Convergence Properties of Natural Gradient Descent for Minimizing KL Divergence

Anonymous authors
Paper under double-blind review

Abstract

The Kullback-Leibler (KL) divergence plays a central role in probabilistic machine learning, where it commonly serves as the canonical loss function. Optimization in such settings is often performed over the probability simplex, where the choice of parameterization significantly impacts convergence. In this work, we study the problem of minimizing the KL divergence and analyze the behavior of gradient-based optimization algorithms under two dual coordinate systems within the framework of information geometry− the exponential family (θ coordinates) and the mixture family (η coordinates). We compare Euclidean gradient descent (GD) in these coordinates with the coordinate-invariant natural gradient descent (NGD), where the natural gradient is a Riemannian gradient that incorporates the intrinsic geometry of the parameter space. In continuous time, we prove that the convergence rates of GD in the θ and η coordinates provide lower and upper bounds, respectively, on the convergence rate of NGD. Moreover, under affine reparameterizations of the dual coordinates, the convergence rates of GD in η and θ coordinates can be scaled to $2c$ and $\frac{2}{c}$, respectively, for any $c > 0$, while NGD maintains a fixed convergence rate of 2, remaining invariant to such transformations and sandwiched between them. Although this suggests that NGD may not exhibit uniformly superior convergence in continuous time, we demonstrate that its advantages become pronounced in discrete time, where it achieves faster convergence and greater robustness to noise, outperforming GD. Our analysis hinges on bounding the spectrum and condition number of the Hessian of the KL divergence at the optimum, which coincides with the Fisher information matrix.

1 Introduction

The convergence properties of the natural gradient descent algorithm, originally introduced in Amari (1996), have been extensively studied in the literature (e.g., Amari (1998); Pascanu & Bengio (2014); Martens (2020)). In particular, the natural policy gradient Kakade (2001) has motivated a rich body of research (see, e.g., Müller & Montúfar (2024); Yuan et al. (2022); Khodadadian et al. (2022).) Beyond this, the natural gradient methods have been applied to diverse problems including Bayesian networks Ay (2023); Ay & van Oostrum (2023), over-parametrized neural networks Zhang et al. (2019); van Oostrum & Ay (2021); van Oostrum et al. (2023) and infinitely-wide networks Karakida et al. (2019); Karakida & Osawa (2021) to name a few. Related to our focus, recent work has also investigated convergence rates of natural gradient flows and their discrete counterparts (see Zhang et al. (2019); Xiao (2022); Yuan et al. (2022); Khodadadian et al. (2022); Müller & Montúfar (2024)). A commonly observed phenomenon is that natural gradient descent outperforms Euclidean gradient descent, albeit at a higher computational cost. In this work, we revisit this comparison in a simple yet illuminating setting: minimizing the Kullback-Leibler (KL) divergence over the probability simplex. The KL divergence is a fundamental loss function in probabilistic machine learning, arising naturally from the maximum likelihood principle and information-theoretic considerations (Mohri et al., 2018, Section 12.1.1). Despite the apparent simplicity of the problem, we observe that natural gradient flows do not universally outperform standard Euclidean gradient flows.

Specifically, we consider two dual parametrizations of the probability simplex: the exponential family representation (the θ coordinates) and the mixture family representation (the η coordinates) Amari (2016). We prove that the natural gradient flow converges faster than the Euclidean gradient flow in the θ coordinates (the *θ-gradient flow*), consistent with results in the literature. However, the natural gradient flow (despite yielding straight-line trajectories) converges more slowly than the Euclidean gradient flow in η coordinates ($\eta-$gradient flow). This demonstrates that the often-reported rapid convergence of natural gradient flow cannot be simplistically attributed to the straightness of its trajectories. Furthermore, leveraging the invariance of the canonical divergence under affine transformations, we show that the convergence rates of the Euclidean gradient flows in η and θ coordinates can be adjusted to $2c$ and $\frac{2}{c}$, respectively, for an arbitrary $c > 0$, while the natural gradient maintains a fixed convergence rate of 2, sandwiched between them. Thus, by setting $c = 1$, we can construct a pair of dual coordinates, $(\bar{\eta}, \bar{\theta})$, that match the convergence rate of the natural gradient flow. Since the advantages of the natural gradient are not immediately apparent in the continuous-time setting, we extend our analysis to the discrete-time case, where the natural gradient demonstrates both faster convergence and greater robustness to noise, outperforming Euclidean gradient descent. We show that the fundamental reason behind the superiority of natural gradient lies in the optimal conditioning of the loss landscape: the natural gradient updates are equivalent to minimizing the loss function $\frac{1}{2}\|\eta - \eta_q\|^2$, whose Hessian has a condition number equal to 1. The main contributions of this paper are summarized as follows:

1. We analyze the convergence rates of Euclidean gradient flows in η and θ coordinates, and of the natural gradient flow (Theorem 3 and Theorem 9). We show that while the natural gradient flow converges faster than the θ-gradient flow, it is slower than the η-gradient flow. This result builds upon bounds on the spectrum of the Hessian of the loss function established in Lemma 2. These theoretical findings are supported by illustrative numerical experiments.
2. Exploiting the duality and the invariance of the canonical divergence under affine transformations, we demonstrate in Theorem 4 that the convergence rates of Euclidean gradient flows in η and θ coordinates can be adjusted to $2c$ and $\frac{2}{c}$, respectively, for an arbitrary $c > 0$, while the natural gradient maintains a fixed convergence rate of 2, sandwiched between them. This shows that there exists a pair of dual coordinates, $(\bar{\eta}, \bar{\theta})$ such that the convergence rate of $\bar{\eta}-$ and $\bar{\theta}-$gradient flows matches the convergence rate of the natural gradient flow.
3. We analyze the discrete-time dynamics in Section 4, where Theorems 7 and 8 establish the superior robustness properties of natural gradient dynamics compared to their Euclidean counterparts. The core reason for this superiority is the optimal conditioning of the underlying loss landscape. In particular, natural gradient updates can be interpreted as minimizing the loss function $\frac{1}{2}\|\eta - \eta_q\|^2$, whose Hessian exhibits the ideal condition number of 1.

Notation

Let $\mathbb{R}$ denote the set of real numbers. For a function g of two variables $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, we use the notation

$$\nabla_x g(x,y) = \begin{bmatrix} \frac{\partial g}{\partial x_1}(x,y) \\ \vdots \\ \frac{\partial g}{\partial x_n}(x,y) \end{bmatrix}, \quad \nabla_x^2 g(x,y) = \begin{bmatrix} \frac{\partial^2 g}{\partial x_1^2}(x,y) & \cdots & \frac{\partial^2 g}{\partial x_1 \partial x_n}(x,y) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 g}{\partial x_n \partial x_1}(x,y) & \cdots & \frac{\partial^2 g}{\partial x_n^2}(x,y) \end{bmatrix}.$$

For functions f of a single variable x, we suppress the subscript and simply write $\nabla f(x)$ and $\nabla^2 f(x)$. For a manifold $\mathcal{M}$ with two global charts $\phi_m : \mathcal{M} \to \phi_m(\mathcal{M}) \subset \mathbb{R}^n$ and $\phi_e : \mathcal{M} \to \phi_e(\mathcal{M}) \subset \mathbb{R}^n$ with coordinates $\eta \in \phi_m(\mathcal{M})$ and $\theta \in \phi_e(\mathcal{M})$, we slightly abuse notation and write for any smooth function $\mathcal{L} : \mathcal{M} \to \mathbb{R}$,

$$\begin{aligned} \mathcal{L}(\eta) &:= \mathcal{L}(\phi_m^{-1}(\eta)), & \mathcal{L}(\theta) &:= \mathcal{L}(\phi_e^{-1}(\theta)), \\ \nabla\mathcal{L}(\eta) &:= \nabla\left(\mathcal{L} \circ \phi_m^{-1}\right)(\eta), & \nabla\mathcal{L}(\theta) &:= \mathcal{L}(\phi_e^{-1}(\theta)), \\ \nabla^2\mathcal{L}(\eta) &:= \nabla^2\left(\mathcal{L} \circ \phi_m^{-1}\right)(\eta), & \nabla^2\mathcal{L}(\theta) &:= \nabla^2\left(\mathcal{L} \circ \phi_e^{-1}\right)(\theta). \end{aligned}$$

For a point $p \in \mathcal{M}$, we write $\eta_p = \phi_m(p)$ and $\theta_p = \phi_e(p)$ to denote the point p in the η and θ coordinates, respectively. For a symmetric matrix Q, we write $Q \succ 0$ (resp. $Q \succeq 0$) to denote that Q is symmetric positive definite (resp. positive semi-definite). Building on this notation, we write $Q \succ P$ (resp. $Q \succeq P$) to mean $Q - P \succ 0$ (resp. $Q - P \succeq 0$). For a symmetric matrix Q, let $\lambda_{\min}(Q)$ and $\lambda_{\max}(Q)$ denote the minimum and maximum eigenvalue of Q. Since $Q \succ 0$ implies that all eigenvalues of Q are positive, we can define the condition number of Q as $\mathrm{cond}(Q) := \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)}$. For any $x \in \mathbb{R}^n$, let $||x||$ denote the standard Euclidean norm, and define the closed norm ball of radius ε centered at x by $\mathcal{B}_\varepsilon(x) := \{y \in \mathbb{R}^n : ||y - x|| \leq \varepsilon\}$. For any real matrix M, let $\|M\|_2 := \sup_{x \neq 0} \frac{\|Mx\|}{\|x\|}$ be the induced matrix 2-norm, which coincides with the maximum eigenvalue of M when $M \succ 0$.

2 Information Geometry Preliminaries and Gradient Flow Dynamics

In this section, we review the information geometric preliminaries and arrive at the continuous-time gradient flow dynamics which are then analyzed in the following section. For further details on the underlying concepts of information geometry, the reader is referred to Amari (2016); Amari & Nagaoka (2000); Ay et al. (2017).

2.1 Discrete Distributions in Mixture and Exponential Coordinates

Consider the family S_n of probability distributions over a discrete random variable X with sample space $\Omega = \{1, 2, \cdots, n, n+1\}$. Let p_i be the probability that X takes the value i. Then, any $p \in S_n$ can be written as

$$p(x) = \sum_{i=1}^{n+1} p_i \delta_i(x),$$

where $\delta_i(x)$ is the delta distribution over Ω, concentrated at i. Thus, S_n can be identified with the $n-$dimensional simplex[1], i.e., $S_n = \left\{(p_1, p_2, \cdots, p_n, p_{n+1}) \in \mathbb{R}^{n+1} | \; p_i > 0, \; \sum_{i=1}^{n+1} p_i = 1\right\}$. This family admits representations both as a mixture family and an exponential family Amari (2016). This can be seen by noticing that any $p \in S_n$ can be written as

$$p(x) = \underbrace{\sum_{i=1}^{n} \eta_i \delta_i(x) + \left(1 - \sum_{k=1}^{n} \eta_k\right) \delta_{n+1}(x)}_{\text{Mixture family representation}} = \underbrace{\exp\left(-\psi(\theta) + \sum_{i=1}^{n} \theta_i \delta_i(x)\right)}_{\text{Exponential family representation}},$$

where $\psi(\theta) = \log\left(1 + \sum_{i=1}^{n} e^{\theta_i}\right)$ is the log-partition function ensuring the normalization constraint $\sum_{x \in \Omega} p_\theta(x) = 1$ for the exponential family representation. With $\eta = (\eta_1, \eta_2, \cdots, \eta_n)$, we obtain a coordinate system for the simplex, with η serving as the natural parameter of the mixture family. We let $\phi_m : S_n \ni p \mapsto \eta = (\eta_1, \eta_2, \cdots, \eta_n) \in \mathbb{R}^n$ denote the global chart for S_n in the mixture family coordinate system. This is depicted in Fig. 1 (left). Similarly, with $\theta := (\theta_1, \theta_2, \cdots, \theta_n)$, we obtain an alternate coordinate system for the simplex with θ being the natural parameter of the exponential family. Define $\phi_e : S_n \ni p \mapsto \theta = (\theta_1, \theta_2, \cdots, \theta_n) \in \mathbb{R}^n$ as the global chart for S_n in the exponential family coordinate system. This is depicted in Fig. 1 (right).

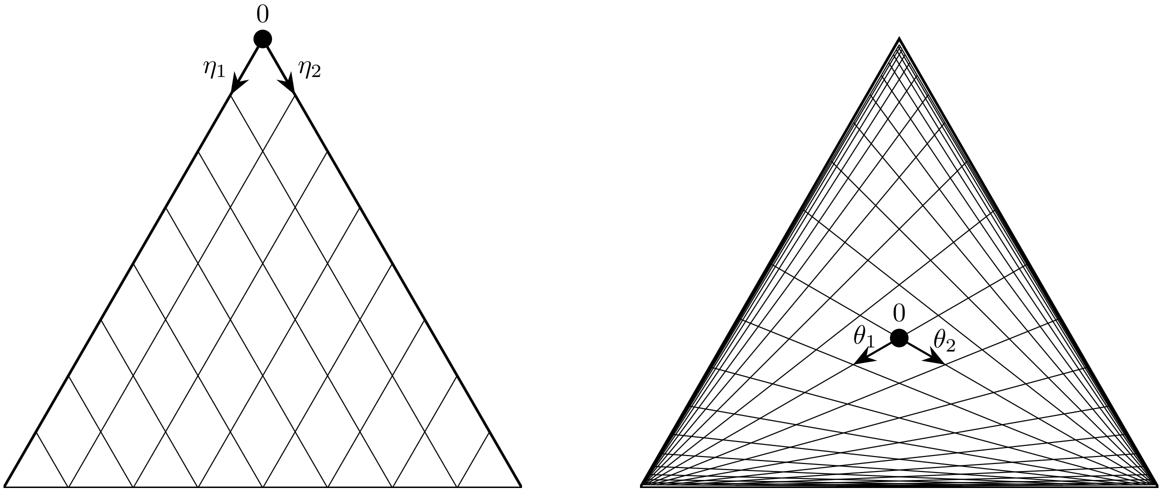
2.2 Convex Duality and Bregman Divergence

For the family S_n, there exists a dual relationship between the coordinates η and θ. Since $\psi(\theta) = \log\left(1 + \sum_{i=1}^{n} e^{\theta_i}\right)$ is strictly convex, it is possible to define its convex conjugate $\varphi(\eta) = \max_\vartheta \left(\eta^T \vartheta - \psi(\vartheta)\right)$. Optimality condition on the maximizer $\vartheta_{\text{opt}} = \theta$ yields the relationship $\nabla\psi(\theta) = \eta$, which can be solved to obtain $\theta_i = \log\left(\frac{\eta_i}{1 - \sum_{i=1}^{n} \eta_i}\right)$. This results in $\varphi(\eta) = \left(\eta^T \theta - \psi(\theta)\right) = \sum_{i=1}^{n+1} \eta_i \log \eta_i$, the negative of Shannon entropy. Conversely, ψ is the convex conjugate of φ leading to $\theta = \nabla\varphi(\eta)$. Since $\nabla_\eta \varphi(\nabla_\theta \psi(\cdot))$ is the identity map, application of the chain rule gives

$$\nabla^2 \varphi(\eta) = \left[\nabla^2 \psi(\theta)\right]^{-1}. \tag{1}$$

---

[1] Note that our definition of the simplex excludes the boundary.



$$\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} \mapsto \begin{bmatrix}\eta_1\\ \eta_2\\ 1-\eta_1-\eta_2\end{bmatrix} \qquad \begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix} \mapsto \frac{1}{1+e^{\theta_1}+e^{\theta_2}}\begin{bmatrix}e^{\theta_1}\\ e^{\theta_2}\\ 1\end{bmatrix}$$

Figure 1: Left: Coordinate system with the natural parameters $\eta = (\eta_1, \eta_2)$ of the mixture family representation of S_2. Right: Coordinate system with the natural parameters $\theta = (\theta_1, \theta_2)$ of the exponential family representation of S_2.

The convex conjugate functions ψ and φ define a pair of Bregman Divergence D_ψ and D_φ satisfying

$$D_\psi(\theta_p \parallel \theta_q) := \psi(\theta_p) - \psi(\theta_q) - \nabla\psi(\theta_q)^T(\theta_p - \theta_q) \tag{2}$$

$$= \varphi(\eta_q) - \varphi(\eta_p) - \nabla\varphi(\eta_p)^T(\eta_q - \eta_p) =: D_\varphi(\eta_q \parallel \eta_p). \tag{3}$$

In our setting of S_n, this Bregman divergence equals the canonical KL-divergence $D(q||p)$ between probability distributions q and p, i.e.,

$$D_\psi(\theta_p \parallel \theta_q) = D_\varphi(\eta_q \parallel \eta_p) = D(q||p) = \sum_{i=1}^{n+1} q_i \log\left(\frac{q_i}{p_i}\right), \tag{4}$$

where (η_q, η_p) and (θ_q, θ_p) are the coordinate representations of (q, p) in the η and θ coordinates, respectively. For further details, the reader is referred to Amari & Nagaoka (2000).

2.3 Gradient and Natural Gradient Dynamics

For a given target distribution $q \in S_n$, let the loss function $\mathcal{L}_q : S_n \to \mathbb{R}$ be defined by $\mathcal{L}_q(p) = D(q||p)$. As discussed in Section 1, we abuse the notation slightly and interchangeably use q, $\theta_q = \phi_e(q)$ or $\eta_q = \phi_m(q)$ to denote probability distribution $q \in S_n$, the same distribution in θ coordinates and in η coordinates, respectively. Thus, we write $\mathcal{L}_q(\eta_p)$ to mean $\mathcal{L}_q(\phi_m^{-1}(\eta_p))$ and $\mathcal{L}_q(\theta_p)$ to mean $\mathcal{L}_q(\phi_e^{-1}(\theta_p))$. The gradient of the loss function can be computed in the different coordinate systems using equation 2 and equation 4 as

$$\nabla\mathcal{L}_q(\eta_p) = -\nabla\varphi(\eta_p) - \nabla^2\varphi(\eta_p)(\eta_q - \eta_p) + \nabla\varphi(\eta_p) = -\nabla^2\varphi(\eta_p)(\eta_q - \eta_p), \tag{5}$$

$$\nabla\mathcal{L}_q(\theta_p) = \nabla\psi(\theta_p) - \nabla\psi(\theta_q). \tag{6}$$

Analogously, for a given target distribution $p \in S_n$, let the loss function $\mathcal{L}_p^* : S_n \to \mathbb{R}$ be defined by $\mathcal{L}_p^*(q) = D(q||p)$. The gradient of this loss function can be computed in the different coordinate systems using equation 2 and equation 4 as

$$\nabla\mathcal{L}_p^*(\eta_q) = \nabla\varphi(\eta_q) - \nabla\varphi(\eta_p), \tag{7}$$

$$\nabla\mathcal{L}_p^*(\theta_q) = -\nabla\psi(\theta_q) - \nabla^2\psi(\theta_q)(\theta_p - \theta_q) + \nabla\psi(\theta_q) = -\nabla^2\psi(\theta_q)(\theta_p - \theta_q). \tag{8}$$

Building on the pair of conjugate dual functions, we define a Riemannian metric g on S_n, which assigns to each point $p \in S_n$ an inner product $\langle \cdot, \cdot \rangle_p$ on the tangent space T_pS_n. The tangent space T_pS_n can be identified with $\left\{(v_1, v_2, \cdots, v_n, v_{n+1}) \in \mathbb{R}^{n+1} \middle| \sum_{i=1}^{n+1} v_i = 0\right\}$ (see Ay et al. (2017)). In coordinates, this Riemannian metric is represented by the matrix $\nabla^2\varphi(\eta)$ in the η coordinates and $\nabla^2\psi(\theta)$ in the θ coordinates. Importantly, this Riemannian metric coincides with the Fisher metric (Amari, 2016, Theorem 2.1). Using this Riemannian metric, we define the Riemannian gradient grad $\mathcal{L}_q(p)$, also called as the natural gradient, at a point $p \in S_n$ through the relation

$$\langle \text{grad } \mathcal{L}_q(p), v \rangle_p = d\mathcal{L}_q(p)[v] \tag{9}$$

for all v in the tangent space of S_n at the base point p. This defining property allows us to compute the natural gradient in η coordinates using equation 5 as

$$\text{grad } \mathcal{L}_q(\eta_p) = \left[\nabla^2\varphi(\eta_p)\right]^{-1} \nabla\mathcal{L}_q(\eta_p) = -(\eta_q - \eta_p).$$

Similarly, grad $\mathcal{L}_p^*(q)$ can be computed in the θ coordinates using equation 8 as grad $\mathcal{L}_p^*(\theta_q) = -(\theta_p - \theta_q)$. Interestingly, the natural gradients when represented in appropriate coordinates take on particularly simple linear forms – they directly point towards the target distributions. Since the situation with the loss function $\mathcal{L}_p^*$ is analogous to that of $\mathcal{L}_q$, for brevity, we will focus our analysis on $\mathcal{L}_q$ for the remainder of the paper[2].

We now introduce the gradient flow dynamics in both η and θ coordinates, as well as the natural gradient flow which will be analyzed in the subsequent sections. Although the natural gradient flow dynamics are coordinate-invariant, we express them in η coordinates to exploit the particularly simple linear structure.

For a given target distribution $q \in S_n$ and an initial distribution $p_0 \in S_n$, consider the gradient flow dynamics described by equation 10 and equation 11 and the natural gradient flow dynamics described by equation 12

$$\dot{\eta}(t) = -\nabla\mathcal{L}_q(\eta(t)), \qquad \eta(0) = \eta_{p_0}, \tag{10}$$

$$\dot{\theta}(t) = -\nabla\mathcal{L}_q(\theta(t)), \qquad \theta(0) = \theta_{p_0}, \tag{11}$$

$$\dot{\eta}_{ng}(t) = -\text{grad } \mathcal{L}_q(\eta_{ng}(t)), \quad \eta_{ng}(0) = \eta_{p_0}. \tag{12}$$

We will analyze these dynamics in the following sections.

3 Convergence Analysis in Continous Time

We start the convergence analysis with a general result which is at the core of the analysis. This result frequently appears in various forms, typically emphasizing the upper bound in inequality 14 (see (Wensing & Slotine, 2020, Proposition 1) for example). We present an adaptation of this result to our setting.

Proposition 1 (Convergence of general gradient flows)**.** *Consider the gradient flow dynamics*

$$\dot{x}(t) = -\nabla f(x(t)), \quad x(t_0) = x_0, \tag{13}$$

where a sufficiently smooth function $f : U \to \mathbb{R}$, with $U \subset \mathbb{R}^n$ being an open neighborhood of x_0, satisfies the following properties:

(a) f has a unique minimizer $x_* \in U$ satisfying $\nabla f(x_*) = 0$.

(b) there exist positive constants m and L such that $m \cdot I \preceq \nabla^2 f(x) \preceq L \cdot I$ for all x in the sublevel set $S := \{x \in U | f(x) \leq f(x_0)\}$.

Then the solution $x : [t_0, \infty) \to U$ of equation 13 satisfies

(i) $x(t) \in S$ for all $t \geq t_0$ and

---

[2] We include the convergence rate analysis of gradient flows for $\mathcal{L}_p^*$ in Appendix A for completeness.

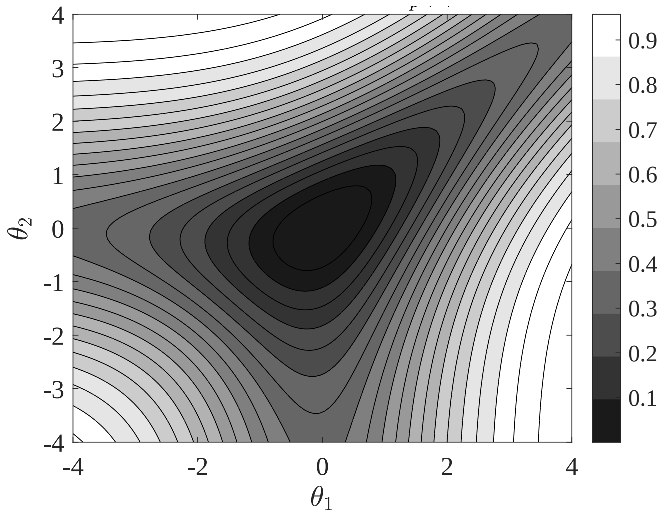


Figure 2: Contour plot of $\mathcal{L}_p^*(\theta)$ showing that the sublevel sets are non-convex.

(ii) With $c = f(x_0) - f(x_*)$*, we get that*

$$c \cdot e^{-2L(t-t_0)} \leq f(x(t)) - f(x_*) \leq c \cdot e^{-2m(t-t_0)} \text{ for all } t \geq t_0, \tag{14}$$

i.e., $f(x(t))$ *converges exponentially to* $f(x_*)$ *with a rate larger than* $2m$ *and smaller than* $2L$*.*

Proof. See Appendix C.1 □

To facilitate the application of Proposition 1 to the dynamics given in equation 10 and equation 11, we establish bounds on the Hessian of the loss function in the following Lemma.

Lemma 2 (Bounds on the Hessian of the loss function)**.** *Let* $p, q \in S_n$*. Then the following statements hold:*

(i) (Global bound) The Hessians of $\mathcal{L}_q$ *and* $\mathcal{L}_p^*$ *satisfy*

$$0 \prec \nabla^2 \mathcal{L}_q(\theta) \prec I \prec \nabla^2 \mathcal{L}_p^*(\eta) \qquad \forall\, \theta \in \phi_e(S_n), \forall\, \eta \in \phi_m(S_n) \qquad \text{and} \tag{15}$$

$$0 \prec \nabla^2 \mathcal{L}_q(\eta) \qquad \forall\, \eta \in \phi_m(S_n). \tag{16}$$

(ii) (Local bound at optimum) The Hessians of $\mathcal{L}_q$ *and* $\mathcal{L}_p^*$ *evaluated at the optimum satisfy*

$$I \prec \nabla^2 \mathcal{L}_q(\eta_q) = \nabla^2 \mathcal{L}_q(\theta_q)^{-1}, \tag{17}$$

$$I \prec \nabla^2 \mathcal{L}_p^*(\eta_p) = \nabla^2 \mathcal{L}_p^*(\theta_p)^{-1}, \tag{18}$$

Proof. See Appendix C.2 □

The positive definiteness of the Hessians established in Lemma 2 (see equation 15 and equation 16) shows that $\mathcal{L}_q$ is convex in the θ coordinates as well as in the η coordinates, and $\mathcal{L}_p^*$ is convex in the η coordinates. One might conjecture that $\mathcal{L}_p^*$ might likewise be convex in the θ coordinates. However, this turns out not to be the case, as illustrated by a counterexample. Figure 2 shows a sample contour plot of $\mathcal{L}_p^*(\theta)$ for $n = 2$. The plot clearly shows that the sublevel sets of $\mathcal{L}_p^*(\theta)$ are non-convex, disproving the conjecture. To summarize, while the KL divergence is geodesically convex along m-geodesics in both of its arguments, it is geodesically convex along e-geodesics only with respect to its second argument (see (Boumal, 2023, Definition 11.3) for definition of geodesic convexity and (Amari, 2016, Section 2.4) for definitions of $e-$ and $m-$ geodesics).

With these uniform bounds on the Hessians in place, we are now equipped to control the exponential decay rates of gradient flows through Proposition 1. This is presented in the next result which is the main result of this section.

Theorem 3 (Convergence analysis)**.** *Let $q \in S_n$ be the target distribution and $p_0 \in S_n$ be the initial distribution. Suppose η, θ and η_{ng} be the solutions to dynamics described by equation 10, equation 11 and equation 12, respectively. Then*

(i) there exist positive constants $1 < m_\eta \leq L_\eta$, c_η, $\bar{c}_\eta$ and T such that

$$c_\eta e^{-2L_\eta t} \leq \mathcal{L}_q(\eta(t)) \leq \bar{c}_\eta e^{-2m_\eta t} \leq \bar{c}_\eta e^{-2t} \qquad \forall t \geq T \tag{19}$$

i.e., $\mathcal{L}_q(\eta(t))$ converges to zero exponentially with rate higher than 2. Furthermore, if $\mathcal{L}_q(\eta(0))$ is sufficiently small, the result holds with $T = 0$.

(ii) there exist positive constants $m_\theta \leq L_\theta < 1$ and c_θ such that

$$c_\theta e^{-2t} \leq c_\theta e^{-2L_\theta t} \leq \mathcal{L}_q(\theta(t)) \leq c_\theta e^{-2m_\theta t} \qquad \forall t \geq 0, \tag{20}$$

i.e., $\mathcal{L}_q(\theta(t))$ converges to zero exponentially with rate lower than 2.

(iii) there exist positive constants c_1 and c_2 such that

$$c_1 e^{-2t} \leq \mathcal{L}_q(\eta_{ng}(t)) \leq c_2 e^{-2t} \qquad \forall t \geq 0, \tag{21}$$

i.e., $\mathcal{L}_q(\eta_{ng}(t))$ converges to zero exponentially with rate 2.

Proof. See Appendix C.3 □

Theorem 3 shows that gradient dynamics in the mixture family coordinates exhibit faster convergence rates than natural gradient dynamics, which, in turn, outperform gradient dynamics in the exponential family coordinates. On one hand, this supports the generally observed superiority of the natural gradient dynamics over gradient dynamics in the exponential family coordinates. On the other hand, it demonstrates that natural gradient dynamics are slower than gradient dynamics in the mixture family coordinates. Note that although we choose to represent the natural gradient dynamics in the η coordinates, the obtained convergence rate bound in equation 21 is independent of this choice. Note that the convergence rate bound for the η coordinates established in equation 19 in Theorem 3 is asymptotic in nature and formally holds only after some time $T > 0$, whose value is not explicitly characterized. This is a common feature of asymptotic convergence rate analyses, where rate guarantees apply beyond some unknown T. Finally, as stated in the last sentence of item (i), Theorem 3, the result can be interpreted as describing local behavior: if the initial distribution lies sufficiently close to the target, the bound holds from the start (i.e., with $T = 0$). More precisely, a sufficient condition for the bound to hold with $T = 0$ is that $\nabla \mathcal{L}_q(\eta) \succ I$ for all η belonging to the sublevel set $S_\eta := \{\eta \in \phi_m(S_n) | \mathcal{L}_q(\eta) \leq \mathcal{L}_q(\eta(0))\}$. Despite the absence of a global estimate on T, our empirical results (see Figure 4 for $n = 2$ and Figure 7 for $n = 10$) indicate that the ordering of convergence rates predicted by theory emerges early in the flow. This supports the practical relevance of the asymptotic comparison.

We now present numerical experiments with $n = 2$ to illustrate the theoretical results developed so far. Figure 3 (left) depicts the trajectories of the $\eta-$gradient flow described by equation 10, $\theta-$gradient flow described by equation 11 and the natural gradient flow described by equation 12 superimposed on the level curves of KL divergence. Although the natural gradient flow follows straight trajectories, it is slower than the $\eta-$gradient flow but faster than the $\theta-$gradient flow, as seen from the unit time markers along the curves. Figure 3 (right) confirms this by plotting the KL divergence over time along these trajectories.

To better highlight the convergence rates, Figure 4 (left) presents the KL divergence on a logarithmic scale, revealing exponential convergence. Best-fit linear curves are superimposed to estimate the slopes, which correspond to the convergence rates. The $\eta-$gradient flow exhibits the fastest convergence (slope ≈ 7), the

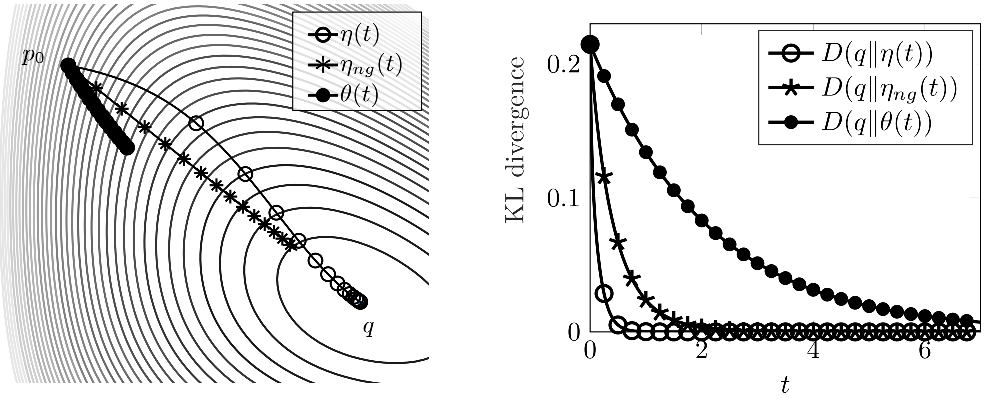


Figure 3: Left: Simulation trajectories of $\eta-$gradient flow described by equation 10, $\theta-$gradient flow described by equation 11 and the natural gradient flow described by equation 12 for $n = 2$ and $t \in [0, 1.5]$ superimposed on the level curves of KL divergence. The markers on the curves show equal time intervals for each curve. Right: KL divergence evaluated along the solutions plotted as a function of time t. The intervals between markers in the left figure are unrelated to the intervals between markers in the right figure.

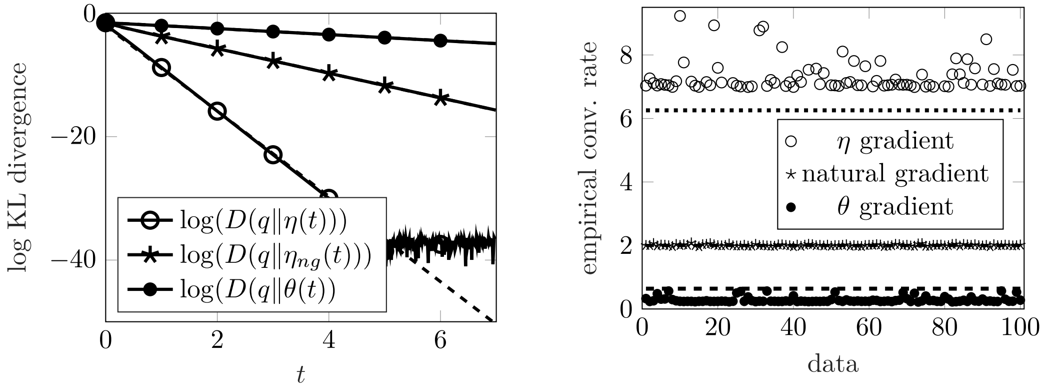


Figure 4: Left: KL divergence evaluated along the solutions to $\eta-$gradient flow described by equation 10, $\theta-$gradient flow described by equation 11 and the natural gradient flow described by equation 12 for $n = 2$ plotted on semi-log scale. The dashed lines show the best-fit linear function used to estimate the slope which gives the convergence rate. Right: Empirical convergence rates for 100 randomly chosen initial distributions and a randomly chosen target distribution for $n = 2$. The dotted line shows the theoretical lower bound for the convergence rate of $\eta-$gradient flows and the dashed line shows the theoretical upper bound for the convergence rate of $\theta-$gradient flows.

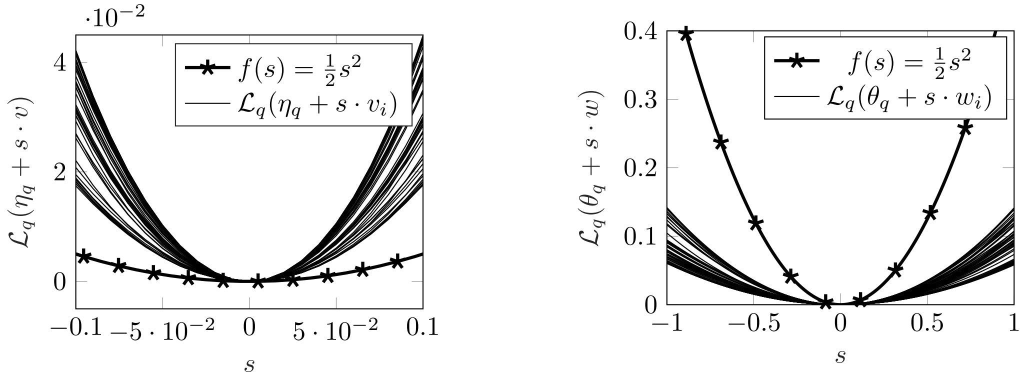


Figure 5: Left: Local sections of KL divergence around the minimizer q plotted as $\mathcal{L}_q(\eta_q + s \cdot v_i)$, where all v_i are unit norm vectors distributed evenly on the unit circle. A quadratic function $f(s) = \frac{1}{2}s^2$ is also shown for reference. Right: Local sections of KL divergence around the minimizer q plotted as $\mathcal{L}_q(\theta_q + s \cdot w_i)$, where all w_i are unit norm vectors distributed evenly on the unit circle. A quadratic function $f(s) = \frac{1}{2}s^2$ is also shown for reference.

$\theta-$gradient flow is slowest (slope ≈ 0.475), and the natural gradient flow lies in between (slope ≈ 2.04), closely matching the theoretical prediction of rate 2. This experiment is repeated over 100 randomly chosen initial conditions and one randomly chosen target distribution, as shown in Figure 4 (right). The empirical convergence rates align well with the theoretical bounds from Theorem 3, confirming that $\eta-$gradient flows exceed rate 2, natural gradient flows converge at rate 2, and $\theta-$gradient flows fall below rate 2.

Finally, note that the natural gradient dynamics governed by equation 12 can be equivalently described as

$$\dot{\eta}_{ng}(t) = -\text{grad } \mathcal{L}_q(\eta_{ng}(t)) = \eta_q - \eta_{ng}(t) = -\nabla f_q(\eta_{ng}(t)),$$

where $f_q(\eta_{ng}(t)) = \frac{1}{2}\|\eta_q - \eta_{ng}(t)\|^2$. The natural gradient dynamics thus correspond to minimizing a convex quadratic function with identity Hessian. Figure 5 offers insight into convergence behaviors by plotting local sections of the KL divergence near the optimum along different directions. This is achieved by plotting the functions $s \mapsto \mathcal{L}_q(\eta_q + s \cdot v_i)$ for different directions v_i on the unit circle for the η coordinates and by plotting the functions $s \mapsto \mathcal{L}_q(\theta_q + s \cdot w_i)$ for different directions w_i on the unit circle for the θ coordinates. This illustrates the local curvature of the function around the optimum. Since the natural gradient dynamics correspond to minimizing a convex quadratic function with identity Hessian as discussed above, we overlay the plots with a quadratic function $f(s) = \frac{1}{2}s^2$ for reference. The plots reveal that the functions $s \mapsto \mathcal{L}_q(\eta_q + s \cdot v_i)$ exhibit higher curvature than the quadratic reference function f, while the functions $s \mapsto \mathcal{L}_q(\theta_q + s \cdot w_i)$ appear flatter. This provides the core intuition behind the fast convergence of the gradient flow in η coordinates in comparison to gradient flow in the θ coordinates and shows why the natural gradient flow falls in between the two extremes.

To extend our numerical study to a higher-dimensional setting, we repeat the experiments from Figures 3 and 4 for the case $n = 10$. Figure 6 shows the KL divergence evaluated along gradient flow trajectories: the left panel shows the divergence on a linear scale, while the right panel uses a semi-logarithmic scale to highlight exponential convergence. As before, best-fit lines are overlaid to estimate the slopes, which correspond to the convergence rates. The $\eta-$gradient flow shows the fastest convergence (slope ≈ 16.239), the $\theta-$gradient flow is the slowest (slope ≈ 0.099), and the natural gradient flow lies in between (slope ≈ 1.974), closely matching the theoretical rate of 2. For greater confidence, we repeat this experiment over 100 randomly chosen initial distributions and one randomly chosen target distribution; the results are shown in Figure 7. The empirical convergence rates align well with the theoretical predictions from Theorem 3, confirming that $\eta-$gradient flows exceed rate 2, natural gradient flows converge at rate 2, and $\theta-$gradient flows fall below rate 2. Moreover, based on these empirical studies, we also observe that the difference in convergence rates is more pronounced for $n = 10$ as compared to $n = 2$.

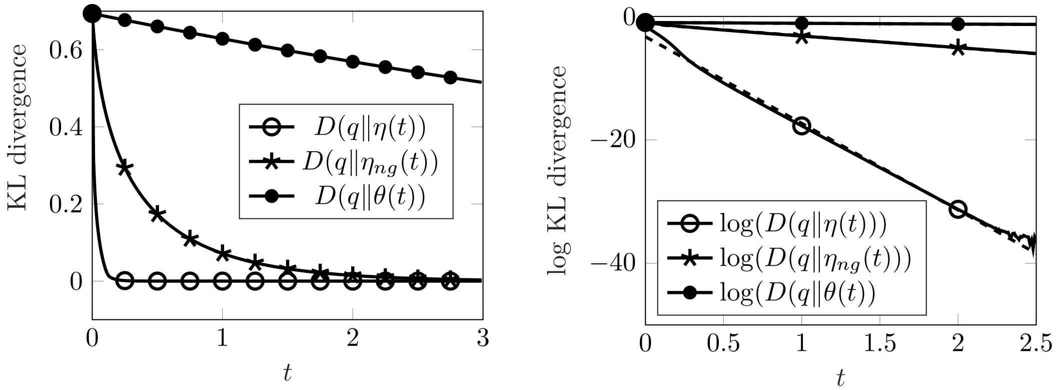


Figure 6: KL divergence evaluated along the solutions to $\eta-$gradient flow described by equation 10, $\theta-$gradient flow described by equation 11 and the natural gradient flow described by equation 12 for $n = 10$ plotted on a linear scale (left) and on a semi-log scale (right). The intervals between markers in the left figure are unrelated to the intervals between markers in the right figure. The dashed lines in the right figure show the best-fit linear function used to estimate the slope which gives the convergence rate.

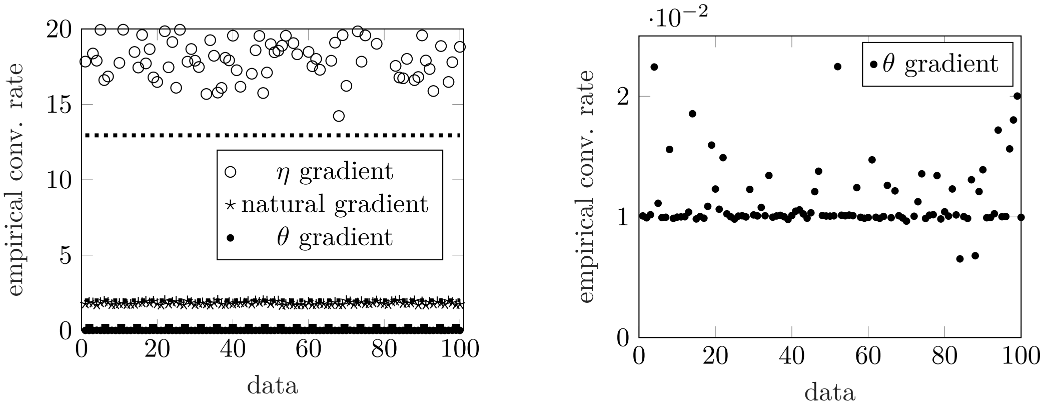


Figure 7: Left: Empirical convergence rates for 100 randomly chosen initial distributions and one randomly chosen target distribution with $n = 10$. The dotted line shows the theoretical lower bound (12.94) for the convergence rate of $\eta-$gradient flows and the dashed line shows the theoretical upper bound (0.31) for the convergence rate of $\theta-$gradient flows. Right: Same data, but with the y-axis restricted to highlight the variation in convergence rates of $\theta-$gradient flows.

Finally, the analysis presented in this section raises an interesting question: Since the dual pairing between the coordinates η and θ is preserved under appropriate affine transformations (as discussed in the following section and in Amari (2016), how do the convergence rates of the resulting dynamics change under such affine coordinate transformations? This question is addressed in the following subsection.

3.1 Convergence Rate Analysis Under Affine Coordinate Transformation

We first review the effect of an affine transformation of coordinates on the duality pairing between θ and η (or equivalently between ψ and φ). Since the convergence rate analysis from the previous section hinges on bounding the Hessian of the loss function, we investigate how the Hessian transforms under an affine change of coordinates. Consider new coordinates $\bar{\theta}$ that are related to the original θ-coordinates via an affine transformation: $\theta = A\bar{\theta} + b$, where $A \in \mathbb{R}^{n\times n}$ is an invertible matrix and $b \in \mathbb{R}^n$ is an arbitrary vector. Let $\bar{\psi}$ be defined as $\bar{\psi}(\bar{\theta}) = \psi(A\bar{\theta} + b)$. A simple application of the chain rule shows $\nabla^2\bar{\psi}(\bar{\theta}) = A^T\nabla^2\psi(\theta)A$. Therefore, $\nabla^2\bar{\psi}(\bar{\theta}) \succ 0$ if and only if $\nabla^2\psi(\theta) \succ 0$, since A is non-singular. This implies the strict convexity of $\bar{\psi}$, and it is possible to define its convex conjugate $\bar{\varphi}(\bar{\eta}) = \max_{\bar{\vartheta}} \left(\bar{\eta}^T\bar{\vartheta} - \bar{\psi}(\bar{\vartheta})\right)$. Optimality condition on the maximizer $\bar{\vartheta}_{\text{opt}} = \bar{\theta}$ yields the relationship $\bar{\eta} = \nabla\bar{\psi}(\bar{\theta}) = A^T\nabla\psi(\theta) = A^T\eta$. It is straight-forward to verify that with these newly defined convex conjugate pairs of functions $\bar{\psi}$ and $\bar{\varphi}$ the Bregman divergence still gives the original KL divergence, i.e.,

$$D_{\bar{\psi}}(\bar{\theta}_p \parallel \bar{\theta}_q) = D_{\bar{\varphi}}(\bar{\eta}_q \parallel \bar{\eta}_p) = D_{\psi}(\theta_p \parallel \theta_q) = D_{\varphi}(\eta_q \parallel \eta_p) = D(q\|p).$$

The Hessians of the loss function when evaluated at the optimum, transform as follows:

$$\nabla^2\mathcal{L}_q(\bar{\eta}_q) = \nabla^2\bar{\varphi}(\bar{\eta}_q) = A^T\nabla^2\varphi(\eta_q)A, \tag{22}$$

$$\nabla^2\mathcal{L}_q(\bar{\theta}_q) = \nabla^2\bar{\psi}(\bar{\theta}_q) = A^{-1}\nabla^2\psi(\theta_q)A^{-T}. \tag{23}$$

This calculation immediately gives us the following theorem.

Theorem 4 (Dual coordinates with identity Hessian)**.** *Let c be a positive constant, $q \in S_n$ be the target distribution and $p_0 \in S_n$ be the initial distribution. There exists a pair of convex conjugate functions $\bar{\psi}$ and $\bar{\varphi}$ inducing the pair of dual coordinates $\bar{\eta}$ and $\bar{\theta}$ for S_n with coordinate maps $\bar{\phi}_m$ and $\bar{\phi}_e$ such that*

$$\nabla^2\mathcal{L}_q(\bar{\eta}_q) = \left[\nabla^2\mathcal{L}_q(\bar{\theta}_q)\right]^{-1} = c \cdot I. \tag{24}$$

Consider the gradient flow dynamics:

$$\begin{aligned} \dot{\bar{\eta}}(t) &= -\nabla\mathcal{L}_q(\bar{\eta}(t)), \qquad & \bar{\eta}(0) &= \bar{\eta}_{p_0}, \\ \dot{\bar{\theta}}(t) &= -\nabla\mathcal{L}_q(\bar{\theta}(t)), \qquad & \bar{\theta}(0) &= \bar{\theta}_{p_0}. \end{aligned}$$

Then for any $\varepsilon > 0$, there exist positive constants c_1, c_2, c_3, c_4 and T such that for all $t \geq T$,

$$c_1 e^{-2(c+\varepsilon)t} \leq \mathcal{L}_q(\bar{\eta}(t)) \leq c_2 e^{-2(c-\varepsilon)t}, \tag{25}$$

$$c_3 e^{-2(\frac{1}{c}+\varepsilon)t} \leq \mathcal{L}_q(\bar{\theta}(t)) \leq c_4 e^{-2(\frac{1}{c}-\varepsilon)t}, \tag{26}$$

i.e., $\mathcal{L}_q(\bar{\eta}(t))$ and $\mathcal{L}_q(\bar{\theta}(t))$ converge exponentially with rate $2c$ and $\frac{2}{c}$, respectively.

Proof. See Appendix C.4 □

Note that by plugging $c = 1$ in Theorem 4, we see that there exists a dual pair of coordinates that achieves the convergence rate of the natural gradient dynamics. However, the affine transformation that leads to this convergence rate depends on the target distribution q and thus cannot be known in advance. Furthermore, Theorem 4 illustrates that the convergence rate of gradient flows in the transformed coordinates can be made arbitrarily small or arbitrarily large by scaling the coordinates. In contrast, the natural gradient flow is independent of the choice of coordinates, and therefore, has a coordinate independent convergence rate.

The superiority of the natural gradient method in terms of convergence rates is not immediately clear from the continuous-time analysis presented so far. In order to facilitate a meaningful discussion of convergence rates in continuous time, Muehlebach & Jordan (2020) propose a particular time-normalization approach that can be deployed in our setting. Alternatively, a more direct comparison between the Euclidean gradient and the natural gradient can be made by studying discrete-time gradient descent iterations. To elaborate this, we turn our attention to the discrete-time setting in the next section, where the advantages of the natural gradient method become evident.

4 Convergence Analysis in Discrete Time

For a given target distribution $q \in S_n$ and an initial distribution $p_0 \in S_n$, the discrete-time gradient dynamics are given by

$$\begin{aligned} \eta(k+1) &= \eta(k) - \alpha_\eta \cdot \nabla \mathcal{L}_q(\eta(k)) = \eta(k) + \alpha_\eta \nabla^2 \varphi(\eta(k))(\eta_q - \eta(k)), & \eta(0) &= \eta_{p_0}, \\ \theta(k+1) &= \theta(k) - \alpha_\theta \cdot \nabla \mathcal{L}_q(\theta(k))) = \theta(k) - \alpha_\theta \nabla \psi(\theta(k)) + \alpha_\theta \nabla \psi(\theta_q), & \theta(0) &= \theta_{p_0}, \\ \eta_{ng}(k+1) &= \eta_{ng}(k) - \alpha_{ng} \cdot \operatorname{grad} \mathcal{L}_q(\eta_{ng}(k)) = \eta_{ng}(k) - \alpha_{ng}\left(\eta_{ng}(k) - \eta_q\right), & \eta_{ng}(0) &= \eta_{p_0}, \end{aligned}$$

where α_η, α_θ and α_{ng} are the learning rates. To simplify the analysis, let us linearize these dynamics around the equilibrium points and examine the local convergence rates of the linearized dynamics. Owing to the already linear natural gradient dynamics, these do not need to be linearized. These linearized dynamics are given by

$$\eta(k+1) = \left(I - \alpha_\eta \nabla^2 \varphi(\eta_q)\right) \eta(k) + \alpha_\eta \nabla^2 \varphi(\eta_q) \eta_q, \qquad \eta(0) = \eta_{p_0}, \tag{27}$$

$$\theta(k+1) = \left(I - \alpha_\theta \nabla^2 \psi(\theta_q)\right) \theta(k) + \alpha_\theta \nabla^2 \psi(\theta_q) \theta_q, \qquad \theta(0) = \theta_{p_0}, \tag{28}$$

$$\eta_{ng}(k+1) = (1 - \alpha_{ng}) \eta_{ng}(k) + \alpha_{ng} \cdot \eta_q, \qquad \eta_{ng}(0) = \eta_{p_0}. \tag{29}$$

Unlike in the continuous-time setting, the choice of coordinates used to represent the natural gradient dynamics in discrete time influences the analysis of convergence rates, primarily due to the presence of the learning rate α in the update equations (see Martens (2020); Song et al. (2018)). However, it turns out that the discrete-time natural gradient dynamics in the θ coordinates, when linearized about the equilibrium θ_q, lead to update equations that are identical to equation 29. This is elaborated in Appendix B. Furthermore, also note that the update equation 29 is invariant to any affine transformation of the coordinates. Therefore, the update equation 29 represents local linearized dynamics for all dual pairs of coordinates.

The dynamics described by equation 27, equation 28 and equation 29 can be written in the general form

$$x(k+1) = (I - \alpha Q)\, x(k) + \alpha Q x^*$$

where Q is a symmetric positive definite matrix. Notice that these dynamics result from gradient descent iterations when optimizing the convex quadratic function $f(x) = \frac{1}{2}(x - x^*)^T Q (x - x^*)$. In this discrete-time setting, we say that a sequence $f(x(k))$ converges exponentially to $f(x_*)$ with rate $\rho \in [0, 1)$ if there exists a positive constant c and an integer k_0 such that $|f(x(k)) - f(x_*)| \leq c\rho^k$ holds for all $k \geq k_0$. A smaller value of ρ corresponds to faster convergence. The convergence rates of gradient descent algorithms for minimizing convex quadratic functions have been extensively studied. For example, the following result from Nesterov (2018) shows that the condition number of Q determines the convergence rates.

Theorem 5 (Nesterov (2018)). *Let $f(x) = \frac{1}{2}(x - x^*)^\top Q (x - x^*)$ with $Q \succ 0$, and let κ denote the condition number of Q. Consider the gradient descent iterations $x(k+1) = x(k) - \alpha \nabla f(x(k))$. Then:*

(i) $f(x(k))$ converges to zero at rate $\left(1 - \frac{1}{\kappa}\right)^2$ when $\alpha = \frac{1}{\lambda_{max}(Q)}$ (standard choice).

(ii) $f(x(k))$ converges to zero at rate $\left(1 - \frac{2}{\kappa+1}\right)^2$ when $\alpha = \frac{2}{\lambda_{min}(Q) + \lambda_{max}(Q)}$ (optimal choice).

Note that the gradient descent dynamics in η and θ coordinates correspond to setting $Q = \nabla^2 \varphi(\eta_q)$ and $Q = \nabla^2 \psi(\theta_q)$, respectively. By directly applying these results to the discrete-time gradient descent dynamics described by equation 27 and equation 28, we observe that poorer conditioning of Q leads to a larger convergence rate ρ, and thus slower convergence. The condition numbers associated with the gradient descent dynamics in the η and θ coordinates can be bounded away from 1 as stated in Lemma 6.

Lemma 6 (Bounds on the condition number of the Hessian)**.** *Let $q \in S_n$. Then,*

$$1 < \kappa_q \leq \operatorname{cond}(\nabla^2 \mathcal{L}_q(\eta_q)) = \operatorname{cond}(\nabla^2 \mathcal{L}_q(\theta_q)), \tag{30}$$

where $\kappa_q = \frac{\eta_{\min,2}}{\eta_{\min}}$ with $\eta_{\min}$ and $\eta_{\min,2}$ being the smallest and the second-smallest element of $\{[\eta_q]_1, \cdots, [\eta_q]_n\}$, respectively.

Proof. See Appendix C.5 ☐

The natural gradient dynamics correspond to setting $Q = I$ which yields optimal conditioning. Observe that the discrete-time natural gradient descent dynamics achieve a convergence rate of $|1-\alpha|$ for $\alpha \in (0, 2)$. Furthermore, it achieves an optimal convergence rate of 0, i.e., convergence in a single step for the optimal learning rate $\alpha = 1$. Note, however, that since this analysis pertains to the linearized system, the actual natural gradient descent does not converge in single step. Finally, with the goal of studying the properties of the stochastic gradient descent (SGD), we examine the robustness of these dynamics to imperfect gradient measurements. Although the noise models studied next do not exactly model the stochastic behavior of the SGD, they take us a step closer to it and provide valuable insight. Furthermore, practical implementations of the natural gradient method involve approximating the Fisher information matrix by an empirical version of it Martens (2020). This can also be captured to some degree by the noise models studied next. Specifically, we study two noise models motivated by (Polyak, 1987, Chapter 4): relative deterministic noise (multiplicative) and absolute random noise (additive). These and other similar noise models have been studied in the optimization literature and they evidently show that the condition number plays a central role in these analyses (see Guille-Escuret et al. (2021); Lessard et al. (2016); Van Scoy & Lessard (2024)).

4.1 Robustness Analysis with Relative Deterministic Noise

Let us first consider the relative deterministic noise (multiplicative) model which replaces the gradient vector v by $(I + \Delta(k))v$ where $\Delta(k) \in \mathbb{R}^{n\times n}$ captures the noise at time instant k. This leads to dynamics

$$\eta(k+1) = \eta(k) - \alpha_\eta \cdot (I + \Delta(k))\nabla^2\varphi(\eta_q)(\eta(k) - \eta_q), \qquad \eta(0) = \eta_{p_0}, \tag{31}$$

$$\theta(k+1) = \theta(k) - \alpha_\theta \cdot (I + \Delta(k))\nabla^2\psi(\theta_q)(\theta(k) - \theta_q), \qquad \theta(0) = \theta_{p_0}, \tag{32}$$

$$\eta_{ng}(k+1) = \eta_{ng}(k) - \alpha_{ng} \cdot (I + \Delta(k))(\eta_{ng}(k) - \eta_q), \qquad \eta_{ng}(0) = \eta_{p_0}, \tag{33}$$

where the learning rates α_η, α_θ and α_{ng} are chosen optimally assuming the noise-free conditions ($\Delta \equiv 0$).

Theorem 7 (Robust stability under relative deterministic noise)**.** *Consider a target distribution $q \in S_n$, an initial distribution $p \in S_n$ and the discrete-time dynamics described by equation 31, equation 32 and equation 33, respectively, where the learning rates α_η, α_θ and α_{ng} are chosen optimally for each case assuming the absence of noise ($\Delta \equiv 0$). Let $\kappa = \operatorname{cond}(\nabla^2\varphi(\eta_q)) = \operatorname{cond}(\nabla^2\psi(\theta_q))$. Then the following statements hold:*

- *(i) If the sequence of perturbations $\Delta(k)$ is such that for some $\varepsilon > 0$, $\|\Delta(k)\|_2 < 1-\varepsilon$ for all $k \geq 0$, then the natural gradient dynamics described by equation 33 are stable, i.e., $\lim_{k\to\infty}\|\eta_{ng}(k) - \eta_q\| = 0$.*
- *(ii) There exist time-invariant perturbations Δ_η and Δ_θ with $\|\Delta_\eta\|_2 = \|\Delta_\theta\|_2 = \frac{1}{\kappa}$ that destabilize the gradient descent dynamics described by equation 31 and equation 32, respectively.*

Proof. See Appendix C.6 ☐

The above result shows that the natural gradient dynamics exhibit a larger robustness margin in comparison to the robustness margin of the η and θ gradient dynamics which depend on the condition number κ. This shows that the superiority of the natural gradient dynamics can be again attributed to the optimal conditioning ($\kappa = 1$). Also note that the above noise model includes time-varying perturbations to the learning rate and shows that the natural gradient dynamics tolerate a much higher deviation from the optimal learning rate. Furthermore, note that statement (i) of the above theorem proves convergence for the situation where the inverse of the Fisher information matrix G^{-1} is replaced by $(I + \Delta(k))G^{-1}$ with $\|\Delta(k)\|_2 < 1 - \varepsilon$ for all $k \geq 0$. This result thus also makes progress towards the more practical implementations of the natural gradient involving an empirical version of the Fisher information matrix Martens (2020).

4.2 Robustness Analysis with Absolute Random Noise

Now let us now consider the absolute random noise (additive) model which perturbs the original dynamics by adding an independent and identically distributed noise signal $\delta(k)$ for $k \in \{0, 1, \cdots\}$. This leads to dynamics

$$\eta(k+1) = \eta(k) - \alpha_\eta \left(\nabla^2\varphi(\eta_q)(\eta(k) - \eta_q)\right) + \delta(k), \qquad \eta(0) = \eta_{p_0} \tag{34}$$

$$\theta(k+1) = \theta(k) - \alpha_\theta \left(\nabla^2\psi(\theta_q)(\theta(k) - \theta_q)\right) + \delta(k), \qquad \theta(0) = \theta_{p_0}, \tag{35}$$

$$\eta_{ng}(k+1) = \eta_{ng}(k) - \alpha_{ng} \left(\eta_{ng}(k) - \eta_q\right) + \delta(k), \qquad \eta_{ng}(0) = \eta_{p_0}, \tag{36}$$

where learning rates α_η, α_θ and α_{ng} are chosen optimally for each case assuming the absence of noise ($\delta(k) \equiv 0$). Furthermore, assume that $\delta(k)$ is an independent and identically distributed stochastic process satisfying $\mathbb{E}[\delta(k)] = 0$ and $\mathbb{E}[\delta(k)\delta(k)^T] = I$ for all $k \geq 0$.

Theorem 8 (Robustness against additive noise)**.** *Consider a target distribution $q \in S_n$, an initial distribution $p_0 \in S_n$ and the discrete-time dynamics described by equation 34, equation 35 and equation 36, respectively, where the learning rates α_η, α_θ and α_{ng} are chosen optimally for each case assuming the absence of noise ($\delta(k) \equiv 0$). Then*

(i) $\lim_{k\to\infty} \mathbb{E}[(\eta(k) - \eta_q)(\eta(k) - \eta_q)^T] = \Sigma_\eta \preceq \frac{(\kappa+1)^2}{4\kappa} I$,

(ii) $\lim_{k\to\infty} \mathbb{E}[(\theta(k) - \theta_q)(\theta(k) - \theta_q)^T] = \Sigma_\theta \preceq \frac{(\kappa+1)^2}{4\kappa} I$,

(iii) $\mathbb{E}[(\eta_{ng}(k) - \eta_q)(\eta_{ng}(k) - \eta_q)^T] = I$ *for all* $k \geq 0$.

The upperbound in (i) and (ii) is tight, i.e., Σ_η and Σ_θ have eigenvalues equal to $\frac{(\kappa+1)^2}{4\kappa}$. Furthermore, for $n = 2$, we get equality in (i) and (ii).

Proof. See Appendix C.7 ☐

The above result explores the effect of adding an independent and identically distributed (i.i.d.) noise signal at every iteration of the dynamics. It establishes that the largest eigenvalues of the steady-state error covariances are given by $\frac{(\kappa+1)^2}{4\kappa} I$ for the η and θ gradient dynamics whereas the error covariance with the natural gradient dynamics equals the noise covariance which corresponds to optimal conditioning ($\kappa = 1$).

5 Conclusions and Outlook

In this work, we revisited the convergence properties of natural gradient flows in comparison to their Euclidean counterparts, focusing on the minimization of the KL divergence over discrete probability distributions. Our analysis revealed a more nuanced picture than the commonly held belief in the universal superiority of natural gradient methods. While the natural gradient flow indeed outperforms the Euclidean gradient flow in the θ coordinates, consistent with traditional expectations, we showed that it converges more slowly than the η-gradient flow, despite following straight-line trajectories in these coordinates. This demonstrates that the commonly observed rapid convergence of natural gradient flow cannot be simplistically attributed to the straightness of its trajectories. Our discrete-time analysis of gradient descent dynamics further clarified that the fundamental reason behind the superiority of natural gradient methods lies in their optimal conditioning: natural gradient updates effectively minimize an optimally conditioned loss landscape, leading to consistently better performance compared to their Euclidean counterparts. Overall, our findings refine the understanding of natural gradient methods and highlight the subtle, yet important, nuances that govern their behavior.

The theoretical results presented in this paper primarily concern exact natural gradient, with robustness analyses taking initial steps toward incorporating noise in gradient measurements and system dynamics. A natural next step is to extend this analysis to practical settings where the Fisher information matrix must be approximated. For instance, Theorem 7 already accommodates structured perturbations of the form $(I + \Delta(k))G^{-1}$, where G is the Fisher information matrix, provided $\|\Delta(k)\|_2 < 1 - \varepsilon$ for all $k \geq 0$, ensuring the stability of the resulting updates. This suggests that if approximations such as K-FAC can be modeled as structured perturbations of G^{-1} satisfying the above condition, then it may be possible to establish convergence guarantees for the resulting approximate natural gradient algorithms. A systematic exploration of the trade-off between the quality of the approximation of the Fisher matrix and the resulting convergence guarantees would be an important direction for bridging theory and practice.

While we focused on the probability simplex equipped with dual coordinate systems, an important extension would be to general dually flat statistical manifolds (Amari & Nagaoka (2000); Ay et al. (2017)), where similar tools may be employed to study optimization dynamics in broader settings. Such manifolds are induced by general Bregman divergences going beyond the setting of a KL divergence. Notably, the identity $\nabla^2\varphi(\eta) = [\nabla^2\psi(\theta)]^{-1}$ holds for any pair of dual Bregman divergences induced by strictly convex functions φ and ψ, respectively Amari & Nagaoka (2000). So it is plausible that one can derive an analogue of Lemma 2 for this general setting provided $\nabla^2\varphi(\eta) \succ I$ holds at the optimum. One can then proceed to derive the "sandwich" ordering as in Theorem 3.

Additionally, our framework may be extended to richer families of probability distributions, such as general exponential families derived from Boltzmann machines without hidden units, and more intricate mixtures of exponential families associated with Boltzmann machines with hidden variables (Amari et al. (1992)). These models exhibit more complex geometries that may reveal deeper interactions between parametrizations and optimization dynamics.

Finally, although our discrete-time analysis highlights robustness advantages of natural gradient methods, it does not fully capture the stochasticity inherent in stochastic gradient descent (SGD). Developing a more precise theoretical model that explicitly incorporates the stochastic dynamics of SGD remains an important avenue for future work.

References

S. Amari, K. Kurata, and H. Nagaoka. Information geometry of Boltzmann machines. *IEEE Transactions on Neural Networks*, 3(2):260–271, March 1992. ISSN 1941-0093. doi: 10.1109/72.125867.

Shun-ichi Amari. Neural Learning in Structured Parameter Spaces - Natural Riemannian Gradient. In *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996.

Shun-ichi Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2):251–276, February 1998. ISSN 0899-7667. doi: 10.1162/089976698300017746.

Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer Japan, Tokyo, 2016. ISBN 978-4-431-55977-1 978-4-431-55978-8. doi: 10.1007/978-4-431-55978-8.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry.* American Mathematical Soc., 2000. ISBN 978-0-8218-4302-4.

Nihat Ay. On the locality of the natural gradient for learning in deep Bayesian networks. *Information Geometry*, 6(1):1–49, June 2023. ISSN 2511-249X. doi: 10.1007/s41884-020-00038-y.

Nihat Ay and Jesse van Oostrum. On the Fisher-Rao Gradient of the Evidence Lower Bound, July 2023.

Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. *Information Geometry*, volume 64 of *Ergebnisse Der Mathematik Und Ihrer Grenzgebiete 34*. Springer International Publishing, Cham, 2017. ISBN 978-3-319-56477-7 978-3-319-56478-4. doi: 10.1007/978-3-319-56478-4.

Rajendra Bhatia. *Perturbation Bounds for Matrix Eigenvalues.* SIAM, 2007.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds.* Cambridge University Press, 1 edition, March 2023. ISBN 978-1-00-916616-4 978-1-00-916617-1 978-1-00-916615-7. doi: 10.1017/9781009166164.

Charles Guille-Escuret, Manuela Girotti, Baptiste Goujaud, and Ioannis Mitliagkas. A Study of Condition Numbers for First-Order Optimization. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 1261–1269. PMLR, March 2021.

Sham M Kakade. A Natural Policy Gradient. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

Ryo Karakida and Kazuki Osawa. Understanding approximate Fisher information for fast convergence of natural gradient descent in wide neural networks*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124010, December 2021. ISSN 1742-5468. doi: 10.1088/1742-5468/ac3ae3.

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal Statistics of Fisher Information in Deep Neural Networks: Mean Field Approach. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pp. 1032–1041. PMLR, April 2019.

Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On linear and super-linear convergence of Natural Policy Gradient algorithm. *Systems & Control Letters*, 164: 105214, June 2022. ISSN 01676911. doi: 10.1016/j.sysconle.2022.105214.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and Design of Optimization Algorithms via Integral Quadratic Constraints. *SIAM Journal on Optimization*, 26(1):57–95, January 2016. ISSN 1052-6234, 1095-7189. doi: 10.1137/15M1009597.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning.* Adaptive Computation and Machine Learning. The MIT Press, Cambridge, Massachusetts, second edition edition, 2018. ISBN 978-0-262-03940-6.

Michael Muehlebach and Michael Jordan. Continuous-time lower bounds for gradient-based algorithms. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7088–7096. PMLR, 2020.

Johannes Müller and Guido Montúfar. Geometry and convergence of natural policy gradient methods. *Information Geometry*, 7(1):485–523, January 2024. ISSN 2511-249X. doi: 10.1007/s41884-023-00106-z.

Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications.* Springer International Publishing, Cham, 2018. ISBN 978-3-319-91577-7 978-3-319-91578-4. doi: 10.1007/978-3-319-91578-4.

Razvan Pascanu and Yoshua Bengio. Revisiting Natural Gradient for Deep Networks, February 2014.

Boris T. Polyak. Introduction to optimization. Translations Series in Mathematics and Engineering., 1987.

Yang Song, Jiaming Song, and Stefano Ermon. Accelerating Natural Gradient with Higher-Order Invariance. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4713–4722. PMLR, July 2018.

Eduardo D. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems.* Springer Science & Business Media, November 2013. ISBN 978-1-4612-0577-7.

Jesse van Oostrum and Nihat Ay. Parametrisation independence of the natural gradient in overparametrised systems. In *International Conference on Geometric Science of Information*, pp. 726–735. Springer, 2021.

Jesse van Oostrum, Johannes Müller, and Nihat Ay. Invariance properties of the natural gradient in overparametrised systems. *Information Geometry*, 6(1):51–67, June 2023. ISSN 2511-249X. doi: 10.1007/s41884-022-00067-9.

Bryan Van Scoy and Laurent Lessard. The Speed-Robustness Trade-Off for First-Order Methods with Additive Gradient Noise, June 2024.

Patrick M. Wensing and Jean-Jacques Slotine. Beyond convexity—Contraction and global convergence of gradient descent. *PLOS ONE*, 15(8):e0236661, August 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0236661.

Lin Xiao. On the Convergence Rates of Policy Gradient Methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022. ISSN 1533-7928.

Rui Yuan, Simon S Du, Robert M Gower, Alessandro Lazaric, and Lin Xiao. Linear convergence of natural policy gradient methods with log-linear policies. *arXiv preprint arXiv:2210.01400*, 2022.

Guodong Zhang, James Martens, and Roger B Grosse. Fast convergence of natural gradient descent for over-parameterized neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

A Gradient Flows for $\mathcal{L}^*_p$

For a given target distribution $p \in S_n$ and an initial distribution $q_0 \in S_n$, consider the gradient flow dynamics described by equation 37 and equation 38, and the natural gradient flow dynamics described by equation 39 given by

$$\dot{\eta}(t) = -\nabla \mathcal{L}^*_p(\eta(t)), \qquad \eta(0) = \eta_{q_0}, \tag{37}$$

$$\dot{\theta}(t) = -\nabla \mathcal{L}^*_p(\theta(t)), \qquad \theta(0) = \theta_{q_0}, \tag{38}$$

$$\dot{\theta}_{ng}(t) = -\text{grad } \mathcal{L}^*_p(\theta_{ng}(t)) = -\theta_{ng}(t) + \theta_p, \quad \theta_{ng}(0) = \theta_{q_0}. \tag{39}$$

The following is an analogue of Theorem 3 applied to the above dynamics.

Theorem 9 (Convergence analysis)**.** *Let $p \in S_n$ and $q_0 \in S_n$ be such that $D(q_0||p)$ is sufficiently small. Suppose η, θ and θ_{ng} be the solutions to dynamics described by equation 37, equation 38 and equation 39, respectively. Then*

(i) there exist positive constants $m^*_\theta \le L^*_\theta < 1$, c^*_θ and $\bar{c}^*_\theta$such that

$$c^*_\theta e^{-2t} \le c^*_\theta e^{-2L^*_\theta t} \le \mathcal{L}^*_p(\theta(t)) \le \bar{c}^*_\theta e^{-2m^*_\theta t} \qquad \forall t \ge 0, \tag{40}$$

i.e., $\mathcal{L}^*_p(\theta(t))$ converges exponentially with rate lower than 2.

(ii) there exist positive constants $1 < m^*_\eta \le L^*_\eta$ and c^*_η such that

$$c^*_\eta e^{-2L^*_\eta t} \le \mathcal{L}^*_p(\eta(t)) \le c^*_\eta e^{-2m^*_\eta t} \le c^*_\eta e^{-2t} \qquad \forall t \ge 0 \tag{41}$$

i.e., $\mathcal{L}_q(\eta(t))$ converges exponentially with rate higher than 2.

(iii) there exist positive constants c^*_1, c^*_2 and T such that

$$c^*_1 e^{-2t} \le \mathcal{L}^*_p(\theta_{ng}(t)) \le c^*_2 e^{-2t} \qquad \forall t \ge 0, \tag{42}$$

i.e., $\mathcal{L}^*_p(\theta_{ng}(t))$ converges exponentially with rate 2.

Proof. Consider the gradient flow dynamics described by equation 37. Analogous to the proof of Theorem 3 (see equation 57 and the discussion thereafter), it can be shown using the Pinsker inequality (Mohri et al., 2018, Proposition E.7) that η_p is the unique minimizer of $\mathcal{L}_p^*$ and $S_\eta^* := \{\eta \in \phi_m(S_n) | \mathcal{L}_p^*(\eta) \leq \mathcal{L}_p^*(\eta(0))\}$ is compact. Furthermore, compactness of S_η^* along with bound 15 implies that

$$I \prec m_\eta^* \cdot I \preceq \nabla^2 \mathcal{L}_p^*(\eta) \preceq L_\eta^* \cdot I \qquad \forall \eta \in S_\eta^*$$

where $m_\eta^* = \min_{\eta \in S_\eta^*} \lambda_{\min}\left(\nabla^2 \mathcal{L}_p^*(\eta)\right) > 1$ and $L_\eta^* = \max_{\eta \in S_\eta^*} \lambda_{\max}\left(\nabla^2 \mathcal{L}_p^*(\eta)\right)$. Applying Proposition 1, we get the desired inequality given in equation 41.

Now consider dynamics described by equation 38. Continuity of $\nabla^2 \mathcal{L}_p^*$ along with equation 18 from Lemma 2 implies that there exists an $\varepsilon > 0$ such that

$$0 \prec m_\theta^* \cdot I \preceq \nabla^2 \mathcal{L}_p^*(\theta) \preceq L_\theta^* \cdot I \prec I \qquad \forall \theta \in \mathcal{B}_\varepsilon(\theta_p) \tag{43}$$

for some positive constants $m_\theta^* \leq L_\theta^* < 1$. Using this local strong convexity condition, it can be shown that if $\mathcal{L}_p^*(\theta(0))$ is sufficiently small, $S_\theta^* := \{\theta \in \phi_e(S_n) | \mathcal{L}_p^*(\theta) \leq \mathcal{L}_p^*(\theta(0))\}$ is contained in $\mathcal{B}_\varepsilon(\theta_p)$. Applying Proposition 1, we get the desired inequality given in equation 40.

Finally consider the gradient flow dynamics described by equation 39 which can be solved exactly to obtain

$$\theta_{ng}(t) = \theta_q + e^{-t}\left(\theta_0 - \theta_q\right). \tag{44}$$

Since $\lim_{t\to\infty} \theta_{ng}(t) = \theta_q$, we can use equation 43 along with (Nesterov, 2018, Theorem 2.1.5 and Theorem 2.1.8) to show that if $\mathcal{L}_p^*(\theta(0))$ is sufficiently small, then

$$\frac{m_\theta^*}{2} ||\theta_{ng}(t) - \eta_q||^2 \leq \mathcal{L}_p^*(\theta_{ng}(t)) \leq \frac{L_\theta^*}{2} ||\theta_{ng}(t) - \theta_q||^2 \tag{45}$$

for $t \geq 0$. Plugging in the exact solution from equation 44, we get the desired inequality given in equation 42. □

B Linearized Discrete-time Natural Gradient Dynamics in θ Coordinates

In this appendix, we show that the discrete-time natural gradient dynamics in the θ coordinates given by

$$\theta_{ng}(k+1) = \theta_{ng}(k) - \alpha_{ng} \cdot \text{grad } \mathcal{L}_q(\theta_{ng}(k)), \qquad \theta_{ng}(0) = \theta_{p_0},$$

when linearized about the equilibrium θ_q, lead to update equations that are identical to the ones in the η coordinates. From the defining property given in equation 9 of the natural gradient, and equation 6, we get that

$$\text{grad } \mathcal{L}_q(\theta) = [\nabla^2\psi(\theta)]^{-1} \nabla \mathcal{L}_q(\theta) = [\nabla^2\psi(\theta)]^{-1}\left(\nabla\psi(\theta) - \nabla\psi(\theta_q)\right) \approx (\theta - \theta_q),$$

where $\approx$ denotes a first-order approximation obtained by linearizing around θ_q. The linearized dynamics in the θ coordinates are thus described by

$$\theta_{ng}(k+1) = \theta_{ng}(k) - \alpha_{ng} \cdot (\theta_{ng}(k) - \theta_q), \qquad \theta_{ng}(0) = \theta_{p_0},$$

which has the same form as in in equation 29.

C Proofs

C.1 Proof of Proposition 1

Proof. [Proposition 1] Let x be a solution of equation 13 and define $E(t) := f(x(t)) - f(x_*)$. Note that

$$\dot{E}(t) = \langle \nabla f(x(t)), \dot{x}(t) \rangle = -\langle \nabla f(x(t)), \nabla f(x(t)) \rangle = -||\nabla f(x(t))||^2 \leq 0. \tag{46}$$

Therefore, for all $t \geq t_0$, $f(x(t)) - f(x_*) = E(t) \leq E(t_0) = f(x(t_0)) - f(x_*)$ which implies statement (i). Furthermore, we get from (Nesterov, 2018, Section 2.1) that $m \cdot I \preceq \nabla^2 f(x) \preceq L \cdot I$ for all $x \in S$ implies

$$2m(f(x) - f(x_*)) \leq ||\nabla f(x)||^2 \leq 2L(f(x) - f(x_*)) \quad \forall x \in S. \tag{47}$$

Since $x(t) \in S$ for all $t \geq t_0$, inequalities given in equation 47 and equation 46 imply that for all $t \geq t_0$,

$$-2L \cdot E(t) \leq \dot{E}(t) \leq -2m \cdot E(t).$$

Integrating from t_0 to t, we get that

$$\begin{aligned} E(t) &= E(t_0) + \int_{t_0}^{t} \dot{E}(s)ds \leq E(t_0) + \int_{t_0}^{t} (-2m) \cdot E(s)ds, \\ -E(t) &= -E(t_0) + \int_{t_0}^{t} -\dot{E}(s)ds \leq -E(t_0) + \int_{t_0}^{t} (-2L) \cdot (-E(s))ds \end{aligned}$$

Finally applying the Bellman-Gronwall Lemma (Sontag, 2013, Lemma C.3.1) to the above two inequalities gives us

$$\begin{aligned} E(t) &\leq E(t_0)e^{-2m(t-t_0)} \\ -E(t) &\leq -E(t_0)e^{-2L(t-t_0)} \end{aligned}$$

which dirctly gives us the desired inequality 14. □

C.2 Proof of Lemma 2

Proof. [Lemma 2] The Hessians of $\mathcal{L}_q$ can be evaluated using equation 5 and equation 6 as

$$\begin{aligned} \nabla^2 \mathcal{L}_q(\eta) &= \nabla^2 \varphi(\eta) - D^3\varphi(\eta)[\eta_q - \eta], \\ \nabla^2 \mathcal{L}_q(\theta) &= \nabla^2 \psi(\theta), \end{aligned} \tag{48}$$

where $D^3\varphi(\eta) : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ is the third order derivative of φ whose action on a vector $v \in \mathbb{R}^n$ is given by $\left(D^3\varphi(\eta)[v]\right)_{ij} = \sum_{k=1}^{n} \frac{\partial^3 \varphi(\eta)}{\partial \eta_i \partial \eta_j \partial \eta_k} v_k$. In particular, note that the inverse relationship given in equation 1 give us

$$\nabla^2 \mathcal{L}_q(\eta_q) = \nabla^2 \varphi(\eta_q) = [\nabla^2 \psi(\theta_q)]^{-1} = [\nabla^2 \mathcal{L}_q(\theta_q)]^{-1}. \tag{49}$$

Similarly, we can compute the Hessians of $\mathcal{L}_p^*$ using equation 7 and equation 8 as

$$\begin{aligned} \nabla^2 \mathcal{L}_p^*(\eta) &= \nabla^2 \varphi(\eta), \\ \nabla^2 \mathcal{L}_p^*(\theta) &= \nabla^2 \psi(\theta) - D^3\psi(\theta)[\theta_p - \theta] \end{aligned} \tag{50}$$

and use equation 1 to obtain the inverse relationship

$$\nabla^2 \mathcal{L}_p^*(\eta_p) = \nabla^2 \varphi(\eta_p) = [\nabla^2 \psi(\theta_p)]^{-1} = [\nabla^2 \mathcal{L}_p^*(\theta_p)]^{-1}. \tag{51}$$

Recall that

$$\varphi(\eta) = \left(\sum_{i=1}^{n} \eta_i \log \eta_i\right) + \left(1 - \sum_{j=1}^{n} \eta_j\right) \log \left(1 - \sum_{k=1}^{n} \eta_k\right)$$

and it's Hessian $\nabla^2 \varphi$ can be explicitly computed to be

$$\nabla^2 \varphi(\eta) = \begin{bmatrix} \frac{1}{\eta_1} & & \\ & \ddots & \\ & & \frac{1}{\eta_n} \end{bmatrix} + \left(\frac{1}{1 - \sum_{i=1}^{n} \eta_i}\right) \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix}. \tag{52}$$

Since the second matrix on the right hand side is positive semi-definite, we get

$$I \prec \frac{1}{\max_i \eta_i} I \preceq \nabla^2 \varphi(\eta). \tag{53}$$

This together with equation 50 gives us

$$I \prec \nabla^2 \mathcal{L}_p^*(\eta) \qquad \forall \eta \in \phi_m(S_n) \tag{54}$$

and together with equation 48 and the inverse relationship given in equation 1 gives us

$$0 \prec \nabla^2 \mathcal{L}_q(\theta) \prec I \qquad \forall \theta \in \phi_e(S_n) \tag{55}$$

proving equation 15. Furthermore, it can be shown by direct computation that

$$\nabla^2 \mathcal{L}_q(\eta) = \begin{bmatrix} \frac{[\eta_q]_1}{\eta_1^2} & & \\ & \ddots & \\ & & \frac{[\eta_q]_n}{\eta_n^2} \end{bmatrix} + \left(\frac{1 - \sum_{i=1}^n [\eta_q]_i}{\left(1 - \sum_{i=1}^n \eta_i\right)^2} \right) \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix} \succ 0 \quad \forall \eta \in \phi_m(S_n). \tag{56}$$

proving equation 16. Finally, evaluating the global bounds from equation 15 at the optimum points and using the inverse relationships given in equation 49 and equation 51 yields the desired local inequalities described by equation 17 and equation 18. □

C.3 Proof of Theorem 3

Proof. [Theorem 3] First consider the gradient flow dynamics described by equation 11 and note that $\mathcal{L}_q(\theta_q) = 0$. Using the Pinsker's inequality (Mohri et al., 2018, Proposition E.7) along with the fact that $\phi_m \circ \phi_e^{-1}$ is bijective, we have that for any $\theta_p \neq \theta_q$, $\mathcal{L}_q(\theta_p) > 0$. Thus θ_q is the unique minimizer of $\mathcal{L}_q$ giving us condition *a*) of Proposition 1. It can be shown that $\mathcal{L}_q$ has bounded sublevel sets in the θ coordinates. To see this, observe that $\|\theta\| \to \infty$ implies that for some i, $|\theta_i| \to \infty$, i.e., either $\theta_i \to -\infty$ or $\theta_i \to \infty$. This implies that

$$p_i = \frac{e^{\theta_i}}{1 + \sum_{j=1}^n e^{\theta_j}} \to 0 \qquad \text{or} \qquad p_{n+1} = \frac{1}{1 + \sum_{j=1}^n e^{\theta_j}} \to 0.$$

Since the $\mathcal{L}_q$ blows up to infinity on the boundary of the simplex, we get that $\mathcal{L}_q$ is coercive, i.e., $\mathcal{L}_q(\theta) = D(q\|p) \to \infty$ as $\|\theta\| \to \infty$. Observe that if some sublevel set of $\mathcal{L}_q$ is unbounded, there exists a sequence of points $\theta_{(1)}, \theta_{(2)}, \cdots$ such that $\|\theta_{(i)}\| \to \infty$ but $\mathcal{L}_q(\theta_{(i)})$ stays bounded. This is a contradiction. Therefore, we have that $\mathcal{L}_q$ has bounded sublevel sets in the θ coordinates. Finally, continuity of $\mathcal{L}_q$ implies that the sublevel sets are closed which implies that S_θ is compact. Using compactness of $S_\theta = \{\theta \in \phi_e(S_n) | \mathcal{L}_q(\theta) \leq \mathcal{L}_q(\theta(0))\}$ and the global bound given in equation 15 from Lemma 2, we get that

$$0 \prec m_\theta \cdot I \preceq \nabla^2 \mathcal{L}_q(\theta) \preceq L_\theta \cdot I \prec I \qquad \forall \theta \in S_\theta$$

where $m_\theta = \min_{\theta \in S_\theta} \lambda_{\min}\left(\nabla^2 \mathcal{L}_q(\theta)\right) > 0$ and $L_\theta = \max_{\theta \in S_\theta} \lambda_{\max}\left(\nabla^2 \mathcal{L}_q(\theta)\right) < 1$. Applying Proposition 1 gives us the desired conclusion in the form of equation 20.

Analogous to the previous case, now consider the gradient flow dynamics described by equation 10. We can use Pinsker's inequality (Mohri et al., 2018, Proposition E.7) to obtain

$$\begin{aligned} \mathcal{L}_q(\eta) &\geq \frac{1}{2}\left(\left(\sum_{i=1}^n |\eta_i - [\eta_q]_i| \right) + \left| \left(1 - \sum_{j=1}^n \eta_j\right) - \left(1 - \sum_{k=1}^n [\eta_q]_k\right) \right| \right)^2 \\ &\geq \frac{1}{2}\left(\sum_{i=1}^n |\eta_i - [\eta_q]_i| \right)^2 = \frac{1}{2}\|\eta_q - \eta\|_1^2 \geq \frac{1}{2}\|\eta_q - \eta\|^2. \end{aligned} \tag{57}$$

This shows that η_q is the unique minimizer of $\mathcal{L}_q$ and together with the continuity of $\mathcal{L}$ implies that

$$S_\eta := \{\eta \in \phi_m(S_n) | \mathcal{L}_q(\eta) \leq \mathcal{L}_q(\eta(0))\}$$

is compact. Using the global bound given in equation 16, we get that

$$0 \prec \bar{m}_\eta \cdot I \preceq \nabla^2 \mathcal{L}_q(\eta) \preceq \bar{L}_\eta \cdot I \qquad \forall \eta \in S_\eta \tag{58}$$

where $\bar{m}_\eta = \min_{\eta \in S_\eta} \lambda_{\min}\left(\nabla^2 \mathcal{L}_q(\eta)\right) > 0$ and $\bar{L}_\eta = \max_{\eta \in S_\eta} \lambda_{\max}\left(\nabla^2 \mathcal{L}_q(\eta)\right)$. Applying Proposition 1, we see that there exists a positive constant c such that

$$ce^{-2\bar{L}_\eta t} \leq \mathcal{L}_q(\eta(t)) \leq ce^{-2\bar{m}_\eta t} \qquad \forall t \geq 0. \tag{59}$$

This implies that $\mathcal{L}_q(\eta(t))$ converges to 0. Thus, for any $\varepsilon > 0$, there exists a $T > 0$ such that $\mathcal{L}_q(\eta(t)) \leq 2\varepsilon$ for all $t \geq T$, which, using the lower bound obtained in equation 57 implies that $\|\eta_q - \eta(t)\| \leq \varepsilon$ for all $t \geq T$. This means that for any $\varepsilon > 0$, there exists a $T > 0$ such that the set $S_T := \{\eta \in \phi_m(S_n) | \mathcal{L}_q(\eta) \leq \mathcal{L}_q(\eta(T))\} \subset \mathcal{B}_\varepsilon(\eta_q)$. Furthermore, continuity of $\nabla^2 \mathcal{L}_q$ along with equation 17 from Lemma 2 implies that there exists an $\varepsilon > 0$ such that

$$I \prec m_\eta \cdot I \preceq \nabla^2 \mathcal{L}_q(\eta) \preceq L_\eta \cdot I \qquad \forall \eta \in \mathcal{B}_\varepsilon(\eta_q).$$

for some positive constants $1 < m_\eta \leq L_\eta$. Putting everything together, we get that there exists a $T \geq 0$ such that $I \prec m_\eta \cdot I \preceq \nabla^2 \mathcal{L}_q(\eta) \preceq L_\eta \cdot I$ holds for all $\eta \in S_T$. Thus, applying Proposition 1 with $t_0 = T$, we get that there exists a positive constant $\bar{c}$ such that

$$\left(\bar{c}e^{2L_\eta T}\right) e^{-2L_\eta t} = \bar{c}e^{-2L_\eta(t-T)} \leq \mathcal{L}_q(\eta(t)) \leq \bar{c}e^{-2m_\eta(t-T)} = \left(\bar{c}e^{2m_\eta T}\right) \bar{c}e^{-2m_\eta t}$$

holds for all $t \geq T$ which is the desired form in equation 19 with $c_\eta = \bar{c}e^{2L_\eta T}$ and $\bar{c}_\eta = \bar{c}e^{2m_\eta T}$. Finally, note that if $\mathcal{L}_q(\eta(0))$ is sufficiently small, then $S_0 \subset \mathcal{B}_\varepsilon(\eta_q)$ implying that the above bound holds with $T = 0$. This completes the proof for statement (i).

Finally consider the gradient flow dynamics described by equation 12 which can be solved exactly to obtain

$$\eta_{ng}(t) = \eta_q + e^{-t}\left(\eta_{p_0} - \eta_q\right). \tag{60}$$

Using (Nesterov, 2018, Theorem 2.1.5 and Theorem 2.1.8) along with equation 58, we get that

$$\frac{\bar{m}_\eta}{2} ||\eta_{ng}(t) - \eta_q||^2 \leq \mathcal{L}_q(\eta_{ng}(t)) \leq \frac{\bar{L}_\eta}{2} ||\eta_{ng}(t) - \eta_q||^2. \tag{61}$$

Plugging in the exact solution from equation 60, we get the desired inequality given in equation 21. ☐

C.4 Proof of Theorem 4

Proof. [Theorem 4] Since $\nabla^2 \psi(\theta_q)$ is symmetric positive definite, we can consider its symmetric matrix square root D_q, such that $\nabla^2 \psi(\theta_q) = D_q \cdot D_q$. Setting $A = \sqrt{c} \cdot D_q$ in equation 22 and equation 23 and using the inverse relationship $\nabla^2 \varphi(\eta_q) = [\nabla^2 \psi(\theta_q)]^{-1}$ we obtain equation 24. Using the continuity of Hessians, we get that for any $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$(c - \varepsilon) \cdot I \preceq \nabla^2 \mathcal{L}_q(\bar{\eta}) \preceq (c + \varepsilon) \cdot I \qquad \forall \bar{\eta} \in \mathcal{B}_\delta(\bar{\eta}_q),$$
$$\left(\frac{1}{c} - \varepsilon\right) \cdot I \preceq \nabla^2 \mathcal{L}_q(\bar{\theta}) \preceq \left(\frac{1}{c} + \varepsilon\right) \cdot I \qquad \forall \bar{\theta} \in \mathcal{B}_\delta(\bar{\theta}_q).$$

Analogous to the proof of Theorem 3, we can apply Proposition 1 to obtain the desired inequalities given in equation 25 and equation 26. ☐

C.5 Proof of Lemma 6

Proof. [Lemma 6] Recall that

$$\nabla^2 \mathcal{L}_q(\eta_q) = \nabla^2 \varphi(\eta_q) = [\nabla^2 \psi(\theta_q)]^{-1} = [\nabla^2 \mathcal{L}_q(\theta_q)]^{-1}.$$

This establishes the equality $\operatorname{cond}(\nabla^2 \mathcal{L}_q(\eta_q)) = \operatorname{cond}(\nabla^2 \mathcal{L}_q(\theta_q))$. Recall that

$$\nabla^2 \varphi(\eta) = \begin{bmatrix} \frac{1}{\eta_1} & & \\ & \ddots & \\ & & \frac{1}{\eta_n} \end{bmatrix} + \left(\frac{1}{1 - \sum_{i=1}^{n} \eta_i} \right) \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix}. \tag{62}$$

Applying Weyl's inequalities Bhatia (2007) to the above rank one perturbation matrix, we get

$$\lambda_{\max}\left(\nabla^2 \varphi(\eta)\right) \geq \frac{1}{\eta_{\min}} \qquad \text{and} \qquad \lambda_{\min}\left(\nabla^2 \varphi(\eta)\right) \leq \frac{1}{\eta_{\min,2}}.$$

This directly implies the desired inequality given in equation 30. □

C.6 Proof of Theorem 7

Proof. [Theorem 7] Consider perturbed dynamics of the form

$$x(k+1) = x(k) - \alpha(I + \Delta(k))Q(x(k) - x^*), \tag{63}$$

where Q is a symmetric positive definite matrix. This encompasses the dynamics described by equation 31, equation 32 and equation 33 by setting Q equal to $\nabla^2\varphi(\eta_q)$, $\nabla^2\psi(\theta_q)$ and I, respectively and setting x^* equal to η_q, θ_q and η_q, respectively.

Let $\kappa = \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)}$. We will now prove that the dynamics described by equation 63 are stable, i.e., $\lim_{k\to\infty} \|x(k) - x^*\| = 0$, if for some $\varepsilon > 0$, $\|\Delta(k)\|_2 < \frac{1}{\kappa} - \varepsilon$ for all $k \geq 0$. This would directly imply statement (i) by plugging in $Q = I$.

By defining the error variable $e(k) := x(k) - x^*$, we get that

$$e(k+1) = (I - \alpha Q)e(k) - \alpha \Delta(k) Q e(k). \tag{64}$$

Note that α is chosen optimally assuming no noise ($\Delta(k) \equiv 0$), i.e., $\alpha = \frac{2}{\lambda_{\max}(Q) + \lambda_{\min}(Q)}$ (see Theorem 5). With this choice of α, we get that $\|I - \alpha Q\|_2 = \frac{\kappa-1}{\kappa+1}$ and $\|\alpha Q\|_2 = \frac{2\kappa}{\kappa+1}$, where the $\|\cdot\|_2$ is the induced 2−norm which coincides in our case to the largest eigenvalue magnitude owing to symmetry. This can be seen most directly by an eigenvalue decomposition of the involved matrices. Therefore, using the triangle inequality and the submultiplicative rule of induced matrix norms, we get that,

$$\|e(k+1)\| \leq \left(\frac{\kappa - 1}{\kappa + 1} + \|\Delta(k)\|_2 \frac{2\kappa}{\kappa+1} \right) \|e(k)\| = \frac{(1 + 2\|\Delta(k)\|_2)\,\kappa - 1}{\kappa+1} \|e(k)\|.$$

Note that if $\|\Delta(k)\|_2 < \frac{1}{\kappa} - \varepsilon$ for all $k \geq 0$, then $\frac{(1+2\|\Delta(k)\|_2)\kappa - 1}{\kappa+1} < \overbrace{1 - \varepsilon \frac{\kappa}{\kappa+1}}^{\rho} < 1$ which gives us the chain of inequalities

$$\|e(k+1)\| \leq \rho \|e(k)\| \leq \rho^2 \|e(k-1)\| \leq \cdots \leq \rho^{k+1} e(0).$$

Since $\rho < 1$, we get that $\lim_{k\to\infty} \|e(k)\| = 0$. Since $Q = I$ implies $\kappa = 1$, this implies statement (i).

Since the above argument only derives a sufficient condition for stability, we still need to prove statement (ii) separately. We now construct a time-invariant perturbation Δ such that $\|\Delta\|_2 = \frac{1}{\kappa}$ and the dynamics described by equation 63 are unstable, i.e., $x(k)$ does not converge to x^*. To this end, let $Q = U\Lambda U^T$ be the eigenvalue decomposition of the symmetric matrix Q where Λ is the diagonal matrix containing eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ in ascending order along the diagonal. Construct Δ as

$$\Delta = U \begin{bmatrix} 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & \frac{1}{\kappa} \end{bmatrix} U^T.$$

Plugging this in equation 64 and using $\alpha = \frac{2}{\lambda_1+\lambda_n}$, we get that

$$e(k+1) = Me(k), \tag{65}$$

where $M = (I - \frac{2}{\lambda_1+\lambda_n}(I+\Delta)Q)$ contains an eigenvalue at -1. Since stability of dynamics described by equation 65 requires the spectral radius of M to be less than 1, and since M contains an eigenvalue at -1, the dynamics are unstable. This completes the proof for statement (*ii*) by applying the constructed Δ to the choices $Q = \nabla^2\varphi(\eta_q)$ and $Q = \nabla^2\psi(\theta_q)$, respectively. □

C.7 Proof of Theorem 8

Proof. [Theorem 8] Following the same strategy as in the proof of Theorem 7, consider perturbed dynamics of the form

$$x(k+1) = x(k) - \alpha Q(x(k) - x^*) + \delta(k), \tag{66}$$

where Q is a symmetric positive definite matrix. This encompasses the dynamics described by equation 34, equation 35 and equation 36 by setting Q equal to $\nabla^2\varphi(\eta_q)$, $\nabla^2\psi(\theta_q)$ and I, respectively and setting x^* equal to η_q, θ_q and η_q, respectively. By defining the error variable $e(k) := x(k) - x^*$, we get that

$$e(k+1) = (I - \alpha Q)e(k) + \delta(k). \tag{67}$$

Note that α is chosen optimally assuming no noise ($\delta(k) \equiv 0$), i.e., $\alpha = \frac{2}{\lambda_{\max}(Q)+\lambda_{\min}(Q)}$ (see Theorem 5). With this choice of α, we get that $\|I - \alpha Q\|_2 = \frac{\kappa-1}{\kappa+1}$, where the $\|\cdot\|_2$ is the induced $2-$norm which coincides in our case to the largest magnitude eigenvalue owing to symmetry. Define $P(k) := \mathbb{E}[e(k)e(k)^T]$ where the expectation is taken over the different realizations of the noise process $\delta(k)$. Using the fact that $\delta(k)$ and $e(k)$ are independent random variables along with $\mathbb{E}[\delta(k)] \equiv 0$ and $\mathbb{E}[\delta(k)\delta(k)^T] \equiv I$, we get that

$$P(k+1) = (I - \alpha Q)P(k)(I - \alpha Q) + I. \tag{68}$$

Since $(I - \alpha Q)$ has all eigenvalues in $(-1, 1)$, it can be shown that $\lim_{k\to\infty} P(k) = P$ where P solves the

$$P = (I - \alpha Q)P(I - \alpha Q) + I. \tag{69}$$

To solve this equation, let $Q = U\Lambda U^T$ be the eigenvalue decomposition of the symmetric matrix Q where Λ is the diagonal matrix containing eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ in ascending order along the diagonal. Note that equation 69 can be solved to obtain

$$P = U \begin{bmatrix} \frac{1}{1-\mu_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{1-\mu_n^2} \end{bmatrix} U^T,$$

where μ_i are the eigenvalues of $(I - \alpha Q)$. Therefore, the largest eigenvalue of P is $\frac{1}{1-\left(\frac{\kappa-1}{\kappa+1}\right)^2} = \frac{(\kappa+1)^2}{2\kappa}$. This proves statements (*i*) and (*ii*) by plugging in Q equal to $\nabla^2\varphi(\eta_q)$ and $\nabla^2\psi(\theta_q)$, respectively. Finally, plugging $Q = I$ and $\alpha = \frac{2}{\lambda_{\min}(Q)+\lambda_{\max}(Q)} = 1$ in equation 68 directly gives statement (*iii*). □