An Information-Geometric Approach to Artificial Curiosity

Alexander Nedergaard
Institute of Neuroinformatics
University of Zurich and ETH Zurich

Pablo A. Morales
Araya

Abstract

Learning in environments with sparse rewards remains a fundamental challenge in reinforcement learning. Artificial curiosity addresses this limitation through intrinsic rewards to guide exploration, however, the precise formulation of these rewards has remained elusive. Ideally, such rewards should depend on the agent's information about the environment, remaining agnostic to its representation—an invariance central to information geometry. Leveraging this, we show that information monotonicity and invariance under the agent-environment interaction uniquely constrains intrinsic rewards to strictly concave functions of the reciprocal occupancy. Requiring these rewards to yield a principled exploration-exploitation trade-off, via information geodesic interpolation on the occupancy manifold, effectively limits the candidates to those determined by a scalar parameter. Remarkably, special values of this parameter are found to correspond to count-based and maximum entropy exploration. This framework provides important constraints to the engineering of intrinsic rewards while integrating foundational exploration methods into a single, cohesive model.

1 INTRODUCTION

Reinforcement learning systems adapt behavior through trial-and-error to maximize rewards. Inspired by the mammalian dopamine system, artificial reinforcement learning systems [Sutton and Barto, 2018] have surpassed humans in many challenging domains [Mnih et al., 2015, Silver et al., 2016, Vinyals et al., 2019, OpenAI et al., 2019, Kaufmann et al., 2023]. Behavior in reinforcement learning is understood as taking actions in response to observed states. In environments with many states, but sparse rewarding states, finding rewards becomes difficult. The problem of finding rewards, exploration, is a main obstacle to real-world application of general-purpose reinforcement learning methods.

Artificial curiosity [Schmidhuber, 1991] proposes to use intrinsic rewards to guide exploration. However, the specific form of the intrinsic rewards is not constrained by the theory. In consequence, specific intrinsic rewards are typically grounded in empirical performance and anthropomorphic explanations such as curiosity or novelty seeking [Pathak et al., 2017, Burda et al., 2018]. Particular intrinsic rewards have been related through bounds to the theoretically grounded approach count-based exploration [Strehl and Littman, 2008, Kolter and Ng, 2009, Bellemare et al., 2016], which uses state counts in finite state spaces to guide exploration, notably achieving super-human performance in challenging boardgames [Silver et al., 2016]. The recent maximum entropy exploration [Hazan et al., 2019] generalizes to infinite state spaces by maximizing entropy over states, with considerable empirical success under some practical deviations from its theoretical foundation [Liu and Abbeel, 2021, Nedergaard and Cook, 2022], but has not been related to intrinsic rewards.

Intrinsic rewards should ideally depend on the agents information about the environment, remaining agnostic to the representation of the information—an invariance central to information geometry, which concerns geometric structures invariant under information-preserving maps. Classically [Shannon, 1948], such geometries are built upon the Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951], but also exist for Rényi divergences [Rényi, 1961] associated to a scalar characterizing the curvature of statistical manifolds [Amari and Nagaoka, 2000]. Recent work in this direction has generalized the maximum entropy

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

principle [Jaynes, 1957] to curved statistical manifolds with applications to various domains including neural networks [Morales and Rosas, 2021, Morales et al., 2023a,b, Aguilera et al., 2025], suggesting to investigate the role of such curvature in exploration.

In this paper, we prove that any representation-invariant intrinsic reward must be a strictly concave function of the reciprocal occupancy. A one-parameter subfamily of these functions uniquely yields a principled exploration–exploitation trade-off, via geodesic interpolation on the occupancy manifold. Moreover, count-based and maximum-entropy exploration arise as special cases of the scalar parameter, corresponding to distinct values of the occupancy manifold's scalar curvature. Our results are proven in the most general setting, establishing the role of these intrinsic rewards in the information-theoretic structure inherently associated to the reinforcement learning problem.

2 PRELIMINARIES

Formally, we use Markov kernels prove our results for the most general state spaces, that is, second-countable measurable topological spaces, beyond which reinforcement learning problems cannot exist because the agent-environment interaction no longer forms a Markov chain [Meyn and Tweedie, 2012]. Furthermore, fundamental invariance results from reinforcement learning (Theorem 2.1) and information geometry (Theorem 2.3), which form the bedrock of our results, are naturally expressed using Markov kernels.

2.1 Markov kernels

Intuitively, Markov kernels extend the concept of transition matrices to spaces of uncountably many elements, such as continuous spaces. Formally, a Markov kernel from measurable space $(\mathcal{X}, \Sigma)$ to measurable space $(\mathcal{Y}, \Lambda)$ is a function $K : \mathcal{X} \times \Lambda \to [0, 1]$ such that for every $x \in \mathcal{X}$, $K(x, \cdot)$ is a probability measure on $(\mathcal{Y}, \Lambda)$, and every $\lambda \in \Lambda$, $K(\cdot, \lambda)$ is Λ-measurable, where $\cdot$ signifies $f(\cdot) \coloneqq x \mapsto f(x)$. For brevity, we refer to K as a Markov kernel from $\mathcal{X}$ to $\mathcal{Y}$. Each Markov kernel K from $\mathcal{X}$ to $\mathcal{Y}$ induces a Markov morphism

$$h_K(P) \coloneqq \int_{\mathcal{X}} K(x, \cdot) dP(x), \tag{1}$$

which maps probability measures to probability measures. Informally, we may also view Markov kernels as operators that transform probability densities, although they technically operate on measures. A probability measure P is said to be invariant under the Markov morphism h if $h_K(P) = P$. For convenience, we often say that P is invariant under K to mean invariant under h_K.

2.2 Reinforcement learning

The heart of reinforcement learning is the agent-environment interaction, where the agent takes an action $a \in \mathcal{A}$ in response to an observed state $s \in \mathcal{S}$, and the environment changes state based on the state and action. Markov kernels formalize this most generally: The state space $\mathcal{S}$ and the action space $\mathcal{A}$ are measurable topological spaces, with $\mathcal{S}$ second-countable. The transition map δ is a Markov kernel from $\mathcal{S} \times \mathcal{A}$ to $\mathcal{S}$. The starting state probability μ is a probability measure in $\mathcal{S}$. The reward function r is a map from $\mathcal{S}$ to $\mathbb{R}$. The policy π is a Markov kernel from $\mathcal{S}$ to $\mathcal{A}$. The agent-environment interaction is the Markov kernel

$$M(s, B) \coloneqq \int_{\mathcal{A},B} d\delta(s, a, s') d\pi(s, a), \quad s \in \mathcal{S}, B \in \mathcal{B}, \tag{2}$$

where $\mathcal{B}$ is the Borel σ-algebra of $\mathcal{S}$. The reinforcement learning problem (Markov decision problem) is to find the optimal policy $\pi^* \coloneqq \arg\max_{\pi \in \Pi} R(\pi)$ which maximizes the return

$$R(\pi) \coloneqq \int_{\mathcal{S}^{n+1}} \sum_{i=0}^{n} r(s_i) d\mu(s_0) \prod_{i=1}^{n} dM(s_{i-1}, s_i), \tag{3}$$

where n is the episode length. In the more general partially observable Markov decision problems, an observation function maps the state to an observation, and the policy instead maps these to actions. We assume that the episode length does not depend on the policy. Practical deep reinforcement learning typically concerns episodic reinforcement learning ($n < \infty$) where the agent-environment interaction is guaranteed convergence to a steady state density, namely the occupancy (state visitation distribution). We prove this here due to some uncertainty in the reinforcement learning community regarding the result [Bojun, 2020], emphasizing that the occupancy exists when the episodic reinforcement learning problem does and is uniquely invariant (stationary) under the agent-environment interaction:

Theorem 2.1. *The Markov chain formed by the time-inhomogeneous Markov kernel*

$$M_t(s, B) \coloneqq \begin{cases} \mu(B), & \text{if } t \text{ divides } n+1 \\ M(s, B), & \text{else} \end{cases} \tag{4}$$

with $n \in \mathbb{Z}_+$, converges to a unique probability measure

$$P_\pi(B) = \frac{1}{n+1} \int_{\mathcal{S}} \sum_{k=0}^{n} M^k(s, B) d\mu(s) \tag{5}$$

that is uniquely invariant under M_t.

Given a Borel measure V on $\mathcal{S}$, we extend the conventional notion of occupancy by defining it either as the measure P_π or as its Radon-Nikodym derivative

$$p_\pi \coloneqq \frac{dP_\pi}{dV}. \tag{6}$$

Intuitively, the occupancy is the generalization of normalized state counts to infinite spaces. For infinite state spaces, the occupancy can be estimated efficiently using a non-parametric density estimator [Hazan et al., 2019, Mutti et al., 2020, Nedergaard and Cook, 2022]. Invariance of the occupancy under the agent-environment interaction provides a useful reformulation of the return in terms of the occupancy:

Lemma 2.2.

$$R(p_\pi) \coloneqq (n+1)\int_{\mathcal{S}} p_\pi(s) r(s) dV(s) = R(\pi). \tag{7}$$

The occupancy reflects the information the agent has about the environment; if the agent interacts with the environment for long enough, the agent's relative information about states eventually stabilizes, depending only on the relative frequency of state visits. The frequency of state visits determines the number of the samples the agent gathers from functions defined over states, such as the reward function, and more samples reduce the agent's uncertainty about those functions. Unsurprisingly, the occupancy forms the basis of some foundational exploration approaches.

2.3 Exploration

Artificial curiosity, count-based exploration and maximum entropy exploration are among the empirically strongest exploration approaches. Taking inspiration from biological curiosity, artificial curiosity [Schmidhuber, 1991] proposes to add an intrinsic reward to the reward:

$$r(s) + \beta \bar{r}(s), \quad \beta \in \mathbb{R}_{\geq 0}. \tag{8}$$

Artificial curiosity with specific intrinsic rewards has had considerable empirical success [Pathak et al., 2017, Burda et al., 2018], but such methods are typically heuristically motivated, making it difficult to formally reason about the form of $\bar{r}$. Furthermore, the very large set of functions $\{\bar{r} : \mathcal{S} \to \mathbb{R}\}$ is not feasible to search over experimentally. Given the precedent of biological curiosity and evidence for surprise and novelty dopamine rewards in the brain [Kakade and Dayan, 2002], artificial curiosity is compelling but requires significant restrictions on the form of intrinsic rewards. A more formally grounded approach is count-based exploration [Kolter and Ng, 2009, Strehl and Littman, 2008, Bellemare et al., 2016], which is based on the principle of optimism in the face of uncertainty. The principle suggests that when estimating the expected value of a random reward function, we should be as optimistic in our estimation as our uncertainty permits. For subgaussian (e.g. bounded) reward functions, the Hoeffding bound [Hoeffding, 1962] shows that the uncertainty is bounded by the reciprocal square root of the number of samples. Count-based exploration proposes to add such an upper confidence bound to the reward:

$$r(s) + \beta \frac{1}{\sqrt{n(s)}}, \quad \beta \in \mathbb{R}_{\geq 0}, \tag{9}$$

where $n(s)$ denotes number of visits to state s. The normalization of $n(s)$ coincides with the occupancy. Count-based exploration has been immensely successful on challenging problems with finite state spaces [Silver et al., 2016]. A recent approach using the occupancy is maximum entropy exploration [Hazan et al., 2019], which based on the principle of maximum entropy for exploration. The principle suggests that in the absence of observed rewards, states should be visited uniformly (c.f. Jaynes [1957]). Using the result that Shannon entropy [Shannon, 1948]

$$H(p) \coloneqq \int_{\mathcal{S}} p(s) \log \frac{1}{p(s)} dV(s) \tag{10}$$

is maximized by the uniform probability density u, maximum entropy exploration proposes to adds the Shannon entropy of the occupancy to the return:

$$R(\pi) + \beta H(p_\pi), \quad \beta \in \mathbb{R}_{\geq 0}. \tag{11}$$

The approach should not be confused with maximum entropy reinforcement learning [Haarnoja et al., 2018] which adds the Shannon entropy of the policy to the return. Practical maximum entropy exploration has had considerable empirical success when replacing the log in Shannon information with a different concave function to improve numerical stability under non-parametric density estimation [Liu and Abbeel, 2021, Nedergaard and Cook, 2022]. We will show that this deviation from theory is not merely a heuristic, but provides a principled unification of count-based and maximum entropy exploration, through artificial curiosity with intrinsic rewards based on the occupancy. However, prior to this, we must establish that spaces of probability densities, such as the occupancy space, possess an invariant geometric structure.

2.4 Information geometry

Information geometry [Amari and Nagaoka, 2000] studies parameterized families of probability densities $\left\{p_\theta : \theta \in \mathbb{R}^d\right\}$ with additional geometric structure that

is invariant under congruent Markov morphisms. Intuitively, congruent Markov morphisms are information-preserving maps, appearing as the equality condition in data processing inequalities such as [Cover and Thomas, 2006, Theorem 2.8.1]. We are interested in structures that are invariant under them, because these structures are agnostic to the representation of information. We will refer to invariance under congruent Markov morphisms as information invariance, and the stronger property of satisfying a data processing inequality as information monotonicity. Formally, a Markov morphism is congruent if it has a left inverse induced by a statistic. Congruent Markov morphisms have a correspondence with sufficient statistics, as follows: A statistic from $\mathcal{X}$ to $\mathcal{Y}$ is a measurable map $\kappa : \mathcal{X} \to \mathcal{Y}$ and induces a Markov morphism

$$h_\kappa(P) \coloneqq \int_\cdot dP\kappa^{-1}(y) = P\kappa^{-1}. \tag{12}$$

The statistic κ is sufficient if there exists a Markov morphism h in the opposite direction from $\mathcal{Y}$ to $\mathcal{X}$, such that h is a right inverse of h_κ. Equivalently, the Markov morphism h has a left inverse h_κ induced by the statistic κ, making h a congruent Markov morphism and κ a sufficient statistic. The invariant geometric structures in information geometry are tensors, which for our purposes can be understood intuitively as mathematical objects that describe geometry in a coordinate-independent manner. A tensor T is invariant under the Markov morphism h if $h^*(T) = T$ where h^* denotes the pullback of h. Pullbacks are maps between additional structure on sets, induced by maps between those sets. A fundamental result in information geometry concerns uniquely invariant tensors:

Theorem 2.3. *[Čencov, 1972, Ay et al., 2015] The unique information invariant (0,2)-tensors and (0,3)-tensors are ηg and αT, where $\eta, \alpha \in \mathbb{R}$, and*

$$g_{ij}(p_\theta) = \int_{\mathcal{S}} \partial_i \log p_\theta(s) \partial_j \log p_\theta(s) dP_\theta(s) \tag{13}$$

is the Fisher-Rao tensor (Fisher information metric), and

$$T_{ijk}(p_\theta) = \int_{\mathcal{S}} \partial_i \log p_\theta(s) \partial_j \log p_\theta(s) \partial_k \log p_\theta(s) dP_\theta(s) \tag{14}$$

is the Amari-Čencov tensor.

where we adopt the notation $\partial_i \coloneqq \frac{\partial}{\partial \theta_i}$ and $\partial_i' \coloneqq \frac{\partial}{\partial \theta_i'}$.

Intuitively, the Fisher-Rao and Amari-Čencov tensors describe a natural geometric structure that is invariant under information-preserving maps. However, this natural geometry is only determined up to the constants $\eta \in \mathbb{R}$ and $\alpha \in \mathbb{R}$, which respectively trivially scale and non-trivially asymmetrically deform the geometry. Asymmetric geometry is already familiar to many machine learning researchers from the asymmetry of the KL-divergence [Kullback and Leibler, 1951]. An important generalization are the f-divergences [Rényi, 1961, Csiszár, 1963]:

$$\mathcal{D}_f(p\|q) \coloneqq \int_{\mathcal{S}} p(s) f\left[\frac{q(s)}{p(s)}\right] dV(s) \tag{15}$$

with f is a thrice-differentiable strictly convex function satisfying $f(1) = 1$, recovering the KL divergence for $f = -\log$. An important example for us is

$$f_\alpha(x) \coloneqq \frac{4}{1-\alpha^2}(1 - x^{\frac{\alpha+1}{2}}), \quad \alpha \in \mathbb{R} \tag{16}$$

which induces the α-divergence [Amari and Nagaoka, 2000]

$$\mathcal{D}_\alpha(p\|q) \coloneqq \frac{4}{1-\alpha^2}\left(1 - \int_{\mathcal{S}} p(s)^{\frac{1-\alpha}{2}} q(s)^{\frac{\alpha+1}{2}} dV(s)\right). \tag{17}$$

Letting $\alpha = 2f''(1) + 3f'''(1)$, which holds naturally for α-divergences, we see that f-divergences induce a geometry fundamentally connected to the information invariant tensors [Eguchi, 1983]:

$$\begin{aligned}
\Gamma^f_{ijk}(p_\theta) &\coloneqq -\partial_i \partial_j \partial_k' \mathcal{D}_f(p_\theta \| p_{\theta'})\Big|_{\theta=\theta'} \\
&= \Gamma^0_{ijk}(p_\theta) - \frac{\alpha}{2} T_{ijk}(p_\theta), && \text{(18a)} \\
\Gamma^{f^*}_{ijk}(p_\theta) &\coloneqq -\partial_i \partial_j \partial_k' \mathcal{D}_f(p_{\theta'} \| p_\theta)\Big|_{\theta=\theta'} \\
&= \Gamma^0_{ijk}(p_\theta) + \frac{\alpha}{2} T_{ijk}(p_\theta), && \text{(18b)} \\
\Gamma^0_{ijk}(p_\theta) &\coloneqq \frac{1}{2}(\partial_i g_{jk}(p_\theta) + \partial_j g_{ki}(p_\theta) - \partial_k g_{ij}(p_\theta)), && \text{(18c)}
\end{aligned}$$

where Γ^f, Γ^{f^*} and $\Gamma^0 = \frac{1}{2}(\Gamma^f + \Gamma^{f^*})$ are the coefficients of the primal, dual and Levi-Civita connections respectively. Connections determine geodesics, generalized straight lines, on curved manifolds. The α-geodesics, or information geodesics, are the curves $\gamma : \mathbb{R} \to \mathcal{P}$ satisfying the geodesic equation:

$$\frac{d^2\gamma_k}{dt^2} + \sum_{i,j} \Gamma^\alpha_{ijk} \frac{d\gamma_i}{dt} \frac{d\gamma_j}{dt} = 0. \tag{19}$$

Intuitively, the Fisher-Rao tensor describes a Riemannian geometry with a symmetric divergence, and α determines its deformation by the Amari-Čencov tensor into a non-Riemannian geometry with an asymmetric divergence. The geometry determined by α has constant sectional curvature $\frac{1}{4}(1-\alpha^2)$ [Amari and Nagaoka, 2000], so the geometry is spherical for $|\alpha| < 1$, flat for $\alpha \pm 1$ and hyperbolic for $|\alpha| > 1$. With (18a-18c), the geometry is thus uniquely flat when $\alpha = \pm 1$

and uniquely Riemannian when $\alpha = 0$. For these special geometries, the α-divergence corresponds to the KL-divergence ($\alpha = -1$), reverse KL-divergence ($\alpha = 1$) and Hellinger divergence [Hellinger, 1909] ($\alpha = 0$). Importantly, curvature depends on the space under consideration. For instance, the geometry is hyperbolic for $\alpha = 0$ in the subspace of Gaussian probability densities, and the geometry is flat for any $\alpha \in \mathbb{R}$ in the ambient space of measures [Amari, 2009]. Closely related to this, α-divergences satisfy the property that their gradients point along the geodesics of their induced geometry in the space of measures, a property which general f-divergences do not satisfy [Ay et al., 2017, Proposition 2.12]. Formalizing this, we say that a divergence is *geodetic* if

$$\operatorname{grad} \mathcal{D}(p\|\cdot)\big|_q = -\eta \left.\frac{d}{dt}\gamma_{q,p}(t)\right|_{t=0} \tag{20}$$

with $\eta \in \mathbb{R}_{>0}$ and the α-geodesic $\gamma_{q,p}$, connecting q and p, induced by $\mathcal{D}$ in the space of measures.

Lemma 2.4. *The geodetic property of divergences is preserved by strictly monotonic functions.*

The Rényi divergences [Rényi, 1961] which are not f-divergences, are geodetic by a monotonic relation to α-divergences through the Box-Cox transformation [Box and Cox, 1964]. Monotonically related divergences $\mathcal{D} = F(\bar{\mathcal{D}})$ induce the same geometry (up to trivial scaling) through (18a-18b), and induce the same constants η and α for the invariant tensors if the monotonic function F additionally satisfies $F'(0) = 1$. We now derive intrinsic rewards based on the geometric structures introduced here and show that artificial curiosity with these rewards unify foundational exploration approaches with the Amari-Čencov tensor's constant α as the sole degree of freedom.

3 INVARIANT INFORMATION REWARDS

To formally ground artificial curiosity as rewarding information, we seek an intrinsic reward that depends on the information that the agent has about the environment and is agnostic to the representation of the information. Formalizing this, we now derive information rewards and prove that they are uniquely information monotonic and invariant under the agent-environment interaction.

3.1 Measuring information

In information theory, information is usually understood through Shannon entropy. For artificial curiosity, we want a notion of information that captures the information received by the agent when a state is observed, ideally without reference to the inner structure of the agent. A candidate function is Shannon information (surprisal) $I(s;p) \coloneqq -\log p(s)$ which can be interpreted as the surprise of observing a state s given a subjective belief encoded by the probability density p. A geometric interpretation follows from the relation

$$\underbrace{\int_{\mathcal{S}} p(s) I(s;p) dV(s)}_{=H(p)} - \log u = \underbrace{-\int_{\mathcal{S}} p(s) \log \frac{p(s)}{u} dV(s)}_{=\mathcal{D}_{\mathrm{KL}}(p\|u)}, \tag{21}$$

where u denotes the uniform probability density, such that Shannon information measures distinctness from maximal uncertainty along each state. Generalizing Shannon information to satisfy an analogous relationship with f-divergences, we define f-information as

$$I_f(s;p) \coloneqq f\left[\frac{1}{p(s)}\right] \tag{22}$$

with f a strictly concave function. f-Information has a similar geometric interpretation to Shannon information when $f(1) = 0$ such that it induces an f-divergence $\mathcal{D}_{-f}$. The induced f-divergence is unchanged by transformations $f(x) \mapsto f(x) + c(x-1)$ with $c \in \mathbb{R}$. We say that a strictly concave function f is geodetic if it induces a geodetic f-divergence. It turns out that the functions f_α, which induce the α-divergences, are uniquely geodetic (c.f. Amari [2009]):

Lemma 3.1. *The family* $\{f_\alpha : \alpha \in \mathbb{R}\}$ *is uniquely geodetic, up to transformations* $f_\alpha(x) \mapsto \eta\{f_\alpha(x) + c(x-1)\}$ *with* $\eta, c \in \mathbb{R}$.

From these functions, we obtain α-information

$$I_\alpha(s;p) = \frac{4}{1-\alpha^2}\left(\left[\frac{1}{p(s)}\right]^{\frac{\alpha+1}{2}} - 1\right), \quad \alpha \neq 1, \tag{23}$$

where, for notational simplicity we write I_α in place of I_{f_α}. While $I_{\alpha\to 1}$ diverges, $I_{-1} \coloneqq I_{\alpha\to -1}$ converges to $-\log$. The distinctive geodetic property of α-information underlies a principled exploration-exploitation trade-off, a fact that we will later prove. However, as we will now show, only strict concavity is required to ensure the invariance of f-information as an intrinsic reward in artificial curiosity—the fact which motivates our choice of definition.

3.2 Artificial curiosity with information rewards

The motivation that intrinsic rewards in artificial curiosity should depend on the information that the agent has about the environment, while remaining agnostic to the representation of the information,

uniquely constrains intrinsic rewards to f-information relative to the occupancy (c.f. Jiao et al. [2014]):

Theorem 3.2. *Intrinsic rewards $\bar{r}(s;p) := \bar{f}[p(s)]$, with any function $\bar{f}:\mathbb{R}\to\mathbb{R}$ and any probability density p, are invariant under the agent-environment interaction M,*

$$\bar{r}(s;p) = \bar{r}(s;h_M(p)), \tag{24}$$

uniquely when $p = p_\pi$. Furthermore, for intrinsic rewards $\bar{r}(s) := \bar{r}(s;p_\pi)$, the intrinsic return

$$\bar{R}(p_\pi) = \int_{\mathcal{S}^{n+1}} \sum_{i=0}^{n} \bar{r}(s_i)\, d\mu(s_0) \prod_{i=1}^{n} dM(s_{i-1},s_i) \tag{25}$$

satisfies information monotonicity

$$\bar{R}(p_\pi) \le \bar{R}(h_\kappa(p_\pi)), \tag{26}$$

with equality if and only if the statistic κ is sufficient, uniquely when $\bar{r}(s) = I_f(s;p_\pi)$.

We refer to intrinsic rewards of the form $I_f(\cdot,p_\pi)$ as f-information rewards or simply information rewards. Artificial curiosity with f-information rewards has the form

$$r(s) + \beta I_f(s;p_\pi), \quad \beta\in\mathbb{R}_{\ge 0}. \tag{27}$$

Crucially, the degrees of freedom in designing intrinsic rewards have been constrained to a strictly concave function f. Moreover, this restriction arises uniquely from the motivation of rewarding information. Further restrictions on f will unify foundational exploration approaches with the Amari-Čencov tensor's constant $\alpha\in\mathbb{R}$ serving as the sole degree of freedom.

3.3 Unifying exploration through geometry

The degrees of freedom in artificial curiosity with information rewards (27) are the strictly concave function f and the parameter $\beta\in\mathbb{R}_{\ge 0}$, which governs the exploration-exploitation trade-off by yielding a distinct optimal occupancy for each value. A principled method to interpolate between these occupancies is along α-geodesics, which are uniquely invariant under information-preserving maps (Theorem 2.3). Since α-information functions are directly linked to these geodesics through their unique geodetic property (Lemma 3.1), α-information rewards provide a principled exploration-exploitation trade-off. Remarkably, artificial curiosity with α-information rewards specializes to count-based and maximum entropy exploration for the uniquely Riemannian and flat geometries:

Theorem 3.3. *Artificial curiosity with*

$$r(s) + \beta I_\alpha(s;p_\pi), \quad \alpha\in\mathbb{R},\ \beta\in\mathbb{R}_{\ge 0} \tag{28}$$

is equivalent to count-based exploration for $\alpha = 0$ and maximum entropy exploration for $\alpha = -1$.

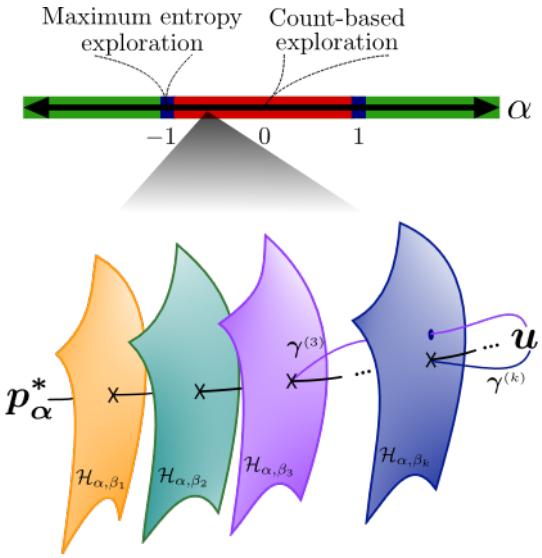


Figure 1: **Artificial curiosity with α-information rewards on the curved occupancy manifold. (Top).** The Amari-Čencov tensor constant $\alpha\in\mathbb{R}$ encodes the occupancy manifold curvature (red–spherical, blue–flat, green–hyperbolic). Count-based exploration corresponds to the Riemannian geometry with $\alpha = 0$, and maximum entropy exploration to the flat geometry with $\alpha = -1$ (Theorem 3.3). **(Bottom)** The intrinsic rewards scaling $\beta\in\mathbb{R}_{\ge 0}$ controls the exploration-exploitation trade-off on the curved occupancy manifold. The optima are α-projections, along $(-\alpha)$-geodesics $\gamma^{(i)}$, from the uniform occupancy u onto isoreturn hyperplanes $\mathcal{H}_{\alpha,\beta_i}$ (Proposition 4.2). The exploration-exploitation trade-off is $(\alpha+2)$-geodesic β-interpolation between the maximally rewarding occupancy p^*_α and the uniform occupancy u (Theorem 4.4).

Equivalence is meant in the sense of having the same optimization objectives under gradients and constant scaling of β. The relationship to Rényi state entropy exploration [Yuan et al., 2022], Boltzmann exploration [Thrun, 1992] and classic artificial curiosity [Schmidhuber, 1991] is given in Appendix B. Strikingly, count-based and maximum entropy exploration are usually derived from completely different principles and frameworks, but coincide with special values of the Amari-Čencov tensor constant α (see Figure 1).

4 EXPLORATION ON CURVED STATISTICAL MANIFOLDS

Artificial curiosity with α-information rewards, exemplified by count-based and maximum entropy exploration, operates under the implicit assumption that

the occupancy space constitutes a statistical manifold induced by the α-divergence. An examination of exploration on this occupancy manifold reveals how these methods trade off exploration and exploitation in a principled manner. This geometric insight serves as the rationale for emphasizing α-information rewards as a distinct and valuable subclass within the broader category of f-information rewards.

4.1 A geometric exploration-exploitation trade-off

We now show that the optima of artificial curiosity with α-information rewards uniquely trace geodesics on the occupancy manifold, reinforcing our claim that α-information rewards yield a principled exploration-exploitation trade-off. Let us consider the optimization objective of artificial curiosity with information rewards:

$$\int_{\mathcal{S}^{n+1}} \sum_{i=0}^{n} \{r(s_i) + \beta I_f(s_i; p_\pi)\}\, d\mu(s_0) \prod_{i=1}^{n} dM(s_{i-1}, s_i). \tag{29}$$

By Lemma 2.2, we may consider the optimization objective as an equivalent function over the occupancy manifold:

$$R_{f,\beta}(p_\pi) \coloneqq (n+1) \int_{\mathcal{S}} p_\pi(s) \{r(s) + \beta I_f(s; p_\pi)\}\, dV(s). \tag{30}$$

The optima of artificial curiosity with f-information

$$p_{f,\beta} \coloneqq \arg\max_{p \in \mathcal{P}_\Pi} R_{f,\beta}(p) \tag{31}$$

are points on the occupancy manifold. The set of occupancies $\mathcal{P}_\Pi \coloneqq \{p_\pi : \pi \in \Pi\}$ may not be the set of all probability densities $\mathcal{P}$, depending on the environment and policy class. Because the return is linear in the occupancy, the set of occupancies where the same return is achieved, $\mathcal{H}(c) \coloneqq \{p \in \mathcal{P}_\Pi : R(p) = c\}$ with $c \in \mathbb{R}$, is a hyperplane in the occupancy space. The optima of artificial curiosity with α-information, $p_{\alpha,\beta} \coloneqq p_{f_\alpha,\beta}$, are the solutions to an α-divergence minimization problem constrained to such a hyperplane:

Lemma 4.1. *If P_Π is convex,*

$$p_{\alpha,\beta} = \arg\min_{p \in \mathcal{H}_{\alpha,\beta}} \mathcal{D}_\alpha(p \| u), \tag{32}$$

where the hyperplane $\mathcal{H}_{\alpha,\beta} \coloneqq \mathcal{H}(c_{\alpha,\beta})$ depends on α and β. The optima can now be understood in terms of pure geometry via geodesics:

Proposition 4.2. *If P_Π is convex, the optima $p_{\alpha,\beta}$ are α-projections from the uniform occupancy u onto isoreturn hyperplanes $\mathcal{H}_{\alpha,\beta}$.*

The α-projections are the points on the hyperplanes where $(-\alpha)$-geodesics, connecting the uniform occupancy and the points, are orthogonal to α-geodesics on the hyperplanes (see Figure 1). Finding these points is equivalent to a generalized maximum entropy problem [Morales and Rosas, 2021] for which the solutions are curved exponential families:

Lemma 4.3. *If $\mathcal{P}_\Pi = \mathcal{P}$,*

$$p_{\alpha,\beta} = \frac{(1 + \frac{1}{\lambda\beta} r)^{-\frac{2}{\alpha+1}}}{\int (1 + \frac{1}{\lambda\beta} r)^{-\frac{2}{\alpha+1}} dV} \tag{33}$$

with $\lambda \in \mathbb{R}$ determined by α and β.

A concrete example of this is given in Appendix 5. Importantly for optimization theory, all states are visited from these optima if $\beta > 0$. As α changes, the isoreturn hyperplane and α-projection change smoothly (c.f. Amari and Shimokawa [2001]). As β changes from 0 to ∞, the isoreturn hyperplane changes smoothly, and the α-projections trace out a geodesic:

Theorem 4.4. *If $\mathcal{P}_\Pi = \mathcal{P}$ and $\beta \geq \beta_0 > 0$, the exploration-exploitation relation $\beta \mapsto p_{f,\beta}$ is an $(\alpha + 2)$-geodesic connecting the maximally rewarding occupancy p_{α,β_0} and the uniform occupancy u, uniquely when $f = f_\alpha$.*

While this establishes the special role of α-information rewards in the information-theoretic structure of the reinforcement learning problem, the result does not hold when states cannot be visited arbitrarily, i.e. $\mathcal{P}_\Pi \subset \mathcal{P}$. Importantly for practice, the topology and geometry of the occupancy manifold determine whether optima can jump when lowering β:

Proposition 4.5. *If $\mathcal{P}_\Pi$ is convex, the exploration-exploitation relation $\beta \mapsto p_{f,\beta}$ is a map. If additionally $\mathcal{P}_\Pi$ is compact, the map is continuous.*

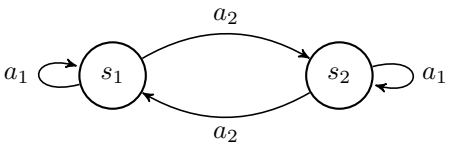
Since the observation function in partially observable Markov decision problems is a statistic, Theorem 3.2 shows that an upper bound on (29) is optimized when using information rewards over observations instead of states. Therefore, Theorem 4.4 holds for partially observable Markov decision problems if the observation function is sufficient, but not in general.

We have demonstrated that, under α-information rewards, artificial curiosity uniquely achieves a principled trade-off between exploration and exploitation: its optima coincide with α-projections, and the exploration-exploitation trade-off is traced by $(\alpha + 2)$-geodesic β-interpolation (see Figure 1). In the special case of maximum entropy exploration ($\alpha = -1$), the α-projection and β-interpolation geodesics coincide. Moreover, these geometric insights extend beyond count-based and maximum entropy exploration

to Rényi state entropy exploration, once the trade-off geodesic is reparameterized (see Appendix B). Because the geodesic can be parameterized by Rényi entropy uniquely for α-information rewards, it is only for these intrinsic rewards that β has a natural information-theoretic interpretation as the *information exploited by the optimal agent to maximize rewards*. By Proposition 4.2, α has a geometric interpretation as the curvature of the occupancy manifold. The importance of the occupancy manifold's curvature for exploration is evident from empirical performance gains when implicitly fine-tuning α to the environment [Yuan et al., 2022]. An interpretation based on α-divergences is that negative α emphasizes the occupancy's peaks while positive α emphasizes the valleys, such that α controls a trade-off between penalizing high and low probability states (see Figure 2), with these behaviors balanced by count-based exploration ($\alpha = 0$). Discussion of the role of α for practical algorithms with density estimation can be found in Appendix A.

5 TOY EXAMPLE

We here show a simple toy example to concretely instantiate concepts and illustrate the exploration-exploitation trade-off in artificial curiosity with α-information rewards. Consider a deterministic environment with two states and two actions, where the state switches when the second action is taken:



Formally, we have a state space $\mathcal{S} = \{s_1, s_2\}$, action space $\mathcal{A} = \{a_1, a_2\}$ and transition probability

$$\delta(s_i, a_k, s_j) = \begin{cases} \mathbb{I}[i = j] & \text{if } k = 1, \\ \mathbb{I}[i \neq j] & \text{if } k = 2, \end{cases} \tag{34}$$

where $\mathbb{I}$ denotes the indicator function. Suppose that only the second state is rewarding, $r(\{s_1, s_2\}) = \{0, 1\}$. Let the episode length be infinite so that the occupancy does not depend on the starting state density. Consider now a stochastic policy $\pi_\theta(s_i, a_1) = \theta_i$ parameterized by $\theta = \{\theta_1, \theta_2\}$ from the parameter space $\Theta = [0, 1]^2$, such that the probability of switching from state s_i is θ_i. Since the state and action spaces are finite, the agent-environment interaction

$$M(s_i, \{s_j\}) = \sum_{k=1}^{|\mathcal{A}|=2} \delta(s_i, a_k, s_j)\pi(s_i, a_k) \tag{35}$$

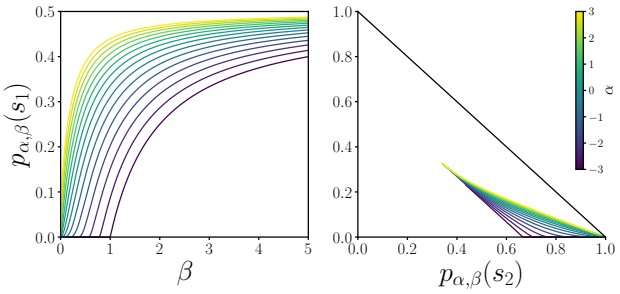


Figure 2: **Exploration-exploitation trade-off in artificial curiosity with α-information rewards. (Left)**. The optimal occupancy $p_{\alpha,\beta}$ for different α and β on a 2-state environment with reward function $r(\{s_1, s_2\}) = \{0, 1\}$. When every occupancy is realizable, the optimal occupancies depend only on the state space and reward function. **(Right)**. The optimal occupancy for different α and β on a 3-state environment with reward function $r(\{s_1, s_2, s_3\}) = \{0, 1, 0.5\}$. The parameter β controls interpolation between the uniform occupancy and the maximally rewarding occupancy, and the parameter α controls the balance between penalizing high and low probability states. Observe how lower α permits very low probability states, while higher α does not.

is conveniently represented as the transition matrix

$$M = \begin{bmatrix} 1 - \theta_1 & \theta_2 \\ \theta_1 & 1 - \theta_2 \end{bmatrix}. \tag{36}$$

Similarly, the occupancy can be represented as a column vector $p = \begin{bmatrix} p_1 & p_2 \end{bmatrix}^\intercal$. From the unique invariance of the occupancy under the agent-environment interaction (Theorem 2.1)

$$\begin{bmatrix} 1 - \theta_1 & \theta_2 \\ \theta_1 & 1 - \theta_2 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}, \tag{37}$$

we obtain a closed-form solution for the occupancy:

$$p_i = \frac{\theta_j}{\theta_i + \theta_j}. \tag{38}$$

We note that it is unusual to have such a closed-form expression of the occupancy for practical problems. By Lemma 4.3 and the particular reward function $r(\{s_1, s_2\}) = \{0, 1\}$, the probability of the first state at the optimal occupancy of artificial curiosity with α-information rewards is given by

$$p_{\alpha,\beta}(s_1) = \frac{1}{1 + (1 + \frac{1}{\lambda\beta})^{-\frac{2}{\alpha+1}}} \tag{39}$$

with $\lambda \in \mathbb{R}$ determined by α and β. Note that the optimal occupancies $p_{\alpha,\beta}$ depend only on the state space

and reward function if every occupancy is realizable. For the special case of $\alpha = -1$, we uniquely have a closed-form solution $p_{-1,\beta}(s_i) = e^{r(s_i)/\beta}$ by Lemma 4.1 and Jaynes [1957], such that we obtain a sigmoid function for the exploration-exploitation trade-off:

$$p_{-1,\beta}(s_1) = \frac{1}{1 + e^{1/\beta}}, \tag{40}$$

By equating this to the occupancy, we see that optimal parameters for $\alpha = -1$ satisfy

$$\theta_1 = \frac{1 + e^{1/\beta}}{1 + e^{-1/\beta}}\theta_2. \tag{41}$$

For $\alpha \neq -1$, we do not have a closed-form solution. We used CVXPY [Diamond and Boyd, 2016, Agrawal et al., 2019] to compute the solutions and plotted them on the left in Figure 2. To illustrate the curvature of the occupancy space, encoded by α, we extended the toy example to have a third state s_3 with reward $r(s_3) = 0.5$ and plotted the optimal occupancies on the right in Figure 2.

6 CONCLUSION

Information rewards—being the only intrinsic rewards that are information monotonic and invariant under agent-environment interactions—specialize artificial curiosity to strictly concave functions of the reciprocal occupancy. This invariance ensures that such rewards depend solely on the agent's information about the environment, independent of any particular representation. In the special case of α-information, artificial curiosity uniquely attains a principled exploration-exploitation trade-off: its optima lie along the $(\alpha + 2)$-geodesic β-interpolation on the occupancy manifold. Notably, the empirically successful count-based $(\alpha = 0)$ and maximum entropy $(\alpha = -1)$ exploration coincide with special values of the Amari-Čencov tensor's constant α, highlighting how the curvature of the occupancy manifold governs effective exploration strategies. Furthermore, $\alpha = 0$ and $\alpha = -1$ also suggest a broader interpretation within biological reinforcement learning, in which novelty and surprise can be understood as different points on a common α-information spectrum (see Appendix C). Our findings unify foundational exploration approaches in a principled formalism and significantly constrain the design of intrinsic rewards for practical exploration in general state spaces.

6.1 Refining the design of intrinsic rewards

Deep reinforcement learning practitioners following our approach should use artificial curiosity with information rewards (27). Pseudocode for this is provided in Appendix A. To apply information-based intrinsic rewards in infinite state spaces, one must first estimate the occupancy density. In practice, one may employ efficient non-parametric estimators drawn from the maximum-entropy exploration literature—e.g., kernel density estimation [Hazan et al., 2019], k-nearest-neighbor density estimation [Mutti et al., 2020] or k-means density estimation [Nedergaard and Cook, 2022]. Moreover, in many applications it is preferable to estimate occupancy over a transformed or lower-dimensional representation of the state space. Although our present focus is on the formal structure of intrinsic rewards, efficient occupancy estimation and the choice of state space representation are essential for effective exploration in practice. We therefore identify these topics—occupancy estimation and state space representation—as important avenues for future research. The form of the strictly concave function f remains unresolved and is primarily an empirical question to be investigated on meaningful reinforcement learning problems. The family $\{f_\alpha : \alpha \in \mathbb{R}\}$, which induces α-information, is promising as it uniquely yields a principled exploration-exploitation trade-off. Furthermore, empirical studies have found superior performance on exploration problems when implicitly using α-information rewards (see Appendix A). Artificial curiosity with α-information shares the same optima as Rényi state entropy exploration with the practical advantage of being computable locally in distributed reinforcement learning. A promising future direction is practical algorithms using α-information rewards with adaptive α and β. In a separate publication, we will empirically investigate artificial curiosity with information rewards on robotics problems, including different methods for occupancy estimation and state space representation. A thorough investigation of the effects of f, α and β on optimization theory bounds, along with the deformation of statistical manifolds in reinforcement learning and scientific domains, has the potential to provide compelling theoretical justification for a specific information reward.

Acknowledgements

P.A.M. acknowledges support by JSPS KAKENHI Grant Number 23K16855, 24K21518. A.N. acknowledges support by SNSF Grant Number 182539 (Neuromorphic Algorithms based on Relational Networks). A.N. was a visiting researcher at Araya for a part of this work.

References

A. Agrawal, S. Diamond, and S. Boyd. Disciplined geometric programming. *Optimization Letters*, 13 (5):961–976, 2019.

M. Aguilera, P. A. Morales, F. E. Rosas, and H. Shimazaki. Explosive neural networks via higher-order interactions in curved statistical manifolds. *Nature Communications*, 16(1):6511, 2025.

S.-I. Amari. α-divergence is unique, belonging to both f-divergence and bregman divergence classes. *IEEE Transactions on Information Theory*, 55(11):4925–4931, 2009.

S.-i. Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

S.-i. Amari and H. Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

S.-i. Amari and H. Shimokawa. Information geometry of alpha-projection in mean field approximation. *Advanced Mean field Methods: Theory and Practice*, 01 2001.

N. Ay, J. Jost, H. V. Lê, and L. Schwachhöfer. Information geometry and sufficient statistics. *Probability Theory and Related Fields*, 162:327–364, 2015.

N. Ay, J. Jost, H. Vân Lê, and L. Schwachhöfer. *Information geometry*, volume 64. Springer, 2017.

A. Barto, M. Mirolli, and G. Baldassarre. Novelty or surprise? *Frontiers in psychology*, 4:907, 2013.

M. G. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. In *Neural Information Processing Systems*, 2016.

R. Bellman. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60 (6):503–515, 1954.

C. Berge. Espaces topologiques: Fonctions multivoques. *Dunod*, 1959.

H. Bojun. Steady state analysis of episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33:9335–9345, 2020.

G. E. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 26(2):211–243, 1964.

Y. Burda, H. Edwards, A. Storkey, and O. Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

T. M. Cover and J. A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006. ISBN 0471241954.

I. Csiszár. An information-theoretic inequality and its application to the proof of ergodicity in markov chains (in german). *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 8:85—108, 1963.

S. Diamond and S. Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83): 1–5, 2016.

S. Eguchi. Second order efficiency of minimum contrast estimators in a curved exponential family. *The Annals of Statistics*, pages 793–803, 1983.

T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/haarnoja18b.html`.

E. Hazan, S. Kakade, K. Singh, and A. Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691. PMLR, 2019.

E. Hellinger. New foundations for the theory of quadratic forms with infinitely many variables (in german). *Journal für die reine und angewandte Mathematik*, 1909(136):210–271, 1909. doi: doi: 10.1515/crll.1909.136.210. URL `https://doi.org/10.1515/crll.1909.136.210`.

W. Hoeffding. Probability inequalities for sums of bounded random variables. 1962.

E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

J. Jiao, T. A. Courtade, A. No, K. Venkat, and T. Weissman. Information measures: the curious case of the binary alphabet. *IEEE Transactions on Information Theory*, 60(12):7616–7626, 2014.

S. Kakade and P. Dayan. Dopamine: generalization and bonuses. *Neural Networks*, 15(4-6):549–559, 2002.

E. Kaufmann, L. Bauersfeld, A. Loquercio, M. Müller, V. Koltun, and D. Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620(7976):982–987, 2023.

J. Z. Kolter and A. Ng. Near-bayesian exploration in polynomial time. In *International Conference on Machine Learning*, 2009. URL `https://api.semanticscholar.org/CorpusID:6463464`.

S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

H. Liu and P. Abbeel. Behavior from the void: Unsupervised active pre-training. *CoRR*, abs/2103.04551, 2021. URL `https://arxiv.org/abs/2103.04551`.

S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

A. Modirshanechi, S. Becker, J. Brea, and W. Gerstner. Surprise and novelty in the brain. *Current opinion in neurobiology*, 82:102758, 2023.

P. A. Morales and F. E. Rosas. Generalization of the maximum entropy principle for curved statistical manifolds. *Physical Review Research*, 3(3), sep 2021. doi: 10.1103/physrevresearch.3.033216. URL https://doi.org/10.1103%2Fphysrevresearch.3.033216.

P. A. Morales, J. Korbel, and F. E. Rosas. Geometric structures induced by deformations of the legendre transform. *Entropy*, 25(4):678, 2023a.

P. A. Morales, J. Korbel, and F. E. Rosas. Thermodynamics of exponential kolmogorov–nagumo averages. *New Journal of Physics*, 25(7):073011, 2023b.

M. Mutti, L. Pratissoli, and M. Restelli. A policy gradient method for task-agnostic exploration. In *4th Lifelong Machine Learning Workshop at ICML 2020*, 2020.

A. Nedergaard and M. Cook. k-means maximum entropy exploration. *arXiv preprint arXiv:2205.15623*, 2022.

OpenAI, C. Berner, G. Brockman, B. Chan, V. Cheung, P. Debiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, R. Józefowicz, S. Gray, C. Olsson, J. Pachocki, M. Petrov, H. P. de Oliveira Pinto, J. Raiman, T. Salimans, J. Schlatter, J. Schneider, S. Sidor, I. Sutskever, J. Tang, F. Wolski, and S. Zhang. Dota 2 with large scale deep reinforcement learning. *CoRR*, abs/1912.06680, 2019. URL http://arxiv.org/abs/1912.06680.

D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction, 2017. URL https://arxiv.org/abs/1705.05363.

A. Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press, 1961.

J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991.

W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275 (5306):1593–1599, 1997.

C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423, 1948.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

A. L. Strehl and M. L. Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

S. B. Thrun. *Efficient exploration in reinforcement learning*. Carnegie Mellon University, 1992.

O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 575(7782):350–354, 2019.

H. A. Xu, A. Modirshanechi, M. P. Lehmann, W. Gerstner, and M. H. Herzog. Novelty is not surprise: Human exploratory and adaptive behavior in sequential decision-making. *PLOS Computational Biology*, 17(6):e1009070, 2021.

M. Yuan, M.-O. Pun, and D. Wang. Rényi state entropy maximization for exploration acceleration in reinforcement learning. *IEEE Transactions on Artificial Intelligence*, 2022.

N. N. Čencov. *Statistical decision rules and optimal inference (in Russian)*. Nauka Moscow, 1972.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/**Not Applicable**]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/**Not Applicable**]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/**Not Applicable**]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/**Not Applicable**]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/**Not Applicable**]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/**Not Applicable**]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/**Not Applicable**]
   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/**Not Applicable**]
   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

An Information-Geometric Approach to Artificial Curiosity: Supplementary Materials

A Practical algorithms

The theory of information rewards leads to a class of practical exploration algorithms with plug-and-play functionality for reinforcement learning and density estimation methods. In finite state spaces, occupancy estimation simply amounts to normalized state counts, while the occupancy must be efficiently estimated using non-parametric density estimation methods when the state space is infinite. For the k-nearest neighbor [Mutti et al., 2020, Liu and Abbeel, 2021] and k-means [Nedergaard and Cook, 2022] density estimators from the maximum entropy exploration literature, the density estimate in a n-dimensional state space has the form

Algorithm 1 Reinforcement learning using artificial curiosity with information rewards.

hyperparameters
  information function f (strictly concave)
  intrinsic reward scaling β (non-negative)
end hyperparameters
variables
  transition function δ
  starting state density μ
  policy π
  occupancy estimate $\hat{p}_\pi$
end variables
for `iteration` **do**
  state $s \sim \mu$
  trajectory $\tau \leftarrow (s)$
  for `step` **do**
    action $a \sim \pi(s)$
    next state $s' \sim \delta(s, a)$
    extrinsic reward $r_e \sim r(s)$
    intrinsic reward $r_i \leftarrow f(\frac{1}{\hat{p}_\pi(s)})$
    $\tau \leftarrow \tau.(a, r_e + \beta r_i, s')$
    $s \leftarrow s'$
  end for
  $\pi \leftarrow \text{UpdatePolicy}(\pi, \tau)$
  $\hat{p}_\pi \leftarrow \text{UpdateDensityEstimate}(\hat{p}_\pi, \tau)$
end for

$$\hat{p}_\pi(s) \equiv \frac{1}{m(s)^n} \tag{42}$$

where $m(s) \in \mathbb{R}_{\geq 0}$ and $\equiv$ denotes equivalence up to affine transformations (such transformations $f(x) \mapsto \eta f(x) + c$ with $\eta, c \in \mathbb{R}$ do not affect policy gradients except for scaling them by η). Using such a density estimate, information rewards have the form $I_f(s; \hat{p}_\pi) \equiv f\left[m(s)^n\right]$ with f strictly concave. For maximum entropy exploration ($f = \log$), this leads to numerical instabilities when $m(s)$ is close to zero. To mitigate this, Liu and Abbeel [2021] implicitly used $f(x) = \log(x + 1)$ as a hack. However, this is a strictly concave function, so the hack is information-theoretically justified by our theory. Similarly, Nedergaard and Cook [2022] reported

best performance when implicitly using f_α with $\alpha = \frac{1}{n} - 1$ through $\sqrt{m(s)}$, since

$$I_\alpha(s; \hat{p}_\pi) \equiv \frac{4}{1-\alpha^2}([m(s)^n]^{\frac{\alpha+1}{2}} - 1) \equiv \sqrt{m(s)} \tag{43}$$

The theory of information rewards provides rigorous foundations for these empirically vital deviations from maximum entropy exploration theory. When maximizing the Renyi entropy of the occupancy, Yuan et al. [2022] found best performance when implicitly fine-tuning α for each environment. Information rewards provide an efficient method of achieving such performance gains, since α-information rewards have equivalent gradients but are computable in a distributed manner. An interesting direction of future research are algorithms that use α-information rewards with adaptive α and β, for example meta learning algorithms.

B Relation to non-equivalent methods

Artificial curiosity with α-information rewards is closely related to Rényi maximum entropy exploration [Yuan et al., 2022], Boltzmann exploration [Thrun, 1992] and classic artificial curiosity [Schmidhuber, 1991]. For Rényi maximum entropy exploration, which uses Rényi entropy instead of Shannon entropy in (11), we have:

Proposition B.1. *Artificial curiosity with α-information is equivalent to Rényi maximum entropy exploration, under non-constant scaling of β.*

From Lemma 4.3, the optima of information rewards with $\alpha = -1$ are equivalent to Boltzmann exploration:

$$\lim_{\alpha \to -1} p_{\alpha,\beta}(s) = \frac{e^{\frac{r(s)}{\beta}}}{\int_{\mathcal{S}} e^{\frac{r}{\beta}} dV} \tag{44}$$

The optima of Rényi maximum entropy exploration are similarly equivalent by Proposition B.1.

By extending α-information rewards with a discrete parameter, it is possible to include the classic artificial curiosity of Schmidhuber [1991]. In particular, consider artificial curiosity with α-information relative to the k-step agent-environment interaction M^k from some state s_0:

$$r(s) + \beta I_\alpha(s; M^k[s_0; \cdot]), \quad \alpha \in \mathbb{R},\ \beta \in \mathbb{R}_{\geq 0},\ k \in \mathbb{Z}_+,\ s_0 \in \mathcal{S}. \tag{45}$$

By definition of the occupancy, we have

$$\lim_{k \to \infty} I_\alpha(s; M^k[s_0; \cdot]) = I_\alpha(s; p_\pi) \tag{46}$$

independent of s_0, such that (45) contains count-based ($k = \infty, \alpha = 0$) and maximum entropy exploration ($k = \infty, \alpha = -1$) by Theorem 3.3. Furthermore, we have

Proposition B.2. *If the agent-environment interaction is a unit variance Gaussian with mean $f_\theta(s_0)$, (45) is equivalent to classic artificial curiosity when $k = 1$ and $\alpha = -1$.*

Count-based and maximum entropy exploration tend to empirically outperform classic artificial curiosity, suggesting that agent-environment interaction invariance is associated with superior empirical performance.

C Novelty and surprise are on a spectrum

Because information-processing in brains and machines are bound by the same mathematical structure, including the unique information-theoretic role of α-information, our results have implications for neuroscience. The firing rates of dopamine neurons in the primate ventral tegmental area are correlated with reward prediction errors [Schultz et al., 1997], but also with novel and surprising stimuli [Kakade and Dayan, 2002], suggestive of a neurobiological basis of curiosity in mammals. Several authors have argued for a functional and mechanistic distinction between dopaminergic novelty and surprise rewards [Barto et al., 2013, Xu et al., 2021, Modirshanechi et al., 2023]. However, our results suggest viewing novelty and surprise as α-information on a continuous spectrum determined by α. In particular, 0-information of neural populations representing generative models corresponds to novelty, and (-1)-information of neural populations representing predictive models corresponds to surprise (if the model is Gaussian, prediction errors). Intriguingly, empirical evidence suggests that human exploratory behavior is best explained by novelty rewards [Xu et al., 2021]. Experimental investigations are warranted on whether α-information rewards explain biological curiosity data better than other theories, and if so, whether inferred parameters for α and β are consistent with empirical and theoretical results from reinforcement learning.

D Proofs

Theorem 2.1. *The Markov chain formed by the time-inhomogeneous Markov kernel*

$$M_t(s, B) \coloneqq \begin{cases} \mu(B), & \text{if } t \text{ divides } n+1 \\ M(s, B), & \text{else} \end{cases} \tag{4}$$

with $n \in \mathbb{Z}_+$, converges to a unique probability measure

$$P_\pi(B) = \frac{1}{n+1} \int_{\mathcal{S}} \sum_{k=0}^{n} M^k(s, B) d\mu(s) \tag{5}$$

that is uniquely invariant under M_t.

Proof. This result is well-known, but we did not find a proof for general state spaces in the literature. We first construct a time-inhomogeneous Markov kernel equivalent to M_t and show that it forms a Markov chain. To do this, we use a counter set $I \coloneqq \{0, ..., n\}$ and define:

$$\tilde{\mathcal{S}} \coloneqq \mathcal{S} \times I \tag{47}$$

$$\tilde{\mu}(B) \coloneqq (\mu(B), 0) \tag{48}$$

$$\tilde{M}(s, B) \coloneqq \begin{cases} (\tilde{\mu}(B), 0) & \text{if } c(s) = n, \\ (M(s, B), c(s) + 1) & \text{else,} \end{cases} \tag{49}$$

where $c(s)$ gives the counter and $B \in \tilde{B}$. To ensure that $\tilde{M}$ forms a Markov chain given that M does, we need to handle a topological technicality: $\tilde{B}$, the Borel σ-algebra of $\tilde{S}$, must be countably generated. Let I have any topology and let $\tilde{\mathcal{S}}$ have the product topology. If M formed a Markov chain, $\mathcal{S}$ must have been second-countable. Since I is second-countable by finiteness, and the property is preserved by the product topology, $\tilde{\mathcal{S}}$ is second-countable. $\tilde{\mathcal{B}}$ is then countably generated and $\tilde{M}$ forms a Markov chain.

We now construct P_π and show that it is the unique invariant probability measure for the Markov chain. Let

$$\tilde{M}^k_{\mathrm{X}}(s, B) \coloneqq \int_{\mathcal{S} \setminus \mathrm{X}} \tilde{M}^{k-1}_{\mathrm{X}}(s, s') \tilde{M}(s', B) dV(s'), \quad \tilde{M}^1_{\mathrm{X}}(s, B) \coloneqq \tilde{M}(s, B) \tag{50}$$

denote the probability of visiting $B \in \tilde{\mathcal{B}}$ from $s \in \mathcal{S}$ in k steps without passing through $\mathrm{X} \subseteq \tilde{\mathcal{B}}$. Let

$$\Xi \coloneqq \{s \in \tilde{\mathcal{S}} : c(s) = n\} \tag{51}$$

denote the terminal states. Now, define the expected number of visits to B without passing through a terminal state:

$$\tilde{P}_\pi(B) \coloneqq \int_{\tilde{\mathcal{S}}} \sum_{k=0}^{\infty} \tilde{M}^k_\Xi(s, B) d\mu(s) \tag{52}$$

Since we are guaranteed to visit Ξ in exactly n steps from a starting state, we have

$$\tilde{P}_\pi(B) = \int_{\tilde{\mathcal{S}}} \Big(\sum_{k=0}^{n} \tilde{M}^k_\Xi(s, B) + \sum_{k=n+1}^{\infty} \tilde{M}^k_\Xi(s, B) \Big) d\mu(s) \tag{53a}$$

$$= \int_{\tilde{\mathcal{S}}} \sum_{k=0}^{n} \tilde{M}(s, B) d\mu(s) \geq 0 \tag{53b}$$

since only the first n-terms of the sum contribute. We can normalize $\tilde{P}_\pi$ to get a probability measure:

$$P_\pi(B) = \frac{\tilde{P}_\pi(B)}{\tilde{P}_\pi(\mathcal{S})} = \frac{\int_{\tilde{\mathcal{S}}} \sum_{k=0}^{n} \tilde{M}^k(s, B) d\tilde{\mu}(s)}{\int_{\tilde{\mathcal{S}}} \sum_{k=0}^{n} \tilde{M}^k(s, \mathcal{S}) d\tilde{\mu}(s)} \tag{54a}$$

$$= \frac{1}{n+1} \int_{\tilde{\mathcal{S}}} \sum_{k=0}^{n} \tilde{M}^k(s, B) d\tilde{\mu}(s) \tag{54b}$$

$\tilde{P}_\pi$-irreducibility and recurrence of the Markov chain implies the existence of a unique, up to constant multiplies, invariant measure equivalent to $\tilde{P}_\pi$ [Meyn and Tweedie, 2012, Theorem 10.4.9]. Therefore, P_π is the unique invariant probability measure if the chain is $\tilde{P}_\pi$-irreducible and recurrent. Aperiodicity is not necessary because we have already constructed the measure as a finite sum. By Meyn and Tweedie [2012, Proposition 4.2.1], $\tilde{P}_\pi$-irreducibility and recurrence follows from

$$\forall s \in \mathcal{S}.\, \forall B \in \tilde{\mathcal{B}}.\, \tilde{P}_\pi(B) > 0 \Rightarrow \sum_{k=1}^{\infty} \tilde{M}^k(s, B) = \infty \tag{55}$$

Since the state resets every $n+1$ steps, we have for any k and $m \in \mathbb{N}$,

$$\tilde{M}^k(s, B) = \tilde{M}^{k+m(n+1)}(s, B) \tag{56}$$

Letting $l \coloneqq n - c(s)$ denote the number of steps until the next state reset, we furthermore have for $k > l$,

$$\tilde{M}^k(s, B) = \int_S \tilde{M}^{k-(l+1)}(s', B) d\tilde{M}^{l+1}(s, s') \tag{57a}$$

$$= \int_S \tilde{M}^{k-(l+1)}(s', B) d\tilde{\mu}(s') \tag{57b}$$

since a state reset happens after l steps. Then,

$$\sum_{k=1}^{\infty} \tilde{M}^k(s, B) = \underbrace{\sum_{k=1}^{l+1} \tilde{M}^k(s, B)}_{\geq 0} + \sum_{k=l+1}^{\infty} \tilde{M}^k(s, B) \geq \sum_{k=l+1}^{\infty} \tilde{M}^k(s, B), \tag{58}$$

as a bound from below,

$$\sum_{k=l+1}^{\infty} \tilde{M}^k(s, B) \overset{(57b)}{=} \sum_{k=l+1}^{\infty} \int_{\tilde{\mathcal{S}}} \tilde{M}^{k-(l+1)}(s', B) d\tilde{\mu}(s') \tag{59a}$$

$$= \sum_{k=0}^{\infty} \int_{\tilde{\mathcal{S}}} \tilde{M}^k(s', B) d\tilde{\mu}(s') \tag{59b}$$

$$= \sum_{m=0}^{\infty} \sum_{k=0}^{n} \int_{\tilde{\mathcal{S}}} \tilde{M}^{k+m(n+1)}(s', B) d\tilde{\mu}(s') \tag{59c}$$

$$\overset{(56)}{=} \sum_{m=0}^{\infty} \sum_{k=0}^{n} \int_{\tilde{\mathcal{S}}} \tilde{M}^k(s', B) d\tilde{\mu}(s') \tag{59d}$$

$$= \lim_{m \to \infty} m \int_{\tilde{\mathcal{S}}} \sum_{k=0}^{n} \tilde{M}^k(s', B) d\tilde{\mu}(s') \tag{59e}$$

$$\overset{(53b)}{=} \lim_{m \to \infty} m \tilde{P}_\pi(B) = \begin{cases} \infty & \text{if } \tilde{P}_\pi(B) > 0, \\ 0 & \text{else.} \end{cases} \tag{59f}$$

The Markov chain is then $\tilde{P}_\pi$-irreducible and recurrent, and P_π is the unique invariant probability measure under $\tilde{M}$. Since $M = \tilde{M}$ when $c(s) < n$, and $c(s) = 0$ for any $s \in \text{supp}(\tilde{\mu})$, P_π is invariant under M and it follows from (54b) that

$$P_\pi(B) = \frac{1}{n+1} \int_{\mathcal{S}} \sum_{k=0}^{n} M^k(s, B) d\mu(s). \tag{60}$$

□

Lemma 2.2.

$$R(p_\pi) \coloneqq (n+1) \int_{\mathcal{S}} p_\pi(s) r(s) dV(s) = R(\pi). \tag{7}$$

Proof. The proof is due to [Bojun, 2020]. Define the state-action value

$$Q(s,a) \coloneqq \int_{\mathcal{S}^{n-k}} \sum_{i=k}^{n} r(s_i) d\delta(s,a,s_k) dM(s_k,s_{k+1}) \cdots dM(s_{n-1},s_n) \tag{61}$$

where the notation $\int_{\mathcal{S}^{n-k}} \coloneqq \prod^{n-k} \int_{\mathcal{S}}$ had been adopted. $Q(s,a)$ satisfies,

$$R(\pi) = \int_{\mathcal{S}} \left\{ r(s) + \int_{\mathcal{A}} Q(s,a) d\pi(s,a) \right\} d\mu(s) \tag{62a}$$

$$= \int_{\mathcal{A}} Q(\xi,a) d\pi(\xi,a) \tag{62b}$$

for any terminal state $\xi \in \Xi$. The Bellman equation [Bellman, 1954] shows that

$$Q(s,a) = \int_{\mathcal{S}} \left\{ r(s') + \gamma(s') \int_{\mathcal{A}} Q(s',a') d\pi(s',a') \right\} d\delta(s,a,s') \tag{63}$$

where

$$\gamma(s) \coloneqq \begin{cases} 1 & \text{if } s \notin \Xi, \\ 0 & \text{else.} \end{cases} \tag{64}$$

Integrating both sides of (63) over dP_π and $d\pi$, we have

$$\int_{\mathcal{S}\times\mathcal{A}} Q(s,a) d\pi(s,a) dP_\pi(s) = \int_{\mathcal{A}\times\mathcal{S}^2} \left\{ r(s') + \gamma(s') \int_{\mathcal{A}} Q(s',a') d\pi(s',a') \right\} d\delta(s,a,s') dP_\pi(s) d\pi(s,a) \tag{65a}$$

$$= \int_{\mathcal{S}^2} \left\{ r(s') + \gamma(s') \int_{\mathcal{A}} Q(s',a') d\pi(s',a') \right\} dM(s,s') dP_\pi(s) \tag{65b}$$

$$= \int_{\mathcal{S}} \left\{ r(s) + \gamma(s) \int_{\mathcal{A}} Q(s,a) d\pi(s,a) \right\} dP_\pi(s) \tag{65c}$$

$$= \int_{\mathcal{S}} r(s) dP_\pi(s) + \int_{\mathcal{S}} \gamma(s) \int_{\mathcal{A}} Q(s,a) d\pi(s,a) dP_\pi(s) \tag{65d}$$

where we used the invariance of P_π by Theorem 2.1 in the second-last line. Subtracting the second term on the right-hand side from both sides,

$$\int_{\mathcal{S}} r(s) dP_\pi(s) = \int_{\mathcal{S}} \left\{ [1-\gamma(s)] \int_{\mathcal{A}} Q(s,a) d\pi(s,a) \right\} dP_\pi(s) \tag{66a}$$

$$= \int_{\Xi\times\mathcal{A}} Q(s,a) d\pi(s,a) dP_\pi(s) \tag{66b}$$

$$\overset{(62b)}{=} R(\pi) \int_{\Xi} dP_\pi(s) = \frac{1}{n+1} R(\pi) \tag{66c}$$

this finally implies that

$$R(\pi) = (n+1) \int_{\mathcal{S}} r(s) dP_\pi(s) \tag{67a}$$

$$= (n+1) \int_{\mathcal{S}} p_\pi(s) r(s) dV(s). \tag{67b}$$

□

Lemma 2.4. *The geodetic property of divergences is preserved by strictly monotonic functions.*

Proof. Let $\bar{\mathcal{D}}$ be a geodetic divergence and $\mathcal{D} = F(\bar{\mathcal{D}})$ with F strictly monotonic function. Using the shorthand $\mathcal{D}_p := \mathcal{D}(p\|\cdot)$, letting g denote the Fisher-Rao tensor, and letting γ denote a unit speed path with arbitrary endpoints p and q,

$$\begin{aligned}\bar{\mathcal{D}}_p(q) &= \bar{\mathcal{D}}_p(q) - \underbrace{\bar{\mathcal{D}}_p(p)}_{=0} \\ &= \int_\gamma d\bar{\mathcal{D}}_p = \int_\gamma g_\gamma(\operatorname{grad}\bar{\mathcal{D}}_p, \gamma')\end{aligned} \tag{68}$$

and

$$\begin{aligned}\bar{\mathcal{D}}_p(q) &= \int_\gamma d\bar{\mathcal{D}}_p = \int_\gamma d(F(\mathcal{D}_p)) = \int_\gamma F'(\mathcal{D}_p) d\mathcal{D}_p \\ &= \int_\gamma F'(\mathcal{D}_p) g_\gamma(\operatorname{grad}\mathcal{D}_p, \gamma') = \int_\gamma g_\gamma(F'(\mathcal{D}_p)\operatorname{grad}\mathcal{D}_p, \gamma')\end{aligned} \tag{69}$$

where $\int_\gamma$ denotes a line integral. Since it then holds for any unit speed path γ with arbitrary endpoints p and q that

$$\int_\gamma g_\gamma(\operatorname{grad}\bar{\mathcal{D}}_p, \gamma') = \int_\gamma g_\gamma(F'(\mathcal{D}_p)\operatorname{grad}\mathcal{D}_p, \gamma') \tag{70}$$

it must be that

$$\operatorname{grad}\bar{\mathcal{D}}(p,\cdot) = F'(\mathcal{D}(p,\cdot))\operatorname{grad}\mathcal{D}(p,\cdot) \tag{71}$$

for any p. Since additionally $F'(\mathcal{D}(p\|\cdot)) > 0$ by strict monotonicity of F, $\mathcal{D}$ must be geodetic if $\bar{\mathcal{D}}$ is. □

Lemma 3.1. *The family $\{f_\alpha : \alpha \in \mathbb{R}\}$ is uniquely geodetic, up to transformations $f_\alpha(x) \mapsto \eta\{f_\alpha(x) + c(x-1)\}$ with $\eta, c \in \mathbb{R}$.*

Proof. Recall that a function f is geodetic if its induced divergence $\mathcal{D}_f$ satisfies:

$$\operatorname{grad}\mathcal{D}_f(p\|\cdot)\big|_q = -\eta\,\frac{d}{dt}\gamma(t)\Big|_{t=0} \tag{72}$$

where $\eta \in \mathbb{R}_{>0}$ and γ is the primal geodesic, connecting q and p, induced by $\mathcal{D}_f$ in the space of measures. By Ay et al. [2017, Equation 2.59], γ is given by

$$\gamma(t) = \left\{(1-t)q^{\frac{1-\alpha}{2}} + tp^{\frac{1-\alpha}{2}}\right\}^{\frac{2}{1-\alpha}} \tag{73}$$

such that

$$\frac{d}{dt}\gamma(t)\Big|_{t=0} = \frac{2}{1-\alpha}\left[(1-t)q^{\frac{1-\alpha}{2}} + tp^{\frac{1-\alpha}{2}}\right]^{\frac{1+\alpha}{1-\alpha}}(p^{\frac{1-\alpha}{2}} - q^{\frac{1-\alpha}{2}})\Big|_{t=0} \tag{74a}$$

$$= \frac{2}{1-\alpha}q^{\frac{1+\alpha}{2}}(p^{\frac{1-\alpha}{2}} - q^{\frac{1-\alpha}{2}}) \tag{74b}$$

$$= \frac{2}{1-\alpha}q\left(\left[\frac{p}{q}\right]^{\frac{1-\alpha}{2}} - 1\right) \tag{74c}$$

We proceed by equating this to the gradient of the f-divergence and showing that this uniquely constrains f. Some technical machinery is required to handle the gradient on a statistical manifold when the underlying state space is infinite. We follow Ay et al. [2017] and consider a 2-integrable parameterized measure model $(\Theta, \mathcal{S}, \mathbf{p})$ with a smooth map $\mathbf{p} : \Theta \to \mathcal{M}$ between the Banach manifolds of parameters Θ and measures $\mathcal{M}$ over the state space $\mathcal{S}$. The Fisher-Rao tensor on Θ, g, is then rigorously defined as the pullback of a natural metric tensor on $\mathcal{M}^{1/2}$ along the smooth map $\sqrt{\mathbf{p}} : \Theta \to \mathcal{M}^{1/2}$:

$$g_\theta(v, w) = 4\int_{\mathcal{S}} d(d_\theta\sqrt{\mathbf{p}}(v) d_\theta\sqrt{\mathbf{p}}(w)) \tag{75}$$

Letting ∂_v denote the Gateaux derivative with respect of to $v \in T_\theta\Theta$, and letting the Radon-Nikodym derivative $q = \frac{d\mathbf{p}(\theta)}{dV}$, the gradient of $\mathcal{D}_f(p\|q)$ at θ is characterized similarly to on Riemannian manifolds:

$$\partial_v \mathcal{D}_f(p\|q) = g_\theta(v, \operatorname{grad} \mathcal{D}_f(p\|q)) \tag{76}$$

Now, observe that

$$\begin{aligned}
\partial_v D_f(p\|q) &= \partial_v \int_{\mathcal{S}} p(s) f\left[\frac{q(s)}{p(s)}\right] dV(s) && \text{(77a)}\\
&= \int_{\mathcal{S}} \{\partial_v q(s)\} f'\left[\frac{q(s)}{p(s)}\right] dV(s) && \text{(77b)}\\
&= 4\int_{\mathcal{S}} d\left(\left\{\frac{\sqrt{V(s)}}{2\sqrt{q(s)}}\partial_v q(s)\right\}\left\{\frac{\sqrt{V(s)}}{2\sqrt{q(s)}} q(s) f'\left[\frac{q(s)}{p(s)}\right]\right\}\right) && \text{(77c)}\\
&= 4\int_{\mathcal{S}} d\left(d_\theta\sqrt{\mathbf{p}}(v) d_\theta\sqrt{\mathbf{p}}(w)\right) && \text{(77d)}
\end{aligned}$$

with $w \in T_\theta\Theta$ satisfying $d_\theta \mathbf{p}(w) = q f'\left[\frac{q}{p}\right] V$. Comparing to (75) and (76), we see that $w = \operatorname{grad} D_f(p\|q)$ on Θ and that the Radon-Nikodym derivative (with respect to V) of its pushforward to $\mathcal{M}$ (along $\mathbf{p}$) is given by $q f'\left[\frac{q}{p}\right]$. We need the gradient on $\mathcal{M}$, rather than Θ, in order to relate it to the derivative $\frac{d}{dt}\gamma(0) \in T_{\gamma(0)}\mathcal{M}$. Now, suppose that the function f is geodetic, then

$$q f'\left[\frac{q}{p}\right] = -\eta \frac{2}{1-\alpha} q\left(\left[\frac{p}{q}\right]^{\frac{1-\alpha}{2}} - 1\right) \Leftrightarrow f'(x) = \eta \frac{2}{1-\alpha}(1 - x^{\frac{\alpha-1}{2}}). \tag{78}$$

Integrating both sides,

$$f(x) = \eta\left(-\frac{4}{1-\alpha^2} x^{\frac{\alpha+1}{2}} + \frac{2}{1-\alpha} x\right) + C. \tag{79}$$

Since $f(1) = 0$ as f induces an f-divergence, we obtain $C = \frac{2\eta}{\alpha+1}$. The theorem then follows from

$$f(x) = \eta\left(-\frac{4}{1-\alpha^2} x^{\frac{\alpha+1}{2}} + \frac{2}{1-\alpha} x + \frac{2}{1+\alpha}\right) = \eta\left\{f_\alpha(x) + \frac{2}{1-\alpha}(x-1)\right\}. \tag{80}$$

□

Theorem 3.2. *Intrinsic rewards $\bar{r}(s;p) \coloneqq \bar{f}[p(s)]$, with any function $\bar{f} : \mathbb{R} \to \mathbb{R}$ and any probability density p, are invariant under the agent-environment interaction M,*

$$\bar{r}(s;p) = \bar{r}(s; h_M(p)), \tag{24}$$

uniquely when $p = p_\pi$. Furthermore, for intrinsic rewards $\bar{r}(s) \coloneqq \bar{r}(s;p_\pi)$, the intrinsic return

$$\bar{R}(p_\pi) = \int_{\mathcal{S}^{n+1}} \sum_{i=0}^{n} \bar{r}(s_i) d\mu(s_0) \prod_{i=1}^{n} dM(s_{i-1}, s_i) \tag{25}$$

satisfies information monotonicity

$$\bar{R}(p_\pi) \leq \bar{R}(h_\kappa(p_\pi)), \tag{26}$$

with equality if and only if the statistic κ is sufficient, uniquely when $\bar{r}(s) = I_f(s;p_\pi)$.

Proof. To be explicit,

$$\bar{R}(p_\pi) \coloneqq (n+1)\int_{\mathcal{S}} p_\pi(s)\bar{r}(s) dV(s) = \int_{\mathcal{S}^{n+1}} \sum_{i=0}^{n} \bar{r}(s_i) d\mu(s_0) \prod_{i=1}^{n} dM(s_{i-1}, s_i) \tag{81}$$

where the equality holds by Lemma 2.2. The unique invariance under the agent-environment interaction when $p = p_\pi$ follows from the unique invariance of p_π under the agent-environment interaction by Theorem 2.1. For

information monotonicity, we may consider $\bar{f}\left[p_\pi(s)\right] = f\left[p_\pi(s)^{-1}\right]$ for some function f without loss of generality, since the correspondence between $\bar{f}$ and f is bijective. The proof amounts to showing that f satisfying information monotonicity is equivalent to strict concavity of f. Define $P' \coloneqq h_\kappa(P)$ and (abusing notation) $p' \coloneqq h_\kappa(p)$, recalling that h_κ denotes the Markov morphism induced by the statistic κ. We need to show that strictly concave f uniquely satisfy

$$\int_{\mathcal{S}} f\left[\frac{1}{p(s)}\right] dP(s) \overset{!}{\leq} \int_{\kappa(\mathcal{S})} f\left[\frac{1}{p'(s)}\right] dP'(s) \tag{82}$$

with equality if and only if the statistic κ is sufficient. We use ! to indicate that a relation does not hold in general. By [Ay et al., 2017, Proposition 5.5-5.6], the statistic κ is sufficient if and only if there exists $\omega : \mathcal{S} \to \mathbb{R}$ such that

$$p_\theta(s) \overset{!}{=} \omega(s) p'_\theta(\kappa(s)) \tag{83}$$

with ω independent of the probability density parameterization θ. If the statistic κ is injective, we trivially have sufficiency of the statistic with equality in (82). Assume therefore that the statistic κ is not injective, such that at least one set of states $\Phi \subseteq \mathcal{S}$ with $|\Phi| > 1$ is mapped to a single state $\kappa(\Phi)$. Note that, by definition, $P(\Phi) = P'(\kappa(\Phi))$, Therefore, it suffices to consider the information monotonicity for the case of Φ, since each such case corresponds to a certain amount of probability measure contributing additively to both sides of (82). Now, observe that

$$\begin{aligned}\int_{\kappa(\Phi)} f\left[\frac{1}{p'(s)}\right] dP'(s) &= \int_{\kappa(\Phi)} f\left[\frac{dV'(s)}{dP'(s)}\right] dP'(s) \\ &= P'(\kappa(\Phi)) f\left[\frac{V'(\kappa(\Phi))}{P'(\kappa(\Phi))}\right] = P(\Phi) f\left[\frac{V(\Phi)}{P(\Phi)}\right]\end{aligned} \tag{84}$$

and notice that the argument of f can be rewritten as

$$\frac{V(\Phi)}{P(\Phi)} = \frac{\int_\Phi dV(s)}{P(\Phi)} = \int_\Phi \frac{p(s)}{P(\Phi)} \frac{1}{p(s)} dV(s) = \int_\Phi \frac{1}{p(s)} \frac{dP(s)}{P(\Phi)} \tag{85}$$

Combining (84) and (85) with (82), we see that theorem follows from showing that strictly concave f uniquely satisfy

$$\int_\Phi f\left[\frac{1}{p(s)}\right] dP(s) = P(\Phi) \int_\Phi f\left[\frac{1}{p(s)}\right] \frac{dP(s)}{P(\Phi)} \tag{86a}$$

$$\overset{!}{\leq} P(\Phi) f\left[\int_\Phi \frac{1}{p(s)} \frac{dP(s)}{P(\Phi)}\right] = \int_{\kappa(\Phi)} f\left[\frac{1}{p'(s)}\right] dP'(s) \tag{86b}$$

with equality if and only if κ is a sufficient statistic. Notice that satisfying the inequality with equality if and only if $p(s_1)^{-1} = p(s_2)^{-1}$ for any $s_1, s_2 \in \Phi$, is equivalent to strict concavity by Jensen's inequality. Therefore, the theorem follows from showing that $p(s_1) = p(s_2)$ for any $s_1, s_2 \in \Phi$ is equivalent to sufficiency of κ. Since our theorem must hold when $\mathcal{P}$ is the space of all probability measures, we may consider arbitrary parameterizations θ in (83) without loss of generality. In particular, consider $s_1, s_2 \in \Phi$ and $p_{\theta_1}, p_{\theta_2} \in \mathcal{P}$ such that $p_{\theta_1}(s_2) = p_{\theta_2}(s_1)$. Then

$$\frac{p_{\theta_1}(s_1)}{p'_{\theta_1}(\kappa(s_1))} \overset{!}{=} \frac{p_{\theta_2}(s_1)}{p'_{\theta_2}(\kappa(s_1))} = \frac{p_{\theta_1}(s_2)}{p'_{\theta_1}(\kappa(s_2))} \Leftrightarrow p_{\theta_1}(s_1) \overset{!}{=} p_{\theta_1}(s_2) \tag{87}$$

since $p'_{\theta_1}(\kappa(s_1)) = p'_{\theta_1}(\kappa(s_2))$, which shows that κ is sufficient if and only if $p(s_1) = p(s_2)$ for any $s_1, s_2 \in \Phi$. □

Theorem 3.3. *Artificial curiosity with*

$$r(s) + \beta I_\alpha(s; p_\pi), \quad \alpha \in \mathbb{R},\ \beta \in \mathbb{R}_{\geq 0} \tag{28}$$

is equivalent to count-based exploration for $\alpha = 0$ and maximum entropy exploration for $\alpha = -1$.

Proof. We first prove the case $\alpha = 0$. From the definition of α-information,

$$I_0(s; p) \coloneqq I_{f_{\alpha=0}}(s; p) = 4\left(\sqrt{\frac{1}{p(s)}} - 1\right). \tag{88}$$

For a finite state space $\mathcal{S}$, we have

$$p_\pi(s) = \frac{n(s)}{n+1} \tag{89}$$

where $n(s)$ is the state count for state s and $(n+1)$ is the total state count. The result then follows from

$$I_0(s, p_\pi) = \sqrt{\frac{16(n+1)}{n(s)}} - 4, \tag{90}$$

since $\sqrt{16(n+1)}$ is a constant which can be absorbed into β and 4 is a constant that disappears under gradients. We now prove the case $\alpha = -1$. From Lemma 2.2, we have

$$\begin{aligned}
R_{\alpha,\beta}(\pi) &= \int_{\mathcal{S}^{n+1}} R_{\alpha,\beta}(s; p_\pi) d\mu(s_0) \prod_{i=1}^{n} dM(s_{i-1}, s_i) && \text{(91a)} \\
&= \int_{\mathcal{S}^{n+1}} \{r(s) + \beta I_\alpha(s, p_\pi)\} \, d\mu(s_0) \prod_{i=1}^{n} dM(s_{i-1}, s_i) && \text{(91b)} \\
&= \int_{\mathcal{S}^{n+1}} r(s) d\mu(s_0) \prod_{i=1}^{n} dM(s_{i-1}, s_i) + \beta \int_{\mathcal{S}^{n+1}} I_\alpha(s, p_\pi) d\mu(s_0) \prod_{i=1}^{n} dM(s_{i-1}, s_i) && \text{(91c)} \\
&= R(\pi) + \beta \int_{\mathcal{S}^{n+1}} I_\alpha(s, p_\pi) d\mu(s_0) \prod_{i=1}^{n} dM(s_{i-1}, s_i) && \text{(91d)} \\
&= R(\pi) + \beta(n+1) \int_{\mathcal{S}} p_\pi(s) I_\alpha(s, p_\pi) dV(s) && \text{(91e)}
\end{aligned}$$

Now,

$$\lim_{\alpha \to -1} \int_{\mathcal{S}} p_\pi(s) I_\alpha(s, p_\pi) dV(s) = \int_{\mathcal{S}} p_\pi(s) \log \frac{1}{p_\pi(s)} dV(s) = H(p_\pi). \tag{92}$$

Using (92) and (91e), the statement then follows from

$$\lim_{\alpha \to -1} R_{\alpha,\beta}(\pi) = R(\pi) + \beta(n+1) H(p_\pi) \tag{93a}$$

since $(n+1)$ is a constant that can be absorbed into β. □

Lemma 4.1. *If P_Π is convex,*

$$p_{\alpha,\beta} = \arg\min_{p \in \mathcal{H}_{\alpha,\beta}} \mathcal{D}_\alpha(p \| u), \tag{32}$$

Proof. From Lemma 2.2, we have using the same derivation as (91e),

$$\begin{aligned}
R_{\alpha,\beta}(p) &= R(\pi) + \beta(n+1) \int_{\mathcal{S}} p_\pi(s) I_\alpha(s, p_\pi) dV(s) && \text{(94a)} \\
&= R(\pi) - \beta(n+1) u^{-\frac{\alpha+1}{2}} \mathcal{D}_\alpha(p \| u) + \beta(n+1)(1 - u^{-\frac{\alpha+1}{2}}) && \text{(94b)}
\end{aligned}$$

Then, the optima

$$\begin{aligned}
p_{\alpha,\beta} &= \arg\max_{p \in \mathcal{P}_\Pi} R_{\alpha,\beta}(p) && \text{(95a)} \\
&= \arg\max_{p \in \mathcal{P}_\Pi} \left\{ R(p) - \beta(n+1) u^{-\frac{\alpha+1}{2}} \mathcal{D}_\alpha(p \| u) \right\} && \text{(95b)} \\
&= \arg\min_{p \in \mathcal{P}_\Pi} \left\{ \mathcal{D}_\alpha(p \| u) - \frac{u^{\frac{\alpha+1}{2}}}{\beta(n+1)} R(p) \right\} && \text{(95c)} \\
&= \arg\min_{p \in \mathcal{P}_\Pi} \left\{ \mathcal{D}_\alpha(p \| u) - \frac{u^{\frac{\alpha+1}{2}}}{\beta(n+1)} [R(p) - c] \right\} && \text{(95d)}
\end{aligned}$$

with $c \in \mathbb{R}$ a constant. Interpreting $-\frac{u^{\frac{\alpha+1}{2}}}{\beta(n+1)}$ as a Lagrange multiplier, we have

$$p_{\alpha,\beta} = \underset{p \in \{p \in \mathcal{P} : R(p) = c_{\alpha,\beta}\}}{\arg\min} \mathcal{D}_\alpha(p\|u) = \underset{p \in \mathcal{H}(c_{\alpha,\beta})}{\arg\min}\ \mathcal{D}_\alpha(p\|u). \tag{96}$$

Note that we used convexity of $\mathcal{P}_\Pi$ to ensure strong duality, establishing the correspondence between the Lagrangian and the constrained optimization problem. □

Proposition 4.2. *If P_Π is convex, the optima $p_{\alpha,\beta}$ are α-projections from the uniform occupancy u onto isoreturn hyperplanes $\mathcal{H}_{\alpha,\beta}$.*

Proof. The result follows from Lemma 4.1 and the α-projection theorem [Amari, 2016, Theorem 4.6]. □

Lemma 4.3. *If $\mathcal{P}_\Pi = \mathcal{P}$,*

$$p_{\alpha,\beta} = \frac{(1 + \frac{1}{\lambda\beta} r)^{-\frac{2}{\alpha+1}}}{\int (1 + \frac{1}{\lambda\beta} r)^{-\frac{2}{\alpha+1}} dV} \tag{33}$$

with $\lambda \in \mathbb{R}$ determined by α and β.

Proof. From (95d), we have

$$p_{\alpha,\beta} = \arg\min_{p \in \mathcal{P}} \left\{ \mathcal{D}_\alpha(p\|u) - \frac{u^{\frac{\alpha+1}{2}}}{\beta(n+1)}[R(p) - c] \right\} \tag{97a}$$

$$= \arg\min_{p \in \mathcal{P}} \left\{ -\frac{4}{1-\alpha^2} \int p(s)^{\frac{1-\alpha}{2}} dV(s) - \frac{1}{\beta}\Big(\int p(s) r(s) dV(s) - c \Big) \right\} \tag{97b}$$

These distributions are determined by minimizing a Lagrangian functional, composed of an entropy function and constrained by a fixed expected reward and a normalization condition.

$$L[p] = -\frac{4}{1-\alpha^2} \int p^{\frac{1-\alpha}{2}} dV - \frac{1}{\beta}\Big(\int pr dV - c \Big) + \lambda \Big(\int p dV - 1 \Big) \tag{98}$$

where λ is a Lagrange multiplier. The minima is obtained by enforcing $\delta L = 0$, where

$$\delta L[p] = \int \Big[-\frac{2}{\alpha+1} p^{-\frac{\alpha+1}{2}} - \frac{r}{\beta} + \lambda \Big] \delta p \, dV = 0 \tag{99}$$

this requires that function inside the squared brackets vanishes, leading to

$$p(s) = \left[\frac{\lambda(\alpha+1)}{2} \right]^{-\frac{2}{\alpha+1}} \Big(1 - \frac{r(s)}{\lambda\beta} \Big)^{-\frac{2}{\alpha+1}}. \tag{100}$$

From the normalization of the probability distribution (100), $\int p dV = 1$, one finds

$$\left[\frac{\lambda(\alpha+1)}{2} \right]^{-\frac{2}{\alpha+1}} = \frac{1}{\int (1 - \frac{r}{\lambda\beta})^{-\frac{2}{\alpha+1}} dV} \tag{101}$$

where we used (100). Substituting (101) into (100), we get

$$p_{\alpha,\beta} = \frac{(1 - \frac{r}{\lambda\beta})^{-\frac{2}{\alpha+1}}}{\int (1 - \frac{r}{\lambda\beta})^{-\frac{2}{\alpha+1}} dV} \tag{102}$$

from which the result follows since a reparameterization $\lambda \leftarrow -\lambda$ does not change the form. □

Theorem 4.4. *If $\mathcal{P}_\Pi = \mathcal{P}$ and $\beta \geq \beta_0 > 0$, the exploration-exploitation relation $\beta \mapsto p_{f,\beta}$ is an $(\alpha+2)$-geodesic connecting the maximally rewarding occupancy p_{α,β_0} and the uniform occupancy u, uniquely when $f = f_\alpha$.*

Proof. We prove first the forward direction by cases, assuming $f = f_\alpha$. Assume first that $\alpha \neq -1$. The $(\alpha+2)$-geodesic connecting p and q is given by

$$\gamma_{p,q}(t) = \left\{(1-t)p^{-\frac{1+\alpha}{2}} + tq^{-\frac{1+\alpha}{2}}\right\}^{-\frac{2}{1+\alpha}} \xi(t) \tag{103}$$

where $t \in [0,1]$ and $\xi(t)$ normalizes $\gamma_{p,q}$ at t [Ay et al., 2017, Equation 2.59]. Let $\beta_0, \beta_1 \in (0,\infty)$ parameterize two points $p_0 = p_{\alpha,\beta_0}$ and $p_1 = p_{\alpha,\beta_1}$ connected by γ_{p_0,p_1}. By Lemma 4.3, $p_{0,1}^{-\frac{1+\alpha}{2}} = \xi_{0,1}(1+\mu_{0,1}r)$

$$\gamma_{p_0,p_1}(t) = \{(1-t)(1+\mu_0 r)\xi_0 + t(1+\mu_1 r)\xi_1\}^{-\frac{2}{1+\alpha}} \xi(t), \tag{104}$$

with $\mu_{0,1} = \frac{1}{\lambda\beta_{0,1}}$ adopted for brevity. Grouping the terms linear in r,

$$\begin{aligned}
\gamma_{p_0,p_1}(t) &= \left\{1 + \left[\left(1 - \frac{t\xi_1}{(1-t)\xi_0 + t\xi_1}\right)\mu_0 + \frac{t\xi_1}{(1-t)\xi_0+t\xi_1}\mu_1\right] r\right\}^{-\frac{2}{1+\alpha}} \\
&\quad [(1-t)\xi_0 + t\xi_1]^{-\frac{2}{1+\alpha}} \xi(t) && \text{(105a)} \\
&= \left\{1 + \left[(1-\hat{t})\mu_0 + \hat{t}\mu_1\right] r\right\}^{-\frac{2}{\alpha+1}} [(1-t)\xi_0 + t\xi_1]^{-\frac{2}{\alpha+1}} \xi(t) && \text{(105b)} \\
&= \left\{1 + \left[(1-\hat{t})\mu_0 + \hat{t}\mu_1\right] r\right\}^{-\frac{2}{\alpha+1}} \hat{\xi}(\hat{t}) && \text{(105c)}
\end{aligned}$$

where $\hat{t} \coloneqq \frac{t\xi_1}{(1-t)\xi_0+t\xi_1}$ and $\hat{\xi}(\hat{t})$ ensures normalization of (105c).

$$\begin{aligned}
\gamma_{p_0,p_1}(\hat{t})^{-\frac{\alpha+1}{2}} &= \left\{1 + \left[(1-\hat{t})\mu_0 + \hat{t}\mu_1\right] r\right\} \hat{\xi}(\hat{t}) && \text{(106a)} \\
&= \left\{1 + \left[(1-\hat{t})\frac{1}{\lambda\beta_0} + \hat{t}\frac{1}{\lambda\beta_1}\right] r\right\} \xi(t) && \text{(106b)}
\end{aligned}$$

leading to

$$\gamma_{p_0,p_1}(\hat{t}) = \left\{1 + \left[(1-\hat{t})\frac{1}{\lambda\beta_0} + \hat{t}\frac{1}{\lambda\beta_1}\right] r\right\}^{-\frac{2}{\alpha+1}} \phi(t) \tag{107}$$

where $\phi(t) \coloneqq \xi(t)^{-\frac{2}{\alpha+1}}$ normalizes $\gamma_{p_0,p_1}(\hat{t})$. Then,

$$\begin{aligned}
\gamma_{p_0,u}(t) &= \lim_{\beta_1\to\infty} \gamma_{p_0,p_1}(t) && \text{(108a)} \\
&= \left[1 + \frac{(1-\hat{t})}{\lambda\beta_0} r\right]^{-\frac{2}{\alpha+1}} \phi(t) && \text{(108b)} \\
&= p_{\alpha,\beta_t} && \text{(108c)}
\end{aligned}$$

where $\beta_t \coloneqq \frac{\beta_0}{1-\hat{t}}$. Therefore,

$$\{\gamma_{p_0,u}(t) : t \in (0,1)\} \subseteq \{p_{\alpha,\beta} : \beta \in (\beta_0,\infty)\}. \tag{109}$$

For any $\beta > \beta_0$, there exists a $\bar{t} \in (0,1)$ by the intermediate value theorem such that

$$p_{\alpha,\beta} = \gamma_{p_0,u}(\bar{t}) \tag{110}$$

since $t \mapsto \beta_t$ is a continuous map between the connected intervals $[0,1)$ and $[\beta_0,\infty)$. Thus,

$$\{p_{\alpha,\beta} : \beta \in (\beta_0,\infty)\} \subseteq \{\gamma_{p_0,u}(t) : t \in (0,1)\}. \tag{111}$$

Taking the closure of (109) and (111), we have the result for $\alpha \neq -1$:

$$\{\gamma_{p_0,u}(t) : t \in [0,1]\} \subseteq \overline{\{p_{\alpha,\beta} : \beta \in (\beta_0,\infty)\}} \subseteq \{\gamma_{p_0,u}(t) : t \in [0,1]\}. \tag{112}$$

For $\alpha = -1$, we have from a reparameterization $\lambda \leftarrow -\frac{2\lambda}{\alpha+1}$ of the optima in Lemma 4.3 (the reparameterization does not change the form) and the definition of Euler's number $e^x \coloneqq \lim_{n\to\infty}(1+\frac{x}{n})^n$,

$$\begin{aligned}
\lim_{\alpha\to-1} p_{\alpha,\beta} &= \lim_{\alpha\to-1} \frac{(1-\frac{\alpha+1}{2\lambda\beta}r)^{-\frac{2}{\alpha+1}}}{\int(1-\frac{\alpha+1}{2\lambda\beta}r)^{-\frac{2}{\alpha+1}}dV} && \text{(113a)} \\
&= \frac{e^{\frac{r}{\lambda\beta}}}{\int e^{\frac{r}{\lambda\beta}}dV}. && \text{(113b)}
\end{aligned}$$

The forward direction result follows from a similar argument to the $\alpha \neq -1$ case.

For the backward direction, assume that $\beta \mapsto p_{f,\beta}$ is an $(\alpha+2)$-geodesic. Let $\beta' > \beta$ and define

$$\bar{\mathcal{D}}_f(\beta) \coloneqq D_f(p_{f,\beta}, u) \tag{114}$$

By optimality of $p_{f,\beta}$ and $p_{f,\beta'}$, we have the inequalities

$$R(p_{f,\beta}) - \beta\bar{\mathcal{D}}_f(\beta) > R(p_{f,\beta'}) - \beta\bar{\mathcal{D}}_f(\beta') \tag{115a}$$
$$R(p_{f,\beta'}) - \beta'\bar{\mathcal{D}}_f(\beta') > R(p_{f,\beta}) - \beta'\bar{\mathcal{D}}_f(\beta) \tag{115b}$$

Adding the inequalities and using $\beta' > \beta$, we get

$$R(p_{f,\beta}) - \beta\bar{\mathcal{D}}_f(\beta) + R(p_{f,\beta'}) - \beta'\bar{\mathcal{D}}_f(\beta') > R(p_{f,\beta'}) - \beta\bar{\mathcal{D}}_f(\beta') + R(p_{f,\beta}) - \beta'\bar{\mathcal{D}}_f(\beta) \tag{116a}$$
$$\therefore -\beta\bar{\mathcal{D}}_f(\beta) - \beta'\bar{\mathcal{D}}_f(\beta') > -\beta\bar{\mathcal{D}}_f(\beta') - \beta'\bar{\mathcal{D}}_f(\beta) \tag{116b}$$
$$\therefore \bar{\mathcal{D}}_f(\beta) > \bar{\mathcal{D}}_f(\beta') \tag{116c}$$

which shows that $\bar{\mathcal{D}}_f$ is strictly monotonic decreasing. By a similar argument, $\bar{\mathcal{D}}_\alpha(\beta) \coloneqq D_\alpha(p_{f_\alpha,\beta}, u)$ is strictly monotonic decreasing. Since $\bar{\mathcal{D}}_f$ is strictly monotonic decreasing, its inverse $\bar{\mathcal{D}}_f^{-1}$ exists and is strictly monotonic decreasing. Then, for any p along the $(\alpha+2)$-geodesic,

$$D_\alpha(p, u) = (\bar{\mathcal{D}}_\alpha \circ \bar{\mathcal{D}}_f^{-1}) \circ D_f(p, u), \quad p \in \gamma^{\alpha+2}_{p^*,u} \tag{117}$$

Since $(\bar{\mathcal{D}}_\alpha \circ \bar{\mathcal{D}}_f^{-1})$ is strictly monotonic increasing, $D_\alpha(\cdot, u)$ is a strictly monotonic increasing function of $D_f(\cdot, u)$ when restricted to the $(\alpha+2)$-geodesic. But then, this restricted D_f is geodetic by the proof of Lemma 2.4. Furthermore, since the proof of Lemma 3.1 requires only the geodetic property along some α-geodesic with distinct endpoints, we have $f = f_\alpha$. □

Proposition 4.5. *If $\mathcal{P}_\Pi$ is convex, the exploration-exploitation relation $\beta \mapsto p_{f,\beta}$ is a map. If additionally $\mathcal{P}_\Pi$ is compact, the map is continuous.*

Proof. It is a standard result that if P_Π is convex, a local optimum $p_{f,\beta} \in \mathcal{P}_\Pi$ is the unique global optimum, such that $\beta \mapsto p_{f,\beta}$ is a map. The result follows from this and Berge's maximum theorem. We give a simplified proof under convexity, following Berge [1959]. We note that $(\beta, p) \mapsto R_{f,\beta}(p)$ is continuous in both β and p. Let $\beta \in \mathbb{R}_{\geq 0}$ and $\varepsilon > 0$. We first show that γ_f is lower semi-continuous. Since $(\beta, p) \mapsto R_{f,\beta}(p)$ is lower semi-continuous, there exists a neighborhood $(U, V) \subseteq \mathbb{R}_{\geq 0} \times \mathcal{P}_\Pi$ of $(\beta, \arg\max_{p \in \mathcal{P}_\Pi} R_{f,\beta}(p))$ such that for any $(\beta', p') \in (U, V)$,

$$\max_{p \in \mathcal{P}_\Pi} R_{f,\beta}(p) < R_{f,\beta'}(p') + \varepsilon \tag{118}$$

Lower semi-continuity of γ_f then follows from

$$\max_{p \in \mathcal{P}_\Pi} R_{f,\beta}(p) < R_{f,\beta'}(p') + \varepsilon \leq \max_{p \in \mathcal{P}_\Pi} R_{f,\beta'}(p) + \varepsilon \tag{119}$$

Next, we show that γ_f is upper semi-continuous. By upper semi-continuity of $R_{f,\beta}$, for each $p \in \mathcal{P}_\Pi$ there exists a neighborhood (U_p, V_p) such that for any $(\beta', p') \in (U_p, V_p)$,

$$R_{f,\beta'}(p') < R_{f,\beta}(p) + \varepsilon \tag{120}$$

The sets $\{V_p : p \in \mathcal{P}_\Pi\}$ are an open cover of $\mathcal{P}_\Pi$. Since $\mathcal{P}_\Pi$ is compact, there exists a finite subcover $\{V_{p_k} : k \in (1, ..., n)\}$. Now, for each $\beta' \in \bigcap_k U_{p_k}$, we have for any $p' \in \mathcal{P}_\Pi$,

$$R_{f,\beta'}(p') < R_{f,\beta}(p_k) + \varepsilon \tag{121}$$

for some k, since $p' \in V_{p_k}$ for some k. Upper semi-continuity of γ_f then follows from

$$\max_{p' \in \mathcal{P}_\Pi} R_{f,\beta'}(p') < \max_{k \in (1,...,n)} R_{f,\beta}(p_k) + \varepsilon \leq \max_{p \in \mathcal{P}_\Pi} R_{f,\beta}(p) + \varepsilon \tag{122}$$

Since γ_f is both lower and upper semi-continuous, it is continuous.

□

Proposition B.1. *Artificial curiosity with α-information is equivalent to Rényi maximum entropy exploration, under non-constant scaling of β.*

Proof. Recall that Rényi maximum entropy exploration uses (11) with Rényi entropy

$$H_\lambda(p) \coloneqq \frac{1}{1-\lambda} \log \int_{\mathcal{X}} p(x)^\lambda dP(x), \quad \lambda \in \mathbb{R}_{\geq 0} \tag{123}$$

instead of Shannon entropy H. Observe the relationship between Rényi entropy and the Rényi divergence

$$D_\lambda(p, u) \coloneqq \frac{1}{\lambda - 1} \log \int_{\mathcal{S}} p(s)^\lambda u^{1-\lambda} dV(s) = -H_\lambda(p) - \log u \tag{124}$$

Since the α-divergences and Rényi divergences are monotonically related by the Box-Cox transform F,

$$D_\alpha(p, u) = F(D_\lambda(p, u)), \tag{125}$$

we have

$$\int_{\mathcal{S}} p(s) I_\alpha(s; p) dV(s) = -u^{-\frac{\alpha+1}{2}} D_\alpha(p, u) - \frac{4}{1-\alpha^2}(1 - u^{-\frac{\alpha+1}{2}}) \tag{126a}$$

$$= -u^{-\frac{\alpha+1}{2}} F(D_\lambda(p, u)) - \frac{4}{1-\alpha^2}(1 - u^{-\frac{\alpha+1}{2}}) \tag{126b}$$

Now, using Lemma 2.2,

$$R_{\alpha,\beta}(\pi) = R(\pi) + \beta(n+1) \int_{\mathcal{S}} p_\pi(s) I_\alpha(s; p_\pi) dV(s) \tag{127a}$$

$$= R(\pi) - \beta(n+1) u^{-\frac{\alpha+1}{2}} F(D_\lambda(p, u)) - \frac{4\beta(n+1)}{1-\alpha^2}(1 - u^{-\frac{\alpha+1}{2}}) \tag{127b}$$

For a parameterized policy π_θ, we then have

$$\partial_i R_{\alpha,\beta}(\pi_\theta) = \partial_i \left\{ R(\pi_\theta) - \beta(n+1) u^{-\frac{\alpha+1}{2}} F(D_\lambda(p_{\pi_\theta}, u)) - \frac{4\beta(n+1)}{1-\alpha^2}(1 - u^{-\frac{\alpha+1}{2}}) \right\} \tag{128a}$$

$$= \partial_i R(\pi_\theta) - \beta(n+1) u^{-\frac{\alpha+1}{2}} F'(D_\lambda(p_{\pi_\theta}, u) \partial_i D_\lambda(p_{\pi_\theta}, u) \tag{128b}$$

$$= \partial_i R(\pi_\theta) + \beta(n+1) u^{-\frac{\alpha+1}{2}} F'(D_\lambda(p_{\pi_\theta}, u) \partial_i H_\lambda(p_{\pi_\theta}) \tag{128c}$$

$$= \partial_i R(\pi_\theta) + C(\theta) \beta \partial_i H_\lambda(p_{\pi_\theta}), \tag{128d}$$

with $C(\theta) \coloneqq (n+1) u^{-\frac{\alpha+1}{2}} F'(D_\lambda(p_{\pi_\theta}, u)$, from which the result follows. □

Proposition B.2. *If the agent-environment interaction is a unit variance Gaussian with mean $f_\theta(s_0)$, (45) is equivalent to classic artificial curiosity when $k = 1$ and $\alpha = -1$.*

Proof. Classic artificial curiosity uses

$$r(s) + \beta \| s - f_\theta(s_0, \pi(s_0)) \|^2 \tag{129}$$

where f_θ is a neural network model of the transition function. By hypothesis, we have

$$M^1(s_0, s) = (2\pi)^{\frac{d}{2}} \exp(-\frac{1}{2}(s - f_\theta(s_0))^\top (s - f_\theta(s_0))), \tag{130}$$

Since $I_{-1}(s; p) \coloneqq \lim_{\alpha \to -1} I_\alpha(s; p) = -\log p(s)$, then

$$I_{-1}(s; M^1(s_0, \cdot)) = -\log((2\pi)^{\frac{d}{2}} \exp(-\frac{1}{2}(s - f_\theta(s_0))^\top (s - f_\theta(s_0)))) \tag{131a}$$

$$= \frac{1}{2}(s - f_\theta(s_0))^\top (f_\theta(s) - s_0) - \frac{d}{2} \log 2\pi \tag{131b}$$

The statement follows since $\frac{1}{2}$ can be absorbed into β and $\frac{d}{2} \log 2\pi$ vanishes under gradients. □