Information Relaxation Bounds for Infinite Horizon Markov Decision Processes

(Online Appendix)

David B. Brown
Fuqua School of Business
Duke University
dbbrown@duke.edu

Martin B. Haugh
Department of IE&OR
Columbia University
mh2078@columbia.edu

This version: April 28, 2016
First version: September 6, 2014

B. Further Examples

We provide some additional instructive examples in this online appendix. In Appendix B.1 we consider a classic DP problem where the optimal policy is easy to find. The goal here is to provide explicit recursions for the inner problems under perfect and imperfect information relaxations for both controlled and uncontrolled formulations. We also demonstrate here the potential of the truncated horizon approach that was discussed in Section 4.4. In Appendix B.2 we consider a simple die-throwing example and a particular information relaxation that renders the inner problem non-Markovian in the original state variables. We therefore need to *expand* the state space in order to obtain a Markovian inner problem that is amenable to value iteration. This example simply makes the point that an injudicious choice of information relaxation can result in an inner problem that is harder to solve than the original primal problem. In Appendix B.3 we provide a simple two-period example which shows (perhaps surprisingly) that uncontrolled formulation bounds can be superior to controlled formulation bounds given the same information relaxation and penalties.

B.1. A Simple Example: Machine Repair

We consider a classical example in which a decision maker is trying to determine the optimal repair policy for a machine that deteriorates over time. The machine can be in any one of n states, denoted by $1, \ldots, n$. In each period the decision maker can either attempt to repair the machine or not. The cost of attempting to repair the machine is $R \geq 0$. In addition, if x is the state in a given period then a cost of $c(x) \geq 0$ is incurred for that period in addition to the possible repair cost. Costs are increasing in x. We let P^r and $P^{\bar{r}}$ denote the transition probability matrices corresponding to the actions "repair" (r) and "do not repair" $(\bar{r})$, respectively. The objective is to minimize the total expected discounted cost over an infinite horizon. If n is not too large, solving for the optimal value function can be done easily (e.g., value iteration or by solving a linear program with n variables). Nonetheless, the example is instructive. The optimal value function $v^\star$ satisfies

$$v^\star(x) = \min\left\{R + c(x) + \delta P^r_x v^\star,\ c(x) + \delta P^{\bar{r}}_x v^\star\right\},$$

where P^r_x and $P^{\bar{r}}_x$ denote the x^{th} rows of P^r and $P^{\bar{r}}$, respectively, and $v^\star \equiv (v^\star(1)\ \ldots\ v^\star(n))^\top$.

In our examples, we use $n = 10$ states, with $c(x) = x$, $R = 10$ and $\delta = 0.75$. If a repair attempt is made, the next state will be state 1 with probability .9. The repair fails with probability .1, in which case the machine remains in its current state. Thus, $P^r(1,1) = 1$, and $P^r(x,1) = .9$ and $P^r(x,x) = .1$ for $x > 1$. If a repair attempt is not made, the machine may deteriorate to a worse (higher) state. We use $P^{\bar{r}}(x,x) = .8$, $P^{\bar{r}}(x,x+1) = P^{\bar{r}}(x,x+2) = .1$ for $x \leq n-2$, $P^{\bar{r}}(n-1,n-1) = .8$, $P^{\bar{r}}(n-1,n) = .2$ and $P^{\bar{r}}(n,n) = 1$.

Information Relaxation Bounds Via the Controlled Formulation

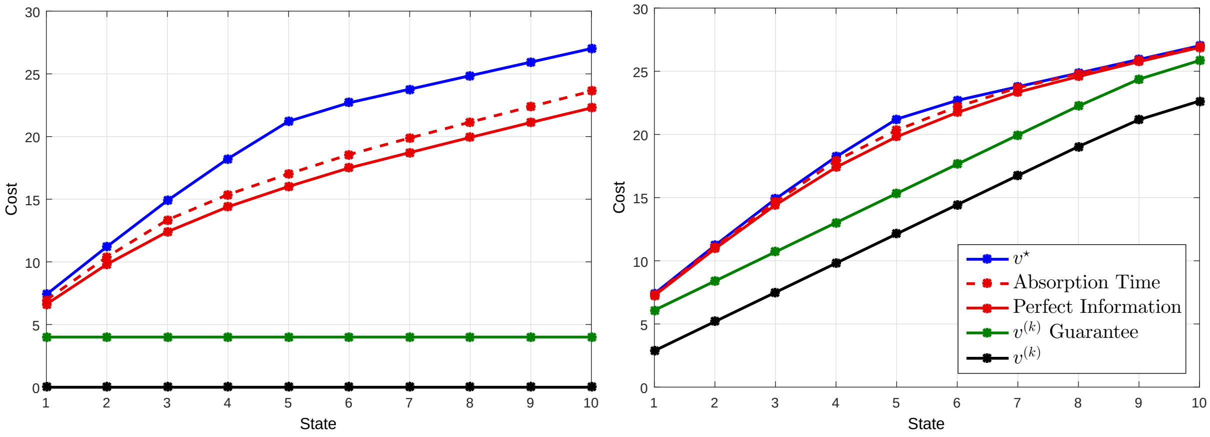
We calculate lower bounds using two information relaxations with the absorption time formulation. For both information relaxations, we use penalties based on approximate value functions calculated from value iteration. Specifically, we take $v = v^{(k)} := \mathscr{B}^k 0$ for some number $k \geq 0$ of iterations of the Bellman operator, $\mathscr{B}$, starting with the zero vector 0 when $k = 0$. Since costs are nonnegative, the choice $v^{(0)} = 0 \in \mathbb{R}^n$ is a subsolution; from monotonicity and convergence of the Bellman operator (e.g., Puterman 1994), v constructed in this way is a sub-solution for any $k \geq 0$.

In an earlier version of this paper, we showed a strengthened version of Proposition 4.1 with slightly stronger guarantees in some special cases. In particular, if v is a subsolution that satisfies $v + \epsilon \leq \mathscr{B}v$ in all states for some $\epsilon \geq 0$, it can be shown that

$$\mathbb{E}[v_0^{\mathbb{G}}(x_0)] \quad \geq \quad v(x_0) + \frac{\epsilon}{1 - \gamma\delta}, \tag{1}$$

where $\gamma := \min_{\{t, x \neq x^a, y \neq x^a, a \in A(x)\}} p(y|x,a)/q_t(y|x,a)$. Thus, if v is a subsolution with "slack" $\epsilon > 0$ and we construct a penalty from v in the information relaxation approach, we are guaranteed to obtain a strict improvement of $\epsilon/(1 - \gamma\delta)$ in the lower bound $v(x_0)$. This result also depends on the discount factor as well as the reformulation: if we use a controlled formulation (i.e., $q_t = p$), then $\gamma = 1$ and the guarantee is stronger. For uncontrolled formulations, γ may be close to zero, and the guarantee may be weak. We will include these bound guarantees from (1) in the example results that follow.

With a perfect information relaxation, the decision maker knows both the time horizon τ as well as the



(a) k=0

(b) k=3

Figure B.1: Lower bounds for the machine repair example using the controlled formulation and perfect and absorption time relaxations.

scenario $(w_1, \ldots, w_{\tau-1})$. An inner problem (8) is therefore obtained by first sampling τ from a geometric distribution with parameter $1 - \delta$ and then sampling $w = (w_1, \ldots, w_{\tau-1})$. Here (and without any loss of generality) we take the w_t's to be IID uniform $[0, 1]$ random variables and use inverse transforms on the state transition matrices $P^{\bar{r}}$ and P^r to generate state transitions for each action in each time period. We use $\Pi_t^{\bar{r}}$ and Π_t^r to denote the associated (deterministic) state transition matrices in each scenario, and again use the subscript x to denote the x^{th} row of these matrices. Solving the inner problem (8) in this case reduces to solving the deterministic DP

$$v_t^{\mathbb{G}}(x) \quad = \quad \min\Big\{c(x) + R + \underbrace{\left(\delta P_x^r - \Pi_{x,t}^r\right) v}_{\text{repair penalty}} + \Pi_{x,t}^r v_{t+1}^{\mathbb{G}},\ c(x) \ + \ \underbrace{\left(\delta P_x^{\bar{r}} - \Pi_{x,t}^{\bar{r}}\right) v}_{\text{do not repair penalty}} \ + \Pi_{x,t}^{\bar{r}} v_{t+1}^{\mathbb{G}}\Big\}$$

for $t = 0, \ldots, \tau - 1$, with $v_\tau^{\mathbb{G}} = 0$.

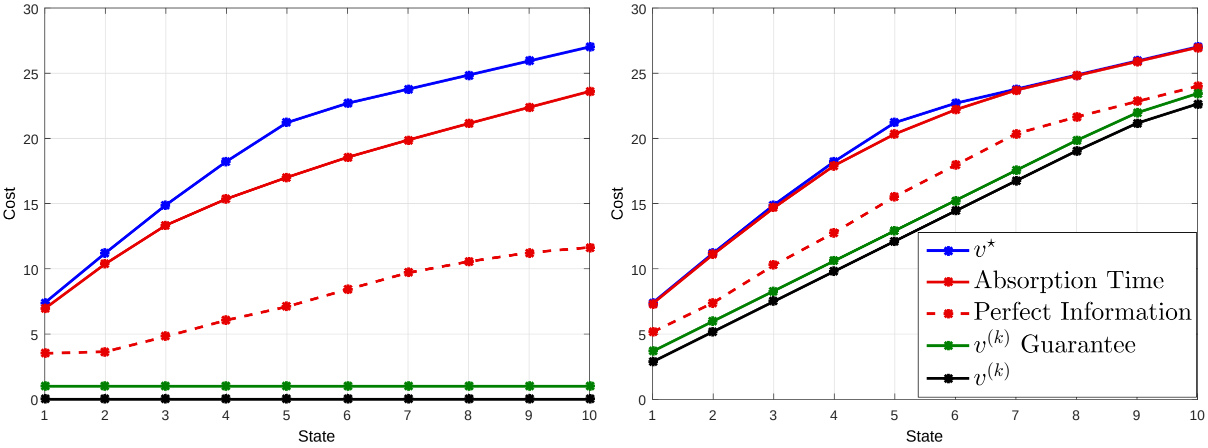
We also consider an imperfect information relaxation in which the decision maker knows τ at all times but does not know any machine state transitions after time t, i.e., $\mathscr{G}_t = \sigma(\mathscr{F}_t, \tau)$ for $t = 0, \ldots, \tau$. We call this the *absorption time relaxation*. This is a tighter information relaxation than perfect information and therefore must lead to better lower bounds. (This follows from a simple extension of Proposition 2.3(i) in BSS 2010). Generating inner problems under this relaxation reduces to only sampling τ from a geometric distribution with parameter $1 - \delta$ and then solving the τ-period inner problem. The inner problems (given by the right-hand side of (6)) are now finite horizon stochastic DPs and have the form

$$v_t^{\mathbb{G}}(x) \quad = \quad \min\Big\{c(x) + R + \underbrace{(\delta - 1) P_x^r v}_{\text{repair penalty}} + P_x^r v_{t+1}^{\mathbb{G}},\ c(x) \ + \ \underbrace{(\delta - 1) P_x^{\bar{r}} v}_{\text{do not repair penalty}} \ + P_x^{\bar{r}} v_{t+1}^{\mathbb{G}}\Big\}$$

for $t = 0, \ldots, \tau - 1$, with $v_\tau^{\mathbb{G}} = 0$.

Figure B.1 shows the bounds for all initial states. The optimal value function is indicated in blue, and the black line indicates $v = v^{(k)}$ for some number k of iterations of value iteration, starting from the zero vector. Since $v^{(k)}$ is a subsolution, we know from part (i) of Proposition 4.1 that the lower bounds from both information relaxations will be at least as good as the lower bounds $v^{(k)}$; the green line shows the bound improvement guarantee in (1). The solid and dashed red lines correspond to the lower bounds from the perfect and absorption time information relaxations, respectively.

There are several noteworthy features in these results. First, even when the penalties are constructed from poor approximations to the optimal value function, the information relaxation bounds are relatively good. Second, the bound improvement guarantee ranges from being quite slack in the $k = 0$ case to (relatively) tight in the $k = 3$ case. Third, even though the absorption time relaxation bounds are tighter than the perfect information bounds (as they must be) there is little difference in these bounds, particularly as k increases. In this example, it appears the benefit of additional information comes largely from knowing the absorption time: if the absorption time is relatively short in a given scenario and the penalty is somewhat weak, then



(a) k=0

(b) k=3

Figure B.2: Lower bounds for the machine repair example using an uncontrolled formulation and the perfect and absorption time relaxations.

regardless of the specific machine state transitions, it may never be optimal to repair in that scenario.

Information Relaxation Bounds Via Uncontrolled Formulations

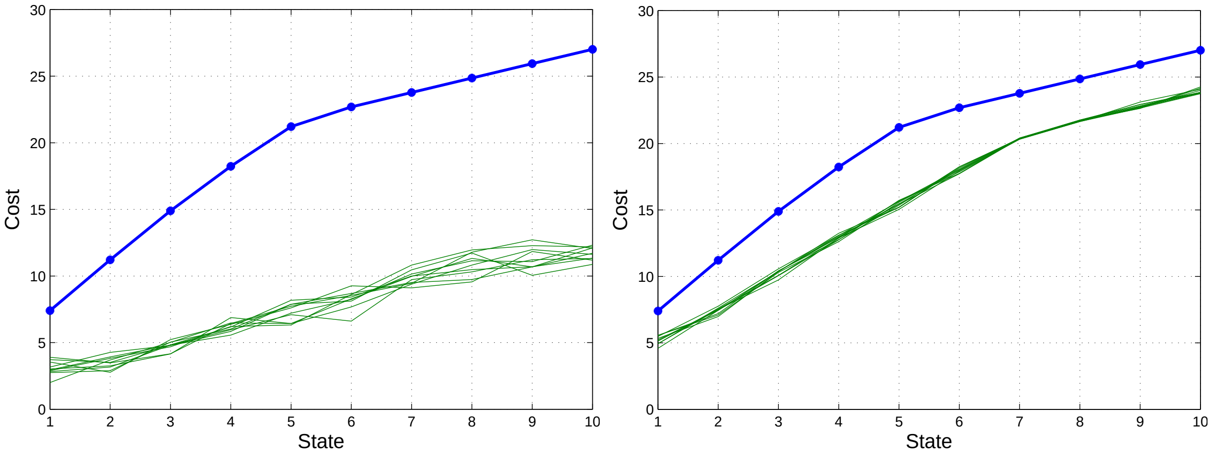
We now demonstrate uncontrolled formulation dual bounds and work directly with the state transition probability matrix (rather than the state transition function), which we denote by Q. Since this is an uncontrolled formulation, state transition probabilities do not depend on actions (i.e., whether a repair attempt is made or not). We assume that absorption occurs with probability $1 - \delta$ under Q from every non-absorbing state. With a perfect information relaxation the uncontrolled formulation inner problems (17) take the form

$$\begin{aligned} v_t^{\mathbb{G}}(x_t) \quad = \quad \min\Bigg\{ & c(x_t) + \delta P^{\bar{r}}_{x_t} v + \frac{P^{\bar{r}}(x_t, x_{t+1})}{Q(x_t, x_{t+1})}\left(v^{\mathbb{G}}_{t+1}(x_{t+1}) - v(x_{t+1})\right), \\ & R + c(x_t) + \delta P^{r}_{x_t} v + \frac{P^{r}(x_t, x_{t+1})}{Q(x_t, x_{t+1})}\left(v^{\mathbb{G}}_{t+1}(x_{t+1}) - v(x_{t+1})\right)\Bigg\} \end{aligned}$$

for $0 \le t \le \tau - 1$ with $v^{\mathbb{G}}_{\tau} = 0$. Under the absorption time relaxation, the lower bounds under the controlled formulation and these uncontrolled formulations coincide. This follows from the fact that only τ is known under the absorption time relaxation and we are using the same distribution of τ in these uncontrolled formulation examples.

Figure B.2 shows the lower bounds. The optimal value function is indicated in blue, and the black line indicates $v = v^{(k)}$ for some number, k, of value iterations beginning with the zero vector. Since $v^{(k)}$ is a subsolution, the results of Proposition 4.1 apply. In particular, the lower bounds from each information relaxation must be strictly better than $v^{(k)}$; the guaranteed improvement over $v^{(k)}$ is indicated with the green line. The red lines represent the lower bounds using the perfect and absorption time information relaxations. The plotted lower bounds are in fact an average of ten separate uncontrolled formulation examples, each resulting from a randomly generated state transition matrix Q. For each uncontrolled formulation example, we used stratified sampling on τ in generating scenarios to help reduce sample variance; for each example, Q is generated uniformly and normalized so that rows sum to one (since every element of each Q was strictly greater than zero, the absolute continuity condition in Theorem 4.1(i) holds relative to any feasible policy; thus even with penalties that were not constructed from subsolutions, we would be ensured lower bounds on the optimal value function here).

A notable feature of Figure B.2 is that the perfect information bound only provides a modest improvement over the aforementioned improvement guarantee, somewhat in contrast to the controlled formulation case in Figure B.1. This is perhaps not surprising given that the bound improvement guarantee (1) is generally weaker for uncontrolled formulations. It is interesting to observe that, unlike upper bounds from any primal feasible policy, the perfect information lower bounds depend on Q. In Figure B.3 we show the perfect information lower bounds for each of the uncontrolled formulation examples, each corresponding to a different Q. The blue line again indicates the optimal value function. More variation in the perfect information lower bounds are apparent in the $k = 0$ case (corresponding to zero penalties) than in the $k = 3$ case. We should expect an interaction between the approximate value function forming the penalties and the uncontrolled formulation's state transition probabilities. In particular, when we have relatively poor approximations to the optimal value function, differences in these state transition probabilities can have a more pronounced effect on the quality of the lower bounds than when the approximations are relatively good. The example in Section B.3 demonstrates explicitly how the state transition probabilities in uncontrolled formulations can affect the lower bounds.



(a) k=0

(b) k=3

Figure B.3: Lower bounds associated with different state transition matrices for the uncontrolled formulation.

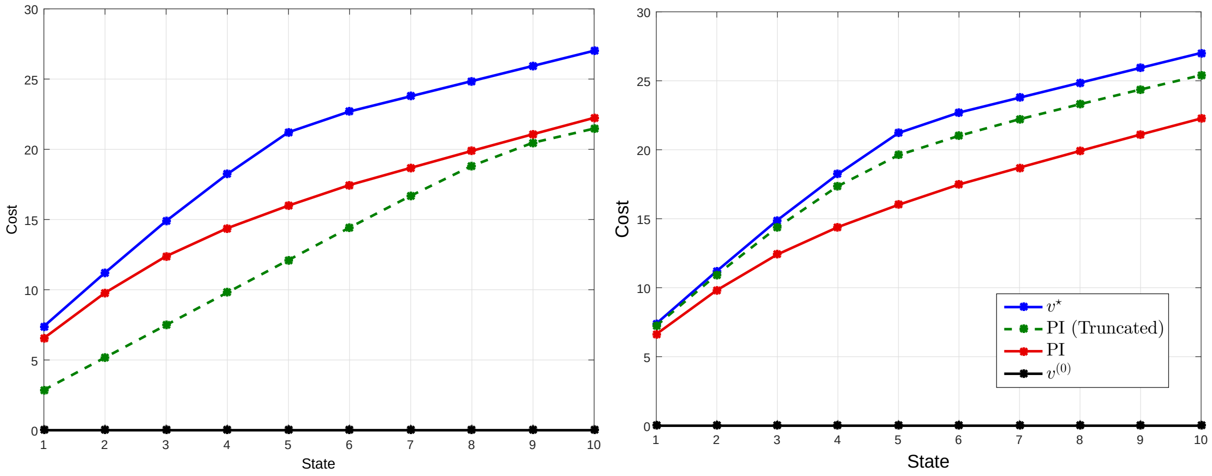
Information Relaxation Bounds Via Truncated Horizons

We can also illustrate the results of Section 4.4 using the machine repair example. In Figure B.4, we show perfect information lower bounds based on a truncated horizon model with $T = 3$ and $T = 25$ periods. Relative to the discount factor of $\delta = 0.75$, the horizon $T = 3$ is short and the horizon $T = 25$ is long. We take $v = 0$ in these examples. We see that that the lower bound from the $T = 3$ truncated model is not as good as the lower bound from the controlled formulation, which uses a geometric absorption time with parameter 0.25. When we increase the horizon to $T = 25$, however, the truncated model fares better than the controlled formulation bounds in all states.

As k increases, $v = v^{(k)}$ converges to the optimal value function, $v^{\star}$. If $v = v^{\star}$, Theorem 4.1(ii) implies that, no matter what distribution of absorption we use, we would obtain tight lower bounds (e.g., even truncated models with very short horizons). When v is a subsolution, the corresponding lower bound can never fall below v, and in general whether or not longer horizons will be worthwhile will depend on the quality of the approximation v.

B.2. Inner Problems May Require Larger State Spaces

Here we provide an illustrative example that shows that for some information relaxations, it may be necessary to expand the state space to solve the inner problems. This example introduces a new imperfect information relaxation, and the calculations with this relaxation can be instructive in terms of the mechanics of the inner problems with imperfect information. For simplicity, we present the example in a finite horizon setting with zero penalty. The example is the following: you can throw a fair n-sided die up to a maximum of T times.



(a) $T = 3$

(b) $T = 25$

Figure B.4: Perfect information bounds for machine repair: controlled formulation vs. truncated horizon formulation.

After each throw, you must choose to stop or continue. If you choose to stop, then the game terminates and you obtain a prize equal in dollars to the value you just threw. For example, if you throw a 4 and then choose to stop, then you obtain 4 dollars and the game is over; otherwise, you continue and can throw again (if you have not already thrown T times). We want to compute the optimal expected value of this game.

We denote by x_t the value of the t^{th} throw, which represents both the uncertainty in the problem as well as the state variable. This problem is a simple dynamic programming problem with an optimal policy that has a threshold structure. When $n = 6$ and $T = 3$, for example, it is easy to show the optimal $t = 0$ value (i.e., the value before the first throw) is $v_0^\star = 4.667$.

Suppose now that we wish to calculate an information relaxation (upper) bound for this problem using the information relaxation $\mathbb{G}$, where $\mathcal{G}_t := \sigma(\mathcal{F}_t, x_{t+1})$ for $t < T$ and $\mathcal{G}_T := \mathcal{F}_T$. That is, the relaxed filtration allows us to see the outcome of the die one period ahead at all times $t < T$. In solving the inner problem recursion under this $\mathbb{G}$, we need to account for both the value x_t (known under $\mathbb{F}$) as well as the value x_{t+1}. This can be handled by augmenting the state space to be all values of $y_t := (x_t, x_{t+1})$ for each period $t < T$. We then can solve a stochastic DP on this increased state space that takes an expectation over x_{t+2} at all periods $t \leq T - 2$. Note that because only x_1 is revealed under $\mathbb{G}$ at $t = 0$, we would average over x_1 in calculating the optimal $t = 0$ value under $\mathbb{G}$; the remaining uncertainty over $(x_2, \ldots, x_T)$ would be "averaged out" in the expectations conditional on $\mathcal{G}_0$ and there is no Monte Carlo simulation that is needed. (We omit the detailed calculations but leave it as an instructive exercise to verify that under this information relaxation with $n = 6$ and $T = 3$, the optimal value at $t = 0$ is 4.870. It is also instructive to confirm that strong duality holds – as it must– when we use a penalty constructed from the optimal value function.)

We emphasize that this example is illustrative: the primary motivation for information relaxations is to lead to problems that are much easier to solve than the primal DP. Thus although information relaxations can in theory lead to problems that are harder than the primal DP itself, it would be unusual to consider such relaxations in applications.

B.3. Bounds from Uncontrolled Formulations Can Outperform Bounds from Controlled Formulations

Given the results of the machine repair example, it is natural to conjecture that, for the same distribution of absorption time, information relaxation, and penalty, the lower bounds from controlled formulations are always at least as good as those from uncontrolled formulations. We now provide a simple counterexample to this conjecture. We consider a simple problem for which there are two periods, $t = 0$ and $t = 1$, two possible states, "good" (g) and "bad" (b), and two possible actions, 1 and 2. Costs are given by $c(g, 1) = 0$, $c(g, 2) = 0.2$, $c(b, 1) = 1$ and $c(b, 2) = 1$. The initial state is g and if the time $t = 0$ action is 1, then the time $t = 1$ state will be g with probability $p_1 = 0.6$ and b otherwise; if the time $t = 0$ action is 2, then the time $t = 1$ state will be g with probability $p_2 = 0.7$ and b otherwise. In time $t = 1$, we can select either action 1 or

2 one final time and the problem then ends. The objective is to minimize the expected costs. The optimal cost c^* is

$$\begin{aligned} c^* &= \min_{i=1,2} \left\{ c(g,i) + p_i \min_{j=1,2} c(g,j) + (1-p_i) \min_{j=1,2} c(b,j) \right\} \\ &= \min \{ 0 + .6 \cdot 0 + .4 \cdot 1, 0.2 + 0.7 \cdot 0 + 0.3 \cdot 1 \} \\ &= 0.4, \end{aligned}$$

an an optimal policy is to always select action 1. (It is irrelevant which action is taken at time $t = 1$ if the state is b). In each of the lower bound calculations below we use a perfect information relaxation and the zero penalty.

Bound from the Controlled Formulation

Using the controlled formulation with a perfect information relaxation, we learn the outcome of the state transition from $t = 0$ to $t = 1$ under each action. With probability $p_1 = 0.6$ we discover that both actions lead to g. With probability $p_2 - p_1 = 0.1$ we learn that action 1 leads to b but action 2 leads to g. With probability $1 - p_2 = 0.3$ we learn that both actions lead to b. In each of these three scenarios, we can choose the best action. The resulting lower bound is thus

$$\begin{aligned} \underline{c}_{\text{controlled}} &= p_1 \min_{i=1,2} \{ 2c(g,i) \} + (p_2 - p_1) \min \{ c(g,1) + c(b,1), c(g,2) + \min_{j=1,2} c(g,j) \} \\ &\quad + (1-p_2) \min \{ c(g,1) + \min_{j=1,2} c(b,j), c(g,2) + \min_{j=1,2} c(b,j) \} \\ &= 0.6(2 \cdot 0) + 0.1(0.2) + 0.3(1) \\ &= 0.32. \end{aligned}$$

Bound from an Uncontrolled Formulation

With an uncontrolled formulation, we draw from an action-independent distribution that transitions to the time $t = 1$ state g with probability q, and b with probability $1 - q$ where $0 < q < 1$. We can therefore express the lower bound in this case as

$$\begin{aligned} \underline{c}_{\text{uncontrolled}} &= q \min_{i=1,2} \left\{ c(g,i) + \frac{p_i}{q} \min_{j=1,2} c(g,j) \right\} + (1-q) \min_{i=1,2} \left\{ c(g,i) + \frac{(1-p_i)}{(1-q)} \min_{j=1,2} c(b,j) \right\} \\ &= 0 + \min \{ 0.2(1-q) + 0.3, 0.4 \} \\ &= \begin{cases} 0.4 & \text{if } q \le 0.5 \\ 0.2(1-q) + 0.3 & \text{otherwise.} \end{cases} \end{aligned}$$

We thus have $\underline{c}_{\text{uncontrolled}} = c^*$ for all $q \le 0.5$ and $\underline{c}_{\text{uncontrolled}} > \underline{c}_{\text{controlled}}$ for all $q < 0.9$. It is interesting to note that even with zero penalty for information, the uncontrolled formulation is capable (with a well-chosen q, i.e. $q \le 0.5$) of providing a tight lower bound. This observation helps highlight the fact that the performance of the lower bounds from uncontrolled (or more generally partially controlled) formulations may depend on both the penalty as well as the way the problem is formulated.