Journal of Machine Learning Research 26 (2025) 1-24 Submitted 7/23; Revised 1/25; Published 5/25

Statistical field theory for Markov decision processes under uncertainty

George Stamatescu GEORGE.STAMATESCU@ADELAIDE.EDU.AU
University of Adelaide
Adelaide, SA, Australia

Editor: George Konidaris

Abstract

A statistical field theory is introduced for finite state and action Markov decision processes with unknown parameters, in a Bayesian setting. The Bellman equation, for policy evaluation and the optimal value function in finite and discounted infinite horizon problems, is studied as a disordered interacting dynamical system. The Markov decision process transition probabilities and mean-rewards are interpreted as quenched random variables and the value functions, or the iterates of the Bellman equation, are deterministic variables that evolve dynamically. The posterior over value functions is then equivalent to the quenched average of Fourier inverse of the Martin-Siggia-Rose-De Dominicis-Janssen generating function. The formalism enables the use of methods from field theory to compute posterior moments of value functions. The paper presents two such methods, corresponding to two distinct asymptotic limits. First, the classical approximation is applied, corresponding to the asymptotic data limit. This approximation recovers so-called plug-in estimators for the mean of the value functions. Second, a dynamic mean field theory is derived, showing that under certain assumptions the state-action values are statistically independent across state-action pairs in the asymptotic state space limit. The state-action value statistics can be computed from a set of self-consistent mean field equations, which we call dynamic mean field programming (DMFP). Collectively, the results provide analytic insight into the structure of model uncertainty in Markov decision processes, and pave the way toward more advanced field theoretic techniques and applications to planning and reinforcement learning problems.

Keywords: Markov decision processes, reinforcement learning, statistical field theory, statistical physics, dynamic programming, disordered systems, mean field theory

1. Introduction

Sequential decision making in finite state and action Markov decision processes (MDPs) with uncertain parameters is a difficult problem. If further estimation or learning can continue while interacting with a particular MDP, one encounters a reinforcement learning (RL) or adaptive control problem. Otherwise, if no further estimation continues, one faces a problem of offline planning or control.

In the case that the MDP parameters are known, one can find a solution to the sequential decision making problem by dynamic programming. That is, one can find the optimal value function, the expected cumulative rewards under an optimal policy over a given horizon, by

©2025 George Stamatescu.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v26/23-0905.html.

solving the Bellman optimality equation. Or, given a specific policy for acting in the MDP, one can again solve the Bellman equation to evaluate the value function under that policy.

In the case that the parameters are unknown, one approach is to adopt a Bayesian view and maintain a set of posterior beliefs over the parameters. In the reinforcement learning setting, a complete Bayesian approach anticipative of future information involves planning in both the state space and the belief space jointly (Duff, 2002). This approach, referred to as dual control, is analytically and computationally intractable in all but the simplest cases, such as the case of a single state MDP (known as the "multi-armed bandit"). An alternative approach is to make sequential decisions based on the current beliefs only. This strategy is suitable for the offline planning problem, and can be deployed in the RL setting by periodically updating beliefs (Ghavamzadeh et al., 2015).

In order to devise a criterion for decision making with the current beliefs, it is useful to translate the posterior over parameters into a posterior over the value functions. The rationale for this translation is that one obtains a distribution over the long term average value of being in a particular state, or being in a state and taking a particular action (O'Donoghue, 2018; Luis et al., 2024). From such a distribution over value functions, one can make better decisions, in terms of robustness or statistically efficient learning (Dearden et al., 2013). Since the value functions are the solutions or fixed points of non-linear or linear Bellman equations, in the optimal and policy evaluation cases respectively, this too is an intractable problem in general.

In this paper we introduce a theoretical formalism from statistical field theory to study analytically the posterior distribution over value functions, both optimal and those under a policy. This formalism leads to exact analytical expressions for the posterior over value functions in the finite horizon case, and for the posterior, over fixed-point iterates of dynamic programming in the infinite horizon case.

Practically, the formalism replaces an intractable integral over the model parameters for an intractable integral over certain complex variables, often termed response fields in the statistical physics literature (Hertz et al., 2016). The utility in the alternative expressions is that they pave the way toward both closed form and computational approximations as well as encourage new theoretical studies of MDPs and their value functions under uncertainty. In particular, the formalism provides a range of methods for calculating the posterior moments of the value functions, such as perturbation theory or approximations derived from mean field theory (Chow and Buice, 2015; Zinn-Justin, 2021).

The contributions and organisation of the paper are as follows. First we introduce the theoretical formalism of generating functionals, or path integrals, for Bayesian MDPs. The introduction of this formalism, involving functional integrals, allows us to write down an analytic form of the posterior distribution over value functions, which are themselves considered as dynamic random variables. Next, we demonstrate how standard methods from perturbation theory can be used to compute uncertainty over value functions. As a simple case, we interpret the leading order term of the semi-classical approximation as corresponding to the asymptotic data limit, recovering a standard "plug-in" estimate for the mean of the value functions.

Unfortunately, while the perturbative approaches are well suited to policy evaluation due to its linearity, these approaches are not suited for the optimality equation, due to the maximum over actions. In this case, borrowing from an analogy with the spin glass and

neural network theory (Helias and Dahmen, 2019), we derive a dynamic mean field theory (DMFT) for the optimality equations. This is a first order self-consistent approximation which takes into account fluctuations for the calculation of the posterior mean. A claim of this paper is that this dynamic mean field result is exact under the limit of an asymptotically large state-space. More formally, a key theoretical result suggested from this mean field theory is a "propagation of chaos" property: the distribution over state-action value functions factorises over state-action pairs in the large state-space limit. This is not proven rigorously here but shown via the method of steepest descents. We then describe dynamic mean field equations for general posterior beliefs on parameters, and provide intuition for the mean field theory, which can be seen as a myopic approximation. Computational and closed-form approximations of the mean field equations and accompanying simulations are provided in the closing sections, validating the key results of the theory. The paper closes by discussing how field theory can be extended and how it can be applied to reinforcement learning problems.

2. Statistical field theory formalism for Bayesian MDPs

In this section we develop the theoretical formalism for Markov decision processes (MDPs) under Bayesian uncertainty. We first introduce our notation for MDPs and the basic assumptions made to establish the formalism. The presentation is provided for the discounted infinite horizon case, since the finite horizon case is captured as a special case.

2.1 Bayesian uncertainty in MDPs

An infinite horizon discounted MDP $M = (\mathcal{S}, \mathcal{A}, P, r, \beta)$, is specified by a state space $\mathcal{S}$, an action space $\mathcal{A}$; a transition function $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ the probability simplex and we say $P_{s'|sa}$ is the probability of transitioning into state s' upon taking action a in state s; a random reward $r_{sa} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for each state s and action a, with mean ρ_{sa} and finite variance, and a discount factor $\beta \in [0, 1)$. We denote the size of the state space as $|\mathcal{S}| = N$ and the action space as $|\mathcal{A}| = A$. A policy specifies a decision-making strategy. We restrict attention to the set of stationary policies $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. Given a fixed policy and fixed starting state $s_0 = s$, we define the state-action or Q-value function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as the discounted sum of future rewards

$$Q^\pi_{sa} = \mathbb{E}[\sum_{t=1}^{\infty} \beta^t r_{s_t a_t} | \pi, s_0 = s, a_0 = a] \tag{1}$$

where expectation is with respect to the randomness of the state trajectory, that is, the randomness in state transitions and π. The optimal state-action value function is defined as

$$Q^*_{sa} = \max_\pi Q^\pi_{sa} \tag{2}$$

The value function from a given state V^π_s is defined similarly, as is its optimal value V^*_s. Consequently we have the relationship $V^*_s = \max_a Q^*_{sa}$.

In the case the parameters are known, we may use dynamic programming methods to compute value functions. The optimal state-action value functions are the fixed points to

the Bellman optimality equation, which we write as

$$Q^*_{sa} = \rho_{sa} + \beta \sum_{s'} P_{s'|s,a} \max_{a'} Q^*_{s'a'} \tag{3}$$

The value function for a given policy in the infinite horizon is the fixed point to the Bellman equation for policy evaluation,

$$V^\pi_s = \sum_a \pi(a|s)\big(\rho_{sa} + \beta \sum_{s'} P_{s'|sa} V^\pi_{s'}\big) \tag{4}$$

The Bellman equation is always a contraction mapping for discount factors $\beta < 1$, for both policy evaluation and the optimality equation. Denoting the iterates of the optimality equations above as Q^k_{sa}, the iteration is

$$Q^{k+1}_{sa} = \rho_{sa} + \beta \sum_{s'} P_{s'|s,a} \max_{a'} Q^k_{s'a'} \tag{5}$$

Note that since the Bellman equation is a contraction mapping, the initial condition can be arbitrarily chosen, for example drawn from a Gaussian distribution or set identically to zero. A corresponding equation can be written for policy evaluation.

The Bayesian approach we adopt to model uncertainty, in the cases of unknown mean-rewards and transition probabilities, makes the following assumptions.

Assumption 1. (*Dirichlet probabilities.*) The transition probabilities are drawn from a Dirichlet distribution, for each state-action pair, $P_{sa} \sim \text{Dirichlet}(\alpha_{sa})$. The vector α_{sa} is N dimensional with elements $\alpha_{sa} = [\alpha_{1|sa}, \ldots, \alpha_{N|sa}]^T$.

Assumption 2. (*Arbitrary mean-reward distribution.*) The mean-reward ρ_{sa} has an arbitrary distribution with a moment generating function that exists. We denote its log moment generating function as $\mathcal{K}_{\rho_{sa}}$.

Assumption 3. (*Bounded effective horizon.*) We assume $\beta \in [0, 1)$ is fixed, meaning that the effective horizon $H_{\text{eff}} = \frac{1}{1-\beta}$ is bounded.

Assumption 4. (*Mean-reward and transition independence assumptions.*) We assume independence across the mean-rewards and transition functions $\rho_{sa} \perp \rho_{(sa)'}$, $P_{sa} \perp P_{(sa)'}$ and $\rho_{sa} \perp P_{(sa)'}$ for all pairs sa and $(sa)'$.

Let us denote the set of parameters with $\Theta = \{\rho, P\}$, where this is a set over all state-action pairs, and the set of hyper parameters as Ψ. Assumptions 1, 2 and 4 allow us to analytically calculate the integral over the parameters Θ.

The assumptions above are standard in the Bayesian reinforcement learning literature, (Dearden et al., 2013), (Poupart et al., 2006). Note that the assumptions may be weakened in certain cases and the analytic handle in the field theory below maintained. For example, the mean-rewards may have a jointly Gaussian posterior that does not factorise over state-action pairs. However, the stronger independence assumptions make the formalism more tractable, and are likely to underpin the asymptotic independence results presented in the following sections.

2.2 Field theoretic expressions for the posterior over value functions

We approach the study of the posterior over value functions by considering the posterior over iterates of the Bellman equation. This means we study distributions over dynamic variables

whose interactions are determined by "quenched" random variables, in the language of statistical physics. Let us denote the set of state action value functions at an iteration k as Q^k for short hand. The posterior distribution we wish to study is then given by,

$$p(Q^{1:K}|\Psi, Q^0) = \int d\Theta p(Q^{1:K}|\Theta, Q^0) p(\Theta|\Psi) \tag{6}$$

where K is fixed, and $p(Q^{1:K}|\Theta, Q^0)$ is a distribution over value function iterates given a realisation of the MDP parameters Θ and an initialisation Q^0. We must define this term in order to compute the posterior over dynamic programming iterates (6).

One approach is to consider a fictitious additive Gaussian noise of variance $\frac{\sigma_W^2}{2}$, added at each iteration to each Q_{sa}^k, and later consider the limit that its variance vanishes.

$$Q_{sa}^{k+1} = \rho_{sa} + \beta \sum_{s'} P_{s'|s,a} \max_{a'} Q_{s'a'}^k + W_{sa}^k \tag{7}$$

We can then define the distribution over the Q-value iterates given the model parameters,

$$p(Q^{1:K}|\Theta, Q^0) = \prod_{k=0}^{K-1} \int \prod_{sa} \delta\big(Q_{sa}^{k+1} - \rho_{sa} - \beta \sum_{s'} P_{s'|s,a} \max_{a'} Q_{s'a'}^k - W_{sa}^k\big) p(W_{sa}^k) dW_{sa}^k \tag{8}$$

In order to take the integral over Θ analytically we introduce a so-called response field $\tilde{Q}^{k+1}$ by representing the Dirac delta by its Fourier integral. This is equivalent to writing the Gaussian density of the fictitious noise as the inverse Fourier transform of its characteristic function, or applying a Hubbard-Stratonovich transformation (Galla, 2024). Once we take the average over the Gaussian perturbation, we obtain the discrete time analogue of the path integral formalism of Martin-Siggia-Rose-De Dominicis-Janssen (MSRDJ),

$$p(Q^{1:K}|\Theta, Q^0) = \int D\tilde{\mathbf{Q}} \exp\left[i\tilde{Q}_{sa}^{k+1}(Q_{sa}^{k+1} - \rho_{sa} + \beta \sum_{s'} P_{s'|sa} \max_{a'} Q_{s'a'}^k) + \frac{\sigma_W^2}{2}(i\tilde{Q}_{sa}^{k+1})^2\right] \tag{9}$$

where we write $\int D\tilde{\mathbf{Q}} = \prod_{t=0}^{T-1} \prod_{sa} \int_\infty^\infty \frac{d\tilde{Q}_{sa}^{k+1}}{2\pi}$. The term inside the exponential is referred to as the *action*,

$$S_\Theta[\tilde{\mathbf{Q}}, \mathbf{Q}] := \sum_{sa,k} i\tilde{Q}_{sa}^{k+1}(Q_{sa}^{k+1} - \rho_{sa} + \beta \sum_{s'} P_{s'|sa} \max_{a'} Q_{s'a'}^k) + \frac{\sigma_W^2}{2}(i\tilde{Q}_{sa}^{k+1})^2 \tag{10}$$

We are now able to take the integral over Θ to compute the posterior (6), by exchanging the Fourier integral and the integral over Θ. This exchange is justified by the presence of the fictitious noise, since this ensures integrability. A justification of this step in the limit that this noise vanishes will require careful justification and is not the subject of the present paper.

Given the independence assumptions over Θ, the term involving ρ_{sa} in (9) presents no difficulty, contributing only the moment generating function of ρ_{sa}. In order to take the average over the Dirichlet distributions however, it is crucial to recognise that what we only

require the moment generating function of the so-called *finite Dirichlet mean*, which in this context we write as

$$M_{sa} := \sum_{s'} P_{s'|sa} \sum_{k=0}^{K-1} i\tilde{Q}_{sa}^{k+1} \max_{a'} Q_{s'a'}^{k} \tag{11}$$

So we need to calculate

$$\int \prod_{(sa)} dP_{\cdot|sa}\ p(P_{\cdot|sa}) \exp\big(-\beta \sum_{s,s'} P_{s'|sa} \sum_{k=1}^{K-1} i\tilde{Q}_{sa}^{k+1} \max_{a'} Q_{s'a'}^{k}\big) = \prod_{(s)} \mathbb{E}_{M_s} \exp(-\beta M_{sa}) \tag{12}$$

The asymptotic expansion of the generating function of a finite Dirichlet mean was developed by (von Neumann, 1941) , although this can be derived more directly using the properties of Gamma random variables (the "Beta-Gamma algebra") (Pitman, 2018; Conniffe and Spencer, 2001). The mean and variance of M_{sa} are given by,

$$\bar{M}_{sa} = \sum_{k=0}^{K-1} \sum_{s'} \bar{P}_{s'|sa} i\tilde{Q}^{k+1} \max_{a'} Q_{s'a'}^{k}$$

$$\mathcal{V}(M_{sa}) = \sum_{s',s''} C_{s',s''|s} \big(\sum_{k=0}^{K-1} i\tilde{Q}_{sa}^{k+1} \max_{a'} Q_{s'a'}^{k} \big) \big(\sum_{k=0}^{K-1} i\tilde{Q}_{sa}^{k+1} \max_{a''} Q_{s''a''}^{k} \big)$$

where $C_{s',s''|s}$ is the covariance of the Dirichlet probabilities. We can thus write the posterior (6),

$$p(Q^{1:K}|\Psi, Q^0) = \int D\tilde{\mathbf{Q}} \int d\Theta \exp\big(S_\Theta[\tilde{\mathbf{Q}}, \mathbf{Q}]\big) p(\Theta|\Psi) = \int D\tilde{\mathbf{Q}} \exp\big(S_\Psi[\tilde{\mathbf{Q}}, \mathbf{Q}]\big) \tag{13}$$

where we have introduced the new action

$$S_\Psi[\tilde{\mathbf{Q}}, \mathbf{Q}] = \sum_{k,sa} i\tilde{Q}_{sa}^{k+1} Q_{sa}^{k+1} + \frac{\sigma_W^2}{2} (\tilde{Q}_{sa}^{k+1})^2 + \sum_{sa} \mathcal{K}_{\rho_{sa}}(-\sum_{k} i\tilde{Q}_{sa}^{k+1}) - \beta \bar{M}_{sa} + \mathcal{O}(\beta^2) \tag{14}$$

with Ψ denoting the dependence on the hyperparameters of the MDP model parameters, that is, the Dirichlet and mean-reward distribution parameters. We can see in S_Ψ the cumulant generating function $\mathcal{K}_{\rho_{sa}}$ of the mean reward, and the moments of the finite Dirichlet mean up to second order, with higher order moments denoted as $\mathcal{O}(\beta^2)$.

A comment on pedagogy may be helpful. From the perspective of the statistical physics of disordered systems, the integration over parameters Θ is identical to the quenched average in the dynamical or path integral approach to disordered systems. This approach was introduced by De Dominicis who studied the one dimensional Landau-Ginzburg model under quenched disorder (De Dominicis, 1978). A general historical overview of the MSRDJ formalism is provided in (Krommes, 2002), which provides a broad context to the original papers of the formalism (Martin et al., 1973; Janssen, 1976; De Dominicis, 1976; De Dominicis and Peliti, 1978). The term quenched refers to variables which are drawn from a distribution but fixed for the duration of some episode of a dynamical process. Here, the

dynamics are of the fixed point iteration on a discrete time interval $[0, T]$, or the dynamics of the Bellman equation in a finite horizon MDP. In contrast to the disordered systems literature, in the case of both uncertain mean-rewards and transition probabilities, we are interested precisely in the disorder averaged calculations, rather than studying this as a proxy due to an overall concentration of measure phenomenon.

3. Perturbation theory

The expression above involving the exponential of the action S_Ψ is the natural starting point for perturbation theory, which provides a collection of techniques for calculating the posterior moments over the state-action values. There are two standard approaches associated to perturbation theory (Chow and Buice, 2015), a weak-coupling expansion or a weak-noise expansion, with the latter going under various names such as the semi-classical or loopwise expansions.

In order to perform a weak-coupling expansion, one splits the action into so-called interacting and non-interacting parts,

$$S_\Psi = S^0_\Psi + S^{\text{int}}_\Psi \tag{15}$$

The non-interacting part is made up of terms in the action that are quadratic in Q and $\tilde{Q}$. These are also referred to as the solvable parts, in reference to the Gaussian integral. In the case of the Bellman equation these are

$$S^0_\Psi[\tilde{\mathbf{Q}}, \mathbf{Q}] = \sum_{k,sa} i\tilde{Q}^{k+1}_{sa} Q^{k+1}_{sa} + \frac{\sigma^2_W}{2}(i\tilde{Q}^{k+1}_{sa})^2 \tag{16}$$

The interacting part consists of the remaining terms,

$$S^{\text{int}}_\Psi[\tilde{\mathbf{Q}}, \mathbf{Q}] = \sum_{sa} \mathcal{K}_{\rho_{sa}}(-\sum_k i\tilde{Q}^{k+1}_{sa}) - \beta \bar{M}_{sa} + \mathcal{O}(\beta^2) \tag{17}$$

One proceeds by expanding the exponential in the coupling strength, controlled by a coupling parameter which may be identified or introduced and then later set to one. For example, let us introduce a parameter g,

$$p(Q^{1:K}|\Psi, Q^0) = \int D\tilde{\mathbf{Q}} \exp\left(S^0_\Psi + gS^{\text{int}}_\Psi\right) \tag{18}$$

$$= \int D\tilde{\mathbf{Q}} \exp\left(S^0_\Psi\right)(1 + gS^{\text{int}}_\Psi + \frac{1}{2}g^2(S^{\text{int}}_\Psi)^2 + \dots) \tag{19}$$

moments of interest can then be approximated by calculating the solvable Gaussian integrals for a given order of the expansion. In order to help with bookkeeping, the use of Feynman diagrams is commonplace.

An example may be instructive. A simple case is that of policy evaluation, with unknown Dirichlet transitions and unknown mean-rewards with a Gaussian posterior. Writing with a simplified notation,

$$V^{k+1}_s = \rho^\pi_s + \beta \sum_{s'} P^\pi_{s'|s} V^k_{s'} + W^k_s \tag{20}$$

where $\rho^\pi_s = \sum_a \pi(a|s)\rho_{sa}$ and $P^\pi_{s'|s} = \sum_a \pi(a|s)P^\pi_{s'|sa}$. In this case the non-interacting action is

$$S^0_\Psi[\tilde{\mathbf{V}}, \mathbf{V}] = \sum_s \Big(\sum_k i\tilde{V}^{k+1}_s V^{k+1}_s - \sum_k \mu_{\rho^\pi_s} i\tilde{V}^{k+1}_s - \beta \sum_{s'} \bar{P}^\pi_{s'|s} \sum_k i\tilde{V}^{k+1}_s V^k_{s'} + \frac{\sigma^2_{\rho^\pi_s}}{2}(\sum_k i\tilde{V}^{k+1}_s)^2 + \sum_k \frac{\sigma^2_W}{2}(i\tilde{V}^{k+1}_s)^2 \Big) \tag{21}$$

The non-quadratic part is linear and due to the mean of the mean-reward and higher terms from the Dirichlet,

$$S^{\text{int}}_\Psi[\tilde{\mathbf{V}}, \mathbf{V}] = -\sum_s \sum_k \mu_{\rho_s} i\tilde{V}^{k+1}_s + \beta^2 \mathcal{V}(M_s) + \mathcal{O}(\beta^3) \tag{22}$$

The linear term due to the mean is something of a special case as compared to other non-quadratic terms (Hertz et al., 2016). One option for this term is to retain it in the interacting part of the action and for it to be expanded as described. Alternatively one can move it into the non-interacting part but expand the exponential about a saddlepoint, since the integrals involved are over the complex plane.

3.1 Weak-noise approximation in the large data limit

The approach of defining a saddlepoint is in fact the starting point of an alternative perturbative approach known as the weak-noise, semi-classical or loopwise expansion (Kleinert, 2006; Zinn-Justin, 2021). Generally, this is derived under the limit of small variance in the additive or driving noise of a stochastic differential equation (Chow and Buice, 2015), or quantum fluctuations (Kleinert, 2006; Zinn-Justin, 2021). This approach corresponds in the large deviations literature to the Freidlin-Wentzell theory (Touchette, 2009). In taking the limit and ignoring fluctuations, we recover what physicists term the classical approximation, which is what we pursue below.

Recall the additive noise is an artefact of the analysis, introduced in order to have a well-defined distribution over value functions conditional on the model parameters. In order to obtain the classical approximation we combine the large data limit with the limit of weak additive noise.

Let us consider the case of Gaussian mean rewards and known transitions for simplicity. The action is

$$S_\Psi[\tilde{\mathbf{Q}}, \mathbf{Q}] = \sum_{k,sa} i\tilde{Q}^k_{sa} Q^{k+1}_{sa} - \mu_{\rho_{sa}} i\tilde{Q}^{k+1}_{sa} - \beta \sum_{s'} P_{s'|sa} i\tilde{Q}^{k+1} \max_{a'} Q^k_{s'a'} + \frac{1}{2}\sigma^2_W (i\tilde{Q}^{k+1}_{sa})^2 + \sum_{sa} \frac{1}{2}\sigma^2_{\rho_{sa}} (\sum_k i\tilde{Q}^{k+1}_{sa})^2 \tag{23}$$

First, the moment generating function of the distribution of interest is considered.

$$Z[\mathbf{j}, \tilde{\mathbf{j}}] = \int p(Q^{1:K}|\Psi, Q^0) = \int D\tilde{\mathbf{Q}} D\mathbf{Q} \exp\left(S_\Psi + \mathbf{j}^T Q + \tilde{\mathbf{j}}^T \tilde{Q}\right) \tag{24}$$

where the vectors $\mathbf{j} = \{j_{k,sa}\}$ and $\tilde{\mathbf{j}} = \{\tilde{j}_{k,sa}\}$ are referred to as source terms, and we define $\mathbf{j}^T Q = \sum_{k,sa} j_{k,sa} Q^k_{sa}$.

Typically the weak-noise limit is then pursued by first considering complex rather than purely imaginary variables, $i\tilde{Q}^k_{sa} \to \tilde{Q}^k_{sa}$. Second, one rescales $\tilde{Q}^k_{sa} \to \frac{\tilde{Q}^k_{sa}}{\sigma_W}$, which allows us to apply the saddlepoint approximation. Here, applying this limit alone will not work since there is another term that is quadratic in the response field $\tilde{Q}^k_{sa}$, corresponding to the variance in the uncertain mean reward. Therefore, we consider the joint limit of both sources of noise going to zero.

For simplicity, let us assume that the variance for all mean-rewards is the same, given by σ^2_ρ. Consider the change of variables $\tilde{Q}^k_{sa} \to \frac{\tilde{Q}^k_{sa}}{\sigma^2_\rho}$. Then we have

$$p(Q^{1:K}|\Psi, Q^0) = \int D\tilde{\mathbf{Q}} \exp\left[\frac{1}{\sigma^2_\rho}\Big(\sum_{k,sa} i\tilde{Q}^{k+1}_{sa} Q^{k+1}_{sa} - \mu_{\rho_{sa}}\tilde{Q}^{k+1}_{sa} - \beta\sum_{s'} P_{s'|sa}\tilde{Q}^{k+1} \max_{a'} Q^k_{s'a'} + \frac{1}{2}\frac{\sigma^2_W}{\sigma^2_\rho}(\tilde{Q}^{k+1}_{sa})^2 + \sum_{sa}\frac{1}{2}\Big(\sum_k \tilde{Q}^{k+1}_{sa}\Big)^2\Big)\right]$$

Allowing $\sigma^2_W \to 0$ at a faster rate than σ^2_ρ, the approximation proceeds via the saddle-point method and thus we seek the stationary point. The conditions are

$$\frac{\delta S_\Psi}{\delta \tilde{Q}^\ell_{ru}} = 0 \tag{25}$$

$$\frac{\delta S_\Psi}{\delta Q^\ell_{ru}} = 0 \tag{26}$$

From the first we have:

$$\frac{\delta S_\Psi}{\delta i\tilde{Q}^\ell_{ru}} = Q^\ell_{ru} - \mu_{\rho_{ru}} + \sigma^2_{\rho_{ru}}\sum_k i\tilde{Q}^k_{ru} - \beta\sum_{s'} P_{s'|ru}\max_{a'} Q^{\ell-1}_{s'a'} \tag{27}$$

and the second:

$$\frac{\delta S_\Psi}{\delta Q^\ell_{ru}} = \tilde{Q}^{\ell+1}_{ru} - \beta\sum_{sa} P_{r|sa}\tilde{Q}^{\ell+1}_{sa}\frac{\delta}{\delta Q^\ell_{ru}}\max_{a'} Q^\ell_{ra'} \tag{28}$$

By inspection we see that $\tilde{Q}^\ell_{ru} = 0$ is a valid solution for the second equation, for all ℓ, r, u. Under this solution the first equation is

$$Q^{\ell+1}_{ru} - \mu_{\rho_{ru}} - \beta\sum_{s'} P_{s'|ru}\max_{a'} Q^\ell_{s'a'} = 0 \tag{29}$$

This is recognisable as standard Q-value iteration, with the mean reward replaced by its posterior mean. The same result can be obtained for uncertain Dirichlet transition probabilities. Approximations of this form, when the systems are unknown, are sometimes referred to as the "plug-in" estimate of the Q-values (corresponding somewhat to the value functions of the maximum likelihood estimate of the MDP, provided one ignores the prior). Of course, such estimates are correct in the large data limit.

To leading order, the saddlepoint approximation ignores all fluctuations. The semi-classical approximation proceeds by considering the second-order or Gaussian fluctuations around the saddlepoint to calculate corrections to the leading term. In practical terms, these are corrections to the plug-in estimate of the Q-values' posterior mean.

Approximations of this form are similar in spirit to the Laplace approximation for a general posterior over parameters. The differences are that we are dealing with complex contour integrals and thus apply the method of steepest descents, and that we are interested in the distribution of a complex, recursive function of the model parameters, rather than constructing a Gaussian approximation of the posterior over the parameters themselves.

Another alternative, as discussed earlier, is to perform a weak-coupling expansion about the saddlepoint. In this approach it is interesting to recognise that in the case the mean-reward posterior is Gaussian[1], we can identify the weak coupling parameter as the discount factor β. This makes sense, as in the limit that the discount factor tends to zero, the problem is indeed non-interacting: one does not care about the future and plans only locally in time (ie. one faces a multi-armed bandit problem). This point will be expanded upon in the discussion.

Unfortunately, due to the nature of the maximum non-linearity, the above perturbative approaches do not enjoy a clean application beyond the leading order. This is in contrast to most applications of the theory in physics, or for the study of neural networks. The problem faced when trying to account for fluctuations beyond the saddlepoint is similar to the case for the dynamics of Ising spin systems studied in (Hertz et al., 2016). In this case, one approach is to develop so-called *effective* field theories that account for fluctuations in a "self-consistent" way (Stapmanns et al., 2020). The simplest example of self-consistent effective field theory in the present context is a dynamic mean-field theory, which we develop in the next section. This theory produces a self-consistent approximation for the mean, which is distinct from the classical approximation. As a final remark, the perturbative approach will still work in the case of policy evaluation, and we leave such investigations to future work.

4. Dynamic mean field theory for uncertain Markov decision processes

In this section we derive a dynamic mean field theory for the Bellman equation, following recent reviews (Helias and Dahmen, 2019; Crisanti and Sompolinsky, 2018). This produces a self-consistent approximation which accounts for fluctuations that are ignored in the leading-order saddlepoint approximation and difficult to correct via higher order expansions due to the maximum non-linearity.

The derivation is based chiefly on the assumption that the Dirichlet distribution over transition probabilities is what we call "isotropic", with Dirichlet parameters $\alpha_{s'|sa}$ at least one, such that the distribution is either peaked in the centre of the simplex or uniform over the simplex. The second restriction, which is for technical convenience and we will argue can weakened significantly, is that the mean-rewards have identical distributions.

1. A Gaussian mean-reward posterior arises when the reward has a Gaussian distribution with unknown mean and known variance, assuming a Gaussian prior on the unknown mean.

Assumption 1.A (*Isotropic Dirichlet probabilities*) The transition probabilities are drawn from a Dirichlet distribution, for each state-action pair, $P_{sa} \sim \text{Dirichlet}(\alpha_{sa})$ with parameters being identical with $\alpha_{s'|sa} = \alpha$ for all s' and (sa) pairs, with $\alpha \geq 1$.

Assumption 2.A (*Identical arbitrary mean-reward distributions.*) The mean-rewards ρ_{sa} have identical arbitrary distribution with a moment generating function that exists. We denote their log moment generating function as $\mathcal{K}_{\rho_{sa}} = \mathcal{K}_{\rho}$.

The identical mean-reward distribution can be weakened in various ways, in a manner similar to the way in which the central limit theorem can be extended to non-identical summands, as in the Lyapunov or Lindeberg theorems.

The flat Dirichlet is an example of the isotropic definition above, with $\alpha_{s'|sa} = 1$ for all s' and (sa) pairs. This is a standard prior for Bayesian reinforcement learning, and is thus not representative of MDPs under general uncertainty. In the flat prior case the Dirichlet mean is $\bar{P}_{s'|sa} = \frac{1}{N}$ and the covariance $C_{s',s''|sa}$ has positive diagonal elements of order $\mathcal{O}(N^{-2})$ and negative off-diagonal of order $\mathcal{O}(N^{-3})$. Note that the flat Dirichlet is thus the weakest assumption of all the isotropic Dirichlet, as defined above.

Under this flat or general isotropic Dirichlet, the higher order interaction terms do not contribute in the large N limit, and one would expect a dynamic mean field theory (DMFT) to hold asymptotically, as as for spin glasses and neural networks (Crisanti and Sompolinsky, 2018; Helias and Dahmen, 2019). In the physics and probability literature, the result that is obtained using DMFT is the so called "propagation of chaos" property.

Theorem (*Propagation of chaos*) Consider a sequence of bounded, continuous functions $f_1, \dots f_n$. Consider a distribution P over variables x_i^k for $i \in \{1, \dots, N\}$, where N denotes the number of variables and k a time index. We say that P satisfies the propagation of chaos property if at a given k and for any $j = 1, \dots, m$ and integers $\ell_1, \dots, \ell_m$,

$$\lim_{N\to\infty} \mathbb{E}_P \prod_{j=1}^{m} f_j(x_{\ell_j}^k) = \prod_{j=1}^{m} \int f_j(x) d\mu(x) \tag{30}$$

for a measure μ. This measure μ is typically called the mean-field solution, and corresponds to the distribution of the system evolving under a mean-field equation.

A claim of this paper is that the time-marginals of the distribution $p(Q^{1:K}|\Psi, Q^0)$, for finite iterations or time steps K of the optimality equation, satisfy a propagation of chaos property. This propagation of chaos, it should be noted, is for the averaged system. This is distinct from a quenched propagation of chaos result. We also claim that the distribution over value function iterates of policy evaluation also have the propagation of chaos property.

On the face of it, this property is surprising, since the recurrent structure of the Bellman equation clearly introduces dependencies between value functions. Of course there is a dependency on the level of the mean field statistics, but given these statistics, the Q-values are statistically independent. We do not prove this independence result rigorously in this paper. Instead, we apply the method of steepest descents by the introduction of two auxiliary fields that "decouple" the interacting variables in the action S_Ψ. The resulting saddlepoint equations give rise to the mean field equations.

4.1 Auxiliary fields and saddlepoint approximation

In this section we write for shorthand the maximum over actions as $\phi^k_{s'} = \max_{a'} Q^k_{s'a'}$. This is of course the standard value function $V^k_{s'}$. Here we consider the flat Dirichlet case of the isotropic Dirichlet assumption. Under this assumption the interacting part of the action S^{int}_Ψ becomes

$$S^{\text{int}}_\Psi[\tilde{\mathbf{Q}}, \mathbf{Q}] = \sum_{sa} \mathcal{K}_{\rho_{sa}}(-\sum_k i\tilde{Q}^{k+1}_{sa}) - \beta \frac{1}{N} \sum_{s'} \sum_{k=0}^{K-1} i\tilde{Q}^{k+1}_{sa} \phi(Q^k_{s':})) + \mathcal{O}(\frac{1}{N^2}) \tag{31}$$

Define the following auxilliary field,

$$\theta^k_1 := \frac{\beta}{NA} \sum_{s'} \phi(Q^k_{s':}) \tag{32}$$

which we can substitute within the action S_Ψ above using the Dirac delta represented in its Fourier form with complex auxiliary field θ^k_2,

$$\delta(-NA\theta^k_1 + \beta \sum_{s'} \phi(Q^k_{s':})) = \int_{-\infty}^{\infty} \frac{1}{2\pi} d\theta^k_2 \exp[\theta^k_2(-NA\theta^k_1 + \beta \sum_{s'} \phi(Q^k_{s':}))] \tag{33}$$

Considering the integral over the exponential of the action, we then have

$$\begin{aligned}
&\exp(S^{\text{int}}_\Psi[\tilde{\mathbf{Q}}, \mathbf{Q}]) = \exp\big[\sum_{k,sa} i\tilde{Q}^{k+1}_{sa} Q^k_{sa} + \mathcal{K}_{\rho_{sa}}(-\sum_k i\tilde{Q}^{k+1}_{sa}) - \beta \frac{1}{N} \sum_{s'} i\tilde{Q}^{k+1}_{sa} \phi(Q^k_{s':}))\big] \\
&= \int D\theta_1 \exp\big[\sum_{k,sa} i\tilde{Q}^{k+1}_{sa} Q^k_{sa} + \mathcal{K}_{\rho_{sa}}(-\sum_k i\tilde{Q}^{k+1}_{sa}) - Ai\tilde{Q}^{k+1}_{sa}\theta^k_1\big] \delta(-NA\theta^k_1 + \beta \sum_{s'} \phi(Q^k_{s':})) \\
&= \int D\theta_1 D\theta_2 \exp\big[\sum_{k,sa} i\tilde{Q}^{k+1}_{sa} Q^{k+1}_{sa} + \mathcal{K}_{\rho_{sa}}(-\sum_k i\tilde{Q}^{k+1}_{sa}) \\
&\qquad - Ai\tilde{Q}^{k+1}_{sa}\theta^k_1 - NA\theta^k_2\theta^k_1 + \beta\theta^k_2 \sum_{s'} \phi(Q^k_{s':}))\big]
\end{aligned}$$

with $\int D\theta_1 D\theta_2 = \prod_{k=1}^{K-1} \int_{-\infty}^{\infty} \frac{1}{2\pi} d\theta^k_1 d\theta^k_2$. If we now bring back in the integral over $\tilde{\mathbf{Q}}$ and the other remaining terms, we can write

$$p(Q^{1:K}|\Psi, Q^0) = \int D\tilde{\mathbf{Q}} \exp\big(S_\Psi[\tilde{\mathbf{Q}}, \mathbf{Q}]\big) \tag{34}$$

$$= \int D\theta_1 D\theta_2 \exp\big[-NA \sum_k \theta^k_2\theta^k_1 + NA \sum_k \ln Z[\theta^k_1, \theta^k_2]\big] \tag{35}$$

$$= \int D\theta_1 D\theta_2 \exp\big[-NA\Gamma(\theta_1, \theta_2)]\big] \tag{36}$$

The term $\Gamma(\theta_1, \theta_2) = \sum_k \theta^k_2\theta^k_1 - \sum_k \ln Z[\theta^k_1, \theta^k_2]$ is known as the *effective* action and is a starting point of an effective field theory which includes the mean field theory developed here

(Stapmanns et al., 2020). This effective action involves a newly defined moment generating function $Z[\theta_1^k, \theta_2^k]$, which depends on the auxiliary fields,

$$
\begin{aligned}
&(Z[\theta_1^k, \theta_2^k])^{NA} \\
&= \int D\tilde{\mathbf{Q}} D\mathbf{Q} \exp\big[\sum_{sa} i\tilde{Q}_{sa}^{k+1} Q_{sa}^{k+1} + \mathcal{K}_{\rho_{sa}}(-\sum_k i\tilde{Q}_{sa}^{k+1}) - Ai\tilde{Q}_{sa}^{k+1}\theta_1^k + \beta\theta_2^k \sum_{s'} \phi(Q_{s':}^k)\big] \qquad (37) \\
&= \prod_{s,a} \int D\tilde{\mathbf{Q}} D\mathbf{Q} \exp\big[i\tilde{Q}_{sa}^{k+1} Q_{sa}^{k+1} + \mathcal{K}_{\rho_{sa}}(-\sum_k i\tilde{Q}_{sa}^{k+1}) - Ai\tilde{Q}_{sa}^{k+1}\theta_1^k - \beta\theta_2^k \frac{1}{A}\phi(Q_{s':}^k)\big] \qquad (38)
\end{aligned}
$$

Note we have dropped the additive noise variance term, for brevity. The factor of NA that comes out of this generating function is due to the summation over states, which have been decoupled via the auxilliary fields. A difference to the neural network and spin glasses cases is the factor of $A = |\mathcal{A}|$, the action space size. This is typically much smaller than N and thus does not alter the method of steepest descents which we now apply.

The leading order approximation in the method of steepest descents, just as in the weak-noise expansion previously, approximates the integral by the value of the integrand at the saddlepoint, ignoring the contributions from higher order terms (i.e. the Hessian and higher terms). We find the stationary point of the effective action in equation (36),

$$\frac{\delta}{\delta\theta_{1,2}^k}\Gamma[\theta_1, \theta_2] = 0 \qquad (39)$$

We now calculate the derivatives,

$$
\begin{aligned}
\frac{\delta}{\delta\theta_1^k}\Gamma[\theta_1, \theta_2] &= -NA\theta_2^k + NA\frac{\delta \ln Z[\theta_1^k, \theta_2^k]}{\delta\theta_1^k} && (40) \\
&= -NA\theta_2^k + NA\frac{1}{Z[\theta_1^k, \theta_2^k]}\frac{\delta Z[\theta_1^k, \theta_2^k]}{\delta\theta_1^k} && (41) \\
&= -NA\theta_2^k + NA\langle -Ai\tilde{Q}_{sa}^{k+1}\rangle_{\theta_1,\theta_2} && (42)
\end{aligned}
$$

where the average $\langle\cdot\rangle_{\theta_1,\theta_2}$ emerges due to the derivative of the log generating function $\ln Z[\theta_1^k, \theta_2^k]$ of equation (37). Setting this derivative to zero, we have

$$\theta_2^k = -A\langle i\tilde{Q}_{sa}^{k+1}\rangle_{\theta_1,\theta_2} \qquad (43)$$

A similar calculation reveals

$$\theta_1^k = \frac{\beta}{A}\langle\phi(Q_{s':}^k\rangle_{\theta_1,\theta_2} \qquad (44)$$

As argued in the physics literature for the saddlepoint (θ_1^*, θ_2^*), the average of the response field $\tilde{Q}^{k+1}$ must be zero due to the normalisation of the probability distribution and that we do not fix both ends of our value function trajectory (Galla, 2024; Stapmanns et al., 2020).

Determining the saddlepoint $(\theta_1^*, \theta_2^* = 0)$ requires solving the saddlepoint equations. Let us first substitute the point $(\theta_1^*, \theta_2^* = 0)$ into (36), noting the integral has now vanished since we have taken the saddlepoint approximation,

$$\begin{aligned}\langle Z[\mathbf{j},\tilde{\mathbf{j}}]\rangle_{P_{sa}} &= \exp\left[N\Gamma(\theta_1^*,\theta_2^*)\right] \\ &= \prod_{s,a}\int D\tilde{\mathbf{Q}}D\mathbf{Q}\exp\big[\sum_k i\tilde{Q}_{sa}^{k+1}Q_{sa}^{k+1} + \mathcal{K}_{\rho_{sa}}(-\sum_k i\tilde{Q}_{sa}^{k+1}) - Ai\tilde{Q}_{sa}^{k+1}(\theta_1^k)^*\big] \\ &= \prod_{s,a}\int D\tilde{\mathbf{Q}}D\mathbf{Q}\exp\big[\sum_k i\tilde{Q}_{sa}^{k+1}Q_{sa}^{k+1} + \mathcal{K}_{\rho_{sa}}(-\sum_k i\tilde{Q}_{sa}^{k+1}) - \beta\langle\phi(Q_{s'}^k)\rangle_{\theta_1^*,\theta_2^*}\big]\end{aligned}$$

If we now consider equation (8) and work backwards to equation (7), we can see that we have a new *effective* Bellman equation where the term $\beta\sum_{s'}P_{s'|sa}\phi_{s'}^k$ has been replaced by $\beta\langle\phi(Q_{s'}^k)\rangle_{\theta_1^*,\theta_2^*}$. We determine this average over $\phi(Q_{s'}^k)$ recursively from the initial iteration and thus find the self-consistent solution. This effective equation over iterations is the dynamic mean field equation. The example below demonstrates how to move between the effective action and the mean field equation.

4.2 Example

In the case that the posterior over the mean reward ρ_{sa} is Gaussian, we have that its moment generating function is

$$\mathcal{K}_{\rho_{sa}}(-\sum_k i\tilde{Q}_{sa}^{k+1}) = -\mu_{\rho_{sa}}\sum_k i\tilde{Q}_{sa}^{k+1} + \frac{1}{2}\sigma_{\rho_{sa}}^2(\sum_k i\tilde{Q}_{sa}^{k+1})^2 \tag{45}$$

substituting this expression for the expression for the density, we find

$$\begin{aligned}&p(Q^{1:K}|\Psi,Q^0) \\ &= \prod_{sa}\int D\tilde{\mathbf{Q}}\exp\big[\sum_k i\tilde{Q}_{sa}^{k+1}Q_{sa}^{k+1} - \mu_{\rho_{sa}}\sum_k i\tilde{Q}_{sa}^{k+1} + \frac{\sigma_{\rho_{sa}}^2}{2}(\sum_k i\tilde{Q}_{sa}^{k+1})^2 - \beta\langle\phi(Q_{s'}^k)\rangle\big]\end{aligned}$$

which we can identify as corresponding to a Gaussian process over the state-action values. Explicitly, the mean is given by

$$\mathbb{E}\big(Q_{sa}^{k+1}\big) = \mu_{\rho_{sa}} + \beta\mathbb{E}\big(\max_{a'}Q_{s'a'}^k\big).$$

where the expectation of the maximum with respect to $Q_{s'a'}^k$, whose mean is given by the previous iteration, and whose variance is just that of the mean-reward, $\sigma_{\rho_{sa}}^2$. The result thus takes a simple form, and is similar to the neural network and spin glass cases, however there is no correlation across time of the Gaussian process, with only the mean propagating forward across iterations. In general of course, the process will not be Gaussian since the mean-rewards can have arbitrary distribution.

5. Dynamic mean field programming

In this section we describe mean field equations for the Bellman equation, whose iteration we call dynamic mean field programming (DMFP). First we consider consider the flat Dirichlet

case for both the optimality equation and for policy evaluation, where we claim DMFP is exact. Then we turn to general belief assumptions, where DMFP can be viewed as a mean field approximation.

The DMFP equation for Q-value iteration, in the flat Dirichlet case with general mean-rewards, is given by

$$\mu_{sa}^{k+1} = \mu_{\rho_{sa}} + \beta \mathbb{E}_{Q^k} \max_{a'} Q_{s'a'}^k \tag{46}$$

$$= \mu_{\rho_{sa}} + \beta \mathbb{E}_{\rho} \max_{a'} \mu_{s'a'}^k + \rho_{s'a'} \tag{47}$$

where we denote the mean of Q_{sa}^{k+1} as μ_{sa}^{k+1}. We see that the expectation $\mathbb{E}_{Q^k}$ is simply that of the mean rewards $\rho_{s'a'}$ with mean shifted by μ_{sa}^k. In general, calculating expectations of the maximum requires us to turn to approximations, which we introduce in the subsections below.

In the case of policy evaluation, there is no need for approximations to propagate the mean. Given a policy π, the policy evaluation equations for the value function are given by

$$V_s^{\pi,k+1} = \sum_a \pi(a|s) \big(\rho_{sa} + \beta \sum_{s'} P_{s'|sa} \, V_{s'}^{\pi,k} \big) \tag{48}$$

Denoting the posterior mean for the value functions $\mu_s^k = \mathbb{E} V_{s'}^{\pi,k}$, the policy evaluation DMFP equations are given in the flat Dirichlet case by,

$$\mu_s^{\pi,k+1} = \sum_a \pi(a|s) \big(\mu_{\rho_{sa}} + \beta \mu_{s'}^{\pi,k} \big) \tag{49}$$

Now let us consider DMFP as a mean field approximation to general posteriors. The heuristic is obvious, with the mean of the Dirichlet transitions substituted, and higher order correction terms for the higher moments, including the variance, ignored. For the optimality equation the DMFP approximation to the mean is then

$$\mu_{sa}^{k+1} = \mu_{\rho_{sa}} + \beta \sum_{s'} \bar{P}_{s'|sa} \mathbb{E}_{\rho} \max_{a'} \mu_{s'a'}^k + \rho_{s'a'} \tag{50}$$

and in the policy evaluation case it is given by

$$\mu_s^{\pi,k+1} = \sum_a \pi(a|s) \big(\mu_{\rho_{sa}} + \beta \sum_{s'} \bar{P}_{s'|sa} \mu_{s'}^{\pi,k} \big) \tag{51}$$

The accuracy of these equations outside the flat Dirichlet case is dependent on the specific structure of the Dirichlet beliefs, or an alternative distribution over the simplex. Similarly, the accuracy as the other assumptions are changed, such as the independence assumptions in Assumption 4, will depend on the specific nature of the dependent beliefs. Of course, even in the isotropic Dirichlet case, the mean field equations will be an approximation due to the finite state space N.

We present numerical simulations in support of the results of the dynamic mean field theory, in particular that the mean field equations provide the mean of the value functions or state-action value functions and that the variance of the state action value functions

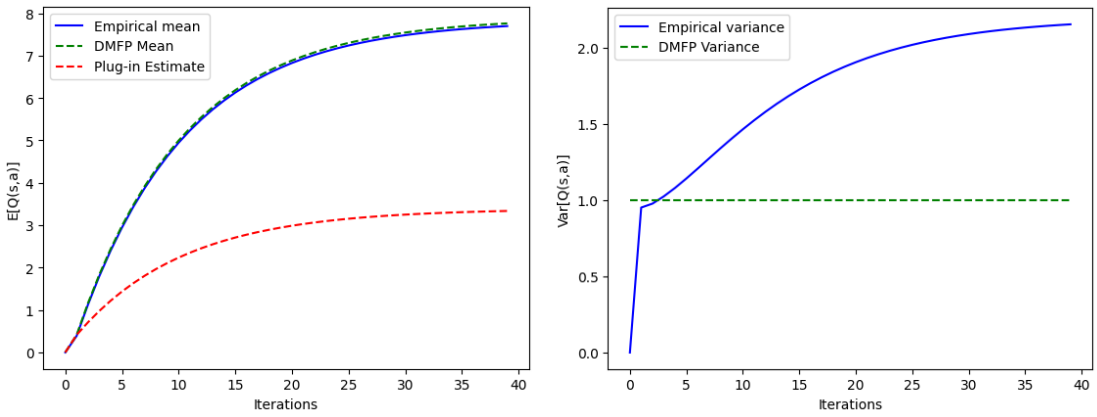

Figure 1: Dynamic mean field programming theory versus simulations of the Bayesian mean and variance of the iterates of a particular Q-value function for an infinite horizon MDP with $N = 50$ states, with $|A| = 2$, discount factor $\beta = 0.9$, with the empirical estimates formed from 500 realisations of the system. The model has a flat Dirichlet prior and Gaussian mean-reward posterior, with uniformly random mean and a variance of 1 for each state-action. The graph on the left shows the empirical mean in agreement with the DMFP mean, while the lower line corresponds to the plug-in estimate (or Jensen lower bound). The graph on the right show that the state-action value's variance is more than double the variance of the mean-reward variance.

reduces to that of the mean-reward in the large state space limit. The simulations in Figure 1. show the accuracy of the DMFP equation for a relatively small MDP, as described. The simulations in Figure 2. show the convergence to the DMFP result depends generally on the effective horizon $\frac{1}{1-\beta}$ in the discounted setting, and correspondingly the horizon T in the finite horizon MDP. More extensive experimental analysis of the DMFP equations, in particular with regard to the computation or approximation of the expected maximum can be found in the thesis (Donnelly, 2023), which investigates the stability of the DMFP equations derived here.

5.1 Approximate statistics of DMFP for the Bellman optimality equation

In order to compute the mean of the optimal state-action values via DMFP, one needs to calculate the mean of the maxima over the actions of Q^k_{sa}. In general, there are no closed form analytic expressions for the maxima, however certain cases are simple enough, such as for the bivariate Gaussian, which was used in the simulations above. The literature on expected maxima and various approximations are reviewed in (Donnelly, 2023). We briefly recapitulate several important cases here.

In the case that the mean-rewards are identical and the action space is very large, one can appeal to extreme value theory, which describes the distribution of the maximum over asymptotically many variables (appropriately re-scaled). This approximation provides

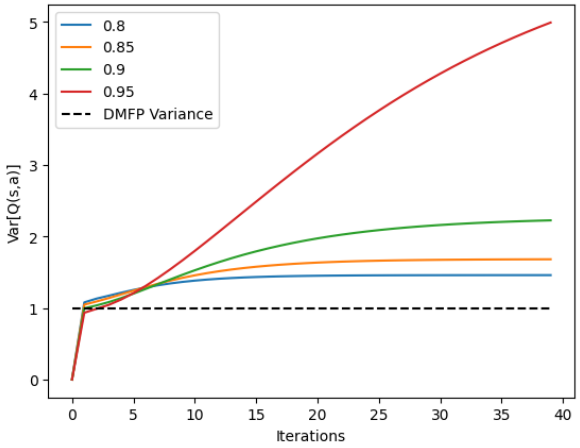

Figure 2: The empirical variance for the same system as presented in Figure 1, but with variable discount factor $\beta = \{0.8, 0.85, 0.9, 0.95\}$. As the effective horizon increases (or the memory of the system increases), correspondingly the posterior variance increases.

closed form equations. For example, assuming the Dirichlet parameters are flat and using a Type-I extreme value or Gumbel distribution as an approximation, the DMFP equation becomes,

$$\mu^{k+1} = \mu_\rho + \beta\big(\mu^k + \sigma_\rho(b_{|\mathcal{A}|} + a_{|\mathcal{A}|}\gamma_{\text{EM}})\big) \tag{52}$$

where γ_{EM} is the Euler-Mascheroni constant, and $a_{|A|}$ and $b_{|A|}$ are constants dependent on the action space size $|A|$, given in (Fisher and Tippett, 1928). The approximation can be improved by using a Type-III or Weibull distribution (Fisher and Tippett, 1928) or other approximations (Cohen, 1982). There exist generalisations to non-independent variables in the case the mean-rewards are Gaussian (Falk et al., 2010; Hüsler, 1994).

For general beliefs on the mean-rewards, one can either appeal to specific cases such as the Gaussian case, or one may appeal to bounds on the expected maximum. For example, the relatively simple bounds of (Aven, 1985) or the tight bounds of (Bertsimas et al., 2006). We also have via Jensen's inequality $\mathbb{E}\max_i X_i \geq \max_i \ m_i$, for random variables X_i with arbitrary dependence and means m_i. This is equivalent to the plug-in estimate, also referred to as the certainty equivalent heuristic.

6. Discussion

This paper has introduced a formalism from statistical field theory for Markov decision processes with Bayesian model uncertainty. The formalism provides techniques for analytically computing posterior moments of value functions, translating uncertainty from the MDP parameters. The weak noise approximation and the dynamic mean field theory developed are examples of such techniques. In both cases, the leading order terms correspond to asymptotic limits, the large data limit and the large state space limit, respectively.

In this final section we discuss the weak noise approximation and dynamic mean field theory results in more detail, in particular the assumptions underpinning the mean field theory and its breakdown. Following this we discuss the next possible steps in the development of an effective field theory. In practical terms this means the computation of higher moments, and how these computations might figure into applications. As we argue, the mean field result presents a source of inspiration for a new class of bandit style algorithms for online learning in MDPs, as well as offline planning.

6.1 Discussion of the weak noise approximation and dynamic mean field theory

In the context of an unknown MDP with Bayesian model uncertainty, we have shown that the weak noise approximation is an expansion around the large data limit. In the classical limit approximation, this consists of replacing the parameters with their posterior means in the Bellman equation, which in the reinforcement learning literature is known as the plug-in estimate. It is important to note that in practice a MDP will not have uniform uncertainty across its state-action space. For example, a reinforcement learning agent may elect not to explore certain parts of the state space, for which it determines that rewards are lower. This means that such expansions may not be uniformly accurate across the state-action space.

Another important point made earlier, is that the semi-classical expansion will more more readily be applied to policy evaluation, due to the way in which the generating function is expanded and the nature of the maximum non-linearity. The rationale is analogous to the discussion regarding the kinetic Ising model in section 10 of (Hertz et al., 2016).

The dynamic mean field theory provides a set of mean field equations, which we call dynamic mean field programming (DMFP). These equations agree with the plug-in estimate for policy evaluation, and under the asymptotic data limit the DMFP equations reduce to the same classical result. However, DMFP produces an approximation outside of the conditions under which it is derived, which accounts for fluctuations in the calculation of the mean. Higher order corrections to the weak-noise limit do not agree with this, suggesting a better path forwards may be effective field theory, as discussed in the next section.

A headline result of the dynamic mean field theory is that the distribution over Q-values factorises across state-action pairs in the asymptotic statespace, under the assumptions of Section 2. This is a surprising result, since the recurrent structure clearly introduces dependencies between Q-values. Clearly there is a dependency on the level of the mean field statistics, but given these statistics the Q-values are statistically independent.

The two key assumptions underpinning the independence result are the assumptions of statistical independence of the parameters across state-action pairs, and the isotropic Dirichlet assumption, the weakest form of which is the flat Dirichlet. The independence assumption allows us to both analytically take the average over model parameters more easily, and builds independence into the model. The isotropic Dirichlet assumption can be thought of as ensuring that in the sum over next transitions no single value function contributes too much to the sum. This is akin to the restrictions assumed to ensure the central limit theorem for non-identical summands.

It is common in the literature to assume the independence assumption for tabular MDPs and to take the flat Dirichlet as a prior. This means that these algorithms start by assuming all Q-values are independent. This would appear to be a strong assumption with which to start learning. The independence between parameters at different state-actions should be contrasted to the linear quadratic regulator, the cornerstone of continuous state control theory. In the linear quadratic case, the state dynamics and cost functions are typically not state-dependent functions, reflecting a causal model of the real world.

How might the mean field theory break down? Several of the key ways this can occur is if the belief structure has (i) finite state space size N, (ii) sparsity in transition functions, (iii) statistical dependence between state-action parameters (mean-rewards and transition dynamics), or (iv) unbounded horizon (eg. scaling with N). We reflect on these in order.

In practice we will have finite size effects of some form. If the independence and isotropic Dirichlet assumptions are maintained, we can study finite size effects by expanding to the next order in the saddelpoint expansion of the auxilliary field (Segadlo et al., 2022). A theoretical description of this breakdown may give further insight into the phenomenology of reinforcement learning problems and the dependence between value functions, at least at the start of learning.

Sparsity in the MDP transition matrices is common in practice, and this would lead to a breakdown in the mean field theory. With regards to the Dirichlet belief structure, if the MDP one is estimating has a sparse transition structure, as more samples are obtained two effects will be seen. First, fewer terms will contribute to the posterior mean of the Dirichlet, meaning that few random variables may dominate the sum, leading to a breakdown in the mean field or central limit type effects. Second, higher order terms from the Dirichlet cumulants will not vanish, which again will see the mean field effects diminish, as more non-linear interactions between value functions emerge.

The introduction of general statistical dependencies between parameters may mean that the average in equation (12) can not be analytically computed. One would also expect more dependence between the state-action value functions, which can be measured by approximated with from higher order expansions.

The length of the horizon or effective horizon in the infinite discounted MDP, brings on different behaviour, as seen for example in the simulations presented in Figure 2. In the neural network literature the scaling of the depth with the width of the networks in constant proportions has been associated with heavy tailed phenomena (Hanin and Nica, 2020). This joint limit corresponds to the horizon scaling with the number of states and warrants investigation.

6.2 Applications and extensions of the theory

The work developed in this paper is suggestive of new approaches to planning or learning in finite state and action MDPs, using field theoretic techniques and inspired by the mean field result. In this final section, we outline one possible approach and the theoretical extensions required to realise it.

A common heuristic in the design of sample efficient reinforcement learning for MDPs is to add an "exploration bonus" to each Q-value, analogous to upper-confidence-bound algorithms for the multi-armed bandit (Osband and Van Roy, 2017; Ghavamzadeh et al., 2015).

This heuristic is applied in both the frequentist and Bayesian settings. The exploration bonus is intended to capture the value of information derived from exploring new parts of a state-action space, and balance this against expected cumulative rewards (Dearden et al., 2013). Since the Q-values are not independent, due to the underlying state evolving as a general Markov process, the standard multi-armed bandit approaches do not apply. Existing methods try to account for this dependence using a range of bounds and concentration inequalities (O'Donoghue et al., 2018; O'Donoghue, 2018; Neu and Pike-Burke, 2020), which may be loose in practice.

Rather than attempting to compute bounds on the posterior variance, one can instead directly estimate the posterior variance analytically using field theory, from which one can derive an exploration bonus. The mean field result suggests a useful way of thinking about this suggested class of bandit approximations for MDPs.

It is helpful to consider that mean field theories for Bayesian statistical inference are generally be derived in several ways. Chief among these are (i) variational techniques with an approximating distribution, (ii) large N expansions and a saddlepoint approximation, and (iii) high temperature expansions.

A variational mean field approximation to statistical inference is obtained by optimising for the "closest" approximating distribution to the true Bayesian posterior, from a family of distributions which often factorise in certain ways (Opper and Saad, 2001), (Wainwright et al., 2008), usually chosen for computational reasons. Although the dynamic mean field theory derived here is the result of a large N expansion coupled with a saddlepoint approximation, the result is a fully factorised distribution over value functions.

Casting the DMFP result as an approximation and using it as the basis for a reinforcement learning strategy is equivalent to assuming an agent faces a set of N independent multi-armed bandit problems in each state. In that case an agent can act and plan locally, in space and in time. Of course this will be suboptimal outside of the DMFP assumptions, however it provides a template and a way of thinking for new bandit algorithms. The DMFP result, of a fully factorised distribution, is then just one extreme amongst a range of possible approximating distributions.

The types of improvements one could seek are ones that estimate the posterior variance and higher order moments, under general belief structures on the transition probabilities and the mean-rewards. The higher moments and higher order expansions would then account for the general non-linear dependence between value functions. In turn, even with a "mean-field" bandit approximation as a basis for the expansions, this would still in principle correspond to an agent planning further ahead since the corrections to higher moments account for the uncertainty in the system and thus the value of information. Note that corrections to fully factorised mean field approximations are commonplace in the theory of disordered systems, for example the Onsager correction terms for the Sherrington-Kirkpatrick model (Mézard et al., 1987; Opper and Saad, 2001).

Given the DMFP result is not derived by a variational approach, and since it is unclear how to formulate the problem studied here in a variational framework in the first place, more rich approximating families of distributions over value functions (Wainwright et al., 2008) cannot be explicitly proposed. Therefore determining the appropriate theoretical framework to calculate higher order corrections and performing these calculations for general conditions on the posterior beliefs will be a challenge.

It is possible that the theory of high-temperature expansions can be adapted to the MDP setting (Yedidia, 2001). One would expect the DMFP result to correspond to the high temperature limit being taken. Otherwise, a range of techniques are available from general field theory (Zinn-Justin, 2021) as well as disordered systems (Hertz et al., 2016), although in the latter case research is generally directed toward quenched systems. In this case the study of the quenched average is an analytic proxy, justified under the "self-averaging" assumption, which corresponds to a large deviations or concentration of measure principle (Mézard et al., 1987; Mezard and Montanari, 2009).

Separate to the above discussion, another important next step is the development of the semi-classical approximation for the case of Bayesian policy evaluation. Although not leading to learning strategies based on multi-armed banddits and the exploration bonuses, policy evaluation is an important problem in many applications. For example in the offline planning setting, one may have competing policies which one might like to compare and choose between, given certain criteria. The semi-classical approach has the advantage of being relatively straightforward due to the linearity of the policy evaluation equation.

As a final comment, we anticipate the theory presented has implications for frequentist as well as Bayesian approaches. In Bayesian statistics, one considers a random variable of interest, for which a prior is assigned and posterior distributions calculated given data. From this posterior one can produce various estimates. In frequentist statistics one produces estimates from estimators, which are random variables. A good example of the difference is the Kalman filter, which can be considered a Bayesian estimate for the state of a linear quadratic Gauss-Markov system, or a minimum variance estimator from the frequentist perspective. Visiting the problem of designing better estimators for the Q-value functions using the techniques developed here could be an interesting line of work.

Acknowledgements

The author would like to acknowledge the support of Hung Nguyen and Langford White, as well as the helpful discussions with colleagues Blake Donnelly, Ian Fuss, Kelli Francis-Staite, Lachlan MacDonald, Jack Valmadre, Federica Gerace and Carlo Lucibello.

References

Terje Aven. Upper (lower) bounds on the mean of the maximum (minimum) of a number of random variables. *Journal of applied probability*, 22(3):723–728, 1985.

Dimitris Bertsimas, Karthik Natarajan, and Chung-Piaw Teo. Tight bounds on expected order statistics. *Probability in the Engineering and Informational Sciences*, 20(4):667–686, 2006.

Carson C Chow and Michael A Buice. Path integral methods for stochastic differential equations. *The Journal of Mathematical Neuroscience (JMN)*, 5:1–35, 2015.

Jonathan P. Cohen. The penultimate form of approximation to normal extremes. *Advances in Applied Probability*, 14(2):324–339, 1982. ISSN 00018678. URL `http://www.jstor.org/stable/1426524`.

Denis Conniffe and John E. Spencer. When moments of ratios are ratios of moments. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 50(2):161–168, 2001. ISSN 00390526, 14679884. URL http://www.jstor.org/stable/2681091.

A Crisanti and H Sompolinsky. Path integral approach to random neural networks. *Physical Review E*, 98(6):062120, 2018.

C De Dominicis. Technics of field renormalization and dynamics of critical phenomena. In *J. Phys.(Paris), Colloq*, pages C1–247, 1976.

C De Dominicis. Dynamics as a substitute for replicas in systems with quenched random impurities. *Physical Review B*, 18(9):4913, 1978.

C De Dominicis and L Peliti. Field-theory renormalization and critical dynamics above t c: Helium, antiferromagnets, and liquid-gas systems. *Physical Review B*, 18(1):353, 1978.

Richard Dearden, Nir Friedman, and David Andre. Model-based bayesian exploration. *arXiv preprint arXiv:1301.6690*, 2013.

Blake Donnelly. *Analysis of New Methods for Inference in Markov Decision Processes*. PhD thesis, University of Adelaide, 2023.

Michael O'Gordon Duff. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. University of Massachusetts Amherst, 2002.

Michael Falk, Jürg Hüsler, and Rolf-Dieter Reiss. *Laws of small numbers: extremes and rare events*. Springer Science & Business Media, 2010.

R. A. Fisher and L. H. C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2):180–190, 1928. doi: 10.1017/S0305004100015681.

Tobias Galla. Generating-functional analysis of random lotka-volterra systems: A step-by-step guide. *arXiv preprint arXiv:2405.14289*, 2024.

Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, Aviv Tamar, et al. Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5-6): 359–483, 2015.

Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, 2020.

Moritz Helias and David Dahmen. Statistical field theory for neural networks, 2019.

John A Hertz, Yasser Roudi, and Peter Sollich. Path integral methods for the dynamics of stochastic and disordered systems. *Journal of Physics A: Mathematical and Theoretical*, 50(3):033001, 2016.

Jürg Hüsler. Extremes: Limit results for univariate and multivariate nonstationary sequences. In *Extreme Value Theory and Applications*, pages 283–304. Springer, 1994.

Hans-Karl Janssen. On a lagrangean for classical field dynamics and renormalization group calculations of dynamical critical properties. *Zeitschrift für Physik B Condensed Matter*, 23(4):377–380, 1976.

Hagen Kleinert. *Path integrals in quantum mechanics, statistics, polymer physics, and financial markets*. World Scientific Publishing Company, 2006.

John A Krommes. Fundamental statistical descriptions of plasma turbulence in magnetic fields. *Physics Reports*, 360(1-4):1–352, 2002.

Carlos E. Luis, Alessandro G. Bottero, Julia Vinogradska, Felix Berkenkamp, and Jan Peters. Value-distributional model-based reinforcement learning. *Journal of Machine Learning Research*, 25(298):1–42, 2024. URL `http://jmlr.org/papers/v25/23-0913.html`.

Paul Cecil Martin, Eric D Siggia, and Harald A Rose. Statistical dynamics of classical systems. *Physical Review A*, 8(1):423, 1973.

Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1392–1403. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/0f0e13216262f4a201bec128044dd30f-Paper.pdf`.

Brendan O'Donoghue. Variational bayesian reinforcement learning with regret bounds. *CoRR*, abs/1807.09647, 2018. URL `http://arxiv.org/abs/1807.09647`.

Brendan O'Donoghue, Ian Osband, Remi Munos, and Vlad Mnih. The uncertainty Bellman equation and exploration. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3839–3848. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/odonoghue18a.html`.

Manfred Opper and David Saad. *From Naive Mean Field Theory to the TAP Equations*, pages 7–20. The MIT Press, 2001.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2701–2710. JMLR.org, 2017.

Jim Pitman. Random weighted averages, partition structures and generalized arcsine laws. *arXiv preprint arXiv:1804.07896*, 2018.

Pascal Poupart, Nikos Vlassis, Jesse Hoey, and Kevin Regan. An analytic solution to discrete bayesian reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 697–704, 2006.

Kai Segadlo, Bastian Epping, Alexander van Meegen, David Dahmen, Michael Krämer, and Moritz Helias. Unified field theoretical approach to deep and recurrent neuronal networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(10):103401, 2022.

Jonas Stapmanns, Tobias Kühn, David Dahmen, Thomas Luu, Carsten Honerkamp, and Moritz Helias. Self-consistent formulations for stochastic nonlinear neuronal dynamics. *Physical Review E*, 101(4):042124, 2020.

Hugo Touchette. The large deviation approach to statistical mechanics. *Physics Reports*, 478(1-3):1–69, Jul 2009. ISSN 0370-1573. doi: 10.1016/j.physrep.2009.05.002. URL `http://dx.doi.org/10.1016/j.physrep.2009.05.002`.

John von Neumann. Distribution of the Ratio of the Mean Square Successive Difference to the Variance. *The Annals of Mathematical Statistics*, 12(4):367 – 395, 1941. doi: 10.1214/aoms/1177731677. URL `https://doi.org/10.1214/aoms/1177731677`.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Jonathan Yedidia. An idiosyncratic journey beyond mean field theory. *Advanced mean field methods: Theory and practice*, pages 21–36, 2001.

Jean Zinn-Justin. *Quantum field theory and critical phenomena*, volume 171. Oxford university press, 2021.