Asymptotic convergence of iterative optimization algorithms

Randal Douc, Sylvain Le Corff

Asymptotic convergence of iterative optimization algorithms

Randal Douc* and Sylvain Le Corff†

*Samovar, Télécom SudParis, Institut Polytechnique de Paris

†LPSM, Sorbonne Université, UMR CNRS 8001, 4 Place Jussieu, 75005 Paris

Abstract

This paper introduces a general framework for iterative optimization algorithms and establishes under general assumptions that their convergence is asymptotically geometric. We also prove that under appropriate assumptions, the rate of convergence can be lower bounded. The convergence is then only geometric, and we provide the exact asymptotic convergence rate. This framework allows to deal with constrained optimization and encompasses the Expectation Maximization algorithm and the mirror descent algorithm, as well as some variants such as the α-Expectation Maximization or the Mirror Prox algorithm. Furthermore, we establish sufficient conditions for the convergence of the Mirror Prox algorithm, under which the method converges systematically to the unique minimizer of a convex function on a convex compact set.

1 Introduction

The minimization of a real-valued function is the most common formulation for mathematical optimization problems. Examples of convex optimization problems in machine learning can be found for instance in [Bubeck, 2015]. For models involving missing or latent data, [Dempster et al., 1977] introduced the modern formulation of the Expectation Maximization (EM) algorithm, whose convergence has been proved under general assumptions in [Wu, 1983].

The asymptotic convergence rate of the EM algorithm has been widely studied and identified as a ratio of missing information from the very beginning [Dempster et al., 1977, Meng and Rubin, 1991, Meng and Rubin, 1993]. Since then, some links with gradient descent approaches have also been drawn, see for instance [Lange, 1995]. Among the most notable recent works, [Balakrishnan et al., 2017] provided quantitative results on the non-asymptotic convergence of the EM algorithm to local optima by considering smoothness and strong-concavity assumptions. In the particular case of exponential families, [Kunstner et al., 2021] show that the M-step is equivalent to a mirror descent update. This allows to obtain non-asymptotic linear convergence rate, which directly depends on the ratio of missing information.

In this paper, instead of casting the EM algorithm into a gradient or a mirror descent framework, we propose an extended formulation to encompass both classes of algorithms, not restricted to exponential families. Indeed, both EM and mirror descent algorithms can be defined using a bivariate function that is iteratively minimized with respect to one coordinate. Such a representation can actually describe any iterative optimization algorithm whose minimization steps are parametrized only by the current parameter estimate. This paper provides the following contributions.

- We prove under general assumptions that the convergence of such iterative optimization algorithms is asymptotically geometric, see Theorem 1. We also provide lower bounds for the rate of convergence,

that allow to prove that the convergence can be only geometric, see Theorem 2, and in some cases to establish the exact asymptotic convergence rate, see Theorem 3. We show that those assumptions are natural either in an EM or in a mirror descent framework, and that they are satisfied generically without requiring any notable technical work, in contrast with non-asymptotic results that tend to be more demanding. Regarding the EM algorithm, we retrieve the well-known ratio of missing information under even more general assumptions, as the minimization mapping is not required to be point-to-point and this framework allows to deal with constrained optimization.

- We derive results for settings with both finite and infinite data, as well as for a variant of the EM algorithm, known as the α-EM algorithm, see [Matsuyama, 2003]. However, the most significant contribution is that brought to the mirror descent framework: under mild assumptions, we prove that its convergence is asymptotically geometric. This also applies to the mirror prox variant. In a general manner, the convergence rates we exhibit are proved to be invariant to C^2-reparametrization.
- Furthermore, we prove that under general assumptions, the convergence of mirror prox is guaranteed for convex functions with a unique minimizer on a convex compact set, and that, without imposing any condition on the initialization.

This paper is organized as follows. Section 2 introduces the general iterative optimization framework we consider and shows how it encompasses classical settings such as the EM algorithm or the mirror descent algorithm. Section 3 states the main general results of this paper on asymptotic convergence rates. Sections 4-6 discuss the assumptions of Theorem 1, illustrating how they are met in those classical settings, but also in variants such as the α-EM or the mirror prox algorithm. Section 7 displays the proof of Theorem 1, and Section 8 is dedicated to the convergence of mirror prox. A discussion follows in Section 9. Additional proofs are postponed to Appendix A, using technical results listed in Appendix B and proved in the Supplementary material.

Notation Throughout this paper, $\mathrm{Spec}(\cdot)$ denotes the spectrum of a matrix and $\varrho(\cdot)$ the spectral radius. The Euclidean norm is denoted by $\|\cdot\|_2$, the spectral norm by $|||\cdot|||_2$, the Frobenius norm by $|||\cdot|||_F$, and for all symmetric positive-definite matrices S, we define the norm $\|\cdot\|_S$ by $\|x\|_S^2 := x^\top S x$. The first derivative (resp. the second) of any univariate function f is written ∂f (resp. $\partial^2 f$). For all bivariate functions $\mathcal{Q}\colon (x_1, x_2) \mapsto \mathcal{Q}_{x_1}(x_2)$ and $i, j \in \{1, 2\}$, we write $\partial_i \mathcal{Q} := \partial \mathcal{Q}/\partial x_i$ and $\partial_{ij}\mathcal{Q} := \partial^2 \mathcal{Q}/\partial x_i \partial x_j$. The maximum of two real numbers a, b is denoted by $a \vee b$. For all topological spaces E, their closure are written $\overline{\mathsf{E}}$ and their interior $\mathring{\mathsf{E}}$. Finally, $\mathrm{Conv}(\cdot)$ stands for convex hull, $\mathrm{Aff}(\cdot)$ for affine hull and $\mathrm{ri}(\cdot)$ for relative interior, i.e. the interior of a set within its affine hull.

2 General framework

Let $q \in \mathbb{N}^*$ and let $\mathcal{Q}$ be a real-valued function defined on $\mathbb{R}^q \times \mathbb{R}^q$:

$$\begin{aligned} \mathcal{Q}\colon \mathbb{R}^q \times \mathbb{R}^q &\longrightarrow \mathbb{R} \\ (\theta, \theta') &\mapsto \mathcal{Q}_\theta(\theta')\,. \end{aligned}$$

Let Θ be a subset of $\mathbb{R}^q$ and $\mathcal{M}$ be the point-to-set map defined on Θ by

$$\mathcal{M}(\theta) := \operatorname*{argmin}_{\theta' \in \Theta} \mathcal{Q}_\theta(\theta')\,.$$

In what follows, provided that $\mathcal{M}(\theta) \neq \emptyset$ for any $\theta \in \Theta$, we let $(\theta_n)_{n\in\mathbb{N}}$ be a sequence defined on Θ such that for all $n \in \mathbb{N}$,

$$\theta_{n+1} \in \mathcal{M}(\theta_n). \tag{1}$$

Example 1 (EM algorithm). Let X and Y be random variables taking values in measurable spaces $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$, respectively. Assume that the pair (X, Y) has a joint density function $p_{\theta_\star}$ with respect to a reference measure μ on $\mathcal{X} \otimes \mathcal{Y}$ that belongs to some parameterized family $\{p_\theta : \theta \in \Theta\}$. Assume also that the state variable X is latent in the sense that the model is only partially observed through the observation Y. In this case, the Expectation Maximization (EM) algorithm, as defined in [Douc et al., 2013, Appendix D.1, p.492], provides an estimate of the unknown parameter $\theta_\star$ by considering a sequence $(\theta_n)_{n\in\mathbb{N}}$ defined on Θ by

$$\theta_{n+1} \in \operatorname*{argmin}_{\theta\in\Theta} \mathcal{Q}_{\theta_n}(\theta), \tag{2}$$

where for all $\theta, \theta' \in \Theta \times \Theta$,

$$\mathcal{Q}_\theta(\theta') := -\mathbb{E}_\theta\left[\log p_{\theta'}(X, Y) \,|Y\right], \tag{3}$$

and $\mathbb{E}_\theta$ denotes the expectation under p_θ. Note that $\mathcal{Q}$ is a random function which depends on the observations we consider. For instance, in a model where $(X_i, Y_i)_{1\leqslant i\leqslant k}$ are independent and identically distributed, with k observations $(Y_i)_{1\leqslant i\leqslant k}$, we define at the **sample level**: $X = (X_1, \ldots, X_k)$ and $Y = (Y_1, \ldots, Y_k)$, and inserting in (3), we obtain up to a multiplicative constant (see [Balakrishnan et al., 2017]):

$$\mathcal{Q}^{\mathrm{samp}}_\theta(\theta') := -\frac{1}{k}\sum_{i=1}^{k}\int_{\mathsf{X}} p_\theta(x|Y_i)\log p_{\theta'}(x, Y_i)\mu(\mathrm{d}x). \tag{4}$$

In the limit of infinite data (i.e. $k \to \infty$), we define at the **population level**:

$$\mathcal{Q}^{\mathrm{pop}}_\theta(\theta') := -\int_{\mathsf{Y}}\left(\int_{\mathsf{X}} p_\theta(x|y)\log p_{\theta'}(x, y)\mu(\mathrm{d}x)\right) p_{\theta_\star}(y)\mu(\mathrm{d}y), \tag{5}$$

where $x \mapsto p_\theta(x|y)$ denotes the conditional density of X given Y when the parameter value is θ and where we assume that $(Y_i)_{i\geqslant 1}$ are iid with density $p_{\theta_\star}$. Both settings are studied in this paper, replacing $\mathcal{Q}$ in (2) by $\mathcal{Q}^{\mathrm{samp}}$ or $\mathcal{Q}^{\mathrm{pop}}$.

Example 2.1 (Mirror descent). Let C be a convex compact set of $\mathbb{R}^q$ and f be a real-valued function defined on C. The mirror descent strategy defined in [Bubeck, 2015, Chapter 4, p.296] considers a convex open set D of $\mathbb{R}^q$ such that C is contained in the closure of D and $\mathsf{C} \cap \mathsf{D} \neq \emptyset$, along with a mirror map $\Phi\colon \mathsf{D} \to \mathbb{R}$, that is,

(i) Φ is strictly convex and differentiable,

(ii) the gradient of Φ takes all possible values: $\partial\Phi(\mathsf{D}) = \mathbb{R}^q$,

(iii) the gradient of Φ diverges on the boundary of D: $\lim_{x\to\partial\mathsf{D}} \|\partial\Phi(x)\| = +\infty$.

Then, the mirror descent algorithm produces two sequences $(\theta_n)_{n\in\mathbb{N}}$ and $(\zeta_n)_{n\in\mathbb{N}}$, defined on C and D respectively by

$$\begin{cases} \partial\Phi(\zeta_{n+1}) = \partial\Phi(\theta_n) - \eta g_n, \quad \text{where } g_n \in \partial f(\theta_n), \\ \theta_{n+1} \in \operatorname{argmin}_{\theta\in\mathsf{C}\cap\mathsf{D}} D_\Phi(\theta, \zeta_{n+1}), \end{cases} \tag{6}$$

where $\eta > 0$ is the step-size, ∂f is the sub-differential of f (by abuse of notation) and D_Φ is the Bregman divergence associated with Φ:

$$\forall x, y \in \mathsf{D}, \quad D_\Phi(x, y) = \Phi(x) - \Phi(y) - \partial\Phi(y)^\top (x - y). \tag{7}$$

Note that gradient descent is a particular case of mirror descent with $\Phi\colon x \mapsto x^\top x/2$. Following [Bubeck, 2015, p.301], mirror descent can be rewritten as

$$\theta_{n+1} \in \operatorname*{argmin}_{\theta \in \mathsf{C} \cap \mathsf{D}} \eta g_n^\top \theta + D_\Phi(\theta, \theta_n),$$

which fits into the general framework (1) with $\Theta := \mathsf{C} \cap \mathsf{D}$ and $\mathcal{Q}$ defined for all $(\theta, \theta') \in \Theta \times \Theta$ by

$$\mathcal{Q}_\theta(\theta') := \eta g^\top \theta' + D_\Phi(\theta', \theta), \quad \text{where } g \in \partial f(\theta). \tag{8}$$

Example 2.2 (Mirror prox). Mirror prox is a variant of mirror descent defined by the following equations [Bubeck, 2015, Chapter 4, p.305]:

$$\begin{aligned}
\partial\Phi(\zeta'_{n+1}) &= \partial\Phi(\theta_n) - \eta\partial f(\theta_n), \\
\zeta_{n+1} &\in \operatorname*{argmin}_{\theta \in \mathsf{C} \cap \mathsf{D}} D_\Phi(\theta, \zeta'_{n+1}), \\
\partial\Phi(\theta'_{n+1}) &= \partial\Phi(\theta_n) - \eta\partial f(\zeta_{n+1}), \\
\theta_{n+1} &\in \operatorname*{argmin}_{\theta \in \mathsf{C} \cap \mathsf{D}} D_\Phi(\theta, \theta'_{n+1}).
\end{aligned}$$

Straightforward algebra yields the equivalent definition:

$$\begin{aligned}
\zeta_{n+1} \in \mathcal{M}(\theta_n) &= \operatorname*{argmin}_{\theta \in \mathsf{C} \cap \mathsf{D}} \eta\partial f(\theta_n)^\top \theta + D_\Phi(\theta, \theta_n), \\
\theta_{n+1} &\in \operatorname*{argmin}_{\theta \in \mathsf{C} \cap \mathsf{D}} \eta\partial f(\zeta_{n+1})^\top \theta + D_\Phi(\theta, \theta_n),
\end{aligned} \tag{9}$$

where $\mathcal{M}$ is defined for mirror descent by (8). We deduce from Example 2.1 that mirror prox fits into the general framework (1) with $\Theta := \mathsf{C} \cap \mathsf{D}$ and $\mathcal{Q}^m$ defined on $\Theta \times \Theta$ by

$$\mathcal{Q}^m_\theta(\theta') := \eta\partial f\left(\mathcal{M}(\theta)\right)^\top \theta' + D_\Phi(\theta', \theta). \tag{10}$$

The fact that $\mathcal{M}(\theta)$ is a singleton is ensured by (i) and (iii) as in this case Φ is a Legendre function, see [Cesa-Bianchi and Lugosi, 2006, Lemma 11.1] or [Bauschke, 1997, Theorem 3.12].

3 Asymptotic convergence rate

Assume there exists $\theta_\star \in \Theta$ such that $\partial_2\mathcal{Q}$ is well-defined in a neighborhood of $\theta_\star$ and differentiable at $(\theta_\star, \theta_\star)$, and write

$$\mathcal{A}_\star := \partial_{22}\mathcal{Q}_{\theta_\star}(\theta_\star), \quad \mathcal{B}_\star := -\partial_{12}\mathcal{Q}_{\theta_\star}(\theta_\star).$$

Let $\mathsf{V} := \operatorname{span}\{\theta - \theta' \,:\, \theta, \theta' \in \Theta\}$ be the direction of $\operatorname{Aff}(\Theta)$. Consider the following set of assumptions.

(H1) The set Θ is convex.

(H2) The sequence $(\theta_n)_{n\in\mathbb{N}}$ converges to $\theta_\star$.

(H3) There exists a neighborhood of $(\theta_\star, \theta_\star)$ on which $\mathcal{Q}$ is continuous and $\partial_2 \mathcal{Q}$ is well-defined and C^1-differentiable.

(H4) The matrix $\mathcal{B}_\star$ is symmetric and for all $v \in \mathsf{V} \setminus \{0\}$, $v^\top \mathcal{A}_\star v > |v^\top \mathcal{B}_\star v|$.

Under (H4), we can define

$$\hat{\rho}_\star := \sup_{v \in \mathsf{V} \setminus \{0\}} \frac{|v^\top \mathcal{B}_\star v|}{v^\top \mathcal{A}_\star v}. \tag{11}$$

In what follows, we set by convention, $\log 0 = -\infty$.

Theorem 1. *Assume that* (H1)-(H4) *hold. Then,* $\hat{\rho}_\star \in [0; 1)$ *and for all* $\rho \in (\hat{\rho}_\star; 1)$,

$$\theta_n - \theta_\star = \mathrm{o}(\rho^n),$$

or equivalently,

$$\limsup_{n \to \infty} \frac{1}{n} \log \|\theta_n - \theta_\star\|_2 \leqslant \log \hat{\rho}_\star.$$

As any two norms on a finite-dimensional linear space are equivalent, Theorem 1 and the next results could be given using another norm on Θ. They are stated here with $\|\cdot\|_2$ for simplicity.

Proof. See Section 7.

In [Balakrishnan et al., 2017, Theorem 1], the authors prove that the population EM algorithm converges geometrically. Their proof rely mainly on convergence results for gradient ascent algorithms applied to the intermediate quantity of the EM algorithm which is assumed to be smooth and strongly concave. Theorem 1 establishes under general assumptions that the convergence of the algorithms introduced in Section 2 is asymptotically geometric. Corollary 1 extends the statement of Theorem 1 to the values taken by the function $(\theta, \theta') \mapsto \mathcal{Q}_\theta(\theta')$ which are invariant to the choice of the parametrisation.

Corollary 1. *Under* (H1)-(H4)*, if* $\partial_1 \mathcal{Q}_{\theta_\star}(\theta_\star)$ *is well-defined, then for all* $\rho \in (\hat{\rho}_\star; 1)$,

$$\mathcal{Q}_{\theta_n}(\theta_{n+1}) - \mathcal{Q}_{\theta_\star}(\theta_\star) = \mathrm{o}(\rho^n).$$

Proof. See Appendix A.1.

Besides, the speed of convergence can be lower-bounded if the limit $\theta_\star$ lies in the relative interior of Θ. Under (H4), we can define

$$\check{\rho}_\star := \inf_{v \in \mathsf{V} \setminus \{0\}} \frac{|v^\top \mathcal{B}_\star v|}{v^\top \mathcal{A}_\star v}. \tag{12}$$

Theorem 2. *Assume that* (H1)-(H4) *hold, that* $\theta_\star \in \mathrm{ri}(\Theta)$*, and that the sequence* $(\theta_n)_{n \in \mathbb{N}}$ *is not eventually equal to* $\theta_\star$*. Then,* $\check{\rho}_\star \in [0; 1)$ *and*

$$\log \check{\rho}_\star \leqslant \liminf_{n \to \infty} \frac{1}{n} \log \|\theta_n - \theta_\star\|_2 .$$

Proof. See Appendix A.1.

If the limit $\theta_\star$ lies in the relative interior of Θ and $\check{\rho}_\star > 0$, the asymptotic convergence is therefore only geometric.

Theorem 3. *Assume that* (H1)-(H4) *hold, that $\partial_2\mathcal{Q}$ is C^2-differentiable in a neighborhood of $(\theta_\star,\theta_\star)$, that $\theta_\star \in \mathrm{ri}(\Theta)$ and that for all $p\in\mathbb{N}$, $\mathrm{Span}(\theta_n-\theta_\star, n\geqslant p)=\mathsf{V}$. Then $\check{\rho}_\star, \hat{\rho}_\star \in [0;1)$ and*

$$\log\min\left(\hat{\rho}_\star, \frac{\check{\rho}_\star}{\hat{\rho}_\star}\right) \leqslant \liminf_{n\to\infty}\frac{1}{n}\log\|\theta_n-\theta_\star\|_2\,,$$

where in the left-hand term we use the convention $0/0=0$ and $\log(0)=-\infty$. In particular, if $\hat{\rho}_\star^2\leqslant\check{\rho}_\star$ then

$$\lim_{n\to\infty}\frac{1}{n}\log\|\theta_n-\theta_\star\|_2 = \log\hat{\rho}_\star.$$

Proof. See Appendix A.1.

4 Comments on H1

The assumption that Θ needs to be convex can be relaxed as follows.

(H'1) There exist $\mathsf{E}\subset\mathbb{R}^q$ and a submanifold $\mathsf{S}\subset\mathbb{R}^q$ of class C^2 such that $\Theta=\mathsf{E}\cap\mathsf{S}$ and $\theta_\star\in\mathring{\mathsf{E}}$.

Under (H'1), if d is the dimension of the submanifold S, for all $x\in\mathsf{S}$, there exist U_1, U_2 two open neighborhoods of x and the null-vector $\mathbf{0}$ in $\mathbb{R}^q$, respectively, and a C^2-diffeomorphism $\psi\colon \mathsf{U}_1\to\mathsf{U}_2$ such that $\psi(x)=\mathbf{0}$ and $\psi(\mathsf{U}_1\cap\mathsf{S})=\mathsf{U}_2\cap(\mathbb{R}^d\times\{0\}^{q-d})$. Note that we identify $\mathbb{R}^q$ and $\mathbb{R}^d\times\mathbb{R}^{q-d}$ in a standard way using $(x_1,\dots,x_q)\mapsto((x_1,\dots,x_d),(x_{d+1},\dots,x_q))$. Write $\mathsf{T}_\star$ the tangent space to S at the point $\theta_\star$. If U_1 and U_2 are two open neighborhoods of $\theta_\star$ and the null-vector $\mathbf{0}$ in $\mathbb{R}^q$, respectively, and $\psi\colon\mathsf{U}_1\to\mathsf{U}_2$ is a C^1-diffeomorphism such that $\psi(\theta_\star)=\mathbf{0}$ and $\psi(\mathsf{U}_1\cap\mathsf{S})=\mathsf{U}_2\cap(\mathbb{R}^d\times\{0\}^{q-d})$, then $\mathsf{T}_\star=\partial\psi_{\theta_\star}^{-1}(\mathbb{R}^d\times\{0\}^{q-d})$, where $\partial\psi$ is the differential of ψ at $\theta_\star$.

(H'4) The matrix $\mathcal{B}_\star$ is symmetric and for all $v\in\mathsf{T}_\star\setminus\{0\}$, $v^\top\mathcal{A}_\star v>|v^\top\mathcal{B}_\star v|$.

Under (H'4), we can define

$$\check{\rho}_\star := \inf_{v\in\mathsf{T}_\star\setminus\{0\}}\frac{|v^\top\mathcal{B}_\star v|}{v^\top\mathcal{A}_\star v} \quad\text{and}\quad \hat{\rho}_\star := \sup_{v\in\mathsf{T}_\star\setminus\{0\}}\frac{|v^\top\mathcal{B}_\star v|}{v^\top\mathcal{A}_\star v}. \tag{13}$$

Theorem 4. *Assume that* (H'1), (H2), (H3) *and* (H'4) *hold. Then, $\hat{\rho}_\star\in[0;1)$ and for all $\rho\in(\hat{\rho}_\star;1)$,*

$$\theta_n-\theta_\star = \mathrm{o}(\rho^n).$$

Furthermore, if the sequence $(\theta_n)_{n\in\mathbb{N}}$ is not eventually equal to $\theta_\star$, then $\check{\rho}_\star\in[0;1)$ and for all $\rho\in(0;\check{\rho}_\star)$,

$$\rho^n = \mathrm{o}\left(\|\theta_n-\theta_\star\|_2\right).$$

Proof. See Appendix A.2. □

Remark 1. If (H1) holds with $\theta_\star\in\mathrm{ri}(\Theta)$, then (H'1) is satisfied with $\mathsf{E}:=\Theta+\mathsf{V}^\perp$ and $\mathsf{S}:=\mathrm{Aff}(\Theta)$.

Remark 2. If $\theta_\star$ does not lie in the relative interior of Θ, the asymptotic convergence rates are not necessarily invariant to C^2-reparametrization. Define for instance $\tilde{\mathcal{Q}}_{\theta_0}(\theta_1)=\theta_0^2-\theta_0\theta_1+\theta_1^2$ with $\Theta=[1;+\infty[$. Using the reparametrization function $\Psi:\theta\mapsto\theta^\alpha$, we set $\check{\mathcal{Q}}_{\theta_0}(\theta_1)=\tilde{\mathcal{Q}}_{\Psi(\theta_0)}(\Psi(\theta_1))=\theta_0^{2\alpha}-\theta_0^\alpha\theta_1^\alpha+\theta_1^{2\alpha}$. Then, with $\mathcal{Q}=\tilde{\mathcal{Q}}$, we get $\check{\rho}_\star=\hat{\rho}_\star=1/2$ whereas with $\mathcal{Q}=\check{\mathcal{Q}}$ and $\alpha=2/5$, we get $\check{\rho}_\star=\hat{\rho}_\star=2$.

5 Comments on H2

The convergence of the sequence $(\theta_n)_{n\in\mathbb{N}}$ to $\theta_\star$ (stated in (H2)) may be the most challenging assumption of Theorem 1. However, we provide alternative sufficient assumptions to establish such convergence.

(H2.1) The set Θ is compact.

(H2.2) The function $\mathcal{Q}$ is continuous on $\Theta \times \Theta$.

(H2.3) The point $\theta_\star$ is a limit point of the sequence $(\theta_n)_{n\in\mathbb{N}}$.

(H2.4) $\mathcal{M}(\theta_\star) = \{\theta_\star\}$.

Theorem 5. *Under* (H1)*,* (H2.1)-(H2.4)*,* (H3) *and* (H4)*, the sequence* $(\theta_n)_{n\in\mathbb{N}}$ *converges to* $\theta_\star$*.*

Proof. See Appendix A.3. □

Remark 3. Assumption (H2.3) weakens (H2) by only requiring that (H3)-(H4) hold for an arbitrary $\theta_\star$ in the limit set of $(\theta_n)_{n\in\mathbb{N}}$, which is non-empty under (H2.1).

Example 2.1 (Mirror descent, cont.). The map $\mathcal{M}$ is point-to-point on Θ under the assumptions of the definition (see Example 2.1 in page 3). Indeed, the surjectivity of the gradient in (ii) provides the existence of ζ_{n+1} in (6), and the strict convexity of Φ in (i) proves its uniqueness. Assumptions (i) and (iii) ensure the existence and the uniqueness of θ_{n+1} in (6) (see [Bauschke, 1997, Theorem 3.12]). Note that if f is convex or differentiable on C, then for all $\theta \in \mathsf{C}$, $\partial f(\theta) \neq \emptyset$ and g_n can be defined in (6).

Moreover, if $\theta_\star$ is a local minimizer of f and f is differentiable at $\theta_\star$, then (H2.4) is met. Indeed, those two assumptions provide that for all $\theta \in \Theta$, $\partial f(\theta_\star)^\top(\theta - \theta_\star) \geqslant 0$, and thus $\mathcal{Q}_{\theta_\star}(\theta) \geqslant \mathcal{Q}_{\theta_\star}(\theta_\star)$ in (8) with equality if and only if $\theta = \theta_\star$.

Proposition 1 establishes that the mirror prox approach described in Example 2.2 statisfies (H2.3) and (H2.4) under additional assumptions.

Proposition 1. *Assume in Example 2.2 that* (H2.1)-(H2.2) *hold and that: (i)* Φ *and* f *are twice differentiable on* Θ*, (ii)* Φ *is* γ*-strongly convex on* $\mathsf{C} \cap \mathsf{D}$ *and* f *is convex and* β*-smooth, with respect to* $\|\cdot\|_2$*, (iii)* $\eta \in (0; \gamma/\beta)$*, (iv)* $\theta_\star$ *is the unique minimizer of* f *on* C*, (v)* $\theta_\star \in \mathrm{ri}(\Theta)$*. Then,* (H2.3) *and* (H2.4) *hold.*

Proof. See Appendix A.3. □

We also provide alternative assumptions to prove (H2.4) in the general case.

($\tilde{\text{H}}$4.1) $\mathcal{M}(\theta_\star)$ is a singleton.

($\tilde{\text{H}}$4.2) There exists a continuous function $\vartheta\colon \Theta \to \mathbb{R}$ such that for all $\theta \in \Theta$, $\theta' \in \mathcal{M}(\theta)$,

$$\vartheta(\theta') \leqslant \vartheta(\theta),$$

with equality if and only if $\theta = \theta'$.

Theorem 6. *Assume* (H2.1)*,* (H2.2)*,* (H2.3)*. Then,* ($\tilde{\text{H}}$4.1) *and* ($\tilde{\text{H}}$4.2) *imply* (H2.4)*.*

Proof. See Appendix A.3. □

Example 1 (EM algorithm, cont.). By definition of the intermediate quantity (3), the EM algorithm monotonically increases the likelihood of the observations and Assumption ($\tilde{\text{H}}$4.2) is satisfied as soon as the log-likelihood is continuous, see for example [Cappé et al., 2005, Proposition 10.1.4, p.350].

6 Comments on H4

First of all, the matrix $\mathcal{B}_\star$ appears to be symmetric in all the examples below. The discussion then focuses on the domination assumption and on the value of the convergence rate $\hat{\rho}_\star$. Note that the domination assumption in (H4) is equivalent to having both $\tilde{\mathcal{A}}_\star \succ \tilde{\mathcal{B}}_\star$ and $\tilde{\mathcal{A}}_\star \succ -\tilde{\mathcal{B}}_\star$. In the case where $\tilde{\mathcal{A}}_\star \succ 0$, it is equivalent to $\hat{\rho}_\star \in [0;1)$.

Example 1.1 (Population EM). Assume that $\theta_\star$ is the true parameter of the model, that for all $x, y \in \mathsf{X}, \mathsf{Y}$, the functions $\theta \mapsto p_\theta(x|y)$ and $\theta \mapsto p_\theta(y)$ are twice differentiable in a neighborhood of $\theta_\star$, and that conditions similar to [Douc et al., 2013, Assumption AD.1, p.492] hold to differentiate under the integral sign. Then, we prove in Appendix A.4, see (61) and (62), that

$$\mathcal{A}_\star^{\mathrm{pop}} = I_{X,Y}(\theta_\star) \quad \text{and} \quad \mathcal{B}_\star^{\mathrm{pop}} = I_{X,Y}(\theta_\star) - I_Y(\theta_\star), \tag{14}$$

where $I_{X,Y}(\theta) := -\mathbb{E}_\theta[\partial_\theta^2 \log p_\theta(X,Y)]$ and $I_Y := -\mathbb{E}_\theta[\partial_\theta^2 \log p_\theta(Y)]$ denote the Fisher information matrices of (X,Y) and Y, respectively. Therefore, (H4) is satisfied as soon as $I_Y(\theta_\star) \succ 0$. Regarding the value of $\hat{\rho}_\star$, the above expressions of $\mathcal{A}_\star^{\mathrm{pop}}$ and $\mathcal{B}_\star^{\mathrm{pop}}$ provide the well-known ratio of missing information $I_{X,Y}(\theta_\star)^{-1} I_{X|Y}(\theta_\star)$ (see [Dempster et al., 1977, Kunstner et al., 2021, Meng and Rubin, 1991, Meng and Rubin, 1993, Orchard and Woodbury, 1972]), where

$$I_{X|Y}(\theta_\star) = \int_{\mathsf{Y}} \int_{\mathsf{X}} p_{\theta_\star}(y) p_\theta(x|y) \partial \log p_{\theta'}(x|y) \left[\partial \log p_\theta(x|y)\right]^\top \mu(\mathrm{d}x)\mu(\mathrm{d}y).$$

Example 1.2 (Sample EM). As for the other examples, all the results below are proved in Appendix A.4. Assume that for all $x, y \in \mathsf{X}, \mathsf{Y}$, the functions $\theta \mapsto p_\theta(x|y)$ and $\theta \mapsto p_\theta(y)$ are twice differentiable in a neighborhood of $\theta_\star$, and that conditions similar to [Douc et al., 2013, Assumption AD.1, p.492] hold to differentiate under the integral sign. Let $(Y_i)_{i\in\mathbb{N}^*}$ be a sequence of independent and identically distributed random variables with probability density function $p_{\theta_\star}$, and write for all $k \in \mathbb{N}^*$, $Y_{1:k} := (Y_i)_{1\leqslant i\leqslant k}$. Then, for all $k \in \mathbb{N}^*$, by (63) and (64),

$$\mathcal{A}_\star^{\mathrm{samp}}(Y_{1:k}) = \frac{1}{k}\sum_{i=1}^k \left(I_{X|Y=Y_i}(\theta_\star) - \partial^2 \log p_{\theta_\star}(Y_i)\right),$$

$$\mathcal{B}_\star^{\mathrm{samp}}(Y_{1:k}) = \frac{1}{k}\sum_{i=1}^k I_{X|Y=Y_i}(\theta_\star),$$

where $I_{X|Y=Y_i}(\theta_\star) = \int_{\mathsf{X}} p_{\theta_\star}(x|Y_i) \partial \log p_{\theta_\star}(x|Y_i) \left[\partial \log p_{\theta_\star}(x|Y_i)\right]^\top \mu(\mathrm{d}x)$.

Note that $\mathcal{A}_\star^{\mathrm{samp}}(Y_{1:k})$ and $\mathcal{B}_\star^{\mathrm{samp}}(Y_{1:k})$ converges almost surely to $\mathcal{A}_\star^{\mathrm{pop}}$ and $\mathcal{B}_\star^{\mathrm{pop}}$. Then, if the corresponding population EM meets (H4), almost surely, for sufficiently large k, the sample EM meets (H4). Denoting by $\hat{\rho}_\star^{\mathrm{pop}}$ and $\hat{\rho}_\star^{\mathrm{samp}}(Y_{1:k})$ their respective rates, as defined in (11), by Lemma A.3, we also have that

$$\hat{\rho}_\star^{\mathrm{samp}}(Y_{1:k}) \xrightarrow{a.s.} \hat{\rho}_\star^{\mathrm{pop}}. \tag{15}$$

Furthermore, if $\partial^2 \log p_{\theta_\star}(X_1, Y_1), \partial^2 \log p_{\theta_\star}(Y_1) \in \mathrm{L}^2(\mathbb{R}^{q\times q})$, Lemma A.3 also establishes that for all $\delta \in (0;1)$ there exists $C_\delta > 0$ such that

$$\liminf_{k\to\infty} \mathbb{P}\left(|\hat{\rho}_\star^{\mathrm{samp}}(Y_{1:k}) - \hat{\rho}_\star^{\mathrm{pop}}| \leqslant \frac{C_\delta}{\sqrt{k}}\right) \geqslant 1-\delta. \tag{16}$$

Example 2.1 (Mirror descent, cont.). If f and Φ are twice differentiable in a neighborhood of $\theta_\star$, we prove in Appendix A.4 that

$$\mathcal{A}_\star = \partial^2\Phi(\theta_\star) \quad \text{and} \quad \mathcal{B}_\star = \partial^2\Phi(\theta_\star) - \eta\partial^2 f(\theta_\star). \tag{17}$$

If for all $v \in \mathsf{V}$, $v^\top \partial^2 f(\theta_\star) v > 0$, the condition $\tilde{\mathcal{A}}_\star \succ \tilde{\mathcal{B}}_\star$ is automatically satisfied. The domination assumption in (H4) then reduces to $\tilde{\mathcal{A}}_\star \succ -\tilde{\mathcal{B}}_\star$, which corresponds to η being small enough. In the particular case of unconstrained gradient descent where $\mathrm{Aff}(\Theta) = \mathbb{R}^q$ and $\Phi\colon x \mapsto x^\top x/2$, as (17) yields $\mathcal{A}_\star = I_q$ the above condition is equivalent to $\eta \in (0; 2/\beta_\star)$, the optimal choice being $\eta = 2/(\alpha_\star + \beta_\star)$ where $\alpha_\star := \min \mathrm{Spec}(\partial^2 f(\theta_\star))$ and $\beta_\star := \max \mathrm{Spec}(\partial^2 f(\theta_\star))$.

Besides, the asymptotic convergence rate $\hat{\rho}_\star$ can be interpreted similarly to the EM framework. Despite not being, strictly speaking, a ratio of missing information, $\hat{\rho}_\star$ still compares the mirror map Φ with the objective function f. Intuitively, the choice of a mirror map with variations closer to those of f provides a better convergence rate. If $\eta = 1$, the extreme case $\Phi = f$ yields $\mathcal{B}_\star = 0$ and $\hat{\rho}_\star = 0$, which is coherent with the fact that, in this case, the mirror descent is defined for all $n \in \mathbb{N}$ by $\theta_{n+1} \in \mathrm{argmin}_{\theta\in\Theta} f(\theta)$.

The following discussion extends the above interpretation to a general class of functions $\mathcal{Q}$ that encompasses both mirror descent and the EM algorithm. The first thing to note is that in both settings the function $\mathcal{Q}$ can be redefined as

$$\mathcal{Q}_\theta(\theta') := f(\theta') + D(\theta, \theta'), \tag{18}$$

where $f\colon \mathbb{R}^q \to \mathbb{R}$ is the objective function and $D\colon \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$ is a function such that for all $\theta \in \Theta$, $\partial_2 D(\theta, \theta) = 0$. Indeed, it is common knowledge that the intermediate quantity of the EM algorithm can be expressed as

$$Q(\theta, \theta') = \log p_{\theta'}(Y) - D_{\mathrm{KL}}(p_\theta(X|Y)||p_{\theta'}(X|Y)),$$

where D_{KL} denotes the Kullback-Leibler divergence (see [Daudel et al., 2020] for example). Regarding mirror descent, if f is twice differentiable in Example 2.1, straightforward computation yields the following equivalent definition for $\mathcal{Q}$:

$$\mathcal{Q}_\theta(\theta') := f(\theta') + \frac{1}{\eta} D_{\Phi - \eta f}(\theta', \theta),$$

where the expression of $D_{\Phi-\eta f}$ follows that of (7) (and defines a Bregman divergence if $\Phi - \eta f$ is strictly convex). Besides, the condition $\partial_2 D(\theta, \theta) = 0$ for all $\theta \in \Theta$ is equivalent to $\partial_2 \mathcal{Q}_\theta(\theta) = \partial f(\theta)$ for all $\theta \in \Theta$, hence

$$\begin{aligned}
\mathcal{A}_\star &= \partial_{22}\mathcal{Q}_{\theta_\star}(\theta_\star), \\
\mathcal{B}_\star &= -\partial_{12}\mathcal{Q}_{\theta_\star}(\theta_\star) = \partial_{22}\mathcal{Q}_{\theta_\star}(\theta_\star) - \partial^2 f(\theta_\star),
\end{aligned}$$

and

$$\hat{\rho}_\star = \sup_{v\in\mathsf{V}} \frac{|v^\top \left(\partial_{22}\mathcal{Q}_{\theta_\star}(\theta_\star) - \partial^2 f(\theta_\star)\right) v|}{v^\top \partial_{22}\mathcal{Q}_{\theta_\star}(\theta_\star)\, v}.$$

In the framework of (18), the convergence rate can thus be viewed as a relative difference between the second-order variations of $\mathcal{Q}$ and f. Computing iteratively $\mathrm{argmin}_\Theta \mathcal{Q}_{\theta_n}(\cdot)$ to estimate $\mathrm{argmin}_\Theta f$ can prove useful if those minimizations are easier to carry out, but the price to pay in terms of iterations (through the convergence rate) is directly related to how far the surrogate function $\mathcal{Q}$ is from the objective function f. If $\tilde{\mathcal{A}}_\star$ is invertible, $\hat{\rho}_\star$ is indeed the spectral radius of $\tilde{\mathcal{B}}_\star \tilde{\mathcal{A}}_\star^{-1} = (\partial_{22}\tilde{\mathcal{Q}}_{\theta_\star}(\theta_\star) - \partial^2 \tilde{f}(\theta_\star))(\partial_{22}\tilde{\mathcal{Q}}_{\theta_\star}(\theta_\star))^{-1}$ by Lemma B.3, and the interpretation of a *ratio of missing information* generalizes to that of a ratio measuring the loss of exactness in the minimization procedure.

Finally, in the particular case where D is a distance or a divergence twice differentiable at $(\theta_\star, \theta_\star)$ with respect to the second argument, $\theta \in \operatorname{argmin}_\Theta D(\theta, \cdot)$ for all $\theta \in \Theta$ implies $\mathcal{B}_\star = \partial_{22} D(\theta_\star, \theta_\star) \succeq 0$. The domination assumption in (H4) then boils down to $\partial^2 \tilde{f}(\theta_\star) \succ 0$.

Example 1.3 (The α-EM algorithm). The above discussion highlighted how the choice of the surrogate function $\mathcal{Q}$ determines the convergence rate $\hat{\rho}_\star$. In the EM algorithm of Example 1, where the function $\mathcal{Q}$ can be defined for all $\theta, \theta' \in \Theta \times \Theta$ as

$$\mathcal{Q}_\theta(\theta') := -\log p_{\theta'}(Y) - \int_{\mathsf{X}} p_\theta(x|Y) \log \frac{p_{\theta'}(x|Y)}{p_\theta(x|Y)} \mu(\mathrm{d}x),$$

the question then rises whether replacing the Kullback-Leibler divergence by an α-divergence (see [Daudel et al., 2020] for example) could provide a better convergence rate. This leads to replacing the previous expression of $\mathcal{Q}$ by:

$$\mathcal{Q}^\alpha_\theta(\theta') := -\int_{\mathsf{X}} p_\theta(x|Y) f_\alpha\left(\frac{p_{\theta'}(x,Y)}{p_\theta(x,Y)}\right) \mu(\mathrm{d}x), \tag{19}$$

where for all $\alpha \in \mathbb{R} \setminus \{0,1\}$, the concave function f_α is defined on $\mathbb{R}_+^*$ by $f_\alpha(x) := (1 - x^\alpha)/\alpha(\alpha - 1)$ and $f_0 := \log$. This approach has been introduced and developed in [Matsuyama, 2003]. We provide further elements for the choice of α by proving (see Appendix A.4) that at a population level, under the assumptions of Example 1.1 with f_α instead of f_0,

$$\mathcal{A}^\alpha_\star = I_{X,Y}(\theta_\star) \quad \text{and} \quad \mathcal{B}^\alpha_\star = I_{X,Y}(\theta_\star) - \frac{1}{1-\alpha} I_Y(\theta_\star). \tag{20}$$

Note that when $\alpha = 0$ we recover the previous quantities $\mathcal{A}_\star$ and $\mathcal{B}_\star$ for the classical EM algorithm. If $I_Y(\theta_\star) \succ 0$, then $\alpha \in (0; 1/2)$ is a sufficient condition to meet (H4). Besides, if $\mathcal{A}^\alpha_\star$ is invertible we can write

$$(\mathcal{A}^\alpha_\star)^{-1} \mathcal{B}^\alpha_\star = I_{X,Y}(\theta_\star)^{-1} I_{X|Y}(\theta_\star) - \frac{\alpha}{1-\alpha} I_{X,Y}(\theta_\star)^{-1} I_Y(\theta_\star). \tag{21}$$

A necessary condition to improve the convergence rate is then $\alpha/(1-\alpha) > 0$, i.e. $\alpha \in (0; 1)$. We can also rewrite (21) as follows:

$$(\mathcal{A}^\alpha_\star)^{-1} \mathcal{B}^\alpha_\star = \frac{1}{1-\alpha} \mathcal{A}_\star^{-1} \mathcal{B}_\star - \frac{\alpha}{1-\alpha} I_q.$$

By the positivity of $\mathcal{B}_\star$ for the original EM algorithm, we deduce that the optimal choice of α corresponds to $\alpha = (\hat{\rho}_\star + \check{\rho}_\star)/2$ and $\hat{\rho}^\alpha_\star = (\hat{\rho}_\star - \check{\rho}_\star)/(2 - \hat{\rho}_\star - \check{\rho}_\star)$, where $\hat{\rho}_\star$ and $\check{\rho}_\star$ are defined in (11) and (12) for the classical EM algorithm.

As a remark, we can see in [Matsuyama, 2003] that the α-EM algorithm does not simply change the D-function in (18), it also replaces the objective function with a different bivariate function.

Example 2.2 (Mirror prox, cont.). Assume that (H2.1) hold, that Φ and f are C^1-differentiable on Θ and twice differentiable at $\theta_\star$, and that the corresponding mirror descent satisfies (H3)-(H4) and $\theta_\star = \mathcal{M}(\theta_\star) \in \mathrm{ri}(\Theta)$. Then, we prove in Appendix A.4 that

$$\tilde{\mathcal{A}}^m_\star = \tilde{\mathcal{A}}_\star \quad \text{and} \quad \tilde{\mathcal{B}}^m_\star = \tilde{\mathcal{A}}_\star + \tilde{\mathcal{B}}_\star \tilde{\mathcal{A}}_\star^{-1} \tilde{\mathcal{B}}_\star - \tilde{\mathcal{B}}_\star,$$

where $\mathcal{A}_\star$, $\mathcal{B}_\star$ are defined in (17) for mirror descent. This provides the symmetry of $\tilde{\mathcal{B}}^m_\star$ and thus of $\mathcal{B}^m_\star$, as well as

$$(\tilde{\mathcal{A}}^m_\star)^{-1} \tilde{\mathcal{B}}^m_\star = (\tilde{\mathcal{A}}_\star^{-1} \tilde{\mathcal{B}}_\star)^2 - \tilde{\mathcal{A}}_\star^{-1} \tilde{\mathcal{B}}_\star + I_d. \tag{22}$$

We deduce that under (H4) for mirror descent, $\hat{\rho}_\star^m < 1$ if and only if $\tilde{\mathcal{B}}_\star \succ 0$, which is met as soon as $\eta \in (0; \gamma_\star/\beta_\star)$, where $\beta_\star := \max \mathrm{Spec}(\partial^2 \tilde{f}(\theta_\star))$ and $\gamma_\star := \min \mathrm{Spec}(\partial^2 \tilde{\Phi}(\theta_\star))$. Besides, a sufficient condition for the C^1-differentiability of $\partial_2 \mathcal{Q}^m$ in a neighborhood of $\theta_\star$ is the C^2-differentiability of Φ and f in a neighborhood of $\theta_\star$. Under all those assumptions, mirror prox thus meets (H3)-(H4).

Note that the above sufficient condition of regularity implies (H3) for mirror descent, and that if $\tilde{\mathcal{B}}_\star \succ 0$, then $\partial^2 \tilde{f}(\theta_\star) \succ 0$ implies (H4) for mirror descent (see Example 2.1 in page 9).

As a remark, (22) yields that the convergence rates defined in (11-12) are always strictly higher for mirror prox than for the corresponding mirror descent. The rate $\check{\rho}_\star^m$ is even lower-bounded by $3/4$ (see Appendix A.4).

Example 3 (Newton's method). Let f be a C^2-differentiable function f whose Hessian is invertible on Θ. Newton's method considers the procedure defined for all $n \in \mathbb{N}$ by

$$\theta_{n+1} = \theta_n - \partial^2 f(\theta_n)^{-1} \partial f(\theta_n).$$

It fits into the general framework of (1) with $\mathcal{Q}$ defined on $\Theta \times \Theta$ by

$$\mathcal{Q}_\theta(\theta') := \frac{1}{2} \left\| \theta' - \theta + \partial^2 f(\theta)^{-1} \partial f(\theta) \right\|_2^2 .$$

If f is thrice differentiable at $\theta_\star$ and $\partial f(\theta_\star) = 0$, straightforward calculus yields $\mathcal{A}_\star = I_q$ and $\mathcal{B}_\star = 0$. Newton's method thus meets (H4) with $\check{\rho}_\star = \hat{\rho}_\star = 0$, which is coherent with the fact that the convergence is quadratic under the assumptions of [Nocedal and J., 2006, Theorem 3.5, p.44].

7 Proof of Theorem 1

We start with some notation that will be used in several parts of the paper. Set $d := \dim(\mathsf{V})$. Let $v_1, \ldots, v_d \in \mathbb{R}^q$ be an orthonormal basis of V and let P be the matrix

$$P := [v_1 | \ldots | v_d] \in \mathbb{R}^{q \times d} \tag{23}$$

so that $\mathsf{V} = P(\mathbb{R}^d)$. For all $x \in \mathbb{R}^q$ and $M \in \mathbb{R}^{q \times q}$, write

$$\tilde{x} := P^\top x \in \mathbb{R}^d , \quad \tilde{M} := P^\top M P \in \mathbb{R}^{d \times d}. \tag{24}$$

Note that for all $v \in \mathsf{V}$, $P\tilde{v} = PP^\top v = v$. Write for all $n \in \mathbb{N}$,

$$\Delta_n := \theta_n - \theta_\star \in \mathsf{V}, \tag{25}$$

and hence

$$\tilde{\Delta}_n := P^\top (\theta_n - \theta_\star) \in \mathbb{R}^d. \tag{26}$$

Then, for all $M \in \mathbb{R}^{q \times q}$ and $n, m \in \mathbb{N}$,

$$\tilde{\Delta}_n^\top \tilde{M} \tilde{\Delta}_m = \tilde{\Delta}_n^\top P^\top M P \tilde{\Delta}_m = \Delta_n^\top M \Delta_m. \tag{27}$$

In particular, with $M = I_q$ the identity matrix, for all $n \in \mathbb{N}$,

$$\|\tilde{\Delta}_n\|_2 = \|\Delta_n\|_2. \tag{28}$$

Proof of Theorem 1. In the definition of $\hat{\rho}_\star$ given in (11), the supremum of $v \mapsto |v^\top \mathcal{B}_\star v|/(v^\top \mathcal{A}_\star v)$ can be taken over the compact set $\{v \in \mathsf{V} \,:\, v^\top v = 1\}$ and it is thus attained. This yields $\hat{\rho}_\star \in [0;1)$ under (H4).

Let $\rho \in (\hat{\rho}_\star; 1)$. Proposition 2 below provides for sufficiently large n,

$$\|\tilde{\Delta}_{n+1}\|_{\tilde{\mathcal{A}}_\star} \leqslant \rho \|\tilde{\Delta}_n\|_{\tilde{\mathcal{A}}_\star}.$$

This yields $\|\tilde{\Delta}_n\|_{\tilde{\mathcal{A}}_\star} = \mathrm{O}(\rho^n)$, and hence $\|\tilde{\Delta}_n\|_2 = \mathrm{O}(\rho^n)$ by the equivalence of norms in finite dimension. The proof is concluded by noting that this holds for any arbitrary $\rho > \hat{\rho}_\star$ and since by (28),

$$\|\tilde{\Delta}_n\|_2 = \|\Delta_n\|_2 = \|\theta_n - \theta_\star\|_2 .$$

□

Lemma 7.1. *Under* (H2)-(H3)*, $\theta_\star$ is a local minimizer on Θ of the function $\theta \mapsto \mathcal{Q}_{\theta_\star}(\theta)$.*

Proof. Let N be a neighborhood of $\theta_\star$ such that $\mathcal{Q}$ is continuous on $\mathsf{N} \times \mathsf{N}$. For all $\theta \in \mathsf{N}$ and $n \in \mathbb{N}$, the definition of $(\theta_n)_{n\in\mathbb{N}}$ in (1) provides $\mathcal{Q}_{\theta_n}(\theta_{n+1}) \leqslant \mathcal{Q}_{\theta_n}(\theta)$. Taking the limit when n goes to infinity yields $\mathcal{Q}_{\theta_\star}(\theta_\star) \leqslant \mathcal{Q}_{\theta_\star}(\theta)$ for all $\theta \in \mathsf{N}$. □

Proposition 2. *Under* (H1)-(H4)*, for all $\rho > \hat{\rho}_\star$, for sufficiently large n,*

$$\|\tilde{\Delta}_{n+1}\|_{\tilde{\mathcal{A}}_\star} \leqslant \rho \|\tilde{\Delta}_n\|_{\tilde{\mathcal{A}}_\star}.$$

Proof. By Lemma 7.1, $\theta_\star$ is a local minimizer of the function $\theta \mapsto \mathcal{Q}_{\theta_\star}(\theta)$. From the differentiability of that function at $\theta_\star$ under (H3), and the convexity of Θ under (H1), we deduce that for all $\theta \in \Theta$,

$$\langle \partial_2 \mathcal{Q}_{\theta_\star}(\theta_\star), \theta - \theta_\star \rangle = \lim_{t\to 0^+} \frac{\partial \mathcal{Q}_{\theta_\star}(t\theta + (1-t)\theta_\star)}{\partial t} \geqslant 0. \tag{29}$$

Similarly, under (H2)-(H3) the function $\theta \mapsto \mathcal{Q}_{\theta_n}(\theta)$ is differentiable at θ_{n+1} for sufficiently large n, which yields for all $\theta \in \Theta$,

$$\langle \partial_2 \mathcal{Q}_{\theta_n}(\theta_{n+1}), \theta - \theta_{n+1} \rangle \geqslant 0. \tag{30}$$

Using (30) with $\theta = \theta_\star$ and (29) with $\theta = \theta_{n+1}$ provides

$$\langle \partial_2 \mathcal{Q}_{\theta_n}(\theta_{n+1}), \theta_{n+1} - \theta_\star \rangle \leqslant 0 \leqslant \langle \partial_2 \mathcal{Q}_{\theta_\star}(\theta_\star), \theta_{n+1} - \theta_\star \rangle,$$

which in turn implies

$$\langle \partial_2 \mathcal{Q}_{\theta_n}(\theta_{n+1}) - \partial_2 \mathcal{Q}_{\theta_n}(\theta_\star), \theta_{n+1} - \theta_\star \rangle \leqslant \langle \partial_2 \mathcal{Q}_{\theta_\star}(\theta_\star) - \partial_2 \mathcal{Q}_{\theta_n}(\theta_\star), \theta_{n+1} - \theta_\star \rangle. \tag{31}$$

Besides, applying Taylor's theorem to $\theta \mapsto \partial_2 \mathcal{Q}_{\theta_n}(\theta)$ and $\theta \mapsto \partial_2 \mathcal{Q}_\theta(\theta_\star)$ yields for sufficiently large n,

$$\partial_2 \mathcal{Q}_{\theta_n}(\theta_{n+1}) - \partial_2 \mathcal{Q}_{\theta_n}(\theta_\star) = \mathcal{A}_n(\theta_{n+1} - \theta_\star), \tag{32}$$

$$\partial_2 \mathcal{Q}_{\theta_\star}(\theta_\star) - \partial_2 \mathcal{Q}_{\theta_n}(\theta_\star) = \mathcal{B}_n(\theta_n - \theta_\star), \tag{33}$$

where

$$\mathcal{A}_n := \int_0^1 \partial_{22} \mathcal{Q}_{\theta_n}(s\theta_{n+1} + (1-s)\theta_\star)\, \mathrm{d}s, \quad \mathcal{B}_n := -\int_0^1 \partial_{12} \mathcal{Q}_{s\theta_\star + (1-s)\theta_n}(\theta_\star)\, \mathrm{d}s. \tag{34}$$

Plugging (32-33) into (31), we deduce

$$(\theta_{n+1}-\theta_\star)^\top \mathcal{A}_n(\theta_{n+1}-\theta_\star) \leqslant (\theta_{n+1}-\theta_\star)^\top \mathcal{B}_n(\theta_n-\theta_\star).$$

Using (27), this can be written as

$$\tilde{\Delta}_{n+1}^\top \tilde{\mathcal{A}}_n \tilde{\Delta}_{n+1} \leqslant \tilde{\Delta}_{n+1}^\top \tilde{\mathcal{B}}_n \tilde{\Delta}_n. \tag{35}$$

Now, by Schwarz's theorem, $\mathcal{A}_\star$ and hence $\tilde{\mathcal{A}}_\star$ are symmetric. Similarly, under (H2)-(H3), $\mathcal{A}_n$ and hence $\tilde{\mathcal{A}}_n$ are symmetric for sufficiently large n. Moreover, (H4) implies the positive-definiteness of $\tilde{\mathcal{A}}_\star$, and by (H2)-(H3) that of $\tilde{\mathcal{A}}_n$ for sufficiently large n (see [Tao, 2012, Section 1.3.4, p.47]). We can thus apply Lemma B.1 to (35) with $x=\tilde{\Delta}_{n+1}$, $y=\tilde{\Delta}_n$, $A=\tilde{\mathcal{A}}_n$ and $B=\tilde{\mathcal{B}}_n$ and we obtain

$$\|\tilde{\Delta}_{n+1}\|_{\tilde{\mathcal{A}}_n} \leqslant \hat{\rho}_n \|\tilde{\Delta}_n\|_{\tilde{\mathcal{A}}_n}, \tag{36}$$

where $\hat{\rho}_n := \|\!|\tilde{\mathcal{A}}_n^{-1/2}\tilde{\mathcal{B}}_n\tilde{\mathcal{A}}_n^{-1/2}|\!\|_2$. Under (H2)-(H3), Lemma B.2 shows that $\hat{\rho}_n$ converges to $\|\!|\tilde{\mathcal{A}}_\star^{-1/2}\tilde{\mathcal{B}}_\star\tilde{\mathcal{A}}_\star^{-1/2}|\!\|_2$ by choosing $A=\tilde{\mathcal{A}}_\star$, $B=\tilde{\mathcal{B}}_\star$, $M=\tilde{\mathcal{A}}_n-\tilde{\mathcal{A}}_\star$ and $N=\tilde{\mathcal{B}}_n-\tilde{\mathcal{B}}_\star$. On the other hand, by Lemma B.3, $\hat{\rho}_\star = \|\!|\tilde{\mathcal{A}}_\star^{-1/2}\tilde{\mathcal{B}}_\star\tilde{\mathcal{A}}_\star^{-1/2}|\!\|_2$.

Let $\rho > \hat{\rho}_\star$. Set $\rho' := (\rho+\hat{\rho}_\star)/2$ and $\varepsilon > 0$ such that $(1+\varepsilon)\rho' \leqslant (1-\varepsilon)\rho$. Under (H2)-(H3), Lemma B.4 yields that for sufficiently large n, for all $u \in \mathbb{R}^d$,

$$(1-\varepsilon)\,\|u\|_{\tilde{\mathcal{A}}_n} \leqslant \|u\|_{\tilde{\mathcal{A}}_\star} \leqslant (1+\varepsilon)\,\|u\|_{\tilde{\mathcal{A}}_n}.$$

Combining with (36) and the convergence of $\hat{\rho}_n$ to $\hat{\rho}_\star$, we deduce for sufficiently large n,

$$\|\tilde{\Delta}_{n+1}\|_{\tilde{\mathcal{A}}_\star} \leqslant (1+\varepsilon)\|\tilde{\Delta}_{n+1}\|_{\tilde{\mathcal{A}}_n} \leqslant (1+\varepsilon)\hat{\rho}_n\|\tilde{\Delta}_n\|_{\tilde{\mathcal{A}}_n} \leqslant \frac{(1+\varepsilon)\rho'}{1-\varepsilon}\|\tilde{\Delta}_n\|_{\tilde{\mathcal{A}}_\star} \leqslant \rho\|\tilde{\Delta}_n\|_{\tilde{\mathcal{A}}_\star}.$$

□

8 Convex constrained optimization

Theorem 7. *Assume that the mirror prox strategy defined in Example 2.2 page 4 satisfies the following assumptions: (i) $\mathsf{C} \subset \mathsf{D}$, (ii) Φ and f are twice differentiable on C and C^2-differentiable in a neighborhood of $\theta_\star$, (iii) Φ is γ-strongly convex on C and f is convex and β-smooth, with respect to $\|\cdot\|_2$, (iv) $\eta \in (0;\gamma/\beta)$, (v) $\theta_\star$ is the unique minimizer of f on C, (vi) $\theta_\star \in \mathrm{ri}(\mathsf{C})$ and $\partial^2 \tilde{f}(\theta_\star) \succ 0$.*

Then, the algorithm converges and the convergence is asymptotically geometric.

Proof. See Proposition 1 in page 7 and Example 2.2 in page 10. □

Even if the convergence rates of mirror descent are always lower than those of mirror prox for the same optimization problem (see Example 2.2 in page 10), the convergence of mirror prox is guaranteed under the assumptions of Theorem 7.

Note that no conditions are imposed on the initialization (see [Bubeck, 2015, Chapter 4, p.299]).

Corollary 2. *Let $q \in \mathbb{N}^*$, $\mathsf{C} \subset \mathbb{R}^q$ be compact set, and f be a function that meets assumptions (ii)-(iii) and (v)-(vi) of Theorem 7. Then, mirror prox provides an algorithm that converges to* $\mathrm{argmin}_{\mathsf{C}} f$.

Proof. Write $R := \max_{x\in\mathsf{C}} \|x\|_2$. For all $R' > R$, the mirror map Φ defined on $\mathsf{D} := \mathbf{B}(0,R') := \left\{x \in \mathbb{R}^d \,:\, \|x\|_2 < R'\right\}$ by $\Phi(x) := \|x\|_2^2/(R'-\|x\|_2^2)$ meets the assumptions of Theorem 7 for all $\eta \in (0; 2(R'\beta)^{-1})$. □

9 Discussion

9.1 Non-asymptotic convergence

We proved the asymptotic geometric convergence in Section 7 by using that for all $n \in \mathbb{N}$,

$$\|\tilde{\theta}_{n+1} - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_n} \leqslant \hat{\rho}_n \|\tilde{\theta}_n - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_n},$$

as soon as we can define the above quantities (see (36)). The question then rises of deriving non-asymptotic convergence rates. Apart from the fact that the norm depends on n, the main issue would be to obtain $\hat{\rho}_n < 1$. However, the ratio $\hat{\rho}_\star$ compares $\partial_{22}\mathcal{Q}_{\theta_\star}(\theta_\star)$ with $\partial_{12}\mathcal{Q}_{\theta_\star}(\theta_\star)$, whereas $\hat{\rho}_n$ compares

$$\mathcal{A}_n = \int_0^1 \partial_{22}\mathcal{Q}_{\theta_n}\left(s\theta_{n+1} + (1-s)\theta_\star\right) \mathrm{d}s \quad \text{with} \quad \mathcal{B}_n = -\int_0^1 \partial_{12}\mathcal{Q}_{s\theta_\star + (1-s)\theta_n}\left(\theta_\star\right) \mathrm{d}s, \tag{37}$$

and the problem is not of the same complexity. A sufficient condition to simplify it can be $\min \mathrm{Spec}(\mathcal{A}_n) > \max |\mathrm{Spec}(\mathcal{B}_n)|$. Despite being less precise than a comparison being matrices, such a condition has the advantage that it is sufficient to verify it for every $s \in [0;1]$ in (37). That pointwise condition essentially corresponds to the conditions 1 and 2 behind $\gamma < \lambda$ in [Balakrishnan et al., 2017, Theorem 1], concerning $\partial_{12}\mathcal{Q}$ and $\partial_{22}\mathcal{Q}$ respectively (the proof of that theorem has besides inspired this work). We can also identify the classical assumptions of smoothness and Lipschitz continuity for gradient descent and its variants.

In light of that remark, we better understand why the framework introduced in this paper yields better results asymptotically, as it allows to work with the true asymptotic convergence rate. We can see it when comparing with the results stated so far in the EM literature [Dempster et al., 1977, Kunstner et al., 2021, Meng and Rubin, 1994] (that generally do not consider constrained optimization and assume that the mapping $\mathcal{M}$ is differentiable, among other things).

9.2 Quadratic convergence

Under the assumptions of Theorem 2, we established in Appendix A.1 that for sufficiently large n, we can write $\tilde{\mathcal{B}}_n \tilde{\Delta}_n = \tilde{\mathcal{A}}_n \tilde{\Delta}_{n+1}$, which is equivalent to

$$\tilde{\theta}_\star = \tilde{\theta}_n + (I - \tilde{\mathcal{A}}_n^{-1}\tilde{\mathcal{B}}_n)^{-1}(\tilde{\theta}_{n+1} - \tilde{\theta}_n). \tag{38}$$

Note that $\tilde{\mathcal{A}}_n$ and $\tilde{\mathcal{B}}_n$ cannot be computed as they depend on $\theta_\star$ (see (34)). However, we can approximate them using only θ_n in order to estimate iteratively $\theta_\star$ with (38). In the example of unconstrained gradient descent with step-size 1, using $\hat{\mathcal{A}}_n = I_q$ and $\hat{\mathcal{B}}_n = I_q - \partial^2 f(\theta_n)$ (see Example 2.1 in page 9) corresponds to Newton's method (see Example 3).

9.3 Non-convex constrained optimization

Considering Corollary 2 and Lemmas B.9 and B.11, the problem of finding the unique minimizer of non-convex functions can be brought down to finding β-smooth approximations of their biconjugates (for an arbitrary $\beta \in \mathbb{R}_+^*$).

A Proofs

A.1 Asymptotic convergence rate

Proof of Corollary 1. Let $\|\cdot\|$ be any norm on $\mathbb{R}^q$. Under (H3), the Taylor expansion with integral remainder yields

$$\begin{aligned}\mathcal{Q}_{\theta_n}(\theta_{n+1}) - \mathcal{Q}_{\theta_n}(\theta_\star) &= \left(\int_0^1 \partial_2 \mathcal{Q}_{\theta_n}(t\theta_{n+1} + (1-t)\theta_\star)\,\mathrm{d}t\right)(\theta_{n+1} - \theta_\star) \\ &= \partial_2 \mathcal{Q}_{\theta_\star}(\theta_\star)(\theta_{n+1} - \theta_\star) + o(\|\theta_{n+1} - \theta_\star\|) \end{aligned} \tag{39}$$

where the last equality follows from the continuity of the function $(\theta, \theta') \mapsto \partial_2 \mathcal{Q}_\theta(\theta')$. Moreover, since $\theta \mapsto \mathcal{Q}_\theta(\theta_\star)$ is differentiable at $\theta_\star$,

$$\mathcal{Q}_{\theta_n}(\theta_\star) - \mathcal{Q}_{\theta_\star}(\theta_\star) = \partial_1 \mathcal{Q}_{\theta_\star}(\theta_\star)(\theta_n - \theta_\star) + o(\|\theta_n - \theta_\star\|) \tag{40}$$

Summing (39) and (40) combined with Theorem 1 yield the expected result. □

Proof of Theorem 2. First, note that the theorem is proved if $\check{\rho}_\star = 0$. We now assume that $\check{\rho}_\star > 0$, which in particular implies that $\tilde{\mathcal{B}}_\star$ is invertible. In what follows, we use the notation introduced in Section 7. Following the proof of Theorem 1, see in particular the proof of Proposition 2, with the additional assumption that $\theta_\star \in \mathrm{ri}(\Theta)$, we can prove that for sufficiently large n, for all $\theta \in \Theta$,

$$\langle \partial_2 \mathcal{Q}_{\theta_n}(\theta_{n+1}), \theta - \theta_{n+1} \rangle = \langle \partial_2 \mathcal{Q}_{\theta_\star}(\theta_\star), \theta - \theta_\star \rangle = 0.$$

In other words, $\partial_2 \mathcal{Q}_{\theta_n}(\theta_{n+1}), \partial_2 \mathcal{Q}_{\theta_\star}(\theta_\star) \in \mathsf{V}^\perp$. This implies that for sufficiently large n,

$$\begin{aligned}\partial_2 \mathcal{Q}_{\theta_\star}(\theta_\star) - \partial_2 \mathcal{Q}_{\theta_n}(\theta_{n+1}) &= \partial_2 \mathcal{Q}_{\theta_\star}(\theta_\star) - \partial_2 \mathcal{Q}_{\theta_n}(\theta_\star) + \partial_2 \mathcal{Q}_{\theta_n}(\theta_\star) - \partial_2 \mathcal{Q}_{\theta_n}(\theta_{n+1}) \\ &= \mathcal{B}_n(\theta_n - \theta_\star) - \mathcal{A}_n(\theta_{n+1} - \theta_\star) \\ &= \mathcal{B}_n P \tilde{\Delta}_n - \mathcal{A}_n P \tilde{\Delta}_{n+1} \in \mathsf{V}^\perp, \end{aligned} \tag{41}$$

where $\mathcal{A}_n$,$\mathcal{B}_n$, P and $\tilde{\Delta}_n$ are defined respectively in (34), (23) and (26). As by definition of P, the condition $v \in \mathsf{V}^\perp$ is equivalent to the identity $P^\top v = 0$, we deduce from (41) that $P^T \mathcal{B}_n P \tilde{\Delta}_n = P^T \mathcal{A}_n P \tilde{\Delta}_{n+1}$, which can be written as

$$\tilde{\mathcal{B}}_n \tilde{\Delta}_n = \tilde{\mathcal{A}}_n \tilde{\Delta}_{n+1}. \tag{42}$$

Moreover, by (H4), $\tilde{\mathcal{A}}_\star$ is positive-definite and by (H2)-(H3), $\tilde{\mathcal{A}}_n$ is also positive-definite for sufficiently large n. Then, the invertibility of $\tilde{\mathcal{B}}_\star$ allows to write for sufficiently large n:

$$\left(\tilde{\mathcal{A}}_n^{-1/2} \tilde{\mathcal{B}}_n \tilde{\mathcal{A}}_n^{-1/2}\right)^{-1} \tilde{\mathcal{A}}_n^{1/2} \tilde{\Delta}_{n+1} = \tilde{\mathcal{A}}_n^{1/2} \tilde{\mathcal{B}}_n^{-1} \tilde{\mathcal{A}}_n \tilde{\Delta}_{n+1} = \tilde{\mathcal{A}}_n^{1/2} \tilde{\mathcal{B}}_n^{-1} \tilde{\mathcal{B}}_n \tilde{\Delta}_n = \tilde{\mathcal{A}}_n^{1/2} \tilde{\Delta}_n,$$

and thus, combining with $\|\tilde{\Delta}_n\|_{\tilde{\mathcal{A}}_n} = \|\tilde{\mathcal{A}}_n^{1/2} \tilde{\Delta}_n\|_2$ and $\|\tilde{\Delta}_{n+1}\|_{\tilde{\mathcal{A}}_n} = \|\tilde{\mathcal{A}}_n^{1/2} \tilde{\Delta}_{n+1}\|_2$, we get

$$\|\tilde{\Delta}_n\|_{\tilde{\mathcal{A}}_n} \leqslant \left|\!\left|\!\left| \left(\tilde{\mathcal{A}}_n^{-1/2} \tilde{\mathcal{B}}_n \tilde{\mathcal{A}}_n^{-1/2}\right)^{-1} \right|\!\right|\!\right|_2 \|\tilde{\Delta}_{n+1}\|_{\tilde{\mathcal{A}}_n}. \tag{43}$$

Besides, by the symmetry of $\tilde{\mathcal{A}}_\star^{-1/2} \tilde{\mathcal{B}}_\star \tilde{\mathcal{A}}_\star^{-1/2}$,

$$\begin{aligned}\left|\!\left|\!\left| \left(\tilde{\mathcal{A}}_\star^{-1/2} \tilde{\mathcal{B}}_\star \tilde{\mathcal{A}}_\star^{-1/2}\right)^{-1} \right|\!\right|\!\right|_2^{-1} &= \varrho\left(\left(\tilde{\mathcal{A}}_\star^{-1/2} \tilde{\mathcal{B}}_\star \tilde{\mathcal{A}}_\star^{-1/2}\right)^{-1}\right)^{-1} \\ &= \inf_{u \in \mathbb{R}^d} \frac{|u^\top \tilde{\mathcal{A}}_\star^{-1/2} \tilde{\mathcal{B}}_\star \tilde{\mathcal{A}}_\star^{-1/2} u|}{u^\top u} = \inf_{u \in \mathbb{R}^d} \frac{|u^\top \tilde{\mathcal{B}}_\star u|}{u^\top \tilde{\mathcal{A}}_\star u} = \check{\rho}_\star. \end{aligned} \tag{44}$$

Let $\rho \in (0; \check{\rho}_\star)$. Following the same steps as for the proof of Theorem 1, we can prove that $|||(\tilde{\mathcal{A}}_n^{-1/2}\tilde{\mathcal{B}}_n\tilde{\mathcal{A}}_n^{-1/2})^{-1}|||_2$ converges to $|||(\tilde{\mathcal{A}}_\star^{-1/2}\tilde{\mathcal{B}}_\star\tilde{\mathcal{A}}_\star^{-1/2})^{-1}|||_2$. Together with (43) and (44), this provides the existence of $n_0 \in \mathbb{N}$ such that for all $n \geqslant n_0$, $\|\tilde{\Delta}_n\|_{\tilde{\mathcal{A}}_\star} \leqslant \rho^{-1}\|\tilde{\Delta}_{n+1}\|_{\tilde{\mathcal{A}}_\star}$. We deduce by induction that for all $n \geqslant n_0$, $\|\tilde{\Delta}_{n_0}\|_{\tilde{\mathcal{A}}_\star} \leqslant \rho^{-(n-n_0)}\|\tilde{\Delta}_n\|_{\tilde{\mathcal{A}}_\star}$. As the sequence $(\theta_n)_{n\in\mathbb{N}}$ is not eventually equal to $\theta_\star$, by (28) we can choose n_0 such that $\|\tilde{\Delta}_{n_0}\|_{\tilde{\mathcal{A}}_\star} \neq 0$. Hence, since all the norms on a finite dimensional space are equivalent, we deduce $\liminf_{n\to\infty} \frac{1}{n}\log\left\|\tilde{\Delta}_n\right\|_2 = \liminf_{n\to\infty} \frac{1}{n}\log\left\|\tilde{\Delta}_n\right\|_{\tilde{\mathcal{A}}_\star} \geqslant \log\rho$. The proof is then concluded by applying (28) and by noting that ρ is arbitrary in $(0; \check{\rho}_\star)$. □

Proof of Theorem 3. In this proof, we use the notation introduced in Section 7. Define

$$\tilde{\mathcal{S}}_\star := \tilde{\mathcal{A}}_\star^{-1}\tilde{\mathcal{B}}_\star.$$

Note that $\tilde{\mathcal{S}}_\star$ is similar to the symmetric matrix $\tilde{\mathcal{A}}_\star^{-1/2}\tilde{\mathcal{B}}_\star\tilde{\mathcal{A}}_\star^{-1/2}$ and is therefore diagonalizable. Therefore, there exists an invertible matrix $R = [R(i,j)]_{1\leqslant i,j\leqslant d} \in \mathbb{R}^{d\times d}$ such that $\tilde{\mathcal{S}}_\star = \tilde{\mathcal{A}}_\star^{-1}\tilde{\mathcal{B}}_\star = R\tilde{\mathcal{D}}_\star R^{-1}$ where $\tilde{\mathcal{D}}_\star$ is a diagonal matrix. For any matrix $S \in \mathbb{R}^{d\times d}$, it is convenient to use the notation $S^R = R^{-1}SR$. In particular, we have $\tilde{\mathcal{S}}_\star^R = \tilde{\mathcal{D}}_\star$. Moreover, for any vector $\Delta \in \mathbb{R}^d$, we use the notation $\Delta^R := R^{-1}\Delta$. Write

$$\tilde{\mathcal{S}}_n := \tilde{\mathcal{A}}_n^{-1}\tilde{\mathcal{B}}_n,$$

where $\mathcal{A}_n$ and $\mathcal{B}_n$ are defined in (34) and $\tilde{\mathcal{S}}_n$ is well-defined for sufficiently large n as $\tilde{\mathcal{A}}_\star$ is positive-definite by (H4), and using (H2)-(H3). Besides, Theorems 1 and 2 provide $\check{\rho}_\star, \hat{\rho}_\star \in [0; 1)$. As the theorem is proved if $\check{\rho}_\star = 0$, we now assume that $\check{\rho}_\star > 0$, which implies the invertibility of $\tilde{\mathcal{B}}_\star$ and hence of $\tilde{\mathcal{S}}_\star$. Moreover, Theorem 1 also yields that for all $\rho \in (\hat{\rho}_\star; 1)$, $\theta_n - \theta_\star = \mathrm{o}(\rho^n)$. We deduce by the C^2-differentiability of $\partial_2\mathcal{Q}$ in a neighborhood of $(\theta_\star, \theta_\star)$ that for all $\rho \in (\hat{\rho}_\star; 1)$,

$$\left(\tilde{\mathcal{S}}_n^R\right)^{-1} - \tilde{\mathcal{D}}_\star^{-1} = \left(\tilde{\mathcal{S}}_n^R\right)^{-1} - \left(\tilde{\mathcal{S}}_\star^R\right)^{-1} = \mathrm{o}(\rho^n). \tag{45}$$

Following the proof of Theorem 2, by (42) there exists $n_0 \in \mathbb{N}$ such that for all $n \geqslant n_0$, $\tilde{\mathcal{B}}_n\tilde{\Delta}_n = \tilde{\mathcal{A}}_n\tilde{\Delta}_{n+1}$, that is, $\tilde{\mathcal{S}}_n\tilde{\Delta}_n = \tilde{\Delta}_{n+1}$ or equivalently $\tilde{\mathcal{S}}_n^R\tilde{\Delta}_n^R = \tilde{\Delta}_{n+1}^R$. This implies for all $m \in \mathbb{N}^*$,

$$\begin{aligned}\tilde{\Delta}_n^R &= \left(\tilde{\mathcal{S}}_n^R\right)^{-1}\tilde{\Delta}_{n+1}^R = \left[\prod_{k=0}^{m-1}\left(\tilde{\mathcal{S}}_{n+k}^R\right)^{-1}\right]\tilde{\Delta}_{n+m}^R, \\ &= \tilde{\mathcal{D}}_\star^{-m}\tilde{\Delta}_{n+m}^R + \sum_{k=0}^{m-1}\left\{\tilde{\mathcal{D}}_\star^{-m+1+k}\left[\left(\tilde{\mathcal{S}}_{n+k}^R\right)^{-1} - \tilde{\mathcal{D}}_\star^{-1}\right]\prod_{l=0}^{k-1}\left(\tilde{\mathcal{S}}_{n+l}^R\right)^{-1}\right\}\tilde{\Delta}_{n+m}^R.\end{aligned}$$

Component-wise, this yields for such $n, m \in \mathbb{N}^*$ that for all $i \in [\![1 : d]\!]$,

$$\tilde{\Delta}_n^R(i) = \tilde{\mathcal{D}}_\star(i,i)^{-m}\tilde{\Delta}_{n+m}^R(i) + \sum_{k=0}^{m-1}\tilde{\mathcal{D}}_\star(i,i)^{-m+1+k}L_{n,m,k}(i)^\top\tilde{\Delta}_{n+m}^R,$$

where for any $k \in [\![0 : m-1]\!]$, $L_{n,m,k}(i)^\top$ denotes the i-th row of the matrix

$$\left[\left(\tilde{\mathcal{S}}_{n+k}^R\right)^{-1} - \tilde{\mathcal{D}}_\star^{-1}\right]\prod_{l=0}^{k-1}\left(\tilde{\mathcal{S}}_{n+l}^R\right)^{-1}.$$

Recalling that $|||\cdot|||_F$ is the Frobenius norm, we let $C_F > 0$ be constant such that $|||\cdot|||_F \leqslant |||\cdot|||_2 C_F$ on $\mathbb{R}^{d\times d}$. Let $i \in [\![1:d]\!]$ such that $|\tilde{\mathcal{D}}_\star(i,i)| = \hat{\rho}_\star$. Using the Cauchy-Schwarz inequality we deduce

$$\begin{aligned}
|\tilde{\Delta}_n^R(i)| &\leqslant \hat{\rho}_\star^{-m}\|\tilde{\Delta}_{n+m}^R\|_2 + \sum_{k=0}^{m-1}\hat{\rho}_\star^{-m+1+k}\|L_{n,m,k}(i)\|_2\|\tilde{\Delta}_{n+m}^R\|_2,\\
&\leqslant \hat{\rho}_\star^{-m}\|\tilde{\Delta}_{n+m}^R\|_2 + \hat{\rho}_\star^{-m}\sum_{k=0}^{m-1}\hat{\rho}_\star^{k+1}\left|\!\left|\!\left|\left[\left(\tilde{\mathcal{S}}_{n+k+1}^R\right)^{-1} - \tilde{\mathcal{D}}_\star^{-1}\right]\prod_{l=0}^{k-1}\left(\tilde{\mathcal{S}}_{n+l}^R\right)^{-1}\right|\!\right|\!\right|_F \|\tilde{\Delta}_{n+m}^R\|_2,\\
&\leqslant \hat{\rho}_\star^{-m}\left[1 + C_F\sum_{k=0}^{m-1}\hat{\rho}_\star^{k+1}|||\left(\tilde{\mathcal{S}}_{n+k+1}^R\right)^{-1} - \tilde{\mathcal{D}}_\star^{-1}|||_2\prod_{l=0}^{k-1}|||\left(\tilde{\mathcal{S}}_{n+l}^R\right)^{-1}|||_2\right]\|\tilde{\Delta}_{n+m}^R\|_2. \qquad (46)
\end{aligned}$$

Let $\delta > \max(\hat{\rho}_\star^{-1}, \hat{\rho}_\star\check{\rho}_\star^{-1})$. Pick $\rho \in (\hat{\rho}_\star, 1)$ and $\varepsilon > 0$ such that $\rho(\check{\rho}_\star^{-1} + \varepsilon) < \delta$. By (45) there exists $C > 0$ and $n_1 \geqslant n_0$ such that for all $n \geqslant n_1$,

$$|||\left(\tilde{\mathcal{S}}_n^R\right)^{-1} - \tilde{\mathcal{D}}_\star^{-1}|||_2 \leqslant C\rho^n \leqslant \varepsilon. \qquad (47)$$

Then, (46) yields for all $n \geqslant n_1$ and $m \in \mathbb{N}^*$, using $\hat{\rho}_\star^{-1} < \delta$ and $\rho(\check{\rho}_\star^{-1} + \varepsilon) < \delta$,

$$\begin{aligned}
|\tilde{\Delta}_n^R(i)| &\leqslant \hat{\rho}_\star^{-m}\left[1 + CC_F\sum_{k=0}^{m-1}\hat{\rho}_\star^{k+1}\rho^{n+k+1}(\check{\rho}_\star^{-1}+\varepsilon)^k\right]\|\tilde{\Delta}_{n+m}^R\|_2\\
&= \left\{\hat{\rho}_\star^{-m} + CC_F\rho^{n+1}\sum_{k=0}^{m-1}\hat{\rho}_\star^{k+1-m}\left[\rho(\check{\rho}_\star^{-1}+\varepsilon)\right]^k\right\}\|\tilde{\Delta}_{n+m}^R\|_2\\
&\leqslant \left(\delta^m + CC_F\rho^{n+1}\sum_{k=0}^{m-1}\delta^{m-1-k}\delta^k\right)\|\tilde{\Delta}_{n+m}^R\|_2\\
&= \left(\delta^m + CC_F\rho^{n+1}\delta^{m-1}m\right)\|\tilde{\Delta}_{n+m}^R\|_2. \qquad (48)
\end{aligned}$$

We now show that there exists $n \geqslant n_1$ such that $\tilde{\Delta}_n^R(i) = R^{-1}P^T\Delta_n(i) \neq 0$. Indeed, otherwise, by Lemma A.1 below, there exists a basis $w_1,\dots,w_d$ of V such that $\Delta_n = \theta_n - \theta_\star \in \mathrm{Span}(w_j, j \in [\![1:d]\!]\backslash\{i\})$ for any $n \geqslant n_1$ which contradicts the assumption $\mathrm{Span}(\Delta_n, n \geqslant n_1) = \mathsf{V}$ in the statement of Theorem 3. Therefore, we can choose $n \geqslant n_1$ such that the lhs of (48) is strictly positive. This n being chosen, take the log in the previous inequality and divide by m. Letting m goes to infinity, we then obtain for any $\delta > \max(\hat{\rho}_\star^{-1}, \hat{\rho}_\star\check{\rho}_\star^{-1})$,

$$\begin{aligned}
\log(\delta^{-1}) &\leqslant \liminf_{m\to\infty}\frac{1}{m}\log\|\tilde{\Delta}_m^R\|_2 \leqslant \liminf_{m\to\infty}\frac{1}{m}\log|||R^{-1}|||_2 \times \|\tilde{\Delta}_m\|_2\\
&= \liminf_{m\to\infty}\frac{1}{m}\log\|\tilde{\Delta}_m\|_2 = \liminf_{m\to\infty}\frac{1}{m}\log\|\Delta_m\|_2,
\end{aligned}$$

where the last equality follows from (28). The proof is completed since δ is arbitrary provided that $\delta > \max(\hat{\rho}_\star^{-1}, \hat{\rho}_\star\check{\rho}_\star^{-1})$, which is equivalent to $\delta^{-1} < \min(\hat{\rho}_\star, \hat{\rho}_\star^{-1}\check{\rho}_\star)$. □

Lemma A.1. *Let V be a d-dimensional linear subspace of $\mathbb{R}^q$. Let $v_1,\dots,v_d \in \mathbb{R}^q$ be an orthonormal basis of V and let $w_1,\dots,w_d \in \mathbb{R}^q$ be another basis obtained from $(v_i)_{1\leqslant i\leqslant d}$ by the change-of-basis matrix R, that is, for any $j \in [\![1:d]\!]$, $w_j = \sum_{i=1}^d R(i,j)v_i$.*

Then, for any $\Delta \in \mathsf{V}$, the i-th component of the decomposition of Δ on the basis $(w_i)_{1 \leqslant i \leqslant d}$ is $R^{-1}P^T\Delta(i)$, where P is the matrix

$$P := [v_1|\dots|v_d] \in \mathbb{R}^{q \times d}.$$

Proof. Decomposing the vector $\Delta \in \mathsf{V}$ on the basis $(v_j)_{1 \leqslant j \leqslant d}$ and using $v_j = \sum_{i=1}^d R^{-1}(i,j)w_i$, we get

$$\Delta = \sum_{j=1}^d (v_j^T\Delta)v_j = \sum_{i=1}^d \left[\sum_{j=1}^d R^{-1}(i,j)(v_j^T\Delta)\right] w_i = \sum_{i=1}^d [R^{-1}P^T\Delta(i)]w_i$$

□

A.2 Comments on (H1)

Proof of Theorem 4. The proof amounts to building an analogous framework that satisfies (H1)-(H4). The first step is to define a suitable minimization set that is convex. Write d the dimension of the submanifold S and for all $x \in \mathbb{R}^q$, $R > 0$, we set $\mathbf{B}(x,R) := \{y \in \mathbb{R}^q \ : \ \|x-y\|_2 < R\}$.

Under (H'1) there exist U_1, U_2 two open neighborhoods of $\theta_\star$ and the null-vector $\mathbf{0}$ in $\mathbb{R}^q$, respectively, and a C^2-diffeomorphism $\psi\colon \mathsf{U}_1 \to \mathsf{U}_2$ such that $\psi(\theta_\star) = \mathbf{0}$ and $\psi(\mathsf{U}_1 \cap \mathsf{S}) = \mathsf{U}_2 \cap (\mathbb{R}^d \times \{0\}^{q-d})$. Let N be an open neighborhood of $\theta_\star$ such that $\mathcal{Q}$ meets the conditions of (H3) on $\mathsf{N} \times \mathsf{N}$. Define $\mathsf{U} := \mathsf{U}_1 \cap \mathsf{N} \cap \mathring{\mathsf{E}}$, which is an open set containing $\theta_\star$ by (H'1). Set $r > 0$ such that $\mathbf{B}(\theta_\star, r) \subset \mathsf{U}$, and $\varepsilon > 0$ such that $\mathbf{B}(\mathbf{0}, \varepsilon) \subset \psi(\mathbf{B}(\theta_\star, r/2))$. Define $\mathsf{W} := \mathbb{R}^d \times \{0\}^{q-d}$ and the convex set

$$\Xi := \mathbf{B}(0, \varepsilon) \cap \mathsf{W}. \tag{49}$$

Write ϕ the corestriction of ψ to Ξ, that is, $\phi\colon \psi^{-1}(\Xi) \to \Xi$ such that $\phi(x) = \psi(x)$ for any $x \in \psi^{-1}(\Xi)$. Note that ϕ is still a C^2-diffeomorphism. Define then

$$\begin{aligned} &\mathcal{R}\colon \Xi \times \Xi \longrightarrow \mathbb{R} \\ &\qquad (\zeta, \zeta') \mapsto \mathcal{R}_\zeta(\zeta') := \mathcal{Q}_{\phi^{-1}(\zeta)}\left(\phi^{-1}(\zeta')\right). \end{aligned} \tag{50}$$

Under (H2) there exists $n_0 \in \mathbb{N}$ such that $\theta_n \in \phi^{-1}(\Xi)$ for all $n \geqslant n_0$, which allows to define on Ξ the sequence

$$(\zeta_n)_{n \in \mathbb{N}} := (\phi(\theta_{n+n_0}))_{n \in \mathbb{N}}. \tag{51}$$

We deduce from the definition of $(\theta_n)_{n \in \mathbb{N}}$ in (1) and from $\phi^{-1}(\Xi) \subset \mathsf{U} \subset \mathsf{E}$ that for all $n \in \mathbb{N}$ and $\zeta \in \Xi$,

$$\mathcal{R}_{\zeta_n}(\zeta_{n+1}) = \mathcal{Q}_{\theta_{n+n_0}}(\theta_{n+n_0+1}) \leqslant \mathcal{Q}_{\theta_{n+n_0}}\left(\phi^{-1}(\zeta)\right) = \mathcal{R}_{\zeta_n}(\zeta). \tag{52}$$

The framework defined by (49-51) thus fits into (1) and meets (H1)-(H3) with $(\theta_n, \theta_\star, \mathcal{Q})$ replaced by $(\zeta_n, \mathbf{0}, \mathcal{R})$. For consistency of notation, we write $\zeta_\star := \phi(\theta_\star) = \mathbf{0}$. We now prove that (H4) is satisfied with $(\theta_\star, \mathcal{Q})$ replaced by $(\zeta_\star, \mathcal{R})$.

Denote by J_ϕ the Jacobian matrix of ϕ. The fact that $\phi^{-1}(\Xi) \subset \mathsf{U} \subset \mathsf{N}$ allows to write for all $\theta, \theta' \in \phi^{-1}(\Xi)$,

$$\partial_2 \mathcal{Q}_\theta(\theta') = J_\phi(\theta')^\top \partial_2 \mathcal{R}_{\phi(\theta)}(\phi(\theta')) = \sum_{i=1}^d \left[\partial_2 \mathcal{R}_{\phi(\theta)}(\phi(\theta'))\right]_i \cdot \partial\phi_i(\theta'),$$

using that the image of ϕ is included in $\mathbb{R}^d \times \{0\}^{q-d}$, i.e. $\phi_i \equiv 0$ for all $i \in [\![d+1 : q]\!]$. This yields

$$\partial_{12}\mathcal{Q}_{\theta_\star}(\theta_\star) = J_\phi(\theta_\star)^\top \times \partial_{12}\mathcal{R}_{\zeta_\star}(\zeta_\star) \times J_\phi(\theta_\star), \tag{53}$$

$$\partial_{22}\mathcal{Q}_{\theta_\star}(\theta_\star) = J_\phi(\theta_\star)^\top \times \partial_{22}\mathcal{R}_{\zeta_\star}(\zeta_\star) \times J_\phi(\theta_\star) + \sum_{i=1}^{d} [\partial_2 \mathcal{R}_{\zeta_\star}(\zeta_\star)]_i \cdot \partial^2 \phi_i(\theta_\star). \tag{54}$$

Besides, (52) and Lemma 7.1 provide that $\zeta_\star$ is a local minimizer of the function $\zeta \mapsto \mathcal{R}_{\zeta_\star}(\zeta)$. Together with the convexity of Ξ, the fact that $\zeta_\star \in \mathrm{ri}(\Xi)$ and that $\mathrm{Aff}(\Xi) = \mathsf{W}$, this implies $\partial_2 \mathcal{R}_{\zeta_\star}(\zeta_\star) \in \mathsf{W}^\perp$. As $\mathsf{W} = \mathbb{R}^d \times \{0\}^{q-d}$, the second term of the rhs in (54) is thus null. Combining with $\mathsf{T}_\star = J_\phi(\theta_\star)^{-1}\mathsf{W}$ by definition of the tangent space of a submanifold, we deduce from (53-54) that the rates $\check{\rho}_\star, \hat{\rho}_\star$ defined in (11-12) for $\mathcal{R}$ are equal to the rates $\check{\rho}_\star, \hat{\rho}_\star$ defined in (13) for $\mathcal{Q}$. Satisfying (H4) for $\mathcal{R}$ is thus equivalent to satisfying (H'4) for $\mathcal{Q}$.

Therefore, we can apply Theorems 1 and 2 to the sequence $(\zeta_n)_{n\in\mathbb{N}}$ and we get for any $(\rho_1, \rho_2) \in (\hat{\rho}_\star, 1) \times (0, \check{\rho}_\star)$

$$\zeta_n - \zeta_\star = \mathrm{o}(\rho_1^n) \quad \text{and} \quad \rho_2^n = \mathrm{o}(\|\zeta_n - \zeta_\star\|_2). \tag{55}$$

To relate with the speed of convergence of $\theta_n - \theta_\star$, note that for all $n \in \mathbb{N}$,

$$\begin{aligned} \zeta_n - \zeta_\star &= \phi(\theta_{n+n_0}) - \phi(\theta_\star) = \psi(\theta_{n+n_0}) - \psi(\theta_\star), \\ \theta_{n+n_0} - \theta_\star &= \phi^{-1}(\zeta_n) - \phi^{-1}(\zeta_\star), \end{aligned} \tag{56}$$

and that $\psi^{-1}(\Xi) \subset \mathbf{B}(\theta_\star, r/2)$ by (49). The C^1-differentiability of ψ on $\bar{\mathbf{B}}(\theta_\star, r/2)$ and of ψ^{-1} on $\overline{\Xi}$ provides the existence of $C > 0$ such that $\sup_{\theta\in\mathbf{B}(\theta_\star, r/2)} \|\mathrm{d}_\psi(\theta)\|_2 \leqslant C$ and $\sup_{\zeta\in\Xi} \|\mathrm{d}_{\phi^{-1}}(\zeta)\|_2 \leqslant C$, where d_ψ and $\mathrm{d}_{\phi^{-1}}$ denote the differentials of ψ and ϕ^{-1}, respectively. Thus ψ is Lipshitz on $\mathbf{B}(\theta_\star, r/2)$ and ϕ^{-1} on Ξ. Combining with (56) and (55) concludes the proof. □

A.3 Comments on (H2)

We first prove a general version of Proposition 2.

Proposition 3. *Assume that* (H1), (H2.4), (H3) *and* (H4) *hold. Let* $\|\cdot\|$ *be a norm on* $\mathbb{R}^q$*. Then, for all* $\rho > \hat{\rho}_\star$*, there exists* $\delta > 0$ *such that for all* $\theta \in \Theta$*,* $\theta' \in \mathcal{M}(\theta)$*,*

$$\|\theta - \theta_\star\| \vee \|\theta' - \theta_\star\| \leqslant \delta \implies \|\tilde{\theta}' - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_\star} \leqslant \rho \|\tilde{\theta} - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_\star},$$

where the notation $\tilde{\theta}$*,* $\tilde{\theta}'$ *and* $\tilde{\theta}_\star$ *are defined in* (24)*.*

Proof. In this proof, we use the notation introduced in Section 7. For all $\delta > 0$, write $\mathbf{B}(\theta_\star, \delta) := \{\theta \in \mathbb{R}^q : \|\theta - \theta_\star\| < \delta\}$. Under (H3) there exists $\delta_0 > 0$ such that $\partial_2 \mathcal{Q}$ is well-defined and C^1-differentiable on $\mathbf{B}(\theta_\star, \delta_0) \times \mathbf{B}(\theta_\star, \delta_0)$. By (H1), (H2.4) and (H3), we can prove, similarly to (31) in the proof of Proposition 2, that for all $\theta, \theta' \in \mathbf{B}(\theta_\star, \delta_0)$ with $\theta' \in \mathcal{M}(\theta)$,

$$\langle \partial_2 \mathcal{Q}_\theta(\theta') - \partial_2 \mathcal{Q}_\theta(\theta_\star), \theta' - \theta_\star \rangle \leqslant \langle \partial_2 \mathcal{Q}_{\theta_\star}(\theta_\star) - \partial_2 \mathcal{Q}_\theta(\theta_\star), \theta' - \theta_\star \rangle,$$

which yields

$$(\theta' - \theta_\star)^\top \mathcal{A}_{\theta\theta'}(\theta' - \theta_\star) \leqslant (\theta' - \theta_\star)^\top \mathcal{B}_{\theta\theta'}(\theta - \theta_\star), \tag{57}$$

where

$$\mathcal{A}_{\theta\theta'} := \int_0^1 \partial_{22}\mathcal{Q}_\theta(s\theta' + (1-s)\theta_\star)\,\mathrm{d}s, \quad \mathcal{B}_{\theta\theta'} := -\int_0^1 \partial_{12}\mathcal{Q}_{s\theta_\star + (1-s)\theta}(\theta_\star)\,\mathrm{d}s.$$

Under (H3)-(H4) there also exists $\delta_1 \in (0;\delta_0)$ such that if $\theta,\theta' \in \mathbf{B}(\theta_\star,\delta_1)$, then the symmetric matrix $\tilde{\mathcal{A}}_{\theta\theta'}$ is positive-definite. Using (27) and applying Lemma B.1 to (57) then provides

$$\|\tilde{\theta}' - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_{\theta\theta'}} \leqslant \hat{\rho}_{\theta\theta'} \|\tilde{\theta} - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_{\theta\theta'}}, \tag{58}$$

where $\hat{\rho}_{\theta\theta'} := \|\tilde{\mathcal{A}}_{\theta\theta'}^{-1/2} \tilde{\mathcal{B}}_{\theta\theta'} \tilde{\mathcal{A}}_{\theta\theta'}^{-1/2}\|_2$.

Let $\rho > \hat{\rho}_\star$. Set $\rho' := (\rho + \hat{\rho}_\star)/2$ and $\varepsilon > 0$ such that $(1+\varepsilon)\rho' \leqslant (1-\varepsilon)\rho$. By (H3)-(H4) and Lemma B.4 there exists $\delta_2 \in (0;\delta_1)$ such that for all $\theta,\theta' \in \mathbf{B}(\theta_\star,\delta_2)$, for all $u \in \mathbb{R}^d$,

$$(1-\varepsilon)\|u\|_{\tilde{\mathcal{A}}_{\theta\theta'}} \leqslant \|u\|_{\tilde{\mathcal{A}}_\star} \leqslant (1+\varepsilon)\|u\|_{\tilde{\mathcal{A}}_{\theta\theta'}}.$$

Moreover, by Lemmas B.2 and B.3, under (H3) there exists $\delta_3 \in (0;\delta_2)$ such that $\hat{\rho}_{\theta\theta'} \leqslant \rho'$ for all $\theta,\theta' \in \mathbf{B}(\theta_\star,\delta_3)$. Combining with (58) yields that for all $\theta,\theta' \in \mathbf{B}(\theta_\star,\delta_3)$ with $\theta' \in \mathcal{M}(\theta)$,

$$\begin{aligned}\|\tilde{\theta}' - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_\star} \leqslant (1+\varepsilon)\|\tilde{\theta}' - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_{\theta\theta'}} &\leqslant (1+\varepsilon)\hat{\rho}_{\theta\theta'}\|\tilde{\theta} - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_{\theta\theta'}} \\ &\leqslant \frac{(1+\varepsilon)\rho'}{1-\varepsilon}\|\tilde{\theta} - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_{\theta\theta'}} \leqslant \rho\|\tilde{\theta} - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_\star}.\end{aligned}$$

□

Proof of Theorem 5. Applying Proposition 3 to $\rho := (1+\hat{\rho}_\star)/2$ and the norm $\|\cdot\|_2$ provides the existence of $\delta_0 > 0$ such that for all $\theta \in \Theta$, $\theta' \in \mathcal{M}(\theta)$,

$$\|\theta - \theta_\star\|_2 \vee \|\theta' - \theta_\star\|_2 \leqslant \delta_0 \implies \|\tilde{\theta}' - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_\star} \leqslant \rho\|\tilde{\theta} - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_\star}. \tag{59}$$

Moreover, by Lemma B.6 under (H2.1)-(H2.2) and by (H2.4), there exists $\delta_1 \in (0;\delta_0)$ such that for all $\theta \in \Theta$, $\theta' \in \mathcal{M}(\theta)$,

$$\|\theta - \theta_\star\|_2 \leqslant \delta_1 \implies \|\theta' - \theta_\star\|_2 \leqslant \delta_0.$$

By (28) and the equivalence of norms in finite dimension, combining with (59) yields the existence of $\delta > 0$ such that for all $\theta \in \Theta$, $\theta' \in \mathcal{M}(\theta)$,

$$\|\tilde{\theta} - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_\star} \leqslant \delta \implies \|\tilde{\theta}' - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_\star} \leqslant \rho\|\tilde{\theta} - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_\star}.$$

Besides, by (H2.3) there exists $n_0 \in \mathbb{N}$ such that $\|\tilde{\theta}_{n_0} - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_\star} \leqslant \delta$. Using that $\rho < 1$ by (H4), we deduce by induction that for all $n \geqslant n_0$, $\|\tilde{\theta}_n - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_\star} \leqslant \delta$, and that

$$\|\tilde{\theta}_{n+1} - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_\star} \leqslant \rho\|\tilde{\theta}_n - \tilde{\theta}_\star\|_{\tilde{\mathcal{A}}_\star},$$

which concludes the proof. □

Lemma A.2. *Assume that $\partial^2\mathcal{Q}$ is well-defined and differentiable on Θ, and that for all $\theta,\theta' \in \Theta$, for all $v \in \mathsf{V}$, $v^\top \partial_{12}\mathcal{Q}_\theta(\theta')\, v < 0$. Then, for all $\theta,\theta' \in \Theta$,*

$$\mathcal{M}(\theta) \cap \mathcal{M}(\theta') \cap \mathrm{ri}(\Theta) \neq \emptyset \implies \theta = \theta'.$$

Proof. Let $\theta'' \in \mathcal{M}(\theta) \cap \mathcal{M}(\theta') \cap \mathrm{ri}(\Theta)$. We can prove as for Theorem 2 that it implies $\tilde{\mathcal{B}}_{\theta\theta'}(\tilde{\theta} - \tilde{\theta}') = \tilde{\mathcal{A}}_{\theta\theta'}(\tilde{\theta}'' - \tilde{\theta}'') = 0$, where

$$\tilde{\mathcal{A}}_{\theta\theta'} = \partial_{22}\tilde{\mathcal{Q}}_\theta(\theta'') \quad \text{and} \quad \tilde{\mathcal{B}}_{\theta\theta'} = -\int_0^1 \partial_{12}\tilde{\mathcal{Q}}_{s\theta'+(1-s)\theta}(\theta'')\,\mathrm{d}s.$$

Besides, by assumption, for all $v \in \mathsf{V}$, $v^\top \mathcal{B}_{\theta\theta'} v = -\int_0^1 v^\top \partial_{12}\mathcal{Q}_{s\theta'+(1-s)\theta}(\theta'')\, v\mathrm{d}s > 0$. This provides the invertibilty of $\tilde{\mathcal{B}}_{\theta\theta'}$ and thus $\tilde{\theta} - \tilde{\theta}' = 0$, which concludes the proof by (28). □

Proof of Proposition 1. It is proved in [Bubeck, 2015, Theorem 4.4, p.305] that under assumptions (ii) and (iii), for all $\theta \in \mathsf{C} \cap \mathsf{D}$ and $n \geqslant 1$,

$$\frac{1}{n}\sum_{i=0}^{n-1} f(\zeta_i) - f(\theta) \leqslant \frac{1}{n}\frac{1}{\eta} D_\Phi(\theta, \zeta_0).$$

We deduce by applying Lemma B.5 under (iv) and the compacity of C that there exists $\varphi\colon \mathbb{N} \to \mathbb{N}$ strictly increasing such that $(\zeta_{\varphi(n)})_{n\in\mathbb{N}}$ converges to $\theta_\star$. By (9) this is equivalent to $(\mathcal{M}(\theta_{\varphi(n)-1}))_{n\in\mathbb{N}^*}$ converging to $\theta_\star$. Besides, (iv) and the differentiability of f provide $\mathcal{M}(\theta_\star) = \{\theta_\star\}$ (see Example 2.1 in page 7), and by Lemma B.7 under (H2.1)-(H2.2) the function $\mathcal{M}$ is continuous on Θ. We deduce that all accumulation points ℓ of the sequence $(\theta_{\varphi(n)-1})_{n\in\mathbb{N}^*}$ verify $\mathcal{M}(\ell) = \theta_\star = \mathcal{M}(\theta_\star)$. By Lemma A.2 under (i), (ii), (iii) and (v) (using (68-69)), this yields $\ell = \theta_\star$ for all accumulation points, and thus the convergence of $(\theta_{\varphi(n)-1})_{n\in\mathbb{N}^*}$ to $\theta_\star$ by the compacity of C.

Using that $\mathcal{M}(\theta_\star) = \{\theta_\star\}$, we can prove as in Example 2.1, page 7, that $\mathcal{M}^m(\theta_\star) = \{\theta_\star\}$, where $\mathcal{M}^m$ is the minimization mapping corresponding to mirror prox (see (10)). □

Proof of Theorem 6. Under (H2.3) there exists $\psi\colon \mathbb{N} \to \mathbb{N}$ strictly increasing such that $(\theta_{\psi(n)})_{n\in\mathbb{N}}$ converges to $\theta_\star$. By the compacity of $\Theta \times \Theta$ under (H2.1) there also exist $\theta_{\star\star} \in \Theta$ and $\varphi\colon \mathbb{N} \to \mathbb{N}$ strictly increasing such that

$$(\theta_{\varphi(n)}, \theta_{\varphi(n)+1}) \underset{n\to\infty}{\longrightarrow} (\theta_\star, \theta_{\star\star}). \tag{60}$$

By the monotonicity of the sequence $(\vartheta(\theta_n))_{n\in\mathbb{N}}$ under ($\tilde{\text{H}}$4.2), for all $n \in \mathbb{N}$, $\vartheta(\theta_{\varphi(n+1)}) \leqslant \vartheta(\theta_{\varphi(n)+1}) \leqslant \vartheta(\theta_{\varphi(n)})$. Together with (60) and the continuity of ϑ this yields

$$\vartheta(\theta_{\star\star}) = \vartheta(\theta_\star).$$

Besides, by the definition of $(\theta_n)_{n\in\mathbb{N}}$, for all $\theta \in \Theta$,

$$\mathcal{Q}_{\theta_{\varphi(n)}}(\theta_{\varphi(n)+1}) \leqslant \mathcal{Q}_{\theta_{\varphi(n)}}(\theta).$$

Using the continuity of $\mathcal{Q}$ under (H2.2), this yields $\theta_{\star\star} \in \mathcal{M}(\theta_\star)$. We deduce under ($\tilde{\text{H}}$4.2) that $\theta_\star = \theta_{\star\star}$, and hence $\mathcal{M}(\theta_\star) = \{\theta_\star\}$ under ($\tilde{\text{H}}$4.1). □

A.4 Comments on (H4)

Proof for Example 1.1 (Population EM). By (5), for all $\theta, \theta' \in \Theta$,

$$\mathcal{Q}_\theta^{\mathrm{pop}}(\theta') = -\int_{\mathsf{Y}}\int_{\mathsf{X}} p_\theta(x|y) \log p_{\theta'}(x,y) p_{\theta_\star}(y) \mu(\mathrm{d}x)\mu(\mathrm{d}y),$$

which yields, in a neighborhood of $(\theta_\star, \theta_\star)$,

$$\partial_{22}\mathcal{Q}_\theta^{\mathrm{pop}}(\theta') = -\int_{\mathsf{Y}}\int_{\mathsf{X}} p_\theta(x|y) p_{\theta_\star}(y) \partial^2 \log p_{\theta'}(x,y) \mu(\mathrm{d}x)\mu(\mathrm{d}y).$$

On the other hand, using that for all $\theta, \theta' \in \Theta$,

$$\mathcal{Q}_\theta^{\mathrm{pop}}(\theta') = -\int_{\mathsf{Y}}\int_{\mathsf{X}} p_\theta(x|y) \log p_{\theta'}(x|y) p_{\theta_\star}(y) \mu(\mathrm{d}x)\mu(\mathrm{d}y) - \int_{\mathsf{Y}} \log p_{\theta'}(y) p_{\theta_\star}(y) \mu(\mathrm{d}y),$$

we deduce that in a neighborhood of $(\theta_\star, \theta_\star)$,

$$\begin{aligned}\partial_{12}\mathcal{Q}_\theta^{\mathrm{pop}}(\theta') &= -\int_{\mathsf{Y}}\int_{\mathsf{X}} p_{\theta_\star}(y)\partial \log p_{\theta'}(x|y)\left[\partial p_\theta(x|y)\right]^\top \mu(\mathrm{d}x)\mu(\mathrm{d}y) \\ &= -\int_{\mathsf{Y}}\int_{\mathsf{X}} p_{\theta_\star}(y)p_\theta(x|y)\partial \log p_{\theta'}(x|y)\left[\partial \log p_\theta(x|y)\right]^\top \mu(\mathrm{d}x)\mu(\mathrm{d}y).\end{aligned}$$

Using the chain rule for Fisher information matrices [Zamir, 1998, Zegers, 2015],

$$\partial_{22}\mathcal{Q}_{\theta_\star}^{\mathrm{pop}}(\theta_\star) = -\int_{\mathsf{Y}}\int_{\mathsf{X}} p_{\theta_\star}(x,y)\partial^2 \log p_{\theta_\star}(x,y)\mu(\mathrm{d}x)\mu(\mathrm{d}y) = I_{X,Y}(\theta_\star), \tag{61}$$

$$\partial_{12}\mathcal{Q}_{\theta_\star}^{\mathrm{pop}}(\theta_\star) = -I_{X|Y}(\theta_\star) = -\left(I_{X,Y}(\theta_\star) - I_Y(\theta_\star)\right). \tag{62}$$

□

Proofs for Example 1.2 (Sample EM). Similarly to the proof of Example 1.1, we deduce from (4) that for all θ, θ' in a neighborhood of $\theta_\star$,

$$\begin{aligned}\partial_{22}\mathcal{Q}_\theta^{\mathrm{samp}}(\theta') &= -\frac{1}{k}\sum_{i=1}^k \int_{\mathsf{X}} p_\theta(x|Y_i)\partial^2 \log p_{\theta'}(x,Y_i)\mu(\mathrm{d}x), \\ \partial_{12}\mathcal{Q}_\theta^{\mathrm{samp}}(\theta') &= -\frac{1}{k}\sum_{i=1}^k \int_{\mathsf{X}} p_\theta(x|Y_i)\partial \log p_{\theta'}(x|Y_i)\left[\partial \log p_\theta(x|Y_i)\right]^\top \mu(\mathrm{d}x),\end{aligned}$$

and therefore,

$$\partial_{22}\mathcal{Q}_{\theta_\star}^{\mathrm{samp}}(\theta_\star) = -\frac{1}{k}\sum_{i=1}^k \int_{\mathsf{X}} p_{\theta_\star}(x|Y_i)\partial^2 \log p_{\theta_\star}(x,Y_i)\mu(\mathrm{d}x), \tag{63}$$

$$\partial_{12}\mathcal{Q}_{\theta_\star}^{\mathrm{samp}}(\theta_\star) = \frac{1}{k}\sum_{i=1}^k \int_{\mathsf{X}} p_{\theta_\star}(x|Y_i)\partial^2 \log p_{\theta_\star}(x|Y_i)\mu(\mathrm{d}x). \tag{64}$$

Lemma A.3. *Assume that $\theta_\star$ is the true parameter of the model, that for all $x, y \in \mathsf{X}, \mathsf{Y}$, the functions $\theta \mapsto p_\theta(x|y)$ and $\theta \mapsto p_\theta(y)$ are twice differentiable in a neighborhood of $\theta_\star$, and that conditions similar to [Douc et al., 2013, Assumption AD.1, p.492] hold to differentiate under the integral sign. Then, if the corresponding population EM meets* (H4)*, almost surely the sample EM meets* (H4) *for sufficiently large k and*

$$\hat{\rho}_\star^{\mathrm{samp}}(Y_{1:k}) \xrightarrow{a.s.} \hat{\rho}_\star^{\mathrm{pop}}.$$

Furthermore, if $\partial^2 \log p_{\theta_\star}(X_1, Y_1), \partial^2 \log p_{\theta_\star}(Y_1) \in \mathrm{L}^2(\mathbb{R}^{q\times q})$, then for all $\delta \in (0;1)$ there exists $C_\delta > 0$ such that

$$\liminf_{k\to\infty} \mathbb{P}\left(|\hat{\rho}_\star^{\mathrm{samp}}(Y_{1:k}) - \hat{\rho}_\star^{\mathrm{pop}}| \leqslant \frac{C_\delta}{\sqrt{k}}\right) \geqslant 1-\delta.$$

Proof. As $\hat{\rho}_\star^{\mathrm{samp}}(Y_{1:k})$ is not necessarily well-defined in (11), using Lemma B.3 we consider the following definition:

$$\hat{\rho}_\star(Y_{1:k}) := \mathbb{1}_{\tilde{\mathcal{A}}_\star(Y_{1:k})\succ 0}\left|\!\left|\!\left|\tilde{\mathcal{A}}_\star(Y_{1:k})^{-1/2}\,\tilde{\mathcal{B}}_\star(Y_{1:k})\,\tilde{\mathcal{A}}_\star(Y_{1:k})^{-1/2}\right|\!\right|\!\right|_2. \tag{65}$$

Note first that $\hat{\rho}_\star$ is a measurable function of $Y_{1:k}$. Besides, we deduce from the assumptions that the following random variables are integrable:

$$Z_1 := \int_{\mathsf{X}} p_{\theta_\star}(x|Y_1)\partial^2 \log p_{\theta_\star}(x|Y_1)\mu(\mathrm{d}x), \quad W_1 := \int_{\mathsf{X}} p_{\theta_\star}(x|Y_1)\partial^2 \log p_{\theta_\star}(x, Y_1)\mu(\mathrm{d}x).$$

Together with (61-62) and (63-64), the strong law of large numbers provides

$$\left(\partial_{22}\mathcal{Q}^{\mathrm{samp}}_{\theta_\star}(\theta_\star), \partial_{12}\mathcal{Q}^{\mathrm{samp}}_{\theta_\star}(\theta_\star)\right) \xrightarrow{a.s.} \left(\partial_{22}\mathcal{Q}^{\mathrm{pop}}_{\theta_\star}(\theta_\star), \partial_{12}\mathcal{Q}^{\mathrm{pop}}_{\theta_\star}(\theta_\star)\right). \tag{66}$$

By the continuity of the functions used in (65), this yields that if the corresponding population EM meets (H4), then, almost surely, the sample EM meets (H4) for sufficiently large k, and

$$\hat{\rho}^{\mathrm{samp}}_\star(Y_{1:k}) \xrightarrow{a.s.} \hat{\rho}^{\mathrm{pop}}_\star.$$

Assume now that $\partial^2 \log p_{\theta_\star}(X_1, Y_1), \partial^2 \log p_{\theta_\star}(Y_1) \in \mathrm{L}^2(\mathbb{R}^{q\times q})$. First, using Jensen's inequality with $\|\cdot\|_2^2$ provides $W_1 \in \mathrm{L}^2(\mathbb{R}^{q\times q})$, and thus $Z_1 = W_1 - \partial^2 \log p_{\theta_\star}(Y_1) \in \mathrm{L}^2(\mathbb{R}^{q\times q})$. Let $\delta \in (0;1)$ and write $\bar{Z}_k = \sum_{i=1}^k Z_i/k$. Set $\delta' \in (0;1)$ such that $4q^2\delta' \leqslant \delta$ and write $x(\delta')$ the quantile of order $1-\delta'$ of the standard Gaussian distribution. Applying the central limit theorem to each component of Z_1 provides for all $i, j \in [\![1 : q]\!]$ and $k \in \mathbb{N}^*$,

$$\mathbb{P}\left(\left|\bar{Z}_k(i,j) - \mu_{ij}\right| \geqslant \frac{x(\delta')\sigma}{\sqrt{k}}\right) \leqslant \mathbb{P}\left(\left|\bar{Z}_k(i,j) - \mu_{ij}\right| \geqslant \frac{x(\delta')\sigma_{ij}}{\sqrt{k}}\right) \underset{k\to\infty}{\longrightarrow} 2\delta',$$

where $\mu_{ij} := \mathbb{E}[Z_1(i,j)]$, $\sigma_{ij}^2 := \mathrm{Var}\left[Z_1(i,j)\right]$ and $\sigma^2 := \mathbb{E}[|||Z_1|||_F^2] \vee \mathbb{E}[|||Z_2|||_F^2]$. This yields

$$\limsup_{k\to\infty} \mathbb{P}\left(\left|\left|\left|\partial_{12}\mathcal{Q}^{\mathrm{samp}}_{\theta_\star}(\theta_\star) - \partial_{12}\mathcal{Q}^{\mathrm{pop}}_{\theta_\star}(\theta_\star)\right|\right|\right|_F \geqslant \frac{qx(\delta')\sigma}{\sqrt{k}}\right) \leqslant 2q^2\delta'.$$

Similarly, the same inequality holds for $\partial_{22}\mathcal{Q}^{\mathrm{samp}}_{\theta_\star}(\theta_\star)$. Let $C_F > 0$ such that $|||\cdot|||_F \geqslant C_F|||\cdot|||_2$ on $\mathbb{R}^{q\times q}$. We deduce using (27) and (28) that

$$\liminf_{k\to\infty} \mathbb{P}\left(\left|\left|\left|\partial_{22}\tilde{\mathcal{Q}}^{\mathrm{samp}}_{\theta_\star}(\theta_\star) - \partial_{22}\tilde{\mathcal{Q}}^{\mathrm{pop}}_{\theta_\star}(\theta_\star)\right|\right|\right|_2 \wedge \left|\left|\left|\partial_{12}\tilde{\mathcal{Q}}^{\mathrm{samp}}_{\theta_\star}(\theta_\star) - \partial_{12}\tilde{\mathcal{Q}}^{\mathrm{pop}}_{\theta_\star}(\theta_\star)\right|\right|\right|_2 < \frac{qx(\delta')\sigma}{C_F\sqrt{k}}\right)$$
$$\geqslant 1 - 4q^2\delta' \geqslant 1 - \delta. \tag{67}$$

Let $\varepsilon > 0$ and $C > 0$ be constants obtained by applying Lemma B.2 to $(\partial_{22}\tilde{\mathcal{Q}}^{\mathrm{pop}}_{\theta_\star}(\theta_\star), \partial_{12}\tilde{\mathcal{Q}}^{\mathrm{pop}}_{\theta_\star}(\theta_\star))$. We deduce from (67) that

$$\liminf_{k\to\infty} \mathbb{P}\left(|\hat{\rho}^{\mathrm{samp}}_\star(Y_{1:k}) - \hat{\rho}^{\mathrm{pop}}_\star| < \frac{2Cqx(\delta')\sigma}{C_F}\frac{1}{\sqrt{k}}\right) \geqslant 1 - \delta,$$

which proves (16). □

□

Proof for Example 2.1 in page 9 (Mirror descent, cont.) For all $\theta, \theta' \in \Theta$ in a neighborhood of $\theta_\star$,

$$\begin{aligned}
\mathcal{Q}_\theta(\theta') &= \eta\partial f(\theta)^\top\theta' + \Phi(\theta') - \Phi(\theta) - \partial\Phi(\theta)^\top(\theta' - \theta),\\
\partial_2\mathcal{Q}_\theta(\theta') &= \eta\partial f(\theta) + \partial\Phi(\theta') - \partial\Phi(\theta),\\
\partial_{22}\mathcal{Q}_\theta(\theta') &= \partial^2\Phi(\theta'), &(68)\\
\partial_{12}\mathcal{Q}_\theta(\theta') &= \eta\partial^2 f(\theta) - \partial^2\Phi(\theta). &(69)
\end{aligned}$$

□

Proof for Example 1.3 (The α-EM algorithm). Let $\alpha \in \mathbb{R} \setminus \{0,1\}$. The function f_α is defined on $\mathbb{R}_+^*$ by $f_\alpha \colon x \mapsto (1-x^\alpha)/(\alpha(\alpha-1))$. Note that $f_\alpha(1) = 0$ and that for all differentiable functions g taking values in $\mathbb{R}_+^*$,

$$\partial\left(f_\alpha \circ g\right) = \frac{1}{1-\alpha}\left[1+\alpha(1-\alpha) f_\alpha \circ g\right] \partial\left(\log \circ g\right).$$

The function $\mathcal{Q}^\alpha$ defined in (19) can be written as $\mathcal{Q}_\theta^\alpha\left(\theta'\right) = -\int_{\mathsf{X}} p_\theta(x|Y) F_\theta^\alpha(\theta') \mu(\mathrm{d}x)$, where $F^\alpha \colon (\theta,\theta') \mapsto f_\alpha\left(p_{\theta'}(x,Y)/p_\theta(x,Y)\right)$. For all $\theta \in \Theta$, $F_\theta^\alpha(\theta) = 0$, and under the assumptions of Example 1.1 with f_α instead of f_0, for all $\theta, \theta' \in \Theta$,

$$\begin{aligned}
\partial_2 F_\theta^\alpha(\theta') &= \frac{1}{1-\alpha}\left[1+\alpha(1-\alpha) F_\theta^\alpha(\theta')\right] \partial_{\theta'} \log p_{\theta'}(x,Y), \\
\partial_1 F_\theta^\alpha(\theta') &= -\frac{1}{1-\alpha}\left[1+\alpha(1-\alpha) F_\theta^\alpha(\theta')\right] \partial_{\theta} \log p_{\theta}(x,Y).
\end{aligned}$$

We deduce

$$\begin{aligned}
\partial_{22} F_\theta^\alpha(\theta') = \frac{1}{1-\alpha}\left[1+\alpha(\alpha-1) F_\theta^\alpha(\theta')\right] \big(&\partial_{\theta'\theta'} \log p_{\theta'}(x,Y) \\
&+ \alpha \partial_{\theta'} \log p_{\theta'}(x,Y) \left[\partial_{\theta'} \log p_{\theta'}(x,Y)\right]^\top \big), \\
\partial_{12} F_\theta^\alpha(\theta') = -\frac{\alpha}{1-\alpha}\left[1+\alpha(\alpha-1) F_\theta^\alpha(\theta')\right] &\partial_{\theta'} \log p_{\theta'}(x,Y) \left[\partial_{\theta} \log p_{\theta}(x,Y)\right]^\top.
\end{aligned}$$

At a population level this yields $\partial_{22} \mathcal{Q}_{\theta_\star}^\alpha\left(\theta_\star\right) = I_{X,Y}(\theta_\star)$ and

$$\partial_{12} \mathcal{Q}_{\theta_\star}^\alpha\left(\theta_\star\right) = \frac{\alpha}{1-\alpha} I_{X,Y}(\theta_\star) - \frac{1}{1-\alpha} I_{X|Y}(\theta_\star) = -I_{X,Y}(\theta_\star) + \frac{1}{1-\alpha} I_Y(\theta_\star).$$

Regarding the value of $\hat{\rho}_\star^\alpha$, note that $\tilde{\mathcal{A}}_\star^{-1}\tilde{\mathcal{B}}_\star$ is equivalent to the symmetric matrix $\tilde{\mathcal{A}}_\star^{-1/2}\tilde{\mathcal{B}}_\star\tilde{\mathcal{A}}_\star^{-1/2}$ and is therefore diagonalizable. Besides, we deduce from (14) and (20) that

$$\tilde{\mathcal{A}}_\star^\alpha = \tilde{\mathcal{A}}_\star \quad \text{and} \quad \tilde{\mathcal{B}}_\star^\alpha = \frac{1}{1-\alpha}\tilde{\mathcal{B}}_\star - \frac{\alpha}{1-\alpha}\tilde{\mathcal{A}}_\star.$$

This yields $\mathrm{Spec}((\tilde{\mathcal{A}}_\star^\alpha)^{-1}\tilde{\mathcal{B}}_\star^\alpha) = g_\alpha(\mathrm{Spec}(\tilde{\mathcal{A}}_\star^{-1}\tilde{\mathcal{B}}_\star))$ where $g_\alpha(x) := (x-\alpha)/(1-\alpha)$. We obtain the optimal α by equating $g_\alpha(\hat{\rho}_\star) = -g_\alpha(\check{\rho}_\star)$. □

Proof for Example 2.2 in page 10 (Mirror prox, cont.) To begin with, the assumptions imply that the corresponding mirror descent meets (H3)-(H4) and (H2.1)-(H2.2), and that $\theta_\star = \mathcal{M}\left(\theta_\star\right) \in \mathrm{ri}(\Theta)$. Together with the fact that the mapping is point-to-point on Θ, (see Example 2.1 in page 7), this allows to apply Lemma B.8 to mirror descent. We deduce the C^1-differentiability of $\mathcal{M}$ in a neighborhood of $\theta_\star$, where we can write

$$\begin{aligned}
\mathcal{Q}_\theta^m\left(\theta'\right) &= \eta \partial f\left(\mathcal{M}\left(\theta\right)\right)^\top \theta' + \Phi(\theta') - \Phi(\theta) - \partial\Phi(\theta)^\top(\theta'-\theta), \\
\partial_2 \mathcal{Q}_\theta^m\left(\theta'\right) &= \eta \partial f\left(\mathcal{M}\left(\theta\right)\right) + \partial\Phi(\theta') - \partial\Phi(\theta), \\
\partial_{22} \mathcal{Q}_\theta^m\left(\theta'\right) &= \partial^2\Phi(\theta'), \\
\partial_{12} \mathcal{Q}_\theta^m\left(\theta'\right) &= -\eta \partial^2 f(\mathcal{M}\left(\theta\right)) P \left[\partial^2\tilde{\Phi}(\mathcal{M}\left(\theta\right))\right]^{-1} P^\top \left[\eta\partial^2 f(\theta) - \partial^2\Phi(\theta)\right] - \partial^2\Phi(\theta).
\end{aligned}$$

This yields $\mathcal{A}_\star^m = \partial^2\Phi(\theta_\star)$ and $\mathcal{B}_\star^m = \partial^2\Phi(\theta_\star) - \eta\partial^2 f(\theta_\star) P\tilde{\mathcal{A}}_\star^{-1}P^\top\mathcal{B}_\star$. Regarding the value of $\hat{\rho}_\star^m$, note that $\tilde{\mathcal{A}}_\star^{-1}\tilde{\mathcal{B}}_\star$ is equivalent to the symmetric matrix $\tilde{\mathcal{A}}_\star^{-1/2}\tilde{\mathcal{B}}_\star\tilde{\mathcal{A}}_\star^{-1/2}$ and is therefore diagonalizable. Together with (22) this yields

$$\mathrm{Spec}((\tilde{\mathcal{A}}_\star^m)^{-1}\tilde{\mathcal{B}}_\star^m) = f\left(\mathrm{Spec}(\tilde{\mathcal{A}}_\star^{-1}\tilde{\mathcal{B}}_\star)\right),$$

where f is defined by $f(x) := x^2 - x + 1$. Note that $|f(x)| < 1$ if and only if $x \in (0;1)$. Besides, $\mathrm{Spec}(\tilde{\mathcal{A}}_\star^{-1}\tilde{\mathcal{B}}_\star) \subset \mathbb{R}_+^*$ if and only if $u^\top\tilde{\mathcal{A}}_\star^{-1/2}\tilde{\mathcal{B}}_\star\tilde{\mathcal{A}}_\star^{-1/2}u > 0$ for all $u \in \mathbb{R}^d$, which is equivalent to $u^\top\tilde{\mathcal{B}}_\star u > 0$ for all $u \in \mathbb{R}^d$, by the symmetry of $\tilde{\mathcal{A}}_\star^{-1/2}$. Finally, $\min_{\mathbb{R}} f = 3/4$, which is attained at $x = 1/2$, and for all $x \in (0;1)$, $f(x) > x$. □

B Technical results

All lemmas below are proved in the Supplementary material (C.1).

B.1 Linear algebra

Lemma B.1. *Let $d \in \mathbb{N}^*$ and $A, B \in \mathbb{R}^{d\times d}$ such that A is symmetric positive-definite. Then, for all $x, y \in \mathbb{R}^d$, $x^\top Ax \leqslant x^\top By \implies \|x\|_A \leqslant \rho\,\|y\|_A$, where $\rho := \|\!|A^{-1/2}BA^{-1/2}|\!\|_2$.*

Lemma B.2. *Let $d \in \mathbb{N}^*$ and $A, B \in \mathbb{R}^{d\times d}$ such that A is symmetric positive-definite. Then, there exist $\varepsilon > 0$ and $C > 0$ such that for all symmetric matrices $M \in \mathbb{R}^{d\times d}$ and for all matrices $N \in \mathbb{R}^{d\times d}$ verifying $\|\!|M|\!\|_2 \vee \|\!|N|\!\|_2 \leqslant \varepsilon$,*

$$\left|\|\!|A^{-1/2}BA^{-1/2}|\!\|_2 - \|\!|(A+M)^{-1/2}(B+N)(A+M)^{-1/2}|\!\|_2\right| \leqslant C\|\!|M|\!\|_2 + C\|\!|N|\!\|_2.$$

Lemma B.3. *Let $d \in \mathbb{N}^*$, $A \in \mathbb{R}^{d\times d}$ be a symmetric positive-definite matrix, and $B \in \mathbb{R}^{d\times d}$ be a symmetric matrix. Then,*

$$\varrho\left(A^{-1}B\right) = \varrho\left(BA^{-1}\right) = \left|\!\left|\!\left|A^{-1/2}BA^{-1/2}\right|\!\right|\!\right|_2 = \sup_{v\in\mathbb{R}^d}\frac{|v^\top Bv|}{v^\top Av}.$$

Lemma B.4. *Let $d \in \mathbb{N}^*$ and $S_\star \in \mathbb{R}^{d\times d}$ be a symmetric positive-definite matrix. Then, for all $\varepsilon > 0$, there exists $\delta > 0$ such that for all symmetric matrices $S \in \mathbb{R}^{d\times d}$ verifying $\|\!|S - S_\star|\!\|_2 < \delta$, and all $x \in \mathbb{R}^d$,*

$$(1-\varepsilon)\,\|x\|_S \leqslant \|x\|_{S_\star} \leqslant (1+\varepsilon)\,\|x\|_S.$$

B.2 Minimization

Lemma B.5. *Let (K, d) be a compact metric space and $f\colon \mathsf{K} \to \mathbb{R}$ be a continuous function. Write $m := \min_K f$ and $\mathsf{M} := \mathrm{argmin}_K f$. Then, for all $\delta > 0$ there exists $\varepsilon > 0$ such that for all $x \in \mathsf{K}$,*

$$f(x) - m < \varepsilon \implies \mathrm{d}(x, \mathsf{M}) < \delta.$$

Lemma B.6. *Let (K, d) be a compact metric space and $\mathcal{Q}\colon \mathsf{K}\times\mathsf{K} \to \mathbb{R}$ be a continuous function. Define for all $x \in \mathsf{K}$, $\mathcal{M}(x) := \mathrm{argmin}_{x'\in\mathsf{K}}\mathcal{Q}_x(x')$. Then, for all $x_\star \in \mathsf{K}$ and $\delta > 0$, there exists $\delta' > 0$ such that for all $x \in \mathsf{K}$,*

$$\mathrm{d}(x, x_\star) < \delta' \implies \sup_{x'\in\mathcal{M}(x)} \mathrm{d}(x', \mathcal{M}(x_\star)) \leqslant \delta.$$

Lemma B.7 (Continuity of the minimization mapping)**.** *Let* (K, d) *be a compact metric space and* $\mathcal{Q}\colon \mathsf{K} \times \mathsf{K} \to \mathbb{R}$ *be a continuous function such that for all* $x \in \mathsf{K}$, $\mathcal{M}(x) := \operatorname{argmin}_{x' \in \mathsf{K}} \mathcal{Q}_x(x')$ *is a singleton. Then, the function* $\mathcal{M}$ *is continuous on* K.

Lemma B.8 (Differentiability of the minimization mapping)**.** *Let* $d \in \mathbb{N}^*$, K *be a compact set of* $\mathbb{R}^d$, *and* $\mathcal{Q}\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *be a function continuous on* $\mathsf{K} \times \mathsf{K}$ *such that for all* $x \in \mathsf{K}$, $\mathcal{M}(x) := \operatorname{argmin}_{x' \in \mathsf{K}} \mathcal{Q}_x(x')$ *is a singleton. Let* $x \in \mathsf{K}$ *such that:*

(i) $x, \mathcal{M}(x) \in \mathrm{ri}(\mathsf{K})$,

(ii) $\partial_2 \mathcal{Q}$ *is well-defined and* C^k*-differentiable in a neighborhood of* $(x, \mathcal{M}(x))$ *for* $k \in \mathbb{N}^*$,

(iii) $\partial_{22} \tilde{\mathcal{Q}}_x(\mathcal{M}(x))$ *is invertible.*

Then, the function $\mathcal{M}$ *is* C^k*-differentiable in a neighborhood of* x *(considering that the domain lies in the ambient space* $\mathrm{Aff}(\mathsf{K})$*).*

Lemma B.9. *Let* (K, d) *be a compact metric space and* $f\colon \mathsf{K} \to \mathbb{R}$ *be a continuous function. Then, for all* $\delta > 0$, *there exists* $\varepsilon > 0$ *such that for all functions* $\hat{f}\colon \mathsf{K} \to \mathbb{R}$,

$$\sup_{\mathsf{K}} \left| \hat{f} - f \right| < \varepsilon \implies \sup_{y \in \hat{\mathsf{M}}} \mathrm{d}(y, \mathsf{M}) \leqslant \delta,$$

with $\hat{\mathsf{M}} := \operatorname{argmin}_{\mathsf{K}} \hat{f}$, *and the convention that the supremum over an empty set is equal to minus infinity.*

Lemma B.10. *Let* (K, d) *be a compact metric space and* $\mathcal{Q}\colon \mathsf{K} \times \mathsf{K} \to \mathbb{R}$ *be a continuous function such that for all* $x \in \mathsf{K}$, $\mathcal{M}(x) := \operatorname{argmin}_{x' \in \mathsf{K}} \mathcal{Q}_x(x')$ *is a singleton. Then, for all* $\delta > 0$, *there exists* $\varepsilon > 0$ *such that for all functions* $\hat{\mathcal{Q}}\colon \mathsf{K} \times \mathsf{K} \to \mathbb{R}$,

$$\sup_{\mathsf{K} \times \mathsf{K}} \left| \hat{\mathcal{Q}} - \mathcal{Q} \right| < \varepsilon \implies \sup_{x \in \mathsf{K}\ ;\ y \in \hat{\mathcal{M}}(x)} \mathrm{d}(y, \mathcal{M}(x)) \leqslant \delta,$$

where $\hat{\mathcal{M}}$ *is defined by* $\hat{\mathcal{M}}(x) := \operatorname{argmin}_{x' \in \mathsf{K}} \hat{\mathcal{Q}}_x(x')$, *and with the convention that the supremum over an empty set is equal to minus infinity.*

B.3 Convexity

Lemma B.11. *Let* $d \in \mathbb{N}^*$, K *be a bounded convex set of* $\mathbb{R}^d$, *and* $f\colon \mathsf{K} \to \mathbb{R}$ *be a continuous function. Assume that* f *has a unique minimizer* $x_\star$ *on* K, *that* $x_\star \in \mathrm{ri}(K)$, *that* f *is* C^2*-differentiable in a neighborhood of* $x_\star$ *and that* $\partial^2 f(x_\star) \succ 0$.

Then, $x_\star$ *is the unique minimizer of* f^{**} *on* K *and* f^{**} *is equal to* f *in a neighborhood of* $x_\star$, *where* f^{**} *denotes the biconjugate of* f *(see [Rockafellar and Wets, 1998, Section 11, p.473]).*

References

[Balakrishnan et al., 2017] Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017). Statistical guarantees for the EM algorithm: from population to sample-based analysis. *Ann. Statist.*, 45:77–120.

[Bauschke, 1997] Bauschke, H. H.and Borwein, J. M. (1997). Legendre functions and the method of random Bregman projections. *J. Convex Anal.*, 4:27–67.

[Bubeck, 2015] Bubeck, S. (2015). Convex optimization: algorithms and complexity. *Found. Trends Mach. Learn.*, 8:231–357.

[Cappé et al., 2005] Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York.

[Cesa-Bianchi and Lugosi, 2006] Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.

[Daudel et al., 2020] Daudel, K., Douc, R., and Portier, F. (2020). Infinite-dimensional gradient-based descent for alpha-divergence. *arXiv:2005.10618v2*.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. A*, 39:1–22.

[Douc et al., 2013] Douc, R., Moulines, E., and Stoffer, D. S. (2013). *Nonlinear time series. Theory, methods, and applications with R examples*. Chapman & Hall/CRC Texts Stat. Sci. Ser.

[Kunstner et al., 2021] Kunstner, F., Kumar, R., and Schmidt, M. (2021). Homeomorphic-invariance of em: Non-asymptotic convergence in kl divergence for exponential families via mirror descent. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3295–3303. PMLR.

[Lange, 1995] Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 57:425–437.

[Matsuyama, 2003] Matsuyama, Y. (2003). The α-EM algorithm: surrogate likelihood maximization using α-logarithmic information measures. *IEEE Trans. Inform. Theory*, 49:692–706.

[Meng and Rubin, 1991] Meng, X.-L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *J. Roy. Statist. Soc.*, 86:899–909.

[Meng and Rubin, 1993] Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80:267—278.

[Meng and Rubin, 1994] Meng, X.-L. and Rubin, D. B. (1994). On the global and componentwise rates of convergence of the EM algorithm. *Linear Algebra Appl.*, 199:413–425.

[Nocedal and J., 2006] Nocedal, O. and J., W. S. (2006). *Numerical optimization. Second edition. Springer series in operations research and financial engineering*. Springer, New York.

[Orchard and Woodbury, 1972] Orchard, T. and Woodbury, M. A. (1972). A missing information principle: Theory and applications. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Theory of Statistics*, 1:697–715.

[Rockafellar and Wets, 1998] Rockafellar, R. T. and Wets, R. J.-B. (1998). *Variational analysis*. Springer-Verlag, Berlin.

[Tao, 2012] Tao, T. (2012). *Topics in random matrix theory. Graduate studies in mathematics, 132*. American Mathematical Society, Providence.

[Wu, 1983] Wu, C.-F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.*, 11:95–103.

[Zamir, 1998] Zamir, R. (1998). A proof of the Fisher information inequality via a data processing argument. *IEEE Trans. Inform. Theory*, 44:1246—-1250.

[Zegers, 2015] Zegers, P. (2015). Fisher information properties. *Entropy*, 17:4918–4939.

C Supplementary material

C.1 Linear algebra

Proof of Lemma B.1. Let $x, y \in \mathbb{R}^d$ such that $0 < x^\top Ax \leqslant x^\top By$. The Cauchy-Schwarz inequality provides

$$\|x\|_A^2 = x^\top Ax \leqslant x^\top By = (A^{1/2}x)^\top A^{-1/2}By \leqslant \|A^{1/2}x\|_2 \|A^{-1/2}By\|_2.$$

Using that $\|A^{-1/2}By\|_2 \leqslant |||A^{-1/2}BA^{-1/2}|||_2 \|A^{1/2}y\|_2$, we deduce

$$\|x\|_A^2 \leqslant \rho \|A^{1/2}x\|_2 \|A^{1/2}y\|_2 = \rho \|x\|_A \|y\|_A ,$$

where $\rho = |||A^{-1/2}BA^{-1/2}|||_2$. □

Lemma C.1. *Let $d \in \mathbb{N}^*$ and $A \in \mathbb{R}^{d\times d}$ be a symmetric positive-definite matrix. Then, there exist $\varepsilon > 0$ and $C > 0$ such that for all symmetric matrices $M \in \mathbb{R}^{d\times d}$ verifying $|||M|||_2 \leqslant \varepsilon$, the matrix $A + M$ is symmetric positive-definite and*

$$\left|\left|\left| A^{1/2} - (A+M)^{1/2} \right|\right|\right|_2 \leqslant C|||M|||_2.$$

Proof. Write $\lambda := \min \mathrm{Spec}(A) > 0$. Let $M \in \mathbb{R}^{d\times d}$ be a symmetric matrix such that $|||M|||_2 \leqslant \lambda/2$. The matrix $A + M$ is then symmetric positive-definite and we can define its square root. By the symmetry of $A_M := (A+M)^{1/2} - A^{1/2}$, there exist $\mu \in \mathrm{Spec}(A_M)$ such that $|\mu| = |||A_M|||_2$ and $x \in \mathbb{R}^{d\times d}$ such that $A_M x = \mu x$ and $\|x\|_2 = 1$. This yields

$$\begin{aligned} x^\top Mx &= x^\top (A+M)^{1/2} \left((A+M)^{1/2} - A^{1/2}\right) x + x^\top \left((A+M)^{1/2} - A^{1/2}\right) A^{1/2}x \\ &= \mu x^\top \left((A+M)^{1/2} + A^{1/2}\right) x. \end{aligned}$$

We deduce, using that $x^T(A+M)^{1/2}x \geqslant 0$ and $\lambda^{1/2} = \min \mathrm{Spec}(A^{1/2})$,

$$|||M|||_2 \geqslant |x^\top Mx| = |\mu| x^\top \left((A+M)^{1/2} + A^{1/2}\right) x \geqslant |||A_M|||_2 x^\top A^{1/2}x \geqslant |||A_M|||_2 \lambda^{1/2},$$

and hence the result with $\varepsilon = \lambda/2$ and $C = \lambda^{-1/2}$. □

Lemma C.2. *Let $d \in \mathbb{N}^*$ and $A \in \mathbb{R}^{d\times d}$ be an invertible matrix. Then, there exist $\varepsilon > 0$ and $C > 0$ such that for all matrices $M \in \mathbb{R}^{d\times d}$ verifying $|||M|||_2 \leqslant \varepsilon$,*

$$\left|\left|\left| A^{-1} - (A+M)^{-1} \right|\right|\right|_2 \leqslant C|||M|||_2.$$

Proof. Let $M \in \mathbb{R}^{d\times d}$ such that $|||M|||_2 \leqslant |||A^{-1}|||_2^{-1}/2$. We can then write

$$A^{-1} - (A+M)^{-1} = A^{-1} - A^{-1}(I + MA^{-1})^{-1} = -A^{-1}MA^{-1}\sum_{k=0}^{\infty}(-MA^{-1})^k.$$

This yields the result with $\varepsilon = |||A^{-1}|||_2^{-1}/2$ and $C = 2|||A^{-1}|||_2^2$. □

Proof of Lemma B.2. Applying Lemma C.1 to A^{-1} and Lemma C.2 to A provides the existence of $\varepsilon > 0$ and $C > 1$ such that for all symmetric matrices $S \in \mathbb{R}^{d\times d}$ verifying $|||S|||_2 \leqslant \varepsilon$, the matrix $A^{-1} + S$ is symmetric positive-definite and

$$\left|\left|\left|\left(A^{-1}\right)^{1/2} - \left(A^{-1}+S\right)^{1/2}\right|\right|\right|_2 \leqslant C|||S|||_2, \tag{70}$$

$$\left|\left|\left|A^{-1} - (A+S)^{-1}\right|\right|\right|_2 \leqslant C|||S|||_2. \tag{71}$$

Let $M \in \mathbb{R}^{d\times d}$ be a symmetric matrix such that $|||M|||_2 \leqslant \varepsilon/C \leqslant \varepsilon$. By (71) we can then define $M' := A^{-1} - (A+M)^{-1}$, which verifies $|||M'|||_2 \leqslant C|||M|||_2 \leqslant \varepsilon$. Choosing $S = -M'$, we deduce that $A^{-1} - M' = (A+M)^{-1}$ is symmetric positive-definite and by (70),

$$\begin{aligned}\left|\left|\left|(A)^{-1/2} - (A+M)^{-1/2}\right|\right|\right|_2 &= \left|\left|\left|\left(A^{-1}\right)^{1/2} - \left((A+M)^{-1}\right)^{1/2}\right|\right|\right|_2 \\ &= \left|\left|\left|\left(A^{-1}\right)^{1/2} - \left(A^{-1}-M'\right)^{1/2}\right|\right|\right|_2 \\ &\leqslant C|||M'|||_2 \leqslant C^2|||M|||_2.\end{aligned}$$

This concludes the proof up to simple algebra, noting that for all matrices $N \in \mathbb{R}^{d\times d}$,

$$\begin{aligned}A^{-1/2}BA^{-1/2} &- (A+M)^{-1/2}(B+N)(A+M)^{-1/2} \\ &= \left(A^{-1/2} - (A+M)^{-1/2}\right)BA^{-1/2} - (A+M)^{-1/2}NA^{-1/2} \\ &\quad + (A+M)^{-1/2}(B+N)\left(A^{-1/2} - (A+M)^{-1/2}\right).\end{aligned}$$

□

Proof of Lemma B.3. The symmetry of A^{-1} and B provides the first equality. As $A^{-1}B = A^{-1/2}(A^{-1/2}BA^{-1/2})A^{1/2}$, $A^{-1}B$ and $A^{-1/2}BA^{-1/2}$ are similar and then

$$\varrho(A^{-1}B) = \varrho(A^{-1/2}BA^{-1/2}).$$

Since $A^{-1/2}BA^{-1/2}$ is symmetric, we can write

$$\begin{aligned}\varrho\left(A^{-1/2}BA^{-1/2}\right) &= \max\left|\mathrm{Spec}(A^{-1/2}BA^{-1/2})\right| \\ &= \sup_{x\in\mathbb{R}^d}\frac{\left|x^\top A^{-1/2}BA^{-1/2}x\right|}{x^\top x} = \left|\left|\left|A^{-1/2}BA^{-1/2}\right|\right|\right|_2.\end{aligned}$$

This provides the second equality, along with the third one by considering the change of variables $v = A^{-1/2}x$. □

Proof of Lemma B.4. For all $x \in \mathbb{R}^d$, we have

$$\|x\|_2 \leqslant |||S_\star^{-1/2}|||_2 \left\|S_\star^{1/2} x\right\|_2 = |||S_\star^{-1/2}|||_2 \, \|x\|_{S_\star}$$

This implies

$$|\, \|x\|_S^2 - \|x\|_{S_\star}^2 \,| = |\langle x, (S - S_\star)x\rangle| \leqslant |||S - S_\star|||_2 \, \|x\|_2^2 \leqslant \|x\|_{S_\star}^2 \, |||S - S_\star|||_2 |||S_\star^{-1/2}|||_2^2,$$

which yields, for all S such that $|||S - S_\star|||_2 \leqslant \delta$,

$$\left(1 + \delta |||S_\star^{-1/2}|||_2^2\right)^{-1/2} \|x\|_S \leqslant \|x\|_{S_\star} \leqslant \left(1 - \delta |||S_\star^{-1/2}|||_2^2\right)^{-1/2} \|x\|_S \,.$$

The proof follows. □

C.2 Minimization

Proof of Lemma B.5. Let $\delta > 0$. The function $\tilde{\mathrm{d}}$ defined on K by $\tilde{\mathrm{d}}(x) := \mathrm{d}(x, \mathsf{M})$ being continuous, the set $\tilde{\mathsf{K}} := \tilde{\mathrm{d}}^{-1}([\delta, +\infty[)$ is compact, as the intersection of a closed set with a compact set. By the continuity of f we deduce the existence of $x_0 \in \tilde{\mathsf{K}}$ such that

$$\varepsilon := \inf_{x \in \tilde{\mathsf{K}}} f - m = f(x_0) - m > 0.$$

□

Proof of Lemma B.6. Let $x_\star \in \mathsf{K}$ and $\delta > 0$. By the compacity of K and the continuity of $\mathcal{Q}_{x_\star}(\cdot)$, there exists $x_{\star\star} \in \mathcal{M}(x_\star)$. Besides, Lemma B.5 applied to $\mathcal{Q}_{x_\star}(\cdot)$ provides the existence of $\varepsilon > 0$ such that for all $x' \in \mathsf{K}$,

$$\mathcal{Q}_{x_\star}(x') - \mathcal{Q}_{x_\star}(x_{\star\star}) < \varepsilon \implies \mathrm{d}(x', \mathcal{M}(x_\star)) < \delta. \tag{72}$$

Moreover, by the uniform continuity of $\mathcal{Q}$ on $(\mathsf{K} \times \mathsf{K}, \tilde{\mathrm{d}})$, where $\tilde{\mathrm{d}}((y, y'), (z, z')) := \mathrm{d}(y, z) + \mathrm{d}(y', z')$, there exists $\delta' > 0$ such that for all $y, y', z, z' \in \mathsf{K}$,

$$\tilde{\mathrm{d}}((y, y'), (z, z')) < \delta' \implies \left|\mathcal{Q}_y(y') - \mathcal{Q}_z(z')\right| < \varepsilon/2.$$

We deduce that for all $x \in \mathsf{K}$ such that $\mathrm{d}(x, x_\star) < \delta'$, for all $x' \in \mathcal{M}(x)$,

$$\mathcal{Q}_{x_\star}(x') < \mathcal{Q}_x(x') + \varepsilon/2 \leqslant \mathcal{Q}_x(x_{\star\star}) + \varepsilon/2 < \mathcal{Q}_{x_\star}(x_{\star\star}) + \varepsilon,$$

and thus $\mathrm{d}(x', \mathcal{M}(x_\star)) < \delta$ by (72). □

Proof of Lemma B.7. This is a direct consequence of Lemma B.6.

Proof of Lemma B.8. We first prove the lemma for $k = 1$. Write V the direction of $\mathrm{Aff}(\mathsf{K})$ and for all $v \in \mathsf{V}, \varepsilon > 0$, write $\mathbf{B}(v, \varepsilon) := \{w \in \mathsf{V} \,:\, \|v - w\|_2 < \varepsilon\}$. Note that the ball is defined as a subset of V.

Under (i) and (ii) there exists $\varepsilon_0 > 0$ such that $\partial_2 \mathcal{Q}$ is C^1-differentiable on $\mathbf{B}(x, 2\varepsilon_0) \times \mathbf{B}(\mathcal{M}(x), 2\varepsilon_0) \subset \mathrm{ri}(\mathsf{K}) \times \mathrm{ri}(\mathsf{K})$. Moreover, the function $\mathcal{M}$ is continuous on K by Lemma B.7, which yields the existence of $\varepsilon_1 \in (0; \varepsilon_0)$ such that $y \in \mathbf{B}(x, \varepsilon_1)$ implies $\mathcal{M}(y) \in \mathbf{B}(\mathcal{M}(x), \varepsilon_0)$. By (iii) there also exists $\varepsilon_2 \in (0; \varepsilon_1)$ such that for all $y \in \mathbf{B}(x, \varepsilon_2)$, the matrix $\partial_{22} \tilde{\mathcal{Q}}_y(\mathcal{M}(y))$ is invertible.

Let $y \in \mathbf{B}(x, \varepsilon_2)$ and set $\varepsilon > 0$ such that $\mathbf{B}(y, \varepsilon) \subset \mathbf{B}(x, \varepsilon_2)$. For all $h \in \mathbf{B}(0, \varepsilon)$, the fact that $\mathcal{M}(y), \mathcal{M}(y+h) \in \mathrm{ri}(\mathsf{K})$ provides (see the proof of Theorem 2):

$$\tilde{\mathcal{A}}(h)\left(\tilde{\mathcal{M}}(y+h) - \tilde{\mathcal{M}}(y)\right) = \tilde{\mathcal{B}}(h)\tilde{h},$$

where the functions $\mathcal{A}$ and $\mathcal{B}$ are defined on $\mathbf{B}(0, \varepsilon)$ by

$$\mathcal{A}(h) := \int_0^1 \partial_{22}\mathcal{Q}_{y+h}(s\mathcal{M}(y+h) + (1-s)\mathcal{M}(y))\,\mathrm{d}s,$$
$$\mathcal{B}(h) := -\int_0^1 \partial_{12}\mathcal{Q}_{sy+(1-s)(y+h)}(\mathcal{M}(y))\,\mathrm{d}s.$$

Besides, $\tilde{\mathcal{A}}(0) = \partial_{22}\tilde{\mathcal{Q}}_y(\mathcal{M}(y))$ is invertible by the definition of ε_2, and the functions $\mathcal{A}$, $\mathcal{B}$ are continuous on $\mathbf{B}(0, \varepsilon)$ by the continuity of $\mathcal{M}$ and the uniform continuity of $\partial_{12}\mathcal{Q}, \partial_{22}\mathcal{Q}$ on $\bar{\mathbf{B}}(x, \varepsilon_0) \times \bar{\mathbf{B}}(\mathcal{M}(x), \varepsilon_0)$. Therefore, there exists $\varepsilon' \in (0; \varepsilon)$ such that on $\mathbf{B}(0, \varepsilon')$,

$$\tilde{\mathcal{M}}(y+h) - \tilde{\mathcal{M}}(y) = \tilde{\mathcal{A}}(h)^{-1}\tilde{\mathcal{B}}(h)\tilde{h} = \tilde{\mathcal{A}}(0)^{-1}\tilde{\mathcal{B}}(0)\tilde{h} + \mathrm{o}(\|h\|).$$

We deduce

$$\partial\mathcal{M}(y) = -P\left[\partial_{22}\tilde{\mathcal{Q}}_y(\mathcal{M}(y))\right]^{-1}\partial_{12}\tilde{\mathcal{Q}}_y(\mathcal{M}(y))\,P^\top,$$

where P is defined as in Section 3. The case $k > 1$ follows by induction, using the above expression and the C^∞-differentiability of the matrix inverse. □

Proof of Lemma B.9. Let $\delta > 0$ and $x_\star \in \mathsf{M}$. By Lemma B.5 there exists $\varepsilon > 0$ such that for all $y \in \mathsf{K}$,

$$f(y) - m < 2\varepsilon \implies \mathrm{d}(y, \mathsf{M}) < \delta. \tag{73}$$

Let $\hat{f}$ be a real-valued function defined on K such that $\sup_{\mathsf{K}} |\hat{f} - f| < \varepsilon$. Then, for all $y \in \hat{\mathsf{M}}$,

$$f(y) < \hat{f}(y) + \varepsilon \leqslant \hat{f}(x_\star) + \varepsilon < f(x_\star) + 2\varepsilon,$$

which implies $\mathrm{d}(y, \mathsf{M}) < \delta$ by (73). □

Lemma C.3. *Let (K, d) be a compact metric space and $\mathcal{Q}\colon \mathsf{K} \times \mathsf{K} \to \mathbb{R}$ be a continuous function such that for all $x \in \mathsf{K}$, $\mathcal{M}(x) := \mathrm{argmin}_{x' \in \mathsf{K}} \mathcal{Q}_x(x')$ is a singleton. Then, for all $\delta > 0$, there exists $\varepsilon > 0$ such that for all $x, x' \in \mathsf{K}$,*

$$\mathcal{Q}_x(x') - \mathcal{Q}_x(\mathcal{M}(x)) < \varepsilon \implies \mathrm{d}(x', \mathcal{M}(x)) < \delta.$$

Proof. By Lemma B.7, the function $\tilde{\mathrm{d}}$ defined on K^2 by $\tilde{\mathrm{d}}(x, x') := \mathrm{d}(\mathcal{M}(x), x')$ is continuous. Let $\delta > 0$. By the compacity of $\tilde{\mathsf{K}} := \tilde{\mathrm{d}}^{-1}([\delta, +\infty[)$ and the continuity of $\mathcal{Q}$, there exists $(x_0, x_0') \in \tilde{\mathsf{K}}$ such that

$$\varepsilon := \inf_{(x,x') \in \tilde{\mathsf{K}}} [\mathcal{Q}_x(x') - \mathcal{Q}_x(\mathcal{M}(x))] = \mathcal{Q}_{x_0}(x_0') - \mathcal{Q}_{x_0}(\mathcal{M}(x_0)) > 0.$$

□

Proof of Lemma B.10. Let $\delta > 0$. By Lemma C.3 there exists $\varepsilon > 0$ such that for all $x, y \in \mathsf{K}$,

$$\mathcal{Q}_x(y) - \mathcal{Q}_x(\mathcal{M}(x)) < 2\varepsilon \implies \mathrm{d}(y, \mathcal{M}(x)) < \delta. \tag{74}$$

Let $\hat{\mathcal{Q}}$ be a real-valued function defined on $\mathsf{K} \times \mathsf{K}$ such that $\sup_{\mathsf{K}\times\mathsf{K}} |\hat{\mathcal{Q}} - \mathcal{Q}| < \varepsilon$. Then, for all $x, y \in \mathsf{K}$ such that $y \in \hat{\mathcal{M}}(x)$,

$$\mathcal{Q}_x(y) < \hat{\mathcal{Q}}_x(y) + \varepsilon \leqslant \hat{\mathcal{Q}}_x(\mathcal{M}(x)) + \varepsilon < \mathcal{Q}_x(\mathcal{M}(x)) + 2\varepsilon,$$

which implies $\mathrm{d}(y, \mathcal{M}(x)) < \delta$ by (74). □

C.3 Convexity

Proof of Lemma B.11. For all $\varepsilon > 0$, write $\mathbf{B}(x_\star, \varepsilon) := \{x \in \mathsf{K} : \|x - x_\star\|_2 < \varepsilon\}$. To begin with, note that $f(x_\star) = f^{**}(x_\star)$. Indeed, $f(x_\star) \geqslant f^{**}(x_\star)$ by definition of the biconjugate, and the constant $f(x_\star)$ is an affine minorant of f, which provides $f^{**} \geqslant f(x_\star)$.

We now prove that $x_\star$ is the unique minimizer of f^{**} on K. By assumption there exists $\varepsilon_0 > 0$ such that f is C^2-differentiable and $\partial^2 f \succ 0$ on $\mathbf{B}(x_\star, \varepsilon_0)$, which implies the convexity of f on that neighborhood. Besides, by Lemma B.5 there exists $\delta > 0$ such that for all $x \in \mathsf{K}$,

$$f(x) - f(x_\star) < \delta \implies \|x - x_\star\|_2 < \varepsilon_0. \tag{75}$$

Let $\varepsilon_1 \in (0; \varepsilon_0)$ such that $f(\mathbf{B}(x_\star, \varepsilon_1)) \subset \mathbf{B}(f(x_\star), \delta/2)$. As $x_\star \in \mathrm{ri}(K)$, for all $x \in \mathsf{K}$, $\partial f(x_\star)^\top (x - x_\star) = 0$. By the boundedness of K and the C^1-differentiability of f on $\mathbf{B}(x_\star, \varepsilon_1)$ this provides the existence of $\varepsilon_2 \in (0; \varepsilon_1)$ such that for all $x \in \mathbf{B}(x_\star, \varepsilon_2)$ and $y \in \mathsf{K}$, $\partial f(x)^\top (y - x) \leqslant \delta/2$. Together with (75) this yields for all $x \in \mathbf{B}(x_\star, \varepsilon_2)$ and $y \in \mathsf{K} \setminus \mathbf{B}(x_\star, \varepsilon_0)$,

$$f(y) \geqslant \delta/2 + \delta/2 + f(x_\star) \geqslant f(x_0) + \partial f(x)^\top (y - x). \tag{76}$$

By the convexity of f on $\mathbf{B}(x_\star, \varepsilon_0)$, the same inequality holds for all $y \in \mathbf{B}(x_\star, \varepsilon_0)$. We deduce from (76) that for all $x \in \mathbf{B}(x_\star, \varepsilon_2)$, $y \in \mathsf{K}$,

$$f^{**}(y) \geqslant f(x) + \partial f(x)^\top (y - x). \tag{77}$$

Let $y \in \mathsf{K} \setminus \{x_\star\}$ and set $t \in (0; 1)$ such that $x(t) := x_\star + t(y - x_\star) \in \mathbf{B}(x_\star, \varepsilon)$. By the convexity of f on $\mathbf{B}(x_\star, \varepsilon)$,

$$\partial f(x(t))^\top (y - x_\star) = \frac{1}{t} \partial f(x(t))^\top (x(t) - x_\star) \geqslant \frac{1}{t} \partial f(x_\star)^\top (x(t) - x_\star) = 0.$$

Using (77) with $x = x(t)$ then yields $f^{**}(y) \geqslant f(x(t)) > f(x_\star) = f^{**}(x_\star)$, which proves that $x_\star$ is the only minimizer of f^{**} on K.

Finally, for all $x \in \mathbf{B}(x_\star, \varepsilon_2)$, using (77) with $y = x$ provides $f^{**}(x) \geqslant f(x)$, and hence $f = f^{**}$ on $\mathbf{B}(x_\star, \varepsilon_2)$. □

Acknowledgement

Many results presented in this paper were obtained during the first year of the Ph.D. of Rayan Charrier, before his resignation for personal reasons.