Selecting Computations: Supplemental proofs

Nicholas Hay and Stuart Russell
Computer Science Division
University of California
Berkeley, CA 94720

Solomon Eyal Shimony and David Tolpin
Department of Computer Science
Ben-Gurion University of the Negev
Beer Sheva, Israel

Included here are all the proofs of the results in the paper. Definitions are included to maintain theorem numbering.

Definition 1. *A **metalevel probability model** is a tuple $(U_1, \ldots, U_k, \mathcal{E})$ consisting of jointly distributed random variables:*

- *Real random variables $U_1, \ldots, U_k$, where U_i is the utility of arm i, and*
- *A countable set $\mathcal{E}$ of random variables, each variable $E \in \mathcal{E}$ being a computation that can be performed and whose value is the result of that computation.*

Definition 2. *A (countable state, undiscounted) **Markov Decision Process** (MDP) is a tuple $M = (S, s_0, A_s, T, R)$ where: S is a countable set of states, $s_0 \in S$ is the fixed initial state, A_s is a countable set of actions available in state $s \in S$, $T(s, a, s')$ is the transition probability from $s \in S$ to $s' \in S$ after performing action $a \in A_s$, and $R(s, a, s')$ is the expected reward received on such a transition.*

Definition 3. *Given a metalevel probability model[1] $(U_1, \ldots, U_k, \mathcal{E})$ and a cost of computation $c > 0$, a corresponding **metalevel decision problem** is any MDP $M = (S, s_0, A_s, T, R)$ such that*

$$
\begin{aligned}
S &= \{\bot\} \cup \{\langle e_1 \ldots, e_n\rangle : e_i \in E_i \text{ for all } i, \\
&\qquad \text{for finite } n \geq 0 \text{ and distinct } E_i \in \mathcal{E}\} \\
s_0 &= \langle\rangle \\
A_s &= \{\bot\} \cup \mathcal{E}_s
\end{aligned}
$$

where $\bot \in S$ is the unique terminal state, where $\mathcal{E}_s \subseteq \mathcal{E}$ is a state-dependent subset of allowed computations, and when given any $s = \langle e_1, \ldots, e_n\rangle \in S$, computational action $E \in \mathcal{E}$, and $s' = \langle e_1, \ldots, e_n, e\rangle \in S$ where $e \in E$, we have:

$$
\begin{aligned}
&T(s, E, s') = P(E = e \mid E_1 = e_1, \ldots, E_n = e_n) \\
&T(s, \bot, \bot) = 1 \\
&R(s, E, s') = -c \\
&R(s, \bot, \bot) = \max_i \mu_i(s)
\end{aligned}
$$

where $\mu_i(s) = \mathbb{E}[U_i \mid E_1 = e_1, \ldots, E_n = e_n]$.

$$
V_M^\pi(s) = \mathbb{E}_M^\pi\left[\sum_{i=0}^{N} R(S_i, \pi(S_i), S_{i+1}) \mid S_0 = s\right] \tag{1}
$$

Theorem 4. *The value function of a metalevel decision process $M = (S, s_0, A_s, T, R)$ is of the form*

$$
V_M^\pi(s) = \mathbb{E}_M^\pi[-c\,N + \max_i \mu_i(S_N) \mid S_0 = s]
$$

where N denotes the (random) total number of computations performed; similarly for $Q_M^\pi(s, a)$.

Proof. Follows immediately from Equation (1) and the definition of the reward function in Definition 3. □

Theorem 5. *The optimal policy's expected number of computations is bounded by the value of perfect information times the inverse cost $1/c$:*

$$
\mathbb{E}^{\pi^*}[N \mid S_0 = s] \leq \frac{1}{c}\left(\mathbb{E}[\max_i U_i \mid S_0 = s] - \max_i \mu_i(s)\right).
$$

Further, any policy π with infinite expected number of computations has negative infinite value, hence the optimal policy stops with probability one.

Proof. The first follows as in state s the optimal policy has value at least that of stopping immediately ($\max_i \mu_i(s)$), and as $\mathbb{E}\max_i \mu_i(S_N) \leq \mathbb{E}\max_i U_i$ by Jensen's inequality. The second from Theorem 4. □

[1] Definition 1 made no assumption about the computational result variables $E_i \in \mathcal{E}$, but for simplicity in the following we'll assume that each E_i takes one of a countable set of values. Without loss of generality, we'll further assume the domains of the computational variables $E \in \mathcal{E}$ are disjoint.

Definition 6. *Given a metalevel decision problem $M = (S, s_0, A_s, T, R)$, the* ***myopic policy*** $\pi^m(s)$ *is defined to equal* $\operatorname{argmax}_{a \in A_s} Q^m(s, a)$ *where* $Q^m(s, \perp) = \max_i \mu_i(s)$ *and*

$$Q^m(s, E) = \mathbb{E}_M[-c + \max_i \mu_i(S_1) \mid S_0 = s, A_0 = E].$$

Theorem 7. *Given a metalevel decision problem $M = (S, s_0, A_s, T, R)$ if the myopic policy performs some computation in state $s \in S$, then the optimal policy does too, i.e., if $\pi^m(s) \neq \perp$ then $\pi^*(s) \neq \perp$.*

Proof. Observe that the myopic Q-function Equation (6) is equivalently given by

$$Q^m(s, a) = Q^{\perp}(s, a)$$

where $\perp$ is the policy which immediately stops $\perp(s) = \perp$. Thus $Q^m(s, a) \leq Q^*(s, a)$. If the optimal policy stops in a state $s \in S$ then

$$Q^{\pi^*}(s, a) \leq \max_i \mu_i(s),$$

and so the same holds for Q^m, showing the myopic stops. □

Definition 8. *Given a metalevel decision problem $M = (S, s_0, A_s, T, R)$, a subset $S' \subseteq S$ of states is* ***closed under transitions*** *if whenever $s' \in S'$, $a \in A_{s'}$, $s'' \in S$, and $T(s', a, s'') > 0$, we have $s'' \in S'$.*

Theorem 9. *Given a metalevel decision problem $M = (S, s_0, A_s, T, R)$ and a subset $S' \subseteq S$ of states closed under transitions, if the myopic policy stops in all states $s' \in S'$ then the optimal policy does too.*

Proof. Take any $s^* \in S'$, and note that all states the chain can transition to from s^* are also in S', by transition closure. Defining $m(s) = \max_i \mu_i(s)$, observe the myopic stopping for all such states implies that

$$\begin{aligned}\mathbb{E}^{\pi}[(m(S_{j+1}) - c)\, 1(j < N) \mid S_0 = s^*] \\ \leq \mathbb{E}^{\pi}[m(S_j)\, 1(j < N) \mid S_0 = s^*]\end{aligned}$$

holds for all j, and as a result:

$$\begin{aligned}V^{\pi}(s) &= \mathbb{E}^{\pi}[-cN + m(S_N) \mid S_0 = s^*] \\ &= \mathbb{E}^{\pi}[m(S_0) + \sum_{j=0}^{N-1}(m(S_{j+1}) - c - m(S_j)) \mid S_0 = s^*] \\ &\leq \max_i \mu_i(s^*)\end{aligned}$$

□

Theorem 10. *The one-armed Bernoulli decision process with constant arm $\lambda \in [0, 1]$ performs at most $\lambda(1 - \lambda)/c - 3 \leq 1/4c - 3$ computations.*

Proof. By Definition 6 and Example **??**, the myopic policy stops in a state (s, f) when

$$c \geq \mu \max(\mu^+, m) + (1 - \mu_i) \max(\mu^-, m) - \max(\mu, m) \tag{2}$$

where $\mu = (s+1)/(n+2)$ is the mean utility for arm 2, where $n = s + f$, $\mu^- = \mu - \mu/(n+3)$, and $\mu^+ = \mu + (1 - \mu)/(n + 3)$ are the posterior means of arm 2 after simulating a failure and a success, respectively. Whenever Equation (2) holds, stopping is preferred to sampling.

Fixing n and maximizing over μ, we get sufficient condition for stopping

$$c \geq \frac{\lambda(1-\lambda)}{(n+3)} \qquad\qquad n \geq \frac{\lambda(1-\lambda)}{c} - 3 \tag{3}$$

Since the set of states satisfying Equation (3) is closed under transitions (n only increases), by Theorem 7. Finally, note $\max_{\lambda \in [0,1]} \lambda(1 - \lambda) = 1/4$. □

Definition 11. *Given a metalevel decision problem $M = (S, s_0, A_s, T, R)$, and a constant $\lambda \in \mathbb{R}$, define $M_\lambda = (S, s_0, A_s, T, R_\lambda)$ to be M with an additional action of known value λ, defined by:*

$$\begin{aligned}R_\lambda(s, E, s') &= R(s, E, s') \\ R_\lambda(s, \perp, \perp) &= \max\{\lambda, R(s, \perp, \perp)\}\end{aligned}$$

Theorem 12. *Given a metalevel decision problem $M = (S, s_0, A_s, T, R)$, there exists a real interval $I(s)$ for every state $s \in S$ such that it is optimal to stop in state s in M_μ iff $\mu \notin I(s)$. Furthermore, $I(s)$ contains $\max_i \mu_i(s)$ whenever it is nonempty.*

Proof. With $m(s) = \max_i \mu_i(s)$, the utility of a policy π starting in state s of M_μ is

$$V^{\pi}_{M_\mu}(s) = \mathbb{E}_{\pi}[-c\,N + \max(\mu, m(S_N)) \mid S_0 = s]$$

and the utility of stopping in this state $\max(\mu, m(s_0))$. We wish to show that the set of μ such that

$$\max_{\pi} \mathbb{E}_{\pi}[-c\,N + \max(\mu, m(S_N)) - \max(\mu, m(S_0)) \mid S_0 = s] \leq 0$$

forms an interval.

Observe that for any random variable X, $\mathbb{E}[\max(\mu, X)]$ is monotonically increasing in μ with subderivative between zero and one. As a result, for any v_1 $\mathbb{E}[\max(\mu, X)] - \max(\mu, v_1)$ is monotonically increasing for $\mu < v$, and monotonically decreasing thereafter. Therefore, the set if μ such that this expression is at most v_2 forms an interval, containing v_1 if non-empty.

Applying this with $v_1 = m(s_0)$ and $v_2 = \mathbb{E}_{\pi}[c\,N]$, and observing that the union of intervals containing a point is an interval containing that point, gives the result. □

Definition 13. *A metalevel probability model* $\mathcal{M} = (U_1, \ldots, U_k, \mathcal{E})$ *has* ***independent actions*** *if the computational variables can be partitioned* $\mathcal{E} = \mathcal{E}_1 \cup \cdots \cup \mathcal{E}_k$ *such that the sets* $\{U_i\} \cup \mathcal{E}_i$ *are independent of each other for different* i.

Definition 14. *Given a metalevel decision problem* $M = (S, s_0, A_s, T, R)$ *with independent actions, the* ***blinkered policy*** π^b *is defined by* $\pi^b(s) = \mathrm{argmax}_{a \in A_s} Q^b(s, a)$ *where* $Q^b(s, \perp) = \perp$ *and*

$$Q^b(s, E_i) = \sup_{\pi \in \Pi_i^b} Q^\pi(s, E_i) \tag{4}$$

for $E_i \in \mathcal{E}_i$*, where* Π_i^b *is the set of policies* π *where* $\pi(s) \in \mathcal{E}_i$ *for all* $s \in S$.

Definition 15. *Given a metalevel decision problem* $M = (S, s_0, A_s, T, R)$ *with independent actions, a* ***one-action metalevel decision problem*** *for* $i = 1, \ldots, k$ *is the metalevel decision problem* $M_{i,m}^1 = (S_i, s_0, A_{s0}, T_i, R_i)$ *defined by the metalevel probability model* $(U_0, U_i, \mathcal{E}_i)$ *with* $U_0 = m$.

Theorem 16. *Given a metalevel decision problem* $M = (S, s_0, A_s, T, R)$ *with independent actions, let* M_{i,λ_i}^1 *be the ith one-action metalevel decision problem for* $i = 1, \ldots, k$*. Then for any* $s \in S$*, whenever* $E_i \in A_s \cap \mathcal{E}_i$ *we have:*

$$Q_M^b(s, E_i) = Q^*_{M^1_{i,\mu^*_{-i}}}(s_i, E_i)$$

where $\mu^*_{-i} = \max_{j \neq i} \mu_j(s)$.

Proof. Fix a state s, a $E_i \in A_s$ and take any $\pi \in \Pi_i^b$. Note that such policies are equivalent to polices π' on $M_{1,m}^1$, and all such policies are represented. Consider $Q^\pi(s, E_i)$. As $\pi(s) \in \mathcal{E}_i$ for all $s \in S$, by action independence $\mu_j(S_n) = \mu_j(s)$. By this and Theorem 4, then,

$$Q_M^\pi(s, E_i) = \mathbb{E}_M^\pi[-c\,N + \max(\mu_i(S_N), m_i) \mid S_0 = s, A_0 = E_i]$$

Noting that $\mu_i(S_N)$ is a function only of $(S_N)_i$, and that since But then this is exactly $Q^*_{M^1_{i,\mu^*_{-i}}}(s_i, E_i)$. Taking the supremum over π gives the result. □

Theorem 17. Λ_i^b *is bounded from above as*

$$\Lambda_\alpha^b \leq \frac{N \overline{X}_\beta^{n_\beta}}{n_\alpha} \Pr(\overline{X}_\alpha^{n_\alpha + N} \leq \overline{X}_\beta^{n_\beta})$$
$$\Lambda_{i|i \neq \alpha}^b \leq \frac{N(1 - \overline{X}_\alpha^{n_\alpha})}{n_i} \Pr(\overline{X}_i^{n_i + N} \geq \overline{X}_\alpha^{n_\alpha}) \tag{5}$$

Proof. For the case $i \neq \alpha$, the probability that the ith arm is finally chosen instead of α is $\Pr(\overline{X}_i^{n_i+N} \geq \overline{X}_\alpha^{n_\alpha})$. $X_i \leq 1$, therefore $\overline{X}_i^{n_i+N} \leq \overline{X}_\alpha^{n_\alpha} + \frac{N(1-\overline{X}_\alpha^{n_\alpha})}{N+n_i}$. Hence, the intrinsic value of blinkered information is at most:

$$\frac{N(1 - \overline{X}_\alpha^{n_\alpha})}{N + n_i} \Pr(\overline{X}_i^{n_i+N} \geq \overline{X}_\alpha^{n_\alpha})$$
$$\leq \frac{N(1 - \overline{X}_\alpha^{n_\alpha})}{n_i} \Pr(\overline{X}_i^{n_i+N} \geq \overline{X}_\alpha^{n_\alpha}) \tag{6}$$

Proof for the case $i = \alpha$ is similar. □

Theorem 18. *The probabilities in Equation (5) are bounded from above as*

$$\Pr(\overline{X}_\alpha^{n_\alpha+N} \leq \overline{X}_\beta^{n_\beta}) \leq 2\exp\left(-\varphi(\overline{X}_\alpha^{n_\alpha} - \overline{X}_\beta^{n_\beta})^2 n_\alpha\right)$$
$$\Pr(\overline{X}_{i|i\neq\alpha}^{n_\alpha+N} \geq \overline{X}_\beta^{n_\beta}) \leq 2\exp\left(-\varphi(\overline{X}_\alpha^{n_\alpha} - \overline{X}_i^{n_i})^2 n_i\right) \tag{7}$$

where $\varphi = \min\left(2(\frac{1+n/N}{1+\sqrt{n/N}})^2\right) = 8(\sqrt{2}-1)^2 > 1.37$.

Proof. Equation (7)) follow from the observation that if $i \neq \alpha$, $\overline{X}_i^{n_i+N} > \overline{X}_\alpha^{n_i}$ if and only if the mean $\overline{X}_i^N$ of N samples from $n_i + 1$ to $n_i + N$ is at least $\overline{X}_\alpha^{n_i} + (\overline{X}_\alpha^{n_i} - \overline{X}_i^{n_i})\frac{n_i}{N}$.

For any δ, the probability that $\overline{X}_i^{n_i+N}$ is greater than $\overline{X}_\alpha^{n_i}$ is less than the probability that $\mathbb{E}[X_i] \geq \overline{X}_i^n + \delta$ *or* $\overline{X}_i^N \geq \mathbb{E}[X_i] + \overline{X}_\alpha^{n_\alpha} - \overline{X}_i^{n_i} - \delta + (\overline{X}_\alpha^{n_\alpha} - \overline{X}_i^{n_i})\frac{n_i}{N}$, thus, by the union bound, less than the sum of the probabilities:

$$\Pr(\overline{X}_i^{n_i+N} \geq \overline{X}_\alpha^{n_i})$$
$$\leq \Pr(\mathbb{E}[X_i] - \overline{X}_i^{n_i} \geq \delta) \tag{8}$$
$$+ \Pr\left(\overline{X}_i^N - \mathbb{E}[X_i] \geq \overline{X}_\alpha^{n_\alpha} - \overline{X}_i^{n_i} - \delta + (\overline{X}_\alpha^{n_\alpha} - \overline{X}_i^{n_i})\frac{n_i}{N}\right)$$

Bounding the probabilities on the right-hand side using the Hoeffding inequality, obtain:

$$\Pr(\overline{X}_i^{n_i+N} \geq \overline{X}_\alpha^{n_\alpha}) \leq$$
$$\exp(-2\delta^2 n_i) +$$
$$\exp\left(-2\left((\overline{X}_\alpha^{n_\alpha} - \overline{X}_i^{n_i})\left(1 + \frac{n_i}{N}\right) - \delta\right)^2 N\right) \tag{9}$$

Find δ for which the two terms on the right-hand side of Equation (9) are equal:

$$\exp(-\delta^2 n) = \exp\left(-2\left((\overline{X}_\alpha - \overline{X}_i)(1 + \frac{n_i}{N}) - \delta\right)^2 N\right) \tag{10}$$

Solve Equation (10) for δ: $\delta = \frac{1+\frac{n_i}{N}}{1+\sqrt{\frac{n_i}{N}}}(\overline{X}_\alpha^{n_\alpha} - \overline{X}_i^{n_i}) \geq 2(\sqrt{2}-1)(\overline{X}_\alpha^{n_\alpha} - \overline{X}_i^{n_i})$. Substitute δ into Equation (9) and obtain

$$
\begin{aligned}
&\Pr(\overline{X}_i^{n_i} \geq \overline{X}_\alpha^{n_\alpha}) \\
&\quad \leq 2\exp\left(-2\left(\frac{1+\frac{n_i}{N}}{1+\sqrt{\frac{n_i}{N}}}(\overline{X}_\alpha^{n_\alpha} - \overline{X}_i^{n_i})\right)^2 n_i\right) \\
&\quad \leq 2\exp(-8(\sqrt{2}-1)^2(\overline{X}_\alpha^{n_\alpha} - \overline{X}_i^{n_i})^2 n_i) \\
&\quad = 2\exp(-\varphi(\overline{X}_\alpha^{n_\alpha} - \overline{X}_i^{n_i})^2 n_i) \qquad (11)
\end{aligned}
$$

Derivation for the case $i = \alpha$ is similar. □

Corollary 19. *An upper bound on the VOI estimate Λ_i^b is obtained by substituting Equation (7) into Equation (5).*

$$
\begin{aligned}
&\Lambda_\alpha^b \leq \hat{\Lambda}_\alpha^b = \frac{2N\overline{X}_\beta^{n_\beta}}{n_\alpha}\exp\left(-\varphi(\overline{X}_\alpha^{n_\alpha} - \overline{X}_\beta^{n_\beta})^2 n_\alpha\right) \\
&\Lambda_{i|i\neq\alpha}^b \leq \hat{\Lambda}_i^b = \frac{2N(1-\overline{X}_\alpha^{n_\alpha})}{n_i}\exp\left(-\varphi(\overline{X}_\alpha^{n_\alpha} - \overline{X}_i^{n_i})^2 n_i\right) \\
&\qquad (12)
\end{aligned}
$$