Proof of the Bound on the Suboptimality of Some Symmetric Policy in an Asymmetric Markov Decision Process (Draft Version)

Martin Zinkevich, Tucker Balch

May 27, 2001

1 Introduction

In this manuscript we present a proof of the theorems presented in (Zinkevich & Balch, 2001a). Specifically, we wish to prove an additive bound on the suboptimality of some symmetric policy for an asymmetric Markov decision process(hereafter referred to as MDP I). The structure of the proof is as follows:

1. Construct a related MDP(called MDP II) which has a symmetric reward function.
   (a) Prove that for every policy, the reward when used in MDP II is no better than in MDP I.
   (b) Prove that for every policy, the reward when used in MDP I is no better than the reward when used in MDP I plus some additive factor.
2. Construct a related MDP(called MDP III) from MDP II which has a symmetric transition function.
   (a) Prove that for every policy, the reward when used in MDP III is no better than in MDP II.
   (b) Prove that for every policy, the reward when used in MDP II is no better than the reward when used in MDP III plus some additive factor.
3. Refer to (Zinkevich & Balch, 2001b) to prove that the there exists a symmetric optimal policy for MDP III.
4. Use this to prove that there exists a symmetric policy in MDP I is within the sum of these two additive factors of the optimal policy for MDP I.

For an introduction to the concept of symmetries in MDPs, and for the definition of terms not introduced here, please see (Zinkevich & Balch 2001b).

2 MDP II: Symmetric Reward Function

How does one construct a symmetric reward function from an asymmetric reward function? Although there are several conceivable ways of doing so, the constraint that the resulting MDP must be strictly worse suggests that we should decrease the reward function at certain points. The restriction that it must be symmetric suggests that we decrease the reward for a state-action pair to the minimum reward for any symmetric state-action pair. The following lemma clarifies this.

Lemma 1 *Given a MDP $(\mathcal{S}, \mathcal{A}, T, R)$, and a symmetry $(E_{\mathcal{S}}, E_{\mathcal{A}})$, one can construct a new MDP $(\mathcal{S}, \mathcal{A}, T, R')$ where*

$$R'(s,a) = \min_{(s',a') \in E_{\mathcal{A}}(s,a)} R(s',a')$$

This new MDP has a symmetric reward function. For any policy σ, for any state $s \in \mathcal{S}$:

$$V_\sigma^I(s) \geq V_\sigma^{II}(s)$$

$$V_\sigma^{II}(s) \geq V_\sigma^I(s) + \frac{\Delta R}{1-\gamma}$$

where V_σ^I is the expected discounted reward function for MDP I, V_σ^{II} is the expected discounted reward function for MDP II, and $\Delta R = \max_{((s,a),(s',a')) \in E_{\mathcal{A}}} (R(s',a') - R(s,a))$.

Proof:

Observe that for all $s \in \mathcal{S}$, for all $a \in \mathcal{A}$:

$$R(s,a) \geq R'(s,a) \geq R(s,a) - \Delta R$$

The remainder of the proof is an extension of this from the reward function to the expected discounted reward function. Suppose that for policy σ and transition function T, $P_{s's} = T(s, \sigma(s), s')$. Then the expected discounted reward function for σ in MDP I is:

$$V_\sigma^I(s) = \sum_{t=0}^{\infty} \sum_{s' \in \mathcal{S}} \gamma^t (P^t)_{s's} R(s', \sigma(s'))$$

$$V_\sigma^I(s) \geq \sum_{t=0}^{\infty} \sum_{s' \in \mathcal{S}} \gamma^t (P^t)_{s's} R'(s', \sigma(s'))$$

$$V_\sigma^I(s) \geq V_\sigma^{II}(s)$$

This was the first part which we wished to prove.

$$V_\sigma^{II}(s) = \sum_{t=0}^{\infty} \sum_{s' \in \mathcal{S}} \gamma^t (P^t)_{s's} R'(s', \sigma(s'))$$

$$V_\sigma^{II}(s) \geq \sum_{t=0}^{\infty} \sum_{s' \in \mathcal{S}} \gamma^t (P^t)_{s's} (R(s', \sigma(s')) - \Delta R)$$

$$V_\sigma^{II}(s) \geq \sum_{t=0}^{\infty} \sum_{s' \in \mathcal{S}} \gamma^t (P^t)_{s's} R(s', \sigma(s')) - \sum_{t=0}^{\infty} \sum_{s' \in \mathcal{S}} \gamma^t (P^t)_{s's} \Delta R$$

$$V_\sigma^{II}(s) \geq V_\sigma^{I}(s) - \Delta R \sum_{t=0}^{\infty} \gamma^t \sum_{s' \in \mathcal{S}} (P^t)_{s's}$$

Because P is a probability transition matrix:

$$V_\sigma^{II}(s) \geq V_\sigma^{I}(s) - \Delta R \sum_{t=0}^{\infty} \gamma^t (1)$$

$$V_\sigma^{II}(s) \geq V_\sigma^{I}(s) - \frac{\Delta R}{1-\gamma}$$

∎

Observe that MDP II has a symmetric reward function, as desired.

3 MDP III: Symmetric Transition Function

Constructing a symmetric transition function is a bit more complex: one cannot simply reduce the probabilities in a distribution, because the result would not be a probability distribution. However, what if just append to this result a "bad" circumstance, a transition to a really "bad" state. How bad can a state be? First of all, the reward in this bad state should be no more than the lowest possible reward in the system. Secondly, the state should be inescapeable. insuring the worst possible future reward.

Before considering how to make the transition function symmetric, first let us consider how to make a probability distribution symmetric.

Definition 1 *Two random variables, $\mathbb{S}$ and $\mathbb{S}'$ defined over the set $\mathcal{S}$ are* ***symmetric*** *if for all $s \in \mathcal{S}$:*

$$Pr[\mathbb{S} \in E_{\mathcal{S}}(s)] = Pr[\mathbb{S}' \in E_{\mathcal{S}}(s)].$$

Two probability distributions are ***symmetric*** *if the associated random variables are symmetric.*

So, in order to make two probability distributions symmetric, their mass on each equivalence class of states must be made equivalent. So, in this simple case, for each equivalence class of states, we look for the probability distribution which has the minimum mass on that class, and then we "reduce" the mass of the other probability distribution on that class to the same value. We then increase the transition probability to the "bad" state by this value.

Observe that there can be multiple probability distributions resulting from reducing the mass, because one could divide the reduction among the probabilities on the states in that reduction in many ways. However, it is important to note that it is unnecessary to make any of these probabilities negative: because there was more mass on this equivalence class than the other to begin with, there will be some mass left over. As opposed to explicitly constructing a probability distribution in this fashion, we simply state that one such probability distribution exists.

Definition 2 *A set R of random variables over the set $\mathcal{S}$ is **Δ-asymmetric** if:*

$$\Delta = \max_{\mathbb{S}\in R}\Big(\sum_{P\in\{E_S(s)|S\in\mathcal{S}\}}(Pr[\mathbb{S}\in P] - \min_{\mathbb{S}'\in R} Pr[\mathbb{S}'\in P]\Big))$$

Lemma 2 *Given a set $\{\mathbb{S}_1,\ldots,\mathbb{S}_n\}$ of Δ-asymmetric random variables over the set $\mathcal{S}$, one can construct a set of symmetric random variables $\{\mathbb{S}'_1,\ldots,\mathbb{S}'_n\}$ over the set $S\cup b$, where for all i, for all $s\in\mathcal{S}$, $Pr[s=\mathbb{S}_i]\geq Pr[s=\mathbb{S}'_i]$, and for all i, $Pr[\mathbb{S}'_i=b]\leq\Delta$.*

Proof: By a generalization of the above argument on two variables. ∎

Lemma 3 *Given a set $\{\mathbb{S}_1,\ldots,\mathbb{S}_n\}$ of Δ-asymmetric random variables over the set $\mathcal{S}$, one can construct a set of symmetric random variables $\{\mathbb{S}'_1,\ldots,\mathbb{S}'_n\}$ over the set $S\cup b$, where for all i, for all $s\in\mathcal{S}$, $Pr[s=\mathbb{S}_i]\geq Pr[s=\mathbb{S}'_i]$, and for all i, $Pr[\mathbb{S}'_i=b]=\Delta$.*

Note that now we claim that the $Pr[\mathbb{S}'_i=b]$ is equal for all i. This follows from the fact that for any i,j, for any $s\in\mathcal{S}$, $Pr[\mathbb{S}'_i\in E_{\mathcal{S}}(s)]=Pr[\mathbb{S}'_j\in E_{\mathcal{S}}(s)]$. Since the probability of b is the "leftover" probability mass, it must be equal. ∎

Definition 3 *For each state-action pair (s,a), define $\mathbb{S}_{s,a}$ to be a random variable such that for all $s'\in\mathcal{S}$, $Pr[\mathbb{S}_{s,a}=s']=T(s,a,s')$. Then, one can consider the set of random variables $\{\mathbb{S}_{s',a'}|(s',a')\in E_{\mathcal{A}}(s,a)\}$. Define ΔT to be the maximum asymmetry of any such set. This is the **asymmetry of the transition function** T.*

Observe that if and only if the sets of random variables are symmetric, the transition function is symmetric. Thus, a symmetric transition function has an asymmetry of 0.

Lemma 4 *Given a MDP II $(\mathcal{S},\mathcal{A},T,R')$, and an equivalence relation $(E_{\mathcal{S}},E_{\mathcal{A}})$, where the asymmetry of T is ΔT, construct a set $\mathcal{S}'=\mathcal{S}\cup b$. For each state-action pair (s,a), define $\mathbb{S}_{s,a}$ to be a random variable such that for all $s'\in\mathcal{S}$, $Pr[\mathbb{S}_{s,a}=s']=T(s,a,s')$. Then, one can consider the set of random variables $\{\mathbb{S}_{s',a'}|(s',a')\in E_{\mathcal{A}}(s,a)\}$. For each such set, construct a set of symmetric random variables according to Lemma 2, and add a state b. For all actions taken at b, there is a probability of 1 of the next state being b. Label the resulting* *transition function $T' : \mathcal{S}' \times \mathcal{A} \to \mathcal{S}'$. Construct a new reward function $R'' : \mathcal{S}' \times \mathcal{A} \to \mathbb{R}$ where for all $s \in \mathcal{S}$, for all $a \in \mathcal{A}$, $R''(s,a) = R'(s,a)$, and $R''(b,a) = \min_{s' \in \mathcal{S}, a' \in \mathcal{A}} R'(s',a')$. Call the MDP $(\mathcal{S}', \mathcal{A}, T', R'')$ MDP III. Then for any policy $\sigma : \mathcal{S}' \to \mathcal{A}$*[1]*, for any state $s \in \mathcal{S}$:*

$$V_\sigma^{II}(s) \geq V_\sigma^{III}(s)$$

$$V_\sigma^{III}(s) \geq V_\sigma^{II}(s) - \frac{|R'|\Delta T \gamma}{(1-\gamma)(1-(1-\Delta T)\gamma)}$$

where $|R'| = (\max_{s \in \mathcal{S}, a \in \mathcal{A}} R'(s,a)) - (\min_{s \in \mathcal{S}, a \in \mathcal{A}} R'(s,a))$.

First, without loss of generality, assume that for all $s \in \mathcal{S}$, for all $a \in \mathcal{A}$, $R'(s,a) \geq 0$ and $\min_{s' \in \mathcal{S}, a' \in \mathcal{A}} R'(s',a') = 0$. The general case can be converted to this case by adding $-\min_{s' \in \mathcal{S}, a' \in \mathcal{A}} R'(s',a')$ to every reward. This does not change $|R'|$, nor does it change the difference between MDP II and MDP III.

For all $s, s' \in \mathcal{S}$, define $P_{s's} = T(s, \sigma(s), s')$. For all $s, s' \in \mathcal{S}'$, define $P'_{s's} = T'(s, \sigma(s), s')$. These are the probability transition matrices. Observe that for all $s, s' \in \mathcal{S}$, $P_{s's} \geq P'_{s's}$. For all $s \in \mathcal{S}$:

$$V_\sigma^{II}(s) - V_\sigma^{III}(s) = \left(\sum_{t=0}^{\infty} \sum_{s' \in \mathcal{S}} \gamma^t (P^t)_{s's} R'(s', \sigma(s')) \right) - \left(\sum_{t=0}^{\infty} \sum_{s' \in \mathcal{S}'} \gamma^t ({P'}^t)_{s's} R''(s', \sigma(s')) \right)$$

Let us construct a matrix which is not a transition matrix: for all $s, s' \in \mathcal{S}$, define $Q_{s,s'} = T'(s, \sigma(s), s')$. The difference between P' and Q is that Q is only defined on $\mathcal{S} \times \mathcal{S}$, and therefore does not consider b. Observe that, since the reward received in b is zero and the state is inescapeable, we can replace P' with Q in the above equation without affecting the value.

$$V_\sigma^{II}(s) - V_\sigma^{III}(s) = \left(\sum_{t=0}^{\infty} \sum_{s' \in \mathcal{S}} \gamma^t (P^t)_{s's} R'(s', \sigma(s')) \right) - \left(\sum_{t=0}^{\infty} \sum_{s' \in \mathcal{S}} \gamma^t (Q^t)_{s's} R''(s', \sigma(s')) \right)$$

For all $s' \in \mathcal{S}$, for all $a \times \mathcal{A}$, $R''(s',a) = R'(s',a)$.

$$V_\sigma^{II}(s) - V_\sigma^{III}(s) = \left(\sum_{t=0}^{\infty} \sum_{s' \in \mathcal{S}} \gamma^t (P^t)_{s's} R'(s', \sigma(s')) \right) - \left(\sum_{t=0}^{\infty} \sum_{s' \in \mathcal{S}} \gamma^t (Q^t)_{s's} R'(s', \sigma(s')) \right)$$

$$V_\sigma^{II}(s) - V_\sigma^{III}(s) = \sum_{t=0}^{\infty} \sum_{s' \in \mathcal{S}} \gamma^t \left((P^t)_{s's} - (Q^t)_{s's} \right) R'(s', \sigma(s'))$$

Observe that for all $s, s' \in \mathcal{S}$, for all $t \geq 0$, $(P^t)_{s's} \geq (Q^t)_{s's}$. This establishes $V_\sigma^{II}(s) - V_\sigma^{III}(s) \geq 0$, so

$$V_\sigma^{II}(s) \geq V_\sigma^{III}(s)$$

[1]Technically, σ is not a policy on MDP II, because its domain is larger than the set of states. $\sigma|_{\mathcal{S}}$, the restriction of σ to the domain $\mathcal{S}$, is a policy on MDP II. However, we surpress this notation because the intention is clear.

However, since for every addend, $(P^t)_{s's} - (Q^t)_{s's} \geq 0$, if we increase R', we will increase the value of the right side. Since $\min_{s' \in \mathcal{S}, a \in \mathcal{A}} R'(s,a) = 0$, $|R'| = \max_{s' \in \mathcal{S}, a \in \mathcal{A}} R'(s,a)$. So:

$$V_\sigma^{II}(s) - V_\sigma^{III}(s) \leq \sum_{t=0}^{\infty} \sum_{s' \in \mathcal{S}} \gamma^t \left((P^t)_{s's} - (Q^t)_{s's}\right) |R'|$$

$$V_\sigma^{II}(s) - V_\sigma^{III}(s) \leq |R'| \sum_{t=0}^{\infty} \gamma^t \left(\left(\sum_{s' \in \mathcal{S}} (P^t)_{s's}\right) - \left(\sum_{s' \in \mathcal{S}} (Q^t)_{s's}\right)\right)$$

Because P is probability transition matrix, for all $s \in \mathcal{S}$, for all $t \geq 0$, $\sum_{s' \in \mathcal{S}} (P^t)_{s's} = 1$.

$$V_\sigma^{II}(s) - V_\sigma^{III}(s) \leq |R'| \sum_{t=0}^{\infty} \gamma^t \left(1 - \left(\sum_{s' \in \mathcal{S}} (Q^t)_{s's}\right)\right)$$

Since in MDP III the probability of transitioning to any state to b is less than ΔT, for all $s \in \mathcal{S}$, for all $t \geq 0$, $\sum_{s' \in \mathcal{S}} (Q^t)_{s's} \geq (1 - \Delta T)^t$.

$$V_\sigma^{II}(s) - V_\sigma^{III}(s) \leq |R'| \sum_{t=0}^{\infty} \gamma^t \left(1 - (1 - \Delta T)^t\right)$$

After some algebraic manipulation:

$$V_\sigma^{III}(s) \leq V_\sigma^{II}(s) - \frac{|R'| \Delta T \gamma}{(1-\gamma)(1-(1-\Delta T)\gamma)}$$

■

Now we will state explicitly the main theorem, followed by the proof we outlined in the introduction.

Now we combine these two lemmas into one cohesive piece.

Lemma 5 *Given an MDP* (S, A, T, R) *and a symmetry* $(E_{\mathcal{S}}, E_{\mathcal{A}})$ *where* T *has an asymmetry of* ΔT, $\Delta R = \max_{((s,a),(s',a')) \in E_{\mathcal{A}}} R(s',a') - R(s,a)$, $|R| = (\max_{s \in \mathcal{S}, a \in \mathcal{A}} R(s,a)) - (\min_{s \in \mathcal{S}, a \in \mathcal{A}} R(s,a))$, *if for any policy* σ V_σ^I *is the reward function for MDP I and* V_σ^{III} *is the reward function for MDP III:*

$$V_\sigma^I(s) \geq V_\sigma^{III}(s)$$

$$V_\sigma^{III}(s) \geq V_\sigma^I(s) - \frac{\Delta R}{1-\gamma} - \frac{|R| \Delta T \gamma}{(1-\gamma)(1-(1-\Delta T)\gamma)}$$

Proof:

By combining the Lemma 1, Lemma 4, and the fact that $|R| \geq |R'|$. ■

Theorem 1 *Given an MDP* (S, A, T, R) *and a symmetry* $(E_{\mathcal{S}}, E_{\mathcal{A}})$ *where* T *has an asymmetry of* ΔT, $\Delta R = \max_{((s,a),(s',a')) \in E_{\mathcal{A}}} R(s',a') - R(s,a)$, $|R| =$

$(\max_{s \in \mathcal{S}, a \in \mathcal{A}} R(s,a)) - (\min_{s \in \mathcal{S}, a \in \mathcal{A}} R(s,a))$, *there exists a symmetric policy sym such that for all* $s \in \mathcal{S}$*:*

$$V_{sym}(s) \geq V^*(s) - \frac{\Delta R}{1-\gamma} - \frac{|R|\Delta T \gamma}{(1-\gamma)(1-(1-\Delta T)\gamma)}$$

$V^*(s)$ *being the optimal discounted reward achievable from state* s *in MDP I.*

So, first we construct MDP II, and then construct MDP III as in the above lemmas. Is MDP III symmetric? Well, strictly speaking $E_{\mathcal{S}}$ is not defined on $\mathcal{S}' \times \mathcal{S}'$. Therefore we must extend the symmetry. Define $E_{\mathcal{S}'} = E_{\mathcal{S}} \cup \{(b,b)\}$. That is, b is only symmetric to itself. Similarly, define $E_{\mathcal{A}'} = E_{\mathcal{A}} \cup \{((b,a),(b,a))|a \in \mathcal{A}\}$. All actions in b are only symmetric to themselves. Hence, if R' is symmetric with respect to $(E_{\mathcal{S}}, E_{\mathcal{A}})$, then R'' is symmetric with respect to $(E_{\mathcal{S}'}, E_{\mathcal{A}'})$. Similarly, due to the result of Lemma 3, T' is symmetric with respect to $(E_{\mathcal{S}'}, E_{\mathcal{A}'})$. Thus, MDP III is symmetric with respect to $(E_{\mathcal{S}'}, E_{\mathcal{A}'})$. So there exists a policy sym which is symmetric with respect to $(E_{\mathcal{S}'}, E_{\mathcal{A}'})$.

Is this policy, when restricted to the domain $\mathcal{S}$, symmetric with respect to $(E_{\mathcal{S}}, E_{\mathcal{A}})$? Since $E_{\mathcal{A}} \subseteq E_{\mathcal{A}'}$, this directly follows from the definition of a symmetric policy. Observe that from Lemma 5, for all $s \in \mathcal{S}$:

$$V^I_{sym}(s) \geq V^{III}_{sym}(s)$$

Consider σ^* to be the optimal policy for MDP I. By Lemma 5, for all $s \in \mathcal{S}$:

$$V^{III}_{\sigma^*}(s) \geq V^I_{\sigma^*}(s) - \frac{\Delta R}{1-\gamma} - \frac{|R|\Delta T \gamma}{(1-\gamma)(1-(1-\Delta T)\gamma)}$$

Because σ^* is optimal in MDP I:

$$V^{III}_{\sigma^*}(s) = V^*(s) - \frac{\Delta R}{1-\gamma} - \frac{|R|\Delta T \gamma}{(1-\gamma)(1-(1-\Delta T)\gamma)}$$

Because sym is optimal in MDP III:

$$V^{III}_{sym}(s) \geq V^{III}_{\sigma^*}(s)$$

Combining:

$$V^I_{sym}(s) \geq V^*(s) - \frac{\Delta R}{1-\gamma} - \frac{|R|\Delta T \gamma}{(1-\gamma)(1-(1-\Delta T)\gamma)}$$

Which was what we intended to prove. ∎

4 Conclusion

This proof is a guideline for the viability of symmetric policies in asymmetric MDPs. I believe that there are a few points that should be highlighted about the nature of the additive bound. First of all, it is not the number of deviations from symmetry which is important, but their magnitude. In order to understand why this is so, consider a line on which there exists one point where one can receive a reward. If one changes that reward, all of the expected discounted rewards from every state change. Similarly, if there were only one transition to that state, changing that transition would again affect every expected discounted reward. On the other hand, perhaps there is some large change in a reward from some inaccessible state, or the transition probabilities change in an action that is not used. It is difficult to clarify which rewards and transitions are "important" and "unimportant" without explicitly solving for all the optimal policies.

Therefore, a more useful form of the above theorem is as follows. If actions which are equivalent under the symmetry have only slight asymmetries in the transition function and the reward function, regardless of the number of such asymmetric actions, there will be a "good" symmetric policy in the Markov decision process.

References

Zinkevich, M. & Balch, T. (2001a). Using Symmetry in Asymmetrical Markov Decision Processes (Extended Abstract). *Fifth International Conference on Autonomous Agents: Workshop on Learning Agents.*

Zinkevich, M. & Balch, T. (2001b). Symmetry in Markov Decision Processes and its Implications for Single Agent and Multiagent Learning.*the Eighteenth International Conference on Machine Learning.*