Conjugate Natural Selection: Fisher-Rao Natural Gradient Descent Optimally Approximates Evolutionary Dynamics and Continuous Bayesian Inference

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

March 10, 2023

ABSTRACT

Rather than refining individual candidate solutions for a general non-convex optimization problem, by analogy to evolution, we consider minimizing the average loss for a parametric distribution over hypotheses. In this setting, we prove that Fisher-Rao natural gradient descent (FR-NGD) optimally approximates the continuous-time replicator equation (an essential model of evolutionary dynamics) by minimizing the mean-squared error for the relative *fitness* of competing hypotheses. We term this finding *conjugate natural selection* and demonstrate its utility by numerically solving an example non-convex optimization problem over a continuous strategy space. Next, by developing known connections between discrete-time replicator dynamics and Bayes's rule, we show that when absolute fitness corresponds to the negative KL-divergence of a hypothesis's predictions from actual observations, FR-NGD provides the optimal approximation of *continuous Bayesian inference*. We use this result to demonstrate a novel method for estimating the parameters of stochastic processes.

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 the distribution of traits, strategies, or *hypotheses* in a population 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. Formal, analytic models of evolution such as the *replicator equation* are indispensable to biology (Sinervo and Calsbeek, 2006; Queller, 2017). In this equation, given below, the absolute fitness 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 $\rho(h) \in [0, 1]$.

Throughout this paper, we identify high fitness with minimal *loss* $\mathscr{L}$, such that evolutionary dynamics selects hypotheses with lower than average loss. In continuous time, the **replicator equation** is

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

For probability distributions over hypothesis space $\mathcal{H}$, this equation induces *replicator dynamics*.

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.

Why should we want analytic evolutionary dynamics for 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 and corresponds 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: In general, the **space requirements** for storing and manipulating arbitrary probability distributions $\rho \in \Delta(\mathcal{H})$, where $\Delta(\mathcal{H})$ denotes the space of probability distributions over $\mathcal{H}$, grow proportional to the size of $\mathcal{H}$ (i.e., asymptotically, as $\Theta(|\mathcal{H}|)$), which may be **continuous and high-dimensional**. The replicator equation describes dynamics in $\mathcal{P} \coloneqq \Delta(\mathcal{H})$.

When a (meta)hypothesis space (e.g., $\mathcal{P}$) is large, it is standard in multiple fields to **parameterize** a subspace or **manifold** $\mathcal{M}(\theta) \subseteq \mathcal{P}$ of tractable solutions $\rho(h;\theta) \in \mathcal{M}(\theta)$, 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 $\mathcal{M}(\theta)$, demanding values of ρ that are not addressable by any θ. To resolve this, we require a model of evolutionary dynamics that is closed on any manifold $\mathcal{M}(\theta)$ — 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** model of **continuous evolutionary dynamics constrained to any manifold** $\mathcal{M}(\theta)$, which we refer to 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 an identification of average loss (i.e., negative fitness) for the distribution $\rho(h;\theta)$ with the divergence of predictions from observations implies that FR-NGD provides an optimal approximation of Bayesian inference for any parametric distribution over hypotheses (Section 4). We demonstrate applications of these findings to a non-convex optimization problem and parameter estimation of a stochastic process.

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 or

Bayesian inference *for all twice-differentiable parameterization schemes* $\rho(h;\theta)$. Additionally, we believe our specific construction of continuous Bayesian inference is novel.

While the exact correspondence between the replicator equation and FR-NGD has been previously identified for tabular or Boltzmann-Gibbs (i.e., softmax) parameterized distributions (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 $\mathcal{M}(\theta) = \Delta(\mathcal{H})$. Notably, Harper (2009b,a) recognizes the replicator equation as both an instance of FR-NGD and as an inference dynamic guided by Fisher information geometry, but does not consider under-parameterized approximations $\mathcal{M}(\theta) \subset \Delta(\mathcal{H})$. While noting connections to Bayes's rule, the cited work stops short of identifying the continuous-time replicator equation as a *generalization* of Bayes's rule to continuous time or identifying 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 very recent work by Nurbekyan et al. (2022), who highlight that natural gradient descent with metric $g(\theta)$ yields the least-squares optimal approximation in $\mathcal{M}(\theta)$ to natural gradient descent with metric $g(\rho)$ on $\mathcal{M}(\rho)$ (Nurbekyan et al., 2022, Eq. 2.2). Alternately stated, 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.

2 Preliminaries

Before detailing our results, we first review necessary background, establishing our setting and notation in Section 2.1. We again introduce the replicator equation in Section 2.2 and use it to derive its discrete form, which we relate to Bayes's rule, and the Price equation. In Section 2.3, we provide essential results from information geometry.

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 $\Delta(\mathcal{H})$ denote the simplex, or space of probability distributions, over $\mathcal{H}$.

We use $\rho \in \Delta(\mathcal{H})$ 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, we use a dot above a symbol to denote its *full* time derivative (that is, $\forall f, \dot{f} \equiv \frac{\mathrm{d}}{\mathrm{d}t} f$) and a bar over a variable to denote its expectation value (explicitly, $\forall f, \overline{f} \equiv \mathrm{E}[f]$). 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 that are repeated twice in a single term (formally, $\forall f, g, f_i g^i \equiv \sum_i f_i g^i$). We also use standard shorthands for partial derivatives, identifying $\partial_i(\cdot) \equiv \frac{\partial}{\partial \theta^i}(\cdot)$ and $\partial_t(\cdot) \equiv \frac{\partial}{\partial t}(\cdot)$. When discussing matrices, we write $M \succeq 0$ to indicate that a matrix M is positive semi-definite (has only non-negative eigenvalues). Finally, we may omit explicit functional dependencies between variables when they are implied or understood.

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 $\overline{\mathscr{L}}(\rho)$.

$$\min_{\rho} \mathop{\mathrm{E}}_{H \sim \rho} \Big[\mathscr{L}(H) \Big]. \tag{1}$$

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-differentiably parameterized manifold $\mathcal{M}(\theta) \subset \mathcal{P}$ of distributions $\rho(h;\theta) \in \mathcal{M}(\theta)$ 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

Replicator dynamics are described by the replicator equation. As background, we restate the continuous-time replicator equation, from which we derive the Price equation. While our treatment throughout this paper assumes a continuous time variable t, we show how to derive the discrete-time form of the replicator equation, taking care to explicitly consider time intervals of the form $[t, t + \Delta t)$.

2.2.1 The Replicator Equation

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

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

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

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

Lemma 1. (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. $\frac{\mathrm{d}}{\mathrm{d}t}\sum_h \rho(h) = \sum_h \dot{\rho}(h) = \sum_h \rho(h)\left[\overline{\mathscr{L}} - \mathscr{L}(h)\right] = \overline{\mathscr{L}} - \sum_h \mathscr{L}(h) = 0.$ □

Remark 2. (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 $\rho_t(h) = \rho_0(h)\exp\int_0^t \left(\overline{\mathscr{L}}_{t'} - \mathscr{L}_{t'}(h)\right)\mathrm{d}t'$, which does not admit roots in finite time unless ρ_0, and therefore ρ_t, for all t, is zero. □

Rem. 2 reveals that the replicator equation cannot generate new hypotheses on its own. For this reason, it is often combined with mutation or diffusion terms in practice, but these terms make the resulting dynamics more difficult to solve (Bloembergen et al., 2015). As an approximation of replicator dynamics, FR-NGD **can avoid this issue** with underparameterized $\rho(\theta)$, since eliminated hypotheses can be reintroduced or provably never be eliminated (e.g., when $\rho(h;\theta)$ is everywhere non-zero by design).

2.2.2 The Discrete-Time Replicator Equation

The replicator equation is frequently encountered in discrete time.

Lemma 3. (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*

$$\rho_{(t+\Delta t)}(h) = \rho_t(h)\frac{r_t(h)^{\Delta t}}{\widetilde{r}_t(\Delta t)}. \tag{3}$$

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

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

After summing over h on both sides of this equation, normalization (Lem. 1) implies that the constant $C_{t,\Delta t}$ is necessarily equal to the multiplicative inverse of $\widetilde{r}_t(\Delta t) \coloneqq \sum_h \rho_t(h) r_t(h)^{\Delta t}$. □

We explore connections between Eq. (3) and Bayes's rule in Section 4.

2.2.3 The Price Equation

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), which may be derived from the replicator equation.

Lemma 4. (The Price Equation). *For any function $f : \mathcal{H} \to \mathbf{R}$, the expected value of f, when h is sampled with probability $\rho(h)$, is*

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathop{\mathrm{E}}_{H \sim \rho} \Big[f(H) \Big] = - \mathop{\mathrm{Cov}}_{H \sim \rho} \Big[f(H), \mathscr{L}(H) \Big] + \mathop{\mathrm{E}}_{H \sim \rho} \Big[\dot{f}(H) \Big]. \tag{4}$$

Proof. By the chain rule, $\forall f, \frac{\mathrm{d}}{\mathrm{d}t} \sum_h \rho(h) f(h) = \sum_h \dot{\rho}(h) f(h) + \sum_h \rho_h \dot{f}(h)$. When we expand $\dot{\rho}(h)$ in terms of the replicator equation, this implies that $\forall f, \frac{\mathrm{d}}{\mathrm{d}t} \sum_h \rho(h) f(h) = \sum_h \rho(h) \left(\overline{\mathscr{L}} - \mathscr{L}(h) \right) f(h) + \sum_h \rho(h) \dot{f}(s)$. The first and last terms appearing in this equation are expectation values, while the middle term the negative covariance of $\mathscr{L}(h)$ and $f(h)$. □

Thm. 2 characterizes the subspace of properties $f(h)$ that obey the Price equation under FR-NGD.

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 parallel approximations (Fradkov, 2020; Tappert). In broad strokes, the technique is to first differentiably parameterize the search space in terms of θ, for example, as $\rho(h; \theta)$, then repeatedly update θ in a direction that approximates the "fastest" decreasing value (i.e., the negative gradient) of the expected loss $\overline{\mathscr{L}}$. For continuous time,

Definition 2. *Naive Gradient Descent* is given by the update rule

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

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 see 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 $\dim(\partial_i \overline{\mathscr{L}}) = \mathsf{L}\mathsf{U}^{-1}$. Even when a learning rate provides a natural conversion between L and T, the powers of U do not balance on each side of Eq. (5), and the equation is dimensionally invalid. This problem is resolved by explicit consideration of a metric g, which assigns distances and angles in the "tangent" space of θ (i.e., where $\mathrm{d}\theta^i$, $\mathrm{d}\theta^j$, etc. live), with units U^2.

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

$$\dot{\theta}^i = -g^{ij} \partial_j \overline{\mathscr{L}}. \tag{6}$$

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 tangent space of θ.

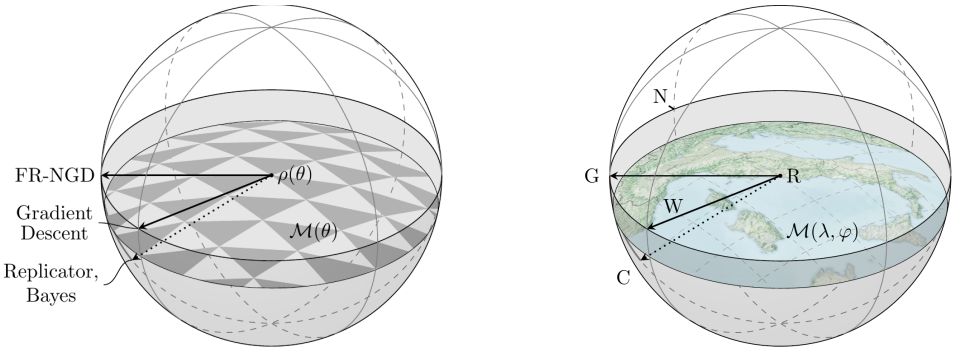

Figure 1: An analogy for how different metrics can suggest different parameter updates in Eq. (6). Earth's curvature is exaggerated to emphasize that vectors W and G are tangent to its surface. The direction of travel from Rome (point R) that most rapidly decreases one's distance from Chicago, when measured 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.

In Fig. 1, we give an analogy for how the choice of metric can affect covariant gradient decent: Many world maps will suggest that the shortest path (i.e., a straight line on the map) from Rome to Chicago on Earth's surface is due west, narrowly missing Barcelona, since all three cities are at nearly the same latitude. Wrap a string around a *globe*, however, and the shortest path is actually closer to Paris (which lies north of Barcelona by over 800 km). Both a flat map and a globe "warp" our perception of the space around Rome to suggest a direction that decreases our distance to Chicago the fastest. In physical space, the optimal direction is actually downward, indicating a straight line through the Earth: this direction in three dimensions is more closely approximated by the path suggested by the globe than the map.

FR-NGD, that is, natural gradient descent with respect to the Fisher-Rao metric, uses a specific choice of g for Eq. (6) 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 the *Fisher*.

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 $F_{ij}(\theta) \coloneqq \mathrm{Cov}_{H\sim\rho}\left[\partial_i \log \rho(H;\theta),\ \partial_j \log \rho(H;\theta)\right]$.

As a covariance matrix, F is symmetric and positive semi-definite. F fails to be fully positive definite (i.e., is degenerate) when parameter updates in different directions (up to constant scaling) affect ρ identically.

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 5. (Zero Expected Score). *The score is the zero vector, in expectation. That is,* $\mathrm{E}_{H\sim\rho}\left[\partial_i \log \rho(H)\right] = 0$.

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

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

$$\dot{\theta}^i = -\left[F^{+}(\theta)\right]^{ij} \partial_j \overline{\mathscr{L}}, \tag{7}$$

where F is the Fisher and F^+ is its Moore-Penrose inverse. While Eq. (7) 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 $\rho \in \mathcal{M}(\theta)$, induced by FR-NGD of $\overline{\mathscr{L}}$ with respect to θ is

$$F_{ij}(\theta)\dot{\theta}^j = -\partial_i \overline{\mathscr{L}}. \tag{8}$$

$F(\theta)$ is positive definite and invertible when θ has only non-degenerate degrees of freedom, in which case $F^+(\theta) = F^{-1}(\theta)$, the nullspace of $F^+(\theta)$ is orthogonal to $\mathcal{M}(\theta)$, and Eq. (8) and Eq. (7) are equivalent. When $F(\theta)$ 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. (7) solves Eq. (8), 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. (8). We show that this conjugate gradient flow is an optimal approximation of the replicator equation (Eq. (2)).

3 Conjugate Natural Selection

In addition to proving an optimal correspondence between natural gradient descent and the continuous-time replicator equation (Thm. 1), we characterize the space of properties $f(\rho)$ that precisely obeys the Price equation, Eq. (4), in terms of score (Def. 5) in Thm. 2. Intuitively, for given parameterization $\mathcal{M}(\theta)$, these are the properties of ρ that precisely obey the replicator dynamics under FR-NGD with respect to θ. Finally, we demonstrate an application of conjugate natural selection (CNS) by experimentally evolving a distribution of continuous hypotheses for a non-convex problem.

Lemma 6. (Functionals). *For any function $f : \mathcal{H} \to \mathbf{R}$, for all parameter components i, the following expressions are equivalent:* $\sum_h \partial_i \rho(h) f(h) = \mathrm{E}_{H\sim\rho}\left[f(H)\partial_i \log \rho(H)\right] = \mathrm{Cov}_{H\sim\rho}\left[\partial_i \log \rho(H), f(H)\right]$.

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

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

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

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. The second follows from rewriting covariance as an explicit sum and pushing the time derivative through the logarithm. □

Lemma 8. (Loss Gradient) *The gradient of expected loss $\overline{\mathscr{L}}$ with respect to θ is*

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

Proof. We may expand $\partial_i \overline{\mathscr{L}} = \sum_h \mathscr{L}(h)\partial_i \rho(h)$, then use Lem. 6 to equate this sum with the desired covariance between $\partial_i \log \rho$ and $\overline{\mathscr{L}}$. We then expand this covariance as an explicit sum. □

Lem. 7 and Lem. 8 allow us to rewrite the conjugate gradient flow (Def. 7) as:

$$\sum_h \Big[\partial_i \log \rho(h)\Big]\dot{\rho}(h) = \sum_h \Big[\partial_i \log \rho(h)\Big]\rho(h)\Big(\overline{\mathscr{L}} - \mathscr{L}(h)\Big). \tag{9}$$

Eq. (9) may be understood as a projection of the replicator dynamics (i.e., FR-NGD unconstrained to $\mathcal{M}(\theta)$) onto the dual space of θ spanned by the score $s(\theta; h)$ (Def. 5) (compare with Nurbekyan et al., 2022, Eqn. 2.3). That is, the score provides a basis for the tangent space of $\mathcal{M}(\theta)$ imbued with the Fisher metric. Intuitively, note that the score describes

a function over h with expectation zero (Lem. 5), so each s_i can be interpreted in terms of a marginal addition to ρ that will still satisfy the normalization condition. Eq. (9) also indicates that such a sum will also follow the manifold: $\rho + s_i \, \mathrm{d}\theta^i \in \mathcal{M}(\theta)$. To reinforce this understanding, Thm. 2 indicates that any function in the score basis defines a property of ρ that obeys the Price equation. First, we prove that the conjugate gradient flow of FR-NGD achieves the natural least-squares fit to the replicator equation.

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

$$\mathcal{E}(\dot{\theta}) \coloneqq \frac{1}{2} \mathop{\mathrm{E}}_{H \sim \rho} \left[\left(\frac{\mathrm{d}}{\mathrm{d}t} \log \rho(H) - \left(\overline{\mathscr{L}} - \mathscr{L}(H) \right) \right)^2 \right]. \tag{10}$$

This quantity corresponds to the distance in the tangent space of $\Delta(\mathcal{H}) = \mathcal{P}$, imposed by the Fisher metric, between the replicator dynamics and realized dynamics $\dot{\rho}$ (compare with Nurbekyan et al., 2022, Eq. 2.15). By inspection, we see that $\mathcal{E}$ is minimized for tabular settings ($\rho \equiv \theta$) if and only if the replicator equation holds: $\dot{\rho} = \rho(\overline{\mathscr{L}} - \mathscr{L}(h))$.

Lemma 9. (Cancel Dots with Chain Rule). $\forall f$ *independent of* $(\dot{\theta}, \dot{\rho})$, $\frac{\partial}{\partial \dot{\theta}^i} \left(\frac{\mathrm{d}}{\mathrm{d}t} f(p) \right) = \partial_i f(p)$.

Proof. By the chain rule,

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

□

Lemma 10. *(Gradient of Natural Deviation)* $\frac{\partial}{\partial \dot{\theta}^i} \mathcal{E}(\dot{\theta}) = F_{ij}(\theta) \dot{\theta}^j + \partial_i \overline{\mathscr{L}}$.

Proof. We may take the gradient of Def. 8 with respect to $\dot{\theta}$, using Lem. 9 (for $f = \log$):

$$\frac{\partial}{\partial \dot{\theta}^i} \mathcal{E}(\dot{\theta}) = \mathop{\mathrm{E}}_{H \sim \rho} \left[\left(\frac{\mathrm{d}}{\mathrm{d}t} \log \rho(H) - \left(\overline{\mathscr{L}} - \mathscr{L}(H) \right) \right) \partial_i \log \rho(H) \right]. \tag{11}$$

Recognizing that $\frac{\mathrm{d}}{\mathrm{d}t} \log \rho = \frac{1}{\rho} \dot{\rho}$, the expectation value on the right side of Eq. (11) 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 of $\mathcal{M}(\theta)$ (Lem. 7) and the loss gradient (Lem. 8), respectively. □

We use Lem. 10 to prove that FR-NGD of $\overline{\mathscr{L}}$ (Def. 7) with respect to θ achieves a local extremum of $\mathcal{E}$ (Thm. 1).

Lemma 11. (Hessian of Natural Deviation) $\mathcal{E}$ *is convex in* $\dot{\theta}$*, that is,* $\frac{\partial^2}{\partial \dot{\theta}^i \partial \dot{\theta}^j} \mathcal{E}(\dot{\theta}) = F_{ij}(\theta) \succeq 0$.

Proof. We differentiate Eq. (11) with respect to $\dot{\theta}$ again, using Lem. 9 for ($f = \log$). Thus, the second derivative of $\mathcal{E}$ is $\frac{\partial^2}{\partial \dot{\theta}^i \partial \dot{\theta}^j} \mathcal{E}(\dot{\theta}) = \mathrm{E}_{H \sim \rho} \left[(\, \partial_i \log \rho(H) \,)(\partial_j \log \rho(H)) \right] = F(\theta)$. As a covariance matrix, $F \succeq 0$. □

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)**.** *Constrained to a given manifold of twice-differentiably parameterized policies* $\rho(h; \theta)$*, FR-NGD of* $\overline{\mathscr{L}}$ *with respect to* θ *achieves the least-squares optimal fit in* $\dot{\theta}$ *to the continuous-time replicator dynamics (i.e., Eq. (2)), as measured by* $\mathcal{E}$ *(Def. 8).*

Proof. Lem. 10 implies that FR-NGD of $\overline{\mathscr{L}}$ (Definition 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. 11, guarantees that any local extremum of $\mathcal{E}$ is a global minimum. □

3.1 The Price Equation

We now renew our discussion regarding score as a family of basis vectors for the tangent space of $\mathcal{M}(\theta)$.

Theorem 2. *Linear combinations of score satisfy the Price equation (Eq. (4)) when θ is updated via to FR-NGD of $\overline{\mathscr{L}}$. That is,*

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

Proof. We differentiate $\mathrm{E}[f] = \sum_h \rho(h) f(h)$ by the chain rule, where $\frac{\mathrm{d}\rho}{\mathrm{d}t} = \rho \frac{\mathrm{d}}{\mathrm{d}t} \log \rho$ implies that

$$\forall f, \quad \frac{\mathrm{d}}{\mathrm{d}t} \mathop{\mathrm{E}}_{H\sim\rho}\Big[f(H)\Big] = \mathop{\mathrm{E}}_{H\sim\rho}\Big[f(H) \frac{\mathrm{d}}{\mathrm{d}t} \log \rho(H)\Big] + \mathop{\mathrm{E}}_{H\sim\rho}\Big[\dot{f}(H)\Big]. \tag{13}$$

Independently, note that Lem. 7 and Lem. 8 allow us to rewrite (Def. 7) as

$$\mathop{\mathrm{Cov}}_{H\sim\rho}\Big[\partial_i \log \rho(H),\ \frac{\mathrm{d}}{\mathrm{d}t} \log \rho(H)\Big] = -\mathop{\mathrm{Cov}}_{H\sim\rho}\Big[\partial_i \log \rho(H),\ \mathscr{L}(H)\Big]. \tag{14}$$

When $f(h) = \alpha^i s_i(\theta; h)$ for some vector $\alpha(t) \in \mathbf{R}^n$, we may take an α-weighted sum over Eq. (14) (recalling by Lem. 5 that $E[f] = 0$) so that the left-hand side may be written as an expectation value. Eq. (14) thus becomes $E_{H\sim\rho}[f(H), \frac{\mathrm{d}}{\mathrm{d}t} \log \rho(H)] = -\operatorname{Cov}_{H\sim\rho}[f(H), \mathscr{L}(H)]$. This implies that the second term in Eq. (13) and the second term in Eq. (12) are equivalent, and, therefore, Eq. (13) implies Eq. (12). □

3.2 Applications

We demonstrate an application of CNS to evolve 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 the analytically-known value of the Fisher for the Gaussian distribution, an Euler discretization of the dynamics, and a non-degenerate parameterization of a 2-dimensional Gaussian distribution using 5 degrees of freedom. We provide our code for this simulation at `https://git.ucsc.edu/rraab/conjugate-natural-selection`.

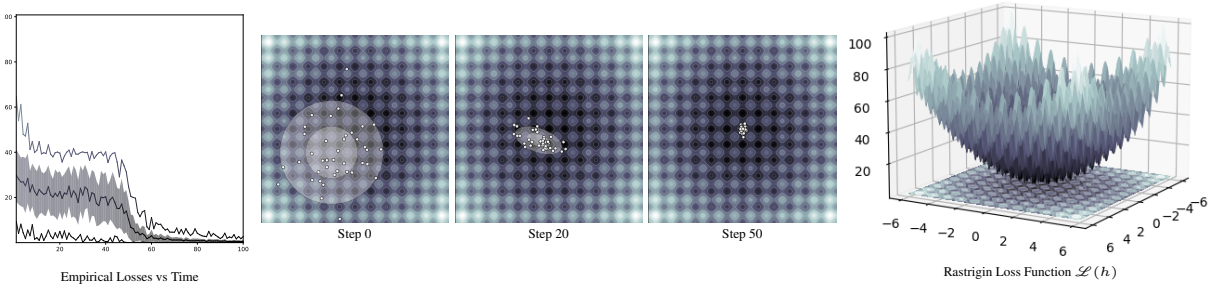

Figure 2: On the left, we plot the mean, standard deviation, and extremal empirical losses for the learned distribution over 100 time steps. On the right, the loss function is visualized as a surface over the domain $[-6, 6] \times [-6, 6]$. In the middle, we represent time steps 0, 20, and 50 of the 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 $1\mathrm{e}^{-3}$ for the Euler update. An animation of the time-evolution of the distribution is available with the linked source code.

Recent characterizations of FR-NGD in 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 $\mathcal{M}(\theta)$ 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.). A more scalable approach, based on solving the convex problem of minimizing $\mathcal{E}(\dot{\theta})$ 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 strategy space is continuous or high-dimensional. As $\overline{\mathscr{L}}$ and $F(\theta)$ 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. Finally, we assert that FR-NGD for a distribution over candidate solutions 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

The replicator equation (Eq. (2)) corresponds to Bayesian inference in continuous time when we perform a change of variables and identify loss with the negative log-likelihood of hypothesis h given observed value x_t. That is, let

$$\mathscr{L}(h,t) = -\log m(x_t \,|\, h;t), \tag{15}$$

where m is a *model* for $\Pr(X_t{=}x_t \,|\, h;t)$. The loss expressed in Eq. (15) 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 should predict observations well and minimize average surprisal. The observable X_t, distributed without reference to m or ρ by Nature as $X_t \sim n_t$, represents a measurement that varies *stochastically* in continuous time: Examples include physical quantities like instantaneous field amplitude; the idealized market price of an asset; Brownian motion; or any continuous-time quantity perturbed by *noise*. For loss given by Eq. (15), the replicator equation Eq. (2) describes stochastic dynamics for ρ that depend on the value of x_t:

$$\dot{\rho}_t(h,x_t) = \rho_t(h)\Big[\overline{\mathscr{L}}(x_t) + \log m\big(x_t \,|\, h;t\big)\Big], \quad \text{for} \quad \overline{\mathscr{L}}(x_t) \coloneqq -\sum_h \rho_t(h)\log m(x_t \,|\, h;t). \tag{16}$$

We will use the gradient of the expected value of $\overline{\mathscr{L}}(x_t)$ to perform FR-NGD. While the expectation value of $\overline{\mathscr{L}}(x_t)$ over $X_t \sim n_t$ might itself be formally defined as a conditional cross-entropy, we compare its *gradient* to the gradient of a Kullback-Leibler divergence (cross-entropy) conditioned on h, which we denote as $\mathscr{D}_n^m(h)$:

$$\mathscr{D}_n^m(h) \coloneqq D_{(n_t \| m)}(X_t \mid h) = -\sum_x n_t(x)\log\frac{m(x \,|\, h;t)}{n_t(x)} \tag{17}$$

Lemma 12. *The gradients of* $\mathrm{E}_{H\sim\rho}[\mathscr{D}_n^m(H)]$ *and* $\mathrm{E}_{X_t\sim n_t}[\overline{\mathscr{L}}(X_t)]$ *are equivalent with respect to* ρ.

Proof. $\mathrm{E}_{H\sim\rho}[\mathscr{D}_n^m(H)] - \mathrm{E}_{X_t\sim n}[\overline{\mathscr{L}}(X_t)] = \sum_{x,h}\rho(h)n_t(x)\log n_t(x) = \sum_x n_t(x)\log n_t(x)$, which, as the negative entropy of $X_t \sim n_t$, is independent of ρ. □

The stochastic dynamics of Eq. (16) subject to $X_t \sim n_t$ admit an analog of the Price equation. Recent work has derived this equation, termed an "observable flow" Berman et al. (2022, Eq. 2.25), from Bayesian updates in the limit of continuous time, whereas we begin with the continuous-time replicator equation for log-likelihood loss and derive Bayes's rule (Thm. 3).

Theorem 3 (Continuous Inference)**.** *Eq. (16) may be used to derive Bayes's rule.*

Proof. To discretize Eq. (16), we first denote the *path* of observations over the time interval from t up to $t + \Delta t$ as $x_t^{\Delta t} := \{x_{t'} : t' \in [t, t + \Delta t)\}$. Next, define the probability density of the path to be proportional to the product of the probabilities of its instantaneous values.

$$\log m(x_t^{\Delta t} \,|\, h; t) := \frac{1}{[\mathrm{t}]} \int_t^{t+\Delta t} \log m(x_{t'} \,|\, h; t') \,\mathrm{d}t' \,. \tag{18}$$

Note that we normalize this equation to make it properly dimensionless by choice of an arbitrary scale, where $[\mathrm{t}]$ 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 $[\mathrm{t}]$ may be considered equal to 1 hereafter.

Subject to the loss of Eq. (15), when $\mathscr{L}(h, t) = -\log m(X_t^{\Delta t} \mid h, t)$, retracing the derivation of the discrete-time replicator equation (Lem. 3) yields

$$\rho_{(t+\Delta t)}(h \,|\, X_t^{\Delta t}) = \rho_t(h) \frac{m(X_t^{\Delta t} \,|\, h, t)}{\Pr_{m,\rho_t}(X_t^{\Delta t})}, \quad \text{from} \quad \rho_{(t+\Delta t)}(h) = \rho_t(h) \frac{r_t(h)^{\Delta t}}{\widetilde{r}_t(\Delta t)}, \tag{19}$$

where $r_t(h)^{\Delta t} = \exp \int_t^{(t+\Delta t)} -\mathscr{L}_{t'}(h) \,\mathrm{d}t' = m(X_t^{\Delta t} \,|\, h, t)$ and $\widetilde{r}_t(\Delta t) := \sum_h \rho_t(h) r_t(h)^{\Delta t} = \Pr_{m,\rho_t}(X_t^{\Delta t})$. 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. □

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 $\mathrm{E}_{H\sim\rho}[\mathscr{D}_n^m(H)]$ (of the ρ-weighted predictions of model $m(X_t \mid h; t)$ from observations $X_t \sim n_t$) optimally approximates Bayesian inference for ρ in continuous time.*

Proof. Assume a twice-differentiable parameterization for probability distribution $\rho(h; \theta)$. FR-NGD of $\mathrm{E}_{H\sim\rho}[\mathscr{D}_n^m(H)]$ Eq. (17)) for this distribution is the same as FR-NGD of $\mathrm{E}_{X_t\sim n_t}[\overline{\mathscr{L}}(X_t)]$ (Eq. (16)), since these quantities have equivalent gradients (Lem. 12). By Thm. 1, FR-NGD of $\mathrm{E}_{H\sim\rho}[\mathscr{D}_n^m(H)]$ with respect to θ, provides an optimal approximation of the replicator dynamics for the corresponding loss $\mathscr{L}(h) = -\log m(x_t \mid h; t) = \mathscr{D}_n^m(h)$ (which is uniquely determined by $\overline{\mathscr{L}}(\rho) = \mathrm{E}_{H\sim\rho}[\mathscr{D}_n^m(H)]$). Finally, because Thm. 3 indicates that the replicator dynamics are consistent with continuous-time Bayesian inference, it follows that FR-NGD of $\mathrm{E}_{H\sim\rho}[\mathscr{D}_n^m(H)]$ with respect to θ is an optimal approximation of Bayesian inference in continuous time. □

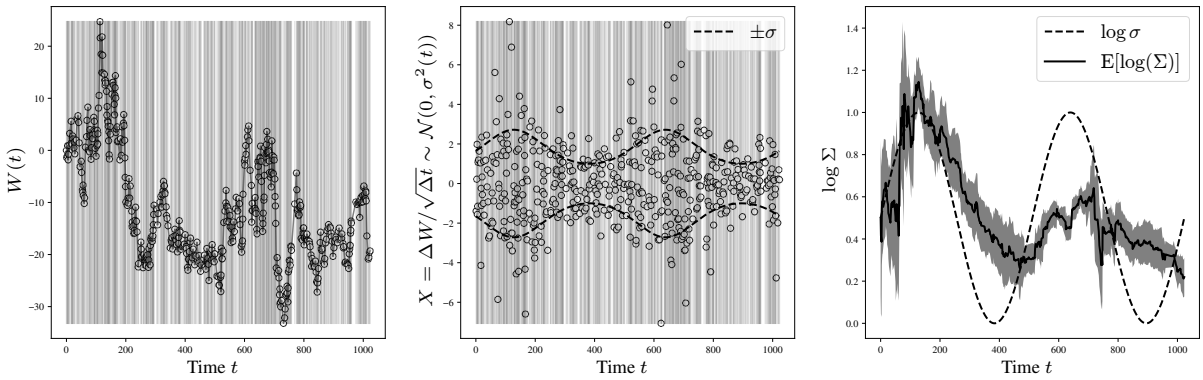

Figure 3: Given randomly-spaced samples of a Wiener process described by $W(t)$ (left) where we define the observable $X = \Delta W/\sqrt{\Delta t} \sim \mathcal{N}(0, \sigma^2(t))$ (center), we learn a Gaussian distribution over the possible values $\log \Sigma$ of $\log \sigma(t)$ (right) using a 40-sample Monte Carlo gradient estimate and FR-NGD of the cross-entropy loss for X_t. The dynamics are discretized in time via Euler update with learning rate $\mathtt{1e}^{-2}$. Vertical lines in the first two images indicate the times of observed samples, while the shaded region in the right image indicates the 1-σ range of the learned distribution.

4.1 Parameter Estimation for a Stochastic Process

In discrete time, the *prior* in Eq. (19), i.e., $\Pr_{m,\rho_t}(X_t^{\Delta t}) = \sum_h \rho_t(h) m(X_t^{\Delta t} \mid h, t)$, is difficult to compute in general when h is high-dimensional. In continuous time, we can avoid the corresponding difficulty of computing $\mathrm{E}_{X_t \sim n_t}[\overline{\mathscr{L}}(X_t)]$ for FR-NGD by using Monte Carlo sampling to estimate its equivalent gradient (Lem. 12), $\partial_i \mathrm{E}_{H\sim\rho}[\mathscr{D}_n^m(H)] = -\mathrm{E}_{X_t,H}\left[\partial_i \log \rho(H;\theta) \log m(X_t \,|\, H, t)\right]$. Using this method, we demonstrate learning a time-varying *distribution* $\log(\Sigma) \sim \mathcal{N}(\theta; t)$, where Σ is an estimator for the time-varying parameter $\sigma(t)$ of a Wiener process $\mathrm{d}W \sim \mathcal{N}(0, \sigma^2(t)\,\mathrm{d}t)$ (Fig. 3).

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 or analogs of genetic mechanisms directly. Moreover, we believe that this approach can be tractable (Nurbekyan et al., 2022), tolerant of noise, and suitable for applications with known domain constraints. Finally, when specialized to cross-entropy loss, we have demonstrated that the same techniques may be used for estimating the time-varying parameters of stochastic processes.

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, and Warren Mardoum for feedback on this or prior versions of our manuscript, and Andrew Warren 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.

David S Berman, Jonathan J Heckman, and Marc Klinger. On the dynamics of inference and learning. *arXiv preprint arXiv:2204.12939*, 2022.

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*.

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.