Supporting derivations: Bayesian methods for efficient Reinforcement Learning in tabular problems

Efstratios Markou
Department of Engineering
University of Cambridge
stratismar@gmail.com

Carl E. Rasmussen
Department of Engineering
University of Cambridge
cer54@cam.ac.uk

Moment matching across the Bellman equations

The state and action-returns are respectively defined as:

$$w^\pi_{\mathbf{s}} \equiv \sum_{t=1}^{T} \gamma^{\,t-1} r_t \big| \pi, \mathbf{s}_1 = \mathbf{s}, \mathcal{T}, \mathcal{R}, \;\text{ and }\; z^\pi_{\mathbf{s},\mathbf{a}} \equiv \sum_{t=1}^{T} \gamma^{\,t-1} r_t \big| \pi, \mathbf{s}_1 = \mathbf{s}, \mathbf{a}_1 = \mathbf{a}, \mathcal{T}, \mathcal{R}.$$

They satisfy corresponding recursive relations:

$$w^\pi_{\mathbf{s}} = r_{\mathbf{s},\mathbf{a},\mathbf{s}'} + \gamma w^\pi_{\mathbf{s}'}, \;\; \mathbf{a} \sim \pi, \mathbf{s}' \sim \mathcal{T} \tag{1}$$

$$z^\pi_{\mathbf{s},\mathbf{a}} = r_{\mathbf{s},\mathbf{a},\mathbf{s}'} + \gamma z^\pi_{\mathbf{s}',\mathbf{a}'}, \;\; \mathbf{a}' \sim \pi, \mathbf{s}' \sim \mathcal{T} \tag{2}$$

where equality is here taken to mean that the two sides are identically distributed. We can explicitly require that the first and second-order moments across these BEs are equal. In subsequent discussion, we use $\boldsymbol{\theta}_{\mathcal{T}}, \boldsymbol{\theta}_{\mathcal{R}}, \boldsymbol{\theta}_{\mathcal{W}^\pi}$ and $\boldsymbol{\theta}_{\mathcal{Z}^\pi}$ to denote the parameters of $\mathcal{T}, \mathcal{R}, \mathcal{W}^\pi$ and $\mathcal{Z}^\pi$ respectively.

First moment matching

Taking expectations of both sides of eq. (1), we obtain

$$\begin{aligned} \mathbb{E}_{w,\boldsymbol{\theta}_{\mathcal{W}}}[w^\pi_{\mathbf{s}}] &= \mathbb{E}_{r,\boldsymbol{\theta}_{\mathcal{R}},\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}},\mathbf{a}}[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}] + \gamma \mathbb{E}_{w,\boldsymbol{\theta}_{\mathcal{W}},\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}[w^\pi_{\mathbf{s}'}], \\ \mathbb{E}_{z,\boldsymbol{\theta}_{\mathcal{Z}}}[z^\pi_{\mathbf{s},\mathbf{a}}] &= \mathbb{E}_{r,\boldsymbol{\theta}_{\mathcal{R}},\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}] + \gamma \mathbb{E}_{z,\boldsymbol{\theta}_{\mathcal{Z}},\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}},\mathbf{a}'}[z^\pi_{\mathbf{s}',\mathbf{a}'}] \end{aligned} \tag{3}$$

where the expectations are taken over the posterior distributions of the subscript variables. We recognise as the familiar BEs in terms of expectations for V^π and Q^π. These encode the requirement that, in expectation, the rewards and state/action-values should be consistent.

Second moment matching

Taking variances of both sides of eq. (1) and eq. (2), and using the laws of total variance and total covariance, we obtain similar consistency requirements for the variances:

$$\begin{aligned} \text{Var}_{w,\boldsymbol{\theta}_{\mathcal{W}}}[w^\pi_{\mathbf{s}}] &= \text{Var}_{r,\boldsymbol{\theta}_{\mathcal{R}},\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}},\mathbf{a}}[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}] + \gamma \, \text{Var}_{w,\boldsymbol{\theta}_{\mathcal{W}},\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}[w^\pi_{\mathbf{s}'}], \\ \text{Var}_{z,\boldsymbol{\theta}_{\mathcal{Z}}}[z^\pi_{\mathbf{s},\mathbf{a}}] &= \text{Var}_{r,\boldsymbol{\theta}_{\mathcal{R}},\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}] + \gamma \, \text{Var}_{z,\boldsymbol{\theta}_{\mathcal{Z}},\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}},\mathbf{a}'}[z^\pi_{\mathbf{s}',\mathbf{a}'}] \end{aligned} \tag{4}$$

where variances are taken over the posterior distributions of the subscript variables. In subsequent discussion we assume a deterministic policy π for simplicity, which implies that variances over $\mathbf{a}$ and $\mathbf{a}'$ are zero. It is straightforward to extend to the general case of a stochastic policy and we refrain from this for simplicity. Starting from the state-returns BE (eq. (1)) and using the law of total variance:

$$\underbrace{\text{Var}_{w,\boldsymbol{\theta}_{\mathcal{W}}}\left[w^\pi_{\mathbf{s}}\right]}_{\text{Total return unc.}} = \underbrace{\text{Var}_{\boldsymbol{\theta}_{\mathcal{W}}}\left[\mathbb{E}_w\left[w^\pi_{\mathbf{s}}|\boldsymbol{\theta}_{\mathcal{W}}\right]\right]}_{\text{Epistemic return unc.}} + \underbrace{\mathbb{E}_{\boldsymbol{\theta}_{\mathcal{W}}}\left[\text{Var}_w\left[w^\pi_{\mathbf{s}}|\boldsymbol{\theta}_{\mathcal{W}}\right]\right]}_{\text{Aleatoric return unc.}}. \tag{5}$$

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

We expand the RHS of eq. (2) to obtain:

$$\begin{aligned}\underbrace{\mathrm{Var}_{r,\boldsymbol{\theta}_{\mathcal{R}},w,\boldsymbol{\theta}_{\mathcal{W}},\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}},\mathbf{a}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}+\gamma w^{\pi}_{\mathbf{s}'}\right]}_{\text{Next-return variance}} =& \underbrace{\mathrm{Var}_{r,\boldsymbol{\theta}_{\mathcal{R}},\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}\right]}_{\text{Reward variance}} \\ &+2\gamma\underbrace{\mathrm{Cov}_{r,\boldsymbol{\theta}_{\mathcal{R}},w,\boldsymbol{\theta}_{\mathcal{W}},\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'},w^{\pi}_{\mathbf{s}'}\right]}_{\text{Reward-return covariance}} \\ &+\gamma^{2}\underbrace{\mathrm{Var}_{w,\boldsymbol{\theta}_{\mathcal{W}},\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[w^{\pi}_{\mathbf{s}'}\right]}_{\text{Next-return variance}}.\end{aligned} \tag{6}$$

The variances in eq. (6) contain both aleatoric and epistemic contributions, which we aim to separate by using the laws of total variance and total covariance (Weiss et al. (2006)). The reward variance in eq. (6) can be expanded as:

$$\begin{aligned}\underbrace{\mathrm{Var}_{r,\boldsymbol{\theta}_{\mathcal{R}},\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}\right]}_{\text{Reward variance}} =& \underbrace{\mathrm{Var}_{\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathbb{E}_{r,\boldsymbol{\theta}_{\mathcal{R}}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}\middle|\mathrm{s}',\boldsymbol{\theta}_{\mathcal{T}}\right]\right]}_{\substack{\text{Reward variance from transition unc.}\\ \text{aleatoric + epistemic}}} \\ &+\underbrace{\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathrm{Var}_{r,\boldsymbol{\theta}_{\mathcal{R}}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}\middle|\mathrm{s}',\boldsymbol{\theta}_{\mathcal{T}}\right]\right]}_{\substack{\text{Reward variance from reward unc.}\\ \text{aleatoric + epistemic}}}.\end{aligned} \tag{7}$$

Applying total variance to the first term in eq. (7), we obtain

$$\begin{aligned}\underbrace{\mathrm{Var}_{\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathbb{E}_{r,\boldsymbol{\theta}_{\mathcal{R}}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}\middle|\mathrm{s}',\boldsymbol{\theta}_{\mathcal{T}}\right]\right]}_{\substack{\text{Reward variance from transition unc.}\\ \text{aleatoric + epistemic}}} =& \underbrace{\mathrm{Var}_{\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathbb{E}_{\mathbf{s}',r,\boldsymbol{\theta}_{\mathcal{R}}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}\middle|\boldsymbol{\theta}_{\mathcal{T}}\right]\right]}_{\substack{\text{Reward variance from transition unc.}\\ \text{purely epistemic}}} \\ &+\underbrace{\mathbb{E}_{\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathrm{Var}_{\mathbf{s}'}\left[\mathbb{E}_{r,\boldsymbol{\theta}_{\mathcal{R}}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}\middle|\mathrm{s}',\boldsymbol{\theta}_{\mathcal{T}}\right]\right]\right]}_{\substack{\text{Reward variance from transition unc.}\\ \text{purely aleatoric}}}\end{aligned} \tag{8}$$

Similarly for the second term in eq. (7):

$$\begin{aligned}\underbrace{\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathrm{Var}_{r,\boldsymbol{\theta}_{\mathcal{R}}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}\middle|\mathrm{s}',\boldsymbol{\theta}_{\mathcal{T}}\right]\right]}_{\substack{\text{Reward variance from reward unc.}\\ \text{aleatoric + epistemic}}} =& \underbrace{\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathrm{Var}_{\boldsymbol{\theta}_{\mathcal{R}}}\left[\mathbb{E}_{r}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}\middle|\mathrm{s}',\boldsymbol{\theta}_{\mathcal{T}},\boldsymbol{\theta}_{\mathcal{R}}\right]\right]\right]}_{\substack{\text{Reward variance from reward unc.}\\ \text{purely epistemic}}} \\ &+\underbrace{\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathbb{E}_{\boldsymbol{\theta}_{\mathcal{R}}}\left[\mathrm{Var}_{r}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}\middle|\mathrm{s}',\boldsymbol{\theta}_{\mathcal{T}},\boldsymbol{\theta}_{\mathcal{R}}\right]\right]\right]}_{\substack{\text{Reward variance from reward unc.}\\ \text{purely aleatoric}}},\end{aligned} \tag{9}$$

with which we conclude the expansion of the reward variance term. We apply the same steps for the state-return variance in eq. (6). By total variance:

$$\underbrace{\mathrm{Var}_{w,\boldsymbol{\theta}_{\mathcal{W}},\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[w^{\pi}_{\mathbf{s}'}\right]}_{\text{Next-step value variance}} = \underbrace{\mathrm{Var}_{\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathbb{E}_{w,\boldsymbol{\theta}_{\mathcal{W}}}\left[w^{\pi}_{\mathbf{s}'}|\mathrm{s}',\boldsymbol{\theta}_{\mathcal{T}}\right]\right]}_{\substack{\text{State-return variance from transition unc.}\\ \text{aleatoric + epistemic}}} + \underbrace{\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathrm{Var}_{w,\boldsymbol{\theta}_{\mathcal{W}}}\left[w^{\pi}_{\mathbf{s}'}|\mathrm{s}',\boldsymbol{\theta}_{\mathcal{T}}\right]\right]}_{\substack{\text{State-return variance from state-return unc.}\\ \text{aleatoric + epistemic}}}. \tag{10}$$

Decomposing each of the two terms in eq. (10) by total variance, we obtain

$$\begin{aligned}\underbrace{\mathrm{Var}_{\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathbb{E}_{w,\boldsymbol{\theta}_{\mathcal{W}}}\left[w^{\pi}_{\mathbf{s}'}|\mathrm{s}',\boldsymbol{\theta}_{\mathcal{T}}\right]\right]}_{\substack{\text{State-return variance from transition unc.}\\ \text{aleatoric + epistemic unc.}}} =& \underbrace{\mathrm{Var}_{\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathbb{E}_{\mathbf{s}',w,\boldsymbol{\theta}_{\mathcal{W}}}\left[w^{\pi}_{\mathbf{s}'}|\boldsymbol{\theta}_{\mathcal{T}}\right]\right]}_{\substack{\text{State-return variance from transition unc.}\\ \text{purely epistemic}}} \\ &+\underbrace{\mathbb{E}_{\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathrm{Var}_{\mathbf{s}'}\left[\mathbb{E}_{w,\boldsymbol{\theta}_{\mathcal{W}}}\left[w^{\pi}_{\mathbf{s}'}|\mathrm{s}',\boldsymbol{\theta}_{\mathcal{T}}\right]\right]\right]}_{\substack{\text{State-return variance from transition unc.}\\ \text{purely aleatoric}}},\end{aligned}$$

for the first term. For the second term:

$$\begin{aligned}\underbrace{\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathrm{Var}_{w,\boldsymbol{\theta}_{\mathcal{W}}}\left[w^{\pi}_{\mathbf{s}'}|\mathrm{s}',\boldsymbol{\theta}_{\mathcal{T}}\right]\right]}_{\substack{\text{State-return variance from state-return unc.}\\ \text{aleatoric + epistemic}}} =& \underbrace{\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathrm{Var}_{\boldsymbol{\theta}_{\mathcal{W}}}\left[\mathbb{E}_{w}\left[w^{\pi}_{\mathbf{s}'}|\mathrm{s}',\boldsymbol{\theta}_{\mathcal{T}}\right]\right]\right]}_{\substack{\text{State-return variance from state-return unc.}\\ \text{purely epistemic}}} + \\ &+\underbrace{\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_{\mathcal{T}}}\left[\mathbb{E}_{\boldsymbol{\theta}_{\mathcal{W}}}\left[\mathrm{Var}_{w}\left[w^{\pi}_{\mathbf{s}'}|\mathrm{s}',\boldsymbol{\theta}_{\mathcal{T}}\right]\right]\right]}_{\substack{\text{State-return variance from state-return unc.}\\ \text{purely aleatoric}}}.\end{aligned}$$

which concludes the decomposition of the value variance terms. For the reward-value covariance term, we use the law of total covariance to obtain

$$\begin{aligned}\underbrace{\mathrm{Cov}_{r,\boldsymbol{\theta}_\mathcal{R},w,\boldsymbol{\theta}_\mathcal{W},\mathbf{s}',\boldsymbol{\theta}_\mathcal{T}}[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}, w^\pi_{\mathbf{s}'}]}_{\text{Reward-return covariance}} &= \underbrace{\mathrm{Cov}_{\mathbf{s}',\boldsymbol{\theta}_\mathcal{T}}\left[\mathbb{E}_{r,\boldsymbol{\theta}_\mathcal{R}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}|\mathrm{s}',\boldsymbol{\theta}_\mathcal{T}\right],\mathbb{E}_{w,\boldsymbol{\theta}_\mathcal{W}}\left[w^\pi_{\mathbf{s}'}|\mathrm{s}',\boldsymbol{\theta}_\mathcal{T}\right]\right]}_{\substack{\text{Reward-return covariance due to transition unc.}\\ \text{epistemic + aleatoric}}} + \\ &+ \mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_\mathcal{T}}\left[\cancel{\mathrm{Cov}_{r,\boldsymbol{\theta}_\mathcal{R},w,\boldsymbol{\theta}_\mathcal{W}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}, w^\pi_{\mathbf{s}'}|\mathrm{s}',\boldsymbol{\theta}_\mathcal{T}\right]}\right]\end{aligned} \tag{11}$$

where the second term is 0 due to the conditional independence of $r_{\mathbf{s},\mathbf{a},\mathbf{s}'}$ and and $w^\pi_{\mathbf{s}'}$ given s'. Applying total covariance to eq. (11):

$$\begin{aligned}\underbrace{\mathrm{Cov}_{\mathbf{s}',\boldsymbol{\theta}_\mathcal{T}}\left[\mathbb{E}_{r,\boldsymbol{\theta}_\mathcal{R}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}|\mathrm{s}',\boldsymbol{\theta}_\mathcal{T}\right],\mathbb{E}_{w,\boldsymbol{\theta}_\mathcal{W}}\left[w^\pi_{\mathbf{s}'}|\mathrm{s}',\boldsymbol{\theta}_\mathcal{T}\right]\right]}_{\substack{\text{Reward-value covariance due to dynamics}\\ \text{epistemic + aleatoric}}} &= \underbrace{\mathrm{Cov}_{\boldsymbol{\theta}_\mathcal{T}}\left[\mathbb{E}_{\mathbf{s}',r,\boldsymbol{\theta}_\mathcal{R}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}|\boldsymbol{\theta}_\mathcal{T}\right],\mathbb{E}_{\mathbf{s}',w,\boldsymbol{\theta}_\mathcal{W}}\left[w^\pi_{\mathbf{s}'}|\boldsymbol{\theta}_\mathcal{T}\right]\right]}_{\substack{\text{Reward-value covariance due to } \boldsymbol{\theta}_\mathcal{T} \text{ uncertainty}\\ \text{purely epistemic}}} \\ &+ \underbrace{\mathbb{E}_{\boldsymbol{\theta}_\mathcal{T}}\left[\mathrm{Cov}_{\mathbf{s}'}\left[\mathbb{E}_{r,\boldsymbol{\theta}_\mathcal{R}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}|\mathrm{s}',\boldsymbol{\theta}_\mathcal{T}\right],\mathbb{E}_{w,\boldsymbol{\theta}_\mathcal{W}}\left[w^\pi_{\mathbf{s}'}|\mathrm{s}',\boldsymbol{\theta}_\mathcal{T}\right]\right]\right]}_{\substack{\text{Reward-value covariance due to dynamics stochasticity}\\ \text{purely aleatoric}}}\end{aligned} \tag{12}$$

This concludes the decomposition of all terms of the RHS of eq. (6). We consider the epistemic uncertainties only, and require that these are equal across eq. (6) to obtain:

$$\begin{aligned}\underbrace{\mathrm{Var}_{\boldsymbol{\theta}_\mathcal{W}}\left[\mathbb{E}_w\left[w^\pi_{\mathbf{s}}|\boldsymbol{\theta}_\mathcal{W}\right]\right]}_{\text{Epistemic state-return unc.}} &= \underbrace{\mathrm{Var}_{\boldsymbol{\theta}_\mathcal{T}}\left[\mathbb{E}_{\mathbf{s}',r,\boldsymbol{\theta}_\mathcal{R}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}|\boldsymbol{\theta}_\mathcal{T}\right]\right]}_{\substack{\text{Epistemic reward unc. from}\\ \text{dynamics unc.}}} + \underbrace{\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_\mathcal{T}}\left[\mathrm{Var}_{\boldsymbol{\theta}_\mathcal{R}}\left[\mathbb{E}_r\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}|\mathrm{s}',\boldsymbol{\theta}_\mathcal{T},\boldsymbol{\theta}_\mathcal{R}\right]\right]\right]}_{\substack{\text{Epistemic rewards unc. from}\\ \text{rewards unc.}}} + \\ &+ 2\gamma\underbrace{\mathrm{Cov}_{\boldsymbol{\theta}_\mathcal{T}}\left[\mathbb{E}_{\mathbf{s}',r,\boldsymbol{\theta}_\mathcal{R}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}|\boldsymbol{\theta}_\mathcal{T}\right],\mathbb{E}_{\mathbf{s}',w,\boldsymbol{\theta}_\mathcal{W}}\left[w^\pi_{\mathbf{s}'}|\boldsymbol{\theta}_\mathcal{T}\right]\right]}_{\substack{\text{Epistemic reward and state-return covariance}\\ \text{from dynamics unc.}}} \\ &+ \gamma^2\underbrace{\mathrm{Var}_{\boldsymbol{\theta}_\mathcal{T}}\left[\mathbb{E}_{\mathbf{s}',w,\boldsymbol{\theta}_\mathcal{W}}\left[w^\pi_{\mathbf{s}'}|\boldsymbol{\theta}_\mathcal{T}\right]\right]}_{\substack{\text{Epistemic state-return unc. from}\\ \text{dynamics unc.}}} + \gamma^2\underbrace{\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_\mathcal{T}}\left[\mathrm{Var}_{\boldsymbol{\theta}_\mathcal{W}}\left[\mathbb{E}_w\left[w^\pi_{\mathbf{s}'}|\mathrm{s}',\boldsymbol{\theta}_\mathcal{W}\right]\right]\right]}_{\substack{\text{Epistemic state-return unc. from}\\ \text{state-return unc.}}}\end{aligned} \tag{13}$$

which concludes our derivation - note that wherever $\mathbf{a}$ is seen as a free variable in the equation above, it is implied that $\mathbf{a} = \pi(\mathrm{s})$. Following the same argument, we obtain a consistency requirement for the action-returns:

$$\begin{aligned}\underbrace{\mathrm{Var}_{\boldsymbol{\theta}_\mathcal{Z}}\left[\mathbb{E}_z\left[z^\pi_{\mathbf{s},\mathbf{a}}|\boldsymbol{\theta}_\mathcal{Z}\right]\right]}_{\text{Epistemic action-return unc.}} &= \underbrace{\mathrm{Var}_{\boldsymbol{\theta}_\mathcal{T}}\left[\mathbb{E}_{\mathbf{s}',r,\boldsymbol{\theta}_\mathcal{R}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}|\boldsymbol{\theta}_\mathcal{T}\right]\right]}_{\substack{\text{Epistemic reward unc. from}\\ \text{dynamics unc.}}} + \underbrace{\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_\mathcal{T}}\left[\mathrm{Var}_{\boldsymbol{\theta}_\mathcal{R}}\left[\mathbb{E}_r\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}|\mathrm{s}',\boldsymbol{\theta}_\mathcal{T},\boldsymbol{\theta}_\mathcal{R}\right]\right]\right]}_{\substack{\text{Epistemic rewards unc. from}\\ \text{rewards unc.}}} + \\ &+ 2\gamma\underbrace{\mathrm{Cov}_{\boldsymbol{\theta}_\mathcal{T}}\left[\mathbb{E}_{\mathbf{s}',r,\boldsymbol{\theta}_\mathcal{R}}\left[r_{\mathbf{s},\mathbf{a},\mathbf{s}'}|\boldsymbol{\theta}_\mathcal{T}\right],\mathbb{E}_{\mathbf{s}',z,\boldsymbol{\theta}_\mathcal{Z}}\left[z^\pi_{\mathbf{s}',\mathbf{a}'}|\boldsymbol{\theta}_\mathcal{T}\right]\right]}_{\substack{\text{Epistemic reward and action-return covariance}\\ \text{from dynamics unc.}}} \\ &+ \gamma^2\underbrace{\mathrm{Var}_{\boldsymbol{\theta}_\mathcal{T}}\left[\mathbb{E}_{\mathbf{s}',z,\boldsymbol{\theta}_\mathcal{Z}}\left[z^\pi_{\mathbf{s}',\mathbf{a}'}|\boldsymbol{\theta}_\mathcal{T}\right]\right]}_{\substack{\text{Epistemic action-return unc. from}\\ \text{dynamics unc.}}} + \gamma^2\underbrace{\mathbb{E}_{\mathbf{s}',\boldsymbol{\theta}_\mathcal{T}}\left[\mathrm{Var}_{\boldsymbol{\theta}_\mathcal{Z}}\left[\mathbb{E}_z\left[z^\pi_{\mathbf{s}',\mathbf{a}'}|\mathrm{s}',\boldsymbol{\theta}_\mathcal{Z}\right]\right]\right]}_{\substack{\text{Epistemic action-return unc. from}\\ \text{state-return unc.}}}\end{aligned} \tag{14}$$

where again we have $\mathbf{a}' = \pi(\mathbf{s}')$ according to the deterministic policy. Note that all terms in the RHS of both eq. (13) and eq. (14) can be obtained either in closed form or by efficient MC integrals once eq. (3) has been solved - except for the very last term in both cases.

We recognise the last term of the RHS is equal to the LHS term, smoothed by the predictive posterior over $\mathbf{s}'$. Therefore eq. (13) and eq. (14) are linear equations in the epistemic uncertainties of the returns which can be solved in $\mathcal{O}(|\mathcal{S}|^3)$ time.

References

Weiss, N., Holmes, P., and Hardy, M. (2006). *A Course in Probability*. Pearson Addison Wesley.