Incoherence in Goal-Conditioned Autoregressive Models

Jacek Karwowski
Department of Computer Science
University of Oxford

Raymond Douglas
Telic Research

Abstract

We investigate mathematically the notion of *incoherence*: a structural issue with reinforcement learning policies derived by naive goal-conditioning of autoregressive models. We focus on the process of re-training models on their own actions, that is, fine-tuning offline-learned policies with online RL. We prove that it decreases incoherence and leads to an improvement in return, and we aim to characterise the resulting trajectory of policies. By re-framing standard notions of control-as-inference and soft Q learning, we establish a three-way correspondence with two other ways of understanding the iterative re-training process: as *folding the posterior into the reward* and, in the deterministic case, as *decreasing the temperature parameter*; the correspondence has computational content via the training-inference trade-off. Through soft-conditioning generative models, we discuss the link between incoherence and the *effective horizon* of Laidlaw et al. (2024).

1 INTRODUCTION

Control-as-inference reframes reinforcement learning as an inference problem: instead of explicitly trying to search for an optimal policy in a given environment, one first constructs a generative model over actions or trajectories, and then conditions it on the goal, deriving the policy from the posterior. Doing so allows for more abstract characterisation of the resulting policies, without reference to the internals of the particular learning algorithm.

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

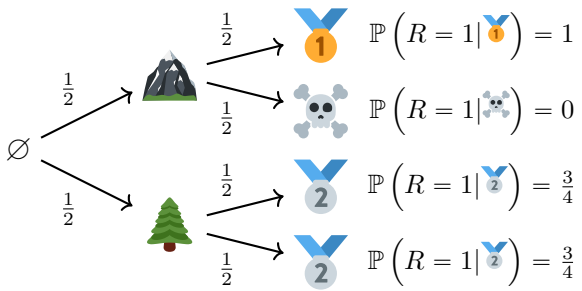


Figure 1: Tree representation of the MDP defining the *mountain race* (Example 1). States are represented as the tree nodes, we put a uniform prior over actions (depicted as arrows ↗, ↘). Rewards for each terminal state (tree leaf) are written on the right.

In this work, we focus on a multi-step environment, where the derived policy is used autoregressively: in each time-step, the generative model is conditioned on both the fixed goal and the current state. In this setup, we refine and characterize the recently introduced notion of *predictor-policy incoherence* (Douglas et al., 2024), or simply *incoherence*. This is a *structural* problem with such policies, which is not fixed by improving predictive accuracy of the underlying model. It stems from the fact that, given a binary reward R and a prior over actions $p(a|s)$, the conditioned policy $\pi(a|s) = p(a|s, R = 1)$ is an answer to the question:

> *(1) Which action to take in state s, such that, if later choices are made according to the prior p, the outcome will lead to R?*

and *not* the question:

> *(2) Which action to take in state s, such that, if later choices are sampled auto-regressively from π, the outcome will lead to R?*

Let us illustrate this problem by looking at a simple deterministic Markov decision process.

Example 1 (Mountain race)**.** *An agent is racing in the mountains. Starting in the state $\varnothing$, it is given a choice between two trails: a slower path down through the forest 🌲, and a faster path up over the ridge ⛰️. Both trails fork in the middle: following the path up on the ⛰️ junction leads quickly to the finish line 🥇, while the path down ends in a chasm ☠️. On the 🌲 path, both choices lead, albeit more slowly, to the finish line 🥈. Full game tree is presented in Figure 1;*

The agent is given binary reward R with probability 1 if finished 🥇, with probability $\frac{3}{4}$ if finished 🥈, and with probability 0 if finished ☠️. Below we denote actions as $\{\nearrow, \searrow\}$. The game is deterministic w.r.t. player's choice, but for the purpose of computing control-as-inference solution, we assume a uniform prior over actions.

We might intuitively understand this game in the following way: an agent is given a choice between "risky" play of ⛰️, which gives it a large payoff of 1 – but only if it chooses the trail correctly later on, and a "safe" play of 🌲, which gives it a smaller but more certain payoff, with probability $\frac{3}{4}$.

We compute the policy given by conditioning on the outcome $R = 1$ as:

$$\pi(a|\varnothing) = p\left(a|\varnothing, R=1\right) = \frac{p\left(R=1|\varnothing, a\right) p\left(a|\varnothing\right)}{p\left(R=1|\varnothing\right)} \propto p\left(R=1|\varnothing, a\right)$$

Since the transition dynamics are deterministic, we plug this into the formulas for π and get:

$$\pi(\nearrow|\varnothing) \propto p\left(R=1|\varnothing, \nearrow\right) = p\left(R=1|⛰️\right) = \frac{1}{2}$$
$$\pi(\searrow|\varnothing) \propto p\left(R=1|\varnothing, \searrow\right) = p\left(R=1|🌲\right) = \frac{3}{4}$$

We now observe that the RHS evaluates simply to the prior probability of attaining R. In other words, it does not take into account that the policy π is to be used autoregressively. On the other hand, a coherent policy $\hat{\pi}$ should have conditioned on the fact that $\hat{\pi}(R = 1|⛰️) = 1$, since, arriving in the state ⛰️, the reward-conditioned policy is also used.

As the above example shows, a naive autoregressive control-as-inference approach might result in incoherent policies. Intuitively, we understand the incoherence as the incapacity of the agent to anticipate its own future actions, in line with the difference between Questions (1) and (2). This also suggests a possible, and indeed, widely-used, fix: to fine-tune the agent on its own actions. Understanding the dynamics of this process, that is, what kinds of policies does it produce along the re-training trajectory, as well as relating it to incoherence, is thus the second point of this work.

1.1 Contributions and outline

The primary contribution of this paper is to properly (re-)define incoherence, and then reframe and unify existing ways of *tightening* goal-conditioned autoregressive policies in that framework. We discuss connections and parallels to many well-known approaches to soft RL in Section 2. After preliminaries in Section 3, Section 4 concerns the issues of incoherence; we refine and generalise the definition given previously by Douglas et al. (2024) to better account for the difference between *coherence* and *optimality*, and describe some properties of coherent policies. In Section 5, we discuss three ways of updating RL policies to remove incoherence:

- In Definition 5.1, as fine-tuning (re-conditioning) policies on their own trajectories, implemented e.g. as collecting trajectories from the model acting in an environment and re-training the model based on the augmented dataset.

- In Definition 5.6 as decreasing the temperature parameter, which can also be understood as manipulating the strength of the entropy regularisation in KL-regularised RL, implemented using e.g. inference-time best-of-n rejection sampling in practice.

- In Definitions 5.7 and 5.8, as folding the posterior over actions into the reward, which resembles the trick of disregarding a prior by folding it into the reward (Levine, 2018). Since the posterior depends on the reward, the process has to be repeated iteratively until convergence. This technically modifies the MDP, instead of modifying the policy. Even if the starting reward was only given sparsely in end-state, this process distributes it around the MDP.

The connections between those perspectives allow us to transfer properties between those formulations, for example, to provide a rate of convergence to the optimal policy of the re-conditioning. We discuss some related perspectives, in particular a connection to effective horizon Laidlaw et al. (2024), and limitations, in Section 6. All proofs are given in Appendix A. We also give a code appendix implementing tabular MDPs experiments and validating our main results numerically at https://github.com/jkarwowski/incoherence.

2 RELATED WORK

Control as Inference view, where optimality is modeled via a binary variable $\mathcal{O}$ with likelihood $p(\mathcal{O}|s,a) \propto e^{\alpha r(s,a)}$, is a central influence on our work. In that setup, deterministic dynamics allow for exact inference; in stochastic dynamics the policy solves a variational problem (Levine, 2018). We emphasize that *coherence* requires computing expectations under the posterior policy, not under a fixed prior; which forces the iterative posterior-folding we analyse here. O'Donoghue et al. (2020) focus on a similarly-termed incoherence in RL-as-inference, having to do with the fact that posterior probabilities do not reflect epistemic uncertainty about actions, and the effect this has on the exploration-exploitation trade-off. The connection between control and inference has been originally studied by Todorov (2006), and in the context of Inverse RL by Ziebart et al. (2008); Gleave and Toyer (2022).

KL-regularized policy search methods optimise expected return with a KL trust region to a prior, yielding E-step that is a Boltzmann distribution over Q and an M-step that updates the actor (Peters et al., 2010; Schulman et al., 2017; Abdolmaleki et al., 2018). Our "temperature" view is the Lagrange multiplier of the KL constraint; our iterative posterior-folding viewpoint recovers the same policy sequence when dynamics are deterministic. We thus reframe KL-regularized policy improvement as *restoring coherence* for goal-conditioned policies, as well as extend it to the stochastic or multi-step environment as compared to Korbak et al. (2022).

RvS (Brandfonbrener et al., 2023) as well as Upside-down RL (Srivastava et al., 2021) and Decision Transformer (Chen et al., 2021) approaches condition actions on desired returns and act autoregressively. Theory and experiments show that naive conditioning yields systematic failures in stochastic environments (trajectory "luck"), and that separating controllable from uncontrollable randomness improves behavior Štrupl et al. (2022); Paster et al. (2022); Yang et al. (2024). Our notion of incoherence is a structural account of the same issue: the posterior used for conditioning assumes futures incompatible with the deployment policy; our equivalence results characterize procedures that realign them.

Expert iteration and MCTS are a widely used and successful method of improving models' performance. Well-known applications include AlphaZero (Silver et al., 2017, 2018) and MuZero (Schrittwieser et al., 2020), algorithms using Monte Carlo Tree Search (Browne et al., 2012). One treatment of those kinds of algorithms combining search and improvement of the policy was proposed by (Anthony et al., 2017) under the name of Expert Iteration. We focus on the abstract properties of the re-training process, in a situation of soft-conditioning policies, a combination which prior work did not address. Self-play in games (Macleod, 2005; OpenAI et al., 2019; Vinyals et al., 2019) is another related strategy of improving the performance of a policy, however, it is distinct from the model learning about its own future policy in a non-competitive setup we study here.

Large language models are capable of general world modelling (Radford et al., 2019; Brown et al., 2020; Bai et al., 2022b; Touvron et al., 2023), and thus capable of simulating agents (Shanahan et al., 2023; Douglas et al., 2024). Although our setup here considers a model trained on a single environment, Andreas (2022) argues that this perspective applies to LLMs whose training corpus comes from human actions in the internet environment. Prior work on eliciting agents through prompting and scaffolding methods (Significant Gravitas, 2024; Yang et al., 2023) conditions the base model in a purely formal sense. The exact nature of this form of conditioning, as well as its connection to RL fine-tuning methods (such as RLHF (Christiano et al., 2017), DPO (Rafailov et al., 2023), GRPO (Shao et al., 2024)) is an open problem, such as e.g. *RLHF Conditioning Hypothesis* Hubinger et al. (2023). It has been argued that the next token prediction objective alone encourages local consistency but not global planning (McCoy et al., 2023). Recent work on Coconut (Hao et al., 2024, Section 5.1, Fig. 7) shows that the answer-token distribution along a latent reasoning tree acts like an implicit value function, an empirical point of support for our energy/value interpretation of goal-conditioning.

Residual Energy-Based Models for text generation add residual energy atop a base autoregressive model to steer sequence probabilities (Deng et al., 2020). Our folding-posterior-into-reward iteration is the control analogue of adding residual energy $\log p_\pi(a|s)$ to the base reward, with dynamics then re-evaluated under the new energy. This clarifies when temperature annealing can substitute for residual terms (deterministic dynamics) and when it cannot (stochastic). Energy-based policies of the general form $\pi(a|s) \propto \exp(E(s,a))$ had been studied by Haarnoja et al. (2017), and in the context of off-policy RL used to develop Soft Actor-Critic algorithm Haarnoja et al. (2018). This line of work is focused on the iterative approach using Bellman updates and countering distributional shift, while we look at the re-training as making policy internally consistent (still requiring computational effort).

3 PRELIMINARIES

A *Markov Decision Process* (MDP) with a time horizon $T \in \mathbb{N}$ is a tuple $\langle S, A, \tau, \mu, R, \gamma \rangle$, where S is a set of *states*, A is a set of *actions*, $\tau : S \times A \to \Delta(S)$ is a transition function, $\mu \in \Delta(S)$ is the initial distribution over states, $r : S \times A \to [-\infty, 0]$ is the reward function (assumed to be non-positive), and $\gamma \in [0, 1]$ is a time discount factor. We will assume $\gamma = 1$ without loss of generality (as one can always convert an MDP with $\gamma < 1$ to $\gamma = 1$ by introducing an auxiliary terminal state). A *trajectory* is a sequence $\xi = (s_0, a_0, s_1, \ldots, s_T, a_T)$ such that $a_i \in A$, $s_i \in S$ for all i. A *policy* is a function $\pi : S \to \Delta(A)$. We say that the policy π is deterministic if for each state s there is some $a \in A$ such that $\pi(s) = \delta_a$. Each policy π on an MDP induces a probability distribution over trajectories $p_\pi(\xi)$; drawing a trajectory $(s_0, a_0 \ldots, s_T, a_T)$ from a policy π means that s_0 is drawn from μ, each a_i is drawn from $\pi(a_i|s_i)$, and s_{i+1} is drawn from $\tau(s_{i+1}|a_i, s_i)$ for each i. For a policy $\pi(a|s)$ and a reward $r(s, a)$ we define the return $J(\pi)$ to be $J(\pi) = \mathbb{E}_{s_t, a_t \sim \pi}\left[\sum_{i=0}^{T} r(s_i, a_i)\right]$. We will sometimes distinguish a policy $\pi(a|s)$ and a prior over actions $p(a|s)$: even though these have the same type, they play different conceptual roles, with the policy being subject to the optimisation process and thus not fixed.

4 THE THEORY OF INCOHERENCE

To build a quantitative measure of incoherence, we first need to define soft Q and V functions, which are better-suited for working in a probabilistic setup than the standard definitions.

Definition 4.1 (Soft Q and V functions)**.** For a policy $\pi(a_t|s_t)$, a reward function $r(s_t, a_t)$ and transition dynamics $\tau(s_{t+1}|s_t, a_t)$, the soft V and Q functions are defined by mutual recursion as:

$$Q^\pi(a_t, s_t) = r(s_t, a_t) + \log \mathbb{E}_{s_{t+1} \sim \tau(s_t, a_t)} \left[\exp V^\pi(s_{t+1})\right]$$
$$V^\pi(s_t) = \log \mathbb{E}_{a_t \sim \pi(a_t|s_t)} \left[\exp Q^\pi(s_t, a_t)\right]$$

We note that our definition is parametrised by a policy π, and thus slightly different from the one given in (Haarnoja et al., 2017; Levine, 2018): the difference is in how we compute the expectation in V^π. Instead of drawing the action from a fixed prior, which is then routinely replaced w.l.o.g. by a uniform distribution over A, we draw it according to the policy π. The point of this subtlety will become apparent in the next section. In any case, above definition gives us the following alternative characterisation.

Proposition 4.2 (Characterisation of V and Q)**.** *For any prior $\pi(a|s)$ and any non-positive reward function $r(a, s)$, we have simple expressions for the soft Q and V functions given by:*

$$Q^\pi(a_t, s_t) = \log p_\pi(\mathcal{O}_{t:T} = 1|s_t, a_t)$$
$$V^\pi(s_t) = \log p_\pi(\mathcal{O}_{t:T} = 1|s_t)$$

where we define the auxiliary optimality variables $\mathcal{O}_t$ *to have Bernoulli distributions with:*

$$p_\pi(\mathcal{O}_t = 1|s_t, a_t) = e^{r(s_t, a_t)}$$

The above characterisation, making use of the auxiliary optimality variables $\mathcal{O}_t$, showcases an approach that we will be utilizing throughout the rest of this paper. To put it differently, one might look at the (non-positive) reward $r(s, a)$ as the probability that taking action a in the state s is *correct* - correctness constituting a latent property, as a more convenient way to cast MDP reward dynamics in a probabilistic setup.

We also have an alternative characterisation. Having a probability distribution over trajectories:

$$p(\xi) = \prod_{t=1}^{T} p(a_t|s_t)\tau(s_{t+1}|a_t, s_t)$$

as in Section 3, we derive the goal-conditioned:

$$p(\xi|\mathcal{O}_{1:T}) = \frac{p(\mathcal{O}_{1:T}|\xi)p(\xi)}{p(\mathcal{O}_{1:T})}$$

On the other hand, we have the distribution over trajectories given by locally goal-conditioning the prior over actions on the future reward:

$$\hat{p}(\xi) = \prod_{t=1}^{T} p(a_t|s_t, \mathcal{O}_{t:T})\tau(s_{t+1}|a_t, s_t)$$

Proposition 4.3. *In case of deterministic dynamics, we have that $p(\xi|\mathcal{O}_{1:T}) = \hat{p}(\xi)$.*

Following Levine (2018), we highlight a potential issue here. Since the policy derived by control-as-inference conditions the prior *trajectories* distribution, (given by (p, τ) jointly) on the goal $\mathcal{O}$, it results in an over-optimistic estimation of the environment dynamics, since it also conditions the stochastic dynamics. This will be the source of various issues separating deterministic and non-deterministic τ we encounter in next sections.

4.1 Incoherence

Incoherence is, in a sense, a statement about a policy's failure to act according to *its* future roll-outs, and

relying on the prior instead, as in the Example 1. Thus, we judge a policy more *coherent* the more its future returns influence the present decision in a consistent way. If we do not want to *a priori* prescribe the exact way of the influence is exerted, we can leave the function as a parameter in the definition below.

Definition 4.4 (Order-respecting f)**.** We say that a function f is order-respecting if for all x and $i \neq j$ we have that $f_i(x)$ is non-decreasing in x_i holding other coordinates fixed; $f_i(x)$ is non-increasing in x_j holding the rest fixed; and if $x_i \geqslant x_j$ then $f_i(x) \geqslant f_j(x)$.

Examples of order-respecting f include softmax with any temperature and the $\arg\max$ indicator (with discontinuities at ties).

Definition 4.5 (f-soft Q policy)**.** For an order-respecting function f and a policy $\pi(a|s)$ its Q, f-soft policy $\pi^{Q,f}(a|s)$ is the probability distribution:

$$\pi^{Q,f}(\cdot|s) \propto f(Q^\pi(s,\cdot))$$

We note that because our MDP setup allowed rewards with zero-probability $p(\mathcal{O}_t = 1|s_t, a_t) = 0$, we allow the rewards to be $-\infty$. We also note that in Example 1 (as well as in examples down below) for clarity of presentation, we have used reward R to mean $\mathcal{O}_{T=2}$.

The incoherence is then the KL divergence between the current policy distribution, and the distribution prescribed by future returns.

Definition 4.6 (f-incoherence)**.** The incoherence of a policy $\pi(a|s)$ with respect to its f-soft Q policy is defined as the KL divergence over trajectories ξ^π:

$$\kappa_f(\pi) = \mathrm{KL}_\xi(\pi(\xi)||\pi^{Q,f}(\xi))$$

We note that by a well-known correspondence (see e.g. Haarnoja et al. (2018) or Belousov (2017) for an informal write up) the KL divergence over trajectories can be rewritten as the per-state KL averaged over occupancy measure.

Proposition 4.7. *If d^t_π is the marginalised occupancy measure $d^t_\pi(s_t) = p_{\xi\sim\pi}(S_t = s_t)$, we have:*

$$\kappa_f(\pi) = \sum_{t=1}^{T} \mathbb{E}_{s_t\sim d^t_\pi} KL(\pi(\cdot|s_t)||\pi^{Q,f}(\cdot|s_t))$$

We say that a policy is *f-coherent*, if $\kappa_f(\pi) = 0$. In particular, coherence does not mean optimality in the usual sense of maximizing the return. Deterministic policies are f-coherent for a single functional f.

Proposition 4.8. *Given deterministic dynamics τ and policy $\pi(a|s)$ and an order-respecting $f : \mathbb{R}^{|A|} \to \Delta(|A|)$, such that for x having a unique maximum, $f(x)$ is an $\arg\max$ indicator, π is f-coherent if and only if it is greedy w.r.t. its own soft-Q function.*

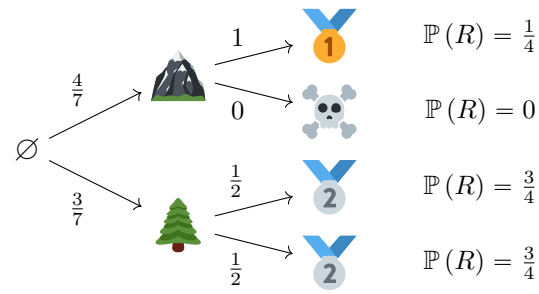


Figure 2: The fixed point of iterated f-coherence achieved after $T = 2$ iterations.

We also note that from Proposition 4.2 we know that:

$$Q^\pi(a_t, s_t) = \log p_\pi(\mathcal{O}_{t:T} = 1)$$

so in this case, maximising $J(\pi)$ is equivalent to maximising the probability of optimality $\mathcal{O}_{1:T} = 1$.

What about the general case of stochastic dynamics? First, we show that we can construct the coherent policy iteratively, starting from any prior $p(a|s)$.

Definition 4.9 (Iterated f-coherence)**.** Given a prior $\pi(a|s)$ and a function f, we define a sequence of policies $\pi^{\mathcal{B}}_i$ recursively by:

$$\pi^{\mathcal{B}}_0(a|s) = \pi(a|s) \qquad \pi^{\mathcal{B}}_{i+1}(a|s) = \pi^{Q,f}(\pi^{\mathcal{B}}_i(a|s))$$

This procedure of iterated f-coherence $\pi^{\mathcal{B}}$ converges to the f-coherent policy after at most T steps. It is the soft value iteration algorithm using an f-transformed soft Q-value.

Proposition 4.10. *The policy $\pi^{\mathcal{B}}_T$ is f-coherent.*

The coherence is defined up to the choice of the particular function f, which dictates how strongly future returns influence the behaviour in a particular step. We will be interested in a special class of coherent policies, given by Boltzmann distributions.

Definition 4.11 (Boltzmann rationality)**.** Given a policy π, the Boltzmann-rational policy π^δ with the parameter $\delta > 0$ is given by:

$$\pi^\delta(a|s) \propto \exp\left(\frac{1}{\delta} Q^\pi(a,s)\right)$$

It is known that in a single-step MDP setup, Boltzmann-rational policies are the solution to the satisficing maximum entropy problem (see, e.g. Jeon et al. (2020, Appendix A)). We will not need this perspective here, but we return to it in Section 5.1.

Example 2 (Boltzmann-coherent mountain race)**.** *We revisit the Example 1. We might compute the Boltzmann-coherent policy for $\delta = 1$ by iterating the f-coherence from Definition 4.9. After $T = 2$ iterations, we arrive at the fixpoint in Figure 2.*

Definition 4.12 (Boltzmann incoherence)**.** Given a policy π and a parameter $\delta > 0$, denoting $g(x) = \exp(x/\delta)$, we define the Boltzmann incoherence (or simply *incoherence*) as $\kappa_\delta(\pi) = \kappa_g(\pi)$.

Unrolling the definition, we have $\kappa_\delta(\pi) = \mathrm{KL}(\pi(a|s)||\pi^\delta(a|s))$. In other words, a policy π is Boltzmann-coherent with parameter δ, if it is coherent with respect to its f-soft Q policy for g, i.e. softmax with the parameter δ.

Proposition 4.13. *Given any prior $\pi(a|s)$, there exists a policy $\pi^*(a|s)$ such that we have convergence in distribution:*

$$\lim_{\delta\to 0} \pi^\delta \to \pi^*$$

where π^δ is defined with respect to the soft Q function induced by the prior π. Moreover, the policy π^* can be explicitly characterised as the uniform distribution over $A^* := \{a^* \in A : Q^\pi(s, a^*) = \max_{a\in A} Q^\pi(s, a)\}$.

We can now tie together the notions of *coherence* and *optimality*.

Corollary 4.14. *If the prior p is an optimal policy for the Markov chain, then limiting policy π^* from Proposition 4.13 is also optimal, which is not necessarily true for non-optimal priors.*

What about other policies? We can still use the limiting distance to the Boltzmann-rational policies to determine the optimality.

Corollary 4.15. *A necessary condition for policy $\pi(a|s)$ to be optimal is that:*

$$\lim_{\delta\to 0} \kappa_\delta(\pi) < \infty$$

We note that the above Corollary 4.15 does not extend to a sufficient condition. To see this, it is enough to consider a case of $T = 2$ MDP, such as e.g. Example 3. If the prior π puts zero weight on the action $\nearrow$ in the state $\varnothing$. Then, even though in state , Q^π is independent of π and therefore being Q^π_t-greedy is equivalent to optimality, it does not extend to $t = 0$.

We also note that the Boltzmann coherence is a combination of being soft-conditioned and coherent. In the case of MDPs with time horizon $T = 1$, the coherence requirement disappears, and the condition comes down to simply optimising a reinforcement learning problem with a KL penalty. We follow Korbak et al. (2022) in stating the following.

Proposition 4.16 (RL with KL penalties)**.** *Given an MDP with time horizon $T = 1$, a reward $r(s, a)$ and a prior $p(a|s)$, the Boltzmann-coherent policy $\pi(a|s)$ can be derived by maximising the KL-regularised return $J(\pi) = \mathbb{E}\left[r(s, a)\right] - KL(\pi(a|s)||p(a|s))$. The resulting policy is of a form:*

$$\pi^*(a|s) \propto p(a|s)\exp(r(s, a))$$

5 REMOVING INCOHERENCE

Having established the notions of incoherence, we now turn to the topic of removing it through the process of fine-tuning policies on their own actions. We first establish the notion of goal-conditioning a prior over actions.

Definition 5.1 (Goal conditioning)**.** Given transition dynamics $\tau(s_{t+1}|a_t, s_t)$, a prior $p(a_t|s_t)$ and a non-positive reward $r(s, a)$, we say that a policy π is *given by goal conditioning*, if we have that:

$$\pi(a_t|s_t) = p(a_t|s_t, \mathcal{O}_{t:T} = 1)$$

where the right-hand side is defined through the joint distribution of (τ, p) and the auxiliary *optimality variable* $\mathcal{O}_t$ has a Bernoulli distribution of:

$$p(\mathcal{O}_t = 1|s_t, a_t) = e^{r(s_t,a_t)}$$

Simply conditioning on the optimality does not guarantee coherence.

Proposition 5.2. *A policy given by goal conditioning is not necessarily coherent for any choice of δ.*

Indeed, the counterexample is given by Example 1. However, we might hope to fix this by iterating the goal conditioning, that is, retraining a policy on its own actions. To formalise this process, we give the following definition.

Definition 5.3 (Control-as-inference)**.** We define the control-as-inference operator $\mathcal{G}$ on the space of probability distributions $p(a_t|s_t)$ as

$$\mathcal{G}(p) = p(a_t|s_t, \mathcal{O}_{t:T} = 1)$$

where the left-hand side is given jointly by $p, \tau, \mathcal{O}$. This gives a sequence of policies $\pi^{\mathcal{G}}_i$ defined recursively as:

$$\pi^{\mathcal{G}}_0(a_t|s_t) = p(a_t|s_t) \qquad \pi^{\mathcal{G}}_{k+1} = \mathcal{G}(\pi^{\mathcal{G}}_k)$$

We note that this process does not condition the dynamics τ of the MDP. In other words, it might be thought to be implemented as iteratively changing the joint by collecting rollouts from the conditioned policy, both those successful and unsuccessful, without rejection sampling (conditioned on optimality). Even still, we have the following property:

Proposition 5.4 (Strong return improvement lemma)**.** *The sequence of policies $(\pi^{\mathcal{G}}_i)_{i=0,1,\dots}$ given by the control-by-inference improves its return monotonically, that is:*

$$J(\pi^{\mathcal{G}}_{i+1}) \geqslant J(\pi^{\mathcal{G}}_i)$$

The fact that the policies monotonically improve gives a reason to suspect that the limiting policy (which exists by a compactness argument) will be an optimal policy for the reward r. Indeed, we can show the following.

Proposition 5.5 ((Douglas et al., 2024, Theorem 3.6))**.** *There exists some policy $\pi^{\mathcal{G}}$ that the sequence $\pi_i^{\mathcal{G}}$ converges to. That is, we have:*

$$\lim_{i\to\infty} KL(\pi_i^{\mathcal{G}}||\pi^{\mathcal{G}}) = 0$$

Moreover, if the prior $p(a|s)$ has full support over all actions a in all states s, then $\pi^{\mathcal{G}}$ is optimal.

This property is a consequence of the equivalence we develop in Section 5.1; a direct proof can be found in (Douglas et al., 2024, Appendix A).

Another formulation of the sequence of increasingly improving policies is given by increasing the temperature of the energy distribution of the optimality variable $\mathcal{O}$.

Definition 5.6 (RL with temperature)**.** Given a prior $p(a_t|s_t)$, a reward $r(s_t, a_t)$ and an *(inverse) temperature* function $\alpha : \mathbb{N} \to \mathbb{R}_+$, we define a sequence of policies:

$$\pi_{\alpha(k)}(a_t|s_t) = p(a_t|s_t, \mathcal{O}_{t:T}^{\alpha(k)} = 1)$$

where:

$$p(\mathcal{O}_t^{\alpha(k)}|s_t, a_t) = \exp\left(\alpha(k) \cdot r(s_t, a_t)\right)$$

We note that the process of increasing temperature in this way behaves differently from what taking $\delta \to 0$ in we explored in the last section's Definition 4.12. There, we took a fixed Q function defined by a particular policy π, and applied softmax. Here, the policy itself is changing, so we see progressively more precise Q functions of a form $Q^{\pi_{\alpha(k)}}$, as (inverse) temperature $\alpha(k)$ increases.

Finally, we talk about folding the posterior into the reward. Levine (2018) discusses a trick for folding *the prior* into the reward. If soft Q and V functions are computed as in Definition 4.1, one can disregard a prior over actions $p(a_t|s_t)$ by modifying the reward function $r(s_t, a_t)$ (specifying the Bernoulli distribution of $\mathcal{O}_t$) to be:

$$\hat{r}(s_t, a_t) = r(s_t, a_t) + \log p(a_t|s_t)$$

This defines an equivalent probabilistic model in which the prior $p(a_t|s_t)$ is uniform over actions, simplifying the calculations. Our point of interest lies in the fact that the coherent definition of the soft V function uses *the posterior* instead of prior:

$$V^{\pi}(s_t) = \log \mathbb{E}_{a_t\sim\pi}[\exp Q^{\pi}(a_t, s_t)]$$

as we discussed in the introduction, pointing to the difference between Questions (1) and (2).

Because the expectation is over the policy $\pi(a_t|s_t) = p(a_t|s_t, \mathcal{O}_{t:T})$ and *not* the prior $p(a_t|s_t)$, it is no longer that simple to fold it into the reward. Since the reward (and therefore, the distributions of $\mathcal{O}_t$'s) changed, it has a downstream effect of changing the posterior. Thus, iterative process is needed.

Definition 5.7 (Folded sequence)**.** For a prior $p(a_t|s_t)$ and a reward $r(s_t, a_t)$ we define a sequence of policies $\pi_k^{\mathcal{F}}$ and rewards r_k recursively, to be

$$r_0 = r \qquad \pi_0^{\mathcal{F}} = p$$

$$r_{k+1}(s_t, a_t) = r_0(s_t, a_t) + \log p(a_t|s_t, \mathcal{O}_{t:T}^{(k)} = 1)$$

$$\pi_{k+1}^{\mathcal{F}}(a_t|s_t) = p(a_t|s_t, \mathcal{O}_{t:T}^{(k)} = 1)$$

$$p(\mathcal{O}_t^{(k)} = 1|a_t, s_t) = \exp\left(r_k(a_t, s_t)\right)$$

It turns out that this iterative process is an equivalent way of formulating the re-training the model on its own output. We can also fold the reward recursively into the previous reward:

Definition 5.8 (Folded sequence, cumulative)**.** For a prior $p(a_t|s_t)$ and a reward $r(s_t, a_t)$ we define a sequence of policies $\pi_k^{\mathcal{H}}$, the optimality variables $\mathcal{O}$ as in Definition 5.7, with the only change being r_{k+1}:

$$r_{k+1}(s_t, a_t) = r_k(s_t, a_t) + \log p(a_t|s_t, \mathcal{O}_{t:T}^{(k)} = 1)$$

This version is, however, only equivalent to the other modes of retraining under the assumption of deterministic transition dynamics τ, as we prove in the next Section 5.1.

5.1 Equivalence and corollaries

After introducing the three ways of constructing a sequence of policies: fine-tuning policies on their own actions, folding the posterior into the reward, and conditioning on the reward with an increased temperature, we are now able to connect those dynamics. In case of deterministic dynamics, all three coincide, giving the same policy trajectories. In case of stochastic dynamics, it is only true for the first two.

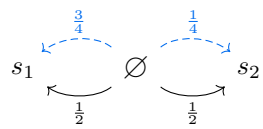
Theorem 5.9 (Optimisation pressure, thrice)**.** *For all priors $p(a|s)$, all non-positive reward functions $r(s_t, a_t)$ and deterministic transition dynamics $\tau(s_{t+1}|s_t, a_t)$, for all $k \in \mathbb{N}_+$ and for $\alpha(k) = 2^k$, we have:*

$$\pi_{\alpha(k)} = \pi_{2^k}^{\mathcal{G}} = \pi_{2^k}^{\mathcal{F}} = \pi_k^{\mathcal{H}}$$

Additionally, for stochastic transition dynamics we have:

$$\pi_k^{\mathcal{F}} = \pi_k^{\mathcal{G}}$$

Example 3. *Transition dynamics such that $\pi_{\alpha(k)} \neq \pi_k^{\mathcal{G}}$ are simple to construct. For example, take a three-state, two action Markov decision process with initial state $\varnothing$ and a uniform prior policy, with transition dynamics depicted on the diagram below:*



where red dotted lines correspond to action a_1, and solid black lines to action a_2. We also assume that:

$$r(s_1) = \log \frac{1}{3} \qquad r(s_2) = \log \frac{2}{3}$$

Computing $\pi_1^{\mathcal{G}}(\cdot|\varnothing)$ and $\pi_{\alpha(2)}(\cdot|\varnothing)$ we get:

$$\frac{25}{61} = \pi_1^{\mathcal{G}}(a_1|\varnothing) \neq \pi_{\alpha(2)}(a_1|\varnothing) = \frac{7}{17}$$

From this equivalence, and additional properties of each of the ways of constructing the sequence, we can easily derive results that would require laborious proofs otherwise. For example, we might easily show that the policies converge to an optimal policy, and explicitly calculate the rate of convergence, using the causal entropy regularisation characterisation of $\pi_\alpha(k)$.

Corollary 5.10 (Return improvement rate)**.** *In case of deterministic dynamics τ, given a sequence of policies $\pi_i^{\mathcal{G}}$, for $k \in \mathbb{N}_+$ have that:*

$$J(\pi_k^{\mathcal{G}}) - J(\pi_{k-1}^{\mathcal{G}}) = \frac{1}{k} \cdot \frac{J'(\pi_k^{\mathcal{G}}) \cdot \hat{\mathcal{H}}'(\pi_k^{\mathcal{G}})}{J''(\pi_k^{\mathcal{G}}) + \frac{1}{k}\hat{\mathcal{H}}''(\pi_k^{\mathcal{G}})} + O\left(\frac{1}{k^2}\right)$$

where $\hat{\mathcal{H}}(\pi)$ denotes the causal entropy of policy π, and the derivatives are taken with respect to the temperature α.

Incoherence disappears at the limit of retraining.

Corollary 5.11. *We have $\lim_{(\delta,i)\to(0,\infty)} \kappa^\delta(\pi_i^{\mathcal{G}}) = 0$.*

The significance of the correspondence should be possible to also be understood through the lenses of training-inference trade-off. Decreasing the temperature in the information-bounded RL can be implemented as top-k rejection-sampling based methods (such as Speculative Rejection (Sun et al., 2024)). On the other hand, we might want to pay the cost of inference once, during training, by retraining the model on its own outputs. Prior work has empirically verified that aligning language models with RLHF produces a similar behavior to best-of-k rejection sampling with respect to the reward model (Bai et al., 2022a; Stiennon et al., 2022). Our results confirm those results, and allow for precise trade-off estimation in the multi-step environment, with no difference on the margin with each phase of re-training equivalent to linearly increasing k.

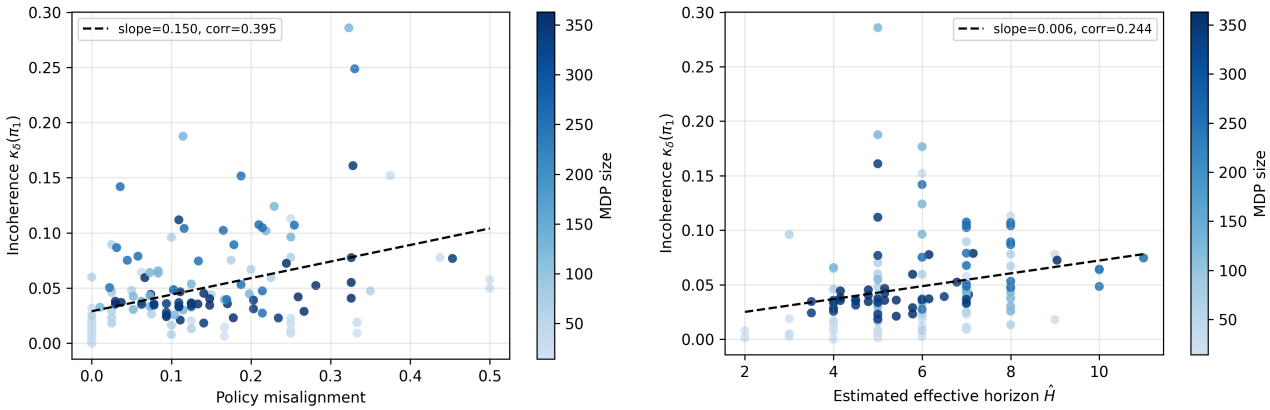
6 INCOHERENCE AND EFFECTIVE HORIZON

We have started the discussion in Section 1 by pointing out the difference between Questions (1) and (2), with Example 1 showing that the autoregressive policy derived by conditioning on the reward answers (1). Concurrently with the development of this work, Laidlaw et al. (2023, 2024) performed extensive experiments to find a reason for which deep RL works in some environments, but not others. They empirically arrived, and later formalised, the notion *effective horizon* (EH): a quantitative measure of the hardness of a given environment. We quote:

> *When actions with the highest Q-values under the random policy also have the highest Q-values under the optimal policy (i.e. when it is optimal to be greedy on the random policy's Q function), deep RL tends to succeed; when they don't, deep RL tends to fail.*

We note the striking similarity to the setup studied here: we rephrase their finding to state that *deep RL tends to succeed when there is no difference between answers to Questions (1) and (2); it tends to fail otherwise.* This suggests that deep RL algorithms approximate naive autoregressive control-as-inference policies, which provides a theoretical justification of the effective horizon result. EH and incoherence are nevertheless different on a technical level, making a direct theoretical comparison difficult. Instead, we opt to compare them experimentally.

Misalignment and incoherence In the first approach, we compare incoherence to the *policy misalignment*, which is one operationalisation of the difference between Questions *(1)* and *(2)* aligned with the EH perspective above. Given an MDP, we define the misalignment $m \in [0, 1]$ to be the proportion of actions being different between π_{rand} and π^*, where π_{rand} selects the greedy action under the random policy's Q-function $Q^{\pi_{\text{rand}}}$, averaged over the random policy occupancy measure, motivated by Laidlaw et al. (2024), and π^* is a deterministic optimal policy (in our experiments, both are a.s. unique). For a set of 140 randomly chosen small MDP, we computed m and Boltzmann incoherence $\kappa_\delta(\pi^G)$ for the goal-conditioned uniform prior $\pi^{\mathcal{G}} = \mathcal{G}(\pi_{\text{rand}})$, for temperature $\delta = 1$ (for a discussion on other temperatures, please see Appendix E). We present this



(a) (b)

Figure 3: Empirically observed significant positive correlations between incoherence and (a) misalignment score, and (b) estimated effective horizon, in a dataset of 140 randomly chosen small MDPs, between 4 and 122 states.

in Figure 3a. The correlation between m and κ is then 0.395 (95% CI $[0.245, 0.526]$), and isn't sensitive to outliers. In other words, environments where being greedy on $Q^{\pi_{\text{rand}}}$ is suboptimal (large misalignment) also tend to be those where naive goal-conditioning produces larger incoherence. This is qualitatively aligned with the hypothesis put forward above.

Effective horizon vs incoherence In the second approach, for a more direct comparison, we implemented a simple Monte Carlo-based effective horizon estimator $\hat{H}$ for small deterministic-transition MDPs (for details on $\hat{H}$, please see Appendix E.2). We reused the same dataset of 140 small random MDPs from the first experiment, and we measured $\kappa_\delta(\pi)$ for the naively goal-conditioned policy $\pi^{\mathcal{G}} = \mathcal{G}(\pi^{\text{unif}})$ on a uniform prior. We obtained the correlation between $\hat{H}$ and $\kappa_\delta(\pi^{\mathcal{G}}) = 0.244$ (95% CI $[0.082, 0.394]$), plotted on Figure 3b. We again checked that the result is not sensitive to outliers. In other words, environments with large effective horizon (where being greedy on $Q^{\pi_{\text{rand}}}$ is misleading) often coincide with those where the conditioned policy is highly incoherent, which is again consistent with the picture put forward at the start of this section.

7 DISCUSSION

Incoherence as a diagnostic metric Incoherence, as a metric of the mismatch between the policy's own Q function and its behavior, could be used as an additional diagnostic for monitoring RL training progress. To test this, we ran an experiment to measure the correlation between return and incoherence. We note that a priori, a policy can be coherent but not optimal and vice versa, but we expect the increase of coherence to lead to better performance on average. Reusing the same dataset of 140 MDPs, starting from uniformly random policy, we applied the operator $\mathcal{G}$ 4 times. Each iteration, we computed the expected return J and incoherence κ_δ, setting $\delta = 1$. The average correlation between J and negative-κ for each experiment (then averaged across experiments) was 0.978. On all trajectories flattened, the average correlation remained high at 0.863, providing some evidence for its validity.

Limitations and future work We have only conducted experiments in small tabular MDP environments, and relied on the already established experimental results from the prior work. Synthetic data experiments with Decision Transformers or other sequence models are a natural next step; by training a model on a game such as chess and controlling the training dataset, it should be possible to verify to what extent do models such as DTs suffer from incoherence, as answers to Questions *(1)* and *(2)* can be computed exactly. On the theory side, extending our treatment to the case of infinite time horizon requires handling discount factor which is known to be more involved (Haarnoja et al., 2017; Levine, 2018). We did not propose any methods of regularising models towards coherence during training, without the expensive re-training procedure, for example by including an auxiliary term in the loss, which would be an interesting direction for future work. Finally, our understanding of stochastic environments is currently limited to negative results, but it seems that in practice, RL still behaves well (both from our limited experiments and in Laidlaw et al. (2023)), which points towards understanding this as an important theoretical direction.

Acknowledgements

The authors would like to thank Victoria Krakovna, Cassidy Laidlaw, Joar Skalse, Sam Staton, participants of the SPAR program, and anonymous reviewers for their feedback and comments on various drafts of this paper. The first author was supported by a grant from Open Philanthropy. Some of this work was performed as part of the SERI MATS program.

References

Abdolmaleki, A., Springenberg, J. T., Tassa, Y., Munos, R., Heess, N., and Riedmiller, M. (2018). Maximum a Posteriori Policy Optimisation. arXiv:1806.06920 [cs].

Andreas, J. (2022). Language Models as Agent Models. arXiv:2212.01681 [cs].

Anthony, T., Tian, Z., and Barber, D. (2017). Thinking Fast and Slow with Deep Learning and Tree Search. arXiv:1705.08439 [cs].

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. (2022a). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862 [cs].

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. (2022b). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs].

Belousov, B. (2017). KL between trajectory distributions vs KL between policies · Boris Belousov.

Brandfonbrener, D., Bietti, A., Buckman, J., Laroche, R., and Bruna, J. (2023). When does return-conditioned supervised learning work for offline reinforcement learning? arXiv:2206.01079 [cs].

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*. arXiv: 2005.14165.

Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., and Colton, S. (2012). A Survey of Monte Carlo Tree Search Methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43. Conference Name: IEEE Transactions on Computational Intelligence and AI in Games.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. (2021). Decision Transformer: Reinforcement Learning via Sequence Modeling. arXiv:2106.01345 [cs].

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4302–4310, Red Hook, NY, USA. Curran Associates Inc.

Deng, Y., Bakhtin, A., Ott, M., Szlam, A., and Ranzato, M. (2020). Residual Energy-Based Models for Text Generation. arXiv:2004.11714 [cs].

Douglas, R., Karwowski, J., Bae, C., Draguns, A., and Krakovna, V. (2024). Limitations of agents simulated by predictive models. In *ICLR 2024 workshop on large language model (LLM) agents*.

Gleave, A. and Toyer, S. (2022). A Primer on Maximum Causal Entropy Inverse Reinforcement Learning. arXiv:2203.11409 [cs].

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. (2017). Reinforcement Learning with Deep Energy-Based Policies. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1352–1361. PMLR. ISSN: 2640-3498.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. arXiv:1801.01290 [cs, stat].

Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., and Tian, Y. (2024). Training Large Language Models to Reason in a Continuous Latent Space. arXiv:2412.06769 [cs].

Hubinger, E., Jermyn, A., Treutlein, J., Hudson, R., and Woolverton, K. (2023). Conditioning Predictive Models: Risks and Strategies. arXiv:2302.00805 [cs].

Jeon, H. J., Milli, S., and Dragan, A. D. (2020). Reward-rational (implicit) choice: A unifying formalism for reward learning. arXiv:2002.04833 [cs].

Korbak, T., Perez, E., and Buckley, C. L. (2022). RL with KL penalties is better viewed as Bayesian inference. arXiv:2205.11275 [cs, stat].

Laidlaw, C., Russell, S., and Dragan, A. (2024). Bridging RL Theory and Practice with the Effective Horizon. arXiv:2304.09853 [cs, stat].

Laidlaw, C., Zhu, B., Russell, S., and Dragan, A. (2023). A Theoretical Explanation of Deep RL Performance in Stochastic Environments. In *The Twelfth International Conference on Learning Representations*.

Levine, S. (2018). Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review. arXiv:1805.00909 [cs, stat].

Macleod, A. (2005). Game design through self-play experiments. In *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*, ACE '05, pages 421–428, New York, NY, USA. Association for Computing Machinery.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., and Griffiths, T. L. (2023). Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve. arXiv:2309.13638 [cs].

O'Donoghue, B., Osband, I., and Ionescu, C. (2020). Making Sense of Reinforcement Learning and Probabilistic Inference. arXiv:2001.00805 [cs].

OpenAI, Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., Józefowicz, R., Gray, S., Olsson, C., Pachocki, J., Petrov, M., Pinto, H. P. d. O., Raiman, J., Salimans, T., Schlatter, J., Schneider, J., Sidor, S., Sutskever, I., Tang, J., Wolski, F., and Zhang, S. (2019). Dota 2 with Large Scale Deep Reinforcement Learning. arXiv:1912.06680 [cs, stat].

Paster, K., McIlraith, S., and Ba, J. (2022). You Can't Count on Luck: Why Decision Transformers and RvS Fail in Stochastic Environments. arXiv:2205.15967 [cs].

Peters, J., Mülling, K., and Altün, Y. (2010). Relative entropy policy search. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'10, pages 1607–1612, Atlanta, Georgia. AAAI Press.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI blog*.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290 [cs].

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., and Silver, D. (2020). Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. *Nature*, 588(7839):604–609. arXiv:1911.08265 [cs, stat].

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal Policy Optimization Algorithms. arXiv:1707.06347 [cs].

Shanahan, M., McDonell, K., and Reynolds, L. (2023). Role-Play with Large Language Models.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. (2024). DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300 [cs].

Significant Gravitas (2024). AutoGPT. original-date: 2023-03-16T09:21:07Z.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144. Publisher: American Association for the Advancement of Science.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359. Number: 7676 Publisher: Nature Publishing Group.

Srivastava, R. K., Shyam, P., Mutz, F., Jaśkowski, W., and Schmidhuber, J. (2021). Training Agents using Upside-Down Reinforcement Learning. arXiv:1912.02877 [cs].

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. (2022). Learning to summarize from human feedback. arXiv:2009.01325 [cs].

Sun, H., Haider, M., Zhang, R., Yang, H., Qiu, J., Yin, M., Wang, M., Bartlett, P., and Zanette, A. (2024).

Fast Best-of-N Decoding via Speculative Rejection. arXiv:2410.20290 [cs].

Todorov, E. (2006). Linearly-solvable Markov decision problems. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs].

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., Vezhnevets, A. S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T. L., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps, C., and Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354. Number: 7782 Publisher: Nature Publishing Group.

Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., and Chen, X. (2024). Large Language Models as Optimizers. arXiv:2309.03409 [cs].

Yang, H., Yue, S., and He, Y. (2023). Auto-GPT for Online Decision Making: Benchmarks and Additional Opinions. arXiv:2306.02224 [cs].

Ziebart, B. D., Maas, A., Bagnell, J. A., and Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 3*, AAAI'08, pages 1433–1438, Chicago, Illinois. AAAI Press.

Štrupl, M., Faccio, F., Ashley, D. R., Schmidhuber, J., and Srivastava, R. K. (2022). Upside-Down Reinforcement Learning Can Diverge in Stochastic Environments With Episodic Resets. arXiv:2205.06595 [stat].

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]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

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]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [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]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [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. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [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. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A PROOFS

A.1 Proofs for Section 4

Proposition A.1 (Characterisation of V and Q)**.** *For any prior $\pi(a|s)$ and any non-positive reward function $r(a, s)$, we have simple expressions for the soft Q and V functions given by:*

$$Q^\pi(a_t, s_t) = \log p_\pi(\mathcal{O}_{t:T} = 1|s_t, a_t)$$
$$V^\pi(s_t) = \log p_\pi(\mathcal{O}_{t:T} = 1|s_t)$$

where we define the auxiliary optimality variables $\mathcal{O}_t$ *to have Bernoulli distributions with:*

$$p_\pi(\mathcal{O}_t = 1|s_t, a_t) = e^{r(s_t,a_t)}$$

Proof. We prove this by backward induction. Base case $t = T$, we have:

$$\begin{aligned} Q^\pi(s_T, a_T) &= r(s_T, a_T) = \log(\exp r(s_T, a_T)) \\ &= \log p_\pi(\mathcal{O}_T = 1|s_T, a_T) \end{aligned}$$

and:

$$\begin{aligned} V^\pi(s_T) &= \log \mathbb{E}_{a_t\sim\pi(a_T|s_T)} \left[\exp Q^\pi(s_t, a_T)\right] \\ &= \log \mathbb{E}_{a_t\sim\pi(a_T|s_T)} \left[p_\pi(\mathcal{O}_T = 1|s_T, a_T]\right) \\ &= \log p_\pi(\mathcal{O}_T = 1|s_T) \end{aligned}$$

Now, assuming that the hypothesis holds for $t < t' \leqslant T$, we compute for t:

$$\begin{aligned} Q^\pi(s_t, a_t) &= r(s_t, a_t) + \log \mathbb{E}_{s_{t+1}} \left[\exp(V^\pi(s_{t+1})]\right) \\ &= \log(p_\pi(\mathcal{O}_t = 1|s_t, a_t) \cdot \mathbb{E}_{s_{t+1}} \left[p_\pi(\mathcal{O}_{t+1:T}|s_{t+1})\right]) \\ &= \log \left(p_\pi(\mathcal{O}_t = 1|s_t, a_t) \cdot p_\pi(\mathcal{O}_{t+1:T}|s_{t+1}, a_{t+1})\right) \\ &= \log p_\pi(\mathcal{O}_{t:T} = 1|s_t, a_t) \end{aligned}$$

and the proof of the inductive step for $V^\pi(s_t, a_t)$ is the same as the base case shown above. □

Proposition A.2. *In case of deterministic dynamics, we have that $p(\xi|\mathcal{O}_{1:T}) = \hat{p}(\xi)$.*

Proof. We compute the policy for time steps t and $t + 1$:

$$\begin{aligned} p(a_t|s_t, \mathcal{O}_{t:T}) &= \frac{p(\mathcal{O}_{t:T}|a_t, s_t)p(a_t|s_t)}{p(\mathcal{O}_{t:T}|s_t)} \\ &= \frac{p(\mathcal{O}_{t:T}|s_{t+1})p(a_t|s_t)}{p(\mathcal{O}_{t:T}|s_t)} \end{aligned}$$

and

$$p(a_{t+1}|s_{t+1}, \mathcal{O}_{t+1:T}) = \frac{p(\mathcal{O}_{t+1:T}|a_{t+1}, s_{t+1})p(a_{t+1}|s_{t+1})}{p(\mathcal{O}_{t+1:T}|s_{t+1})}$$

where $\tau(s_{t+1}|s_t, a_t) = 1$. Thus, considering $\log \hat{p}(\xi)$ we have telescoping series:

$$\begin{aligned} \log \hat{p}(\xi) &= \sum_{t=1}^{T} \log p(\mathcal{O}_{t:T}|s_t, a_t) - \log p(\mathcal{O}_{t:T}|s_t) + \log p(a_t|s_t) \\ &= \log(\mathcal{O}_{T:T}|a_T, s_T) - \log p(\mathcal{O}_{1:T}|s_1) + \sum_{t=1}^{T} \log p(a_t|s_t) \end{aligned}$$

which is exactly:

$$\log p(\xi|\mathcal{O}_{1:T}) = \log p(\mathcal{O}_{1:T}|\xi) - \log(\mathcal{O}_{1:T}) + \log p(\xi)$$

□

Proposition A.3. *Given deterministic dynamics τ and policy $\pi(a|s)$ and an order-respecting $f : \mathbb{R}^{|A|} \to \Delta(|A|)$, such that for x having a unique maximum, $f(x)$ is an* $\arg\max$ *indicator, π is f-coherent if and only if it is greedy w.r.t. its own soft-Q function.*

Proof. We first show argmax necessity under order-respecting f: let $x \in \mathbb{R}^{|A|}$ have a unique maximizer $m = \arg\max_i x_i$. If f is order-respecting and outputs a point mass at x, then $f(x) = \delta_m$. Indeed, suppose $f(x) = \delta_i$ with $i \neq m$. Then $x_m > x_i$, but the last condition in the order-preservation definition demands $f_m(x) \geqslant f_i(x) = 1$, a contradiction.

Now, the forward direction. Assume π is f-coherent, and fix the state s. If $Q^\pi(s, \cdot)$ has a unique maximizer, from what we proved above we know that $f(Q^\pi(s, \cdot)) = \delta_{\arg\max}$, i.e. $\pi(s)$ must be an argmax. If there are ties, coherence still forces $\pi(s)$ to be one of the maximizers, which is consistent with the selector's tie-break, as otherwise the per-state KL contributes ∞. So $\pi(s) \in \arg\max Q^\pi(s, \cdot)$.

For the reverse direction, if $\pi(s) \in \arg\max_a Q^\pi(s, a))$ for all s, we pick f to be the arg max selector $f(Q_\pi(s, \cdot)) = \delta_{\pi(s)}$ for all s. Hence $\mathrm{KL}(\pi||\pi^{Q,f}) = 0$ by Proposition 4.7, i.e., π is f-coherent.

□

Proposition A.4. *The policy $\pi_T^{\mathcal{B}}$ is f-coherent.*

Proof. We prove by backwards induction that $\pi(s_t|a_t)$ is fixed in iteration t and does not change afterwards. In the base case $t = T$, the Q function $Q^\pi(s_T, \cdot)$ does not depend on π, so it is fixed in the first iteration and remains unchanged. The inductive case follows because $\pi(s_t|a_t)$ only depends on future times $t' > t$. □

Proposition A.5. *Given any prior $\pi(a|s)$, there exists a policy $\pi^*(a|s)$ such that we have convergence in distribution:*

$$\lim_{\delta \to 0} \pi^\delta \to \pi^*$$

where π^δ is defined with respect to the soft Q function induced by the prior π. Moreover, the policy π^* can be explicitly characterised as the uniform distribution over $A^* := \{a^* \in A : Q^\pi(s, a^*) = \max_{a \in A} Q^\pi(s, a)\}$.

Proof. Let us denote $A_t^* := \{a^* \in A : Q^\pi(s_t, a^*) = \max_{a \in A} Q^\pi(s_t, a)\}$ and $a_t^* \in A^*$ arbitrarily chosen.
For any $1 \leqslant t \leqslant T$ we have:

$$\begin{aligned}
\lim_{\delta \to 0} \pi^\delta(a_t^*|s_t) &= \lim_{\delta \to 0} \frac{\exp\left(\frac{1}{\delta} Q^\pi(s_t, a_t^*)\right)}{\sum_{a \in A} \exp\left(\frac{1}{\delta} Q^\pi(s_t, a)\right)} \\
&= \lim_{\delta \to 0} \frac{\exp\left(\frac{1}{\delta} Q^\pi(s_t, a^*) - Q^\pi(s_t, a^*)\right)}{\sum_{a \in A} \exp\left(\frac{1}{\delta} Q^\pi(s_t, a) - Q^\pi(s_t, a^*)\right)} \\
&= \lim_{\delta \to 0} \frac{1}{\sum_{a \in A} \exp\left(\frac{1}{\delta} Q^\pi(s_t, a) - Q^\pi(s_t, a^*)\right)}
\end{aligned}$$

where the denominator now splits into:

$$\lim_{\delta \to 0} \sum_{a \in A_t^*} \exp\left(\frac{1}{\delta} Q^\pi(s_t, a) - Q^\pi(s_t, a^*)\right) = |A_t^*|$$

$$\lim_{\delta \to 0} \sum_{a \in A \setminus A_t^*} \exp\left(\frac{1}{\delta} Q^\pi(s_t, a) - Q^\pi(s_t, a^*)\right) = 0$$

giving $\pi^\delta(a|s_t)$ uniform on A_t^*. □

Corollary A.6. *A necessary condition for policy $\pi(a|s)$ to be optimal is that:*

$$\lim_{\delta \to 0} \kappa_\delta(\pi) < \infty$$

Proof. From Definition 4.12, Proposition 4.7 and Proposition 4.13, we write:

$$\begin{aligned}
\lim_{\delta\to 0}\kappa_\delta(\pi) &= \lim_{\delta\to 0}\mathrm{KL}_\tau(\pi(\tau)|\pi^\delta(\tau)) \\
&= \lim_{\delta\to 0}\sum_{t=1}^{T}\mathbb{E}_{s_t\sim d_\pi^t}\mathrm{KL}(\pi(\cdot|s_t)||\pi^\delta(\cdot|s_t)) \\
&= \sum_{t=1}^{T}\mathbb{E}_{s_t\sim d_\pi^t}\mathrm{KL}(\pi(\cdot|s_t)||\lim_{\delta\to 0}\pi^\delta(\cdot|s_t)) \\
&= \sum_{t=1}^{T}\mathbb{E}_{s_t\sim d_\pi^t}\mathrm{KL}(\pi(\cdot|s_t)||\pi^*(\cdot|s_t))
\end{aligned}$$

Thus, for any $s_t \sim d_\pi^t$, the $\mathrm{KL}(\pi||\pi^*)$ is finite if and only $\mathrm{supp}(\pi^*) \subseteq \mathrm{supp}(\pi)$. Meaning that π must assign positive mass to every action in A_t^* for each t. But this is a necessary condition for optimality (by backwards induction on t). □

A.2 Proofs for Section 5

Proposition A.7 (Strong return improvement lemma)**.** *The sequence of policies $(\pi_i^{\mathcal{G}})_{i=0,1,\dots}$ given by the control-by-inference improves its return monotonically, that is:*

$$J(\pi_{i+1}^{\mathcal{G}}) \geqslant J(\pi_i^{\mathcal{G}})$$

Proof. We prove this by backwards induction on T. First, fix some index i and denote for brevity:

$$\pi = \pi_i^{\mathcal{G}} \qquad \pi' = \pi_{i+1}^{\mathcal{G}}$$

We also define the future probability of success in a state s_t as in Definition 4.1:

$$\begin{aligned}
V^\pi(s_t) &= \log p_\pi(\mathcal{O}_{t:T} = 1|s_t) \\
Q^\pi(s_t, a_t) &= \log p_\pi(\mathcal{O}_{t:T} = 1|s_t, a_t)
\end{aligned}$$

Now, we proceed by induction to show that:

$$V^{\pi'}(s_t) \geqslant V^\pi(s_t)$$

The policy $\pi' = \mathcal{G}(\pi)$ can be written from Bayes theorem as:

$$\pi'(a_t|s_t) = \frac{\pi(a_t|s_t)\exp Q^\pi(a_t|s_t)}{\mathbb{E}_{a_t\sim\pi(\cdot|s_t)}[\exp Q^\pi(a|s_t)]} \tag{1}$$

Now:

$$\begin{aligned}
\mathbb{E}_{a_t\sim\pi'(\cdot|s_t)}[\exp Q^\pi(a|s_t)] &= \frac{\mathbb{E}_{a_t\sim\pi(\cdot|s_t)}\left[(\exp Q^\pi(a|s_t))^2\right]}{\mathbb{E}_{a_t\sim\pi(\cdot|s_t)}[\exp Q^\pi(a|s_t)]} \\
&\geqslant \mathbb{E}_{a_t\sim\pi(\cdot|s_t)}[\exp Q^\pi(a|s_t)]
\end{aligned} \tag{2}$$

where the first line follows from Equation (1), and the second line follows from Cauchy–Schwarz inequality.

First, assume $t = T$. Then, $Q^\pi(a_T|s_T)$ does not depend on π and is simply given by $r(s_t, a_t)$. We then immediately have from Equation (2) that:

$$V^{\pi'}(\mathcal{O}_{T:T}|s_t) \geqslant V^\pi(\mathcal{O}_{T:T}|s_t)$$

establishing the base of the induction. For the inductive case $t < T$, assume that the above holds for all $t < t' \leqslant T$

and write:

$$\begin{aligned}
\exp V^{\pi'}(s_t) &= p_{\pi'}(\mathcal{O}_{t:T} = 1|s_t) \\
&= \mathbb{E}_{a_t \sim \pi'(\cdot|s_t)} \left[e^{r(s_t,a_t)} \mathbb{E}_{s_{t+1}} \exp V^{\pi'}(s_{t+1}) \right] \\
&\geqslant \mathbb{E}_{a_t \sim \pi'(\cdot|s_t)} \left[e^{r(s_t,a_t)} \mathbb{E}_{s_{t+1}} \exp V^{\pi}(s_{t+1}) \right] \\
&= \mathbb{E}_{a_t \sim \pi'(\cdot|s_t)} \left[\exp Q^{\pi}(s_t, a_t) \right] \\
&\geqslant \mathbb{E}_{a_t \sim \pi(\cdot|s_t)} \left[\exp Q^{\pi}(s_t, a_t) \right] \\
&= \exp V^{\pi}(s_t)
\end{aligned}$$

where the first inequality comes from the inductive hypothesis, and the second from Equation (2). From the Cauchy–Schwarz, inequalities are strict unless $Q^{\pi}(s_t, \cdot)$ is $\pi(\cdot|s_t)$-almost-surely constant. Intuitively, the update is non-trivial exactly when the old policy is not already proportional to the frozen-future success factors. □

A.3 Proof of Section 5.1

We split the proof of Theorem 5.9 into several lemmas.

Lemma A.8. *For any prior $p(a|s)$, non-positive reward function $r(s_t, a_t)$, deterministic dynamics $\tau(s_{t+1}|s_t, a_t)$, and $k \in \mathbb{N}_+$ we have $\pi_{\alpha(k)} = \pi_k^{\mathcal{H}}$, for $\alpha(k) = 2^k$.*

Proof. Throughout the proof, we will assume that the prior is uniform; otherwise the prior is folded into the reward. Let us first consider the case of $T = 1$. We have that:

$$\begin{aligned}
\log p(a|s, \mathcal{O}^{(k)}) &= \log \left(\frac{p(\mathcal{O}^{(k)}|a, s) p(a|s)}{p(\mathcal{O}^{(k)}|s)} \right) \\
&= r_k(s, a) + C
\end{aligned}$$

where the constant

$$C = \log |A| - \log \frac{\sum_a \exp r_k(s, a)}{|A|}$$

does not depend on a. We also observe that shifting reward by any additive constant does not change the policy (a soft version of the reward shaping lemma): for $r'(s, a) = r(s, a) + C$ we have

$$\begin{aligned}
p(a|s, \mathcal{O}') &= \frac{p(a|s) e^{r(s,a)+C}}{\sum_{a'} p(a'|s) e^{r(s,a)+C}} \\
&= \frac{p(a|s) e^{r(s,a)}}{\sum_{a'} p(a'|s) e^{r(s,a)}}
\end{aligned}$$

To ignore the constants, we write $r \equiv r'$ when $r = r' + C$. Thus, the iterative process of adding the posterior to the reward results in a sequence:

$$\begin{aligned}
r_0(s, a) &\equiv r(s, a) \\
r_1(s, a) &\equiv r_0(s, a) + \log p(a|s, \mathcal{O}^0) \\
&\equiv 2r(s, a) + C_1 \equiv 2r(s, a) \\
&\dots \\
r_k(s, a) &\equiv r_{k-1}(s, a) + \log p(a|s, \mathcal{O}^{(k-1)}) \\
&\equiv 2r_{k-1}(s, a) + C_k \equiv 2^k r(s, a)
\end{aligned}$$

This proves, by induction on k, that $\pi_k^{\mathcal{H}} = \pi_{\alpha(k)}$.
The case of $T > 1$ is then done by backwards induction over the time horizon. We are given $1 \leqslant t < T$ and we assume that the following inductive hypothesis holds for all $t < t'$ and for all k:

$$r_k(s, a) = 2^k r_0(s, a) \qquad p(\mathcal{O}_{t':T}^{\alpha(k)}|s_{t'}) = p(\mathcal{O}_{t':T}^{(k)}|s_{t'})$$

In case of the policy obtained by decreasing the temperature, we have:

$$\begin{aligned}\pi_{\alpha(k+1)}(a_t|s_t) &= \log p(a_t|s_t, \mathcal{O}_{t:T}^{\alpha(k+1)}) \\ &\equiv \log p(\mathcal{O}_{t:T}^{\alpha(k+1)}|a_t, s_t) + \log p(a_t|s_t) - \log p(\mathcal{O}_{t:T}|s_t) \\ &\equiv \log p(\mathcal{O}_{t}^{\alpha(k+1)}|a_t, s_t) + \log p(\mathcal{O}_{t+1:T}^{\alpha(k+1)}|a_t, s_t) \\ &= 2^{k+1} r_0(a_t, s_t) + \log p(\mathcal{O}_{t+1:T}^{\alpha(k+1)}|s_{t+1})\end{aligned}$$

where the first line follows from the definition, the next one from the Bayes' law, the next one from the assumption that the prior is uniform, and the next one from the definition of the $\mathcal{O}_1^{\alpha(k+1)}$ and using the fact that there is a unique s_{t+1} which follows $\tau(a_t, s_t)$.

We look at the policy obtained by folding the posterior into the reward:

$$\begin{aligned}\pi_{(k+1)}^{\mathcal{H}}(a_t|s_t) &= \log p(a_t|s_t, \mathcal{O}_{t:T}^{(k+1)}) \\ &\equiv \log p(\mathcal{O}_{t}^{(k+1)}|a_t, s_t) + \log p(\mathcal{O}_{t+1:T}^{(k+1)}|a_t, s_t) \\ &= r_{k+1}(a_t, s_t) + \log p(\mathcal{O}_{t+1:T}^{(k+1)}|s_{t+1}) \\ &= 2^{k+1} r_0(a_t, s_t) + \log p(\mathcal{O}_{t+1:T}^{\alpha(k+1)}|s_{t+1})\end{aligned}$$

where the last line follows from the inductive hypothesis. □

Lemma A.9. *For any prior $p(a|s)$, non-positive reward function $r(s_t, a_t)$, arbitrary dynamics $\tau(s_{t+1}|s_t, a_t)$, and $k \in \mathbb{N}_+$ we have $\pi_k^{\mathcal{G}} = \pi_k^{\mathcal{F}}$.*

Proof. Proof by induction on k. Assume without loss of generality uniform prior p, otherwise fold it into the reward. For $k = 0$ they coincide by definition. Now, assume that $\pi_k^{\mathcal{G}} = \pi_k^{\mathcal{F}}$, and we show the inductive step for $k+1$. Let us denote the corresponding Q functions by $Q_k^{\mathcal{G}}$ and $Q_k^{\mathcal{F}}$. First, exactly as in the proof of Proposition 5.4, we know that:

$$\pi_{k+1}^{\mathcal{G}}(a_t|s_t) = \frac{\pi_k^{\mathcal{G}}(a_t|s_t) \exp Q_k^{\mathcal{G}}(a_t|s_t)}{\mathbb{E}_{a_t \sim \pi(\cdot|s_t)}[\exp Q_k^{\mathcal{G}}(a|s_t)]} \propto \pi_k^{\mathcal{G}}(a_t|s_t) \exp Q^{\mathcal{G}}(a_t|s_t)$$

and on the other hand, we also know from the definition that:

$$\pi_{k+1}^{\mathcal{F}}(a_t|s_t) \propto p(s_t|a_t) \exp Q_k^{\mathcal{F}}(a_t|s_t)$$

Thus, we aim to show that:

$$p(s_t|a_t) \exp Q_k^{\mathcal{F}}(a_t|s_t) = C_t \pi_k^{\mathcal{G}}(a_t|s_t) \exp Q_k^{\mathcal{G}}(a_t|s_t)$$

up to some multiplicative constant C_t. We do this again by induction, this time on a time horizon t. Using Definition 4.1, we have:

$$\exp Q_k^{\mathcal{F}}(s_T, a_T) = \exp r_k(s_T, a_T) = \pi_k^{\mathcal{G}}(a_T|s_T) \exp r_0(s_T, a_T) = \pi_k^{\mathcal{G}}(a_T|s_T) \exp Q_k^{\mathcal{G}}(s_T, a_T)$$

which proves this in case $t = T$. Inductive step: assume that the hypothesis holds for $t + 1$, and write:

$$\exp V^{\mathcal{F}}(s_{t+1}) = \mathbb{E}_{a_{t+1} \sim p}[\exp Q_k^{\mathcal{F}}(s_{t+1}(a_{t+1})] = \frac{C_t}{|A|} \mathbb{E}_{a_{t+1} \sim \pi_k^{\mathcal{G}}}[\exp Q_k^{\mathcal{G}}(s_{t+1}(a_{t+1})]$$

Taking log and bringing the factor outside the expectation shows that:

$$\mathbb{E}_{s_{t+1}}[\exp V_k^{\mathcal{F}}(s_{t+1})] = \frac{C_t}{|A|} \mathbb{E}_{s_{t+1}}[\exp V_k^{\mathcal{G}}(s_{t+1})]$$

which means:

$$\exp Q_k^{\mathcal{F}}(s_t, a_t) = \exp r_k(s_t, a_t) \cdot \frac{C_t}{|A|} \mathbb{E}_{s_{t+1}}[\exp V_k^{\mathcal{G}}(s_{t+1})]$$

proving the lemma. □

Lemma A.10. *For any prior $p(a|s)$, non-positive reward function $r(s_t, a_t)$, deterministic dynamics $\tau(s_{t+1}|s_t, a_t)$, and $k \in \mathbb{N}_+$ we have $\pi_k^{\mathcal{F}} = \pi_{\alpha(k)}$ for schedule $\alpha(k) = k$.*

Proof. The proof follows almost exactly like the one in Lemma A.8. The base case is identical. In the inductive case, we write analogously:

$$\begin{aligned}
r_0(s,a) &\equiv r(s,a) \\
r_1(s,a) &\equiv r_0(s,a) + \log p(a|s, \mathcal{O}^0) \\
&\equiv 2r(s,a) + C_1 \equiv 2r(s,a) \\
&\dots \\
r_k(s,a) &\equiv r_0(s,a) + \log p(a|s, \mathcal{O}^{(k-1)}) \\
&\equiv kr(s,a) + C_k \equiv kr(s,a)
\end{aligned}$$

□

and the rest of the proof follows as before.

A.4 Proof of Corollary 5.10

Lemma A.11. *Given a function $h(x) = f(\arg\max_t(f(t) + xg(t)))$ for some differentiable functions f, g, the derivative of $h(x)$ is:*

$$\frac{dh}{dx} = -\frac{f'(t^*(x))g'(t^*(x))}{f''(t^*(x)) + xg''(t^*(x))}$$

where $t^*(x) = \arg\max_t f(t) + xg(t)$.

Proof. Let $t^*(x) = \arg\max_t(f(t) + xg(t))$. Therefore, we have:

$$\frac{d}{dt}\left(f(t) + xg(t)\right)\Big|_{t=t^*(x)} = 0 = f'(t^*(x)) + xg'(t^*(x))$$

Thus:

$$x = -\frac{f'(t^*(x))}{g'(t^*(x))}$$

To differentiate $h(x)$, we use the fact that $h(x) = f(t^*(x))$. This gives us:

$$\frac{dh}{dx} = \frac{d}{dx} f(t^*(x))$$

By the chain rule:

$$\frac{dh}{dx} = f'(t^*(x)) \cdot \frac{dt^*(x)}{dx}$$

To find $\frac{dt^*(x)}{dx}$, we differentiate the first-order condition $f'(t^*(x)) + xg'(t^*(x)) = 0$ with respect to x:

$$\frac{d}{dx}\left(f'(t^*(x)) + xg'(t^*(x))\right) = 0$$

Applying the chain rule:

$$f''(t^*(x)) \cdot \frac{dt^*(x)}{dx} + g'(t^*(x)) + xg''(t^*(x)) \cdot \frac{dt^*(x)}{dx} = 0$$

Rearrange to solve for $\frac{dt^*(x)}{dx}$:

$$\left(f''(t^*(x)) + xg''(t^*(x))\right)\frac{dt^*(x)}{dx} = -g'(t^*(x))$$

$$\frac{dt^*(x)}{dx} = -\frac{g'(t^*(x))}{f''(t^*(x)) + xg''(t^*(x))}$$

Finally, substituting this back into the expression for $\frac{dh}{dx}$:

$$\frac{dh}{dx} = f'(t^*(x)) \cdot \left(-\frac{g'(t^*(x))}{f''(t^*(x)) + xg''(t^*(x))} \right)$$

□

Corollary A.12 (Return improvement rate)**.** *In case of deterministic dynamics τ, given a sequence of policies $\pi_i^{\mathcal{G}}$, for $k \in \mathbb{N}_+$ have that:*

$$J(\pi_k^{\mathcal{G}}) - J(\pi_{k-1}^{\mathcal{G}}) = \frac{1}{k} \cdot \frac{J'(\pi_k^{\mathcal{G}}) \cdot \hat{\mathcal{H}}'(\pi_k^{\mathcal{G}})}{J''(\pi_k^{\mathcal{G}}) + \frac{1}{k}\hat{\mathcal{H}}''(\pi_k^{\mathcal{G}})} + O\left(\frac{1}{k^2}\right)$$

where $\hat{\mathcal{H}}(\pi)$ denotes the causal entropy of policy π, and the derivatives are taken with respect to the temperature α.

Proof. To prove this, we use an alternative characterization of the maximum entropy policies (Haarnoja et al., 2017; Levine, 2018). For a fixed temperature $\alpha \in \mathbb{R}_+$, the policy π_α maximizes the functional:

$$J_\alpha(\pi) = \mathbb{E}_{s_t, a_t}[r(s_t, a_t) + \frac{1}{\alpha}\mathcal{H}(\pi(\cdot|s_t))] = J(\pi) + \frac{1}{\alpha}\hat{\mathcal{H}}(\pi)$$

where the expectation is taken over the trajectory induced by the policy π, $\mathcal{H}(\pi(\cdot|s))$ denotes the Shannon entropy of the policy in state s, and $\hat{\mathcal{H}}$ denotes the causal entropy of the policy π. By Lemma A.10 we know that $\pi_j^{\mathcal{G}} = \pi_{\alpha(j)}$ for $\alpha(j) = j$.

From the Taylor expansion, for any twice-differentiable $u(\alpha)$ we have:

$$u(\alpha + \eta) - u(\alpha) = \eta u'(\alpha) + O(\eta^2)$$

Substituting $u(\alpha) = J(\arg\max_\pi J_{\frac{1}{\alpha}}(\pi))$, we obtain:

$$\begin{aligned} J(\pi_{k-1}^{\mathcal{G}}) - J(\pi_k^{\mathcal{G}}) &= J(\pi_{\alpha(k-1)}) - J(\pi_{\alpha(k)}) \\ &= u\left(\frac{1}{k} + \eta\right) - u\left(\frac{1}{k}\right) \\ &= \eta u'\left(\frac{1}{k}\right) + O\left(\eta^2\right) \end{aligned}$$

for $\eta = \frac{1}{k(k-1)}$. Now, using Lemma A.11, we derive:

$$u'(x) = -\frac{J'(\pi^*(x))\hat{\mathcal{H}}'(\pi^*(x))}{J''(\pi^*(x)) + x\hat{\mathcal{H}}''(\pi^*(x))}$$

and substituting $x = \frac{1}{k}$, and therefore $\pi^*(x) = \pi_{\alpha(k)} = \pi_k^{\mathcal{G}}$, we get:

$$u'(x) = -\frac{J'(\pi_{\alpha(k)})\hat{\mathcal{H}}'(\pi_{\alpha(k)})}{J''(\pi_k^{\mathcal{G}}) + \frac{1}{k}\hat{\mathcal{H}}''(\pi_{\alpha(k)})}$$

which gives us the final equation:

$$J(\pi_k^{\mathcal{G}}) - J(\pi_{k-1}^{\mathcal{G}}) = \frac{1}{k}\frac{J'(\pi_k^{\mathcal{G}})\hat{\mathcal{H}}'(\pi_k^{\mathcal{G}})}{J''(\pi_k^{\mathcal{G}})) + \frac{1}{k}\hat{\mathcal{H}}''(\pi_k^{\mathcal{G}})} + O\left(\frac{1}{k}^2\right)$$

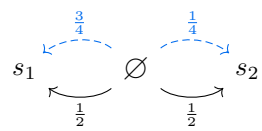
□

B EXAMPLE 3 EXPLICIT CALCULATION

Calculation for the Example 3 in detail.
We take a three-state, two action Markov decision process with initial state $\varnothing$ and a uniform prior policy:

$$\pi(a_1|\varnothing) = \pi(a_2|\varnothing) = \frac{1}{2}$$

with transition dynamics depicted on the diagram below:



where red dotted lines correspond to action a_1, and solid black lines to action a_2. We also assume that:

$$r(s_0, \cdot) = \log 1 \qquad r(s_1, \cdot) = \log \frac{1}{3} \qquad r(s_2, \cdot) = \log \frac{2}{3}$$

B.1 Control-as-inference operator

The first iteration of $\pi_1^{\mathcal{G}}(\cdot|\varnothing)$ computes the probability of getting reward for action a_1:

$$\pi_1^{\mathcal{G}}(a_1|\varnothing) = \mathbb{P}(a_1|\mathcal{O}_{1:2}^2, \varnothing) = \frac{\mathbb{P}(\mathcal{O}_{1:2}^2|a_1, \varnothing)\mathbb{P}(a_1|\varnothing)}{\mathbb{P}(\mathcal{O}_{1:2}^2)} = \left(\frac{1}{2}\cdot\frac{3}{4}\cdot\frac{1}{3} + \frac{1}{2}\cdot\frac{1}{4}\cdot\frac{2}{3}\right)/\mathbb{P}(\mathcal{O}_{1:2}^2) = \frac{5}{24}/\mathbb{P}(\mathcal{O}_{1:2}^2)$$

and for a_2:

$$\pi_1^{\mathcal{G}}(a_2|\varnothing) = \mathbb{P}(a_2|\mathcal{O}_{1:2}^2, \varnothing) = \frac{\mathbb{P}(\mathcal{O}_{1:2}^2|a_2, \varnothing)\mathbb{P}(a_2|\varnothing)}{\mathbb{P}(\mathcal{O}_{1:2}^2)} = \left(\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{3} + \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{2}{3}\right)/\mathbb{P}(\mathcal{O}_{1:2}^2) = \frac{6}{24}/\mathbb{P}(\mathcal{O}_{1:2}^2)$$

After normalisation, we get:

$$\pi_1^{\mathcal{G}}(a_1|\varnothing) = \frac{5}{11} \qquad \pi_1^{\mathcal{G}}(a_2|\varnothing) = \frac{6}{11}$$

Now, applying the same computation second time - for a_1:

$$\pi_2^{\mathcal{G}}(a_1|\varnothing) = \mathbb{P}(a_1|\mathcal{O}_{1:2}^2, \varnothing) = \frac{\mathbb{P}(\mathcal{O}_{1:2}^2|a_1, \varnothing)\pi_1^{\mathcal{G}}(a_1|\varnothing)}{\mathbb{P}(\mathcal{O}_{1:2}^2)} = \left(\frac{5}{11}\cdot\frac{3}{4}\cdot\frac{1}{3} + \frac{5}{11}\cdot\frac{1}{4}\cdot\frac{2}{3}\right)/\mathbb{P}(\mathcal{O}_{1:2}^2) = \frac{25}{11\cdot 12}/\mathbb{P}(\mathcal{O}_{1:2}^2)$$

and for a_2:

$$\pi_2^{\mathcal{G}}(a_2|\varnothing) = \mathbb{P}(a_1|\mathcal{O}_{1:2}^2, \varnothing) = \frac{\mathbb{P}(\mathcal{O}_{1:2}^2|a_2, \varnothing)\pi_1^{\mathcal{G}}(a_2|\varnothing)}{\mathbb{P}(\mathcal{O}_{1:2}^2)} = \left(\frac{6}{11}\cdot\frac{3}{4}\cdot\frac{1}{3} + \frac{6}{11}\cdot\frac{1}{4}\cdot\frac{2}{3}\right)/\mathbb{P}(\mathcal{O}_{1:2}^2) = \frac{36}{11\cdot 12}/\mathbb{P}(\mathcal{O}_{1:2}^2)$$

Again applying normalisation:

$$\pi_2^{\mathcal{G}}(a_1|\varnothing) = \frac{25}{61} \qquad \pi_2^{\mathcal{G}}(a_2|\varnothing) = \frac{36}{61}$$

B.2 Raising temperature

Raising temperature policy $\pi_{\alpha(2)}(\cdot|\varnothing)$ first modifies the MDP by setting the rewards:

$$r_2(s_1) = \log \frac{1}{9} \qquad r_2(s_2) = \log \frac{4}{9}$$

and then recomputes the posterior for a_1:

$$\pi_{\alpha(2)}(a_1|\varnothing) = \mathbb{P}(a_1|\mathcal{O}_{1:2}^2, \varnothing) = \frac{\mathbb{P}(\mathcal{O}_{1:2}^2|a_1, \varnothing)\mathbb{P}(a_1|\varnothing)}{\mathbb{P}(\mathcal{O}_{1:2}^2)} = \left(\frac{1}{2}\cdot\frac{3}{4}\cdot\frac{1}{9} + \frac{1}{2}\cdot\frac{1}{4}\cdot\frac{4}{9}\right)/\mathbb{P}(\mathcal{O}_{1:2}^2) = \frac{7}{72}/\mathbb{P}(\mathcal{O}_{1:2}^2)$$

and a_2:

$$\pi_{\alpha(2)}(a_2|\varnothing) = \mathbb{P}(a_1|\mathcal{O}_{1:2}^2, \varnothing) = \frac{\mathbb{P}(\mathcal{O}_{1:2}^2|a_2, \varnothing)\mathbb{P}(a_2|\varnothing)}{\mathbb{P}(\mathcal{O}_{1:2}^2)} = \left(\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{9} + \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{4}{9}\right)/\mathbb{P}(\mathcal{O}_{1:2}^2) = \frac{10}{72}/\mathbb{P}(\mathcal{O}_{1:2}^2)$$

Again applying normalisation:

$$\pi_{\alpha(2)}(a_1|\varnothing) = \frac{7}{17} \qquad \pi_{\alpha(2)}(a_2|\varnothing) = \frac{10}{17}$$

C ITERATED BOLTZMANN CONVERGENCE

Example 4 (Boltzmann-coherent mountain race)**.** *Let us revisit the Example 1. We might compute the Boltzmann-coherent policy by using the construction from Definition 4.9.*

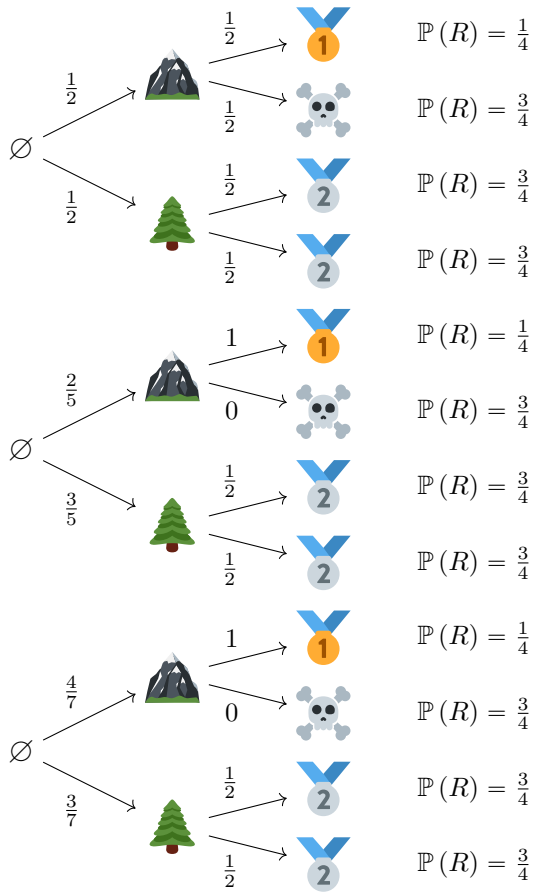


Figure 4: The fixed point of iterated f-coherence achieved after $t = 2 = T$ iterations.

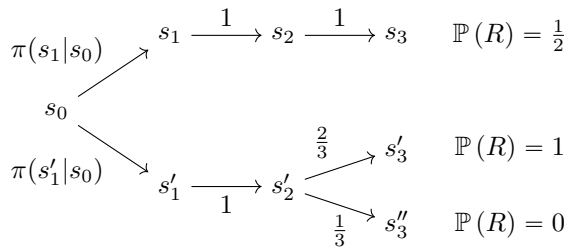


Figure 5: π cannot satisfy both 2-policy-stability and 1-policy-stability.

D POLICY STABILITY

In search of the sufficient conditions for the autoregressive goal-conditioned policy to be optimal, one point of focus might be the question of how to *extend* trajectories: we might hope to stitch the optimal trajectory from actions that are optimal at a k-step-lookahead; optimal policies should then be those that behave consistently, in the sense that it would make the same decisions as if it were allowed to take actions looking k steps into the future. Formally, we have the following definition.

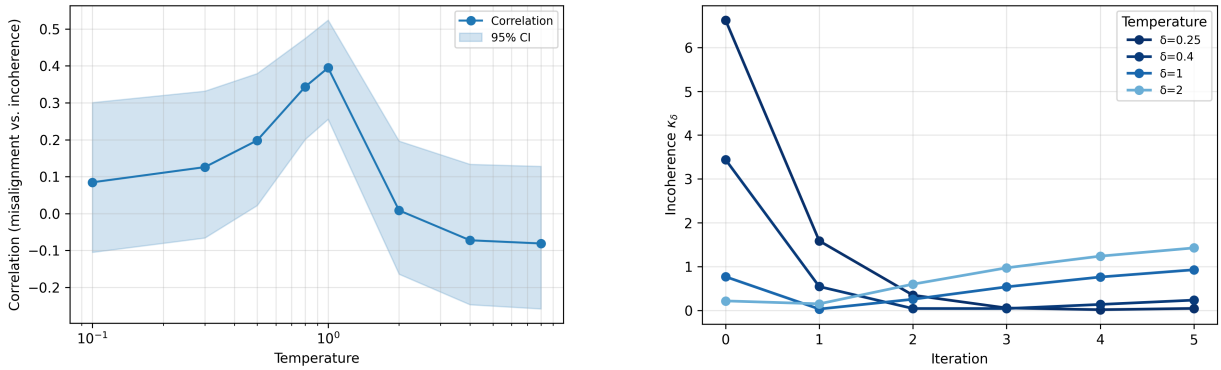
Definition D.1 (Policy-stable reasoning)**.** Given deterministic dynamics τ and a prior $p(a_t|s_t)$, we say that a policy $\pi(a_t|s_t) \propto p(\mathcal{O}_{t:T}|s_t, a_t)$ is *(1-)policy-stable*, if for any actions $a_t^1, a_{t+1}^1, a_t^2, a_{t+1}^2$, with $a_{t+1}^i = \arg\max_{a_{t+1}^i} p(\mathcal{O}_{t:T}|a_{t,t+1}^1, s_t)$ (where $s_{t+1}^i = \tau(s_t, a_t^i)$), we have that:

$$p(\mathcal{O}_{t:T}|a_{t,t+1}^1, s_t) > p(\mathcal{O}_{t:T}|a_{t,t+1}^2, s_t) \implies p(\mathcal{O}_{t:T}|a_t^1, s_t) > p(\mathcal{O}_{t:T}|a_t^2, s_t)$$

In other words: given that that a_{t+1}^i best continues a_t^i, if (a_t^1, a_{t+1}^1) is preferred to (a_t^2, a_{t+1}^2), then a_t^1 should be preferred to a_t^2.

Unfortunately, not only this is not a sufficient condition - in some instances, it is even contradicting optimality. This is because it is impossible to properly extend the the policy-stability over an arbitrary number of actions. To show that, we can introduce a notion of n-policy-stable predictor, which says that for any sequences $a_{1:k}, a'_{1:k}$ and actions a, a', such that the sequence $a_{1:k}$ best continues a and sequence $a'_{1:k}$ best continues a', we have that $(a, a_{1:k})$ being preferred over $(a', a'_{1:k})$ implies that a is preferred to a'.

From direct calculation, it can be seen that, if a policy derived by control-as-inference from a prior over the MDP shown in the Figure 5 is to be 1-policy-stable, then it must be the case that $\pi(s_1|s_0) \geqslant \pi(s_1'|s_0)$, while considering two actions into the future and then following the prior, moving into s_2 is a better choice. However, to be 2-policy-stable, it has to satisfy $\pi(s_1|s_0) < \pi(s_1'|s_0)$, since after considering three moves in the future, moving to s_1' becomes the better choice. Since 1-policy-stable and 2-policy-stable are mutually exclusive for a policy derived by control-as-inference from this prior, one has to make a choice as to which n-policy-stability to require, which, in practice, requires the knowledge of the time horizon (and for it to be fixed and finite).



(a) (b)

Figure 6: (a) An aggregated version of the graph in Figure 3a, each point corresponding to a particular choice of temperature. Correlation between misalignment and incoherence is only meaningful in moderate temperatures near $\delta = 1$. (b) Experimental illustration of Corollary 5.11, only in the simultaneous limit of $(\delta, i) \to (0, \infty)$ incoherence vanishes.

E EXPERIMENTAL DETAILS

E.1 MDP environments

We perform all experiments on a suite of 140 small randomly generated tree MDPs of varying sizes. We set the number of actions $|A|$ between 2 and 4, and time horizon T between 3 and 6, resulting in environments ranging in state space sizes between 4 and 122, and total state-action sizes between 8 and 366. For the misalignment experiment, we used both deterministic and stochastic environments in equal proportions, with transition probabilities and rewards sampled uniformly from relevant distributions. For the estimated effective horizon experiment, we used only deterministic environments, because otherwise the estimator was difficult to compute. Full details of the environments are available in the released code repository available at `https://github.com/jkarwowski/incoherence`, in the config file for each experiment.

E.2 Estimating effective horizon

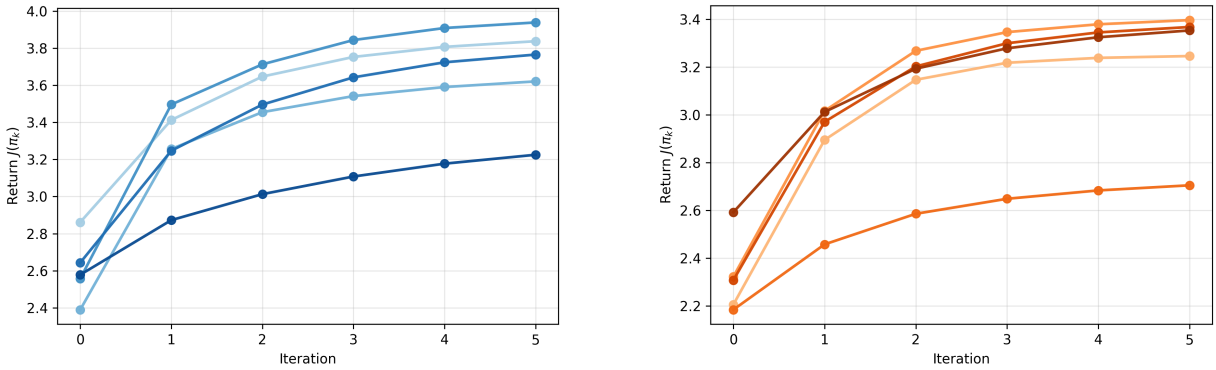
Effective horizon is a metric that is difficult to compute exactly even for small environments. We approximate $\hat{H}$ using Monte Carlo simulation and , to output $\min_k H_k$, where $H_k = k + \log_A m$ for which $\mathrm{GORP}(k, m)$ recovers the optimal policy with probability at least 1/2, as in Laidlaw et al. (2024, Definition 5.2). We note this is a different estimator from the one used by the original paper - see Appendix C and their codebase. This is because they work on much larger environments such as Atari games, for which our approach would be impractical. On the other hand, their upper bounds such as those using Laidlaw et al. (2024, Theorem 5.4) turned to be insufficiently tight for our purposes.

E.3 Misalignment versus incoherence

The correlation between incoherence and misalignment depends on the specific value of temperature δ. We present the graph of correlations in Figure 6a. For small values of δ, the policy is converges to deterministic, making incoherence very large variance and correlations disappear. On the other hand, in the limit of $\delta \to \infty$, the policy tends to uniform distribution and again rendering incoherence useless. We note that this is only a heuristic explanation, note and we do not have a satisfactory theoretical explanation for the exact shape of the graph.

E.4 Experimental validation of Corollary 5.11 and Proposition 5.4

Corollary 5.11 states that in the limit of both $(\delta, i) \to (0, \infty)$, incoherence $\kappa_\delta(\pi^i)$ vanishes. We note that taking both limits simultaneously is essential: indeed, holding one coordinate fixed, we do not get convergence to 0. We



(a) (b)

Figure 7: Experimental validation of Corollary 5.11, (a) on 5 randomly chosen deterministic MDPs, and (b) on 5 randomly chosen stochastic environments.

illustrate this point experimentally in Figure 6b. Each line shows incoherence for a different choice of temperature δ. Holding δ fixed and increasing iteration i, the incoherence of π^i initially decreases, such that at a certain time it dips below every higher temperature, but later starts to increase. This is because Proposition 5.4 guarantees the improvement in return, which is not equal to the KL-regularised return optimised for by the soft-conditioned policies.

Proposition 5.4 states that regardless of the environment, we should see a consistent improvement in return. We validate this empirically for 5 randomly chosen deterministic-dynamics MDPs and 5 randomly chosen stochastic-dynamics MDPs, $|A| = 2$, $|T| = 4$, both with $i = 5$ iteration steps, with results conforming to the theory.

E.5 Confidence intervals

All confidence intervals in the paper are constructed with the bootstrap method, using $B = 2000$ samples.