Hippocampal Contributions to Control: The Third Way – Supporting Material –

Máté Lengyel
Collegium Budapest Institute for Advanced Study
2 Szentháromság u, Budapest, H-1014, Hungary
and
Computational & Biological Learning Lab
Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ, UK
lmate@gatsby.ucl.ac.uk

Peter Dayan
Gatsby Computational Neuroscience Unit, UCL
17 Queen Square, London WC1N 3AR, UK
dayan@gatsby.ucl.ac.uk

Abstract

Our aim is to derive analytical approximations for the performance of a model-based controller in tree-structured Markov decision processes.

1 Paradigm for analysis

A controller navigating an initially unknown environment (defined as a Markov decision process, MDP) needs to have two component algorithms: a *learning algorithm* that defines a mapping from the history of experienced state-action transitions and obtained rewards to some internal representation (*eg* a value function), and an *action selection algorithm* (policy) that defines a mapping from this internal representation and the current state to an action (or, in the case of stochastic action choice, a distribution over actions).[1] Thus, the performance of a controller depends on how close to optimal its two component algorithms are: how efficiently it uses sample data points when learning, and how accurately it selects actions based on the knowledge it acquired about its environment so far. There is a theoretically optimal, but of course hugely impractical, solution to this problem: a controller which has an explicit statistical model of its environment, learns it by Bayesian updating (assuming it has the correct priors), and then uses this model to exhaustively simulate all possible outcomes of its actions. Here we analyse the performance of such a controller to derive upper bounds on the performance of any other controller, and also investigate how the performance of this model-based controller breaks down if action selection (but not learning!) involves approximations rendering it imperfect – which is one obvious way to approach practicality.

Since our ultimate interest lies in analyzing the trade-offs between multiple alternative controllers that might be available to the same agent, we consider off-policy controllers that are able to learn from experience even if it was gathered while following a policy other than the controller's own.

[1] For a general introduction to the topic and definitions of all terms not defined here see [1]. Direct policy improvement algorithms are also possible and if the authors were hard-pressed those would be considered as a special case of this general scheme, in which the 'policy' that is being improved is the internal representation in our terms, and action selection just looks up the action for the current state.

The model-based controller investigated here is such an off-policy controller. Furthermore, because the problem presented by the exploration-exploitation dilemma is notoriously difficult to solve in a well-founded way, we will focus on exploitation performance. In order to separate the effects of exploitation from those of exploration, the performance of a controller is analysed using a two-stage procedure. First, in the learning stage, the controller learns from experience gathered while following for a number of steps a 'parallel sampler' which guarantees an unbiased coverage of state-action pairs. Second, in the test stage, the controller is allowed to make a number of consecutive action-choices with the objective that it should maximize return in these few steps (finite horizon) using its knowledge about the environment acquired during the learning stage (there is no learning during the test stage).[2]

For mathematical tractability, we analyze the case of discrete-time episodic tasks without temporal discounting in fully observable discrete-state tree-structured MDPs (tMPDs). Let $\mathcal{S}_s$ denote the set of successor states of state s, *ie* states for which there is an action that takes the agent there from state s with some non-zero probability, and $\mathcal{A}_s$ denote the set of available actions at s. TMDPs are defined as MDPs for which, under any policy, the directed graph of the corresponding Markov process is a rooted tree ($|\{s' \colon s \in \mathcal{S}_{s'}\}| = 1$ for all states s, except for the root state for which $\{s' \colon s \in \mathcal{S}_{s'}\} = \emptyset$). In particular, we consider 'regular' tMDPs in which all terminal states $s^\bullet$ ($\mathcal{S}_{s^\bullet} = \emptyset$) have the same distance from the root (the depth of the tMDP, D), all non-terminal states, s°, have the same number of successor states (the branching factor of the tMDP, $B = |\mathcal{S}_{s^\circ}| > 0$ for all s°), and there is an identical number of possible actions available in each of them ($A = |\mathcal{A}_{s^\circ}|$ for all s°).

Furthermore, we assume that the vector of transition probability parameters, $\mathbf{p}^a_{s^\circ}$, for the (multinomial) transition probability distribution $\mathbb{P}[\mathrm{state}(t+1) = s' \in \mathcal{S}_{s^\circ} \mid \mathrm{state}(t) = s^\circ, \mathrm{action}(t) = a] = p^a_{s^\circ s'}$ of each (non-terminal) state-action pair, s° and a, are sampled *iid* from the same Dirichlet distribution $\mathbb{P}[\mathbf{p}^a_{s^\circ}] = \mathrm{Dirichlet}(\mathbf{p}^a_{s^\circ}; \boldsymbol{\alpha})$, and rewards are only available at the B^D terminal states, $s^\bullet$, stochastically from $\mathbb{P}[\mathrm{reward}(t) = r \mid \mathrm{state}(t) = s^\bullet] = \mathcal{N}(r; \bar{r}_{s^\bullet}, 1)$ normal distributions, where the mean reward $\bar{r}_{s^\bullet}$ for each terminal state is sampled *iid* from the same normal distribution $\mathbb{P}[\bar{r}_{s^\bullet}] = \mathcal{N}\big(\bar{r}_{s^\bullet}; \mu_{\bar{r}}, \sigma^2_{\bar{r}}\big)$.

We assume that the hyperparameters, $\Theta = \big\{D, B, A, \boldsymbol{\alpha}, \mu_{\bar{r}}, \sigma^2_{\bar{r}}, \big\}$, and the structure, $\mathcal{S}_{s^\circ}$ and $\mathcal{A}_{s^\circ}$ for all $s^\circ \in \mathcal{T}$, of the tMDP representing the current environment, $\mathcal{T}$, are known *a priori* to the controller, but its actual parameters, $\theta = \Big\{\{\mathbf{p}^a_{s^\circ}\}_{a \in \mathcal{A}_{s^\circ}, s^\circ \in \mathcal{T}}, \{\bar{r}_{s^\bullet}\}_{s^\bullet \in \mathcal{T}}\Big\}$, are unknown. Of particular interest is the performance of each controller as a function of the amount of experience available during the learning stage and the hyperparameters of the tMDP.

2 Exploration by the 'parallel sampler'

In our paradigm, the task of the parallel sampler is to collect experience from which the other controllers, whose exploitation performance is our main interest, can learn. As noted in Section 1, we do not deal with the problem of exploration in any substance here. Our aim is neither to define an optimal exploratory algorithm, nor to make a statement about possible actual exploration strategies used by animals / humans. Rather, we choose an algorithm which has minimal bias to favour any particular exploitation algorithm for a fair comparison, and makes it relatively easy to analyse the effects of learning, similar to [2].

We define the parallel sampler with the objective that in each iteration it should sample all non-terminal state-action pairs n° times (at least on average). Each iteration consists of several trials. On each trial, the controller is started from some non-terminal state s°, which is not necessarily the root of the tMDP, and made to select action $a \in \mathcal{A}_{s^\circ}$. This will take it to another state, and from then on it selects actions randomly until it reaches a terminal state, which ends the trial. The question is how many trials should be started on each iteration from the same state to satisfy our objective? We need to take into account that each state-action pair (but those for the root state) can be visited in two ways: directly, by starting a trial there, and indirectly, when a trial started 'upstream' of it happens to go through it. This means that for each state-action pair at level $d > 0$ (distance from

[2] A corollary of concentrating on exploitation performance is that only deterministic policies need to be considered.

the root state), the number of times a trial should be started from it, p_d, should satisfy the following constraint:

$$n^\circ = \underbrace{p_d}_{\text{direct visits}} + \underbrace{n^\circ\, A\, \frac{1}{B}\, \frac{1}{A}}_{\text{indirect visits}} \tag{1}$$

which readily gives us

$$p_d = \begin{cases} n^\circ & \text{for } d = 0 \\ n^\circ \left(1 - \frac{1}{B}\right) & \text{for } 0 < d < D \end{cases} \tag{2}$$

(We deal with fractional starts by stochastically deciding whether to skip a trial for a state-action pair.)

We also note that this exploration algorithm will sample each terminal state $n^\bullet = \frac{A}{B}\, n^\circ$ times.

We measure 'learning time' in units of iterations of the parallel sampler. Thus, at the end of time step T, each action at non-terminal states and each terminal state has been probed $N^\circ = Tn^\circ$ and $N^\bullet = Tn^\bullet = \frac{A}{B} N^\circ$ times on average. One needs to keep in mind that the computational complexity of a *single* iteration of the parallel sampler scales as $A \cdot \frac{B^{D+1}-1}{B-1}$. Hence, the actual amount of time required to collect the same relative amount of experience (number of samples per state-action pair) will be considerably larger for more complex environments (characterized by larger A, B or D). Our measure of learning time strips away this factor which can always be included as an appropriate rescaling of the time axis.

3 Model-based control

A model-based controller maintains an explicit model of the environment. In order to achieve optimality, this model has to be in the form of a distribution over possible tMDPs. This makes learning maximally data-efficient but, as we will see, makes action choice computationally extremely expensive.

3.1 Learning

Learning in the model-based controller amounts to updating the distribution of possible tMDPs as experience accumulates. Given the prior knowledge of hyperparameters and structure, its task is to infer from experience a posterior distribution over all the parameters θ.

Taking advantage of the Dirichlet distribution being the conjugate prior for the parameters of a multinomial observation distribution, and the normal distributions being the conjugate prior for the mean of a normal observation distribution (with known variance), the posterior distribution over tMDPs can be conveniently parametrised by three sets of parameters: $\{\hat{\boldsymbol{\alpha}}^a_{s^\circ}\}_{a \in \mathcal{A}_{s^\circ}, s^\circ \in \mathcal{T}}$ for transition probabilities $\{\mathbf{p}^a_{s^\circ}\}_{a \in \mathcal{A}_{s^\circ}, s^\circ \in \mathcal{T}}$

$$\mathbb{P}[\hat{\mathbf{p}}^a_{s^\circ}] = \mathrm{Dirichlet}(\hat{\mathbf{p}}^a_{s^\circ}; \hat{\boldsymbol{\alpha}}^a_{s^\circ}) \tag{3a}$$

and $\{\hat{\nu}_{s^\bullet}\}_{s^\bullet \in \mathcal{T}}$ and $\{\hat{\rho}^2_{s^\bullet}\}_{s^\bullet \in \mathcal{T}}$ for mean rewards $\{\bar{r}_{s^\bullet}\}_{s^\bullet \in \mathcal{T}}$

$$\mathbb{P}\left[\hat{\bar{r}}_{s^\bullet}\right] = \mathcal{N}\left(\hat{\bar{r}}_{s^\bullet}; \hat{\mu}_{s^\bullet}, \hat{\rho}^{-2}_{s^\bullet}\right), \text{ where } \hat{\mu}_{s^\bullet} = \hat{\nu}_{s^\bullet}\, \hat{\rho}^{-2}_{s^\bullet} \tag{3b}$$

These posterior parameters are initialised according to their prior counterparts

$$\hat{\boldsymbol{\alpha}}^a_{s^\circ} := \boldsymbol{\alpha} \tag{4a}$$

$$\hat{\nu}_{s^\bullet} := \mu_{\bar{r}}\, \sigma^{-2}_{\bar{r}} \tag{4b}$$

$$\hat{\rho}^2_{s^\bullet} := \sigma^{-2}_{\bar{r}} \tag{4c}$$

and updated whenever state s' is reached from non-terminal state s° by taking action a

$$\hat{\alpha}^a_{s^\circ s'} \rightarrow \hat{\alpha}^a_{s^\circ s'} + 1 \tag{5a}$$

or reward r is experienced in terminal state $s^\bullet$

$$\hat{\nu}_{s^\bullet} \rightarrow \hat{\nu}_{s^\bullet} + r \tag{5b}$$

$$\hat{\rho}^2_{s^\bullet} \rightarrow \hat{\rho}^2_{s^\bullet} + 1 \tag{5c}$$

3.2 Action selection

Given its current knowledge about the environment, represented by the three sets of posterior parameters, $\hat{\Theta} = \left\{ \{\hat{\boldsymbol{\alpha}}^a_{s^\circ}\}_{a \in \mathcal{A}_{s^\circ}, s^\circ \in \mathcal{T}}, \{\hat{\nu}_{s^\bullet}\}_{s^\bullet \in \mathcal{T}}, \{\hat{\rho}^2_{s^\bullet}\}_{s^\bullet \in \mathcal{T}} \right\}$, selecting the optimal action a^* that maximises expected return when in (non-terminal) state s° requires a recursive 'mental simulation' of the consequences of taking each possible action in each state, which is – at least formally – a relatively straight-forward exercise in dynamic programming:[3]

$$a^* = \operatorname*{argmax}_{a \in \mathcal{A}_{s^\circ}} \hat{Q}^a_{s^\circ} \tag{6a}$$

$$\hat{Q}^a_{s^\circ} = \sum_{s' \in \mathcal{S}_{s^\circ}} \bar{p}^a_{s^\circ s'} \hat{V}_{s'}, \text{ where } \bar{p}^a_{s^\circ s'} = \langle \hat{p}^a_{s^\circ s'} \rangle = \frac{\hat{\alpha}^a_{s^\circ s'}}{\sum_{s'' \in \mathcal{S}_{s^\circ}} \hat{\alpha}^a_{s^\circ s''}} \tag{6b}$$

$$\hat{V}_s = \begin{cases} \max_{a \in \mathcal{A}_{s^\circ}} \hat{Q}^a_{s^\circ} & \text{if } s \text{ is non-terminal} \\ \hat{\mu}_{s^\bullet} & \text{if } s \text{ is terminal} \end{cases} \tag{6c}$$

Straight-forward this procedure may seem, the recursion implied by Equations 6b-6c means that the computational complexity of optimal (exploitative) action selection is immense: it is potentially exponential in D.

4 Performance analysis

4.1 Asymptotic performance

We are interested in the expected return harvested by the model-based controller averaged over a distribution of possible tMDPs defined by hyperparameters Θ.

As a first step, we calculate the asymptotic performance when infinitely many data have been collected so that $\mathbb{P}[\hat{\mathbf{p}}^a_{s^\circ}] \to \delta(\hat{\mathbf{p}}^a_{s^\circ} - \mathbf{p}^a_{s^\circ})$ and $\mathbb{P}\left[\hat{\bar{r}}_{s^\bullet}\right] \to \delta\left(\hat{\bar{r}}_{s^\bullet} - \bar{r}_{s^\bullet}\right)$. Therefore, in this limit, transition probabilities $\mathbf{p}^a_{s^\circ}$ and mean rewards $\bar{r}_{s^\bullet}$ can be treated as known quantities, and the expected return for starting from state s is given simply by its value, V_s, which can be computed using direct analogues of Equations 6b-6c:

$$V_s = \begin{cases} \max_{a \in \mathcal{A}_{s^\circ}} Q^a_{s^\circ} & \text{if } s \text{ is non-terminal} \\ \bar{r}_{s^\bullet} & \text{if } s \text{ is terminal} \end{cases} \tag{7a}$$

$$Q^a_{s^\circ} = \sum_{s' \in \mathcal{S}_{s^\circ}} p^a_{s^\circ s'} V_{s'} \tag{7b}$$

In order to compute the *average* return, we first compute the distribution of this value given the hyperparameters by marginalizing over possible tMDP instances defined by θ

$$\mathbb{P}[V_s \mid \Theta] = \int \mathrm{d}\theta \, \mathbb{P}[V_s \mid \theta, \Theta] \, \mathbb{P}[\theta \mid \Theta] \tag{8}$$

where the first term is a delta distribution concentrated on the value computed in Equations 7a-7b.

If s is terminal than this distribution is simply

$$\mathbb{P}[V_{s^\bullet} \mid \Theta] = \mathcal{N}\left(V_{s^\bullet}; \mu_{\bar{r}}, \sigma^2_{\bar{r}}\right) \tag{9}$$

[3] In principle, optimal action selection should also take advantage of the internal representation of the environment being fully probabilistic, and take into account the *uncertainties* associated with the values, in the form of computing some sort of 'exploration bonuses' [3] in each step of the recursion in Equations 6b-6c, or – to be strictly optimal – by performing dynamic programming on the corresponding belief-state MDP. However, in the case of concentrating exclusively on exploitation – which we study –, only the expected values matter.

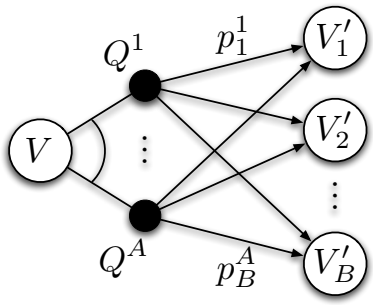

Figure 1: A sub-treelet of a regular tMDP, in which the value distribution of the root state can be computed by Equations 12a-12b.

In the case when s is non-terminal, we can take advantage of the Markov property of values implied by the recursive nature of Equations 7a-7b and that parameters are *iid*, and write

$$\mathbb{P}[V_{s^\circ} \mid \Theta] = \int \mathrm{d}\mathbf{P}_{s^\circ}\, \mathrm{d}\left\{V_{s'}\right\}_{s' \in \mathcal{S}_{s^\circ}} \mathbb{P}\Big[V_{s^\circ} \mid \mathbf{P}_{s^\circ}, \left\{V_{s'}\right\}_{s' \in \mathcal{S}_{s^\circ}}\Big] \prod_{a \in \mathcal{A}_{s^\circ}} \mathrm{Dirichlet}(\mathbf{p}_{s^\circ}^a; \boldsymbol{\alpha}) \prod_{s' \in \mathcal{S}_{s^\circ}} \mathbb{P}[V_{s'} \mid \Theta] \tag{10}$$

where $\mathbf{P}_{s^\circ} = \left\{\mathbf{p}_{s^\circ}^a\right\}_{a \in \mathcal{A}_{s^\circ}}$ contains the transition probability parameters for all actions available at s°.

Equation 10 gives a recursive definition of value distributions in a tMDP that closely parallels the recursion in Equations 7a-7b. The computational complexity of this recursion is still exponential in D. However, in regular tMDPs, all successor states of the same state have the same value distribution. More generally, it is easy to see that all states $s^{(d)}$ at the same level d in a regular tMDP will have the same distribution because they are also at the same distance from the terminal states, and their subtrees (of which they are the root) are identical in their structure and parameter priors. By defining $\mathrm{P}_{\mathrm{V}}^{(d)}(V) := \mathbb{P}[V_{s^{(d)}} = V \mid \Theta]$, this allows us to rewrite Equations 9-10 more conveniently:

$$\mathrm{P}_{\mathrm{V}}^{(d)}(V) = \begin{cases} \displaystyle\int \quad \mathrm{d}\mathbf{p}^1 \ldots \mathrm{d}\mathbf{p}^A\, \mathrm{d}V_1' \ldots \mathrm{d}V_B'\, \mathbb{P}\big[V \mid \mathbf{p}^1 \ldots \mathbf{p}^A, V_1' \ldots V_B'\big] \cdot \\ \qquad \cdot \displaystyle\prod_{a=1}^{A} \mathrm{Dirichlet}(\mathbf{p}^a; \boldsymbol{\alpha}) \prod_{b=1}^{B} \mathrm{P}_{\mathrm{V}}^{(d+1)}(V_b') & \text{for } d < D \quad \text{(11a)} \\ \mathcal{N}\big(V; \mu_{\bar{r}}, \sigma_{\bar{r}}^2\big) & \text{for } d = D \quad \text{(11b)} \end{cases}$$

Thus, it is possible to compute the value distribution of a state at a given level in the tMDP iteratively, by proceeding 'backwards' from terminal states ($d = D$) to the state (level) in question. Note that the complexity of this computation is only *linear* in D. Although this is already a substantial improvement, the integrals in Equation 11a will generally be analytically intractable, so the complexity of computing them numerically is still exponential in A and B (at worst).

We now rewrite the value distribution of a state at a non-terminal level $d < D$ (Eq. 11a) making use of action values (Q) defined in Equation 7b and the Markov properties within a 'sub-treelet' of a tMDP upon which the state value in question depends (see Fig. 1):

$$\mathrm{P}_{\mathrm{V}}^{(d)}(V) = \int \mathrm{d}Q^1 \ldots \mathrm{d}Q^A\, \mathbb{P}\big[V \mid Q^1 \ldots Q^A\big]\ \mathrm{P}_{\mathrm{Q}}^{(d)}\big(Q^1 \ldots Q^A\big) \tag{12a}$$

$$\mathrm{P}_{\mathrm{Q}}^{(d)}\big(Q^1 \ldots Q^A\big) = \int \mathrm{d}\mathbf{p}^1 \ldots \mathrm{d}\mathbf{p}^A \int \mathrm{d}V_1' \ldots \mathrm{d}V_B' \prod_{a=1}^{A} \mathrm{Dirichlet}(\mathbf{p}^a; \boldsymbol{\alpha}) \prod_{b=1}^{B} \mathrm{P}_{\mathrm{V}}^{(d+1)}(V_b')\ \mathbb{P}\big[Q^1 \ldots Q^A \mid \mathbf{p}^1 \ldots \mathbf{p}^A, V_1' \ldots V_B'\big] \tag{12b}$$

For analytical tractability, we approximate $\mathrm{P}_{\mathrm{Q}}^{(d)}\big(Q^1 \dots Q^A\big)$ with a multivariate normal distribution

$$\mathrm{P}_{\mathrm{Q}}^{(d)}\big(Q^1 \dots Q^A\big) \simeq \mathcal{N}\Big(Q^1 \dots Q^A; \boldsymbol{\nu}^{(d)}, \boldsymbol{\Xi}^{(d)}\Big) \tag{13a}$$

where we need to determine the elements of the mean vector, $\boldsymbol{\nu}^{(d)}$, and covariance matrix, $\boldsymbol{\Xi}^{(d)}$. We know from Equation 7b that $\mathbb{P}[Q^a \mid \mathbf{p}^a, V_1' \dots V_B']$ in fact expresses each Q^a as a *deterministic* sum of successor state values weighted by transition probabilities, and so Equation 12b entails the averaging of this weighted sum over a distribution of each term in the sum. This argument readily gives us

$$\nu_a^{(d)} = \langle Q_a \rangle = \mu_{V_{d+1}} \tag{13b}$$

$$\Xi_{aa'}^{(d)} = \langle Q_a Q_{a'} \rangle - \langle Q_a \rangle \langle Q_{a'} \rangle = \begin{cases} \sigma_{V_{d+1}}^2 \, \xi_0 & \text{for } a = a' \\ \sigma_{V_{d+1}}^2 \, \xi_1 & \text{otherwise} \end{cases} \tag{13c}$$

$$\text{with} \quad \xi_0 = \sum_b \frac{\alpha_b^2 + \alpha_b}{\alpha_0^2 + \alpha_0}, \quad \xi_1 = \sum_b \frac{\alpha_b^2}{\alpha_0^2}, \quad \text{and} \quad \alpha_0 = \sum_b \alpha_b \tag{13d}$$

where μ_{V_d} and $\sigma_{V_d}^2$ are the mean and variance of $\mathrm{P}_{\mathrm{V}}^{(d)}(V)$, respectively.

Finally, since $\mathbb{P}\big[V \mid Q^1 \dots Q^A\big]$ in Equation 12a simply entails choosing *deterministically* the maximum of $Q^1 \dots Q^A$ (Eq. 7a), we can also observe that $\mathrm{P}_{\mathrm{V}}^{(d)}(V)$ just expresses the Ath order statistic of $\mathrm{P}_{\mathrm{Q}}^{(d)}\big(Q^1 \dots Q^A\big)$, which we also approximate with a normal distribution

$$\mathrm{P}_{\mathrm{V}}^{(d)}(V) = \mathbb{P}_{\mathrm{OS}}\Big[V; \mathrm{P}_{\mathrm{Q}}^{(d)}\big(Q^1 \dots Q^A\big), A\Big] \simeq \mathcal{N}\big(V; \mu_{V_d}, \sigma_{V_d}^2\big) \tag{14}$$

whose mean, μ_{V_d} and variance, $\sigma_{V_d}^2$, we need to determine.

The order statistics of *uncorrelated* variables can be conveniently expressed through a Beta distribution, so that the distribution of $Z_A^{(k)}$, the kth order statistic of A *iid* standard normal variates, $W_1 \dots W_A$, takes the form:

$$\mathbb{P}_{\mathrm{OS}}\left[Z_A^{(k)}; \prod_{i=1}^{A} \mathcal{N}^*(W_i), k\right] = \mathcal{N}^*\Big(Z_A^{(k)}\Big) \; \mathrm{Beta}\Big(\Psi^*\Big(Z_A^{(k)}\Big); k, A + 1 - k\Big) \tag{15}$$

where $\mathcal{N}^*$ and Ψ^* are the p.d.f. and c.d.f. of the standard normal distribution, and $k = A$ in our case because we are interested in the distribution of the maximum.

Using this formula and slightly extending the results of [4] on the moments of order statistics from the *equicorrelated* multivariate normal distribution (which is the case of $\mathrm{P}_{\mathrm{Q}}^{(d)}\big(Q^1 \dots Q^A\big)$, as can be seen from Eq. 13c), the mean and variance of $\mathrm{P}_{\mathrm{V}}^{(d)}(V)$ are

$$\mu_{V_d} = \mu_{V_{d+1}} + \lambda_1 \, \sigma_{V_{d+1}} \tag{16a}$$

$$\sigma_{V_d}^2 = \lambda_2 \, \sigma_{V_{d+1}}^2 \tag{16b}$$

$$\text{with} \quad \lambda_1 = (\xi_0 - \xi_1)^{1/2} \; \overline{Z_A^1}, \quad \lambda_2 = \xi_1 + (\xi_0 - \xi_1)\left(\overline{Z_A^2} - \overline{Z_A^1}^2\right), \quad \text{and} \tag{16c}$$

$$\overline{Z_A^i} = \int \mathrm{d}Z \, Z^i \, \mathcal{N}^*(Z) \; \mathrm{Beta}(\Psi^*(Z); A, 1) \tag{16d}$$

which can be re-expressed in closed form, without any recursion, as

$$\mu_{V_d} = \mu_{\bar{r}} + \frac{1-\lambda_2^{(D-d)/2}}{1-\lambda_2^{1/2}}\,\lambda_1\,\sigma_{\bar{r}} \tag{17a}$$

$$\sigma^2_{V_d} = \lambda_2^{(D-d)}\,\sigma^2_{\bar{r}} \tag{17b}$$

This completes our derivation: the average expected return of the model-based controller in tMDPs of hyperparameters Θ is simply given by substituting $d = 0$ to Equation 17a (and as a bonus, Eq. 17b also gives us the variance of the expected return).

4.2 Effects of computational noise

As we noted in Section 3.2, the computational complexity of optimal action selection is immense. In any practical case, it will be so overwhelming that *approximations* will be required while computing the values of different options, such as pruning (*ie* ignoring parts of the tree), or sampling, etc. Clearly, these approximations will have adverse effects on the performance of the controller.

We treat the effects of all sort of possible approximations as if the action selection algorithm of Equations 6a-6c was modified so that instead of having access to the true values of available actions, Q, the controller would only have access, and thus base its decision on, their *noisy* versions, Q':

$$a^*_{s^\circ} = \underset{a\in\mathcal{A}_{s^\circ}}{\operatorname{argmax}}\, Q'^a_{s^\circ} \tag{18a}$$

$$Q'^a_{s^\circ} = \eta_1 Q^a_{s^\circ} + \eta_2 z, \quad \text{where } \mathbb{P}[z] = \mathcal{N}(z;0,1) \tag{18b}$$

$$Q^a_{s^\circ} = \sum_{s'\in\mathcal{S}_{s^\circ}} \bar{p}^a_{s^\circ,s'} V_{s'} \tag{18c}$$

$$V_s = \begin{cases} Q^{a^*_{s^\circ}}_{s^\circ} & \text{if } s \text{ is non-terminal} \\ \hat{\mu}_{s^\bullet} & \text{if } s \text{ is terminal} \end{cases} \tag{18d}$$

This means that the achievable value of a state is not simply the maximum of the values of the actions available in it. Rather, it will be the true value of the action that happens to have the greatest noisy value. Thus, the state value distribution of a state will no longer be the maximal order statistic distribution of the corresponding action value distributions as in Equation 14, but will be something slightly more complicated:

$$\mathrm{P}^{(d)}_{\mathrm{V}}(V) = A\;\mathbb{P}\big[Q^1 = V\big]\;\mathbb{P}\big[Q'^1 \text{ is greater than any of } Q'^2 \dots Q'^A \mid Q^1 = V\big] \tag{19}$$

where the middle term on the RHS can be expressed directly from Equations 13a-13c as

$$\mathbb{P}\big[Q^1 = V\big] = \mathrm{P}^{(d)}_{\mathrm{Q}}(V) \simeq \mathcal{N}\Big(V; \mu_{V_{d+1}}, \xi_0\,\sigma^2_{V_{d+1}}\Big) \tag{20}$$

and the last term as

$$\mathbb{P}\big[Q'^1 \text{ is greater than any of } Q'^2 \dots Q'^A \mid Q^1 = V\big] = \\ = \int \mathrm{d}Q'^1\,\mathbb{P}\big[Q'^1 \mid Q^1 = V\big] \int_{-\infty}^{Q'^1} \mathrm{d}Q^*\,\mathbb{P}_{\mathrm{OS}}\big[Q^*; \mathbb{P}\big[Q'^2 \dots Q'^A \mid Q^1 = V\big], A-1\big] \tag{21}$$

Since we have already approximated the distribution of true action values with a normal in Equation 13a, it is fair to use a similar normal approximation for the distribution of noisy action values, where

$$\mathbb{P}\big[Q'^1 \mid Q^1 = V\big] \simeq \mathcal{N}\big(Q'^1; \eta_1\,V, \eta_2^2\big) \tag{22}$$

and

$$\mathbb{P}\big[Q'^2 \dots Q'^A \mid Q^1 = V\big] = \int \mathrm{d}Q^2 \dots \mathrm{d}Q^A\,\mathbb{P}\big[Q'^2 \dots Q'^A \mid Q^2 \dots Q^A\big]\;\mathbb{P}\big[Q^2 \dots Q^A \mid Q^1 = V\big] \\ \simeq \mathcal{N}\Big(Q'^2 \dots Q'^A; \boldsymbol{\nu}'^{(d)}, \boldsymbol{\Xi}'^{(d)}\Big) \tag{23a}$$

is a multivariate normal with mean

$$\nu_a'^{(d)} = \eta_1 \left\langle Q^{a+1} \mid Q^1 = V \right\rangle \tag{23b}$$

and (co)variance

$$\Xi_{aa'}'^{(d)} = \begin{cases} \eta_1^2 \operatorname{Var}\left[Q^{a+1} \mid Q^1 = V\right] + \eta_2^2 & \text{for } a = a' \\ \eta_1^2 \operatorname{Cov}\left[Q^{a+1}, Q^{a'+1} \mid Q^1 = V\right] & \text{otherwise} \end{cases} \tag{23c}$$

where, from Equations 13a-13c,

$$\left\langle Q^a \mid Q^1 = V \right\rangle = \mu_{V_{d+1}} + \gamma \left(V - \mu_{V_{d+1}}\right) \tag{24a}$$

$$\text{with} \quad \gamma = \frac{(A-1)\,\frac{\xi_1}{\xi_0} + 1}{\frac{\xi_0}{\xi_1} + A - 1} \tag{24b}$$

$$\operatorname{Var}\left[Q^a \mid Q^1 = V\right] = \sigma^2_{V_{d+1}} \left[\xi_0 - \frac{\xi_1^2}{\xi_0}\right] \tag{24c}$$

$$\operatorname{Cov}\left[Q^a, Q^{a'} \mid Q^1 = V\right] = \sigma^2_{V_{d+1}} \left[\xi_1 - \frac{\xi_1^2}{\xi_0}\right] \tag{24d}$$

Thus, $\mathbb{P}\left[Q'^2 \dots Q'^A \mid Q^1 = V\right]$ is equicorrelated, and so the first two moments of its maximal order statistic can be computed in the same way we obtained Equations 16a-16b, and these in turn can be used for a Gaussian approximation

$$\mathbb{P}_{\mathrm{OS}}\left[Q^*; \mathbb{P}\left[Q'^2 \dots Q'^A \mid Q^1 = V\right], A-1\right] \simeq \mathcal{N}\Big(Q^*; \mu^*, {\sigma^*}^2\Big) \tag{25a}$$

with mean and variance

$$\mu^* = \eta_1 \left(\mu_{V_{d+1}} + \gamma \left(V - \mu_{V_{d+1}}\right)\right) + \lambda_1' \, \sigma_{V_{d+1}} + \lambda_1'' \tag{25b}$$

$$\sigma^{*2} = \lambda_2' \, \sigma^2_{V_{d+1}} + \lambda_2'' \tag{25c}$$

with

$$\lambda_1' = \eta_1 \left(\xi_0 - \xi_1\right)^{1/2} \overline{Z^1_{A-1}}, \quad \lambda_1'' = \eta_2 \, \overline{Z^1_{A-1}}, \tag{25d}$$

$$\begin{aligned} \lambda_2' &= \eta_1^2 \left[\left(\xi_1 - \tfrac{\xi_1^2}{\xi_0}\right) + \left(\xi_0 - \xi_1\right)\left(\overline{Z^2_{A-1}} - \overline{Z^1_{A-1}}^2\right)\right], \text{ and} \\ \lambda_2'' &= \eta_2^2 \left(\overline{Z^2_{A-1}} - \overline{Z^1_{A-1}}^2\right) \end{aligned} \tag{25e}$$

As a consequence, the second integral in Equation 21 can be simply expressed with the c.d.f. of the standard normal:

$$\int_{-\infty}^{Q'^1} \mathrm{d}Q^* \, \mathbb{P}_{\mathrm{OS}}\left[Q^*; \mathbb{P}\left[Q'^2 \dots Q'^A \mid Q^1 = V\right], A-1\right] \simeq \Psi^*\left(\frac{Q'^1 - \mu^*}{\sigma^*}\right) \tag{26}$$

Thus, Equation 21 entails the convolution of a normal p.d.f with a normal c.d.f. Unfortunately, this does not have a convenient analytical form. We could integrate it numerically, but then it would be hard to work out the dependence of this integral on V, which we need in order to be able to have an expression for Equation 19. As a compromise, we approximate Ψ by two Gaussian tails (one of them upside-down) glued together at their inflection points. Since the mean of the normal c.d.f., μ^* (from Eq. 25b), is *usually* larger than the mean of the normal p.d.f, $\eta_1 \, V$ (from Eq. 22), the contribution of the second half of the normal c.d.f. to the integral will be negligible, and so we finally approximate the normal c.d.f. rather crudely with a single (unnormalized) Gaussian that we obtain by Taylor expanding its first half in the log domain around 0:

$$\Psi^*(x) \approx \frac{1}{2} e^{\sqrt{\frac{2}{\pi}} x - \frac{1}{\pi} x^2} \text{ for } x \leq 0 \tag{27}$$

This approximation allows us to rewrite the integrand of the outer integral in Equation 21 as a single unnormalized Gaussian (the product of two Gaussians), which is further simplified by the observation that the integral is invariant under a change of variables $Q'^1 \rightarrow Q'^1 + \eta_1 V$:

$$\mathbb{P}\big[Q'^1 \text{ is greater than any of } Q'^2 \dots Q'^A \mid Q^1 = V\big] \simeq$$

$$\simeq \underbrace{\int \mathrm{d}Q'^1 \mathcal{N}\Big(Q'^1; \frac{\kappa_1}{2\kappa_2}, \frac{1}{2\kappa_2}\Big)}_{=1} \sqrt{\frac{1}{2\kappa_2 \eta_2^2}} \frac{1}{2} e^{\frac{\kappa_1^2}{4\kappa_2} - \kappa_0} \tag{28a}$$

with

$$\kappa_0 = \sqrt{\frac{2}{\pi}} \frac{\mu^* - \eta_1 V}{\sigma^*} + \frac{1}{\pi} \frac{(\mu^* - \eta_1 V)^2}{\sigma^{*2}} \tag{28b}$$

$$\kappa_1 = \sqrt{\frac{2}{\pi}} \frac{1}{\sigma^*} + \frac{2}{\pi} \frac{\mu^* - \eta_1 V}{\sigma^{*2}} \tag{28c}$$

$$\kappa_2 = \frac{1}{2\eta_2^2} + \frac{1}{\pi} \frac{1}{\sigma^{*2}} \tag{28d}$$

Concentrating only on the last term, which is the only one that depends on V, since the rest will be lumped into the normalization factor of Equation 19, we can write that

$$\mathbb{P}\big[Q'^1 \text{ is greater than any of } Q'^2 \dots Q'^A \mid Q^1 = V\big] \propto e^{\frac{\kappa_1^2}{4\kappa_2} - \kappa_0} \propto \mathcal{N}\Big(V; \mu'_V, {\sigma'}^2_V\Big) \tag{29a}$$

where

$$\mu'_V = \frac{1}{\kappa''_1} \left(\sqrt{\frac{\pi}{2}} \sigma^* + \kappa''_0 \right) \tag{29b}$$

$${\sigma'}^2_V = \frac{1}{2 {\kappa''_1}^2 \kappa'_2} \tag{29c}$$

with

$$\kappa'_2 = \frac{1}{\pi\, \sigma^{*2}} \left(1 - \frac{1}{\pi\, \kappa_2 \sigma^{*2}} \right) \tag{29d}$$

$$\kappa''_0 = \eta_1 (1 - \gamma)\, \mu_{V_{d+1}} + \lambda'_1\, \sigma_{V_{d+1}} + \lambda''_1 \tag{29e}$$

$$\kappa''_1 = \eta_1 (1 - \gamma) \tag{29f}$$

Finally, we observe that with the above approximations $\mathrm{P}_{\mathrm{V}}^{(d)}(V)$ (Eq. 19) is just the renormalized product of two Gaussians (one defined in Eq. 20, the other in Eq. 29a), and thus can be written as

$$\mathrm{P}_{\mathrm{V}}^{(d)}(V) \simeq \mathcal{N}\big(V; \mu_{V_d}, \sigma^2_{V_d}\big) \tag{30a}$$

where

$$\mu_{V_d} = \frac{\frac{\mu_{V_{d+1}}}{\xi_0\, \sigma^2_{V_{d+1}}} + \frac{\mu'_V}{{\sigma'}^2_V}}{\frac{1}{\sigma^2_{V_d}}} \tag{30b}$$

$$\sigma^2_{V_d} = \frac{1}{\frac{1}{\xi_0\, \sigma^2_{V_{d+1}}} + \frac{1}{{\sigma'}^2_V}} \tag{30c}$$

Equations 13d, 16d, 24b, 25c-25e, 28d, 29b-29f, 30b-30c jointly define a recursive relationship between the mean and variance of $\mathrm{P}_{\mathrm{V}}^{(d)}(V)$ and $\mathrm{P}_{\mathrm{V}}^{(d+1)}(V)$. Unfortunately, there is no obvious way to

rewrite it in closed form without recursion, as we could do for the asymptotic case in Equations 17a-17b. Another somewhat disappointing consequence of the numerous, and sometimes rather crude, approximations that we made along the way is that Equations 16a-16b describing the asymptotic case do not seem to be the limiting cases of Equations 30b-30c with computational noise $\eta_2 \to 0$, as they should. However, numerical simulations show that the two solutions do become approximately equal. Fortunately, in the noise-dominated limit, $\eta_1 \to 0$, Equations 30b-30c behave appropriately, as they yield $\mu_{V_d} \to \mu_{V_{d+1}}$ and $\sigma^2_{V_d} \to \xi_0 \sigma^2_{V_{d+1}}$ (because $\kappa''_1 \to 0$, and thus $\mu'_V / {\sigma'_V}^2$ and $1/{\sigma'_V}^2 \to 0$), *ie* equivalence with a random controller.

4.3 Learning curve

Up to this point, we assumed that the controller had perfect knowledge of the environment, more precisely its parameters, θ. In this section, we analyse the effects of incomplete information due to limited experience. The ansatz of our approach is that these effects can be modeled by assuming that the controller does not have access to the true values of actions, for which it would need perfect knowledge of all transition and reward probabilities, only to their 'noisy' versions. This 'noise' comes from the fact that action values are based on *estimates* of transition probabilities, $\hat{\mathbf{p}}^a_{s^\circ}$ (Eq. 6b), and mean rewards, $\hat{\mu}_{s^\bullet}$ (Eq. 6c). These estimates are inherently stochastic themselves as they are based on stochastic experience (Eqs. 5a-5c). As the final step, we show that the form of the resulting 'noise' in the action values can be well approximated by the form of computational noise we have already treated in Section 4.2, and so we only need to determine how the amount of learning experience, expressed by N° and $N^\bullet$ (defined in Section 2), maps on to the parameters of effective computational noise, η_1 and η_2 (Eq. 18b).

For a 'non-*pre*terminal' state, s°, which is at depth $d < D-1$, the only source of stochasticity in action value estimates that is not due to propagated effects of 'noise' in the action value estimates of its successor states is the error in the estimate of transition probabilities. The learning rules of Equations 4a and 5a result in $\hat{\boldsymbol{\alpha}}^a_{s^\circ} = \boldsymbol{\alpha} + \sum_{t=1}^{T} \mathbf{x}_t$ where $\mathbf{x}_t$ is a vector of binary variables indicating which state s' was reached in learning time step t when action a was taken in state s°. Since estimated action values are based on estimated transition probabilities $\bar{\mathbf{p}}^a_{s^\circ} = \frac{\hat{\boldsymbol{\alpha}}^a_{s^\circ}}{\alpha_0 + N^\circ}$ (Eq. 6b), we are interested in the distribution of these estimates after N° learning experience with non-terminal states given the true transition probabilities, $\mathbb{P}[\bar{\mathbf{p}}^a_{s^\circ} \mid \mathbf{p}^a_{s^\circ}, N^\circ, \Theta]$ (and marginalizing over possible sequences of $\mathbf{x}_t$), and in particular the mean and (co)variance of this distribution

$$\langle \bar{p}^a_{s^\circ b} \mid \mathbf{p}^a_{s^\circ}, N^\circ, \Theta \rangle = \frac{\alpha_b + N^\circ \cdot p^a_{s^\circ b}}{\alpha_0 + N^\circ} \tag{31a}$$

$$\mathrm{Var}[\bar{p}^a_{s^\circ b} \mid \mathbf{p}^a_{s^\circ}, N^\circ, \Theta] = \frac{N^\circ \cdot p^a_{s^\circ b}\,(1 - p^a_{s^\circ b})}{(\alpha_0 + N^\circ)^2} \tag{31b}$$

$$\mathrm{Cov}[\bar{p}^a_{s^\circ b}, \bar{p}^a_{s^\circ b'} \mid \mathbf{p}^a_{s^\circ}, N^\circ, \Theta] = -\frac{N^\circ \cdot p^a_{s^\circ b}\, p^a_{s^\circ b'}}{(\alpha_0 + N^\circ)^2} \tag{31c}$$

Estimated action values, $\hat{Q}^a_{s^\circ}$, are just a linear combination of estimated transition probabilities with (estimated) successor state values ($\hat{V}_{s'}$ where $s' \in \mathcal{S}_{s^\circ}$, Eq. 6b), and so the mean and variance of the distribution of estimated action values, $\mathbb{P}\left[\hat{Q}^a_{s^\circ} \mid \mathbf{p}^a_{s^\circ}, \{V'_{s'}\}_{s' \in \mathcal{S}_{s^\circ}}, N^\circ, \Theta\right]$, are

$$\left\langle \hat{Q}^a_{s^\circ} \mid \mathbf{p}^a_{s^\circ}, \{V'_{s'}\}_{s' \in \mathcal{S}_{s^\circ}}, N^\circ, \Theta \right\rangle = \sum_b \frac{\alpha_b}{\alpha_0 + N^\circ} V'_b + \underbrace{\frac{N^\circ}{\alpha_0 + N^\circ}}_{\eta_1} \overbrace{\sum_b p^a_{s^\circ b} V'_b}^{Q^a_{s^\circ}} \tag{32a}$$

$$\mathrm{Var}\left[\hat{Q}^a_{s^\circ} \mid \mathbf{p}^a_{s^\circ}, \{V'_{s'}\}_{s' \in \mathcal{S}_{s^\circ}}, N^\circ, \Theta\right] = \frac{N^\circ}{(\alpha_0 + N^\circ)^2} \left(\sum_b p^a_{s^\circ b} V'^2_b - \overbrace{\sum_b \sum_{b'} p^a_{s^\circ b} p^a_{s^\circ b'} V'^2_b V'^2_{b'}}^{{Q^a_{s^\circ}}^2} \right) \tag{32b}$$

Thus, the mean of the estimated action value (Eq. 32a) is the sum of a constant term, which is independent of the action, and a term that is proportional to the true action value. When we analysed the effects of noise in Section 4.2, this was exactly our assumption about the mean of noisy action values (Eq. 18b), and the constant of proportionality for the action value-dependent term was η_1. (We ignored the constant term because it made no difference to any of the results.) However, the form of the variance of estimated action values (Eq. 32b) does not seem to be compatible with what we assumed about noisy action values: rather than being an action-independent constant η_2^2, the term in the bracket does depend on the action. In order to reconcile this discrepancy, we approximate this term with its expected value under our generative distribution for tMDPs, and obtain

$$\mathrm{Var}\Big[\hat{Q}^a_{s^\circ} \mid \mathbf{p}^a_{s^\circ}, \{V'_{s'}\}_{s'\in\mathcal{S}_{s^\circ}}, N^\circ, \Theta\Big] \simeq \Big\langle \mathrm{Var}\Big[\hat{Q}^a\Big] \mid N^\circ, \Theta\Big\rangle = \underbrace{\frac{N^\circ}{(\alpha_0+N^\circ)^2}\,(1-\xi_0)\,\sigma^2_{V_{d+1}}}_{\eta_2^2} \tag{32c}$$

Numerical simulation shows that the inaccuracy introduced by this approximation is tolerable. Furthermore, although we did not explicitly allow for depth-dependence of η_2 (such as that in the last $\sigma^2_{V_{d+1}}$ term of Eq. 32c) when treating the case of noisy action values, it makes no changes to the results derived therein. We also note that, conveniently, the evolution of η_1 and η_2 with learning (*ie* as a function of N°) follows intuition: the former grows from 0 to 1, while the latter starts at 0 (because all estimates are concentrated on the prior parameters), then attains positive values, but returns ultimately again to 0 (as all estimates become increasingly concentrated on the corresponding true action values).

The case of 'preterminal' states, $s^\odot$, those that are at depth $d = D-1$, is slightly more complicated. Since we want to lump all the estimation noise into noisy action values, this is where we need to take into account the uncertainty that the controller has during learning about the mean reward probabilities of its terminal states, $\bar{r}_{s^\bullet}$. The learning rules for this estimate (Eqs. 4b-4c, and 5b-5c) yield $\hat{\mu}_{s^\bullet} = \frac{\frac{1}{\sigma^2_{\bar{r}}}\mu_{\bar{r}}+\sum_t r_t}{\frac{1}{\sigma^2_{\bar{r}}}+N^\bullet}$, after $N^\bullet$ learning experience with terminal states, where r_t is the reward harvested in this state in learning time step t. Thus, the posterior for this estimate, $\mathbb{P}[\hat{\mu}_{s^\bullet} \mid \bar{r}_{s^\bullet}, N^\bullet, \Theta]$ (after marginalizing over possible experienced reward sequences), happens to be truly normal and its mean and variance are

$$\langle \hat{\mu}_{s^\bullet} \mid \bar{r}_{s^\bullet}, N^\bullet, \Theta\rangle = \frac{\frac{1}{\sigma^2_{\bar{r}}}\mu_{\bar{r}} + N^\bullet\,\bar{r}_{s^\bullet}}{\frac{1}{\sigma^2_{\bar{r}}}+N^\bullet} \tag{33a}$$

$$\mathrm{Var}[\hat{\mu}_{s^\bullet} \mid \bar{r}_{s^\bullet}, N^\bullet, \Theta] = \frac{N^\bullet}{\Big(\frac{1}{\sigma^2_{\bar{r}}}+N^\bullet\Big)^2} \tag{33b}$$

The noise due to uncertainty about terminal state mean rewards (Eqs. 33a-33b) needs to be combined with that due to stochastic transition probability estimates at preterminal states (Eqs 31a-31c) in order to obtain an expression for the mean and variance of action value estimates of preterminal states

$$\Big\langle \hat{Q}^a_{s^\odot} \mid \mathbf{p}^a_{s^\odot}, \{\bar{r}_{s^\bullet}\}_{s^\bullet\in\mathcal{S}_{s^\odot}}, N^\circ, N^\bullet, \Theta\Big\rangle = \frac{\frac{1}{\sigma^2_{\bar{r}}}\mu_{\bar{r}}}{\frac{1}{\sigma^2_{\bar{r}}}+N^\bullet} + \frac{N^\bullet}{\frac{1}{\sigma^2_{\bar{r}}}+N^\bullet}\sum_b \frac{\alpha_b}{\alpha_0+N^\circ}\bar{r}_b + \underbrace{\frac{N^\circ}{\alpha_0+N^\circ}\,\frac{N^\bullet}{\frac{1}{\sigma^2_{\bar{r}}}+N^\bullet}}_{\eta_1}\overbrace{\sum_b p^a_{s^\odot b}\,\bar{r}_b}^{Q^a_{s^\circ}} \tag{34a}$$

$$
\begin{aligned}
&\mathrm{Var}\left[\hat{Q}^a_{s^\odot} \mid \mathbf{p}^a_{s^\odot}, \{\bar{r}_{s^\bullet}\}_{s^\bullet \in \mathcal{S}_{s^\odot}}, N^\circ, N^\bullet, \Theta\right] \simeq \\
&\quad \simeq \quad \left\langle \mathrm{Var}\left[\hat{Q}^a_{s^\odot}\right] \mid N^\circ, N^\bullet, \Theta \right\rangle \\
&\quad = \quad \frac{N^\bullet}{\left(\frac{1}{\sigma^2_{\bar{r}}} + N^\bullet\right)^2} \frac{1}{\alpha_0 + N^\circ} \frac{1}{\alpha_0 + 1} \cdot \\
&\qquad \cdot \left[\frac{\sum_b \alpha_b^2}{\alpha_0}\left(\alpha_0 + N^\circ + 1 - \frac{N^\circ N^\bullet \sigma^2_{\bar{r}}}{\alpha_0 + N^\circ}\right) + N^\circ\left(1 + \frac{\alpha_0 N^\bullet \sigma^2_{\bar{r}}}{\alpha_0 + N^\circ}\right)\right] \\
&\quad = \quad \eta_2^2 \qquad\qquad (34b)
\end{aligned}
$$

where we replaced the variance by its expectation under the generative distribution for tMDPs for the same reason as before. These formulæ again behave as expected intuitively, and numerical simulations also confirm their validity.

In summary, the performance of the model-based controller at limited experience can be evaluated by using Equations 34a-34b and 32a, 32c to determine η_1 and η_2 at the preterminal and non-preterminal levels. These parameters then can be substituted to the derivation of Section 4.2 (Eqs. 25d-25e, 28d, 29e-29f) (true computational noise can be added to taste) and used to compute the value of progressively higher levels in the tMDP by proceeding backwards from terminal states to the root state. The value of the root state yields the expected performance of the controller.

References

[1] Sutton, R.S. & Barto, A.G. *Reinforcement Learning* (MIT Press, 1998).

[2] Kearns, M. & Singh, S. Finite-sample convergence rates for Q-learning and indirect algorithms. in *Advances in Neural Information Processing Systems* Vol. 11 (eds. Kearns, M.S., Solla, S.A. & Cohn, D.A.), Vol. 11, 996–1002 (MIT Press, Cambridge, MA, 1999).

[3] Meuleau, N. & Bourgine, P. Exploration of multi-state environments: local measures and back-propagation of uncertainty. *Machine Learning* **35**, 117–154 (1999).

[4] Owen, D.B. & Steck, G.P. Moments of order statistics from the equicorrelated multivariate normal distribution. *Ann Math Stat* **33**, 1286–1291 (1962).