Expectation of the maximum of Normal random variables with applications to reinforcement learning

Anonymous authors
Paper under double-blind review

Abstract

We explore the application of expressions for the expected maxima of Normal random variables to compute the mean of the distribution of fixed points of the Bellman optimality equation for large state space Markov decision processes (MDPs), under a Bayesian framework. Current approaches to computing the statistics of the value functions in reinforcement learning rely on bounds and estimates that do not exploit statistical properties of the Bellman equation which can arise in the large system limit. Specifically, we utilise a recently developed mean field theory called *dynamic mean field programming*, which in principle allows us to compute exactly the the prior and posterior moments of the value functions, under certain conditions on the MDP structure and beliefs. Computing the solution to the mean field equations however relies on computing expected maxima, and current approaches are limited to identically distributed rewards. We apply expressions for expected maxima, in general settings, to compute the mean field solutions of the Bellman equation *at the start of learning.* We analyse the resulting approximations to the mean field equations, establishing Lyapunov stability and contractive properties.

1 Introduction

An outstanding challenge in Bayesian reinforcement learning is how to accurately compute or approximate the posterior distribution of solutions to the Bellman optimality equation. A recent approach to analysing this problem introduces a statistical mean field theory for Bayesian reinforcement learning called dynamic mean field theory (DMFT), Stamatescu (2022). The DMFT results in a set of *mean field equations* dubbed *dynamic mean field programming* (DMFP), that propagate parameter uncertainty through the Bellman optimality equation exactly, under certain conditions.

It is common for reinforcement learning algorithms based on multi-armed bandit theory to use concentration inequalities to derive so-called index policies to design "low regret" algorithms (e.g., upper confidence based algorithms calculate an index for each action), see Lattimore and Szepesvári (2020). Index policies provide excellent solutions for the multi-armed bandit, a Markov decision process with a single state, where the index is over the value of a particular action. Current extensions to general MDPs, where the index is over state-action values, has resulted in both biased estimates for the mean and in loose concentration bounds. In contrast, the DMFP approach is *exact* under certain assumptions, providing a means of quantifying biases in these reinforcement learning algorithms. By solving the mean field equations, we are able to measure the bias present in Bayesian algorithms.

The mean field or DMFP equations, as the name suggests, propagate the *moments* of the Bellman optimality equation. In the finite horizon case, these are the moments of the time-dependent value functions, and in the infinite horizon case, these are the moments of the *iterates* of the Bellman equation, whose fixed point is the optimal value, see Stamatescu (2022). In order to solve the equations, one must compute expectations over the maximum of the state-action value functions. The approach presented in Stamatescu (2022) applies an extreme value theory approximation, and is thus restricted to the large action space limit and the identical reward case. The task of computing the expected maxima under general conditions is the starting point for the present paper, from which our main contributions follow. We anticipate the results presented here will find use in *corrections* to the mean field theory, such as finite size corrections Helias and Dahmen (2020).

After introducing some background material on MDPs and the DMFT, the paper divides into three parts. The first contains an exhaustive review of formulas, approximations and bounds for the expected value of the maximum of a set of random variables. We provide a simple comparative analysis of the bounds, exact formulae and sample estimation to obtain insights into the numerical performance of each approach. The second part of this paper turns to the reinforcement learning problem described above. We simulate the DMFP equation to high accuracy, computing the mean of the Bellman optimality equation, at the start of learning, based on our comparative analysis of formulae for the expected maximum. From this we then compare the bias of existing Bayesian RL algorithms. The third and final part of the paper is dedicated to establishing the stability and contractive properties of the DMFP equations and their approximations. We present both positive and negative results and an open problem.

1.1 Summary of the contributions

- (Expected maximum)

  Our analytical based study of the expected maximum provides exact posterior statistics for the distribution of solutions to the Markov decision process in more general settings than Stamatescu (2022). New simulation studies compute the prior statistics of the distribution, and give further empirical support for validity of the asymptotically exact DMFT for reinforcement learning. Along the way, our study also provides a comprehensive review of formulae and bounds of the expected maximum from the statistics and machine learning literature.

- (Bayesian RL)

  The utility of the DMFP equations and the expressions for expected maxima is demonstrated by revealing the bias in both the prior mean and variance estimates that different Bayesian RL algorithms implicitly assume at the start of learning.

- (Certainty equivalence as a lower bound)

  Notably, we show that, under the conditions of the DMFP theory the *certainty equivalence* heuristic, which replaces the model parameters with their means, is in fact a lower bound to the mean of the value functions. This simple fact was not reported earlier, and follows from an application of Jensen's inequality.

- (Contraction mappings and stability)

  We explicate the difficulty in analysing the DMFP equation for the posterior mean outside of numerical simulations. A suite of modified mean field equations and special cases are proven to be either contraction mappings, stable, or both.

2 Background

We briefly discuss the foundational concepts underlying this paper. These concepts are Markov decision processes and the dynamic mean field theory. In both of these cases our interest is not to develop the general theory of these ideas, but to provide the context within which our work makes sense. We provide significantly more detail on the dynamic mean field theory than on Markov decision processes, as the former is a relatively recent development, see Stamatescu (2022). For more details on Markov decision processes see Kallenberg (2011), Thomas (2007) or Bertsekas (2022).

2.1 Markov decision processes and Bayesian reinforcement learning

We define a Markov decision process (MDP) as a tuple; $\mathcal{M} = (S, A, P, R, \gamma)$. The sets

$$S = \{s : s = 1, ..., N\} \quad \text{and} \quad A = \{a : a = 1, ..., K\}$$

are the *state space* and the *action space*, respectively. The mapping $P : S \times A \to \Delta(S)$ denotes the transition dynamics of the system. For each state-action pair, $P(s, a) = P_{s,a}$, is a probability distribution over states. We write

$$P_{s,a,s'} = Pr[s_{t+1} = s' : s_t = s, a_t = a]$$

for the probability of transitioning, at time t, from state s to state s' under action a. The mapping $R : S \times A \to \mathbb{R}$ defines the reward function for each state-action pair. The rewards are bounded, as are the variances of the reward distributions. The mean of the reward function at a specific state-action pair is written $\rho_{s,a}$. The number $\gamma \in [0, 1]$ is the discount factor.

We use Q-learning to solve our MDP, Thomas (2007). The equation; $V^t_s = \max_a Q^t_{s,a}$, Bertsekas (2012), connecting standard value functions to Q-value functions. The Bellman optimality equations for the optimal Q-value functions is given by

$$Q^{t+1}_{s,a} = \rho_{s,a} + \gamma \sum_{s'} P_{s,a,s'} \cdot \max_{a'} Q^t_{s',a'}, \tag{1}$$

see also Bellman (1966).

If we consider parameter uncertainty in the distribution rewards and transition probabilities, and take a Bayesian approach, the Q-value functions are then random variables. We denote the mean, variance and covariance of the random Q-values as: $\mathbb{E}(Q^t_{s,a}) = \mu^t_{s,a}, \mathrm{Var}(Q^t_{s,a}) = \nu^t_{s,a}$ and $\mathrm{Cov}(Q^t_{s,a}, Q^t_{s,a'}) = \Sigma^t_{a,a'}$ respectively. The mean field theory, from Stamatescu (2022) provides a way of approximating the posterior parameters over state-action values. Under certain conditions, which we specify below, this theory tells us that the posterior statistics for the state-action values is given by the solution to a set of mean field equations. It is these equations that we refer to when we use the term *dynamic mean field programming.*

2.2 Dynamic mean field programming

Dynamic mean field theory is an application of statistical field theory to the equations of dynamic programming, specifically in the case that the underlying MDP is uncertain and one maintains a set of Bayesian beliefs on its parameters. The general motivation behind the development of the dynamic mean field theory comes from the successful application of methods from statistical physics to topics in computer science, and specifically neural network theory. Such interaction between the fields of computer science and statistical field theory has already occurred with results in both theoretical neuroscience, see Sompolinsky et al. (1988), and deep neural networks, see Poole et al. (2016).

One way to visualise the dynamic mean field theory for reinforcement learning, and one that will especially serve the purpose of illustrating how we utilize it within this paper, is to envision a large, highly connected, Markov decision process whose transition probabilities, rewards, and thus value functions are random variables due to Bayesian uncertainty. By high connectivity it is meant that the agent believes it is possible to transition from one state to another state from a large set of potential next states. Thus, the influence of the *uncertainty* of any state-action value function on another is negligible. As the number of states increases this influence of the uncertainty approaches zero. In other words an effective independence between value functions at different state-action pairs is obtained in the large state space limit.

Our goal as stated is to find the parameters of the distribution of fixed points of the Bellman optimality equation when the rewards and transition probabilities are unknown. The core idea that the mean field theory brings to helping solve this problem is that, in the large state space limit, the correlations between the Q-value functions are effectively zero. Therefore we are allowed to consider the Q-value functions as independent random variables.

Assumptions of DMFP. For the dynamic mean field theory to hold a set of assumptions on the distributions of the mean rewards and transition probabilities are required. These assumptions can be seen as analagous to assumptions made in statistical physics, specifically spin glasses, and dynamical systems subject to quenched disorder, see Crisanti and Sompolinsky (2018), and Castellani and Cavagna (2005) respectively. For a full explanation of the specific connection to reinforcement learning see Stamatescu (2022), for our purposes we accept these assumptions justified in the RL context so long as they make sense or people use them. We proceed as follows.

The first assumption required is that the posterior of the means of the rewards, $\rho_{s,a}$, has a finite moment generating function. This condition ensures that the posterior distribution of rewards has the parameters that we would like to compute. The second assumption is that the posterior distributions over states, $P_{s,a} \sim \text{Dir}(\boldsymbol{\alpha}_{s,a})$ are Dirichlet distributions, Kotz et al. (2004), as is the standard in Bayesian reinforcement learning, Poupart et al. (2006), Strens (2000). We add the condition that

$$\frac{\alpha_{s,a,s'}}{\sum_{s'} \alpha_{s,a,s'}} \sim \mathcal{O}\left(\frac{1}{N}\right)$$

The last assumption is that, for non-equal state-action pairs $(s,a) \neq (s',a')$, independence holds between the means $\rho_{s,a}$ and $\rho_{s',a'}$, and between the Dirichlet vectors $P_{s,a}$ and $P_{s',a'}$. $P_{s,a}$ is the Dirichlet random variable, and $P_{s,a,s'}$ is the probability of transition, which is taken as the s'th entry in $P_{s,a}$; $\sum_{s'} P_{s,a,s'} = P_{s,a}$.

Results of the DMFP theory. With the above condition satisfied, one obtains the following consequences. The first consequence is a set of independence results between the Q-value functions, the reward means and the transition probabilities respectively. The first independence result is that the Q-value functions are independent across different state-action pairs, for a given stage t in the iteration process. The second and third independence results are that the Q-value functions are independent of both the reward means and the Dirichlet transition probabilities, across all state-action pairs. Such a result is known as the *propagation of chaos* property. As a consequence of these independence assumptions it is possible, when taking the expectation of an arbitrary Q-value function, to effectively pass the expectation over each of the terms in Equation (1), giving the following evaluation:

$$\begin{aligned}\mathbb{E}\big(Q^{t+1}_{s,a}\big) &= \mathbb{E}\big(\,\rho_{s,a} + \gamma \sum_{s'} P_{s,a,s'} \cdot \max_{a'} Q^t_{s',a'}\,\big) \\ &= \mathbb{E}(\rho_{s,a}) + \gamma \sum_{s'} \mathbb{E}(P_{s,a,s'}) \cdot \mathbb{E}\big(\max_{a'} Q^t_{s',a'}\big) \\ &= \mu_{\rho_{s,a}} + \gamma \sum_{s'} \bar{P}_{s,a,a'} \cdot \mathbb{E}\big(\max_{a'} Q^t_{s',a'}\big).\end{aligned}$$

Note that the expectation does not move through the maximum function. Here, $\mu_{\rho_{s,a}}$ is the mean of the distribution of the means of the rewards, while $\bar{P}_{s,a,s'}$ is the mean over the distribution of the random variable $P_{s,a,s'}$. Additionally note that $P_{s,a}$ is the Dirichlet vector, and $P_{s,a,s'}$ is an element of that vector. The independence across vectors is needed for the factorisation of the expectation as the Q-value function at time t is a function of all of the Dirichlet vectors, i.e., the transition probabilities.

We write the expectation of the Q-value function as: $\mathbb{E}\big(Q^{t+1}_{s,a}\big) = \mu^{t+1}_{s,a}$, giving us:

$$\mu^{t+1}_{s,a} = \mu_{\rho_{s,a}} + \gamma \sum_{s'} \bar{P}_{s,a,s'} \cdot \mathbb{E}\big(\max_{a'} Q^t_{s',a'}\big) \tag{2}$$

which we refer to as the mean field equation.

Discussion of DMFP theory. When running simulations of the mean field equation we are aware of the initial parameters of both the distribution of the mean rewards and the distribution of the Dirichlet transition probabilities. Something we see immediately from the mean field equation is that the value of $\mu^{t+1}_{s,a}$ will be determined primarily by the value $\mu_{\rho_{s,a}}$, that is the mean of the reward means. Therefore, in empirical simulations we should expect to see that the mean value will be equal to $\mu_{\rho_{s,a}} + C$ for some constant amount C. One of our main objectives in our work is to determine this constant precisely, so that we may obtain exact numerical values for the posterior mean of the fixed points of the Bellman equation at the start of learning. This objective motivates our study of the expected value of the maximum of sets of Normally distributed random variables.

Example. Suppose that the mean rewards are Normal, that is $\rho_{s,a} \sim \mathcal{N}(\mu_{\rho_{s,a}}, \nu_{s,a})$, then for each state-action pair, the fixed point $Q^*_{s,a} \sim \mathcal{N}(\mu_{\rho_{s,a}} + C, \nu_{s,a})$. Similarly we could take mean rewards from any other univariate distribution, such as Bernoulli or Beta distributions, and our fixed points would be distributed in the same way.

Regarding the variance of the posterior distribution, the DMFP equation for the variance of the posterior distribution is equal to the variance of the mean rewards plus some amount dependent upon the correlations between the Q-value functions. However, as the Q-value functions are independent under our assumptions, this term is effectively zero, and so the posterior variance of the distribution of fixed points of the Bellman equation is simply the variance of the distribution of the mean rewards. Hence we have justification for only studying the mean in this paper.

Referring back to Equation (2), the mean field equation, we see that in order to obtain accurate numerical results for the posterior mean we have to be able to compute the expected value of the maximum of a set of Normally distributed random variables. This can be done with the use of approximations, upper and lower bounds on the maximum, and with the use of exact formulas for the expected maximum. In this paper we will discuss methods that fit into all of these categories, and explore both the advantages and disadvantages of the different approaches. Some of the criteria we are are interested in is the accuracy of the numerical result, the ease of computation, and the applicability of the method of computation.

The DMFT derives the DMFP equations from a saddle-point equation using *field theoretic* techniques, and becomes exact in the large state space limit, Stamatescu (2022). The derivation also depends upon a large deviations principle established as a propagation of chaos result Sznitman (1991). It is beyond the scope of this paper to elaborate on these topics further.

3 Comparison of bounds and estimates

In this section we introduce the different methods for computing and approximating the expected value of a set of independently distributed random variables. We focus in particular on Normal random variables, though some of the formulae hold for arbitrary distributions.

Formulas for the expected maxima that have closed forms, including special functions, are referred to as ‘exact’. Upper and lower bounds are referred to as either ‘bounds’ or ‘approximations’. Anything requiring numerical integration is strictly an approximation, and Monte-Carlo samples are ‘estimates’

We introduce equations for the expectation of the maximum, which we have termed ‘exact’ formulas. We detail their parameter constraints, derivations, and we provide a short summary of some of the limitations we have found through simulation work and also in theoretical work. The different upper and lower bounds that we study are also introduced, and are given the same treatment as the exact formulae. Our only approximation method that is not a bound or exact formula is the method of Monte-Carlo sample estimation. While we ultimately settle on one of the exact formulas as our ‘standard’, against which we measure the rest, the Monte-Carlo sample estimations also provide a decent benchmark for determining the general degree of accuracy that each bound and exact value provides.

Lastly we summarise the results of our simulations. We tabulate the numerical results of each expression for a quick evaluation of the difference in precision between them, and we generate plots for selected expressions, showing the change in the numerical values as the number of random variables is increased.

3.1 Exact formula for independent, non-identical random variables

For independent, non-identical random variables from an arbitrary distribution we can derive a useful formula for the expected maximum, as follows. Let X be a random variable with cumulative distribution function $F(x)$; by definition

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \, dF(x). \tag{3}$$

A well known way to write the expectation, see Feller (1991), is as

$$\mathbb{E}(X) = \int_0^\infty 1 - F(x)\,dx - \int_{-\infty}^0 F(x)\,dx. \tag{4}$$

Equation (4) can then be written as

$$\mathbb{E}(X) = \int_0^\infty 1 - F(x) - F(-x)\,dx, \tag{5}$$

which can be found in a more general form in David (1981), which can be derived from formulas in Tippett (1925) and Cox (1954). It is then shown in Ross (2003) that if $X_1, ..., X_k$ are independent random variables the above formula can be used to write,

$$\mathbb{E}(\max_i X_i) = \int_0^\infty 1 - \prod_{i=1}^k F_{X_i}(x) - \prod_{i=1}^k F_{X_i}(-x)\,dx. \tag{6}$$

In our simulations the cumulative distribution functions will be Normal c.d.f.'s written as error functions. We will sometimes also refer to this equation as the "Ross" exact formula.

3.2 Exact formula for multivariate Normal variables

A less well known exact formula for the expected maximum can be found in a paper by Biyi Afonja, Afonja (1972). For a multivariate Normal random variable of dimension k, we write $\mathbf{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with k-dimensional probability density function $f_k(\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ to emphasise the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ used in the formula for the multidimensional Normal p.d.f., Moran (1968). We denote the standardized p.d.f. for the standardised variable $z_i = (X_i - \mu_i)/\sigma_i$ by $f_k(\mathbf{z}; \mathbf{R})$ where $\mathbf{R}$ is the correlation matrix of $\mathbf{X}$. We write $\Sigma_{i,j}$ for the covariance between X_i and X_j, and $\sigma_i^2 = \Sigma_{i,i}$.

For a k-dimensional vector $\mathbf{b} = (b_1, ..., b_k)$ we write the complementary cumulative distribution function for the standardised probability density function as

$$\bar{F}(\mathbf{b}; \mathbf{R}) = \int_{\mathbf{b}}^\infty f_k(\mathbf{z}; \mathbf{R})\,d\mathbf{z} = \int_{b_1}^\infty \cdots \int_{b_k}^\infty f_k(\mathbf{z}; \mathbf{R})\,d\mathbf{z}. \tag{7}$$

We now define the following piece-wise function:

$$\lambda_{i,j} := \begin{cases} (\mu_j - \mu_i)/\sqrt{\text{var}(X_i - X_j)} & i \neq j \\ -\infty & i = j \end{cases}. \tag{8}$$

If we then define a vector $\lambda_i = (\lambda_{i,1}, ..., \lambda_{i,k})$ in $\mathbb{R}^k$, we can then define another vector in $\mathbb{R}^{k-1}$ of the form $\boldsymbol{\lambda}_i = \lambda_i$ with the ith element removed. For example: $\boldsymbol{\lambda}_1 = (\lambda_{1,2}, ..., \lambda_{1,k})$, $\boldsymbol{\lambda}_2 = (\lambda_{2,1}, \lambda_{2,3}, ..., \lambda_{2,k})$, and so on.

We also define another vector, this time in $\mathbb{R}^{k-2}$, as $\boldsymbol{\lambda}_{i,j} = \{\lambda_{i,jj'}\}$ for $j \neq j'$, $j, j' \neq i$, where each element is defined by the equation:

$$\lambda_{i,jj'} := \frac{\lambda_{i,j'} - \lambda_{i,j} r_{i,jj'}}{\sqrt{1 - r_{i,jj'}^2}} \tag{9}$$

with the symbol $r_{i,jj'}$ defined in accordance with the matrix $\mathbf{R}_i$ below.

We construct a new correlation matrix of correlations between differences of the random variables: $\mathbf{R}_i = \{r_{i,jj'}\}_{j,j'=1}^k$, where for $j, j' \neq i$, the symbol $r_{i,jj'}$ denoted the correlation between $X_i - X_j$ and $X_i - X_{j'}$. Similar to before, we can then define another matrix: $\mathbf{R}_i^+$ as the matrix $\mathbf{R}_i$ with the ith row and column removed. Lastly, we define the matrix $\mathbf{R}_{i,j} = r_{i,qs \cdot j}$, where $r_{i,qs \cdot j}$ is the partial correlation between $X_i - X_q$ and $X_i - X_s$ given $X_i - X_j$. It is shown in Baba et al. (2004) that for multivariate Normal distributions the partial correlation is equivalent to the conditional correlation, a fact that can be useful for calculating the values of $\mathbf{R}_{i,j}$.

Letting $f_1(x)$ be the standard Normal probability density function in 1-dimension, we define a new symbol

$$\Lambda_{i,j}(x) = \frac{\sigma_i^2 - \mathbf{\Sigma}_{ij}}{\sqrt{\sigma_i^2 + \sigma_j^2 - 2\mathbf{\Sigma}_{ij}}} f_1(x) \tag{10}$$

which allows us to write the "Afonja" equation for the expected maximum of a set of Normal random variables, as given in Equation (3.1) of Afonja (1972) as

$$\mathbb{E}(\max_i X_i) = \sum_{i=1}^{k} \mu_i \bar{F}_{k-1}(\boldsymbol{\lambda}_i; \mathbf{R}_i^+) + \sum_{i=1}^{k} \sum_{j \neq i} \Lambda_{i,j}(\lambda_{i,j}) \bar{F}_{k-2}(\boldsymbol{\lambda}_{i,j}; \mathbf{R}_{i,j}) \tag{11}$$

where

$$\bar{F}_{k-1}(\boldsymbol{\lambda}_i; \mathbf{R}_i^+) = \int_{\lambda_{i,n_1}}^{\infty} \cdots \int_{\lambda_{i,n_{k-1}}}^{\infty} \frac{1}{\sqrt{(2\pi)^{k-1} |\, \mathbf{R}_i^+ \,|}} \exp\Big[-\frac{1}{2} (\mathbf{z})^t \left[\mathbf{R}_i^+\right]^{-1} (\mathbf{z}) \Big] \, d\mathbf{z} \tag{12}$$

$$\bar{F}_{k-2}(\boldsymbol{\lambda}_{i,j}; \mathbf{R}_{i,j}) = \int_{\lambda_{i,m_1}}^{\infty} \cdots \int_{\lambda_{i,m_{k-2}}}^{\infty} \frac{1}{\sqrt{(2\pi)^{k-2} |\, \mathbf{R}_{i,j} \,|}} \exp\Big[-\frac{1}{2} (\mathbf{z})^t \left[\mathbf{R}_{i,j}\right]^{-1} (\mathbf{z}) \Big] \, d\mathbf{z} \tag{13}$$

$n_1, ..., n_{k-1}$ and $m_1, ..., m_{k-2}$ being the appropriate indices, according to the expansion of (11). This is why we specified the matrix included in the definitions of the probability density function and the cumulative distribution function earlier. Equation (11) is equivalent to *Equation 3.1* in Afonja (1972), we have simply altered the notation for succinctness.

3.3 Exact formula for standardised, independent and identical Normal variables

In the more restricted setting of standardised I.I.D. Normal random variables, an exact formula for the expected maximum is given by Kamath (2015). Suppose $X_i \sim \mathcal{N}(0, 1)$ for $i = 1, ..., k$, then the equation

$$\mathbb{E}(\max_i X_i) = \int_{-\infty}^{\infty} xk f_X(x) F_X(x)^{k-1} \, dx \tag{14}$$

can be derived by taking the derivative of the product of cumulative distribution functions. Substituting the particular expressions for a Normal distribution gives the equation

$$\mathbb{E}(\max_i X_i) = \int_{-\infty}^{\infty} \frac{xk}{2\pi} e^{-\frac{x^2}{2}} \cdot \left[\frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x}{\sqrt{2}} \right) \right]^{k+1} \, dx. \tag{15}$$

Because this expression for the expected maximum is only applicable for I.I.D. Normal variables we do not include the results of our simulations of it in this paper. However we note that during those simulations Eq. (15) agreed with both Afonja (11) and Ross (6) to at least 16 decimal places.

3.4 Limitations of the exact formulae

Each of the exact formulas above has its own advantages and disadvantages. The exact formula given by Afonja (1972) works for multivariate correlated Normal variables, that is, it places no restrictions on the parameters of the distribution. It also has many simplifications when the means of the random variables are equal. One unfortunate drawback of Equation (11) is that there seems to be no way to avoid higher dimensional numerical integration. This prevents us from effectively studying the DMFP using this formula when there are more than 3 random variables. Even with the simplifications when the means are equal, we are limited to $k \leq 5$ random variables when doing our expected maximum comparisons. As found in Afonja (1972), Gupta (1963) and Tallis (1961), formulas exist for computing the multiple integrals in Equation (11) for values of k greater then 5, however they are largely impractical for simulation work and so we do not use them. Equation (15), as we have already stated, is restricted by its parameters, however it is much faster to compute. Equation (6) is the most useful of the three that we have found. It has broader applicability than (15), and can be applied in higher dimensions without any significant decrease in accuracy, unlike (11).

As previously mentioned, in our case the cumulative distribution functions are those of a Normal distribution and we can write (6) as the integral over products of error functions. The reason for the accuracy given by this formula is that error functions can be computed to very high precision numerically. Further, since we only have a single integral to compute regardless, we are not hampered by any dimensionality issues which gives us the relatively greater accuracy. Additionally, these facts hold true for any number of random variables, meaning we can use this formula to run DMFP simulations in much higher dimensions, when our action space is larger than 2. We will see the effect that this has on through our simulations later in Section 3 and in Section 4.

Before we discuss the upper and lower bounds we would like to address the question as to why, if exact formulae for the expected maximum exist, do people still consider bounds and other approximations. Other than the limitations already specified there are two broad issues that the exact formulas face. The first is that the numerical computations are not simple when using the exact formulas, unless the number of variables is very small, or the parameters are heavily restricted. Even Equation (6), which was found to be the most applicable throughout our work, requires one to integrate over products of cumulative distribution functions, which in our simulations almost always meant error functions. When compared to some of the bounds this extra computational work may be excessive. A second possibility is that the exact formulas do not lend themselves to the simple algebraic manipulations desirable in theoretical work. As we will make clear in the later sections using the exact formulas is technically more difficult than using many of the bounds.

3.5 A lower bound via Jensen's inequality and the certainty equivalence heuristic

Jensen's inequality provides us a straightforward lower bound on the expected maximum:

$$\mathbb{E}(\max_i X_i) \geq \max_i \mu_i. \tag{16}$$

In the context of the DMFP equations, we note that this implies that the Jensen lower bound provides the so-called "certainty equivalent" control Duff (2002). Explicitly, the DMFP equations under this approximation read as,

$$\mu^{t+1}_{s,a} = \mu_{\rho_{s,a}} + \gamma \sum_{s'} \bar{P}_{s,a,s'} \cdot \max_{a'} \mu^t_{s',a'}. \tag{17}$$

We see that this expression simply replaces the model parameters by their averages. General in discrete state and action RL this is considered a good approximation to the mean of the value functions, which we can interpret as the certainty equivalent (CE) approximation. Originally, CE was derived for the case of linear quadratic Gaussian control, where the state variable is uncertain but the model is known Theil (1957), in which case it is the optimal strategy. In the case of model uncertainty it has heuristically been applied for decades, and more recently adapted to provide algorithms with learning guarantees for linear quadratic control Mania et al. (2019).

The relationship between CE and the posterior mean, for which it is a lower bound, was not noted in the earlier work introducing DMFP Stamatescu (2022).

3.6 Upper bounds for multivariate Normal variables

The first upper bound we consider is given by Ross (2003):

$$\mathbb{E}(\max_i X_i) \leq c + \sum_i^k \int_c^\infty Pr[X_i \geq x]\, dx \tag{18}$$

where the optimal value of c, that is, the one that gives the lowest upper bound, is found by solving: $\sum_i Pr[X_i > c] = 1$. The practice of finding c makes this upper bound somewhat slower and more complicated than the exact equation given in Ross (2003), Equation (6). We determined c exactly for the 2-dimensional cases only, in all other simulations $c = 2$.

One upper bound that we study, and will be useful in a later analysis of the exact expression of the expected maximum for bivariate Normal random variables, is given in a paper by Aven (1985). We have for an arbitrary set of Normal random variables, "Aven's" upper bound:

$$\mathbb{E}(\max_i X_i) \leq \max_i \mu_i + \frac{1}{\sqrt{k}}\Big((k-1)\sum_i \sigma_i^2\Big)^{\frac{1}{2}}. \tag{19}$$

which importantly for our simulations, is true for Normal random variables under any parameter domain. We will use this bound repeatedly when drawing conclusions from the DMFP simulations and comparing it to the exact formula (6).

From Bertsimas et al. (2006) we find the following upper bounds for the expected maxima when $\mathbf{X} \sim_{\boldsymbol{\theta}} (\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. While it is not specified in Bertsimas et al. (2006) exactly what type of distribution $\boldsymbol{\theta}$ is, in this paper we will consider it a Normal distribution with arbitrary means and variances with all variables independent. Writing

$$\alpha_i = \mu_i + \frac{k-2}{2\sigma_i\sqrt{k-1}},$$

we can then write the two upper bounds as:

$$\mathbb{E}(\max_i X_i) \leq \frac{1}{2}\sum_i \left(\mu_i + \sqrt{(\mu_i - \max_i \alpha_i)^2 + \sigma_i^2}\right) + (2-k)\cdot \max_i \alpha_i \tag{20}$$

and

$$\mathbb{E}(\max_i X_i) \leq \frac{1}{2}\sum_i \left(\mu_i + \sqrt{(\mu_i - \min_i \alpha_i)^2 + \sigma_i^2}\right) + (2-k)\cdot \min_i \alpha_i. \tag{21}$$

Under certain parameter conditions described below, these two upper bounds can be reduced to equation (19).

3.7 Tight upper bound for identical Normal variables

Bertsimas et al. (2006) also give us a tight upper bound on the expected maxima for variables with equal means and variances, that is, when $X_i \sim_{\boldsymbol{\theta}} (\mu, \sigma^2)$ for all i:

$$\mathbb{E}(\max_i X_i) \leq \mu + \sigma\sqrt{k-1} \tag{22}$$

An important point to make is that because our MDP is not asymptotic in the action space, tightness of the upper bound is not relevant.

3.8 Bounds for zero mean, i.i.d. Normal variables

From Kamath (2015) we get an upper and lower bound for k many independent Normal variables $X_i \sim \mathcal{N}(0, \sigma^2)$:

$$\mathbb{E}(\max_i X_i) \leq \sigma\sqrt{\log k^2} \tag{23}$$

and

$$\mathbb{E}(\max_i X_i) \geq \frac{1}{\sqrt{\pi \log 2}}\sigma\sqrt{\log k}. \tag{24}$$

While the parameter domain of these bounds are limited to zero mean and equal variances, we will find in our analysis that the lower bound performs much better, both to Equation (23) and to some of the other upper bounds.

Since the different bounds above have different parameter constraints we compute numerical values for each of them over specific parameter domains on which they are well defined. Numerical results for all allowed parameter domains under consideration are given for completeness, and for a more comprehensive comparison of the bounds, approximations and exact values.

3.9 Numerical simulation results

The Monte-Carlo sample estimation results are given in the table below for 10^8 samples of maxima of k-many random variables. It is important to note that the Monte-Carlo sample estimation has a very slow rate of convergence to the true value, and so a high number of samples is required even for the generally weak agreement we see between the Monte-Carlo and exact values in the tabulated data below.

As mentioned earlier, there are certain parameter limitations for which the various exact values and bounds are defined. For this reason we compute numerical values for the formulas across a range of different parameter conditions. Again, throughout this paper we are interested in the DMFP for Normally distributed random variables, and this section is no different. Due to a desire to obtain as many numerical results for the exact formula given by Afonja (11), we generated tables for $k = 2, 3, 4$ and 5 many random variables. All of the random variables come from a k-dimensional Normal distribution, in each column our means and variances, when not 0 or 1, were randomly generated from a Uniform$(0, 1)$ distribution, and kept the same for each equation in that column.

# r.v.’s $= 3$	$\mathcal{N}(0, \sigma^2)$	$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{N}(\mu, \sigma_i^2)$	$\mathcal{N}(\mu_i, \sigma_i^2)$
Monte-Carlo	**0.7450792647**	**1.3157341145**	**1.8704639466**	**1.4338316337**
eq. (6) (Ross)	**0.7450909840**	**1.3157845430**	**1.8705236213**	**1.4338546792**
eq. (11) (Afonja)	**0.7450909840**	**1.3157845430**	**1.8705236213**	-
eq. (18)	1.1125162138	1.3970498097	1.9120496507	2.1262286577
eq. (19) (Aven)	**1.2451107518**	**1.5389876683**	**2.4853683500**	**2.2635242888**
eq. (20) (Bert MAX)	1.2451107518	1.5389876683	2.4277949344	2.0283510222
eq. (21) (Bert MIN)	1.2451107518	1.5389876683	2.4157786926	2.0066731675
eq. (22)	1.2451107518	1.5389876683	-	-
eq. (23) (K. upper)	1.3050591913	-	-	-
eq. (24) (K. lower)	0.6253563224	-	-	-

Figure 1: *Numerical values for* $\mathbb{E}(\max_i X_i)$ *using exact values, bounds and Monte-Carlo estimation under different parameters constraints on the Normally distribution random variables.* $k = 3$

# r.v.'s $= 5$	$\mathcal{N}(0, \sigma^2)$	$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{N}(\mu, \sigma_i^2)$	$\mathcal{N}(\mu_i, \sigma_i^2)$
Monte-Carlo	**1.1261417367**	**1.4278245252**	**1.9965215452**	**1.4667868827**
eq. (6) (Ross)	**1.1261270919**	**1.4278005320**	**1.9965940928**	**1.4668164213**
eq. (11) (Afonja)	**1.1261270919**	**1.4278005320**	**1.9965940928**	-
eq. (18)	1.1515498802	1.4885300836	2.6062061621	2.0133558752
eq. (19) (Aven)	**1.9366491711**	**1.7477860941**	**2.7272233051**	**2.5949895751**
eq. (20) (Bert MAX)	1.9366491711	1.7477860941	2.7231229198	2.204443024
eq. (21) (Bert MIN)	1.9366491711	1.7477860941	2.7244871754	2.3534389117
eq. (22)	1.9366491711	1.7477860941	-	-
eq. (23) (K. upper)	1.7372930018	-	-	-
eq. (24) (K. lower)	0.8324734769	-	-	-

Table 2: *Numerical values for* $\mathbb{E}(\max_i X_i)$ *using exact values, bounds and Monte-Carlo estimation under different parameters constraints on the Normally distribution random variables.* $k = 5$

Aside from the convergence of the Monte-Carlo sample estimation, there are a few key features that can be observed from the information given in the table above. The first point of interest is the equivalent numerical values given by Equations (19), (20), (21), and (22) under the parameter condition of all random variables having equal means and variances, as alluded to in the previous subsection. The second point is the accuracy of the upper bound given in Ross (2003), Equation (18).

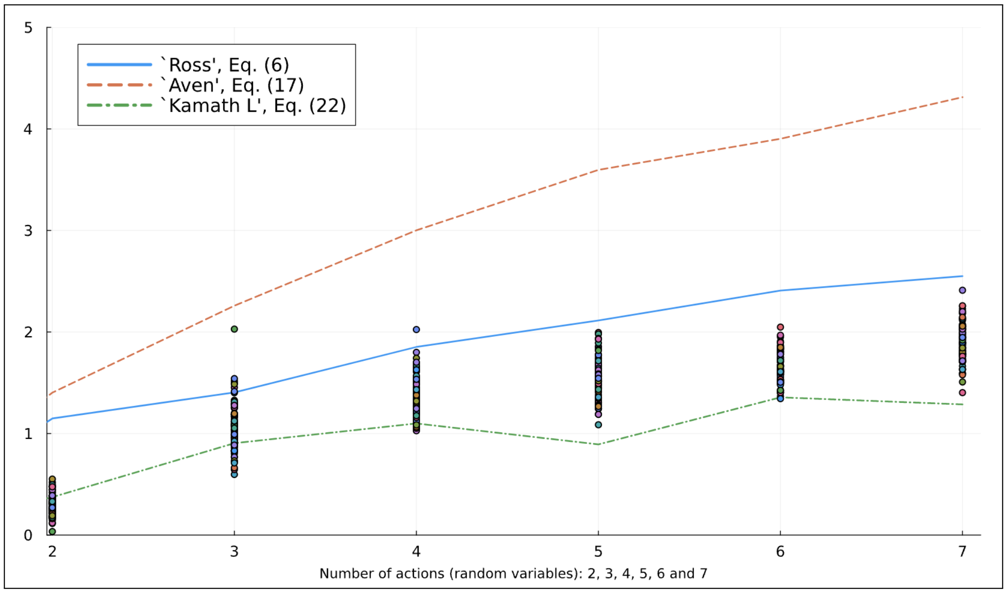

Figure 3: *Numerical values of eq. (6), eq. (19) and eq. (24), and the spread of Monte-Carlo sample estimates.*

In Figure 3 we visualise some of the results above for higher dimensional Normal random variables. Due to the relatively similar performance of the expressions (19), (20), (21) and (22) we only simulate (19) among these bounds. Of the three exact formulas for the expected maximum we only simulate Equation (6) as it can be extended to higher dimensional spaces and holds for arbitrary means and variances. Lastly we have added the lower bound given by Kamath (2015), and a spread of samples obtained via Monte-Carlo sample estimation, represented by a series of points distributed above the number of random variables in the Normal distribution.

Only 20 samples are used to obtain our Monte-Carlo estimates as anything above this reduces the spread in the empirical distribution to the point of collapsing them onto one another.

4 DMFP simulations

In this section we discuss our simulations of the DMFP mean field equation (2), and present our results for the numerical evaluation of the posterior mean of the distribution of fixed points of the Bellman optimality equation, using our insights from the previous section. This simulation work provides the first instance of this posterior mean being computed exactly (up to numerical integration error). It also provides a concrete demonstration of the utility of the dynamic mean field theory for reinforcement learning.

4.1 Empirical mean field equation

In accordance with the DMFT we obtain the posterior mean through an application of value iteration on the mean field equation. Simply put we iterate Equation (2) and it converges to the the true posterior mean.

Algorithm 1 below describes a function that takes as input the defining parameters of a Markov decision process and outputs the posterior mean of the distribution of fixed points of the corresponding Bellman optimality equation of the same MDP, for a specific state-action pair. The expected maximum is presented in mathematical form, as this term can be encoded using a number of different formulas.

Algorithm 1 DMFP - Iteration

```
function DMFP($N, K, T, \gamma, \mu, \Sigma$)
    Initialise. $\bar{P}$ :: $(N \times K \times N)$ − array, every element $= 1/N$
    Initialise. $\mu_\rho$ :: $(N \times K)$ − mean rewards
    Emax :: $(N \times T)$ − array of zeros
    mean :: $(N \times K \times T)$ − array of zeros
    for s = 1 : N do
        for a = 1 : K do
            mean$[s, a, 1] = \mu[a]$
        end for
    end for
    for t = 2 : T do
        for s = 1 : N do
            Emax$[s, t-1] = \mathbb{E}(\max_{a'} Q^t_{s',a'})$
        end for
        for s = 1 : N do
            for a = 1 : K do
                mean$[s, a, t] = \rho[s, a] + \gamma * P[:, a, s]' *$ Emax$[:, t-1]$
            end for
        end for
    end for
    return: mean$[:, :, T]$
end function
```

Figure 4: *The DMFP function. Returns the posterior means for all state-action pairs.*

During our simulations we run the above algorithm for the posterior mean while simultaneously generating a large sample set of $Q^*_{s,a}$ values. We generate this sample set by re-initialising the realisations of the transition probabilities, rewards, and initial $Q^0_{s,a}$ values. This gives us a Monte-Carlo sample estimate of the posterior mean to compare with the DMFP derived posterior mean.

For our Monte-Carlo sample estimation we take 500 samples. This limitation on the number of samples is due to our observation of an accurate res presentation posterior distribution with this many samples. Our investigations showed that the trade-off between decimal precision and computation time did not favour, nor really require, taking any more than this.

As mentioned, we can take many different formulas for the expected maximum and input them into Algorithm 1. To determine the best formula from our collection in Section 3 we do a simple comparison, measuring the average absolute difference between the empirical posterior mean derived from the sample estimates and the posterior mean derived from the DMFP equation. Tabulated below is the average absolute difference corresponding to the formula used in place of the expected maximum in Algorithm 1. We restrict ourselves to only $k = 2$ actions, since this allows use to use a closed form exact value for the expected maximum.

Equation used in DMFP	**Avg absolute difference**
eq. (6) 'Ross'	0.7924783869403459
eq. (11) 'Afonja'	0.7805012326416166
eq. (19) 'Aven'	1.1031464317479736
eq. (20) 'Bert MAX'	1.159607798255533
eq. (21) 'Bert MIN'	1.2431677387725635

Table 5: $k = 2$. *Average absolute differences between posterior means computed using the DMFP and Monte-Carlo sample estimation.* 500 *samples, using the same IID Normal parameters for each DMFP expression.*

While it is clear from the numerical results that Equation (6) provides the most accurate results, we point out that when compared in the previous analysis, both the exact formula from Afonja (1972), Equation (11) and the exact formula of Equation (6) gave almost identical results. This tells us that the discrepancy in outcome between the two versions of the mean field equation is a result of the code that we used in simulation, and more specifically, the numerical methods employed in that code.

It is important to note that when running the DMFP simulations and computing the posterior mean via Equation (2), we have multiple sources of error accumulation to consider. The first is that the numerical integration required to evaluate the expected maximum using Ross (6) is not truly exact, and that any numerical discrepancy will accumulate throughout the iterations. Additionally we have observed certain limitations with the method of numerical integration used in our simulations. All of our simulations have been run using the *Julia programming language*, and our numerical integration is performed by using the QuadGK.jl package. This package uses the Gauss-Kronrod quadrature method of numerical integration, see Calvetti et al. (2000). While this method is very accurate and we have attuned its parameters for the best possible evaluations, we have also found that around the point of inflection of error functions, if the tangent becomes too steep, there is a noticeable break down in the accuracy of the numerical integration as the number of actions increases. The decrease in accuracy is not very strong however, and is consistent with the slow convergence of the expected maximum to the Gumbel distribution. Coupled with the accumulation of error that the iterations entail, and this could contribute to significant differences between the DMFP derived posterior mean and the Monte-Carlo sample estimation of the posterior mean.

4.2 Computation of the posterior distribution

After seeing how well each substitute for the expected maximum performs, and by qualifying potential error risks, we would also like to see how well (6) performs in larger actions spaces, and observe this against some of the bounds.

In Figure 6 we show the overlay of the Normal distribution defined using the parameters $\mu^*_{s,a}$ and $\nu^*_{s,a}$ from the mean field equations, placed on top of the empirical distribution consisting of the samples of $Q^*_{s,a}$. We know that $\nu^*_{s,a}$ is unchanged, and $\mu^*_{s,a}$ is derived using Algorithm 1 and Equation (6), as well as Equation (19).

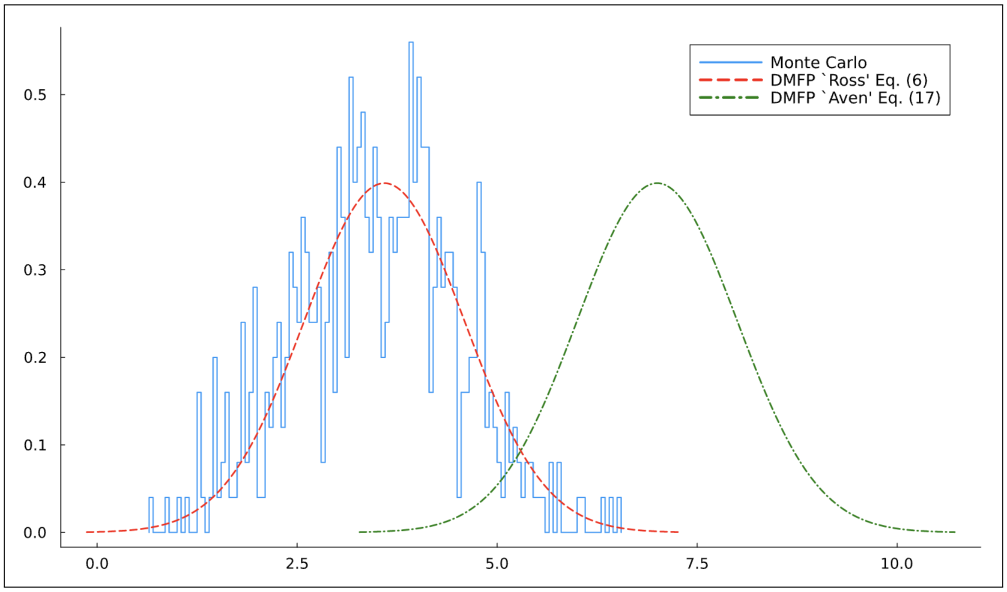

Figure 6: *Overlay of the Normal distribution with DMFP derived parameters, using the 'Ross' exact formula (6) 'Aven's' upper bound Eq. (19) for the expected maximum, onto the empirical distribution of the fixed points Q^* using Monte Carlo sample estimation (*500 *samples). Number of actions $k = 10$.*

We also provide a comparison plot of the accuracy of the DMFP derived posterior mean using a suite of different formulas for the expected maximum. We use Ross (6), Aven (19), and two bounds from Bertsimas et al. (2006), 'Bertsimas MAX' (20) and 'Bertsimas MIN' (21). We add the lines $y = x$ and $y = \ln(x)$ to provide an idea of the inaccuracy of the bounds as the number of actions increases.

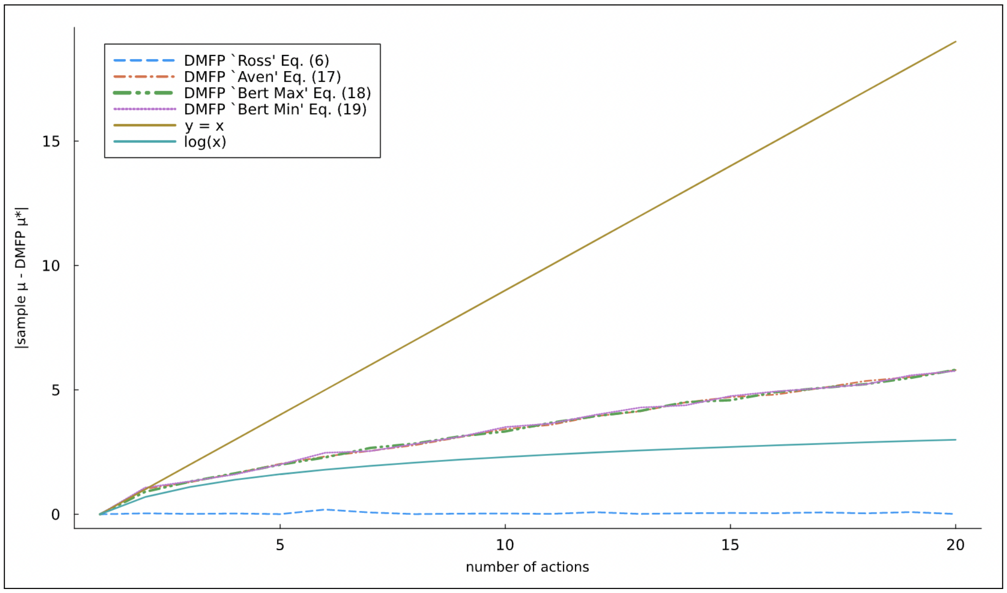

Figure 7: *Numerical values of the absolute difference between the Monte-Carlo sample posterior means and the DMFP posterior mean using Eq.'s (6, (19), (20), and (21) in Algorithm 1. $y = \ln(x)$ and $y = x$ are also plotted for comparison.*

We can see from the results of the simulations above that the 'exact' DMFP using (6) in place of the expected maximum, computes the posterior mean almost perfectly. Even with the qualifications on the actual and potential sources of error that we mentioned at the beginning of this section. This is achieved by iterating the mean field equation, Equation (2), a total of 30 times for each formula. By comparison, at least 500 samples for Monte-Carlo estimation were required to achieve a small difference between posterior mean values. Additionally, as we have seen with the Monte-Carlo convergence rate to the exact value of the DMFP, it would require a much larger number of samples than 500, to see significant agreement between the two posterior means, proving the utility of the DMFP.

5 RL algorithm comparison

Throughout this paper our studies have mostly been concerned with the mean field equation, its numerical approximation and measures of its performance for computing the posterior mean of the distribution of Q^*'s.

Now we wish to provide some more direct insight into the role that the DMFP can provide in other reinforcement learning algorithms, besides Q-learning. We begin by comparing the posterior mean of the fixed points of optimal solutions computed using a suite of different RL algorithms in the finite horizon setting, to the DMFP posterior mean calculation. We compare the posterior means given by solving the two RL algorithms: BEB, see Kolter and Ng (2009), and VBRB, see Sorg et al. (2012), to the posterior mean given by DMFP. This gives us an indication of the accumulation of inaccuracy of the modified value functions, and transition dynamics. In other words we can measure how loose these algorithms are by computing the posterior distribution of the system exactly, and computing the posterior distribution using the alternative value functions.

While the BEB paper establishes that they consider known rewards, they make it clear that their algorithm works for unknown rewards as well. In all of these comparison the RL algorithms we are comparing to a finite horizon setting. BEB and VBRB add a term to the value function, while BOLT modifies the transition probabilities. This means that the empirical distribution of fixed points from the first two algorithms should be shifted by some pre-determined amount. Note that as the trajectories of the three upper bounds: 'Aven' (19), 'Bert MAX' (20) and 'Bert MIN' (21), are so close that we do not use a different line style for our plot.

5.1 BEB algorithm

The *Bayesian exploration bonus* algorithm from Kolter and Ng (2009) chooses actions greedily, following a value function:

$$\tilde{V}_t^*(s,\phi) = \max_{a\in A}\left\{R(s,a) + \frac{\beta}{1+\phi_0(s,a)} + \sum_{s'} P(s' \mid \phi, s, a)\tilde{V}_{t-1}^*(s',\phi)\right\} \tag{25}$$

where $\phi_0(s,a) = \sum_{s'} \phi_{s,a,s'}$. As stated in Kolter and Ng (2009), the parameter ϕ, representing the posterior, does not change throughout the iteration process, and so we can solve for the optimal values by using value iteration.

Through the relationship: $V(s) = \max_a Q(s,a)$, we write a $\hat{Q}$-value function for the approximation as

$$\hat{Q}_{s,a}^t = \mu_{\rho_{s,a}} + \frac{\beta}{1+N} + \sum_{s'} \bar{P}_{s,a,s'} \cdot \max_{a'} \hat{Q}_{s',a'}^{t-1} \tag{26}$$

for the finite horizon setting. Here the discount factor γ is set to 1, and since $\phi_{s,a,s'}$ are in our system all $1/N$, the sum in the exploration bonus equals N.

The results of our simulations presented in Figure 8 demonstrate the slight bias introduced through the modified value function in BEB. When compared to the DMFP equation for the posterior mean we can see that this bias is relatively significant, especially in the large system limit, where the mean field equation approaches perfect numerical accuracy.

The absolute difference between $\hat{Q}^*_{s,a}$ and $\mu^*_{s,a}$ is

$$\left| \frac{\gamma}{N} \sum_{s'} \Big(\max_{a'} \hat{Q}^*(s', a') - \mathbb{E}(\max_{a'} Q^*(s', a')) \Big) + \frac{\beta}{1+N} \right|, \tag{27}$$

setting this equal to zero, we can determine the value of β that will equalize the two fixed points: $\beta = \frac{\gamma(1+N)}{N} \sum_{s'} \Big(\max_{a'} \hat{Q}^*(s', a') - \mathbb{E}(\max_{a'} Q^*(s', a')) \Big)$. However this value of β would not be able to be determined until after the fixed points have been found, hence we are unable to assign this to β. Additionally β is dependent upon the values N, K, γ and the variance of the reward distribution, meaning for a given set of these values a value of β could be found such that the two fixed points, $\hat{Q}^*_{s,a}$ and $\mu^*_{s,a}$, were made to be equal, or at least extremely close.

We can also quantify the difference between $\hat{Q}^*_{s,a}$ and $\mu^*_{s,a}$ in the case of Normal reward distributions by finding a scale d, where $\hat{Q}^*_{s,a} = \mu^*_{s,a} + d \cdot \sigma^2_{s,a}$. It is immediately seen that

$$d = \frac{\frac{\gamma}{N} \sum_{s'} \Big(\max_{a'} \hat{Q}^*_{s,a} - \mathbb{E}(\max_{a'} Q^*(s, a)) \Big) + \frac{\beta}{1+N}}{\sigma^2_{s,a}}. \tag{28}$$

Unlike the value of β in the BEB algorithm which is fixed, we are free to consider a value d_t that can be calculated at each time step t: $\hat{Q}^t_{s,a} = \mu^t_{s,a} + d_t \cdot \sigma^2_{s,a}$. The resulting formula for d_t is given by simply replacing the fixed point values of Equation (28) with $\hat{Q}^t_{s',a'}$ and $Q^t(s', a')$ respectively.

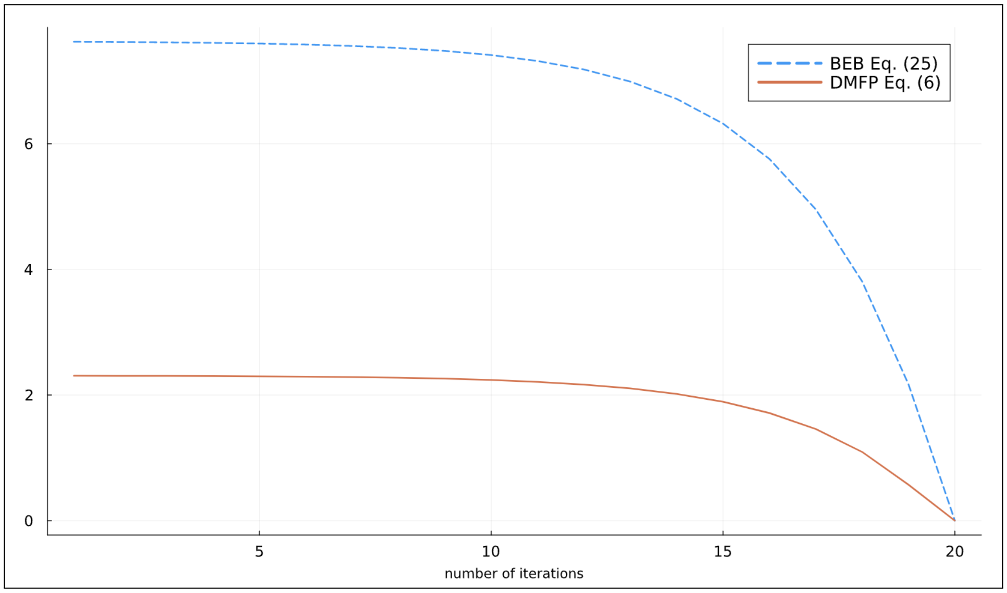

Figure 8: *Convergence of Equations (26) and (2) using (6) for the expected maximum.* $N = 500$*,* $K = 3$*,* $\gamma = 0.7$*, and* $\beta = 2(20)^2$*.*

In Figure 8, our comparison of the DMFP equation and the BEB Equation (26) is under the condition that the rewards are known. Ordinarily the DMFP considers unknown rewards, but we simulate a known reward DMFP equation to provide a valid comparison with BEB.

5.2 Variance based reward bonus (VBRB)

The *variance based reward bonus* algorithm, found in Ghavamzadeh et al. (2016) and Sorg et al. (2012), has a value function

$$\tilde{V}_t^*(s,\phi) = \max_{a\in A}\left\{R(s,\phi,a) + \hat{R}_{s,\phi,a} + \sum_{s'\in S} P(s' \mid s,\phi,a)\tilde{V}^*_{t-1}(s',\phi)\right\} \tag{29}$$

where

$$\hat{R}_{s,\phi,a} = \beta_R\,\sigma_{R(s,\phi,a)} + \beta_P\sqrt{\sum_{s'\in S}\sigma^2_{P(s'|s,\phi,a)}}. \tag{30}$$

The constants β_R and β_P control the magnitude of the exploration bonus, while the variances in Equation (30) are defined respectively as:

$$\sigma^2_{R(s,\phi,a)} = \int_\theta R(s,\theta,a)^2\,b(\theta)\,d\theta - R(s,\phi,a)^2, \tag{31}$$

$$\sigma^2_{P(s'|s,\phi,a)} = \int_\theta P(s' \mid s,\theta,a)^2\,b(\theta)\,d\theta - P(s' \mid s,\phi,a)^2. \tag{32}$$

which are the variances of the rewards and Dirichlet transition probabilities respectively. We can construct a Q-value function similar to the one above for the BEB value function:

$$\hat{Q}^t(s,\phi,a) = R(s,\phi,a) + \hat{R}_{s,\phi,a} + \sum_{s'} P(s' \mid s,\phi,a)\cdot\max_{a'}\left(\hat{Q}^{t-1}(s',\phi,a')\right), \tag{33}$$

For our comparisons we use values for the parameters β_R and β_P taken from the respective paper. From, Sorg et al. (2012), we set $\beta_R = \frac{1}{\sqrt{p}}$ and $\beta_P = \frac{\gamma N}{1-\gamma}\frac{1}{\sqrt{p}}$ with the probability $p = 0.5$.

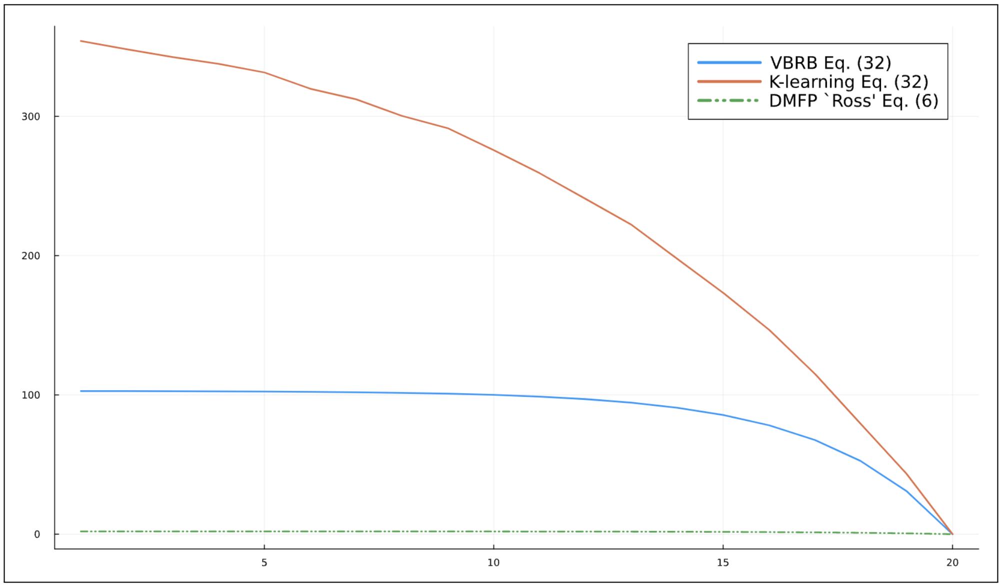

Figure 9: *Convergence of Equations (33), (2) using (6) for the expected maximum, and Eq. (34). $N = 500$, $K = 3$, $\gamma = 0.7$, and $\sigma^2_{s,a} = 1$ and $\tau = 5$.*

P

In Figure (9) above we compare the Variance based reward bonus equation for a finite horizon, to the DMFP equation, and we include the 'K-learning' Bellman optimality equation given by O'Donoghue (2021), Equation 6, for a single episode.

$$\mathcal{B}_l^1(\tau, K_l)(s,a) = \mathbb{E}^1\mu_l(s,a) + \frac{\sigma^2 + (L-l)^2}{2\tau} + \sum_{s' \in S} \mathbb{E}^1 P_l(s' \mid s,a)(\tau \log \sum_{a' \in A} \exp(K_l(s',a')/\tau)). \tag{34}$$

Note that Equation (10) from O'Donoghue (2021) provides a formula for the "temperature" parameter τ that gives a bound on the posterior expected maximum. With the parameters of the MDP that we establish for our simulation inf Figure 9, this formula comes out to approximately 370, well above the value $\tau = 5$ in our simulation, from which we see already that the function (34) upper bounds both the DMFP equation and the VBRB equation.

5.3 Additional Bayesian RL algorithms

There are many other deterministic Bayesian reinforcement learning algorithms that work by altering the value function in some way, with the goal of improving some learning outcome. See Ghavamzadeh et al. (2016) for a large collection of Bayesian reinforcement learning algorithms. Some of the algorithms seem more obviously amenable to treatment with use of the DMFT than others. One of these is *Bayesian Optimistic Local Transitions*, or BOLT, which is a natural development of the process of including exploration bonuses found in the previous two algorithms, where the BOLT algorithm places an exploration bonus in the transition probabilities, see Araya et al. (2012). Of course, there exist non-deterministic Bayesian RL algorithms. Most prominent are those based on Thompson sampling Osband and Van Roy (2017), but other approaches exist as well, such as the *Best Of Sampled Set* (BOSS) algorithm Asmuth et al. (2012).

6 Stability analysis

The correspondence between dynamical systems and function iteration allows us to study the stability of the dynamical system represented by the mean field equation. As we will see the point around which we expect stability is the fixed point $\mu^*_{s,a}$, which we do not prove exists, but can observe empirically through our simulations. Throughout this section we will simply assume that the mean field equation has a fixed point, $\mu^*_{s,a}$, and that this fixed point is arrived at through iterations in exactly the same way as when finding the fixed point of the standard Bellman equation.

Formally we examine the stability of the equilibrium solution to the dynamical system corresponding to the function iteration of $\mu^{t+1}_{s,a}$. As we have considered different bounds on the expected maximum in our DMFP simulations we consider the stability of the mean field equation when the expected max is replaced by one of these formulas, specifically we look at the exact formula given in Afonja (1972) when the number of actions is $k = 2$, the upper bound approximation to the expected maximum from Aven (1985), Equation (19). We also consider the stability of the mean field equation in the large action space limit, where our expected maximum becomes a parameter for a Gumbel distribution, Gumbel (1954).

To show that a system is stable at a fixed point we have to compute the Jacobian matrix of the system, and determine the corresponding eigenvalues of the Jacobian. If they have absolute values less than 1, then the system is stable at that fixed point. It is clear that our fixed point is the optimal value, obtained via the iteration process on the mean field equation.

6.1 Avens upper bound Aven (1985)

Proposition 1. If the expected maximum term of the mean field equation (2) is substituted by the upper bound (19), then the resulting iterative function is stable around an equilibrium point is stable provided the ratio γ/N is strictly less than 1.

Proof. Consider the DMFP equation with the upper bound given by Aven (1985) in place of the expected value of the maximum:

$$\mu^{t+1}_{s,a} = \mu_{\rho_{s,a}} + \gamma \sum_{s'} \bar{P}_{s'|s,a} \cdot \Big(\max_{a'} \mu^t_{s',a'} + \sqrt{\frac{n-1}{n} \sum_{a'} (\sigma^2_{s',a'})^t} \Big). \tag{35}$$

In order to show stability we compute eigenvalues of the Jacobian and show that they are strictly less than 1. We compute an arbitrary element of the Jacobian matrix, taking derivatives as follows:

$$\frac{\partial \mu^{t+1}_{s,a}}{\partial \mu^t_{(s,a)'}} = \frac{\partial}{\partial \mu^t_{(s,a)'}} \left[\mu_{\rho_{s,a}} + \gamma \sum_{s'} \bar{P}_{s'|s,a} \cdot \Big(\max_{a'} \mu^t_{s',a'} + \sqrt{\frac{n-1}{n} \sum_{a'} (\sigma^2_{s',a'})^t} \Big) \right] \tag{36}$$

Here we have taken the partial derivative with respect to the previous time mean, at an arbitrary state-action pair.

$$\begin{aligned} \frac{\partial \mu^{t+1}_{s,a}}{\partial \mu^t_{(s,a)'}} &= \frac{\gamma}{N} \sum_{s'} \frac{\partial}{\partial \mu^t_{(s,a)'}} \Big(\max_{a'} \mu^t_{s',a'} + \sqrt{\frac{n-1}{n} \sum_{a'} (\sigma^2_{s',a'})^t} \Big) & (37) \\ &= \frac{\gamma}{N} \sum_{s'} \frac{\partial}{\partial \mu^t_{(s,a)'}} \big(\max_{a'} \mu^t_{s',a'} \big) & (38) \\ &= \frac{\gamma}{N} \delta_{\mu^t_{(s,a)'}, \max_{a'} \mu^t_{s',a'}} & (39) \end{aligned}$$

where $\delta_{\mu^t_{(s,a)'}, \max_{a'} \mu^t_{s',a'}}$ is the Kronecker delta function, defined here as

$$\delta_{\mu^t_{(s,a)'}, \max_{a'} \mu^t_{s',a'}} = \begin{cases} 1 & \max_{a'} \mu^t_{s',a'} = \mu^t_{(s,a)'} \\ 0 & \text{otherwise.} \end{cases} \tag{40}$$

We again write the Jacobian of the system using a different indexing system, for greater clarity. The Jacobian is

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mu^{t+1}_1}{\partial \mu^t_1} & \cdots & \frac{\partial \mu^{t+1}_1}{\partial \mu^t_{|S|\times|A|}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mu^{t+1}_{|S|\times|A|}}{\partial \mu^t_1} & \cdots & \frac{\partial \mu^{t+1}_{|S|\times|A|}}{\partial \mu^t_{|S|\times|A|}} \end{bmatrix} \tag{41}$$

and then once again becomes a matrix consisting of a single column of γ/N with every other entry 0. To see this, suppose that $\max_{a'} \mu^t_{s',a'} = \mu^t_k$ for some k. Then the indicator function will return 1, only when the derivative is taken with respect to μ^t_k, and it will do this for μ^{t+1}_i for all i. This is true also at the fixed point μ^*_t, the point around which we expect stability. Writing out the Jacobian in terms of basis vectors we will have

$$\mathbf{J} = \frac{\gamma}{N} \mathbf{e}_k + 0 \cdot \sum_{j \neq k} \mathbf{e}_j. \tag{42}$$

Taking the determinant of $\mathbf{J} - \lambda \mathbf{I}$ to find the eigenvalues, will always give us $\gamma/N - \lambda$ times the determinant of a minor matrix equal to $-\mathbf{I}$. This gives

$$\det(\mathbf{J} - \lambda \mathbf{I}) = (\gamma/N - \lambda)(-\lambda)^{|S|\times|A|-1}$$

and subsequently, eigenvalues of $\lambda_1 = \gamma/N$ with all others equal to 0. As γ and N are always positive but less than 1, the eigenvalues of $\mathbf{J}$ have absolute value strictly less than 1, so long as both γ and N are not both equal, and the ratio remains strictly less than 1. □

6.2 Exact formula for $k = 2$

The analysis of the stability of the mean field equation for an arbitrary number of random variables runs into technical challenges, specifically the ordering of the expectations does not satisfy many of the usual inequalities of absolute differences. Some work in this are has been done. For this reason we study the Lyapunov stability of the mean field equation for bivariate normal random variables. We use a special form of the expected maximum for bivariate random variables that can be derived directly from Equation (11), but can also be found in Clark (1961), Ker (2001) and Ross (2003).

Proposition 2. The mean field equation (2) is stable around an equilibrium point when the action space of our MDP has cardinality 2; $|A| = 2$, and the discount factor γ is strictly less than 1.

Proof. Suppose we have a random variable $\mathbf{X} = (X_1, X_2) \sim \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $X_1 \perp X_2$, we can write

$$\mathbb{E}(\max_i X_i) = \mu_1 \Phi\left(\frac{\mu_1 - \mu_2}{\alpha}\right) + \mu_2 \Phi\left(\frac{\mu_2 - \mu_1}{\alpha}\right) + \alpha \phi_1 \left(\frac{\mu_2 - \mu_1}{\alpha}\right) \tag{43}$$

where $\alpha = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\Sigma_{1,2}}$. Putting this expression into the mean field equation and then taking a partial derivative with respect to the mean of the first variable we have

$$\frac{\gamma}{N} \frac{\partial}{\partial \mu^t_{s',a_1}} \left[\mu^t_{s',a_1} \Phi\left(\frac{\mu^t_{s',a_1} - \mu^t_{s',a_2}}{\alpha}\right) + \mu^t_{s',a_2} \Phi\left(\frac{\mu^t_{s',a_2} - \mu^t_{s',a_1}}{\alpha}\right) + \alpha \phi_1 \left(\frac{\mu^t_{s',a_2} - \mu^t_{s',a_1}}{\alpha}\right) \right]. \tag{44}$$

Consider the partial derivative of the first cumulative distribution function. We have

$$\begin{aligned} \frac{\partial}{\partial \mu^t_{s',a_1}} \Phi\left(\frac{\mu^t_{s',a_1} - \mu^t_{s',a_2}}{\alpha}\right) &= \frac{\partial}{\partial \mu^t_{s',a_1}} \int_{-\infty}^{\frac{\mu^t_{s',a_1} - \mu^t_{s',a_2}}{\alpha}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\, dt \qquad (45) \\ &= \frac{1}{\alpha\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{\mu^t_{s',a_1} - \mu^t_{s',a_2}}{\alpha}\right)^2 \right\}. \qquad (46) \end{aligned}$$

From which it is easy to see that the derivative of the second cumulative distribution function is

$$\frac{\partial}{\partial \mu^t_{s',a_1}} \Phi\left(\frac{\mu_2 - \mu_1}{\alpha}\right) = -\frac{1}{\alpha\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{\mu^t_{s',a_2} - \mu^t_{s',a_1}}{\alpha}\right)^2 \right\} \tag{47}$$

and for the other derivative with respect to μ^t_{s',a_2} we have

$$\begin{aligned} \frac{\partial}{\partial \mu^t_{s',a_2}} \Phi\left(\frac{\mu_2 - \mu_1}{\alpha}\right) &= \frac{1}{\alpha\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{\mu^t_{s',a_2} - \mu^t_{s',a_1}}{\alpha}\right)^2 \right\} \qquad (48) \\ \frac{\partial}{\partial \mu^t_{s',a_2}} \Phi\left(\frac{\mu_1 - \mu_2}{\alpha}\right) &= -\frac{1}{\alpha\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{\mu^t_{s',a_1} - \mu^t_{s',a_2}}{\alpha}\right)^2 \right\}. \qquad (49) \end{aligned}$$

The derivative of the probability density function is simply

$$\frac{\partial}{\partial \mu^t_{s',a_1}} \phi\left(\frac{\mu^t_{s',a_2} - \mu^t_{s',a_1}}{\alpha}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\mu^t_{s',a_2})^2}{2\alpha^2}} \cdot \frac{\partial}{\partial \mu^t_{s',a_1}} \left[e^{-\frac{(\mu^t_{s',a_1})^2}{2\alpha^2}} e^{\frac{\mu^t_{s',a_1}\mu^t_{s',a_2}}{\alpha^2}} \right] \tag{50}$$

where

$$\frac{\partial}{\partial \mu^t_{s',a_1}} \left[e^{-\frac{(\mu^t_{s',a_1})^2}{2\alpha^2}} e^{\frac{\mu^t_{s',a_1}\mu^t_{s',a_2}}{\alpha^2}} \right] = \frac{\mu^t_{s',a_2}}{\alpha^2} e^{-\frac{(\mu^t_{s',a_1})^2}{2\alpha^2}} e^{\frac{\mu^t_{s',a_1}\mu^t_{s',a_2}}{\alpha^2}} - \frac{\mu^t_{s',a_1}}{\alpha^2} e^{-\frac{(\mu^t_{s',a_1})^2}{2\alpha^2}} e^{\frac{\mu^t_{s',a_1}\mu^t_{s',a_2}}{\alpha^2}}. \tag{51}$$

We simplify notation, letting $\mu^t_{s',a_1} = \mu_1$ and $\mu^t_{s',a_2} = \mu_2$, and writing $g_{12} = \frac{\mu_1-\mu_2}{\alpha}$ and $g_{21} = \frac{\mu_2-\mu_1}{\alpha}$. This allows us to write the full partial derivative with respect to μ_1 as:

$$\frac{\gamma}{N}\sum_{s'}\left[\Phi(g_{12}) + \frac{1}{\alpha\sqrt{2\pi}}\left(\mu_1 e^{-\frac{1}{2}g_{12}^2} - \mu_2 e^{-\frac{1}{2}g_{21}^2}\right) + \frac{\alpha}{\sqrt{2\pi}}\left(\frac{\mu_2}{\alpha^2} - \frac{\mu_1}{\alpha^2}\right)e^{-\frac{1}{2}(\mu_1-\mu_2)^2}\right] \tag{52}$$

The corresponding partial derivative with respect to μ_2 is

$$\frac{\gamma}{N}\sum_{s'}\left[\Phi(g_{21}) + \frac{1}{\alpha\sqrt{2\pi}}\left(\mu_2 e^{-\frac{1}{2}g_{21}^2} - \mu_1 e^{-\frac{1}{2}g_{12}^2}\right) + \frac{\alpha}{\sqrt{2\pi}}\left(\frac{\mu_1}{\alpha^2} - \frac{\mu_2}{\alpha^2}\right)e^{-\frac{1}{2}(\mu_1-\mu_2)^2}\right]. \tag{53}$$

Writing Equation (52) at the fixed point with A, and Equation (53) at the fixed point with B, the Jacobian matrix and characteristic equation are:

$$\begin{bmatrix} A & B \\ A & B \end{bmatrix}; \quad \lambda^2 - \lambda(A+B) = 0 \tag{54}$$

with eigenvalues of 0 and $A+B$. In the notation defined above, $A+B = \gamma/N \cdot \sum_{s'}[\Phi(g_{12}) + \Phi(g_{21})]$, which is further expanded as

$$A + B = \frac{\gamma}{N}\sum_{s'}\left[\Phi\left(\frac{\mu^*_{s',a_1} - \mu^*_{s',a_2}}{\alpha}\right) + \Phi\left(\frac{\mu^*_{s',a_2} - \mu^*_{s',a_1}}{\alpha}\right)\right]. \tag{55}$$

It is clear that the expression inside the summation can be written as

$$\Phi(g_{12}) + \Phi(-g_{12}) = 1 - \frac{1}{\sqrt{2\pi}}\int_{-g_{12}}^{g_{12}} e^{-\frac{1}{2}t^2}\,dt \tag{56}$$

assuming $g_{12} \geq 0$. If $g_{21} \leq 0$ we integrate over $[-g_{21}, g_{21}]$ instead. It is clear that (56) is bounded above by 1 (true when $g_{21} \leq 0$ also), which implies

$$A + B = \frac{\gamma}{N}\sum_{s'}\left[\Phi\left(\frac{\mu^*_{s',a_1} - \mu^*_{s',a_2}}{\alpha}\right) + \Phi\left(\frac{\mu^*_{s',a_2} - \mu^*_{s',a_1}}{\alpha}\right)\right] \tag{57}$$

$$\leq \frac{\gamma}{N}\sum_{s'} 1 \tag{58}$$

$$= \gamma. \tag{59}$$

Therefore, so long as the discount factor is strictly less than 1, the 2-action space mean field equation is stable. □

6.3 Gumbel distribution

Consider a set $X_1, ..., X_k$ of Normal IID random variables. As the number of random variables increases to infinity, that is, as $i \to \infty$, the cumulative distribution function of the maximum of this set of random variables tends to the cumulative distribution function of the Gumbel distribution, see Von Mises (1936) and Fisher and Tippett (1928).

Define constants $a_n = 1/n\phi(b_n)$ and $b_n = \Phi^{-1}(1 - 1/n)$, then

$$\lim_{k\to\infty} F(a_n x + b_n)^k = e^{-e^{-x}} = G(x; 0, 1) \tag{60}$$

where $G(x; 0, 1)$ is the cumulative distribution function of the Gumbel distribution with location parameter $\nu = 0$ and scale parameter $\beta = 1$. Therefore if we have IID Normal random variables, constants exist such that

$$Pr[\max_i X_i \leq x] = \prod_i Pr[X_i \leq x] = F(x)^k \to_{k\to\infty} G(x; \nu, \beta). \tag{61}$$

We can absorb the constants a_n and b_n into the random variables X_i and so maintain convergence to the standardized Gumbel distribution.

Proposition 3. As the number of actions tends to infinity, the mean field equation (2) is stable.

Proof. As the number of actions in our MDP model tends to infinity, $\max_a Q^t(s,a)$ becomes a Gumbel distributed random variable, and hence its expected value is just the mean value $\nu + \beta\gamma_{\text{em}}$, where γ_{em} is our notation for the Euler-Mascheroni constant. With $\nu = 0$ and $\beta = 1$ we can substitute this into the mean field equation as

$$\mu^{t+1}_{s,a} = \mu_{\rho_{s,a}} + \gamma \sum_{s'} \bar{P}_{s,a,s'} \cdot \gamma_{\text{em}} \tag{62}$$

$$= \mu_{\rho_{s,a}} + \frac{\gamma}{N} \cdot \frac{N}{\gamma_{\text{em}}} \tag{63}$$

$$= \mu_{\rho_{s,a}} + \frac{\gamma}{\gamma_{\text{em}}}. \tag{64}$$

It is easy to see that all partial derivatives will be zero, hence all eigenvalues of the Jacobian matrix will be zero, that is, strictly less than 1 and hence stable. □

7 Contraction mappings

Throughout this section we use the standard definition of a contraction mapping defined on a metric space: Given a map $F : X \to X$ on a metric space X, we call F a contraction if for all $x, y \in X$, there exists some real $k \in [0, 1)$ such that

$$d_X(F(x), F(y)) \leq k\, d_X(x, y) \tag{65}$$

where d_X is the metric defined on X.

In this section we show that different forms of the mean field equation are contractions. We do this by considering the mean field equation over all state-action pairs as an operator

$$\mathbf{F}_{\text{DMFP}} : \mathbb{R}^{|S|\times|A|} \to \mathbb{R}^{|S|\times|A|}, \quad \boldsymbol{\mu}^{t+1} = \mathbf{F}_{\text{DMFP}}(\boldsymbol{\mu}^t).$$

Similar to the consideration of stability, we begin by using the upper bound (19) in place of the expected maximum. After this we consider the case when the number of action goes to infinity $k \to \infty$, and we have a Gumbel distributed random variable for our expected maximum.

Proposition 4. If the expected maximum term in the mean field equation is substituted with the upper bound (19), then this form of the mean field equation is a contraction.

Proof. Let us write $\nu^t(s) = \max_a \mathbb{E}\left(Q^t(s,a)\right) = \max_a \mu^t_{s,a}$, and $(\nu')^t(s) = \max_a (\mu'_{s,a})^t$. The proof of the contraction property follows identically the one given by Bertsekas (2012), and found in Agarwal et al. (2019). As in those proofs, we have

$$|\max_a \mu^t_{s,a} - \max_a (\mu'_{s,a})^t\,| \leq \max_a |\,\mu^t_{s,a} - (\mu'_{s,a})^t\,|. \tag{66}$$

For a moment let us consider $\mu^{t+1}_{s,a}$ and $(\mu'_{s,a})^{t+1}$ from $\boldsymbol{\mu}^{t+1}$ and $(\boldsymbol{\mu}')^{t+1}$ respectively.

$$\mu_{\rho_{s,a}} + \gamma \sum_{s'} \overline{P}_{s'|s,a} \cdot \Big(\max_{a'} \mu^t_{s',a'} + \sqrt{\frac{k-1}{k} \sum_{a'} (\sigma^2_{s',a'})^t} \Big) \tag{67}$$

$$\mu_{\rho_{s,a}} + \gamma \sum_{s'} \overline{P}_{s'|s,a} \cdot \Big(\max_{a'} (\mu'_{s',a'})^t + \sqrt{\frac{k-1}{k} \sum_{a'} (\sigma^2_{s',a'})^t} \Big). \tag{68}$$

As we can see, the mean of the mean rewards is the same for both expressions. Additionally, the sum over the square root term is the same in both expressions, and since the summation is distributive, the sum over the square root term will be the same in both expressions, hence when we take the absolute difference of the two we get

$$\left| \gamma \sum_{s'} \bar{P}_{s'|s,a} \cdot \max_{a'} \mu^t_{s',a'} - \gamma \sum_{s'} \bar{P}_{s'|s,a} \cdot \max_{a'} (\mu'_{s',a'})^t \right| \tag{69}$$

which we write succinctly as $\left| \gamma \bar{P}_{s,a} \nu^t(s) - \gamma \bar{P}_{s,a} (\nu')^t(s) \right|$. With this notation specified we can now prove the main result. Using the ℓ^∞-norm we have

$$\begin{aligned} \left\| \boldsymbol{\mu}^{t+1} - (\boldsymbol{\mu}')^{t+1} \right\|_\infty &= \gamma \left\| \bar{P} \nu^t - \bar{P} (\nu')^t \right\| && (70) \\ &\leq \gamma \left\| \nu^t - (\nu')^t \right\|_\infty && (71) \\ &= \gamma \max_s | \, \nu^t(s) - (\nu')^t(s) \, | && (72) \\ &\leq \gamma \max_s \max_a | \, \mu^t_{s,a} - (\mu'_{s,a})^t \, | && (73) \\ &= \gamma \left\| \boldsymbol{\mu}^t - (\boldsymbol{\mu}')^t \right\|_\infty && (74) \end{aligned}$$

which concludes the proof. □

A trivial lower bound on the expected maxima that we have not covered so far is one we call the Jensen lower bound, since via the convexity of the maximum function, the expected maximum is greater than the maximum of the expected values, see Feller (1971). It is trivial to see that the mean field equation is a contraction when the expected maximum is substituted with the maximum of the expected values. As both the ‘Aven’ and Jensen mean field equation variants are contractions, they have fixed points, and since these variants bound the exact mean field equation, we know that the sequence of points of the exact mean field equation over iterations is bounded. If it could be shown that the sequence $\{\mu^0_{s,a}, \mu^1_{s,a}, ...\}$ is monotonic, this would also furnish a proof that the mean field equation, (2), has a fixed point, without resorting to a proof that it is a contraction.

Similar to the Jensen bound form of the mean field equation, the mean field equation where $k \to \infty$ and we use the Gumbel mean for the expected maximum is trivially a contraction. Looking at the Gumbel mean field equation from the previous section:

$$\mu^{t+1}_{s,a} = \mu_{\rho_{s,a}} + \frac{\gamma}{\gamma_{\text{em}}} \tag{75}$$

we can see that there is no dependency on the previous mean values, and so the difference between any two μ’s will be zero. This satisfies the definition of a contraction for any value of $k \in [0, 1)$.

Similar to our examination of stability, our results for contraction mappings are limited to a handful of special cases and estimations of the mean field equation. While extending these results to the general mean field equation is beyond the scope of this paper, we note that the cumulative evidence for this conjecture being true prompts us to study this in future work.

8 Final remarks

Improving methods of sample efficiency in Bayesian reinforcement learning is an ongoing process. Our study has provided simulation evidence of the utility of the DMFP for this purpose, and validation of the prerequisite assumptions required to formulate the mean field equation. Additionally we have managed to measure the accuracy of different estimation formulas for the expected maximum of a set of Normal random variables, a comparative study that we have then shown significantly effects the computation of posterior statistics of the Bellman optimality equation.

We hope that this work sparks further study into both the dynamic mean field theory and more generally the application of methods from statistical physics to reinforcement learning. As noted in the introduction, more advanced field theory, beyond mean field theory, may require the use of the analytic and approximate forms for the expected maximum surveyed here. Specifically, we anticipate that second and higher order corrections to the “first order" DMFP will require derivatives of the expected maximum Helias and Dahmen (2020).

References

Biyi Afonja. The moments of the maximum of correlated normal and t-variates. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):251–262, 1972.

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, pages 10–4, 2019.

Mauricio Araya, Vincent Thomas, and Olivier Buffet. *Near-Optimal BRL using Optimistic Local Transitions (Extended Version)*. PhD thesis, INRIA, 2012.

John Asmuth, Lihong Li, Michael L Littman, Ali Nouri, and David Wingate. A bayesian sampling approach to exploration in reinforcement learning. *arXiv preprint arXiv:1205.2664*, 2012.

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.

Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664, 2004.

Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 1. Athena scientific, 2012.

Dimitri Bertsekas. *Abstract dynamic programming*. Athena Scientific, 2022.

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.

Daniela Calvetti, G Golub, W Gragg, and Lothar Reichel. Computation of gauss-kronrod quadrature rules. *Mathematics of computation*, 69(231):1035–1052, 2000.

Tommaso Castellani and Andrea Cavagna. Spin-glass theory for pedestrians. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(05):P05012, 2005.

Charles E Clark. The greatest of a finite set of random variables. *Operations Research*, 9(2):145–162, 1961.

DR Cox. The mean and coefficient of variation of range in small samples from non-normal populations. *Biometrika*, 41(3-4):469–481, 1954.

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

Herbert A David. *Order statistics*. John Wiley & Sons, 1981.

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

William Feller. Probability theory and its applications, volume ii, 1971.

William Feller. *An introduction to probability theory and its applications, Volume 2*, volume 81. John Wiley & Sons, 1991.

Ronald Aylmer Fisher and Leonard Henry Caleb Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical proceedings of the Cambridge philosophical society*, volume 24, pages 180–190. Cambridge University Press, 1928.

Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar. Bayesian reinforcement learning: A survey. *arXiv:1609.04436*, 2016.

EJ Gumbel. The maxima of the mean largest value and of the range. *The Annals of Mathematical Statistics*, pages 76–84, 1954.

Shanti S Gupta. Probability integrals of multivariate normal and multivariate t1. *The Annals of mathematical statistics*, pages 792–828, 1963.

Moritz Helias and David Dahmen. *Statistical field theory for neural networks*, volume 970. Springer, 2020.

Lodewijk Kallenberg. Markov decision processes. *Lecture Notes. University of Leiden*, pages 2–5, 2011.

Gautam Kamath. Bounds on the expectation of the maximum of samples from a gaussian. *URL http://www. gautamkamath. com/writings/gaussian max. pdf*, 2015.

Alan P Ker. On the maximum of bivariate normal random variables. *Extremes*, 4(2):185–190, 2001.

J Zico Kolter and Andrew Y Ng. Near-bayesian exploration in polynomial time. In *Proceedings of the 26th annual international conference on machine learning*, pages 513–520, 2009.

Samuel Kotz, Narayanaswamy Balakrishnan, and Norman L Johnson. *Continuous multivariate distributions, Volume 1: Models and applications*, volume 1. John Wiley & Sons, 2004.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms.* Cambridge University Press, 2020.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. *Advances in Neural Information Processing Systems*, 32, 2019.

Patrick Alfred Pierce Moran. *An introduction to probability theory.* Clarendon Press, Oxford, 1968.

Brendan O'Donoghue. Variational bayesian reinforcement learning with regret bounds. *Advances in Neural Information Processing Systems*, 34:28208–28221, 2021.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pages 2701–2710. PMLR, 2017.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. *Advances in neural information processing systems*, 29, 2016.

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.

Andrew M Ross. Useful bounds on the expected maximum of correlated normal variables. *Industrial and Systems Engineering*, 2003.

Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural networks. *Physical review letters*, 61(3):259, 1988.

Jonathan Sorg, Satinder Singh, and Richard L Lewis. Variance-based rewards for approximate bayesian reinforcement learning. *arXiv preprint arXiv:1203.3518*, 2012.

George Stamatescu. Dynamic mean field programming. *arXiv preprint arXiv:2206.05200*, 2022.

Malcolm Strens. A bayesian framework for reinforcement learning. In *ICML*, volume 2000, pages 943–950, 2000.

Alain-Sol Sznitman. Topics in propagation of chaos. *Ecole d'été de probabilités de Saint-Flour XIX—1989*, 1464:165–251, 1991.

Georges M Tallis. The moment generating function of the truncated multi-normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, 23(1):223–229, 1961.

Henri Theil. A note on certainty equivalence in dynamic planning. *Econometrica: Journal of the Econometric Society*, pages 346–349, 1957.

Garrett Thomas. Markov decision processes. *https://ai.stanford.edu/ gwthomas/notes/*, 2007.

Leonard HC Tippett. On the extreme individuals and the range of samples taken from a normal population. *Biometrika*, pages 364–387, 1925.

Richard Von Mises. La distribution de la plus grande de n valuers. *Rev. math. Union interbalcanique*, 1: 141–160, 1936.