Appendices

A AUSO instance from Section 4.1

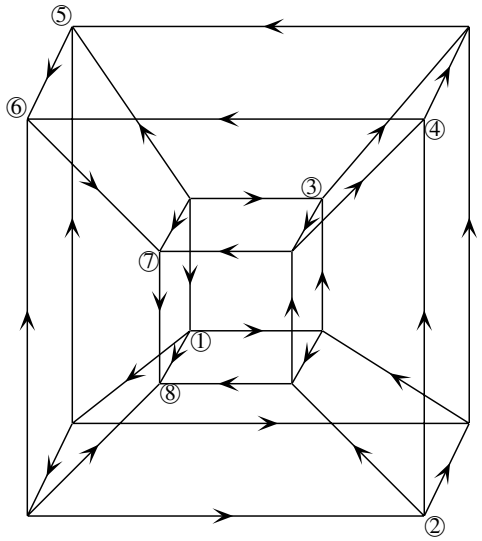

Figure 5: The only 4-AUSO (up to an isomorphism) on which HPI performs 8 vertex evaluations. The 8 vertices are numbered in sequence. This AUSO does not satisfy the Holt-Klee conditions. Notice, for example, that the inner 3-AUSO does not have 3 vertex-disjoint paths from source to sink.

B Proofs from Section 5

We provide a proof of Theorem 9, which uses the MDP designed by Melekopoglou and Condon [1994], shown in Figure 3. Recall from Section 5 that we only consider states $s \in \{1, 2, \ldots, n\}$ as a part of our analysis.

For this proof, we find it convenient to consider a slight modification to RPI. If a policy π has $m > 1$ improvable states, note that RPI obtains $\pi' \succ \pi$ by picking uniformly at random among the $2^m - 1$ improving policies in $I(\pi)$. We consider an algorithm RPI1 that instead picks π' uniformly at random from $I(\pi) \cup \{\pi\}$. The reason for so doing is that RPI1 can be implemented by independently switching each improvable state with probability $1/2$, which simplifies our analysis. The consequence, though, is that RPI1 is not strictly a PI algorithm, since with a finite probability, we can get $\pi' = \pi$. This probability is at most $1/2$, and therefore, the expected number of policies visited by RPI1 (which might contain repetitions) is at most twice the expected number of policies visited by RPI. To prove the theorem, we show below that the former quantity is at least $n + 1$.

Building on Melekopoglou and Condon [1994], first we obtain a simple rule to check if a state s is switchable.

Lemma 10. *For a policy π for M_n, a state s is switchable if and only if*

$$\sum_{s' \leq s} \pi(s) \equiv 0 \mod 2.$$

Proof. For states $s \in \{1, 2, \ldots, n\}$, Melekopoglou and Condon [1994] define

$$a(1) = -\frac{1}{2} \text{ and } a(s+1) = a(s)\left(\frac{1}{2} - \pi(s)\right).$$

It is easy to verify from the definition that $a(s+1)$ is negative if and only if $\sum_{s' \leq s} \pi(s) \equiv 0 \mod 2$ [Melekopoglou and Condon, 1994, see Corollary 2.3]. Since $a(s+1) = a(s)(\frac{1}{2} - \pi(s))$, $a(s+1)$ is negative if and only if $\pi(s) = 0$ and $a(s) < 0$, or $\pi(s) = 1$ and $a(s) > 0$. Based on the structure of M_n, Melekopoglou and Condon[1994, see Corollary 2.4] show that the latter condition is equivalent to s being switchable. □

The crucial step in our proof is to define a *progress* function f on the policy space, which is then shown to be non-increasing with respect to PI updates.

Definition 11. *For a policy π for M_n,*

$$f(\pi) \stackrel{\text{def}}{=} \min\left(\textit{states}(T^\pi) \cup \{n+1\}\right).$$

In other words, $f(\pi)$ is defined to be the smallest switchable state if π is not optimal, and $n + 1$ if it is $\pi^\star$. The lemma below establishes the monotonicity of f.

Lemma 12. *If RPI1 visits the policies $\pi^0, \pi^1, \ldots, \pi^m$ in sequence, then for $1 \leq i \leq m$, $f(\pi^{i-1}) \leq f(\pi^i)$.*

Proof. Since we stop when there are no improvable states, $f(\pi^{m-1}) \leq f(\pi^m) = n + 1$. Otherwise assume that $i < m$. Let $f(\pi^{i-1}) = s$. Since vertex s is the smallest switchable state in π^{i-1}, any state s' will not be switched in π^{i-1} for $1 \leq s' < s$, and hence $\pi^i(s') = \pi^{i-1}(s')$. It follows from Lemma 10 that states $1, 2, \ldots, s-1$ are not switchable in π^i. Thus, $f(\pi^i) \geq s = f(\pi^{i-1})$. □

Next we show that as RPI1 proceeds, with sufficiently high probability f increases quite slowly. It follows thereafter that at least $n + 1$ policy evaluations must be made in expectation if $\pi^0 = 0^n$ is the initial policy ($f(\pi^0)$ and $f(\pi^\star)$ differ by n).

Lemma 13. *If RPI1 visits the policies $\pi^0, \pi^1, \ldots, \pi^m$ in sequence, then for $1 \leq i \leq m$, $t \geq 0$,*

$$\mathbb{P}\{f(\pi^i) - f(\pi^{i-1}) \geq t\} \leq \frac{1}{2^t}.$$