Proceedings of the Thirty-Fifth International Conference on Automated Planning and Scheduling (ICAPS 2025)

Going Topological in Multi-risk Extended Markov Ratio Decision Processes

Alexander Zadorojniy[1], **Orit Davidovich**[1], **Takayuki Osogami**[2]

[1]IBM Research - Israel
[2]IBM Research - Tokyo
zalex@il.ibm.com, orit.davidovich@ibm.com, osogami@jp.ibm.com

Abstract

Incorporating risk into decision making is natural, if one is to address safety concerns or operational limitations. In the context of risk-aware Markov Decision Processes (MDPs), one identifies a notion of risk which is uncertainty driven (e.g., CVaR). Risk, however, may also be inherent to the MDP setup itself, e.g., to taking certain actions. In that case, we would consider a decision policy to be better, if it either increases reward or reduces risk (or both). A simple mathematical formulation that expresses such a notion of improvement is the ratio of reward over risk. Though intuitive, this ratio is inherently non-linear, which introduces challenges for optimization. We provide an algorithm that solves this non-linear problem in the context of multiple risk aspects, extending upon single-risk Extended Markov Ratio Decision Processes (EMRDPs). We show that it is strongly polynomial under a monotonicity assumption over actions, satisfied, for example, in financial market applications (e.g., Quasi-Sharpe Ratio). We tackle non-linearity by integrating Walkup-Wets' topological view of parametric LPs. This topological framework highlights the non-trivial move from a single (EMRDP) to multiple risk aspects, once it is interpreted as moving from triangulations of 1-dimensional to those of m-dimensional polyhedra, with all the topological (and combinatorial) complexities this entails.

1 Introduction

A Markov Decision Process (MDP) (Puterman 1994, Sutton and Barto 2018) is a fundamental model for sequential decision making. Incorporating a notion of risk into decision making is natural if one is to address safety concerns or operational limitations. In the context of risk-aware MDP extensions, one identifies in the literature a notion of risk which is uncertainty driven, stemming, say, from reward fluctuations or modelling errors. One chief example is Conditional Value-at-Risk (CVaR) MDP (Rockafellar, Uryasev et al. 2000) which is widely used in financial markets. Risk, however, may also be inherent to taking certain types of actions at certain types of states. A real-life example here is high- vs. low-volume investment.

Whether in traditional or risk-aware MDPs, finding the optimal policy efficiently is paramount to applicability. In this work we regard efficiency through the lens of strong polynomiality (Zadorojniy, Osogami, and Davidovich 2023, Numerical Stability §7.4).[1] First, for traditional MDPs, optimal policy algorithms were found in strongly-polynomial time in the size of the MDP in the discounted cost model and, in some cases, in the expected average cost model. Since risk-aware MDP extensions are of particular interest in real-world applications, a key question is whether strongly-polynomial optimal policy algorithms exist there as well.

Considering algorithmic efficiency for uncertainty-driven risk, the literature is vast. To highlight just a few results, Meggendorfer (2022) address CVaR minimization, where the algorithm overall runtime is exponential. Chow et al. (2015) minimize an approximate CVaR, while their finite-time convergence error bound cannot be made zero. Borkar and Jain (2014) maximize the expected return such that CVaR is below a threshold; though their algorithm does not hold for infinite horizon problems. For finite horizon problems, they require the separability of the cost function as well as some additional function approximations.

If we regard risk as inherent to the MDP setup itself, we would naturally consider a decision policy to be better, if it either increases reward or reduces risk or both. A simple mathematical formulation that expresses such a notion of improvement most obviously is the ratio of reward over (some expression of) risk. Though intuitive, this ratio is inherently non-linear, which introduces challenges for optimization. Previous work (Zadorojniy, Osogami, and Davidovich 2023) has provided a solution to a special case, known as an Extended Markov Ratio Decision Process (EMRDP), in which a single risk aspect in the denominator is exponentiated.

In this work, we provide an algorithm that solves this non-linear problem in far greater generality, incorporating multiple risk aspects, and show that it is strongly polynomial under a monotonicity assumption over actions. Our solution imposes natural requirements on the denominator, $\varphi(\boldsymbol{\alpha})$, that aggregates multiple risk aspects $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$, namely, that it be positive, monotonic, and concave. Two immediate examples that satisfy these requirements are (i) the linear aggregator $\varphi(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \omega_i \alpha_i$, for $\omega_i > 0$, and (ii)

[1]The running time of a strongly polynomial-time algorithm is bounded polynomially as a function of problem size, independent of numerical input size.

the power aggregator $\varphi(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i^{\omega_i}$, for $\omega_i \in (0, 1]$. The latter (ii) incorporates the MDP ($\omega_i \equiv 0$), the Markov Ratio Decision Process (MRDP) ($m = 1, \omega_1 = 1$, (Derman 1962)), and the quasi-Sharpe ratio ($m = 1, \omega_1 = 0.5$). The linear aggregator (i) can be shown, in fact, to translate back to MRDP and hence reduces to the latter example (ii).

We tackle non-linearity by integrating two crucial components. The first is CMDP theory (Altman 1999), which we use to formulate a parametric linear programme (LP) reduction of the non-linear problem. The second is Walkup-Wets' topological formulation of parametric LPs (Walkup and Wets 1969). Such topological formulation reveals the inherent combinatorial nature of the problem and, crucially, localizes the search for the global and (what turns out to be) deterministic optimal policy. It also highlights the non-trivial move from a single to multiple risk aspects, once it is interpreted as moving from triangulations of 1-dimensional polyhedra to those of m-dimensional polyhedra, with all the topological (and combinatorial) complexities this entails. Our topological interpretation exposes the relation of multi-risk EMRDP to deep questions concerning the statistics of graph diameters, which explains why strong polynomiality for single or multi-risk EMRDP, in its utmost generality, would be a hard nut to crack.

A high level outline of our arguments is as follows. CMDP theory (5) allows us to cast EMRDP (9) in terms of a parametric LP (Proposition 2). Using Walkup-Wets theory (Theorem 1) applicable to parametric LPs (11), we can establish a minimal search set (Corollary 6) to solve (9) per reward vector r. Moreover, the triangulation arising from Walkup-Wets theory allows us to establish a search procedure (25) via a local criterion (Theorem 7), which is key both for establishing and proving the correctness of Algorithm 1 (Theorem 8) as well as its strong polynomiality and runtime under a monotinicity assumption (Theorem 10).

2 Background

Markov Decision Processes

A Markov Decision Process (MDP) is a tuple $\langle \mathcal{S}, \mathcal{A}, P, r \rangle$ that consists of finite sets of states $\mathcal{S}$ and actions $\mathcal{A}$, a transition probability matrix P of size $n \times n\,k$, where $n = |\mathcal{S}|$ and $k = |\mathcal{A}|$, and an immediate reward vector $r \in \mathbb{R}^{n\,k}$. Let S_t and A_t denote the random variables representing the state and action of the process at time t and consider an infinite time horizon $t \to \infty$. We define $P = [P(a_1), \cdots, P(a_k)]$ to be the concatenation of the k square matrices $P(a)$ of size $n \times n$, for $a \in \mathcal{A}$, where the entries of $P(a)$ equal $P(a)_{s',s} = \mathbb{P}(S_{t+1} = s' | S_t = s, A_t = a)$. We let $r(s, a)$ denote the immediate reward for taking action a at state s.

Throughout this paper, we consider stationary policies $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$, where $\Delta_{\mathcal{A}}$ denotes the set of probability measures over $\mathcal{A}$, and denote their collection by Π. A deterministic policy is a stationary policy whose image is contained in the subset of Dirac measures over $\mathcal{A}$. We denote the subset of deterministic policies by $\Pi^{\text{det}} \subset \Pi$. A strictly m-randomized policy is a stationary policy $\pi \in \Pi$ whose support satisfies $|\text{Supp}(\pi)| = n + m$. We denote their collection by $\Pi^{m-\text{rnd}} \subset \Pi$.

Fix an initial state distribution $\nu \in \Delta_{\mathcal{S}}$ and a discount factor $\beta \in (0, 1]$. We consider two reward models in the infinite horizon setting: discounted and expected average. We use the occupation measure to represent these two reward models in a unified manner. Specifically, for $\pi \in \Pi$, let

$$\rho^{\pi}(s, a) = (1 - \beta) \sum_{t=0}^{\infty} \beta^t \, \mathbb{P}^{\pi}(S_t = s, A_t = a) \quad (1)$$

be the occupation measure for the discounted reward model, where $\beta \in (0, 1)$, and let

$$\rho^{\pi}(s, a) = \lim_{T \to \infty} \frac{1}{T} \sum_{t<T} \mathbb{P}^{\pi}(S_t = s, A_t = a) \quad (2)$$

be the occupation measure for the expected average reward model where we regard β as equal to 1. Then, the discounted reward is given by

$$\mathfrak{r}(\pi) := r^{\top} \rho^{\pi} = (1 - \beta) \sum_{t=0}^{\infty} \beta^t \, \mathbb{E}_t^{\pi}[r(S_t, A_t)],$$

where $\mathbb{E}_t^{\pi}$ denotes the expectation with respect to the corresponding occupation measure induced by π and ν at time t, and the expected average reward is given by

$$\mathfrak{r}(\pi) := r^{\top} \rho^{\pi} = \lim_{T \to \infty} \frac{1}{T} \sum_{t<T} \mathbb{E}_t^{\pi}[r(S_t, A_t)].$$

Constrained Markov Decision Processes

Constrained MDPs (CMDPs) incorporate multiple constraints into the MDP framework (Altman 1999). A CMDP is a tuple $\langle \mathcal{S}, \mathcal{A}, P, r, d_1, \ldots, d_m \rangle$ ($n + m < nk$) where $d_i \in \mathbb{R}^{n\,k}$ are immediate risk vectors; $d_i(s, a)$ represents the i-th immediate risk for taking action a at state s. The i-th risk aspect of a policy π is given by

$$\mathfrak{d}_i(\pi) = d_i^{\top} \rho^{\pi}, \quad (3)$$

where the occupation measure ρ^{π} of $\pi \in \Pi$ is defined according to the model of choice (1) or (2).

The objective of the CMDP is to maximize the reward while constraining the i-th risk aspect to a target value α_i.

$$\text{CMDP}(\alpha_1, \ldots, \alpha_m) : \quad \max_{\pi \in \Pi} \{ \, \mathfrak{r}(\pi) \mid \mathfrak{d}_i(\pi) = \alpha_i, i = 1, \ldots, m \, \}. \quad (4)$$

Both reward $\mathfrak{r}(\pi)$ and risks $\mathfrak{d}_i(\pi)$ depend on ν and β, though we suppress this dependence in the formulation of (4).

We adopt two standard assumptions: uniqueness and irreducibility. A CMDP is *irreducible* if every deterministic policy of its underlying MDP induces an irreducible Markov chain.

Assumption 1 (Uniqueness)**.** *If feasible, the optimal policy for CMDP*$(\alpha_1, \ldots, \alpha_m)$ *is unique.*

Assumption 2 (Irreducibility)**.** *CMDP*$(\alpha_1, \ldots, \alpha_m)$ *is irreducible.*

Uniqueness can be achieved by perturbation of r and d_i (Megiddo and Chandrasekaran 1989, Zadorojniy, Even, and Shwartz 2009). Irreducibility can be achieved by bidirectionally connecting an arbitrary state, typically an initial state s_0 (i.e., $\nu = \delta_{s_0}$), with all other states under all actions with "small" transition probabilities.

Theorems 3.3 and 4.3 of Altman (1999) establish the equivalence between (4) and the LP

$$\text{CMDP}(\alpha_1, \dots, \alpha_m) : \\ \max_{\rho \in \mathbb{R}^{nk}} \left\{ r^T \rho \;\middle|\; Q\,\rho = \mu, d_i^T \rho = \alpha_i, i = 1, \dots, m, \rho \geq 0 \right\} \tag{5}$$

The matrix Q and vector μ are defined according to the reward/risk model of choice. First, denote

$$\tilde{Q}(\beta) = [I - \beta\, P(a_1), \cdots, I - \beta\, P(a_k)], \tag{6}$$

where I is the $n \times n$ identity matrix and $\beta \in (0, 1]$. For the discounted model, $\beta \in (0, 1)$, we have $Q = \tilde{Q}(\beta)$ and $\mu = (1 - \beta)\nu \in \mathbb{R}^n$, while, for the expected average model, $\beta = 1$, Q is the $(n + 1) \times n\,k$ matrix defined by appending a row of 1s to $\tilde{Q}(1)$, and $\mu = e_{n+1} \in \mathbb{R}^{n+1}$ is the standard basis vector with a single 1 at the $(n + 1)$-th entry.

The map, defined for $\nu \in \Delta_{\mathcal{S}}$ and $\beta \in (0, 1]$,

$$\rho_{\nu,\beta} : \Pi \to \Delta_{\mathcal{S}\times\mathcal{A}}, \quad \rho_{\nu,\beta} : \pi \mapsto \rho^\pi, \tag{7}$$

sending π to its occupation measure ρ^π, is bijective and establishes the equivalence between (4) and (5). The feasibility set

$$\Omega_{\nu,\beta} = \left\{ \rho \;\middle|\; \tilde{Q}(\beta)\rho = (1 - \beta)\nu, \mathbf{1}^T\rho = 1, \rho \geq 0 \right\} \tag{8}$$

constitutes the necessary and sufficient condition for $\rho \in \Delta_{\mathcal{S}\times\mathcal{A}}$ to be in the image of $\rho_{\nu,\beta}$ (Altman 1999, Theorem 3.2, Theorem 4.2), i.e., for ρ to be the occupation measure of some stationary policy in the reward/risk model of choice. In particular, $\Omega_{\nu,\beta}$ is non-empty.

Employing the notation $\mathbb{R}_{\geq 0} = \{\, x \in \mathbb{R} \mid x \geq 0 \,\}$ and $\mathbb{R}_{>0} = \{\, x \in \mathbb{R} \mid x > 0 \,\}$, which extends coordinate-wise to $\mathbb{R}^N_{\geq 0}$ and $\mathbb{R}^N_{>0}$, we assume further that any stationary policy yields non-negative reward and positive risks.

Assumption 3. *(Positiveness) For any $\pi \in \Pi$, we have $\mathfrak{r}(\pi) \in \mathbb{R}_{\geq 0}$ and $\mathfrak{d}_i(\pi) \in \mathbb{R}_{>0}$ for $i = 1, \dots, m$.*

3 Multi-risk Extended Markov Ratio Decision Processes

We first introduce the notion of a *risk aggregator* that incorporates all risk aspects of a CMDP §2. We denote a vector of multi-risk values by $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)$. Risk inequalities $\boldsymbol{\alpha} \leq \boldsymbol{\alpha'}$ (or $\boldsymbol{\alpha} < \boldsymbol{\alpha'}$) are satisfied component wise.

Definition 1. *A strictly positive function $\varphi : \mathbb{R}^m_{\geq 0} \to \mathbb{R}_{>0}$ is a risk aggregator if it is strictly monotonic, i.e., $\boldsymbol{\alpha} < \boldsymbol{\alpha'}$ implies $\varphi(\boldsymbol{\alpha}) < \varphi(\boldsymbol{\alpha'})$, and concave.*

Let $\varphi : \mathbb{R}^m_{\geq 0} \to \mathbb{R}_{>0}$ be a risk aggregator. Extending upon Zadorojniy, Osogami, and Davidovich (2023), we define a φ-aggregated multi-risk Extended Markov Ratio Decision Process (EMRDP) using the same underlying tuple $\langle \mathcal{S}, \mathcal{A}, P, r, d_1, \dots, d_m \rangle$ of a CMDP. However, in contrast with CMDPs, its objective is to maximize the ratio of the reward to the φ-aggregated risk. Explicitly, the optimization problem for a φ-aggregated multi-risk EMRDP is

$$\text{EMRDP}(\varphi) : \\ \max_{\pi \in \Pi} \left\{ \frac{\mathfrak{r}(\pi)}{\varphi(\mathfrak{d}_1(\pi), \dots, \mathfrak{d}_m(\pi))} \;\middle|\; Q\,\rho^\pi = \mu, \rho^\pi \geq 0 \right\} \tag{9}$$

As a key example for $\boldsymbol{\omega} = (\omega_1, \dots, \omega_m)$, $\omega_i \in [0, 1]$,

$$\varphi_{\boldsymbol{\omega}}(\mathfrak{d}_1(\pi), \dots, \mathfrak{d}_m(\pi)) = \sum_{i=1}^m \mathfrak{d}_i^{\omega_i}(\pi), \tag{10}$$

provides us with a strictly positive, monotonic, and concave risk aggregator. When $\omega_i \equiv 0$, this reduces to a MDP (Puterman 1994), and, when $m = 1$ and $\omega_1 = 1$, it reduces to a MRDP (Derman 1962, Aggarwal, Chandrasekaran, and Nair 1977). The risk aggregator (10) thus interpolates between MDP and MRDP. The case $m = 1$ for (10) has been resolved (Zadorojniy, Osogami, and Davidovich 2023). though the general question of strong polynomiality remains open, even for the discounted reward model. Note that the linear aggregator $\sum_{i=1}^m \omega_i \mathfrak{d}_i(\pi) = \sum_{i=1}^m \omega_i d_i^\top \rho^\pi$ is, in fact, a special case of (10). Applying (3), we can replace $\sum_{i=1}^m \omega_i d_i$ with a single d and use 1 in the exponent.

4 A Topological Perspective

We provide a topological perspective of multi-risk EMRDP that allows us to establish an algorithm to solve (9) with a strongly polynomial guarantee under a monotonicity assumption. We briefly define the notion of a (conic) triangulation and then explain how it fits into the multi-risk EMRDP picture. An exhaustive and rigorous review of triangulations can be found in de Loera, Rambau, and Santos (2010).

Triangulations

Consider a real full-rank matrix $A = (a_1, \dots, a_N)$ of size $D \times N$, with $a_j \in \mathbb{R}^D$ denoting its columns. Define the *cone* of A to be $\text{Cone}(A) = \{Ax \mid x \in \mathbb{R}^N_{\geq 0}\} \subseteq \mathbb{R}^D$. Let $J = \{j_1, \dots, j_l\} \subseteq [N]$ be a subset of column indices, and let $A_J = (a_{j_1}, \dots, a_{j_l})$ denote the sub-matrix of size $D \times |J|$. $\text{Cone}(A_J)$ is considered a *sub-configuration* of $\text{Cone}(A)$. Additionally, we allow $J = \emptyset$ with $\text{Cone}(A_J) := \{0\}$ as a sub-configuration of $\text{Cone}(A)$.

Definition 2 (de Loera, Rambau, and Santos (2010); Definition 2.5.7)**.** *A polyhedral subdivision of $\text{Cone}(A)$ is a collection $\mathcal{C}$ of its sub-configurations that covers $\text{Cone}(A)$ and is closed w.r.t. faces and intersections.*

Each element $\text{Cone}(A_J)$ of a polyhedral subdivision $\mathcal{C}$ of $\text{Cone}(A)$ is called a *cell*. The dimension k of a cell equals the rank $\text{rk}(A_J)$. We regard the dimension of $\text{Cone}(A_\emptyset)$ to be zero. A k-cell $\text{Cone}(A_J)$ is a cell of dimension k. A k-cell is *simplicial* if $k = |J|$. Its faces are cells of the form $\text{Cone}(A_J)$ for $\emptyset \neq J' \subseteq J$.

Definition 3. *A polyhedral subdivision of $\text{Cone}(A)$ is a triangulation if all of its cells are simplicial.*

We refer to the cells of a triangulation $\mathcal{T}$ as *simplices*. A k-simplex of $\mathcal{T}$ is a simplicial $(k+1)$-cell. We denote the set of all top-dimensional simplices, i.e., its D-cells, by $\mathcal{T}^{\rm top}$ and the set of all co-dimension-1 simplices, i.e., its $(D-1)$-cells by $\mathcal{T}^{\rm codim-1}$. For $k \geq 0$, the *k-skeleton* $\mathcal{T}^k$ of a triangulation $\mathcal{T}$ is defined as the union of all l-simplices of $\mathcal{T}$ for $0 \leq l \leq k$. In particular, the 0-skeleton consists of all *vertices*, namely, 0-simplices of $\mathcal{T}$. We also employ the notation $V(\mathcal{T})$ for $\mathcal{T}^0$. Figure 1 shows a triangulation of a cone spanned by vectors $u_1, \ldots, u_9$. The triangulation involves ten top dimensional simplices. As an example, Cone(A_{J_0}), Cone(A_{J_1}), and Cone(A_{J_2}) are 0-, 1-, and 2-simplices, respectively, for $J_0 = \{u_1\}$, $J_1 = \{u_1, u_2\}$, and $J_2 = \{u_1, u_2, u_4\}$.

Basis Decomposition

Fix a real full-rank matrix $A = (a_1, \ldots, a_N)$ of size $D \times N$, with $a_j \in \mathbb{R}^D$ denoting its columns, as in §4, and a vector $r \in \mathbb{R}^N$ and consider the parametric LP

$$LP_{A,r}(b) : \max_{\rho \in \mathbb{R}^N} \{\, r^T\rho \;\big|\; A\rho = b, \rho \geq \mathbf{0} \,\} \tag{11}$$

where $b \in \mathbb{R}^D$ is allowed to vary. The Basis Decomposition Theorem of Walkup and Wets (1969) captures the behavior of $LP_{A,r}(b)$ in terms of b.

Theorem 1 (Basis Decomposition)**.** *Consider $LP_{A,r}(b)$ in* (11) *and assume rk$(A) = D$ (full rank). Then:*

(i) $LP_{A,r}(b)$ is feasible iff $b \in$ Cone(A).

(ii) $LP_{A,r}(b)$ is bounded for all $b \in$ Cone(A) iff $\ker A \cap \mathbb{R}^N_{\geq 0} = \{0\}$.

(iii) If $LP_{A,r}(b)$ is bounded, then there exists a triangulation $\mathcal{T}$ of Cone(A) such that each D-cell Cone(A_J) of $\mathcal{T}$ with $J = \{j_1, \ldots, j_D\} \subseteq [N]$ constitutes an optimal basis for all $b \in$ Cone(A_J).

Assuming $\mathrm{rk}(A) = D$ and $\ker A \cap \mathbb{R}^N_{\geq 0} = \{0\}$, Theorem 1(i)-(ii) stipulates that the primal problem (11) is feasible and bounded for $b \in \mathrm{Cone}(A)$. The dual problem

$$LP^*_{A,r}(b) : \min_{y} \{\, y^T b \;\big|\; y^T A \geq r^T \,\} \tag{12}$$

is feasible and bounded and Strong Duality holds. It provides us with a well-defined map

$$\rho_* : \mathrm{Cone}(A) \to \mathbb{R}^N, \quad \rho_*(b) := \mathrm{argmax}\, LP_{A,r}(b). \tag{13}$$

Part (iii) of Theorem 1 stipulates the existence of a unique triangulation $\mathcal{T}_r$ of Cone(A) determined by r that encodes the optimal solutions of $LP_{A,r}(b)$. The triangulation $\mathcal{T}_r$ in part (iii) is defined so that Cone(A_J) is a top-dimensional simplex of $\mathcal{T}_r$ iff there exists $y_* \in \mathbb{R}^D$ such that

$$y_*^T A_J = r_J^T, \quad y_*^T A_{[N]\setminus J} > r^T_{[N]\setminus J} \tag{14}$$

(Sturmfels and Thomas 1997), which implies that $y_* \in \mathbb{R}^D$ is both feasible and optimal for (12), since, by definition, the index set J constitutes an optimal basis for $b \in \mathrm{Cone}(A_J)$ iff A_J is invertible and an optimal solution $\rho_*(b)$ of $LP_{A,r}(b)$ for $b \in \mathrm{Cone}(A_J)$ is given by

$$\rho_*(b)_J = A_J^{-1} b, \quad \rho_*(b)_{[N]\setminus J} = \mathbf{0}. \tag{15}$$

Lemma 1. *Assuming rk$(A) = D$ and* $\ker A \cap \mathbb{R}^N_{\geq 0} = \{0\}$, *the function $LP_{A,r}$: Cone$(A) \to \mathbb{R}$ mapping $b \in$ Cone(A) to $r^T\rho_*(b)$ is concave, continuous, and piece-wise linear (PWL).*

Parametric LP

In this section, we introduce a unified parametric LP formulation for CMDP$(\alpha_1, \ldots, \alpha_m)$ (5) for both reward/risk models and show that it satisfies the conditions of Theorem 1.

Recall our definition (6) of the $n \times nk$-matrix $\tilde{Q}(\beta)$ for $\beta \in (0, 1]$. By the Levy–Desplanques Theorem (Horn and Johnson 2013, Theorem 6.1.10), $I - \beta\, P(a)$ is non-singular for every $a \in \mathcal{A}$ and $\beta \in (0, 1)$. Additionally, since $P(a)$ is a transition probability matrix, $\mathbf{1}^T\tilde{Q}(1) = \mathbf{0}^T$. Hence

$$\mathrm{rk}(\tilde{Q}(\beta)) = n, \quad \beta \in (0, 1), \tag{16}$$

$$\mathrm{rk}(\tilde{Q}(1)) \leq n - 1, \tag{17}$$

The homogenization of $\tilde{Q}(\beta)$,

$$\tilde{Q}(\beta)_{\rm hom} = \begin{pmatrix} \tilde{Q}(\beta) \\ \hline \mathbf{1}^T \end{pmatrix},$$

allows us to combine both equality constraints in $\Omega_{\nu,\beta}$ (8). By Assumption 2, whereby the Perron-Frobenius Theorem applies, we have

$$\mathrm{rk}(\tilde{Q}(1)_{\rm hom}) = n, \tag{18}$$

which also implies equality in (17). We can thus choose a new representation of fixed rank.

Lemma 2. *Let $\beta \in (0, 1]$. There exist a parametric family of matrices $Q(\beta)$ of size $n \times nk$ and a parametric family of vectors $\mu(\beta) \in \mathbb{R}^n$ such that*

$$\Omega_{\nu,\beta} = \{\, \rho \mid Q(\beta)\rho = \mu(\beta), \rho \geq \mathbf{0} \,\},$$

for $\Omega_{\nu,\beta}$ in (8)*, and rk$(Q(\beta)) = n$ for all $\beta \in (0, 1]$.*

From this point on, we will assume $\nu \in \Delta_{\mathcal{S}}$ is fixed and suppress it in our notation. As an immediate consequence of Lemma 2, we have the following proposition.

Proposition 2. *Let $\beta \in (0, 1]$. For every immediate reward vector $r \in \mathbb{R}^{n\,k}$, CMDP$(\alpha_1, \ldots, \alpha_m)$ is equivalent to $LP_{A,r}(b)$ with A and b given by*

$$A \equiv A_{\beta,d_1,\ldots,d_m} = \begin{pmatrix} Q(\beta) \\ \hline d_1^T \\ \vdots \\ d_m^T \end{pmatrix}, b \equiv b_\beta = \begin{pmatrix} \mu(\beta) \\ \hline \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} \tag{19}$$

Following our notation in §4 applied to A and b in (19), we have $D = n + m$ and $N = nk$. We can show that $LP_{A,r}(b)$ satisfies the conditions of Theorem 1. First, by Assumption 1, $A_{\beta,d_1,\ldots,d_m}$ is full rank,

$$\mathrm{rk}(A_{\beta,d_1,\ldots,d_m}) = n + m. \tag{20}$$

Second, we have the following lemma.

Lemma 3. $\ker A_{\beta,d_1,\ldots,d_m} \cap \mathbb{R}^{(n+m)\times nk}_{\geq 0} = \{\mathbf{0}\}$.

Applying Theorem 1, we have the following corollary.

Corollary 3. *For A as in (19), each immediate reward vector r determines a unique triangulation $\mathcal{T}_r$ of Cone(A).*

The Induced Triangulation

Proposition 2 establishes the equivalence between CMDP$(\alpha_1, \ldots, \alpha_m)$ and $LP_{A,r}(b)$ for A and b in (19). In §4, we considered the function $LP_{A,r} : \text{Cone}(A) \to \mathbb{R}$ assigning to $b \in \text{Cone}(A)$ the value $LP_{A,r}(b)$ (11). Given this equivalence, we may now regard CMDP as a function by restricting $LP_{A,r}$ to b of the form $b_\beta = (\mu(\beta), \boldsymbol{\alpha})$. The domain for CMDP is thus determined by the intersection of the m-dimensional affine subspace

$$H \equiv H_\beta = \{\, b_\beta \mid \boldsymbol{\alpha} \in \mathbb{R}^m \,\} \subset \mathbb{R}^{n+m} \tag{21}$$

with Cone(A) (see H of dimension $m = 2$ in Figure 1). This intersection is an m-dimensional convex polyhedron

$$P = H \cap \text{Cone}(A) \subset \mathbb{R}^{n+m}. \tag{22}$$

We may assume with probability 1 that H cuts Cone(A) transversally at the interior of each $\mathcal{T}_r$ cell.[2] Considering the $\boldsymbol{\alpha}$-parametrization of $b = (\mu(\beta), \boldsymbol{\alpha})$, we can alternately regard P as a convex polytope in $\mathbb{R}^m$. Thus, we can either regard CMDP as a function that evaluates $b \in P \subset \mathbb{R}^{n+m}$ at $LP_{A,r}(b)$ or regard it as a function that evaluates $\boldsymbol{\alpha} \in P \subset \mathbb{R}^m$ at $LP_{A,r}((\mu(\beta), \boldsymbol{\alpha}))$. Either way, by Lemma 1, we have the following.

Corollary 4. *CMDP is concave and PWL continuous over P.*

By Corollary 3, we have a unique triangulation $\mathcal{T}_r$ of Cone(A). Thus, we get an induced triangulation $\mathcal{T}_P$ of P. By dimension considerations, the vertices $V(\mathcal{T}_P)$ of $\mathcal{T}_P$ (i.e., its 0-simplices) correspond to optimal deterministic policies. The 0-simplices of $\mathcal{T}_P$ correspond to those $D - m = (n + m) - m = n$-simplices of $\mathcal{T}_r$ that intersect H (trasversally), namely, those occupation measures with $|\text{Supp}(\rho)| = n$. By similar dimension considerations, the interiors of edges of $\mathcal{T}_P$ (i.e., 1-simplices) correspond to optimal strictly 1-randomized policies, and, so on, the interiors of m-simplices of $\mathcal{T}_P$ correspond to strictly m-randomized policies.

Proposition 5. *The expression $\frac{\mathfrak{r}(\pi)}{\varphi(\mathfrak{d}_1(\pi),\ldots,\mathfrak{d}_m(\pi))}$ restricted to an m-simplex τ of $\mathcal{T}_P$ is quasiconvex as a continuous function of $(\alpha_1, \ldots, \alpha_m) \in \tau$.*

Corollary 6. *The expression $\frac{\mathfrak{r}(\pi)}{\varphi(\mathfrak{d}_1(\pi),\ldots,\mathfrak{d}_m(\pi))}$ defined over P achieves its optima at $V(\mathcal{T}_P)$, namely, at deterministic policies $\pi \in \Pi^{\text{det}}$.*

5 Solving Multi-risk EMRDP

The triangulation of Cone$(A_{\beta,d_1,\ldots,d_m})$ projected onto the affine risk subspace H_β is inherent to the optimization landscape of the objective (9) in that its vertices constitute the minimal search set as compared to the maximal search set over all deterministic policies.

[2]Transversal intersection captures the notion of "generic" intersection. Specifically, H intersects Cone(A) transversally if its intersection with each n-cell τ of $\mathcal{T}_r$ is either empty or satisfies $T_p\tau + T_pH \cong \mathbb{R}^{n+m}$ at an interior intersection point p.

Vertex Order

Following Corollary 6, it is possible to find the optimal policy for the φ-aggregated multi-risk EMRDP (9) via its parametric LP formulation $LP_{\beta,r,d_1,\ldots,d_m}(b) := LP_{A,r}(b)$ with A and b as in (19). To that end, we need a systematic and efficient way to traverse the vertices in $V(\mathcal{T}_P)$. To avoid cumbersome notation, we continue to identify $A \equiv A_{\beta,d_1,\ldots,d_m}$, $b \equiv b_\beta$, $L \equiv L_\beta$, and $H \equiv H_\beta$ throughout this section.

By Definition 1, a risk aggregator φ is concave. Its *level sets* are subsets of the form $\varphi^{-1}(c)$ for $c \in \mathbb{R}_{>0}$. Considering the $\boldsymbol{\alpha}$-parametrization of $b = (\mu(\beta), \boldsymbol{\alpha}) \in P \subset \mathbb{R}^{n+m}$, we overload our notation to define φ over P:

$$\varphi(b) := \varphi(\boldsymbol{\alpha}), \quad b = (\mu(\beta), \boldsymbol{\alpha}) \in P. \tag{23}$$

Let $V_P = V(\mathcal{T}_P)$ be the set of vertices of the induced triangulation $\mathcal{T}_P$. Recall that a vertex $v \in V_P$ corresponds to a unique deterministic policy with its occupation measure ρ. For $b = A\rho \in P$, we employ the same notation to define

$$\varphi(v) := \varphi(b) \quad \text{or} \quad \varphi(\rho) := \varphi(b).$$

We may assume with probability 1 that V_P is in general position w.r.t. the level sets of φ, namely, no two vertices lie on the same level set. This results in a complete order on V_P where $v \preceq v'$ iff $\varphi(v) \leq \varphi(v')$ with equality iff $v = v'$. The complete order $\preceq$ establishes V_P as sequence $v_1, \ldots, v_{|V_P|}$:

$$\begin{aligned} v_1 &= \operatorname*{argmin}_{v \in V_P} \varphi(v) \\ v_i &= \operatorname*{argmin}_{v \in V_P \setminus \{v_1,\ldots,v_{i-1}\}} \varphi(v), \quad i \geq 2 \end{aligned} \tag{24}$$

We can establish V_P as a sequence via an alternate construction,

$$\begin{aligned} &u_1 = \operatorname*{argmin}_{v \in V_P} \varphi(v), \quad \mathcal{P}_1 = \{u_1\} \\ &\mathcal{P}_{i+1} = \mathcal{P}_i \setminus \{u_i\} \cup V(\text{st}_{u_i}) \setminus \{u_1, \ldots, u_i\}, \quad i \geq 1 \\ &u_i = \operatorname*{argmin}_{v \in \mathcal{P}_i} \varphi(v), \quad i \geq 2 \end{aligned} \tag{25}$$

where the star $V(\text{st}_{u_i})$ of u_i consists of the vertices of all top-dimensional simplicies of $\mathcal{T}_P$ incident to $u_i \in V_P$ (see §4). By construction, there exists a minimal index $i_{\min}$ for which $\mathcal{P}_i = \emptyset$ for all $i > i_{\min}$, at which point the sequence of vertices terminates. As an illustration, the level sets plotted on the intersecting plane in Figure 1, superimposed on a triangulation of a 2-dimensional P, establish $\mathcal{P}_1 = \{u_1\}, \mathcal{P}_2 = \{u_2, u_3, u_4\}$, where u_2 is the minimum of $\mathcal{P}_2$ and its star encompasses $\{u_1, u_4, u_6, u_8\}$, thus $\mathcal{P}_3 = \{u_3, u_4, u_6, u_8\}$, and, similarly, $\mathcal{P}_4 = \{u_4, u_5, u_6, u_7, u_8\}, \ldots, \mathcal{P}_9 = \{u_9\}$ and $\mathcal{P}_i = \emptyset, \forall i > i_{\min} = 9$.

Theorem 7. $v_i = u_i, i = 1, \ldots, |V_P|$.

Algorithm Description

Following Theorem 7, we provide Algorithm 1 to traverse $V_P = V(\mathcal{T}_P)$ by order of φ-aggregated risk (25). Corollary 6 guarantees that we will find the optimal policy for the φ-aggregated multi-risk EMRDP (9) within this sequence.

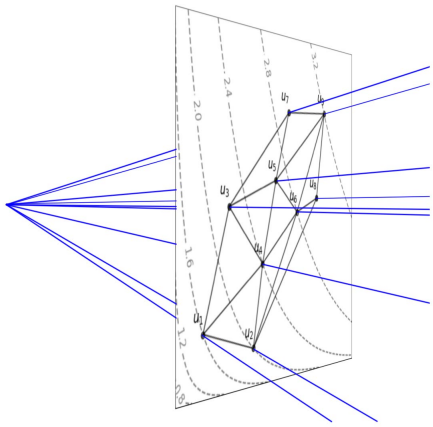

Figure 1: The level sets of the risk aggregator $\varphi_{\boldsymbol{\omega}}(\alpha_1, \ldots, \alpha_m) = \sum_{i=1}^{m} \alpha_i^{\omega_i}$ (10), with $\omega_1 = 0.5, \omega_2 = 0.3$ ($m = 2$), determine a complete order $v_1, \ldots, v_9$ on the vertices $V_P = V(\mathcal{T}_P)$ of the induced triangulation $\mathcal{T}_P$ of P.

Algorithm 1 is initialized (Line 16) with the minimal vertex, i.e., the deterministic policy whose occupation measure is given by

$$\rho_{\min} = \operatorname*{argmin}_{\rho} \{\varphi(\rho) \mid Q(\beta)\rho = \mu(\beta), \rho \geq 0\} \qquad (26)$$

The set $\mathcal{P}$, updated at each iteration (Lines 21-22), corresponds to $\mathcal{P}_i$, beginning with $\mathcal{P}_1$. By definition (25), the optimal occupation measure ρ^* (Line 18) we get at each iteration corresponds to u_i. The necessary and sufficient conditions for m-randomized stationary policies to come next along the Pareto front, that is, for a top-dimensional simplex of $\mathcal{T}_P$ to be in st_{u_i}, are evaluated in Lines 2-4. Note that the limit on the number of such optimal bases, i.e., top-dimensional simplices, is triangulation specific. Strong Duality allows us to combine the top-dimensional simplex condition (14) together with (15) to establish the Pareto front. The condition in Line 4, in particular, is a consequence of (14) and (15). We add the next vertices, i.e., deterministic policies, to $\mathcal{P}$ in Line 6, so long as the φ-aggregated risk does not decrease (Line 9). As noted in Corollary 4, CMDP is concave and PWL continuous over P, thus, Algorithm 1 terminates with the stopping criterion in Line 23.

Theorem 8. *Under Assumptions 1-3, Algorithm 1 returns the policy π^* that maximizes $\frac{\mathfrak{r}(\pi)}{\varphi(\mathfrak{d}_1(\pi), \ldots, \mathfrak{d}_m(\pi))}$ over all stationary policies $\pi \in \Pi$.*

Strong Polynomiality

Algorithm 1 runs in strongly polynomial time with the following additional assumption.

Assumption 4 (Monotonicity)**.** *There is a linear order $(a_1, \ldots, a_k)$ on $\mathcal{A}$ such that for each $\pi \in \Pi^{\mathrm{det}}$ and $s \in \mathcal{S}$*

$$(\mathfrak{d}_1(\pi^{s,a_1}), \ldots, \mathfrak{d}_m(\pi^{s,a_1})) < \ldots < (\mathfrak{d}_1(\pi^{s,a_k}), \ldots, \mathfrak{d}_m(\pi^{s,a_k})). \qquad (27)$$

Algorithm 1: Multi-risk EMRDP ($A \equiv A_{\beta,d_1,\ldots,d_m}$, $H \equiv H_\beta$, $\mathcal{J} \equiv [nk]$)

1: **procedure** NEXTPOLICY($\mathcal{P}, \rho^*, \alpha^*$)
2: **for** $I \subset \mathcal{J} \setminus \mathrm{Supp}(\rho^*)$, $|I| = m$ **do**
3: $J \leftarrow I \cup \mathrm{Supp}(\rho^*)$
4: **if** $r_J^T A_J^{-1} A - r^T \geq 0$ **then** ▷ $\det(A_J) \neq 0$
5: $\mathcal{P}_J \leftarrow \{\}$
6: **for** $b' = H \cap \mathrm{Cone} A_{J'}$, $|J'| = n$, $J' \subset J$, $J' \neq \mathrm{Supp}(\rho^*)$ **do**
7: Add ρ' to $\mathcal{P}_J$ s.t. $\rho'_J = A_J^{-1} b'$ and 0 elsewhere
8: **end for**
9: **if** $\forall \rho' \in \mathcal{P}_J : \varphi(\rho') \geq \alpha^*$ **then** ▷ $|\mathcal{P}_J| = m$
10: $\mathcal{P} \leftarrow \mathcal{P} \cup \mathcal{P}_J$
11: **end if**
12: **end if**
13: **end for**
14: **return** $\mathcal{P}$
15: **end procedure**
16: Initialize: $\Pi \leftarrow \{\}$, $\mathcal{P} \leftarrow \{\rho_{\min}\}$ ▷ Eq. (26)
17: **do**
18: $\rho^* \leftarrow \operatorname*{argmin}_{\rho \in \mathcal{P}} \varphi(\rho)$, $\alpha^* \leftarrow \varphi(\rho^*)$, $r^* \leftarrow r^T \rho^*$
19: Add ρ^* to Π
20: $r' \leftarrow \max_{\rho \in \mathcal{P}} r^T \rho$
21: $\mathcal{P} \leftarrow$ NEXTPOLICY($\mathcal{P}, \rho^*, \alpha^*$)
22: $\mathcal{P} \leftarrow \mathcal{P} \setminus \{\rho^*\}$
23: **while** $\mathcal{P} \neq \emptyset$ and $\max_{\rho \in \mathcal{P}} r^T \rho \geq r'$ ▷ CMDP concave over P
24: **return** $\operatorname*{argmax}_{\rho \in \Pi} r^T \rho / \varphi(\rho)$

Since a risk aggregator φ is strictly monotonic by definition, Assumption 4 implies

$$\varphi(\mathfrak{d}_1(\pi^{s,a_1}), \ldots, \mathfrak{d}_m(\pi^{s,a_1})) < \ldots < \varphi(\mathfrak{d}_1(\pi^{s,a_k}), \ldots, \mathfrak{d}_m(\pi^{s,a_k})). \qquad (28)$$

for every $\pi \in \Pi^{\mathrm{det}}$ and $s \in \mathcal{S}$. Assumption 4 often holds in financial market applications, where "larger" actions, such as investing a larger amount, are fundamentally riskier than "smaller" ones, irrespective of the risk aspect.

Proposition 9. *Consider m fixed. Given Assumption 4, covering over all $I \subset \mathcal{J}$, such that $|I| = m$ and $\mathrm{Supp}(\rho^*) \subset J \setminus \mathrm{Supp}(\rho^*)$ (Line 2 of Algorithm 1) will run worst case in $O(n^m k^m)$ iteration.*

Theorem 10. *Consider m fixed. Given Assumption 4, the number of arithmetic operations of Algorithm 1 in the worst case is bounded by $O(n^{2m+3} k^{m+1})$.*

6 Single-risk EMRDP

An approach for solving single-risk ($m = 1$) EMRDP with the risk aggregator (10) was introduced in Zadorojniy, Osogami, and Davidovich (2023). It employs an algorithm that looks for the risk-sensitive optimal policy through a discrete form of gradient ascent in the policy space that is the graph

Algorithm 2: (Zadorojniy, Osogami, and Davidovich 2023) Algorithm for single-risk EMRDP.

1: Initialize: $\pi \leftarrow \operatorname{argmin}_\pi \{\mathfrak{d}(\pi) \mid Q\,\rho^\pi = \mu, \rho^\pi \geq 0\}$
2: $T \leftarrow \{(\pi, \mathfrak{r}(\pi), \mathfrak{d}(\pi))\}$
3: **while** $\exists (s,a)$ s.t. $\mathfrak{d}(\pi^{s,a}) > \mathfrak{d}(\pi)$ **do**
4: $\quad \pi^{s,a} \leftarrow \operatorname{argmax}_{\pi^{s,a}} \{\nabla_{s,a}(\pi) \mid \mathfrak{d}(\pi^{s,a}) > \mathfrak{d}(\pi)\}$
5: $\quad$ Add $(\pi^{s,a}, \mathfrak{r}(\pi^{s,a}), \mathfrak{d}(\pi^{s,a}))$ into T
6: $\quad \pi \leftarrow \pi^{s,a}$
7: **end while**
8: **return** $\operatorname{argmax}_\pi \{\mathfrak{r}(\pi)/\mathfrak{d}^\omega(\pi) \mid (\pi, \mathfrak{r}(\pi), \mathfrak{d}(\pi)) \in T\}$

of deterministic and 1-randomized policies. We will demonstrate its correspondence to Algorithm 1 for the risk aggregator (10) with $m = 1$. One can, therefore, regard Algorithm 1 as a natural extension of Zadorojniy, Osogami, and Davidovich (2023) to multi-risk EMRDP.

An Alternative Viewpoint

Zadorojniy, Osogami, and Davidovich (2023) introduced Algorithm 2 for solving (9) in the single-risk case ($m = 1$) with the risk aggregator (10) under the same Assumptions 1-3. Here, we will recap some of their main results. To that end, for $\pi \in \Pi^{\rm det}$, let $\pi^{s,a}$ be the deterministic policy that follows π, except at the state $s \in \mathcal{S}$ where the specified action $a \in \mathcal{A}$ is taken instead of $\pi(s)$. Denote the difference in reward over the difference in risk by

$$\nabla_{s,a}(\pi) \equiv \frac{\mathfrak{r}(\pi^{s,a}) - \mathfrak{r}(\pi)}{\mathfrak{d}(\pi^{s,a}) - \mathfrak{d}(\pi)} = \frac{r^\top \rho^{\pi^{s,a}} - r^\top \rho^\pi}{d^\top \rho^{\pi^{s,a}} - d^\top \rho^\pi}. \quad (29)$$

Algorithm 2 of Zadorojniy, Osogami, and Davidovich (2023) initializes π (Line 1) with a feasible policy that minimizes $\mathfrak{d}(\pi)$, with Q and μ as in (5). At each while-loop iteration, the algorithm considers policies $\pi^{s,a}$ with $\mathfrak{d}(\pi^{s,a}) > \mathfrak{d}(\pi)$. In Line 4, $\pi^{s,a}$ is chosen to maximize $\nabla_{s,a}(\pi)$. In Line 6, the policy π is updated with $\pi^{s,a}$ for the next iteration of the while loop. Once done, the policy that corresponds to the maximal ratio is returned (Line 8).

Theorem 11 (Theorem 1, Zadorojniy, Osogami, and Davidovich (2023))**.** *Let $\omega \in (0, 1]$. Under Assumptions 1-3, Algorithm 2 returns the policy π^* that maximizes $\mathfrak{r}(\pi)/\mathfrak{d}^\omega(\pi)$ over all stationary policies π.*

Let $G = G(V, E)$ be the policy graph, where V is the set of deterministic policies and E is the set of pairs of neighbouring deterministic policies (i.e., policies that disagree in exactly one state). Let Γ_α be the set of deterministic and strictly 1-randomized policies (see §2) for CMDP(α). Let $\Gamma^*_\alpha \subseteq \Gamma_\alpha$ be the set of optimal policies in Γ_α, and let $\Gamma^* \equiv \cup_\alpha \Gamma^*_\alpha$. The set Γ^* forms a path in the policy graph G (Zadorojniy, Even, and Shwartz 2009, Lemma 5.4). Let π^0 and π^1 be any two neighbouring deterministic policies in Γ^*. Denote a strictly 1-randomized policy on the edge between π^0 and π^1 by π^q, that is, $\pi^q = (1-q)\,\pi^0 + q\,\pi^1 \in \Pi^{1-\rm rnd}$, where $q \in (0, 1)$. Let $\Gamma^{\rm det}$ be the set of deterministic policies returned by Algorithm 2 and $\Gamma^{1-\rm rnd}$ be the set of strictly 1-randomized policies on the edges between the neighboring policies in $\Gamma^{\rm det}$. The following proposition identifies the policy path that Algorithm 2 traverses.

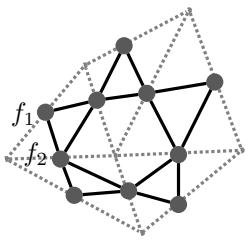

Figure 2: A 2-dimensional section (the dotted lines) of a 3-dimensional (conic) triangulation $\mathcal{T}_r$. The triangulation graph G_r is marked (in bold) with two adjacent vertices $f_1, f_2 \in V(G_r)$.

Proposition 12 (Corollary 1, Zadorojniy, Osogami, and Davidovich (2023))**.** $\Gamma^{\rm det} \cup \Gamma^{1-\rm rnd} = \Gamma^*$.

Zadorojniy, Osogami, and Davidovich (2023) introduced a variant of Algorithm 2 that runs in strongly-polynomial time when Assumption 4 holds. It is identical to Algorithm 2 except in Line 1, where π is initialized with the risk-minimizing policy that takes the minimal action a_1, well-defined by Assumption 4, at every state.

Proposition 13 (Lemma 3, Zadorojniy, Osogami, and Davidovich (2023))**.** *Under Assumption 4, the strongly-polynomial variant of Algorithm 2 runs in $O(n^4\,k^2)$ time.*

Correspondence of Viewpoints

Our goal is to establish a correspondence between Algorithm 1 and Algorithm 2 of Zadorojniy, Osogami, and Davidovich (2023) with $m = 1$ and $\varphi \equiv \varphi_\omega$ (10).

Setup. Recall our notation for the constituents of the parametric LP in Proposition 2. For the single-risk case:

$$A_{\beta,d} = \begin{pmatrix} Q(\beta) \\ d^T \end{pmatrix}, \quad b_\beta = \begin{pmatrix} \mu(\beta) \\ \alpha \end{pmatrix}, \quad (30)$$

Following Proposition 2, CMDP(α) is equivalent to the parametric LP $LP_{\beta,r,d}(b_\beta) := LP_{A_{\beta,d},r}(b_\beta)$ and $\text{rk}(A_{\beta,d}) = n + 1$ (20). The domain for CMDP is determined by the intersection of the line

$$L_\beta = \left\{ \begin{pmatrix} \mu(\beta) \\ \alpha \end{pmatrix} \,\middle|\, \alpha \in \mathbb{R} \right\} \subseteq \mathbb{R}^{n+1}. \quad (31)$$

with Cone($A_{\beta,d}$). We may assume with probability 1 that L_β is transverse to Cone($A_{\beta,d}$), namely, that L_β intersects Cone($A_{\beta,d}$) at interior points of co-dimension 1 simplices. By Lemma 3, the condition in part (ii) of Theorem 1 holds, meaning that $LP_{\beta,r,d}(b)$ is bounded for all $b \in$ Cone($A_{\beta,d}$), and, in particular, for all $b \in L_\beta \cap$ Cone($A_{\beta,d}$).

Triangulation Graph. Following Corollary 3, let $\mathcal{T}_r$ be the unique triangulation of Cone($A_{\beta,d}$) corresponding to the immediate reward vector $r \in \mathbb{R}^{nk}$. $\mathcal{T}_r^{\rm top}$ denotes the set of top-dimensional simplices, i.e., the $(n+1)$-simplices, of $\mathcal{T}_r$. The co-dimension 1 simplices $\mathcal{T}_r^{\rm codim-1}$ of $\mathcal{T}_r$ are its n-simplices. The boundary $\partial\tau$ of a simplex $\tau =$ Cone(A_J) in $\mathcal{T}_r$ is the union of all sub-simplices $f =$ Cone($A_{J'}$) such that $J' \subset J$ and $|J \setminus J'| = 1$.

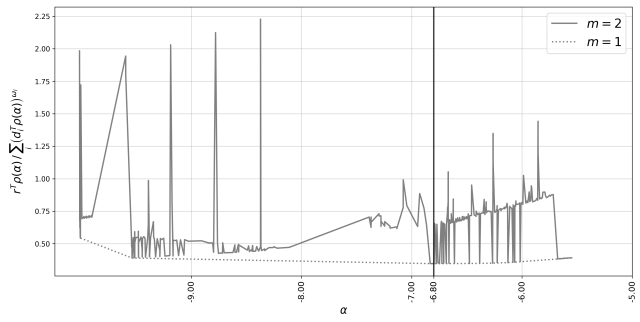

Figure 3: The deterministic Pareto fronts of equivalent single ($m = 1$) vs. double ($m = 2$) risk EMRDPs.

Define the (undirected) graph G_r of the triangulation $\mathcal{T}_r$ as follows.

$$V(G_r) := \mathcal{T}_r^{\mathrm{codim}-1}$$
$$E(G_r) := \left\{ e = \{f_1, f_2\} \;\middle|\; \exists \tau \in \mathcal{T}_r^{\mathrm{top}} : f_1, f_2 \subset \partial\tau \right\}$$

Figure 2 illustrates a 2-dimensional section (the dotted lines) of a 3-dimensional conic triangulation $\mathcal{T}_r$. Thus, $n + 1 = 3$ and the single-risk EMRDP has two states. The triangulation graph G_r in Figure 2 (in bold) is marked with two adjacent $f_1, f_2 \in V(G_r)$. By definition, the valence of a graph node $f \in \mathcal{T}_r^{\mathrm{codim}-1}$ is either $2n$ or n, depending whether f is internal or not.

Proposition 14. *L_β determines a simple path Γ_r in G_r.*

By a standard procedure, we can metrize G_r, setting the length of each edge to be 1. We will denote this metrized graph by G_r^{met} to distinguish it as a metric space. This induces a metric on Γ_r as well. Recall that in §6 we let G denote the policy graph whose vertices were Π^{det} and whose edges E were pairs of deterministic policies that differ at a single state. Introducing 1-randomized policy on the edges between vertices of G effectively metrized it and its subgraph $\Gamma^* \subseteq G^{\mathrm{met}}$.

Theorem 15. *$\Gamma_r^{\mathrm{met}} \cong \Gamma^*$ as metric graphs.*

As a consequence of Theorem 15, Algorithm 2 of Zadorojniy, Osogami, and Davidovich (2023) and the single-risk reduction of Algorithm 1 traverse the same optimal policy path, albeit within different graphs.

Single-risk Reduction. The single-risk reduction of Algorithm 1 is designed to traverse Γ_r (and equivalently Γ^*). As in (26), it is initialized with

$$\rho_{\min} = \operatorname*{argmin}_{\rho} \left\{ d^T \rho \mid Q(\beta)\rho = \mu(\beta), \rho \geq 0 \right\} \qquad (32)$$

where we utilize the fact that φ is strictly monotonic (see also Line 1 of Algorithm 2). Following Theorem 1, Strong Duality allows us to combine (14) together with (15) to establish the Pareto front. For deterministic policies ρ^*, this translates, at each iteration, into finding the optimal basis $J \subseteq [nk]$ that satisfies the conditions: (i) $\mathrm{Supp}(\rho^*) \subset J$, (ii) A_J is invertible, and (iii) $r_J^T A_J^{-1} A - r^T \geq 0$ (Lines 2-4). There can be at most *two* such bases corresponding to at most two neighboring top-dimensional simplices. This is a key feature of single-risk EMRDP: the bound on the number of possible J's is not triangulation specific, unlike in the multi-risk case.

Now, even with an oracle that provides us with a minimal risk deterministic policy (32), the single-risk reduction of Algorithm 1 still fails to be strongly polynomial. This would depend, for each $j \in [nk]$, on a polynomial bound on the number of iterations for which j will be present in J that satisfies the condition in Line 4. This touches on deep questions of upper bounds and statistics of triangulations and graph diameters (Santos 2012). Assumption 4 operates as an upper bound of order $O(1)$.

When Assumption 4 holds, Algorithm 2 is more efficient than the single-risk reduction of Algorithm 1. It runs in $O(n^4\, k^2)$ (Proposition 13), while the latter runs in $O(n^5\, k^2)$ (Theorem 10). This is due to its two-phase search, first for incident top-dimensional simplices (Lines 2-4) and then for their co-dimension 1 faces (Line 6), while Algorithm 2 looks for the best next deterministic policy, i.e., neighboring co-dimension 1 face, in one fell swoop. However, extending Algorithm 2 from single to multiple φ-aggregated risks by greedy gradient ascent will not work, even though Corollary 6 ensures us that the optimal policy for multiple-risk EMRDP is to be found at deterministic policies. We will illustrate the challenge of adding risk dimensions with the Grid World benchmark.

Grid World. Solving the benchmark Grid World problem requires finding the optimal policy for traversing an $h \times w$ grid from an initial to a terminal cell in a minimal number of moves in the presence of obstacles.

To the standard benchmark Grid World setup, introduced thus far, Zadorojniy, Osogami, and Davidovich (2023) add a notion of risk: it is defined as high for walking along the top or bottom boundaries of the grid and as medium for walking left into the left boundary or right into the right boundary. Risk is otherwise set to a baseline low value. Consequently, the goal is to *minimize* the ratio $\mathfrak{r}(\pi)/\mathfrak{d}(\pi)$, thereby increasing (negative) reward while decreasing (negative) risk. This extends Grid World to a single-risk EMRDP problem where we take $\varphi \equiv \varphi_\omega$ and $\omega = 1$.

One can easily turn this single-risk EMRDP into a (degenerate) double-risk EMRDP by splitting its immediate risk vector d into d_1 and d_2, so that d_1 is risk concentrated in the first action (Up), while d_2 is risk concentrated at the latter actions (Down, Left, Right, None).[3] Together, $d = d_1 + d_2$, and the single and double risk EMRDPs have the exact same optimal solution. Figure 3 illustrates the difference between its single ($m = 1$) and double ($m = 2$) risk incarnations. Both dotted and solid lines plot the EMRDP objective to the aggregated risk for each π_* along the Pareto. While the Pareto for the single-risk EMRDP (dotted) is convex, the one for double-risk (solid) is clearly not, indicating that a straightforward approach involving greedy gradient descent will likely fail to reach the EMRDP optimum.

[3]Notebook available on https://github.com/IBM/IBM-Extended-Markov-Ratio-Decision-Process/blob/main/notebooks/grid_world_multi.ipynb.

References

Aggarwal, V.; Chandrasekaran, R.; and Nair, K. P. 1977. Markov ratio decision processes. *Journal of Optimization Theory and Applications*, 21(1): 27–37.

Altman, E. 1999. *Constrained Markov Decision Processes*, volume 7. CRC Press.

Borkar, V.; and Jain, R. 2014. Risk-constrained Markov decision processes. *IEEE Transactions on Automatic Control*, 59(9): 2574–2579.

Chow, Y.; Tamar, A.; Mannor, S.; and Pavone, M. 2015. Risk-Sensitive and Robust Decision-Making: a CVaR Optimization Approach. In Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

de Loera, J.; Rambau, J.; and Santos, F. 2010. *Triangulations: Structures for Algorithms and Applications*, volume 25 of *Algorithms and Computation in Mathematics*. Springer-Verlag Berlin Heidelberg.

Derman, C. 1962. On sequential decisions and Markov chains. *Management Science*, 16–24.

Horn, R. A.; and Johnson, C. R. 2013. *Matrix analysis*. Cambridge University Press, 2. edition.

Meggendorfer, T. 2022. Risk-Aware Stochastic Shortest Path. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9858–9867.

Megiddo, N.; and Chandrasekaran, R. 1989. On the ε-perturbation method for avoiding degeneracy. *Operations Research Letters*, 8(6): 305–308.

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

Rockafellar, R. T.; Uryasev, S.; et al. 2000. Optimization of conditional value-at-risk. *Journal of risk*, 2: 21–42.

Santos, F. 2012. A Counterexample to the Hirsch Conjecture. *Annals of Mathematics*, 176: 383–412.

Sturmfels, B.; and Thomas, R. R. 1997. Variation of cost functions in integer programming. *Mathematical Programming*, 77(2): 357–387.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: The MIT Press. ISBN 0262039249.

Walkup, D. W.; and Wets, R. J.-B. 1969. Lifting projections of convex polyhedra. *Pacific Journal of Mathematics*, 28(2): 465–475.

Zadorojniy, A.; Even, G.; and Shwartz, A. 2009. A Strongly Polynomial Algorithm for Controlled Queues. *Mathematics of Operations Research*, 34(4): 992–1007.

Zadorojniy, A.; Osogami, T.; and Davidovich, O. 2023. A Rigorous Risk-aware Linear Approach to Extended Markov Ratio Decision Processes with Embedded Learning. In Elkind, E., ed., *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, 5475–5483. International Joint Conferences on Artificial Intelligence Organization. Main Track.