^3).$$

Now $\|G\| \lesssim \|\delta z\| \lesssim \eta\|\nabla F(z^0)\|$ so $(I+H^{-1}G)^{-1} = I - H^{-1}G + O\left(\eta^2\|\nabla F(z^0)\|^2\right)$, and by our order-1 estimates from (40) we have $\|v\| \lesssim \|P\nabla F(z^1)\| + \frac{1}{\eta}\|\delta z\|^2 \lesssim \|\nabla F(z^0)\|$ using that $\|P\bullet\| \lesssim \|\bullet\|_{*\mathcal{Z}}$ by Remark 51. So

$$\delta z = -\eta(I - H^{-1}G)v + O(\eta^3\|\nabla F(z^0)\|^3). \tag{41}$$

It remains to estimate v, and namely the $A^\top\lambda^2$ term, up to $O\left(\eta^2\|\nabla F(z^0)\|\right)$ error terms. To do this just write

$$A\delta z = 0 = -\eta A(I - H^{-1}G)v + O\left(\eta^3\|\nabla F(z^0)\|^3\right)$$
$$A(I - H^{-1}G)v = O\left(\eta^2\|\nabla F(z^0)\|^3\right).$$

We already have an estimate of G of the correct order thanks to (40):

$$G_{ij} = \frac{1}{2}K_{ijk}\delta z^k$$
$$= \underbrace{\frac{1}{2}K_{ijk}\left[-\eta H^{-1}P\nabla F(z^1)\right]^k}_{=:\widetilde{G}_{ij}} + O(\|\delta z\|^2). \tag{42}$$

Since $\|v\| \lesssim \|\nabla F(z^0)\|$ we just need to solve for $A^\top\lambda^2$ in $A(I - H^{-1}\widetilde{G})v = O\left(\eta^2\|\nabla F(z^0)\|^3\right)$:

$$A(I - H^{-1}\widetilde{G})H^{-1}\left(\nabla F(z^1) - A^\top\lambda^2\right) = O\left(\eta^2\|\nabla F(z^0)\|^3\right)$$
$$A(I - H^{-1}\widetilde{G})H^{-1}A^\top\lambda^2 = A(I - H^{-1}\widetilde{G})H^{-1}\nabla F(z^1) + O\left(\eta^2\|\nabla F(z^0)\|^3\right)$$
$$\lambda^2 = \left[A(I - H^{-1}\widetilde{G})H^{-1}A^\top\right]^{-1}A(I - H^{-1}\widetilde{G})H^{-1}\nabla F(z^1)$$
$$+ O\left(\eta^2\|\nabla F(z^0)\|^3\right).$$

Since $\|\widetilde{G}\| \lesssim \eta\|\nabla F(z^0)\|$, we have the expansion

$$A(I - H^{-1}\widetilde{G})H^{-1}A^\top = AH^{-1}A^\top - AH^{-1}\widetilde{G}H^{-1}A^\top$$
$$= AH^{-1}A^\top\left(I - \left[AH^{-1}A^\top\right]^{-1}AH^{-1}\widetilde{G}H^{-1}A^\top\right)$$
$$\left[A(I - H^{-1}\widetilde{G})H^{-1}A^\top\right]^{-1} = \left(I - \left[AH^{-1}A^\top\right]^{-1}AH^{-1}\widetilde{G}H^{-1}A^\top\right)^{-1}\left[AH^{-1}A^\top\right]^{-1}$$
$$= \left(I + \left[AH^{-1}A^\top\right]^{-1}AH^{-1}\widetilde{G}H^{-1}A^\top\right)\left[AH^{-1}A^\top\right]^{-1} + O\left(\eta^2\|\nabla F(z^0)\|^2\right)$$
$$= \left(\left[AH^{-1}A^\top\right]^{-1} + \left[AH^{-1}A^\top\right]^{-1}AH^{-1}\widetilde{G}H^{-1}A^\top\left[AH^{-1}A^\top\right]^{-1}\right)$$
$$+ O\left(\eta^2\|\nabla F(z^0)\|^2\right).$$

Substituting and expanding the product, and neglecting the terms in $\|\widetilde{G}\|^2\|\nabla F(z^0)\|$, we get the following expression for λ^2; as a sanity-check, when we neglect the terms in $\|\widetilde{G}\|\|\nabla F(z^0)\|$, we recover the estimate from (40).

$$\begin{aligned}
&\lambda^2 + O\left(\eta^2\|\nabla F(z^0)\|^3\right)\\
&= \left(\left[AH^{-1}A^\top\right]^{-1} + \left[AH^{-1}A^\top\right]^{-1} AH^{-1}\widetilde{G}H^{-1}A^\top\left[AH^{-1}A^\top\right]^{-1}\right) A(I - H^{-1}\widetilde{G})H^{-1}\nabla F(z^1)\\
&= \left(\left[AH^{-1}A^\top\right]^{-1} A + \left[AH^{-1}A^\top\right]^{-1} AH^{-1}\widetilde{G}H^{-1}A^\top\left[AH^{-1}A^\top\right]^{-1} A - \left[AH^{-1}A^\top\right]^{-1} AH^{-1}\widetilde{G}\right) H^{-1}\nabla F(z^1)\\
&= \left(\left[AH^{-1}A^\top\right]^{-1} A - \left[AH^{-1}A^\top\right]^{-1} AH^{-1}\widetilde{G}\left(I - H^{-1}A^\top\left[AH^{-1}A^\top\right]^{-1} A\right)\right) H^{-1}\nabla F(z^1)\\
&= \left(\left[AH^{-1}A^\top\right]^{-1} AH^{-1} - \left[AH^{-1}A^\top\right]^{-1} AH^{-1}\widetilde{G}H^{-1}\left(I - A^\top\left[AH^{-1}A^\top\right]^{-1} AH^{-1}\right)\right)\nabla F(z^1)\\
&= \left(\left[AH^{-1}A^\top\right]^{-1} AH^{-1} - \left[AH^{-1}A^\top\right]^{-1} AH^{-1}\widetilde{G}H^{-1}P\right)\nabla F(z^1).
\end{aligned}$$

Substituting, we get the following expression for $v = H^{-1}\left(\nabla F(z^1) - A^\top\lambda^2\right)$:

$$\begin{aligned}
-Hv &= A^\top\lambda^2 - \nabla F(z^1)\\
&= \left(A^\top\left[AH^{-1}A^\top\right]^{-1} AH^{-1} - A^\top\left[AH^{-1}A^\top\right]^{-1} AH^{-1}\widetilde{G}H^{-1}P - I\right)\nabla F(z^1) + O\left(\eta^2\|\nabla F(z^0)\|^3\right)\\
&= \left(-P - A^\top\left[AH^{-1}A^\top\right]^{-1} AH^{-1}\widetilde{G}H^{-1}P\right)\nabla F(z^1) + O\left(\eta^2\|\nabla F(z^0)\|^3\right)\\
v &= H^{-1}\left(I + A^\top\left[AH^{-1}A^\top\right]^{-1} AH^{-1}\widetilde{G}H^{-1}\right)P\nabla F(z^1) + O\left(\eta^2\|\nabla F(z^0)\|^3\right).
\end{aligned}$$

Substituting into (41) and using $\|G - \widetilde{G}\| \lesssim \eta^2\|\nabla F(z^0)\|^2$, we get the following expression for δz:

$$\begin{aligned}
\delta z &= -\eta(I - H^{-1}\widetilde{G})v + O\left(\eta^3\|\nabla F(z^0)\|^3\right)\\
&= -\eta(I - H^{-1}\widetilde{G})H^{-1}\left(I + A^\top\left[AH^{-1}A^\top\right]^{-1} AH^{-1}\widetilde{G}H^{-1}\right)P\nabla F(z^1) + O\left(\eta^3\|\nabla F(z^0)\|^3\right)\\
&= -\eta H^{-1}(I - \widetilde{G}H^{-1})\left(I + A^\top\left[AH^{-1}A^\top\right]^{-1} AH^{-1}\widetilde{G}H^{-1}\right)P\nabla F(z^1) + O\left(\eta^3\|\nabla F(z^0)\|^3\right)\\
&= -\eta H^{-1}\left(I - \left(I - A^\top\left[AH^{-1}A^\top\right]^{-1} AH^{-1}\right)\widetilde{G}H^{-1}\right)P\nabla F(z^1) + O\left(\eta^3\|\nabla F(z^0)\|^3\right)\\
&= -\eta H^{-1}\left(I - P\widetilde{G}H^{-1}\right)P\nabla F(z^1) + O\left(\eta^3\|\nabla F(z^0)\|^3\right).
\end{aligned}$$

Expanding and recalling the definition of $\widetilde{G}$ (42), we get that

$$\begin{aligned}
\left[\widetilde{G}H^{-1}P\nabla F(z^1)\right]_i &= \widetilde{G}_{ij}\left[H^{-1}P\nabla F(z^1)\right]^j\\
&= \frac{1}{2}K_{ijk}\left[-\eta H^{-1}P\nabla F(z^1)\right]^k\left[H^{-1}P\nabla F(z^1)\right]^j\\
&= -\eta\frac{1}{2}K_{ijk}\left[H^{-1}P\nabla F(z^1)\right]^k\left[H^{-1}P\nabla F(z^1)\right]^j
\end{aligned}$$

or in shorthand, $\widetilde{G}H^{-1}P\nabla F(z^1) = -\eta\frac{1}{2}K\left[H^{-1}P\nabla F(z^1)\right]^2$.

So finally, we can write δz as

$$\delta z = -\eta H^{-1}P\nabla F(z^1) - \eta^2 H^{-1}P\cdot\frac{1}{2}K\left[H^{-1}P\nabla F(z^1)\right]^2 + O\left(\eta^3\|\nabla F(z^0)\|^3\right). \tag{43}$$

To make the expression of δz fully explicit and conclude the analysis, let us substitute the expression of $\nabla F(z^1)$ from (39). Note that doing so makes us lose an order of precision in $\|\nabla F(z^0)\|$ for the first term.

$$\begin{aligned}
\delta z &= -\eta H^{-1}P\left[\nabla F(z^0) - \eta\nabla^2 F(z^0)H^{-1}P\nabla F(z^0)\right] + \eta^2 H^{-1}P\cdot\frac{1}{2}K\left[H^{-1}P\nabla F(z^0)\right]^2 + O\Big(\eta^3\|\nabla F(z^0)\|^2\Big)\\
&= -\eta H^{-1}P\nabla F(z^0) + \eta^2 H^{-1}P\cdot\left(\nabla^2 F(z^0)H^{-1}P\nabla F(z^0) - \frac{1}{2}K\left[H^{-1}P\nabla F(z^0)\right]^2\right) + O\Big(\eta^3\|\nabla F(z^0)\|^2\Big).
\end{aligned} \tag{44}$$

H.3 Proof of Lemma 50

We keep the notations of the previous section, and this time we are interested in getting a similar Taylor expansion for the Proximal Point (PP) update $\mathrm{PP}(z^0;\eta) = z^\infty$ defined by

$$z^\infty = \operatorname*{arg\,min}_{Az=b} F(z) + \frac{1}{\eta} D(z, z^0). \tag{U∞}$$

The goal is to get an order-2 expansion for $\delta z \coloneqq z^\infty - z^0$.

Also again to lighten notation, we will write $\left\|\nabla F(z^0)\right\|$ for $\left\|\nabla F(z^0)\right\|_{*\mathcal{Z}}$ and similarly for $\nabla F(z^\infty)$.

By Lagrangian duality, there exists $\lambda \in \mathbb{R}^m$ such that

$$\nabla F(z^\infty) + \frac{1}{\eta}\nabla D^0(z^\infty) - A^\top \lambda = 0. \tag{S∞}$$

By similar calculations as for MP, we get that

$$\|\delta z\| \lesssim \eta \left\|\nabla F(z^\infty)\right\| \lesssim \eta \left\|\nabla F(z^0)\right\|.$$

An order-1 expansion of δz

From (S∞) and an order-1 expansion of ∇D^0, we have

$$\begin{aligned}
\eta\nabla F(z^\infty) + H\delta z - \eta A^\top\lambda &= O(\|\delta z\|^2) \\
\delta z &= \eta H^{-1}\left(-\nabla F(z^\infty) + A^\top\lambda\right) + O(\|\delta z\|^2).
\end{aligned}$$

We can get an approximate expression of $A^\top\lambda$ by

$$\begin{aligned}
\frac{1}{\eta} A\delta z = 0 &= AH^{-1}\left(-\nabla F(z^\infty) + A^\top\lambda\right) + \frac{1}{\eta} O(\|\delta z\|^2) \\
AH^{-1}A^\top\lambda &= AH^{-1}\nabla F(z^\infty) + \frac{1}{\eta} O(\|\delta z\|) \\
\lambda &= \left[AH^{-1}A^\top\right]^{-1} AH^{-1}\nabla F(z^\infty) + \frac{1}{\eta} O(\|\delta z\|^2).
\end{aligned}$$

Thus we get

$$\delta z = -\eta H^{-1} \underbrace{\left(I - A^\top\left[AH^{-1}A^\top\right]^{-1} AH^{-1}\right)}_{=P} \nabla F(z^\infty) + O(\|\delta z\|^2).$$

This limited-order expansion is sufficient for us to get an approximate expression of $\nabla F(z^\infty)$ in terms of z^0. Indeed,

$$\begin{aligned}
\nabla F(z^\infty) &= \nabla F(z^0) + \nabla^2 F(z^0)\delta z + O(\|\delta z\|^2) \\
&= \nabla F(z^0) - \eta\nabla^2 F(z^0) H^{-1} P \nabla F(z^\infty) + O(\|\delta z\|^2) \\
\left(I + \eta\nabla^2 F(z^0) H^{-1} P\right)\nabla F(z^\infty) &= \nabla F(z^0) + O(\|\delta z\|^2) \\
\nabla F(z^\infty) &= \left(I - \eta\nabla^2 F(z^0) H^{-1} P\right)\nabla F(z^0) + O\left(\eta^2 \left\|\nabla F(z^0)\right\|\right)
\end{aligned} \tag{45}$$

using Remark 51 to control the error in the last line.

An order-2 expansion of δz.

From (S∞) and this time an order-2 expansion of ∇D^0, we have more precisely

$$\begin{aligned}
\eta\nabla F(z^\infty) + H_{ij}\delta z^j + \frac{1}{2} K_{ijk}\delta z^j \delta z^k - \eta A^\top\lambda &= O(\|\delta z\|^3) \\
\left(H_{ij} + \frac{1}{2} K_{ijk}\delta z^k\right)\delta z^j &= -\eta\nabla F(z^\infty) + \eta A^\top\lambda + O(\|\delta z\|^3)
\end{aligned} \tag{46}$$

where unmarked vectors are implicitly indexed by subscript i. Denote for concision

$$G_{ij} \coloneqq \frac{1}{2} K_{ijk}\delta z^k \qquad \text{and} \qquad v \coloneqq H^{-1}\left(\nabla F(z^\infty) - A^\top\lambda\right).$$

We can unroll the exact same calculations as in the last paragraph of Appendix H.2 with $\nabla F(z^1)$ replaced by $\nabla F(z^\infty)$, and we obtain an equivalent of (43) for PP:

$$\delta z = -\eta H^{-1} P \nabla F(z^\infty) - \eta^2 H^{-1} P \cdot \frac{1}{2} K \left[H^{-1} P \nabla F(z^\infty)\right]^2 + O\left(\eta^3 \left\|\nabla F(z^0)\right\|^3\right).$$

Since the expression of $\nabla F(z^\infty)$ in terms of z^0 for PP (45) is exactly the same as the one of $\nabla F(z^1)$ for MD (39), the very last step of the calculations is also the same, and we get

$$\delta z = -\eta H^{-1} P \nabla F(z^0) + \eta^2 H^{-1} P \cdot \left(\nabla^2 F(z^0) H^{-1} P \nabla F(z^0) - \frac{1}{2} K \left[H^{-1} P \nabla F(z^0)\right]^2\right) + O\left(\eta^3 \left\|\nabla F(z^0)\right\|\right). \quad (47)$$

The only difference is that we lose yet another order of precision in $\left\|\nabla F(z^0)\right\|$ in the error term compared to (44).

I Proof of the main result

In this section we show in detail how our main result Theorem 2 follows from combining Proposition 7, Proposition 8 and Theorem 9.

Proof. Fix $\Gamma_0 \geq 1$. Choose η_0, σ_0 as in Theorem 9. Fix any $\eta \leq \eta_0, \sigma \leq \sigma_0$ with $\Gamma_0^{-1} \leq \frac{\sigma}{\eta} \leq \Gamma_0$. Let λ, τ as in (23), let $(\varphi_I)_I, (\psi_J)_J$ as in (13), and let V_1 and V as in (12). Let $\widetilde{C}, \widetilde{C}', \widetilde{r}$ as in Proposition 7. Let $\widetilde{\widetilde{C}}, \widetilde{\widetilde{r}}$ as in Proposition 8. Let K resp. R_0 the quantities denoted κ resp. r_0 in Theorem 9.

Let $r_0 = \min\left\{\widetilde{r}, \widetilde{\widetilde{r}}, \widetilde{C}'\left(\frac{R_0}{\Gamma_0}\right)^{5/4}\right\}$. Denote $(z^k)_k$ the CP-PP iterates and $\mu^k = \sum_{i=1}^n a_i^k \delta_{x_i^k}$, $\nu^k = \sum_{j=1}^m b_j^k \delta_{y_j^k}$. Suppose $\mathrm{NI}(\mu^0, \nu^0) \leq r_0$, then by the second part of Proposition 7,

$$\begin{aligned}
&\widetilde{C}' V_1(z^0)^{5/4} \leq \mathrm{NI}(\mu^0, \nu^0) \leq r_0 \\
&\implies V_1(z^0) \leq \left(\frac{r_0}{\widetilde{C}'}\right)^{4/5} \\
&\implies V(z^0) \leq \Gamma_0 V_1(z^0) \leq \Gamma_0 \left(\frac{r_0}{\widetilde{C}'}\right)^{4/5} \leq R_0.
\end{aligned}$$

So by Theorem 9, $V(z^k) \leq V(z^0)(1-K)^k$, and so by the first part of Proposition 7,

$$\mathrm{NI}(\mu^k, \nu^k) \leq \widetilde{C}\sqrt{V(z^k)} \leq \widetilde{C}\sqrt{V(z^0)}\left(\sqrt{1-K}\right)^k \leq \widetilde{C}\sqrt{V(z^0)}\left(1 - \frac{K}{2}\right)^k.$$

This proves the first inequality of Theorem 2 by letting $C = \widetilde{C}\sqrt{R_0}$ and $\kappa = \frac{K}{2}$.

Moreover, by Proposition 8,

$$\begin{aligned}
\mathrm{WFR}_2^2(\mu^k, \mu^*) + \mathrm{WFR}_2^2(\nu^k, \nu^*) &\leq 2V(z^k)\left(1 + \widetilde{\widetilde{C}}\frac{\eta}{\sigma}(\lambda\tau)^2\right) \\
&\leq 2V(z^0)(1-K)^k\left(1 + \widetilde{\widetilde{C}}\,\Gamma_0^{-1}(\lambda\tau)^2\right) \\
&\leq \underbrace{2R_0\left(1 + \widetilde{\widetilde{C}}\,\Gamma_0^{-1}(\lambda\tau)^2\right)}(1-\kappa)^k,
\end{aligned}$$

which proves the second inequality of Theorem 2 by letting C' be the underbraced expression. ◀