Geometry of the Space of Value Functions in POMDPs

Ryan A. Anderson

Department of Statistics and Data Science
University of California, Los Angeles

July 9, 2025

Table of Contents

Table of Contents

Intro to Reinforcement Learning

- Sutton and Barto [11] distinguish *reinforcement learning* from both supervised and unsupervised learning — agents in RL must balance *exploitation* with *exploration*
- RL approaches have led to state of the art ML models such as DeepSeek-R1 [2]; finding approaches to tractably solving POMDPs is a hot problem at the moment [3]
- Today: introduce interval arithmetic and algebraic methods to understand geometry of optimization in partially observable RL problems

Parameterizing the RL Problem

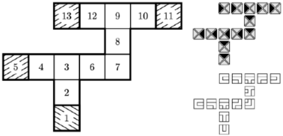


Figure: Taken from [6]

Definition (Markov Decision Process)

An MDP is a tuple $(S, A, O, \alpha, \beta, r, \gamma)$ where:

- S, A, O: finite set of states, actions, observations
- $\beta(s, o), \alpha(s, a, s')$: probability of observing o given state s, probability of moving to s' given (s, a)
- $r(s, a)$: reward of action a in state s with discount factor $\gamma \in [0, 1)$

Agents choose policies $\pi : S \to \Delta_A$ to navigate state space. Optimal policy π^* maximizes total rewards

Differences Between MDPs

- Researchers distinguish between *fully observable* and *partially observable* MDPs
  - For fully observable MDPs, finding the optimal policy can be done in polynomial time — this is impossible in partially observable setting [8]
- Also distinguish between *finite horizon* and *infinite horizon, discounted rewards* settings as well as *memoryless policies* vs *policies with memory*
  - Policies with memory allow one to model a POMDP as an MDP, which allows for finding an optimal policy via methods like policy iteration [12]
- For fully observable MDPs, the optimal memoryless policy is deterministic, but for POMDPs the generic optimal memoryless policy is stochastic
  - With conditions on the entries of the observation kernel, then there may exist an optimal memoryless deterministic policy for a POMDP [6]

Role of the Value Function in Solving MDPs

- Define the *value function* $V^\pi(s)$ as the expected discounted sum of future rewards from starting in state s:

$$V^\pi(s) = \mathbb{E}_{P^\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \;\middle|\; s_0 = s\right].$$

- By conditioning on the first observation-action step we get $V^\pi = (I - \gamma P^\pi)^{-1} r^\pi$ – this is the *Bellman equation*
- Given an initial state distribution ρ, we can find the optimal policy by solving a linear program in terms of the value function V^π [9]

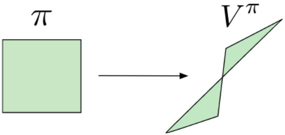
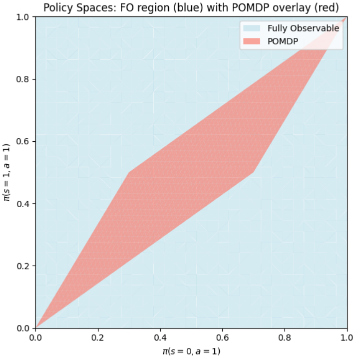
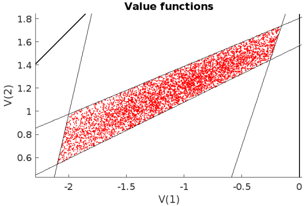


Figure: Taken from [1]

Geometry of the Value Function in MDPs

For FOMDPs, the space of value functions is a union of polytopes [1].
For POMDPs, the space of value functions is **not** a union of polytopes — instead rational functions in policy entries [7]





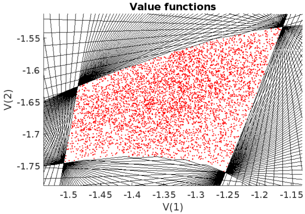


Table of Contents

Interval Matrix Systems

- Let $[A]$ be a *matrix of intervals*, i.e., entries satisfy $A_{ij} \in [\underline{a_{ij}}, \overline{a_{ij}}]$). Define A^c as the matrix of interval centers (i.e., entries $A^c_{ij} = \frac{1}{2}(\underline{a_{ij}} + \overline{a_{ij}})$), and A^Δ as the matrix of interval lengths.
- **Theorem** (Oettli-Prager 1964) [10]: For an interval matrix $[A]$ and vector $[b]$ with centers A^c, b^c and lengths A^Δ, b^Δ, x satisfies $Ax = b$ for some $A \in [A], b \in [b]$ iff

$$|A^c x - b^c| \leq |A^\Delta x| + |b^\Delta|.$$

Proof sketch of $\Rightarrow$: Let there exist $A \in [A], b \in [b]$ such that $Ax - b = 0$. Then by triangle inequality we have

$$|A^c x - b^c| = |A^c x - b^c - (Ax - b)|$$
$$= |(A^c - A)x + (b - b^c)| \leq |A^\Delta x| + |b^\Delta|$$

Parametric Matrix Systems

- More useful class of matrix systems to consider are *parametric*: Consider a parameter space formed by a product of intervals, $[p] = \times_{k=1}^{K} [\underline{p_k}, \overline{p_k}]$, with $p_k^c = \frac{1}{2}(\overline{p_k} + \underline{p_k})$ and $p_k^\Delta = \frac{1}{2}(\overline{p_k} - \underline{p_k})$
- Consider a parametrization of a family of matrices via $A(p) = \sum_{k=1}^{K} p_k A^k$, $b(p) = \sum_{k=1}^{K} p_k b^k$ for $p \in [p]$.
- Analog of the Oettli-Prager theorem gives only necessary condition, which constructs a loose enclosure of the solution set: if $x \in \Sigma$, then

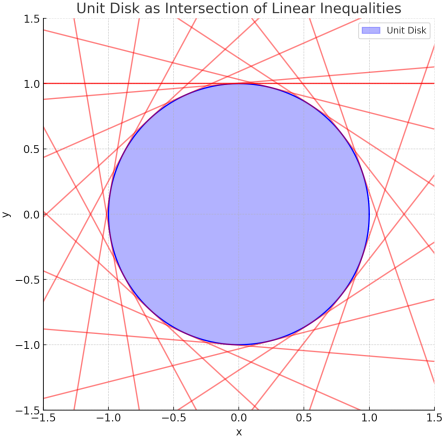
$$|A(p^c)x - b(p^c)| \leq \sum_{k=1}^{K} p_k^\Delta |A^k x - b^k|.$$

- **Theorem** [5]: $x \in \Sigma = \{x \in \mathbb{R}^n : A(p)x = b(p), p \in [p]\}$ if and only if for every $y \in \mathbb{R}^n$ we have

$$y^\top (A(p^c)x - b(p^c)) \leq \sum_{k=1}^{K} p_k^\Delta |y^\top (A^k x - b^k)|.$$

Infinitely Many Linear Ineqs vs Finitely Many Poly. Ineqs

$$\Sigma = \{(x, y) : x^2 + y^2 \leq 1\} = \cap_{\theta \in [0, 2\pi)} \{(x, y) : x \cos\theta + y \sin\theta \leq 1\}$$



Finite Characterization of Solutions to Parametric Systems

- From above: if x solves $A(p)x - b(p) = 0$, then $|A(p^c)x - b(p^c)| \leq \sum_{k=1}^{K} p_k^{\Delta}|A^k x - b^k|$
- This implies x solves $A(p)x - b(p)$ if and only if there exists $q = (q_1, \ldots, q_K), q_i \in [-1, 1]$ such that

$$A(p^c)x - b(p^c) = \sum_{k=1}^{K} q_k p_k^{\Delta}(A^k x - b^k). \tag{1}$$

- Let D be the matrix of *deviations* that appear on the RHS of above equation, with the kth column of D given as

$$D_k(x) = p_k^{\Delta}(A^k x - b^k).$$

- Let R_c be the vector of *midpoint residuals* that appear in the LHS:

$$R_c(x) = A(p^c)x - b(p^c).$$

Finite Characterization of Solutions to Parametric Systems

Theorem

A vector x solves a parametric matrix system $A(p)x - b(p) = 0$ if and only if there exist $q_k \in [-1, 1]$ such that

$$A(p^c)x - b(p^c) = \sum_k q_k p_k^\Delta (A^k x - b^k)$$

and

$$R_c^\perp = (I - DD^\dagger)R_c = 0,$$

where $D^\dagger$ is the pseudo-inverse of D.

In particular, we have two conditions involving R_c, D:

1. Zonotope condition, which is a finite set of inequalities
2. Non-orthogonality condition, which is a finite set of polynomial equations

The Zonotope Condition

- Recall that a zonotope $Z(g_1, \ldots, g_k)$ generated by vectors $(g_1, \ldots, g_k)$ is defined as $Z = \{\sum_{i=1} \alpha_i g_i : \alpha_i \in [-1, 1]\}$. They arise from the Minkowski sum of line segments: $Z = [-1, 1]g_1 + [-1, 1]g_2 + \cdots + [-1, 1]g_k$ [4].
- Here we need R_c to remain in the zonotope generated by the columns of D, which correspond to possible directions of deviation away from the solution set of the parametric system

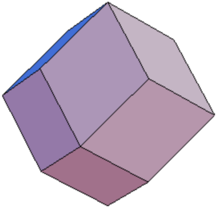


Figure: From the Geometry Junkyard

Polynomials from Non-Orthogonality Condition

- Condition $R_c^\perp = 0$ comes from the fact that if there exists $q \in [-1, 1]^K$ with

$$A(p^c)x - b(p^c) = \sum_{k=1}^{K} q_k p_k^\Delta (A^k x - b^k).$$

then $R_c = A(p^c)x - b(p^c)$ must be in column space of $D = [D_1 | \cdots | D_K], D_k = p_k^\Delta (A^k x - b^k)$.
- Component of R_c orthogonal to colspace(D) must vanish.
- Each element of $R_c^\perp$ can be written as polynomials in x

$$(R_c^\perp)_i = ((I - DD^T)R_c)_i = (R_c)_i - \sum_k^m D_{k,i} \langle D_k, R_c \rangle,$$

$$D_{k,i} = (A(p^c)x - b(p^c))_i,$$

$$\langle D_k, R_c \rangle = p_k^\Delta \sum_{j=1}^{m} (A^k x - b^k)_j (A(p^c)x - b(p^c))_j.$$

Conditions on 2x2 Real Parametric Systems

Consider a 2×2 real parametric system in 3 parameters, so that we have

$$A(p) = A^0 + p_1 A^1 + p_2 A^2, b(p) = b^0 + p_1 b^1 + p_2 b^2$$

For $k = 0, 1, 2$, let

$$A^k = \begin{pmatrix} a_{11}^k & a_{12}^k \\ a_{21}^k & a_{22}^k \end{pmatrix}, \quad b^k = \begin{pmatrix} b_1^k \\ b_2^k \end{pmatrix}, \quad \Delta_k = p_k^\Delta.$$

For $x = (x_1, x_2)$, define

$$R_c(x) = A^0 x - b^0 = \begin{pmatrix} r_1(x) \\ r_2(x) \end{pmatrix} = \begin{pmatrix} a_{11}^0 x_1 + a_{12}^0 x_2 \ - \ b_1^0 \\ a_{21}^0 x_1 + a_{22}^0 x_2 \ - \ b_2^0 \end{pmatrix}.$$

Each column of $D(x) \in \mathbb{R}^{2 \times 2}$ is

$$d^k(x) = \Delta_k \left(A^k x - b^k\right) = \Delta_k \begin{pmatrix} a_{11}^k x_1 + a_{12}^k x_2 - b_1^k \\ a_{21}^k x_1 + a_{22}^k x_2 - b_2^k \end{pmatrix},$$

for $k = 1, 2$, so $D(x) = \left[\, d^1(x) \;\; d^2(x) \right]$.

Conditions on 2x2 Real Parametric Systems 2

The condition $(I - D\,D^{\dagger})R_c = 0$ is equivalent to the vanishing of the two minors of $\big[D \mid R_c\big]$. Define

$$f_1(x_1, x_2) = \det \begin{pmatrix} d_1^1(x) & r_1(x) \\ d_2^1(x) & r_2(x) \end{pmatrix} = d_1^1(x)\, r_2(x) - d_2^1(x)\, r_1(x),$$

$$f_2(x_1, x_2) = \det \begin{pmatrix} d_1^2(x) & r_1(x) \\ d_2^2(x) & r_2(x) \end{pmatrix} = d_1^2(x)\, r_2(x) - d_2^2(x)\, r_1(x).$$

$$\begin{aligned} f_1(x_1, x_2) = \Delta_1 \Big[&\left(a_{11}^1 x_1 + a_{12}^1 x_2 - b_1^1\right) \left(a_{21}^0 x_1 + a_{22}^0 x_2 - b_2^0\right) \\ &- \left(a_{21}^1 x_1 + a_{22}^1 x_2 - b_2^1\right) \left(a_{11}^0 x_1 + a_{12}^0 x_2 - b_1^0\right) \Big] = 0, \end{aligned}$$

$$\begin{aligned} f_2(x_1, x_2) = \Delta_2 \Big[&\left(a_{11}^2 x_1 + a_{12}^2 x_2 - b_1^2\right) \left(a_{21}^0 x_1 + a_{22}^0 x_2 - b_2^0\right) \\ &- \left(a_{21}^2 x_1 + a_{22}^2 x_2 - b_2^2\right) \left(a_{11}^0 x_1 + a_{12}^0 x_2 - b_1^0\right) \Big] = 0. \end{aligned}$$

Conditions on 2x2 Real Parametric Systems 3

Here we simulate $A(p) = A^0 + p_1 A^1 + p_2 A^2, b(p) = b^0 + p_1 b^1 + p_2 b^2$, with all matrices drawn from $N(0, 1)$

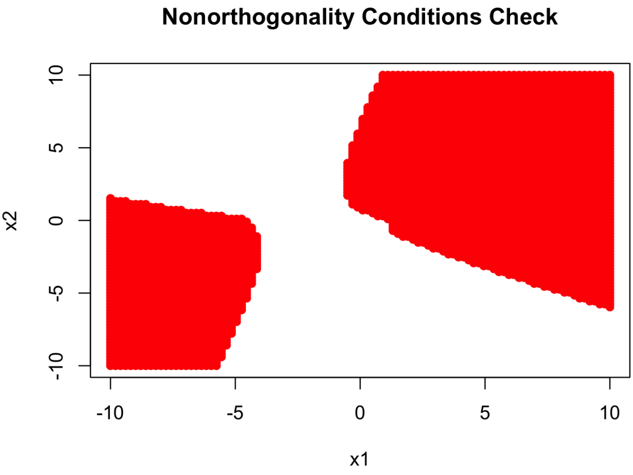


Table of Contents

Translating Bellman Equation into Parametric System

- Need to be careful about reparametrizing Bellman equation for MDPs ($V^\pi = (I - \gamma P^\pi)^{-1} r^\pi$) into a parametric system ($A(p)x - b(p) = 0$)
- Want entries of policy π to be our parameters, but policy-weighted transition matrix P^π is row-stochastic, so there are dependencies between parameter entries
- Technical solve — take intersection of two hyperrectangle parametrizations which together cut out the correct policy simplex — see Appendix

Infinitely Many Linear Inequalities for Value Functions

Theorem

Consider a (PO)MDP. Then $x \in \mathbb{R}^{\mathcal{S}}$ is a feasible value function, meaning that it solves the Bellman equation $(I - \gamma P^{\pi})x - r^{\pi} = 0$ for some $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$, if and only if it solves

$$y^{\top}(A(p^c)x - b(p^c)) \leq \sum_{(o,a)\in\mathcal{O}\times\mathcal{A}} p^{\Delta}_{(o,a)} |y^{\top}(A^{(o,a)}x - b^{(o,a)})|$$

$$y^{\top}(B(v^c)x - c(v^c)) \leq \sum_{(o,a)\in\mathcal{O}\times\mathcal{A}\setminus\{a_o\}} v^{\Delta}_{(o,a)} |y^{\top}(B^{(o,a)}x - c^{(o,a)})|$$

for every $y \in \mathbb{R}^n$, where the matrices are defined in the technical sense given above.

Finitely Many Poly. Inequalities for Value Functions

Theorem

$x \in \mathbb{R}^{\mathcal{S}}$ is a feasible value function, meaning that it solves the Bellman equation $(I - \gamma P^{\pi})x - r_{\pi} = 0$ for some $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$, if and only if

1. *The zonotope condition is satisfied*
2. *The non-orthogonality condition is satisfied*

Table of Contents

Example for 2x2 Real Parametric Systems

$$A^0 = \begin{bmatrix} -0.93 & -2.19 \\ 0.01 & 0.37 \end{bmatrix}, A^1 = \begin{bmatrix} -0.38 & 0.2 \\ 0.41 & -0.16 \end{bmatrix}, A^2 = \begin{bmatrix} -1.25 & 1.60 \\ 0.57 & 2.45 \end{bmatrix},$$

$$\mathrm{b}^0 = \begin{bmatrix} 0.58 \\ 1.44 \end{bmatrix}, b^1 = \begin{bmatrix} -0.80 \\ -0.87 \end{bmatrix}, b^2 = \begin{bmatrix} -1.35 \\ 0.61 \end{bmatrix}, p_1, p_2 \in [0, 1]$$

Connecting POMDPs and Parametric Systems

To use the characterization above for solution sets of parametric systems, we need a parametrization by a hyperrectangle. We can express our parameter set $\Delta_{\mathcal{A}}^{\mathcal{O}}$ as the intersection of two sets parametrized by hyperrectangles, and infer a result by taking the intersection of the solution sets.

First hyperrectangle:

$$A(p) = A^0 + \sum_{o,a} A^{(o,a)} p_{o,a}, \quad \text{with}$$

$$(A^0)_{s,s'} = I_{s,s'}, (A^{(o,a)})_{s,s'} = -\gamma\alpha(s,a;s')\beta(s;o), (o,a) \in \mathcal{O} \times \mathcal{A}$$

$$b(p) = b^0 + \sum_{o,a} b^{(o,a)} p_{o,a}, \quad \text{with}$$

$$(b^0)_s = 0, (b^{(o,a)})_s = r(s;a)\beta(s;o),\ (o,a) \in \mathcal{O} \times \mathcal{A},$$

with parameter $p \in \Delta_{\mathcal{A}}^{\mathcal{O}} \subseteq \mathbb{R}^{\mathcal{O} \times \mathcal{A}}$.

Connecting POMDPs and Parametric Systems 2

Second hyperrectangle: fix some $a_o \in \mathcal{A}$ for each $o \in \mathcal{O}$ and take $\forall (o, a) \in \mathcal{O} \times \mathcal{A} \setminus \{a_o\}$

$$B(v) = B^0 + \sum_{o, a \neq a_o} B^{(o,a)} v_{o,a}, \quad \text{with}$$

$$(B^0)_{s,s'} = (A^0)_{s,s'} + \sum_o (A^{(o,a_o)})_{s,s'},$$

$$(B^{(o,a)})_{s,s'} = (A^{(o,a)})_{s,s'} - (A^{(o,a_o)})_{s,s'},$$

and

$$c(v) = c^0 + \sum_{o, a \neq a_o} c^{(o,a)} v_{o,a}, \quad \text{with}$$

$$(c^0)_s = \sum_o (b^{(o,a_o)})_s,$$

$$(c^{(o,a)})_s = (b^{(o,a)})_s - (b^{(o,a_o)})_s,$$

with parameter $v \in [0,1]^{\mathcal{O} \times \mathcal{A} \setminus \{a_o\}}$.

References I

[1] Robert Dadashi, Adrien Ali Taïga, Nicolas Le Roux, Dale Schuurmans, and Marc G. Bellemare. The Value Function Polytope in Reinforcement Learning, May 2019. arXiv:1901.11524 [cs, stat].

[2] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang,

Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, January 2025. arXiv:2501.12948 [cs].

References IV

[3] Simon S. Du, Sham M. Kakade, Ruosong Wang, and Lin F. Yang. Is a Good Representation Sufficient for Sample Efficient Reinforcement Learning?, February 2020. arXiv:1910.03016 [cs, math, stat].

[4] Leonidas J. Guibas, An Nguyen, and Li Zhang. Zonotopes as bounding volumes. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '03, page 803–812, USA, 2003. Society for Industrial and Applied Mathematics.

[5] Milan Hladík. Enclosures for the solution set of parametric interval linear systems. *Int. J. Appl. Math. Comput. Sci.*, 22(3):561–574, sep 2012.

[6] Guido Montúfar, Keyan Ghazi-Zahedi, and Nihat Ay. Geometry and Determinism of Optimal Stationary Control in Partially Observable Markov Decision Processes. *arXiv:1503.07206*, 2015.

References V

[7] Johannes Müller and Guido Montúfar. The geometry of memoryless stochastic policy optimization in infinite-horizon POMDPs. In *International Conference on Learning Representations*, 2022.

[8] Christos H Papadimitriou and John N Tsitsiklis. The complexity of Markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

[9] Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[10] Jiri Rohn. Systems of linear interval equations. *Linear Algebra and its Applications*, 126:39–78, December 1989.

[11] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

[12] K.J Åström. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, February 1965.