Highlights

A Note on the Bias and Kemeny's Constant in Markov Reward Processes with an Application to Markov Chain Perturbation

Ronald Ortner

- presents a new identity for the bias in Markov reward processes
- generalizes results for Markov chain perturbation from irreducible to arbitrary chains
- gives a new improved perturbation bound in 1-norm for the stationary distribution
- provides a new intuitive explanation why Kemeny's constant is a constant

A Note on the Bias and Kemeny's Constant in Markov Reward Processes with an Application to Markov Chain Perturbation

Ronald Ortner

[a]*Montanuniversitaet Leoben, Franz-Joseph-Strasse 18, Leoben, 8700, Austria*

Abstract

Given a unichain Markov reward process (MRP), we provide an explicit expression for the bias values in terms of mean first passage times. This result implies a generalization of known Markov chain perturbation bounds for the stationary distribution to the case where the perturbed chain is not irreducible. It further yields an improved perturbation bound in 1-norm. As a special case, Kemeny's constant can be interpreted as the translated bias in an MRP with constant reward -1, which offers an intuitive explanation why it is a constant.

Keywords: Markov reward process, Markov chain, bias, perturbation theory, stationary distribution, mean first passage times, Kemeny's constant
2020 MSC: 60J10, 90C40

1. Preliminaries

1.1. Markov reward processes

Consider a Markov chain over a finite state space S with states $1, 2, \ldots, N$ and transition probabilities p_{ij} $(1 \leq i, j \leq N)$. We assume in what follows that the Markov chain is unichain, that is, it consists of a single recurrent class and a possibly empty set of transient states. Equipping the Markov chain with a reward function $r : S \to \mathbb{R}$ yields a Markov reward process (MRP), cf. Section 8.2 of [1] for the following facts. Usually, it is assumed that the reward[1] r_i in each state i is the mean of some fixed reward distribution. The

[1]For functions $f : S \to \mathbb{R}$ we write in the following short f_i instead of $f(i)$.

average reward ρ in the MRP is then defined as

$$\rho = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[r_{S_t} | S_1 = i],$$

where S_t is a random variable for the state at step t. In the assumed case of a unichain MRP the value ρ is independent of the initial state i. In fact, ρ can be written in terms of the stationary distribution μ as

$$\rho = \sum_{i=1}^{N} \mu_i \, r_i,$$

noting that a unchain Markov chain has a unqiue stationary distribution μ with $\mu_i = 0$ for transient states i.

1.2. The bias

While ρ is the average reward in the limit, the actual collected rewards will differ depending on the initial state. This is made precise by the notion of *bias*, which for each state i is defined as

$$\lambda_i = \mathbb{E}\left[\sum_{t=1}^{\infty} \left(r_{S_t} - \rho \right) \,\middle|\, S_1 = i \right] \tag{1}$$

in MRPs with underlying aperiodic chain, while in general one sets

$$\lambda_i = \lim_{T\to\infty} \frac{1}{T} \sum_{\tau=1}^{T} \mathbb{E}\left[\sum_{t=1}^{\tau} \left(r_{S_t} - \rho \right) \,\middle|\, S_1 = i \right]. \tag{2}$$

By definition, the difference of two bias values $\lambda_i - \lambda_j$ quantifies the advantage in accumulated reward when starting in state i over starting in state j.

Example 1. *Assume that all states i have the same mean reward $r_i = r$. Then the average reward $\rho = r$ and ρ is attained from the first step, independent of the initial state. Accordingly, the difference of any two bias values $\lambda_i - \lambda_j$ has to be 0. Indeed, the bias is 0 for all states.*

The bias values of an MRP are a solution of the *Poisson equation*, that is, for all i,

$$\rho + \lambda_i = r_i + \sum_{j=1}^{N} p_{ij} \, \lambda_j. \tag{3}$$

The bias values λ_i in addition satisfy $\sum_i \mu_i \lambda_i = 0$, which can be achieved for any solution of (3) by adding a suitable vector with identical entries. On the other hand, any respective translation $\lambda'_i := \lambda_i + c$ of the bias values λ_i still fulfills the Poisson equation.

1.3. Mean first passage times

The *mean passage time* τ_{ij} $(i \neq j)$ is defined as the expected time it takes to first visit j when starting in i. Further, the *mean return time* τ_{ii} is the first time i is revisited again when starting in i. It is well-known [2] that in irreducible Markov chains, for $1 \leq i, j \leq N$

$$\tau_{ij} = 1 + \sum_{k \neq j} p_{ik}\, \tau_{kj}, \text{ and} \tag{4}$$

$$\tau_{ii} = \frac{1}{\mu_i}. \tag{5}$$

In unichain Markov chains τ_{ij} can be infinite for transient states j. However, all τ_{ij} are finite for recurrent states j and for these (4) and (5) still hold.

2. Main result

Our main result gives an explicit expression for the bias values in terms of the mean first passage times of an MRP.

Theorem 2. *The values*

$$\lambda'_i := -\sum_{j \neq i} \mu_j\, r_j\, \tau_{ij}$$

satisfy the Poisson equation (3)*.*

Proof. Inserting the defined values λ'_i in the right hand side of the the Poisson equation (3), we obtain, using (4) and (5),

$$\begin{aligned}
r_i + \sum_j p_{ij}\, \lambda'_j &= r_i - \sum_j p_{ij} \sum_{k \neq j} \mu_k\, r_k\, \tau_{jk} \\
&= r_i - \sum_j p_{ij} \sum_k \mu_k\, r_k\, \tau_{jk} + \sum_j p_{ij}\, \mu_j\, r_j\, \tau_{jj} \\
&= r_i - \sum_k \mu_k\, r_k \sum_j p_{ij}\, \tau_{jk} + \sum_j p_{ij}\, r_j \\
&= r_i - \sum_k \mu_k\, r_k \big(\tau_{ik} - 1 + p_{ik}\, \tau_{kk}\big) + \sum_j p_{ij}\, r_j
\end{aligned}$$

$$
\begin{aligned}
&= r_i - \sum_k \mu_k\, r_k \tau_{ik} + \sum_k \mu_k\, r_k - \sum_k \mu_k\, r_k\, p_{ik}\, \tau_{kk} + \sum_j p_{ij}\, r_j \\
&= r_i + \lambda'_i - \mu_i\, r_i\, \tau_{ii} + \sum_k \mu_k\, r_k - \sum_k r_k\, p_{ik} + \sum_j p_{ij}\, r_j \\
&= \lambda'_i + \rho,
\end{aligned}
$$

which concludes the proof. □

In order to obtain the actual bias values λ_i from the λ'_i defined in Theorem 2, these have to be translated, cf. the remark after (3).

3. Implications

While Theorem 2 is quite simple, it has some interesting implications discussed in the following.

3.1. Bias span

A known connection between the bias and transition times is the following. We define the *diameter* $D := \max_{i \neq j} \tau_{ij}$ to be the maximal mean first passage time between two states. Then for rewards bounded in $[0, 1]$ the *bias span* $\mathrm{span}(\lambda)$ is upper bounded as

$$
\mathrm{span}(\lambda) \;:= \max_i \lambda_i - \min_i \lambda_i \;\leq\; D. \tag{6}
$$

This observation has been made in the more general context of Markov decision processes (MDPs), see [3]. Theorem 2 makes the connection between bias and transition times precise. Note that (6) is a straightforward consequence of Theorem 2.

3.2. Markov chain perturbation

Let us consider a Markov chain with transition matrix $P = (p_{ij})_{i,j=1}^N$ and a perturbed chain with transition matrix $\tilde{P} = (\tilde{p}_{ij})_{i,j=1}^N$. Perturbation bounds for the stationary distribution provide inequalities of the form

$$
\|\mu - \tilde{\mu}\|_p \;\leq\; \kappa \cdot \|P - \tilde{P}\|_q
$$

for so-called *condition numbers* κ (i.e., parameters of the unperturbed chain), most commonly for $p = 1, \infty$ and $q = \infty$, cf. [4] for an overview. The condition numbers of the following two bounds involve mean first passage times and are closely related to the bias values of Theorem 2.

Theorem 3 (Cho & Meyer [5]). *Let P, $\tilde{P}$ be the transition matrices of two irreducible Markov chains with stationary distributions μ and $\tilde{\mu}$. Then*

$$|\mu_i - \tilde{\mu}_i| \;\leq\; \frac{\mu_i}{2} \cdot \max_{j \neq i} \tau_{ji} \cdot \|P - \tilde{P}\|_\infty .$$

The condition number of the following bound uses Kemeny's constant η, defined as

$$\eta \;:=\; \eta_i \;:=\; \sum_{j \neq i} \mu_j \, \tau_{ij} .$$

It can be shown that η_i is indeed independent of i (cf. also next section below). Note that η_i coincides with λ'_i when all rewards are -1.

Theorem 4 (Hunter [6]). *Let P, $\tilde{P}$ be the transition matrices of two irreducible Markov chains with stationary distributions μ and $\tilde{\mu}$. Then*

$$\|\mu - \tilde{\mu}\|_1 \;\leq\; \frac{\eta}{2} \cdot \|P - \tilde{P}\|_\infty .$$

The bounds of Theorems 3 and 4 have been shown for irreducible Markov chains. In the more general setting of MDPs, perturbation bounds are known that hold more generally in structures that need not be irreducible [7]. The respective condition number is the diameter, which is larger than the condition numbers used in Theorems 3 and 4. However, the diameter only serves as an upper bound on the bias span as in (6). Accordingly, with the result of Theorem 2, we can obtain perturbation bounds which are not only more general but also sharper.

Let us first restate the perturbation bound of [7] for the case of MRPs, a proof is given in the appendix.[2]

Theorem 5 (Ortner et al. [7]). *Consider a unichain MRP with transition matrix P and another MRP with the same reward function r but a (possibly not irreducible) perturbed matrix $\tilde{P}$. Then, independent of the initial state, the difference of the average rewards ρ, $\tilde{\rho}$ of the two MRPs is upper bounded as*

$$|\rho - \tilde{\rho}| \;\leq\; \tfrac{1}{2} \cdot \operatorname{span}(\lambda) \cdot \|P - \tilde{P}\|_\infty .$$

[2]The proof of the original bound is contained in an unpublished appendix of [8]. This bound is stated in a very general context when the state spaces of the original and the perturbed MDP need not coincide and also the reward function may be perturbed. For the case of two MDPs with the same state space, the proof has been restated in [8].

Theorem 5 easily implies Theorems 3 and 4, but now these hold more generally for the case when the original Markov chain is unichain, and there are no conditions on the perturbed chain.

Proof of Theorem 3 from Theorem 5. We fix an initial state and note that $\tilde{\mu}$ and $\tilde{\rho}$ depend on this initial state in the following. Set the reward function in Theorem 5 to be $r_i = 1$ and $r_j = 0$ for all $j \neq i$. Then by definition of λ'_i,

$$\lambda'_j = \begin{cases} 0 & \text{for } j = i, \\ -\mu_i \tau_{ji} & \text{for } j \neq i, \end{cases} \tag{7}$$

so that

$$\operatorname{span}(\lambda') = \mu_i \max_{j \neq i} \tau_{ji}.$$

By Theorems 5 and 2,

$$\begin{aligned} |\mu_i - \tilde{\mu}_i| \;=\; |\rho - \tilde{\rho}\,| \;&\leq\; \tfrac{1}{2} \operatorname{span}(\lambda) \cdot \|P - \tilde{P}\|_\infty, \\ &=\; \tfrac{1}{2} \operatorname{span}(\lambda') \cdot \|P - \tilde{P}\|_\infty, \\ &=\; \tfrac{1}{2}\, \mu_i \max_{j \neq i} \tau_{ji} \cdot \|P - \tilde{P}\|_\infty, \end{aligned}$$

which is precisely the bound of Theorem 3 and holds independent of the chosen initial state. □

Proof of Theorem 4 from Theorem 5. Again we fix an initial state on which $\tilde{\mu}$ and $\tilde{\rho}$ depend in the following. We define a reward function

$$r_i = \begin{cases} 1 & \text{if } \mu_i \geq \tilde{\mu}_i, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Then the difference of the average rewards ρ, $\tilde{\rho}$ of the original and the perturbed MRP is the total variation distance between μ and $\tilde{\mu}$, which is known to be $\frac{1}{2}\|\mu - \tilde{\mu}\|_1$. Therefore, we get by Theorems 5, independent of the initial state,

$$\tfrac{1}{2}\, \|\mu - \tilde{\mu}\|_1 \;=\; |\rho - \tilde{\rho}\,| \;\leq\; \tfrac{1}{2} \operatorname{span}(\lambda) \cdot \|P - \tilde{P}\|_\infty. \tag{9}$$

For $\operatorname{span}(\lambda)$ we have by Theorem 2,

$$\begin{aligned} \operatorname{span}(\lambda) = \operatorname{span}(\lambda') \;&=\; \max_i \sum_{\substack{j \neq i \\ \mu_j \geq \tilde{\mu}_j}} \mu_j \tau_{ij} - \min_i \sum_{\substack{j \neq i \\ \mu_j \geq \tilde{\mu}_j}} \mu_j \tau_{ij} \\ &\leq\; \max_i \sum_{\substack{j \neq i \\ \mu_j \geq \tilde{\mu}_j}} \mu_j \tau_{ij} \;\leq\; \max_i \sum_{j \neq i} \mu_j \tau_{ij} \;=\; \eta, \end{aligned} \tag{10}$$

which finishes the proof. □

Looking at the proofs, we see that while we precisely obtain the bound of Theorem 3, the bound of Theorem 4 is a bit loose when compared to the bound implied by Theorems 5 and 2. The following corollary to Theorem 5 summarizes our findings and presents a respective improved bound on $\|\mu - \tilde{\mu}\|_1$.

Corollary 6. *Consider a unichain Markov chain with transition matrix P and stationary distribution μ, and a perturbed Markov chain with transition matrix $\tilde{P}$, which may be not irreducible. Then independent of the initial state, the stationary distribution μ' of the perturbed chain satisfies*

$$\begin{aligned} |\mu_i - \tilde{\mu}_i| &\leq \frac{\mu_i}{2} \cdot \max_{j \neq i} \tau_{ji} \cdot \|P - \tilde{P}\|_\infty, \text{ and} \\ \|\mu - \tilde{\mu}\|_1 &\leq \tfrac{1}{2} \cdot \max_{A \subseteq \{1,2,\ldots,N\}} \left\{ \max_i \sum_{j \in A \setminus \{i\}} \mu_j \tau_{ij} - \min_i \sum_{j \in A \setminus \{i\}} \mu_j \tau_{ij} \right\} \cdot \|P - \tilde{P}\|_\infty. \end{aligned}$$

Proof. The first statement is just Theorem 3 generalized, which we have shown before. The second statement follows from the proof of Theorem 4 above, considering the maximal possible expression on the right hand side of (10). □

3.3. Kemeny's constant

Our results also give another answer to the question why Kemeny's constant is a constant [2, 9, 10]. When all rewards r_i are -1, then $\lambda'_i = \eta_i$ for all i. As discussed in Example 1, identical rewards imply identical bias values so that it follows that all the η_i have to be identical. This not only provides a short proof that $\eta_i = \eta$ for all i, it also gives a simple explanation why Kemeny's constant is a constant: The η_i are the translated bias values in an MRP with identical rewards and hence have to be identical, too.

Appendix A. Proof of Theorem 5

We start with a result that after taking ℓ steps in the perturbed MRP compares the accumulated rewards to the quantity $\ell\rho$.

Lemma 7. *Consider a unichain MRP with transition matrix P, stationary distribution μ, and bias λ, and let another MRP have the same reward function r but a perturbed transition matrix $\tilde{P}$. We take ℓ steps in the perturbed*

MRP and write v_i for the number of visits in state i. Then it holds with probability at least $1-\delta$ that

$$\ell\rho - \sum_i v_i \cdot r_i \;\le\; \tfrac{\ell}{2}\,\mathrm{span}(\lambda)\cdot\|P-\tilde{P}\|_\infty + \mathrm{span}(\lambda)\left(1+\sqrt{2\ell\log(1/\delta)}\right), \quad \text{(A.1)}$$

independent of the initial state.

Proof. We first apply a translation $\bar{\lambda}_i := \lambda_i - \frac{1}{2}(\max_j \lambda_j + \min_j \lambda_j)$ to the bias values λ_i. Then

$$\|\bar{\lambda}\|_\infty \;= \max_j \bar{\lambda}_j \;=\; \tfrac{1}{2}\,\mathrm{span}(\lambda) \;=\; \tfrac{1}{2}\,\mathrm{span}(\bar{\lambda}). \quad \text{(A.2)}$$

Further, the $\bar{\lambda}_i$ still satisfy the Poisson equation (3), so that

$$\begin{aligned}\ell\rho - \sum_i v_i \cdot r_i &= \sum_i v_i\,(\rho - r_i) = \sum_i v_i \Big(\sum_j p_{ij}\,\bar{\lambda}_j - \bar{\lambda}_i\Big)\\ &= \sum_i v_i \Big(\sum_j \tilde{p}_{ij}\,\bar{\lambda}_j - \bar{\lambda}_i\Big) + \sum_i v_i \cdot \sum_j \big(p_{ij}-\tilde{p}_{ij}\big)\,\bar{\lambda}_j. \quad \text{(A.3)}\end{aligned}$$

Writing S_t for the state at step t we obtain for the first term in (A.3)

$$\begin{aligned}\sum_i v_i \Big(\sum_j \tilde{p}_{ij}\,\bar{\lambda}_j - \bar{\lambda}_i\Big) &= \sum_{t=1}^{\ell} \Big(\sum_j \tilde{p}_{S_t,j}\,\bar{\lambda}_j - \bar{\lambda}_{S_t}\Big)\\ &= \sum_{t=1}^{\ell} \Big(\sum_j \tilde{p}_{S_t,j}\,\bar{\lambda}_j - \bar{\lambda}_{S_{t+1}}\Big) + \bar{\lambda}_{S_{\ell+1}} - \bar{\lambda}_{S_1}. \quad \text{(A.4)}\end{aligned}$$

The sequence

$$X_t := \sum_j \tilde{p}_{S_t,j}\,\bar{\lambda}_j - \bar{\lambda}_{S_t+1}$$

is a martingale difference sequence with $|X_t| \le \mathrm{span}(\lambda)$, so that by Azuma-Hoeffding's inequality (e.g., Lemma 10 of [3]) with probability $1-\delta$,

$$\sum_{t=1}^{\ell} \Big(\sum_j \tilde{p}_{S_t,j}\,\bar{\lambda}_j - \bar{\lambda}_{S_{t+1}}\Big) \;\le\; \mathrm{span}(\lambda)\sqrt{2\ell\log(1/\delta)}. \quad \text{(A.5)}$$

Hence we obtain from (A.4)

$$\sum_i v_i \Big(\sum_j \tilde{p}_{ij}\,\bar{\lambda}_j - \bar{\lambda}_i\Big) \;\le\; \mathrm{span}(\lambda)\sqrt{2\ell\log(1/\delta)} + \mathrm{span}(\lambda). \quad \text{(A.6)}$$

The second term in (A.3) can be bounded by (A.2) as

$$\begin{aligned}\sum_i v_i \cdot \sum_j \left(p_{ij} - \tilde{p}_{ij}\right) \bar{\lambda}_j &\leq \sum_i v_i \cdot \sum_j |p_{ij} - \tilde{p}_{ij}| \cdot \left\|\bar{\lambda}\right\|_\infty \\ &\leq \ell \cdot \|P - \tilde{P}\|_\infty \cdot \tfrac{1}{2}\operatorname{span}(\lambda). \qquad \text{(A.7)}\end{aligned}$$

Combining (A.3), (A.6), and (A.7) gives the claimed

$$\ell\rho - \sum_i v_i \cdot r_i \;\leq\; \tfrac{\ell}{2}\operatorname{span}(\lambda) \cdot \|P - \tilde{P}\|_\infty + \operatorname{span}(\lambda)\left(1 + \sqrt{2\ell\log(1/\delta)}\right). \quad \square$$

Now Theorem 5 follows from Lemma 7 by dividing (A.1) by ℓ, choosing $\delta = 1/\ell$, and letting $\ell \to \infty$. $\square$

References

[1] M. L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley & Sons, Inc., New York, NY, USA, 1994.

[2] J. Kemeny, J. Snell, Finite Markov Chains, Van Nostrand, 1960.

[3] T. Jaksch, R. Ortner, P. Auer, Near-optimal regret bounds for reinforcement learning, J. Mach. Learn. Res. 11 (2010) 1563–1600.

[4] G. E. Cho, C. D. Meyer, Comparison of perturbation bounds for the stationary distribution of a Markov chain, Linear Algebra Appl. 335 (2001) 137–150.

[5] G. E. Cho, C. D. Meyer, Markov chain sensitivity measured by mean first passage times, Linear Algebra Appl. 316 (2000) 21–28.

[6] J. J. Hunter, Mixing times with applications to perturbed Markov chains, Linear Algebra Appl. 417 (2006) 108–123.

[7] R. Ortner, O. Maillard, D. Ryabko, Selecting near-optimal approximate state representations in reinforcement learning, in: Algorithmic Learning Theory – 25th International Conference, ALT 2014, 2014, pp. 140–154.

[8] R. Ortner, Markov chain estimation, approximation, and aggregation for average reward Markov decision processes and reinforcement learning, submitted to Handbook of Statistics (2024).

[9] K. Gustafson, J. J. Hunter, Why the Kemeny time is a constant, Special Matrices 4 (1) (2016) 176–180.

[10] D. Bini, J. J. Hunter, G. Latouche, B. Meini, P. G. Taylor, Why is Kemeny's constant a constant?, J. Appl. Probab. 55 (4) (2018) 1025–1036.