Conjugate Natural Selection

A PREPRINT

Reilly Raab
Computer Science and Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
reilly@ucsc.edu

Luca de Alfaro
Computer Science and Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
luca@ucsc.edu

Yang Liu
Computer Science and Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
yangliu@ucsc.edu

June 14, 2023

ABSTRACT

We prove that Fisher-Rao natural gradient descent (FR-NGD) optimally approximates the continuous time replicator equation (an essential model of evolutionary dynamics), and term this correspondence "conjugate natural selection". This correspondence promises alternative approaches for evolutionary computation over continuous or high-dimensional hypothesis spaces. As a special case, FR-NGD also provides the optimal approximation of continuous Bayesian inference when hypotheses compete on the basis of predicting actual observations. In this case, the method avoids the need to compute prior probabilities. We demonstrate our findings on a non-convex optimization problem and a system identification task for a stochastic process with time-varying parameters.

Keywords evolution · natural selection · Bayesian inference · Fisher information · non-convex optimization · stochastic process · parameter estimation

1 Introduction

Evolution describes how distributions change. Specifically, evolution provides a model for how a population's distribution of traits or strategies (hereafter *hypotheses*) changes over time as an environment modulates reproduction rates (i.e., of individuals or of hypotheses; Lloyd, 2020): Hypotheses that have higher *fitness* are "selected" by the environment and, in expectation, become more popular with time. **The replicator equation is a formal, analytic model of evolution** and is indispensable to biology (Sinervo and Calsbeek, 2006; Queller, 2017).

In the replicator equation, the absolute fitness (in this paper, the negative *loss* $\mathscr{L}$) of hypotheses $h \in \mathcal{H}$ is identified with its rate of *replication*: exponential growth (or decline) in a population where different hypotheses compete for relative frequency $\rho(h) \in [0, 1]$. For probability distributions over hypothesis space $\mathcal{H}$, this equation induces *replicator dynamics*, selecting hypotheses with lower than average loss. In continuous time, the replicator equation i

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(h) = \rho(h)\Big[\overline{\mathscr{L}}_\rho - \mathscr{L}(h)\Big], \quad \text{where} \quad \overline{\mathscr{L}}_\rho \coloneqq \sum_h \rho(h)\mathscr{L}(h), \quad \sum_h \rho(h) = 1.$$

The replicator equation has been applied to game theory (Hofbauer et al., 1998; Sandholm, 2010; Cressman and Tao, 2014; Friedman and Sinervo, 2016), economics (Friedman, 1991), and machine learning (Hennes et al., 2019). For many real-world applications, however, **an exceedingly large or continuous hypothesis space $\mathcal{H}$ presents challenges** (Bloembergen et al., 2015, Sec 4.2), and standard techniques in evolutionary computation resort to modeling finite populations of individuals directly (Bäck et al., 2018). Our results show promise in alleviating these challenges. Specifically, we show that Fisher-Rao Natural Gradient Descent (FR-NGD) optimally approximates replicator dynamics for a tractable, lower dimensional representation of an evolving distribution over $\mathcal{H}$.

Why should we want analytic model of evolutionary dynamics with large hypothesis spaces? Suppose we wish to find the minimum value of a (possibly non-convex) function $\mathscr{L}\colon \mathbf{R}^d \to \mathbf{R}$, where we must make (possibly noisy, expensive) queries for $\mathscr{L}(h)$ at any input $h \in \mathbf{R}^d$. This problem formulation is foundational to machine learning. First, we relax our attention to *individual* hypotheses h. Evolution describes how *distributions* change, and the replicator equation gives us a way to model how any probability distribution ρ over hypothesis space $\mathcal{H} = \mathbf{R}^d$ should evolve when hypotheses are selected according to the loss function $\mathscr{L}$: Under replicator dynamics, hypotheses with minimal $\mathscr{L}$ will eventually out-compete all others, and any initial ρ that assigns non-zero probability (density) to all of $\mathcal{H}$ will converge to a distribution over globally optimal hypotheses. We call ρ a *metahypothesis*.

For *small* (finite) hypothesis spaces $\mathcal{H}$, the replicator dynamics may be simulated directly, corresponding to the method of multiplicative weight updates (Littlestone and Warmuth, 1994; Friedman and Sinervo, 2016). For *large* $\mathcal{H}$, however, our need to independently track $\rho(h)$ and sample $\mathscr{L}(h)$ for each value of $h \in \mathcal{H}$ causes issues: Note that the replicator equation describes dynamics in $\mathcal{P}$, where $\mathcal{P}$ denotes the space of probability distributions over $\mathcal{H}$. In general, the space requirements for storing and manipulating arbitrary probability distributions $\rho \in \mathcal{P}$ grow proportional to the size of $\mathcal{H}$ (i.e., asymptotically, as $\Theta(|\mathcal{H}|)$). This is not feasible when $\mathcal{H}$ is continuous or high-dimensional.

When a (meta)hypothesis space (e.g., $\mathcal{P}$) is large, it is standard in multiple fields to choose a parametric manifold $\mathcal{M} \coloneqq \{\rho(\,\cdot\,;\theta)\colon \theta \in \mathbf{R}^n\} \subseteq \mathcal{P}$ of tractable solutions, using a parameter vector $\theta \in \mathbf{R}^n$, where n dictates computational space requirements. For example, we often linearize dynamical systems near equilibrium using a manifold described by eigenvalues and eigenvectors. Similarly, neural networks parameterize manifolds in function space using weights and biases. Unfortunately, replicator dynamics need not respect our chosen manifold: the replicator equation may force us off of $\mathcal{M}$, demanding values of ρ that are not addressable by any θ. To resolve this, we desire a model of evolutionary dynamics that is closed on any parametric manifold $\mathcal{M}$ — ideally one that approximates replicator dynamics as closely as possible...

Our central result is that Fisher-Rao Natural Gradient Descent (FR-NGD) provides an optimal approximation of continuous replicator dynamics constrained to any twice-differentiable parametric manifold $\mathcal{M}$ (Thm. 1). We refer to this correspondence as *conjugate natural selection* (CNS; Section 3). As an extension of this result, by building on known connections between the replicator equation and Bayes's rule, we prove that FR-NGD also provides an optimal approximation of Bayesian inference when loss is identified with a Kullback-Leibler divergence between predictions and observations (Section 4). We demonstrate applications of these findings to a non-convex optimization problem (Section 3.1) and parameter estimation of a stochastic process (Section 4.1).

By calling attention to the special case of FR-NGD among metrics for natural gradient descent, our work highlights beauty in the natural world and provides immediate applications. First, our result indicates a provocative correspondence between learning algorithms informed by information geometry and evolutionary processes driven by natural selection. Second, our experiments indicate that CNS provides an alternative approach to evolutionary computation for non-convex optimization and may be used for Bayesian system identification and parameter estimation.

1.1 Related Work

Prior work has discussed mutual connections between natural gradient descent, replicator dynamics, and Bayesian inference, though even when cast as a synthesis of these previous results, our results retain novelty. In particular, we are aware of no prior work that explicitly identifies FR-NGD as the best approximation of evolutionary dynamics nor Bayesian inference *for all twice-differentiable parameterization schemes* $\rho(h;\theta)$. Additionally, we believe our specific construction of continuous Bayesian inference is novel.

While exact correspondence between the replicator equation and FR-NGD has been previously identified for tabular or Boltzmann-Gibbs parameterized distributions (as the corresponding mirror-descent update) (Harper, 2009a, 2011;

Harper and Safyan, 2020; Bloembergen et al., 2015; Gao and Pavel, 2017; Hennes et al., 2019; Otwinowski et al., 2020; Chalub et al., 2021), this identity is limited to the case where $\mathcal{M} = \mathcal{P}$, and we are aware of no prior work that extends this correspondence to under-parameterized approximations $\mathcal{M} \subset \mathcal{P}$.

While Harper (2009b,a) recognizes the replicator equation as both an instance of FR-NGD and as an inference dynamic guided by Fisher information geometry and connected to Bayes's rule in discrete time, the cited work stops short of identifying the continuous time replicator equation as a *generalization* of Bayes's rule to continuous time, nor does it identify average fitness with a (negative) Kullback-Leibler divergence (despite recognizing the latter as a Lyapunov function for the replicator equation). Achieving deeper connections to Bayesian inference, recent work by Khan and Rue (2021) shows that FR-NGD gives rise to optimal Bayesian inference even for underparameterized distributions, but the provided analysis assumes exponential families, rather than arbitrary, twice-differentiable parameterizations.

The previous results cited above all generalize to arbitrary twice-differentiable parameterizations in light of recent work by Nurbekyan et al. (2022), who observe that natural gradient descent with any metric $g(\theta)$ yields the least-squares optimal approximation in $\mathcal{M}$ to natural gradient descent with metric $g(\rho)$ on $\mathcal{P}$ (Nurbekyan et al., 2022, Eq. 2.2). For our purposes: natural gradient descent on a parametric manifold always yields an optimal approximation of natural gradient descent in the continuous analog of the tabular setting (i.e., in the case of FR-NGD, the replicator equation, where, for every h, $\rho(h)$ may be independently specified). Nonetheless, to our knowledge, this observation has not been synthesized with the aforementioned results, nor have the implications for evolutionary dynamics been previously explored.

2 Preliminaries

Before detailing our results, we first review necessary background, establishing our setting in Section 2.1. We briefly discuss properties of the replicator equation in Section 2.2 and provide essential results from information geometry in Section 2.3.

2.1 Setting

In this paper, we denote a *hypothesis* as h, which we identify with a "strategy" in the evolutionary game theory literature (Friedman and Sinervo, 2016): For example, h may represent a combination of genes, a behavior, a belief, or a machine learning policy. Let $\mathcal{H}$ denote the space of possible hypotheses, such that $h \in \mathcal{H}$ for all h. For example, $\mathcal{H}$ might represent a population's genome, a set of competing social norms, an array of alternative beliefs, or the parameter space of a neural network. Finally, let $\mathcal{P}$ denote the simplex, or space of probability distributions, over $\mathcal{H}$.

Table 1: Variables

h	a hypothesis. $h \in \mathcal{H}$.
$\mathcal{H}$	hypothesis space (arbitrary in size and dimension).
$\mathcal{P}$	the space of probability distributions over $\mathcal{H}$.
ρ	a probability distribution over hypotheses. $\rho \in \mathcal{P}$.
H	a random hypothesis. $H \sim \rho$.
$\mathscr{L}$	a (possibly noisy) loss function. $\mathscr{L} \colon \mathcal{H} \to \mathbf{R}$.
$\overline{\mathscr{L}}_\rho$	the expected value of $\mathscr{L}$ according to distribution ρ.
θ	a parameter value. $\theta \in \mathbf{R}^n$. Components indexed, e.g., as θ^i.
$\rho(h;\theta)$	the probability (density) at h, parameterized by θ.
$\mathcal{M}$	the parametric manifold $\{\rho(\,\cdot\,;\theta) \colon \theta \in \mathbf{R}^n\}$. $\mathcal{M} \subseteq \mathcal{P}$.
$\mathcal{I}$	the Fisher (Section 2.3.1).
s	score (Def. 5), e.g., $s_i(\theta;h)$.
$\mathcal{E}$	"natural deviation" (Def. 8).

We use $\rho \in \mathcal{P}$ to represent an individual probability distribution over $\mathcal{H}$, and denote a hypothesis sampled at random from this distribution as $H \sim \rho$. In a slight abuse of notation, we denote the probability (density) associated with h in ρ as $\rho(h)$. When $\mathcal{H}$ is discrete, $\rho(h)$ corresponds to the relative frequency of h in a given population. When $\mathcal{H}$ is

continuous, $\rho(h)$ generalizes to a probability density, while sums over h generalize to integrals. While all equations in this paper readily generalize to continuous $\mathcal{H}$, we write our equations as if h were discrete for consistency and convenience. This is not a limitation of our results.

For ease of notation, let a dot above a symbol to denote its *full* time derivative (that is, $\forall u, \dot{u} \equiv \mathrm{d}u/\mathrm{d}t$) and a bar over a variable to denote its expectation value (explicitly, $\forall u, \overline{u} \equiv \mathrm{E}[u]$). We denote contravariant vectors with an upper index and covariant vectors with a lower index, using the Einstein summation convention of implicitly summing over matching upper and lower indices in a single term (formally, $\forall u, v, u_i v^i \equiv \sum_i u_i v^i$), but we will not use this convention for time index t. We also use standard shorthands for partial derivatives, identifying $\partial_i(\cdot) \equiv \partial(\cdot)/\partial\theta^i$ and $\partial_t(\cdot) \equiv \partial(\cdot)/\partial t$.

The **motivating problem** we consider is the minimization, over ρ, of expected *loss*, for some loss function $\mathscr{L}\colon \mathcal{H} \to \mathbf{R}$, when $H \sim \rho$. Put simply, we wish to select the distribution of hypotheses $\rho^\star$ with the smallest average loss.

$$\rho^\star = \operatorname*{argmin}_\rho \overline{\mathscr{L}}_\rho \quad ; \quad \overline{\mathscr{L}}_\rho \coloneqq \mathop{\mathrm{E}}_{H\sim\rho}\Big[\mathscr{L}(H)\Big] = \sum_h \rho(h)\mathscr{L}(h). \tag{1}$$

In general, we assume that $\mathscr{L}$ may be a non-convex function. For example, $\mathscr{L}(h)$ could represent the rate of excess deaths compared to births for a genotype h, the negative rate of total returns for an investment portfolio h, or the expected loss of machine learning policy h on a given task. For now, we assume that we may sample $\mathscr{L}$ without noise, but this condition is easily relaxed as long as noise remains unbiased.

Ultimately, our proposed solution for Prob. (1) involves a **twice-differentiable parametric manifold** $\mathcal{M} \subset \mathcal{P}$ of distributions $\rho(h;\theta) \in \mathcal{M}$ where $\theta \in \mathbf{R}^n$ is a parameter vector for integer n greater than zero. We analyze continuous time equations of motion for ρ and θ: the replicator equation for ρ and FR-NGD for θ. In Section 3, we show that the latter optimally approximates the former.

2.2 Replicator Dynamics

The replicator equation describes *replicator dynamics*. As background, we restate the continuous time replicator equation and introduce Price equation (Lem. 2). While our treatment throughout this paper assumes a continuous time variable t, we also derive the discrete-time form of the replicator equation in the supplementary material (Lem. 12), taking care to explicitly consider time intervals of the form $[t, t+\Delta t)$.

We have already introduced the continuous time replicator equation in a form adapted to the notation of machine learning literature:

Definition 1. *The replicator dynamics* are governed by the equation

$$\dot{\rho}(h) = \rho(h)\Big[\overline{\mathscr{L}}_\rho - \mathscr{L}(h)\Big], \quad \text{where} \quad \overline{\mathscr{L}}_\rho \coloneqq \sum_h \rho(h)\mathscr{L}(h), \quad \sum_h \rho(h) = 1. \tag{2}$$

Although we allow ourselves to omit time-indexing for ρ and $\mathscr{L}$, these quantities are time-varying.

Remark 1. (No new hypotheses). *For any finite times t and t', $\rho_t(h) = 0$ iff $\rho_{t'}(h) = 0$.*

Proof. The replicator equation has solutions of the form $\rho_t(h) = \rho_0(h) \exp \int_0^t \left(\overline{\mathscr{L}}_{\rho_{t'}} - \mathscr{L}_{t'}(h)\right) \mathrm{d}t'$, which does not admit roots in finite time unless ρ_0 (and therefore ρ_t, for all t), is zero. □

Rem. 1 reveals that **the replicator equation cannot generate new hypotheses**. For this reason, it is often combined with mutation or diffusion terms in practice, but the resulting dynamics are more difficult to solve (Bloembergen et al., 2015). As an approximation of replicator dynamics with under parameterized $\rho(\theta)$, **FR-NGD can avoid this issue**, since eliminated hypotheses can be reintroduced or provably never be eliminated (e.g., when $\rho(h;\theta)$ is everywhere non-zero by design).

Lemma 2. (The Price Equation). *For any function or real-valued property of hypotheses $u\colon \mathcal{H} \to \mathbf{R}$, the expected value of u, denoted $\overline{u}_\rho$ when h is sampled with probability $\rho(h)$, evolves according to*

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{u}_\rho = -\mathop{\mathrm{Cov}}_{H\sim\rho}\Big[u(H), \mathscr{L}(H)\Big] + \mathop{\mathrm{E}}_{H\sim\rho}\Big[\dot{u}(H)\Big] \quad ; \quad \overline{u}_\rho \coloneqq \mathop{\mathrm{E}}_{H\sim\rho}\Big[u(H)\Big]. \tag{3}$$

Many key results of evolutionary dynamics, such as fundamental theorems for gene and phenotype selection or heritability, may be derived from the Price equation (Queller, 2017). We provide a derivation in the supplementary material (Proof of Lem. 2), though this is a standard result.

2.3 Fisher-Rao Natural Gradient Descent

Variations of gradient descent have recently become standard techniques for non-convex optimization problems like Prob. (1), facilitated by automatic differentiation and parallelized updates (Fradkov, 2020; Tappert, 2020). In broad strokes, the technique is to first differentiably parameterize a search space $\mathcal{M}$ with a mapping $\theta \mapsto \rho(\,\cdot\,;\theta)$, for example, then repeatedly update θ in a direction that approximates the "fastest" decreasing value (i.e., the negative gradient) of the expected loss $\overline{\mathscr{L}}$.

Definition 2. *Naive Gradient Descent*, in continuous time, is given by the update rule

$$\dot{\theta}^i = -\partial_i \overline{\mathscr{L}}. \tag{4}$$

There is a problem with this update that is often unacknowledged in machine learning pedagogy: One side of this equation is *contravariant*, while the other is *covariant*. To understand this intuitively, assign units to the quantities such that $\dim(\theta) = \mathsf{U}$, $\dim(\mathscr{L}) = \mathsf{L}$, and $\dim(t) = \mathsf{T}$. It follows that $\dim(\dot{\theta}^i) = \mathsf{U}\mathsf{T}^{-1}$ while $\dim(\partial_i \overline{\mathscr{L}}) = \mathsf{L}\mathsf{U}^{-1}$. While a learning rate provides a natural conversion between L and T^{-1}, the powers of U do not balance on each side of Eq. (4), and the equation is dimensionally invalid. This problem is resolved by explicit consideration of an (inverse) metric g^{ij} with units U^2 that assigns distances and angles in the cotangent space of θ (i.e., where $\partial_i(\cdot), \partial_j(\cdot)$, etc. live). The metric g_{ij} applies to the tangent space of θ (i.e., where $\mathrm{d}\theta^i$, $\mathrm{d}\theta^j$, etc. live).

Definition 3. *Covariant Gradient Descent* is given by the update rule

$$\dot{\theta}^i = -g^{ij}\partial_j \overline{\mathscr{L}}. \tag{5}$$

Importantly, **the choice of metric g can strongly influence the dynamics of gradient descent**, which we call the gradient *flow*. That is, the direction of the "fastest" decreasing value of $\overline{\mathscr{L}}$ depends on how the tangent space of θ is measured by g. The implicit assumption of naive gradient descent is that $g^{ij} = \delta^{ij}$ for Kronecker delta δ with the appropriate units, i.e., a Euclidean metric for the (co)tangent space of θ. FR-NGD, that is, natural gradient descent with respect to the Fisher-Rao metric, uses a specific, alternative choice of g in Eq. (5) that derives from information geometry (Amari, 1998; Martens, 2020). It (un)warps the space around any given parameter value θ before performing the gradient update, so that small updates of θ in any direction all contain the same marginal information about the new distribution $\rho(\theta + \mathrm{d}\theta)$. The metric it uses is known as the *Fisher*.

In Fig. 1, we give an analogy for how the choice of metric can affect covariant gradient decent. Both a flat map and a globe "warp" our perception of local distances and angles, and can mislead us when finding the "fastest" route between two points on Earth.

2.3.1 The Fisher

The Fisher-Rao information metric tensor, Fisher information matrix (FIM), or "Fisher" may be expressed in multiple ways, but is defined thus:

Definition 4. *The Fisher* for θ is $\mathcal{I}_{ij}(\theta) \coloneqq \mathrm{Cov}_{H\sim\rho}\left[\partial_i \log \rho(H;\theta),\ \partial_j \log \rho(H;\theta)\right]$.

As a covariance matrix, $\mathcal{I}$ is symmetric and positive semi-definite. $\mathcal{I}$ fails to be fully positive definite (i.e., is degenerate) when parameter updates in different directions (up to constant scaling) affect ρ identically. Note that the quantity $\partial_i \log \rho(h;\theta)$ appearing in Def. 4 is called the *score*.

Definition 5. *The score* $s_i(\theta;h)$ is defined as $\partial_i \log \rho(h;\theta)$.

Lemma 3. (Zero Expected Score). *The expected score is zero. That is,* $\mathrm{E}_{H\sim\rho}\left[\partial_i \log \rho(H)\right] = 0$.

Proof. $\forall i,\ \mathrm{E}_{H\sim\rho}\left[\partial_i \log \rho(H)\right] = \sum_h \rho(h)\partial_i \log \rho(h) = \sum_h \partial_i \rho(h) = \partial_i \sum_h \rho(h) = 0.$ □

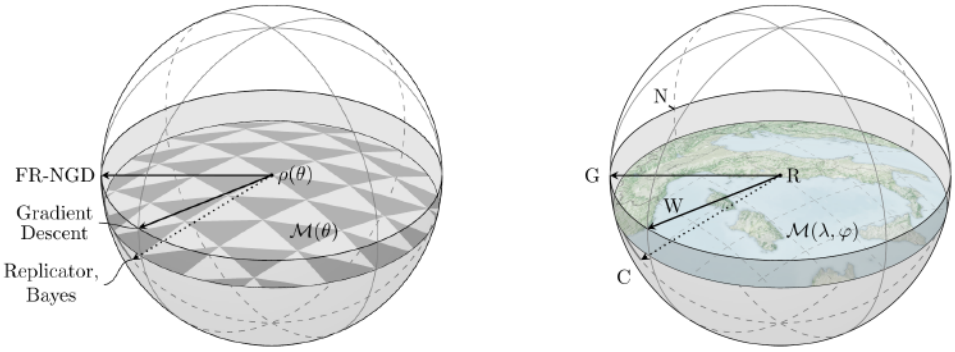

Figure 1: An analogy for how different metrics can suggest different parameter updates in Eq. (5). Earth's curvature is exaggerated to emphasize that vectors W and G are tangent to its surface. The direction of travel from Rome (point R) that most rapidly decreases one's distance from Chicago, when measured in the Euclidean space of latitude-longitude pairs (λ, φ), is nearly due west (vector W), because Rome and Chicago have nearly the same latitude λ. Performing gradient descent with an implicit Euclidean metric for parameter space is similarly naive. Vector C is tangent to the true shortest path in physical space: north-west at an angle of nearly 35 degrees downwards. Like the update given by the replicator equation or a continuous generalization of Bayes's rule, this direction may not be tangent to the manifold $\mathcal{M}$. Constrained to $\mathcal{M}$, the optimal approximation of the direction of C is its projection, G: north-west, tangent to the surface, and tangent to the geodesic from Rome to Chicago on the surface of the sphere. Map Data Credit: NASA Visible Earth.

2.3.2 Primal Gradient Flow

Definition 6. *The primal gradient flow*, for θ, induced by FR-NGD of $\overline{\mathscr{L}}$ with respect to θ is

$$\dot{\theta}^i = -\left[\mathcal{I}^\dagger(\theta)\right]^{ij} \partial_j \overline{\mathscr{L}}, \tag{6}$$

where $\mathcal{I}$ is the Fisher and $\mathcal{I}^\dagger$ is its Moore-Penrose inverse. While Eq. (6) is often used in practice to update θ, the dynamics of the *distribution* ρ are of ultimate, material consequence. The dynamics of ρ are described by the conjugate gradient flow.

2.3.3 Conjugate Gradient Flow

Definition 7. *The conjugate gradient flow*, for $\rho \in \mathcal{M}$, induced by FR-NGD of $\overline{\mathscr{L}}$ with respect to θ is

$$\mathcal{I}_{ij}(\theta)\dot{\theta}^j = -\partial_i \overline{\mathscr{L}}. \tag{7}$$

$\mathcal{I}(\theta)$ is positive definite and invertible when θ has only non-degenerate degrees of freedom, in which case $\mathcal{I}^\dagger(\theta) = \mathcal{I}^{-1}(\theta)$, the nullspace of $\mathcal{I}^\dagger(\theta)$ is orthogonal to the tangent space of $\mathcal{M}$, and Eq. (7) and Eq. (6) are equivalent. When $\mathcal{I}(\theta)$ is not invertible, which occurs when θ has redundant degrees of freedom (e.g., in a state of gimbal lock), the properties of the Moore-Penrose inverse imply that Eq. (6) solves Eq. (7), producing, among under-determined solutions for $\dot{\theta}$, the one with minimal Euclidean norm. In this case, the *conjugate* gradient flow induced by FR-NGD with respect to θ is still described by Eq. (7). We show that this conjugate gradient flow is an optimal approximation of the replicator equation (Thm. 1).

To prime intuition, using results provided in the supplementary material (Lem. 6 and Lem. 7), we may rewrite the conjugate gradient flow (Def. 7) as

$$\sum_h \Big[\underbrace{\partial_i \log \rho(h)}_{s_i}\Big] \dot{\rho}(h) = \sum_h \Big[\underbrace{\partial_i \log \rho(h)}_{s_i}\Big] \rho(h) \Big(\overline{\mathscr{L}} - \mathscr{L}(h)\Big). \tag{8}$$

Comparing Eq. (2) and Eq. (8), we **recognize conjugate gradient flow as a projection of the replicator dynamics** onto the dual space of θ spanned by the score (Def. 5), where the score provides a basis for the tangent space of $\mathcal{M}$ at θ

such that, by Lem. 3, local motion along each basis vector introduces zero marginal entropy relative to $\rho(\,\cdot\,;\theta)$:

$$\mathop{\mathrm{E}}_{H\sim\rho}\Big[\log\rho(H)+s_i(H)\,\mathrm{d}\theta^i\Big]=\mathop{\mathrm{E}}_{H\sim\rho}\Big[\log\rho(H)\Big] \tag{9}$$

As a projection of replicator dynamics onto $\mathcal{M}$, we naturally expect the conjugate gradient flow to **minimize an appropriately defined distance** from the replicator dynamics. Indeed, we define such a distance with Def. 8 and realize this expectation with Thm. 1.

3 Conjugate Natural Selection

In this section, we state our primary results. We make use of the Fisher metric for ρ, also known as the Shahshahani metric (Harper, 2009a), which follows from Def. 4 when $\theta\equiv\rho$:

$$\mathcal{I}_{ij}(\rho)=\mathop{\mathrm{E}}_{H\sim\rho}\left[\frac{\delta_{ij}}{\rho(H)^2}\right]. \tag{10}$$

Definition 8. *The natural deviation* $\mathcal{E}$ of $\dot\rho$, induced by $\dot\theta$, from its nominal value under the replicator equation is given by the corresponding mean-squared error in realized relative fitness $\mathrm{d}/\mathrm{d}t\log\rho$.

$$\mathcal{E}\big(\dot\theta\big)\coloneqq\frac{1}{2}\mathop{\mathrm{E}}_{H\sim\rho}\left[\Big(\underbrace{\frac{\mathrm{d}}{\mathrm{d}t}\log\rho(H)}_{\dot\rho(H)/\rho(H)}-\underbrace{\big(\overline{\mathscr{L}}-\mathscr{L}(H)\big)}_{\dot\rho^\star(H)/\rho(H)}\Big)^2\right]=\frac{1}{2}\Big(\dot\rho-\dot\rho^\star\Big)^i\mathcal{I}_{ij}(\rho)\Big(\dot\rho-\dot\rho^\star\Big)^j. \tag{11}$$

As desired, $\mathcal{E}$ defines a distance in the tangent space of $\mathcal{P}$, imposed by the Fisher metric $\mathcal{I}(\rho)$ between the replicator dynamics $\dot\rho^\star$ and the dynamics $\dot\rho$ realized by FR-NGD with respect to θ. By inspection, we see that $\mathcal{E}$ is minimized for tabular settings ($\rho\equiv\theta$) if and only if the replicator equation holds (i.e., $\dot\rho=\rho(\overline{\mathscr{L}}-\mathscr{L}(h))$). The minimization of $\mathcal{E}$ by FR-NGD generalizes to any twice-differentiable parameterization of ρ by θ.

Theorem 1 (**Conjugate Natural Selection; Main Result**). *Constrained to a given manifold of twice-differentiable parametric policies $\rho(h;\theta)$, FR-NGD of $\overline{\mathscr{L}}$ with respect to θ (Eqs. (6) and (7)) achieves the least-squares optimal fit in $\dot\theta$ to the continuous time replicator dynamics (i.e., Eq. (2)), as measured by the natural deviation $\mathcal{E}$ (Def. 8).*

Our Proof of Thm. 1, provided in the supplementary material, proceeds by establishing that FR-NGD of $\overline{\mathscr{L}}$ with respect to θ induces a stationary point of $\mathcal{E}$, such that $\partial\mathcal{E}/\partial\dot\theta=0$, with a Hessian that is positive semi-definite everywhere, implying a global minimum.

In addition to proving an optimal correspondence between FR-NGD with respect to θ and the continuous time replicator equation (Thm. 1), we may characterize the space of functions $u\colon\mathcal{H}\to\mathbf{R}$ that undergo the same dynamics under either update rule as linear combinations of score (Thm. 2). Finally, we demonstrate an application of conjugate natural selection (CNS) by experimentally evolving a distribution of continuous hypotheses for a non-convex problem.

Theorem 2 (Preserved Dynamics). *Linear combinations of score satisfy the Price equation (Eq. (3)) when θ is updated via FR-NGD of $\overline{\mathscr{L}}$. That is,*

$$\forall\alpha^i\in\mathbf{R},u=\alpha^i s_i(\theta;h),\quad\frac{\mathrm{d}}{\mathrm{d}t}\mathop{\mathrm{E}}_{H\sim\rho}\Big[u(H)\Big]=-\mathop{\mathrm{Cov}}_{H\sim\rho}\Big[u(H),\mathscr{L}(H)\Big]+\mathop{\mathrm{E}}_{H\sim\rho}\Big[\dot u(H)\Big]. \tag{12}$$

We include Proof of Thm. 2 in the supplementary material.

3.1 Applications

We demonstrate an application of CNS by evolving a Gaussian distribution over candidate solutions for a non-convex optimization problem: namely, unconstrained minimization of the Rastrigin function,

$$\mathscr{L}(h_x,h_y)=20+h_x^2+h_y^2-10\cos(2\pi h_x)-10\cos(2\pi h_y),$$

depicted in the rightmost pane of Fig. 2.

At each time step, $N{=}40$ hypotheses h are sampled from ρ_t and the loss for each h is calculated, yielding a Monte Carlo estimate of the loss gradient $\partial_i \overline{\mathscr{L}} \approx \frac{1}{N} \sum_{k=1}^{N} \mathscr{L}(h_k) \partial_i \log \rho(h_k)$. We use an analytically-known form of the Fisher for a non-degenerate parameterization of a 2-dimensional Gaussian distribution using 5 degrees of freedom and an Euler discretization of the dynamics. We provide our code for this simulation in the supplementary material.

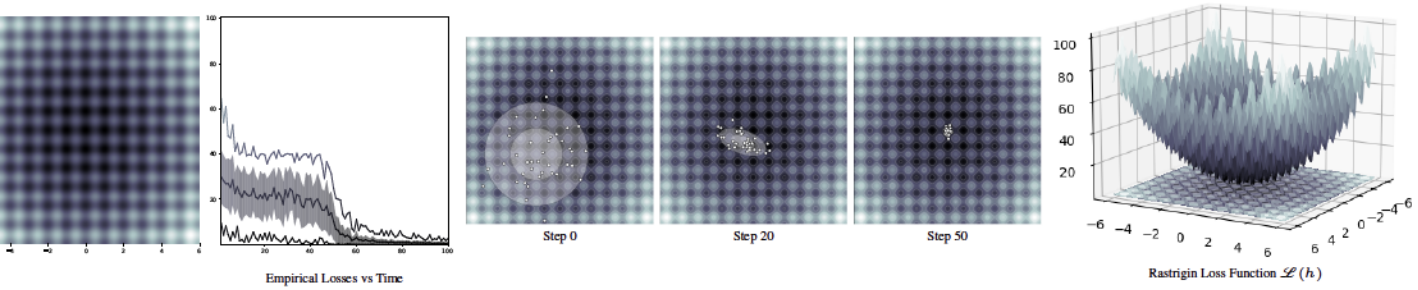

Figure 2: On the left, we plot the mean, standard deviation, and extremal empirical losses for the learned distribution over 100 time steps. On the right, the loss function is visualized as a surface over the domain $[-6, 6] \times [-6, 6]$. In the middle, we represent time steps 0, 20, and 50 of the evolution: The Rastrigin function is visualized with shading and highlighted level sets; the sampled hypothesis are represented by white dots; and the 1- and 2-σ ellipses for the evolving Gaussian distribution ρ are shaded white with partial transparency. The distribution is initialized with mean at $[-1.5, -1.5]$ and identity covariance, and we use a constant learning rate of $1e^{-3}$ for the Euler update. An animation of the time-evolution of the distribution is available with the included source code.

3.2 Limitations

Recent characterizations of FR-NGD have suggested quadratic convergence rates under certain conditions (Müller and Montúfar, 2022; Hu et al., 2022), though care must be taken to choose appropriate parametric manifolds $\mathcal{M}$ for $\mathscr{L}$. (Intuitively, it is possible to "optimally" fit data to any model, but the model must be appropriate to the domain for the optimal fit to be useful).

Until recently, the bottleneck for applying FR-NGD in practice was the $\mathcal{O}(n^{\approx 2.37})$ cost of Fisher matrix inversion, prompting alternative empirical approximations of natural gradient descent (Martens, 2020; Hennes et al., 2019; Peirson et al., 2022). A more scalable approach, based on solving the corresponding least-squares problem directly, has been recently proposed by Nurbekyan et al. (2022).

Our simple demonstration in Fig. 2 indicates that conjugate natural selection provides an alternative approach to standard approaches in evolutionary computation: Rather than directly simulate a population, we may use FR-NGD to update a parametric *distribution* of candidate solutions and ultimately solve a non-convex optimization problem, even when the hypothesis space is continuous or high-dimensional. As $\overline{\mathscr{L}}$ and $\mathcal{I}(\theta)$ are often approximated by empirical averages, we also comment that this approach readily extends in the presence of noise that is independent of ρ or θ. In particular, the hypothesis space $\mathcal{H}$ may correspond to the space of *functions* over a random input, as in Section 4. Finally, we assert that using FR-NGD for evolutionary computation may be suitable for *constrained* optimization in practice, because simple sample rejection can guarantee that domain constraints for h are satisfied (although sample rejection will distort the corresponding Fisher information matrix).

4 Continuous Bayesian Inference

For this section, let us interpret h as a predictive model yielding a probability (density) for a stochastic process X_t observed in continuous time t and distributed by Nature as $\mathfrak{n}$. Examples of such processes X_t include physical quantities like instantaneous field amplitude; the idealized market price of an asset; Brownian motion; or any continuous time quantity perturbed by *noise*. In this context, the replicator equation (Eq. (2)) corresponds to *continuous Bayesian inference* if we identify loss with the negative log-likelihood of hypothesis h given x_t. That is, let

$$h(x_t; t) = \Pr_h(X_t{=}x_t \,|\, t) \quad ; \quad \mathfrak{n}(x_t; t) = \Pr_{\mathfrak{n}}(X_t{=}x_t \,|\, t) \quad ; \quad \mathscr{L}(h, t) = -\log h(x_t; t). \tag{13}$$

The loss $\mathscr{L}$ expressed in Eq. (13) corresponds to *surprisal*, or the amount of information about X_t revealed under hypothesis h by the observation $X_t{=}x_t$. A good hypothesis minimizes average surprisal by correctly predicting the process X_t. For this loss, the replicator equation (Eq. (2)) describes stochastic dynamics for ρ that depend on the instantaneous value of x_t:

$$\dot{\rho}_t(h, x_t) = \rho_t(h)\Big[\overline{\mathscr{L}}_{\rho_t}(x_t) + \log h(x_t; t)\Big], \quad \text{where} \quad \overline{\mathscr{L}}_{\rho_t}(x_t) = -\sum_h \rho_t(h) \log h(x_t; t). \tag{14}$$

We will use the gradient of the expected value of $\overline{\mathscr{L}}_{\rho_t}(\cdot)$ over $X_t \sim \mathfrak{n}$ and perform FR-NGD to evolve ρ. While $\overline{\mathscr{L}}_{\rho_t}(\cdot)$ might be formally defined as a cross-entropy term, its *gradient* with respect to ρ_t is the same as the gradient of the expected Kullback-Leibler divergence (relative-entropy) $\mathscr{D}$ from h to $\mathfrak{n}$, defined

$$\mathscr{D}_t(\mathfrak{n} \parallel h) \coloneqq -\sum_x \mathfrak{n}(x; t) \log \frac{h(x; t)}{\mathfrak{n}(x; t)}. \tag{15}$$

Lemma 4. *The gradients of* $\mathrm{E}_{X_t \sim \mathfrak{n}}[\overline{\mathscr{L}}_{\rho_t}(X_t)]$ *and* $\mathrm{E}_{H \sim \rho_t}[\mathscr{D}_t(\mathfrak{n} \parallel H)]$ *with respect to* ρ_t *are equal.*

A proof of Lem. 4 is provided in the supplementary material.

Theorem 3 (Continuous Inference)**.** *Eq. (14) may be used to derive Bayes's rule.*

We provide Proof of Thm. 3 in the supplementary material. Unfortunately, when $\mathcal{H}$ is large, it is **difficult to compute the *prior*** for an observable, as this calculation requires integrating over $\mathcal{H}$ (i.e., $\Pr_{\rho_t}(X_t^{\Delta t}) = \sum_h \rho_t(h) h(X_t^{\Delta t}, t)$ in Eq. (21)). This difficulty is used to justify approximate Bayesian inference based on variational bounds, such as the Evidence Lower Bound (ELBO). We may **avoid the corresponding difficulty via FR-NGD**, however, by using Monte Carlo sampling to estimate the necessary gradient, $\partial_i \mathrm{E}_{H \sim \rho}[\mathscr{D}_t(\mathfrak{n} \parallel H)] = -\mathrm{E}_{X_t, H}\left[\partial_i \log \rho(H; \theta) \log H(X_t, t)\right]$.

Theorem 4 (FR-NGD Yields Optimal Continuous Inference)**.** *For any probability distribution* $\rho(h; \theta)$ *that is twice-differentiable with respect to parameters* θ*, FR-NGD of the expected divergence* $\mathrm{E}_{H \sim \rho_t}[\mathscr{D}_t(\mathfrak{n} \parallel H)]$ *(of the* ρ*-weighted predictions of model* $h(X_t; t)$ *for* $X_t \sim \mathfrak{n}$*) with respect to* θ *optimally approximates Bayesian inference for* ρ *in continuous time, by minimizing* $\mathcal{E}$ *(Def. 8).*

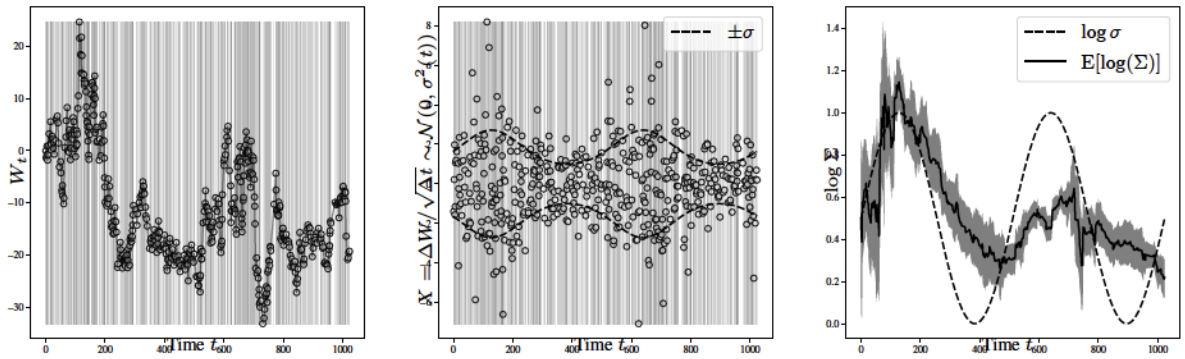

Figure 3: Given randomly-spaced samples of a Wiener process W_t with time-dependent variance (left), where vertical lines indicate the times of observed samples (left, center), we define the observable $X_t = \Delta W/\sqrt{\Delta t} \sim \mathfrak{n} = \mathcal{N}(0, \sigma^2(t))$ between measurements of W_t (center). We update a θ-parameterized Gaussian distribution ρ over $H = \log \Sigma$ via FR-NGD, with respect to θ, of the expected Kullback-Leibler divergence between n and $H \sim \rho$, interpreting Σ as an estimate of the time-evolving parameter $\sigma(t)$. We visualize the mean and standard-deviation of $\log \Sigma \sim \rho$ with time, compared to the actual value of $\log \sigma$ (right).

4.1 Parameter Estimation for a Stochastic Process

Using FR-NGD, we demonstrate learning a time-varying *distribution* $\rho = \mathcal{N}(\theta; t)$ over $H = \log(\Sigma)$, where Σ is an estimator for the time-varying parameter $\sigma(t)$ of an observable Wiener process $\mathrm{d}W_t \sim \mathcal{N}(0, \sigma^2(t)\,\mathrm{d}t)$ (Fig. 3). Our example uses a 40-sample Monte Carlo gradient estimate and an Euler discretization of the dynamics $\theta_{t+1} = \theta_t + \eta\dot{\theta}_t$ with constant learning rate $\eta = \mathtt{1e}^{-2}$.

5 Conclusion

We have shown that FR-NGD optimally approximates evolutionary and Bayesian dynamics for any twice-differentiable parameterization of a distribution over hypotheses. We believe it is remarkable that the essential dynamics of evolution by natural selection share such close relationships with the fundamentals of information theory, and that the unifying theoretical machinery is widely applicable to machine learning and optimization in practice.

In the case of the correspondence between FR-NGD and evolutionary dynamics, we have termed our finding "conjugate natural selection" and demonstrated its application to a non-convex optimization problem over a continuous hypothesis space. We assert that this approach provides an alternative to existing methods of evolutionary computation by dispensing with the need to simulate populations

Acknowledgments This work is partially supported by the National Science Foundation (NSF) under grants IIS-2143895 and IIS-2040800 (FAI program in collaboration with Amazon), and CCF-2023495. We thank Shadi Haddad, Yin Lin, Andrew Warren, Warren Mardoum, Ehsan Amid, and Abhishek Halder for feedback on this or prior versions of our manuscript or for consultation on technical details.

References

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Thomas Bäck, David B Fogel, and Zbigniew Michalewicz. *Evolutionary computation 1: Basic algorithms and operators*. CRC press, 2018.

Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 53:659–697, 2015.

Fabio ACC Chalub, Léonard Monsaingeon, Ana Margarida Ribeiro, and Max O Souza. Gradient flow formulations of discrete and continuous evolutionary models: a unifying perspective. *Acta Applicandae Mathematicae*, 171(1):1–50, 2021.

Ross Cressman and Yi Tao. The replicator equation and other game dynamics. *Proceedings of the National Academy of Sciences*, 111(supplement_3):10810–10817, 2014.

Alexander L Fradkov. Early history of machine learning. *IFAC-PapersOnLine*, 53(2):1385–1390, 2020.

Daniel Friedman. Evolutionary games in economics. *Econometrica: journal of the econometric society*, pages 637–666, 1991.

Daniel Friedman and Barry Sinervo. *Evolutionary games in natural, social, and virtual worlds*. Oxford University Press, 2016.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Marc Harper. Information geometry and evolutionary game theory. *arXiv preprint arXiv:0911.1383*, 2009a.

Marc Harper. The replicator equation as an inference dynamic. *arXiv preprint arXiv:0911.1763*, 2009b.

Marc Harper. Escort evolutionary game theory. *Physica D: Nonlinear Phenomena*, 240(18):1411–1415, 2011.

Marc Harper and Joshua Safyan. Momentum accelerates evolutionary dynamics. *arXiv preprint arXiv:2007.02449*, 2020.

Daniel Hennes, Dustin Morrill, Shayegan Omidshafiei, Remi Munos, Julien Perolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, Paavo Parmas, Edgar Duenez-Guzman, et al. Neural replicator dynamics. *arXiv preprint arXiv:1906.00190*, 2019.

Josef Hofbauer, Karl Sigmund, et al. *Evolutionary games and population dynamics*. Cambridge university press, 1998.

Jiang Hu, Ruicheng Ao, Anthony Man-Cho So, Minghan Yang, and Zaiwen Wen. Riemannian natural gradient methods. *arXiv preprint arXiv:2207.07287*, 2022.

Mohammad Emtiyaz Khan and Hravard Rue. The Bayesian learning rule. *arXiv preprint arXiv:2107.04562*, 2021.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2): 212–261, 1994.

Elisabeth Lloyd. Units and levels of selection. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2020 edition, 2020.

James Martens. New insights and perspectives on the natural gradient method. *The Journal of Machine Learning Research*, 21(1):5776–5851, 2020.

Johannes Müller and Guido Montúfar. Geometry and convergence of natural policy gradient methods. *arXiv preprint arXiv:2211.02105*, 2022.

Levon Nurbekyan, Wanzhou Lei, and Yunan Yang. Efficient natural gradient descent methods for large-scale optimization problems. *arXiv preprint arXiv:2202.06236*, 2022.

Jakub Otwinowski, Colin H LaMont, and Armita Nourmohammad. Information-geometric optimization with natural selection. *Entropy*, 22(9):967, 2020.

Abel Peirson, Ehsan Amid, Yatong Chen, Vladimir Feinberg, Manfred K Warmuth, and Rohan Anil. Fishy: Layerwise fisher approximation for higher-order neural network optimization. In *Has it Trained Yet? NeurIPS 2022 Workshop*, 2022.

David C Queller. Fundamental theorems of evolution. *The American Naturalist*, 189(4):345–353, 2017.

William H Sandholm. *Population games and evolutionary dynamics*. MIT press, 2010.

Barry Sinervo and Ryan Calsbeek. The developmental, physiological, neural, and genetical causes and consequences of frequency-dependent selection in the wild. *Annu. Rev. Ecol. Evol. Syst.*, 37:581–610, 2006.

Charles C Tappert. Frank Rosenblatt, the father of deep learning. *Proceedings of Student-Faculty Research Day, CSIS, Pace University*, 2020.

A Deferred Proofs

We organize our deferred proof thematically. In Appendix A.1, we provide lemmas that allow for more succinct proofs throughout the rest of this section. In Appendix A.2, we derive Eq. (8), appearing in the main text. In Appendix A.3, we prove our primary result, deemed "conjugate natural selection" (Thm. 1). In Appendix A.4, we derive the Price equation and our result regarding dynamics preserved between replicator dynamics and FR-NGD on $\mathcal{M}$ (Thm. 2). Finally, in Appendix A.5, we derive the discrete replicator equation and Bayes's rule, and we prove the optimality of FR-NGD for continuous Bayesian inference (Thm. 3).

A.1 Four Useful Lemmas (5-8)

Lems. 5 to 7 provide equivalent expressions for quantities that frequently appear in our problem domain: Lem. 5 identifies a simple term-rewriting rule, while Lem. 6 and Lem. 7 address either side of the conjugate gradient flow equation (Def. 7), restated below. Lem. 8 proves a simple identity relying on repeated application of the chain rule.

Lemma 5. (Functionals). *For any function $u\colon \mathcal{H} \to \mathbf{R}$, for all parameter components i, the following expressions are equivalent:*

$$\sum_h \partial_i \rho(h) u(h) = \mathop{\mathrm{E}}_{H\sim\rho}\left[u(H)\partial_i \log \rho(H)\right] = \mathop{\mathrm{Cov}}_{H\sim\rho}\left[\partial_i \log \rho(H), u(H)\right].$$

Proof. The first equality is an instance of the "log-derivative trick", while the second follows from Lem. 3:

$$\forall u, \quad \sum_h \partial_i \rho(h) u(h) = \underbrace{\sum_h \rho(h)\partial_i \log \rho(h) u(h)}_{\mathrm{E}_{H\sim\rho}[u(H)\partial_i \log \rho(H)]} = \mathop{\mathrm{Cov}}_{H\sim\rho}\left[\partial_i \log \rho(H), u(H)\right].$$

□

Briefly recall the definitions for the Fisher (Def. 4) and conjugate gradient flow under FR-NGD (Def. 7), which we reference in our proof of Lem. 6:

Definition 4. *The Fisher* for θ is $\mathcal{I}_{ij}(\theta) \coloneqq \mathrm{Cov}_{H\sim\rho}\left[\partial_i \log \rho(H;\theta),\ \partial_j \log \rho(H;\theta)\right]$.

Definition 7. *The conjugate gradient flow*, for $\rho \in \mathcal{M}$, induced by FR-NGD of $\overline{\mathscr{L}}$ with respect to θ is

$$\mathcal{I}_{ij}(\theta)\dot{\theta}^j = -\partial_i \overline{\mathscr{L}}. \tag{7}$$

Lemma 6. (Conjugate Flow) *The dynamics of $\rho \in \mathcal{M}$ under FR-NGD of $\overline{\mathscr{L}}$ with respect to θ (Def. 7 are*

$$\mathcal{I}_{ij}(\theta)\dot{\theta}^j = \mathop{\mathrm{Cov}}_{H\sim\rho}\left[\partial_i \log \rho(H),\ \frac{\mathrm{d}}{\mathrm{d}t}\log \rho(H)\right] = \sum_h \Big[\partial_i \log \rho(h)\Big]\dot{\rho}(h).$$

where we elide the explicit dependence of ρ on θ for compactness.

Proof. The first equality follows from substituting the definition of Fisher information (Def. 4) into Def. 7 and summing over j, by the linearity of covariance.

$$\begin{aligned}\mathcal{I}_{ij}(\theta)\dot{\theta}^j &= \mathop{\mathrm{Cov}}_{H\sim\rho}\Big[\partial_i \log \rho(H),\ \partial_j \log \rho(H)\Big]\dot{\theta}^j \\ &= \mathop{\mathrm{Cov}}_{H\sim\rho}\left[\partial_i \log \rho(H),\ \frac{\mathrm{d}}{\mathrm{d}t}\log \rho(H)\right].\end{aligned}$$

The second follows from rewriting covariance as an explicit sum over h while invoking Lem. 3 and pushing the time derivative through the logarithm, that is,

$$\begin{aligned}\mathop{\mathrm{Cov}}_{H\sim\rho}\left[\partial_i \log \rho(H),\ \frac{\mathrm{d}}{\mathrm{d}t}\log \rho(H)\right] &= \mathop{\mathrm{Cov}}_{H\sim\rho}\left[\partial_i \log \rho(H),\ \frac{1}{\rho(H)}\frac{\mathrm{d}}{\mathrm{d}t}\rho(H)\right] \\ &= \sum_h \rho(h)\Big[\partial_i \log \rho(h)\Big]\frac{1}{\rho(h)}\frac{\mathrm{d}}{\mathrm{d}t}\rho(h) \\ &= \sum_h \Big[\partial_i \log \rho(h)\Big]\dot{\rho}(h).\end{aligned}$$

□

Lemma 7. (Loss Gradient) *The gradient of expected loss $\overline{\mathscr{L}}$ with respect to θ is*

$$\partial_i \overline{\mathscr{L}} = \operatorname*{Cov}_{H\sim\rho}\Big[\partial_i \log \rho(H),\ \mathscr{L}(H)\Big] = -\sum_h \Big[\partial_i \log \rho(h)\Big]\rho(h)\Big(\overline{\mathscr{L}} - \mathscr{L}(h)\Big).$$

Proof. As the function $\mathscr{L}\colon \mathcal{H} \to \mathbf{R}$ has no explicit dependence on θ, we may expand

$$\partial_i \overline{\mathscr{L}} = \sum_h \mathscr{L}(h)\partial_i \rho(h).$$

By Lem. 5, when we choose $u = \mathscr{L}$, we may equate this sum with the desired covariance between $\partial_i \log \rho$ and $\overline{\mathscr{L}}$:

$$\sum_h \mathscr{L}(h)\partial_i \rho(h) = \operatorname*{Cov}_{H\sim\rho}[\partial_i \log \rho(H), \mathscr{L}(H)].$$

Finally, we re-express this covariance as an explicit sum:

$$\operatorname*{Cov}_{H\sim\rho}\Big[\partial_i \log \rho(H),\ \mathscr{L}(H)\Big] = -\sum_h \Big[\partial_i \log \rho(h)\Big]\rho(h)\Big(\overline{\mathscr{L}} - \mathscr{L}(h)\Big).$$

□

Lemma 8. (Cancel Dots with Chain Rule). $\forall u$ *independent of* $(\dot\theta, \dot\rho)$,

$$\frac{\partial}{\partial \dot\theta^i}\left(\frac{\mathrm{d}}{\mathrm{d}t}u(\rho)\right) = \partial_i u(\rho).$$

Proof. By the chain rule,

$$\frac{\partial}{\partial \dot\theta^i}\left(\frac{\mathrm{d}}{\mathrm{d}t}u(\rho)\right) = \frac{\partial}{\partial \dot\theta^i}\left(u'(\rho)\dot\rho\right) = u'(\rho)\frac{\partial \dot\rho}{\partial \dot\theta^i} = u'(\rho)\frac{\partial}{\partial \dot\theta^i}\left(\dot\theta^i \partial_i \rho + \partial_t \rho\right) = u'(\rho)\partial_i \rho = \partial_i u(\rho).$$

□

A.2 Proof: Projection of Replicator Dynamics (Eq. (8))

In the main text, we claim that the conjugate gradient flow (Def. 7) may be re-written as

$$\sum_h \Big[\underbrace{\partial_i \log \rho(h)}_{s_i}\Big]\dot\rho(h) = \sum_h \Big[\underbrace{\partial_i \log \rho(h)}_{s_i}\Big]\rho(h)\Big(\overline{\mathscr{L}} - \mathscr{L}(h)\Big). \tag{7}$$

Proof. By direction application of Lem. 6 and Lem. 7, we compare Eq. (8) to Eq. (7):

$$\begin{aligned}
\mathcal{I}_{ij}(\theta)\dot\theta^j &= -\partial_i \overline{\mathscr{L}}. \\
\sum_h \Big[\partial_i \log \rho(h)\Big]\dot\rho(h) &= \mathcal{I}_{ij}(\theta)\dot\theta^j. \\
-\partial_i \overline{\mathscr{L}} &= \sum_h \Big[\partial_i \log \rho(h)\Big]\rho(h)\Big(\overline{\mathscr{L}} - \mathscr{L}(h)\Big). \\
\therefore\quad \sum_h \Big[\partial_i \log \rho(h)\Big]\dot\rho(h) &= \sum_h \Big[\partial_i \log \rho(h)\Big]\rho(h)\Big(\overline{\mathscr{L}} - \mathscr{L}(h)\Big).
\end{aligned}$$

□

A.3 Proof: Conjugate Natural Selection

With Lem. 9, we establish that FR-NGD of $\overline{\mathscr{L}}$ with respect to θ (Def. 7) induces a local extremum of $\mathcal{E}$ (Def. 8) with respect to $\dot{\theta}$. With Lem. 10, we establish that this local extremum is a global minimum, by the fact that the Hessian of $\mathcal{E}$ with respect to $\dot{\theta}$ is everywhere positive semi-definite; In fact, this Hessian is the Fisher. We conclude that FR-NGD of $\overline{\mathscr{L}}$ with respect to θ is an optimal approximation of the replicator dynamics, a finding we term "conjugate natural selection" (Thm. 1).

We first restate Def. 8:

Definition 8. *The natural deviation* $\mathcal{E}$ of $\dot{\rho}$, induced by $\dot{\theta}$, from its nominal value under the replicator equation is given by the corresponding mean-squared error in realized relative fitness $\mathrm{d}/\mathrm{d}t \log \rho$.

$$\mathcal{E}(\dot{\theta}) \coloneqq \frac{1}{2} \mathop{\mathrm{E}}_{H\sim\rho}\left[\Big(\underbrace{\frac{\mathrm{d}}{\mathrm{d}t}\log\rho(H)}_{\dot{\rho}(H)/\rho(H)} - \underbrace{\big(\overline{\mathscr{L}} - \mathscr{L}(H)\big)}_{\dot{\rho}^\star(H)/\rho(H)}\Big)^2\right] = \frac{1}{2}\Big(\dot{\rho}-\dot{\rho}^\star\Big)^i \mathcal{I}_{ij}(\rho)\Big(\dot{\rho}-\dot{\rho}^\star\Big)^j. \tag{11}$$

We additionally recall that $\mathcal{I}(\rho)$ is given by

$$\mathcal{I}_{ij}(\rho) = \mathop{\mathrm{E}}_{H\sim\rho}\left[\frac{\delta_{ij}}{\rho(H)^2}\right]. \tag{10}$$

Lemma 9. *(Gradient of Natural Deviation)*

$$\frac{\partial}{\partial\dot{\theta}^i}\mathcal{E}(\dot{\theta}) = \mathcal{I}_{ij}(\theta)\dot{\theta}^j + \partial_i\overline{\mathscr{L}}.$$

Proof. We may take the gradient of Eq. (11) with respect to $\dot{\theta}$. We use Lem. 8 (for $u = \log$) to first write

$$\frac{\partial}{\partial\dot{\theta}^i}\left(\frac{\mathrm{d}}{\mathrm{d}t}\log\rho(h)\right) = \partial_i \log\rho(h),$$

thus,

$$\frac{\partial}{\partial\dot{\theta}^i}\mathcal{E}(\dot{\theta}) = \mathop{\mathrm{E}}_{H\sim\rho}\left[\left(\frac{\mathrm{d}}{\mathrm{d}t}\log\rho(H) - \big(\overline{\mathscr{L}} - \mathscr{L}(H)\big)\right)\partial_i\log\rho(H)\right]. \tag{16}$$

Recognizing that $\frac{\mathrm{d}}{\mathrm{d}t}\log\rho = \frac{1}{\rho}\dot{\rho}$, the expectation value on the right side of Eq. (16) separates into two explicit sums; i.e.,

$$\frac{\partial}{\partial\dot{\theta}^i}\mathcal{E}(\dot{\theta}) = \sum_h \left[\partial_i\log\rho(h)\right]\dot{\rho}(h) - \sum_h \left[\partial_i\log\rho(h)\right]\rho(h)\left(\overline{\mathscr{L}} - \mathscr{L}(h)\right).$$

These sums correspond to the gradient flow on $\mathcal{M}$ (Lem. 6) and the loss gradient (Lem. 7), respectively. □

Lemma 10. (Hessian of Natural Deviation) *$\mathcal{E}$ is convex in $\dot{\theta}$, that is,*

$$\frac{\partial^2}{\partial\dot{\theta}^i\partial\dot{\theta}^j}\mathcal{E}(\dot{\theta}) = \mathcal{I}_{ij}(\theta) \succeq 0,$$

where $\mathcal{I}_{ij}(\theta) \succeq 0$ denotes that $\mathcal{I}_{ij}(\theta)$ is positive semi-definite (has only non-negative eigenvalues).

Proof. We differentiate Eq. (16) with respect to $\dot{\theta}$, again using Lem. 8 (for $u = \log$). Thus, the second derivative of $\mathcal{E}$ is

$$\begin{aligned}\frac{\partial^2}{\partial\dot{\theta}^i\partial\dot{\theta}^j}\mathcal{E}(\dot{\theta}) &= \mathop{\mathrm{E}}_{H\sim\rho}\left[\left(\partial_i\log\rho(H)\right)\left(\partial_j\log\rho(H)\right)\right] \\ &= \mathop{\mathrm{Cov}}_{H\sim\rho}\left[\partial_i\log\rho(H;\theta),\ \partial_j\log\rho(H;\theta)\right] \\ &= \mathcal{I}_{ij}(\theta).\end{aligned}$$

Where the last equality relies on Def. 4. As a covariance matrix, $\mathcal{I}$ is positive semi-definite (i.e., $\mathcal{I} \succeq 0$). □

That the Fisher is the Hessian of $\mathcal{E}$ with respect to $\dot{\theta}$ is unsurprising, since $\mathcal{E}$ is ultimately a distance measured by the Fisher metric in the tangent space of ρ. This underlying reason is shared with characterizations of the Fisher as the Hessian of the loss surface (Martens, 2020).

Theorem 1 (Conjugate Natural Selection; Main Result). *Constrained to a given manifold of twice-differentiable parametric policies $\rho(h;\theta)$, FR-NGD of $\overline{\mathscr{L}}$ with respect to θ (Eqs. (6) and (7)) achieves the least-squares optimal fit in $\dot{\theta}$ to the continuous time replicator dynamics (i.e., Eq. (2)), as measured by the natural deviation $\mathcal{E}$ (Def. 8).*

Proof of Thm. 1. Lem. 9 implies that FR-NGD of $\overline{\mathscr{L}}$ (Def. 7) with respect to θ achieves a local extremum of $\mathcal{E}$ (i.e., $\frac{\partial}{\partial \dot{\theta}^i}\mathcal{E}(\dot{\theta}) = 0$), while convexity (Lem. 10) guarantees that any local extremum of $\mathcal{E}$ is a global minimum. □

A.4 Proof: The Price Equation and Preserved Dynamics

Thm. 2 characterizes the subspace of properties $u(h)$ that obey the Price equation under FR-NGD: linear combinations of the score. This result provides a direct route for determining how an approximation of the replicator dynamics by FR-NGD in θ for some chosen parameterization affects quantities of interest: if a property is naturally expressed as a linear combination of score, there is no resultant distortion of the dynamics of the property in question when using a lower-dimensional representation θ with FR-NGD when compared to replicator dynamics in $\mathcal{P}$.

We first restate the replicator equation (Eq. (2)) for local reference:

$$\dot{\rho}(h) = \rho(h)\Big[\overline{\mathscr{L}}_\rho - \mathscr{L}(h)\Big], \quad \text{where} \quad \overline{\mathscr{L}}_\rho \coloneqq \sum_h \rho(h)\mathscr{L}(h), \quad \sum_h \rho(h) = 1. \tag{2}$$

Next, we restate and prove Lem. 2, defining the price equation, before proving our characterization of the space of properties preserved by FR-NGD on $\mathcal{M}$ when compared to the replicator dynamics:

Lemma 2. (The Price Equation). *For any function or real-valued property of hypotheses $u\colon \mathcal{H} \to \mathbf{R}$, the expected value of u, denoted $\overline{u}_\rho$ when h is sampled with probability $\rho(h)$, evolves according to*

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{u}_\rho = -\operatorname*{Cov}_{H\sim\rho}\Big[u(H), \mathscr{L}(H)\Big] + \operatorname*{E}_{H\sim\rho}\Big[\dot{u}(H)\Big] \quad ; \quad \overline{u}_\rho \coloneqq \operatorname*{E}_{H\sim\rho}\Big[u(H)\Big]. \tag{3}$$

Proof of Lem. 2. By the chain rule,

$$\forall u, \quad \frac{\mathrm{d}}{\mathrm{d}t}\sum_h \rho(h)u(h) = \sum_h \dot{\rho}(h)u(h) + \sum_h \rho_h \dot{u}(h).$$

Expanding $\dot{\rho}(h)$ in terms of the replicator equation (Eq. (2)) and recognizing the terms of the equation as expectation values, we have that

$$\forall u, \quad \underbrace{\frac{\mathrm{d}}{\mathrm{d}t}\sum_h \rho(h)u(h)}_{\overline{u}_\rho} = \underbrace{\sum_h \rho(h)\Big[\overline{\mathscr{L}}_\rho - \mathscr{L}(h)\Big]u(h)}_{-\operatorname{Cov}_{H\sim\rho}[u(H),\mathscr{L}(H)]} + \underbrace{\sum_h \rho(h)\dot{u}(h)}_{\operatorname{E}_{H\sim\rho}[\dot{u}(H)]}.$$

□

Theorem 2 (Preserved Dynamics). *Linear combinations of score satisfy the Price equation (Eq. (3)) when θ is updated via FR-NGD of $\overline{\mathscr{L}}$. That is,*

$$\forall \alpha^i \in \mathbf{R}, u = \alpha^i s_i(\theta;h), \quad \frac{\mathrm{d}}{\mathrm{d}t}\operatorname*{E}_{H\sim\rho}\Big[u(H)\Big] = -\operatorname*{Cov}_{H\sim\rho}\Big[u(H), \mathscr{L}(H)\Big] + \operatorname*{E}_{H\sim\rho}\Big[\dot{u}(H)\Big]. \tag{12}$$

Proof of Thm. 2. Differentiate $\mathrm{E}[u] = \sum_h \rho(h)u(h)$ by the chain rule, where $\frac{\mathrm{d}\rho}{\mathrm{d}t} = \rho\frac{\mathrm{d}}{\mathrm{d}t}\log\rho$ implies that

$$\forall u, \quad \frac{\mathrm{d}}{\mathrm{d}t}\operatorname*{E}_{H\sim\rho}\Big[u(H)\Big] = \operatorname*{E}_{H\sim\rho}\Big[u(H)\frac{\mathrm{d}}{\mathrm{d}t}\log\rho(H)\Big] + \operatorname*{E}_{H\sim\rho}\Big[\dot{u}(H)\Big]. \tag{17}$$

Independently, note that Lem. 6 and Lem. 7 allow us to rewrite (Def. 7) as

$$\operatorname*{Cov}_{H\sim\rho}\Big[\partial_i \log \rho(H),\ \frac{\mathrm{d}}{\mathrm{d}t}\log \rho(H)\Big] = -\operatorname*{Cov}_{H\sim\rho}\Big[\partial_i \log \rho(H),\ \mathscr{L}(H)\Big]. \tag{18}$$

When $u(h) = \alpha^i s_i(\theta; h) = \alpha^i \partial_i \log \rho(h)$ for some vector $\alpha(t) \in \mathbf{R}^n$, we may take an α-weighted sum over Eq. (18). By Lem. 3, $E[u] = 0$, thus Eq. (18) becomes

$$E_{H\sim\rho}\Big[u(H)\frac{\mathrm{d}}{\mathrm{d}t}\log \rho(H)\Big] = -\operatorname*{Cov}_{H\sim\rho}\Big[u(H), \mathscr{L}(H)\Big].$$

This implies that the second term (first term on the right) in Eq. (17) and the second term (first term on the right) in Eq. (12) are equivalent, and, therefore, Eq. (17) implies Eq. (12). □

A.5 Proof: Continuous Inference

In this section, we give a derivation of the discrete-time replicator equation as background. We reference this derivation while also proving that continuous Bayesian inference (Eq. (14)) may be used to derive Bayes's rule (Thm. 3) and that FR-NGD provides an optimal approximation of continuous Bayesian inference (Thm. 4).

Let us first establish that the replicator dynamics preserve the normalization condition necessary for a proper probability distribution, restating the replicator equation (Eq. (2)) for local reference:

$$\dot{\rho}(h) = \rho(h)\Big[\overline{\mathscr{L}}_\rho - \mathscr{L}(h)\Big], \quad \text{where} \quad \overline{\mathscr{L}}_\rho \coloneqq \sum_h \rho(h)\mathscr{L}(h), \quad \sum_h \rho(h) = 1. \tag{2}$$

Lemma 11. (Preservation of Normalization). *The dynamics of the continuous time replicator equation preserve the normalization of ρ (i.e., $\sum_h \rho(h) = 1$). That is, $\frac{\mathrm{d}}{\mathrm{d}t}\sum_h \rho(h) = 0$.*

Proof.

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum_h \rho(h) = \sum_h \dot{\rho}(h) = \sum_h \rho(h)\left[\overline{\mathscr{L}}_\rho - \mathscr{L}(h)\right] = \overline{\mathscr{L}}_\rho - \sum_h \rho(h)\mathscr{L}(h) = 0.$$

□

The replicator equation is frequently encountered in discrete time.

Lemma 12. (The Discrete-Time Replicator Equation). *Define*

$$\log r_t(h) = -\frac{1}{\Delta t}\int_t^{t+\Delta t} \mathscr{L}_{t'}(h)\,\mathrm{d}t',$$

for each h, as the time-average of $-\mathscr{L}_{t'}(h)$ over $[t, t+\Delta t)$, and let

$$\widetilde{r}_t(\Delta t) \coloneqq \sum_h \rho_t(h) r_t(h)^{\Delta t}.$$

It follows that

$$\rho_{(t+\Delta t)}(h) = \rho_t(h)\frac{r_t(h)^{\Delta t}}{\widetilde{r}_t(\Delta t)}. \tag{19}$$

Proof. The solution of Eq. (2) (which may be verified by differentiating with respect to time) is

$$\rho_{(t+\Delta t)}(h) = \rho_t(h)\underbrace{\Big(\exp\int_t^{(t+\Delta t)} \overline{\mathscr{L}}_{\rho_{t'}}\,\mathrm{d}t'\Big)}_{C_{t,\Delta t}}\underbrace{\Big(\exp\int_t^{(t+\Delta t)} -\mathscr{L}_{t'}(h)\,\mathrm{d}t'\Big)}_{r_t(h)^{\Delta t}}.$$

After summing over h on both sides of this equation, normalization (Lem. 11) implies that the constant $C_{t,\Delta t}$ is necessarily equal to the multiplicative inverse of $\widetilde{r}_t(\Delta t) \coloneqq \sum_h \rho_t(h) r_t(h)^{\Delta t}$. □

Our formulation of "continuous Bayesian inference" in the main text defines, for local reference,

$$h(x_t;t) = \Pr_h(X_t{=}x_t\,|\,t) \quad ; \quad \mathfrak{n}(x_t;t) = \Pr_{\mathfrak{n}}(X_t{=}x_t\,|\,t) \quad ; \quad \mathscr{L}(h,t) = -\log h(x_t;t). \tag{13}$$

such that

$$\dot{\rho}_t(h,x_t) = \rho_t(h)\Big[\overline{\mathscr{L}}_{\rho_t}(x_t) + \log h(x_t;t)\Big], \quad \text{where} \quad \overline{\mathscr{L}}_{\rho_t}(x_t) = -\sum_h \rho_t(h)\log h(x_t;t). \tag{14}$$

Theorem 3 (Continuous Inference)**.** *Eq. (14) may be used to derive Bayes's rule.*

Proof of Thm. 3. To discretize Eq. (14), we first denote the *path* of observations over the time interval from t up to $t+\Delta t$ as $x_t^{\Delta t} \coloneqq \{x_{t'} : t' \in [t, t+\Delta t)\}$. Next, define the probability density of the path to be proportional to the product of the probabilities of its instantaneous values.

$$\log h(x_t^{\Delta t};t) \coloneqq \frac{1}{[\mathrm{t}]}\int_t^{t+\Delta t} \log h(x_{t'};t')\,\mathrm{d}t'\,. \tag{20}$$

Note that we normalize this equation to make it properly dimensionless by choice of an arbitrary scale, where [t] denotes units of time. The choice of an arbitrary scale is integral to the definition of differential entropy, as it allows us to establish a volume of configuration space (in this case, with units of time) to correspond to unit entropy. For a motivating example, we must choose how many units of (differential) entropy correspond to the space of possible paths over 1s, when each X_t is a uniformly distributed Bernoulli random variable. We choose the same units that we use to measure Δt, so that [t] may be considered equal to 1 hereafter.

Subject to the loss of Eq. (13), when $\mathscr{L}(h,t) = -\log h(X_t^{\Delta t}, t)$, retracing the derivation of the discrete-time replicator equation (Lem. 12) yields Bayes's rule, i.e.,

$$\rho_{(t+\Delta t)}(h\,|\,X_t^{\Delta t}) = \rho_t(h)\frac{h(X_t^{\Delta t},t)}{\Pr_{\rho_t}(X_t^{\Delta t})}, \quad \text{from} \quad \rho_{(t+\Delta t)}(h) = \rho_t(h)\frac{r_t(h)^{\Delta t}}{\widetilde{r}_t(\Delta t)}, \tag{21}$$

where

$$r_t(h)^{\Delta t} = \exp\int_t^{(t+\Delta t)} -\mathscr{L}_{t'}(h)\,\mathrm{d}t' = h(X_t^{\Delta t},t)$$

and

$$\widetilde{r}_t(\Delta t) \coloneqq \sum_h \rho_t(h) r_t(h)^{\Delta t} = \Pr_{\rho_t}(X_t^{\Delta t}).$$

We identify $\rho_{t+\Delta(t)}(h)$ as the posterior $\rho_{t+\Delta(t)}(h\,|\,X_t^{\Delta t})$ when path $X_t^{\Delta t}$ is observed. □

Having established that the replicator equation with a specific loss based on *surprisal* (Eq. (13)) may be identified with "continuous Bayesian inference", we next prove that continuous Bayesian inference is optimally approximated by FR-NGD (Thm. 4). Let us first restate the definition of Kullback-Leibler divergence of $\mathfrak{n}$ from h, to which we relate the gradient of the surprisal-based loss:

$$\mathscr{D}_t(\mathfrak{n} \parallel h) \coloneqq -\sum_x \mathfrak{n}(x;t)\log\frac{h(x;t)}{\mathfrak{n}(x;t)}. \tag{15}$$

Lemma 4. *The gradients of* $\mathrm{E}_{X_t\sim\mathfrak{n}}[\overline{\mathscr{L}}_{\rho_t}(X_t)]$ *and* $\mathrm{E}_{H\sim\rho_t}[\mathscr{D}_t(\mathfrak{n} \parallel H)]$ *with respect to* ρ_t *are equal.*

Proof of Lem. 4.

$$\begin{aligned}
&\mathop{\mathrm{E}}_{H\sim\rho_t}[\mathscr{D}(H,t)] - \mathop{\mathrm{E}}_{X_t\sim\mathfrak{n}}[\overline{\mathscr{L}}_{\rho_t}(X_t)] \\
&\quad= -\sum_{x,h}\rho_t(h)\mathfrak{n}(x;t)\log\frac{h(x;t)}{\mathfrak{n}(x;t)} + \sum_{x,h}\rho_t(h)\mathfrak{n}(x;t)\log h(x_t;t) \\
&\quad= \sum_{x,h}\rho_t(h)\mathfrak{n}(x,t)\log\mathfrak{n}(x,t) \\
&\quad= \sum_x \mathfrak{n}(x,t)\log\mathfrak{n}(x,t).
\end{aligned}$$

As the negative entropy of $X_t \sim \mathfrak{n}$, this difference is independent of ρ_t and therefore has zero gradient with respect to ρ_t. Each term of the original expression must therefore have the same gradient. □

We conclude with a restatement and proof of the optimal correspondence of FR-NGD with the appropriate loss to continuous Bayesian inference:

Theorem 4 (FR-NGD Yields Optimal Continuous Inference)**.** *For any probability distribution $\rho(h;\theta)$ that is twice-differentiable with respect to parameters θ, FR-NGD of the expected divergence $\mathrm{E}_{H\sim\rho_t}[\mathscr{D}_t(\mathfrak{n} \parallel H)]$ (of the ρ-weighted predictions of model $h(X_t;t)$ for $X_t \sim \mathfrak{n}$) with respect to θ optimally approximates Bayesian inference for ρ in continuous time, by minimizing $\mathcal{E}$ (Def. 8).*

Proof of Thm. 4. Assume a twice-differentiable parameterization for probability distribution $\rho(h;\theta)$. FR-NGD of $\mathrm{E}_{H\sim\rho}[\mathscr{D}(H,t)]$ (Eq. (15)) is the same as FR-NGD of $\mathrm{E}_{X_t\sim\mathfrak{n}}[\overline{\mathscr{L}}_{\rho_t}(X_t)]$ (Eq. (14)), since these quantities have equivalent gradients (Lem. 4). We will treat the latter.

By Thm. 1, FR-NGD of $\mathrm{E}_{X_t\sim\mathfrak{n}}[\overline{\mathscr{L}}_{\rho_t}(X_t)]$ with respect to θ provides an optimal approximation of the replicator dynamics (Eq. (2)) with the corresponding stochastic loss $\mathscr{L}(h) = -\log h(x_t;t)$.

Finally, because Thm. 3 indicates that the replicator dynamics are consistent with continuous time Bayesian inference, it follows that FR-NGD of $\mathrm{E}_{H\sim\rho_t}[\mathscr{D}(H),t]$ with respect to θ is an optimal approximation of Bayesian inference in continuous time. □