Multitasking: Supplementary Material

June 8, 2015

Abstract

This supplementary material contains proofs for the theorems in *Multitasking: Efficient Optimal Planning for Bandit Superprocesses*. This document is largely self-contained and includes a copy of the relevant notation and definitions from the paper.

1 Model Definitions & Notation

Definition 1. *(Markov Decision Process [Puterman, 2009]) A (finite-state, discounted) MDP, M, is a tuple $M = \langle \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle$. $\mathcal{S}$ is a set of states. $\mathcal{A}$ is a set of actions. $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is a function that assigns probability to state transitions for each state–action pair. R is a (bounded) reward function that maps state–action pairs to (positive) rewards $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$. $\gamma \in [0, 1)$ is a discount factor.*

A solution to M is a policy, π, that maps states to actions. The value of a state, s, under π is the sum of expected discounted rewards received by starting in s and selecting actions according to π:

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) | s_0 = s, \pi\right].$$

The optimal policy, π^*, maximizes this value. In the above definition, and the ones that follow, we use superscripts to indicate dependence on the agent's policy. To simplify notation, we will omit these superscripts when the policy referred to is the optimal policy (e.g., $V(s) = V^{\pi^*}(s)$). The Q-function for the state–action pair, (s, a), is the value of taking a in s and selecting future actions according to π^*.

Definition 2. *(Bandit Superprocess [Nash, 1973]) Given k MDPs, $\{M_i = \langle \mathcal{S}_i, \mathcal{A}_i, T_i, R_i, \gamma \rangle\}$, we define*

$$M = \sum_i M_i = \langle \mathcal{S},\ \mathcal{A},\ T,\ R,\ \gamma \rangle$$

to be the bandit superprocess (BSP) with arms $\{M_i\}$. M has a state for each combination of arm states and a action for each arm action:

$$\mathcal{S} = \mathop{\times}_i \mathcal{S}_i \qquad\qquad \mathcal{A} = \mathop{\cup}_i \mathcal{A}_i.$$

The transition distribution is stationary for arms that are not selected and follows the identical reward and transition distributions for the selected arm.

The classic multi-armed bandit (MAB) consists of a set of Markov reward processes (MRPs). An MRP is a degenerate MDP with $|\mathcal{A}| = 1$, and so is a special case of a bandit superprocess. MABs and BSPs exhibit highly factored transition dynamics, a useful tool to leverage this decomposition is a retirement process.

Definition 3. *(Retirement Process [Whittle, 1980]) Let M be an MDP. For $\rho \geq 0$, the retirement process for M with retirement reward, ρ, is an MDP, M_ρ, with a single additional state, s_R, and action, a_R. a_R transitions deterministically to s_R and receives reward ρ. s_R is a sink state that accrues zero reward.*

We will denote the retirement process value function as a function of a state and retirement reward, $V(s, \rho)$. We let the optimal policy for retirement reward ρ be π^*_ρ. For any fixed policy, π, and retirement reward, ρ, we write the set of states where retirement is optimal as τ^π_ρ. We drop the superscript in the case where the policy is the optimal policy under ρ: $\tau_\rho = \tau_\rho^{\pi^*_\rho}$. We adopt a convention from the MAB literature and abuse notation somewhat to denote the (random) number of actions taken prior to retirement, given that the agent is in state s, as $\tau^\pi_\rho(s)$. For $s' \in \tau^\pi_\rho$ we let $P_{retire}(s'|s, \rho, \pi)$ be the probability that s' is the first state in τ^π_ρ the agent will reach given that it is in state s and executes policy π. We denote the expected discounted reward accrued prior to retirement as $R^\pi_\rho(s)$. We omit the superscript in the case where the policy is π^*_ρ. This allows us to write the following expression for the retirement process value function:

$$V(s, \rho) = R_\rho(s) + \mathbb{E}[\gamma^{\tau_\rho(s)}]\rho. \tag{1}$$

$V(s, \rho)$ is piecewise linear in ρ. In regions of retirement reward where the optimal policy and stopping rule do not change, $\frac{\partial V}{\partial \rho}(s, \rho)$ is defined and is equal to the expected value of the discount parameter at retirement. Policies whose optimality is independent of ρ are called *dominating policies*.

Definition 4. *(Dominating Policy) Let π be a policy for an MDP, M. π is a* dominating policy *iff*

$$\forall \rho \geq 0 \, \forall s \notin \tau_\rho \; \pi(s) = \pi^*_\rho(s).$$

The values of retirement reward where the optimal stopping rule or the optimal policy changes will characterize the interactions between arms of a BSP or MAB:

Definition 5. *(Critical Values of an MDP) Let M be a Markov decision process. The interaction values of M, $\mathcal{C}(M) = \{\rho_i\}$, are the values of retirement reward such that optimal stopping rule or policy changes.*

2 Whittle Integral

In this section, we define the Whittle integral and state the Whittle condition. Then we prove that it is equivalent to taking a weighted combination of retirement process values for a single arm.

Definition 6. *(Whittle Integral) Let M be a BSP. Let i index the arms of M. For any state, $s = \{s_i\}$, and $\rho \geq 0$, the Whittle integral of s is defined as follows:*

$$\hat{V}(s, \rho) = I - \int_{\alpha=\rho}^{I} d\alpha \prod_i \frac{\partial V_i}{\partial \rho}(s_i, \alpha). \tag{2}$$

Where $I \geq \max_i I_{s_i}$.

Theorem 1. *(Whittle Condition [Whittle, 1980]) Let M be a k-armed BSP with components $\{M_i\}$ and state space $\mathcal{S}$. If each M_i has a dominating policy, then*

$$\forall s \in \mathcal{S}, \forall \rho \geq 0, \hat{V}(s, \rho) = V(s, \rho).$$

Theorem 2. *Let X, Y be Markov reward processes. Let $Z = X + Y$ be the 2-armed bandit that corresponds to their sum. $\forall s = \{s_X, s_Y\} \in \mathcal{S}_Z$,*

$$V_Z(s) = \sum_{\rho \in \mathcal{C}(Y)} V_X(s_X, \rho) \Delta^Y(s_Y, \rho). \tag{3}$$

Proof. For a fixed policy and stopping rule, $V_Y(s_Y, \cdot)$ is linear. Thus, $\rho \notin \mathcal{C}(Y)$ implies $\Delta^Y(s_Y, \rho) = 0$.

$$\Delta_\rho^Y(s, x) = \left\{ \begin{array}{cc} \Delta^Y(s, x) & x > \rho \\ \sum_{x' \leq x} \Delta^Y(s, x') & x = \rho \\ 0 & x < \rho \end{array} \right. .$$

$$I_s = \max_{s' \in \{s_X, s_Y\}} I_{s'}$$

Integrating Eq. 2 by parts, we have

$$\begin{aligned}
\hat{V}_Z(s, \rho) &= V_X(s_X, \rho) \frac{\partial V_Y}{\partial \rho}(s_Y, \rho) + \int_\rho^{I_s} V_X(s_x, \alpha) \frac{\partial^2 V_Y}{\partial \rho^2}(s_Y, \alpha) d\alpha \\
&= V_X(s_X, \rho) \frac{\partial V_Y}{\partial \rho}(s_Y, \rho) + \sum_{\{\alpha \in \mathcal{C}(Y) | \alpha \geq \rho\}} V_x(s_x, \alpha) \lim_{\delta \to 0} \int_{\alpha - \delta}^{\alpha + \delta} \frac{\partial^2 V_Y}{\partial \rho^2}(s, \beta) d\beta \\
&= V_X(s_X, \rho) \frac{\partial V_Y}{\partial \rho}(s_Y, \rho) + \sum_{\{\alpha \in \mathcal{C}(Y) | \alpha \geq \rho\}} V_X(s_X, \alpha) \Delta^Y(s_Y, \alpha)
\end{aligned}$$

$$\frac{\partial V_Y}{\partial \rho}(s_Y, \rho) = \frac{\partial V_Y}{\partial \rho}(s_Y, \rho) - \frac{\partial V_Y}{\partial \rho}(s_Y, 0) = \int_0^\rho \frac{\partial^2 V}{\partial \rho^2}(s_Y, \alpha) d\alpha = \Delta_\rho^Y(s, \rho) \tag{4}$$

$$V_Z(s, \rho) = \sum_{\alpha \in \mathcal{C}(Y)} V_X(s_X, \alpha) \Delta_\rho^Y(s, \alpha)$$

Where the equality in 4 holds because $\frac{\partial V_Y}{\partial \rho}(s_Y, 0) = 0$. Furthermore,

$$\frac{\partial V_Y}{\partial \rho}(s_Y, I_s) = 1.$$

Thus

$$\sum_{0\leq m\leq I_s} \Delta^Y_\rho(s_Y, m) = \sum_{0\leq m\leq I_s} \Delta^Y(s_Y, m) = \int_{m=0}^{I_s} \frac{\partial^2 V}{\partial \rho^2}(s_Y, m)dm = 1.$$

Thus, we can see that $V_Z(s, \rho)$ is a convex combination of $V_X(s_X, \alpha)$ for $\alpha \in \rho$, the relative weights of which are specified by Δ^Y_ρ. Taking $\rho = 0$ and invoking Theorem 1 shows the result. □

3 Dominated Relaxation

In this section, we define our primary theoretical result: a relaxation for the arms of a BSP so that a dominating policy exists. We prove that the retirement process value function of the relaxed MDP is equivalent to the Whittle integral. This allows us to derive a novel proof that the Whittle integral is an upper bound.

Definition 7. *(Dominated Relaxation of an MDP) Let M be an MDP with discount factor γ and state space $\mathcal{S}$. Let s be a state in M. The dominated relaxation of P for s, $M_D(s)$ is a semi-Markov decision process that fixes s as an initial state. Let we let $\{\pi_i\}$ be polices that are optimal for some ρ: $\{\pi^*_\rho | \rho \in \mathcal{C}(M)\}$. This sequence is ordered so that $\rho_-(\pi_i)$ is increasing in i.*

For each i, we introduce a copy of the state space, $\mathcal{S}_i$, where the agent is restricted to following π_i. Let s'_i be the analogue of s' in $\mathcal{S}_i$. For $s'_i \in \tau_{\rho_-(\pi_i)}$, we introduce a single durative action, a_i, that takes the agent from $\mathcal{S}_i$ to $\mathcal{S}_{i-1}$ and characterize it as follows:

- $R(s'_i) = R_{\rho_+(\pi_{i-1})}(s) - R_{\rho_-(\pi_i)}(s)$
- $T(s'_i, a_i, s''_{i-1}) = P_{retire}(s'' | \pi_{i-1}, \rho_+(\pi_{i-1}), s)$
- $\delta(a_i) = \log \mathbb{E}[\gamma^{\tau_{\rho_+(\pi_{i-1})}(s)}] - \log \mathbb{E}[\gamma^{\tau_{\rho_-(\pi_i)}(s)}]$

Finally, for each i, we introduce an action that transitions from s to s_i with $\delta = 0$. We will write $V_D(s)$ to represent the value of s in $M_D(s)$.

Theorem 3. *Let M be an MDP with state space $\mathcal{S}$. The following statements are true for $s \in \mathcal{S}$ and $\rho \geq 0$:*

1. *$M_D(s)$ satisfies the Whittle condition.*
2. *$V_D(s, \rho) = \hat{V}(s, \rho)$.*
3. *$\hat{V}(s, \rho) \geq V(s, \rho)$*

Proof. 1. By construction, $\rho_-(\pi_i) = \rho_+(\pi_{i-1})$. We denote this as ρ_i. Let π_i' be a policy in M_D that initially transitions to s_i, and then transitions to $\mathcal{S}_{i-1}$. We have

$$\begin{aligned}
V^{\pi_i'}(s, \rho_i) &= R_{\rho_-(\pi_i)}(s) + \mathbb{E}[R(s_i') + \gamma^{\tau_{\rho_i}^{\pi_i}(s)+\delta(a')}\rho_i] \\
&= R_{\rho_-(\pi_i)}(s) + R_{\rho_+(\pi_{i-1})}(s) - R_{\rho_-(\pi_i)}(s) + \mathbb{E}[\gamma^{\tau_{\rho_i}^{\pi_i}(s)}]\mathbb{E}[\gamma^{\delta(a')}]\rho_i \\
&= R_{\rho_+(\pi_{i-1})}(s) + \mathbb{E}[\gamma^{\tau_{\rho_i}^{\pi_i}(s)}]\frac{\mathbb{E}[\gamma^{\tau_{\rho_+(\pi_{i-1})}(s)}]}{\mathbb{E}[\gamma^{\tau_{\rho_-(\pi_i)}(s)}]}\rho_i \\
&= V^{\pi_{i-1}}(s, \rho_i)
\end{aligned}$$

Thus, the agent is indifferent between π_i, π_{i-1}, and π_i' at ρ_i. By definition, the distribution over states for π_i' after transitioning to $\mathcal{S}_{i-1}$ is identical to that of π_{i-1}. The above reasoning also shows that the value of the discount parameter is the same. This is sufficient to conclude that

$$\forall \rho \in [\rho_{i-1}, \rho_{i-2}],\ V^{\pi_i'}(s, \rho) = V^{\pi_{i-1}}(s, \rho). \tag{5}$$

This is sufficient to show that a policy that transitions to s_0 initially and then takes a durative action at each stopping state is optimal for all ρ. Thus, M_D satisfies the Whittle condition.

2. Eq. 5, it is easy to see that $V^{\pi_k'}(s, \rho) = V(s, \rho)$. An appeal to Eq. 2 allows us to conclude (2).

3. The optimal policy for M is a feasible policy for M_D. Thus, V is a lower bound on V_D. This allows us to conclude (3) from (2).

□

4 Branch and Bound Value Iteration

In this section, we prove correctness for Branch and Bound Value Iteration (BBVI). Pseudocode for BBVI can be found in Algorithm 2.

First, we show that the upper and lower bounds BBVI computes (Q^+, Q^-) actually bound the value function.

Theorem 4. *Let M be an MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$. Let $\mathcal{S}' \subseteq \mathcal{S}$ where $\mathcal{S}' = \mathcal{U} \cup \mathcal{E}$ and $\forall\ s \in \mathcal{E}, T(s, a, s') \neq 0 \Rightarrow s' \in \mathcal{S}'$. Let M_{UB} be an MDP with state space $\mathcal{S}' \cup \{\alpha\}$, with identical transition distribution and rewards for states in $\mathcal{E}$(expanded states), with each state in $s \in \mathcal{U}$ transitioning deterministically to α and receiving reward that is an upper bound on $V^*(s)$. Let α be a sink state with 0 reward. The value of any state in P_{UB} is an upper bound on the value of the corresponding state in the original MDP.*

Proof. Initialize value iteration with upper bounds on the value at each state: $\forall s \in \mathcal{S}', V_0(s) \geq V^*(s)$. Initialize $V_0(\alpha) = 0$. Consider the initial Bellman backup for state

Algorithm 2 Branch-and-Bound Value Iteration

```
Define:BBVI($\langle M_0, \ldots, M_K \rangle, s_0, \epsilon$)
Input: BSP arms, $M_k$; Initial state, $s_0$; Tolerance, $\epsilon$
# Lower bound $M$ by fixing a policy for each arm
$LB_k \leftarrow$toMRP($M_k$)
Compute critical points for $M_k, LB_k$
Compute bounds on $Q(s_0, \cdot)$ with Whittle integrals for $M$ and $LB$.
# Keep track of expanded region of state space
$\mathcal{E} \leftarrow \emptyset$
# Keep track of states at boundary of $M^+, M^-$
$\mathcal{B} \leftarrow \{s_0\}$
$a^* \leftarrow \underset{a}{\text{argmax}} Q^-(s_0, a)$
while $\exists a' \neq a^*\ s.t\ Q^+(s_0, a') \geq Q^-(s_0, a^*)$ do
  $s \leftarrow$pop($\mathcal{B}$)
  $\mathcal{E} \leftarrow \mathcal{E} \cup \{s\}$
  for $a \in \mathcal{A}$ do
    for $s' \in successors(s)$ do
      Compute upper and lower bounds for $Q(s', \cdot)$
      $\mathcal{B} \leftarrow \mathcal{B} \cup \{s'\}$
    end for
  end for
  $Q^+ \leftarrow$ SOLVE(BOUNDMDP($\mathcal{E}, \mathcal{B}, Q^+$))
  $Q^- \leftarrow$ SOLVE(BOUNDMDP($\mathcal{E}, \mathcal{B}, Q^-$))
  $a^* \leftarrow \underset{a}{\text{argmax}} Q^-(s_0, a)$
end while
return $\epsilon$-optimal action $a^*$
```

$s \in \mathcal{E}$.

$$\begin{aligned}
V_1(s) &= \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}'} T(s,a,s')[R(s,a,s') + \gamma V_0(s')] \\
&\geq \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')] \\
&= \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} T(s,a,s')[R(s,a,s') + \gamma V^*(s')] \\
&= V^*(s)
\end{aligned}$$

Thus, for expanded states, the backup operator preserves upper bounds. Furthermore, the backup operation for states in $\mathcal{U}$ does not change the value and so trivially preserves bounds. Thus at all iterations, the current estimate of the value function is an upper bound. The convergence of value iteration shows the desired result. □

Theorem 5. *Let M be a BSP and let s be a state in M. Let a be the action returned by $BBVI(M, s, \epsilon)$.*

$$V(s) - Q(s,a) < \epsilon.$$

Proof. We proceed by contradiction. This implies that there exists and a' such that $Q(s,a') - Q(s,a) \geq \epsilon$. At termination, all actions other that a have been pruned from s. Thus, $Q_U(s,a') - Q_L(s,a) < \epsilon$. This implies that $Q_L(s,a) + \epsilon > Q_U(s,a') = V(s)$. Thus,

$$\begin{aligned}
Q_U(s,a) &> V(s) - \epsilon \\
Q_U(s,a) - V(s) &> -\epsilon \\
V(s) - Q_U(s,a) &< \epsilon
\end{aligned}$$

But this contradicts Theorem 4, which implies that $\epsilon > V(s) - Q_U(s,a)$. □

References

Peter Nash. *Optimal allocation of Resources Between Research Projects.* PhD thesis, University of Cambridge, 1973.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2009.

Peter Whittle. Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 143–149, 1980.