

Robustness of Stable Matchings When Attributes and Saliency Determine Preferences

Amit Ronen
Bar-Ilan University
Ramat Gan, Israel
amit.ronen@biu.ac.il

S. S. Ravi
University of Virginia
Charlottesville, USA
ssravi0@gmail.com

Sarit Kraus
Bar-Ilan University
Ramat Gan, Israel
sarit@cs.biu.ac.il

ABSTRACT

In many matching markets—such as athlete recruitment or academic admissions—participants on one side are evaluated by attribute vectors known to the other side, which in turn applies individual *saliency vectors* to assign relative importance to these attributes. Since saliencies are known to change in practice, a central question arises: how robust is a stable matching to such perturbations?

First, we formalize robustness as a radius within which a stable matching remains immune to blocking pairs under any admissible perturbation of saliency vectors (which are assumed to be normalized). Given a stable matching and a radius, we present a polynomial-time algorithm to verify whether the matching is stable within the specified radius. We also give a polynomial-time algorithm for computing the maximum robustness radius of a given stable matching. Further, we design an anytime search algorithm that uses certified lower and upper bounds to approximate the most robust stable matching, and we characterize the robustness-cost relationship through efficiently computable bounds that delineate the achievable tradeoff between robustness and cost. Finally, for each stable matching, we show that the saliency profiles preserving its stability form a product of low-dimensional polytopes within the simplex. This characterization enables efficient volume computation, with approximations available in higher dimensions, linking robustness analysis in matching markets to convex geometry.

KEYWORDS

Stable matching; Attributes; Saliency; Robustness; Mathematical programs; Algorithms; Convex optimization; Rotation Poset

ACM Reference Format:

Amit Ronen, S. S. Ravi, and Sarit Kraus. 2026. Robustness of Stable Matchings When Attributes and Saliency Determine Preferences. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026)*, Paphos, Cyprus, May 25 – 29, 2026, IFAAMAS, 9 pages. <https://doi.org/10.65109/IQTL2250>

1 INTRODUCTION

1.1 Background and Motivation

Stable matching is a foundational model in market design [27]. It underpins real-world allocation systems such as college admissions [15], medical residency programs [28], school choice [2], roommate allocation [5], student-project-resource-allocation [18],

and kidney exchange [21, 29]. In the classic bipartite matching problem, two sets of agents must be paired, with each agent having preferences over potential partners on the other side. A matching is considered *stable* if no pair of agents would both prefer each other to their assigned partners. Traditionally, the model assumes that agents’ preferences are fixed, complete, and independent [17].

However, evidence from behavioral science challenges these assumptions [34]. Human preferences are not arbitrary; they are often dynamic, evolving with context, experience, and time. This raises an important question: if preferences change, can a previously stable matching dissolve (i.e., become unstable)?

Prior research has shown that stable matchings, when preferences are independent of one another, are surprisingly fragile to changes in preferences. Even when preferences are altered in very small ways, stable matchings that remain stable despite such perturbations - that is, matchings that are robust to preference changes - turn out to be exceedingly rare [7, 13]. We present some results in an extended version of this paper [26] to reinforce this point.

However, as noted above, agents’ preferences are not arbitrary. In this paper, we focus on settings where agents on one side evaluate potential counterparts through saliency-weighted computations, where each agent assigns *saliency weights* to the observable attributes of potential counterparts [31], reflecting their perceived importance. Crucially, the saliency of attributes, and thus the agents’ preferences, may shift over time [8, 9].

For instance, in college admissions, applicants are characterized by attributes such as Scholastic Aptitude Test (SAT) scores and GPA, which are known to colleges. Each college assigns varying levels of importance to these attributes and ranks applicants accordingly. This motivates a shift in perspective. Rather than thinking about robustness in terms of small swaps in a ranked list, we propose viewing it through the lens of attribute saliency. Robustness then becomes a geometric question: how far can these weights shift, within some tolerance, while still preserving stability? In this sense, robustness can be viewed as a *second-order stability* property - a form of “stability of stability” that captures whether stable outcomes persist under changes in preferences.

Model overview. We study balanced two-sided markets where side A has static strict preferences and side B induces strict lists by *saliency-weighted scores*. Each agent $a \in A$ has a vector $\mathbf{u}(a)$ with m non-negative attribute values, for a constant number m of attributes. Let Λ_{m-1} denote the $(m-1)$ -dimensional probability simplex consisting of all the normalized vectors with m non-negative components. Each agent $b \in B$ has a saliency vector $\mathbf{s}(b) \in \Lambda_{m-1}$, and b ranks each $a \in A$ by the scalar product $\mathbf{s}(b) \cdot \mathbf{u}(a)$ along with a given tie-breaking rule.

The model is *hybrid*: continuous in saliency space yet ordinal in realized lists, letting us bring geometric tools to classical structures



This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). <https://doi.org/10.65109/IQTL2250>

of stable matchings (e.g., deferred acceptance, the lattice of stable matchings, and rotations [15, 17]). The set of all stable matchings forms a distributive lattice ordered by the preferences of one side, ranging from the B -optimal to the A -optimal outcome. Within this lattice, *rotations* represent minimal cyclic exchanges of partners that move monotonically from one stable matching to another, and serve as the fundamental units in our robustness analysis. Robustness in this framework is quantified geometrically by a radius r in an ℓ_p norm, capturing the maximum perturbation under which stability persists. In practice, decision makers rarely alter all *salience weights* simultaneously, motivating a *support budget* k that limits how many coordinates of $s(b)$ may change at once.

1.2 Summary of Contributions

While previous work on the robustness of stable matchings (e.g., [13]) considered swapping entries in preference lists, our work considers robustness when preference lists are modified by perturbing salience vectors. We study some fundamental questions under this notion of robustness. Our primary contributions are listed below.

1. A new notion of robustness using salience. When a salience vector $s(b)$ of an agent $b \in B$ is perturbed to $\hat{s}(b)$, the distance r between $s(b)$ and $\hat{s}(b)$ under some ℓ_p metric is the *radius* of the perturbation. Such perturbations could change the preference list of b and may cause a matching to become unstable. We define a stable matching μ to be (k, r, p) -robust if it remains stable when at most $k \leq m$ components of the salience vector of an agent are modified and the radius of this perturbation under the ℓ_p metric is at most r . We examine several algorithmic and geometric problems based on this notion of robustness.

2. Algorithm for robustness verification. Given a stable matching μ , a non-negative real number r , a non-negative integer $k \leq m$ and a value $p \in \{1, 2, \infty\}$, we present an algorithm to verify whether μ is (k, r, p) -robust. For fixed m , our algorithm runs in polynomial time as the verification problem reduces to solving a polynomial number of convex programs (linear program (LP) or second-order cone program (SOCP) feasibility checks), each solvable in polynomial time (see e.g., [4, 10]).

3. Algorithm for finding the maximum robustness radius. Extending the ideas used to solve the verification problem, we show that for any given a stable matching μ , the maximum radius $r^*(\mu)$ such that μ is $(k, r^*(\mu), p)$ -robust can be computed in polynomial time for fixed m and any $p \in \{1, 2, \infty\}$.

4. Finding the most robust stable matching. Given the feature vectors of agents in A and the salience vectors of agents in B , we show that lower and upper bounds on the maximum robustness radius r^* of a most robust stable matching can be computed efficiently. Using these bounds, we present an anytime search algorithm that finds a stable matching whose robustness radius is within the computed bounds. This search algorithm relies on several concepts associated with stable matchings (e.g., deferred acceptance, rotation poset; see [17, 23] or [26] for definitions of these concepts).

5. Robustness-cost tradeoffs. We also examine an extension of robustness that incorporates a cost function capturing welfare, fairness, or other priorities relevant to applications. We provide a polynomial-time algorithm for finding a stable matching whose

robustness radius exceeds a given threshold while minimizing the cost among all matchings meeting that robustness requirement.

6. Describing the robustness region of a given matching. We use the term *salience profile* to denote the matrix $S \in \mathbb{R}^{n \times m}$, where the b -th row is the salience vector $s(b)^\top$. Given a matching μ , the *robustness region* of μ consists of all salience profiles S under which μ remains stable. We show that this robustness region factors as a product of n low-dimensional polytopes within the simplex, fully characterizing its geometry. Reference [26] further derives how the volume of this region can be computed in polynomial time.

1.3 Related Work

Foundations and structure. We rely on the classical lattice and rotation frameworks of stable matchings [17, 23], which provide the structural basis for our analysis. Polyhedral characterizations, in particular Rothblum’s stable-marriage polytope [30], provide the geometric foundation for our optimization formulations of robustness and radius computation.

Robustness under ordinal perturbations. A major line of research explores robustness under ordinal perturbations. Chen et al. [13] formalize d -robustness via Kendall- τ distance (adjacent swaps) and develop polynomial algorithms to *find* and *optimize* d -robust matchings using the rotation partially ordered set (poset), with hardness emerging in the presence of ties. Mai and Vazirani [22] extend this perspective by considering robustness to *uncertainty in the preference lists themselves*, developing algorithms that preserve stability under bounded perturbations of linear orders. In sharp contrast, Boehmer et al. [7] show that on large random instances, a stable matching typically fails to survive even a single adjacent swap. Earlier work introduced the notion of *supermatches* [16], defining robustness in terms of repairability: a matching is robust if small disruptions (i.e., a limited number of pair breakups) can be fixed by a nearby stable matching. More recent rotation-based presolver algorithms [11] use the rotation poset to efficiently identify matchings with maximal robustness under such ordinal formulations. In contrast, our work introduces a continuous, attribute-based notion of robustness, where stability is preserved against perturbations of agents’ salience weights rather than discrete swaps in their ordinal preference lists.

Behavioral motivation. Beyond purely ordinal models, prior work links matching behavior to context-dependent *salience* of attributes in behavioral decision theory [6]. Similarly, many real-world matching systems implicitly rely on attribute-based evaluations. For example, in New York City’s centralized high-school match [1], programs assess applicants using transparent, multi-criteria rubrics (e.g., test performance, attendance, or neighborhood priority), and different schools emphasize these features to varying degrees – effectively corresponding to distinct salience vectors.

Preference evolution and dynamic settings. Another line of work studies how preference profiles evolve over time. Echenique et al. [14] provide experimental evidence on decentralized matching processes with endogenously changing choices. Other recent models, such as Bredereck et al. [12] and Alimudin and Ishida [3], explore algorithmic adaptation of stable matchings under dynamically changing preference profiles. Similar adaptive-weight dynamics also appear in multi-agent learning, where agents adjust

feature weights over fixed attribute spaces to adapt to changing environments or interaction patterns (e.g., [19, 24]). Our framework complements these approaches by capturing preference change as a *structured geometric perturbation* in salience space, bridging dynamic evolution and robust stability within a unified model.

Algorithmic and empirical perspectives beyond stability. Several works study matching beyond classical stability. These include facilitation mechanisms that advise agents to relax constraints [32], multi-round models focusing on general welfare [33], and online matching with delays prioritizing competitive guarantees over stability [35]. Empirical studies of reciprocal recommenders further highlight behavioral heterogeneity [20], while [25] introduce a formal model of loyalty in random matching markets, analyzing how endowment-driven acceptance thresholds interact with competition and stability.

Roadmap. We begin by defining the attribute-salience model and introducing our notation (Section 2). We then study the robustness of a given matching (Section 3) and show how to compute its maximum robustness radius (Section 4). Next, we develop algorithms for finding a stable matching that attains the maximal robustness radius (Section 5) and derive bounds on the relationship between robustness and cost (Section 6). Finally, we explore the geometric structure of robustness regions (Section 7).

2 MODEL AND NOTATION

We define the attribute-salience matching setup (agents, preferences, attributes of A , salience of B), robustness notions, and notational conventions used throughout this paper. A complete numerical example illustrating all the definitions is provided in [26].

Agents and matchings. Let $A = \{a_1, \dots, a_n\}$ and $B = \{b_1, \dots, b_n\}$, with $|A| = |B| = n$, be two disjoint sets of agents. A *matching* is a bijection $\mu : A \rightarrow B$; write $\mu^{-1}(b)$ for b 's partner in A .

Side A: static preferences. Each agent $a \in A$ has a permanent, strict, complete order over B that never changes. We use the notation $b \succ_a b'$ to indicate that a prefers b to b' .

Attributes of A. For each agent $a \in A$, there is a vector of attributes $\mathbf{u}(a)$ that characterizes a 's properties. These attributes are observable to the agents in B . The vector $\mathbf{u}(a)$ consists of $m \geq 2$ observable attributes. Throughout this paper, we assume that m is a *fixed constant*. In our notation, $\mathbf{u}(a) = (u_1(a), \dots, u_m(a)) \in \mathbb{R}_{\geq 0}^m$.

Side B: salience vectors. Each agent $b \in B$ evaluates candidates by assigning weights to the observable attributes of A . Formally, we represent these priorities by a salience vector

$$\mathbf{s}(b) = (s_1(b), \dots, s_m(b)) \in \Lambda_{m-1} := \{\mathbf{s}(b) \in \mathbb{R}_{\geq 0}^m : \sum_i s_i(b) = 1\}.$$

We denote the *salience profile* of side B by the $n \times m$ matrix $S = (\mathbf{s}(b))_{b \in B} \in (\Lambda_{m-1})^n$, which collects the salience vectors of all agents in B . Agent $b \in B$ assigns to each $a \in A$ the score $\mathbf{s}(b) \cdot \mathbf{u}(a) = \sum_{i=1}^m s_i(b) u_i(a)$. Agents in B rank candidates by decreasing score, breaking ties using a fixed, public, agent-specific strict order \succ_b^{tie} . Thus, each $b \in B$ induces a strict total order over A . We write $a \succ_b a'$ when b prefers a over a' .

Ordinal ranks of A and B. For each agent $a \in A$, let $\text{rank}_a(b) \in \{1, \dots, n\}$ denote the position of b in a 's strict preference list (with a lower value indicating a more preferred agent). For each agent

$b \in B$, let $\text{rank}_b(a) \in \{1, \dots, n\}$ denote the position of a in b 's strict list induced by the salience rule (sorting A by decreasing $\mathbf{s}(b) \cdot \mathbf{u}(\cdot)$ and breaking ties by \succ_b^{tie}). Thus $b \succ_a b'$ iff $\text{rank}_a(b) < \text{rank}_a(b')$, and $a \succ_b a'$ iff $\text{rank}_b(a) < \text{rank}_b(a')$.

Blocking pairs and stability. A pair $(a, b) \in A \times B$ is a *blocking pair* for a matching μ if $b \succ_a \mu(a)$ and $a \succ_b \mu^{-1}(b)$. A matching μ is *stable* when no blocking pair exists.

Radius parameters and worst-case perturbations. Fix a support budget $k \in \{1, \dots, m\}$, a radius $r \geq 0$, and a norm $p \in \{1, 2, \infty\}$. We allow a single agent $b \in B$ to *modify* up to k components of its salience vector. Formally, b selects additive perturbations $\delta \in \mathbb{R}^m$ with support $Q \subseteq [m]$, $|Q| \leq k$, satisfying $s_i(b) + \delta_i \geq 0$ for all $i \in Q$, and defines $T := \sum_{i=1}^m (s_i(b) + \delta_i) > 0$. The perturbed vector is then normalized as $\hat{\mathbf{s}}(b) = \frac{\mathbf{s}(b) + \delta}{T} \in \Lambda_{m-1}$. The perturbation radius, computed post-normalization, is $\|\hat{\mathbf{s}}(b) - \mathbf{s}(b)\|_p$. We use \hat{S} to denote the modified profile.

A matching μ is said to be (k, r, p) -robust if it remains stable under all perturbations of the above form satisfying this bound for every agent $b \in B$. Note that μ is $(k, 0, p)$ -robust iff μ is stable at the salience profile S .

Pre- vs. post-normalization. An equivalent description replaces T by its reciprocal $\lambda = 1/T > 0$. In this *post-normalized* view, perturbations are specified by a vector $\hat{\mathbf{s}}(b) \in \Lambda_{m-1}$ and a scalar $\lambda > 0$ such that

$$\hat{s}_i(b) = \lambda s_i(b) \quad (i \notin Q), \quad \|\hat{\mathbf{s}}(b) - \mathbf{s}(b)\|_p \leq r.$$

Thus pre- and post-normalization are simply two parameterizations of the same admissible set, and they yield the same robustness radius. Throughout the paper, we adopt the post-normalized form for clarity and consistency.

Score margin. To quantify how strongly b prefers its current partner over another candidate a , we define the *attribute-gap vector* $\Delta(b; a | \mu) := \mathbf{u}(\mu^{-1}(b)) - \mathbf{u}(a)$.

Given a salience profile S , the corresponding *score margin* is $\gamma_S(b; a | \mu) := \mathbf{s}(b) \cdot \Delta(b; a | \mu)$.

The value $\gamma_S(b; a | \mu)$ is positive when b prefers its current partner $\mu^{-1}(b)$ to a under profile S . When μ and/or S are clear from the context, we write simply $\gamma(b; a)$.

Base radius. The *base inner radius* $r^{\text{base}}(\mu)$ is a baseline robustness guarantee: it denotes a perturbation level that is sufficient to keep μ stable. By construction $r^{\text{base}}(\mu) \leq r^*(\mu)$, and it depends only on score margins and dual-norm attribute gaps, independent of the support budget k (see Section 5.1).

Stability region. For a given matching μ , the *stability region* \mathcal{P}_μ is the set of all salience profiles under which μ is stable. Formally, $\mathcal{P}_\mu \subseteq (\Lambda_{m-1})^n$; see Section 7 for the explicit polyhedral form and product structure of \mathcal{P}_μ .

Remark ($k = m - 1$ vs. $k = m$). Since unchanged coordinates automatically adjust to preserve normalization, varying $m - 1$ coordinates already spans all feasible perturbations. A higher support size $k = m$ matters only in boundary cases where the support itself changes (e.g., $m = 2$, $(1, 0) \rightarrow (0, 1)$).

3 VERIFICATION FOR A GIVEN MATCHING

Given a salience profile $S = \{s(b)\}_{b \in B}$ and a matching μ , the verification problem asks whether μ remains stable under *every* admissible perturbation of the salience vector of a single agent $b \in B$ under the worst-case perturbation model. The main result of this section is that the verification problem can be solved in polynomial time.

For $b \in B$, let $\mathcal{H}_\mu(b) := \{a \in A : b \succ_a \mu(a)\}$ be the set of A -agents who prefer b to their current partner. We can now formalize the verification task more precisely by defining the robustness verification problem, which captures the requirement that no new blocking pair can emerge under bounded perturbations of the salience profile.

Definition 3.1 (Robustness Verification (RV)). Given a stable matching μ , attributes $\{u(a)\}_{a \in A}$, salience profile S , norm $p \in \{1, 2, \infty\}$, support budget $k \leq m$, and radius $r \geq 0$, decide if, for all $b \in B$, $a \in \mathcal{H}_\mu(b)$, and post-normalized perturbations $\hat{s}(b) \in \Lambda_{m-1}$, the score margin satisfies the condition $\gamma_S(b; a) \geq 0$.

Robustness verification ensures that no blocking pair can arise under any admissible perturbation of the salience vectors. The next lemma expresses this condition as a local check: for each b and each possible perturbation direction, stability holds iff all resulting score margins remain nonnegative.

Lemma 3.1. For any integers m and k with $k \leq m$, a value $p \in \{1, 2, \infty\}$, and a rational value $r \geq 0$, a stable matching μ is (k, r, p) -robust if and only if, for every $b \in B$, $a \in \mathcal{H}_\mu(b)$, and support set $Q \subseteq [m]$ with $|Q| \leq k$, no admissible perturbation $\hat{s}(b)$ yields a negative score margin, i.e., $\gamma_S(b; a) < 0$.

PROOF. (Only if.) Assume μ is (k, r, p) -robust. Assume otherwise that there exist $b \in B$, $a \in \mathcal{H}_\mu(b)$, and an admissible perturbation $\hat{s}(b)$ with $\|\hat{s}(b) - s(b)\|_p \leq r$ such that $\gamma_S(b; a) < 0$. Then (a, b) becomes a blocking pair after the perturbation, contradicting robustness. Hence no such perturbation exists.

(If.) Conversely, assume that for all $b \in B$, $a \in \mathcal{H}_\mu(b)$, and supports Q with $|Q| \leq k$, every admissible perturbation $\hat{s}(b)$ satisfies $\gamma_S(b; a) \geq 0$. Thus each b still weakly prefers its partner over any $a \in \mathcal{H}_\mu(b)$, and strict tie-breaking ensures no new blocking pair can arise. Therefore μ is (k, r, p) -robust. \square

Having reduced the condition to finitely many checks, we now establish that the overall verification task admits a polynomial-time algorithm.

THEOREM 3.2 (POLYNOMIAL-TIME VERIFICATION VIA SUPPORT ENUMERATION). Let m and $k \leq m$ be fixed constants, $p \in \{1, 2, \infty\}$, and $r \geq 0$. Verifying whether a given stable matching μ is (k, r, p) -robust can be done in polynomial time.

PROOF. The key idea is to translate each potential blocking deviation into a convex feasibility problem that tests whether a perturbation within distance r can violate stability. In particular, using Lemma 3.1, we reduce RV to testing the infeasibility of $O(n^2 m^k)$ convex optimization instances: a linear program (LP) when $p \in \{1, \infty\}$, or a second-order cone program (SOCP) when $p = 2$. Since m and k are constants and the infeasibility of convex instances can be determined in polynomial time [4, 10], We obtain a polynomial-time algorithm for RV.

Explicit LP/SOCP formulations. For each $b \in B$, $a \in \mathcal{H}_\mu(b)$, and support $Q \subseteq [m]$ with $|Q| \leq k$, we test whether there exists an admissible perturbation $\hat{s}(b) \in \Lambda_{m-1}$ satisfying $\|\hat{s}(b) - s(b)\|_p \leq r$ that makes (a, b) a blocking pair, i.e., violates stability. Each feasibility check is a convex program - an LP when $p \in \{1, \infty\}$ or an SOCP when $p = 2$. For illustration, the LP for $p = \infty$ is shown below; the formulations for $p = 1$ and $p = 2$ follow by modifying only the distance constraint, as described thereafter.

(a) $p = \infty$ (box distance) - LP feasibility for (a, b, Q)

$$\begin{aligned} & \text{find } \hat{s}(b), \lambda > 0 \\ & \text{s.t. } \sum_i \hat{s}_i(b) = 1, \quad \hat{s}_i(b) \geq 0, \\ & \hat{s}_i(b) = \lambda s_i(b) \quad (i \notin Q), \\ & \hat{s}(b) \cdot \Delta(b; a \mid \mu) \leq 0, \\ & -r \leq \hat{s}_i(b) - s_i(b) \leq r \quad (\forall i). \end{aligned}$$

Variables: $\hat{s}(b) \in \mathbb{R}_{\geq 0}^m$, $\lambda \in \mathbb{R}_{> 0}$.

The formulations for $p = 1$ and $p = 2$ differ only in the way the distance constraint $\|\hat{s}(b) - s(b)\|_p \leq r$ is enforced. For $p = 1$, this constraint is linearized by introducing additional variables $z_i \geq 0$ with $z_i \geq \hat{s}_i(b) - s_i(b)$ and $z_i \geq s_i(b) - \hat{s}_i(b)$ for all i , together with $\sum_i z_i \leq r$, yielding an LP. For $p = 2$, the distance constraint is replaced by the second-order cone constraint $\|\hat{s}(b) - s(b)\|_2 \leq r$, resulting in an SOCP. All other constraints remain unchanged.

Collectively, these feasibility programs cover all potential deviations across agents and supports. Each instance tests whether a perturbation of magnitude at most r exists that makes (a, b) a blocking pair. If all are infeasible, μ is (k, r, p) -robust; otherwise, a feasible instance yields a blocking witness. Since LP and SOCP infeasibility can be decided in polynomial time by interior-point methods [4, 10], and since m and k are constants, the overall runtime is polynomial in n . \square

When $k = m$ (full-support moves), support enumeration vanishes and the procedure reduces to $O(n^2)$ independent LP/SOCP checks indexed by (a, b) .

Remark 3.1 (Strict vs. non-strict inequalities). We use non-strict inequalities (\leq) in the LP/SOCP formulations for solver compatibility. Strict tie-breaking ensures that this is equivalent to strict stability, since every unperturbed score margin $\gamma(b; a \mid \mu)$ is positive.

4 MAXIMUM RADIUS OF A MATCHING

Having established how to verify robustness for a given radius r , we now turn to the complementary task: determining the largest radius for which stability still holds. In other words, for a given stable matching μ , support size k , and norm p , we ask: *what is the maximum perturbation radius under which μ remains robust?* This value quantifies the exact tolerance of μ to worst-case changes and serves as a natural definition of robustness. Intuitively, the robustness radius is determined by the weakest link: the smallest perturbation, over all potentially blocking pairs (a, b) and admissible supports Q , that makes b prefer a over its current partner $\mu^{-1}(b)$.

Definition 4.1 (Pairwise thresholds and robustness radius). For each (a, b, Q) , let $r^{\min}(b; a \mid Q)$ denote the optimal value of the

corresponding LP (for $p \in \{1, \infty\}$) or SOCP (for $p = 2$). Aggregating over all admissible supports, define $r^{\min}(b; a)$ by

$$r^{\min}(b; a) := \min_{\substack{Q \subseteq [m] \\ |Q| \leq k}} r^{\min}(b; a \mid Q).$$

The *maximum robustness radius* of μ is then

$$r^*(\mu) := \sup\{r \geq 0 : \mu \text{ is } (k, r, p)\text{-robust}\} = \min_{\substack{b \in B \\ a \in \mathcal{H}_\mu(b)}} r^{\min}(b; a).$$

The following lemma provides the basis for our polynomial-time algorithm for computing the maximum radius.

Lemma 4.1. *For each fixed m and k with $k \leq m$, any norm $p \in \{1, 2, \infty\}$, and a stable matching μ , each term $r^{\min}(b; a)$ from Definition 4.1 can be computed in polynomial time.*

PROOF. For a given a, b , and support $Q \subseteq [m]$, $r^{\min}(b; a \mid Q)$ is obtained by minimizing the radius r over admissible perturbations of $s(b)$ supported on Q that make (a, b) a blocking pair. Formally, each $r^{\min}(b; a \mid Q)$ is the optimal value of a convex optimization instance, that is: a linear program (LP) when $p \in \{1, \infty\}$, or a second-order cone program (SOCP) when $p = 2$.

The explicit LP/SOCP formulations for all cases are provided in [26]. Their construction is identical to the verification instances of Section 3, except for the following modifications:

- (1) the radius r is now a decision variable,
- (2) the objective is to minimize r , and
- (3) the additional constraint $r \geq 0$ is included.

Since the feasibility of each convex optimization instance is decidable in polynomial time and there are $O(m^k)$ supports per pair (a, b) for given $k < m$, each $r^{\min}(b; a)$ can be computed in polynomial time. When $k = m$, support enumeration collapses to a single program per (a, b) , so the complexity reduces to the time needed to solve one convex program. \square

Building on Lemma 4.1, we can now establish the main result of this section.

THEOREM 4.2 (POLYNOMIAL-TIME COMPUTATION VIA PAIRWISE THRESHOLDS). *Let m and $k \leq m$ be fixed constants, and let $p \in \{1, 2, \infty\}$. For any stable matching μ , the robustness radius $r^*(\mu)$ is computable in polynomial time.*

PROOF. For any stable matching μ , the robustness radius $r^*(\mu)$ is given by Definition 4.1. As shown above, for each pair (a, b) , the value $r^{\min}(b; a)$ can be computed by solving $O(m^k)$ convex optimization instances. Since there are $O(n^2)$ pairs $r^*(\mu)$ can be found by solving $O(n^2 m^k)$ convex instances. Since m and k are fixed, and each convex instance can be solved in polynomial time, the theorem follows. As explained earlier, the number of convex instances reduces to $O(n^2)$ when $k = m$. \square

5 FINDING THE MOST ROBUST MATCHING: BOUNDS AND ANYTIME SEARCH

In the previous sections, we assumed that a matching had already been selected and examined its robustness properties. However, it is often more advantageous to identify in advance the most robust stable matching. Therefore, our next goal is to *select the stable matching with the largest robustness radius*, namely $\mu^* \in$

$\arg \max_{\mu \in \mathcal{SM}} r^*(\mu)$, where \mathcal{SM} denotes the set of all stable matchings of the market and $r^*(\mu)$ is the exact robustness radius from Section 4. A challenge arises because the size of \mathcal{SM} can be exponential in n . A classical way to explore this space is via the *rotation poset* [17, 23], which compactly represents exponentially many stable matchings and supports polynomial-time traversal and optimization under structural criteria (see Section 5.1 for details). Unfortunately, $r^*(\mu)$ itself does not align with the rotation-poset structure: perturbations that create blocking pairs need not correspond to a single rotation or any set of rotations. To address this misalignment, we introduce the *base inner radius* $r^{\text{base}}(\mu)$, a conservative proxy satisfying $r^{\text{base}}(\mu) < r^*(\mu)$. This proxy can be expressed in closed form from local score margins and dual gaps, and its constraints align with the rotation poset, enabling efficient search for highly robust matchings. While $r^*(\mu)$ can be computed by solving $O(n^2 m^k)$ convex optimization instances, r^{base} can be found in $O(n^2 m)$ time.

In this section, Section 5.1 recalls the classical rotation-poset machinery and defines r^{base} , together with its structural and computational properties. In Sections 5.2–5.4, we tackle the main problem of finding the most robust stable matching, combining lower/upper bounds with an anytime search.

5.1 Rotations and the base inner radius

Classical background. We recall standard notions from stable matching; full details appear in [26] (see also [17, 23]). B -proposing DA (deferred acceptance) returns a B -optimal matching μ_B . The set of stable matchings forms a distributive lattice with $\mu_A \preceq \mu \preceq \mu_B$. A *rotation ρ exposed at μ* is a cyclic sequence of pairs $((a_1, b_1), \dots, (a_\nu, b_\nu))$ whose elimination yields another stable matching $\mu' = \text{elim}(\mu, \rho)$. Along a rotation, the B -side weakly improves and the A -side weakly worsens. Rotations admit a partial order; every stable μ can be written as $\mu = \text{elim}(\mu_A, D)$ for a unique down-set D ; there are $O(n^2)$ rotations in total, and the rotation poset can be built in $O(n^2)$ time.

Deterministic inner radius from score margins. Set $p \in \{1, 2, \infty\}$ and let p^* denote its dual norm, defined by $1/p + 1/p^* = 1$. Recall the concept of score margin $\gamma(b; a \mid \mu)$ from Section 2, which captures how strongly b prefers its partner over a . Together with the partner-dependent dual gap $U_{p^*}(b) := \max_{a' \neq \mu^{-1}(b)} \|\mathbf{u}(\mu^{-1}(b)) - \mathbf{u}(a')\|_{p^*}$, we obtain the *base inner radius*:

$$r^{\text{base}}(\mu) := (1 - \varepsilon_{\text{base}}) \min_{b \in B} \min_{a: \mu^{-1}(b) \succ_b a} \frac{\gamma(b; a)}{U_{p^*}(b)}. \quad (1)$$

Here $\varepsilon_{\text{base}} \in (0, 1)$ is a small fixed constant used only to ensure strict feasibility of inequalities. We also assume that the attribute vectors $\{\mathbf{u}(a) : a \in A\}$ are not all identical in their attributes, so that $U_{p^*}(b) > 0$ for every $b \in B$.

Lemma 5.1 (Margin \Rightarrow radius). *If $r \leq r^{\text{base}}(\mu)$, then μ remains stable under all ℓ_p perturbations of radius r .*

PROOF. For each b and $a \neq \mu^{-1}(b)$, write

$$\hat{s}(b) \cdot (\mathbf{u}(\mu^{-1}(b)) - \mathbf{u}(a)) = \gamma(b; a) + (\hat{s}(b) - \mathbf{s}(b)) \cdot \Delta(b; a \mid \mu).$$

By Hölder’s inequality [10, App. A.1.6], $|(\hat{s}(b) - \mathbf{s}(b)) \cdot \Delta(b; a \mid \mu)| \leq \|\hat{s}(b) - \mathbf{s}(b)\|_p \|\Delta(b; a \mid \mu)\|_{p^*} \leq r U_{p^*}(b)$. Hence the perturbation term is at least $-r U_{p^*}(b)$. Thus the residual margin is $\geq \gamma(b; a) -$

$rU_{p^*}(b)$. When $r \leq r^{\text{base}}(\mu)$, this remains positive, hence stability is preserved \square

As an immediate consequence, we obtain the following corollary.

Corollary 5.2 (Base inner radius). *If $r \leq r^{\text{base}}(\mu)$, then μ remains stable under all perturbations $\hat{s}(b)$ with $\|\hat{s}(b) - s(b)\|_p \leq r$ for every $b \in B$. Equivalently, $r^*(\mu) > r^{\text{base}}(\mu)$.*

5.2 A polynomial-time computable lower bound via B -optimal matching

To search for the most robust stable matching, it is essential to start from a guaranteed baseline. A natural choice is the B -optimal matching μ_B , the outcome of B -proposing deferred acceptance. By construction, each $b \in B$ receives its most preferred attainable partner among all stable matchings (a classical property of μ_B) [23]. This maximizes the score margins $\gamma(b; a)$ against less-preferred candidates, and hence yields the largest base inner radius $r^{\text{base}}(\mu)$ across the lattice of stable matchings.

Since μ_B is itself a stable matching, its robustness radius directly yields a lower bound:

$$\text{LB} := r^*(\mu_B) \leq \max_{\mu \in \text{SM}} r^*(\mu). \quad (2)$$

Proposition 5.1 (Polynomial-time LB). *The lower bound $\text{LB} = r^*(\mu_B)$ is computable in polynomial time by solving $O(n^2 \binom{m}{k})$ LP/SOCP instances from Section 4, which reduces to $O(n^2)$ when $k = m$.*

This gives a tractable starting point: although μ_B need not maximize robustness, its radius is efficiently computable and provides a guaranteed baseline against which all further improvements can be evaluated.

5.3 A relaxation-based global upper bound

To upper-bound $\max_{\mu} r^*(\mu)$, we relax integrality, working with the *stable-marriage polytope* $\text{conv}(\text{SM})$ [30]. This polytope is the convex hull of incidence matrices $X^\mu \in \{0, 1\}^{A \times B}$ that encode stable matchings μ , where $X_{ab}^\mu = 1$ if a is matched to b under μ and 0 otherwise.

Although the convex hull may have exponentially many vertices (since the number of stable matchings can be exponential in n), the stable-matching polytope admits a compact linear description of polynomial size [30]. Hence, all computations over this convex relaxation remain tractable. Thus, instead of a single integral matching, we allow fractional convex combinations of stable matchings, which yields a tractable convex relaxation.

For each $b \in B$, define the *fractional partner attribute vector*

$$\bar{\mathbf{u}}_b = \sum_{a \in A} X_{ab} u(a), \quad X \in \text{conv}(\text{SM}).$$

Intuitively, $\bar{\mathbf{u}}_b$ averages the attribute vectors of b 's partners across the mixture X . Robustness of radius $r \geq 0$ requires that even under post-normalized ℓ_p perturbations of radius r , each b prefers $\bar{\mathbf{u}}_b$ to every alternative $a \in A$.

Case $k = m$. When perturbations may reweight all m coordinates, Hölder's inequality yields the exact no-blocking constraint

$$s(b) \cdot (\bar{\mathbf{u}}_b - \mathbf{u}(a)) \geq r \left\| (\mathbf{u}(a) - \bar{\mathbf{u}}_b)_+ \right\|_{p^*}, \quad \forall a \in A, b \in B, \quad (3)$$

where $(\cdot)_+$ denotes coordinate-wise positive part. Unlike the absolute value, negative coordinates are set to zero, capturing only attributes where a dominates $\bar{\mathbf{u}}_b$. The right-hand side represents the largest margin loss that b could suffer if a perturbation concentrates all weight on attributes where a is stronger than $\bar{\mathbf{u}}_b$. Thus, if (3) holds, b cannot be tempted by a under any ℓ_p perturbation of radius r .

Case $k < m$. When perturbations may alter at most k coordinates of $s(b)$, the worst-case loss is smaller. Let $\mathbf{v} = \mathbf{u}(a) - \bar{\mathbf{u}}_b$ and let $v_{[1]} \geq v_{[2]} \geq \dots \geq v_{[m]}$ denote the coordinates of \mathbf{v}_+ sorted in nonincreasing order. The maximum reduction in margin under ℓ_p radius r and support budget k is then $r \cdot \left\| (v_{[1]}, \dots, v_{[k]}) \right\|_{p^*}$. Accordingly, the no-blocking constraint becomes

$$s(b) \cdot (\bar{\mathbf{u}}_b - \mathbf{u}(a)) \geq r \left\| (\mathbf{u}(a) - \bar{\mathbf{u}}_b)_+^{(k)} \right\|_{p^*}, \quad \forall a \in A, b \in B, \quad (4)$$

where $(\cdot)_+^{(k)}$ denotes the vector formed by the k largest positive coordinates of $\mathbf{u}(a) - \bar{\mathbf{u}}_b$ (zeros elsewhere). This is the k -aware support norm, which interpolates between the unrestricted case ($k = m$) and the degenerate case $k = 1$.

Feasibility and optimization. Checking feasibility of (3)–(4) reduces to convex optimization: linear programming for $p \in \{1, \infty\}$ and second-order cone programming for $p = 2$, both solvable in polynomial time. We therefore define the relaxation-based upper bound as

$$\text{UB} := \max \left\{ r \geq 0 : \exists X \in \text{conv}(\text{SM}) \text{ satisfying (3) or (4)} \right\}. \quad (5)$$

THEOREM 5.3 (RELAXATION-BASED UB). *UB is computable to additive accuracy ϵ_{UB} in polynomial time via bisection with LP/SOCP feasibility tests. Moreover, $\max_{\mu \in \text{SM}} r^*(\mu) \leq \text{UB}$. If the optimizer X^* is integral (i.e., corresponds to some stable matching μ^*), then $\text{UB} = r^*(\mu^*)$ and the maximizing stable matching μ^* is recovered explicitly, certifying exact optimality.*

PROOF. Feasibility for a fixed r is a convex program, an LP for $p \in \{1, \infty\}$ and an SOCP for $p = 2$, and thus solvable in polynomial time. Bisection on r yields an additive ϵ_{UB} -approximation in polynomial time. Since $\text{conv}(\text{SM})$ contains all integral stable matchings, $\max_{\mu \in \text{SM}} r^*(\mu) \leq \text{UB}$. If the optimal X^* is integral, it corresponds to a stable matching μ^* achieving equality. \square

Local frontier bounds. Beyond the global value UB , the same relaxation can be restricted to sublattices of SM defined by downsets D in the rotation poset. This yields local bounds $UB(D)$ that refine UB during the anytime search. Formally, $UB(D)$ is obtained by adding the rotation constraints of D to the global relaxation, so the feasible region only shrinks. These refinements are computable in polynomial time without recomputing from scratch, and they integrate directly into the anytime search (Section 5.4), where they ensure that the global LB/UB relation continues to hold.

5.4 Anytime search on the rotation poset

We now describe an anytime search procedure that combines the lower bound from μ_B and the relaxation-based upper bounds to progressively narrow the gap between them.

Algorithm. We maintain a priority queue of nodes D in the rotation poset, each corresponding to a downset of rotations and its associated stable matching $\mu_D = \text{elim}(\mu_A, D)$. The queue is in descending order of $UB(D)$ values, so that at each step we explore the matching whose sublattice still allows the largest possible robustness radius.

- (1) **Initialization.** Insert the root node $D = \emptyset$ (corresponding to μ_A) into the queue, and set $\mathbf{LB} \leftarrow r^*(\mu_B)$.
- (2) **Node extraction.** Remove from the queue the node D with the largest $UB(D)$.
- (3) **Pruning.** If $UB(D) \leq \mathbf{LB}$, discard D , since no matching reachable from it can improve the current best robustness.
- (4) **Exact evaluation.** Compute $r^*(\mu_D)$ exactly (via the LP/SOCP formulations from Section 4) and update \mathbf{LB} if improved.
- (5) **Expansion.** For each rotation ρ exposed at D , form $D' = D \cup \{\rho\}$, compute $UB(D')$, and insert it into the priority queue.

The search continues until either $\mathbf{LB} = \mathbf{UB}_{\text{frontier}}$ (certifying exact optimality) or until a predefined expansion budget is reached, after which the best matching found so far is returned.

THEOREM 5.4 (ANYTIME CORRECTNESS). *At all times,*

$$\mathbf{LB} \leq \max_{\mu} r^*(\mu) \leq \mathbf{UB}_{\text{frontier}} \leq \mathbf{UB}.$$

\mathbf{LB} increases monotonically while $\mathbf{UB}_{\text{frontier}}$ decreases monotonically. If some frontier node attains an integral relaxation, exact optimality is certified: $\mathbf{LB} = \mathbf{UB}_{\text{frontier}} = r^*(\mu^*)$.

PROOF. Each node D corresponds to a downset of rotations and a stable matching μ_D . Since $UB(D)$ is derived from a relaxation of the exact feasibility region, we have $r^*(\mu_D) \leq UB(D) \leq \mathbf{UB}$ for all D . \mathbf{LB} records the best exact radius found so far and can only increase, while $\mathbf{UB}_{\text{frontier}} = \max_D UB(D)$ over the active nodes can only decrease as nodes are expanded or pruned. Thus the stated inequalities and monotonicity follow. If an integral relaxation is reached, it corresponds to a stable matching μ^* with $\mathbf{LB} = \mathbf{UB}_{\text{frontier}} = r^*(\mu^*)$, certifying exact optimality. \square

Although this anytime search is not polynomial-time in the worst case, it always maintains certified bounds $[\mathbf{LB}, \mathbf{UB}_{\text{frontier}}]$ on the optimum. Both bounds evolve monotonically (\mathbf{LB} increases, $\mathbf{UB}_{\text{frontier}}$ decreases), so the procedure converges toward the true optimum and can be stopped at any point with provable bounds on the achieved robustness.

6 ROBUSTNESS–COST TRADEOFFS

In the previous section, we studied robustness alone, aiming to find a stable matching with the largest radius. A natural next step is to combine robustness with a *cost function* which can model welfare, fairness, or other priorities in applications. Fixing a robustness at a given value $\tau \geq 0$, we ask for the minimum cost achievable under this requirement. Formally, for separable costs $C(\mu) = \sum_{a \in A} c_{a, \mu(a)}$ and robustness requirement $\tau \geq 0$, the target is $C^*(\tau) = \min\{C(\mu) : r^*(\mu) \geq \tau\}$.

Direct optimization under $r^*(\mu)$ is computationally difficult, since the exact radius does not align with the rotation-poset structure. Instead, we develop polynomial-time proxy quantities that yield certified (i.e., provable) upper and lower bounds on $C^*(\tau)$.

6.1 Upper bound via the base radius

The base radius $r^{\text{base}}(\mu)$ (Section 5.1) is computable in closed form and always satisfies $r^{\text{base}}(\mu) \leq r^*(\mu)$. Requiring $r^{\text{base}}(\mu) \geq \tau$ is equivalent to insisting that each $b \in B$ is matched only to candidates a for which $\gamma(b; a)/U_{\rho^*}(b) \geq \tau$, where $U_{\rho^*}(b)$ is the dual-norm attribute gap defined in Section 5.1. In terms of the standard rotation poset, this condition simply prunes all stable matchings that violate the base-radius threshold, leaving a distributive sublattice over which optimization can be carried out.

THEOREM 6.1 (POLYNOMIAL-TIME BASE FRONTIER). *For any $\tau \geq 0$, the minimum-cost stable matching with $r^{\text{base}}(\mu) \geq \tau$ is computable in polynomial time. Hence*

$$C^*(\tau) \leq C^{\text{UB}}(\tau) := \min\{C(\mu) : r^{\text{base}}(\mu) \geq \tau\}.$$

PROOF. Fix $\tau \geq 0$. For any stable matching μ , the constraint $r^{\text{base}}(\mu) \geq \tau$ requires that every eliminated rotation ρ satisfy

$$\min_{(a,b) \in \rho} [\mathbf{s}(b) \cdot (\bar{\mathbf{u}}_b - \mathbf{u}(a)) - \tau \|\mathbf{u}(a) - \bar{\mathbf{u}}_b\|_{\rho^*}] \geq 0,$$

where $(\cdot)_+$ denotes the coordinate-wise positive part. Hence, the set of rotations that can be eliminated while maintaining stability at level τ forms a *downward-closed subset* of the rotation poset.

We assign each rotation a modified weight

$$\Delta_{\tau}(\rho) = \min_{(a,b) \in \rho} [\mathbf{s}(b) \cdot (\bar{\mathbf{u}}_b - \mathbf{u}(a)) - \tau \|\mathbf{u}(a) - \bar{\mathbf{u}}_b\|_{\rho^*}] - \sum_{(a,b) \in \rho} c_{a,b}.$$

Minimizing $C(\mu)$ subject to $r^{\text{base}}(\mu) \geq \tau$ is therefore equivalent to finding a *maximum-weight closure* in the rotation poset. By the classical reduction of maximum-weight closure to a single s - t min-cut [17], an optimal stable matching for each fixed τ can be computed in polynomial time.

Distinct threshold values of τ arise only when some base constraint becomes tight, i.e., $\mathbf{s}(b) \cdot (\bar{\mathbf{u}}_b - \mathbf{u}(a)) = \tau \|\mathbf{u}(a) - \bar{\mathbf{u}}_b\|_{\rho^*}$. Since there are $O(n^2)$ score margins (a, b) and $O(n)$ rotations in the poset, the number of breakpoints is $O(n^3)$. Sweeping over these values yields the exact (r^{base}, C) frontier. \square

6.2 Lower bound via relaxation

To obtain a lower bound, we again work with Rothblum’s stable-marriage polytope $\text{conv}(\mathcal{SM})$ introduced in Section 5.3, but now augment it with *vulnerability cuts*: for each cross pair (a, b') and (a', b) that could block within radius τ , we add the constraint $y_{ab'} + y_{a'b} \leq 1$. Here $y_{ab} \in [0, 1]$ denotes the standard assignment variable, and $\text{conv}(\mathcal{SM})$ admits a polynomial-size linear description [30]. Formally, we solve

$$\min \sum_{a,b} c_{ab} y_{ab}$$

s.t. $y \in \text{conv}(\mathcal{SM})$, y satisfies all vulnerability cuts for τ .

Every τ -robust stable matching μ induces a feasible point in this relaxation, so the LP optimum $C^{\text{LB}}(\tau)$ is a valid lower bound.

Proposition 6.1 (LP lower bound). *Every stable μ with $r^*(\mu) \geq \tau$ induces a feasible LP solution. Therefore $C^*(\tau) \geq C^{\text{LB}}(\tau)$.*

The vulnerable set changes only at $O(n^4)$ thresholds, since each arises from a cross pair (a, b') , (a', b) . Thus, at most $O(n^4)$ distinct

LPs need to be solved. Each such LP carries $O(m^k)$ additional constraints from the support budget; hence, $C^{LB}(\tau)$ can be computed by solving a total of $O(n^4 m^k)$ LP instances.

6.3 Certified tradeoff bounds

For each robustness target τ , our two constructions yield $C^{LB}(\tau) \leq C^*(\tau) \leq C^{UB}(\tau)$. Here $C^{UB}(\tau)$ comes from requiring the base radius $r^{base}(\mu) \geq \tau$, which produces an explicit stable matching and a constructive upper bound. The lower bound $C^{LB}(\tau)$ arises from the LP relaxation with vulnerability cuts, giving a valid numerical bound that coincides with a true stable matching whenever the LP optimum is integral. Varying τ reveals the robustness-cost tradeoff: stricter robustness raises the minimum cost, and the bounds capture this relationship.

7 GEOMETRY OF THE ROBUSTNESS REGION

Up to this point, we examined robustness under a fixed salience profile S . When both a matching μ and a profile S are given, we asked whether μ remains stable under a given perturbation radius and computed its exact robustness radius. When only S is given, we searched for the most robust matching and studied the tradeoff between robustness and cost. We now take the dual view: fix a matching μ and characterize, geometrically, its *robustness region* – the set of all salience profiles under which μ remains stable. A quantitative analysis of the volume of this region appears in [26].

While the robustness radius captures the most fragile perturbation direction (the minimal deviation that breaks stability), the robustness region reveals the full multidimensional structure of stability: its *geometry* indicates which perturbation directions are tolerated, and its *volume* (analyzed in [26]) quantifies the total range of salience profiles for which μ remains stable. This allows us to compare matchings not only by their most fragile direction, but also by the overall extent of their stability region in salience space. This contrasts with Rothblum’s *stable-marriage polytope* – a polytope in matching space whose vertices are stable matchings. Our robustness region instead lives in the salience space and forms the *profile polytope* of μ .

From Section 2, the robustness region \mathcal{P}_μ of a matching μ is defined by the set of salience profiles under which no blocking pair arises. Equivalently, stability reduces to the linear inequalities

$$\mathbf{s}(b) \cdot (\mathbf{u}(\mu^{-1}(b)) - \mathbf{u}(a)) \geq 0 \quad \forall b \in B, a \in \mathcal{H}_\mu(b), \quad (6)$$

where $\mathcal{H}_\mu(b)$ is the set of candidate blockers as introduced in Section 3.

Lemma 7.1 (Factorization). *The robustness region factorizes across B : $\mathcal{P}_\mu = \prod_{b \in B} \mathcal{P}_\mu(b)$, where each $\mathcal{P}_\mu(b)$ is the polytope*

$$\mathcal{P}_\mu(b) := \{\mathbf{s} \in \Lambda_{m-1} : \mathbf{s} \cdot (\mathbf{u}(\mu^{-1}(b)) - \mathbf{u}(a)) \geq 0, \forall a \in \mathcal{H}_\mu(b)\}.$$

PROOF. This lemma follows directly from the fact that each stability constraint involves only the salience vector $\mathbf{s}(b)$ of a single agent b . \square

Proposition 7.1 (Polyhedral structure). *Each factor $\mathcal{P}_\mu(b)$ is a convex polytope in Λ_{m-1} defined by at most $|\mathcal{H}_\mu(b)| \leq n - 1$ linear inequalities. Therefore \mathcal{P}_μ is a polytope in $(\Lambda_{m-1})^n$ defined by $O(n^2)$ inequalities in total.*

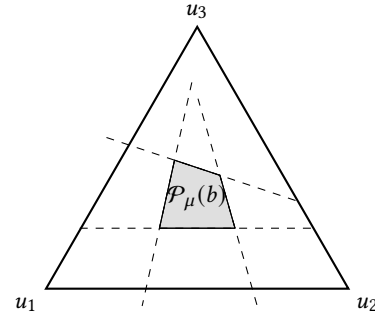


Figure 1: Example of $\mathcal{P}_\mu(b)$ for $m = 3$: the shaded polygon inside the simplex Λ_2 .

PROOF. Each inequality in (6) is linear in $\mathbf{s}(b)$ and together with the simplex constraints $\mathbf{s}(b) \geq 0, \sum_i s_i = 1$, they define a bounded convex polyhedron. Hence $\mathcal{P}_\mu(b)$ is a polytope, i.e., a bounded intersection of finitely many half-spaces inside the simplex. Summing over all b , the total number of inequalities is $O(n^2)$. \square

Illustration. Figure 1 shows an example of a $\mathcal{P}_\mu(b)$ for a single agent b when $m = 3$. The simplex Λ_2 is visualized as an equilateral triangle in the plane, where each interior point corresponds to a valid salience vector of b . In this example, $H_\mu(b) = 4$, so four potential blocking agents induce linear indifference constraints, drawn as dashed lines. The shaded 2-dimensional polytope (i.e., polygon) is their feasible intersection, consisting of all salience vectors for which b does not deviate from μ .

Testing whether $S \in \mathcal{P}_\mu$ is simply the standard stability check for μ under the salience profile S , a polynomial-time check over all $O(n^2)$ pairs. This observation motivates a geometric analysis of the full stability region. A complete analysis of the polyhedral structure and volume computation of \mathcal{P}_μ is provided in [26].

8 SUMMARY AND CONCLUSIONS

We introduced a new notion of robustness for stable matchings based on attributes and salience vectors. We presented polynomial-time algorithms for several problems in this context, including verifying whether a given matching remains stable for a specified radius, computing its maximum stability radius, and approximating the most robust stable matching through efficiently computable bounds. We also extended the framework to incorporate costs, deriving computable upper and lower bounds that relate robustness to the cost of stability. Finally, we showed that the robustness region of a given matching factorizes as a product of low-dimensional polytopes within the simplex. Overall, our results establish a unified geometric and algorithmic foundation for analyzing stability under structured preference perturbations and open several promising directions for further study.

Acknowledgments. We thank the reviewers of AAMAS for providing valuable comments which improved our paper. This research has been partially supported by the Israel Science Foundation under grant 2544/24.

REFERENCES

- [1] Atila Abdulkadiroğlu, Parag A Pathak, and Alvin E Roth. 2005. The new york city high school match. *American Economic Review* 95, 2 (2005), 364–367.
- [2] Atila Abdulkadiroğlu and Tayfun Sönmez. 2003. School choice: A mechanism design approach. *American economic review* 93, 3 (2003), 729–747.
- [3] Akhmad Alimudin, Yoshiteru Ishida, and Koutarou Suzuki. 2023. Maintaining Stability for a Matching Problem Under Dynamic Preference. *IEEE Access* 11 (2023), 24203–24215. <https://doi.org/10.1109/ACCESS.2023.3243245>
- [4] Farid Alizadeh and Donald Goldfarb. 2003. Second-order cone programming. *Mathematical programming* 95, 1 (2003), 3–51.
- [5] Haris Aziz. 2013. Stable marriage and roommate problems with individual-based stability. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. IFAAMAS, Richland, SC, USA, 287–294.
- [6] Nayantara Bhatnagar, Sam Greenberg, and Dana Randall. 2008. Sampling stable marriages: why spouse-swapping won't work. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2008, San Francisco, California, USA, January 20-22, 2008*, Shang-Hua Teng (Ed.). SIAM, Philadelphia, PA, 1223–1232. <http://dl.acm.org/citation.cfm?id=1347082.1347215>
- [7] Kimon Boehmer and Niclas Boehmer. 2024. Worst- and Average-Case Robustness of Stable Matchings: (Counting) Complexity and Experiments. In *ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024) (Frontiers in Artificial Intelligence and Applications, Vol. 392)*, Ulle Endriss, Francisco S. Melo, Kerstin Bach, Alberto José Bugarin Diz, Jose Maria Alonso-Moral, Senén Barro, and Fredrik Heintz (Eds.). IOS Press, Amsterdam, The Netherlands, 3469–3476. <https://doi.org/10.3233/FAIA240899>
- [8] Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. 2012. Salience theory of choice under risk. *The Quarterly journal of economics* 127, 3 (2012), 1243–1285.
- [9] Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. 2013. Salience and consumer choice. *Journal of Political Economy* 121, 5 (2013), 803–843.
- [10] Stephen Boyd and Lieven Vandenbergh. 2004. *Convex Optimization*. Cambridge University Press, New York, NY.
- [11] Sulian Le Bozec-Chiffolleau, Charles Prud'homme, and Gilles Simonin. 2024. Polynomial Time Presolve Algorithms for Rotation-Based Models Solving the Robust Stable Matching Problem. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*. ijcai.org, Online, 2860–2867. <https://www.ijcai.org/proceedings/2024/317>
- [12] Robert Bredereck, Jiehua Chen, Dusan Knop, Junjie Luo, and Rolf Niedermeier. 2020. Adapting Stable Matchings to Evolving Preferences. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, Washington, DC, 1830–1837. <https://doi.org/10.1609/AAAI.V34I02.5550>
- [13] Jiehua Chen, Piotr Skowron, and Manuel Sorge. 2021. Matchings under Preferences: Strength of Stability and Tradeoffs. *ACM Trans. Economics and Comput.* 9, 4 (2021), 20:1–20:55. <https://doi.org/10.1145/3485000>
- [14] Federico Echenique, Alejandro Robinson-Cortés, and Leeat Yariv. 2025. An experimental study of decentralized matching. *Quantitative Economics* 16, 2 (2025), 497–533.
- [15] David Gale and Lloyd S Shapley. 1962. College admissions and the stability of marriage. *The American mathematical monthly* 69, 1 (1962), 9–15.
- [16] Begum Genc, Mohamed Siala, Barry O'Sullivan, and Gilles Simonin. 2017. Finding Robust Solutions to Stable Marriage. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, Carles Sierra (Ed.). ijcai.org, Online, 631–637. <https://doi.org/10.24963/IJCAI.2017/88>
- [17] Dan Gusfield and Robert W. Irving. 1989. *The Stable marriage problem - structure and algorithms*. MIT Press, Cambridge, MA.
- [18] Anisse Ismaili, Kentaro Yahiro, Tomoaki Yamaguchi, and Makoto Yokoo. 2019. Student-Project-Resource Matching-Allocation Problems: Two-Sided Matching Meets Resource Allocation.. In *AAMAS. IFAAMAS, Richland, SC, USA, 2033–2035*.
- [19] Cristian Jimenez-Romero, Alper Yegenoglu, and Christian Blum. 2025. Multi-agent systems powered by large language models: applications in swarm intelligence. *Frontiers Artif. Intell.* 8 (2025), 1–24. <https://doi.org/10.3389/FRAI.2025.1593017>
- [20] Akiva Kleinerman, Ariel Rosenfeld, Francesco Ricci, and Sarit Kraus. 2021. Supporting users in finding successful matches in reciprocal recommender systems. *User Model. User Adapt. Interact.* 31, 3 (2021), 541–589. <https://doi.org/10.1007/S11257-020-09279-Z>
- [21] Yicheng Liu, Pingzhong Tang, and Wenyi Fang. 2014. Internally stable matchings and exchanges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 28. AAAI Press, Washington, DC, 1433–1439.
- [22] Tung Mai and Vijay V. Vazirani. 2018. Finding Stable Matchings That Are Robust to Errors in the Input. In *26th Annual European Symposium on Algorithms, ESA 2018, August 20-22, 2018, Helsinki, Finland (LIPIcs, Vol. 112)*, Yossi Azar, Hannah Bast, and Grzegorz Herman (Eds.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Wadern, Germany, 60:1–60:11. <https://doi.org/10.4230/LIPICS.ESA.2018.60>
- [23] David F. Manlove. 2013. *Algorithmics of Matching Under Preferences*. Series on Theoretical Computer Science, Vol. 2. WorldScientific, Singapore. <https://doi.org/10.1142/8591>
- [24] Calarina Muslimani, Bram Grooten, Deepak Ranganatha Sastry Mamillapalli, Mykola Pechenizkiy, Decebal Constantin Mocanu, and Matthew E. Taylor. 2025. Boosting Robustness in Preference-Based Reinforcement Learning with Dynamic Sparsity. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2025, Detroit, MI, USA, May 19-23, 2025*, Sanmay Das, Ann Nowé, and Yevgeniy Vorobeychik (Eds.). International Foundation for Autonomous Agents and Multiagent Systems / ACM, Online, 2687–2689. <https://doi.org/10.5555/3709347.3743979>
- [25] Amit Ronen, Jonah Evan Hess, Yael Belfer, Simon Muraas, and Alon Eden. 2025. Stable Marriage: Loyalty vs. Competition. *CoRR abs/2501.18442* (2025), 1–28. <https://doi.org/10.48550/ARXIV.2501.18442> arXiv:2501.18442
- [26] Amit Ronen, S. S. Ravi, and Sarit Kraus. 2026. Robustness of Stable Matchings When Attributes and Salience Determine Preferences. *ArXiv:2602.04115 [cs.GT]*. <https://doi.org/10.48550/arXiv.2602.04115>
- [27] Alvin E. Roth. 1982. The Economics of Matching: Stability and Incentives. *Math. Oper. Res.* 7, 4 (1982), 617–628. <https://doi.org/10.1287/MOOR.7.4.617>
- [28] Alvin E Roth and Elliott Peranson. 1999. The redesign of the matching market for American physicians: Some engineering aspects of economic design. *American economic review* 89, 4 (1999), 748–780.
- [29] Alvin E. Roth, Tayfun Sönmez, and M. Utku Ünver. 2005. Pairwise kidney exchange. *J. Econ. Theory* 125, 2 (2005), 151–188. <https://doi.org/10.1016/J.JET.2005.04.004>
- [30] Uriel G. Rothblum. 1992. Characterization of stable matchings as extreme points of a polytope. *Math. Program.* 54 (1992), 57–67. <https://doi.org/10.1007/BF01586041>
- [31] Jay Sethuraman, Chung-Piaw Teo, and Liwen Qian. 2006. Many-to-One Stable Matching: Geometry and Fairness. *Math. Oper. Res.* 31, 3 (2006), 581–596. <https://doi.org/10.1287/MOOR.1060.0207>
- [32] Yohai Trabelsi, Abhijin Adiga, Yonatan Aumann, Sarit Kraus, and S. S. Ravi. 2025. Facilitating Matches on Allocation Platforms. In *Proceedings ECAI 2025*. IOS press, Online, 3862–3870.
- [33] Yohai Trabelsi, Abhijin Adiga, Sarit Kraus, S. S. Ravi, and Daniel J. Rosenkrantz. 2023. Resource Sharing through Multi-Round Matchings. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, Brian Williams, Yiling Chen, and Jennifer Neville (Eds.). AAAI Press, Online, 11681–11690. <https://doi.org/10.1609/AAAI.V37I10.26380>
- [34] Amos Tversky and Itamar Simonson. 1993. Context-dependent preferences. *Management science* 39, 10 (1993), 1179–1189.
- [35] Osnat Ackerman Viden, Yohai Trabelsi, Pan Xu, Karthik Abinav Sankaraman, Oleg Maksimov, and Sarit Kraus. 2023. Allocation Problem in Remote Teleoperation: Online Matching with Offline Reusable Resources and Delayed Assignments. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023 - 2 June 2023*, Noa Agmon, Bo An, Alessandro Ricci, and William Yeoh (Eds.). ACM, New York, NY, 513–521. <https://doi.org/10.5555/3545946.3598678>