Appendices

Section EC.1 gives complete proofs for results stated in the main text of the paper. Section EC.2 discusses an extension of the Bayesian learning model to handle unknown variance. Section EC.3 provides the full implementation details for our numerical examples.

EC.1. Appendix: proofs

Below, we give the full technical proofs for results that were stated in the main text.

EC.1.1. Proof of Proposition 1

From (9), it is clear that $\Sigma^n(S^x, S^x)$ is decreasing in n, and therefore must have a limit. To show the first statement, fix ω and let $S^x \in E^{\pi,x}(\omega)$. Let (n_k) be a subsequence, converging to infinity, such that under policy π, $S^{x,n_k}(\omega) = S^x$ for all k. Then, we rewrite (9) as

$$\Sigma^{n_k+1}(S^x, S^x, \omega) = \left(1 - \frac{\Sigma^{n_k}(S^x, S^x, \omega)}{\sigma_\varepsilon^2 + \Sigma^{n_k}(S^x, S^x, \omega)}\right) \Sigma^{n_k}(S^x, S^x, \omega). \tag{EC.1}$$

Suppose that $\lim_{n\to\infty} \Sigma^n(S^x, S^x, \omega) > 0$. Then, it follows from (EC.1) that

$$\lim_{k\to\infty} \Sigma^{n_k+1}(S^x, S^x, \omega) < \lim_{n\to\infty} \Sigma^n(S^x, S^x, \omega),$$

contradicting the uniqueness of limits. Thus, $\Sigma^n(S^x, S^x, \omega) \to 0$. By the Cauchy-Schwarz inequality, it follows that $\Sigma^n(S^x, S^y, \omega) \to 0$ as well for all $S^y \in \mathcal{S}^x$.

We will now show the second statement. If $E^{\pi,x}(\omega) \neq \mathcal{S}^x$, we can partition $\Sigma^n(\omega)$ as

$$\Sigma^n(\omega) = \begin{pmatrix} \Sigma^{E,n}(\omega) & \Sigma^{cross,n}(\omega) \\ \left(\Sigma^{cross,n}(\omega)\right)^T & \Sigma^{E^c,n}(\omega) \end{pmatrix}, \tag{EC.2}$$

where $\Sigma^{E,n}(\omega)$ contains variances and covariances for only those states in $E^{\pi,x}(\omega)$, while $\Sigma^{E^c,n}(\omega)$ contains covariance information only for states in the complement of $E^{\pi,x}(\omega)$. We define τ to be the last time the policy π visits a state $S^x \notin E^{\pi,x}(\omega)$ on ω. Without loss of generality, we can assume $\tau = -1$, since we can just take $\Sigma^0 = \Sigma^{\tau+1}$.

By Assumption 2, Σ^0 is invertible. As in (EC.2), we can partition $(\Sigma^0)^{-1}$ to obtain

$$\left(\Sigma^0\right)^{-1}(\omega) = \begin{pmatrix} \bar{\Sigma}^{E,0}(\omega) & \bar{\Sigma}^{cross,0}(\omega) \\ \left(\bar{\Sigma}^{cross,0}(\omega)\right)^T & \bar{\Sigma}^{E^c,0}(\omega) \end{pmatrix}.$$

If $\tau = 0$, we can apply the Sherman-Morrison formula to (9) to write

$$\lim_{n\to\infty} \left(\Sigma^n\right)^{-1}(\omega) = \left(\Sigma^0\right)^{-1}(\omega) + \sigma_\varepsilon^2 \lim_{n\to\infty} nD(\omega),$$

where D is a diagonal matrix with

$$D(S^x, S^x, \omega) = \begin{cases} 1 & S^x \in E^{\pi,x}(\omega) \\ 0 & \text{otherwise.} \end{cases}$$

From the preceding discussion, we know that $\Sigma^{E,n}(\omega) \to 0$ and $\Sigma^{cross,n}(\omega) \to 0$. We now apply the matrix inversion lemma, and the continuity of the matrix inverse over invertible matrices, to obtain

$$\lim_{n\to\infty} \Sigma^{E^c,n}(\omega) = \left(\bar{\Sigma}^{E^c,0}\right)^{-1}(\omega).$$

We thus conclude that $\Sigma^n(\omega)$ converges componentwise to a limit $\Sigma^\infty(\omega)$.

We argue that $\Sigma^\infty(S^x, S^y, \omega) \neq 0$ for $S^x, S^y \notin E^{\pi,x}(\omega)$. By Assumption 2, Σ^0 has full rank. Let $M = |\mathcal{S}^x|$. We can view Σ^0 as the covariance matrix in a ranking and selection problem (Chau et al. 2014) with M alternatives. In this problem, let μ denote the vector of true values of these alternatives (analogous to V in the DP), and suppose that we have a multivariate Gaussian prior with covariance matrix Σ^0 on their values.

Suppose that we can collect unbiased Gaussian observations of the unknown values (as in Assumption 1). Assume, furthermore, that we sequentially collect these observations according to a deterministic policy ρ_ω, which measures the alternatives in the same order in which the policy π visits post-decision states in the DP for the sample path ω. The update (9) does not depend on the value of the observation. In fact, if we know which alternative was observed at time n, the covariance matrix is updated deterministically. Thus, the sequence of posterior covariance matrices Σ^n in the ranking and selection problem will be identical to the sequence observed in the DP for the sample path ω. The limiting behaviour of these two sequences will also be identical.

Now suppose that $S^x, S^y \notin E^{\pi,x}(\omega)$, and consider the two corresponding alternatives in the ranking and selection problem. For notational simplicity, we still label these as S^x and S^y. By Assumption 2, our prior beliefs about these two alternatives are correlated. We can then express the true values of S^x and S^y as

$$\mu(S^x) = a_x \cdot C + b_x \cdot Z_x + c \cdot Z_{x,y}$$

$$\mu(S^y) = a_y \cdot C + b_y \cdot Z_y + d \cdot Z_{x,y},$$

where C, Z_x, Z_y and $Z_{x,y}$ are mutually independent Gaussian random variables, each with strictly positive variance. Suppose that we never collect any observations for the alternatives S^x and S^y, analogously to our earlier assumption that $\Sigma^0 = \Sigma^{\tau+1}$. Then, the conditional variance of Z_x, Z_y and $Z_{x,y}$ remains unchanged, and the resulting correlation between S^x and S^y remains non-zero even in the limit. Since the limiting behaviour of Σ^n is identical for this problem and for the DP, we conclude that $\Sigma^\infty(S^x, S^y, \omega) \neq 0$.

EC.1.2. Proof of Proposition 2

Recall from (15) that

$$\nu^{KG}(S^x, S) = \sum_{y_i \in \mathcal{A}(S)} [b(S, y_{i+1}) - b(S, y_i)] f(-|c_i|),$$

where a and b are computed using (13) and (14), and the values c_{y_i} are the breakpoints between non-dominated lines of the form $a_i + b_i \cdot z$. These breakpoints have the form

$$c_i = \frac{a(S, y_i) - a(S, y_{i+1})}{b(S, y_{i+1}) - b(S, y_i)}, \tag{EC.3}$$

which is a rational function of the components of a and b. The denominator in (EC.3) is non-zero because the set $\mathcal{A}(S)$ has already removed all dominated actions. The function f is continuous. Thus, to show the continuity of $\nu^{KG}(S^x, S)$, it is enough to consider only those parameters $\left(\bar{V}, \Sigma\right)$ such that, for either $y = \arg\max_{y_i \in \mathcal{A}(S)}$ or $y = \arg\min_{y_i \in \mathcal{A}(S)}$, there exists some action z such that $a(S, z) < a(S, y)$ and $b(S, z) = b(S, y)$. This means that there is an action z that is dominated,

but that the line corresponding to this action has the same slope as the line for action y. A slight change in the line corresponding to action z will add another action to the set $\mathcal{A}(S)$.

Let $\varepsilon > 0$ and choose δ to satisfy

$$\delta f(0) < \varepsilon.$$

In addition, δ should be small enough so that changing the slope and intercept of the line corresponding to action z will only add another action to the beginning or end of the list $\mathcal{A}(S)$. Now, let $\left(\bar{V}', \Sigma'\right)$ be such that $|a(S,z) - a'(S,z)| < \delta$ and $|b(S,z) - b'(S,z)| < \delta$. The only change in $\nu^{KG,n}(S^x, S)$ will be a new breakpoint, so that

$$\begin{aligned}
\left|\nu^{KG,n}\left(S^x, S; \bar{V}, \Sigma\right) - \nu^{KG,n}\left(S^x, S; \bar{V}'\Sigma'\right)\right| &= |b(S,z) - b(S,y)| \cdot f\left(-\left|\frac{a(S,y) - a(S,z)}{b(S,z) - b(S,y)}\right|\right) \\
&\leq |b(S,z) - b(S,y)| f(0) \\
&< \delta f(0) \\
&< \varepsilon.
\end{aligned}$$

The second line is due to the fact that f is increasing. We conclude that the KG factor is continuous in the belief parameters.

EC.1.3. Proof of Theorem 1

We assume that a suitable set of measure zero has been removed from the outcome space. Suppose that $E^{KG}(\omega) \neq \mathcal{S}$. As in Proposition 1, we partition $\Sigma^n(\omega)$ as

$$\Sigma^n(\omega) = \begin{pmatrix} \Sigma^{E,n}(\omega) & \Sigma^{cross,n}(\omega) \\ \left(\Sigma^{cross,n}(\omega)\right)^T & \Sigma^{E^c,n}(\omega) \end{pmatrix}.$$

By Proposition 1, we have

$$\Sigma^n(\omega) \to \begin{pmatrix} 0 & 0 \\ 0 & \Sigma^\infty(\omega) \end{pmatrix}, \tag{EC.4}$$

where all components of $\Sigma^\infty(\omega)$ are non-zero.

By Assumption 3, there must exist states $S \in E^{KG}(\omega)$ and $\bar{S} \notin E^{KG}(\omega)$ such that, for at least one action $x \in \mathcal{X}$, we have $P\left(\bar{S} \,|\, S, x\right) > 0$. Furthermore, by assumption, the offline KG policy must

take action x out of state S only finitely many times. If this were not the case, it would follow that $\bar{S} \in E^{KG}(\omega)$. Thus, we have $(S,x) \notin E^{KG,x}(\omega)$. Furthermore, $\left(\bar{S},\bar{x}\right) \notin E^{KG,x}(\omega)$ for all $\bar{x}$, since $\bar{S}$ is visited only finitely many times.

From (EC.4), we have

$$\Sigma^{\infty}\left(S^{M,x}\left(\bar{S},\bar{x}\right),S^{x}\right) \neq 0, \tag{EC.5}$$

$$\Sigma^{\infty}\left(S^{x},S^{x}\right) \neq 0. \tag{EC.6}$$

By Proposition 2, we have

$$\nu^{KG,n}\left(S^{x},\bar{S},\omega\right) \rightarrow \nu^{KG,\infty}\left(S^{x},\bar{S},\omega\right),$$

where $\nu^{KG,\infty}\left(S^{x},\bar{S},\omega\right) > 0$. The fact that the limit is strictly positive is ensured by Assumption 4. In (15), one component of the vector b^n will always be zero for all n, corresponding to the action Δ. In the limit, $b^n(\omega) \rightarrow b^{\infty}(\omega)$ where $b^{\infty}(\omega)$ has at least one zero component (due to Δ) and at least one non-zero component due to (EC.5). There must therefore be at least one breakpoint. The function f in (15) has no zeros on the real line, so we conclude that $\nu^{KG,\infty}\left(S^{x},\bar{S},\omega\right) > 0$. Since $P\left(\bar{S} \mid S,x\right) > 0$, it follows that

$$\lim_{n\rightarrow\infty} \sum_{S'} P\left(S' \mid S,x\right) \nu^{KG,n}\left(S^{x},S',\omega\right) \geq P\left(\bar{S} \mid S,x\right) \nu^{KG,\infty}\left(S^{x},\bar{S},\omega\right) > 0. \tag{EC.7}$$

Now, let y be an action taken infinitely often by the offline KG policy out of state S. Such an action must exist because S is visited infinitely often. From the preceding discussion, it follows that $P\left(S' \mid S,y\right) = 0$ for all $S' \notin E^{KG}(\omega)$. Furthermore, for any $S' \in E^{KG}(\omega)$, we have

$$\Sigma^{n}\left(S^{M,x}\left(S',x'\right),S^{y}\right) \rightarrow 0$$

due to (EC.4). By Proposition 2, it follows that $\nu^{KG,n}\left(S^{y},S',\omega\right) \rightarrow 0$, whence

$$\sum_{S'} P\left(S' \mid S,y\right) \nu^{KG,n}\left(S^{y},S',\omega\right) \rightarrow 0. \tag{EC.8}$$

We now put together (EC.7) and (EC.8). Let $\varepsilon = \nu^{KG,\infty}(S^x, S', \omega)$. Then, there exists an integer K_ω such that, for all $n \geq K_\omega$, we have

$$\left| \sum_{S'} P(S' \mid S, x) \nu^{KG,n}(S^x, S', \omega) - \lim_{n\to\infty} \sum_{S'} P(S' \mid S, x) \nu^{KG,n}(S^x, S', \omega) \right| < \frac{\varepsilon}{2},$$
$$\sum_{S'} P(S' \mid S, y) \nu^{KG,n}(S^y, S', \omega) < \frac{\varepsilon}{2}.$$

Consequently, at all times after K_ω, the offline KG policy will prefer action x to action y out of state S. This contradicts the assumption that $\bar{S}$ is visited finitely many times. We conclude that $E^{KG}(\omega) = \mathcal{S}$.

EC.1.4. Proof of Proposition 3

By Proposition 2, every KG factor converges to a limit. Suppose that $\nu^{KG,n}\left(S^x, \bar{S}, \omega\right) \to \nu^{KG,\infty}\left(S^x, \bar{S}, \omega\right)$ for some $S^x \in \mathcal{S}^x$ and $\bar{S} \in \mathcal{S}$. It follows that $S^x \notin E^{KG,x}(\omega)$, otherwise we would have $\Sigma^n(S^x, S^x, \omega) \to 0$, which would imply that $\nu^{KG,n}\left(S^x, \bar{S}, \omega\right) \to 0$ by continuity.

Because S is visited infinitely often by Theorem 1, there must be at least one action y such that $S^y \in E^{KG,x}(\omega)$. For this action, $\nu^{KG,n}(S^y, S', \omega) \to 0$. We can then repeat the argument concluding the proof of Theorem 1 to find that, after some time K_ω, the offline KG policy will prefer action x to action y, which implies that $S^x \in E^{KG,x}(\omega)$ and therefore $\nu^{KG,n}\left(S^x, \bar{S}, \omega\right) \to 0$. We conclude that every KG factor must converge to zero under the offline KG policy.

EC.1.5. Proof of Proposition 4

Suppose that $S^x, S^{M,x}\left(\bar{S}, \bar{x}\right) \notin E^{KG,x}(\omega)$. By Proposition 1, it must be the case that $\Sigma^n\left(S^{M,x}\left(\bar{S}, \bar{x}\right), S^x, \omega\right)$, $\Sigma^n\left(S^{M,x}\left(\bar{S}, \bar{x}\right), S^{M,x}\left(\bar{S}, \bar{x}\right), \omega\right)$, and $\Sigma^n(S^x, S^x, \omega)$ converge to non-zero limits, as in (EC.5) and (EC.6). Applying Assumption 4 as in the proof of Theorem 1, we find that $\nu^{KG,n}\left(S^x, \bar{S}\right)$ converges to a strictly positive limit. This contradicts Proposition 3, which states that all KG factors must converge to zero.

EC.2. Appendix: learning the unknown noise variance

One limitation of the Bayesian model from Section 2.2 is that the observations $\hat{v}^{n+1}$ are assumed to have known variance σ_ε^2. In practice, this creates a tunable parameter that requires additional computational effort to optimize. In the following, we present an extension that explicitly models the noise variance as a random variable, and updates a set of beliefs about this quantity over time.

Let $\rho = \sigma_\varepsilon^{-2}$ be the *precision* of the observations. We assume that ρ is unknown and impose the prior distribution $\rho \sim Gamma\,(\alpha^0, \beta^0)$, where $\alpha^0, \beta^0 > 0$ are fixed. We then suppose that the *conditional* prior distribution of V, given ρ, is multivariate normal with mean $\bar{V}^0$ and covariance matrix $\rho^{-1}\Sigma^0$. In Bayesian statistics, the joint distribution of (V, ρ) is known as a "multivariate normal-gamma" prior. It can be shown (DeGroot 1970) that the marginal distribution of V under this model is a multivariate Student's t-distribution (Kotz and Nadarajah 2004), by analogy with classical statistics where this distribution is used to model observations with unknown variance. In this setting, Assumption 1 reads as follows.

Assumption EC.1. *Given V and ρ, the ADP observation $\hat{v}^{n+1}$ follows the conditional distribution $\mathcal{N}\left(V\left(S^{x,n}\right), \rho^{-1}\right)$ and is conditionally independent of past observations.*

As in Section 2.2, Assumption EC.1 allows us to update our entire approximation $\bar{V}^n$ using a single scalar observation $\hat{v}^{n+1}$. Equations (8)-(9) now become

$$\begin{aligned}
\bar{V}^{n+1}\left(S^x\right) &= \bar{V}^n\left(S^x\right) + \frac{\hat{v}^{n+1} - \bar{V}^n\left(S^x\right)}{1 + \Sigma^n\left(S^{x,n}, S^{x,n}\right)} \Sigma^n\left(S^x, S^{x,n}\right), \\
\Sigma^{n+1}\left(S^x, S^y\right) &= \Sigma^n\left(S^x, S^y\right) - \frac{\Sigma^n\left(S^x, S^{x,n}\right)\Sigma^n\left(S^{x,n}, S^y\right)}{1 + \Sigma^n\left(S^{x,n}, S^{x,n}\right)}, \\
\alpha^{n+1} &= \alpha^n + \frac{1}{2}, \\
\beta^{n+1} &= \beta^n + \frac{\left(\hat{v}^{n+1} - \bar{V}^n\left(S^x\right)\right)^2}{2\left(1 + \Sigma^n\left(S^{x,n}, S^{x,n}\right)\right)}.
\end{aligned}$$

This learning model thus has essentially the same complexity as the one in Section 2.2; the only addition consists of two scalar posterior parameters α^n, β^n.

The KG logic of Section 2.3 can now be applied. The only difference (Han et al. 2016) is that now (15) should be rewritten as

$$\mathbb{E}_x^n \max_y \bar{Q}^{n+1}\left(S^{n+1},y\right) = \left(\max_y a^n\left(S^{n+1},y\right)\right) + \sum_{y_i \in \mathcal{A}\left(S^{n+1}\right)} \left[b^n\left(S^{n+1},y_{i+1}\right) - b^n\left(S^{n+1},y_i\right)\right] g_{2\alpha^n}\left(-|c_i|\right), \quad \text{(EC.9)}$$

where

$$g_s(t) = \frac{s+t^2}{s-1}\psi_s(t) + t\Psi_s(t)$$

and ψ_s, Ψ_s are the pdf and cdf of Student's t-distribution with s degrees of freedom. Equations (13)-(14), which define the vectors a^n and b^n in (15), are now replaced by

$$a^n\left(S^{n+1},y\right) = C\left(S^{n+1},y\right) + \gamma \bar{V}^n\left(S^{M,x}\left(S^{n+1},y\right)\right), \quad \text{(EC.10)}$$

$$b^n\left(S^{n+1},y\right) = \gamma \Sigma^n\left(S^{M,x}\left(S^{n+1},y\right), S^{x,n}\right)\sqrt{\frac{\beta^n}{\alpha^n\left(1+\Sigma^n\left(S^{x,n},S^{x,n}\right)\right)}}. \quad \text{(EC.11)}$$

After this, the computation of the KG policy proceeds as before (in particular, the breakpoints c_i in (EC.9) are computed from a^n and b^n in the same way). The only difference is that the knowledge gradient computations now use the tail properties of the t-distribution, rather than the standard normal.

In fact, the multivariate normal-gamma prior may also be used in conjunction with the linear VFA of Section 3.1. Once again, we assume that $V(S^x) = \theta^T \phi(S^x)$ and impose the prior distribution $\rho \sim Gamma(\alpha^0, \beta^0)$. The conditional prior distribution of θ, given ρ, is $\mathcal{N}(\theta^0, \rho^{-1}\Lambda^0)$. Under Assumption EC.1, this produces the update

$$\begin{aligned}
\theta^{n+1} &= \theta^n + \frac{\hat{v}^{n+1} - \left(\theta^n\right)^T \phi\left(S^{x,n}\right)}{1+\phi\left(S^{x,n}\right)^T \Lambda^n \phi\left(S^{x,n}\right)} \Lambda^n \phi\left(S^{x,n}\right), \\
\Lambda^{n+1} &= \Lambda^n - \frac{\Lambda^n \phi\left(S^{x,n}\right)\phi\left(S^{x,n}\right)^T \Lambda^n}{1+\phi\left(S^{x,n}\right)^T \Lambda^n \phi\left(S^{x,n}\right)}, \\
\alpha^{n+1} &= \alpha^n + \frac{1}{2}, \\
\beta^{n+1} &= \beta^n + \frac{\left(\hat{v}^{n+1} - \left(\theta^n\right)^T \phi\left(S^{x,n}\right)\right)^2}{2\left(1+\phi\left(S^{x,n}\right)^T \Lambda^n \phi\left(S^{x,n}\right)\right)}.
\end{aligned}$$

The KG computation is again given by (EC.9), but (EC.10)-(EC.11) are replaced by

$$a^n\left(S^{n+1},y\right)=C\left(S^{n+1},y\right)+\gamma\left(\theta^n\right)^T\phi\left(S^{M,x}\left(S^{n+1},y\right)\right),$$
$$b^n\left(S^{n+1},y\right)=\gamma\phi\left(S^{M,x}\left(S^{n+1},y\right)\right)^T\Lambda^n\phi\left(S^{x,n}\right)\sqrt{\frac{\beta^n}{\alpha^n\left(1+\phi\left(S^{x,n}\right)^T\Lambda^n\phi\left(S^{x,n}\right)\right)}}.$$

Although the performance of this model may still be influenced by the initialization of α^0,β^0, these parameters will be adjusted over time and thus the model is less susceptible to misspecification of the noise variance. Furthermore, since Student's t-distribution has heavier tails than the normal distribution, the value of information will tend to be higher under this framework and thus KG will conduct more exploration. Moreover, KG will incur virtually the same computational cost here as in the known-variance model, since the main computational bottleneck in (EC.9) is the calculation of the breakpoints, which remains unchanged.

EC.3. Appendix: details of experimental settings

In this section, we give more details on the benchmark policies and test problems from Section 5.2.

EC.3.1. Description of learning policies

Five different types of policies were implemented; their descriptions are as follows.

Knowledge gradient (KG). We tested both the online and offline versions of the KG policy, from (17) and (19), with sample size $K=30$.

Value of perfect information (VPI). To our knowledge, the VPI policy by Dearden et al. (1998) was the first exploration strategy to be used together with a Bayesian prior on the value function. The original definition of VPI is designed for discrete state spaces with normal-gamma priors. However, the policy easily carries over to the VFA structures from Section 3. The decision rule is given by

$$X^{VPI,n}\left(S^n,K^n\right)=\arg\max_x C\left(S^n,x\right)+\gamma\bar{V}^n\left(S^{x,n}\right)+\gamma\nu^{VPI,n}\left(S^{x,n}\right)$$

where

$$\nu^{VPI,n}\left(S^{x,n}\right)=\sqrt{\Sigma^n\left(S^{x,n},S^{x,n}\right)}f\left(-\frac{\left|\bar{V}^n\left(S^{x,n}\right)-\max_{y\neq x}\bar{V}^n\left(S^{y,n}\right)\right|}{\sqrt{\Sigma^n\left(S^{x,n},S^{x,n}\right)}}\right),$$

with f remaining the same as in Section 2.3. If we use a VFA from Section 3, we replace $\Sigma^n\left(S^{x,n},S^{x,n}\right)$ by the corresponding expression for the prior variance of $V\left(S^{x,n}\right)$. For example, in Section 3.1, this is $\phi\left(S^{x,n}\right)^T\Lambda^n\phi\left(S^{x,n}\right)$, and in Section 3.2, this is $\left(\sum_g\lambda^{g,n}\left(S^{x,n}\right)+\delta^{g,n}\left(S^{x,n}\right)\right)^{-1}$.

VPI can be viewed as a version of the expected improvement policy (Jones et al. 1998, Gramacy and Lee 2011) from the global optimization literature. However, its adaptation to ADP with correlated beliefs is completely new to this paper. Like KG, the VPI policy can work well with both basis functions and hierarchical aggregation. We view this as an additional argument in favour of our Bayesian framework; although we mainly focus on the KG policy in this paper, our Bayesian models have even broader potential since they may be combined with other algorithms.

Epsilon-greedy. The ε-greedy policy chooses the action $\arg\max_x C\left(S^n,x\right)+\gamma\bar{V}^n\left(S^{x,n}\right)$ with probability $1-\varepsilon$, and a random action with probability ε. We tuned the parameter ε in our experiments. Clearly, this policy can be used with any representation of $\bar{V}^n$.

R-max. The R-max policy of Brafman and Tennenholtz (2003) has attracted considerable attention in the reinforcement learning literature. Essentially, the policy classifies the states based on whether or not we have "enough" knowledge of their values. The decision rule is given by

$$X^{Rmax,n}\left(S^n,K^n\right)=\arg\max_x C\left(S^n,x\right)+\gamma F^n\left(S^{x,n}\right)$$

where

$$F^n\left(S^{x,n}\right)=\begin{cases}R^{\max} & \text{if } S^{x,n} \text{ has been visited fewer than } m \text{ times,}\\ \bar{V}^n\left(S^{x,n}\right) & \text{if } S^{x,n} \text{ has been visited at least } m \text{ times.}\end{cases}$$

The integer m is a tunable parameter that represents the number of times we need to visit a post-decision state to gain enough knowledge about it. The value $R^{\max}$ is an arbitrarily large number. Thus, we are encouraged to explore actions with which we are unfamiliar. Because this policy does not easily extend to hierarchical aggregation (the number of times we have visited a state is ambiguous in the hierarchical model), we implemented it with a lookup-table approximation in one of our test problems. We did not implement R-max in problems where the lookup-table approximation did not scale.

E^3. The E^3 policy of Kearns and Singh (2002) is somewhat similar to R-max. If we visit a state that we have never visited before, we choose a random action. If we have visited the state at least once, but fewer than m times, we choose the action that has been tried the fewest number of times out of all the times we have previously visited the state. Lastly, if we have visited the state more than m times, we take the greedy action $\arg\max_x C(S^n, x) + \gamma \bar{V}^n(S^{x,n})$. We implemented E^3 in our first test problem, but not in subsequent problems, for the same reasons as R-max.

EC.3.2. Commodity storage problem with stochastic price

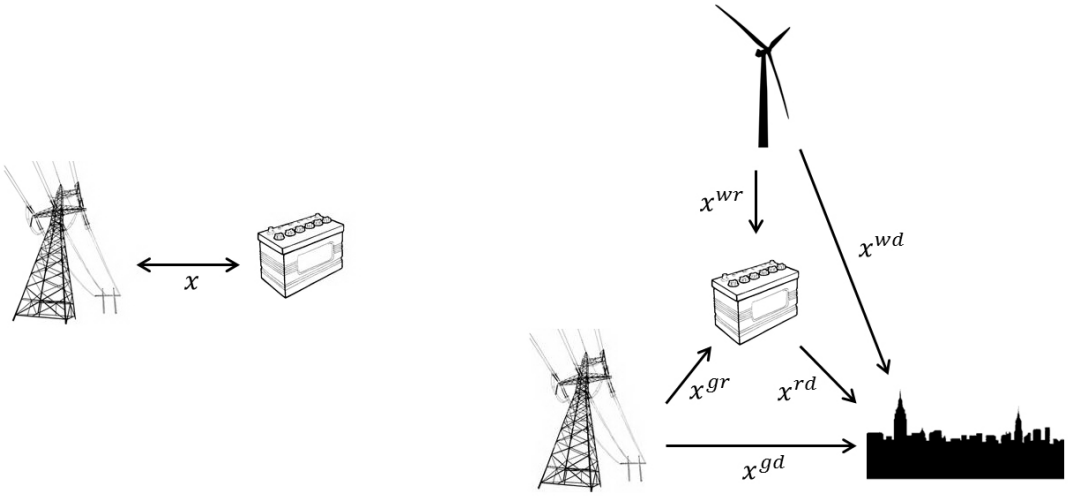
We considered two versions of a stylized inventory problem motivated by commodity storage and trading, an application that has seen recent attention, e.g., in Secomandi (2010). We have simplified this problem for easier benchmarking, to make some of the competing policies computationally tractable, and to allow us to run more simulations for statistically valid comparisons.

The first version of the problem considers an asset held in storage. At any point in time, we can buy from or sell to the spot market. Figure 1(a) illustrates the decision variable in this problem using the example of electricity stored in a battery. Our decision x^n at time n depends on the state variable $S^n = (R^n, P^n)$, where R^n is the amount currently in storage and P^n is the current spot price. If $x^n \geq 0$, we buy energy from the market, whereas we sell if $x^n < 0$. The single-period revenue or cost is then

$$C(S^n, x^n) = -P^n x^n,$$

since we pay a cost to buy, and receive revenue from selling. Our objective is to maximize the long-term profit. The spot price was assumed to follow a geometric Ornstein-Uhlenbeck process with mean reversion parameter 0.0633 and volatility parameter 0.2, ensuring a sufficient level of noise in the problem. The post-decision state, given a decision x^n, is computed by $R^{x,n} = R^n + x^n$ and $P^{x,n} = P^n$. To find the next pre-decision state, we take $R^{n+1} = R^{x,n}$ and simulate P^{n+1} from our price process.

We discretized R^n on the interval $[0, 1]$ in increments of 0.01, representing the percentage of the battery's capacity that is being used. The decision x^n was similarly discretized on $[-R^n, 1 - R^n]$.



(a) Commodity storage with scalar decision. (b) Commodity storage with vector decision.

Figure EC.1 Illustrations of the decision variable in two types of storage problems.

However, we did not discretize the price variable in the problem. When running a policy, we always kept track of the continuous value of P^n, and only discretized prices when calling the value function approximation. Thus, we are still solving a continuous problem; see Section 8.1.1 of Powell (2011) for a discussion of this point. The prior approximation $\bar{V}^0$ was uniformly set to a very large value, as discussed in Section 4.9 of Powell (2011), and (7) was used to set a prior covariance for the lookup table VFA.

For the hierarchical VFA of Section 3.2, each level of aggregation partitioned the state space into rectangles. At the finest levels of aggregation, we added more bins to the price variable, while the coarser levels focused primarily on the resource variable. We found that, as the discretization became finer, it was more important to distinguish between similar prices than similar storage quantities. This VFA was used for the hierarchical KG method, as well as the ε-greedy and VPI policies. Thus, all competing policies had access to the benefits of correlated beliefs.

EC.3.3. Commodity storage problem with stochastic supply

The second version of our commodity storage problem has a vector decision variable. Figure 1(b) illustrates using an example where electricity is obtained exogenously from a wind farm, stored in

a battery, and then used to satisfy demand. If there is not enough energy in storage to cover all the demand, we have to purchase the remainder from the spot market. This problem is similar to the one in Section EC.3.2, but now we can also receive supply from an exogenous process.

We considered a simplified version of the problem where the demand in each time period was set to a fixed number D. Our state variable was thus $S^n = (R^n, W^n)$, where $0 \leq R^n \leq \bar{R}$ is the amount in storage as before, and W^n is the stochastic supply process. In our simulations, we designed W^n by fitting a mean-reverting process to historical wind speed data, then converting these wind speeds to power output.

From Figure 1(b), we see that the decision is a vector $x^n = (x^{wr,n}, x^{wd,n}, x^{gr,n}, x^{gd,n}, x^{rd,n})$. We reduce this vector to three dimensions by observing that

$$\begin{aligned} E^n &= x^{wr,n} + x^{wd,n} \\ D &= x^{wd,n} + x^{gd,n} + x^{rd,n}, \end{aligned}$$

and setting $x^{wd,n} = E^n - x^{wr,n}$, and $x^{gd,n} = D - (E^n - x^{wr,n}) - x^{rd,n}$. We constrain $0 \leq x^{wr,n} \leq E^n$ to ensure the positivity of $x^{wd,n}$. We assume that oversupplying demand is not allowed, so we also constrain $0 \leq (E^n - x^{wr,n}) + x^{rd,n} \leq D$. Finally, we add the constraints

$$\begin{aligned} 0 \leq \rho^c x^{gr,n} &\leq \bar{R} - R^n \\ 0 \leq \rho^c x^{wr,n} &\leq \bar{R} - R^n \\ 0 \leq x^{rd,n} &\leq R^n + \rho^c \left(x^{wr,n} + x^{gd,n}\right) \\ \rho^d x^{rd,n} &\leq D, \end{aligned}$$

where ρ^c and ρ^d represent loss rates for the storage device (e.g., energy dissipation from the battery). Once the decision has been made, the post-decision storage level is given by

$$R^{x,n} = \min\left(R^n + \rho^c \left(x^{gr,n} + x^{wr,n}\right) - x^{rd,n}, \bar{R}\right),$$

indicating that there is a cap on how much we can store.

The single-period contribution function depends on a selling price P^s, a buying price P^g (if we buy from the spot market), and an extra penalty cost P^p imposed when our purchase $x^{gr,n} + x^{gd,n}$ exceeds some level x^{max}. To reduce the size of the state variable, we assumed constant prices $P^s = 0.14$, $P^g = 0.12$ and $P^p = 0.5$. We also let $D = 100$ be the constant demand, with $x^{max} = 75$. The contribution is then calculated as

$$C(S^n, x^n) = P^s D - P^g \left(x^{gr,n} + x^{gd,n}\right) - P^p \left(x^{gr,n} + x^{gd,n} - x^{max}\right)^+.$$

Maximizing the long-term contribution is equivalent to minimizing the long-term cost of meeting demand.

Even with these simplifications, we found that this discretized problem was too large to maintain a covariance matrix directly on the state space. For this reason, we only used the hierarchical VFA of Section 3.2 and compared KG, VPI, and ε-greedy. The R-max and E^3 policies are tied to a discrete value function representation and do not translate easily to the hierarchical model.

EC.3.4. Nomadic trucker problem

In the nomadic trucker problem (Powell 2011), a single truck observes demands that arise randomly in different locations (e.g., cities in the US), modeled as elements of a set $\mathcal{L}$. The trucker travels between these locations to accept those loads that maximize the long-term reward. The state $S^n = (S^{l,n}, S^{d,n}, S^{k,n})$ is a vector of attributes representing, respectively, the current location of the trucker, the current day of the week, and the trailer type.

The decision x^n is modeled as a binary vector with $x_i^n = 1$ if we choose the ith possible decision and $\sum_i x_i^n = 1$. The set of possible decisions depends on the current state. For example, we can choose to accept a currently available load of a particular type in some location, or we can choose to move to a different location without accepting any load. For each location $i \in \mathcal{L}$, the number $0 \leq b_i \leq 1$ represents the probability that a load originating at location i will appear at a given time step. We let $p_{ij}^d = p_d b_i (1 - b_j)$ be the probability that, on a given day d of the week, a load from i to j will appear, where p_d is the probability of loads appearing on day d.

When we select the decision x^n, our post-decision state $S^{x,n}$ is determined as if we had already arrived at the destination (that is, we change the location and day-of-week components of the state to the values they will have upon our arrival). The costs $C(S^n, x^n)$ generally depend on the distance d_{ij} that we travel between the current location i and the chosen destination j, either with or without a load.

In our implementation, locations lie on a 16×16 grid placed on a square area of 1000×1000 miles. Each location is described by coordinates (x_i, y_i), and the origin probabilities are given by

$$b_i = \rho \left(1 - \frac{h(x_i, y_i) - h^{min}}{h^{max} - h^{min}}\right),$$

where ρ is the arrival intensity of loads and h is the six-hump camelback function

$$h(x, y) = 4x^2 - 2.1x^4 + \frac{1}{3}x^6 + xy - 4y^2 + 4y^4$$

on the domain $[1.5, 2] \times [1, 1]$, properly scaled to the domain $[0, 0] \times [1000, 1000]$. The values $h^{min} = \min_{i \in \mathcal{L}} h(x_i, y_i)$ and $h^{max} = \max_{i \in \mathcal{L}} h(x_i, y_i)$ are used to scale $h(x_i, y_i)$ between $[0, 1]$. We set $\rho = 1$, which corresponds to an average of approximately 93 outgoing loads from the most popular origin location on the busiest day of the week. We use the load probabilities $p^d = (1, 0.8, 0.6, 0.7, 0.9, 0.2, 0.1)$ for d from Monday to Sunday, representing a situation where loads are more likely to appear at the beginning of the week and toward the end. The cost function is given by

$$C(S, x) = \begin{cases} -d_{ij} & \text{if we choose to move from } i \text{ to } j \text{ without taking a load} \\ r^k d_{ij} b_i & \text{if we move from } i \text{ to } j \text{ with a load of type } k. \end{cases}$$

The trailer type attribute can be either small, medium, or large, and varies in a cyclic fashion, irrespective of the remaining attributes. Larger trailer types result in higher rewards, with $r^1 = 1$, $r^2 = 1.5$, and $r^3 = 2$. Rewards are discounted using $\gamma = 0.95$.

Table EC.1 gives an overview of the aggregation structure used by the VFA, with '*' corresponding to a dimension that was included in the aggregation level, and '-' corresponding to a dimension that was aggregated out. Trailer type and day-of-week are either included or left out, while location is represented with an increasingly fine grid at the more disaggregate levels.

Level	Location	Trailer type	Day-of-week	Size of state space
0	16×16	*	*	$256 \cdot 3 \cdot 7 = 5376$
1	8×8	*	*	$64 \cdot 3 \cdot 7 = 1344$
2	4×4	*	*	$16 \cdot 3 \cdot 7 = 336$
3	4×4	-	*	$16 \cdot 7 = 112$
4	2×2	-	*	$4 \cdot 7 = 28$
5	-	-	*	7
6	-	-	-	1

Table EC.1 Aggregation structure for the nomadic trucker problem.

EC.3.5. Freight consolidation problem

In the freight consolidation problem, a decision-maker receives loads to be transported and periodically decides which of these loads should be consolidated in a high-capacity vehicle to be dispatched in the current period. The full mathematical model for this problem is given in Pérez Rivera and Mes (2017) and so we do not repeat it here; for example, the state variable is formally defined in eq. (1) of Pérez Rivera and Mes (2017), the decision is described by eqs. (2a)-(2f), etc. Below, we describe the modifications made to this model in order to adapt it to the setting of our paper.

Our implementation is based on Instance I_6^L from Pérez Rivera and Mes (2017), a delivery-only variant (no pickups). As originally presented, this instance was a finite-horizon minimization problem without exploration; we adapted it to the setting of infinite-horizon, discounted maximization with exploration as follows. First, the discount factor was chosen to be $\gamma = 0.99$. Second, we introduced revenues per container shipped, varying between 325 and 825 depending on the destination. These numbers are chosen in order to be comparable to the costs in Pérez Rivera and Mes (2017); for instance, travel costs are between 250 and 1000 depending on which destinations are visited, while the variable cost per container is between 50 and 100. The probabilities of freights of various types were taken from Pérez Rivera and Mes (2017).

Instance I_6^L has 12 different destinations, 3 possible values for the release day, and 3 possible values for the time window. Because of the large number of dimensions in this problem, we

implemented the method of basis functions as laid out in Section 3.1 and considered two VFAs. The more informative VFA, denoted by VFA2 in Section 5.2, includes a dummy variable for each destination/time-window combination, as well as features counting the numbers of loads headed to each destination. This second set of basis functions distinguishes between (i.e., uses separate features for) urgent loads, less urgent loads, and loads that have not yet been released. This VFA is identical to "VFA3" in Pérez Rivera and Mes (2017). In Section 5.2 we also considered a less informative VFA, denoted by VFA1, which omits the dummy variables for destination/time-window combinations.

The implementation of the KG and VPI policies is new to the present paper, as exploration was not considered in Pérez Rivera and Mes (2017). As mentioned in Section 5.2, although KG with VFA2 became computationally expensive, KG with VFA1 turned out to perform better than the other policies did with the richer VFA2.

References

Brafman RI, Tennenholtz M (2003) R-max – a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research* 3:213–231.

Chau M, Fu MC, Qu H, Ryzhov IO (2014) Simulation optimization: a tutorial overview and recent developments in gradient-based methods. Tolk A, Diallo SY, Ryzhov IO, Yilmaz L, Buckley S, Miller JA, eds., *Proceedings of the 2014 Winter Simulation Conference*, 21–35.

Dearden R, Friedman N, Russell S (1998) Bayesian Q-learning. *Proceedings of the 15th National Conference on Artificial Intelligence*, 761–768.

DeGroot MH (2011) *Optimal statistical decisions* (John Wiley and Sons).

Gramacy RB, Lee HKH (2011) Optimization under unknown constraints. Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M, eds., *Bayesian Statistics 9: Proceedings of the Ninth Valencia International Meeting*, 229–256.

Han B, Ryzhov IO, Defourny B (2016) Optimal learning in linear regression with combinatorial feature selection. *INFORMS Journal on Computing* 28(4):721–735.

Jones DR, Schonlau M, Welch WJ (1998) Efficient global optimization of expensive black-box functions. *Journal of Global Optimization* 13(4):455–492.

Kearns M, Singh S (2002) Near-optimal reinforcement learning in polynomial time. *Machine Learning* 49(2):209–232.

Kotz S, Nadarajah S (2004) *Multivariate t-distributions and their applications* (Cambridge University Press).

Pérez Rivera AE, Mes MRK (2017) Anticipatory freight selection in intermodal long-haul round-trips. *Transportation Research Part E: Logistics and Transportation Review* 105:176–194.

Powell WB (2011) *Approximate dynamic programming: solving the curses of dimensionality (2nd ed.)* (John Wiley and Sons).

Secomandi N (2010) Optimal commodity trading with a capacitated storage asset. *Management Science* 56(3):449–467.