PLOS ONE

RESEARCH ARTICLE

Beyond convexity—Contraction and global convergence of gradient descent

Patrick M. Wensing [1] *, **Jean-Jacques Slotine** [2]

1 Department of Aerospace and Mechanical Engineering, University of Notre Dame, Notre Dame, IN, United States of America, **2** Department of Mechanical Engineering, Department of Brain and Cognitive Sciences, and Nonlinear Systems Laboratory, Massachusetts Institute of Technology, Cambridge, MA, United States of America

* pwensing@nd.edu
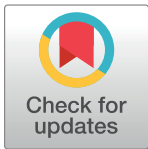
Abstract

This paper considers the analysis of continuous time gradient-based optimization algorithms through the lens of nonlinear contraction theory. It demonstrates that in the case of a time-invariant objective, most elementary results on gradient descent based on convexity can be replaced by much more general results based on contraction. In particular, gradient descent converges to a unique equilibrium if its dynamics are contracting in any metric, with convexity of the cost corresponding to the special case of contraction in the identity metric. More broadly, contraction analysis provides new insights for the case of geodesically-convex optimization, wherein non-convex problems in Euclidean space can be transformed to convex ones posed over a Riemannian manifold. In this case, natural gradient descent converges to a unique equilibrium if it is contracting in any metric, with geodesic convexity of the cost corresponding to contraction in the natural metric. New results using semi-contraction provide additional insights into the topology of the set of optimizers in the case when multiple optima exist. Furthermore, they show how semi-contraction may be combined with specific additional information to reach broad conclusions about a dynamical system. The contraction perspective also easily extends to time-varying optimization settings and allows one to recursively build large optimization structures out of simpler elements. Extensions to natural primal-dual optimization and game-theoretic contexts further illustrate the potential reach of these new perspectives.

1 Introduction

This paper considers the analysis of continuous-time gradient-based optimization through the lens of nonlinear contraction theory. It is motivated, in part, by recent observations in machine learning that arise in the application of gradient descent (or its stochastic counterpart) for the training of over-parameterized networks [1]. Modern networks often possess many more parameters than training examples and can fit the labels perfectly, resulting in submanifold valleys of the parameter space with equal cost [2–4]. Moreover, recent results suggest that highly-redundant networks experience few to no local optima that are not global optima

[5–7]. These observations may be surprising in light of the fact that the loss landscapes for these problems are rarely convex.

Although convex problems admit provable globally optimal solutions, other broader classes of functions share this same property. For example, Invex functions [8] guarantee that any local optimum is a global optimum, although the utility of invexity conditions remains a point of contention [9]. Functions satisfying the Polyak-Lojasiewicz (PL) inequality [1, 3, 10, 11] give rise to exponentially convergent gradient descent to a provably optimal solution. While the PL condition is, in general, difficult to verify without an a-priori known globally optimal solution, the existence of zero-loss solutions in over-parameterized learning [3, 6, 7] makes it tractable in important special cases. Geodesic convexity [12, 13] generalizes convexity to a Riemannian setting, with applicability to optimization on manifolds [14], as well as to conventional Euclidean settings where $\mathbb{R}^n$ is endowed with a manifold structure through the definition of a metric. Here, we consider another class of conditions for the convergence of gradient and natural gradient descent to a globally optimal point. We do so through adopting the perspective of nonlinear contraction theory and analyzing gradient descent in continuous time.

Contraction theory [15] allows the stability of nonlinear non-autonomous systems to be characterized through linear time-varying dynamics describing the propagation of infinitesimally small displacements along the systems' flow. The existence of a Riemannian metric that contracts these virtual displacements (i.e., elements in the tangent space) is necessary and sufficient for exponential convergence of any pair of trajectories. Contraction naturally yields methods for constructing stable systems of systems, including synchronization phenomena [16] and consensus [17, 18] as well as other key building blocks that allow the construction of large contracting systems out of simpler elements [19]. These properties provide opportunities to construct larger optimization structures from simpler elements (e.g., in distributed or competitive optimization settings).

The contribution of this paper is to apply these contraction tools for the analysis of gradient and natural gradient optimization. We consider optimization problems posed over $\mathbb{R}n$ wherein no explicit manifold structure necessarily exists a-priori. Instead, we consider the analysis of optimization following endowing these problems with additional structure (a Riemannian metric), analyzing their convergence, and considering the use of contraction tools to build larger optimization structures out of smaller ones. Analysis proceeds in continuous time. While this approach is limited, in part, by the fact that computational optimization algorithms require a discrete implementation, a continuous perspective has yielded insight on important phenomena such as in the analysis [20], discrete implementations [21], and extensions [22, 23] of Nesterov's accelerated gradient descent method [24]. It has also enabled analysis of primal-dual algorithms [25], where an absolute time reference is obtained by introducing additional fast dynamics or delays using a singular perturbation framework. Recent results [26] provide principled tools to derive discrete-time implementations that preserve specific continuous-time convergence rates.

The paper is organized as follows. Section 2 provides our main results, detailing the applicability of contraction theory to analyze gradient descent in continuous time. We show that convex functions represent the special case of contraction in the identity metric. The flexibility afforded by state-dependent contraction metrics, however, enables significant extra freedom for guaranteeing that all local optima are globally optimal. We then consider the extensions of these results to natural gradient descent, where geodesic convexity of a function corresponds to contraction of its natural gradient system in the natural metric. In both cases, results highlight the topology of the set of optimizers in the case of semi-contraction, which would have most direct applicability to over-parameterized networks. New results also show how semi-

contraction may be combined with specific additional information to reach broad conclusions about a dynamical system. Section 3 details extensions of these results to the case of primal-dual type dynamics that appear in mixed convex/concave saddle systems, and shows how a broad class of natural adaptive control laws can be interpreted as a primal-dual system. Section 4 discusses the special case of g-convex functions and associated combination properties for interfacing with other models. Section 5 provides an outlook on potential future advances that may stem from these connections.

2 Contraction analysis of gradient systems

We first recall basic definitions and facts on convex optimization and show how a contraction analysis of gradient-based optimization considerably generalizes the class of functions that admit a unique global optimum. Following this presentation, results are generalized to the case of geodesically-convex optimization, which is particularly suited to analysis via contraction tools. Throughout this analysis, given a differentiable function $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^m$, we denote the Jacobian of $\mathbf{h}(\mathbf{x})$ by

$$\frac{\partial \mathbf{h}}{\partial \mathbf{x}} = \left[\frac{\partial \mathbf{h}}{\partial x_1} \quad \cdots \quad \frac{\partial \mathbf{h}}{\partial x_n} \right] \in \mathbb{R}^{m \times n}$$

In the special case of a scalar-valued function $f : \mathbb{R}^n \to \mathbb{R}$ we denote the gradient of $f(\mathbf{x})$ by

$$\nabla f(\mathbf{x}) = \left[\frac{\partial f}{\partial \mathbf{x}} \right]^\top \in \mathbb{R}^n$$

and its Hessian by $\nabla^2 f(\mathbf{x})$. Unless otherwise stated, we assume all functions are sufficiently smooth such that derivatives of the necessary order exist and are continuous.

Before we embark on this discussion, let us note that of course, as illustrated, e.g., in [20] and in the following example, continuous-time analysis tools in general may be used to conceptually illuminate the mechanisms involved in discrete-time algorithms. As this paper will show, contraction tools give particularly simple insights into important classes of optimization problems, such as, e.g., geodesically-convex optimization.

Example 1. *The Polyak-Lojasiewicz (PL) inequality is one of the most general sufficient conditions for discrete-time gradient descent to exhibit linear convergence rates without strong convexity of the cost* [10, 11]. *A function is said to satisfy the PL inequality if it has a (typically unknown) global minimum value f^* and there exists a constant $\mu > 0$ such that*

$$\forall \mathbf{x}, \qquad \|\nabla f(\mathbf{x})\|^2 \geq \mu(f(\mathbf{x}) - f^*)$$

Consider gradient descent on the cost function $f(\mathbf{x})$ from a continuous-time point of view,

$$\dot{\mathbf{x}} = -\nabla f(\mathbf{x})$$

Using $V = f(\mathbf{x}) - f^*$ as a Lyapunov-like function, and then requiring that V converges exponentially with rate μ, yields

$$\dot{V} = -\|\nabla f(\mathbf{x})\|^2 \leq -\mu V$$

The inequality above is exactly the PL condition. Thus, we see that the PL condition is nothing but the condition for exponential convergence of the residual cost $V = f(\mathbf{x}) - f^*$.

Similarly, imposing $\dot{V} \le -\mu\sqrt{V}$, corresponding to finite-time convergence (in time less than $2\sqrt{V(0)}/\mu$ [27]), would require a modified PL-like condition

$$\forall \mathbf{x}, \qquad \|\nabla f(\mathbf{x})\|^2 \;\ge\; \mu \; \sqrt{f(\mathbf{x}) - f^*}$$

while imposing $\dot{V} \le -\mu V^2$ would require

$$\forall \mathbf{x}, \qquad \|\nabla f(\mathbf{x})\| \;\ge\; \sqrt{\mu} \; (f(\mathbf{x}) - f^*)$$

By comparison, the results pursued via contraction analysis in this paper will ensure exponential convergence of any pair of trajectories for gradient descent, but likewise will ensure convergence of those solutions to a global optimum.

2.1 Relationships between convexity and contraction

Definition 1 (Strong Convexity). *A twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is α-strongly convex with $\alpha > 0$ if its Hessian matrix $\nabla^2 f(\mathbf{x})$ satisfies the matrix inequality*

$$\nabla^2 f(\mathbf{x}) \succcurlyeq \alpha\,\mathbf{I} \qquad \forall \mathbf{x} \in \mathbb{R}^n$$

As its name suggests, a function that is strongly convex is convex in the usual sense, while the converse is not always true. From a dynamic systems perspective, strong convexity provides exponential convergence of gradient flows:

Proposition 1 (Exponential Convergence of Gradient Systems for Strongly Convex Functions). *If a twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is α-strongly convex, then its gradient system*

$$\dot{\mathbf{x}} = -\nabla f(\mathbf{x}) \tag{1}$$

converges to the unique global minimum of f exponentially with rate α.

Toward proving this proposition, we will consider stability analysis through the application of nonlinear contraction theory.

Definition 2 (Contraction Metric [15]). *A system $\dot{\mathbf{x}} = \mathbf{h}(\mathbf{x}, t)$ is said to be contracting at rate $\alpha > 0$ with respect to a symmetric positive definite metric $\mathbf{M} : \mathbb{R}^n \to \mathbb{R}^{n \times n}$, if for all $t \in \mathbb{R}$ and all $\mathbf{x} \in \mathbb{R}^n$,*

$$\dot{\mathbf{M}} + \mathbf{A}^\top \mathbf{M} + \mathbf{M}\,\mathbf{A} \preccurlyeq -2\alpha\mathbf{M} \tag{2}$$

where $\mathbf{A}(\mathbf{x}, t) = \frac{\partial \mathbf{h}}{\partial \mathbf{x}}$ is the system Jacobian and $\dot{\mathbf{M}} = \sum_i (\partial\mathbf{M}/\partial x_i)h_i(\mathbf{x}, t)$. The system is said to be semi-contracting with respect to $\mathbf{M}$ when (2) holds with $\alpha = 0$.

Given an α-contracting system and an arbitrary pair of initial conditions $\mathbf{x}_1(0)$ and $\mathbf{x}_2(0)$, the solutions $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ converge to one another exponentially

$$d_{\mathcal{M}}(\mathbf{x}_1(t), \mathbf{x}_2(t)) \;\le\; e^{-\alpha t} \; d_{\mathcal{M}}(\mathbf{x}_1(0), \mathbf{x}_2(0)) \tag{3}$$

where $d_{\mathcal{M}}(\cdot, \cdot)$ denotes the geodesic distance on the Riemannian manifold $\mathcal{M} = (\mathbb{R}^n, \mathbf{M})$. This property can be shown by considering the evolution of differential displacements $\delta\mathbf{x}$, which describe the evolution of nearby trajectories and coincide with the notion of virtual displacements in Lagrangian mechanics. More precisely, letting $\mathbf{x}(t; \mathbf{x}_0, t_0)$ denote the solution of $\dot{\mathbf{x}} = \mathbf{h}(\mathbf{x}, t)$ from initial condition $\mathbf{x}(t_0) = \mathbf{x}_0$, differential displacements evolve according to

$$\delta\mathbf{x}(t) = \frac{\partial\mathbf{x}(t; \mathbf{x}_0, t_0)}{\partial\mathbf{x}_0} \; \delta\mathbf{x}(t_0)$$

Property (3) follows from the evolution of the squared length of these differential

displacements [15], which verifies

$$\frac{\mathrm{d}}{\mathrm{d}t}(\delta\mathbf{x}^\top \mathbf{M}\delta\mathbf{x}) \leq -2\alpha(\delta\mathbf{x}^\top \mathbf{M}\delta\mathbf{x}) \tag{4}$$

Furthermore, if a system is α-contracting in a metric $\mathbf{M}$ that satisfies $\mathbf{M}(\mathbf{x}) \succcurlyeq \beta\,\mathbf{I}$ uniformly for some constant $\beta > 0$, then any two solutions verify

$$\|\mathbf{x}_1(t) - \mathbf{x}_2(t)\| \ \leq \ \frac{1}{\sqrt{\beta}}\ \mathrm{e}^{-\alpha t}d_{\mathcal{M}}(\mathbf{x}_1(0), \mathbf{x}_2(0))$$

Example 2. *Consider an α-strongly convex function f and its associated gradient descent system* (1). *Since f is strongly convex, it has a unique global minimum* $\mathbf{x}^*$, *which is a equilibrium point of* (1). *It can be verified that the gradient descent dynamics of f are contracting in the identity metric* $\mathbf{M} = \mathbf{I}$ *with rate α. Since geodesic distances are just Euclidean distances in this metric,* (3) *immediately implies that*

$$\forall t \geq 0, \quad \|\mathbf{x}(t) - \mathbf{x}^*\| \ \leq \ \mathrm{e}^{-\alpha t}\,\|\mathbf{x}(0) - \mathbf{x}^*\|$$

thus proving Proposition 1.

From this example, it is clear that strongly convex functions are a special case of ones whose gradient systems are contracting. The following proposition shows that one does not lose the convergence properties to a global optimum on this more general class of functions.

Proposition 2 (Exponential Convergence of Contracting Gradient Systems). *Consider again gradient descent as in* Eq (1). *The system converges exponentially to a unique global minimum if it is contracting in* some *metric.*

Proof. Because (1) is autonomous and contracting, it converges exponentially to a unique equilibrium $\mathbf{x}^*$ [15]. Furthermore, this equilibrium must be a global minimum since f can only decrease along trajectories, with $\dot{f} = -\nabla f(\mathbf{x})^\top \mathbf{M}(\mathbf{x})^{-1}\nabla f(\mathbf{x}) < 0$ for $\mathbf{x} \neq \mathbf{x}^*$.

The above result, which emphasizes contraction rather than convexity as a sufficient condition to converge to a global minimum, can be extended to the semi-contracting case as follows.

Proposition 3 (Asymptotic Convergence of Semi-Contracting Gradient Systems). *Consider a twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, a symmetric positive definite metric $\mathbf{M} : \mathbb{R}^n \to \mathbb{R}^{n\times n}$, and the associated gradient system*

$$\dot{\mathbf{x}} = -\nabla f(\mathbf{x}) \tag{5}$$

Assume that dynamics (5) *is semi-contracting in* some *metric, and furthermore that one trajectory of the system is known to be bounded. Then, (a) f has at least one stationary point, (b) any local minimum of f is a global minimum, (c) all global minima of f are path-connected, and (d) all trajectories asymptotically converge to a global minimum of f.*

Proof. (a) By assumption, there exists some initial condition $\mathbf{x}_0$ such that $\mathbf{x}(t; \mathbf{x}_0)$ remains bounded. This, in turn, implies that the ω-limit set $\omega[\mathbf{x}(t; \mathbf{x}_0)]$ is non-empty, compact, forward invariant, and that

$$d(\mathbf{x}(t; \mathbf{x}_0), \omega[\mathbf{x}(t; \mathbf{x}_0)]) \to 0 \quad \text{as} \quad t \to +\infty$$

Let $\mathbf{x}^*$ denote an element of $\omega[\mathbf{x}(t; \mathbf{x}_0)]$. Since (5) is a gradient system, Theorem 15.0.3 of [28] guarantees that $\mathbf{x}^*$ must be an equilibrium point of (5). This proves that f has at least one stationary point.

Let us now show that $\omega[\mathbf{x}(t; \mathbf{x}_0)]$ consists only of the single point $\mathbf{x}^*$, by contradiction. Let $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ be distinct elements in $\omega[\mathbf{x}(t; \mathbf{x}_0)]$. Further let $\epsilon = d_{\mathbf{M}}(\mathbf{x}_1^*, \mathbf{x}_2^*)$ the geodesic distance

between $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$. Then, the geodesic balls $\mathcal{B}_1 := \mathcal{B}_{\mathbf{M}}\left(\mathbf{x}_1^*, \frac{\epsilon}{3}\right)$ and $\mathcal{B}_{\mathbf{M}}\left(\mathbf{x}_2^*, \frac{\epsilon}{3}\right)$ are disjoint. Further, since the system is semi-contracting these geodesic balls are forward invariant. Yet, since $\mathbf{x}_1^*$ a limit point, $\mathbf{x}(t;\mathbf{x}_0)$ arrives within $\mathcal{B}_1$ at some point, and never leaves. Likewise, since $\mathbf{x}_2^*$ is a limit point, $\mathbf{x}(t;\mathbf{x}_0)$ arrives within $\mathcal{B}_2$ at some point, and never leaves. Thus, we have a contradiction, and the limit set must consist of a single point.

(b) and (c): Consider now two equilibrium points of (5), $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$, and a smooth path $\gamma(s)$ such that $\gamma(0) = \mathbf{x}_1^*$ and $\gamma(1) = \mathbf{x}_2^*$. Since the gradient dynamics are semi-contracting, for each s the solution $\mathbf{x}(t;\gamma(s))$ remains bounded. Thus, by the same reasoning as above, each $\mathbf{x}(t;\gamma(s))$ converges to some equilibrium $\mathbf{x}^*(s)$ as $t \to +\infty$. Since $\nabla f(\mathbf{x}^*(s)) = 0$ for each s, and $\mathbf{x}^*(s)$ smoothly connects $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$, it follows that $f(\mathbf{x}_1^*) = f(\mathbf{x}_2^*)$. That is, all solutions converge to the same value for f.

(d): That all solutions of (1) asymptotically converge to a global minimum of f follows from that fact that f decreases along all solutions, and all solutions converge to the same value for f.

Remark 1. *In the case that a contraction metric needs to be found numerically, note that the conditions* (2) *for certifying contraction or semi-contraction are convex criteria. Thus, in many instances, the process of finding a metric numerically to verify contraction may be accomplished via convex optimization approaches, such as those based on sums-of-squares programming* [29].

2.2 Relationship between geodesic convexity and contraction

Geodesic convexity [12] generalizes conventional notions of convexity to the case where the domain of a function is equipped with a Riemannian metric. A special case occurs in geometric programming (GP) [30]. In GP, a non-convex problem over positive variables $\{x_i\}_{i=1}^N$ can be transformed into a convex problem by a change of variables $y_i = \log(x_i)$. Alternately GP can be formulated over the positive reals viewed as a Riemannian manifold by measuring differential length elements ds in a relative sense

$$ds^2 = \sum_{i=1}^{N} \left(\frac{dx_i}{x_i}\right)^2 = \sum_{i=1}^{N} dy_i^2 \qquad (6)$$

Geodesically-convex optimization generalizes this transformation strategy to a broader class of problems [13]. However, beyond special cases (see, e.g., [31]), generative procedures remain lacking to formulate g-convex optimization problems or recognize g-convexity.

To introduce g-convexity more formally, consider a function $f : \mathbb{R}^n \to \mathbb{R}$ and a positive definite metric $\mathbf{M} : \mathbb{R}^n \to \mathbb{R}^{n \times n}$. We note that geodesic convexity of f is not an intrinsic property of the function itself, but rather is a property of f defined on the Riemannian manifold $(\mathbb{R}^n, \mathbf{M})$.

Definition 3 (g-Strong Convexity [32]). *A twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be geodesically α-strongly convex (with $\alpha > 0$) in a symmetric positive definite metric $\mathbf{M}$ if its Riemannian Hessian matrix $\mathbf{H}(\mathbf{x})$ satisfies:*

$$\mathbf{H}(\mathbf{x}) \succeq \alpha \, \mathbf{M}(\mathbf{x}) \qquad \forall \mathbf{x} \in \mathbb{R}^n \qquad (7)$$

The elements of the Riemannian Hessian are given as [32]

$$H_{ij} = \partial_{ij} f - \Gamma_{ij}^k \partial_k f \qquad (8)$$

where $\partial_{ij} f = \frac{\partial^2 f}{\partial x_i \partial x_j}$ provide the elements of the conventional (Euclidean) Hessian and Γ_{ij}^k denotes

the Christoffel symbols of the second kind

$$\Gamma_{ij}^{m} = \frac{1}{2} \sum_{k=1}^{n} \left[M^{mk} \left(\partial_j M_{ik} + \partial_i M_{jk} - \partial_k M_{ij} \right) \right]$$

with $M^{ij}(\mathbf{x}) = (M(\mathbf{x})^{-1})_{ij}$. The function f is g-convex when (7) holds with $\alpha = 0$.

The Riemannian Hessian generalizes the notion of the Hessian from a Euclidean context and captures the curvature of f along geodesics. Likewise, the natural gradient generalizes the notion of a Euclidean gradient to the Riemannian context in the following sense.

Definition 4 (Natural Gradient [33]). *Consider $\mathbb{R}^n$ equipped with a Riemannian metric $\mathbf{M}$. The natural gradient of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is the direction of steepest ascent on the manifold and is given in coordinates by $\mathbf{M}(\mathbf{x})^{-1} \nabla f(\mathbf{x})$.*

Remark 2. *When $\mathbf{M}(\mathbf{x})$ is the Hessian of some twice differentiable strictly convex scalar function $\psi(\mathbf{x})$, natural gradient descent coincides with the continuous-time limit of mirror descent [34, Sec. 2.3] with potential $\psi(\mathbf{x})$.*

Remark 3. *From a differential geometric viewpoint, the first covariant derivative of f is a covector field given in coordinates by $\nabla f(\mathbf{x})$, while the natural gradient is a vector field given in coordinates by $\mathbf{M}(\mathbf{x})^{-1} \nabla f(\mathbf{x})$ [33]. In a Euclidean context, where $\mathbf{M}(\mathbf{x})$ is identity, this distinction between covariant (covector) and contravariant (vector) representations of the gradient is immaterial.*

Similarly, the Riemannian Hessian $\mathbf{H}$ represents in coordinates the second covariant derivative of f.

When $\mathbf{M}$ is the identity metric, geodesic α-strong convexity naturally coincides with the definition of α-strong convexity in Definition 1. The natural gradient can be used to directly mirror Proposition 1 within the Riemannian context.

Theorem 1 (Equivalence between g-Strong Convexity and Contraction of Natural Gradient). *Consider a twice differentiable function $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$, a symmetric positive definite metric $\mathbf{M} : \mathbb{R}^n \to \mathbb{R}^{n \times n}$, and the natural gradient system [33]*

$$\dot{\mathbf{x}} = \mathbf{h}(\mathbf{x}, t) = -\mathbf{M}(\mathbf{x})^{-1} \nabla_{\mathbf{x}} f(\mathbf{x}, t) \tag{9}$$

Then, f is α-strongly g-convex in the metric $\mathbf{M}$ for each t if and only if (9) is contracting with rate α in the metric $\mathbf{M}$. More specifically, the Riemannian Hessian verifies

$$\mathbf{H} = -\frac{1}{2} \left(\mathbf{M}\mathbf{A} + \mathbf{A}^{\top}\mathbf{M} + \dot{\mathbf{M}} \right) \tag{10}$$

where $\mathbf{A} = \frac{\partial \mathbf{h}}{\partial \mathbf{x}}$.

Appendix 1 provides a self-contained proof using conventional tensor analysis methods [35], whose relationship with contraction conditions have been noted previously [36, 37]. The same relationships drive coordinate-free versions of the result in [38].

Remark 4. *Theorem 1 can also be viewed as a special case of contraction analysis for complex Hamilton-Jacobi dynamics [37]. A reorganization of (9) as*

$$\mathbf{M}(\mathbf{x}) \ \dot{\mathbf{x}} \ = -\nabla_{\mathbf{x}} f(\mathbf{x}, t)$$

may be recognized as the generalized momentum being the negative covariant gradient within a Hamiltonian mechanics context.

Remark 5. *While Theorem 1 applies to α-strong convexity, the link between the Riemannian Hessian and the contraction condition (2) also provides immediate equivalence between g-convexity of a function and semi-contraction of its natural gradient dynamics.*

Remark 6. Eq (10) *provides an alternate way to compute the geodesic Hessian* **H**, *and, as expected, leaves it invariant when the metric* **M** *is scaled by a strictly positive constant. Because of the structure of the natural gradient dynamics, scaling* **M** *is akin to scaling time and implies inversely scaling the contraction rate* α, *consistently with* (7).

By contrast, note that given a fixed dynamics **H**, *the contraction metric analyzing it can always be arbitrarily scaled while leaving the contraction rate unchanged.*

Similar to in Section 2.1 where convexity corresponded to contraction of gradient in the identity metric, we likewise see that Thm. 1 imposes g-convexity via a particular choice of contraction metric for the natural gradient dynamics. Mirroring Prop. 2, removing this restriction on the contraction metric leads to significant additional flexibility for guaranteeing convergence to a globally optimal point.

Proposition 4 (Exponential Convergence of Contracting Natural Gradient Systems). *Consider again natural gradient descent as in* Eq (9). *The system converges exponentially to a unique global minimum if it is contracting in* some *metric.*

Proof. The proof follows immediately from the same logic as the proof of Proposition 2.

Remark 7. *Note that contraction also provides robustness. Consider perturbed dynamics* $\dot{\mathbf{x}} = \mathbf{h}(\mathbf{x}, t) + \mathbf{d}(t)$ *with* $\sqrt{\mathbf{d}(t)^{\top}\mathbf{M}(\mathbf{x})\mathbf{d}(t)} < R$ *uniformly. If the dynamics are contracting with rate* λ, *then all trajectories contract to a geodesic ball of radius* R/λ [15]. *This observation implies favorable properties for algorithms where an exact gradient may be difficult or intractable to compute, with approximation methods used in their place.*

Theorem 2 (Semi-Contraction for Natural Gradient). *Consider a twice differentiable function* $f : \mathbb{R}^n \to \mathbb{R}$, *a symmetric positive definite metric* $\mathbf{M} : \mathbb{R}^n \to \mathbb{R}^{n \times n}$, *and the associated natural gradient system*

$$\dot{\mathbf{x}} = -\mathbf{M}(\mathbf{x})^{-1}\nabla f(\mathbf{x}) \tag{11}$$

Assume that dynamics (11) *is semi-contracting in* some *metric, and furthermore that one trajectory of the system is known to be bounded. Then, (a) f has at least one stationary point, (b) any local minimum of f is a global minimum, (c) all global minima of f are path-connected, and (d) all trajectories asymptotically converge to a global minimum of f.*

Proof. The proof follows the exact same line of logic as the proof to Prop. 3. The result of Theorem 15.0.3 of [28], which guarantees that any ω-limit point of gradient descent (5) is an equilibrium point, generalizes immediately to the case of natural gradient descent (11).

Remark 8. *The topology of global optimizers satisfying this semi-contraction condition is the same as those observed when training over-parameterized networks* [2, 3]. *However, empirical loss functions in these networks often also experience multiple saddle points* [39, 40]. *The attractor sets associated with strict saddles have measure zero* [41, 42] *under discrete gradient descent with sufficiently small stepsize (i.e., with adequately close approximation to the continuous time case), while the dimensionality of the attractor sets can be further reduced via smoothed versions of the gradient* [43].

While the presence of strict saddles precludes the ability of a gradient system to be globally semi-contracting, any of the results given here can be generalized to forward invariant contraction or semi-contraction regions [15]. *In principle, saddles could then be treated by excluding their measure zero attractor sets from suitably chosen contraction or semi-contraction regions.*

The topology of equilibria in semi-contracting gradient systems immediately implies the following result.

Corollary 1. *Consider an autonomous, semi-contracting natural gradient system. If the linearization at some equilibrium point is strictly stable, then all system trajectories tend to this global minimizer.*

More generally, if some equilibrium is locally asymptotically stable, all trajectories tend to this global minimizer.

Proof. We prove the second part, the first then follows directly from Lyapunov's linearization method. Existence of an equilibrium implies existence of a bounded trajectory. Furthermore, by definition, there exists a ball around the equilibrium point $\mathbf{x}^*$ such that all trajectories initiated in that ball tend to $\mathbf{x}^*$. If there was another equilibrium, the path connecting it to $\mathbf{x}^*$ would intersect that ball, which is a contradiction since the path is itself composed of equilibria via Thm. 2.

Remark 9. *Strict stability of a natural gradient system at an equilibrium point can of course be established simply by ensuring that all eigenvalues of its Jacobian at this point are strictly in the left-half complex plane. This condition is equivalent to requiring that the Hessian of the objective function is positive definite at $\mathbf{x}^*$.*

Indeed, given the natural gradient dynamics (11) with $\mathbf{h}(\mathbf{x}) = -\mathbf{M}(\mathbf{x})^{-1}\nabla f(\mathbf{x})$, the Jacobian at any equilibrium $\mathbf{x}^*$ is

$$\left.\frac{\partial \mathbf{h}}{\partial \mathbf{x}}\right|_{\mathbf{x}^\star} = -\left.\frac{\partial[\mathbf{M}^{-1}]}{\partial \mathbf{x}}\right|_{\mathbf{x}^\star} \cdot \nabla f(\mathbf{x}^\star) - \mathbf{M}(\mathbf{x}^\star)^{-1}\nabla^2 f(\mathbf{x}^\star)$$

$$= -\mathbf{M}(\mathbf{x}^\star)^{-1}\nabla^2 f(\mathbf{x}^\star)$$

Applying a similarity transformation with the symmetric square root of $\mathbf{M}(\mathbf{x}^*)$ yields

$$\mathbf{M}^{\frac{1}{2}}(\mathbf{x}^\star)\left.\frac{\partial \mathbf{h}}{\partial \mathbf{x}}\right|_{\mathbf{x}^\star} \mathbf{M}^{-\frac{1}{2}}(\mathbf{x}^\star)$$

$$= -\mathbf{M}(\mathbf{x}^\star)^{-\frac{1}{2}}[\nabla^2 f(\mathbf{x}^\star)]\mathbf{M}(\mathbf{x}^\star)^{-\frac{1}{2}}$$

All eigenvalues of the symmetric matrix above are real, and they are all strictly negative if and only if the Hessian $\nabla^2 f(\mathbf{x}^*)$ is positive definite.

Note that this condition is equivalent to the geodesic Hessian at $\mathbf{x}^*$ being positive definite in any metric, as the Euclidean Hessian is numerically equal to the geodesic Hessian in any metric in this case, due to all terms multiplying the Christoffel symbols in (8) being zero.

Corollary 2. *Consider an autonomous semi-contracting natural gradient system, and assume that the system has more than one equilibrium. Then, at any equilibrium, both the Jacobian matrix of the dynamics and the Hessian of the objective have at least one zero eigenvalue.*

Proof. Consider an equilibrium $\mathbf{x}^*$, and an equilibrium path connecting it to some other equilibrium. The unit tangent vector at $\mathbf{x}^*$ along this path is an eigenvector of the Jacobian with eigenvalue zero. Given the algebraic relation between the Jacobian and the objective Hessian pointed out in Remark 9, this shows in turn that the objective Hessian has a zero eigenvalue.

2.3 Examples

Let us illustrate Theorem 1 using the classical nonconvex Rosenbrock function:

$$f(\mathbf{x}) = 100(x_1^2 - x_2)^2 + (x_1 - 1)^2 \tag{12}$$

This function has a unique global optimum at $\mathbf{x}^* = [1, 1]^\top$, which is located along a long, shallow, parabolic-shaped valley.

Example 3 *Consider the Rosenbrock function* (12) *and the metric* [32]

$$\mathbf{M}(\mathbf{x}) = \begin{bmatrix} 400x_1^2 + 1 & -200x_1 \\ -200x_1 & 100 \end{bmatrix}$$

The metric $\mathbf{M}(\mathbf{x})$ *satisfies* $\mathrm{tr}(\mathbf{M}(\mathbf{x})) = 400x_1^2 + 101 > 0$ *and* $\det(\mathbf{M}(\mathbf{x})) = 100 > 0$, *and thus* $\mathbf{M}(\mathbf{x}) \succ 0$. *Note that* $\mathbf{M}(\mathbf{x})$ *is not the Hessian of* $f(\mathbf{x})$. *The natural gradient dynamics follows*

$$\dot{\mathbf{x}} = \mathbf{h}(\mathbf{x}) = -\mathbf{M}(\mathbf{x})^{-1}\nabla f(\mathbf{x}) = -2 \begin{bmatrix} x_1 - 1 \\ x_1^2 - 2x_1 + x_2 \end{bmatrix}$$

It can be verified algebraically that

$$\mathbf{M}\left(\frac{\partial \mathbf{h}}{\partial \mathbf{x}}\right) + \left(\frac{\partial \mathbf{h}}{\partial \mathbf{x}}\right)^{\top} \mathbf{M} + \dot{\mathbf{M}} = -4\mathbf{M}$$

which shows that natural gradient descent is contracting with rate $\alpha = 2$. *This implies that the natural gradient dynamics satisfy*

$$d_{\mathcal{M}}(\mathbf{x}(t), \mathbf{x}^*) \leq e^{-2t} \, d_{\mathcal{M}}(\mathbf{x}(0), \mathbf{x}^*)$$

where $\mathbf{x}^* = [1, 1]^{\top}$. *Equivalently, the Rosenbrock function is geodesically* α-*strongly convex with* $\alpha = 2$.

The Rosenbrock metric $\mathbf{M}(\mathbf{x})$ can be viewed as following from a differential change of variables

$$\delta \mathbf{z} = \Theta(\mathbf{x})\delta \mathbf{x} = \begin{bmatrix} 20x_1 & -10 \\ 1 & 0 \end{bmatrix} \delta \mathbf{x}$$

where $\mathbf{M} = \Theta^{\top}\Theta$ yields $\delta \mathbf{x}^{\top} \mathbf{M} \, \delta \mathbf{x} = \delta \mathbf{z}^{\top} \delta \mathbf{z}$. This differential change of variables is integrable, so that g-convexity of the Rosenbrock can be shown using the explicit nonlinear coordinate change $z_1 = 10x_1^2 - 10x_2$ and $z_2 = x_1 - 1$ that provides $f = z_1^2 + z_2^2$.

Example 4. *Mirror descent provides another example of a metric corresponding to an explicit state transformation, with Newton's method as a special case.*

Consider a twice differentiable scalar objective function $f(\mathbf{x})$, *and a smooth strictly convex scalar function* $\psi(\mathbf{x})$. *Denoting by* $\mathbf{H}_f(\mathbf{x}) = \nabla^2 f(\mathbf{x})$ *and* $\mathbf{H}_\psi = \nabla^2 \psi$ *the Hessians of these functions, continuous-time mirror descent of* $f(\mathbf{x})$ *under potential* $\psi(\mathbf{x})$ *corresponds to natural gradient in the Hessian metric* $\mathbf{H}_\psi$ [34, Sec. 2.3]

$$\dot{\mathbf{x}} = -\mathbf{H}_\psi^{-1}\nabla f(\mathbf{x}) \tag{13}$$

Consider the explicit change of variables $\mathbf{z} = \nabla \psi(\mathbf{x})$, *which can be written in differential form as* $\delta \mathbf{z} = \mathbf{H}_\psi \, \delta \mathbf{x}$. *The dynamics* (13) *can be viewed in the mirror space as*

$$\dot{\mathbf{z}} = \mathbf{H}_\psi \, \dot{\mathbf{x}} = -\nabla f(\mathbf{x})$$

and therefore

$$\frac{\mathrm{d}}{\mathrm{d}t} \, \delta \mathbf{z} = -\mathbf{H}_f \delta \mathbf{x}$$

Letting $\mathbf{M}(\mathbf{x}) = \mathbf{H}_\psi^2$, *this yields*

$$\frac{\mathrm{d}}{\mathrm{d}t}[\delta\mathbf{x}^T\mathbf{M}(\mathbf{x})\delta\mathbf{x}] = \frac{\mathrm{d}}{\mathrm{d}t}[\delta\mathbf{z}^T\delta\mathbf{z}] = -2\ \delta\mathbf{z}^\top\mathbf{H}_f\ \delta\mathbf{x} = -\delta\mathbf{x}^\top[\mathbf{H}_f\mathbf{H}_\psi + \mathbf{H}_\psi\mathbf{H}_f]\delta\mathbf{x}^\top$$

Thus, continuous mirror descent (13) *is contracting with rate* $\lambda > 0$ *in the metric* $\mathbf{M}(\mathbf{x}) = \mathbf{H}_\psi^2$ *if*

$$\mathbf{H}_f\mathbf{H}_\psi + \mathbf{H}_\psi\mathbf{H}_f \succcurlyeq 2\lambda\mathbf{H}_\psi^2 \tag{14}$$

In the particular case when f is α-strongly convex and the potential function is chosen as $\psi(\mathbf{x}) = f(\mathbf{x})$, Eq (13) *simply corresponds to Newton's method, and* (14) *verifies that Newton's method is contracting with rate 1 in the squared Hessian metric* $\mathbf{M}(\mathbf{x}) = \mathbf{H}_f^2(\mathbf{x})$ [44].

Note that the well-known result that the transformation $\mathbf{z} = \nabla\phi(\mathbf{x})$ *is one-to-one (given the strict convexity of* ψ*) can also be shown by constructing, for a given* $\mathbf{z}$*, the system*

$$\dot{\mathbf{x}} + \nabla\psi(\mathbf{x}) = \mathbf{z} \tag{15}$$

which is autonomous and contracting in the identity metric and thus must reach a unique equilibrium point.

The following proposition provides further insight into the case when the contraction metric is related to an explicit change of variables more generally.

Proposition 5 (Relationship between gradient and natural gradient under a diffeomorphic change of variables). *Consider a diffeomorphic change of variables* $\mathbf{z} = \mathbf{g}(\mathbf{x})$, *and the associated metric* $\mathbf{M}(\mathbf{x}) = \boldsymbol{\Theta}(\mathbf{x})^\top\boldsymbol{\Theta}(\mathbf{x})$, *with* $\boldsymbol{\Theta}(\mathbf{x}) = \frac{\partial\mathbf{g}}{\partial\mathbf{x}}$. *For any twice differentiable function* $f : \mathbb{R}^n \to \mathbb{R}$, *natural gradient descent in* $\mathbf{x}$

$$\dot{\mathbf{x}} = -\mathbf{M}(\mathbf{x})^{-1}\nabla f(\mathbf{x})$$

is equivalent to gradient descent in $\mathbf{z}$

$$\dot{\mathbf{z}} = -\nabla[f \circ \mathbf{g}^{-1}](\mathbf{z})$$

Proof. In the $\mathbf{z}$ coordinates we have

$$\dot{\mathbf{z}} = \frac{\partial\mathbf{g}}{\partial\mathbf{x}}\dot{\mathbf{x}} = -\boldsymbol{\Theta}\ \mathbf{M}^{-1}\nabla f(\mathbf{x}) = -\left[\frac{\partial f}{\partial\mathbf{x}}\boldsymbol{\Theta}^{-1}\right]^\top$$
$$= -\left[\frac{\partial f \circ \mathbf{g}^{-1}}{\partial\mathbf{z}}\right]^\top = -\nabla[f \circ \mathbf{g}^{-1}](\mathbf{z})$$

Proposition 6. *Consider a metric* $\mathbf{M}(\mathbf{x})$ *and suppose there exists a diffeomorphic change of variables* $\mathbf{z} = \mathbf{g}(\mathbf{x})$ *such that* $\mathbf{M}(\mathbf{x}) = \boldsymbol{\Theta}(\mathbf{x})^\top\boldsymbol{\Theta}(\mathbf{x})$, *with* $\boldsymbol{\Theta}(\mathbf{x}) = \frac{\partial\mathbf{g}}{\partial\mathbf{x}}$. *Then, the associated Riemannian curvature tensor with components* $R_{ik\ell m}$ *must be identically zero.*

Proof. Note that since $\delta\mathbf{z} = \boldsymbol{\Theta}(\mathbf{x})\delta\mathbf{x}$, it follows that $\delta\mathbf{z}^\top\ \delta\mathbf{z} = \delta\mathbf{x}^\top\ \mathbf{M}(\mathbf{x})\delta\mathbf{x}$ and thus the Riemannian metric tensor expressed in the $\mathbf{z}$ coordinates is the identity. Since the components of the Riemannian metric tensor are constant in these transformed coordinates, it follows that the components of the Riemannian curvature tensor are identically zero [32]. Transformation laws for tensors ensure that the components of the curvature tensor remain zero under arbitrary coordinate change, thus $R_{ik\ell m} = 0$.

The general freedom to consider differential changes of coordinates $\delta\mathbf{z} = \boldsymbol{\Theta}(\mathbf{x})\delta\mathbf{x}$ where $\boldsymbol{\Theta}$ is non-integrable provides additional flexibility and generality to both contraction analysis and g-convexity, as illustrated by the following examples.
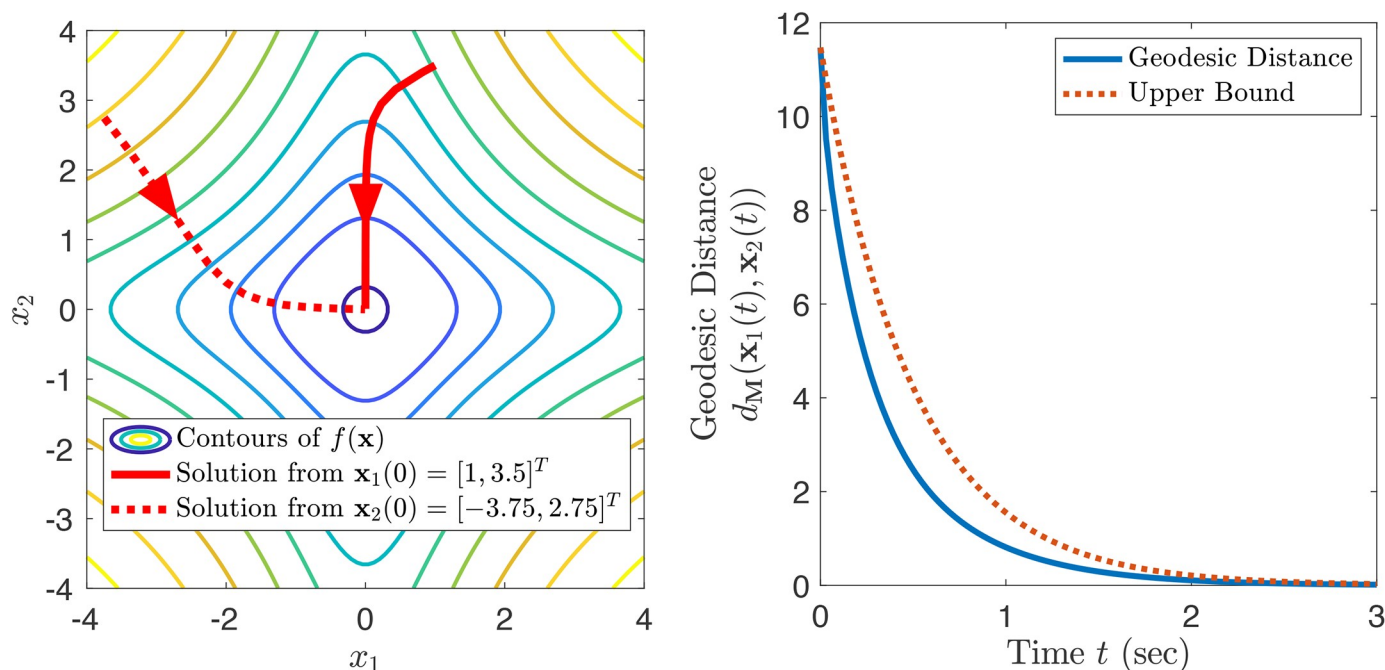
Fig 1. Contracting gradient descent corresponding to Example 5.

Example 5. Consider the non-convex function

$$f(\mathbf{x}) = x_1^2 + x_2^2 + x_1^2 x_2^2$$

which has a global minimum at $\mathbf{x} = \mathbf{0}$. *Contours of the function are shown in* Fig 1. *Gradient descent*

$$\dot{\mathbf{x}} = -\nabla f(\mathbf{x}) = -2 \begin{bmatrix} x_1(1 + x_2^2) \\ x_2(1 + x_1^2) \end{bmatrix}$$

can be shown to be contracting at rate $\lambda = 2$ *in the metric*

$$\mathbf{M}(\mathbf{x}) = \begin{bmatrix} 2 + x_1^2 & -x_1 x_2 \\ -x_1 x_2 & 2 + x_2^2 \end{bmatrix}$$

Fig 1 *shows two solutions and plots their geodesic distance. The decay is, as expected, at a rate faster than the exponentially decreasing upper bound as derived from* (3). *The curvature tensor for this metric has some non-zero components, such as*

$$R_{1221} = \frac{2}{2 + x_1^2 + x_2^2}$$

From Proposition 6, this shows that this metric cannot be derived from an explicit change of coordinates.

Example 6. *Consider the function*

$$f(\mathbf{z}) = z_1^8 - z_1^6 - 2z_1^5 z_2 - z_1^4 + z_1^2 z_2^2 + z_1^2 + 2z_1 z_2 + z_2^2$$

and natural gradient descent with a given natural metric $\Theta(\mathbf{z})^{\top} \Theta(\mathbf{z})$, where

$$\Theta(\mathbf{z}) = \begin{bmatrix} 1 & 0 \\ 1 - 3z_1^2 & 1 \end{bmatrix}$$

This natural gradient dynamics is verified semi-contracting in the metric

$$\mathbf{M}(\mathbf{z}) = \Theta(\mathbf{z})^{\top} \begin{bmatrix} 1 + z_2^2 - 2z_1^3 z_2 + z_1^6 & 0 \\ 0 & 1 + z_1^2 \end{bmatrix} \Theta(z)$$

Similar to Example 5, this metric has non-zero Riemannian curvature, and thus cannot be derived from a change of coordinates. Fig 2 shows the contours of f and two solutions of natural gradient descent. Fig 3 shows that the the geodesic distance between these two solutions is non-increasing. Since the system is only semi-contracting, the distance between solutions does not tend toward zero. It can be verified that f is a sum of squares and thus $f(\mathbf{z}) \geq 0$, and that $f(\mathbf{z}) = 0$ when $z_2 = z_1^3 - z_1$. Both initial conditions asymptotically lead to this path connected set of global optima.

Example 7. *Geodesically-convex optimization can also be used to carry out manifold-constrained optimization in an unconstrained fashion via recasting problems over a Riemannian manifold directly [14, 31]. Taking an intrinsic view of the manifold, coordinate free results are available [38], however, for the purposes of computation, we assume a global coordinate chart here. Consider for instance optimization over the set $\mathbb{S}_+^n$ of $n \times n$ positive definite matrices, and specifically the problem of finding the Karcher mean of m matrices $\mathbf{A}_i \in \mathbb{S}_+^n$ [13], which*
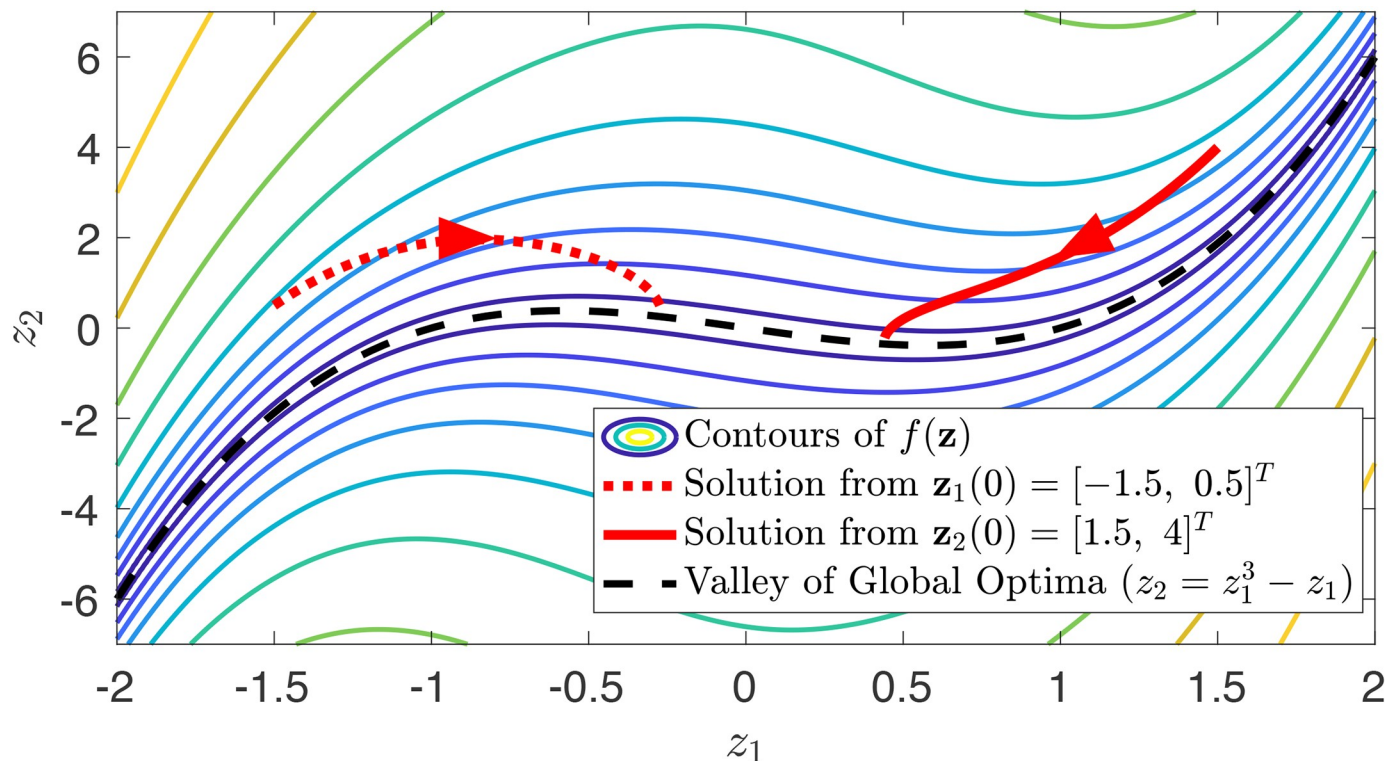

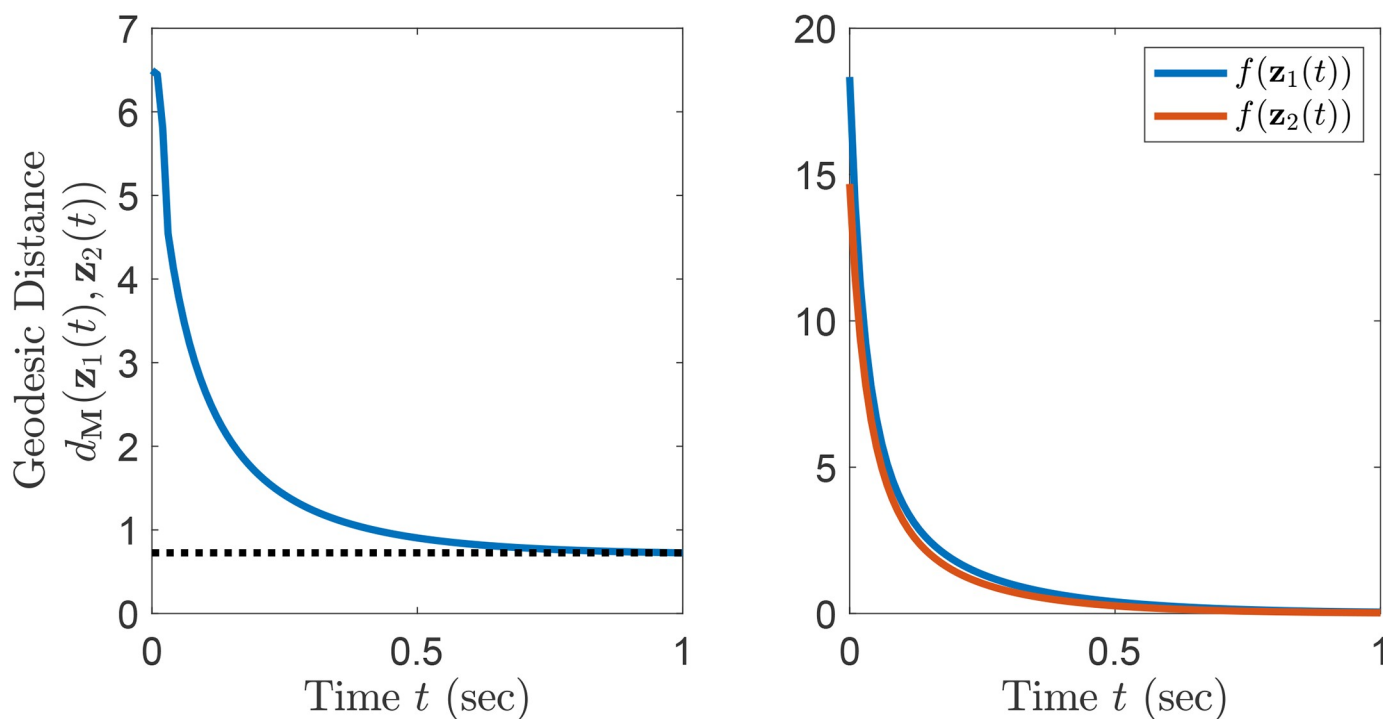**Fig 2. Semi-contracting natural gradient descent for Example 6.**

Fig 3. Semi-contracting natural gradient descent for Example 6.

minimizes the objective function

$$f(\mathbf{X}) = \frac{1}{2}\sum_{i=1}^{m}\|\log(\mathbf{A}_i^{-1}\mathbf{X})\|_F^2$$

where $\|\mathbf{A}\|_F = \sqrt{\mathrm{tr}(\mathbf{A}^\top\mathbf{A})}$ *denotes the Frobenius norm of a matrix* $\mathbf{A}$. *The function* $f(\mathbf{X})$ *is m-strongly convex* [13] *on* $\mathbb{S}_+^n$ *in the metric that measures symmetric differential displacements as*

$$ds^2 = \mathrm{tr}((\delta\mathbf{X}\,\mathbf{X}^{-1})^2) \tag{16}$$

Naturally, the requirement that $\delta\mathbf{X}$ *be symmetric makes it an element of the tangent space to the manifold of symmetric positive definite matrices.*

This metric generalizes the GP case (6), *and coincides with the second-order terms in the Taylor series of the log barrier* $-\log\det(\mathbf{X})$ [45]. *The gradient of* $f(\mathbf{x})$ *can be written*

$$\nabla f(\mathbf{X}) = \sum_{i=1}^{m}\log(\mathbf{A}_i^{-1}\mathbf{X})\,\mathbf{X}^{-1}$$

and accordingly the natural gradient can be shown to satisfy

$$\sum_{i=1}^{m}\mathbf{X}\log(\mathbf{A}_i^{-1}\mathbf{X}) = \mathbf{X}\,\nabla f(\mathbf{X})\,\mathbf{X}$$

From Theorem 1, any trajectory with arbitrary initial condition $\mathbf{X}(0) \in \mathbb{S}_+^n$ *will remain within*

$\mathbb{S}^n_+$ *under the natural gradient descent dynamics*

$$\dot{\mathbf{X}} = -\sum_{i=1}^m \mathbf{X} \log(\mathbf{A}_i^{-1} \mathbf{X})$$

since (intuitively) the Riemannian metric (16) *makes any element on the boundary of the positive definite cone an infinite distance away from any one in the interior, and contraction of the natural gradient dynamics ensures that geodesic distances decrease exponentially.*

    Example 8. *An approximation to the Riemannian distance of two positive definite (PD) matrices on the PD cone is given by the Bregman LogDet divergence on* $\mathbb{S}^n_+$

$$d(\mathbf{A}\|\mathbf{X}) = \operatorname{logdet}(\mathbf{A}^{-1}\mathbf{X}) + \operatorname{tr}(\mathbf{X}^{-1}\mathbf{A}) - n \tag{17}$$

The metric is convex in its first argument, and can be shown to be geodesically convex in the second. We illustrate the connection with contraction to show this property. Note that

$$\nabla d(\mathbf{A}\|\mathbf{X})\mathbf{X} = \mathbf{X}^{-1} - \mathbf{X}^{-1}\mathbf{A}\mathbf{X}^{-1}$$

so that the natural gradient descent dynamics are simply

$$\dot{\mathbf{X}} = -\mathbf{X}[\nabla d(\mathbf{A}\|\mathbf{X})\mathbf{X}]\mathbf{X} = -\mathbf{X} + \mathbf{A}$$

with differential dynamics

$$\delta\dot{\mathbf{X}} = -\delta\mathbf{X}$$

where the differential displacement $\delta\mathbf{X}$ *must be symmetric. Considering the rate of change in length of these differential displacements*

$$\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{tr}\big((\mathbf{X}^{-1}\,\delta\mathbf{X})^2)\big) = -2\operatorname{tr}(\mathbf{X}^{-1}\mathbf{A}(\mathbf{X}^{-1}\delta\mathbf{X}^{-1})^2\big)$$

and defining the differential change of variables $\delta\mathbf{Z} = \mathbf{X}^{-\frac{1}{2}}\delta\mathbf{X}\,\mathbf{X}^{-\frac{1}{2}}$, *one has* $\operatorname{tr}(\delta\mathbf{Z}^2) = \operatorname{tr}((\mathbf{X}^{-1}\,\delta\mathbf{X})^2)$ *and*

$$\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{tr}(\delta\mathbf{Z}^2) = -2\operatorname{tr}\Big(\delta\mathbf{Z}\,\mathbf{X}^{-\frac{1}{2}}\mathbf{A}\mathbf{X}^{-\frac{1}{2}}\,\delta\mathbf{Z}\Big) = -2\|\mathbf{A}^{\frac{1}{2}}\mathbf{X}^{-\frac{1}{2}}\,\delta\mathbf{Z}\|_F^2 < 0 \tag{18}$$

for all $\delta\mathbf{Z} \neq \mathbf{0}$. *Hence, considering only the second argument to LogDet divergence, its Riemannian Hessian is positive definite, thus proving g-convexity via Thm. 1.*

2.4 Non-autonomous systems and virtual systems

In our optimization context, the fact that contraction analysis is directly applicable to non-autonomous systems can be exploited in a variety of ways. As we shall detail later, a key aspect is that it allows feedback combinations or hierarchies of contracting modules to be exploited to address more elaborate optimization problems or architectures. Also, it makes the construction of *virtual* systems [46] possible to potentially extend results beyond natural-gradient descent.

    Remark 10. *The natural gradient* $\mathbf{M}^{-1}(\mathbf{x})\nabla_{\mathbf{x}} f(\mathbf{x}, t)$ *represents the direction of steepest ascent on the manifold at any given time. With this in mind, Remark 7 on robustness enables convergence analysis for natural gradient descent within time-varying optimization contexts* [25]. *Let* $\mathbf{x}^*(t)$ *denote the optimum of a time-varying* α-strongly g-convex function. If

$\sqrt{\dot{\mathbf{x}}^*(t)^\top \mathbf{M}(\mathbf{x})\dot{\mathbf{x}}^*(t)} < R$, *then the natural gradient will track* $\dot{\mathbf{x}}^*(t)$ *with accuracy* R/α *after exponential transient.*

Remark 11. *Consider a contracting natural gradient system of the form* (9). *In the autonomous case, equations governing the differential displacement follow*

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\delta\mathbf{x} \;=\; \frac{\partial\mathbf{h}}{\partial\mathbf{x}}\,\delta\mathbf{x} \tag{19}$$

which has a similar structure to the time evolution of $\mathbf{h}(\mathbf{x})$

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbf{h}(\mathbf{x}) \;=\; (\nabla\mathbf{h}(\mathbf{x}))\,\mathbf{h}(\mathbf{x}) \tag{20}$$

Thus, for natural gradient descent of an α*-strong g-convex function* $f(\mathbf{x})$*, the same algebra leading to* (4) *also gives*

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{h}^\top\mathbf{M}\mathbf{h}) = -2\mathbf{h}^\top\mathbf{H}\mathbf{h} \leq -2\alpha(\mathbf{h}^\top\mathbf{M}\mathbf{h})$$

so that the Krasovskii-like function

$$V(\mathbf{x}) = \mathbf{h}(\mathbf{x})^\top\mathbf{M}(\mathbf{x})\mathbf{h}(\mathbf{x}) = \nabla f(\mathbf{x})^\top\mathbf{M}(\mathbf{x})^{-1}\nabla f(\mathbf{x})$$

can be viewed as an exponentially converging Lyapunov function, with global minimum $V = 0$ *at the unique minimum of* $f(\mathbf{x})$*. Of course,* (19) *remains valid for non-autonomous systems as well, while* (20) *does not.*

The use of virtual contracting systems [46–48] allows guaranteed exponential convergence to a unique minimum to be extended to classes of dynamics which are not pure natural gradient. For instance, it is common in optimization to adjust the learning rate as the descent progresses. Consider a natural gradient descent with the function $f(\mathbf{x})$ α-strongly g-convex in metric $\mathbf{M}(\mathbf{x})$, and define the new system

$$\dot{\mathbf{x}} = -p(\mathbf{x},t)\,\mathbf{M}(\mathbf{x})^{-1}\,\nabla f(\mathbf{x}) \tag{21}$$

where the smooth scalar function $p(\mathbf{x}, t)$ modulates the learning rate [33] and is uniformly positive definite,

$$\exists\, p_{min} > 0,\ \ \forall t \geq 0,\ \ \forall\mathbf{x},\qquad p(\mathbf{x},t) \geq p_{min}$$

Let us show that this system tends exponentially to the minimum $\mathbf{x}^*$ of $f(\mathbf{x})$.

Consider the auxiliary, *virtual* system,

$$\dot{\mathbf{y}} = -\mathbf{M}(\mathbf{y})^{-1}\,\nabla_{\mathbf{y}}(\,p(\mathbf{x},t)f(\mathbf{y})\,) \tag{22}$$

For this system, $p(\mathbf{x}(t), t)$ is an external, uniformly positive definite function of time, and thus

$$\nabla_{\mathbf{y}}(\,p(\mathbf{x},t)f(\mathbf{y})\,)\mathbf{y} \;=\; p(\mathbf{x},t)\,\nabla f(\mathbf{y})$$

so that the contraction of (9) with rate α implies the contraction of (22) with rate αp_{min}. Since both $\mathbf{x}(t)$ and $\mathbf{x}^*$ are particular solutions of (22), this implies in turn that $\mathbf{x}(t)$ tends to $\mathbf{x}^*$ with rate αp_{min}.

Note that since we only assumed that $p(\mathbf{x}, t)$ is uniformly positive definite, in general the actual system (21) is not contracting with respect to the metric $\mathbf{M}(\mathbf{x})$.

Remark 12. *The learning rate may also be selected to improve the numerical properties of the algorithm in a discrete time implementation. For example,* $p(\mathbf{x}, t)$ *could vary as the inverse of the*

condition number of $\nabla^2 f(\mathbf{x})$ to improve numeric conditioning without impact on stability guarantees.

2.5 Contraction +

Corollary 1 above points to a more general class of results where contraction or semi-contraction properties are combined with other information, such as a stable local linearization or a decreasing cost, to provide global results.

2.5.1 Contraction is attractive.   As we now show, Corollary 1 extends more generally to autonomous semi-contracting systems. An instance of this result in the case of an identity metric was derived in [49].

Proposition 7. *Consider an autonomous system*

$$\dot{\mathbf{x}} = \mathbf{h}(\mathbf{x}) \tag{23}$$

semi-contracting in a bounded metric $\mathbf{M}(\mathbf{x})$,

$$\mathbf{0} \prec \beta_0^2 \mathbf{I} \preccurlyeq \mathbf{M}(\mathbf{x}) \preccurlyeq \beta_1^2 \mathbf{I} \qquad \forall \mathbf{x}$$

If a system equilibrium is locally asymptotically stable, then it is globally asymptotically stable. In particular, if the system linearization at some equilibrium point is strictly stable, then all system trajectories tend to this equilibrium.

Proof. The result is a particular case of Theorem 3, to be discussed next.

Theorem 3. *Consider a non-autonomous system, semi-contracting in a bounded metric* $\mathbf{M}(\mathbf{x})$,

$$\mathbf{0} \prec \beta_0^2 \mathbf{I} \preccurlyeq \mathbf{M}(\mathbf{x}) \preccurlyeq \beta_1^2 \mathbf{I} \qquad \forall \mathbf{x} \tag{24}$$

Assume that a specific *trajectory* $\mathbf{x}^*(t)$ *is locally attractive. Then all trajectories tend asymptotically to* $\mathbf{x}^*(t)$.

In particular, if contraction holds (possibly in a different bounded metric) along a specific trajectory $\mathbf{x}^*(t)$, *and within a tube of constant size around it, then all trajectories tend asymptotically to* $\mathbf{x}^*(t)$.

Proof. The first part generalizes the equilibrium argument from [49] to arbitrary trajectories and arbitrary metrics. Assume that $\mathbf{x}^*(t)$ is locally attractive, by which we mean there exists some $\epsilon > 0$ such that, for any initial time t_0 and initial condition $\mathbf{x}_0 \in \mathcal{B}_{\mathbf{I}}(\mathbf{x}^*(t_0), \epsilon)$, one has $\mathbf{x}(t_0 + T; \mathbf{x}_0, t_0) \to \mathbf{x}^*(t_0 + t)$ as $T \to +\infty$. Without loss of generality, we assume $t_0 = 0$.

Consider some generic initial condition $\mathbf{x}_0$ with $d_{\mathbf{I}}(\mathbf{x}_0, \mathbf{x}^*(0)) > \epsilon$. We will argue that there is always a finite time window over which the geodesic distance from $\mathbf{x}(t; \mathbf{x}_0)$ to $\mathbf{x}^*(t)$ decreases by a fixed finite increment.

Consider a geodesic connecting $\mathbf{x}_0$ and $\mathbf{x}^*(0)$ and denote by $\underline{\mathbf{x}}$ the unique point on this geodesic that is a geodesic distance $\beta_0 \epsilon$ away from $\mathbf{x}^*(0)$. Due to the uniform positive definiteness of $\mathbf{M}$, this condition implies that $\underline{\mathbf{x}} \in \mathcal{B}_{\mathbf{I}}(\mathbf{x}^*(0), \epsilon)$.

Because of the local attractivity of $\mathbf{x}^*(t)$, there exists a time $t_1 > 0$ such that

$$d_{\mathbf{I}}(\mathbf{x}(t_1; \underline{\mathbf{x}}), \mathbf{x}^*(t_1)) \leq \frac{\epsilon \beta_0}{2\beta_1}$$

which further implies that

$$d_{\mathbf{M}}(\mathbf{x}(t_1; \underline{\mathbf{x}}), \mathbf{x}^*(t_1)) \leq \beta_0 \frac{\epsilon}{2}$$

In addition, since the system is semi-contracting, we have

$$d_{\mathbf{M}}(\mathbf{x}(t_1; \underline{\mathbf{x}}), \mathbf{x}(t_1; \mathbf{x}_0)) \leq d_{\mathbf{M}}(\mathbf{x}, \mathbf{x}_0)$$

and so by the triangle inequality

$$
\begin{aligned}
d_{\mathbf{M}}(\mathbf{x}^*(t_1), \mathbf{x}(t_1; \mathbf{x}_0)) \leq d_{\mathbf{M}}(\underline{\mathbf{x}}, \mathbf{x}_0) + \beta_0 \frac{\epsilon}{2} \quad &= d_{\mathbf{M}}(\underline{\mathbf{x}}, \mathbf{x}_0) + d_{\mathbf{M}}(\mathbf{x}^*(0), \underline{\mathbf{x}}) \\
&\quad - d_{\mathbf{M}}(\mathbf{x}^*(0), \mathbf{x}) + \beta_0 \frac{\epsilon}{2} \\
&= d_{\mathbf{M}}(\mathbf{x}^*(0), \mathbf{x}_0) - \beta_0 \frac{\epsilon}{2}
\end{aligned}
$$

This implies that so long as $d_{\mathbf{I}}(\mathbf{x}_0, \mathbf{x}^*(0)) > \epsilon$, the trajectory from $\mathbf{x}_0$ will eventually decrease its geodesic distance by a fixed finite increment. Since this process can be repeated, it follows that there must exist some time T such that $\mathbf{x}(T; \mathbf{x}_0) \in \mathcal{B}_{\mathbf{I}}(\mathbf{x}^*(T), \epsilon)$.

To complete the second part of the proof we proceed to show that if contraction (2) holds within a tube of constant size around trajectory $\mathbf{x}^*(t)$, for some bounded metric $(\beta_0^\star)^2 \mathbf{I} \preccurlyeq; M^\star(\mathbf{x}) \preccurlyeq (\beta_1^\star)^2 \mathbf{I}$ and some rate $\alpha^* > 0$, then that trajectory is locally attractive. By condition (2) holding within a tube we mean that there exists some $\epsilon > 0$ such that (2) holds for any time t and any $\mathbf{x} \in \mathcal{B}_{\mathbf{I}}(\mathbf{x}^*(t), \epsilon)$. From boundedness of the metric, we have

$$\beta_0^\star d_{\mathbf{I}}(\mathbf{x}^*(t), \mathbf{x}) \leq d_{\mathbf{M}^*}(\mathbf{x}^*(t), \mathbf{x})$$

so that any initial condition $\mathbf{x}_0$ satisfying

$$d_{\mathbf{M}^*}(\mathbf{x}^*(0), \mathbf{x}_0) \leq \epsilon \beta_0^\star$$

necessarily starts within this tube. Further, since (2) holds within the tube, it follows that the geodesic ball of radius $\epsilon \beta_0^\star$ around $\mathbf{x}^*(t)$ is forward invariant. Since this ball is contained within a contraction region, this implies that for any $\mathbf{x}_0 \in \mathcal{B}_{\mathbf{M}}(\mathbf{x}^\star(0), \beta_0^\star \epsilon)$

$$d_{\mathbf{M}^*}(\mathbf{x}^*(t), \mathbf{x}(t; \mathbf{x}_0)) \leq e^{-\alpha^\star t} \, d_{\mathbf{M}^*}(\mathbf{x}^*(0), \mathbf{x}_0)$$

which proves local asymptotic stability of $\mathbf{x}^*(t)$.

Remark 13. *The condition regarding contraction within a tube of fixed size is included to avoid pathological cases where the region of contraction shrinks to zero as $t \to +\infty$. For example, the system $\dot{x} = -x + tx^3$ is contracting with rate 1 at the origin for all time, yet the origin is not locally asymptotically stable.*

Remark 14. *Intuitively, the result can be understood by analogy with a shrinking rope. Consider a path of initial conditions connecting $\mathbf{x}^*(0)$ to any $\mathbf{x}_0$. As this path flows forward in time, at $t = 0$, only a portion of this path of states is within the basin of attraction for $\mathbf{x}^*(t)$. Viewing this path as a rope, the semi-contraction property ensures that no part of the rope can increase in length as it flows forward through the dynamics. Yet, due to local attractivity at one end of the rope, a portion of it is guaranteed to have shrinking length, pulling the rest of the rope toward the the region of attraction.*

Remark 15. *Numerical tools for determining contraction metrics [29] are based on the fact that contraction conditions (2) are convex in the metric for a fixed contraction rate. In practice, these methods often involve an outer search procedure for the contraction rate (e.g., via a binary search). In this sense, the use of semi-contraction is desirable as it does not require this additional search.*

Remark 16. *These results have analogs in the context of controller design using control contraction metrics (CCMs) [48, 50]. In this setting, one can impose a semi-contracting closed-loop*

metric everywhere, except in a tube along a desired trajectory where a strict contraction condition would be required, possibly in a different metric. Since the existence of an exponential (resp. semi) CCM implies that the closed-loop plant can be rendered contracting (resp. semi-contracting), Theorem 3 would then imply asymptotic stabilizability of the desired trajectory.

This extension likewise has analogs for manifold convergence results [48] and convergence to a limit cycle by transverse contraction [51], both of which are special cases of CCM results applied to suitably constructed virtual control systems [48]. In either case, a semi-contracting CCM everywhere can be combined with a contracting CCM condition on the manifold (or limit cycle) and within a neighborhood of it to assert asymptotic stability of the manifold (or limit cycle). In the limit cycle case for autonomous systems, the contracting CCM condition needs only be enforced on the limit cycle itself, as its satisfaction within some neighborhood is then guaranteed by compactness. Likewise, for convergence to a compact manifold (e.g., an eggshell) in an autonomous system, the contracting CCM condition needs only be considered on the manifold itself.

2.5.2 Contraction as minimization. Similarly, Proposition 2 may be viewed as a particular instance of the following results, which use contraction properties to minimize a cost or Lyapunov-like function.

Proposition 8 (Exponential Cost Minimization). *Consider an autonomous contracting system* (23), *and a scalar cost function* $V(\mathbf{x})$ *such that* $\dot{V}(\mathbf{x}) \leq 0$ *for all* x. *Then all trajectories tend exponentially to a global minimum of V.*

Proof. Because the system is contracting and autonomous, it tends exponentially to a unique equilibrium $\mathbf{x}^*$ [15]. Consider now an arbitrary x, and the system trajectory initialized at x. Since the cost V can only decrease along the trajectory, this implies that $V(\mathbf{x}^*) \leq V(\mathbf{x})$, for all x.

Proposition 9. *Consider an autonomous semi-contracting system* (23) *in a bounded metric* $\mathbf{M}(\mathbf{x})$, *and a scalar cost function* $V(\mathbf{x})$ *such that* $\dot{V}(\mathbf{x}) \leq 0$ *for all* x. *Assume that one system equilibrium* $\mathbf{x}^*$ *is locally attractive (e.g., that linearization at* $\mathbf{x}^*$ *is strictly stable). Then this equilibrium is unique, it is a global minimum of V, and all trajectories converge to it asymptotically.*

Proof. Applying Proposition 7 shows that all trajectories asymptotically tend to $\mathbf{x}^*$, which also implies that the equilibrium is unique. By the same reasoning as in Proposition 8, since V can only decrease, $V(\mathbf{x}^*)$ must be a global minimum.

Remark 17. *These results extend readily to the case where a system is semi-contracting within some forward invariant region, as opposed to globally. These generalizations may have applicability e.g., to the continuous-time limit of trained neural networks [52–54], wherein semi-contraction regions represent basins of attraction that are free of saddles. Metrics may become singular as they approach the boundary of these open sets [15], allowing the semi-contraction region to cover the entire basin.*

Proposition 9 can be stated more generally as follows.

Theorem 4 (Asymptotic Cost Minimization). *Consider an autonomous semi-contracting system in a bounded metric* $\mathbf{M}(\mathbf{x})$, *and a scalar cost function* $V(\mathbf{x})$ *such that* $\dot{V}(\mathbf{x}) \leq 0$ *for all* x. *Assume that one trajectory is known to be bounded. Let* $\mathcal{I}$ *be a forward invariant set where* $\dot{V} = 0$, *and assume that the contraction condition* (2) *holds on* $\mathcal{I}$ *for some (possibly different) metric.*

Then $\mathcal{I}$ *is path connected, all system trajectories converge to a unique equilibrium* $\mathbf{x}^* \in \mathcal{I}$ *and V is globally minimized at* $\mathbf{x}^*$.

Proof. Let us first show that $\mathcal{I}$ is path connected, by contradiction. Assume $\mathcal{I}$ is not path connected, then it can be decomposed into two disjoints subsets, $\mathcal{I}_1$ and $\mathcal{I}_2$. Because $\mathcal{I}$ is invariant and the subsets are disjoint, each of the subsets must be invariant. Strict contraction on $\mathcal{I}_1$ and $\mathcal{I}_2$ then implies that each subset contains at least one locally stable equilibrium point (note that each of the subsets may themselves be disconnected and thus may contain more than one

stable equilibrium point). The existence of two equilibrium points contradicts Proposition 9, and thus $\mathcal{I}$ is path connected.

Next, on the connected invariant set $\mathcal{I}$, contraction implies that the geodesic distance between any two points shrinks exponentially. By the same reasoning as in Proposition 8, this in turn implies convergence to a global minimum of V.

Remark 18. *Note that for a system where a scalar cost V satisfies $\dot{V} \leq 0$, radial unboundedness of V is a sufficient condition for all trajectories to be bounded, ensuring the existence of a bounded trajectory as necessary in Thm. 4.*

Remark 19. *In the case of mechanical systems, V may often be chosen as the total energy of the system, so that Proposition 8 implies exponential convergence of the total energy, and, in turn, that potential energy is exponentially minimized. Similarly, Theorem 4 implies that potential energy is asymptotically minimized.*

Remark 20. *Contraction criteria can also be expressed in non-Euclidean norms and their associated matrix measures ([15], section 3.7.ii). The results above extend immediately to these representations.*

3 Primal-dual optimization

Primal-Dual algorithms are widely used in optimization to determine saddle points and also appear naturally in constrained optimization [45], where Lagrange parameters play the role of dual variables. When a function is strictly convex in a subset of its variables, and strictly concave in the remaining, gradient descent/ascent dynamics converge to a unique saddle equilibrium [55, 56]. Within the context of constrained optimization, these dynamics are known as the primal-dual dynamics. Such dynamics play an important role e.g., in machine learning, for instance in adversarial training [57], in the information theory [58] of deep networks, in reinforcement learning [59] and actor-critic methods, and in support vector machine representations [60]. More generally, they are central to a large class of practical min-max problems, such as problems in physics involving free energy, or, e.g., nonlinear electrical networks modeled in terms of Brayton-Moser mixed potentials [25, 61, 62].

Consider a scalar function $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, t)$, possibly time-dependent, and metrics $\mathbf{M_x}(\mathbf{x})$ and $\mathbf{M}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})$. Consider the natural primal-dual dynamics, which we define as

$$\mathbf{M_x}(\mathbf{x})\,\dot{\mathbf{x}} = -\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, t) \tag{25a}$$

$$\mathbf{M}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})\,\dot{\boldsymbol{\lambda}} = \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, t) \tag{25b}$$

In contrast to Remark 8, wherein spurious saddle equilibrium points presented an obstacle to global contraction, here the target equilibrium points of these dynamics are, by construction, chosen to be the saddle points of the function $\mathcal{L}$. Using the metrics $\mathbf{M_x}(\mathbf{x})$ and $\mathbf{M}_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})$ extends the standard case [25], where they would be replaced by constant, symmetric positive definite matrices. The practical relevance of this extension is illustrated by the following example in the case of natural adaptive control.

3.1 Primal dual dynamics in natural adaptive control

This section illustrates the presence of natural primal-dual dynamics embedded in the application of natural adaptive control laws. Consider a system given by

$$\mathbf{J}(\mathbf{x}, \ldots, \mathbf{x}^{(n-2)}, \mathbf{a})\mathbf{x}^{(n)} + \mathbf{Y}(\mathbf{x}, \ldots, \mathbf{x}^{(n-1)})\mathbf{a} = \mathbf{u} \tag{26}$$

with configuration $\mathbf{x} \in \mathbb{R}^N$, control $\mathbf{u} \in \mathbb{R}^N$, and unknown parameters $\mathbf{a} \in \mathcal{A} \subset \mathbb{R}^p$. The

regressor $\mathbf{Y} \in \mathbb{R}^{N \times p}$ and symmetric matrix $\mathbf{J} \in \mathbb{R}^{N \times N}$ may depend nonlinearly on the state and its derivatives. We assume that the matrix $\mathbf{J}$ remains positive definite for all $\mathbf{a} \in \mathcal{A}$ and that it is linear in $\mathbf{a}$. As a result, there exists a regressor function $\mathbf{W}$ such that

$$\mathbf{W}(\mathbf{x}, \ldots, \mathbf{x}^{(n-1)}, \mathbf{x}^{(n)})\mathbf{a} = \mathbf{J}(\mathbf{x}, \ldots, \mathbf{x}^{(n-2)}, \mathbf{a})\mathbf{x}^{(n)} \tag{27}$$

$$+\mathbf{Y}(\mathbf{x}, \ldots, \mathbf{x}^{(n-1)})\mathbf{a} \tag{28}$$

and a regressor function $\mathbf{Q}$ such that for any $\mathbf{s} \in \mathbb{R}^N$

$$-\frac{1}{2}\dot{\mathbf{J}}\mathbf{s} = \mathbf{Q}(\mathbf{x}, \ldots, \mathbf{x}^{(n-1)}, \mathbf{s})\,\mathbf{a}$$

Consider a desired trajectory $\mathbf{x}_d(t)$ and the associated sliding variable [27, 63]

$$\mathbf{s} = \left(\frac{\mathrm{d}}{\mathrm{d}t} + \lambda\right)^{n-1}\tilde{\mathbf{x}} = \mathbf{x}^{(n-1)} - \mathbf{x}_r^{(n-1)} \tag{29}$$

where $\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{x}_\mathrm{d}$. With this sliding variable, we define a reference $\mathbf{x}_r^{(n-1)}$ for the order $n-1$ derivative of the state.

Choosing the control law

$$\mathbf{u} = -\mathbf{K}\mathbf{s} + \hat{\mathbf{J}}\mathbf{x}_r^{(n)} + \mathbf{Y}(\mathbf{x}, \ldots, \mathbf{x}^{(n-1)})\hat{\mathbf{a}} - \frac{1}{2}\dot{\hat{\mathbf{J}}}\,\mathbf{s}$$

where $\mathbf{x}_r^{(n)} = \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}_r^{(n-1)}$ provides the closed-loop dynamics

$$\mathbf{J}\dot{\mathbf{s}} + \mathbf{K}\mathbf{s} = \mathbf{W}\tilde{\mathbf{a}} + \mathbf{Q}\hat{\mathbf{a}} \tag{30}$$

Inspired by the elegant modification of the Slotine and Li adaptive robot controller introduced by Lee et al. [64, 65], we consider the Lyapunov-like function

$$V = \frac{1}{2}\mathbf{s}^\top\mathbf{J}\mathbf{s} + d_f(\mathbf{a}\|\hat{\mathbf{a}})$$

where $d_f(\mathbf{a}\|\hat{\mathbf{a}})$ denotes the Bregman divergence of a function f assumed convex on $\mathcal{A}$ and given by

$$d_f(\mathbf{a}\|\hat{\mathbf{a}}) = f(\mathbf{a}) - f(\hat{\mathbf{a}}) - <\nabla f(\hat{\mathbf{a}}), \mathbf{a} - \hat{\mathbf{a}}>$$

Note that the LogDet divergence from (17) follows this form for $f(\mathbf{X}) = -\mathrm{logdet}(\mathbf{X})$. Here, we consider the case when $\mathcal{A}$ is open and f is chosen as a convex barrier function on $\mathcal{A}$, such that the Hessian metric $\mathbf{H} = \nabla^2 f(\mathbf{x})$ endows $\mathcal{A}$ with a barrier Hessian manifold structure [65, 66]. Note that if f is a second-order function $\frac{1}{2}\mathbf{a}^T\mathbf{P}\mathbf{a}$, the Bregman divergence is simply $\frac{1}{2}\tilde{\mathbf{a}}^\top\mathbf{P}\tilde{\mathbf{a}}$, with $\mathbf{H}^{-1}$ equal to the constant matrix $\mathbf{P}^{-1}$ similar to the standard adaptive algorithm [63].

A quick calculation shows that the derivative of the Bregman divergence is simply $\dot{\hat{\mathbf{a}}}^\top\mathbf{H}\tilde{\mathbf{a}}$, so that the adaptation law

$$\dot{\hat{\mathbf{a}}} = -\mathbf{H}^{-1}(\mathbf{W} + \mathbf{Q})^\top\,\mathbf{s} \tag{31}$$

yields

$$\dot{V} = -\mathbf{s}^\top\mathbf{K}\mathbf{s} + \mathbf{s}^\top(\mathbf{W} + \mathbf{Q})\tilde{\mathbf{a}} + \dot{\hat{\mathbf{a}}}^\top\mathbf{H}\,\tilde{\mathbf{a}} = -\mathbf{s}^\top\mathbf{K}\mathbf{s} \leq 0$$

Considering a virtual system with $\mathbf{W}$ and $\mathbf{Q}$ as externally provided functions of time, the

dynamics (30) and (31) are equivalent to natural primal-dual over the function

$$\mathcal{L} = \frac{1}{2}\mathbf{s}^\top \mathbf{K}\mathbf{s} - \mathbf{s}^\top \mathbf{W}\tilde{\mathbf{a}} - \mathbf{s}^\top \mathbf{Q}\,\hat{\mathbf{a}}$$

in the decoupled metric $\mathbf{M_s} = \mathbf{J}$ and $\mathbf{M}\hat{\mathbf{a}} = \mathbf{H}(\hat{\mathbf{a}})$. Overall, this construction enables the results in natural adaptive robot control [64, 65] to be extended to the broader class (26).

Remark 21. *Note that a similar construction could be applied to provide natural adaptation within recent applications of nonlinear adaptive control [67, Thm. 2] based on control contraction metrics [48, 50].*

3.2 Natural primal dual

Continuous-time convex primal-dual optimization is analyzed from a nonlinear contraction perspective in [25], building on a earlier result of [19]. As we now show, Theorem 1 yields a natural extension to geodesic primal-dual optimization, where convexity in terms of primal and dual variables is replaced by g-convexity, thus broadening the above results to state-dependent metrics.

Theorem 5. *Consider a scalar function $\mathcal{L}(\mathbf{x}, \lambda, t)$, with $\mathcal{L}$ g-strongly convex over $\mathbf{x}$ and g-strongly concave over λ in metrics $\mathbf{M_x}(\mathbf{x})$ and $\mathbf{M}_\lambda(\lambda)$ respectively. Then, the geodesic primal-dual dynamics (25) is globally contracting, in metric*

$$\mathbf{M}(\mathbf{x}, \lambda) = \begin{bmatrix} \mathbf{M_x}(\mathbf{x}) & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_\lambda(\lambda) \end{bmatrix} \tag{32}$$

Proof. Letting $\mathbf{z} = [\mathbf{x}^\top, \lambda^\top]^\top$ and $\dot{\mathbf{z}} = \mathbf{f}(\mathbf{z}, t)$ denote the overall system dynamics, the system's Jacobian can be written

$$\mathbf{A}(\mathbf{x}, \lambda, t) = \frac{\partial \mathbf{f}}{\partial \mathbf{z}} = \begin{bmatrix} -\dfrac{\partial}{\partial \mathbf{x}}\left(\mathbf{M_x}^{-1}\nabla_{\mathbf{x}}\mathcal{L}\right) & -\,\mathbf{M_x}^{-1}\,\nabla_{\mathbf{x}\lambda}\mathcal{L} \\[2mm] \mathbf{M}_\lambda^{-1}\,\nabla_{\lambda\mathbf{x}}\mathcal{L} & \dfrac{\partial}{\partial \lambda}\left(\mathbf{M}_\lambda^{-1}\nabla_\lambda\mathcal{L}\right) \end{bmatrix} \tag{33}$$

so that, using Theorem 1,

$$\mathbf{M}\mathbf{A} + \mathbf{A}^\top\mathbf{M} + \dot{\mathbf{M}} = -2\begin{bmatrix} \mathbf{H_x} & \mathbf{0} \\ \mathbf{0} & -\mathbf{H}_\lambda \end{bmatrix} < \mathbf{0}$$

Proposition 10. *Consider the primal dual dynamics (25) for a scalar cost function $\mathcal{L}(\mathbf{x}, \lambda)$, with $\mathcal{L}$ g-strongly convex over $\mathbf{x}$ and g-concave (not necessarily strongly so) over λ in metrics $\mathbf{M_x}(\mathbf{x})$ and $\mathbf{M}_\lambda(\lambda)$ respectively. Suppose also that one solution of (25) is known to be bounded. Then, for any initial condition, the geodesic primal-dual dynamics (25) converge to an equilibrium $\mathbf{x}^*, \lambda^*$. Moreover, $\mathbf{x}^*$ is independent of initial conditions.*

Proof. The proof is given as a corollary to Theorem 6 in the next section.

Remark 22. *The above proposition highlights that contraction of the PD dynamics (e.g., as developed in [25]) is not necessary to guarantee convergence to a unique primal solution. Note however, that the above results only guarantee asymptotic convergence toward the unique primal equilibrium, as opposed to exponential convergence [25] when contraction can be shown for the PD dynamics as a whole.*

This observation is reminiscent of results in adaptive control wherein the error dynamics of a certainty-equivalent controller may be asymptotically stable despite the fact that an

associated adaptation law may not converge to the actual unknown parameters [63, 68], with adaptation occurring on a "need-to-know" basis in that sense. Conceptually, this principle can apply to more general contexts involving concurrent control and learning, when effective control is the main goal (e.g., in reinforcement learning).

4 Applying contraction tools to g-convex optimization

Theorem 1 immediately implies that existing combination properties [15, 19] from contraction analysis can be directly applied in the context of g-convex optimization. While these properties derive from simple matrix algebra and in principle could be proven directly from the definition of geodesic convexity, as we will see most rely for their practical relevance on the flexibility afforded by the contraction analysis point of view.

4.1 Sum of g-convex

If two functions $f_1(\mathbf{x}, t)$ and $f_2(\mathbf{x}, t)$ are g-convex in the same metric for each t, then their sum $f_1(\mathbf{x}, t) + f_2(\mathbf{x}, t)$ is g-convex in the same metric.

Example 9. *Consider a function $f_1(\mathbf{x}_1, \mathbf{y}_1, t)$ g-convex for each t in a block diagonal metric* $\mathrm{BlkDiag}(\mathbf{M}_{x_1}(\mathbf{x}_1), \mathbf{M}_y(\mathbf{y}_1))$ *and a function $f_2(\mathbf{x}_2, \mathbf{y}_2, t)$ g-convex for each t in a block diagonal metric* $\mathrm{BlkDiag}(\mathbf{M}_{x_2}(\mathbf{x}_2), \mathbf{M}_y(\mathbf{y}_2))$. *Then, the function*:

$$f(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}, t) = f_1(\mathbf{x}_1, \mathbf{y}, t) + f_2(\mathbf{x}_2, \mathbf{y}, t)$$

is g-convex in metric $\mathrm{BlkDiag}(\mathbf{M}_{x_1}, \mathbf{M}_{x_2}, \mathbf{M}_y)$ *for each t.*

4.2 Skew-symmetric feedback coupling

Assume that a scalar function $f_1(\mathbf{x}_1, \mathbf{x}_2)$ is α_1-strongly g-convex in $\mathbf{x}_1$ in a metric $\mathbf{M}_1(\mathbf{x}_1)$ for each fixed $\mathbf{x}_2$, and similarly that a scalar function $f_2(\mathbf{x}_1, \mathbf{x}_2)$ is α_2-strongly g-convex in a metric $\mathbf{M}_2(\mathbf{x}_2)$ for each fixed $\mathbf{x}_1$. If f_1 and f_2 satisfy the scaled skew-symmetry property

$$\nabla_{\mathbf{x}_1\mathbf{x}_2} f_1 \ = \ - \ k \ \nabla_{\mathbf{x}_1\mathbf{x}_2} f_2 \tag{34}$$

where k is some strictly positive constant, then the natural gradient dynamics

$$\begin{aligned}
\dot{\mathbf{x}}_1 &= -\mathbf{M}_1(\mathbf{x}_1)^{-1} \ \nabla_{\mathbf{x}_1} f_1(\mathbf{x}_1, \mathbf{x}_2) \\
\dot{\mathbf{x}}_2 &= -\mathbf{M}_2(\mathbf{x}_2)^{-1} \ \nabla_{\mathbf{x}_2} f_2(\mathbf{x}_1, \mathbf{x}_2)
\end{aligned} \tag{35}$$

is contracting with rate $\min(\alpha_1, \alpha_2)$ in metric $\mathbf{M}(\mathbf{x}_1, \mathbf{x}_2) = \mathrm{BlkDiag}(\mathbf{M}_1(\mathbf{x}_1), k\,\mathbf{M}_2(\mathbf{x}_2))$. Since the *overall* system is both contracting and autonomous, it tends to a unique equilibrium [15]$(\mathbf{x}_1^*, \mathbf{x}_2^*)$ which satisfies the Nash-like conditions

$$\begin{aligned}
\mathbf{x}_1^* &= \mathrm{argmin}_{\mathbf{x}_1} f_1(\mathbf{x}_1, \mathbf{x}_2^*) \\
\mathbf{x}_2^* &= \mathrm{argmin}_{\mathbf{x}_2} f_2(\mathbf{x}_1^*, \mathbf{x}_2)
\end{aligned}$$

Note that the result can be broadened to cases where the scaled skew-symmetry property is not exactly satisfied, by using the small-gain extension in [19]. Taking again the machine learning context as a potential example, such two-player game dynamics can occur in certain types of adversarial training.

The result extends to a game with an arbitrary number of players. Consider n functions $\{f_i(\mathbf{x}_1, \dots, \mathbf{x}_n)\}_{i=1}^n$ such that each f_i is α_i-strongly g-convex over $\mathbf{x}_i$ in a metric $\mathbf{M}_i(\mathbf{x}_i)$. If the

functions satisfy the skew-symmetry conditions

$$\nabla_{\mathbf{x}_i \mathbf{x}_j} f_i = -k_j \nabla_{\mathbf{x}_i \mathbf{x}_j} f_j$$

for each $j > i$, then the suitable generalizations of (35) result in a coupled system that is contracting with rate $\min(\alpha_1, \ldots, \alpha_n)$ in the metric

$$\mathbf{M} = \mathrm{BlkDiag}(\mathbf{M}_1, k_2 \mathbf{M}_2, \ldots, k_n \mathbf{M}_n).$$

The overall system converges to a unique Nash-like equilibrium satisfying

$$\mathbf{x}_1^* = \mathrm{argmin}_{\mathbf{x}_1} f_1(\mathbf{x}_1, \mathbf{x}_2^*, \ldots, \mathbf{x}_n^*)$$

and a similar relation for each other player.

Likewise, the result can be extended to the case when the natural gradient dynamics for each individual player may only be semi-contracting.

Theorem 6. *Consider the two player case* (35), *wherein (a) f_1 is α_1-strongly g-convex with $\alpha_1 > 0$ in a uniformly positive definite metric $\mathbf{M}_1(\mathbf{x}_1)$ for each $\mathbf{x}_2$ (b) the Riemannian Hessian $\mathbf{H}_2(\mathbf{x}_1, \mathbf{x}_2)$ of $f_2(\mathbf{x}_1, \mathbf{x}_2)$ in $\mathbf{x}_2$ is only positive semi-definite for each $\mathbf{x}_1$ in a uniformly positive definite metric $\mathbf{M}_2(\mathbf{x}_2)$ and (c) the skew-symmetry property (34) holds. Assume that one trajectory of (35) is known to be bounded. Then, every trajectory of (35) converges to a Nash equilibrium $\mathbf{x}_1^*$, $\mathbf{x}_2^*$. Moreover, $\mathbf{x}_1^*$ does not depend on initial conditions (i.e., every Nash has the same strategy for player 1).*

Proof. It can be shown that virtual displacements evolve such that

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\delta\mathbf{x}_1^\top \mathbf{M}_1(\mathbf{x}_1)\delta\mathbf{x}_1 + k\delta\mathbf{x}_2^\top \mathbf{M}_2(\mathbf{x}_2)\delta\mathbf{x}_2 \right)$$
$$\leq -2\,\alpha_1 \delta\mathbf{x}_1^\top \mathbf{M}_1(\mathbf{x}_1)\delta\mathbf{x}_1 - 2\,k\,\delta\mathbf{x}_2^\top \mathbf{H}_2(\mathbf{x}_1, \mathbf{x}_2)\delta\mathbf{x}_2$$

which implies, by Barbalat's lemma,

$$\delta\mathbf{x}_1 \to 0 \quad \text{and} \quad \mathbf{H}_2(\mathbf{x}_1, \mathbf{x}_2)\delta\mathbf{x}_2 \to 0$$

Via the same argument as follows from (20), it follows that $\nabla_{\mathbf{x}_1}\mathbf{x}_1(\mathbf{x}_1(t), \mathbf{x}_2(t)) \to 0$ as $t \to \infty$. So, for each initial condition, $\mathbf{x}_1(t)$ must converge to some equilibrium $\mathbf{x}_1^*$ of the $\mathbf{x}_1$ dynamics. Furthermore, since any $\delta\mathbf{x}_1 \to 0$, $\mathbf{x}_1^*$ must be *unique* and independent of initial conditions. Let us now turn to the behavior of the $\mathbf{x}_2$ dynamics. Given an arbitrary initial condition $(\mathbf{x}_{1,0}, \mathbf{x}_{2,0})$ for (35), let L^+ denote its ω-limit set. Any point in $(\mathbf{x}_1, \mathbf{x}_2) \in L^+$ must satisfy $\mathbf{x}_1 = \mathbf{x}_1^*$. Since the dynamics are autonomous, L^+ is composed of trajectories of the system

$$\dot{\mathbf{x}}_1 = \mathbf{0} \tag{36}$$

$$\dot{\mathbf{x}}_2 = -\mathbf{M}_2(\mathbf{x}_2)^{-1}\nabla_{\mathbf{x}_2} f_2(\mathbf{x}_1^*, \mathbf{x}_2) \tag{37}$$

Moreover, L^+ must be closed and bounded. Since (37) is a natural gradient system of a g-convex function, Thm. 2 ensures that any trajectory of (37) must converge to an equilibrium point that is a global minimizer for $f_2(\mathbf{x}_1^*, \mathbf{x}_2)$. Considering any initial condition of (37) that begins in L^+, we denote $(\mathbf{x}_1^*, \mathbf{x}_2^*) \in L^+$ as the resulting equilibrium point. However, since (35) is semi-contracting, any geodesic ball around $(\mathbf{x}_1^*, \mathbf{x}_2^*)$ is forward invariant for (35), which implies that $L^+ = \{(\mathbf{x}_1^*, \mathbf{x}_2^*)\}$. Thus, $\mathbf{x}_2(t) \to \mathbf{x}_2^*$ as $t \to \infty$. Note again that, while $\mathbf{x}_2^*$ depends on initial conditions, $\mathbf{x}_1^*$ does not.

Corollary 3. *Consider any two equilibrium points $(\mathbf{x}_1^*, \mathbf{x}_{21}^*)$ and $(\mathbf{x}_1^*, \mathbf{x}_{22}^*)$ for a system that satisfies the condition of Theorem 6. Then, the geodesic between these points is comprised of extremal Nash equilibrium points, all of which have the same cost.*

Further, if one equilibrium of (37) *is locally asymptotically stable, then it is necessarily globally attractive, and thus all trajectories of* (37) *converge to this unique equilibrium $(\mathbf{x}_1^*, \mathbf{x}_2^*)$ regardless of initial conditions.*

Proof. Proof of the first part follows immediately from applying Corollary 3.1 of [12] to the function $f_2(\mathbf{x}_1^*, \mathbf{x}_2)$. Proof of the second part follows immediately from the application of Corollary 1 herein.

Corollary 4. *Proposition 10 is true.*

Proof. Consider Thm. 6 with $f_1 = \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ and $f_2 = -\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$.

4.3 Hierarchical natural gradient

Consider a function $f_1(\mathbf{x}_1)$ α_1-strongly g-convex in a metric $\mathbf{M}_1(\mathbf{x}_1)$, and a function $f_2(\mathbf{x}_1, \mathbf{x}_2)$ α_2-strongly g-convex in a metric $\mathbf{M}_2(\mathbf{x}_2)$ for each given $\mathbf{x}_1$. Then, the hierarchical natural gradient dynamics

$$
\begin{aligned}
\dot{\mathbf{x}}_1 &= -\mathbf{M}_1(\mathbf{x}_1)^{-1} \, \nabla_{\mathbf{x}_1} f_1(\mathbf{x}_1) \\
\dot{\mathbf{x}}_2 &= -\mathbf{M}_2(\mathbf{x}_2)^{-1} \, \nabla_{\mathbf{x}_2} f_2(\mathbf{x}_1, \mathbf{x}_2)
\end{aligned}
$$

is contracting with rate $\min(\alpha_1, \alpha_2)$ in metric $\mathbf{M}(\mathbf{x}_1, \mathbf{x}_2) = \text{BlkDiag}(\mathbf{M}_1(\mathbf{x}_1), \mathbf{M}_2(\mathbf{x}_2))$, under the mild assumption that the coupling Jacobian is bounded [15]. Since the *overall* system is both contracting and autonomous, it tends to a unique equilibrium [15] at rate $\min(\alpha_1, \alpha_2)$, and thus to the unique solution of

$$
\begin{aligned}
\nabla_{\mathbf{x}_1} f_1(\mathbf{x}_1) &= \mathbf{0} \\
\nabla_{\mathbf{x}_2} f_2(\mathbf{x}_1, \mathbf{x}_2) &= \mathbf{0}
\end{aligned}
$$

By recursion, this structure can be chained an arbitrary number of times, or applied to any cascade or directed acyclic graph of natural gradient dynamics. Such hierarchical optimization may play a role, for instance, in backpropagation of natural gradients in machine learning, with all descents occurring concurrently rather than in sequence.

Remark 23. *In large-scale optimization settings such as those appearing commonly in machine learning, natural gradient with a fully-dense metric can become intractable. In specific cases, such as natural gradient descent based on Fisher information* [33], *computationally effective approximations have been derived* [69, 70]. *In addition, the combination of simple (e.g., diagonal) metrics through hierarchical structures lends an opportunity to recover significant complexity at broad scale − see, e.g., the hierarchical combination of scalar metrics to learn hierarchical representations of symbolic data in* [71, 72]. *Such simpler metrics are also well motivated in the context of positive or monotone systems* [73, 74]. *In special cases of a dense Hessian metric $\mathbf{M}(\mathbf{x}) = \nabla^2 \psi(\mathbf{x})$ from a potential $\psi(\mathbf{x})$, note that continuous mirror descent (see also Proposition 5 and Example 4) provides an alternate method to compute continuous natural gradient. These methods can avoid the need to invert the metric in cases where there is an explicit inverse exists for the change of variables $\mathbf{z} = \nabla \psi(\mathbf{x})$, or when* (15) *can be run at a fast time scale to invert the gradient map through dynamics.*

5 Conclusions

Overall, this paper has demonstrated that nonlinear contraction analysis provides a general perspective for analyzing and certifying the global convergence properties of gradient-based

optimization algorithms. The common case of strong convexity corresponds to the special case of contracting gradient descent in the identity metric, while our analysis admits global convergence results in the significantly broader case of state-dependent metrics. This result has clear links to the case of geodesically-convex optimization wherein natural gradient descent converges to a unique equilibrium if it is contracting in any metric, broadening from the special case of g-convexity corresponding to contraction in the natural metric. Our analysis of semi-contraction of gradient systems, and the resulting smoothly connected sets of global optima may shed additional light on applications in learning with over parameterized networks [2] where the set of optimizers is recognized to take the form of a low-dimensional manifold. Results on natural primal dual and the convergence to Nash equilibria showcase the broad reach of these fundamental results, where they may serve as the basis for the generation of larger scale distributed optimization algorithms in future work. A framework we call Contraction + shows how contraction or semi-contraction properties can be combined with specific but coarse information on a system, such as the local stability of a particular equilibrium or the weak decreasing of a cost or a Lyapunov-like function, to conclude on global convergence or minimization.

A natural next step for the application of contraction in optimization is to design geodesic quorum sensing [16, 75] algorithms for synchronization [76], as well as other consensus mechanisms considering time-delays [17, 18], which may serve as the basis for distributed and large-scale optimization techniques on Riemannian manifolds. Other future applications will consider stochastic gradient descent in the Riemannian setting [77] with quorum sensing extensions (as, e.g., in [78, 79]). Such advances could have direct applications, e.g., in the context of machine learning, among others.

Supporting information

S1 File.
(PDF)

Acknowledgments

Author Contributions

Conceptualization: Patrick M. Wensing, Jean-Jacques Slotine.

Formal analysis: Patrick M. Wensing, Jean-Jacques Slotine.

Methodology: Patrick M. Wensing, Jean-Jacques Slotine.

Writing – original draft: Patrick M. Wensing, Jean-Jacques Slotine.

Writing – review & editing: Patrick M. Wensing, Jean-Jacques Slotine.

References

1. Bassily R, Belkin M, Ma S. On exponential convergence of sgd in non-convex over-parametrized learning. arXiv preprint arXiv:181102564. 2018.

2. Cooper Y. The loss landscape of overparameterized neural networks. arXiv preprint arXiv:1804.10200. 2018.

3. Liu C, Zhu L, Belkin M. Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning. arXiv preprint arXiv:2003.00307. 2020.

4. Brea J, Simsek B, Illing B, Gerstner W. Weight-space symmetry in deep networks gives rise to permutation saddles, connected by equal-loss valleys across the loss landscape. arXiv preprint arXiv:190702911. 2019.

5. Sagun L, Evci U, Guney VU, Dauphin Y, Bottou L. Empirical analysis of the hessian of over-parametrized neural networks. arXiv preprint arXiv:170604454. 2017.

6. Allen-Zhu Z, Li Y, Song Z. A convergence theory for deep learning via over-parameterization. arXiv preprint arXiv:181103962. 2018.

7. Du SS, Zhai X, Poczos B, Singh A. Gradient descent provably optimizes over-parameterized neural networks. arXiv preprint arXiv:181002054. 2018.

8. Hanson MA. On sufficiency of the Kuhn-Tucker conditions. Journal of Mathematical Analysis and Applications. 1981; 80(2):545–550.

9. Zalinescu C. A critical view on invexity. Journal of Optimization Theory and Applications. 2014; 162 (3):695–704.

10. Polyak BT. Gradient methods for minimizing functionals. Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki. 1963; 3(4):643–653.

11. Karimi H, Nutini J, Schmidt M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer; 2016. p. 795–811.

12. Rapcsak T. Geodesic Convexity in Nonlinear Optimization. Journal of Optimization Theory and Applications. 1991; 69(1):169–183.

13. Zhang H, Sra S. First-order methods for geodesically convex optimization. In: Conference on Learning Theory; 2016. p. 1617–1638.

14. Absil PA, Mahony R, Sepulchre R. Optimization Algorithms on Matrix Manifolds. Princeton University Press; 2008.

15. Lohmiller W, Slotine JJE. On Contraction Analysis for Non-linear Systems. Automatica. 1998; 34 (6):683–696.

16. Tabareau N, Slotine JJ, Pham QC. How synchronization protects from noise. PLoS computational biology. 2010; 6(1):e1000637. https://doi.org/10.1371/journal.pcbi.1000637 PMID: 20090826

17. Wang W, Slotine JJE. Contraction analysis of time-delayed communications and group cooperation. IEEE Transactions on Automatic Control. 2006; 51(4):712–717.

18. Wensing PM, Slotine JJ. Cooperative Adaptive Control for Cloud-Based Robotics. Proceedings of the IEEE International Conference on Robotics and Automation. 2018.

19. Slotine JJE. Modular stability tools for distributed computation and control. International Journal of Adaptive Control and Signal Processing. 2003; 17(6):397–416.

20. Su W, Boyd S, Candes E. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In: Advances in Neural Information Processing Systems; 2014. p. 2510–2518.

21. Zhang J, Mokhtari A, Sra S, Jadbabaie A. Direct Runge-Kutta Discretization Achieves Acceleration. ArXiv e-prints. 2018.

22. Wibisono A, Wilson AC, Jordan MI. A variational perspective on accelerated methods in optimization. Proceedings of the National Academy of Sciences. 2016; 113(47):E7351–E7358. https://doi.org/10.1073/pnas.1614734113 PMID: 27834219

23. Krichene W, Bayen A, Bartlett PL. Accelerated mirror descent in continuous and discrete time. In: Advances in neural information processing systems; 2015. p. 2845–2853.

24. Nesterov Y. Introductory lectures on convex programming—A Basic course. Springer; 1998.

25. Nguyen HD, Vu TL, Turitsyn K, Slotine JJ. Contraction and Robustness of Continuous Time Primal-Dual Dynamics. IEEE Control Systems Letters. 2018; 2(4).

26. França G, Jordan MI, Vidal R. On Dissipative Symplectic Integration with Applications to Gradient-Based Optimization. arXiv preprint arXiv:2004.06840. 2020.

27. Slotine JJE, Li W. Applied nonlinear control. Prentice hall Englewood Cliffs, NJ; 1991.

28. Wiggins S. Gradient Vector Fields. In: Introduction to Applied Nonlinear Dynamical Systems and Chaos. New York, NY: Springer New York; 2003. p. 231–233.

29. Aylward EM, Parrilo PA, Slotine JJE. Stability and robustness analysis of nonlinear systems via contraction metrics and SOS programming. Automatica. 2008; 44(8):2163–2170.

30. Boyd S, Kim SJ, Vandenberghe L, Hassibi A. A tutorial on geometric programming. Optimization and Engineering. 2007; 8(1):67.

31. Sra S, Hosseini R. Conic Geometric Optimization on the Manifold of Positive Definite Matrices. SIAM Journal on Optimization. 2015; 25(1):713–739.

32. Udriste C. Convex functions and optimization methods on Riemannian manifolds. vol. 297. Springer Science & Business Media; 1994.

33. Amari SI. Natural Gradient Works Efficiently in Learning. Neural Comput. 1998; 10(2):251–276.

34. Gunasekar S, Lee J, Soudry D, Srebro N. Characterizing implicit bias in terms of optimization geometry. arXiv preprint arXiv:180208246. 2018.

35. Lovelock D, Rund H. Tensors, differential forms, and variational principles. Courier Corporation; 1989.

36. Lohmiller W, Slotine JJ. Exact decomposition and contraction analysis of nonlinear hamiltonian systems. In: AIAA Guidance, Navigation, and Control (GNC) Conference; 2013. p. 4931.

37. Lohmiller W, Slotine JJE. Exact Modal Decomposition of Nonlinear Hamiltonian Systems. In: AIAA Guidance, Navigation, and Control Conference; 2009. p. 5792:1–18.

38. Simpson-Porco JW, Bullo F. Contraction theory on Riemannian manifolds. Systems & Control Letters. 2014; 65:74–80.

39. Dauphin YN, Pascanu R, Gulcehre C, Cho K, Ganguli S, Bengio Y. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In: Advances in neural information processing systems; 2014. p. 2933–2941.

40. Jin C, Ge R, Netrapalli P, Kakade SM, Jordan MI. How to escape saddle points efficiently. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org; 2017. p. 1724–1732.

41. Lee JD, Simchowitz M, Jordan MI, Recht B. Gradient Descent Converges to Minimizers; 2016.

42. Lee JD, Panageas I, Piliouras G, Simchowitz M, Jordan MI, Recht B. First-order methods almost always avoid saddle points. arXiv preprint arXiv:171007406. 2017.

43. Kreusser LM, Osher SJ, Wang B. A Deterministic Approach to Avoid Saddle Points. arXiv preprint arXiv:1901.06827. 2019.

44. Lohmiller W, Slotine JJ. Exact Modal Decomposition of Nonlinear Hamiltonian Systems. In: AIAA Guidance, Navigation, and Control Conference; 2009. p. 5792.

45. Boyd S, Vandenberghe L. Convex Optimization. Cambridge, U.K.: Cambridge Univ. Press; 2004.

46. Wang W, Slotine JJE. On partial contraction analysis for coupled nonlinear oscillators. Biological cybernetics. 2005; 92(1):38–53. PMID: 15650898

47. Jouffroy J, Slotine JJE. Methodological remarks on contraction theory. In: IEEE Conference on Decision and Control. vol. 3; 2004. p. 2537–2543 Vol.3.

48. Manchester IR, Slotine JJE. Control Contraction Metrics: Convex and Intrinsic Criteria for Nonlinear Feedback Design. IEEE Transactions on Automatic Control. 2017; 62(6):3046–3053.

49. Cisneros-Velarde P, Jafarpour S, Bullo F. Distributed and time-varying primal-dual dynamics via contraction analysis. arXiv preprint arXiv:2003.12665. 2020.

50. Singh S, Majumdar A, Slotine JJ, Pavone M. Robust online motion planning via contraction theory and convex optimization. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE; 2017. p. 5883–5890.

51. Manchester IR, Slotine JJE. Transverse contraction criteria for existence, stability, and robustness of a limit cycle. Systems & Control Letters. 2014; 63:32–38.

52. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.

53. Chen TQ, Rubanova Y, Bettencourt J, Duvenaud DK. Neural ordinary differential equations. In: Advances in neural information processing systems; 2018. p. 6571–6583.

54. Dupont E, Doucet A, Teh YW. Augmented neural odes. In: Advances in Neural Information Processing Systems; 2019. p. 3134–3144.

55. Arrow KJ, Hurwicz L, and Uzawa H. Studies in Linear and Non-linear Programming. Stanford University Press, Stanford, CA; 1958.

56. Feijer D, Paganini F. Stability of primal–dual gradient dynamics and applications to network optimization. Automatica. 2010; 46(12):1974–1981.

57. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:170606083. 2017.

58. Tishby N, Zaslavsky N. Deep learning and the information bottleneck principle. In: 2015 IEEE Information Theory Workshop (ITW). IEEE; 2015. p. 1–5.

59. Cho WS, Wang M. Deep Primal-Dual Reinforcement Learning: Accelerating Actor-Critic using Bellman Duality. CoRR. 2017;abs/1712.02467.

60. Kosaraju KC, Mohan S, Pasumarthy R. On the primal-dual dynamics of Support Vector Machines. International Symposium on Mathematical Theory of Networks and Systems. 2018; p. 468–474.

61. Ortega R, Jeltsema D, Scherpen JM. Power shaping: A new paradigm for stabilization of nonlinear RLC circuits. IEEE Transactions on Automatic Control. 2003; 48(10):1762–1767.

62. Cavanagh K, Belk JA, Turitsyn K. Transient stability guarantees for ad hoc DC microgrids. IEEE Control Systems Letters. 2018; 2(1):139–144.

63. Slotine JJE, Li W. On the adaptive control of robot manipulators. The international journal of robotics research. 1987; 6(3):49–59.

64. Lee T, Kwon J, Park FC. A Natural Adaptive Control Law for Robot Manipulators. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2018. p. 1–9.

65. Lee T. Geometric Methods for Dynamic Model-Based Identification and Control of Multibody Systems. Seoul National University; 2019.

66. Nesterov YE, Todd MJ, et al. On the Riemannian geometry defined by self-concordant barriers and interior-point methods. Foundations of Computational Mathematics. 2002; 2(4):333–361.

67. Lopez BT, Slotine JJE. Contraction Metrics in Adaptive Nonlinear Control. arXiv e-prints. 2019; p. arXiv:1912.13138.

68. Lee T, Kwon J, Park FC. A Natural Adaptive Control Law for Robot Manipulators. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2018. p. 1–9.

69. Martens J, Grosse R. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning; 2015. p. 2408–2417.

70. Amari Si, Karakida R, Oizumi M. Information geometry connecting Wasserstein distance and Kullback–Leibler divergence via the entropy-relaxed transportation problem. Information Geometry. 2018; 1 (1):13–37. PMID: 30883281

71. Nickel M, Kiela D. Poincaré Embeddings for Learning Hierarchical Representations. CoRR. 2017;abs/ 1705.08039.

72. Nickel M, Kiela D. Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry. ArXiv e-prints. 2018.

73. Rantzer A. Scalable control of positive systems. European Journal of Control. 2015; 24:72–80.

74. Manchester IR, Slotine JJE. On Existence of Separable Contraction Metrics for Monotone Nonlinear Systems. IFAC-PapersOnLine. 2017; 50(1):8226–8231.

75. Russo G, Slotine JJE. Global convergence of quorum-sensing networks. Physical Review E. 2010; 82-(4):041919. PMID: 21230325

76. Bouvrie J, Slotine JJ. Synchronization Can Control Regularization in Neural Systems via Correlated Noise Processes. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. USA; 2012. p. 854–862.

77. Bonnabel S. Stochastic Gradient Descent on Riemannian Manifolds. IEEE Transactions on Automatic Control. 2013; 58(9):2217–2229.

78. Zhang S, Choromanska A, LeCun Y. Deep Learning with Elastic Averaging SGD. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press; 2015. p. 685–693.

79. Boffi NM, Slotine JJE. A Continuous-Time Analysis of Distributed Stochastic Gradient. Neural Computation. 2020; 32(1):36–96. PMID: 31703177